A Cluster Analysis on the Structural Diversity of Protein Crystals, Exemplified by Human Immunodeficiency Virus Type 1 Protease

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

Graduate School of Pharmaceutical Sciences, Chiba University; 1–8–1 Inohana, Chuo-ku, Chiba 260–8675, Japan.
Received January 28, 2014; accepted March 30, 2014

Information on many protein crystal structures has recently become available due to developments in crystallographic techniques. Even for a single kind of protein, several and sometimes many crystal structures are available. Human immunodeficiency virus type 1 (HIV-1) protease is one of the most extensively studied viral proteins, and about six hundred crystal structures have been determined. In this work, we examined the structural diversity of HIV-1 protease, classifying crystal structures into several groups from the viewpoint of similarity in atom geometry. Using 499 crystal structures downloaded from the Protein Data Bank (PDB), cluster analysis was applied to the whole body of HIV-1 protease and also to a limited number of residues at the binding pocket. As a consequence of clustering with regard to the whole body, 499 crystal structures were separated into 6 groups. It was found that a major factor for this separation is the space group of the crystals and that the space group strongly depends on the agents used in the protein crystallization. Amino acid mutation is a minor factor for separation in clustering. In cluster analysis for a limited number of residues at the binding pocket, crystal structures were not distinctly separated, and no clear factor linked to the separation was clarified. The results suggest that amino acid mutations have little effect on the coordinates of the main-chain atoms of HIV-1 protease. Hence, the changes in drug efficacy or substrate fitness caused by mutations are mainly due to the physicochemical features of amino acid side chains.

Key words clustering analysis; crystal structure; amino acid mutation; drug resistance; space group

Due to the progress in techniques for protein crystallization and in software for model building, crystal structures of many proteins have been elucidated. The reliability of solved protein structures has also increased.[1] The development of a high-energy X-ray source has also been important for advancing crystallographic studies.[2] Due to the progress in crystallographic technology, the availability of crystal structures in good quality has enabled us to examine structural differences in proteins in detail. Even for a single kind of protein, an amino acid mutation will cause the difference in its activity. Mutagenesis is one of the best approaches for clarifying the relationship between the protein activity and the mutated amino acid residue. Since protein activity has a close relation with its structure,[3,4] it is of great interest in molecular biology that the influence of amino acid mutation induces the change in protein structure.

Apart from artificial mutagenesis, amino acid mutation ceaselessly occurs in the process of evolution. There is a common consensus that the accumulation of amino acid mutations generates a genetic diversity and that the diversity is observed as a difference in protein function in phenotype.[5,6] In general, the mutation rate of viral proteins is much higher than that of protein in eukaryotes. Variants of a virus sometimes cause acute respiratory infection in humans and occasionally progress to a severe pandemic as seen in the recent emergence of highly pathogenic influenza viruses.[7] For the purpose of obtaining information that would be useful for the development of antiviral agents, crystallographic study of viral proteins has been extensively performed. Due to the broad diversity of mutations of viral proteins and the discovery of novel classes of inhibitor compounds, many crystal structures have been obtained on a target protein of a virus for the wild type and its variants in the *apo*-form as well as in the complex form with its inhibitor.[8] The acquirement of crystal structures has enabled detailed comparisons of structural changes in proteins.[9]

Human immunodeficiency virus type 1 (HIV-1) protease is one of the most extensively studied viral proteins because of its importance as a target of antiviral drugs.[4,5] Since inhibition of the enzymatic activity hinders maturation of the viral precursor and leads to incomplete replication of the virus, inhibitors of HIV-1 protease are effective in chemotherapy for HIV-1 infectious diseases. HIV-1 protease is coded in the *pol* region of the viral genome and is expressed in a form of the *pol* precursor peptide. With the formation of a homo-dimer in which each monomer consists of 99 amino residues, the mature HIV-1 protease is produced through processing of the precursor peptide in the viral particle. Polymorphisms are frequently observed,[10,11] and about 500000 sequences have been registered in the Los Alamos HIV database.[12] Polymorphism and/or amino acid mutation have a close relation with the emergence of drug resistance. For example, D30N mutation is one of the well-known primary mutations for drug resistance.[13,14] L90M mutation has been frequently detected after introduction of the protease inhibitor Nelfinavir in chemotherapy.[15] The crystal structure of HIV-1 protease in the dimer form was first deposited in the protein data bank (PDB) in 1989.[16] This first crystal structure, PDB code: 4HVP, was obtained as a complex with a peptide inhibitor. Due to the accumulation of crystal structures of HIV-1 protease, about 600 structures are currently available at the PDB site.[17] This abundance of plenty of information on the structure of HIV-1 protease has been utilized for understanding the roles of respective amino acid residues in enzymatic activity and the fitness to the substrate, and then for the design of novel antiviral agents.[18]

Drug resistance is one of the most serious problems in chemotherapy for HIV-1 infectious diseases. The virus bears amino acid mutations to diminish the binding affinity to an

The authors declare no conflict of interest.

* To whom correspondence should be addressed. e-mail: hoshino@chiba-u.jp
© 2014 The Pharmaceutical Society of Japan

inhibitor, leading to a reduction in the efficacy of the inhibitor.[19,20] While several compounds have been approved as HIV-1 protease inhibitors by the U.S. Food and Drug Administration (FDA), the emergence of resistant variants has been reported for every inhibitor. A question of great interest is how a variety of structural changes appear in protease. That is, the following two cases will be assumed. One case is that the change in protease structure is limited so that the structures can be classified into several groups, and the other case is that HIV-1 protease can change its structure into a wide variety of forms so that the structures are difficult to classify. If protease structures can be classified into several groups as in the former case, a small number of inhibitors will be sufficient for chemotherapy. Given that sufficient kinds of inhibitors susceptible to the respective structural groups are available, we can devise an effective protocol depending on the virus infecting an individual patient. Therefore, it is important to examine the classification of the protease in terms of structure, namely, to investigate the diversity of HIV-1 protease.

In this study, we performed cluster analysis of the crystal structures of HIV-1 protease registered in PDB. The aim of this work was to determine the structural diversity of the protease. If the crystal structures can be classified into a few groups, variation of the protease structure would not be large and conformational change of HIV-1 protease would be limited. If the crystal structures are scattered and difficult to assign them to several groups, this protein would have many conformational variations which could easily diminish the affinity of any approved inhibitors. The major factor for group separation in cluster analysis was discussed in terms of amino acid mutations and experimental differences in protein crystallization.

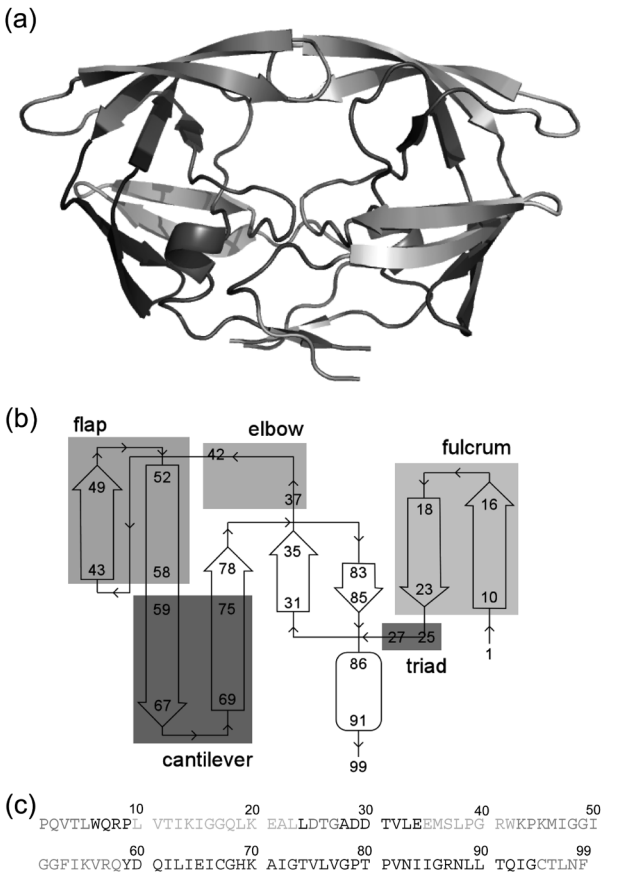

Fig. 1. (a) Structure of the HIV-1 Protease

Several functional regions are indicated by colors. (b) Schematic representation of the secondary structure of HIV-1 protease. Colors correspond to the regions shown in (a). (c) Amino sequence of HIV-1 protease for HXB2 strain. (Color images were converted into grey scale.)

Experimental

Structure of HIV-1 Protease A significant feature of HIV-1 protease is homo-dimer formation, in which each monomer consists of 99 amino acid residues (Fig. 1a). Each monomer contains one α-helix, three β hairpins, and seven β strands that compose two antiparallel β-sheets. Functional regions have various names including flap, cantilever, elbow, catalytic triad, and fulcrum as shown in the topological diagram of Fig. 1b. The flap region, located at the residues from codons 43 to 58, composes a β strand–β hairpin–β strand configuration and has a role in holding a substrate precursor peptide inside the binding pocket. Asp25 is located at the beginning of the catalytic triad and is the most important residue responsible for enzymatic activity.[21] Two Asp residues from the respective monomers are positioned close to each other at the center of the binding pocket and functionalize the hydrolysis of the amide bond of the precursor peptide. The cantilever region from codons 59 to 75 composes a long antiparallel β-sheet and has a mechanical role for opening and closing of the flap region as well as for stabilizing the protein shape.[22] The region at codons 83–92 has been suggested to be involved in the initial stage of folding of the monomer,[23] and the regions at the N- and C-termini play a role in maintaining the stability of the dimerization interface.[22]

HXB2 is one of the most major wild-type strains of HIV-1 and it has been employed in many experimental studies.[24] Hence, the sequence of this strain was used as a principal reference of HIV-1 protease in this work. The amino sequence of HXB2 is shown in Fig. 1c for reference, indicating the functional regions by colors.

Cluster Analysis A search for crystal structures of HIV-1 protease was made by using the query "human immunodeficiency virus type 1 protease" at the PDB site. The result of the query search contained 79 structures that are not relevant to the HIV-1 protease. Consequently, 618 structures were available in total,[17] and all of the structures were downloaded. The downloaded data were filtered in terms of the following five points. (1) Atom coordinates are provided in the dimer form. Namely, crystal structures carrying information only on the monomer were omitted even if the dimer coordinates could be generated by symmetry operation, because they are a minor form and have a risk for making an unfavorable specific group in cluster analysis. Forty three crystal structures of HIV-1 protease were excluded by this criterion. (2) There are no missing residues for the protease part. Fifty three crystal structures of HIV-1 protease have the missing residues. (3) No chemical modification was applied for the residues of the protease part. Fortunately, every crystal structure met this criterion. (4) The α-helix and β-sheets are neatly visible in PyMOL.[25] Ten structures of HIV-1 protease were excluded by the judgment. (5) The protease monomer consists of more than 99 residues. There were 11 structures fell in this condition. After applying the filter, 499 crystal structures were selected as a data set for cluster analysis.

The coordinates of substrate, inhibitor, ions, water and all

other hetero-atoms were removed from all of the 499 crystal structures. Hence, only the protease part was focused on in this work. If a crystal structure contained several dimers, the first one, usually labeled chain A and chain B, was selected. If atom coordinates were not uniquely determined, the coordinate in the highest occupancy, usually labeled A, was selected. The main-chain atoms were extracted from every structure and a set of the extracted coordinates was used as the base material for cluster analysis. First, the average structure of the 499 structures was obtained. Second, all of the structures were fitted to the average structure to unify the molecular orientations. Third, the root mean square deviations (RMSDs) with respect to all of the main-chain atoms were calculated among all structures. Fourth, based on the RMSDs values, the structures were classified into several groups by performing cluster analysis with the nearest neighboring method using the "hclust" function of R software.[26,27] Finally, amino acid mutation, space group of the protein crystal, category of substrate or inhibitor, and crystallization condition were surveyed for every structure of the respective clusters. The nearest neighboring method produces a dendrogram of crystal structures in a hierarchical manner. At first, all the structures are assumed as singleton clusters. A pair of clusters that give the nearest distance is searched among the all the combinations of the clusters. The selected two clusters are merged, which results in the decrease of the number of clusters by one. By repeating the search of the nearest two clusters and the merge of them, all the structures are connected as a spanning tree.

Cluster analysis was also carried out using only the residues at the binding pocket. The same procedure was performed with extraction of the coordinates only for the main-chain atoms of the residues: 23–32, 47–50, 81–84, 23′–32′, 47′–50′, and 81′–84′. The average structures and their superimposition were visualized by PyMOL.[25]

Results

Clustering with the Whole-Body Structure of the Protease A dendrogram deduced from cluster analysis is shown in Fig. 2. Judging from the shape of the tree and the branches, we separated the protein crystal structures into 6 clusters. The clusters are labeled from group 1 to group 5 which are from major to minor in number of members. Group 6 consists of structures that were not assigned to groups 1–5. A list of cluster members of the respective groups is provided in Supplementary Information (List S1). The numbers of member structures for groups 1–6 are 173, 71, 64, 44, 17, and 130, respectively. One third of all the structures belong to the largest cluster, group 1. The second largest cluster, group 2, is about two fifths the size of group 1. The smallest cluster, group 5, contains less than one thirtieth of all structures.

The average structure was obtained from each cluster group, and the respective average structures were compared and superimposed on that of group 1 (Fig. 3). A comparison of groups 1 and 3 indicates a structural difference in the flap region and in the cantilever (Fig. 3b). The deviations are distributed symmetrically to both monomers. In contrast, a structural displacement is asymmetric in the average structure of group 4 (Fig. 3c). Large deviations are observed in the elbow region of one monomer and in the flap region of the opposite monomer. The amplitude of structural deviation between groups 1 and 2 (Fig. 3a) is obviously larger than that between groups 1 and 3 and that between groups 1 and 4. In the comparison of groups 1 and 2, differences are observed in one monomer and the deviations in the flap regions are remarkable. A marked structural deviation is also seen in the side area of the binding pocket as well as in the cantilever region. A comparison of groups 1 and 5 demonstrates a large displacement of the flap region (Fig. 3d). The flap region is just open and the whole configuration of the protease is quite different from that of other groups. Furthermore, considerably large deviations are observed in the contact area of two monomers. In the comparison of groups 1 and 6, the amplitude of the deviation is moderate, while a notable deviation is seen in the β hairpin area of the flap regions (Fig. 3e).

In order to find a critical factor to characterize the respective clusters, the properties of the cluster members were examined in terms of amino acid mutation, substrate or inhibitors, crystallization condition, pH condition and resolution. That is, we tried to clarify a common feature among the members of each group. It was found that the primary factor to distinguish the crystal structures of HIV-1 protease was the space group of protein crystals as shown in Table 1. Space group represents the positional and directional arrangement of molecules in a unit cell of protein crystal. A unit cell is comprised of one or several asymmetric units, arranged in pattern characterized by symmetry. The space group is determined by a combination of rotational and translational operations that give rise to a unique pattern of symmetry elements. The cluster members of each group were classified with respect to the space group. The most major space group for the crystal of HIV-1 protease

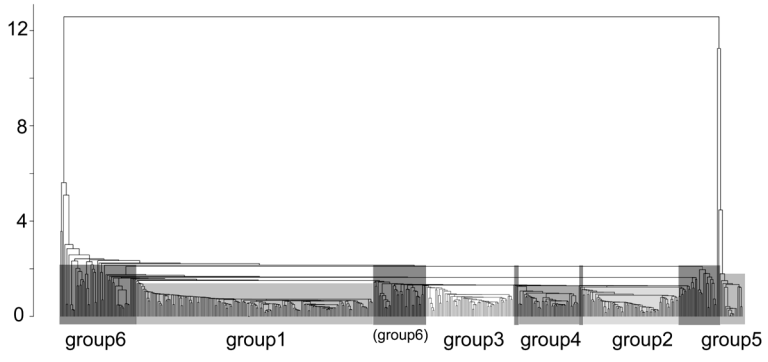

Fig. 2. Dendrogram for Cluster Analysis of the Whole-Body Structure of HIV-1 Protease

Totally, 499 structures were grouped into six clusters. The height of the tree reflects the root mean square distance of the main-chain atoms among the crystal structures.

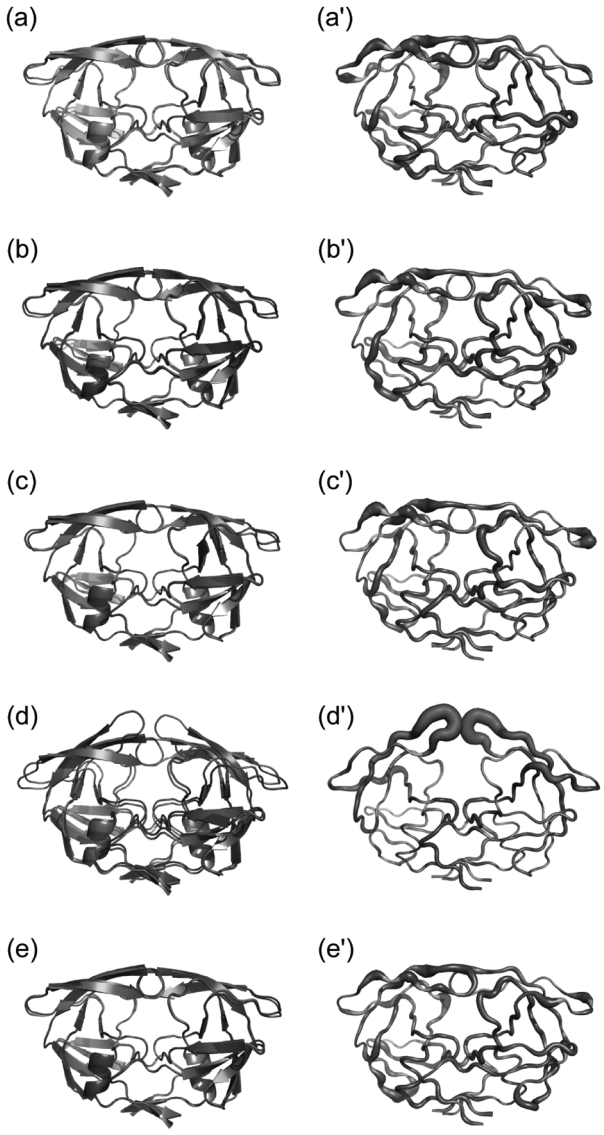

Fig. 3. Comparison of the Average Structures of Clusters Grouped with Respect to Whole-Body Structure

Superimpositions of the average structure of (a) group 2, (b) group 3, (c) group 4, (d) group 5, or (e) group 6 on that of group 1. The structure of group 1 is shown in grey. Deviation of the main-chain atoms in the average structure of (a′) group 2, (b′) group 3, (c′) group 4, (d′) group 5 or (e′) group 6, measured from that of group 1. The deviation increases as the color changes from cyan to red. (Color images were converted into grey scale.)

is P21 21 2, and the second major space group is P21 21 21. The members of these two space groups reach 72.5% of all the crystal structures. It is notable that most of the crystal structures bearing the P21 21 2 space group belong to group 1. Most of the crystal structures bearing the P61 space group belong to group 3. Most of the crystal structures bearing the P41 space group belong to group 5. Furthermore, all of the cluster members of group 4 and most of the members of group 2 bear the P21 21 21 space group. This fact clearly indicates that the whole-body structure of HIV-1 protease revealed by X-ray crystallographic analysis is dominantly determined by the symmetry under which the proteins have been piled up in the crystal growth. The crystal structures bearing the P21 21 21 space group are mainly restricted to groups 2 and 4. A factor to differentiate groups 2 and 4 is not clear at this stage. Since group 6 consists of structures that were not assigned to other groups, it is natural that the cluster members of group 6 are broadly distributed over the space groups.

Amino acid mutations in the cluster members of the respective groups were surveyed as shown in Table 2. It is interesting to focus on the primary resistant mutations for protease inhibitors. The D30N mutation, which is known as a resistant mutation for Nelfinavir,[15] is seen in every cluster. Because a limited number of crystal structures contain the G48V and I50V mutations, it cannot be concluded that these mutations have an influence on the protein whole-body structure. The V82, A, F, T and I84V mutations, which are also known as primary resistant mutations occurring at residues located at the binding pocket,[28] are distributed across all cluster groups. Judging from the results for the D30N, V82A, F, T and I84V mutations, a resistant mutation at the binding pocket has little effect on the whole-body structure of HIV-1 protease.

The L90M mutation is known as a resistant mutation for Saquinavir and Nelfinavir[14] and is located outside the binding pocket, and this mutation also scatters over all groups in Table 2. The ratio of this mutation in group 5 is, however, high. The members of group 5 contain not only L90M but also other mutations including I84V, I54V, L and their average structure has a semi-open configuration that is quite different from that in other groups. It should be noted that all of the crystal structures in group 5 were provided by two research groups.[29–31] The L63P mutation that is an amino substitution sometimes appeared in drug-resistant viruses[32,33] is seen in every cluster except for group 3.

The Q7K mutation that is seen in one third of the crystal structures scatters over all cluster groups, which is sometimes

Table 1. Number of Crystal Structures Classified by the Space Group for Clusters Grouped with Respect to Whole-Body Structure

Space group	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
P21 21 2	172 (0.99)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	19 (0.15)	191 (0.38)
P21 21 21	1 (0.01)	69 (0.97)	0 (0.00)	44 (1.00)	1 (0.06)	56 (0.43)	171 (0.34)
P61	0 (0.00)	0 (0.00)	59 (0.92)	0 (0.00)	0 (0.00)	23 (0.18)	82 (0.16)
P41	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	16 (0.94)	2 (0.02)	18 (0.04)
P1 21 1	0 (0.00)	0 (0.00)	5 (0.08)	0 (0.00)	0 (0.00)	13 (0.10)	18 (0.04)
Others[a)]	0 (0.00)	2 (0.03)	0 (0.00)	0 (0.00)	0 (0.00)	17 (0.13)	19 (0.04)
Total	173	71	64	44	17	130	499

a) Others is the sum of those for C1 2 1, I 41 2 2, I2 2 2, P1, P1 1 21, P43, P41 21 2 and P43 21 2 space groups. *b*) The value in parenthesis represents the ratio relative to the total number of crystal structures in each group. For example, group1 contains 173 structures and 99% of the 173 structures bear P21 21 2 space group.

Table 2. Number of Crystal Structures Bearing Amino Mutations for Clusters Grouped with Respect to Whole-Body Structure

Mutation	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Q7K	95 (0.55)	71 (1.00)	5 (0.08)	43 (0.98)	1 (0.06)	78 (0.60)	293 (0.59)
D25N	2 (0.01)	7 (0.10)	0 (0.00)	7 (0.16)	16 (0.94)	15 (0.12)	47 (0.09)
D30N	4 (0.02)	1 (0.01)	4 (0.06)	2 (0.05)	1 (0.06)	6 (0.05)	18 (0.04)
S37N, E	9 (0.05)	69 (0.97)	8 (0.13)	23 (0.52)	17 (1.00)	72 (0.55)	198 (0.40)
G48V	1 (0.01)	0 (0.00)	1 (0.02)	0 (0.00)	0 (0.00)	5 (0.04)	7 (0.01)
I50V	8 (0.05)	0 (0.00)	0 (0.00)	1 (0.02)	0 (0.00)	5 (0.04)	14 (0.03)
I54V, L	4 (0.02)	0 (0.00)	1 (0.02)	0 (0.00)	17 (1.00)	26 (0.20)	48 (0.10)
L63P	9 (0.05)	69 (0.97)	0 (0.00)	17 (0.39)	17 (1.00)	47 (0.36)	159 (0.32)
C67A	64 (0.36)	1 (0.01)	5 (0.08)	27 (0.61)	1 (0.06)	26 (0.20)	124 (0.25)
V82A, F, T	18 (0.10)	4 (0.06)	11 (0.17)	12 (0.27)	16 (0.94)	31 (0.24)	92 (0.18)
I84V	22 (0.13)	3 (0.04)	9 (0.14)	7 (0.16)	17 (1.00)	18 (0.14)	76 (0.15)
L90M	4 (0.02)	0 (0.00)	2 (0.03)	5 (0.11)	17 (1.00)	20 (0.15)	48 (0.10)
C95A	64 (0.37)	1 (0.01)	10 (0.16)	27 (0.61)	1 (0.06)	28 (0.22)	131 (0.26)

a) The value in parenthesis represents the ratio relative to the total number of crystal structures in each group.

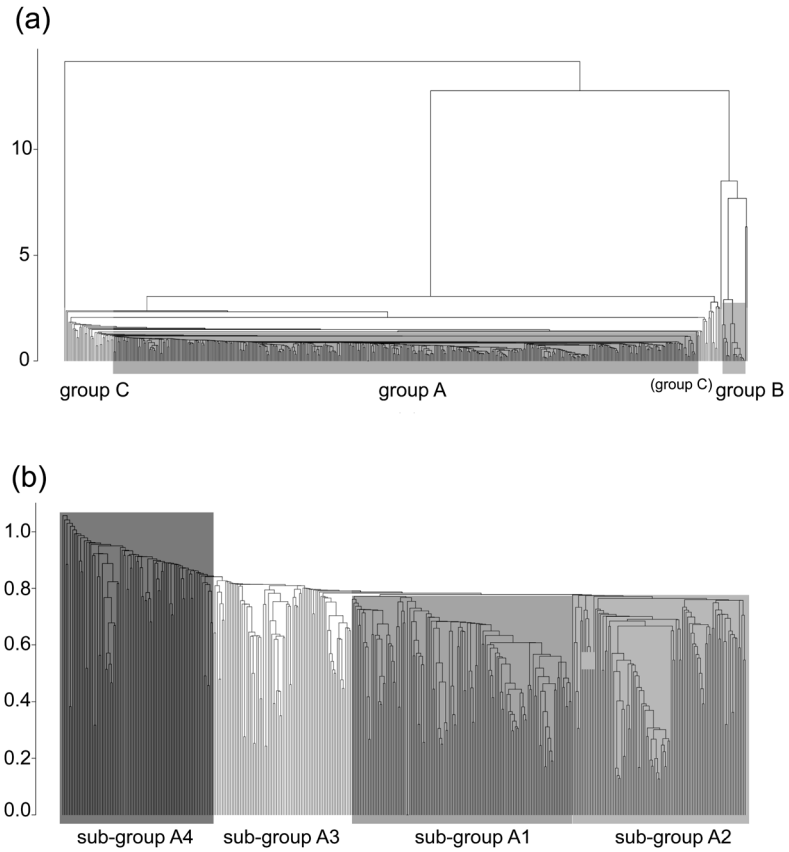

Fig. 4. Dendrogram for Cluster Analysis of Residues at the Binding Pocket of HIV-1 Protease

(a) Totally, 499 structures were grouped into three clusters. (b) Separation of 416 structures of group (A) in (a) into 4 sub-groups.

used in experiments to increase the expression efficiency of HIV-1 protease with *Escherichia coli*, by suppressing autolysis.[34] This finding suggests that the Q7K mutation has little effect on the whole-body structure of HIV-1 protease. Ser37 is sometimes substituted into Asn or Glu and these amino mutations are also distributed over all clusters. The C67A and C95A double mutations are sometimes introduced to prevent degradation of HIV-1 protease by avoiding oxidation of the thiol group of Cys.[35] The redox states of these cysteine thiol groups were reported to be involved in the regulation of enzymatic activity for Gag processing.[36,37] In the crystal structures, the C67A mutation always appeared together with the C95A mutation, and about one fourth of all structures carry the C67A and C95A double mutations. Over 70% of the crystal structures bearing the $P2_12_12_1$ space group in group 4 have these double mutations. In contrast, crystal structures bearing the $P2_12_12_1$ space group in group 2 scarcely include these mutations. Accordingly, the double mutation in the crystal with the $P2_12_12_1$ space group is one of the critical factors to distinguish protease structures. Since the conversion of Asp25 which is the most important catalytic residue generates the protease without catalytic activity, the D25N mutation is sometimes introduced to investigate the binding mode of the substrate. The D25N mutation is distributed all over the

clusters, and thus this mutation also has little influence on the whole-body structure of HIV-1 protease.

Clustering with Residues at the Binding Pocket A dendrogram obtained from cluster analysis using a limited number of residues located at the binding pocket is shown in Fig. 4a. Judging from the shape of the tree, we separated 499 structures into 3 clusters (List S2). The clusters are labeled group A, group B, and group C, and the numbers of group members are 416, 17, and 66, respectively. Namely, 83% of the structures belong to one major group. The two minor groups contain about 3 and 13% of the structures in each. It should be noted that the members of group B are completely identical to those of group 5 of Fig. 2 derived from cluster analysis for the whole-body structure.

The average structure was obtained and a comparison was made with the average structure of group A (Fig. 5). A comparison of groups A and B shows deviations in many areas around the binding pocket. Hence, remarkable structural deviations are observed not only in the flap regions but also in the active site. In contrast, a comparison of groups A and C indicates that the deviations are limited to the β hairpin area of the flap region of both monomers.

The members of the respective clusters were classified in terms of space group (Table S1) and amino acid mutation (Table S2). Since group B is identical to group 5 in cluster analysis with the whole-body structure, the same result was obtained for group B. Namely, most of the members of group B have the P41 space group. The number of crystal structures carrying specific amino acid mutations in Table S2 indicates that group A contains a variety of amino acid mutations. The amino acid sequence of group B is considerably different from those in other groups. The amino acid sequence of group C has a difference from other groups, too. A relatively large number of group C carry the S37N,E and/or L63P mutations. Two fifths of all crystal structures contain S37N,E mutation, while two thirds of the members of group C carry this mutation. Another noticeable difference is the L63P mutation.[32,33]

Clustering of Group A into Sub-groups by Analysis with Residues at the Binding Pocket Since group A derived from cluster analysis with residues at the binding pocket constitutes a large cluster, we further applied cluster analysis for group A and obtained sub-groups of group A with respect to residues at the binding pocket (Fig. 4b). Judging from the shape of the dendrogram, we managed to separate 416 structures into 4 clusters (List S3). The clusters are labeled from sub-group A1 to sub-group A4, and the numbers of group members are 134, 106, 84, and 92, respectively. The longest cluster, sub-group A1, contains one third of the structures. Three minor groups contain about 20–25% of the structures in each.

The properties of the cluster members were examined in terms of space group (Table 3) and amino acid mutation (Table 4). As shown in Table 3, 97% of the sub-structure A1 have the P21212 space group. In sub-group A1, there is no

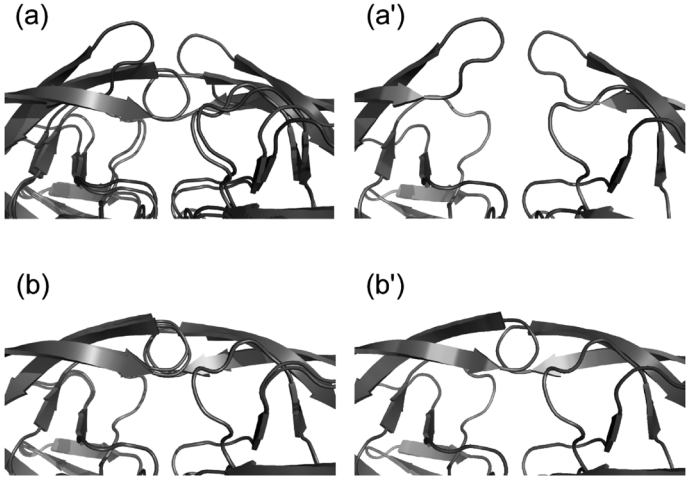

Fig. 5. Comparison of the Average Structures of the Clusters Grouped with Respect to the Residues at the Binding Pocket

(a, b) Superimpositions of the average structure of group (B) and group (C) on that of the group (A). (a′, b′) Deviation of the main-chain atom in the average structure of group (B) or group (C), measured from those of group (A). The deviation increases with the color change from cyan to red. (Color images were converted into grey scale.)

Table 3. Number of Crystal Structures Classified by Space Group in Regard to the Sub-groups of Group A in Cluster Analysis with Respect to Residues at the Binding Pocket

Space group	Sub-group A1	Sub-group A2	Sub-group A3	Sub-group A4	Total
P21212	130 (0.97)	2 (0.02)	18 (0.21)	30 (0.33)	180 (0.43)
P212121	2 (0.01)	60 (0.57)	39 (0.46)	32 (0.35)	133 (0.32)
P61	0 (0.00)	39 (0.37)	20 (0.24)	18 (0.20)	77 (0.19)
P41	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.01)	1 (0.00)
P1211	1 (0.01)	5 (0.05)	2 (0.02)	7 (0.08)	15 (0.04)
Others[a)]	1 (0.01)	0 (0.00)	5 (0.06)	4 (0.04)	10 (0.02)
Total	134	106	84	92	416

a) Others is the sum of those for C121, I4122, I222, P1, P1121, P43, P41212 and P43212 space groups. *b*) The value in parenthesis represents the ratio relative to the total number of crystal structures in each sub-group.

Table 4. Number of Crystal Structures Bearing Amino Mutations in Regard to the Sub-groups of Group A in Cluster Analysis with Respect to Residues at the Binding Pocket

Mutation	Sub-group A1	Sub-group A2	Sub-group A3	Sub-group A4	Total
Q7K	68 (0.51)	69 (0.65)	59 (0.70)	51 (0.55)	247 (0.59)
D25N	2 (0.01)	5 (0.05)	9 (0.11)	7 (0.08)	23 (0.06)
D30N	3 (0.02)	5 (0.05)	4 (0.05)	0 (0.00)	12 (0.03)
S37N, E	10 (0.07)	63 (0.59)	35 (0.42)	28 (0.30)	136 (0.33)
G48V	0 (0.00)	0 (0.00)	0 (0.00)	5 (0.05)	5 (0.01)
I50V	4 (0.03)	2 (0.02)	6 (0.07)	2 (0.02)	14 (0.03)
I54V, L	2 (0.01)	3 (0.03)	1 (0.01)	13 (0.14)	19 (0.05)
L63P	6 (0.04)	62 (0.58)	24 (0.29)	21 (0.23)	113 (0.27)
C67A	48 (0.36)	8 (0.08)	31 (0.37)	23 (0.25)	110 (0.26)
V82A, F, T	12 (0.09)	15 (0.14)	14 (0.17)	21 (0.23)	62 (0.15)
I84V	14 (0.10)	14 (0.13)	8 (0.10)	12 (0.13)	48 (0.12)
L90M	3 (0.02)	7 (0.07)	5 (0.06)	6 (0.07)	21 (0.05)
C95A	48 (0.36)	12 (0.11)	33 (0.39)	23 (0.25)	116 (0.28)

a) The value in parenthesis represents the ratio relative to the total number of crystal structures in each sub-group.

Table 5. Number of Structures Classified by the Primary Agent for Protein Crystallization for Clusters Grouped with Respect to Whole-Body Structure

Agent	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Ammonium sulfate	14	62	20	41	0	48	185
Polyethyleneglycol[a)]	2	0	4	0	0	5	11
Sodium chloride	114	1	6	1	17	29	168
Others[b)]	15	0	1	0	0	14	30
Data not shown	28	8	33	2	0	34	105

a) Polyethyleneglycol 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.

crystal structure bearing P61 space group. This finding indicates that the space group of a protein crystal has an influence on the structural characteristic of the residues at the binding pocket to some extent.

The number of crystal structures carrying specific amino acid mutations in Table 4 suggests that every amino acid mutation cannot be assigned to any single sub-group exclusively. Structures carrying drug-resistance mutations such as I84V and L90M scatter across all of the sub-groups. The D30N drug-resistance mutation is seen in every sub-group except for sub-group A4. The D25N mutation, which is occasionally introduced to generate a protease without catalytic activity, is distributed all over the sub-groups. Consequently, no amino acid mutation is a determining factor to differentiate the coordinates of the main-chain atoms at the binding pocket.

Discussion

The primary chemical agents used in protein crystallization were surveyed as shown in Table 5. The most dominant agent was selected from the description in the PDB file for each crystal structure and then the structures of each group were classified in terms of selected agents. Table 5 clearly indicates that ammonium sulfate and sodium chloride are the two major agents for crystallization of HIV-1 protease. About 90% of the crystals were obtained from these two chemicals if the crystal structures without information on crystallization agent were omitted. Polyethyleneglycol, which is frequently used in many kinds of proteins, was scarcely used in the crystallization of HIV-1 protease. In group 1, many of the protein crystallizations were achieved by using sodium chloride. Furthermore, sodium chloride was exclusively employed in group 5. In contrast, most of the crystal structures of groups 2, 3, and 4 were obtained by using Ammonium sulfate, which was dominantly employed in groups 2 and 4. A variety of agents were used in group 6. This large variation of chemicals in group 6 is natural because the cluster members have the structures not assigned to other groups. As shown in Table 1, the clusters deduced from cluster analysis of the whole-body structure mainly depend on the space group of crystals. The survey of crystallization agents in Table 5 indicates a compatibility between cluster group and chemical agent. These results suggest that the crystallization agent strongly influences the space group of the protein crystal and that the space group of crystals largely reflects the structure of the protease.

A major factor to distinguish the respective cluster groups is the space group of the crystals in whole-body structure of the protease. However, at the binding pocket, the relationship between the space group of crystals and the cluster groups is limited. In protein crystals, one protein molecule makes contacts with other surrounding molecules. The pattern of the contact with other molecules is determined by the space group of crystal. Therefore the difference in space group will be strongly reflected on the residue located at the contact area. Accordingly the outer part of proteins will be largely influenced by the difference in space group. This is reason why the whole body-structure of protein crystals strongly depends on the space group of protein crystals and the structure of the residues at the binding pocket does not so strongly.

The relationship between the cluster group and the pH condition in protein crystallization was surveyed as shown in Table S3. The averaged pH values are within the range of 5.4 to 6.2 in all clusters. In cluster analysis with the whole-

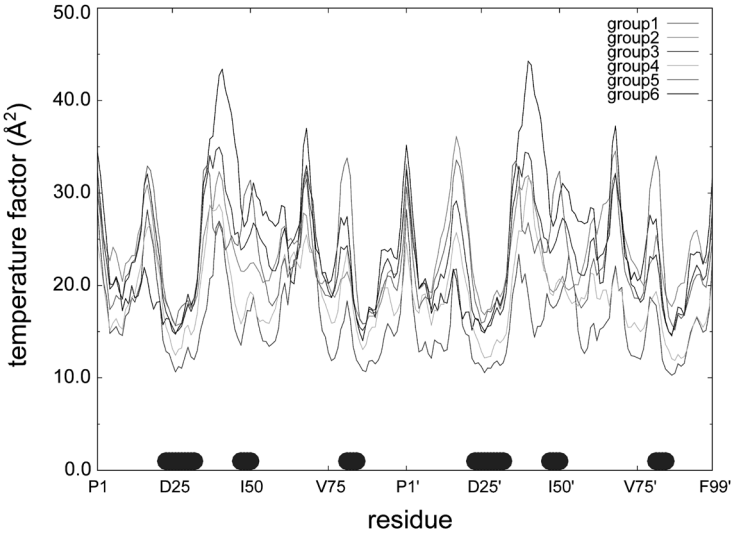

Fig. 6. Temperature Factors (B-Factors) of Individual Residues of HIV-1 Protease for Every Cluster Group

The bottom thick lines represent the location of residues at the binding pocket (L23–V32, I47–I50, P81–I84, L23′–V32′, I47′–I50′, P81′–I84′).

body structure, the standard deviations of the pH values are more than 0.7 in three groups. Standard deviations over 0.7 are also seen in two groups derived from cluster analysis with residues at the binding pocket. Judging from these results, it is concluded that the pH condition have little effect on structural diversity of the protein crystal of HIV-1 protease.

The results of a survey of the resolution of X-ray diffraction are summarized in Table S4. The resolutions of crystal structures of HIV-1 protease are fine, and the averaged resolutions are in the range of 1.6 to 2.0 Å in all the groups except for group 3. The averaged value of the resolutions of crystal structures in group 3 is 2.07 Å. Since the standard deviations are about 0.4 for groups 1–5, group 3 is definitely inferior to the other groups in resolution of X-ray diffraction.

The temperature factors were extracted from the PDB files and the averaged values were calculated for the main-chain atoms with regard to respective amino acid residues. The averaged temperature factors were summed up for all of the members of each group and then the average values were obtained as shown in Table S5. The temperature factors for groups 1 and 4 are small compared to those for the other groups. The temperature factor for group **3** is the largest among the groups derived from cluster analysis with the whole-body structure, while the standard deviation is considerably large. The average temperature factors for the respective amino acid residues are shown for each cluster group in Fig. 6. This diagram shows that the motions of the regions for the fulcrum (10–23), elbow (37–42), and cantilever (59–75) are large in all of the groups. The movement at the elbow region is markedly large in group 3. On the other hand, the motions of the residues around the binding pocket are small in all groups. In groups 1 and 4, the temperature factors are significantly small at the binding pocket. These low temperature factors are reflected in the small values in Table S5.

It is interesting to note that B-factor values are not exactly the same between two monomers. HIV-1 protease is a homodimer of the peptides, each of which consists of 99 amino acid residues. Although the sequence of each peptide is the same to the other, the crystals structures of HIV-1 protease are usually not symmetrical in each monomer. The reason of the asymmetry is that almost all of the crystal structures contain a substrate or an inhibitor at the binding pocket. The chemical structure of the inhibitor is not symmetrical, which will be reflected in the difference of crystal structures between two monomers.

The substrates or the inhibitors in complex with HIV-1 protease were surveyed as shown in Table S6. Many kinds of inhibitors have been examined in crystallographic studies. Hence, it is difficult to classify inhibitor compounds into several groups. Therefore, approved inhibitory agents and peptidomimetic molecules were focused on in this work. The number of crystal structures in complex with Darnavir is the largest among the approved inhibitors. Structures containing Darnavir scatter over the groups except for group 5 obtained by cluster analysis for the whole-body structure. The number of structures in complex with Saquinavir which is seen in four groups is the second largest. Crystal structures containing one of the approved inhibitors except for Tipranavir are distributed over at least three groups. This suggests that the whole-body crystal structure have little dependence on the inhibitory agents. The crystal structures in complex with peptides scatter over the groups. Peptidomimetic inhibitors are also not restricted to any single group. Hence, the inhibitory molecules in the crystal structure have little effect on the whole-body structure of HIV-1 protease. It is informative to note that two structures in group 5 contain neither substrate nor inhibitor. Therefore, the absence of a ligand may have a slight influence on the largely-deviated conformation of group 5. As for the flap conformation of one structure in group 5, the influence of the crystal packing was previously pointed out by a computational study.[38)]

The distribution of inhibitors was also examined for the groups obtained by cluster analysis with residues at the binding pocket (Table S6b). No characteristic distribution was observed for the inhibitory agents and peptidomimetic molecules between groups A and C. Group B, which is identical to group 5 in cluster analysis with the whole-body structure, shows a high population in the complex with peptides and also contains crystal structures without a substrate. In the subgroups obtained by cluster analysis with residues at the bind-

Table 6. Number of Crystal Structures Containing Protease Inhibitors or Peptidomimetic Substrates in Regard to the Sub-groups of Group A in Cluster Analysis with Respect to Residues at the Binding Pocket

Molecule	Sub-group A1	Sub-group A2	Sub-group A3	Sub-group A4	Total
Amprenavir	7 (0.05)	3 (0.03)	3 (0.04)	2 (0.02)	15 (0.04)
Atazanavir	0 (0.00)	4 (0.04)	3 (0.04)	1 (0.01)	8 (0.02)
Darnavir	10 (0.07)	7 (0.07)	9 (0.11)	5 (0.05)	31 (0.07)
Indinavir	0 (0.00)	3 (0.03)	4 (0.05)	6 (0.07)	13 (0.03)
Lopinavir	0 (0.00)	1 (0.01)	0 (0.00)	3 (0.03)	4 (0.01)
Nelfinavir	0 (0.00)	4 (0.04)	2 (0.02)	3 (0.03)	9 (0.02)
Ritonavir	0 (0.00)	1 (0.01)	1 (0.01)	3 (0.03)	5 (0.01)
Saquinavir	1 (0.01)	5 (0.05)	0 (0.00)	12 (0.13)	18 (0.04)
Tipranavir	0 (0.00)	0 (0.00)	1 (0.01)	4 (0.04)	5 (0.01)
TL-3_inhibitor	2 (0.01)	0 (0.00)	2 (0.02)	4 (0.04)	8 (0.02)
CA-p2_analogue	2 (0.01)	0 (0.00)	7 (0.08)	3 (0.03)	12 (0.03)
p2-NC_analogue	8 (0.06)	0 (0.00)	3 (0.04)	0 (0.00)	11 (0.03)
Peptide	4 (0.03)	10 (0.09)	13 (0.15)	12 (0.13)	39 (0.09)
No substrate	0 (0.00)	2 (0.02)	3 (0.04)	1 (0.01)	6 (0.01)
Others	100 (0.75)	66 (0.62)	33 (0.39)	33 (0.36)	232 (0.56)

a) The value in parenthesis represents the ratio relative to the total number of crystal structures in each sub-group.

ing pocket, no clear unbalanced population was observed in the distribution of inhibitory agents and peptidomimetic molecules, as shown in Table 6. Accordingly, the substrate or the inhibitors in complex with HIV-1 protease scarcely displaces the coordinates of the main-chain atoms even for the residues at the binding pocket.

To deduce the sub-groups of the primary cluster derived from analysis with residues at the binding pocket, we performed an additional cluster analysis of group A in Fig. 4a. As seen in Fig. 4b, separation of the branches is not clear and the sub-groups are not finely classified. Judging from the results shown in Tables 3, 4, and 6, the clustering of group A seems no more efficacious. Ko *et al.* recently performed data mining analysis of the binding pocket using crystal structures.[13] They searched for adequate chemical descriptors to identify the physiochemical features of the binding pocket of HIV-1 protease and tried to classify the pockets using the selected descriptors. As a consequence of their analysis, 70 binding pockets of the crystal structures were primarily grouped into two clusters. The inhibitors Atazanavir, Indinavir, and Saquinavir mainly belonged to one cluster and the inhibitors Darnavir, Nelfinavir, Amprenavir, Lopinavir, Ritonavir, and Tipranavir belonged to the other cluster. In our examination of the distribution of approved inhibitors over the sub-groups in Table 6, the inhibitors; Atazanavir, Indinavir, and Saquinavir, hardly show a distinguishable tendency from the other agents. Hence, the coordinates of the main-chain atoms at the binding pocket are not sensitive to the chemical species of the inhibitors. Instead, the geometric and polar features of side chains of the residues will be important to identify patterns of the binding pockets if the pockets can be classified. In the data mining study by Ko *et al.*,[13] they suggested that an important descriptor to classify the binding pockets of HIV-1 protease was the sum of exchange energy and electron–electron repulsion energy obtained by quantum chemical calculation. This finding indicates that the energetic aspect is critical for the binding pocket classification as well as the geometrical one. Further, not only static properties such as crystal structure but also dynamic features of the side-chain atoms of residues at the binding pocket may be of great importance as pointed out by Ai *et al.*[39]

Conclusion

Cluster analysis was applied to the crystal structures of HIV-1 protease. For the X-ray crystallographic data downloaded from the PDB site, 499 structures were selected as a base set for the cluster analysis. The crystal structures were classified into several groups by the nearest neighboring method. In the clustering with regard to the whole-body structure of the HIV-1 protease, 6 cluster groups were obtained. Information on amino mutation, space group of the crystals, category of substrate or inhibitor, and crystallization condition was surveyed for every member of the respective clusters. A major factor to distinguish the respective cluster groups is the space group of the crystals. Chemical agents used in the protein crystallization have a critical influence on the structural difference in the crystals. Amino acid mutations had little influence on the separation of cluster groups. As for the cluster analysis with a limited number of amino acid residues at the binding pocket, there was no clear relationship observed between the space group of crystals and the cluster groups. The crystal structures cannot be classified well by cluster analysis only with amino acid residues at the binding pocket. That is, the coordinates of the main-chain atoms of the residues at the binding pocket have no clear differences in the crystal structures. This means that amino acid mutations, especially drug-resistant mutations, have little effect on the coordinates of the main-chain atoms. Instead, mutations are closely related to the geometric and polar features of the side-chain atoms in HIV-1 protease.

Acknowledgments This work was supported by a Grant from the Ministry of Health and Labour of Japan. A part of this work was supported by a Grant for Scientific Research C from the Japan Society for the Promotion of Science.

References

1) Berger I., Mayr L. M., *Curr. Opin. Struct. Biol.*, **23**, 317–318 (2013).
2) Yumoto H., Mimura H., Koyama T., Matsuyama S., Tono K., Togashi T., Inubushi Y., Sato T., Tanaka T., Kimura T., Yokoyama H.,

Kim J., Sano Y., Hachisu Y., Yabashi M., Ohashi H., Ohmori H., Ishikawa T., Yamauchi K., *Nat. Photonics*, **7**, 43–47 (2012).
3) de Vera I. M. S., Smith A. N., Dancel M. C. A., Huang X., Dunn B. M., Fanucci G. E., *Biochemistry*, **52**, 3278–3288 (2013).
4) Singh M. K., Streu K., McCrone A. J., Dominy B. N., *J. Mol. Biol.*, **408**, 792–805 (2011).
5) Fernàndez G., Clotet B., Martınez M. A., *J. Virol.*, **81**, 2485–2496 (2007).
6) Matsuyama S., Aydan A., Ode H., Hata M., Sugiura W., Hoshino T., *J. Phys. Chem. B*, **114**, 521–530 (2010), and references therein.
7) Gao R., Cao B., Hu Y., Feng Z., Wang D., Hu W., Chen J., Jie Z., Qiu H., Xu K., Xu X., Lu H., Zhu W., Gao Z., Xiang N., Shen Y., He Z., Gu Y., Zhang Z., Yang Y., Zhao X., Zhou L., Li X., Zou S., Zhang Y., Li X., Yang L., Guo J., Dong J., Li Q., Dong L., Zhu Y., Bai T., Wang S., Hao P., Yang W., Zhang Y., Han J., Yu H., Li D., Gao G. F., Wu G., Wang Y., Yuan Z., Shu Y., *N. Engl. J. Med.*, **368**, 1888–1897 (2013).
8) Ali A., Bandaranayake R. M., Cai Y., King N. M., Kolli M., Mittal S., Murzycki J. F., Nalam M. N. L., Nalivaika E. A., Özen A., Prabu-Jeyabalan M. M., Thayer K., Schiffer C. A., *Viruses*, **2**, 2509–2535 (2010), and references therein.
9) Agniswamy J., Shen C.-H., Aniana A., Sayer J. M., Louis J. M., Weber I. T., *Biochemistry*, **51**, 2819–2828 (2012).
10) For example: Bandaranayake R. M., Kolli M., King N. M., Nalivaika E. A., Heroux A., Kakizawa J., Sugiura W., Schiffer C. A., *J. Virol.*, **84**, 9995–10003 (2010).
11) Heaslet H., Kutilek V., Morris G. M., Lin Y. C., Elder J. H., Torbett B. E., Stout C. D., *J. Mol. Biol.*, **356**, 967–981 (2006).
12) Foley B., Leitner T., Apetrei C., Hahn B., Mizrachi I., Mullins J., Rambaut A., Wolinsky S., Korber B., Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, "HIV Sequence Compendium 2013.": ‹http://www.hiv.lanl.gov/›, 2013.
13) Ko G. M., Reddy A. S., Kumar S., Bailey B. A., Garg R., *J. Chem. Inf. Model.*, **50**, 1759–1771 (2010).
14) Ode H., Neya S., Hata M., Sugiura W., Hoshino T., *J. Am. Chem. Soc.*, **128**, 7887–7895 (2006), and references therein.
15) Ode H., Ota M., Neya S., Hata M., Sugiura W., Hoshino T., *J. Phys. Chem. B*, **109**, 565–574 (2005), and references therein.
16) Miller M., Schneider J., Sathyanarayana B. K., Toth M. V., Marshall G. R., Clawson L., Selk L., Kent S. B., Wlodawer A., *Science*, **246**, 1149–1152 (1989).
17) "RCSB Protein Data Bank.": ‹http://www.rcsb.org/pdb›, cited March 13, 2014.
18) Aoki M., Danish M. L., Aoki-Ogata H., Amano M., Ide K., Das D., Koh Y., Mitsuya H., *J. Virol.*, **86**, 13384–13396 (2012).
19) Ode H., Matsuyama S., Hata M., Neya S., Kakizawa J., Sugiura W., Hoshino T., *J. Mol. Biol.*, **370**, 598–607 (2007).
20) Ode H., Matsuyama S., Hata M., Hoshino T., Kakizawa J., Sugiura W., *J. Med. Chem.*, **50**, 1768–1777 (2007), and references therein.
21) Mager P. P., *Med. Res. Rev.*, **21**, 348–353 (2001).
22) Braz A. S. K., Tufanetto P., Perahia D., Scott L. P. B., *Proteins*, **80**, 2680–2691 (2012).
23) Broglia R., Levy Y., Tiana G., *Curr. Opin. Struct. Biol.*, **18**, 60–66 (2008).
24) Korber B. T., Foley B. T., Kuiken C. L., Pillai S. K., Sodroski J. G., "Numbering Positions in HIV Relative to HXB2CG," Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 1998, pp. III-102–III-111.
25) DeLano W. L., The PyMOL Molecular Graphics System, Schrödinger, LLC.
26) Chambers J. M., Software for Data Analysis: Programming with R., Springer, New York, 2008.
27) Gentleman R., "R Programming for Bioinformatics," Chapman & Hall/CRC, Boca Raton, FL, 2008.
28) Perryman A. L., Lin J.-H., Mccammon J. A., *Protein Sci.*, **13**, 1108–1123 (2004), and references therein.
29) Logsdon B. C., Vickrey J. F., Martin P., Proteasa G., Koepke J. I., Terlecky S. R., Wawrzak Z., Winters M. A., Merigan T. C., Kovari L. C., *J. Virol.*, **78**, 3123–3132 (2004).
30) Liu Z., Wang Y., Brunzelle J., Kovari I. A., Kovari L. C., *Protein J.*, **30**, 173–183 (2011).
31) Agniswamy J., Shen C. H., Aniana A., Sayer J. M., Louis J. M., Weber I. T., *Biochemistry*, **51**, 2819–2828 (2012).
32) Piana S., Carloni P., Rothlisberger U., *Protein Sci.*, **11**, 2393–2402 (2002), and references therein.
33) Clemente J. C., Moose R. E., Hemrajani R., Whitford L. R., Govindasamy L., Reutzel R., McKenna R., Agbandje-McKenna M., Goodenow M. M., Dunn B. M., *Biochemistry*, **43**, 12141–12151 (2004).
34) Xie D., Gulnik S., Gustchina E., Yu B., Shao W., Qoronfleh W., Nathan A., Erickson J. W., *Protein Sci.*, **8**, 1702–1707 (1999), and references therein.
35) Rout M. K., Hosur R. V., *Arch. Biochem. Biophys.*, **482**, 33–41 (2009).
36) Davis D. A., Dorsey K., Wingfield P. T., Stahl S. J., Kaufman J., Fales H. M., Levine R. L., *Biochemistry*, **35**, 2482–2488 (1996).
37) Davis D. A., Yusa K., Gillim L. A., Newcomb F. M., Mitsuya H., Yarchoan R., *J. Virol.*, **73**, 1156–1164 (1999).
38) Layten M., Hornak V., Simmerling C., *J. Am. Chem. Soc.*, **128**, 13360–13361 (2006).
39) Ai R., Qaiser Fatmi M., Chang C. A., *J. Comput. Aided Mol. Des.*, **24**, 819–827 (2010).