Viruses **2012**, *4*, 348-362; doi:10.3390/v4030348

OPEN ACCESS

viruses

ISSN 1999-4915
www.mdpi.com/journal/viruses

Article

Mining the Protein Data Bank to Differentiate Error from Structural Variation in Clustered Static Structures: An Examination of HIV Protease

Balasubramanian Venkatakrishnan [†], **Miorel-Lucian Palii** [†], **Mavis Agbandje-McKenna** and **Robert McKenna** *

Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, FL 32610, USA; E-Mails: balavenkat@ufl.edu (B.V.); mlpalii@gmail.com (M.-L.P.); mckenna@ufl.edu (M.A.-McK.)

[†] These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: rmckenna@ufl.edu; Tel.: +1-352-392-5696; Fax: +1-352-392-3422.

Received: 5 February 2012; in revised form: 29 February 2012 / Accepted: 1 March 2012 / Published: 5 March 2012

Abstract: The Protein Data Bank (PDB) contains over 71,000 structures. Extensively studied proteins have hundreds of submissions available, including mutations, different complexes, and space groups, allowing for application of data-mining algorithms to analyze an array of static structures and gain insight about a protein's structural variation and possibly its dynamics. This investigation is a case study of HIV protease (PR) using in-house algorithms for data mining and structure superposition through generalized formulæ that account for multiple conformations and fractional occupancies. Temperature factors (*B*-factors) are compared with spatial displacement from the mean structure over the entire study set and separately over bound and ligand-free structures, to assess the significance of structural deviation in a statistical context. Space group differences are also examined.

Keywords: B-factor and spatial variation; data mining; HIV protease; structure superposition

1. Introduction

1.1. The Protein Data Bank

Established in 1971, the Protein Data Bank (PDB) has proved invaluable not only to the research community but also to students and educators [1]. The PDB has outgrown its initial purpose as a repository of the atomic coordinates of protein structures [2] and now contributes to the understanding of biological function by structural genomics and similar initiatives.

Over 71,000 structures were in the database at the time of this writing, and more are deposited weekly [3]. Considerable molecular dynamics work has been done to assess conformational changes and mobility in individual macromolecules, but looking at an entire array of static structures is an untapped approach. Extensively characterized proteins have hundreds of structures available in the PDB, including mutations, various inhibitor complexes, and different resolutions or crystallographic space groups. This provides a unique opportunity to apply data-mining algorithms to the multitude of static coordinates deposited in the PDB, obtaining a measure of reliability when deciding on the significance of a structural change, as well as possibly revealing an alternative, dynamic view of a protein.

1.2. HIV protease

The protein chosen as a case study to demonstrate this approach is the protease of the human immunodeficiency virus (HIV), the causative agent of acquired imunodeficiency syndrome (AIDS). The role of HIV protease (PR) in the maturation of the virus to an infective state has made it an attractive drug target. Several PR inhibitors have been used in AIDS therapy [4,5]. Highly Active Anti-Retroviral Therapy (HAART) combines multiple drugs, significantly improving prognoses [6]. However, the high mutation rate of the virus has given rise to a number of polymorphs, including drug-resistant mutants [5,7,8]. This has prompted extensive study of the mechanisms of inhibitor action and resistance in the various polymorphs. Recent investigations have looked at the structure of PR in different polymorphic forms, in both the bound and ligand-free state [8,9].

PR is a homodimer, as shown in figure 1. Each of the 99-residue monomers contributes a catalytic aspartate (D25) to the active site, located above the dimeric interface and enclosed by a pair of flaps [10]. The dynamic nature of the flaps has not prevented crystallographic examinations of PR, and numerous studies have worked toward elucidating the structural basis of drug action and resistance [11].

Figure 1. The PR dimer. Cartoon diagram of PDB ID 3hvp showing the monomers in orange and blue. Regions of PR structure are labeled, and relevant residue numbers are given in parenthesis. Rendered using PyMOL.

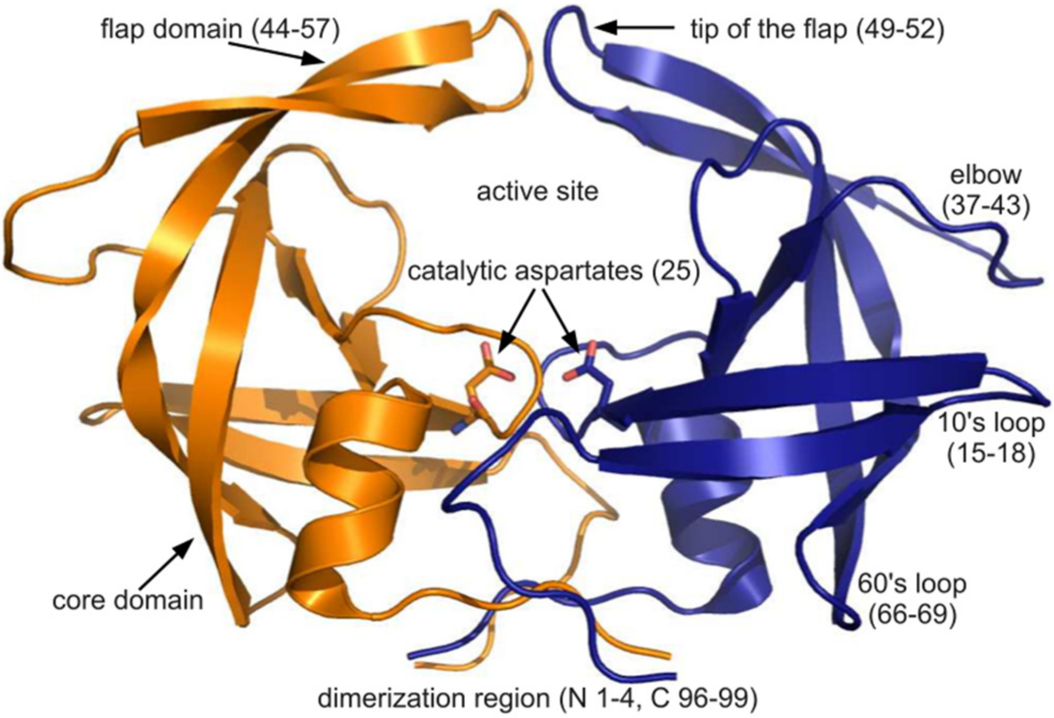
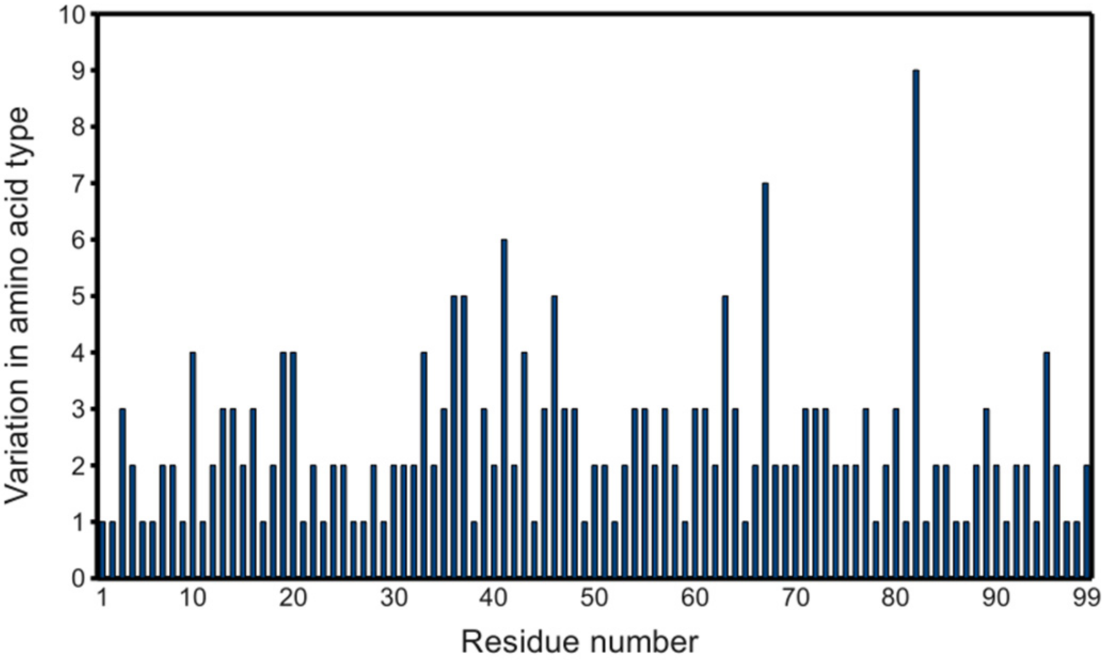


Figure 2. Variability of PR primary structures in the PDB. Graph considers the 811 PR monomers obtained as described in the experimental section, except PDB ID 2rkg, which contains an insertion. Non-standard residues (e.g., norleucine) are also included.



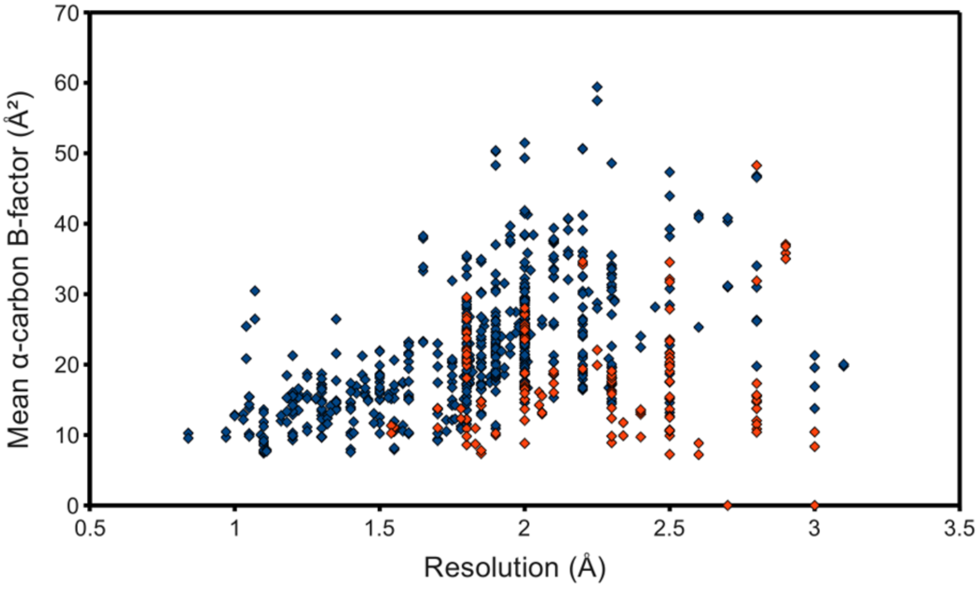
Since the first complete crystal structure of HIV-1 protease was solved [10], several hundred PR structures have been deposited in the PDB, covering significant variation in amino acid sequence, as shown in figure 2. The availability of such a substantial data set makes possible statistical probing into the properties of PR.

Presented here is an investigation using a set of in-house tools to data-mine the PDB, superpose structures, and calculate various parameters by residue or by structure. Mean temperature factors (*B*-factors) and spatial displacements were examined and correlated to resolution, ligand presence, and space group, and the obtained results were compared to current biological views.

2. Results and Discussion

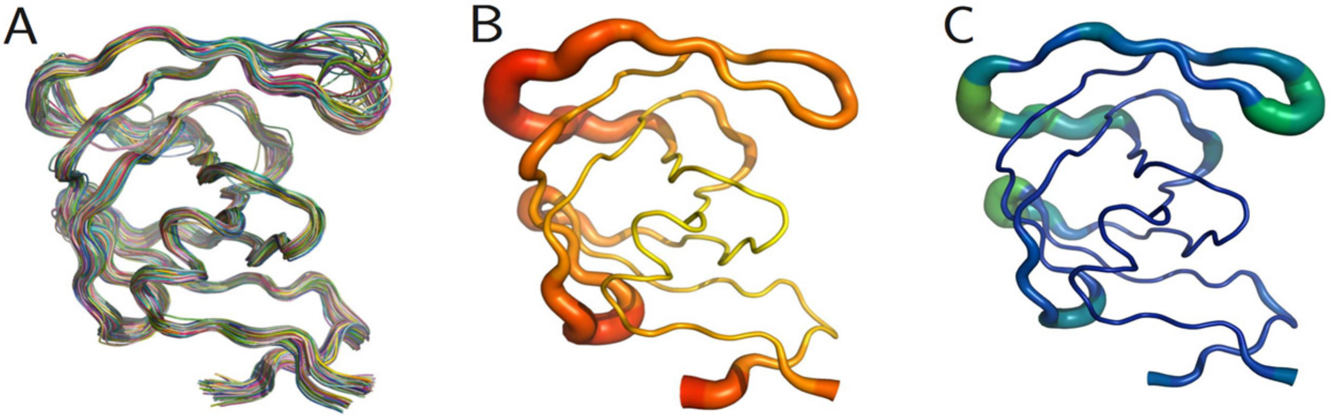
The occupancy-weighted average α-carbon *B*-factor was calculated for each of the resulting chains and plotted as the ordinate of a graph using structure resolution as the abscissa, in the hope of observing an association, even if possibly a weak one. However, figure 3 shows at best a resolution-dependent upper bound for the mean C *B*-factor. The lack of a stronger association can at least partially be attributed to the surprisingly low *B*-factors reported by some structures. Many files contained atomic coordinates with *B*-factors of 2 Å^2 and below, and several included negative *B*-factors. It was therefore necessary to remove from the study-set any structures containing *B*-factors lower than some reasonable value. This cut-off value came from a high resolution structure with reliable low *B*-factors, with no negative values. The highest resolution structure of lysozyme in the PDB at the time of this writing, PDB ID 2vb1 [14], lists no *B*-factors lower than 2.15 Å^2, so this was selected as the cut-off. 597 HIV protease chains passed, represented in figure 3 as blue points. This filtering step noticeably improved the linearity of the relationship between resolution and C *B*-factors.

Figure 3. Quality of deposited PR structures. Monomers that passed the *B*-factor cut-off of 2.15 Å^2 are marked blue, whereas those that failed are red. NMR structures, to which the concept of resolution does not apply, were not used for this plot.



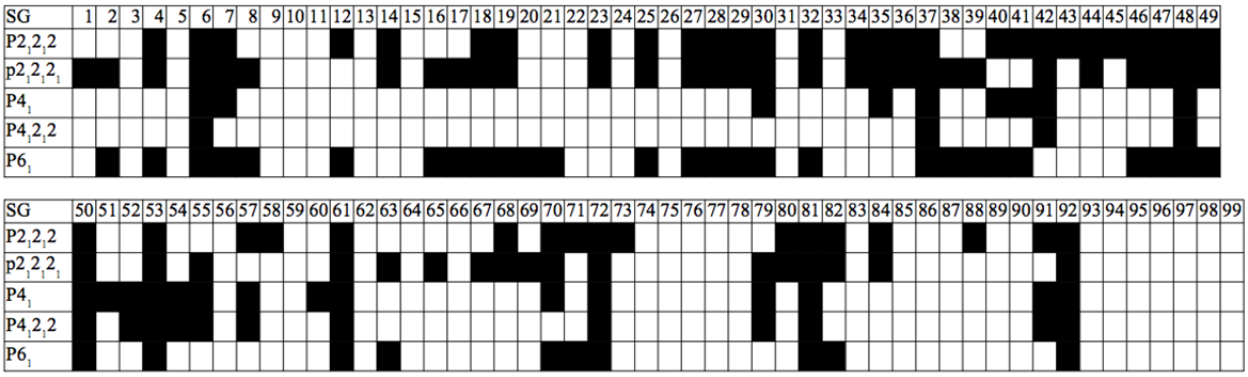
The PR monomers composing the study-set were superposed by the least squares method. Shown in figure 4A, the ribbon diagram representation of this superposition resembles an ensemble of NMR structures, even though no NMR structures were present in the data set. Though motion cannot be inferred directly from crystallographic data, it is worth noting that the greatest variation is observed in the flap and elbow, supporting the findings of NMR and molecular dynamics studies that have described these regions as the most dynamic[18]. Interestingly, the flap region showed a greater relative thermal stability. Figure 4B, a putty cartoon based on mean Cα *B*-factors, shows much greater values in the elbow and 60's loop than in the flap. However, when considering spatial displacement from the mean monomer (as in figure 4C), the tip of the flap joins the elbow and the 10's and 60's loops (defined as in figure 1) as one of the most variable regions, even though some of the range suggested by figure 4A has been averaged out.

Figure 4. PR monomers. (**A**) Ribbon diagram of the final data set superposed by least squares. (**B**) Putty cartoon of *B*-factor variation on the mean structure, colored from low to high (yellow to red). (**C**) putty cartoon of spatial variation on the mean structure, colored from low to high (blue to green). Refer to figure 1 for definitions of PR regions. Rendered using PyMOL (DeLano, 2002)



A possible explanation would be the existence of two distinct conformations of the enzyme: open and closed. In the latter, the presence of a ligand would enable interactions that hold the flap closed, ensuring its stability. In the former, steric clashes with symmetry-related molecules may limit flap opening and movement, or alternate conformers may be induced by amino acid variation. An analysis of crystal contacts across the various space groups mentioned in Table 1 affirms that there are several crystal contacts on the flap and elbow regions. Residues that formed crystal contacts in all the structures within each space group were used to calculate consensus contact regions within the space group. Though there are regions of contact that are specific to some space groups, the elbow and flap stretch and a number of other key contact points were common for all the space groups.

Table 1. Distribution of crystal contacts by residue in representative structures from each of the space groups reported for PR. This is not a table but a figure. Author need to use Word Table tools to format table.



SG	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
$P2_12_12$				■		■	■					■		■				■	■				■		■		■	■	■	■		■		■	■	■	■			■	■	■	■	■	■	■	■	■	■
$p2_12_12_1$	■	■		■		■	■	■						■		■	■	■	■				■		■		■	■	■	■		■		■	■	■	■	■	■			■		■		■	■	■	■
$P4_1$						■	■																							■					■		■			■	■	■						■	
$P4_12_12$						■																															■					■						■	
$P6_1$		■		■		■	■	■				■				■	■	■	■	■	■				■		■	■	■	■		■					■	■	■	■	■					■	■	■	■

SG	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
$P2_12_12$	■			■				■	■			■							■		■	■	■	■							■	■	■		■				■			■	■							
$p2_12_12_1$	■			■		■						■		■		■		■	■	■	■		■							■	■	■	■		■								■							
$P4_1$	■	■	■	■	■	■		■			■	■									■		■							■		■										■	■							
$P4_12_12$	■		■	■	■	■		■				■											■							■		■										■	■							
$P6_1$	■			■								■		■							■	■	■									■	■										■							

Figure 5A, a plot of the range of temperature factors and spatial displacement by residue, confirms the elbow, the tip of the flap, and the other loops as maxima. Overall, the *B*-factor seems to adequately fulfill its role as a *de facto* measure of spatial variation because there is high agreement in the location of the extrema of the two data series. However, the *B*-factor is not as reliable in predicting the magnitude of these extrema. Crystal contacts deduced from structures in different space groups seemingly coincide with regions of higher *B*-factors which may support the effect of crystal artifacts on the actual dynamics and *B*-factor values. The unreliability is especially apparent when separately treating ligand-bound (figure 5B) and ligand-free (figure 5C) monomers. The majority of PR structures in the PDB are bound to a ligand, so figure 5A,B do not differ significantly.

Figure 5. PR *B*-factor (orange) and spatial displacement (blue) variation with residue number. (**A**) Final data set, (**B**) bound monomers, (**C**) ligand-free monomers. Values were normalized for comparison purposes. Mean and standard deviation values are given in Table 2. Secondary structure elements are identified.

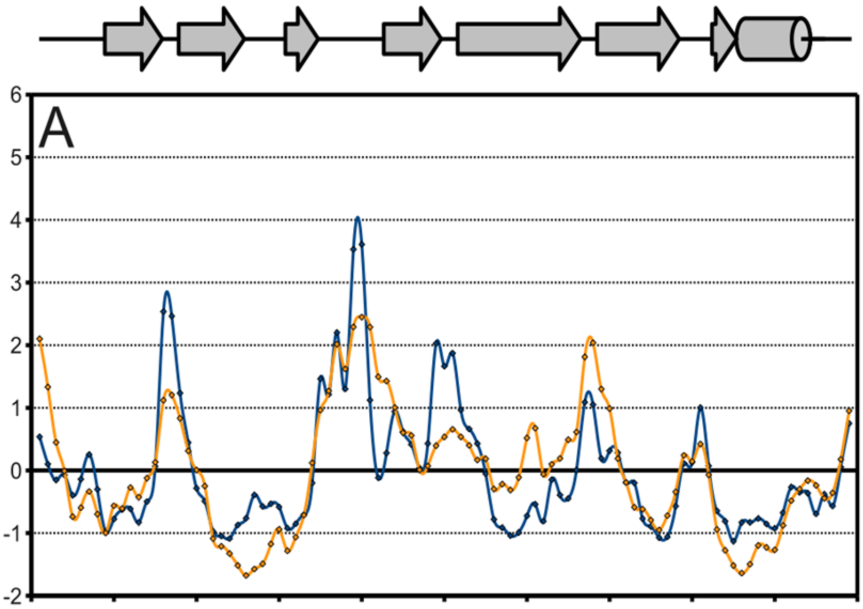
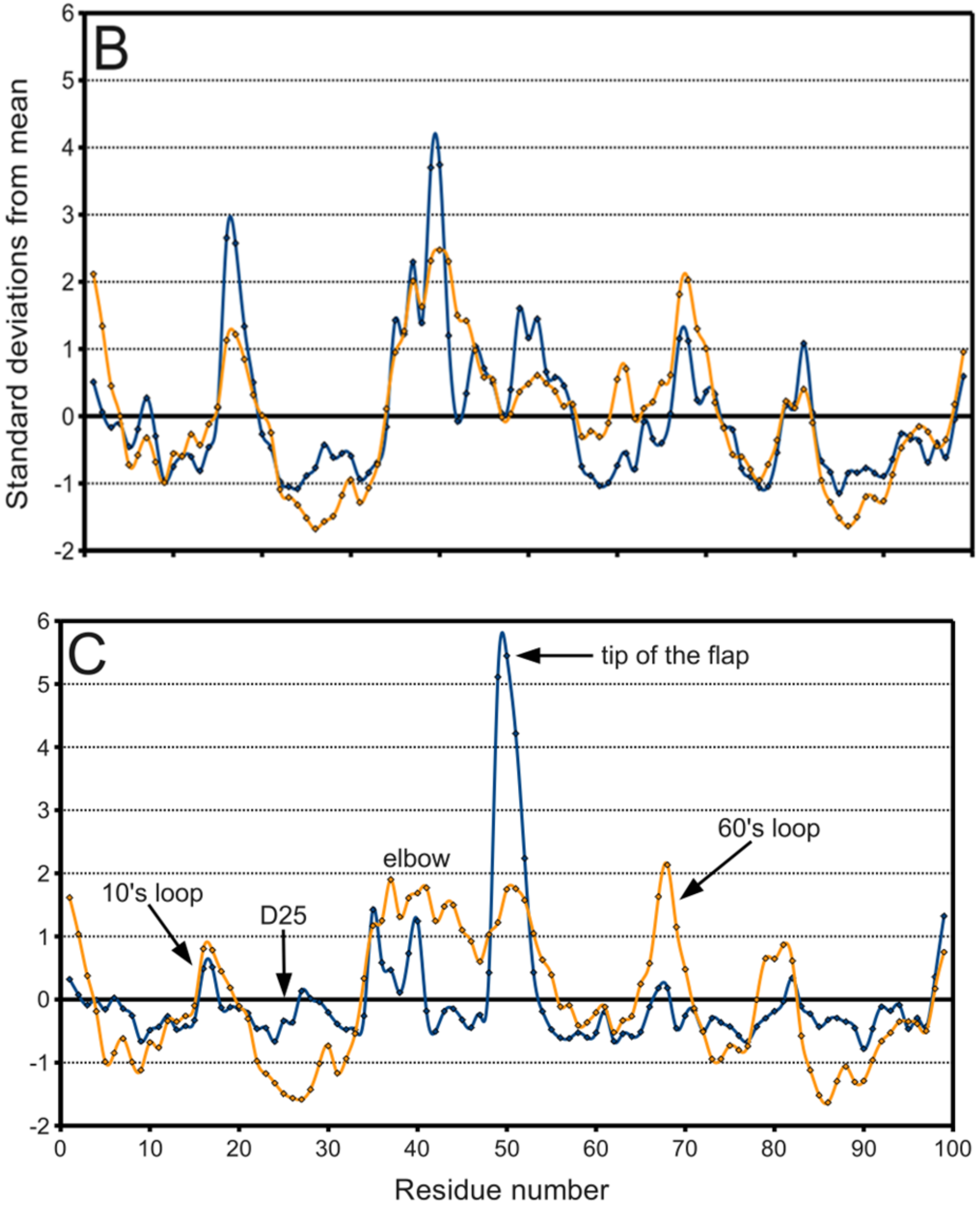
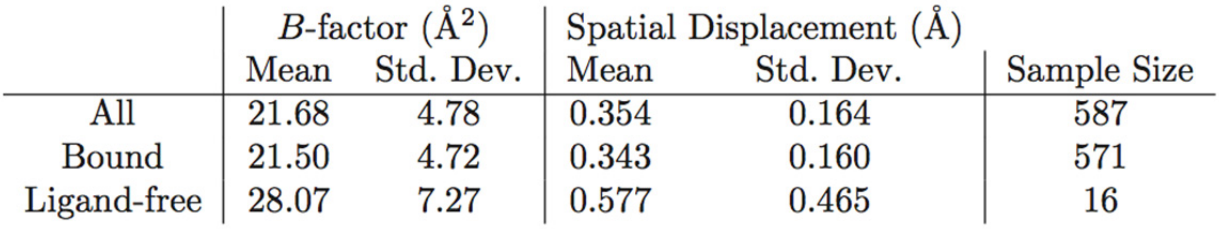


Figure 5. *Cont.*



In figure 5C on the other hand, the tip of the flap exhibits by far the greatest displacement from the mean ligand-free structure, and this value is much larger than the corresponding *B*-factor might indicate. Furthermore, the distribution of ligand-free structures is as a whole more variable spatially than that of bound ones, as described in Table 2. Spatial displacement over ligand-free structures has both a greater mean, 0.577 Å, and a greater standard deviation, 0.465 Å, than over bound structures (mean = 0.343 Å, standard deviation = 0.160 Å). The difference may be partially due to the discrepancy in sample sizes, but it nevertheless suggests the possibility of multiple PR conformations in the absence of a ligand.

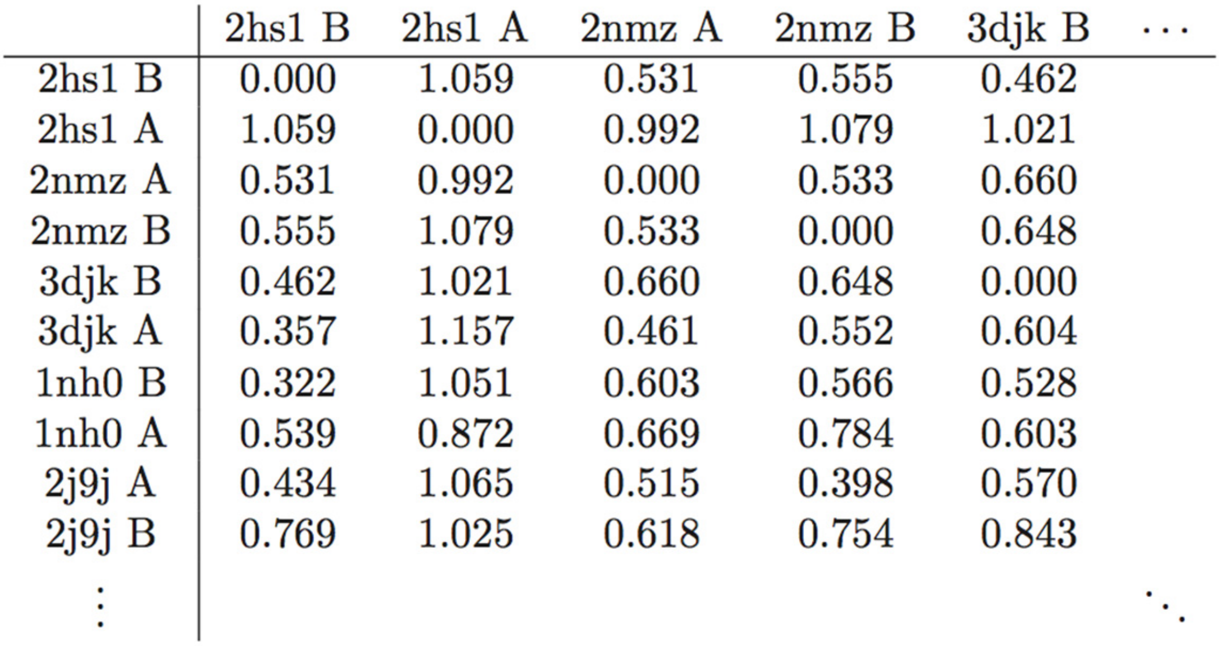
Table 2. PR B-factor and spatial displacement distribution. This table shows the mean spatial displacement observed in the ligand-bound Vs ligand-free PR structures. This table is a figure.



	B-factor (Å^2)		Spatial Displacement (Å)		
	Mean	Std. Dev.	Mean	Std. Dev.	Sample Size
All	21.68	4.78	0.354	0.164	587
Bound	21.50	4.72	0.343	0.160	571
Ligand-free	28.07	7.27	0.577	0.465	16

From this analysis of an entire array of structures, it is also possible to obtain an estimate of what Å value represents a significant conformational change. Referring to the statistics in Table 2, a change of 0.5 Å or below is within error range. A spatial displacement of 1.0 Å, or approximately four standard deviations from the mean of the entire study-set, is more convincing. In the distance matrix of pairwise *RMSD*s, of which a small segment is given as Table 3, several structures, namely PDB IDs 1xl2, 2fns, 2fnt, 2hs1, 2hs2, and 5upj, have *RMSD*s of 0.95 Å and above separating the monomers that compose them, supporting the finding that the two monomers adopt different conformational states when PR binds an asymmetric ligand [19]. The initial work of Prabu-Jeyabalan *et al.* [19] was on an inactivated HIV-1 PR-substrate complex, and the two monomers in the reported structure (PDB ID 1f7a) have an *RMSD* of only 0.34 Å. However, of the aforementioned structures, only PDB IDs 2fns and 2fnt have peptide ligands whereas the rest are bound to non-peptides, and PDB ID 5upj is an HIV-2 PR. The observation may therefore be conjectured to hold generally for HIV proteases and asymmetric ligands.

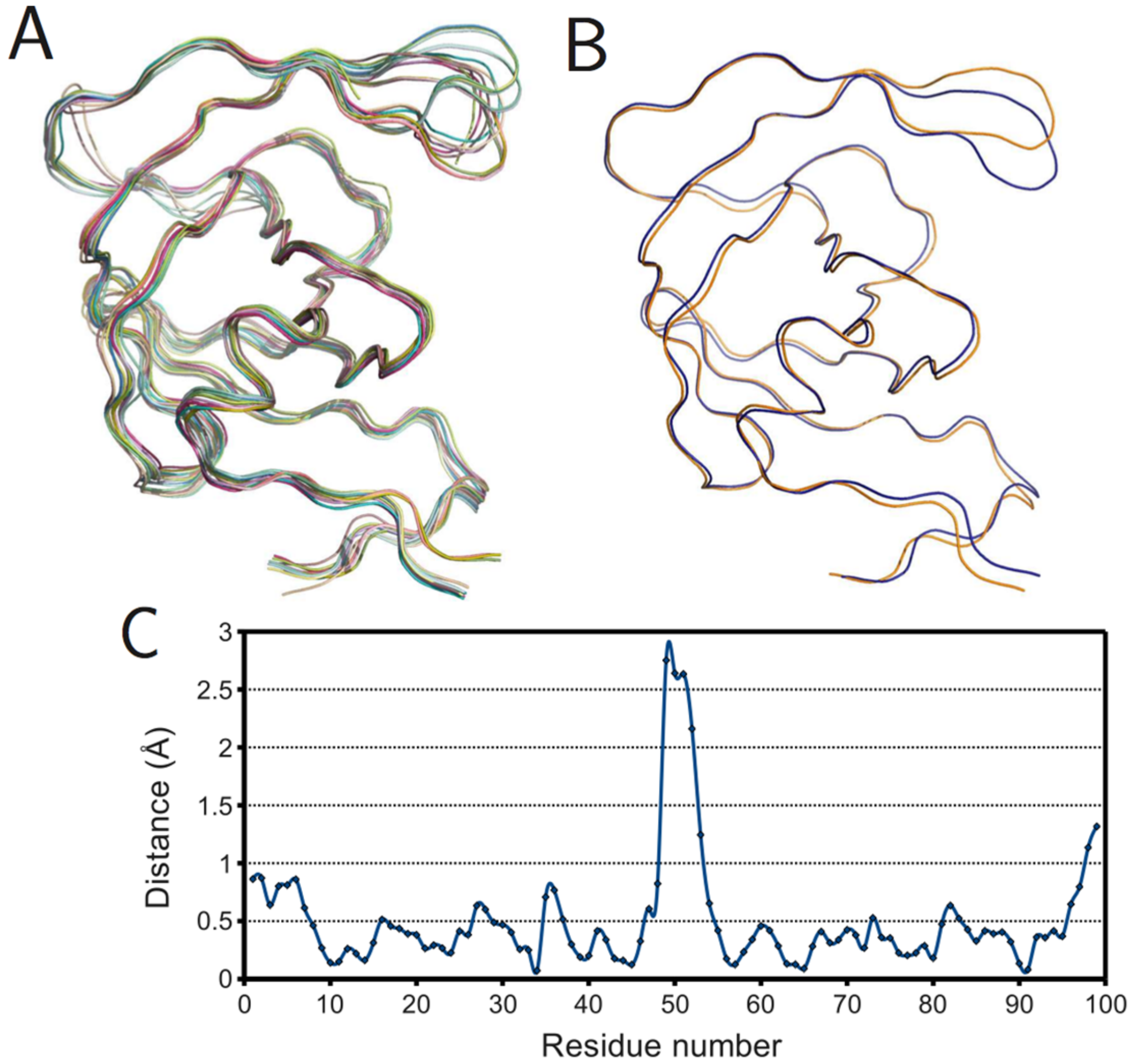
Table 3. Representative table of the pairwise RMSD distance (Å) matrix of the 587 monomers in the study set. Rows and columns are labeled with the PDB ID and chain identifier. This is a figure.



	2hs1 B	2hs1 A	2nmz A	2nmz B	3djk B	···
2hs1 B	0.000	1.059	0.531	0.555	0.462	
2hs1 A	1.059	0.000	0.992	1.079	1.021	
2nmz A	0.531	0.992	0.000	0.533	0.660	
2nmz B	0.555	1.079	0.533	0.000	0.648	
3djk B	0.462	1.021	0.660	0.648	0.000	
3djk A	0.357	1.157	0.461	0.552	0.604	
1nh0 B	0.322	1.051	0.603	0.566	0.528	
1nh0 A	0.539	0.872	0.669	0.784	0.603	
2j9j A	0.434	1.065	0.515	0.398	0.570	
2j9j B	0.769	1.025	0.618	0.754	0.843	
⋮						⋱

To further understand the effects of ligand binding on PR structure, monomers were superposed within the bound and ligand-free subsets. Figure 6A shows the result for the ligand-free monomers. Surprisingly, not all structures had flaps in the "semi-open" or open conformations. Several exhibited the closed flap conformation, though closer examination revealed these to belong to covalently-bonded PR dimers (PDB IDs 1g6l and 1lv1) that were split into monomers by removing the bridge of connecting residues. This also explains why these structures differ noticeably at the C-terminus from the other ligand-free structures. The superposition of the mean bound and mean ligand-free monomers rendered as a ribbon diagram in figure 6B, shows a prominent difference in the tip of the flap but little variation elsewhere. Figure 6C gives the same information as a plot of spatial difference by residue, and the spike corresponding to the tip of the flap is unmistakable. However, the maximal distance, 2.75 Å, is smaller than the actual deviation between the open and closed conformations, because the mean ligand-free structure is closer to the semi-open state due to averaging.

Figure 6. Ligand effects on PR monomer structure. (A) Superposition of ligand-free monomers; (B) superposition of mean ligand-free (orange) and bound (blue) monomers; (C) plot of spatial difference Vs residue number for mean ligand-free and bound monomers. Ribbon diagrams rendered using PyMOL.



Monomers were also organized on the basis of crystallographic space group and superposed to obtain mean monomers. Most of the representative monomers were in the closed conformation, as shown in figure 7. Exceptions were the monomers corresponding to the C2, $P4_12_12$, and $P4_1$ space groups. Interestingly, all structures that crystallized in the C2 space group were bound to a ligand, as listed in Table 4. This surprising observation may be due to the fact that all C2 structures except PDB ID 1ztz were of HIV-2 PR and solved during the same study. The deviation of the $P4_12_12$ space group is accounted for by noting that all its monomers are ligand-free, except one (PDB ID 3bc4 [20]) whose flaps are prevented from closing by two non-peptide inhibitors that pack the active site, acting as a wedge. Finally, the $P4_1$ space group has an almost equal distribution of bound and ligand-free monomers but differs the most from the closed conformation, which would be expected of a predominantly ligand-free space group. This may be because most of the $P4_1$ structures have mutations at residues 82 and 84, which are essential to ligand binding and structural stability in the active site [21].

Figure 7. Superposition of mean PR monomer structures for the space groups: $P2_1$ (orange), C2 (red), $P2_12_12$ (chartreuse), $P2_12_12_1$ (yellow), I222 (purple), $P4_1$ (cyan), $P4_3$ (lime green), $P4_12_12$ (blue), $P4_32_12$ (magenta), $I4_122$ (salmon), $P6_1$ (olive), $P6_5$ (brown), $P6_122$ (pink) and $I2_13$ (green). Ribbon diagram rendered using PyMOL. Table 3 describes the distribution of space groups in the final data set.

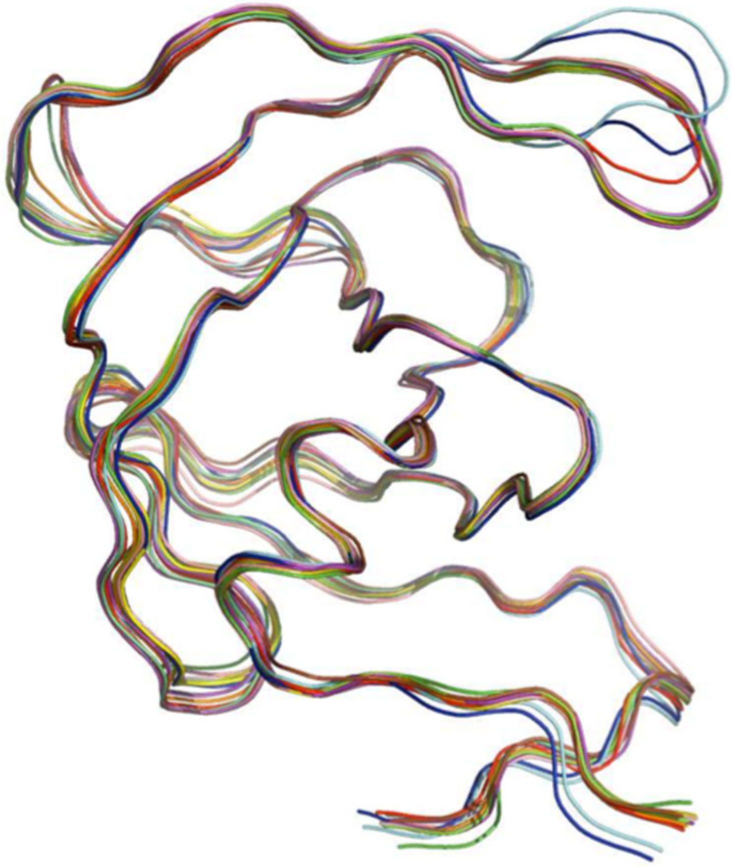


Table 4. Distribution of PR structures by space group. In the strictest sense, the distribution should be further subdivided because not all structures belonging to the same space group have isomorphous unit cells.

Space Group	Ligand-free	Bound	Total
$P2_1$	2	22	24
C2	0	8	8
$P2_12_12$	0	212	212
$P2_12_12_1$	0	184	184
I222	0	4	4
$P4_1$	4	6	10
$P4_3$	0	6	6
$P4_12_12$	6	1	7
$P4_32_12$	0	2	2
$I4_122$	0	2	2
$P6_1$	4	112	116
$P6_5$	0	2	2
$P6_122$	0	9	9
$I2_13$	0	1	1
Total	16	571	587

3. Experimental Section

3.1. Data-mining the Protein Data Bank

A list of relevant PDB IDs was obtained by querying the PDB for structures matching the keywords "HIV protease." The same 174 hits were returned regardless of whether the query was executed programmatically or through the PDB's web interface. Expanding the search parameters to include "human immunodeficiency virus protease" increased the number of results to 405. However, a review of the literature revealed that the true number of relevant structures in the PDB is greater still. Using the aforementioned search parameters, a home-grown script that does not limit itself to phrases explicitly declared as keywords found an additional 34 PDB IDs. Several of the previously-missed files were in fact PR structures that had managed to evade normal search mechanisms by having only general keywords, such as hydrolase.

3.2. Refining the search results

Of the 439 PDB IDs obtained, many corresponded not to PR, but to other proteins, including but not limited to integrase, reverse transcriptase, and proteases of the simian immunodeficiency viruses and feline immunodeficiency viruses (SIV and FIV, respectively). Therefore, a screening tool was implemented to omit structures containing no chains with primary structures within a specified edit distance [12] of several reference PR sequences from both HIV-1 and HIV-2. The two variants exhibit very similar overall structure despite having only about 50% sequence identity [13]. The results of the screen still had to be checked manually; SIV protease structures passed this test due to high sequence homology with PR. Conversely, covalently bonded PR dimers failed due to high edit distances; one of the monomers as well as any connecting residues had to be deleted to match the reference sequences. Hence, SIV proteases were removed, and tethered HIV proteases were readded. Despite the limitations, developing tools to facilitate the tedious task of selecting search results is an essential step towards the ultimate goal of having complete data mining packages to take advantage of the ever-increasing volume of information contained in the PDB.

3.3. Quality control

The 368 structure files remaining after the previous step included several structures from NMR experiments and were split into 811 PR monomers. In the case of the covalently bonded PR dimers, this involved stripping any connecting residues. NMR structures were excluded from the study-set. The final data-set can be found in the supplementary data section.

3.4. Structure superposition

The problem of superposing two sets of three-dimensional coordinates (atomic or not) reduces to identifying the rotation and translation that minimize some error function, usually the root-mean-square deviation (*RMSD*) between the coordinate sets in question:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(\Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2)}{N}} \tag{1}$$

Because *N* is just a constant, and the square root is a strictly increasing function, this is equivalent to the least squares criterion. In this case, the optimal translation can be calculated independent of rotation and is well known as the vector separating the centroids of the two coordinate sets [15].

Several methods are known for computing the optimal rotation. The most popular seems to be that of Kabsch [16], but this approach represents rotations as 3 x 3 matrices and can give rise to rotoinversions. For this investigation, the method of Horn [17] was chosen because it circumvents this problem by instead representing rotations as unit quaternions. Letting (x_{Ai}, y_{Ai}, z_{Ai}) denote the displacement of the *i*th point in set *A* from its centroid, the optimal rotation becomes the eigenvector corresponding to the most positive eigenvalue of the symmetric 4 X 4 matrix:

$$\begin{bmatrix} (S_{xx} + S_{yy} + S_{zz}) & S_{yz} - S_{zy} & S_{zx} - S_{xz} & S_{xy} - S_{yx} \\ S_{yz} - S_{zy} & (S_{xx} - S_{yy} - S_{zz}) & S_{xy} + S_{yx} & S_{zx} + S_{xz} \\ S_{zx} - S_{xz} & S_{xy} + S_{yx} & (-S_{xx} + S_{yy} - S_{zz}) & S_{yz} + S_{zy} \\ S_{xy} - S_{yx} & S_{zx} + S_{xz} & S_{yz} + S_{zy} & (-S_{xx} - S_{yy} + S_{zz}) \end{bmatrix} \tag{2}$$

where $S_{xx} = \sum_i x_{Ai}x_{Bi}$, $S_{xy} = \sum_i x_{Ai}y_{Bi}$, and so on.

Applying this algorithm to the atomic coordinates of protein structures involves accounting for fractional occupancies and multiple conformations. This amounts to using the occupancy-weighted centroids for the translation step and generalizing x_{Ai} to the occupancy-weighted average of its conformations *j*,

$$x_{Ai} = \frac{\sum_j occ_{Aij} x_{Aij}}{occ_{Ai}} \tag{3}$$

and similarly for *y* and *z*, where occ_{Aij} denotes the occupancy of the *j*th conformation and $occ_{Ai} = \Sigma_j occ_{Aij}$ is the total occupancy corresponding to (x_{Ai}, y_{Ai}, z_{Ai}) . However, the *S* summations are also occupancy-weighted, which factors out the denominators in (3), and S_{xy}, for example, becomes

$$S_{xy} = \sum_{ijk} (occ_{Aij} x_{Aij})(occ_{Bik} y_{Bik}) \tag{4}$$

In the simple case of a single conformation with full occupancy, the formula reduces to that of Horn [17].

Unfortunately, a closed-form recipe for the superposition of multiple structures does not exist. A first approach might be to superpose all structures onto the same reference structure, but the results may be erroneous if the reference is poorly chosen. A possible improvement would be to superpose each structure onto the mean of those already considered, but even this strategy is vulnerable to the effects of an arbitrary order of superposition because structures considered earlier are inherently attributed more importance. To alleviate this problem, the 587 monomers to be analyzed were sorted from highest to lowest resolution, with ties being broken in favor of the structure with the lowest mean C^{α} *B*-factor. Equal treatment of all structures would be ideal, but preferring "better" monomers is acceptable.

4. Conclusions

Analysis of a static array of PDB structures to gain further insight about a protein has great potential as a method to deduce a statistical bar for structural variation, as demonstrated by this PR case study. While there are other algorithms to data-mine the PDB, it is clear from this study that quality control is required before using a data-set for analysis. There also exist algorithms to superpose structures, but to our knowledge, this is the first method that also occupancy-weighs available conformations for the superposition. The algorithms described here were used to data-mine the PDB, filter search results, perform quality control, and superpose structures. This made possible a comparison of *B*-factors and spatial variation over the entire study-set of PR monomers, the bound and ligand-free subsets, and the different represented space groups. Examination of the resulting distributions is an alternative way of identifying a protein's most variable regions and qualifying spatial displacement as significant or within the range of error.

However, such an approach to protein study is made more difficult by the many different practices of PDB depositors even within the limits of a file format with a detailed specification. Choice of title, choice of keywords, numbering of residues, and organization into models and chains are often overlooked. This is unnoticeable to a human user, but it makes selection of the study-set the most complex and error-prone step of a data-mining endeavor. Additionally, many structures abuse *B*-factors, occupancies, and other parameters, or assign them special meaningless values not specified by the PDB file format. Therefore quality controls must be implemented to exclude from such investigations any structures with statistics that might bias results. For data-mining investigations to be successful, a paradigm shift will be required of depositors to the PDB: to stop treating the painstaking process of preparing a structure for submission as an unnecessary complication and see the PDB itself not just as a collection of coordinates, but as a tool that could shed light on many of the questions of structural biology.

Acknowledgments

This project could not have happened without the work of all those who have deposited structures to the PDB and those who continue to maintain a smoothly-running database.

Conflict of Interest

The authors declare no conflict of interest.

References and Notes

1. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; *et al.* The Protein Data Bank. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 899-907.
2. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.; Meyer, E.F., Jr; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* **1978**, *185*, 584-591.
3. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
4. Debouck, C. The HIV-1 protease as a therapeutic target for AIDS. *AIDS Res. Hum. Retroviruses* **1992**, *8*, 153-164.
5. Flexner, C. HIV-protease inhibitors. *N. Engl. J. Med.* **1998**, *338*, 1281-1292.
6. Carpenter, C.C.; Fischl, M.A.; Hammer, S.M.; Hirsch, M.S.; Jacobsen, D.M.; Katzenstein, D.A.; Montaner, J.S.; Richman, D.D.; Saag, M.S.; Schooley, R.T.; *et al.* Antiretroviral therapy for HIV infection in 1997. Updated recommendations of the International AIDS Society-USA panel. *JAMA* **1997**, *277*, 1962-1969.
7. Condra, J.H.; Schleif, W.A.; Blahy, O.M.; Gabryelski, L.J.; Graham, D.J.; Quintero, J.C.; Rhodes, A.; Robbins, H.L.; Roth, E.; Shivaprakash, M. *In vivo* emergence of HIV-1 variants resistant to multiple protease inhibitors. *Nature* **1995**, *374*, 569-571.
8. Clemente, J.C.; Robbins, A.; Graña, P.; Paleo, M.R.; Correa, J.F.; Villaverde, M.C.; Sardina, F. J.; Govindasamy, L.; Agbandje-McKenna, M.; McKenna, R.; *et al.* Design, synthesis, evaluation, and crystallographic-based structural studies of HIV-1 protease inhibitors with reduced response to the V82A mutation. *J. Med. Chem.* **2008**, *51*, 852-860.
9. Heaslet, H.; Rosenfeld, R.; Giffin, M.; Lin, Y.C.; Tam, K.; Torbett, B.E.; Elder, J.H.; McRee, D. E.; Stout, C.D. Conformational flexibility in the flap domains of ligand-free HIV protease. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 866-875.

10. Wlodawer, A.; Miller, M.; Jaskólski, M.; Sathyanarayana, B.K.; Baldwin, E.; Weber, I.T.; Selk, L.M.; Clawson, L.; Schneider, J.; Kent, S.B. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* **1989**, *245*, 616-621.
11. Wlodawer, A.; Vondrasek, J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Ann. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249-284.
12. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707-710.
13. Priestle, J.P.; Fässler, A.; Rösel, J.; Tintelnot-Blomley, M.; Strop, P.; Grütter, M.G. Comparative analysis of the X-ray structures of HIV-1 and HIV-2 proteases in complex with CGP 53820, a novel pseudosymmetric inhibitor. *Struct.* **1995**, *3*, 381-389.
14. Wang, J.; Dauter, M.; Alkire, R.; Joachimiak, A.; Dauter, Z. Triclinic lysozyme at 0.65 A resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007**, *63*, 1254-1268.
15. Flower, D.R. Rotational superposition: a review of methods. *J. Mol. Graphics Modell.* **1999**, *17*, 238-244.
16. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1978**, *34*, 827-828.
17. Horn, B.K.P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A:* **1987**, *4*, 629-642.
18. Collins, J.R.; Burt, S.K.; Erickson, J.W. Flap opening in HIV-1 protease simulated by “activated” molecular dynamics. *Nat. Struct. Biol.* **1995**, *2*, 334-338.
19. Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C.A. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *J. Mol. Biol.* **2000**, *301*, 1207-1220.
20. Böttcher, J.; Blum, A.; Dörr, S.; Heine, A.; Diederich, W.E.; Klebe, G. Targeting the open-flap conformation of HIV-1 protease with pyrrolidine-based inhibitors. *ChemMedChem* **2008**, *3*, 1337-1344.
21. 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. Crystal structures of a multidrug-resistant human immunodeficiency virus type 1 protease reveal an expanded active-site cavity. *J. Virol.* **2004**, *78*, 3123-3132.