Characterizing the structural variability of HIV-2 protease upon the binding of diverse ligands using a structural alphabet approach

Dhoha Triki[1], Sandrine Fartek[1], Benoit Visseaux[2], Diane Descamps[2], Anne-Claude Camproux[1], Leslie Regad[1,*].

[1]*MTi, UMR-S973, Université Paris Diderot, Sorbonne Paris Cité, INSERM, F-75013, Paris, France*

[2] *IAME, UMR 1137, Université Paris Diderot, Sorbonne Paris Cité, INSERM, AP-HP, Hôpital Bichat, Virologie, F-75018 Paris, France*

leslie.regad@univ-paris-diderot.fr

Characterizing the structural variability of HIV-2 protease upon the binding of diverse ligands using a structural alphabet approach

The HIV-2 protease (PR2) is an important target for designing new drugs against the HIV-2 infection. In this study, we explored the structural backbone variability of all available PR2 structures complexed with various inhibitors using a structural alphabet approach. 77% of PR2 positions are structurally variable, meaning they exhibit different local conformations in PR2 structures. This variability was observed all along the structure, particularly in the elbow and flap regions. A part of these backbone changes observed between the 18 PR2 are induced by intrinsic flexibility and ligand binding putatively induces others occurring in the binding pocket. These latter changes could be important for PR2 adaptation to diverse ligands and are accompanied by changes outside of the binding pocket. In addition, the study of the link between structural variability of the pocket and PR2-ligand interactions allowed us to localize pocket regions important for ligand binding and catalytic function, regions important for ligand recognition that adjust their backbone in response to ligand binding, and regions important for the pocket opening and closing that have large intrinsic flexibility. Finally, we suggested that differences in ligand effectiveness for PR2 could be partially explained by different backbone deformations induced by these ligands. To conclude, this study is the first characterization of the PR2 structural variability considering ligand diversity. It provides information about recognition of PR2 to various ligands and its mechanisms to adapt its local conformation to bound ligands that could help understanding the resistance of PR2 to its inhibitors, a major antiretroviral class.

HIV-2 protease; structural variability; structural deformation upon ligand binding; structural alphabet; ligand specificity.

Introduction

There are two types of HIV: HIV type 1 (HIV-1) and HIV type 2 (HIV-2) corresponding to distinct simian origins. The drugs used for the treatment of HIV-2

correspond to those developed for HIV-1, targeting various viral proteins: fusion complex, integrase, reverse transcriptase, and protease. HIV-2 is naturally resistant to all non-nucleoside inhibitors of reverse transcriptase or fusion inhibitors (Ren et al., 2002; Poveda et al., 2005) and also demonstrated reduced susceptibility to protease inhibitors (PIs) (Rodés et al., 2006; Masse et al., 2007; Desbois et al., 2008; Brower et al., 2008; Cavaco-Silva et al., 2013; Raugi et al., 2013 ; 2016; Visseaux et al., 2016). In vivo studies showed that HIV-2 does not produce a stronger immunological response to the more recently developed class of integrase inhibitors than previously observed with PIs (Ntemgwa et al., 2009). Thus, it is still necessary to develop new molecules specifically designed for HIV-2. One approach is based on the identification of new molecules inhibiting HIV-2 protease (PR2).

PR2 is an aspartic protease hydrolyzing the viral Gag and the Gag-Pol precursor polyproteins during the maturation of viral particles. It is a homodimer of 99 residues in each monomer, including the catalytic triplet Asp-Thr-Gly conserved in all aspartic proteases. Substrates and inhibitors bind the PR2 at the interface of the two monomers. Their bindings are associated with large conformational changes in PR2 resulting in a transition from a semi-open form, allowing the ligand entry, to a closed form, allowing the catalytic action (Menéndez-Arias & Álvarez, 2014). Currently, available data about the PR2 deformation upon ligand binding are very limited. A comparison of the PR2 crystallographic structures showed that PR2 in complex with different inhibitors exhibit similar fold with large structural deviations occurring in the elbow region (Tong et al., 1995). More recently, molecular dynamics simulations of PR2 in complex with darunavir (DRV) and amprenavir (APV) – i.e., two FDA-approved drugs – showed that residues near the catalytic position D25A/B present a high degree of rigidity, whereas regions around residues 17, 40, 65, and 80 show big dynamics fluctuations (Kar &

Knecht 2012; Chen et al., 2014). They also showed that the binding of these two drugs does not produce the same flap move (Chen et al., 2014). This different flexible behavior induced by these two drugs could be linked to their different effectiveness against PR2. However, these studies were performed only on PR2 in complex with DRV and APV and did not yield information about deformation induced by other ligands. Thus, a better understanding of flexible behaviors of PR2 for all PIs could help to provide new insights on PI binding and PR2 resistance against commercially available PIs and is important for the design of new fully active PIs.

Currently 19 structures of the wild-type PR2 are available in the Protein Data Bank (Berman et al., 2000). These structures are in different forms (unbound and bound) and are in complex with diverse ligands. Comparing such multiple conformations of a given target is an easy and efficient method to provide information about the target structural variability (Zoete et al., 2002; van Westen et al., 2010; Venkatakrishnan et al., 2012; Regad et al., 2017). In our previous study, we characterized the structural asymmetry of PR2 by locating positions exhibiting different local conformations in the two chains of the 19 available PR2 structures (Triki et al., 2018). To do so, we developed a method based on the structural alphabet HMM-SA (Hidden Markov Model – Structural Alphabet), a tool allowing the simplification of protein 3D structures (Camproux et al., 2004; Regad et al., 2008). Applying this tool on this large and diverse set of PR2 structures allowed us to localize (i) structural asymmetry specific to particular ligands and (ii) the one conserved across most PR2 structures. In addition, we highlighted the structural asymmetry induced by ligand binding that is important for the ligand recognition and specificity (Triki et al., 2018).

In this study, we focused our investigation on the analysis of PR2 flexibility by exploring the conformational space sampled by the 18 available PR2 structures

complexed with various ligands. The SA-conf tool (Regad et al., 2017) was used to compare local structures of each PR2 structures resulting in the location of PR2 structurally variable positions, i.e. positions exhibiting different local conformations in the 18 PR2 structures. Crossing detected PR2 structural variability with its flexibility, quantified by crystallographic B-factor values, allowed us to localize the structural variability induced by PR2 intrinsic flexibility. The study of structural variability of the PR2 binding pocket highlighted the structural variability induced by ligand binding and important for ligand recognition and the adaptation of PR2 to these ligands. We finally explored the link between structural variability and the bound ligands by building a PR2-ligand interaction network. The analysis of this network provided the first PR2 pocket annotation that allows localizing residues for which the conserved conformation is important for ligand binding, as well as residues undergoing structural changes upon ligand binding. Our results could improve the understanding of structural changes of PR2 and its adjustments to recognize and bind various inhibitors, as well as the understanding of PR2 determinants important to explain its resistance to some FDA-approved drugs.

2. Material and Methods

2.1. PR2 set composition

For this study, we used a set developed in our previous study (Triki et al., 2018), composed of the 19 crystallographic structures of wild-type PR2 available in the PDB (Berman et al., 2000). These structures have a good resolution ranging from 1.18 Å to 3 Å and present the same amino-acid sequence except eight PR2 sequences having the mutation K57L experimentally introduced to help the crystallographic process. This

originally set contains one unbound PR2 structure, i.e., not in complex with ligand (PDB code: 1HSI), and 18 bound structures. These 18 bound PR2 structures, referred to as the PR2 set, are in complex with various ligands corresponding to three FDA-approved drugs (DRV, APV, and indinavir), two molecules chemically related to DRV, three peptides (including two non-determined ones), and nine chemical molecules (Figure 1).

2.2. Quantification of PR2 structural variability using SA-conf tool

We used SA-conf tool (Regad et al., 2017) to extract structurally variable positions defined as positions exhibiting different local backbone conformations in the PR2 set, i.e. the 18 bound PR2 structures. As PR2 is a homodimer, we separately applied SA-conf on the 18 chains A and the 18 chains B of each PR2 structure (Figure 2). Supplementary Appendix S1 presents the different steps of the SA-conf software, which are briefly explained in below. Using the structural alphabet HMM-SA (Camproux et al., 2004 ; Regad et al., 2008), SA-conf simplifies the 3D structure of each 99-residue chain into sequences of 96 structural letters, where each structural letter describes the local geometry of each four-Cα fragment (Figure 2). Based on these structural-letter sequences, SA-conf quantifies the structural variability of each position by computing the number of structural letters seen at each position within the 18 structures using the Shannon entropy criterion (neq_{SL}). The higher the value of neq_{SL} is, the more structurally variable the position is, as more PR2 structures exhibit different local conformations (structural letters) at this position.

2.3. Classification of PR2 residues according to their variability or/and their flexibility

2.3.1- Classification of PR2 residues according to their structural variability

First, we classified PR2 positions according to their structural variability quantified by their neq_{SL} value (Figure 2). We identified two position classes: (i) structurally conserved positions and (ii) structurally variable positions. Structurally conserved positions are positions with a neq_{SL} value equal to 1 and they correspond to positions where the 18 PR2 structures exhibit the same conformation. Structurally variable positions have a neq_{SL} value higher than 1 and they correspond to positions where the 18 PR2 structures exhibit different conformations.

2.3.2- Classification of PR2 residues according to their flexibility

Second, we classified PR2 residues according to their flexibility quantified using B-factor values (temperature factor/atomic displacement factor) extracted from each PDB file. This B-factor value reflects the degree of isotropic smearing of electron density around its center (Drenth, 1995). We had two choices to quantify the flexibility of PR2 residues. First, we assigned to each residue the B-factor values of its Cα atom, noted $B\alpha$. This allowed simplifying PR2 structures using only their Cα atoms as in the quantification of the structural variability of PR2 set. Second, we computed the average B-factor value for each residue using all residue atoms, noted $B_{allAtoms}$. This allowed considering the flexibility of residue side-chains, which is important in the protein

deformation induced by ligand binding. Figure S1 presents the $B\alpha$ and $B_{allAtoms}$ values for each PR2 residue and shows that these two flexibility parameters are strongly correlated (Pearson correlation coefficient = 0.94). As the induced-fit deformation is important in this study, we chose to quantify the residue flexibility using $B_{allAtoms}$. First, the $B_{allAtoms}$ values of all residues in the 18 PR2 structures were computed and then normalized using Equation 1.

$$B_{norm}(i,j) = \frac{[B(i,j) - < Bj >]}{\sigma_{Bj}} \qquad \text{Equation 1}$$

$B_{norm}(i,j)$ corresponds to the normalized $B_{allAtoms}$ value of the residue *i* of the PR2 structure *j*. $B(i,j)$ is the $B_{allAtoms}$ value of the residue *i* of the PR2 structure *j*. $<Bj>$ and σ_{Bj} are the average value and the standard deviation of the $B_{allAtoms}$ values of residues in the PR2 structure *j*. Average B_{norm} values were then calculated for each PR2 position using B_{norm} values of corresponding residues in the 18 PR2 structures.

A flexible position is defined as a position with an average B_{norm} higher than 0. A rigid position is defined as a position with an average B_{norm} smaller than 0.

2.3.3. Extraction of the different types of PR2 positions/residues according to their variability and their flexibility

Finally, we mixed these two residue classifications to define four residue types:

- Type I: structurally conserved and rigid residues that have a neq_{SL} value equal to 1 and B_{norm} smaller than 0,

- Type II: structurally conserved and flexible residues that have a neq_{SL} value equal to 1 and B_{norm} higher than 0,
- Type III: structurally variable and rigid residues that have a neq_{SL} value higher than 1 and B_{norm} smaller than 0,
- Type IV: structurally variable and flexible residues that have a neq_{SL} value higher than 1 and B_{norm} higher than 0.

The structural variability observed for type IV residues is considered as resulting from their intrinsic flexibility. In contrast, structural variability of type III residues could result from induced-fit effects, such as ligand binding or different experimental conditions.

2.4. Characterization of PR2 structure

We localized PR2 pocket residues by defining the consensus pocket across the 18 structures as we previously done (Triki et al., 2018). To do so, for each bound PR2 structure, we extracted the ligand-binding pockets by determining PR2 atoms at less than 4.5 Å from the co-crystallized ligand. The consensus-pocket residues across the 18 PR2 complexes were then defined as the 38 residues involved in at least one extracted ligand-binding pocket.

The PR2 structure was also divided into eight structural and functional regions as previously described (Sadiq et al., 2004): the dimerization (1-5 + 96-99), fulcrum (10-23), catalytic (24-30), elbow (37-42), flap (43-58), cantilever (59-72), wall (80-83), and α-helix (87-95) regions (Figure 2).

To study the closeness of each PR2 region, we built the intraproteic hydrogen-bond (H-bond) network of PR2. It is based on the 212 unique intraproteic H-bonds occurring

between two PR2 residues and extracted from the 18 PR2 structures using PyMoL software and a distance threshold of 3.5 Å. This network links two PR2 residues if they establish together an H-bond in at least one PR2 structure. It was drawn using the *igraph* library of R software (Csárdi, & Nepusz, 2006).

2.5. Construction and analysis of the PR2-ligand interaction network

2.5.1. Construction of the PR2-ligand interaction network and extraction of PR2 residue communities

To study the link between the four defined PR2 residue types and ligand-binding modes, we analyzed the PR2-ligand interaction network according to these four residue types. The PR2-ligand interaction network was built based on 1451 interactions (H-bonds and non-bonded interactions) established between PR2 residues and co-crystallized ligands in the 18 PR2 complexes. These interactions were extracted using LigPlot software (Wallace et al., 1996), see Supplementary Appendix S2.

To facilitate the visualization of the network, the 725 ligand atoms extracted from the 18 bound structures were classified using a hierarchical classification according to their closeness in the 3D space. The obtained tree was cut using an optimal distance criteria of 0.925 to 216 atom clusters. Ligand atoms grouped in the same cluster were named equivalent ligand-atoms and we supposed that they establish similar interactions with PR2 (see Supplementary appendix S2). Using these data interactions were described by four arguments (i) the complex name (PDB code) from the interaction was extracted, (ii) the name ("residue number"_chain) of the PR2 residue involved in the interaction, (iii) the name ("atom number"_chain) of the ligand atom involved in the interaction, and (iv) the cluster to which the ligand atom belongs. We drew the PR2-ligand

interaction network using the igraph library of R software (Csárdi & Nepusz, 2006). A PR2 residue and a ligand-atom cluster was linked by an edge if the residue established an interaction with one ligand atom of this cluster. From this PR2-ligand interaction network, we extracted ten communities – i.e., sets of nodes strongly connected internally – using the multi-level modularity optimization algorithm (Blondel et al., 2008). In this network, a community groups PR2 pocket residues that interact with similar ligand atoms.

2.5.2- Analysis of the PR2-ligand interaction network according to the residue types.

We analyzed the PR2-ligand interaction network according to the four residue types to assess the link between residue types, the ligand structure, and the established interactions. We first studied the link between residue types and their capacity to establish interactions with ligand by comparing the distribution of the 1451 interactions in the four residue types using a Chi-squared test. Then, we focused on the link between residue types and ligand structure conservation by comparing average size of ligand-atom clusters for each residue type using a Kruskal-Wallis test. Finally, we analyzed the link between residue types and the conservation of the PR2-ligand interactions across the 18 PR2 complexes by comparing the average number of complexes exhibiting an interaction for each residue type using a Kruskal-Wallis test. For more explanations, see Supplementary Appendix S3.

3. Results & Discussion

3.1. Localization and quantification of PR2 structural variability

Local structural changes observed across the 18 available structures of bound PR2 were detected using SA-conf tool (Regad et al., 2017). SA-conf compares local conformations of each PR2 residue and computes the number of structural conformations observed in each position using the neq_{SL} parameter that quantifies its structural variability (Figure 3).

The PR2 set contains much more structurally variable positions (150 positions with neq_{SL} higher than 1 that exhibit several conformations in the PR2 structures) than conserved positions (42 positions with a neq_{SL} value equal to 1 that exhibit only one conformation in the 18 PR2 structures). A total of 34% of the observed structurally variable positions are highly variable ($neq_{SL} \geq 2$), meaning that at least two local conformations are observed at these positions in the PR2 set. Thus, despite their conserved global fold previously described (Raugi et al., 2013), the 18 PR2 structures still present structural variability in terms of local conformations. In agreement with previous results about PR2 structural asymmetry (Tong et al., 1993; Mulichak et al., 1993; Tong et al., 1995; Chen et al., 2014), SA-conf results highlight an asymmetric behavior of PR2. Indeed, we noted that chain A contains less structural conserved positions than chain B (81 versus 72), showing that chain B is more conserved than chain A.

Structurally variable positions are located along the sequence with a strong frequency in the tail, elbow, flap regions, and in the region between the cantilever and the wall (Figure 3). These results are in agreement with the flexible behavior of the regions around residues 17, 40, and 80 observed during molecular dynamics simulations

of the PR2 complexed with DRV (Kar & Knecht, 2012). However, we also identified a high frequency of variable positions in the α-helix region (90-93) that have not been detected previously. This highlights the relevance of analyzing several structures of the same target together, as allowed by SA-conf tool, to detect structural variability and changes, particularly in regular secondary structures.

3.2- Putative factors explaining structural variability of the PR2

According to the PR2 set composition, structural changes observed across the 18 structures of PR2 could be explained by several factors: the intrinsic flexibility of PR2, the binding of diverse ligands, the different experimental conditions (pH, space group, …), and crystal packing used to solve structures. To locate structurally variable positions resulting from PR2 intrinsic flexibility, we crossed the structural variability, quantified by the neq_{SL} parameter, and the flexibility, measured by B_{norm} values, of each PR2 position. This resulted in the differentiation of four position/residue types (Table 1).

A total of 66 positions are of type IV, i.e., exhibiting different local conformations in the 18 PR2 structures and a large intrinsic flexibility. The structural changes observed at these positions across the 18 PR2 structures could result from an intrinsic property of PR2 and not from ligands or experimental conditions. 33% of these positions are located in the elbow and flap regions, this is expected due to their implication in the transition from the semi-open (unbound dimer) to closed (bound dimer) conformations of PR2 (Kar & Knecht, 2012; Chen et al., 2014). They were also

located in the fulcrum and cantilever regions (Figure 3 and Table 1). The intraproteic H-bond network between PR2 residues highlights H-bond interactions between some elbow residues (37A-42A), cantilever residues (61A, 63A, 70A, and 72A), and fulcrum residues (14A, 16A, and 18A) (Figure 4). This could indicate that conformational transition in flap regions affects other PR2 regions.

To locate structural variability induced by ligand binding, we analyzed structural variability of the PR2 consensus pocket. This PR2 pocket has an average neq_{SL} of 1.61 ± 0.65, revealing a strong structural variability in this region. A total of 63% of pocket residues are structurally variable (24 positions), with 7 residues having a strong structural variability ($neq_{SL} \geq 2$, Figure 3). Five of these latter positions are in the flap region, one in the catalytic region and one in the wall region (Figure 3 and Table 1). Amongst the 38 pocket residues, 18 are of type III (structurally variable and rigid) suggesting that the different conformations occurred at these positions result from induced-fit effects. As these positions are within the pocket, we conclude that these structural changes observed across the 18 bound PR2 are directly involved by ligand binding.

In addition, a total of 66 type III positions are located outside the pocket (Table 1). Amongst of them, 10 are neighbors of pocket residues (Figure 3 and Table 1) and their structural variability can be induced directly by ligand binding. Others are located all along the sequence, particularly in the fulcrum, flap, cantilever, and α-helix regions (Figure 3 and Table 1). The observed structural changes at these positions across the 18 PR2 structures could result from crystal packing, different experimental conditions or indirect effects of ligand binding through cooperative moves. Figure 4 highlights a network of intraproteic H-bonds occurring between the pocket residues (23B, 29A, 47B, and 49A) and flap residues (52A and 54B), α-helix residue 88A, and wall residue (83B).

This corroborates the putative structural-changes induced by ligand binding at these positions. Thus, the structural deformation of pocket residues caused by ligand binding may be accompanied by changes in other regions underlying cooperation between these regions upon ligand binding. Our observations reinforce with a larger dataset and a new approach, the cooperation in motions previously observed using molecular dynamics simulations of PR2 complexed with APV and DRV (Kar & Knecht, 2012).

3.3. Localization of structurally conserved residues

The PR2 set contains 42 structurally conserved positions – i.e., positions with a neq_{SL} value equal to 1 – exhibiting only one local conformation in the 18 structures whatever their diversity (different crystal space groups and ligands). According to B_{norm} values, 13 of these conserved positions are flexible (type II residues, Figure 3 and Table 1). Three of these positions have conformations encoded by one of the structural letters exhibiting the largest structural diversity (F, R, and U) (Camproux et al., 2004). Thus, the structural conservation of these positions may be a methodological artefact. The remaining 29 structurally conserved positions correspond to rigid residues; they were characterized as type I residues. Amongst them, 11 are located in the binding pocket, including five in the catalytic region (Figure 3 and Table 1). These positions are structurally conserved in all bound PR2 structures regardless of the co-crystallized ligands. In addition, the structural-letter map of these pockets (Figure 5) showed that these positions exhibit the same local conformation in the unbound PR2 structure. This underlies the important role of these particular local conformations in PR2 structure and function. The structural conservation of these catalytic residues (27B, 28A/B, and 30A/B) is in agreement with the high degree of rigidity previously observed near the

catalytic residues D25A/B (Chen et al., 2014). Other type I residues, which are also conserved in the unbound PR2 structures (data not shown), are observed all along the PR2 structure: in the flap, fulcrum, cantilever, and α-helix regions (Figure 3 and Table 1). Figure 4 highlights that some type I residues established H-bond with binding-pocket residues, such as α-helix residue 87A/B, residue 33B, and residue 85B that interact with pocket residues 28A/B, 76B, and 33B, respectively. These results suggest that the conserved conformation of these residues is also of importance for the maintenance of the PR2 catalytic-site structure.

3.4 –Relationship between binding pocket residue types and ligand diversity and binding mode

We analyzed the link between (i) PR2 variability and ligand structure and (ii) between PR2 variability and ligand binding mode. To do so, we first classified the 725 atoms of the 18 co-crystallized ligands into 216 atom clusters based on their 3D coordinates. We then built the PR2-ligand interaction network based on the 1451 PR2-ligand interactions extracted from the 18 PR2-ligand complexes (Material & Method). This network linked a PR2 residue (square node) to a ligand-atom cluster (circle node) if the PR2 residue establishes an interaction with a ligand atom belonging to the cluster (Figure 6).

The obtained network has a modularity value of 0.62 indicating the strength of division into communities, i.e. PR2 residues and ligand-atom clusters sharing similar interactions. As expected, this network contains most of the 38 ligand-binding pocket residues: only five residues do not establish interaction with at least one ligand (Figure

6). To analyze the link between residue type and interactions, we determined the total number of interactions established by each residue type in the network (Table 2).

By taking into account the number of residues in each type, we observed that the four residue types do not establish the same number of interactions: variable residues (type II and type IV) are involved in more interactions than conserved residues but they do not establish more interactions with ligand atoms on average (Table 2). No link was observed neither between residue types and size of cluster of atoms involved in interactions with these residues (Table 2) nor between the residue types and the conservation of interaction across the 18 complexes (Table 2). This indicates that regions conserved across the 18 ligands bound the four residue types and not only PR2 residues with a conserved conformation, i.e., structurally conserved residues (type I and type II residues).

3.5 –Relationship between binding pocket residues, their variability, and ligand interactions

In the PR2-ligand interaction network, we identified ten communities, named G1 to G10, with three singletons (G7, G8, and G10) using the multi-level modularity optimization algorithm (Blondel et al., 2008) (Figure 7A).

Each community corresponds to a set of nodes strongly connected internally. The composition of communities in terms of residue types is different. For example,

communities G2, G3, G5, G7, G8, and G10 contain only rigid residues suggesting that these regions are important for ligand binding and PR2 activity. Figure 7B shows that communities G3 and G5 residues form the pocket floor, communities G2 and G10 correspond to the right side of the pocket entrance, community G9 constitutes the left wall of the pocket, and the community G1 forms a part of the pocket tip. Community G3 residues interact with ligand regions densely populated – i.e., corresponding to larger atom-clusters – and establish most of interactions strongly conserved across the 18 PR2 complexes (Figure 7A). This means that the community G3 residues, corresponding to catalytic residues, bind conserved regions across the 18 ligands with similar interactions. In contrast, communities G4 and G6 residues establish very few interactions with ligands. Most community G1 residues are flexible (residue of type II and IV) and correspond to flap residues. These residues interact with conserved ligand regions through many conserved interactions (Figure 7A). Thus, we conclude that these residues and their flexibility are important for the ligand interaction. Communities G2, G5, and G9 residues establish many interactions with less conserved ligand regions than community G3. Thus, we conclude that these three regions are important for ligand recognition and for the specificity of ligand interactions. For example, the three PR2 structures binding molecules chemically related to the 4-hydroxycoumarin (PDB codes: 3UPJ, 5UPJ, and 6UPJ (Thaisrivongs et al., 1995))) exhibit a particular structural letter at position 84A (community G9 residue of type III) compared to other PR2 structures (Figure 5). This suggests that the specific local conformation observed at position 84A results from backbone deformation specifically induced by the binding of molecules chemically related to the 4-hydroxycoumarin. In addition, the pocket extracted from PR2-APV and PR2-DRV complexes – i.e., two PR2 complexed with very similar FDA-approved drugs but presenting totally different effectiveness against PR2 (with a Ki

value of 4.4 and 0.17 nM, respectively (Tong et al., 1993; Brower et al., 2008) – have pockets with similar structural-letter profiles. The differences between these two pockets are located at structurally variable positions 47A (community G9 residue of type IV) and 48A (community G2 residue of type III) and at two type III residues: 23B (community G3) and 31A (Figure 5). These results suggest that the binding of DRV and APV causes different backbone deformations at these four positions that could lead to the modification of interaction networks. In consequences, they provide a new explanation for their differential action on PR2. This is confirmed by the fact that residue 48A backbone establishes two not bounded interactions with DRV but only one with APV.

4. Conclusion

In this study, we provided for the first time a large and robust description and characterization of the PR2 structural variability. Using SA-conf tool, we detected the structural variability in a set containing the 18 available PR2 crystallographic structures complexed with various ligands and presenting different experimental conditions (X-ray space group, and resolution). Our study demonstrated that PR2 presents a large structural variability with 66% of its positions characterized as structurally variable. These structurally variable positions are observed all along the PR2 structures, mainly in the tail, flap, and α-helix regions.

Moreover, our results confirmed that the PR2 pocket is composed of three types of residues: (i) residues having a well-defined conserved conformation that are important for ligand binding and catalytic function, (ii) those that adjust their backbone conformation in response to ligand binding that are important for ligand recognition,

and (iii) those having intrinsic flexibility that could be important for the pocket opening and closing. The analysis of the PR2-ligand interaction patterns allowed us to characterize the ligand-binding site. We showed that the pocket floor, which contains the catalytic region, is the region that establishes the most interactions with similar regions of the diverse ligands. In addition, the left side and entrance of the pocket is important for ligand recognition and for the specificity of ligand interactions. Our results suggested that the different drug effectiveness observed for the various PIs against PR2 could be partially explained by different induced backbone deformations in the pocket. For example, we suggested that DRV and APV binding do not cause similar backbone deformations at 31A, 48A, and 23B positions. This could modify their binding mode and partially explain the differences of DRV and APV effectiveness on HIV-2 (Raugi et al., 2013). Furthermore, it would be interesting to study the relationship between drug effectiveness and local conformations of PR2 positions by considering an even larger number of drugs. This could help to better understand the PR2 resistance and the optimization of new PIs.

In addition, our results suggested that the conformation of some residues of the α-helix, flap, and fulcrum region is important for maintaining the conserved structure of certain catalytic-pocket residues. We also showed that the ligand-induced deformation in the binding pocket seems to be accompanied by structural changes of residues outside the binding pocket, particularly in the α-helix, the end of flaps, and the beginning of the fulcrum regions. This underlies the cooperative movements in the PR2 structure upon ligand binding that needs to be taken into account for the comprehension of ligand binding. These results could help to develop new allosteric PIs with original mode of action. Indeed, the usual strategy consists in identifying chemical molecules binding catalytic-site to prevent substrate binding. However, to develop small molecules able to

bind positions involving pocket deformation through allosteric effects or positions important for the maintain of the catalytic-site conformation could efficiently alter PR2 substrate binding and catalytic activity.

To conclude, our results provide new insights about PR2 structural changes upon ligand binding and mechanism of PR2 ligand recognition. Understanding and taking advantage of such conformational flexibility will be important for understanding the natural resistance of PR2 to PIs as well as for the design and optimization of new PR2 inhibitors.

Supplementary Materials

S1 appendix: Presentation of SA-conf tool

S2 appendix: Construction of the PR2-ligand interaction network

S3 appendix: Analysis of the PR2-ligand interaction network

S1 Figure: Comparison between $B\alpha$ and $B_{allAtoms}$ values

Acknowledgments

We thank Yasaman Karami and Natacha Cerisier for proofreading the manuscript.

Funding

This work was supported by an ANRS Grant to B.V., D.D., A.C.C. and L.R. D.T. was supported by an ANRS fellowship.

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., & Bourne P.E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28,* 235–242.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment,* P10008.

Brower, E. T., Bacha, U. M., Kawasaki, Y., & Freire, E. (2008). Inhibition of HIV-2 protease by HIV-1 protease inhibitors in clinical use. *Chemical Biology & Drug Design, 71,* 298–305.

Camproux, A.C., Gautier, R., & Tuffery, P. (2004). A hidden markov model derivated structural alphabet for proteins. *Journal of Molecular Biology, 339,* 561–605.

Cavaco-Silva, J. Aleixo, M.J., Van Laethem, K., Faria, D., Valadas, E., Gonçalves Mde F., Gomes, P., Vandamme, A.M., Cunha, C., & Camacho, R.J.; Portuguese HIV-2 Resistance Study Group. (2013). Mutations selected in HIV-2-infected patients failing a regimen including atazanavir. *Antimicrobial Agents and Chemotherapy, 68,* 190–192.

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems, 1695,* 1-9.

Chen, J., Liang, Z., Wang, W., Yi, C., Zhang, S., & Zhang, Q. (2014). Revealing origin of decrease in potency of darunavir and amprenavir against HIV-2 relative to HIV-1 protease by molecular dynamics simulations. *Scientific Reports, 4,*6872.

Desbois, D., Roquebert B., Peytavin G., Damond F., Collin G., Bénard A., Campa P., Matheron S., Chêne G., Brun-Vézinet F., & Descamps D; French ANRS HIV-2 Cohort (ANRS CO 05 VIH-2). (2008). In vitro phenotypic susceptibility of human immunodeficiency virus type 2 clinical isolates to protease inhibitors. *Antimicrobial Agents and Chemotherapy, 52,* 1545–1548.

Drenth J. (1995). Principles of protein X-ray crystallography. *Acta Crystallographica, D51,* 248.

Kar, P., & Knecht, V.J. (2012). Origin of decrease in potency of darunavir and two related antiviral inhibitors against HIV-2 compared to HIV-1 protease. *Journal of physical chemistry B, 116,* 2605-2614.

Masse, S., Lu, X., Dekhtyar, T., Lu, L., Koev, G., Gao, F., Mo, H., Kempf, D., Bernstein, B., Hanna, G.J., & Molla, A. (2007). In vitro selection and characterization

of human immunodeficiency virus type 2 with decreased susceptibility to lopinavir. *Antimicrobial Agents and Chemotherapy, 51*, 3075–3080.

Menéndez-Arias, L., & Álvarez, M. (2014). Antiretroviral therapy and drug resistance in human immunodeficiency virus type 2 infection. *Antiviral Research, 102*, 70–86.

Mulichak, A.M., Hui, J.O., Tomasselli, A.G., Heinrikson, R.L., Curry, K.A., Tomich, C.S., Thaisrivongs, S., Sawyer, T.K., & Watenpaugh, K.D. (1993). The crystallographic structure of the protease from human immunodeficiency virus type 2 with two synthetic peptidic transition state analog inhibitors. *Journal of Biological Chemistry, 268*, 13103–13109.

Ntemgwa, M.L., Toni, T.D., Brenner, B.G., Oliveira, M., Asahchop, E.L., Moisi, D., & Wainberg, M.A. (2009). Nucleoside and nucleotide analogs select in culture for different patterns of drug resistance in human immunodeficiency virus types 1 and 2. *Antimicrobial Agents and Chemotherapy, 53*, 708–715.

Poveda, E., Briz, V., & Soriano, V. (2005). Enfuvirtide, the first fusion inhibitor to treat HIV infection. *AIDS Reviews, 7*, 139–147.

Raugi, D. N., Smith, R.A., Ba, S., Toure, M., Traore, F., Sall, F., Pan, C., Blankenship, L., Montano, A., Olson, J., Dia Badiane, N.M., Mullins, J.I., Kiviat, N.B., Hawes, S.E., Sow, P.S., & Gottlieb, G.S.; University of Washington-Dakar HIV-2 Study Group. (2013). Complex patterns of protease inhibitor resistance among antiretroviral treatment-experienced HIV-2 patients from senegal: Implications for second-line therapy. *Antimicrobial Agents and Chemotherapy, 57*, 2751–2760.

Raugi, D. N., Smith, R. A., & Gottlieb, G. S.; the University of Washington-Dakar HIV-2 Study Group. (2016). Four Amino Acid Changes in HIV-2 Protease Confer Class-Wide Sensitivity to Protease Inhibitors. *Journal of Virology, 90*, 1062–1069.

Regad, L., Guyon, F., Maupetit, J., Tuffery, P., & Camproux, A.C. (2008). A hidden Markov model applied to the protein 3D structure analysis. *Computational Statistics & Data Analysis, 52*, 3198–3207.

Regad, L., Chéron J.B., Triki, D. , Senac, C., Flatters, D., & Camproux, A.C. (2017). Exploring the potential of a structural alphabet-based tool for mining multiple target conformations and target flexibility insight. *PLoS One, 12*:e0182972.

Ren, J., Bird, L.E., Chamberlain, P.P., Stewart-Jones, G.B., Stuart, D.I., & Stammers, D.K. (2002). Structure of HIV-2 reverse transcriptase at 2.35-A resolution and the

mechanism of resistance to non-nucleoside inhibitors. *Proceedings of the National Academy of Sciences USA, 99*, 14410–14415.

Rodés, B.; Sheldon, J., Toro, C., Jiménez, V., Alvarez, M.A., & Soriano V. (2006). Susceptibility to protease inhibitors in HIV-2 primary isolates from patients failing antiretroviral therapy. *Antimicrobial Agents and Chemotherapy, 57*, 709–713.

Sadiq, S.K., & de Fabritiis, G. (2010). Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. *Proteins, 78*, 2873–2885.

Thaisrivongs, S., Watenpaugh, K.D., Howe, W.J., Tomich, P.K., Dolak, L.A. et al. (1995). Structure-Based Design of Novel HIV Protease Inhibitors: Carboxamide-Containing 4-Hydroxycoumarins and 4-Hydroxy-2-pyrones as Potent Nonpeptidic Inhibitors. *Journal of Medicinal Chemistry*, 38, 3624–3637.

Tong, L., Pav, S., Pargellis, C., Dô, F., Anderson, P. C., & Lamarre, D. (1993). Crystal structure of human immunodeficiency virus (HIV) type 2 protease in complex with a reduced amide inhibitor and comparison with HIV-1 protease structures. *Proceedings of the National Academy of Sciences USA, 90*, 8387–8391.

Tong, L., Pav, S., Mui, S., Lamarre, D., Yoakim, C., Beaulieu, P., & Anderson, P.C. (1995). Crystal structures of HIV-2 protease in complex with inhibitors containing the hydroxyethylamine dipeptide isostere. *Structure, 3*, 33–40.

Triki, D., Cano Contreras, M.E., Flatters, D., Visseaux, B., Descamps, D., Camproux, A.C., & Regad, L. (2018). Analysis of the HIV-2 protease's adaptation to various ligands: characterization of backbone asymmetry using a structural alphabet. *Scientific Reports, 8*:710.

van Westen, G.J., Wegner, J.K., Bender, A., Ijzerman, A.P., & van Vlijmen, H.W. (2010). Mining protein dynamics from sets of crystal structures using "consensus structures". *Protein Sciences, 19*, 742-52.

Venkatakrishnan, B., Palii, M.L., Agbandje-McKenna, M., & McKenna, R. (2012). Mining the protein data bank to differentiate error from structural variation in clustered static structures: an examination of HIV protease. *Viruses, 4*,48-62.

Visseaux, B., Damond, F., Matheron, S., Descamps, D., & Charpentier, C. (2016). Hiv-2 molecular epidemiology. *Infection, Genetics and Evolution, 46*, 233–240.

Wallace, A.C., Laskowski, R.A., & Thornton, J.M. (1996). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering, 8*, 127-134.

Zoete, V., Michielin, O., & Karplus, M. (2002). Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *Journal of Molecular Biology, 315*, 21-52.

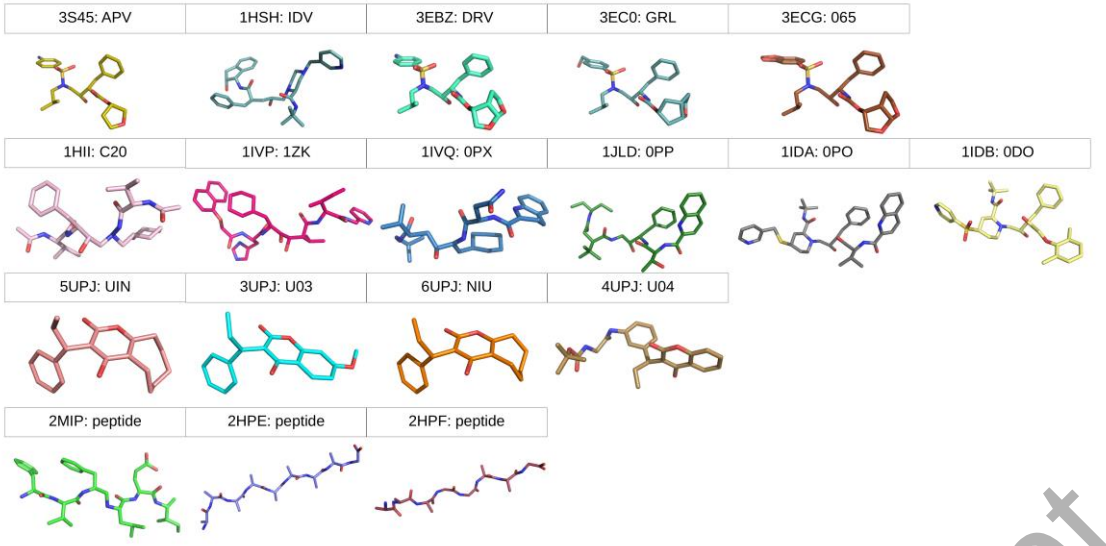

Figure 1: Description of the ligands bound to the18 PR2 structures. For each co-crystallized ligand, it is provided the PDB code of the structure where the ligand was extracted, the name HETAM code of the ligand and its 3D representation.

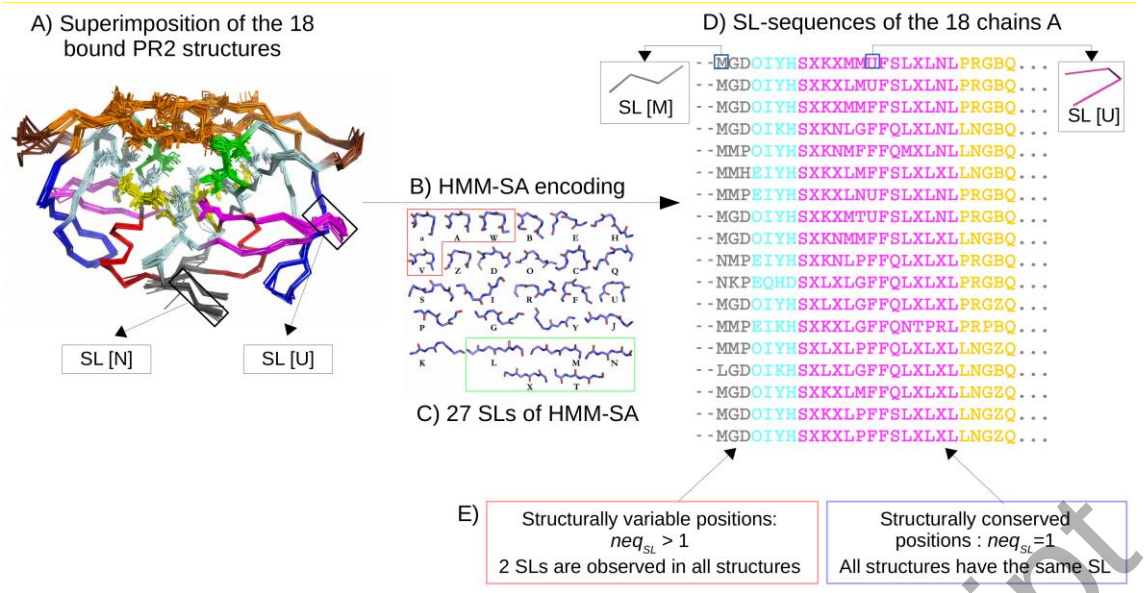

Figure 2: Detection of structural variability in the PR2 set using HMM-SA (Camproux et al., 2004 ; Regad et al., 2008). A) Superimposition of the 18 bound PR2 structures onto the unbound PR2 (PDB code: 1HSI). Proteins are displayed in ribbon mode and colored according to structural regions: the Nter and Cter regions are colored in grey, the fulcrum region in magenta, the catalytic site region in yellow, the elbow region in brown, the flap region in orange, the cantilever region in blue, the wall region in green, and the α-helix in red. Pocket residues are shown in lines. B) HMM-SA was used to simplify the 3D structure of each PR2 chain into sequence of structural letters. C) Presentation of the geometry of the 27 structural letters (SL) of HMM-SA. The four structural letters specific to helices are framed in red and the five structural letters specific to β-strands are framed by a blue rectangle. D) Part of the 18 structural-letter sequences corresponding to the 3D structures of the 18 PR2 chains. Each structural letter represents the geometry of a four-Cα fragment. Structural letters are colored according to the PR2 structural regions. E) From the set of 18 structural-letter sequences, structural variable and conserved positions were located.

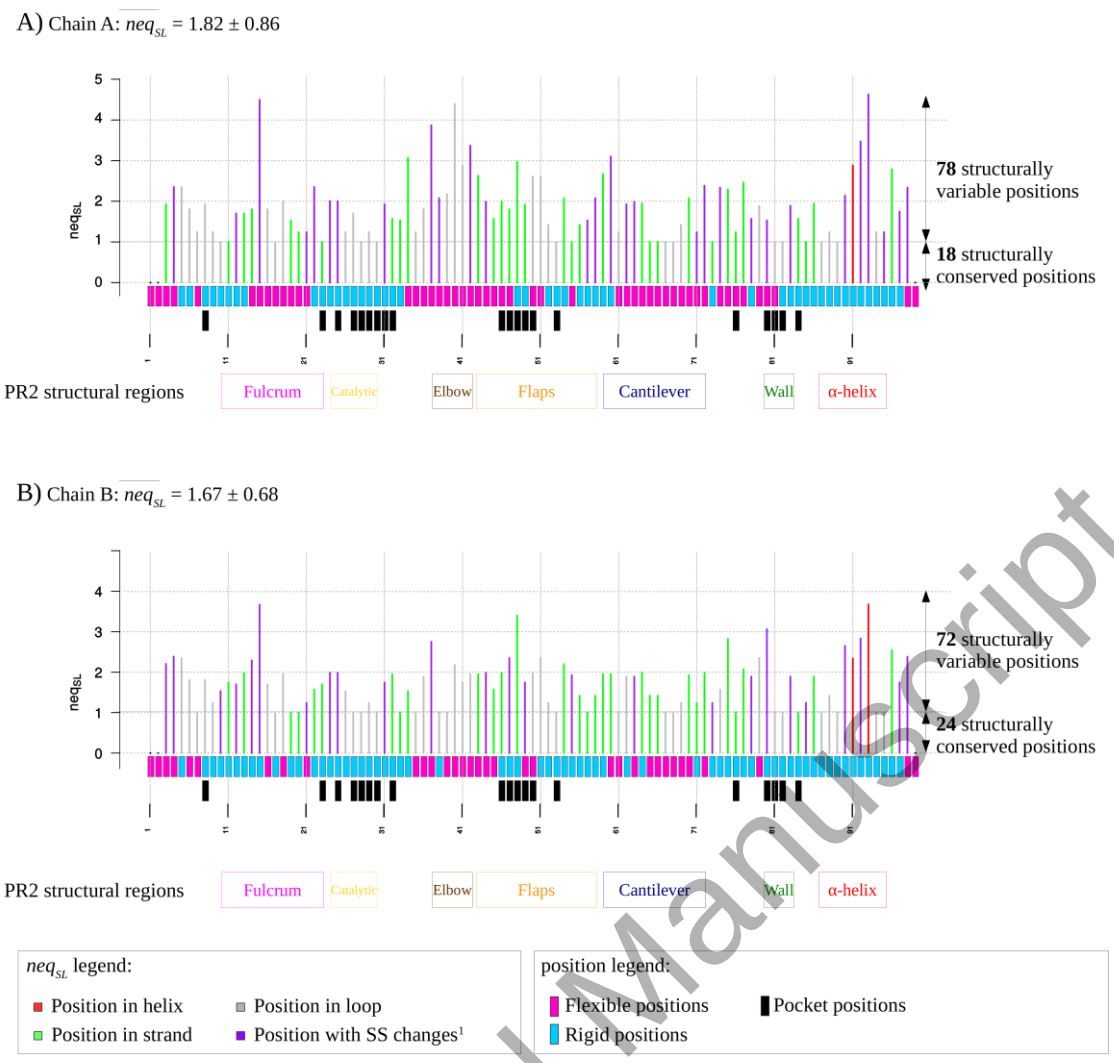

Figure 3: Structural variability of all positions of PR2 chains A and B quantified by the neq_{SL} parameter. Each neq_{SL} value is colored according to the secondary structure (SS) statue. The first line of rectangles indicates the position flexibility quantified by the B_{norm} parameter. The second line localizes situates pocket residues.

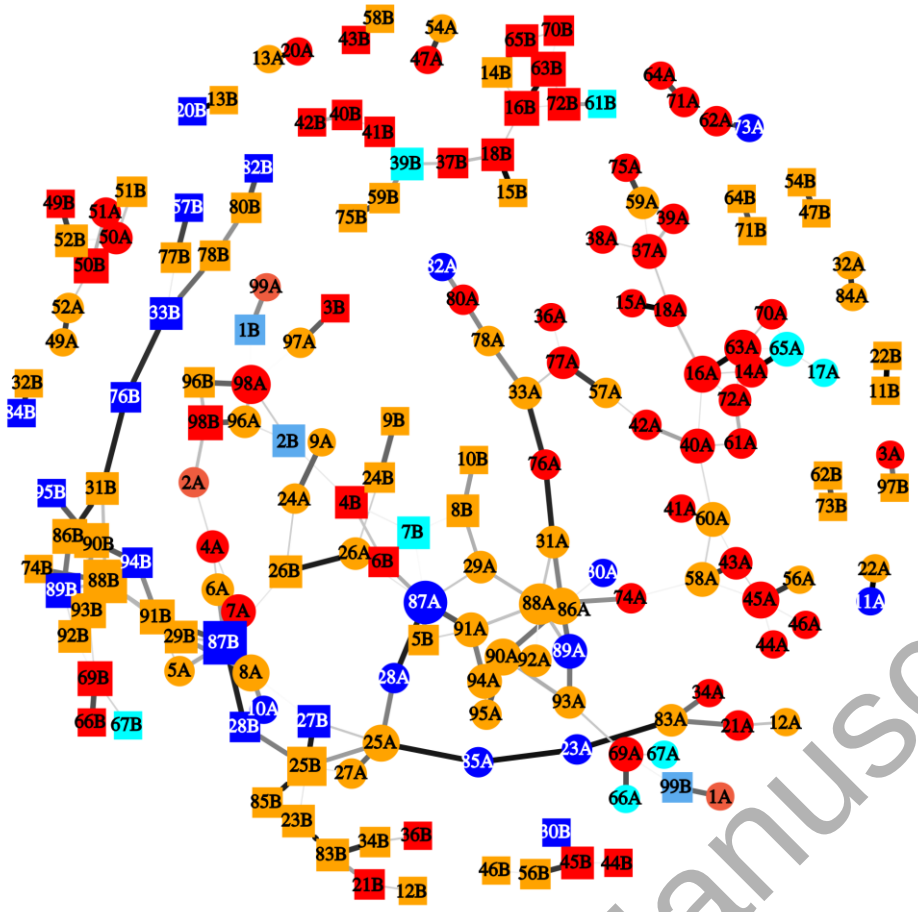

Figure 4: The intraproteic H-bond network of PR2. H-bonds were extracted from each PR2 structure using PyMoL software. Circle nodes represent PR2 residues of chain A, and square nodes represent PR2 residues of chain B. Nodes are colored according to residue types: type I residues (structurally conserved and rigid residues) in blue, type II residues (structurally conserved and flexible residues) in cyan, type III residues (structurally variable and rigid residues) in orange, and type IV residues (structurally variable and flexible) in red. Light blue nodes correspond to residues 1, 2, and 99 of both chains, for which structural letter has not been defined. Node size is proportional to the number of intraproteic H-bonds established by the corresponding residue. Edge thickness and color are proportional to the number of structures exhibiting the interactions represented by the edge.

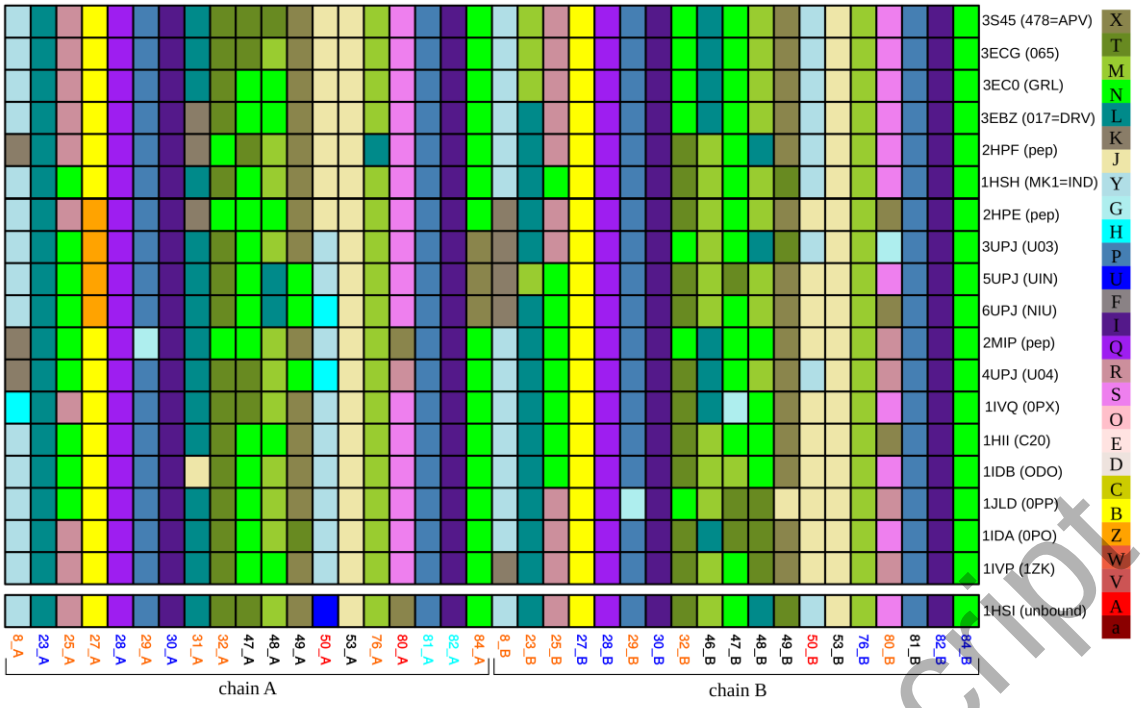

Figure 5: Structural-letter map of pocket residues in the 18 PR2 structures. In this structural-letter map, the structural-letter sequences of each pocket are presented in rows, and positions are presented in columns. Positions are colored according to the 27 structural letters.

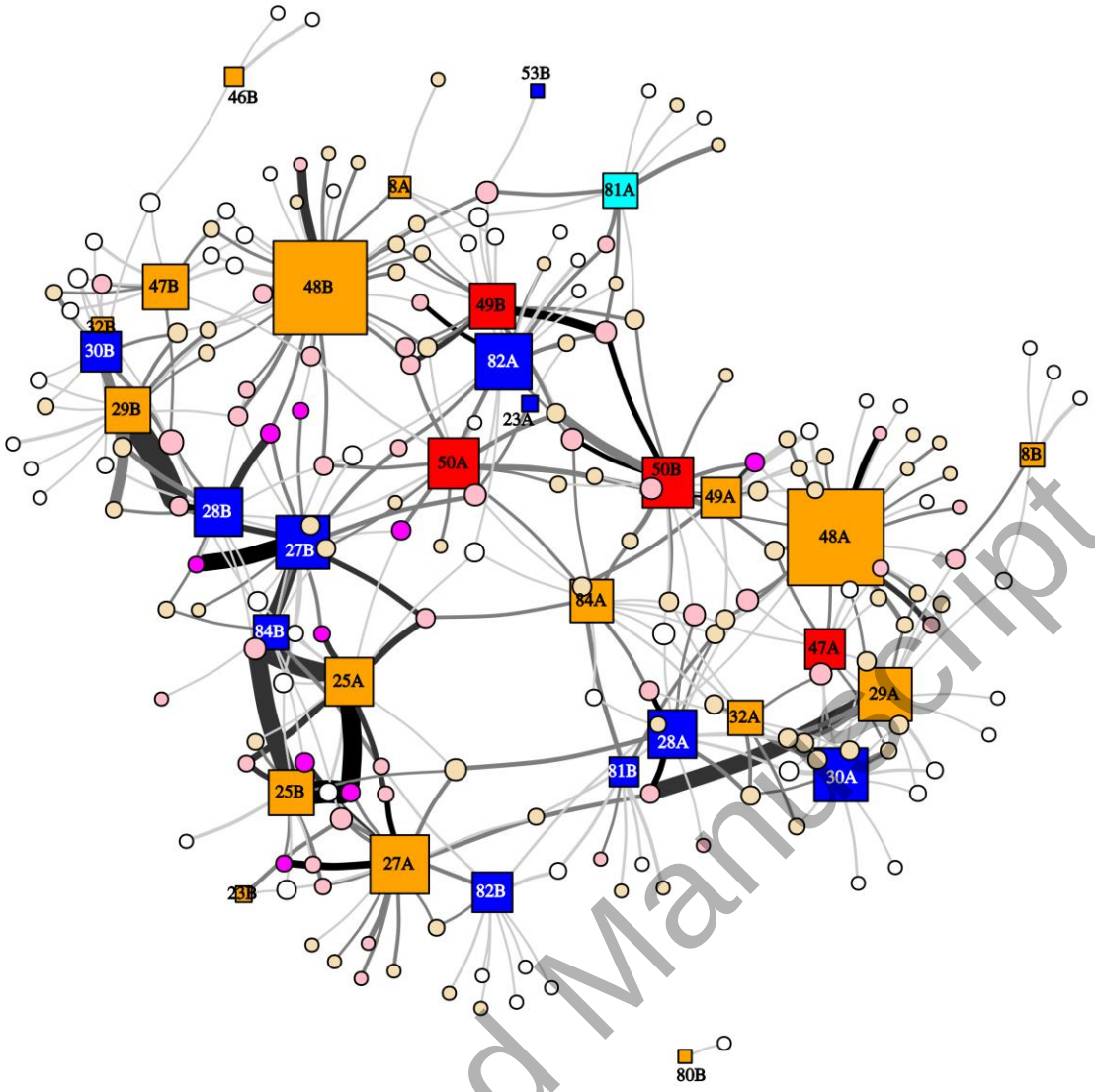

Figure 6: The network of PR2-ligand interactions. H-bonds and non-bonded interactions between PR2 residues and ligands were extracted from each PR2 structure using LigPlot software (Wallace et al., 1996). Square nodes represent PR2 residues and circle nodes represent ligand-atom clusters grouping ligand atoms according to their 3D closeness. Square nodes are colored according to residue types: type I residues in blue, type II residues in cyan, type III residues in orange, and type IV residues in red. Circle nodes are colored according to the number of ligand atoms contained in the ligand-atom clusters: clusters containing only one ligand-atom in white, clusters containing two to four ligand-atoms in wheat, clusters containing five to eleven ligand-atoms in pink, and clusters containing at least twelve ligand-atoms in magenta. The color of circle node informs about the conservation of ligand regions symbolized by the nodes across the 18

ligands. An edge represents an interaction (H-bonds or non-bonded interactions) between a PR2 residue and a ligand atom belonging to a cluster. Edge thickness is proportional to the number of structures exhibiting the interaction.

A)

communities	PR2 residues	atom cluster size linked to residues belonging to the community	Total number of Interactions established by residues and atoms of cluster in the community	average number of Interactions established by residues and atoms of cluster in the community	Interaction conservation
G1	48B, 49B, 50A, 53B, 81A	4.22 (3.03)	214	42.8 (31.87)	2.13 (1.44)
G2	48A, 49A, 81B	3.44 (2.94)	153	51 (35.55)	2.08 (1.98)
G3	23B, 25A, 25B, 27A, 27B	6.54 (4.92)	439	87.8 (48.69)	4.08 (3.69)
G4	8A, 23A, 82A	3.87 (3.31)	60	20 (23.52)	2.13 (1.55)
G5	28B, 29B, 84B	4.97 (2.42)	165	55 (40.58)	2.17 (2.06)
G6	30B, 32B, 46B, 47B	2.42 (2.08)	74	18.5 (16.62)	1.39 (0.72)
G7	80B	1	1	1	1
G8	8B	1.80	11	11	1.4 (0.89)
G9	28A, 29A, 30A, 32A, 47A, 50B, 84A	3.54 (2.43)	312	44.57 (25.96)	1.87 (1.43)
G10	82B	2.45 (1.63)	22	22	1.36 (0.92)
∉	31A, 53A, 76A, 76B, 80A	-	-	-	-

B)

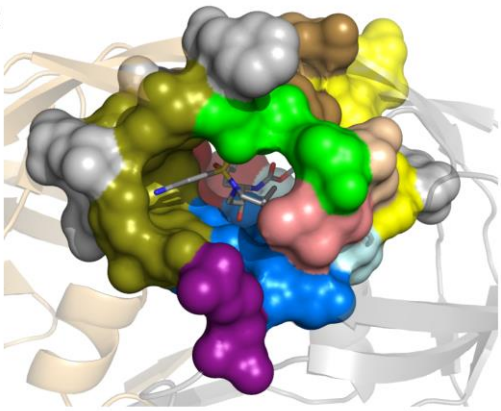

Figure 7: A) Description of the 10 communities extracted from the PR2-ligand interaction network using the multi-level modularity optimization algorithm (Blondel et al., 2008). B) Representation of the consensus PR2 pocket colored according to the communities from the network of the PR2-ligand interactions. PR2 protein (PDB code: 3S45) is displayed in cartoons, the bound ligand (APV) in sticks, and the pocket in surface mode.

Table 1: Classification of PR2 residues according to their structural variability quantified by the neq_{SL} parameter and their flexibility quantified by the B_{norm} parameter. Underlined positions highlights pocket residues.

Residue types	neq_{SL}	B_{norm}	Number of positions	Positions
Type I: Structurally conserved and rigid positions	$neq_{SL} = 1$	$B_{norm} < 0$	24	• **fulcrum** (6): 10A, 11A, 17B, 19B, 20B, <u>23A</u>, • **catalytic** (5): <u>27B</u>, <u>28A/B</u>, <u>30A/B</u>, • **elbow** (1): 38B, • **flaps** (3): <u>53A/B</u>, 57B, • **cantilever** (1): 73A, • **wall** (3): <u>81B</u>, <u>82A/B</u>, • **α-helix** (6): 87A, 89A/B, 87B, 94B, 95B, • **other positions** (3): 85A, 33B, <u>76B</u>, <u>84B</u>
Type II: Structurally conserved and flexible positions		$B_{norm} > 0$	15	• **fulcrum** (1): 17A • **elbow** (1): 39B, • **flaps** (1): 55A, • **cantilever** (7): 61B, 65A, 66A, 67A/B, 68A/B, • **wall** (1): <u>81A</u>,

				• **other position**s (2): 7B, 35B,
Type III: Structurally variable and rigid positions	$neq_{SL} > 1$	$B_{norm} < 0$	84	• **dimer** (6): 5A/B, 96A/B, 97A/B, • **fulcrum** (11): 10B, 11B, 12A/B, 13A/B, 14B, 15B, 22A/B, 23B, • **catalytic** (9): 24A/B, 25A/B, 26A/B, 27A, 29A/B, • **flaps** (16): 46B, 47B, 48A/B, 49A, 51B, 52A/B, 54A/B, 55B, 56A/B, 57A, 58A/B, • **cantilever** (9): 59A/B, 60A, 62B, 64B, 71B, 73B, 74B, 75B, • **α-helix** (12): 88A/B, 90A/B, 91A/B, 92A/B, 93A/B, 94A, 95A, • **wall** (3): 80B, 83A/B, • **other positions** (18): 6A, 8A/B, 9A/B, 31A/B, 32A/B, 33A, 34B, 77B

				78A/B, 84A, 85B, 86A/B,
Type IV: Structurally variable and flexible positions		$B_{norm} > 0$	69	• **dimer** (6): 3A/B, 4A/B,98A/B, • **fulcrum** (10): 14A, 15A, 16A/B, 18A/B, 19A, 20A, 21A/B, • **elbow** (10): 37A/B, 38A, 39A, 40A/B, 41A/B, 42A/B, • **flaps** (12): 43A/B, 44A/B, 45A/B, 46A, 47A, 49B, 50A/B, 51A • **cantilever** (17): 60B, 61A, 62A, 63A/B, 64A, 65B, 66B, 69A/B, 70A/B, 71A, 72A /B, 74A, 75A, • **wall** (1): 80A, • **other positions** (10): 6B, 7A, 34A, 35A, 36A/B, 76A, 77A, 79A/B

Table 2: Link between residue types and ligand structures and PR2-ligand interactions presented in Figure 6. PR2-ligand interactions. Standard deviation values of average values were reported in brackets. The "*" symbol indicates the significant p-value, i.e. higher than 0.05.

		p-value	Type I residues	Type II residues	Type III residues	Type IV residues
Number of positions in the PR2-ligand interaction network			11	1	17	4
Link between residue types and the capacity to establish interaction with ligands	Total number of residue-atom interactions established by each residue type	9e-04*	432	28	808	183
		2e-04*	460		991	
	Average number of residue-atom interactions established by each residue type	0.91	3.66 (4.64)	3.11 (2.20)	4.01 (6.82)	3.39 (3.28)
		0.637	3.62 (4.51)		3.88 (6.24)	
Link between residue types and ligand structure	Average size of ligand-atom clusters involved linked to residues of	0.35	4.25 (3.62)	3.11 (2.15)	4.11 (3.63)	4.24 (2.70)
		0.94	4.17 (3.54)		4.14 (3.45)	

conservation	each type					
Link between residue types and the conservation of the PR2-ligand interactions	Average number of PR2-ligand complexes exhibiting each interaction established by each residue types	0.74	2.15 (1.94)	2.11 (1.36)	2.41 (2.52)	2.17 (1.53)
		0.35	2.36 (1.90)		2.12 (2.32)	