On Defining the Dynamics of Hydrophobic Patches on Protein Surfaces

Philip Lijnzaad∗†, K. Anton Feenstra‡, Jaap Heringa‡, Frank C.P. Holstege†

October 31, 2007

Abstract

We present a simple and efficient method called PATCHTRACK, for studying the dynamics of hydrophobic surface patches. It tracks the patches on snapshot structures taken from a Molecular Dynamics simulation. They are connected into so-called patch runs, which are subsequently clustered into so-called recurrent patches. The method is applied to simulations of three different proteins. Protein motion causes addition and removal of one or more atoms to a patch, resulting in size fluctuations of around 25%. The fluctuations eventually lead to the break-up of a patch, and their average life span is therefore remarkably short at around 4 ps. However, some patch runs are much more stable, lasting hundreds of picoseconds. One such case is the largest patch in amicyanin that is known to be biologically relevant. Another case, previously not reported, is found

∗Corresponding author
†Department of Physiological Chemistry, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands
‡Centre for Integrative Bioinformatics VU (IBIVU), Vrije Universiteit, De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
in phospholipase A$_2$, where the functional significance of a large recurrent patch formed by Leu58 and Phe94 seems likely. This patch appears to have been overlooked as it is relatively small in the X-ray structure, demonstrating the utility of the current method. The most frequently occurring patch size is 40-60 Å$^2$, but sizes of up to 500 Å$^2$ are also observed. There is no clear relation between patch run durations and their average size. However, long-lasting patch runs tend not to have large fluctuations. The recurrent patches have alternating periods of “liveness” and “dormancy”; around 25% of them is predominantly in the live state.

**Introduction**

Hydrophobic patches on protein surfaces are crucial in protein-protein binding,$^{1,2}$ protein-ligand binding,$^{3,4}$ and protein folding.$^{5,6}$ Hydrophobic interactions are often functionally significant, but they may also be inadvertent: a number of pathologies are explained by such hydrophobic aggregation.$^7$

It has long been realized that the mobility of parts or all of the protein structure is essential in protein function.$^{8,9}$ Given the importance of protein structural dynamics, it is interesting to probe the dynamics of hydrophobic patches using Molecular Dynamics simulations. However, the dynamics of hydrophobic patches have, to our knowledge, never been defined well, nor studied in detail.

In order to explore the dynamics of patches, a definition of hydrophobic patches is required. We use the QUILT method proposed in earlier work,$^{10}$ because it is, to our knowledge, the only method that yields well defined hydrophobic patches in atomic detail.

The current work describes, firstly, how the QUILT patches can be used to study the patch dynamics. Secondly, their general trends in simulations of proteins are established, to serve as a baseline against which to judge the behaviour of hydrophobic patches in the in-depth studies of protein-ligand dynamics.
Methods

Protein structures

We chose three different proteins, spanning a reasonable range of sizes and architectures: the immunoglobulin G-binding domain B1 of streptococcal protein G (PDB code 1pgb); amicyanin (PDB code: 1aaj), and the closed, inactive form of phospholipase A$_2$ (PDB code: 1p2p). Of the three structures, only amicyanin (1aaj) has a known functional hydrophobic patch, which is involved in binding to methylamine dehydrogenase during electron transfer. For brevity, the proteins are referred to by their PDB accession code in the remainder of the text. The protein structures used in this work were taken from the Protein Data Bank. Some essential details of these structures are given in Table 1.

Simulations

All simulations were carried out with the GROMACS package. The simulation protocol was as follows: atomic coordinates were extracted from the PDB file, and a molecular topology was generated. Periodic boundary conditions were employed, using a rhombic dodecahedron cell with a separation of at least 8 Å between protein atoms and the nearest side of the cell. Crystallographically resolved water molecules were kept, and the system was immersed in a box of pre-equilibrated SPC water molecules at 300 K. For both 1pgb and 1aaj, the system had a net charge of 4e$^-$, so in both cases, 4 water molecules with high electrostatic potential were replaced by Na$^+$ ions. The half-occupied Ca$^{2+}$ site in the 1p2p structure was left out to achieve electric neutrality. The 53A6 forcefield was used.

Bond lengths were constrained with the Lincs algorithm. The time step of the integrator was 2 fs. Neighbour lists for the calculation of non-bonded interactions were recalculated every 10 time steps, and a twin-range cut-off radius was used for evaluating them. Within 9 Å, they were calculated every time step, whereas those within 12 Å were only evaluated every 10 time steps.
Temperature was kept constant at 300 K by coupling to a thermal bath with coupling constant $\tau_T=0.1$ ps; pressure was kept constant by coupling with constant $\tau_P=0.5$ ps to atmospheric pressure, $10^5$ Pa.

The system’s energy was minimized until convergence to machine precision using the steepest descent optimizer, and the water was subsequently equilibrated by restraining the protein atoms to their initial positions while integrating the system for 10 ps at 300 K. All restraints were removed, and simulation was continued. After 20 ps of Molecular Dynamics integration, equilibrium was deemed to have been achieved by visual inspection of the potential energy terms plotted against time (protein-protein, protein-solvent as well as solvent-solvent), as well as of the radius of gyration. The runs were resumed to obtain a trajectory of 300 ps, and from these trajectories, snapshots taken at 0.1 ps intervals were used in the subsequent analyses.

**Patch detection**

The analysis of the dynamics of the patches requires the following ingredients: the detection of patches on each static structure using QUILT; tracking each QUILT patch over time, by assembling them into so-called patch runs; and lastly, assembling patch runs into so-called recurrent patches.

A brief overview of QUILT$^{10}$ is as follows. The solvent accessible surface of a protein is considered a contiguous surface of adjacent hydrophobic (carbon and sulfur) atoms, surrounded by strings of hydrophilic (nitrogen and oxygen) atoms. Most proteins surfaces are relatively apolar (around 60% by solvent accessible surface area), so the connection of neighbouring solvent-accessible parts of hydrophobic atoms would result in one large surface patch spanning the entire protein. This surface is dotted with polar “islands” formed by the hydrophilic atoms, with hydrophobic connections through variously sized “channels” between these islands. To delineate the hydrophobic patches, the channels are closed off by temporarily expanding all solvent-accessible polar atoms by a fixed “polar expansion radius”. Thus, the surface is divided into isolated...
proper patches. They are enumerated, and adjacent surface area lost due to the polar expansion is added back.

For QUILT, the default probe and polar expansion radius (1.4 Å) were used.

**Tracking patches over time**

Given QUILT’s patch definition, each particular patch changes over time: it changes size and shape, it may loose or gain atoms (typically at the edges), it may split or merge; it may even disappear and re-emerge altogether. That is, a patch is a dynamic entity with varying constituents. To study the dynamic behaviour of patches, the terminology has to be made more precise. In the current context, a QUILT patch is the occurrence of a patch on one snapshot of a structure taken from the Molecular Dynamics trajectory. We will call this a *snapshot patch*.

A snapshot patch may be followed across subsequent time frames of a trajectory. We will call one such run of connected snapshot patches a *patch run*. Establishing them is the second step of the procedure.

As will be discussed below, patch runs are generally limited in duration, as protein motion causes changes in their size, eventually leading to their breakup. Although patch runs do not last very long, they often re-emerge a while after their disappearance. We will call such a series of patch runs a *recurrent patch*, and they are obtained in the last step of the procedure.

**Identification of patch runs**

The method used to assemble snapshot patches into patch runs uses no geometrical criteria. Instead, it is based solely on the identities of the atoms constituting the snapshot patches. By optimizing the matching between all QUILT patches on bothsnapshot structures, correspondences between snapshot patches on either structure are obtained. This allows the connection of each snapshot patch with its predecessor and successor, yielding a chain of snapshot patches that constitute a patch run.
To judge the matching of patches on two structures, we measure the overlap between two patches \( p \) and \( q \) on structures \( A \) and \( B \) using the so-called *Jaccard coefficient*\(^{16} \) of the two sets of atoms \( p \) and \( q \):

\[
J_{AB}(p, q) = \frac{|p \cap q|}{|p \cup q|}.
\]

That is, the overlap coefficient is defined as the number of atoms present in both snapshot patches, divided by the number of atoms in their union. \( J_{AB}(p, q) \) is 1 when \( p \) and \( q \) are identical, and 0 when they are fully disjunct.

For speed, the lists of snapshot patches \( p \) and \( q \) are each ordered by decreasing size, and for each of the snapshot patches \( p \), the overlap with all snapshot patches \( q \) is calculated. The pair \((p_i, q_j)\) for which \( J_{AB}(p_i, q_j) \) is maximum, is declared a match; \( p_i \) and \( q_j \) are removed from both lists, and the process is reiterated until either list of \( p_i \)'s or \( q_j \)'s is empty (i.e., this is a greedy algorithm).

To avoid poor matches causing spurious matchings, a cut-off value \( J_{rej} \) is applied below which a match \((p_i, q_j)\) is rejected, even if it has the maximum overlap coefficient between \( p_i \) and any from the list of \( q \) patches. Without a cut-off, a poor \( J_{AB}(p_m, q_j) \) early in the iteration over the \( p \)'s occasionally rules out a better \( J_{AB}(p_m, q_j) \) later in the iteration. Increasing the \( J_{rej} \) makes the matching process less greedy. This improves the overlap coefficient for matched snapshot patches, but it reduces the overall matching because the number of snapshot patches that fail to be matched altogether also rises. A \( J_{rej} \) between 3% and 12% was determined to be optimal (data not shown); we used 5% in this work.

**Identification of recurrent patches**

In the last step, patch runs are assembled into recurrent patches. As with the patch run identification described above, only atom identities are used. To this end, the set of “core atoms” of a patch run has to be defined. The obvious choice would be the set of atoms that is present in all snapshot patches of the patch run. We will call these atoms the *fully persistent atoms* of a patch run. For brevity, the number of fully persistent atoms will be called \( N_{fpa} \). They are found by simply taking the intersection amongst all sets of atoms of all snapshot patches of a patch run. However, the resulting set of fully persistent
atoms is frequently small \((N_{fpa} \leq 1)\) as a result of fluctuations. The converse option is to use the union of the atoms in all snapshot patches of the patch run. This becomes unwieldy however, because these sets can become large. We take the middle ground, by defining the core atoms of a patch run as those that are present during at least half of its duration. The correlation of their number with the area of a patch run is greater than 0.75 in the three systems studied.

The next step relates the patch runs by clustering them based on their core atoms. The distance measure is the inverse of the formula used for the snapshot patch overlap: 
\[ D = \frac{1}{J(p,q)} = \frac{|p \cup q|}{|p \cap q|}, \]
with \(p\) and \(q\) the respective sets of core atoms of the patch runs being compared. Whenever the patch run overlap is zero, the distance was arbitrarily set to \(f_{\text{max}} = 10\) times the maximum distance found amongst all pairs of patch runs, to avoid infinite values in the distance matrix. The data is clustered using group-average linkage (UPGMA), and cutting the dendrogram at an ultrametric distance \(d_{\text{cut}} = 10\) produces tight clusters of patch runs. These clusters are the recurrent patches.

Unlike their constituent patch runs, the recurrent patches do not have a duration; they are features of the dynamical protein structure in general. They do however have alternating periods of presence and absence of the patch runs that constitute it. We will call these “states” of the recurrent patch live and dormant, respectively.

Programs and scripts (available from
http://bioinformatics.holstegelab.nl/publications/lijnzaad/patchdynamics/) were written in Perl,\(^7\) R,\(^8\) and standard Unix utilities. In addition to these scripts, use has been made of xmgr/grace,\(^9\) RasMol,\(^{10}\) PyMol (DeLano Scientific LLC), and some of the programs in the GROMACS suite.\(^{12}\)
Results

Matching procedure

The degree of matching can be defined as the fraction of patches mapped between one snapshot and its successor. We find an average, over all time frames, of around 80%. The time interval $\Delta t$ between the snapshots influences the degree of matching, as shown in Fig. 1. The data is obtained by averaging over matchings between all pairs of snapshots from a 10 ps interval from the trajectory of structure 1aaj.

Patch runs

Visual inspections of the patch runs using movies of the molecular trajectory showed that the matching behaves as expected: a snapshot patch in one frame looks largely the same as its match in the next frame, providing a smooth dynamic picture of its behaviour.

An impression of how the size of a patch evolves during the patch run is given in Fig. 2, which shows the surface area of the three longest-lasting patch runs in each simulation. The longest-lasting of all patch runs lasts around 250 ps; it is also the largest amongst all patch runs. It was found in the simulation of 1aaj, and corresponds to the hydrophobic patch known to be involved in the interaction of this protein with methylamine dehydrogenase.

As mentioned in Methods, the number of patch runs with an appreciable set of fully recurrent atoms ($N_{fpa} > 1$) is generally low (1pgb: 28%; 1aaj: 25%; 1p2p: 32%).

An important characteristic of a patch run is its “lifespan”, called its duration. Since there is little to conclude about the “real” duration of a patch run that is already existent when the simulation starts, or is still ongoing when it ends, patch runs crossing the temporal boundaries of the simulation were discarded, as were those lasting shorter than 1 ps.

We define the size of a patch run simply as the average area of its consti-
tuting snapshot patches; their standard deviation will be called its fluctuation. Statistics (e.g., average etc.) and distributions on the size and fluctuation of a patch run (used below) ignore the fact that they themselves are also defined as an average. Some statistics on the resulting patch runs are given in Table 2.

The average patch run lasts around 4 ps, although a substantial number is more persistent, exceeding 100 ps. Their mean size is around 90 Å², but the maximum patch run sizes are a few hundred square Ångstroms.

Fig. 3 shows the distributions of duration, size and fluctuation in more detail. It shows that short-lived patch runs are much more prevalent than longer-lasting ones. This trend is so strong that the only meaningful presentation of the distribution is as a log-log plot. The distributions of patch run sizes shows that the most frequently occurring size is 40-60 Å². The fluctuations depicted in Fig. 3 are shown as a relative value (standard deviation divided by mean, also known as the coefficient of variation) to avoid trends being obscured by the mass of small patches having a low absolute fluctuation. The most frequently occurring value for the relative fluctuations is around 25%.

Size, fluctuation and duration

Fig. 4 shows scatter plots of the average size of patch runs, their life spans, and fluctuations.

There is hardly any relation between duration and size. Only some mid-sized patch runs reach long life-spans; large patches do not, although there are a few exceptions. There is a roughly linear relationship between the average size of a patch run and its fluctuations. The last column of scatter plots in Fig. 4 depicts the relation between fluctuation and duration of a patch. Its correlation is even lower than that between duration and size. We also looked into the following patch characteristics:

- fraction of hydrophobic amino acids
- fraction of backbone atoms
• fraction of sequentially remote amino acids

• shape (ratio of the largest two semi-axes of a fitted ellipsoid)

None of these correlated to an appreciable extent with duration, size or fluctuation (the highest correlation observed was 0.42).

**Recurrent patches**

To obtain the recurrent patches, the patch runs are clustered and the resulting tree is cut at an ultrametric distance of 10, as described under *Methods*. Visual inspection showed that this works well. The cut-height is not critical, because the clusters are very tight, and the distance to the next clustering level is large (data not shown). This is also clear from the Silhouette width, a measure of cluster compactness. Its value is around 0.7, averaged over all clusters obtained by cutting the tree, and depends little on the choice of the cut-height. The clustering method is also not critical; group-mean average, complete, median, centroid and Ward’s method all give similarly compact clusters.

To get an impression of how patch runs appear and disappear in the recurrent patches, they are plotted in Fig. 5. It shows how recurrent patches alternate between the “dormant” and “live” states. The latter are indicated by the blocks in the figure. Most recurrent patches are dormant most of the time; few are predominantly in the live state. For many recurrent patches, some of their patch runs overlap briefly in time (data not shown). On average, such temporal overlaps occur during approximately 3% of the simulation time.

For most recurrent patches, the coefficient of variation of the size of their constituting patch runs comes out at around 25%, the same as that found for the patch runs.

We define the *liveness* of a recurrent patch as the percentage of simulation time during which it is in the live state. Statistics on the liveness of recurrent patches are given in Table 3. As mentioned, few recurrent patches are mostly live. Therefore, the table also shows the statistics for a high-liveness (≥ 50%
live) and a low-liveness (< 50% live) group.

The number of recurrent patches is somewhat larger than the number of QUILT patches in the static structure of a protein (see Table 1). The high-liveness group is 2 to 3 times smaller than the low-liveness group. The contrast in the liveness of these groups is quite stark: the high-liveness recurrent patches are 5 times as often in the live state.

Discussion

Matching procedure

The matching procedure works well, based on visual inspection and as judged from the fraction of the number of patches matched (around 80%). The snapshot patches tend to be of similar size during the course of the patch run. This is achieved by the use of the Jaccard coefficient. For instance, it classifies the containment of a small snapshot patch by a large one as poor, because the large snapshot patch still contributes to the denominator of the coefficient. As a result, the case of one snapshot patch \( p_i \) simultaneously containing two smaller patches \( q_j \) and \( q_k \) is dealt with correctly, because the larger of the two \( q \)'s will have the higher overlap coefficient.

Visual inspection of patch runs showed that the patches do not “wander” over the protein surface. This is confirmed by the clustering, into recurrent patches, of patch runs that are widely separated in time (see Fig. 5). The PATCH-TRACK algorithm would, in fact, accommodate and therefore allow such wandering of patches. The fact that this does not occur suggests that they are genuine protein surface features.

As the time interval \( \Delta t \) between two snapshot structures increases, the root mean squared deviation (RMSD) between the two structures, especially that of solvent accessible atoms, increases likewise. Consequently, the matching becomes more difficult as can be seen in Fig. 1. It shows that a sampling frequency of less than once per 0.1 ps results in a considerably lower degree of
matching. Conversely, sampling at a higher frequency should increase the degree of matching, but this is rarely used in practice.

Theoretically, following patches can be done in a geometrical fashion, for instance by clustering the coordinates of the patch centers. This was attempted initially, but it is slow due to the large amount of data, and very sensitive to the choice of parameters. A more serious drawback is that geometric matching may not be possible, or be very difficult to parameterize, if there are large scale changes in the proteins structure over the course of the simulation. PATCH-TRACK deals with this automatically by using atom identities only.

QUILT uses two parameters: the probe radius and polar expansion radius. The PATCHTRACK method introduces three more: the patch overlap cut-off \( r_{\text{ov}} \) for the matching procedure, and, for the clustering step, the maximum distance multiplication factor \( f_{\text{max}} \) and the cut-height \( d_{\text{cut}} \). Their default values should not need adjustment in the large majority of cases, as results are not sensitive to the exact settings. The matching and clustering procedures perform comparably for the three different protein structures, indicating that the method is robust.

**Patch runs**

The fluctuations in patch size can be substantial, as shown in Fig. 2. They occur at different time scales, small fluctuations lasting tenths of picoseconds, superimposed on larger fluctuations lasting up to 100 times longer. This reflects the hierarchy in the motion of atoms, side-chains and secondary structure elements.\(^{22,23}\)

Many natural entities (e.g., populations) behave gradually: they start small, grow in size steadily, persist for a while, shrink slowly, then cease existing. This is not the case for patch runs; they start and stop fairly abruptly. Although the areas of the snapshot patches at either end of the life span is a bit below average, they are by no means atypical, given the large fluctuations displayed during the patch run. However, the fluctuations do, of course, cause the end
of patch runs, in that it is matter of time until the fluctuation is so large that it destroys the patch altogether.

There are two mechanisms behind the fluctuations and the volatility of the patch run. When an atom joins a patch, it can easily add 20-30 Å² to its area. With average patch sizes of around 100 Å² (cf. Table 2), such additions (and likewise, removals) are substantial. This is also visible in the distribution of fluctuations shown in Fig. 3, which shows that a relative fluctuation of around 25% is the most commonly occurring one. Atoms can join and leave patches due to brief burials, but more importantly, due to the fact that they frequently alternate between neighbouring patches. This, in turn, is caused by small shifts in the positions of the polar atoms used by QUILT to delineate patches. It is then a matter of time until so many atoms are by chance simultaneously absent from the patch under consideration, that it has become too small to be matched properly to the corresponding snapshot patch in the next time frame. The result is the end of the patch run. The shifts of the atoms leading to reassignment happen at the sub-picosecond time scale, explaining the time scale of the fluctuations.

A second, for large patches more important mechanism generating large fluctuations is a scaled-up version of the first mechanism: the addition or removal of a complete group of atoms to the patch under consideration. Typically, the atoms in this group belong to one amino-acid side chain. An example of this is shown in Fig. 6. It depicts the solvent accessible surface of structure 1aa at time points 174.2, 174.3 and 174.4 ps. Over these 200 femtoseconds, the patch area plummets from 945, to 567, to 400 Å². This is the result of the removal of both Phe97 and Lys27 in the first time step, and Lys73 in the second time step. The overall changes in the structure are minimal, but they conspire to reduce the surface of this patch by more than half.

Looking at the set of atoms contributing to a patch run over its life time, we find similar volatility, as few patch runs have many fully persistent atoms. \( N_{fpa} \) is not correlated with the patch run size, nor with patch run duration (data not
The long-lasting patch runs shown in Fig. 2 last between 100 and 250 ps. If longer simulation times are used, the maximum duration will probably be greater, especially so for the patches already existing at \( t = 0 \) or still existing at \( t = 300 \). The long-lasting patches have sizes well above the average (cf. Table 2), but they are mostly not the largest patches. The exceptions are those in the 1aaJ (amicyanin) and 1p2p (phospholipase A2) simulations, at around 600 and 400 Å\(^2\), respectively. They are both the longest-lasting and largest patches in their simulation. The large recurrent patch on the structure of amicyanin is known to interact with methylamine dehydrogenase. The large recurrent patch on phospholipase A2 is not known to be functionally significant, but may well be so (see below). If this is indeed the case, then these two examples suggest a hypothesis for further enquiry, namely that functional patches tend to be not only large, but also persistent.

The graph for amicyanin (Fig. 2; 1aaJ) shows two smaller patch runs with an area fluctuating around 200 Å\(^2\), separated by around 10 ps. The atoms making up these patch runs are nearly identical, and this is a clear example of two patch runs that are part of one recurrent patch.

None of the patch runs exceeds our 300 ps simulation time (Fig. 2), showing that the relatively short simulation times used in this study are sufficient to capture the essentials of the dynamics of the QUILT patches.

To see how the procedure held up with larger systems and longer simulation times, we also simulated the heme binding domain of the Cytochrome P450-BM3 (pdb code 1bu7) for 1 ns, using the same parameters as the three systems described in the preceding. The longest-lasting patch run in this case was around 280 ps, not significantly longer than those found earlier.

The average patch run duration is about 4 ps, a time scale similar to that found for simulated and experimental residence times of water molecules in the first hydration shell of proteins.

There is generally no relation between the duration and size of a patch run,
as shown in Fig. 4. We conclude that the duration and size of a patch run are so dependent on the details of the local mobility that no general trend emerges.

In general, larger quantities usually display larger fluctuations, and this is also observed for patch runs and their fluctuations. It is interesting to see that this relation is roughly linear, as one would expect a patch to grow or shrink predominantly at its perimeter. The area changes must then be proportional to the length of the perimeter, and therefore proportional to the square root of the patch area. This, however, ignores the roughness of the patch perimeter. If this would be taken into account, it is likely that the perimeter as a function of area varies with an exponent closer to 1, explaining the apparent linear relationship between patch size and fluctuation.

Although there is no linear correlation between duration and fluctuation (Fig. 4), there is an obvious relationship, namely, that long-lasting patch runs rarely occur in patch runs that exhibit high fluctuations. This is not surprising: the chances of a patch run surviving for hundreds of picoseconds are slim if the magnitude of the fluctuations increases the chances of disruption.

Recurrent patches

The recurrent patches were subdivided by their average liveness, using a liveness of 50% as the dividing line. This results in a relatively small (20-30%) group of recurrent patches with strong liveness (≥ 75%; Table 3). The large remaining group is predominantly dormant, and consist of relatively small and brief patch runs caused by splits from their “main” patch. One consequence of this is that the total number of persistent patches exceeds the number of patches on the original crystal structure (Table 1). These small, mostly dormant patches are the inevitable “noise” inherent in a patch detection method based on contiguous atomic surface area, and are best ignored when focussing on recurrent patches. Because of the small number of highly live recurrent patches in the current work, we have not attempted to relate the liveness to the size or duration of the constituting patch runs.
A novel phospholipase patch

The large, persistent patch on phospholipase A_2, visible in Fig. 2, is an interesting case. In the crystal structure, it is much smaller (220 Å²) than in the simulation (400 Å²). It is formed by Leu58 at the C-terminal end of helix C, and Phe94 at the C-terminal end of helix E, at the “back” of the protein. It is far removed from the active site, and is not part of the ligand binding cleft. The patch could be a by-product of an interaction, between said residues, that stabilizes the two helices. However, this would seem unnecessary, as the helices are already tethered by a total of three disulfide bridges. The hydrophobic nature of the two residues at this position appears to be conserved across much of the Pfam alignment (accession PF00068, Phospholipase A_2). Based on this and on the fact that the patch is both large and persistent, it is likely to be biologically relevant. Phospholipases require association with membranes or micelles for their activation, but this interaction is ionic in nature and takes place at the other side of the molecule. The Leu58,Phe94 patch, then, must be important for a different reason, such as association with lipophilic moieties of membrane compounds, or the stabilization of the short D helix in the, as yet unknown, membrane-associated form of the protein. We reiterate that this patch is not the largest patch on the X-ray structure, and therefore appears to have been overlooked in previous studies. Although elucidating the function will require experimental work, the case clearly demonstrates the value of the PatchTrack approach in generating hypotheses.

Conclusion

PatchTrack is a robust, fast procedure to follow hydrophobic patches on Molecular Dynamics trajectories of proteins. It connects the snapshot patches given by QUILT into patch runs using a simple matching procedure based on atom identities. The patch runs are subsequently clustered into recurrent patches, again using only atom identities.
The patches do not “wander” over the protein surface, but the set of atoms constituting a patch run is not very stable, owing to protein mobility. Protein motion causes small fluctuations in patch size, whereas large fluctuations arise from the addition or removal of groups of atoms, typically from the same residue. The most commonly observed mean size of patch runs is 40-60 Å², but they can be much larger, up to a few hundred Å². The patch runs can last up to approximately 150 ps, but are generally much shorter; the average patch run lasts around 4 ps. Size fluctuations during a patch run are roughly proportional to the patch size. They are substantial, at around 25% of the patch size. The relation between patch run duration and the average size is not clear, but that between patch size fluctuations and their duration is: large fluctuations preclude long durations.

Clustering the patch runs into recurrent patches uncovers the really persistent patches. These are relatively few, as on average, the recurrent patches are in the live state during 25% of the time. However, the 50% most “live” patches have an average liveness of around 75%.

Amongst the persistent large patches in the systems studied, two stand out. The first is the well-known patch on amicyanin that is involved in the interaction with methylamine dehydrogenase. It has a size fluctuating around 600 Å², and lasts for 250 ps. Secondly, in phospholipase A₂, a novel patch that is probably functionally important is found. It consists of Leu58 and Phe94, measures around 400 Å² and lasts for over 220 ps.

In addition to QUILT’s probe radius and polar expansion radius, PATCHTRACK introduces three parameters: the cut-off value for patch overlaps, and the maximum distance and cut-height for clustering. These parameters are not critical, and the default settings should suffice in almost all circumstances. PATCHTRACK should prove valuable as a an aid in exploration and visualization, particularly in the context of studies involving large-scale movements. The results of the QUILT/PATCHTRACK analysis are consistent across the three proteins studied, suggesting that the trends found are general.
References


Figures and captions
Figure 1: Percentage of patches matched between two time frames as a function of the time step $\Delta t$ between them.
Figure 2: The surface area of the three longest-lasting patch runs in each simulation. Different shades of grey are used to help distinguish the graphs. PDB codes of the structures in each row are indicated on the right.
Figure 3: Distributions of patch run duration, size, and size fluctuations. The columns depict the distributions of the duration of patch runs (left), their size (middle), and their fluctuation (right). The data in the left column are binned and plotted logarithmically. Those in the middle and right columns contain linear plots, with vertical axes at the same scale. PDB codes of the structures in each row are indicated on the right.
Figure 4: Scatter plots of the patch run size and duration (left), size and fluctuation (middle) and fluctuations and duration (right). Linear correlation coefficients are indicated in each graph. PDB codes of the structures in each row are indicated on the right.
Figure 5: Recurrent patches on structure 1aa]. Each track represents one recurrent patch. Each block is one patch run, with thickness proportional to its average size. The tracks of the recurrent patches are stacked in order of their liveness; the vertical separation is 500 Å. Recurrent patches with an average area below 50 Å² or consisting of one patch run are excluded.
Figure 6: Three successive time frames from the 1aaj simulation. The figure shows the solvent accessible surface area of the complete protein. White: oxygen and nitrogen atoms; light-grey: carbon and sulfur atoms; dark-grey: atoms belonging to largest and longest-lasting patch. Time points and surface area of the snapshot patch are indicated below the snapshots.
Tables and captions
<table>
<thead>
<tr>
<th></th>
<th>protein G Ig-binding domain B1</th>
<th>Amicyanin</th>
<th>Phospholipase A$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PDB code</strong></td>
<td>1pgb</td>
<td>1aaj</td>
<td>1p2p</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
<td>32</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td><strong>Resolution [Å]</strong></td>
<td>1.92</td>
<td>1.8</td>
<td>2.6</td>
</tr>
<tr>
<td><strong>Residues</strong></td>
<td>56</td>
<td>105</td>
<td>124</td>
</tr>
<tr>
<td><strong>Heavy atoms</strong></td>
<td>436</td>
<td>807</td>
<td>971</td>
</tr>
<tr>
<td><strong>Crystallographic waters</strong></td>
<td>24</td>
<td>98</td>
<td>5</td>
</tr>
<tr>
<td><strong>Waters added</strong></td>
<td>2168</td>
<td>2937</td>
<td>5025</td>
</tr>
<tr>
<td><strong>SAS [Å$^2$]</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>X-ray</strong></td>
<td>3668</td>
<td>5454</td>
<td>7089</td>
</tr>
<tr>
<td><strong>MD$^\ddagger$</strong></td>
<td>3802 ± 72</td>
<td>5848 ± 72</td>
<td>7653 ± 103</td>
</tr>
<tr>
<td><strong>Percentage hydrophobic</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>X-ray</strong></td>
<td>58</td>
<td>62</td>
<td>60</td>
</tr>
<tr>
<td><strong>MD$^\ddagger$</strong></td>
<td>59 ± 0.9</td>
<td>61 ± 0.8</td>
<td>58 ± 0.8</td>
</tr>
<tr>
<td><strong>Number of patches</strong></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>X-ray</strong></td>
<td>13</td>
<td>21</td>
<td>31</td>
</tr>
<tr>
<td><strong>MD$^\ddagger$</strong></td>
<td>16.9 ± 1.9</td>
<td>26.9 ± 2.5</td>
<td>28.5 ± 2.5</td>
</tr>
<tr>
<td><strong>min;max</strong></td>
<td>10; 23</td>
<td>19; 35</td>
<td>19; 36</td>
</tr>
</tbody>
</table>

* for simulation purposes
†Solvent accessible surface area; probe radius 1.4 Å
‡average ± standard deviation, over 300 ps simulation
§QUILT patches larger than 10 Å$^2$

Table 1: Details of the protein structures.
<table>
<thead>
<tr>
<th>PDB code</th>
<th>1pgb</th>
<th>1aaj</th>
<th>1p2p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patch runs</td>
<td>788</td>
<td>1479</td>
<td>1365</td>
</tr>
<tr>
<td>Duration [ ps ]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>4.0</td>
<td>3.7</td>
<td>3.9</td>
</tr>
<tr>
<td>standard dev.</td>
<td>9.1</td>
<td>5.9</td>
<td>7.3</td>
</tr>
<tr>
<td>median</td>
<td>1.9</td>
<td>1.9</td>
<td>1.7</td>
</tr>
<tr>
<td>maximum</td>
<td>145</td>
<td>104</td>
<td>138</td>
</tr>
<tr>
<td>Size [ Å² ]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>97</td>
<td>89</td>
<td>108</td>
</tr>
<tr>
<td>standard dev.</td>
<td>69</td>
<td>47</td>
<td>78</td>
</tr>
<tr>
<td>minimum</td>
<td>12</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>median</td>
<td>81</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>maximum</td>
<td>484</td>
<td>320</td>
<td>426</td>
</tr>
</tbody>
</table>

Table 2: Statistics on the patch runs. Those lasting shorter than 1 ps or existing at \( t = 0 \) or at \( t = 300 \) ps are excluded.
<table>
<thead>
<tr>
<th>PDB code</th>
<th>1pgb</th>
<th>1aaj</th>
<th>1p2p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of recurrent patches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>liveness ≥ 50%</td>
<td>35</td>
<td>55</td>
<td>63</td>
</tr>
<tr>
<td>liveness &lt; 50%</td>
<td>9</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td><strong>Liveness [ % ]</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>29</td>
<td>32</td>
<td>26</td>
</tr>
<tr>
<td>≥ 50%</td>
<td>78</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>&lt; 50%*</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>minimum</td>
<td>0.4</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>maximum</td>
<td>94</td>
<td>96</td>
<td>96</td>
</tr>
</tbody>
</table>

*The subset-averages need not sum to unity as the liveness distribution is asymmetric

Table 3: Statistics of recurrent patches. Overall statistics are given, as well as those obtained by first separating the recurrent patches into a “high-liveness” (≥ 50% live) and “low-liveness” (< 50% live) group.