Analyses of displacements resulting from a point mutation in proteins

Mathilde Carpentier, Jacques Chomilier

Analyses of displacements resulting from a point mutation in proteins

Mathilde Carpentier[1*], Jacques Chomilier[2]

1. Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, 57 rue Cuvier, CP 50, 75005 Paris, France.

Tel : +33 (0)1 40 79 34 73

2. Sorbonne Université, BiBiP IMPMC UMR 7590, CNRS, MNHN, Paris, France

Corresponding authors: mathilde.carpentier@upmc.fr; jacques.chomilier@upmc.fr

Abstract

The effects of a single residue substitution on the protein backbone are frequently quite small but there are many other potential sources of structural variation for protein. We present here a methodology considering different sources of distortions in order to isolate the very effect of the mutation. To validate our methodology, we consider a well-studied family with many single mutants: the human lysozyme. Most of the perturbations are expected to be at the very localisation of the mutation, but in many cases the effects are propagated at long range. We show that the distances between the mutated residue and the 5% most disturbed residues, exponentially decreases. One third of the affected residues are in direct contact with the mutated position; the remaining two thirds are potential allosteric effects. We confirm the reliability of the residues assigned as significantly perturbed by comparing our results to experimental studies. We confirm with the present method all the previously identified perturbations. This study shows that mutations have long-range impact on protein backbone that can be detected, although the displacement of the affected atoms is small.

Keywords: point mutation; protein structure; allosteric effect; structure comparison; backbone flexibility

Abbreviations. ASA: Accessible Surface Area; AUC: Area Under the Curve; Ca: alpha carbon; PDB: Protein Data Bank; RMS: root mean square deviation; ROC: Receiver Operating Characteristic.

Introduction

Single point mutations, i.e. substitutions of one amino acid side chain, have variable consequences on proteins. On the functional side, the majority of mutations are considered neutral even if some of them are at the origin of diseases (Studer et al., 2013). Sometimes, local changes may occur in the binding sites and therefore impact the protein function (Bartlett et al., 2003; Gong et al., 2009). On the structural side, most of the mutations are neutral (Bastolla et al., 2003; Shakhnovich and Gutin, 1991) and have a local effect but longer range interactions may also be disrupted (Zhou et al., 2007). However, small structural variation of the backbone are crucial for the protein conformation (Berkholz et al., 2009) and if a mutation has a significant negative effect on expression, folding, stability, crystallization, etc., the structure may completely change or cannot be determined. A sticking example of tremendous structural change is the mutation L16A in the Engrailed homeodomain DNA binding protein (PDB code 1ztr), that deeply modifies the structures of the native one (PDB code 1enh) (Religa et al., 2005). Several methods have been developed to predict of the effect of mutations on protein functionality, using sequence information (Adzhubai et al., 2010), but also sometimes structural information (Ponzoni et al., 2020).

Protein structure resolutions are of the order of 1Å nowadays, and the displacement of the backbone atoms caused by a substitution is typically one order of magnitude smaller (Bordner and Abagyan, 2004). Moreover, it has been noticed that even for identical sequences, protein structures often vary, mainly because of different protein–protein interactions, different ligand or protein-ligand interactions or different solvent (Kosloff and Kolodny, 2008). Nevertheless, provided a sufficient number of structures is available, the effect of the mutation can be distinguished from the "noise", intended as the statistical fluctuations due to other sources, such as solvent exposure or secondary structure soaking (Shanthirabalan et al., 2018).

In the present paper, we are interested in evaluating the localisation and the amplitude of the structural perturbations due to a single point mutation. A structural distortion may occur at long range, the so-called allosteric effect. The difficulty is to make the difference between structural perturbations unrelated to mutation, and the real effect of the substitution. We bypassed this issue by analyzing a set of structures from the human lysozyme family, differing one from another by only one mutation. This allowed us to calibrate a method dedicated to understand if a displacement at a position away from the mutation is due to an intrinsic local flexibility or if it is a long-range effect of the side chain replacement.

Human lysozyme has been chosen because of the large number of structures available in the Protein Data Bank (Berman et al., 2000). Lysozyme is a member of the glycoside hydrolase enzyme family that hydrolases the peptidoglycans of cell walls of gram-positive bacteria. Its chain has a length of 130 residues and is divided into two domains: an alpha-domain (residues 1–40 and 83–130) and a beta-domain (residues 41–82) (Dumoulin et al., 2007). The protein contains four disulphide bonds, one of them linking the two domains. The active site of the enzyme is formed in the cleft between the two domains and residues E35 and D53 play important roles in the catalytic mechanism (Muraki et al., 1987). Five naturally occurring mutations, located in the beta domain, result in non-neuropathic amyloidosis (Ahn et al., 2016). In this paper, we are interested in the comparison of the final states of

the folding process, at equilibrium, and we do not address the effect of mutations on the population of intermediate species which may be important in the formation of amyloid fibrils (Buell et al., 2011).
In a previous article, we proposed a method to study in detail the effect on the backbone of a mutation at its very location (Shanthirabalan et al., 2018). We showed that it is possible to identify small perturbations, even in rigid and buried regions, by separately considering protein variations due to structural determination experiments and also due to the protein flexibility. This was achieved by calculating, for a position in one structure, a p-value indicating the rank of the displacement compared to all the structures at this position. This method intrinsically needs a high number of structures with a single mutation each. It actually converts an absolute difference into a relative one, emphasizing by this mean small effects that may nevertheless be significant because they are located in rigid regions. We wish here to apply the same method not only at the very position of the mutation but also at all the positions in order to fully identify the effect of single substitution on the entire structure.

Material and methods

Data set

We have retrieved all the structures of the human lysozymes from the PDB and selected all those with one single mutation relatively to the wild type. Structures determined from NMR experiments have been discarded because of their intrinsic dynamical properties. The chain A of the PDB code 2nwd is taken as the reference for the family because it has the best resolution (1.04 Å) among the 45 structures with the same sequence. All sequences of the dataset (207 PDB entries, 219 sequences of distinct chains) have been aligned with the program MAFFT (Katoh and Standley, 2013). If several chains with identical sequences were available, we selected the structure with the best quality score (defined as resolution minus R-factor) resulting in a dataset of 123 chains. Their resolutions range from 1.04Å to 2.5Å with a median at 1.8Å. The selected chains have the same length and the numbering is taken from the reference structure. The PDB codes and the mutations are listed in Supplementary Table 1.

Structural comparison

We have superimposed all corresponding alpha carbons (Ca) of each mutant on the reference structure with the QBestFit library available at http://bioserv.rpbs.univ-paris-diderot.fr/software.html (Alland et al., 2005). For each pair, the root mean square (RMS) is calculated over a window of 3 consecutive residues. We will call these RMS "RMS3G", to remind that they result from a global superposition and that they are calculated on three residues. We have also computed a local RMS7L, where the superposition is made upon 7 consecutive Ca before calculation of the RMS on this fragment. These are the best window lengths according to (Shanthirabalan et al., 2018). To characterize the perturbation relative to the reference, we will use in this paper, for each Ca pair (mutated vs native), either RMS3G or RMS7L.

P-values, P-RANK and Normalisation

We wish to know whether a given RMS (3G or 7L) is significantly higher than expected, but the values depend on the localisation of the considered fragment and on external factors, such as the

experimental conditions of structure determination or the presence of a ligand. We took into account the two sources of variations (experiment and position) either by ranking these RMS (as in (Shanthirabalan et al., 2018)) or by normalising them. The effects of ligands are analysed in the results.

We have sorted by decreasing values all the RMS3G and RMS7L of each mutated protein and divided each rank by the length of the chain, resulting in a unitless empirical p-value. All positions are taken into account for the calculation of p-values as recommended in (North et al., 2002) and (Davison and Hinkley, 1997). There is no empirical p-value of 0. The RMS (3G or 7L) variations depend on the 3D localisation of the considered position: they are larger in flexible regions, and consequently, the associated p-values are smaller in these regions. To finally assess whether a mutation has a significant effect on the backbone, despite this structural heterogeneity, we needed to compare the p-value at a given position of one protein to the p-values of the other proteins of the dataset at the same position. Then, at each position, we have once again sorted the p-values of all the proteins and divided their rank by the number of proteins in the family. It results in a new empirical p-value that we will call from here on P-RANK. It is meaningful if the number of structures is sufficient to provide statistical significance. This can be seen as an a posteriori methodological justification of the choice of the human lysozyme.

We also calculated the mean and the variance of the RMS (3G or 7L) along the chain for each protein and transformed them into a z-score (named z-score 1). In a second step, we calculated the mean and the variance of these z-scores at each position over all the proteins of the dataset and transformed them again into a second z-score (named z-score 2). For this analysis, we have taken the top 5% of each RMS or transformed RMS: the 5% highest RMS (3G or 7L), the 5% smallest p-values or P-RANK, and the 5% highest z-scores. This is not the best threshold according to ROC curves and Youden index but we chosed it to avoid the selection of too many residues. For example, the best threshold for RMS7L P-RANK is 0.24 (see Figure S6 of supplementary) meaning that one fourth of all positions would be selected, which is too much to analyse. All plots and statistics have been computed with R (R Core Team, 2017). The ROC curves have been calculated and plotted with the pRoc R package (Robin et al., 2011).

Structural characteristics

We have determined the neighbours in 3D space with the NeigbhourSearch function from the PDB module of BioPython (Cock et al., 2009). Two residues are considered as neighbours if there is at least one atom of each separated by a distance less than 4Å. Solvent accessibility (ASA) has been calculated with the program Naccess (Petersen et al., 2011) for all the proteins. The amino acids have been categorized either as buried it the relative ASA is smaller than 25% (Levy, 2010) or exposed otherwise. This separation in two classes as a function of solvent accessibility is also related to the chemical nature of the residues, the bulk of the globular proteins being preferentially occupied by hydrophobic amino acids (Angelov et al., 2002). Secondary structure assignments have been performed with the Stride algorithm (Frishman and Argos, 1995). The six classes given in the output by Stride are back coded in three classes: helices, strands and coils. All assignments are given

according to the residue indexes in the reference protein (2nwd). We have retrieved the ligands and their contacts from the PDBSum records (Laskowski et al., 2018).

Results

RMS

The RMS3G range from 0.02Å to 7.96Å and the RMS7L range from 0.02Å to 2.15Å. The RMS at the mutated positions are slightly larger than those at unmutated positions, as shown in the boxplots of Figure 1. The median is 0.2Å for the RMS3G for the unmutated positions, while it reaches 0.3Å for the mutated positions Although small, a variation of 0.2Å can be considered as significant, as it is highlighted in several studies performed in the group of Takano and coll. (Luzzati, 1952; Takano et al., 1995).

For each pair of proteins, the computed RMS can be represented as profiles (see Figure 2) and the interpretation of the RMS3G or RMS7L is rather difficult because the profile amplitudes can significantly differ between structures determined in various experimental conditions. Moreover, the largest RMS for a protein does not necessarily occurs at the mutated position. Figure 2 shows an example of two possible profiles. In protein 1b5x the largest deviation is located far from the mutation site (Figure 2a), regardless of the superposition (local or global). On the opposite, for 2hee (Figure 2b) the largest deviation occurs at the very position of the mutation.

We present in Figure 2 (C) all the RMS7L distribution along the sequence for all the proteins of the human lysozyme family, sorted according to their median RMS7L values. The most distorted proteins are therefore in the background and the closest to the reference structure in the foreground, along the protein axis. Variable regions of the lysozyme are clearly visible: for example, the region around position 70 has a large RMS7L for almost every protein of this family, whereas the region around residue 90 has a low RMS7L, whichever mutated protein is considered. Taking into account the protein flexibility, a small variation in this rigid region may be significant, whereas the same variation in a more flexible region may not be significant due to the inherent variability of the local substructure.

P-RANK

We have selected the top 5% positions for each measure: the 5% largest RMS or z-scores, or the 5% smallest p-values or P-RANK. Over the whole human lysozyme family, 52% of the residues are exposed and 48% buried. Focusing on the positions populating the 5% highest RMS7L, 74% are exposed at 26% and buried, indicating a significant bias toward the most external regions when the RMS is applied. On the contrary, among the 5% smallest P-RANK, the distribution is 52% of exposed, 48% of buried, like in the full dataset. This illustrates that the P-RANK criterion is able to remove the bias due to 3D localisation.

RMS	Transformation	Mutated	Contact	Ligand	Sum	Other	Total

	ALL	122 (100%)	1222 (100%)	149 (100%)	1493 (100%)	13635 (100%)	15128 (100%)
RMS-3G	RMS	3 (2%)	31 (3%)	0 (0%)	34 (2%)	723 (5%)	757 (5%)
	empirical p-value	25 (20%)	107 (9%)	13 (9%)	145 (10%)	516 (4%)	661 (4%)
	P-RANK	36 (30%)	205 (17%)	3 (2%)	244 (16%)	624 (5%)	868 (6%)
	z-score 1	30 (25%)	118 (10%)	14 (9%)	162 (11%)	595 (4%)	757 (5%)
	z-score 2	38 (31%)	205 (17%)	2 (1%)	245 (16%)	512 (4%)	757 (5%)
RMS-7L	RMS	2 (2%)	31 (3%)	1 (1%)	34 (2%)	723 (5%)	757 (5%)
	empirical p-value	20 (16%)	113 (9%)	4 (3%)	137 (9%)	595 (4%)	732 (5%)
	P-RANK	41 (34%)	241 (20%)	3 (2%)	285 (19%)	614 (5%)	899 (6%)
	z-score 1	24 (20%)	139 (11%)	3 (2%)	166 (11%)	591 (4%)	757 (5%)
	z-score 2	40 (33%)	231 (19%)	3 (2%)	274 (18%)	483 (4%)	757 (5%)

Table 1. Evaluation of the significant displacements in numbers and percentage for the RMS3G and RMS7L, restricted to the top 5% of the five methods. The first line (ALL) reports the number of positions in the entire dataset that are either mutated, or in contact with the mutated residue or with a ligand, then the Sum of the three previous columns. Other is the remaining number of positions not reported by any of the previous measurements. For the RMS3G and RMS7L the same numbers and their proportions are reported, limited to the top 5% of the five methods. The total number of selected residues may differ (last column) because of ex aequo scores.

Table 1 shows the numbers and proportions of the positions with a score in the top 5% according to the various methods used to quantify the structural perturbation. In the dataset there are 122 mutated positions, 1222 residues in 3D contact with the mutated residue and 149 residues in contact with a ligand absent or different from the reference (2nwd). Selecting 5% of the highest RMS (3G or 7L), corresponds to only 2% of the mutated residues. This indicates an un-appropriate tool for evaluating the distortion induced by the mutation. The proportion of high measures of the displacements located at the mutated position increases from RMS to p-value and P-RANK, both for RMS3G and RMS7L. For RMS7L, among the 122 mutated residues that may be submitted to a displacement, 34%, of them are considered as significant, while it would only be 2% with the sole RMS. The simple normalisation (z-score 1) has similar results to the empirical p-values, while the double normalization (z-score 2) is similar to the P-RANK. These trends of the methods are similar with the residues in contacts to the mutation but the proportion of selected residues is smaller (at most 20%). For the ligand effect, all methods show a very smaller proportion of measured displacements of the backbone (between 0% and 9%). These enrichments in mutated residues or residues in direct contact with the mutation in the selected residues with this method demonstrate its ability to highlight outstanding displacements, even if they are in rigid regions and consequently very small. However, the mutated residues or residues in direct contact with the mutation only represent one third of the selected residues (for example 285/899 for P-RANK and RMS7L, the percentages are the same with RMS3G and the z-scores 2). The remaining two thirds are potential allosteric effects.

In order to assess the best method we calculated the ROC (Receiver Operating Characteristic) curves, plotting the sensitivity as a function of specificity for several evaluators of the perturbation (Figure 3).

The residues in direct contact with a source of potential perturbation (mutation or ligand) are considered in the contingency table as positives, resulting in an overestimation of the number of residues really displaced. If a residue is top ranked (5%) by a method, it is considered as true *i.e.* we detect an effect of the mutation. Sensitivity or true positive (TP) rate is the ratio TP/(TP+FN) with FN the number of false negative, i.e. not significantly displaced and unrelated to mutation. Specificity, or false positive rate, is defined as the ratio TN/(TN+FP) with TN the number of true positive, and FP the number of false positive. The maximum area under the curve, i.e. the best evaluator, is obtained with the P-RANK calculated over RMS7L. The top 5% positions selected with P-RANK or z-score 2 of RM7L are mostly the same: 57% of all selected positions by the two methods are common. The area under the curves obtained with P-RANK are larger than those computed for empirical p-values which are themselves larger than those calculated with the RMS. The areas of RMS7L are larger than those of RMS3G, but these two measures don't highlight the same displacements: RMS7L is strictly local whereas RMS3G relies on a global superposition of the structures and may therefore show more global displacements of residues. We will mainly analyse from now on the structural perturbations with the P-RANK calculated from RMS7L and RMS3G.

Amino acid classes

We have clustered the residues in three categories: hydrophobic (F, I, L, M, V, W, Y), loop residues (D, G, N, P, S) (Wojcik et al., 1999) and the third category containing all the others (A, R, C, Q, E, H, K, T). Then the mutations have been clustered in all the 9 possible arrangements of the pairs of wild and mutated classes. We have performed a chi square test to check the distribution in the 9 arrangements in all the mutations on one hand and in the top 5% selected mutations according to the RMS7L P-RANK on the other hand. The test is significant (p-value: 0.0017) :the repartition of the significantly disturbed mutated residues is not uniform within these categories. The main over-represented categories are other->loop and hydrophobic->other and the main under-represented categories are hydrophobic>hydrophobic and loop->loop. According to the P-RANK, a mutation of one residue toward a residue of another category has therefore more effect on the structure than a mutation within the same category.

Distance from the mutation

The distribution of the internal distances between all Ca of the reference protein (2nwd) is presented in Figure 4 left (white bars). From it, we have extracted the distances between otherwise mutated in the database and the significantly displaced Ca according to the top 5% RMS7L P-RANK; they are represented in blue in Figure 4 left. To better evaluate the trend of this distribution, these quantities are divided by the total number of distances in bins of 1Å. This normalised distribution is presented in Figure 4 right: the proportion of selected positions decreases exponentially with the distance to the mutation. The parameter of this exponential is -3.52. This result is in qualitative agreement with the distribution of the propagation of mutational effects found by Naganathan et coll. for T4 lysozyme (Naganathan, 2019; Rajasekaran et al., 2017) with a parameter of -4.7.

Structural and overall analysis

We analysed in details some mutants available in our dataset and compared our results to the ones published by the Dobson group (Ahn et al., 2016; Booth et al., 1997). The two natural mutants, D67H and I56T form amyloid fibrils in the extracellular space of multiple organs and tissues, resulting in systemic non-neuropathic amyloidosis. According to (Ahn et al., 2016), D67H and I56T mutations decrease the native state stability so transient intermediate states can be populated.

The D67H mutant (1lyyA) is one of the most distorted structures relative to the wild type in our dataset, with a global RMS for the whole structure of 1.94Å. These changes mainly occur in the two loops of the beta domain (residues 45-54 and 67-75 (Booth et al., 1997)). Figure 5 shows the superposed structures of the wild type and the D67H mutant. Four segments of 7 residues centred at 48, 53, 67 and 72 are shown as close-up in Figure 5 because of their major distortion between local and global superpositions. We can also notice that these fragments harbour all the residues selected in the top 5% RMS7L P-RANK (in orange and in pink). The first set of residues (66 to 69, in orange), includes the mutation and it is located just before one of the two loops that greatly varies according to global superposition. The highest RMS3G of the D67H structure occurs at the position 70 of the mutant (7.96Å) .The local superposition to the wild type over a 7-residue window around position 67 is shown in the close-up of the lower left corner of Figure 5. The loop centred at 72, even with an important difference compared to the wild type when both structures are globally superposed, results in a very small local variability in all the mutants (upper left close-up of Figure 5). Therefore, it is not selected with P-RANK (RMS7L) at a threshold of 5%. Residues 51 to 54 are identified as significantly disturbed (pink fragment in the bottom right close-up of Figure 5). They are located just after the second loop, highly distorted according to the global superposition. When locally superimposed, this loop, centred at 48, varies much less (upper right close-up). This analysis emphasizes the advantage of using RMS7L P-RANK which allows identifying the hinges, i.e. areas where disturbances occur, but not those impacted by a crankshaft effect.

The network of interactions has been calculated with Arpeggio (Jubb et al., 2017) at the position of the residue 67 for the native and the mutant (Figure S1). It results that the mutation removes residue 67 from the network present in the native structure and forms a single interaction with the cycle of the opposite residue T54. These observations are consistent with those of (Booth et al., 1997) who have shown that the network of H-bonds between the alpha and beta domains is strongly disrupted by this mutation.

The structural effect of the second known amyloid natural mutation I56T (1lozA) is less obvious with a RMS over the whole protein smaller than 1Å from the wild type. Their superposed structures are shown in Figure 6. Two fragments are selected according to the 5% threshold of the RMS7L P-RANK. The first group, residues 36 and 37, is in pink in the close-up on the right of Figure 6. It is in direct 3D contact with the mutated residue, not significantly displaced itself. The structural variations of the backbone are very weak but the interaction network around residue 56 is slightly modified (Figure S2). The second group, residues 73 to 76, is in orange in the close-up on the left of Figure 6. It is not in contact with the mutation but is localised in one of the two loops of the beta domain. The displacements of the backbone are more visible.

In order to easily visualise the location of the significant displacements over the whole dataset, we present in Figure 7 all the most perturbed positions according to the RMS7L P-RANK. The selected positions appear with bright colours in Figure 7, corresponding either to mutations (cyan) or to residues in contact with mutations (magenta). The maps obtained with the RMS3G P-RANK (Figure S3) or with the double normalization of the RMS7L (z-score 2) (Figure S4) are very similar to the map of RMS7L P-RANK. One can notice that when several mutations are available at a given position, they produce a displacement in the same region. For example, mutations at positions 21 and 23 result in a significant displacement at long range, on the fragment 95 to 106. The 3D localization of these residues in the reference structure is shown in Figure 8. The four mutations available for positions 21 and 23 (R21A, R21G, I23A and I23V) mainly influence the structure of the N-terminal end of the facing alpha helix.

We finally compared our results with the articles published by researchers at the Institute for Protein Research at Osaka University who designed and crystallized many mutants of human lysozyme (Takano et al., 1995, 1997, 2000, 2001a, 2001b, 2001c; Funahashi et al., 2000, 2002). In most cases, they did not identify noticeable structural variation in the main chain, but we have extracted the 11 proteins where it was the case (see Table 2). When these authors observe variations in the mutant structures, we always also detect them with RMS7L and sometimes also with RMS3G (bold residue indexes in Table 2). Additionally, we find also other perturbations.

In the 1995 paper, Takano and col. (Takano et al., 1995) studied the effect of Isoleucine to Valine mutations at positions 23, 56, 59, 89 and 106 (see Table 2 and Figure S5 of supplementary materials). For the mutant I23V, Takano and col. found that the region 99-105 facing the mutation is displaced, but not the location of the mutation itself. In mutant I23V, we select residues 70, 95, 96, 100, and 101 as disturbed (by RMS7L). Takano observed the C-ter and N-ter extremities are disturbed by the mutations I56V and I89V and I106V; we select the same two regions. The mutated residue is selected by Takano and also with both P-RANK.

In (Takano et al., 1997) the authors designed 9 valine to alanine mutants to study the hydrophobic effect on stability; we have 7 of these mutants in our dataset (we cannot compute the RMS for the mutations V2A and V130A). The mutations V93A, V100A and V125A create a cavity within the protein toward which the residues move. We also identify these mutated residues and their surrounding residues as disrupted (see Figure 9, the cyan positions surrounded by pink positions). For the V99A mutation, we also select residues after the mutation as in (Takano et al., 1997), but besides we detect a disruption at the very location of the mutation, as for V100A. The three remaining mutations (V74A, V110A, V121A) cause no noticeable backbone structural variations according to Takano and col. We also found a few numbers of perturbed residues that are far from the mutation for V74A and V110A, but for V121A we find two blocks of displaced residues that are very similar to the residues identified for the V125A mutant (Figure 9).

Discussion

The initial question that motivated this work was whether small structural variations in the peptide backbone caused by single point mutations were measurable with present techniques. In a previous

paper (Shanthirabalan et al., 2018) we developed a method focusing on the very location of the mutation itself. We showed that 38% of all mutations in 11 protein families produce a significant effect on the displacement. Even if the RMSD is greater when the mutation is in loops than in regular secondary structure, the relative effect is actually more important for regular secondary structures and buried positions, i.e. is for rigid regions. This is coherent with previous studies on Bacteriophage T4 Lysozyme showing that mutations on lysozyme that impaired function exhibited lower than average temperature factors (Alber et al., 1987). Here we want to look at the structural variations spread on the whole protein. There are many potential sources of structural variation for proteins, other than the mutation itself. In fact, protein–protein interactions, ligand interactions or different solvent have often a greater impact on the overall structure than a substitution (Kosloff and Kolodny, 2008; Shanthirabalan et al., 2018). We have minimized their impact working on a well-studied family with a large number of mutants: the human lysozyme. We have collected 122 different mutants with known structures, each one differing from the wild type by a single replacement.

It is very uncommon that the entire structure of a protein varies greatly as a result of a single substitution. Therefore, we did not measure a global RMS on the whole structures but rather local variations by calculating RMS on fragments of 3 residues after optimal superposition of whole structures (RMS3G) or on fragments of 7 residues after optimal superposition on these 7 residues (RMS7L). These two RMS allow measuring the local variations of the protein chain but the RMS3G measures the displacement of the fragment relative to the whole molecule - there can therefore be lever-arm effects - while the RMS7L measures the deformation of the fragment of 7 residues. We chose these lengths of fragments because they produce the strongest effect (Shanthirabalan et al., 2018).

RMS3G and RMS7L are computed at each position (except the extremities) of the mutants and can be represented as profiles (Figure 2). The profile amplitude varies depending on the mutants: the structure of some proteins fluctuates more than others and the RMS for these proteins are all large. It is also clear that they vary according to the positions: at each position, the RMS are strongly correlated to the flexibility of the molecule. These two sources of variation mask the effect of the mutation itself: if we select the largest RMS, the only information we get is the detection of the most flexible regions of the proteins. We can tackle this issue thanks to the large number of mutants of the lysozyme family: we estimated the variability of RMS for each protein and at each position and took them into account by calculating either P-RANKs or normalized values (z-score2) from these RMS. We then analysed the distribution queues, i.e. the 5% of the lowest P-RANKs or the 5% of the largest z-scores and compared them to the 5% of the highest RMS. The objective is then to evaluate if this methodology better allows identifying regions where the main chain is deformed (with the RMS7L) or displaced (with the RMS3G), or if the regions are randomly selected.

The place where the chain is most likely to be disrupted is the exact location of the mutation. We therefore first checked the proportion of mutated residues selected in the top 5% (Table 1). P-RANK and z-score2 allow finding the highest number of mutations among the most significant 5% perturbations. These results are also confirmed with the ROC curve (Figure 3). For the other selected positions, we also looked at potential sources of disturbance such as ligand difference, or direct contact with the mutated residue. The proportions of residues belonging to these two categories are

also highest for P-RANK and z-score2 compared to RMS (Table 1). It shows that these variables are more relevant than the others to evidence small distortions in the structures. These three categories, entitled as mutated, contact or ligand in Table 1, only cover one third of the residues selected as significantly disturbed. What about the other residues selected? Is there a long-range propagation of the effect of the mutation?

The main displacement in the structures of an ensemble of human lysozyme differing one from another by a single point mutation does not necessarily occur at the position of the mutation. This is in agreement with the results of the group of Matthews on the case of T4 lysozyme (Eriksson et al., 1993). We find that mutations seem to have long-range consequences as in (Verma et al., 2012). By calculating the distribution of the distance to mutation for all the residues selected as significantly displaced, we find, as in (Naganathan, 2019; Rajasekaran et al., 2017), that the distribution of the proportion of disturbed residues exponentially decreases with distance. These selected residues therefore have a consistent distribution according to this criterion. We can also notice that they are often grouped together: several successive residues are often selected (Figure 7) or that they are geometrically grouped (Figure 8).

We have studied in more detail the two natural mutations, D67H and I56T, related to non-neuropathic amyloid. There are two side-chain hydrogen bonding networks in the wild-type protein (Artymiuk and Blake, 1981). Interestingly, both of these networks stabilize the beta-domain, and cover the diverse regions of the primary structure that are brought into close proximity by the native fold. In the first case, the structure is disturbed in two very flexible loops (Figure 5), that we find with the P-RANK calculated from the RMS3G. This is due to a leverage effect because with local superposition (RMS7L), these loops remain unchanged although some residues at the base of these loops are locally disturbed. These results are consistent with the work of Booth and. coll. where they have shown that the weak bond network of these residues is strongly disturbed. The structure of the I56T variant is almost identical to that of the wild-type protein (Booth et al., 1997) (Figure 6) but we also identify displaced residues belonging to the weak bond network which is disturbed according to (Booth et al., 1997).

Conclusion

We present in this paper a method to identify and isolate protein backbone perturbations caused by a point mutation. We show that it allows identifying disturbances in direct contact with the mutation but also distant disturbances, sign of an allosteric effect of the mutation. However, this method requires a large number of mutant structures, which is not available for the majority of protein families, but this work must be seen as a milestone for identifying allosteric effects in protein structures. Now that we have shown on a test case that it is possible to identify these perturbations on a large dataset of mutated structures, we wish to extend the method to be able to evidence the structural displacements from a single structure. In order to do this, one assumption is to generate multiple conformations of the molecule thanks to several methods: B-factors, normal modes, model the expected variations of Ca positions with ellipses (Gerstein and Altman, 1995; Taylor et al., 1983), use of the several models usually reported with NMR structures, generation of conformations fulfilling distance constraints, such as CONCOORD (de Groot et al. 1999), or some of the various developments of the elastic network models, such as, among others, PROPHET (Lavery and Sacquin-Mora, 2007).

Acknowledgements

JC wishes to deeply thank Bernard Lahire for drawing his attention on the fact that folding also may concern the social world, and not only proteins.

References

Ahn, M., Hagan, C.L., Bernardo-Gancedo, A., De Genst, E., Newby, F.N., Christodoulou, J., Dhulesia, A., Dumoulin, M., Robinson, C.V., Dobson, C.M., Kumita, J.R., 2016. The Significance of the Location of Mutations for the Native-State Dynamics of Human Lysozyme. Biophys. J. 111, 2358–2367. https://doi.org/10.1016/j.bpj.2016.10.028

Adzhubai, IA, Schmidt, S., Pershkin, L., Ramensky, VE., Gerasimova, A., Bork, P., Kondrashov, AS., Sunyaev, SR. 2020. A method and server for predicting damaging missense mutations. Nat Methods. 7:248-249

Alber, T., Sun, D.P., Nye, J.A., Muchmore, D.C., Matthews, B.W., 1987. Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. Biochemistry 26, 3754–3758. https://doi.org/10.1021/bi00387a002

Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., Renault, N., Wong, Y., Cantalloube, H., Chomilier, J., Hochez, J., Pothier, J., Villoutreix, B.O., Zagury, J.F., Tuffery, P., 2005. RPBS: a web resource for structural bioinformatics. Nucleic Acids Res. 33, W44-9.

Angelov, B., Sadoc, J.-F., Jullien, R., Soyer, A., Mornon, J.-P., Chomilier, J., 2002. Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds. Proteins 49, 446–456. https://doi.org/10.1002/prot.10220

Artymiuk, P.J., Blake, C.C.F., 1981. Refinement of human lysozyme at 1.5 Å resolution analysis of non-bonded and hydrogen-bond interactions. J. Mol. Biol. 152, 737–762. https://doi.org/10.1016/0022-2836(81)90125-X

Bartlett, G.J., Borkakoti, N., Thornton, J.M., 2003. Catalysing new reactions during evolution: economy of residues and mechanism. J. Mol. Biol. 331, 829–860. https://doi.org/10.1016/s0022-2836(03)00734-4

Bastolla, U., Porto, M., Eduardo Roman, M.H., Vendruscolo, M.H., 2003. Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J. Mol. Evol. 56, 243–254. https://doi.org/10.1007/s00239-002-2350-0

Berkholz, D.S., Shapovalov, M.V., Dunbrack, R.L., Karplus, P.A., 2009. Conformation Dependence

of Backbone Geometry in Proteins. Structure 17, 1316–1325. https://doi.org/10.1016/j.str.2009.08.012

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 Res. 28, 235–242.

Booth, D.R., Sunde, M., Bellotti, V., Robinson, C.V., Hutchinson, W.L., Fraser, P.E., Hawkins, P.N., Dobson, C.M., Radford, S.E., Blake, C.C., Pepys, M.B., 1997. Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. Nature 385, 787–793. https://doi.org/10.1038/385787a0

Bordner, A.J., Abagyan, R.A., 2004. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 57, 400–413. https://doi.org/10.1002/prot.20185

Buell, A.K., Dhulesia, A., Mossuto, M.F., Cremades, N., Kumita, J.R., Dumoulin, M., Welland, M.E., Knowles, T.P.J., Salvatella, X., Dobson, C.M., 2011. Population of Nonnative States of Lysozyme Variants Drives Amyloid Fibril Formation. J. Am. Chem. Soc. 133, 7737–7743. https://doi.org/10.1021/ja109620d

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinforma. Oxf. Engl. 25, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press.

Dumoulin, M., Johnson, R.J.K., Bellotti, V., Dobson, C.M., 2007. Human Lysozyme, in: Uversky, V.N., Fink, A.L. (Eds.), Protein Misfolding, Aggregation, and Conformational Diseases: Part B: Molecular Mechanisms of Conformational Diseases, Protein Reviews. Springer US, Boston, MA, pp. 285–308. https://doi.org/10.1007/978-0-387-36534-3_14

Eriksson, A.E., Baase, W.A., Matthews, B.W., 1993. Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. J. Mol. Biol. 229, 747–769. https://doi.org/10.1006/jmbi.1993.1077

Frishman, D., Argos, P., 1995. Knowledge-based protein secondary structure assignment. Proteins 23, 566–579. https://doi.org/10.1002/prot.340230412

Funahashi, J., Takano, K., Yamagata, Y., Yutani, K., 2002. Positive Contribution of Hydration Structure on the Surface of Human Lysozyme to the Conformational Stability. J. Biol. Chem. 277, 21792–21800. https://doi.org/10.1074/jbc.M110728200

Funahashi, J., Takano, K., Yamagata, Y., Yutani, K., 2000. Role of Surface Hydrophobic Residues in the Conformational Stability of Human Lysozyme at Three Different Positions,. Biochemistry 39, 14448–14456. https://doi.org/10.1021/bi0015717

Gerstein, M., Altman, R., 1995. Using a measure of structural variation to define a core for the globins. Comput. Appl. Biosci. CABIOS 11, 633–644.

Gong, S., Worth, C.L., Bickerton, G.R.J., Lee, S., Tanramluk, D., Blundell, T.L., 2009. Structural and functional restraints in the evolution of protein families and superfamilies. Biochem Soc Trans 37, 727–733. https://doi.org/10.1042/BST0370727

de Groot, B., Vriend, G., Berendsen, HJ, 1999. Conformational changes in the chaperonin GroEL: new insights into the allosteric mechanism. J. Mol. Biol. 286, 1241-1249

Jubb, H.C., Higueruelo, A.P., Ochoa-Montaño, B., Pitt, W.R., Ascher, D.B., Blundell, T.L., 2017. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J. Mol. Biol. 429, 365–371. https://doi.org/10.1016/j.jmb.2016.12.004

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30, 772–780. https://doi.org/10.1093/molbev/mst010

Kosloff, M., Kolodny, R., 2008. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins 71, 891–902. https://doi.org/10.1002/prot.21770

Laskowski, R.A., Jabłońska, J., Pravda, L., Vařeková, R.S., Thornton, J.M., 2018. PDBsum: Structural summaries of PDB entries. Protein Sci. Publ. Protein Soc. 27, 129–134. https://doi.org/10.1002/pro.3289

Lavery, R. Sacquin-Mora, S. 2007. Protein mechanics: a route from structure to function. J. Biosci. 32, 891-898.

Levy, E.D., 2010. A simple definition of structural regions in proteins and its use in analyzing interface evolution. J. Mol. Biol. 403, 660–670. https://doi.org/10.1016/j.jmb.2010.09.028

Luzzati, V., 1952. Traitement statistique des erreurs dans la determination des structures cristallines. Acta Crystallogr. 5, 802–810. https://doi.org/10.1107/S0365110X52002161

Muraki, M., Morikawa, M., Jigami, Y., Tanaka, H., 1987. The roles of conserved aromatic amino-acid residues in the active site of human lysozyme: a site-specific mutagenesis study. Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol. 916, 66–75. https://doi.org/10.1016/0167-4838(87)90211-1

Naganathan, A.N., 2019. Modulation of allosteric coupling by mutations: from protein dynamics and packing to altered native ensembles and function. Curr. Opin. Struct. Biol. 54, 1–9. https://doi.org/10.1016/j.sbi.2018.09.004

North, B.V., Curtis, D., Sham, P.C., 2002. A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. Am. J. Hum. Genet. 71, 439–441.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786. https://doi.org/10.1038/nmeth.1701

Ponzoni, L., Penahererra, DA., Oltvai, Z., Bahar, I., 2020. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa127

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R Taylor, W., M Thornton, J., G Turnell, W., 1983. An ellipsoidal approximation of protein shape. J. Mol. Graph. 1, 30–38. https://doi.org/10.1016/0263-7855(83)80001-0

Rajasekaran, N., Suresh, S., Gopi, S., Raman, K., Naganathan, A.N., 2017. A General Mechanism for the Propagation of Mutational Effects in Proteins. Biochemistry 56, 294–305. https://doi.org/10.1021/acs.biochem.6b00798

Religa, T.L., Markson, J.S., Mayor, U., Freund, S.M.V., Fersht, A.R., 2005. Solution structure of a protein denatured state and folding intermediate. Nature 437, 1053–1056. https://doi.org/10.1038/nature04054

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77. https://doi.org/10.1186/1471-2105-12-77

Shakhnovich, E.I., Gutin, A.M., 1991. Influence of point mutations on protein structure: probability of a neutral mutation. J. Theor. Biol. 149, 537–546.

Shanthirabalan, S., Chomilier, J., Carpentier, M., 2018. Structural effects of point mutations in proteins. Proteins Struct. Funct. Bioinforma. 86, 853–867. https://doi.org/10.1002/prot.25499

Studer, R.A., Dessailly, B.H., Orengo, C.A., 2013. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem. J. 449, 581–594. https://doi.org/10.1042/BJ20121221

Takano, K., Ogasahara, K., Kaneda, H., Yamagata, Y., Fujii, S., Kanaya, E., Kikuchi, M., Oobatake, M., Yutani, K., 1995. Contribution of Hydrophobic Residues to the Stability of Human Lysozyme:

Calorimetric Studies and X-ray Structural Analysis of the Five Isoleucine to Valine Mutants. J. Mol. Biol. 254, 62–76. https://doi.org/10.1006/jmbi.1995.0599

Takano, K., Yamagata, Y., Fujii, S., Yutani, K., 1997. Contribution of the Hydrophobic Effect to the Stability of Human Lysozyme: Calorimetric Studies and X-ray Structural Analyses of the Nine Valine to Alanine Mutants,. Biochemistry 36, 688–698. https://doi.org/10.1021/bi9621829

Takano, K., Yamagata, Y., Yutani, K., 2001a. Contribution of Polar Groups in the Interior of a Protein to the Conformational Stability,. Biochemistry 40, 4853–4858. https://doi.org/10.1021/bi002792f

Takano, K., Yamagata, Y., Yutani, K., 2001b. Role of amino acid residues in left-handed helical conformation for the conformational stability of a protein. Proteins Struct. Funct. Bioinforma. 45, 274–280. https://doi.org/10.1002/prot.1147

Takano, K., Yamagata, Y., Yutani, K., 2001c. Role of non-glycine residues in left-handed helical conformation for the conformational stability of human lysozyme. Proteins Struct. Funct. Genet. 44, 233–243. https://doi.org/10.1002/prot.1088

Takano, K., Yamagata, Y., Yutani, K., 2000. Role of Amino Acid Residues at Turns in the Conformational Stability and Folding of Human Lysozyme,. Biochemistry 39, 8655–8665. https://doi.org/10.1021/bi9928694

Verma, D., Jacobs, D.J., Livesay, D.R., 2012. Changes in Lysozyme Flexibility upon Mutation Are Frequent, Large and Long-Ranged. PLoS Comput. Biol. 8, e1002409. https://doi.org/10.1371/journal.pcbi.1002409

Wojcik, J., Mornon, J.P., Chomilier, J., 1999. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J. Mol. Biol. 289, 1469–1490. https://doi.org/10.1006/jmbi.1999.2826

Zhou, R., Eleftheriou, M., Royyuru, A.K., Berne, B.J., 2007. Destruction of long-range interactions by a single mutation in lysozyme. Proc. Natl. Acad. Sci. U. S. A. 104, 5824–5829. https://doi.org/10.1073/pnas.0701249104