Structural effects of point mutations in proteins

Suvethigaa Shanthirabalan, Jacques Chomilier, Mathilde Carpentier

Structural effects of point mutations in proteins

Suvethigaa Shanthirabalan[1], Jacques Chomilier[2], Mathilde Carpentier[1,2]

1. Institut Systématique Evolution Biodiversité (ISYEB), Sorbonne Université, MNHN, CNRS, EPHE, Paris, France.

2. Sorbonne Université, CNRS, MNHN, IRD, IMPMC, BiBiP, Paris, France

Corresponding author: Mathilde.Carpentier@upmc.fr

Mail: s.suvethigaa@hotmail.fr; jacques.chomilier@upmc.fr; Mathilde.Carpentier@upmc.fr

Abstract

A structural database of eleven families of chains differing by a single amino acid substitution has been built. Another structural dataset of 5 families with identical sequences has been used for comparison. The RMSD computed after a global superimposition of the mutated protein on each native one is smaller than the RMSD calculated among proteins of identical sequences. The effect of the perturbation is very local, and not necessarily the highest at the position of the mutation. A RMSD between mutated and native proteins is computed over a 3-residue or a 7-residue window at each position. To separate the effects of structural fluctuations due to point mutations from other sources, pair RMSD have been translated into p-values which themselves are included in a score called P-RANK. This score allows highlighting small backbone distortions by comparing these RMSD between mutated and native positions to the RMSD at the same positions in the absence of a mutation. It results from the P-RANK that 38% of all mutations produce a significant effect on the displacement. Compared to a random distribution of RMSD at un-mutated positions, we show that, even if the RMSD is greater when the mutation is in loops than in regular secondary structure, the relative effect is more important for regular secondary structures and for buried positions. We confirm the absence of correlation between RMSD and the predicted variation of free energy

of folding but we found a small correlation between high RMSD and the error in the prediction of ΔΔG.

Abbreviations: ASA: Accessible Solvent Area; PDB: Protein Data Bank; RMSD: root mean square deviation, RMSD-n: root mean square deviation on n adjacent amino acids.

Short title: Structural effects of point mutations in proteins

Introduction

Amino acid substitution, a type of genetic point mutation, is one of the basic events that can drive evolution, leading to a variety of downstream consequences that are only moderately predicted, with a spectrum that goes from neutrality to deleterious effects on the proteins. On the one hand, the building of phylogenies relies on the assumption that the majority of the observed genetic mutations are neutral, with no effect on the phenotype. On the other hand mutations can be at the origin of tremendous diseases [1]. From a structural point of view, the issue of the neutrality of an amino acid substitution is difficult to predict from scratch, because it may have huge effects on the protein structure and consequently on the function itself, especially if it occurs in the active site of enzymes or in the binding pockets of receptors [2–4]. The mutation may also have effects at a long range position [5]. One of the most exciting examples of a rare event 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) [6]. However, a body of studies agrees that the majority of substitutions has no significant effect on the global structure, stability and function of the protein. Indeed, it has been shown that 50 to 80% of the amino acids may be changed without altering significantly the protein structure [7–10]. These observations corroborate the neutral hypothesis of point mutations but it is important to keep in mind that it does not mean that all mutations are

neutral: the majority of the point mutations are indeed counter-selected. For example, the probability that a human DNA repair enzyme 3-methyladenine DNA glycosylase becomes non-functional upon random mutation is 34% (±6%) and this proportion can be extended to other families [11].

Amino acid side chain substitutions have a non negligible effect on protein stability, since proteins are only marginally stable, typically between 3 and 7 kcal/mol between folded and unfolded conformations [12] and the effect of a single mutation is around -0,95 kcal/mol (average of the Protherm database collected experimental data [13]). This weak stability of proteins is supposed to be the either the result of a balance between function and stability [14] or the result of a balance between destabilizing mutations and highly unstable proteins ([15–18], and [19] for a very interesting review on protein structural bioinformatics). Many algorithms and web servers have been developed to provide estimation of the variation of the Gibbs free energy upon the effect of a given point mutation, such as FoldX [20], I-Mutant [21], PoPMusic [22], Maestro [23], Cupsat [24], mCSM [25], NeEMO [26], Strum [27], RosettaMP [28] or SPROUTS [29] which is a web server combining the results of several methods. The Gibbs free energy estimations are scaled with experimental measures of the stability (ΔΔG) recorded in the Protherm database [13]. From the calibration with Protherm (1866 proteins with both ΔΔG and PDB structures in 2017), these methods try to predict if a given mutation is going to be destabilizing, neutral or stabilizing. Comparison between predicted and experimental energy variation, globally gives satisfactory results but in some cases the discrepancy can be rather high [29]. Similarly, in most of the previously mentioned methods, the backbone displacement is not taken into account: only side chains are allowed moving [30]. The amplitude of side chain moves are retrieved from rotamer libraries and their degrees of freedom are highly dependent on backbone conformation [31–33]. Therefore, taking into account backbone deformation should improve our understanding of substitution consequences. Two models have been defined for

backbone accommodation under the effect of an amino acid substitution. The first one is the backrub motion [34], derived from the observation of alternative positions of side chain atoms in ultra-high-resolution crystallographic structures. A backrub motion can be represented by a rotation around the pseudo-dihedral angles of three adjacent peptides, with no effect on the rest of the protein. It has been successfully used to improve Rosetta ΔΔG computation [35]. The authors concluded that protein backbone deformations are influenced by side chain conformations in a predictable manner [36] and that the side chain conformation is backbone dependent [37]. Another model has been designed by Bordner and Abagyan [38] which was refined upon a database of 2141 pairs of protein structures differing only by a single point mutation. The backbone is first transformed in an idealized covalent geometry, and after introduction of the mutation, the side chain χ dihedral angles are minimized. This model has also been successfully used to improve Gibbs free energy prediction after a mutation.

We aim in this article at analysing the very effect of a mutation on the backbone conformation, while taking into account other parameters such as the variability due to different experimental conditions but also the localisation of the mutation in the protein structure. Is there now enough data to statistically quantify the small effect of a point mutation on the backbone? Protein structures are routinely determined at a resolution of 1 Å at best, while the displacement of the backbone atoms caused by a substitution is typically 10 times smaller. Therefore, what we are observing in protein crystal is at least noisy, or maybe the variations are too small to be observed. Moreover, it has been remarked that even if the protein sequences are strictly identical, their structures often vary [4]. From this knowledge, our first question was: can we observe, in protein crystal structures, variations of the backbone caused by an amino acid substitution? How to make the difference between the noise (assumed as structural fluctuations independent of the mutation) and the actual effect of the substitution. We tackled this problem by constructing families of at least 20 protein structures

differing one from another by only one mutation. We evaluated in these families the effect of the mutations on the structural variations and compared them to other sources of displacements. Having established that we definitely observe an effect, we then tried to quantify and characterize the perturbation of amino acid substitution on the backbone conformation of proteins.

We first extracted from the PDB [39] 11 families of proteins bringing together 591 structures, with a single mutation relative to a reference sequence (mainly, the native one). The superposition of mutated structures on the native one is performed in a second step and the structural variations are locally quantified along the chains. To derive a statistical analysis of the data, in a third step the variations in structures are evaluated within a fixed number of residues window and compared to random distributions. Contribution of the residue burying and of the secondary structure assignment has been investigated. Besides, the assumption that replacement of side chains can affect the position of an amino acid distant in sequence, but close in the 3D structure, has been raised. The impact of the mutations on the stability has been investigated from experimental values (Protherm) and predicted values (with FoldX).

Materials and methods

Data set

Our goal is to observe the effect of unique single point mutations on a protein structure, so we built a dataset of families of proteins with only one mutation per chain within one family. In order to evaluate the statistical significance of the distribution of structural perturbation introduced at various positions, we only kept families containing at least 20 proteins. This threshold drastically reduces the number of mutants to work on: we only found 591 proteins in the whole PDB satisfying our criteria. The description of the methodology is the following. The PISCES server [40] has been scanned to filter proteins according to several criteria, such as

protein length (more than 30 amino acids long), no missing backbone residues in the middle of the chains (some side chains may be missing). Structures from NMR experiments have been discarded because of their intrinsic dynamical properties. We clustered the obtained proteins at two thresholds of sequence identity: 100% and 99% with the CDHit sequence clustering program [41]. The next step is the subtraction of the PDB codes present in the 100% list from the 99% list, in order to remove strictly identical sequences. In case of two proteins with the same sequence, the one with the best quality is kept, defined as the resolution minus R-value, according to the formula from the PDB site (http://www.rcsb.org/pdb/statistics/clusterStatistics.do). In cases where several identical chains are present in a given protein, only the first one is kept. Multi-domain chains are conserved. The length of the chains can slightly differ within a cluster, due to the allowed extensions either at N or C terminal ends. The maximum variation in length is 12 residues, corresponding to a relative difference of less than 4%. At this step, from the 47,311 clusters (6,576 with at least two chains), only 24 with at least 20 chains are retained, sharing at least 99% of sequence identity.

Within each cluster, all sequences are aligned using MAFFT multiple alignment program [42]. Within each cluster, a reference is considered, defined as the protein sharing only one mutation with the greatest number of sequences of the cluster. It generally corresponds to the native protein, as reported in the PDB files, but for one cluster, the T4 lysozyme cluster, the reference protein is 1lw9, which is not a native sequence, since it is annotated as a two mutations sequence. Nevertheless, all the 146 proteins in this cluster contain one mutation relative to 1lw9. This choice has been done because it produces a cluster with a larger size. As already noticed by [38], the identity of the wild type is unimportant for the purpose of comparing the conformations. All chains with more than one single mutation per chain (with the exception of the N and C termini residues) are discarded. Then, it has been checked that

the number of chains is constant for all members of the clusters. The presence of ligands is known to be a difficulty [38], and we tried at best to minimize the difference of ligands. When a ligand was found for a protein in one cluster, we reconsidered the list of identical chains (CD-Hit 100%) and we replaced this member by another one with a ligand identical to the reference one. The criterion of resolution was hence discarded in this particular case.

In order to keep a sufficient number of members within one cluster, the threshold has been placed at 20 members, and the final number of families is 11, listed in Table I. The total number of chains is 591, and it could not be increased because of the lack of entries in the PDB fulfilling the previous criteria. It is smaller than the recent database from Zhang lab [27], with 3,421 mutations from 150 proteins, or the one by Kosloff [4], but in our case we have at least 20 chains per family, allowing a statistical analysis of the actual effect of the mutation. This dataset will be called from hereon the mutated dataset.

As a control sample, a second dataset of families of proteins with 100% sequence identity among members has been constructed with the chains issued from the CDHit procedure of previous step. All pairs within one cluster are then strictly identical. Since the threshold of at least 20 members in any family is conserved for statistical reasons, this dataset is composed of only five clusters, of the following references: 2nwd (lysozyme), 2e3w (RNaseA), 2vb1(lysozyme), 4fi8 (transthyretin) and 2j8c (reaction centre). This second dataset, containing the same functions as the mutated dataset, will be called in the rest of the paper the identical dataset. Modified amino acids have been accepted in this second dataset of 510 chains.

Structural comparison

We have computed, within each family, the RMSD between each reference and mutated structures, after a superposition of all alpha carbons (Ca) with the QBestFit algorithm, downloaded from http://bioserv.rpbs.univ-paris-diderot.fr/software.html [44]. However, the

effects of a point mutation are rather small (0.36 Å on average), so we calculated a local RMSD on a window of 1 to 20 residues after the global superposition. We will call these RMSD "RMSD-nG" (RMSD-3G for example), to remind that they result from a global superposition and that they are calculated on n residues. It is possible to locally superpose only these n Ca, and to compute a "RMSD-nL" after this local superposition. The putative effect of the window size has been checked by increasing its size up to 20 positions; involving all atoms of the backbone instead of limitation to alpha carbons have also been considered. However, analyses have been conducted with RMSD-3G because the length of 3 is in phase with the backrub concept of local deformation due to the accommodation of diverse side chains [34, 36, 45].To appreciate the relative importance of the value of the RMSD-n(G/L) at the position of the mutation, compared to the RMSD-n(G/L) over the rest of the positions of the protein, all RMSD-n(G/L) are translated into unit less empirical p-values. For a pair composed of one mutated structure and the reference, the p-value of a residue position (mutated or not) corresponds to the rank of its RMSD-n(G/L) within all RMSD-n(G/L) of each residue position in the pair, divided by the total number of residue positions in the pair. It is the relative area of the empirical distribution of the RMSD-n(G/L) of a pair higher than the RMSD-n(G/L) value at the considered residue position. In other words, it calibrates the importance of the RMSD-n(G/L) at the place of the considered residue in regard of all RMSD-n(G/L) along the sequence. In order to appreciate the effect of mutations, all p-values calculated for the RMSD-n(G/L) centred on the mutations are ranked by increasing order. Empirical p-values give an overview of the structural perturbation at the position of the mutation, compared to all displacements of the same mutated protein. The advantage of a p-value instead of the raw RMSD-n(G/L) or of the rank, is that it is independent of the length of the protein, and therefore it is more of comparable character. It allows separating the effect of the mutation from the effect of the modification of structures due to different experimental

conditions or alternative conformations. The aim of this method of calculating an empirical p-value is to remove the RMSD-n(G/L) due to global protein structure variations not originating in mutations.

Random distributions of RMSD

In order to separate the effect of the mutation from the effect of the modification of structures at this position regardless of a mutation, actual p-values need to be compared to a set of p-values calculated in chains without mutation at these positions. Therefore, we have generated a set of 200 random distributions of RMSD-n(G/L) for each cluster: each one contains the same number of RMSD-n(G/L) as the number of members in the cluster, randomly chosen at positions where at least one member of the family is mutated, but only in proteins not mutated at that position They are also converted into p-values. From hereon, these reference distributions will be called random. It is also possible to quantify the tendency of the RMSD-n(G/L) relative to a random draw by computing the cumulative distance (or the surface) between the diagonal of the p-values against rank plot and the p-value curve (see for instance Figure 5). These distances are positive if the p-value curve is under the diagonal and negative if it is above.

Those random distributions allow us to outline that the RMSD-n(G/L) variations are depending on the 3D localisation of the considered position. It is possible to take into account the differences due to the 3D localisation by performing an extension of this p-value method. Position by position, we ranked p-values for all the proteins of the family. This rank is then divided by the number of proteins in the family, resulting in a new empirical p-value, that we will call from hereon P-RANK. Therefore, one can use a threshold for the P-RANK that will be fixed at 0.05 as usually.

Structural characteristics

To determine the neighbours in 3D space, a contact is defined if any pair of atoms from different amino acids is separated by less than 4 Å as in [38]. This has been calculated by the NeigbhourSearch function from the PDB module of Biopython [46]. We then analysed the RMSD-n(G/L) of all neighbouring residues that are not necessarily close in sequence.

Solvent accessibility has been calculated with the program Naccess for all the proteins, in order to separate the amino acids in two classes: either buried (relative ASA < 25%) or otherwise exposed [47]. 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 [48].

Secondary structure assignments have been performed with the Stride algorithm [49]. The six classes given in the output by Stride are back coded in three classes: helices, strands and coils. All analyses have been made according to these characteristics computed for the reference proteins in order to have a coherent dataset.

Free energy calculation

To compute the free energy of folding of a protein, one needs to check if any atomic coordinates are missing in the side chains of some PDB files. If it is the case, missing atoms have been added with the PDB_Hydro program [50]. When the structures are complete, the procedure described in the literature for free energy calculation by FoldX has been followed [20, 51–53]. First, 3D structures were optimized using the RepairPDB module of FoldX in order to obtain a canonical stereochemistry for the proteins. Second, structures corresponding to single point mutants were generated by the BuildModel function. It includes back-mutated structures, i.e. structures generated from the mutant, on which a mutation leading to the native sequence is applied. For each mutation of the mutated dataset, the $\Delta\Delta G$ estimated by FoldX has been computed, and compared to experimental values extracted from Protherm when available. Protherm has been mined with the sequence of each of the 591

proteins as an input, resulting in 147 entries with a ΔΔG value documented. None of them was in the cluster of transmembrane proteins (reaction center). We have excluded 18 proteins with positive estimates of ΔG by FoldX, because it would be too different from the folded structure such that the effect of a small perturbation would be hidden. Finally, we obtained a working set of 129 PDB codes in the mutated dataset with experimental ΔΔG values reported in Protherm for a sequence 100% identical to one of the mutated dataset but possibly from another PDB code.

Results

Estimation of the magnitude of the global distortion

The histogram of the global RMSD (therefore for a window of the length of each chain) between all pairs of proteins from the 11 clusters of the mutated dataset is represented in Figure 1a. The mean value of the global RMSD for this dataset is 0.36 Å and the values range from 0.1 Å to 2.1 Å. This is coherent with the results by Bordner and Abagyan [38] who obtained a mean RMSD of 0.21 Å after removing the pairs with an RMSD higher than 0.5 Å because they wished to consider only displacements attributed to mutation, and not to other factors. We did not use such a threshold because 1) it would remove the cases where a mutation has a large effect that is quite interesting and 2) we used other methods (p-values and random distributions) to remove the other factors unrelated to mutation. To estimate the magnitude of the global structural perturbation without mutations, Figure 1b shows the global RMSD distribution for the five clusters of the identical dataset (grey in Figure 1b). The mean RMSD is higher, at 0.62 Å. Therefore, high perturbations can occur in some proteins, even in the case of strictly identical sequences. Since all clusters are not present in both datasets, the five clusters with same folds as in the identical dataset are extracted from the mutated database. This subset of 5 clusters of common reference is presented Figure 1b by the dashed

curve. The mean RMSD of this subset is rather similar (0.33 Å). Actually, all the proteins with a global RMSD higher than 1 Å in Figure 1b originate from the cluster of reference 4fi8 (transthyretin). The high mean value of structural displacements among transthyretins may be due to the presence of a ligand in the identical dataset. Interestingly, mutation perturbation is reduced, and only five tranthyretins contain a ligand (1iii, 1etb, 3a4e, 3i9a and 4fi8) in the mutated dataset. There may be several reasons one can advance to explain the global higher RMSD of the identical proteins: slightly different ligands, crystallisation conditions, pH, alternative side chain conformation… These effects have already been highlighted by Kosloff et al. [4]. All these arguments can be put under the topic of environmental reasons and only a thorough and detailed analysis, beyond the scope of this paper, would decide between options. Nevertheless, for our study, only 79 of 580 proteins of the mutated dataset have a RMSD greater than 0.5 Å so we can assume that we have minimized the effect of these factors.

From the smaller mean value of the global RMSD in the mutated dataset, compared to the identical dataset, we can conclude that we cannot observe a strong effect of the mutations on whole structures. This is in line with the literature on a different dataset, with 2,141 pairs of completely unrelated proteins [38]. We have then conducted a thorough analysis of the local perturbations of punctual mutations on the backbone by considering only RMSD calculated on n adjacent residues, n ranging from 1 to 20 residues.

Localisation of the distortion

We first wish to know whether the RMSD-n(G/L) at the position of the mutation significantly deviates from the RMSD-n(G/L) at un-mutated positions. We show in this section only the RMSD-3G analyses, but the conclusions about the localisation of the distortion are the same whatever the window size. The RMSD-3G ranges from 0.01Å to 12.8Å (this high value is computed for 3nrb at the N-terminal) with a mean of 0.25Å and a median of 0.19Å (see Table II). The interpretation of the RMSD-3G is rather difficult because

its amplitude can significantly differ between structures determined in various experimental conditions. If we select the top 5% RMSD-3G, we favour mutated proteins with large mean RMSD-3G, instead of deciphering a set of specific positions of high relative distortion. The correlation between the number of the top 5% selected RMSD-3G for each protein and the mean of the RMSD-3G of the protein is 0.72: selecting high value RMSD-3G favors proteins which structure is more globally different from the reference. The p-value removes this bias. If we now select the 5% smallest p-values, then the correlation coefficient is -0.06. Therefore, we do not select positions in proteins that have been globally perturbed, but also positions in proteins that are globally very structurally similar. We believe this effect is quite interesting because experimental conditions of structure determination as pH or temperature are different and may have a global effect on the structure.

For a given protein, the highest RMSD-3G can either be located at the position of the mutation, or elsewhere along the sequence. These two possibilities are shown for the example of transferases in Figure 2. The highest RMSD-3G is located at the very position of the mutation in 2e8r, but not in 2dv7. In the entire dataset, the highest RMSD-3G is only located in 9 cases at the very position of the mutation. Nevertheless, the empirical p-values and their distribution at the mutations are very different from the random selection of the RMSD-3G. The lowest 10% of the p-value distribution (Figure 3) contains 162/580 (28%) RMSD-3G for the mutated positions compared to 9781/116000 (8%) RMSD-3G for the random selections. It means that the RMSD-3G at a given position are often greater with a mutation than without (Figure 3).

Window size effect

To investigate the effect of the window size, we have computed the cumulative distance (or the surface) between the diagonal of the p-values against rank plot and the p-value curve for all windows ranging from 1 to 20 residues for RMSD-nG and ranging from 3 to 20 residues

for RMSD-nL (for 1 and 2 residues, the RMSD is of course null after local superimposition). These values are plotted in Figure 4 as a function of the window size for all the mutations and for 3 subsets of the mutated dataset according to the secondary structure of the mutated residue. The largest surface, i.e. the largest effect of the mutation on the structure as measured by the RMSD-nG, is reached for n=2 (0.15) (Figure 4 black continuous line). This maximum is smaller amplitude for regular secondary structures than for loops. The surface is close to the maximum for n=3 table III and (0.15, Table III) and its decrease is clear from 4 residue windows. In order to keep coherence with other studies (in particular the backrub moves) and because there is no significant between windows of 2 and 3, we decided to do all the calculations with the 3 residues window for RMSD-nG. For the RMSD-nL, a window of size 7 produces the largest effect (grey continuous line, value of 0.15, see table III), and this maximum also depends upon the secondary structure type. The effect is higher for loops (dashed dotted line) than for regular secondary structures, for any window size, as it was the case for RMSD-G. Therefore it was decided from hereon to use a 7 residue window for local RMSD, RMSD-7L.These windows being chosen, it results that the absolute values for RMSD-7L are smaller than those for RMSD-3G (see table II) because the fragment length upon which the RMSD-7L is calculated is exactly the one over which the superposition is performed. RMSD-7L values range from 0.0Å to 3.08Å with a mean of 0.11Å and a median of 0.09Å.

Comparison with random distributions of RMSD

The effect of the mutation on the structure being very small for a local measurement, we raised the question whether the perturbation lies within the statistical variations or not. Therefore, the RMSD-3G and RMSD-7L between all pairs of the mutated dataset have been converted into p-values and sorted by increasing values. They are plotted in Figure 5 (black dots) for the mutated dataset, with all the random distributions of p-values as grey lines. To

understand the p-value plots as a function of their ranks (Figures 4, 5 and 6), one must remember that the greater the RMSD, the smaller its rank, and the smaller its p-value. The diagonal of this plot represents a regular constant increment from one position to the next one. It somehow constitutes a neutral hypothesis of a uniform distribution of RMSD. When the p-value curves of random draws within mutated positions are slightly above (resp. below) this diagonal (see for example Figure 5a, grey curves), then the corresponding random RMSD are smaller (resp. greater) than expected from a neutral assumption. If the actual mutated RMSD (see for example black curve, Figure 5a) are smaller (greater) than the random ones, the resulting p-values will be above (under) the random distribution of p-values.

For the whole mutated dataset, the random p-values (corresponding to particular draws at mutated positions) are slightly above the diagonal for the RMSD-3G, which means that the associated RMSD-3G are slightly smaller than expected from an increasingly uniform distribution (Figure 5a). This effect is similar for the RMSD-7L since the distance of the mean random to the diagonal is -0.04 for RMSD-3G and -0.01 for RMSD-7L (Figure 5b). The curve of the sorted p-values for the RMSD-3G or RMSD-7L at the mutated positions (black curves) is under the diagonal, indicating that these RMSD are statistically greater at the place of the mutation compared to all other positions. Nevertheless, they must be analysed not as a crude value, but instead by comparison to the set of random curves. This is the major interest of our method, which allows such comparisons in dealing with large families of single mutations. The actual distributions at the mutated positions lie outside the random curves, indicating that the structural perturbations due to mutation are not of the same amplitude as the variations due to statistical perturbations, regardless of the type of superimposition.

S*econdary structure*

We have further analysed our results by categorizing the positions of mutations according to the type of the secondary structure where they are located. Results are presented in Figure 6

for strands, helices and other (mainly loops) for RMSD-3G. In both cases, the area between the diagonal and the p-values sorted according to their rank is presented in Table III and split into various characteristics (secondary structures and level of solvent exposure). The random distributions of p-values go closer to the main diagonal from strands to helices and loops. It indicates that these respective locations are more and more flexible and mobile as it is already established [55–57]. The actual experimental distribution of displacements at the mutated positions, as measured by the p-values, remains in all three cases smaller than random and under the diagonal, which corresponds to a higher RMSD-3G. The distortions are greater in loops than in helices and strands (the distance to the diagonal is greater, see Figure 6 and table III). The areas to diagonal going from strands to helices, and to loops are respectively 0.09, 0.12, 0.21. Nevertheless one cannot conclude that the effect of mutations is of higher effect in loops than in regular secondary structure. Actually, one needs to calibrate the RMSD-3G and RMSD-7L in regard to mutation independent fluctuations at these positions. If one calculates the distance between the experimental curve and the mean random curve, it decreases for RMSD-3G when going from strands to helices, and to loops: the absolute values are respectively 0.26, 0.18 and 0.17 (Table III). In other words, the relative effect of a mutation is more important in positions of strands than in the flexible loops but the RMSD-3G in the loops are greater. The order of the areas between the experimental and the mean random are different for RMSD-7L: 0.18 for strands, 0.11 for helices and 0.20 for loops. The areas are then higher for loops than for regular secondary structures with a local superimposition and it is the contrary with a global superimposition. This is coherent with the fact that values of the pseudo dihedral angles in alpha helices occur in a smaller range than for beta strands[58], which means that helices are more locally rigid than strands and than loops.

Accessibility to the solvent

When p-values are analysed according to solvent accessibility of the mutated positions, the importance of the perturbation is clearly not uniform, as it is presented in Figure 7. The random RMSD-3G distribution expressed as sorted p-values (Figure 7a, in grey) for buried residues is clearly above the diagonal, which means that the RMSD-3G of buried segments are among the smallest RMSD-3G. On the contrary, the random RMSD-3G distribution in the RMSD of sorted p-values for exposed residues (Figure 7b, in grey) lies clearly under the diagonal: the RMSD-3G are then among the largest. In the two cases, the actual mutated RMSD-3G are under the diagonal, and even if the distance to the diagonal is greater for the exposed residues, the distance to the mean random curve is smaller for exposed than for buried residues (surfaces of 0.16 for exposed residues and 0.21 for buried residues, Table III). In other words, the relative effect is smaller for exposed residues than for buried ones even if the RMSD-3G are bigger in exposed residues. This effect may be explained by an effect of excluded volume on the surface. This is also coherent with the vision of an outer shell of globular proteins more flexible than the core [59]. The effect is the same but slightly smaller in the case of local superimposition (Table III).

Energy

One now may ask whether the structural distortion introduced by one single point mutation is important enough to be significant in the variation of the free energy change. This needs checking if there is any correlation between structural displacements and variation in free energy. From the mutated database, the correlation coefficient between ΔΔG and RMSD-3G has been computed, and it is 0.03. The correlation is -0.06 between the ΔΔG and the p-value. This absence of any significant correlation has been confirmed by taking the experimental ΔΔG published in the ProTherm database for proteins that are present in the mutated database (129 entries). A linear fit regression has been calculated between RMSD-3G and predicted ΔΔG on the 129 common entries and the correlation coefficient is 0.14. Both correlation

coefficients are not significant according to the Pearson correlation coefficient test. This result is not surprising as the group of Brian Matthews[60] who studied the effect of single mutations within the core of T4 lysozyme, concluded that the free energy of unfolding relative to wild type did not correlate with the packing density, defined as the number of methyl and methylene groups within 6 Å of the removed atoms, and therefore related to the adjustment of the backbone.

One may now wonder whether some correlation would not exist between the RMSD-3G and the errors on the ΔΔG prediction. The error on ΔΔG prediction is defined as the difference between the ΔΔG calculated by FoldX and the experimental one reported in Protherm. Then, the 129 entries set have been split in two groups of equivalent populations, according to the RMSD-3G at the position of the mutation. It results in 64 RMSD-3G assigned as "low" and 65 assigned as "high" values. Then, for each group, the mean error over the ΔΔG calculation is performed: 2.00 kcal/mol for the low RMSD-3G and 3.13 kcal/mol for the high RMSD-3G group. A Student t-test at 5% insures that these two values are significantly different. In other words, one can conclude that the larger the effect of the mutation on the structure, the larger the error in the prediction of ΔΔG.

Sequence neighbourhood

We have investigated two kinds of neighbourhood: in sequence and in 3D-structure. Let us start to examine the effect of the sequence neighbourhood. After a global superposition, we have calculated the RMSD-3G after shifting the window up to 20 residues on both sides of the mutated residue. This has been done for all mutations of the mutated database. We present in Figure 8 the distance to the diagonal for the random (points and grey area) and the mutated (continuous line) p-values as a function of the shift of the position where the RMSD-3G is calculated for helices; strands and loops. For the strands, the effect of the mutation propagates up to the third neighbours. It is noticeable that this effect is of the same magnitude at the very

location of the mutation and for the first backbone neighbours in each direction. The damping is much smaller for helices and loops since the effect is still appreciable up to 5 residues. One also notices that the highest effects are at the point of mutation and on the first neighbours.

3D space neighbourhood

In seminal papers, the group of Matthews concluded on the case of T4 lysozyme that the major effect of the mutation does not necessarily occurs at the mutated position, after a superposition performed on half the protein [54]. We wished to study long-range effect defined as a displacement of two positions distant along the sequence, but close in the 3D space. To this goal, we have calculated the RMSD-3G for the central alpha carbons of all residues in contact with the mutated ones that are not adjacent on the sequence. It thus collects all amino acids in a sphere centred on the mutated residue that are at long range in sequence. There is actually an effect of the mutation on the backbone of these long-range residues but it is weaker than for the mutated residue and its adjacent residues. The p-value curve is around the diagonal (distance to the diagonal of -0.016), meaning that the RMSD-3G are not globally greater than the other RMSD-3G of the protein pairs. The random p-values (drawn from the set of all RMSD-3G at 3D neighbour positions by the same procedure as for mutated positions) are above the diagonal and the area between the mean random and the diagonal is 0.08. This means that the RMSD-3G for these positions are affected by their contact to a mutated position. This small effect is in phase with 2D lattice simulations by Liu et al. that have concluded to an exponential decay of the coupling between residues when their separation increases [61].

P-RANK

As explained in methods, we have computed the p-value at every position of every protein and we have calculated an empirical p-value named P-RANK. This P-RANK allows us to take into account this difference of RMSD-3G or RMSD-7L due to the 3D localization. As we

have seen previously, the RMSD-3G and RMSD-7L are quite different for buried or exposed positions, or for positions in strands, helices or loops. In the whole mutated dataset (*i.e*. all positions regardless of a mutation), 18% of residues are in strands, 42% in helices and 40% in loops; there are also 50% of buried residues and 50% of exposed. This over representation of helices is due to the high number of proteins from alpha class. If we select the 5% smallest p-values computed with RMSD-3G, the proportions are: Strands: 8%; Helices: 30%; Loops: 62%; Buried: 33% and Exposed: 67%. There is a bias in this 5% selection because the RMSD-3G is smaller for positions in strands, in helices or buried; consequently, the p-values are bigger and less selected. If we now select the 5% smallest P-RANK, the proportions are: Strands: 17%; Helices: 44%; Loops: 39%; Buried: 49% and Exposed: 51%. It is clear that the bias due to 3D localisation is removed by using the P-RANK method.

We can now look at the mutated positions among those 5% smallest P-RANK. There are 144 significant mutations (110 if P-RANK are computed from RMSD-7L). In the mutated positions, the distribution gives: Strands: 18%; Helices: 44%; Loops: 38%; Buried: 56%, Exposed: 44%. Among this set of positions, if we restrict to the significant ones according to the P-RANK-3G, considering both mutated and native residues, the proportions are: Strands: 21%; Helices: 53%; Loops: 26%; Buried: 69%; Exposed: 31%. We can conclude that even if the RMSD-3G is larger in loops and exposed positions, the relative effect of a mutation is more important in regular secondary structures and buried positions if one considers instead the P-RANK. If P-RANK are computed from RMSD-7L the proportions are: Strands: 17%; Helices: 51%; Loops: 32%; Buried: 66%, Exposed: 34%. Both RMSD-3G and RMSD-7L conclude to the reduction of significant displacements in loops. So we decided to keep both RMSD-3G and RMSD-7L because they capture different mutated positions: 74 are below the 5% threshold with RMSD-7L only and 108 with RMSD-3G only.

If we plot the P-RANK as a function of the RMSD-3G (Figure 9), the 0.05 threshold in the P-RANK (represented by the horizontal line) allows considering RMSD as small as 0.09Å value that would have been disregarded as below most of the thresholds in the literature (for example 0.2Å with Schaefer and Rost pentamers[7]). This proves the interest of the use of P-RANK instead of RMSD as it is illustrated in Figure 9. P-RANK allows considering as significant displacements mutations that give rise to very small RMSD.

We present here 3 examples to illustrate the effect of the secondary structure location of the mutation. The first case is a mutation in the Anhydrase II (PDB code 4jswA, mutation H94C), located in a strand. The RMSD-3G is 0.8Å and the RMSD-7L 0.16Å. Both are significant (<0.05) according to the p-value (0.02 and 0.008 respectively) and are also significant for the P-RANK (0.015 and 0.015). The deviation of the backbone at the very position of the mutation (in red) is clearly visible in Figure 10A and very distant from all other backbones of proteins which are not mutated at this position (in grey). This is an “easy” case. The second case (PDB code 2pb6A, mutation Y220C) is less easy because the RMSD-3G is smaller: the values are 0.29Å for the RMSD-3G but 0.20Å for the RMSD-7L. However, the P-RANK are both significant (0.013 and 0.04) because the RMSD of the proteins of the family are small at this position and the deviation is large relatively to the other proteins which are un-mutated at this position. In Figure 10B, the backbone of 2pb6A (in red) is visible and not buried in the other backbones of the family, but its distance with the other backbones is smaller than in the first case. In the last case (PDB code 2egbA, mutation E140N), both P-RANK are not significant even if the p-values are below 0.05: 0.03 for a RMSD-3G of 0.25Å and 0.02 for a RMSD-7L of 0.16Å. As the mutation is in a loop, such RMSD are frequent at this position and the P-RANK are ranked 19th and 5th (P-RANK of 0.22 and 0.06) among the p-values of the family at this position. In Figure 10C, the red backbone of the 2egbA is buried

among the other backbones and sometimes visible. These examples and the previous global results show that our method is very sensitive and accurate to detect perturbation of the backbone, even if the global structure moves or if the perturbations are in the core of the protein and very small in RMSD absolute value.

School case: lysozyme

We show in Figure 11 the trace of the alpha carbons for the Human Lysozyme family (reference 2nwd). There are 123 mutants superimposed on the reference protein. All the alpha carbons at positions where there was at least one mutation are represented by spheres. Two colours are used for the spheres: red if the alpha carbon is the one of a mutated residue; blue if it is not mutated. In all positions with spheres, there is at least one red sphere, even if it is hidden by the blue ones. This is intended to emphasize the importance of the displacement due to the mutation: the red spheres often lie outside of the cloud of the blue non-mutated alpha carbon. We can also note that the alpha carbon cloud is wider in the loops than in regular secondary structures.

Discussion

The effect of a single point mutation is tackled in this paper, in order to evidence the contribution of the side chain replacement, either locally or at long range, on the backbone displacement. A dataset has been built of eleven families containing each one more than 20 protein structures, differing one from the other by a single mutation. Previous studies, for instance the one by the Abagyan group [38], used a larger number of proteins, but the clusters were reduced to pairs of proteins, the native and one mutant. The size of the clusters is important to bypass the bottleneck of pair comparisons that are too small to use them to calibrate backrubs motions [62]. Backrub is defined in the literature as the relaxation of the main chain under the effect of side chain multiple conformations, as detected from high resolution

structures [34, 36]. To separate the effect of mutation on the backbone adjustment due to side chain diversity of conformations (backrub motion), from other effects, a database was also built, composed of clusters containing only structures with sequence perfectly identical. The comparison of both data sets on the full length of the chains results in an average move that is higher in the identical dataset (mean global RMSD 0.62 Å) than in the mutated one (mean RMSD 0.36 Å). In other words, adjustment of the backbone under the effect of a point mutation is smaller, on the average over all proteins belonging to both datasets, than the adjustment due to fluctuations of conformation of the side chains. Kosloff et al. found numbers of protein pairs with sequence identity in the range 50-100% that had dissimilar structures, with global RMSD greater than 3Å [4]. For sequence pairs at 100% identical, one can still find a significant amount (more than 20) of pairs with a global RMSD of 7Å. These authors also hypothesized several causes of structural differences, among which we can retain four: 1) ligand binding; 2) crystallization conditions; 3) alternative conformations; 4) point mutations. Comparison of the two datasets used in our study and keen analysis of the ligand presence allow focusing on the two remaining factors.

To further analyse the small amplitudes of the mutation induced moves, comparisons must be performed on short fragments of the chains. Two ways of superimpositions are then explored: one over a 7-residue window centred on the mutation (called local), and one over the full length of the protein (called global). The structural comparison is measured as an RMSD computed on a 7 or 3 residue window (RMSD-7L or RMSD-3G). This length of the fragment along which the RMSD-3G is calculated is consistent with the definition of a backrub motion [34]. A global superimposition is expected to enhance the effect of the perturbation in cases where it is limited on a fragment of small size. This is why, for most of the results presented in this study, global superposition has been preferred.

The window size effect has been explored and the backbone disturbance at the position of the mutation are best captured with a window of size 2 or 3 after a global superimposition and a window of size 7 after a local superimposition of the 7 Ca in the window. It is interesting to note that the effect is stronger with a window of size 5 for strands, which is the mean length of a strand, and with a window of size 8 for loops and helices for RMSD-7L.

The displacements in the backbone, as measured by the RMSD-3G and RMSD-7L relative to the native structure, are higher than expected at the very position of the mutation, but not the highest among the RMSD of the considered protein pair.

To separate between mutation induced and unrelated backbone moves, a threshold in the RMSD-3G or RMSD-7L would have no absolute meaning. To account for the various conditions of experimental determination present in large families with single mutations among members, a high value of RMSD may be due to chance because all the alpha carbons have moved.

We have developed a method to study small local variations of the backbone by calculating two empirical p-values from the RMSD values. This method does not fix an RMSD threshold but takes into account the variability of the RMSD inside the two aligned proteins and also the variability of the RMSD inside the family at the same position. We first calculate the ratio of the rank of the RMSD among the RMSD for the considered protein pair divided by the number of RMSD for this pair. This value is an empirical p-value. Second, we compute the rank of this p-value among the p-values calculated for all the proteins of the family at the same position. This value is also an empirical p-value and we name it P-RANK. The p-value allows us to remove the bias due to RMSD amplitude variability between the proteins due to global structural variations. The P-RANK allows us to remove the RMSD amplitude variability due to 3D localisation (the RMSD of buried residues are smaller than those calculated for exposed residues).

In the paper from Schaefer and Rost [7], by comparing pairs of pentamers, it is admitted that an RMSD above 0.4 Å is an evidence of a local change in three-dimensional structure. If it is below 0.2 Å, then the mutation is considered as not altering the structure [7]. This value may be seen as a lower limit under which an RMSD can be considered as negligible between two structures. But, if a small mutation induced RMSD is analysed by comparison with a random distribution of RMSD-3, then one can conclude even in the case of much smaller RMSD values. This is shown in Figure 5, where the actual distribution of RMSD at the mutated positions in the whole dataset, translated into sorted p-values, are clearly distinguished from the random distribution of RMSD due to other causes than mutations. The random distributions are calculated only at the positions where at least one of the chains of the cluster is mutated. The actual distribution of p-values lies out of a random selection in all localisations of the mutation: in helices, strands, coils, at buried or accessible positions. The RMSD are more important in loops than in regular secondary structures. It is coherent with the fact that moves are more important for exposed than for buried residues; this actually matches the distribution of the various kinds of secondary structures. However, we found that relatively to variations of un-mutated positions, the mutations have a greater effect for the buried residues and in the strands.

Concerning the extension of the effect of the mutation, it can be estimated from two points of view. The diffusing effect along the backbone is restricted to less than five residues around the mutated position; actually it is depending upon the nature of the secondary structure where the mutation occurs. To evidence this fact, considering a large family of structures with one single mutation in each entry is an advantage compared to a study performed only on pairs of mutated structures. The long-range effect of the mutation can also be evidenced on residues in 3D neighbourhood of the mutation, but it is seriously damped compared to the effect at the position of the mutation.

Most of the free energy difference calculation (ΔΔG) make the assumption that the free energy difference between the unfolded states of the mutated and native structures are generally assumed to be equal [38]. On the mutated dataset, evaluation of ΔΔG by Fold-X does not show any correlation with the movements that the backbone completes to accommodate the substitution of one side chain. The relation between mutation and stability evaluated by means of ΔΔG is difficult and Funahashi et al. concluded to the absence of relation between stability and structure [63]. Nevertheless, it is interesting to notice that the error in ΔΔG, i.e. the difference between predicted (Fold-X) and experimental (Protherm) ΔΔG values, increases when the RMSD between mutated and native structures becomes more important. This may help to give a range of acceptable quality for the algorithms dedicated to the prediction of the stability upon a point mutation.

Conclusion

In this paper, we proposed to consider the problem of the effect of a point mutation on the structure of a globular protein. Eleven families of at least twenty members with one single mutation compared to the native protein were retrieved from the PDB. We hope that our dataset will help to improve the predictions of the impact of punctual mutations. We also have developed a method in order to measure the effect of a mutation even if the structures are determined in different conditions or if the mutation is localised in the rigid core of the protein. We hope that this method will allow giving a better insight in the consequences of substitutions on the backbone and the whole protein structures.

It is important to have in mind that our study only takes into account the amplitude of the backbone deformation but not the direction of it. Therefore, even if the deformation at a given position is of the same amplitude for the mutated protein and for all the un-mutated others, it is possible that the direction of the deformation is different. Therefore, it will also be

interesting to take in our study all the mutated positions and compare their directions to the un-mutated ones. We will also investigate the backrub model or the Abagyan and al. model for the modelisation of the variations of the backbone at the mutated positions.

We did not study here the dynamical aspects of the structures. In particular, long range perturbation, such as the propagation of the destabilisation into the second shell, at the origin of the allosteric effects, that have been evidence by [64]. We will further investigate these effects by analysing all significant RMSD whether the positions are mutated or not.

It could be worth using NMR data and also to extend this study to double mutations. Is there a compensation that occurs or on the contrary is there a synergic effect? These aspects need to be further investigated.

Acknowledgements

This work has been partially funded by a grant from the ANR Tempomut (ANR-12-JSV7-0007). Are particularly acknowledged: Guillaume Achaz on the way to drive statistical analysis; Joel Pothier for energies interpretation; Marc Delarue for PDB files missing atoms correction; Marc Baaden and Raphael Guérois for free energy computation.

The authors recognize they have no conflict of interest.

Availability

The PDB codes for both the datasets used in this study are available upon request to the authors.

Bibliography

1. Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem J 2013;449(3):581–594.
2. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. Nat Rev Mol Cell Biol 2009;10(10):709–720.
3. Bartlett GJ, Borkakoti N, Thornton JM. Catalysing New Reactions during Evolution: Economy of Residues and Mechanism. J Mol Biol 2003;331(4):829–860.
4. Kosloff M, Kolodny R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins Struct Funct Bioinforma 2008;71(2):891–902.
5. Sinha N, Nussinov R. Point mutations and sequence variability in proteins: redistributions of preexisting populations. Proc Natl Acad Sci 2001;98(6):3139–3144.
6. Religa T, Markson J, Mayor U, Freund S, Fersht A. Solution structure of a protein denatured state and folding intermediate. Nature 2005;437:1053–1056.
7. Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. BMC Genomics 2012;13 Suppl 4:S4.
8. Shakhnovich EI, Gutin AM. Influence of point mutations on protein structure: probability of a neutral mutation. J Theor Biol 1991;149(4):537–546.
9. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
10. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
11. Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A 2004;101(25):9205–9210.
12. Privalov PL. Stability of proteins: small globular proteins. Adv Protein Chem 1979;33:167–241.

13. Gromiha MM, Sarai A. Thermodynamic Database for Proteins: Features and Applications. In: Carugo O, Eisenhaber F, editors. Data Mining Techniques for the Life Sciences. , Methods in Molecular Biology. Humana Press; 2010. p 97–112.
14. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 2005;6(9):678–687.
15. Taverna DM, Goldstein RA. Why are proteins marginally stable. Proteins 2002;46:105–109.
16. Zeldovich K, Chen P, Shakhnovich E. Protein stability imposes limits on organism complexity and speed of molecular evolution. PNAS 2007;104:16152–16157.
17. Wylie CS, Shakhnovich EI. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci U A 2011.
18. Bloom JD, Raval A, Wilke CO. Thermodynamics of Neutral Protein Evolution. Genetics 2007;175(1):255–266.
19. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, Koning APJ de, Dokholyan NV, Echave J, Elofsson A, Gerloff DL, Goldstein RA, Grahnen JA, Holder MT, Lakner C, Lartillot N, Lovell SC, Naylor G, Perica T, Pollock DD, Pupko T, Regan L, Roger A, Rubinstein N, Shakhnovich E, Sjölander K, Sunyaev S, Teufel AI, Thorne JL, Thornton JW, Weinreich DM, Whelan S. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci Publ Protein Soc 2012;21(6):769–785.
20. Guerois R, Nielsen J, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. JMB 2002;320:369–387.
21. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306–W310.

22. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 2009;25(19):2537–2543.
23. Laimer J, Hofer H, Fritz M, Wegenkittl S, Lackner P. MAESTRO--multi agent stability prediction upon point mutations. BMC Bioinformatics 2015;16:116.
24. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 2006;34(suppl 2):W239–W242.
25. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30(3):335–342.
26. Giollo M, Martin AJM, Walsh I, Ferrari C, Tosatto SCE. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. BMC Genomics 2014;15 Suppl 4:S7.
27. Quan L, Lv Q, Zhang Y. STRUM: Structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016:btw361.
28. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins Struct Funct Bioinforma 2011;79(3):830–838.
29. Lonquety M, Chomilier J, Papandreou N, Lacroix Z. SPROUTS: a database for evaluation of the protein stability upon point mutation. Nucleic Acids Res 2008;37:D374–D379.
30. Gautier R, Tufféry P. Critical assessment of side chain conformational space sampling procedures designed for quantigying the effect of side chain environment. J Comput Chem 2003;24:1950–1961.
31. Dunbrack R. Rotamer libraries in the 21st century. COSB 2002;12:431–440.

32. Lee C, Levitt M. Packing as a structural basis of protein stability: understanding mutant properties from wildtype structure. Pac Symp Biocomput Pac Symp Biocomput 1997:245–255.

33. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: the endgame. Annu Rev Biochem 1997;66:549–579.

34. Davis IW, Arendall III WB, Richardson DC, Richardson JS. The Backrub Motion: How Protein Backbone Shrugs When a Sidechain Dances. Structure 2006;14(2):265–274.

35. Lauck F, Smith CA, Friedland GF, Humphris EL, Kortemme T. RosettaBackrub—a web server for flexible backbone protein structure modeling and design. Nucleic Acids Res 2010;38(suppl_2):W569–W575.

36. Smith CA, Kortemme T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. J Mol Biol 2008;380(4):742–756.

37. Dunbrack RL, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1994;1(5):334–340.

38. Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 2004;57(2):400–413.

39. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank archive as an open data resource. J Comput Aided Mol Des 2014;28(10):1009–1014.

40. Wang G, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33(suppl 2):W94–W98.

41. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150–3152.

42. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 2013;30(4):772–780.

43. Cuff A, Sillitoe I, Lewis T, Redfern O, Garratt R, Thornton J, Orengo C. The CATH classification revisited - architectures reviewed and new ways to characterize structural divergence in superfamilies. NAR 2009;37:D301–D3014.

44. 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, Zagury J-F, Tufféry. P. RPBS: a web resource for structural bioinformatics. Nucleic Acids Res 2005;33:W44–W49.

45. Smith CA, Kortemme T. Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. PLOS ONE 2011;6(7):e20451.

46. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, Hoon MJL de. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25(11):1422–1423.

47. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 2009;9:51.

48. Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon J-P. Deciphering protein sequence information through Hydrophobic Cluster Analysis (HCA) : current status and perspectives. Cell Mol Life Sci 1997;53:621–645.

49. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.

50. Azuara C, Lindahl E, Koehl P, Orland H, Delarue M. PDB_Hydro: incorporating dipolar solvents with variable density in the Poisson-Boltzmann treatment of macromolecule electrostatics. Nucleic Acids Res 2006;34(Web Server issue):W38-42.

51. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res 2005;33(Web Server issue):W382-8.
52. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. J Mol Biol 2007;369(5):1318–1332.
53. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How protein stability and new functions trade off. PLoS Comput Biol 2008;4(2):e1000002.
54. Eriksson AE, Baase WA, Matthews BW. Similar Hydrophobic Replacements of Leu99 and Phe153 within the Core of T4 Lysozyme Have Different Structural and Thermodynamic Consequences. J Mol Biol 1993;229(3):747–769.
55. Hrabe T, Li Z, Sedova M, Rotkiewicz P, Jaroszewski L, Godzik A. PDBFlex: exploring flexibility in protein structures. Nucleic Acids Res 2016;44(D1):D423–D428.
56. Znamenskiy D, Chomilier J, Tuan KL, Mornon JP. A new protein folding algorithm based on hydrophobic compactness : Rigid Unconnected Secondary Structure Iterative Assembly (RUSSIA). I : Methodology. Prot Engng 2003;16:925–935.
57. Znamenskiy D, Letuan K, Mornon J, Chomilier J. A new protein folding algorithm based on hydrophobic compactness : Rigid Unconnected Secondary Structure Iterative Assembly (RUSSIA).. II : applications. Prot Engng 2003;16:937–948.
58. Carpentier M, Brouillet S, Pothier J. YAKUSA: a fast structural database scanning method. Proteins 2005;61(1):137–151.
59. Papandreou N, Eliopoulos E, Berezovsky I, Lopes A, Chomilier J. Universal positions in globular proteins : observation to simulation. Eur J Biochem 2004;271:4762–4768.
60. Xu J, Baase WA, Baldwin E, Matthews BW. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. Protein Sci Publ Protein Soc 1998;7(1):158–177.

61. Liu Z, Chen J, Thirumalai D. On the accuracy of inferring energetic coupling between distant sites in protein families from evolutionary imprints: illustrations using lattice model. Proteins 2009;77(4):823–831.
62. Keedy DA, Georgiev I, Triplett EB, Donald BR, Richardson DC, Richardson JS. The Role of Local Backrub Motions in Evolved and Designed Mutations. PLOS Comput Biol 2012;8(8):e1002629.
63. Funahashi J, Takano K, Yutani K. Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? Protein Eng 2001;14(2):127–134.
64. Rajasekaran N, Suresh S, Gopi S, Raman K, Naganathan AN. A General Mechanism for the Propagation of Mutational Effects in Proteins. Biochemistry (Mosc) 2017;56(1):294–305.

Figure captions

Figure 1. Distribution of the global RMSD between all pairs. A): Mutated dataset. B): Grey: identical dataset. Dashed: 5 clusters of the mutated dataset also present in the identical dataset.

Figure 2. Distribution of the RMSD-3G for two cases taken from cluster with 2dek as a reference (transferase): 2e8r (A) and 2dv7 (B). The dashed upper dotted line indicates the location of the mutation.

Figure 3. A) p-value distribution for the RMSD-3G at the mutated positions for the whole mutated dataset. B) Same distribution but for the random selection of RMSD-3G (same positions but not mutated).

Figure 4. Distance to the diagonal of the p-values for the mutated RMSD-nG (black lines) or RMSD-nL (grey lines) as a function of the window size. Continuous lines are for all the mutated dataset; dashed lines for extended; dotted lines for helices and dashed-dotted lines for loops.

Figure 5. Plot of the empirical p-value for RMSD-3G (A) and RMSD-7L (B) as a function of their rank, for all the members of the 11 clusters (dark) and the random distribution (light grey).

Figure 6. Curves of the p-values of the RMSD-3G sorted as a function of their rank for mutations occurring in: A) strands; B) helices; C) loops. Superposition is global in the three cases.

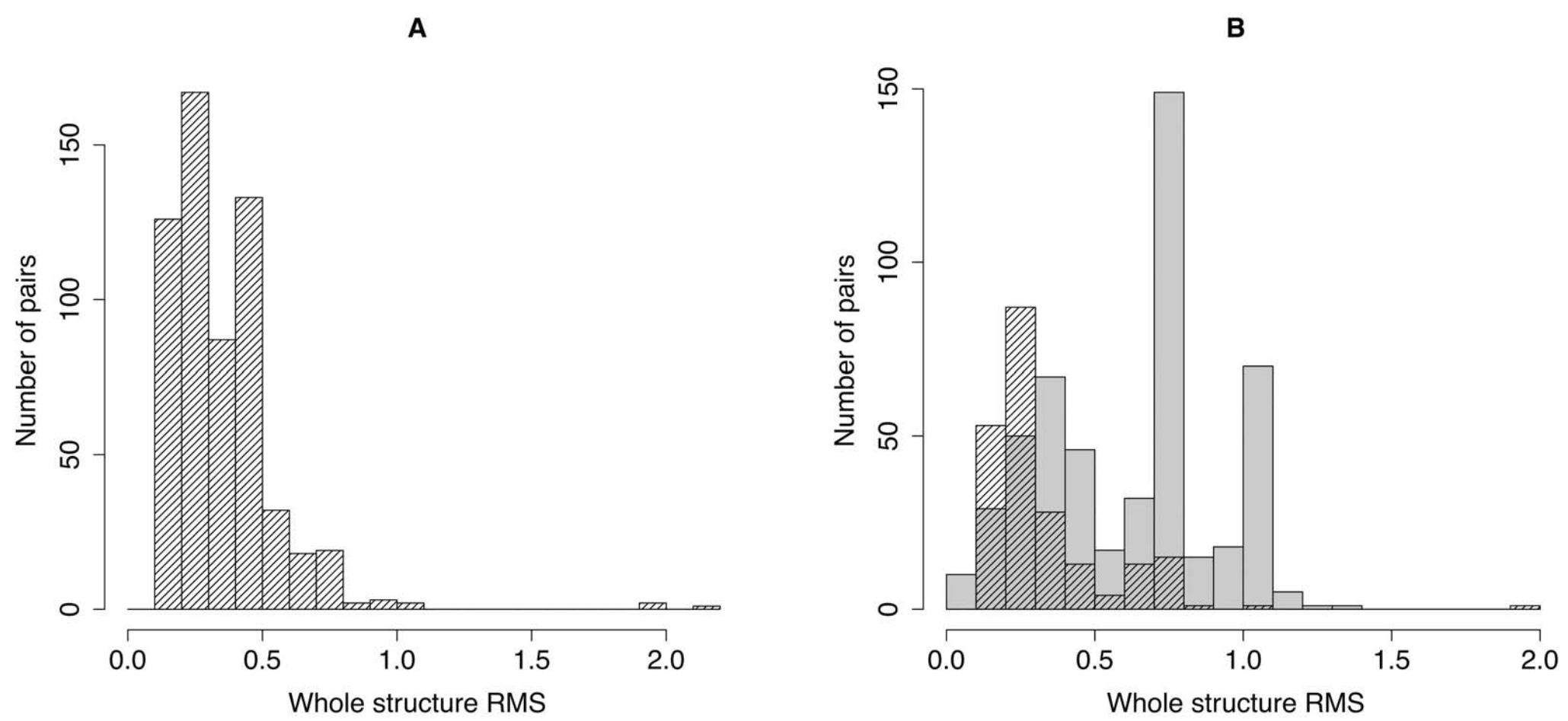
Figure 7. Distribution of the p-value of RMSD-3G as a function of their rank, among all members of the mutated dataset, according to their accessibility: a) buried; b) exposed.

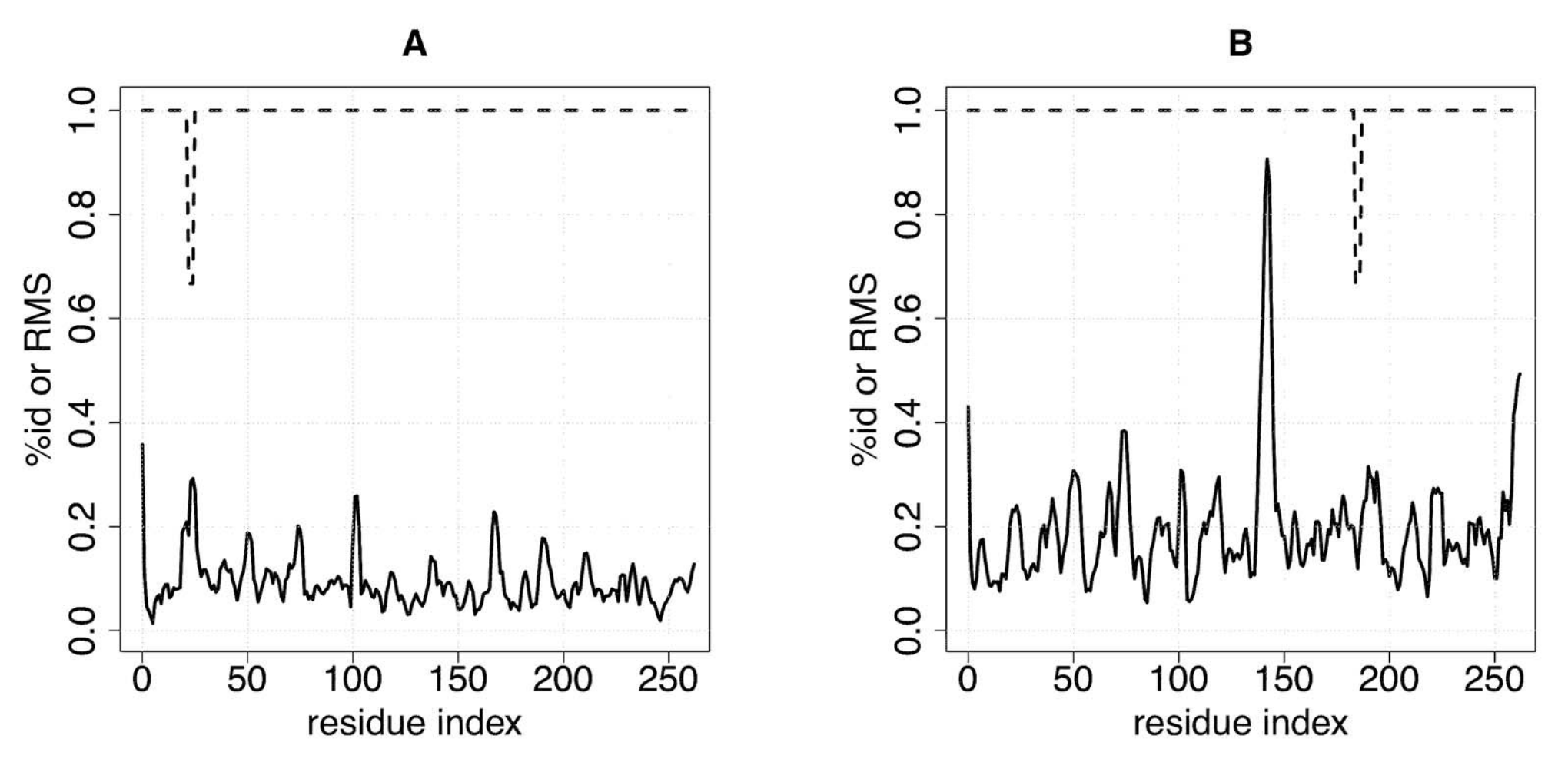
Figure 8. Distance to the diagonal of the p-values for the mutated RMSD-3G (continuous line) and the median random distribution (dashed line), the grey surface represents 95% of the random values. A) strands; B) helices; C) loops.

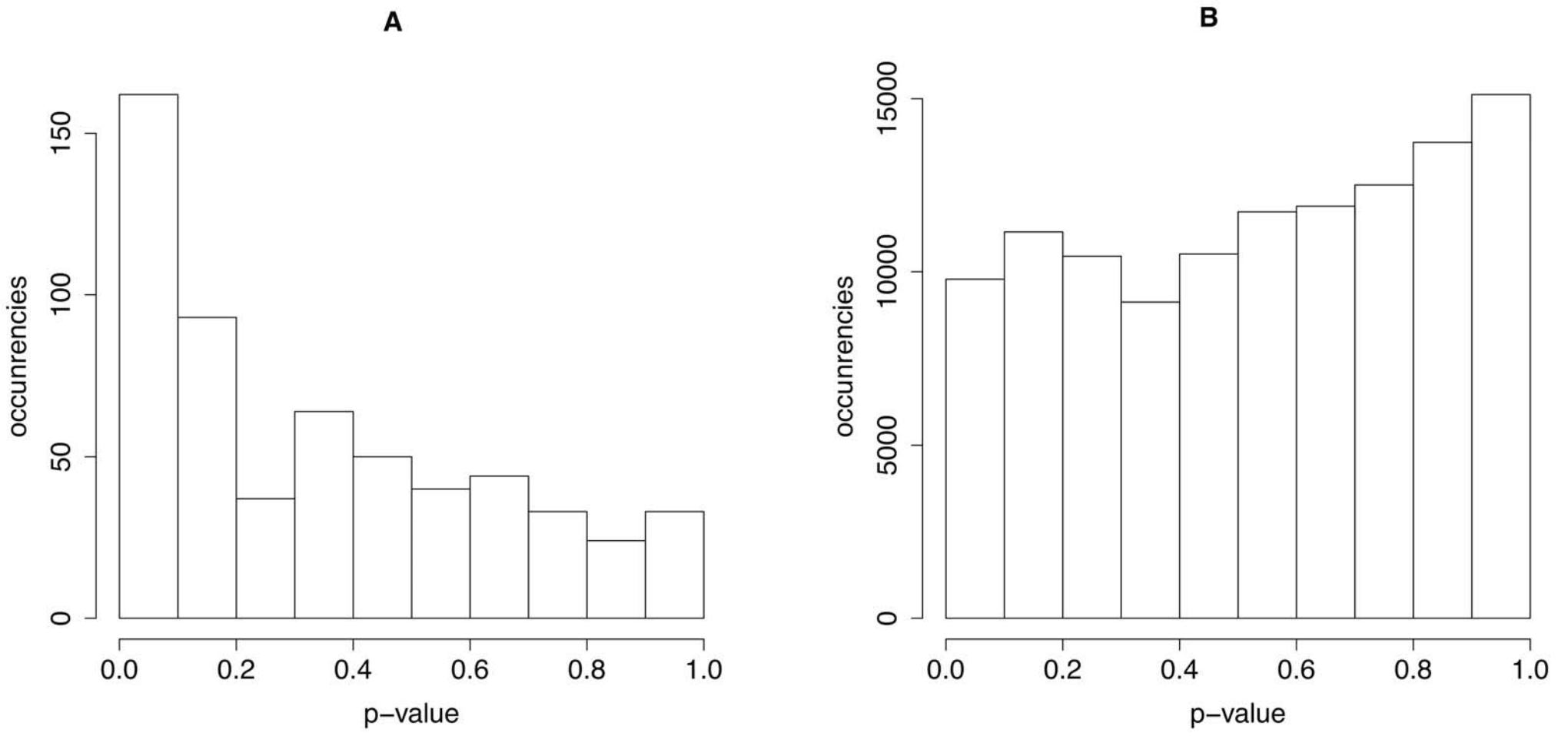
Figure 9. P-RANK for the mutated positions as a function of RMS-3G for the whole mutated dataset. The horizontal line is the 5% threshold of P-RANK. The non-significant displacements according to P-RANK computed from RMSD-3G or RMSD-7L are represented by a grey circle; the significant displacements according to P-RANK computed from RMSD-3G are indicated by black circle and located below threshold; the significant displacements according to P-RANK computed from RMSD-7L are indicated by symbol + and located above threshold; the significant displacements according to P-RANK computed both from RMSD-3G and RMSD-7L are indicated by diamonds and located below threshold.

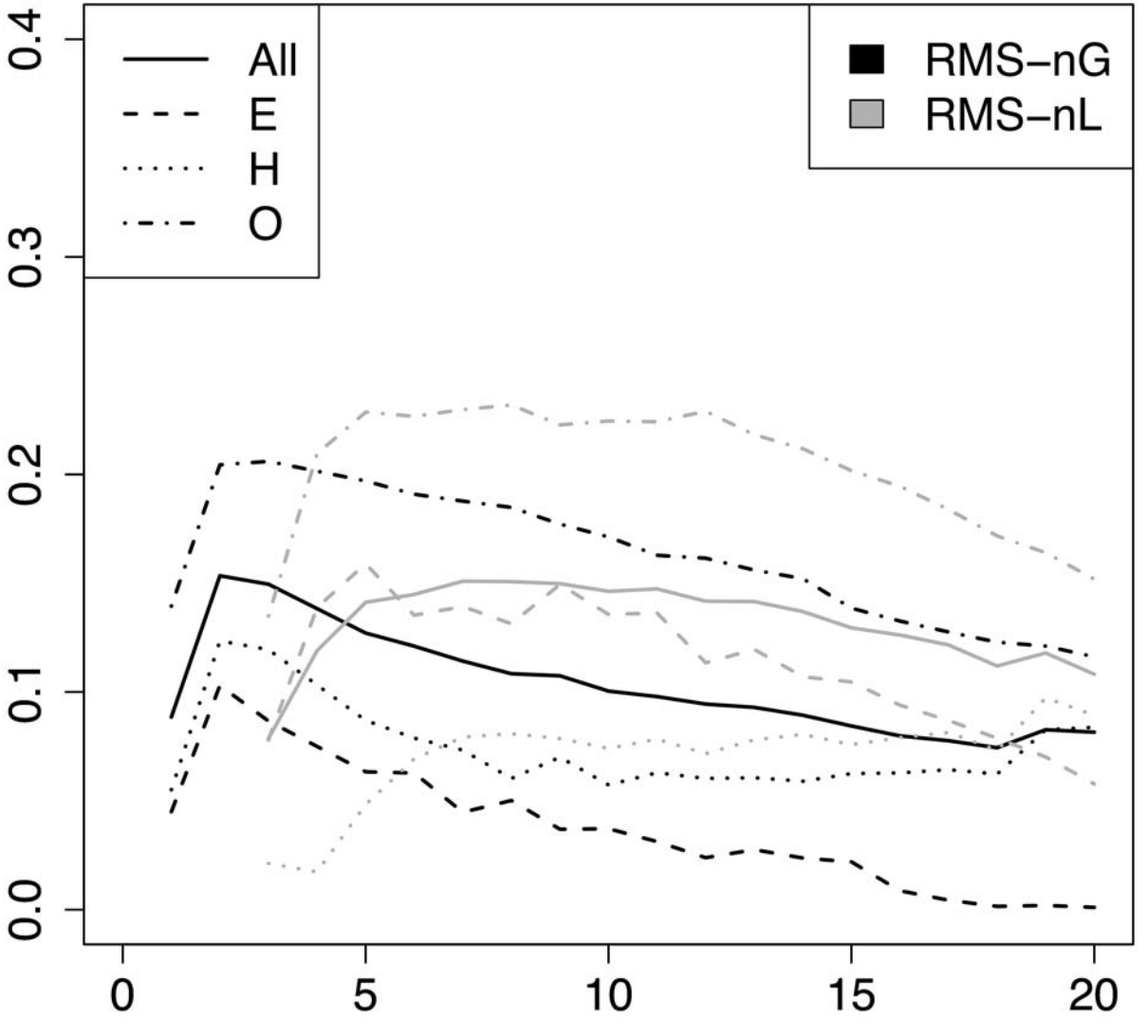
Figure 10. Three examples of mutations. A) 4sjwA (Anhydrase II); B) 2pb6A (transferase); C) 2egbA (transferase). The mutated protein of interest is in red for the main chain and brown for the side chain. The reference protein is in blue. The backbone of all other proteins, which are mutated at another position, are in grey. Backbones are represented by lines. The blue backbone of the reference is buried among all other backbones and is generally not visible, but the blue reference side chain at the position of the mutation is shown. The red backbone of the mutated protein of interest is visible for 4jswA, 2pb6A but is mostly buried among others for 2egbA. Pictures are made with Pymol.

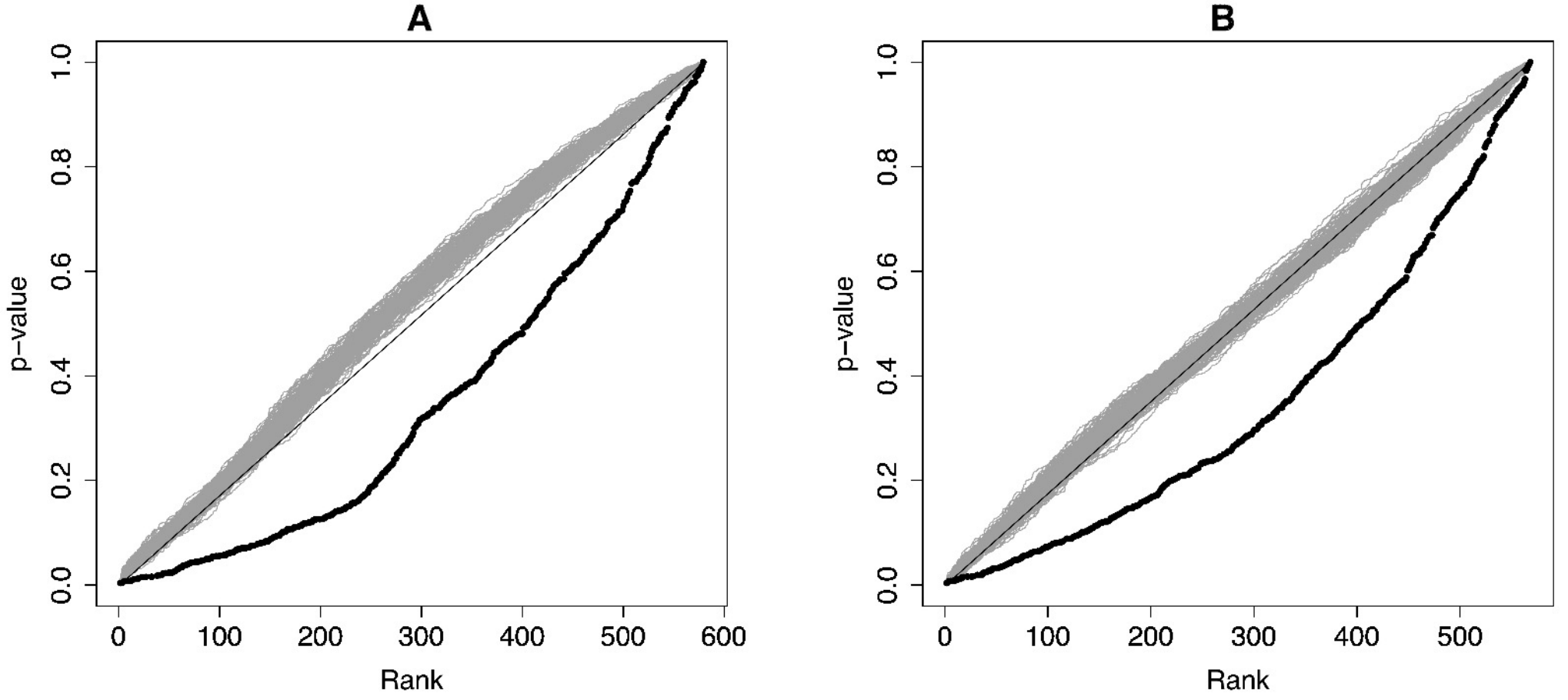
Figure 11. Human lysozyme mutants, all superimposed on the reference (2nwd). The spheres are the alpha carbons at the mutated positions. They are red when the residue is mutated and blue otherwise.

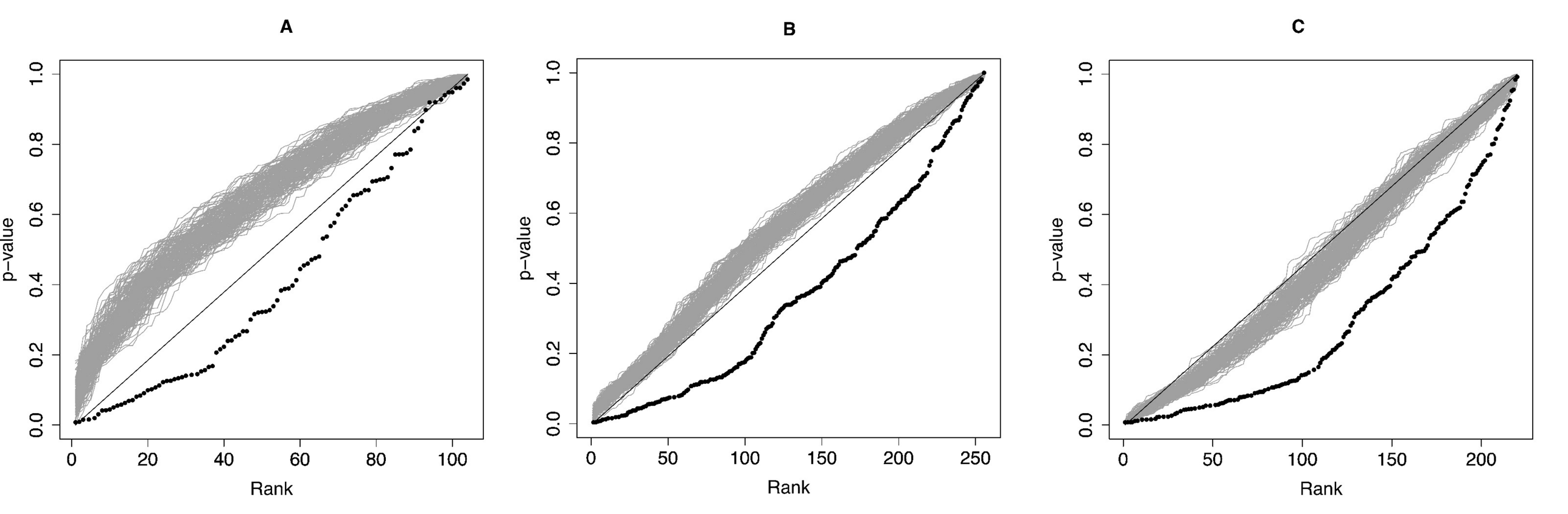

A
B
Number of pairs
0
50
100
150
0.0
0.5
1.0
1.5
2.0
Whole structure RMS

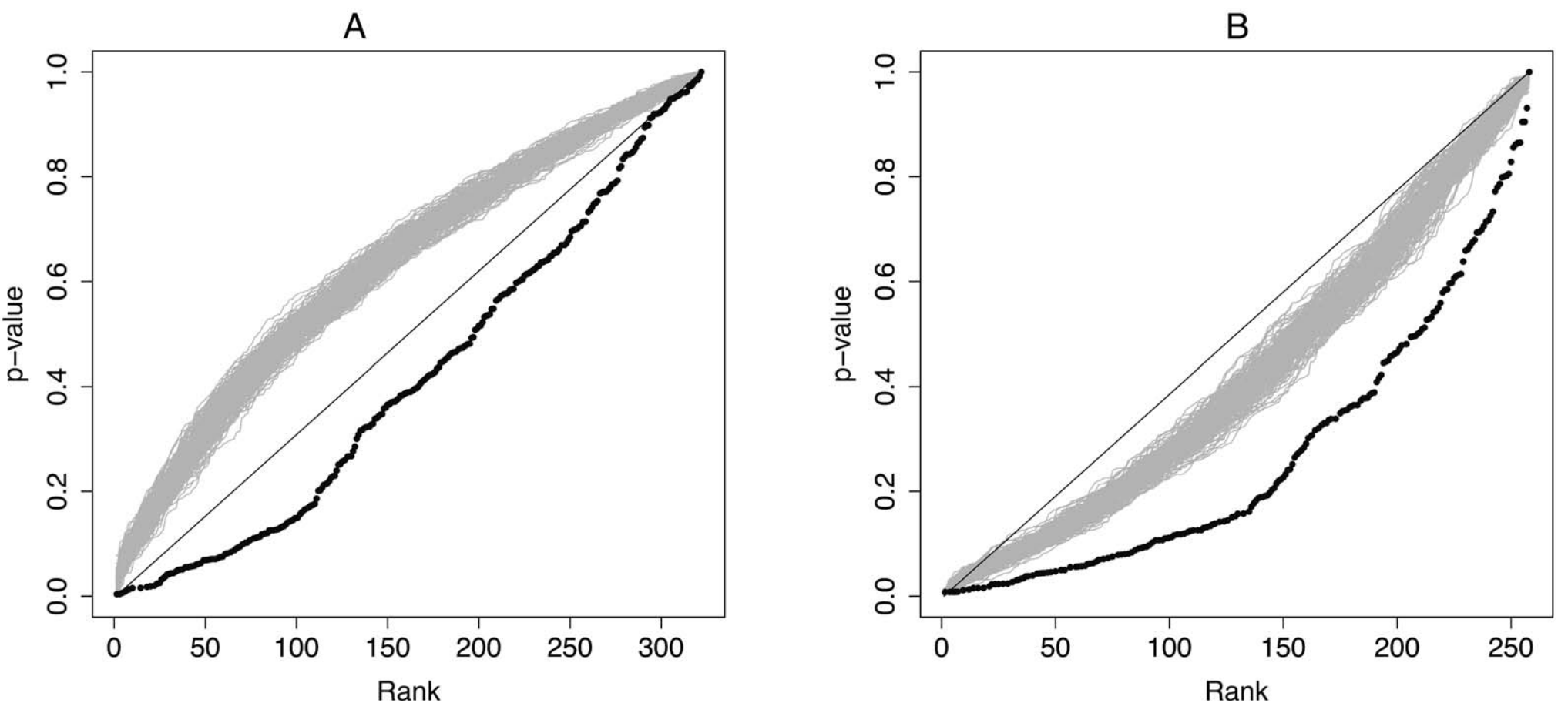

A
B
%id or RMS
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
residue index

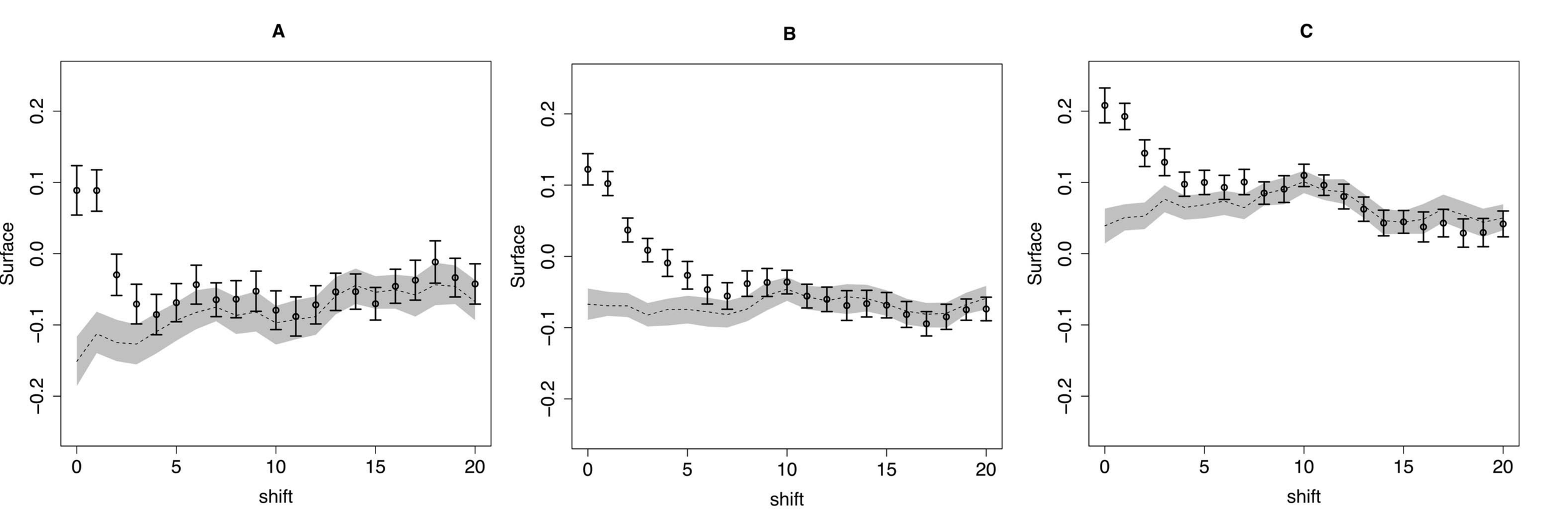

A
occunrencies
0
50
100
150
0.0
0.2
0.4
0.6
0.8
1.0
p−value
B
occunrencies
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
p−value

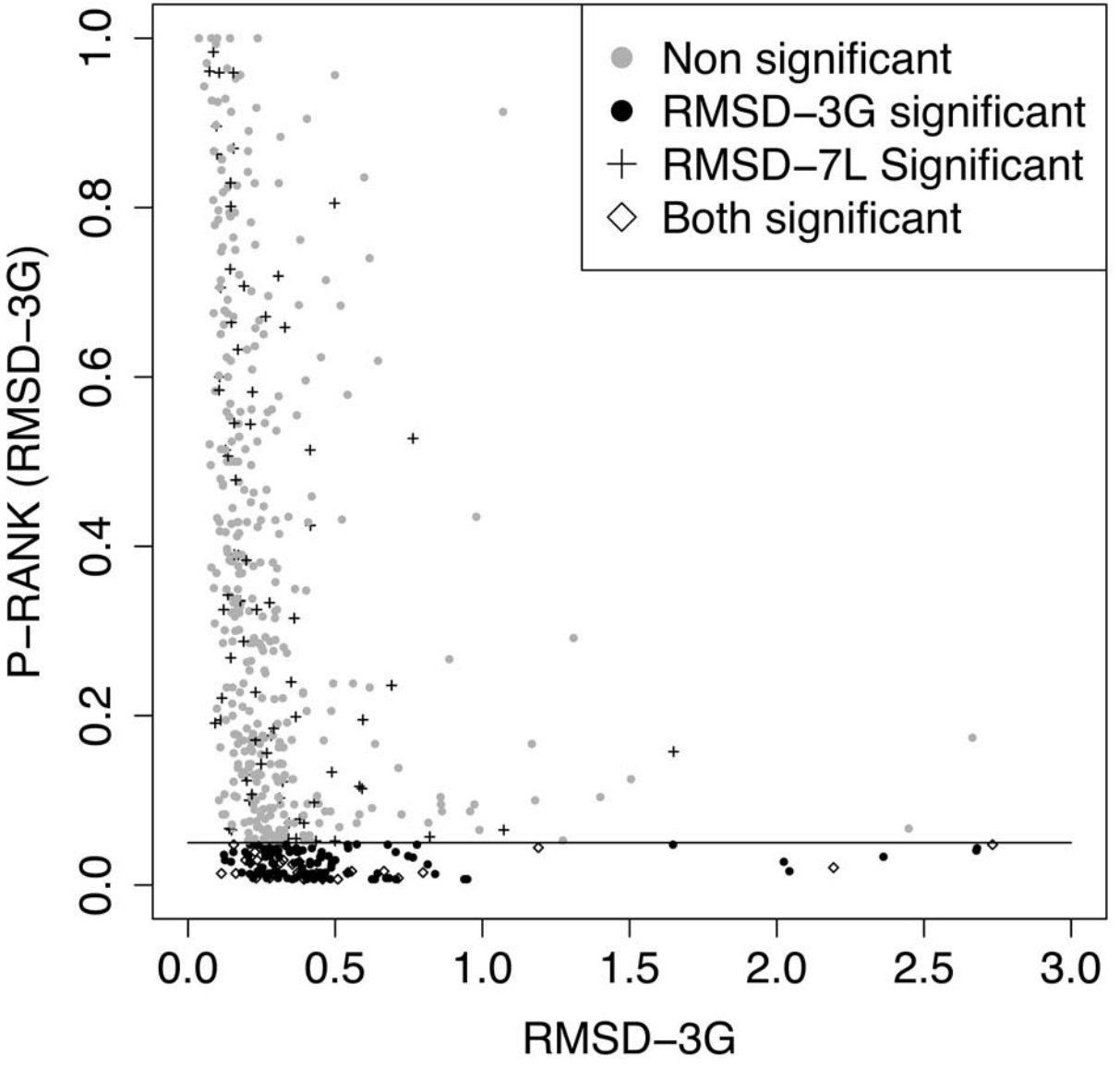

All
E
H
O
RMS−nG
RMS−nL
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20

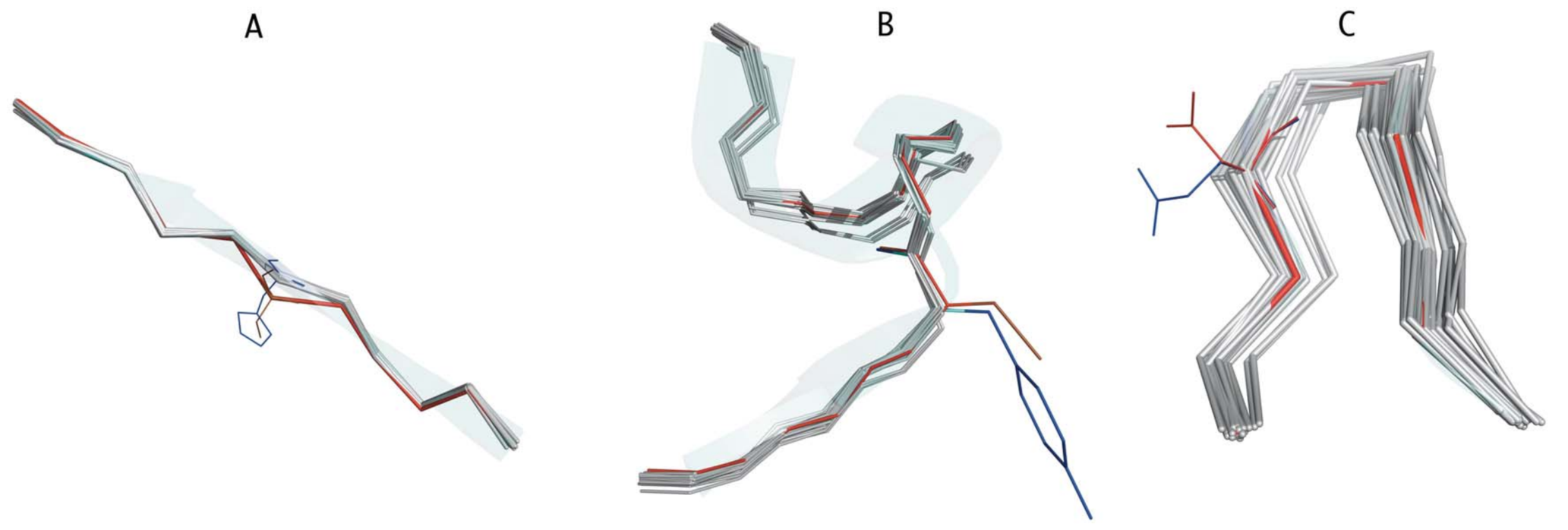

A
B
p−value
Rank
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
600

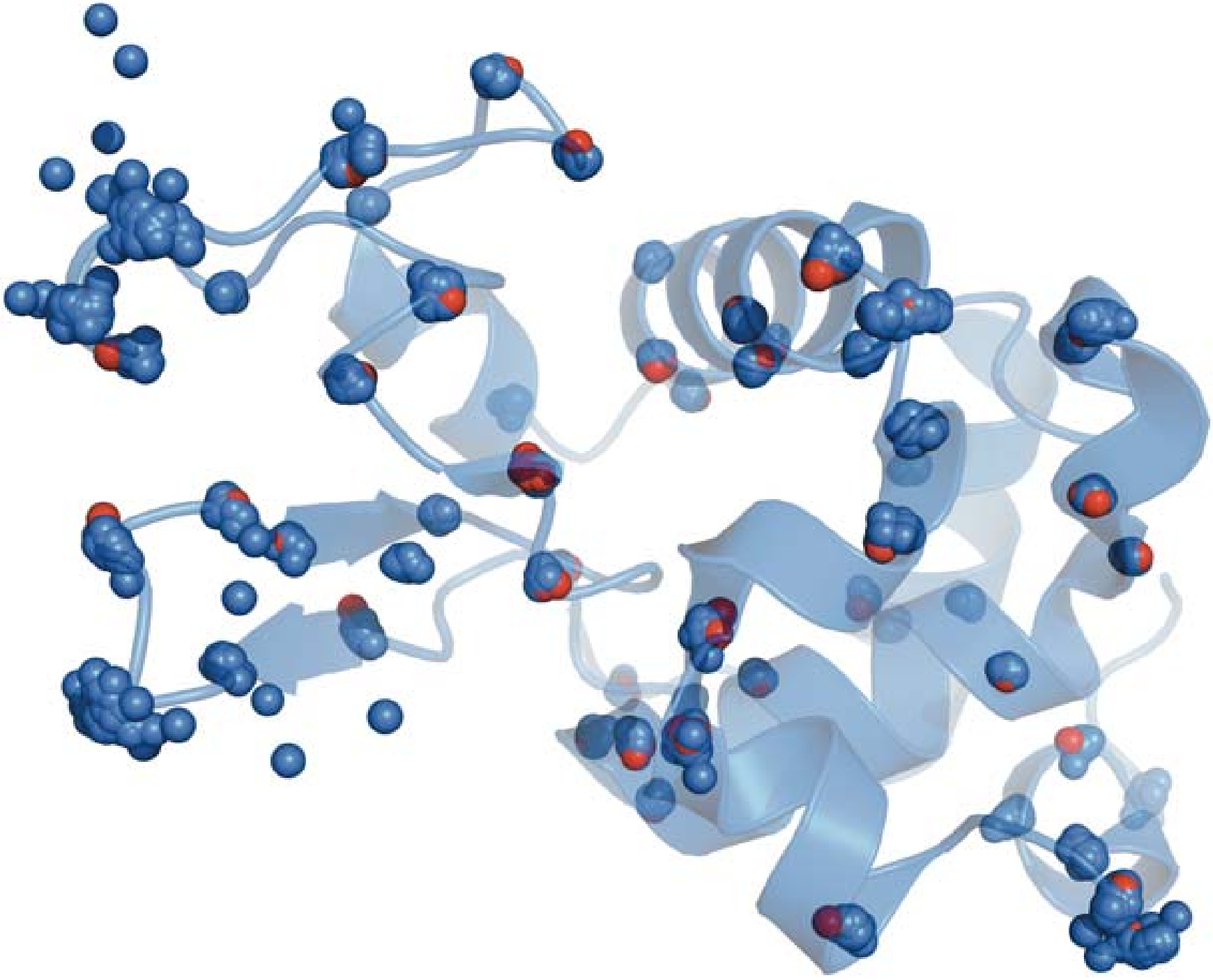

A
B
C
p−value
Rank
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
50
150
200
250


A
B
p−value
Rank
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300


A
B
C
Surface
shift
0.2
0.1
0.0
−0.1
−0.2
0
5
10
15
20


Non significant
RMSD−3G significant
RMSD−7L Significant
Both significant
P−RANK (RMSD−3G)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
RMSD−3G


A
B
C



Rank	Members	Reference	Length	Protein	Class	organism
1	147	1lw9A	164	T4 lysozyme	Alpha	Phage T4
2	124	**2nwdX**	130	Human lysozyme	Alpha	H. sapiens
3	78	2dekA	265	Transferase	Alpha beta	P. Horikoshii
4	69	2iliA	255-260	Anhydrase II	Alpha beta	H. sapiens
5	31	1ey0A	149	Staphylococcal nuclease	Beta	S. aureus
6	29	4bflA	753	Catalase HPII	Multi-domain	E. coli
7	25	**2e3wA**	124	Ribonuclease A	Alpha beta	B. taurus
8	24	**2vb1A**	129	Hen lysozyme	Alpha	G. gallus
9	22	**4fi8A**	126-127	Transthyretin	Beta	H. sapiens
10	22	**2j8cM**	302-314	Reaction centre	Alpha beta	R. sphaeroides
11	20	5deiA	524-536	Benzoylformate decarboxylase	Alpha beta	P. putida
Total	591					

Table I. List of the 11 families of proteins differing by one single point mutation. They are sorted by decreasing number of members. The PDB code and the chain of the reference protein is given, with the protein name, the class assigned by CATH [43] and the organism of origin. In bold the PDB codes of the reference protein in the identical dataset.

		Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
RMS-3G	Mutated	0.04	0.17	0.25	0.36	0.38	11.8
	All	0.01	0.13	0.19	0.24	0.27	12.76
RMS-7L	Mutated	0.04	0.08	0.12	0.16	0.16	1.83
	All	0.00	0.07	0.09	0.11	0.125	3.08

Table II. Descriptive values of the distributions of RMSD-3G and RMSD-7L (in Å) from mutated positions only or all positions of the mutated dataset.

	RMSD-3G			RMSD-7L		
	Random	mutated	Random - mutated	Random	mutated	Random - mutated
All	-0.04	0.15	0.19	-0.01	0.15	0.16
Strands	-0.17	0.09	0.26	-0.04	0.14	0.18
Helices	-0.06	0.12	0.18	-0.03	0.08	0.11
Loops	0.03	0.21	0.17	0.03	0.23	0.20
Buried	-0.13	0.08	0.21	-0.04	0.10	0.14
Exposed	0.07	0.23	0.16	0.04	0.22	0.18

Table III. Cumulative distances between the diagonal and the p-values sorted according to their rank curves for the random dataset and the mutated dataset and their difference (random - mutated), after global and local superimposition. The distances are negative if the p-value curve is above the diagonal and positive if it is under.