Impact of Single Amino Acid Substitution Upon Protein Structure

Mark Livingstone, Lukas Folkman and Bela Stantic
School of Information and Communication Technology, Griffith University, Gold Coast, Qld, 4222, Australia
{*mark.livingstone, lukas.folkman*}*@griffithuni.edu.au, b.stantic@griffith.edu.au*

Keywords: Protein Mutations, Structural Changes.

Abstract: In the biological sciences, one of the most fundamental operations is that of comparison. As we strive to further understand the constituent parts of living tissue, we need to examine proteins and their many mutations. Indeed, characterising mutations is an important part of proteomics, because a seemingly trivial mutation can sometimes stand between creating a life-saving drug on one hand, or blocking a vital receptor inactivating that same drug on the other. In this work we examined single point mutations to characterise their effects on outwardly expanding neighbourhood ranges. As the shape of a protein is very important, we examined how mutations can make subtle changes to the protein shape as well as investigated the implications both for backbone and side-chain residues. Our findings suggest that structural changes upon a mutation are significantly influenced by the protein shape, which allows for the prediction of the impact brought about the mutation by looking only into the protein shape. Surprisingly, we found that there was very little variation between wild type and mutant protein structures close to the mutation site. Also, in contrast with what was expected, the largest structural variations were found when deleted and introduced residues had similar hydrophobicity.

1 INTRODUCTION

Proteomics is one of the most exciting and important areas of study in the biological sciences. It is of importance to areas as diverse as pathology, medicine, and drug design amongst others. With proteins being the building block of living tissue, the need to understand how they operate and inter-operate is of the utmost importance. Over the past several decades, there have been many algorithms introduced which allow us to analyse residue chains on both local and global basis. The International CASP (Critical Assessment of protein Structure Prediction) competition was initiated in 1994. Protein structure prediction methods are competing in CASP and then, prediction results are evaluated using different *scoring algorithms*. While some scoring algorithms have only lasted one CASP cycle (2 years) and then been superseded, others like TM-score (Zhang and Skolnick, 2004) are still being used. However, the *Root Mean Square Deviation* (RMSD) algorithm predates the CASP competition. It was first described back in 1972 by McLachlan (McLachlan, 1972) and then further elucidated by Kabsch in (Kabsch, 1976). Later, Kabsch corrected the calculation in (Kabsch, 1978) where it was used in minimising the distances between aligned atoms using least square minimisation when superimposing two residue chains. Indeed, the whole concept of *least squared deviation* has been used in standard statistics for standard deviations and for regression line analysis since first applied by Sir Francis Galton (Moore, 2004).

While there have been various attempts to modify the RMSD algorithm for various related purposes (e.g., URMS (Kedem et al., 1999), normalised RMSD (Carugo and Pongor, 2001), URMS-RMS (Yona and Kedem, 2005), iRMSD (Armougom et al., 2006)), the original algorithm has never been replaced because of its simplicity, and because it gives a simple distance-based result (generally in Ångstroms (Å)) describing the deviation of one structure from the other. This deviation-based result is much more informative than many of the more recent RMSD variants and other scoring algorithms which are deriving probability-based results, with the most well-known example being TM-score (Zhang and Skolnick, 2004). To further confirm this trend, we examined the CASP 9 proceedings, and found that of 17 papers which dealt with algorithms relevant to this work, nine mentioned RMSD, five mentioned TM-score, and no other relevant algorithm was mentioned more than twice.

In this work we examined single point mutations to characterise their effects on outwardly expanding neighbourhood ranges. As the shape of a protein

Livingstone M., Folkman L. and Stantic B. (2014).
Impact of Single Amino Acid Substitution Upon Protein Structure.
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, pages 123-129
DOI: 10.5220/0004792501230129

is very important, we examined how mutations can make subtle changes to the protein three-dimensional shape as well as investigated the implications both for the backbone and side-chain residues. By using the *Root Mean Square Deviation* (RMSD) we examined mutation neighbourhoods. Furthermore, we derived custom statistics *Shape Ratio* (SR), *Cubic Volume* (CV), and *Ring of the Sums* (RoS) to beneficially describe protein conformational shapes.

Our findings indicate that the mutation has bigger influence in cases when the differences in SR and RoS for the wild type and mutant protein structures are bigger, which allows prediction of structural influence upon a mutation by looking only into the protein shape—the value of SR or RoS. Surprisingly, we found that there was little RMSD variation between the wild type and mutant close to the mutation site. Also, in contrast with what was expected, the largest structural variations were found when deleted and introduced residues had similar hydrophobicity.

2 METHODS

2.1 Data Sets

We employed the data set compiled previously in (Bordner and Abagyan, 2004). This data set includes 2141 pairs of protein structures which differ in a single amino acid position. Protein structures were downloaded from the *Protein Data Bank* (PDB) (Berman et al., 2000). After removing a small number of pairs with missing atoms, ligand and extraneous proteins, and a few with multiple mutations, the data set contained 2,067 pairs of protein structures.

To investigate how a single mutation influences a protein structure, and if the influence is dependent on the length of the protein, we initially grouped proteins as *small*, *medium*, and *large*, where every group had approximately the same number of proteins. As can be seen in Table 1, we ensured that a sufficient gap existed between these classes. It must be noted however that in our results, only Figures 2, and 3 represent these size divisions. For all other results, the entire data set was used.

Table 1: Three groups of proteins with different sequence lengths.

Group	Length	Count	Std. deviation
Small	29–100	170	17.02
Medium	165–210	170	12.62
Large	450+	159	119.65

2.2 Software

The proprietary programs for calculating the RMSDs of different neighbourhood ranges were developed using Python v2.7.3 (van Rossum, 2007) and OpenStructure framework v1.3.1 (Biasini et al., 2010). The analysis of the results was done on a 3.4GHz Intel Core i7 8GB RAM Apple iMac running OS X v10.8.2. For alignment purposes, we used the Smith-Waterman algorithm, with SVD superposition as available in OpenStructure. This worked well given our structures varied mainly due to the effect of the mutation site side chains.

2.3 Calculation of Different RMSD Variants

We analysed the structural changes in protein mutants using the *Root Mean Square Deviation* (RMSD). In Equation 1, wild type and mutant structures are denoted as *M* and *WT*, respectively:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||M_{atom}(i) - WT_{atom}(i)||^2} \qquad (1)$$

To determine which variants of the RMSD calculation were best-suited for analysis of the structural changes in protein mutants, we considered six different RMSD variants:

- *all-atom* RMSD
- *side-chain* RMSD
- *range all-atom* RMSD
- *range side-chain* RMSD
- *range* Cα RMSD
- *range* Cα/Cβ RMSD

All-atom RMSD involved calculating the deviations between all wild type and mutant atoms pairs, whereas *side-chain* RMSD involved side-chain atoms only. In the case of Cα and Cα/Cβ RMSDs, we calculated the deviations between the Cα and Cα/Cβ atom pairs, respectively. In the last four RMSD variants listed above, the *range* refers to the spatial extent in which RMSD was calculated. We considered different *neighbourhood ranges* (e.g., 0–5 Å, 5–10 Å, ...) centred on the mutation site. The smallest range was within 5 Å which is slightly larger than the mean Cα-Cα distance in a protein chain (3.84 Å) (Kedem et al., 1999). We created a set containing all residues within the given neighbourhood range in the wild type structure and calculated the RMSD to the matching atoms in the mutant structure. In our implementation, each range was treated discretely. This means that when

we examined residues in any given RMSD range (e.g., 10–15 Å), the residues in the ranges closer to the mutation site were *not* included (0–5 Å and 5–10 Å).

2.4 Shape Statistics

We introduced three *shape statistics* which we refer to as the *Shape Ratio* (SR), *Cubic Volume* (CV), and *Ring of the Sums* (RoS). To calculate the SR, we calculated the *minimum bounding box* for the structure in a three-dimensional coordinate space (e.g., 8 Å×4 Å×2 Å). We then took the largest dimension and divided it by the shortest dimension ($\frac{8}{2} = 4$). The SR statistic exploits the observation that proteins which are spherical have equal-sided bounding boxes, thus, resulting in an SR of ∼1. The longer (and narrower) the bounding box, the larger the SR will become. In calculating the bounding boxes as described above, we excluded hetamer, ligand, and solvent atoms. The CV statistic is simply the cubic volume of the minimum bounding box of the protein structure.

Finally, the third statistic is the *Ring of the Sums* (RoS). It considers a protein as an undirected complete graph having Cα atoms as its vertices, and edges between every possible Cα atoms pairs. From each vertex, we determine the Euclidean distance (Deza, 2009) to every other vertex (as can be seen in Figure 1). We then sum all distances and divide the result by the number of edges including the given vertex arriving at a mean distance. The number of edges is dependent on the number of residues (n) (in Figure 1, n is 4). Equation 2 gives the formal definition of the RoS statistic (*WT* refers to the wild type structure).

$$RoS = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} ||WT_{residue}(i) - WT_{residue}(j)|| \quad (2)$$

The RoS is in units of 'Ångstroms per residue', and it contains information about the density of the given protein structure. We also considered the *radius of gyration vector* as another possible shape analogue, but did not proceed with it since it cannot be straightforwardly reduced to a single-number value. For the same reason we also dismissed statistics involving the superposition *rotation and translation vector*.

3 RMSD AND NEIGHBOURHOOD RANGES

We commenced our investigation by examining how the structural effects of single point mutations in proteins of different lengths could be quantified. For this purpose, we studied the RMSDs for proteins belonging to three different groups: *small*, *medium*, and *large* (Table 1). We examined how the *range* Cα RMSD and *range side-chain* RMSD vary for a number of discrete outwardly extending neighbourhoods with a radius of 5 Å.

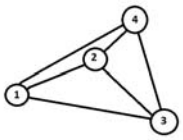

Figure 1: Ring of the Sums (RoS). For the RoS calculation, the protein structure is represented as an undirected complete graph.

Figure 2 shows the mean *range* Cα RMSD as the function of the distance from the mutation site (discrete intervals of 5 Å). Surprisingly we found that when we examined directly next to the mutation site (0–5 Å range), there was little deviation between the wild type and mutant proteins. This is likely due to the small volume of the first neighbourhood range. More significant deviation can be observed in the neighbourhood ranges which were situated further from the mutation site (neighbourhood ranges 10–15 and 15–20 Å).

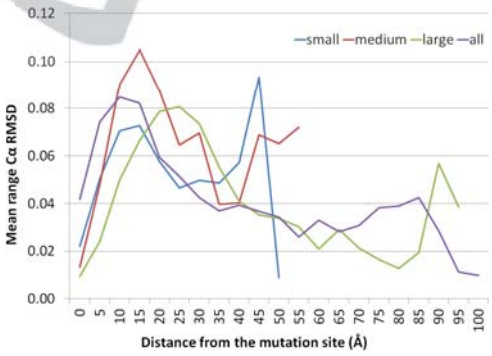

Figure 2: Mean range Cα RMSD as the function of the distance from the mutation site for small, medium, large, and all proteins.

To get a comparison for our data set as a whole, we also plotted a *weighted mean* of the *range* Cα RMSDs for proteins in all three groups (Figure 2). The weighted mean was calculated as the sum of *range* Cα RMSDs over all structure pairs for each neighbourhood range divided by the total number of residues found in the given neighbourhood. We observed that there is a characteristic bi-peaked curve in common to all three groups as well as to the weighted mean results in Figure 2. The initial peak is relatively close to the mutation site, followed by a smaller peak situated further.

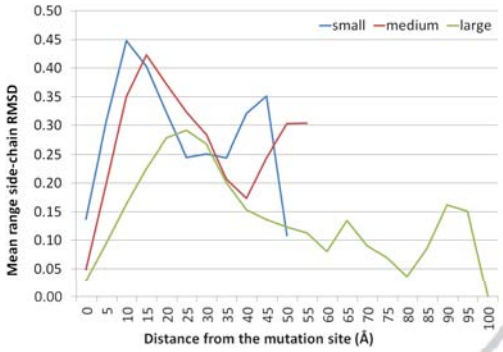

Figure 3: Mean range side-chain RMSD as the function of the distance from the mutation site for small, medium, and large proteins.

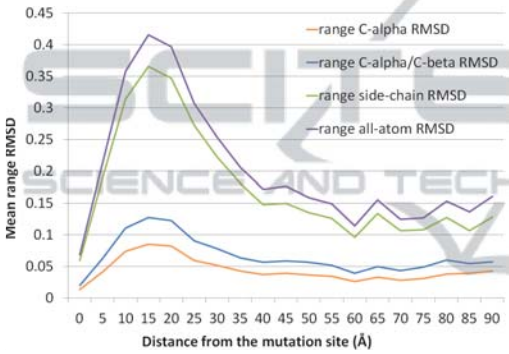

Figure 4: Mean range RMSDs for different atom types as the function of the distance from the mutation site.

The same bi-peaked curve can be seen also in Figure 3 which shows the mean *range side-chain* RMSD as the function of the distance from the mutation site. The most significant structural changes occur in the neighbourhood ranges between 10 and 35 Å. It seems that the longer the protein chain, the further from the mutation site these significant changes occur (for the *small* group it was 10–15 Å neighbourhood range, while for the *large* group 25–30 Å range).

Since we had found that the weighted mean deviations for proteins of all lengths can closely approximate the individual curves for each protein group (Figure 2), we examined how the deviations varied depending on whether we calculated the *range all-atom*, *range side-chain*, *range* Cα, or *range* Cα/Cβ RMSD. As we expected, when we considered more atom types in the calculation, the RMSD values increased. This is apparent from the Figure 4 which shows the RMSD as a function of the distance from the mutation site for different types of *range* RMSD calculations.

Upon further examining Figure 4, the main peak in the plot occurs slightly further out than the mutation site. We believe that this peak is caused by a strain in the protein structure brought about the mutation. Then, in a flexible region of the protein structure (possibly further away from the mutation site), the strain is released via structural reconfiguration. To examine the significance of this, we produced three plots which shows the different types of *range* RMSDs as a function of the change in hydrophobicity. The change in hydrophobicity, denoted as Δ *hydrophobicity*, is equal to the difference in the hydrophobicity of the introduced and deleted amino acids (Kyte and Doolittle, 1982). We inspected Δ *hydrophobicity* at three different neighbourhood ranges: 0–5, 10–15, and 25–30 Å (Figures 5, 6, and 7, respectively). These were selected to inspect the mutation site as well as each side of the main peak in Figure 4.

Surprisingly, as can be seen in Figures 5 and 6, the largest structural deviations were in cases when introduced amino acid had similar hydrophobicity to the deleted amino acid. Furthermore, as the absolute value of Δ *hydrophobicity* increased, the RMSD decreased. This trend is obvious for all four types of *range* RMSD calculations considered.

When we look at the 25–30 Å neighbourhood range shown in Figure 7 where we are now on the decreasing side of the peak (the main peak from Figure

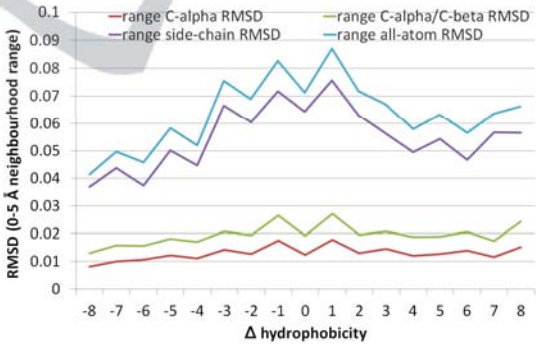

Figure 5: RMSDs for different atom types in the 0–5 Å range as a function of Δ hydrophobicity.

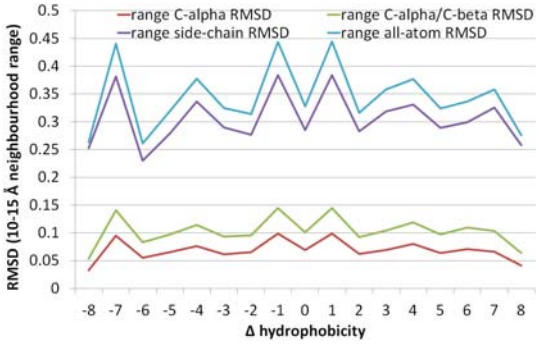

Figure 6: RMSDs for different atom types in the 10–15 Å range as a function of Δ hydrophobicity.

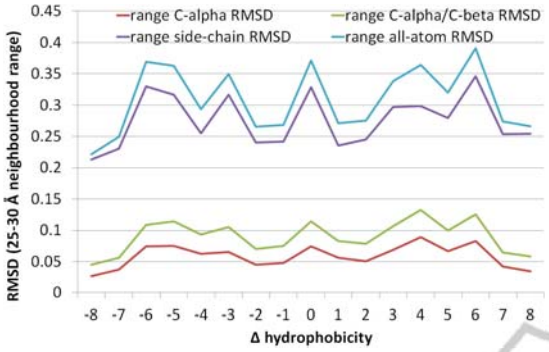

Figure 7: RMSDs for different atom types in the 25–30 Å range as a function of Δ hydrophobicity.

4), we see that the 0 Δ *hydrophobicity* point is in fact a local maximum. Here we find that on either side of the 0 Δ *hydrophobicity* point, we have a decrease to ±1 Δ *hydrophobicity*, and a flat section out to ±2 Δ *hydrophobicity*. This represents a reversal of what we had found in Figures 5 and 6.

Next, we inspected the relationship between the structural deviations and the statistics that we proposed in this study (Section 2). Figures 8, 9, and 10 show the *all-atom* RMSD as a function of the *Ring of the Sums* (RoS), *Delta Shape Ratio* (ΔSR), and *Delta Cubic Volume* (ΔCV), respectively. We defined the ΔSR and ΔCV as the difference of the mutant and wild type values for the SR and CV, respectively. Our results indicate that a mutation has bigger influence in cases when RoS, ΔSR, and ΔCV are bigger, particularly in the case of ΔSR. This is a significant finding meaning that just from inspecting the protein shape or the value of ΔSR, we can predict the impact of a single-site amino acid substitution.

Table 2 gives a summary of the shape statistics and the mean values for the data set used in this study. To relatively compare RoS with ΔSR and ΔCV, we calculated ΔRoS which refers to the difference between the

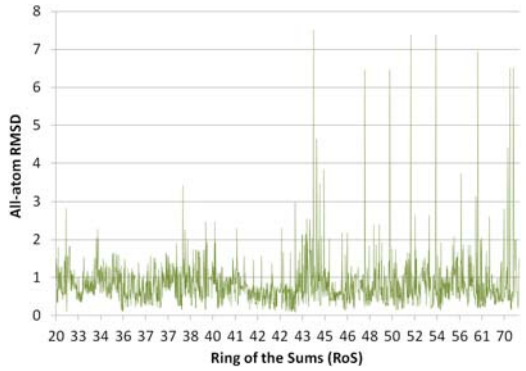

Figure 8: All-atom RMSD as a function of the Ring of the Sums (RoS).

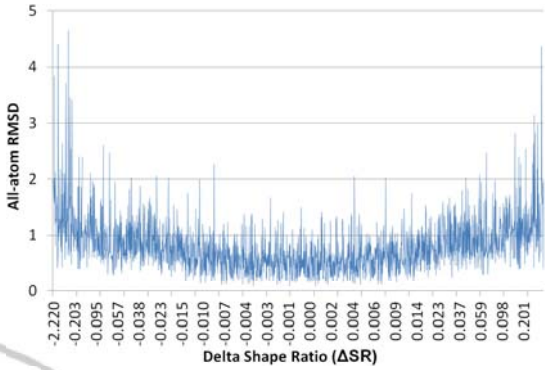

Figure 9: All-atom RMSD as a function of the Delta Shape Ratio (ΔSR).

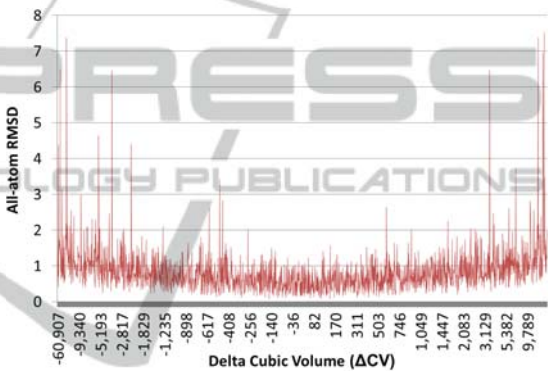

Figure 10: All-atom RMSD as a function of the Delta Cubic Volume (ΔCV).

RoS of the mutant and wild type structures. Thus, a positive mean value implies that the mutation caused an increase in the given statistic.

3.1 Protein Shape Statistics

The second goal of this paper was to find one or more ways of numerically describing the shape of a protein conformation. Such numerical statistic could be then used to effectively describe the conformational change caused by an amino acid substitution. Moreover, knowing the overall shape of a protein is very important because it determines the surface accessible area. Also, shape statistics could be used to compare and classify proteins.

Table 2: Changes in the shape statistics upon a mutation.

Statistic	Mean value
ΔCV	12.06
ΔSR	0.46
ΔRoS	-0.004

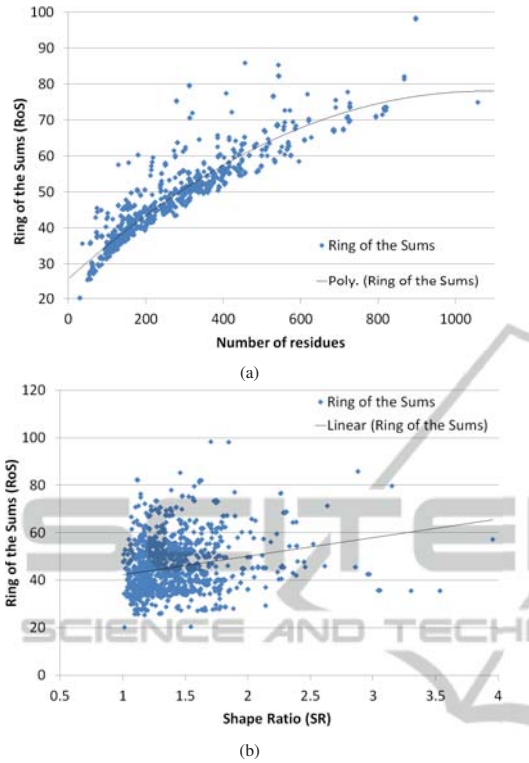

Figure 11: Ring of the Sums (RoS) as a function of (a) the number of residues and (b) the Shape Ratio (SR).

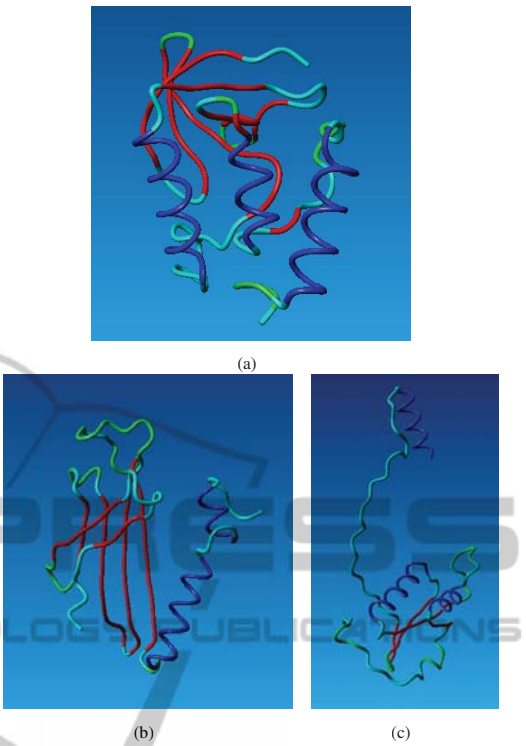

Figure 12: Three different protein structures with a length of 129 residues. (a) Protein 1EY8 with an SR of 1.08, RoS of 35.91, (b) protein 1E6T with an SR of 1.80 and RoS of 46.31, and (c) protein 1M0D with an SR of 3.95 and RoS of 57.56.

First, we compared the *Ring of the Sums* (RoS) with the number of residues. As can be seen in Figure 11(a) by the polynomial regression (a correlation coefficient of $R^2 = 0.866$), there is a strong correlation. It was also noticeable that the lower edge of the data plot has a characteristic curve very much like the polynomial regression line plotted in the same graph which indicates how good the fit is.

Analysing the *Shape Ratio* (SR) statistic, we found that the SR mean value is 1.37 for our data set. Since an SR of 1.0 implies a spherical protein, we can infer that a 'mean' structure of the data set was slightly oblong in shape. The minimum SR was 1.003, and the maximum was 3.95. This implies that the most oblong structure had its maximum dimension almost four times its minimum bounding dimension.

We then examined the relationship between the RoS and SR. In Figure 11(b) we can see a proportional relationship: as the RoS increases, the SR increases slightly. However, it is apparent from this figure that most of the data used in this study had predominantly spherical shape. Therefore, the predictive power of the RoS and SR was quite limited ($R^2 = 0.0424$)

Finally, let us examine visually three representative protein structures which were chosen on the basis that they all have 129 residues (Figure 12). Figure 12(a) shows a structure which is visually fairly spherical with an SR of 1.08. This fairly compact structure with a small number of residues has a RoS of 35.91. Figure 12(b) depicts a structure which is slightly more elongated with an SR of 1.80. Compared with Figure 12(a) we see a more spread out and sparse arrangement, which results in a larger RoS value of 46.31. Finally, the structure in Figure 12(c) is long and stringy. Hence, it has SR and RoS values of 3.95 and 57.56, respectively.

Comparing the two extreme structures (Figure 12(a) and 12(c)), we have a difference in the SR of 365%, which is indicative of the increased gross size of the bounding boxes required to encompass the structures. Also, there is a difference in the RoS of 160%.

4 CONCLUSIONS

In this paper we set out to examine structural deviations and changes in the overall shape of a protein upon a single amino acid substitution. We introduced three shape statistics which we referred to as the *Shape Ratio* (SR), *Cubic Volume* (CV), and *Ring of the Sums* (RoS).

We showed that there is a good relationship between the SR and RoS. While the SR is significantly easier to calculate, the RoS gives us density of residues as part of its value which may be useful in some cases. We have also seen 86.6% correlation between the number of residues and the RoS statistic.

Furthermore, we have demonstrated that there is a characteristic curve that is shared for all *range* RMSD variants that we investigated, regardless of the size of the protein.

Our results indicate that the mutation has bigger influence in cases when the ΔSR, ΔCV, and RoS are bigger. This finding implies that the prediction of structural impact upon a mutation might be possible simply by inspecting protein shape—the value of ΔSR or RoS.

Surprisingly, we found that there was very little variation between wild type and mutant protein structures close to the mutation site. Also, in contrast with what was expected, the largest structural variations were found when deleted and introduced residues had similar hydrophobicity.

REFERENCES

Armougom, F., Moretti, S., Keduas, V., and Notredame, C. (2006). The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, 22(14):e35–9.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic Acids Research*, 28(1):235–242.

Biasini, M., Mariani, V., Haas, J., Scheuber, S., Schenk, A. D., Schwede, T., and Philippsen, A. (2010). OpenStructure: a flexible software framework for computational structural biology. *Bioinformatics*, 26(20):2626–2628.

Bordner, A. and Abagyan, R. (2004). Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins: Structure, Function, and Bioinformatics*, 57(2):400–413.

Carugo, O. and Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science*, 10(7):1470–1473.

Deza, M. M. (2009). *Encyclopedia of Distances*. Springer.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, 32(5):922–923.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica*, 34(5):827–828.

Kedem, K., Chew, L. P., and Elber, R. (1999). Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins: Structure, Function, and Bioinformatics*, 37(4):554–564.

Kyte, J. and Doolittle, R. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132.

McLachlan, A. (1972). A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallographica*, 28(6):656–657.

Moore, D. (2004). *The basic practice of statistics*. W. H. Freeman, 3rd edition.

van Rossum, G. (2007). *Python language website*. `http://www.python.org`.

Yona, G. and Kedem, K. (2005). The URMS-RMS hybrid algorithm for fast and sensitive local protein structure alignment. *Journal of Computational Biology*, 12(1):12–32.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710.