Regions of Similarity: A Novel Graph Theoretical Protein Structure Comparison and Analysis Technique

Aaron Maus and Christopher Summa

November 3, 2019

Regions of Similarity: A Novel Graph Theoretical Protein Structure Comparison and Analysis Technique

Aaron P. Maus[*] and Christopher M. Summa[†]

[*]Department of Computer Science, Tulane University
New Orleans, Louisiana, 70118, United States
amaus@tulane.edu

[†]Department of Computer Science, University of New Orleans
New Orleans, Louisiana, 70148, United States
csumma@uno.edu

Abstract

All existing protein structure comparison methods return a score for similarity, but few give a deep underlying look at the parts of the structures which match. Zemla's Global Distance Test (GDT) [1] partially does by identifying the largest region of a pair of structures whose superposition errors all fall under some threshold, but the region and its errors are dependent on that superposition, and smaller regions are not identified. By converting the Cα distances matrices of two structures into a graph, a maximum clique analysis can be used to identify the largest non-overlapping regions of similarity between structures. These regions can easily be visualized, and they lend themselves to a deep analysis of the underlying similarities between structures, complementing existing methods of comparison by providing additional information that is not readily available. Additionally, when applied to an analysis such as that performed for each CASP experiment, models which correctly represent each domain in a multi-domain structure but whose orientations differ from the native will be immediately apparent. A regions of similarity analysis can be performed on multi-domain targets without *a priori* knowledge of the domains.

keywords: structural bioinformatics, protein structure comparison; max clique; protein structure prediction; conformational comparative analysis; CASP

1 Introduction

Whether for analyzing the results of different protein structure predictors, different conformations of the same protein, or similar conformations of related proteins, the comparison and analysis of complex three-dimensional structures is a difficult yet fundamental task. Comparison of two or more protein structures requires a correspondence between reference points (usually the α

backbone carbon atoms, or Cαs) in one structure to reference points in the other. It is based on these correspondences that differences and similarities in the two structures can be assessed. Broadly speaking, there are two major categories of methods for protein structure comparison: superposition-based methods and contact-based methods.

In superposition-based methods, the correspondences between structures are the distances between analogous Cαs following a superposition of one structure onto another. Two such methods are the Global Distance Test (GDT) [1] and the Template Modelling Score (TM-Score) [2]. TM-Score is a score originally designed to rank templates for the application of protein threading. It is the score given to the optimal superposition of a template onto a reference structure that minimizes its scoring function. Another method, GDT, identifies four sets of residues, a set for each of four thresholds: {1.0, 2.0, 4.0, and 8.0 Å}. Each set is the largest set of residues that can be superposed whose superposition errors all fall under its threshold. These residues can be highlighted in visualizations, but they are not encompassing of all of the potential similarity between two structures. Since GDT relies on superposition, if there are multiple areas of similarity – for example, two domains, each well-modelled (in the case of protein structure prediction efforts) but shifted relative to each other – GDT will, in most cases, only identify the largest domain. The difficulty ultimately is that the underlying information, the residue errors, are dependent on the superposition. Does an individual residue have a large superposition error because it is in a part of the structure that is not well-modelled or because of an unfavorable superposition?

In contact-based methods, the correspondences are analogous contacts within the structures. For example, the Contact Area Difference score (CAD) examines the contact surface areas of pairs of residues within both structures [3], and the local Distance Difference Test (lDDT) method compares pairwise atomic distances within both structures [4]. lDDT measures the fraction of pairwise distances, all less than an inclusion radius R_o (by default 15 Å) and not within the same residue, in a reference structure that are preserved in a model. By restricting the distances analyzed to those under the inclusion radius, lDDT acts a measure of the local accuracy of the structures.

It is important to note that in protein comparison there is a distinction between the global and local accuracy of structures and that these two directions of structural analysis are often orthogonal. Globally accurate structures are those which orient the tertiary components of structures, such as domains, correctly relative to each other while locally accurate structures are those that get the details of the components correct. Structures which are globally accurate might not be locally accurate and vice versa. Domain movements in multi-domain structures will contribute to a poor global score even if the domains themselves are locally accurate. In general, superposition-based methods tend to favor global accuracy since only those structures which are globally accurate will receive high scores for some optimal superposition. Likewise, contact-based methods tend to evaluate local accuracy since they rely on contacts – local measures which are reproduced in both a reference and a model. Balancing the orthogonal pull of the analysis global versus local accuracy remains a key difficulty in protein structural analysis.

Ideally, methods to analyze the similarities of and differences between protein structures should have certain properties [5]: They should be quantitative and visualizable (*i.e.* they should produce an overall metric but rely on underlying information that can easily be visualized in a meaningful way). They should not only allow analysis across large data sets, but also allow insightful analysis into individual comparisons. They should be stable against large variations in small parts of the structures (*i.e.* large swings in variable loops or at the termini of a structure should not result in large leaps in the similarity score). Finally, any new method should provide information that is not easily accessible from other measures, and their assessments should be intuitive to understand.

Inspired by the work done on GDT and lDDT we propose a novel method that we believe supplements these techniques and allows a more detailed analysis of the similarities of and differences between protein structures. Regions of Similarity, the method proposed in this paper, is a contact-based method most like lDDT in spirit. Like lDDT, it is based on pairwise distances, but instead of finding the fractions of preserved local distances between two structures, it uses the same distance information to perform a graph analysis of the similarity between the two structures. It is through this graph analysis that a detailed assessment of the similarities and differences between structures can be performed.

2 Materials & Methods

2.1 Definition of Regions of Similarity

A Region of Similarity is a set of aligned residues between two protein structures whose intra-structure Cα distances are all the same – within a tolerance threshold – in both structures and which all form a cohesive unit within the structures. Rigorously defined, given a reference and a model structure whose residues have been aligned, a region of similarity is a set of residues whose:

1. Size is at least 10 residues.
2. Pairwise Cα atomic distances are all the same, within a tolerance threshold, in both structures.

3. Contact map in the model forms a connected graph.

The third condition ensures that the residues in a region all come from some local part of the model. It forces a region to contain contiguous residues in three-dimensional space and enforces the idea that a region should represent a set of residues that take the shape they do because they are strongly interacting with one another. Without this condition, it would be possible to have residues from distant parts of the structures forming a region because they are coincidentally the same distance apart in both structures.

2.2 Finding Regions of Similarity

To find the largest region of similarity between two protein structures, first their sequences are aligned. Then the distance differences matrix is calculated: $D_{i,j} = R_{i,j} - M_{i,j}$ where i and j are aligned residues, R is the Cα distance matrix for the reference structure, M is the Cα distance matrix for the model structure, and D is the distance differences matrix. A similarity graph is then built from $D_{i,j}$. Every residue is a vertex, and there is an edge between two vertices if their value in $D_{i,j}$ is less than a tolerance threshold, $t = 1.0$Å by default. The maximum clique of this graph reveals the set of potential residues for the region of similarity. The last step is to select only those which form the largest spatially contiguous region in the model. To find this region, a graph is built from the contact map of the model (all residues are vertices and there is an edge between two residues if their Cαs are less than 10.0 Å apart), and the largest component found by a depth-first search of this graph reveals the final residues in this region. If this region contains at least 10 residues, return it, otherwise there is no region of similarity between the structures.

A disjoint set of regions of similarity (denoted simply as RoS) can be found by iteratively identifying regions on the same similarity graph G. After each region is found, its residues are removed from G to prevent residues from being assigned into multiple regions. This continues until no more regions are found. If the two structures are identical, there will be a single region containing all residues. If the structures consist of two identical domains that are shifted relative to each other, then there will be two regions of similarity, one for each domain.

Regions of similarity can also be used to perform a threshold tiered test inspired by GDT: RoS-GDT. Given a set of thresholds {1.0, 2.0, 4.0, and 8.0 Å}, four regions of similarity are identified: $R_{1.0}$, $R_{2.0}$, $R_{4.0}$, and $R_{8.0}$. Each region is the largest region of similarity in the similarity graph built under its threshold which, for each threshold except the first, completely encompasses the region of similarity found for the previous threshold. To find these regions, four similarity graphs, $G_{1.0}$, $G_{2.0}$, $G_{4.0}$, and $G_{8.0}$, are constructed as described above. To start, the largest region of similarity in $G_{1.0}$ is found. This is $R_{1.0}$. Then, the subgraph in $G_{2.0}$ consisting of the residues from $R_{1.0}$ is identified and all residues which are neighbors of this subgraph and which have an edge to every residue in this subgraph are selected. The maximum clique found within these residues in $G_{2.0}$ is the maximum set of residues which can be combined with those in $R_{1.0}$ and still form a clique in $G_{2.0}$. Within this combined set of residues, the largest connected component in the contact map graph is found, and the residues in this component are returned as $R_{2.0}$. The same process is repeated for $R_{4.0}$ and $R_{8.0}$. The thresholds {0.5, 1.0, 2.0, and 4.0 Å} can be used to perform an RoS-GDT-HA test. The set of regions found by RoS-GDT is called an expanded region of similarity.

The regions found by RoS-GDT show tiers of modelling quality, but they only encompass one part of a pair of structures. Like the original GDT, in a multi-domain structure where separate domains are well modelled but shifted relative to each other, RoS-GDT will identify only the largest domain. To identify multiple areas of a pair of structures that are similar, a disjoint set of Expanded Regions of Similarity (ERoS) can be identified. Each expanded region of similarity has tiers of residues found using the thresholds {1.0, 2.0, 4.0, and 8.0 Å}. To start, a set of disjoint regions of similarity is identified under the first threshold. Then, for each subsequent threshold, each region of similarity, in the order of initial discovery, is expanded to the next threshold using the similarity graph for that threshold omitting all residues found in all other regions so far. At the end of the process, a set of Expanded Regions of Similarity is returned. A score similar to GDT_TS can be calculated from this set: the average of the percent of residues under each threshold. $ERoS\_score = \frac{1}{4}\left(R_{t_1} + R_{t_2} + R_{t_3} + R_{t_4}\right)$, where R_{t_n} is the sum of the fractions of residues that fall under the n^{th} threshold over all of the expanded regions of similarity. Each fraction is calculated with respect to the number of residues in the reference structure.

Expanded Regions of Similarity can also be generated using twenty thresholds: {0.5, 1.0, 1.5, …, 10.0}. The fraction of residues under each threshold can be used to generate plots which show the percent of the structures which match under decreasing levels of accuracy. This technique is denoted as ERoS-Plot.

2.3 Visualizations

Local accuracy maps can be generated from regions of similarity. They show, at the sequence level, which residues in a model are within which region of similarity. Up to five regions can be colored: blue, green, purple, brown, and yellow. If a single threshold is used, such as when finding disjoint regions of similarity, the region with the largest number of residues is colored blue and the region with the smallest number of residues is colored yellow. If expanded regions of similarity are being visualized, the colors are determined in the same order by the size of the regions identified using the most stringent threshold. Residues which are not in any of the top five regions are colored red, and those that are not in the reference or the model are colored white. The colors have been chosen to be visually distinct. If expanded regions of similarity are being visualized, within each color, the shades vary uniformly in saturation and luminosity to indicate under which threshold that residue was added to the region. Darker shades indicate more stringent thresholds. Finally, if RoS-GDT regions are being represented, a divergent color scheme from blue to peach is used. Red residues are not in any of the regions.

ERoS plots can be generated from the ERoS-Plot data. For each model, the total fraction of residues identified under each threshold is plotted and the result shows how well that model represents the target. Those models which include larger portions of their structure within regions of similarity under tighter thresholds are the better models.

Regions of similarity can also be visualized on the three-dimensional structural representations of proteins as well. Both PyMOL [6] and Chimera [7] scripts can be generated to select and color residues belonging to each region and threshold so that individual structure pairs can be examined in detail.

2.4 Feasibility Study

Identifying regions of similarity relies on solving instances of the NP-complete problem of finding maximum cliques. To ensure the feasibility of the technique, a study was performed on a set of 88,758 pairs of different experimentally determined structures for identical proteins provided by Kufareva [5]. This dataset contains a variety of structures of varying sizes and levels of similarity. The smallest structures contain less than 20 residues and the largest over 1000. Measured by LGA_S, the least similar pairs have scores less than 10 and the most similar have scores of 100. For each pair, RoS, RoS-GDT, RoS-GDT-HA, ERoS, and ERoS-Plot were generated. The runtimes were recorded and are presented below.

2.5 Software & Hardware

All algorithms for finding regions of similarity have been implemented in jProt, a java protein comparisons library freely available at https://github.com/amaus/jProt. Maximum cliques are found using Li, Fang, and Xu's C program implementation of their IncMaxCLQ algorithm [8]. Molecular graphics were produced with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. Local accuracy maps and ERoS plots were generated using gnuplot.

The feasibility study was performed on the lee2 cluster at the University of New Orleans. This cluster consists of 36 compute nodes, each with dual XEON X5650 CPUs. Lee2 has a total of 1.1 TB of RAM.

3 Results & Discussion

3.1 Illustrating Regions through Local Accuracy Maps

Local accuracy maps can be generated using each of three major techniques: RoS, ERoS, and RoS-GDT. Figure 1 illustrates the differences between them using the two-domain target T0976 from the CASP13 experiment [9]. This target was chosen because most models roughly represent each domain (and some do accurately), but they generally shift the domains relative to each other with respect to the reference structure. In these plots, the top four models ranked according to their *ERoS_Score* are displayed.

The regions identified by RoS and ERoS show that in these structures, there are two large regions, blue and green, that are well-modelled. Since the residues in these regions are not sequential, it is likely that these are elements of secondary structure that are accurately representing parts of the tertiary structure of the reference. Additionally, in the top model, in each half there are sequential segments of the sequence, brown and yellow, that are likely secondary structures shifted relative to the others. Comparing these plots against the three-dimensional structures illustrated in Figure 2, the two large regions correspond to the two domains and the yellow and brown regions are alpha helices shifted relative to their domains.

The information in these maps is information that regions of similarity can present in addition to the information provided by other methods of comparison. For example, while lDDT gives each residue a local accuracy score, regions of similarity can identify the sets of residues that together are all locally accurate as a group. While regions of similarity, like lDDT, is a measure of local accuracy, GDT is a measure of global accuracy. It tends to rank structures favorably that are globally accurate since structures with accurate global orientations are more likely to capture larger parts of the structures in an optimal superposition. In the case of T0976, GDT will rank well

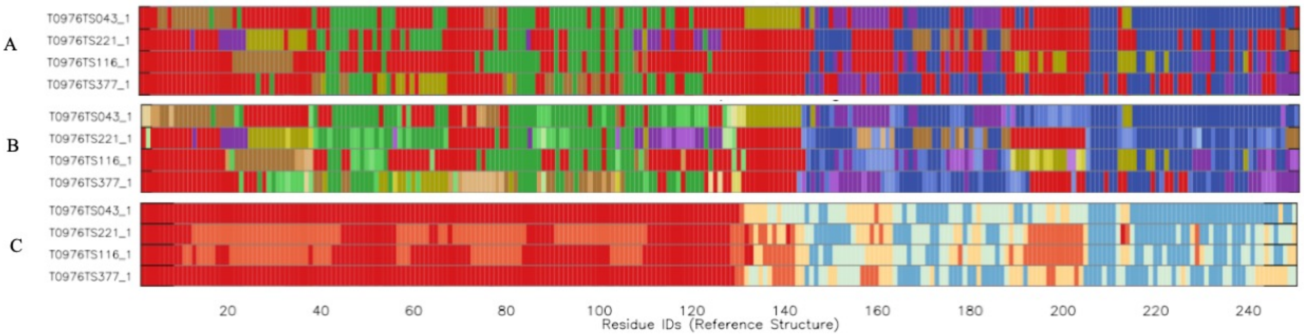

Figure 1: Local Accuracy Map showing the comparison of the three Regions of Similarity methods on target T0976 from CASP13 (A) RoS: A disjoint set of regions of similarity (identified under the default threshold of 1.0 Å), colored in order of largest to smallest: blue, green, purple, brown, then yellow. Red indicates that a residue is not in any of the five largest regions highlighted. **(B)** ERoS: The Expanded Regions of Similarity. Starting from those regions found by RoS, each region has been expanded in turn to include residues at looser thresholds. The coloring is the same except that different shades indicate under which threshold the residue was added to the region. Darker shades indicate more stringent thresholds. **(C)** RoS-GDT: A test analogous to GDT. The largest region of similarity is identified and expanded through the GDT thresholds. The divergent color scheme indicates decreasing modeling accuracy from blue to light red for this region. Bold red indicates that a residue is not included under any of the thresholds.

the models which have the domains in the same orientation as the reference structure. In conjunction with GDT, regions of similarity can then identify which parts of the structures that are globally accurate are locally accurate as well.

3.2 ERoS Plots

ERoS plots can be generated for one or more models of some reference structure. They show how well each structure models the reference by plotting the percent of residues within all regions of similarity under each of twenty thresholds {0.5, 1.0, 1.5, …, 10.0 Å}. The larger the fraction of a structure that is included within regions of similarity under each of the thresholds, the better that structure will perform in the plot. Given that the underlying analysis relies on regions of similarity, ERoS Plots illustrate how well each of a set of structures match their reference structure locally across the whole of their structures.

Figure 3 shows the ERoS plot for the "first models" submitted for the dual-domain CASP13 target T0976 shown in Figure 2. In a CASP experiment, each group may submit multiple models for each target. The models plotted in Figure 3 are those each group submitted as their "first model", the model they wish to be included in the default rankings for the experiment. The curves of the models T0976TS043_1, T0976TS472_1, and T0976TS322_1 are highlighted in blue, green, and purple respectively. The first is the top ranked model by ERoS_Score. It should also be noted that this model is ranked first by lDDT as well [10]. This is not surprising given the similarity between these two scores, but the scores are not directly analogous. The next two models are those ranked as the first and second place models respectively according to GDT_TS. The plot shows that while TS472_1 has a better global score, TS322_1 has more of its structure within regions of similarity across the majority of the thresholds. In other words, TS322_1's local geometries are a better representation of the native.

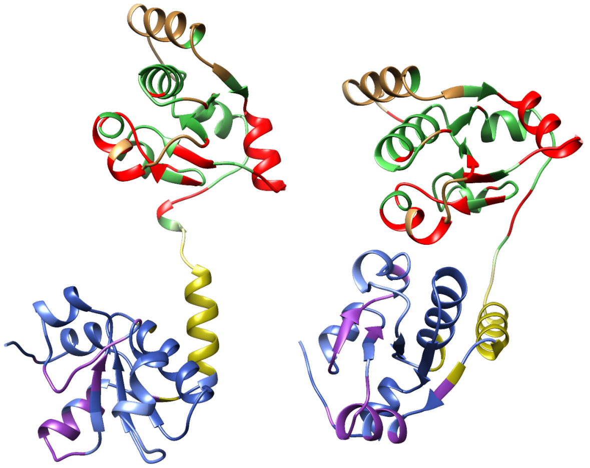

Figure 2: Regions of similarity identified for T0976 and T0976TS043_1. Left: T0976 (the reference) and on the right is T0976TS043_1 (the model) colored according to the expanded regions of similarity illustrated in Figure 1 **Right:** Despite the fact that the two domains in this structure are oriented differently between the reference and the model, the regions of similarity can still be identified and the overall similarity between the structures is apparent.

In any structural comparison, structures with a high degree of global similarity, such as domains being in proper orientations, may not have a high degree of local similarity and vice versa. ERoS plots can be used in conjunction with global measures such as GDT or TM-Score to identify those structures which not only match globally but locally as well. Combined with local accuracy maps and three-dimensional representations, the structures which exhibit both global and local similarity can then be further analyzed to identify exactly which parts of the structures match.

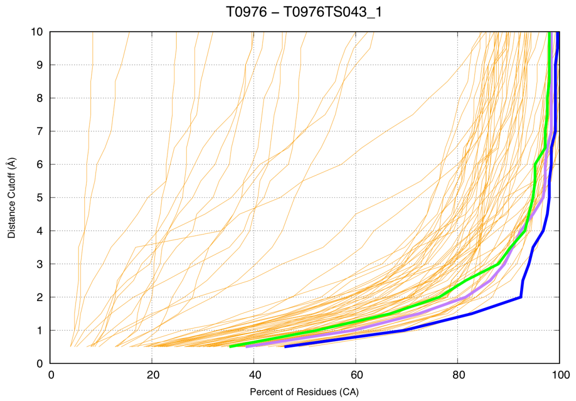

Figure 3: ERoS Plot for CASP13 target T0976. T0976TS043_1 (blue), T0976TS472_1 (green), and T0976TS322_1 (purple) are highlighted. The first is the model ranked best by ERoS_Score. The next two are the top two models ranked by GDT_TS. While TS472_1 is a slightly better global representation of the target (GDT_TS score of 59.2 vs 58.2 for TS322_1), the plot shows that TS322_1 is a better local representation.

3.3 Feasibility Analysis

Since finding regions of similarity relies on solutions to instances of an NP-complete problem (finding the maximum clique of a graph), these techniques were rigorously tested on a set of 88,758 pairs of different structures for identical proteins [5]. Table 1 summarizes the results.

In Table 1, the runtime statistics for five different comparison techniques are presented. As the table shows, the most intensive technique is ERoS-Plot. This matches expectations as ERoS-Plot has the largest number of thresholds to evaluate and therefore depends on solving more instances of the maximum clique problem than any other method. Its average runtime is 7.3 seconds. The maximum time recorded for any individual comparison is 238 seconds. This time is for the structure pair 2drd_C and 2j8s_A. Three of the largest runtimes in Table 1, those for RoS, RoS-GDT, and ERoS, are all for the same pair of structures, 3hhm_A and 2rd0_A. These results speak to the nature of instances of NP-complete problems. For many cases, the solution will be easy, but for some, the solution will be difficult. For the majority of the comparisons, the solutions took on the order of seconds. For a few, the time required was on the order of minutes.

Table 1: Regions of Similarity Techniques Runtimes (ms)

Technique	**RoS**	**RoS-GDT**	**RoS-GDT-HA**	**ERoS**	**ERoS-Plot**
Average	1352	964	935	1749	7315
Median	991	620	539	1226	4350
Max	90457	89791	17813	98558	237509

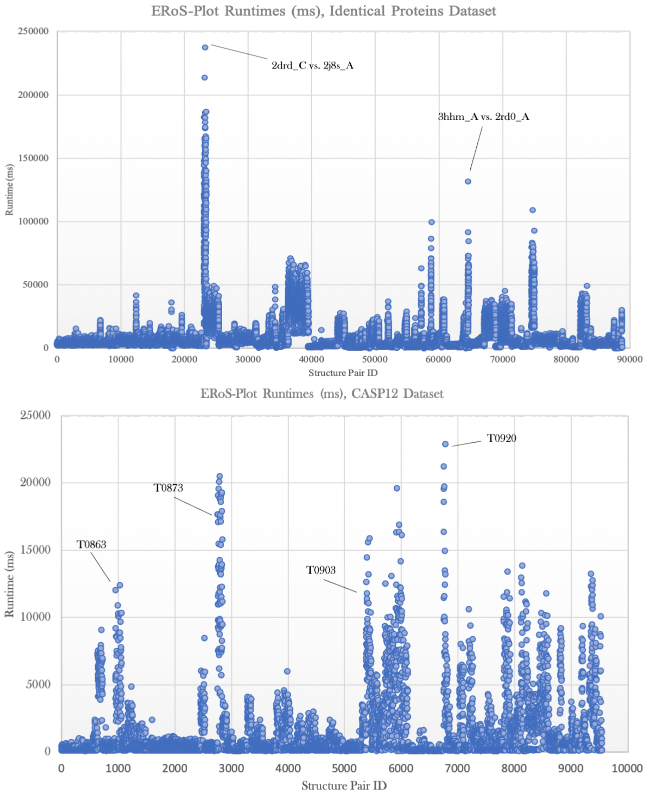

Figure 4: ERoS-Plot Runtimes for the identical proteins dataset and the CASP12 dataset. Top: The identical proteins dataset. Two outlying structure pairs are labeled. The “spikes” are sets of identical structures all pairwise compared with each other. Identical sets tend to have similar runtimes. There is some undetermined property of their underlying similarity graphs that make them difficult instances of the max clique problem. **Bottom:** The CASP12 dataset. The most prominent “spikes” are labeled by the CASP target the structure pairs in it belong to. Note the scale for the runtimes. The range is 0-25 seconds, compared against Figure 4.7 with a runtime range of 0-250 seconds. Evaluation of the CASP12 dataset is feasible with this technique.

The identical proteins dataset is a rigorous test of these techniques. As an example of a practical application, the most intensive technique, ERoS-Plot, was run on the CASP12 dataset containing 131 targets with a total of 9545 models. The average runtime was 1.5 seconds with a median runtime of 553 ms and a maximum runtime of 23 seconds.

Figure 4 shows the ERoS-Plot runtimes for both the identical proteins and CASP12 datasets in Figures 4A and 4B respectively. In the plots, the structure pairs are ordered by groups of identical proteins in Figure 4A and by models for a given target in Figure 4B. In the identical proteins plot, the outlying runtimes group together. These runtimes are from comparisons within sets of multiple structures of the same protein. While a full discussion is beyond the scope of this paper, it should be noted that there is some feature within the similarity graphs that were constructed for these structures that make them difficult instances of the max clique problem. No simple correlation was found between the size or the density of the graph and the runtime, but it can be noted that the longest runtimes tend to belong to large structures that are very similar.

4 Conclusion

Many protein structure comparison methods provide an overall similarity score for structure pairs, but few take an in-depth look at the underlying information of the comparison. GDT [1] partially does by allowing the largest set of residues from a model whose superposition errors on some reference are all under some threshold, but the set identified depends on the superposition and multiple sets are not identified. lDDT [4] allows for an in-depth look at the residues of the structures. It gives each residue a score, measuring how well its local environment (defined as all atoms within some radius of the that residue) is reproduced in a model by finding the fraction of preserved contacts within that environment. Both methods provide scores for individual residues, but they do not identify sets of residues whose environment as a whole is reproduced.

Regions of Similarity is a contact-based protein structure comparison suite which provides a graphical analysis of the similarities between protein structures by performing a detailed analysis of the contacts between them. A region of similarity is a set of residues that together are geometrically similar in both structures. That is, all of their inter-residue distances are the same, within some tolerance threshold. Based on a maximum clique analysis on the graph representing pairwise residue contact similarities between a pair of structures, regions are found independently of the superposition of the structures. Disjoint regions of similarity, those which are independent of each other and possibly shifted relative to each other, can be found. As a result, regions of similarity can be identified in multi-domain structures irrespective of domain movements. It must also be noted that while this method relies on solutions to the NP-complete problem of finding maximum cliques, it has been tested against a rigorous dataset of similar proteins and found to be feasible.

Regions of similarity can easily and meaningfully be visualized. At the sequence level, residues can be colored according to their region and the tolerance threshold at which they were added to that region, showing not only which parts of the sequence form regions of similarity, but also giving an indication of the relative local accuracy of each residue. These local accuracy maps can be generated for sets of structures, allowing a group of models to be compared against some reference structure. These same regions can also be visualized on the individual three-dimensional structures using either PyMOL or Chimera. Lastly, overall accuracy plots (ERoS-Plots) can be produced. These plots show, for each structure in some set compared against a reference, how the fraction of residues identified within regions of similarity changes as the tolerance threshold of similarity is increased from 0.5 Å to 10.0 Å in increments of 0.5 Å. These plots allow for a whole set of structures to be quickly evaluated and for different models within a set to be compared against each other. Those models which are locally accurate over larger portions of the structures will be evident.

Regions of similarity evaluates the local accuracy of pairs of protein structures. While different use cases may have different requirements, binding site analysis may require high levels of local similarity and conformational analysis may focus more on global similarity, in general, when evaluating models against some reference structure, the best models are those which exhibit both global and local accuracy, orthogonal modes of comparison. Only by combining both global and local methods can the similarities and differences between protein structures be fully explored. In conjunction with global measures such as GDT_TS and TM-Score, regions of similarity can be used to identify which of the models that are globally accurate are also locally accurate and furthermore, exactly which parts of models are accurate representations of their corresponding parts in their reference structures. By providing access to information that was not previously available, Regions of Similarity allows for a novel and intuitive look into the similarities and differences between protein structures and can be used in concert with existing metrics to provide a complete global and local comparative analysis of proteins structures.

5 Acknowledgements

We thank Tamjidul Hoque, Michael Levitt, and Krzysztof Fidelis for insightful discussion and critical analysis.

References

[1] A. Zemla, “LGA: a method for finding 3D similarities in protein structures,” *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, Jul. 2003.

[2] Y. Zhang and J. Skolnick, “Scoring function for automated assessment of protein structure template quality,” *Proteins*, vol. 57, no. 4, pp. 702–710, 2004.

[3] R. A. Abagyan and M. M. Totrov, “Contact area difference (CAD): a robust measure to evaluate accuracy of protein models.,” *Journal of Molecular Biology*, vol. 268, no. 3, pp. 678–685, May 1997.

[4] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, “lDDT: a local superposition-free score for comparing protein structures and models using

distance difference tests.," *Bioinformatics*, vol. 29, no. 21, pp. 2722–2728, Nov. 2013.

[5] I. Kufareva and R. Abagyan, "Methods of Protein Structure Comparison," in *Methods in Molecular Biology*, vol. 857, no. 10, Totowa, NJ: Humana Press, 2012, pp. 231–257.

[6] Schrödinger, LLC, "The PyMOL Molecular Graphics System, Version 1.8," Nov. 2015.

[7] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004.

[8] C.-M. Li, Z. Fang, and K. Xu, "Combining MaxSAT Reasoning and Incremental Upper Bound for the Maximum Clique Problem," presented at the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), 2013, pp. 939–946.

[9] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)-Round XII," *Proteins*, vol. 86, no. 1, pp. 7–15, Dec. 2017.

[10] Protein Structure Prediction Center, "13th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction," http://www.predictioncenter.org/casp13, 2018.