Computational Methods for Evaluation of Protein Structural Models

by

Rahul Alapati

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 4, 2019

Keywords: Protein Structure Comparison, Protein Decoy Clustering, Integration of side-chain orientation and global distance-based measures, CASP13, clustQ, SPECS

Approved by

Debswapna Bhattacharya, Chair, Assistant Professor of Computer Science and Software Engineering
Dean Hendrix, Associate Professor of Computer Science and Software Engineering
Gerry Dozier, Professor of Computer Science and Software Engineering

Abstract

Proteins are essential parts of organisms and participate in virtually every process within the cells. The function of a protein is closely related to its structure than to its amino acid sequence. Hence, the study of the protein's structure can give us valuable information about its functions. Due to the complex and expensive nature of the experimental techniques, computational methods are often the only possibility to obtain structural information of a protein. Major advancements in the field of protein structure prediction have made it possible to generate a large number of models for a given protein in a short amount of time. Hence, to assess the accuracy of any computational protein structure prediction method, evaluation of the similarity between the predicted protein models and the experimentally determined native structure is one of the most important tasks. Existing approaches in model quality assessment suffer from two key challenges: (1) difficulty in efficiently ranking and selecting optimal models from a large number of protein structures (2) lack of a similarity measure that takes into consideration the side-chain orientation along with main chain Carbon alpha (Cα) and Side-Chain (SC) atoms for comparing two protein structures.

This thesis attempts to address these challenges by (1) developing a rapid protein decoy clustering algorithm, called clustQ, that employs a multi-model pairwise comparison approach for model quality assessment, based on weighted internal distance comparisons and (2) developing a Superposition-based Protein Embedded Cα-SC (SPECS) score, that integrates the high accuracy version of the Global Distance Test (GDT-HA) metric, and side-chain distance and orientation in a singular framework for protein structure comparison. We show that our methods outperform many traditional and state-of-the-art model quality assessment approaches and similarity measures in terms of accuracy, speed and robustness. In particular, the clustQ method was ranked 6th among the model quality estimators in the 13th edition of the Critical Assessment of Techniques for Protein

Structure Prediction. All of these methods are freely available to the scientific community in the form of software and web-servers.

To

My parents

Mr. Venkateswara Rao, Mrs. Siva Parvathi

and

My advisor

Dr. Debswapna Bhattacharya

Acknowledgments

I would like to express my heartfelt gratitude to my advisor, Dr. Debswapna Bhattacharya, without whose help this work wouldn't have seen the light of the day. I am greatly indebted to Dr. Bhattacharya, for his invaluable guidance, support and enthusiastic encouragement. I would also like to thank the other members of my committee, Dr. Dean Hendrix and Dr. Gerry Dozier for their suggestions and guidance which have greatly improved the quality of my work. Dr. Hendrix has long been an inspiration to me. His teaching style and organizational skills have proved to be one of my best learning experiences. Dr. Dozier has been a great teacher and the numerous interactions I had with him have enlightened me in many ways.

I owe my gratitude to all my professors at Auburn University from whose courses I have acquired the required knowledge for my research. I would also like to thank my colleagues at the Bhattacharya Laboratory for their support. I extend my sincere thanks to the Department of Computer Science and Software Engineering, Auburn University for my Graduate Teaching Assistantship. From the bottom of my heart, I want to thank my parents and my extended family for their love and support. Finally, I am thankful to the Almighty for giving me this worthwhile opportunity.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

RMSD	Root Mean Square Deviation
GDT	Global Distance Test
TM	Template Modeling
CAD	Contact Area Difference
LDDT	Local Distance Difference Test
SPECS	Superposition-based Protein Embedded CA SC score
Cα	Carbon Alpha
SC	Side-Chain
3D	Three-Dimensional
CASP	Critical Assessment of protein Structure Prediction
QA	Quality Assessment
MQAPs	Model Quality Assessment Programs

CHAPTER 1: INTRODUCTION

In this chapter, we will give an overview of the protein model quality assessment problem that we will consider in the thesis. We also describe some of the issues faced in the existing approaches for protein model quality assessment and protein structure comparison that we will try to address and give an outline for the rest of this thesis.

1.1 Protein Model Quality Assessment Problem

1.1.1. Overview of Proteins

Proteins are large, complex molecules that are responsible for doing most of the work in the cells. They are required for the structure, function, and regulation of the body's tissues and organs [1]. They are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-D structure and its specific function.

1.1.2. Protein Structure Prediction

The function of a protein is closely related to its structure and its study can give us valuable information about its function. Hence, protein structure prediction is one of the most important problems in bioinformatics, drug design and in the design of novel enzymes. Protein structure prediction is the inference of the three-dimensional structure of a protein from its amino acid sequence.

Massive amounts of protein sequence data are being produced using the next generation sequencing technologies and this has led to a major gap between the available sequences and the

experimentally determined structures. The existing experimental techniques for protein structure prediction like X-ray crystallography or NMR spectroscopy are both time-consuming as well as expensive. Hence, there is an increasing emphasis on the development of computational protein structure prediction methods, that are much cheaper and faster than the experimental methods.

1.1.3. Protein Structure Comparison

The advancements in computational protein structure prediction methods have led to a growth in the number of structures being determined. Structural comparison methods are thus highly desirable for comparing three-dimensional structures of proteins. A large number of similarity measures have been developed to compare the models with their natives. The aim of these measures is to quantify the correctness of the computationally determined models when compared to the actual native structures.

There are two types of similarity measures for protein structure comparison, namely superposition-based and superposition free similarity measures. Most of the existing similarity measures like GDT-TS, GDT-HA [2] and TMScore [3] are superposition based i.e. they are based on the structural alignment of the proteins. In these measures, an optimal alignment of the protein structures which results in the lowest RMSD is obtained and then the similarity is measured using the distances between the aligned residues. Some of the challenges of superposition-based similarity measures are time-consuming alignment process, not so efficient in case of Free Modeling targets, strongly influenced by domain motions and do not assess the accuracy of local atomic details in the model. Whereas superposition free similarity measures like LDDT [4] and CAD [5] doesn't need any structural alignment and hence are less time consuming and are well suited to assess local model quality.

1.1.4 Model Quality Assessment of Protein Models

The existing computational methods for protein structure prediction have made it possible to generate a large number of models in a short period of time. Therefore, it becomes critical to be able to judge and rank these models based on their quality. This has led to the development of Model Quality Assessment Programs (MQAPs), for evaluating the correctness of predicted protein models.

In general, there are two different kinds of protein quality assessment (QA) methods: single-model quality assessment like ProQ2 [6], QAcon [7] and consensus/multi-model quality assessment like clustQ [8], MUFOLD-WQA [9]. The multi-model QA methods rank and select models using pairwise comparison between all the models in a pool, predicted by different protein structure prediction methods. The single-model QA methods determine protein model quality based on a single model itself, without using the information of other models.

1.2 Existing Approaches and Challenges in Protein Model Quality Assessment and Protein Structure Comparison

The ability to reliably estimate the quality of computationally predicted protein models without comparing them with the native structure, is called the Quality Assessment Problem. The tremendous rise in the computational power has made it possible to produce tens of thousands of models from a single sequence in a day. The availability of thousands of models for a given protein, has made the ability of ranking and selecting optimal models, a challenging task for the MQAPs. The most popular approach is to employ a clustering based multi-model QA program to estimate quality of the models by using pairwise comparisons between the candidate structures available in the pool of predicted models. However, a major challenge faced by the clustering based multi-

model QA programs is that they can be time consuming, which can hinder their ability to be applied to large datasets containing thousands of predicted models.

Most of the existing model-native similarity scores use by default either the main chain Cα or side-chain atoms for quantitating structural similarity. However, protein side-chains play a major role in defining its biologically relevant conformation. Therefore, quantifying the side chain similarities or differences can improve the sensitivities of model-native similarity metrics. In most of the cases, the similarity measures like RMSD, GDT-TS, GDT-HA and TMScore only reflect the confirmation of protein backbone and not the rotameric states of the side chains. There are some similarity metrics like Global Distance Calculation for Side-Chains (GDC-SC) [13] which determines the correctness of the side chain positioning. Although, the GDC-SC measure quantifies the positioning of the side-chain, it only takes into consideration the distances between the side-chain atoms and not their orientation with respect to the backbone. Therefore, there is a lack of a similarity measure, which considers the main chain Cα atoms, SC atoms as well as their orientation for comparing two protein structures. Development of methods capable of tacking these two problems is, therefore, a crucial step forward for solving protein model quality assessment problem and more generally, towards the improvement of computational protein structure prediction.

1.3 Thesis Outline and Contributions

The remainder of this thesis is structured as follows. In chapter 2, we begin by attempting to address the first challenge associated with protein model quality assessment – the need for an efficient protein decoy clustering algorithm. We propose a rapid protein decoy clustering algorithm, called clustQ, that employs a multi-model pairwise comparison approach for model

quality assessment, based on weighted internal distance comparisons. The contents of Chapter 2 are mostly from the manuscript published as:

Alapati, R., Bhattacharya, D. "clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons", In Proceedings of the 2018 ***ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics****, pp. 307-314., ACM, 2018.*

In chapter 3, we discuss the performance of clustQ in CASP13 Quality Assessment category. As per the official results released by CASP13 assessors, clustQ was ranked 6^{th} among the model accuracy estimators from all over the world. The contents of Chapter 3 are mostly from the official results released by the assessors of CASP13.

In chapter 4, we turn our attention to address the second challenge associated with protein structure comparison – the need for a similarity measure that takes into consideration the main chain Cα and SC atoms along with their orientation for comparing two protein structures. We propose a Superposition-based Protein Embedded Cα-SC score, that integrates the high accuracy version of the Global Distance Test (GDT-HA) metric, and side-chain distance and orientation in a singular framework for protein structure comparison. The contents of Chapter 4 are mostly from the manuscript submitted as:

Alapati, R., Bhattacharya, D. "SPECS: Integration of side-chain orientation and global distance-based measures for improved evaluation of protein structural models", ***Proteins: Structure, Function, and Bioinformatics****, 2019.*

Finally, in Appendix A, we provide a brief overview of the freely available software and web-services developed based on the aforementioned methods for the scientific community. These freely available software and web servers would allow researchers from around the world to apply these methods to their own data and these fully automated and computationally inexpensive systems provide a suitable framework for high-throughput proteomics and protein engineering projects.

To summarize, the contributions of this thesis are two-fold: (1) attempting to address two key issues in existing approaches of protein model quality assessment and protein structure comparison – lack of an efficient protein decoy clustering method and lack of a similarity measure that takes into consideration both the Cα and SC atoms along with their orientation for comparing two protein structures and (2) providing the scientific community with access to fast, reliable and freely available software and web-services to facilitate biomedical research.

CHAPTER 2:

clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons

2.1 ABSTRACT

Structure of a protein largely determines its functional properties. Hence, the knowledge of the protein's 3D structure is an important aspect in determining solutions to fundamental biological problems. Structure prediction algorithms generally employ clustering algorithm to select the optimal model for a target from a large number of predicted confirmations (a.k.a. decoy). Despite significant advancement in clustering-based optimal decoy selection methods, these approaches often cannot deliver high performance in terms of the time taken to cluster large number of protein structures owing to the computational cost associated with pairwise structural superpositions. Here, we propose a superposition-free approach to protein decoy clustering, called clustQ, based on weighted internal distance comparisons. Experimental results suggest that the novel weighing scheme is helpful in both reproducing the decoy-native similarity score and estimating pairwise clustering based predicted quality score in a computationally efficient manner. clustQ attains performance comparable to the state-of-the-art multi-model decoy quality estimation methods participating in the latest Critical Assessment of protein Structure Prediction (CASP) experiments irrespective of target difficulty. Moreover, clustQ predicted score offers a unique way to reliably estimate target difficulty without the knowledge of the experimental structure.

clustQ is freely available at http://watson.cse.eng.auburn.edu/clustQ/.

2.2 INTRODUCTION

Knowledge of the three-dimensional (3D) structure of a protein molecule provides crucial insights for addressing fundamental problems in biomedical research. Due to the rapid progress in sequencing technologies, we already have far more sequences than experimental structures, and this gap is likely to grow with the development of next generation sequencing. Hence, there is an increasing emphasis on the development of computational protein structure prediction methods in bioinformatics and computational biology that are much cheaper and faster than the experimental methods. To assess the accuracy of any computational protein structure prediction method, evaluation of the similarity between the predicted protein models and the experimentally determined native structure is one of the most important tasks.

Global Distance Test (GDT-TS) [2], a structural superposition-based approach, is a widely used measure of evaluating model-native similarity and has been a major assessment metric over the last several Critical Assessment of protein Structure Prediction (CASP) experiments [10–13], particularly for evaluating high-accuracy Template Based Modeling (TBM) category. However, in case of moderate to low accuracy Free Modeling (FM) targets, GDT-TS often does not correlate well with the true model-native similarity that is otherwise apparent through visual inspection. Hence, the assessors had to rely on alternative ways by including visual assessments to determine the best model [15]. To overcome the shortcomings of GDT-TS, an alternative metric named Q-score was developed in CASP8 [13] to directly compare the internal distances of a model to its experimentally determined native structure without requiring any structural superposition. The superposition-free nature of Q-score makes the comparison computationally efficient.

In the absence of the native structure, reliably estimating quality of computationally predicted protein models without comparing them with the native structure, the so-called Quality

Assessment (QA) problem is another key challenge in protein structure prediction. One way to tackle this problem is to employ clustering based multi-model QA approach to estimate quality of the models by using pairwise comparisons between the candidate structures available in the pool of predicted models (a.k.a. decoy). Intuitively, structural superposition based multi-model QA approaches [16–22] can be time consuming, hindering the ability to be applied to large datasets containing thousands of decoys.

Here, we first develop an extended version of the original Qscore, called WQ-score, based on weighted internal distance comparisons at four different sequence separations in a superposition-free way. When benchmarked on popular Rosetta [23], I-TASSER [24] and Modeller [25] decoy datasets, WQ-score shows high correlations with most existing model-native similarity metrics [26-28], particularly GDT-TS score [2]. Second, we employ WQ-score based multi-model pairwise comparisons for model quality estimation to develop a rapid protein decoy clustering algorithm called clustQ. Unlike, ModFOLDclustQ [26] which is a multi-model pairwise comparison approach employing the original Q-score for model comparisons, clustQ uses WQ-score that captures both short and long range intra-residue interactions weighted by sequence separation. When benchmarked on CASP 10, 11 and 12 targets, clustQ and the state of the art of the model-native similarity measures show similar correlation with superposition-based metrics such as GDT-TS and GDT-HA [2], with the added advantage of clustQ being computationally much more efficient, particularly when there are large number of decoys. Third, we propose a way to reliably estimate whether a given protein target belongs to the TBM or FM category directly from the result of clustQ without the knowledge of the native structure. Finally, clustQ, delivers comparable performance with the state-of-the-art multi-model QA methods [16–18, 27, 28]

participating in CASP12 [10] as well as model selection approach used in the popular Zhang [30] and QUARK [31] pipelines during CASP10 [12] and CASP11 [11].

2.3 METHODS

2.3.1 WQ-score

Most of the existing model-native similarity metrics, such as GDT-TS, are based on structural alignment i.e. superposition of proteins to perform protein structure comparison. However, GDT-TS performs poorly in case of moderate to low accuracy FM targets [15]. To overcome the limitation, Q-score was introduced during CASP8 as an alternative to accurately highlight the successes and failures of the FM predictions. Qscore estimates the structural similarity between two given protein structures based on comparing their internal distances, thereby overcoming the need for structural alignment. Originally, Q-score is calculated based on the weighted internal distance comparisons at two different sequence separations namely, Q_{short} and Q_{long}. The internal distances are calculated between the Cα atom of each residue i and all N-1 other Cα atoms in the protein, obtaining a matrix $\{r_{ij}\}$. The matrix for the target is designated as $\{r_{ij}^0\}$. For each pair of residues (i-j > 0), Q_{ij} is calculated as in (1).

$$Q_{ij} = \exp[-(r_{ij} - r_{ij}^0)^2] \quad (1)$$

Thus, for a very good prediction, $[(r_{ij} - r_{ij}^0)] = 0$, and $Q_{ij} = 1$. For a very poor prediction, $[(r_{ij} - r_{ij}^0)] >> 0$, and $Q_{ij} = 0$ [7]. Q_{short} measure of a given prediction is calculated by averaging the Q_{ij}, when the best pair and 20, 40, 60, 80 and 100 percent of the ranked pairs that satisfy |i-j| <= 20 were included. Q_{long} was similarly calculated on the lines of Q_{short} for all the pairs that satisfy |i-j| > 20. Q_{short} and Q_{long} indicate the quality of the secondary and tertiary structure of the prediction. WQ-score extends the concept of internal distance comparisons at two different sequence separations

in Q-score to four different sequence separations. It is calculated based on the weighted internal distance comparisons at four different sequence separations namely Q_{narrow}, Q_{short}, Q_{medium} and Q_{long}. Q_{narrow}, Q_{short}, Q_{medium} and Q_{long} are obtained by averaging the Q_{ij} for each pair of residues i, j that satisfy $|i-j| < 6$, $6 \leq |i-j| < 12$, $12 \leq |i-j| < 24$ and $24 \leq |i-j|$ respectively. The weights are assigned as 1, 2, 4 and 8 for Q_{narrow}, Q_{short}, Q_{medium} and Q_{long} respectively. Higher weights are assigned to residues far away from each other in the sequence because such long-range interactions carry more information about the overall protein fold than local shortrange interactions. The WQ-score is calculated as in (2).

$$WQ - Score = \frac{1.0*Q_{narrow}+2.0*Q_{short}+ 4.0*Q_{medium}+8.0*Q_{long}}{15.0} \qquad (2)$$

2.3.2 clustQ

clustQ is a rapid protein decoy clustering algorithm that employs a WQ-score based multi-model pairwise comparison approach for model quality assessment. In clustQ method, we carry out all against all comparisons of server models in order to determine predicted model WQ-scores for individual models. clustQ assigns a score to individual decoy in the pool, on the basis of the average pairwise WQ-score of a decoy when compared against all other decoys. Consequently, clustQ scores range between 0 and 1.

2.4 RESULTS AND DISCUSSION

2.4.1 Comparison between WQ-score and other model-native similarity metrics

We compared WQ-score on the popular Modeller [25], I-TASSER [24] and Rosetta [23] decoy datasets, to determine its correlation with the existing popular model-native similarity metrics. The Modeller decoy set consists of 20 protein targets, each consisting of 300 decoys of 51 to 568 residues in length. The I-TASSER decoy set consists of 56 non-homologous small proteins, each consisting of 300-500 decoys of 47 to 118 residues in length. The Rosetta decoy set consists of 32 proteins, each consisting of 100 decoys of 32 to 85 residues in length. The average Pearson and Spearman correlation coefficients in the Figure 2.1, indicate high correlations between WQ-score and most of the existing model-native similarity metrics.

When benchmarked on I-TASSER and Rosetta datasets, WQ-score exhibits high correlation with all the model-native similarity metrics namely TMScore [3], MaxSub [29], GDT-HA [2], LDDT [4] especially GDT-TS [15] except CAD [5]. Whereas when benchmarked on Modeller decoy dataset, interestingly WQ-score exhibits high correlation with both superposition-based and superposition-free model-native similarity measures [2, 26]. Due to high correlation with GDT-TS and similarity with Q-Score, WQ-score can be used as a reliable measure for model-native similarity in case of both high-accuracy TBM Targets and moderate to low accuracy FM Targets.

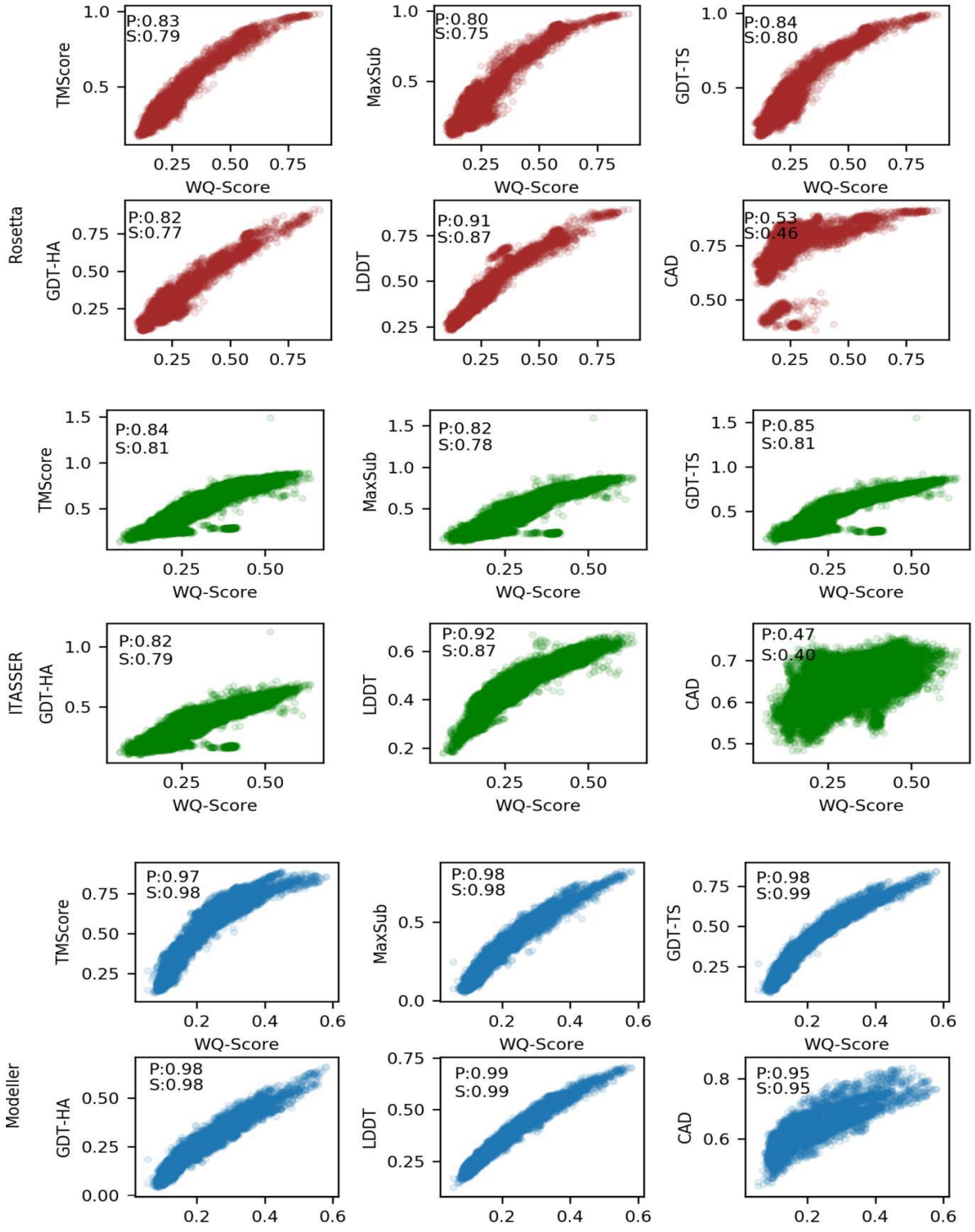

Figure 2.1. Comparisons between WQ-score (horizontal axis) and existing model-native similarity metrics (vertical axis) using Modeller, I-TASSER and Rosetta models. Average Pearson (P) and Spearman (S) correlation coefficients are shown for each plot.

2.4.2 Performance of clustQ in CASP10, 11, 12 datasets

To test the effectiveness of clustQ in scoring and ranking decoys, we benchmarked it on CASP 10 [12], 11 [11] and 12 [10] Stage 1 and Stage 2 targets. The CASP 10 Stage 2 decoy set consists of 113 protein targets, each consisting of 150 decoys. The CASP 11 Stage 2 decoy set consists of 98 protein targets, each consisting of 150 decoys. The CASP 12 Stage 2 decoy set consists of 86 protein targets, each consisting of 150 decoys. We carried out all against all comparisons of server models i.e. pairwise clustering in order to determine predicted model similarity scores like TMScore, GDT-TS, GDT-HA, LDDT and CAD Score for individual models. We compared clustQ and the existing model-native similarity metrics with the True GDT-TS and GDT-HA scores.

Figure 2.2 shows that the clustQ score and the state of the art model-native similarity measures show similar correlation with superposition-based metrics such as GDT-TS and GDT-HA, even though the nature of the scores is different. GDT-HA is a more stringent version of GDT-TS score. Like GDT-TS, GDT-HA is also derived from 4 independent superposition-based scores, but their threshold distances of 0.5, 1, 2 and 4 Å are half the size of those used for GDT-TS. Notably, this is true for both Pearson's correlation coefficient, which depends on the linear relationship between the two scores and also on the Spearman's rank correlation, which indicates the extent to which ranking by True GDT-TS or GDT-HA agrees with ranking by clustQ score without the assumption of the linear relationship between the two scores. Surprisingly, the correlation between the True GDT-TS and CAD is very low, even though LDDT, a superposition free model-native similarity metric like CAD, is highly correlated with True GDT-TS. The low correlation of CAD Score with True GDT-TS and GDT-HA can be attributed to the way CAD score treats the missing residues in

the model. Both the failure to include the residue into the model and the failure to predict all of its contacts are treated identically by the CAD Score.

Furthermore, we compared the performance of clustQ with the existing model-native similarity metrics, in terms of time taken to rank the decoys. We tracked the average time taken by clustQ and other model-native similarity metrics to compare a decoy with all the other decoys in the protein target. We observed that in case of CASP10, 11, 12 Stage 1 datasets, on an average clustQ is 4.8 times faster than TMScore, 264 times faster than LDDT and 173 times faster than CAD Score. The performance difference becomes even more pronounced in case of CASP10, 11, 12 Stage 2 datasets (as shown in Figure 2.3), on an average clustQ is 5.2 times faster than TMScore, 291 times faster than LDDT and 168 times faster than CAD Score. Hence, clustQ is computationally much more efficient when compared to others, particularly when there are large number of decoys. This premise holds true in case of 16,950 decoys in CASP10 dataset, 14,700 decoys in CASP11 dataset and 7,500 decoys in CASP12. Collectively, the results demonstrate that clustQ is rapid offering orders of magnitude speedup compared to TMScore, CAD and lDDT.

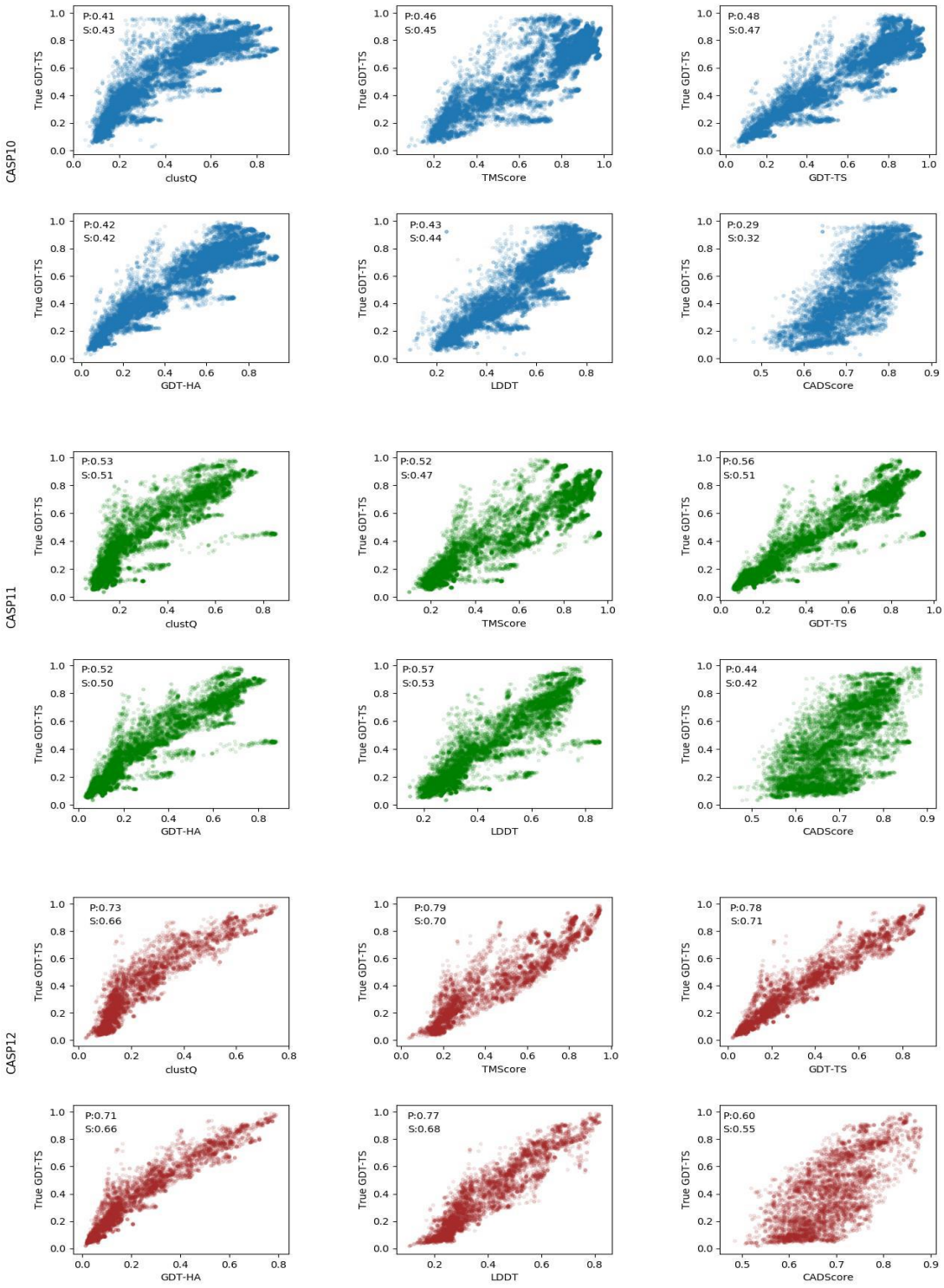

Figure 2.2. Comparisons between True GDT-TS Score (vertical axis) and clustQ as well as pairwise clustering using other similarity metrics (horizontal axis) in CASP 10, 11 and 12 Stage 2. Average Pearson (P) and Spearman (S) correlation coefficients are shown for each plot.

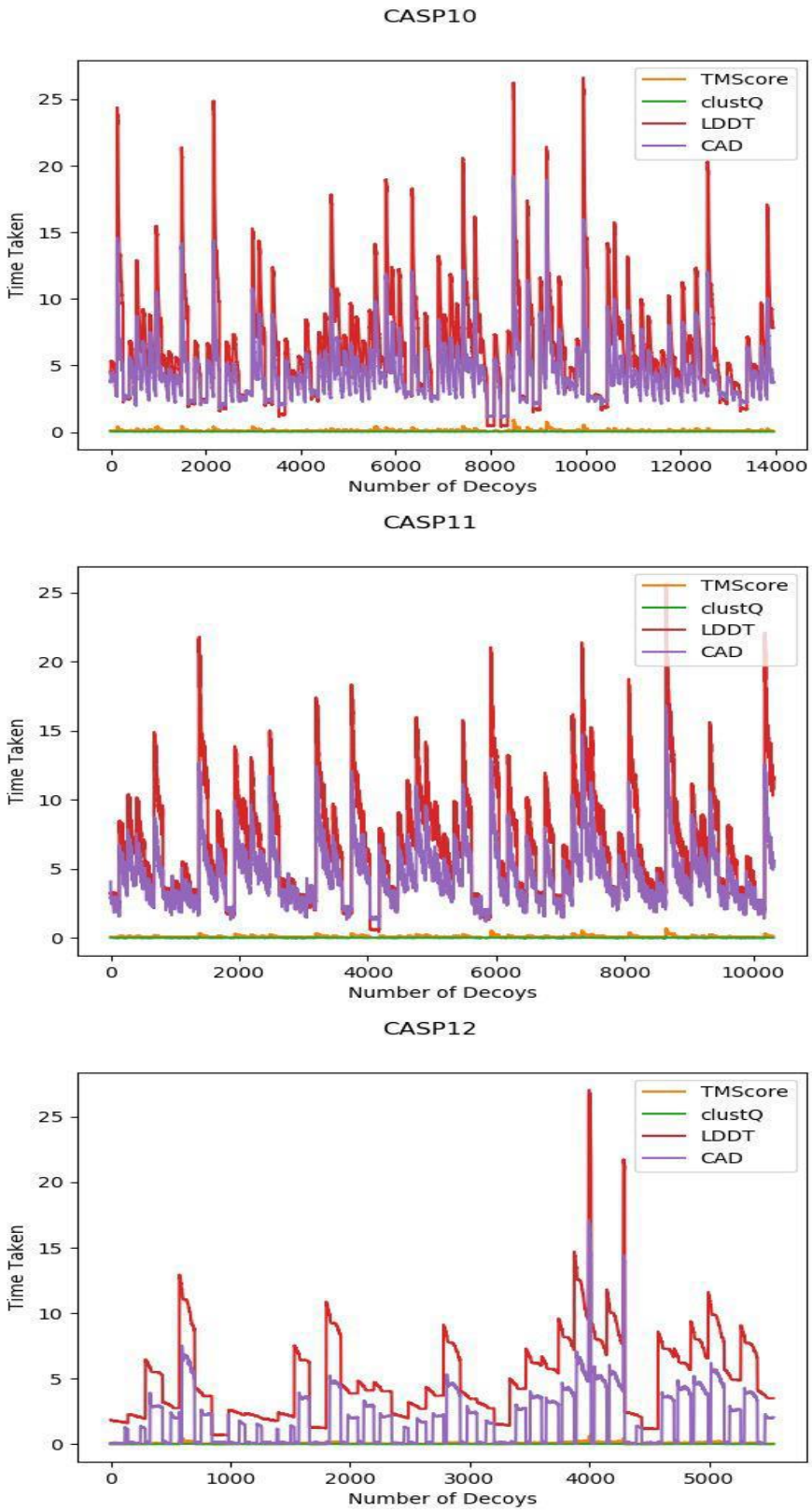

Figure 2.3. Comparisons between clustQ and pairwise clustering using other similarity metrics in terms of time taken in CASP 10, 11 and 12 Stage 2 datasets.

2.4.3 Predicting target difficulty using clustQ

We further investigated the possibility of applying clustQ in classifying a given target as TBM or FM without the knowledge of its native structure, using CASP 10, 11 and 12 Stage 2 targets. When the best predicted clustQ score for a given target is greater than a specified threshold value, it is predicted as TBM. To determine the threshold value for target classification, we varied the threshold from 0.3 to 0.6 in step size of 0.1. For each threshold value, we calculated the True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Precision, Recall and Accuracy. A classification is considered as TP if both CASP and clustQ term it as "EASY", as TN if both CASP and clustQ term it as "HARD", as FN if CASP terms it as "EASY" and clustQ terms it as "HARD" and as FP if CASP terms it as "HARD" and clustQ terms it as "EASY".

Precision, Recall and Accuracy are calculated as in (3).

$$\text{Precision} = TP / (TP + FP)$$
$$\text{Recall} = TP / (TP + FN) \quad (3)$$
$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

As shown in Figure 2.4, threshold values greater than 0.4 result in a precision of 1. Increasing threshold values results in lower recall and accuracies. We therefore, predict that a target to be TBM if clustQ score is greater than 0.4, otherwise FM. The results demonstrate that the top clustQ score is a reliable estimate for target difficulty without the knowledge of native structure.

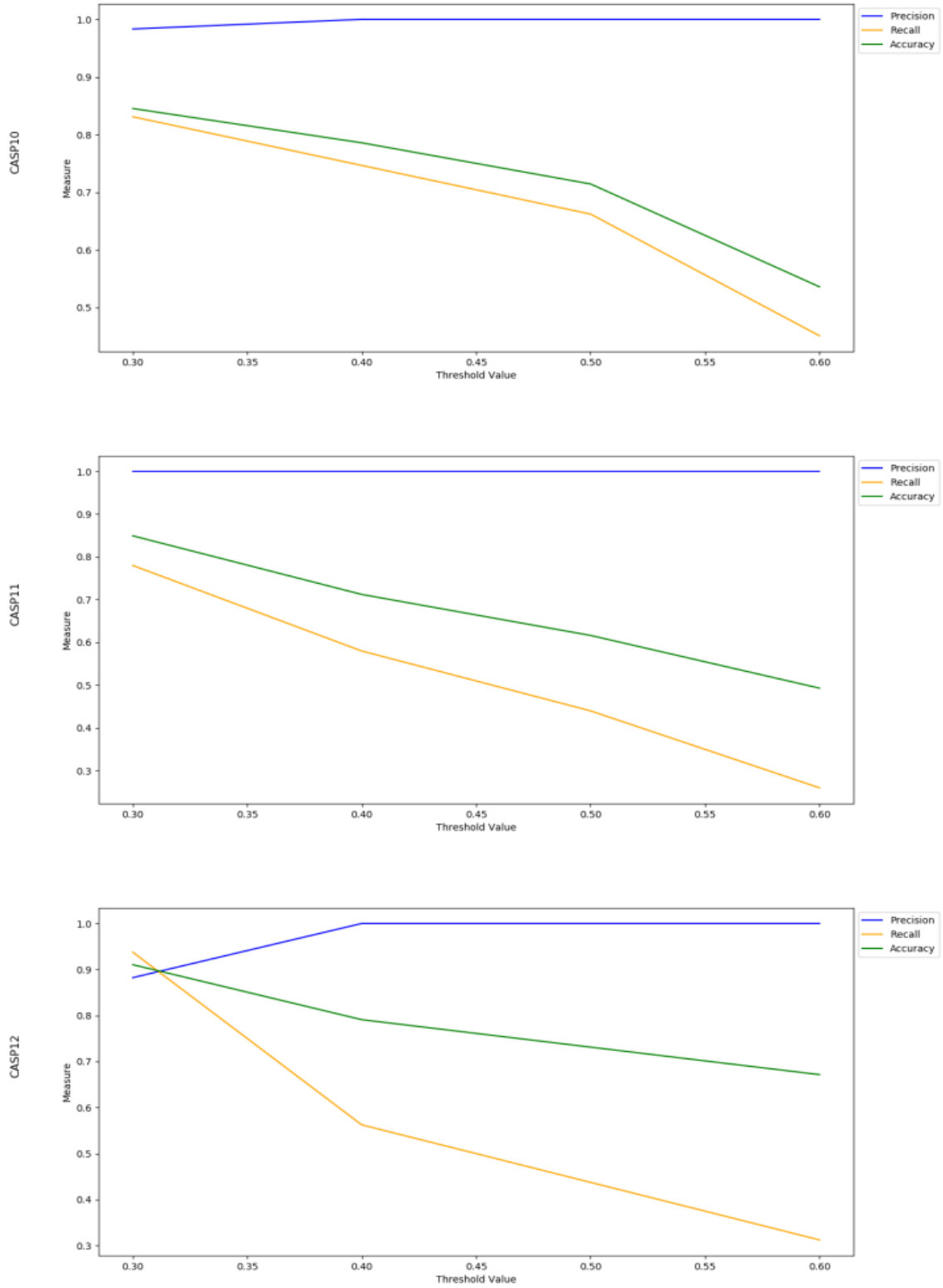

Figure 2.4. Effect of threshold on clustQ score in target difficulty classification in CASP 10, 11 and 12 datasets.

2.4.4 Comparison between clustQ and top multi model QA methods participating in CASP12

We benchmarked the performance of clustQ against top performing multi-model QA approaches participating in CASP12 [10] namely, Meshi_con server, Pcomb_domain, ModFoldclust2 and Pcons [16–18, 27, 28]. Pcomb_domain, ModFoldclust2 and Pcons are consensus methods that are highly accurate in identifying good models over bad models and in reliably predicting regions of models. The quasi-single methods from the ModFOLD 6 family being the next best choice [27]. Meshi_con server is less dependent on the dataset composition than that of other clustering methods [27].

We selected the top model ranked by each of these QA methods and calculated its True GDT-TS and GDT-HA scores [2] by comparing each decoy against its native. Then we calculated the average True GDT-TS and GDT-HA scores [2] of top models ranked by each of these QA methods over all the targets in CASP 12 Stage 2 and compared them against the average True GDT-TS and GDT-HA scores of top models ranked by clustQ.

Table 2.1 shows that clustQ delivers a comparable performance with the state of the art QA methods and is also better than Meshi_con and Pcons QA methods. The performance of clustQ is consistent across both the GDT-TS and GDT-HA scores [2].

Table 2.1: Comparison between clustQ and state of art QA Methods in CASP 12. Average GDT-TS and GDT-HA comparison between clustQ and state of the art QA Methods participating in CASP 12

QA Method	Average GDT-TS	Average GDT-HA
Pcomb_domain	0.5098725	0.3652375
ModFoldclust2	0.4968475	0.3572975
clustQ	0.4761375	0.3395075
Meshi_con	0.332503571	0.239453571
Pcons	0.3304925	0.232665

2.4.5 Performance of clustQ in Zhang and QUARK CASP decoy dataset

We evaluated the performance of clustQ in Zhang [30] and QUARK [31] decoy datasets and compared it directly with the model selection approach used in Zhang [30] and QUARK [31] pipelines. The Zhang and QUARK decoy set for single domain targets in CASP 10 consists of 43 protein targets, each consisting of 145-825 decoys of 67 to 540 residues in length. The Zhang and QUARK decoy set for single domain targets in CASP 11 consists of 64 protein targets, each consisting of 170-1550 decoys of 44 to 525 residues in length.

We selected the decoy pool ranked by the Zhang and QUARK pipelines during CASP 10 and 11. Then, we applied clustQ to the rank the decoys in Zhang and QUARK datasets. We selected the top models ranked by each of the Zhang and QUARK pipelines and calculated its True GDT-TS and GDT-HA scores by comparing it against its native. We also calculated the True GDT-TS and GDT-HA score of top model selected by clustQ, by comparing it against its native. Then, we

calculated the average of the True GDT-TS and GDT-HA scores of top models selected by clustQ, Zhang and QUARK pipelines over all the protein targets. We performed the above experiment by dividing the targets based on their difficulty. Finally, we compared the average True GDT-TS and GDT-HA scores of top models selected by Zhang and QUARK pipelines with that of the top models selected by clustQ to determine the performance of clustQ as a model selection approach.

As shown in Table 2.2, for “Trivial”, “Easy”, “Hard”, and “Very Hard” target categories, clustQ attaints similar performance when compared to the QA methods employed in Zhang and QUARK pipelines, while outperforming them in few cases. To investigate the statistical significance of the performance difference between clustQ versus Zhang and QUARK. We performed “one sample T-test” against Zhang and QUARK versus clustQ with the null hypothesis that the GDT-TS score difference between the top model selected by clustQ and Zhang or QUARK is zero. As shown in Table 2.3, clustQ is statistically indistinguishable compared to Zhang and QUARK at 95% confidence level, warranting the comparable performance.

Table 2.2: Comparison between clustQ and state of art QA Methods used in Zhang, Quark Pipelines. Average GDT-TS and GDT-HA comparison between clustQ and state-of-the-art QA Methods used in Zhang, Quark pipelines for Easy, Trivial, Hard and Very Hard Targets

		QA Method	GDT-TS	GDT-HA
EASY Targets	**CASP10**	clustQ	0.6737	0.4897
		Zhang	0.6736	0.4903
		QUARK	0.6728	0.4901
	CASP11	clustQ	0.5420	0.3951
		Zhang	0.5451	0.3940
		QUARK	0.5363	0.3874
Trivial Targets	**CASP10**	clustQ	0.7851	0.5932
		Zhang	0.7889	0.5989
		QUARK	0.7846	0.5908
	CASP11	clustQ	0.7618	0.5795
		Zhang	0.7523	0.5687
		QUARK	0.7494	0.5653
Hard Targets	**CASP10**	clustQ	0.4059	0.2723
		Zhang	0.3963	0.2585
		QUARK	0.3869	0.2480
	CASP11	clustQ	0.4659	0.3042
		Zhang	0.4651	0.3130
		QUARK	0.5044	0.3409

Very Hard Targets	**CASP10**	clustQ	0.3371	0.1954
		Zhang	0.3584	0.2036
		QUARK	0.3200	0.1832
	CASP11	clustQ	0.2769	0.1633
		Zhang	0.2724	0.1653
		QUARK	0.2644	0.1606

Table 2.3: Statistical significance test. One Sample t-test for CASP 10 and 11 Zhang and QUARK Targets.

	CASP round	**t-value**	**p-value**
Zhang	CASP10	-0.36643	0.7159
	CASP11	1.1417	0.2586
QUARK	CASP10	0.70359	0.4856
	CASP11	0.19091	0.8493

2.5 CONCLUSION

We developed a decoy clustering method clustQ that employs a WQ-score based pairwise comparison approach to rank the decoys. WQ-score is calculated based on weighted internal distance comparisons at four different sequence separations. Higher weights are assigned to residues far away from each other in the sequence because such long-range interactions carry more information about the overall protein fold than local short-range interactions. The experimental results suggest that WQ-score is highly correlated with the existing model-native similarity

metrics, especially the GDT-TS with the added advantage of clustQ being computationally much more efficient, particularly when there are a large number of decoys. clustQ delivers comparable performance with the state of the art QA methods participating in the recent CASP experiments (CASP10, 11, 12) in addition to being comparable to the decoy selection method employed in the popular Zhang and QUARK pipelines in CASP10 and CASP11. Moreover, clustQ offers a unique way to reliably estimate difficulty of a target without knowledge of the native. Collectively, these results demonstrate that clustQ is an important addition to protein decoy clustering in particular and protein structure modeling in general.

CHAPTER 3:

clustQ's Assessment in CASP13

In this chapter, we present an evaluation of the performance of clustQ in CASP13 quality assessment experiment.

3.1 INTRODUCTION

Critical Assessment of protein Structure Prediction (CASP) [10-13], is a community-wide, worldwide experiment for protein structure prediction that takes place once every two years. It provides researchers with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state of the art methods in protein structure modeling to the research community. The main aim of CASP is to help advance the methods of identifying protein three-dimensional structure from its amino acid sequence.

Model Quality Assessment Programs (MQAPs) [6-9] are developed in order to rank and select the computationally predicted protein models. The increase in the number of protein structure prediction methods, has necessitated the need for MQAPs that can rapidly assign a quality accuracy to each computationally predicted model. This quality accuracy can then be used to estimate the accuracy of a specific model and to rank many alternative models to select the most accurate model. MQAPs can be divided into two categories based on the information they use. Consensus based methods use pairwise comparison between all the models in a pool, to estimate the quality of the models, whereas the single-model QA methods use various features calculated from the structure.

Estimation of Model Accuracy (EMA, a.k.a. QA) category in CASP, evaluates the ability of MQAPs in providing useful accuracy estimates for the overall accuracy of models.

3.2 Performance of clustQ in CASP13

The consensus based QA method, clustQ [8], was first blindly tested in CASP13 QA experiment, 2018 with the group name Bhattacharya-ClustQ (Server group 014). Here, we perform a comparative analysis of clustQ against all the groups participating in CASP13 QA category. A total of fifty methods participated in CASP13 QA experiment including both seventeen consensus based and thirty three single-model QA methods. All the fifty QA methods were given three days per target to rank the models of sixty four targets.

In the recent community-wide experiment, CASP13, clustQ was ranked sixth among the forty eight QA groups as per the official assessment of CASP13 experiment. In Table 3.1, we summarize the rankings of the forty eight groups participating in CASP13 based on the average difference in accuracy between the models predicted to be the best and the actual best according to the GDT_TS score [2]. For each group, the differences are averaged over all predicted targets for which at least one structural model had a GDT_TS score above 40.

Table 3.1: Rankings of CASP13 QA methods based on the Average difference in accuracy between the models predicted to be the best and the actual best according to the GDT_TS score over best150 dataset

Rank	Group	Number of Targets	Avg.GDT_TS
1	MULTICOM_CLUSTER	64	5.162
2	UOSHAN	63	5.555
3	MUFoldQA_M	64	6.264
4	MULTICOM-CONSTRUCT	64	6.865
5	Davis-EMAconsensus	64	6.888
6	**Bhattacharya-ClustQ**	64	7.071
7	ModFOLDclust2	64	7.178
8	ModFOLD7_rank	64	7.525
9	MUfoldQA_T	64	7.536
10	RaptorX-DeepQA	60	7.595
11	SBROD-plus	60	7.875

12	ProQ3D	64	8.259
13	SBROD-server	60	8.345
14	SBROD	57	8.394
15	CPClab	60	8.639
16	Wallner	63	8.657
17	ProQ3	64	8.843
18	Grudinin	57	8.963
19	ProQ3D-lDDT	64	9.132
20	Pcomb	64	9.136
21	MESHI	62	9.21
22	MESHI-enrich-server	64	9.311
23	FaeNNz	64	9.355
24	Pcons	64	9.495
25	ProQ3D-TM	64	9.986
26	ModFOLD7	64	10.589
27	VoroMQA-B	64	10.742
28	MUFold_server	64	10.782
29	Bhattacharya-SingQ	64	10.799
30	FALCON-QA	59	10.821
31	ProQ3D-CAD	64	10.951
32	ModFOLD7_cor	64	10.989
33	MULTICOM-NOVEL	64	11.016
34	ProQ4	64	11.323
35	VoroMQA-A	64	11.506
36	MASS1	64	11.749
37	Bhattacharya-Server	64	11.99
38	LamoureuxLab	62	12.383
39	Kiharalab	64	12.527
40	MASS2	64	12.984
41	ProQ2	63	13.018
42	SASHAN	63	13.33
43	3DCNN	46	13.566
44	MUfoldQA_S2	64	14.123
45	PLU-AngularQA	64	14.765
46	Jagodzinski-Cao-QA	64	17.585
47	MESHI-server	60	20.882
48	PLU-TopQA	64	21.443

From Table 3.1, it can be noted that clustQ shows a relatively good performance when compared to all the single-model QA methods. To further analyze the performance of clustQ on individual targets, we calculated the Pearson and Spearman correlations between the scores assigned by clustQ, GDT-TS and GDT-HA. In the Figure 3.1, we present the per-target Pearson and Spearman correlation for clustQ with respect to GDT-TS and GDT-HA for 34 targets whose natives could be identified as of writing this chapter. It shows that clustQ is well correlated with both GDT-TS and GDT-HA (average per-target Pearson correlation ~ 0.88 and Spearman correlation ~ 0.87).

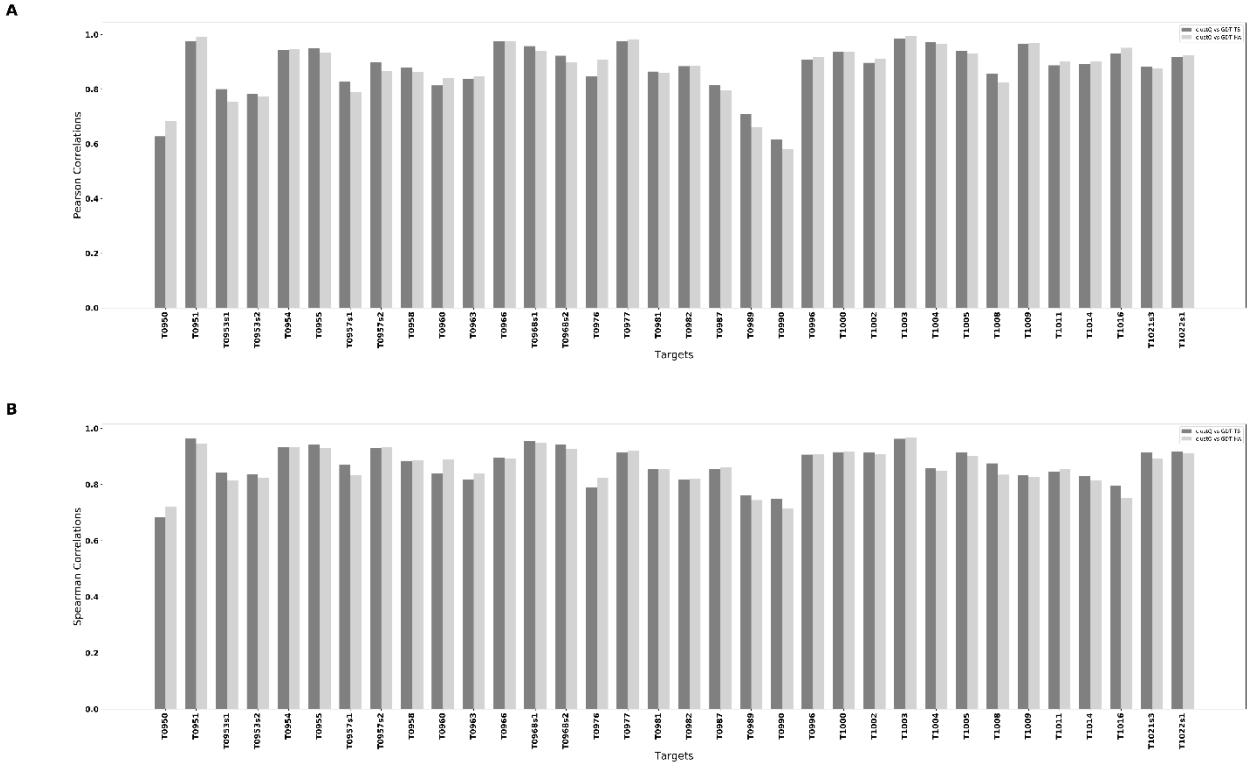

Figure 3.1. Performance of multi-model QA method clustQ for 34 CASP13 targets. (A) Per-target Pearson correlation with respect to GDT-TS and GDT-HA, (B) Per-target Spearman correlation with respect to GDT-TS and GDT-HA.

3.3 CONCLUSION

In this chapter, we systematically analyze the performance of a rapid protein decoy clustering algorithm, clustQ in a completely blind mode on the targets issued for model accuracy estimation in recently concluded CASP13 experiment. When compared with the state-of-the-art QA methods participating in CASP13, clustQ is observed to perform more consistently than the all single-model QA methods and most of the consensus based QA methods.

CHAPTER 4:

SPECS: Integration of side-chain orientation and global distance-based measures for improved evaluation of protein structural models

4.1 ABSTRACT

Protein structure prediction is an important yet highly challenging open problem in structural bioinformatics. Significant advancements in the field of protein structure prediction, have necessitated the need for accurate, reliable and efficient protein structure comparison methods. Despite its apparent simplicity, evaluation of protein models against its native structure is quite complex and non-trivial in nature. A number of protein structure similarity measures proposed till date, either consider the backbone Cα atoms or the Side-Chain atoms for comparing a model against its native. However, the true nativity of a protein model can only be determined based on both the backbone and side-chain conformations. Here, we propose a superposition-based evaluation measure, called SPECS, which determines the similarity of two protein structures by comparing both the backbone and side-chain atoms. Experimental results show that SPECS is a reliable measure for model-native similarity in case of both regular domain models and high accuracy refined models. In addition to high correlations with TMScore and GDT-HA score, SPECS also demonstrates a strong affinity in promoting the physical realism of structural models. When benchmarked on a special monomer dataset, SPECS is found to be a reliable model-native similarity measure for side-chain conformations. Moreover, our study illustrates that the usage of both the backbone and the side-chain conformations improves protein structure comparison. SPECS web server is freely available at http://watson.cse.eng.auburn.edu/SPECS/.

4.2 INTRODUCTION

The biological function of a protein is determined from its three dimensional structure. The knowledge of the three dimensional structure of a protein helps us in understanding its function, and also helps us in modifying and controlling it. However, from an experimental standpoint determining the three dimensional structure of a protein is expensive, time consuming and requires high levels of expertise. Due to these challenges and the rate at which new protein sequences are being discovered, it is practically impossible to solve the structures of all the proteins experimentally [44]. The high demand for protein structures has given rise to the development of a large number of protein structure prediction methods, which computationally predict the three dimensional structures of proteins. These computationally predicted three dimensional protein structures are used in many areas of biomedicine, ranging from approximate family assignments to precise drug screening.

The cheaper and the computationally efficient protein structure prediction methods predict a large number of protein models for a given protein. This increase in the number of computationally predicted models has placed the protein structure comparison methods, the only way to determine the accuracy of the predicted models, at an unprecedented critical position [45]. The protein structure comparison methods assess the protein models against their experimentally determined native structure and helps us in identifying the predicted models which are useful for biomedical research. Despite its complex and non-trivial nature, many protein structure evaluation measures have been developed over the last few years and till date it still continues to be an important yet challenging line of research [46].

Most of the existing model-native similarity measures either superposition-based or superposition free are distance-based measures [2-4, 47, 48], which determine the level of similarity between

two protein models based on the distance between either the backbone Cα atoms or the side-chain atoms. Root Mean Square Deviation (RMSD) [49] is the most commonly used superposition-based model-native similarity score. It is the measure of the average distance between the backbone Cα atoms of the superimposed proteins. The lower the RMSD, the better the model is in comparison to the native. RMSD between two sets of superimposed atomic coordinates t and u, is defined as follows:

$$RMSD(t,u) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((t_{ix} - u_{ix})^2 + \left(t_{iy} - u_{iy}\right)^2 + (t_{iz} - u_{iz})^2)} \quad (1)$$

where n is the total number of superimposed atoms.

One major drawback of the RMSD, is that it is heavily dependent on the quality of the superposition of the protein structures and is also sensitive to the outlier regions created by poor modeling of the individual loop regions in the structures [46].

Global Distance Test (GDT) [2], a structural superposition-based approach and a widely used assessment metric in Critical Assessment of protein Structure Prediction (CASP) [50, 51], is a more accurate measurement than RMSD [46]. It is defined as the largest set of amino acid residue's backbone Cα atoms in the model falling within a defined distance cutoff of their position in the native. Here, multiple superpositions of two protein structures, each including the largest set of superimposable atoms are considered and the maximal residue set for each cutoff is selected, followed by averaging over several predetermined cutoffs. For GDT-TS [2] measure, predetermined cutoffs of 1, 2 , 4 and 8 Å are considered for calculation of the maximal residue set. The high accuracy version of the GDT measure, GDT-HA [48], uses smaller cutoffs of 0.5, 1, 2 and 4 Å for the calculation of the maximal residue set. The range of GDT-TS and GDT-HA measures are from 0 to 1, higher the score, the better the model is in comparison to the native.

All the above mentioned standard model-native similarity measures only consider the backbone Cα atoms for determining the structural similarity. However, it is also known that the protein side-chains play a major role in protein-protein interactions and are closely related to their biological function. Therefore, quantifying the side-chain variations can improve our understanding of the side-chain conformations, which in turn will help improve the quality of protein structure prediction methods [52]. Global distance calculation for sidechains (GDC-SC) [53] is a measure, which determines the correctness of the side-chain positioning. GDC-SC metric is similar to GDT-TS, while backbone Cα atoms are used in GDT-TS calculation, a single reference atom from each sidechain is used in GDC-SC calculation. Like GDT-TS, GDC-SC will first determine the optimal superposition between the backbone Cα atoms of the model and native, and then the distance between the side-chain reference atoms in model and native is calculated. Finally, the distances are assigned to ten different bins ranging from 0.5 Å to 5.0 Å with a step size of 0.5 Å. GDC-SC measure is calculated as follows:

$$GDC - sc = \frac{200 \sum_{i=1}^{10} (k - i + 1) P_i}{k(k+1)} \quad (2)$$

where k = 10 is the number of bins and P_i is the fraction of reference atoms assigned to bin i. The range of GDC-SC measure is from 0 to 100, higher the score, lower the distance between the atoms in the model and native.

Although, the GDC-SC measure quantifies the positioning of the side-chain, it only takes into consideration the distances between the side-chain atoms and not their orientation with the backbone atoms. Orientation also plays a major role in the positioning of the side-chain and the backbone Cα atoms and the structure of a protein model is only accurate, when the orientation between the backbone and side-chain atoms in the model matches with that in the native [52]. Hence, model-native similarity measures which takes into consideration the backbone Cα atoms,

the side-chain atoms and their orientations are the right choice for determining the true nativity of a computationally predicted protein model.

Here, we develop a superposition-based model-native similarity measure, Superposition-based Protein Embedded CA SC (SPECS) score, which extends the concept of GDT-HA score, GDC-SC score and also quantifies the backbone side-chain orientations in model and native structures. Firstly, SPECS is highly correlated with both backbone Cα based and side-chain based model-native similarity metrics. When benchmarked on the CASP 12, 13 regular domain targets and the CASP 12, 13 high accuracy refinement targets [54], SPECS shows high correlations with both the backbone Cα based and side-chain based model-native similarity metrics [2-4, 48, 53, 55], particularly the GDT-TS, GDT-HA and GDC-SC scores. Secondly, SPECS is fairly sensitive to structural features such as steric clashes or deviations in residue geometries [56]. When compared with GDT-HA score and TMScore using MolProbity [57] as a structure quality evaluation suite, SPECS is more consistent with the physical realism of the models in the 3D Robot set [58]. Thirdly, SPECS also acts as a reliable evaluation metric, for determining the correctness of the side-chain positioning. When benchmarked on the monomer proteins side-chain positions predicted by three most widely used side-chain conformation prediction programs [59], SPECS is moderately correlated with their angular RMSDs, when all the backbone Cα based scores are perfect.

4.3 MATERIALS AND METHODS

4.3.1 Parameterization of united-residue model of superimposed protein structures

We use the united-residue representation in Figure 1, to parameterize the backbone Cα atoms and the side-chain reference atoms in a protein structure. The protein structure consists of a sequence of Cα atoms and the side-chain reference atoms which are attached to the Cα atoms. All the atoms in the protein structure are connected using virtual bonds which are denoted by a thick black line in Figure 4.1. The Cα position of the residue i in the protein is represented by $C\alpha_i$ and the corresponding side-chain reference atom attached to $C\alpha_i$ is represented by SC_i.

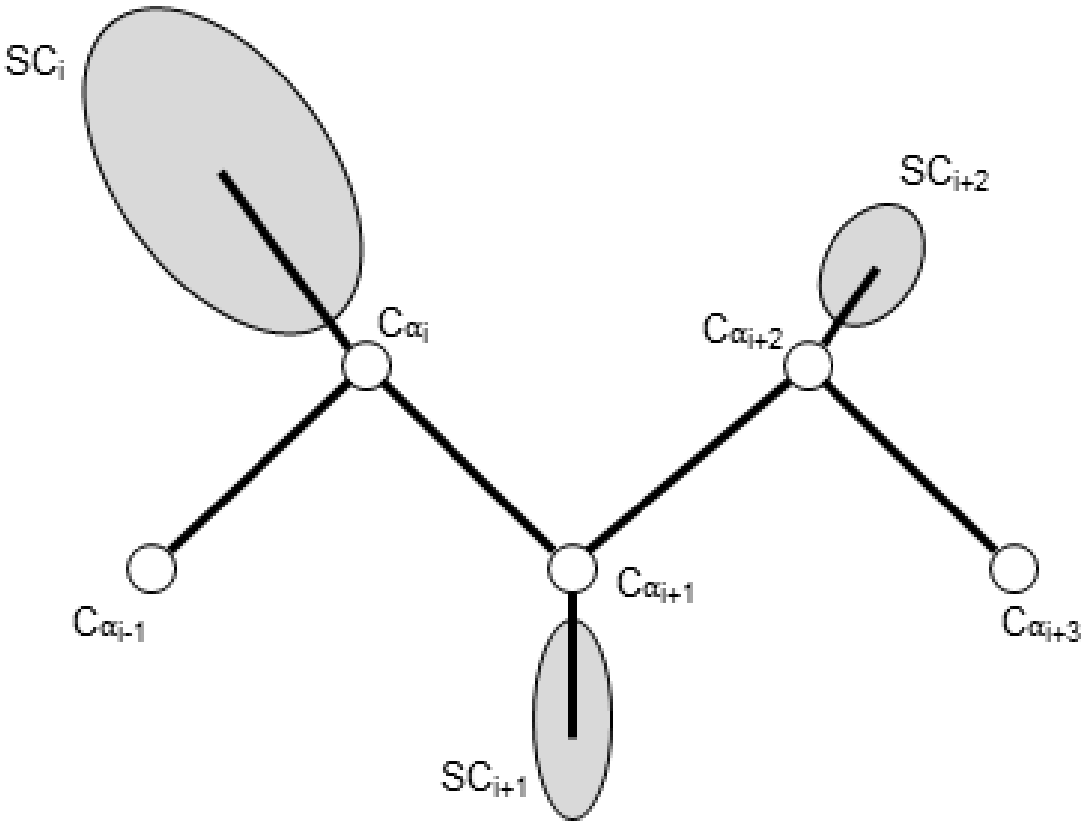

Figure 4.1. Parameterization of united-residue model of a protein structure. United-residue model parameterized using virtual lengths for backbone and side-chain.

Now, we use our united-residue model (Figure 4.2) [60, 61], to represent the superimposed protein structures, in which the C-alpha (C_α) and side-chain reference atoms (SC) are linked by virtual bonds. A side-chain reference atom is obtained by computing the centroid of all the atoms in the side-chain cloud. We parameterize the backbone Cα position of the residue i in the model as $C\alpha_i$ and the residue j in the native as $C\alpha_j$. The corresponding side-chain reference atom i in the model is represented as SC_i and the reference atom j in the native is represented as SC_j. The distance between the side-chain reference atoms is denoted using r_{ij} and $\vec{r}_{ij}$ is a vector which determines the relative position of the side-chain reference atoms and also links them. $\hat{u}_{ij}^{(1)}$, $\hat{u}_{ij}^{(2)}$ are the unit vectors which represent the direction of the C_α and SC virtual bonds in the model and native, respectively. $\theta_{ij}^{(1)}$ is the virtual planar angle between $\hat{u}_{ij}^{(1)}$ and $\vec{r}_{ij}$ in the model and $\theta_{ij}^{(2)}$ is the virtual planar angle between $\hat{u}_{ij}^{(2)}$ and $\vec{r}_{ij}$ in the native. Φ_{ij} is the virtual angle of counterclockwise rotation between $\hat{u}_{ij}^{(2)}$ and $\vec{r}_{ij}$ in the plane defined by $\hat{u}_{ij}^{(1)}$ and $\vec{r}_{ij}$.

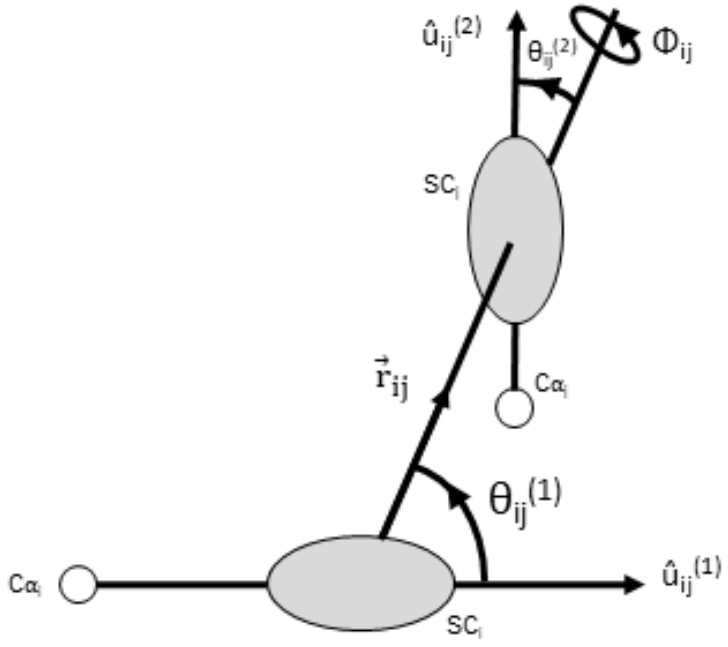

Figure 4.2. Parameterization of united-residue model of superimposed protein structures. United-residue model parameterized using virtual lengths and virtual angle pairs for backbone and side-chain.

4.3.2 SPECS : Superposition-based Protein Embedded CA SC score

SPECS extends the concept of GDT-HA score [48], GDC-SC score [53] and also quantifies the backbone side-chain orientations in the model and native structures. It consists of two distance based components which quantify the positioning of the CA and SC atoms in the model and native and three angle based components which quantify the orientations of backbone and side-chain in the model and native. By quantifying both the backbone and side-chain conformations of the model against the native, SPECS is highly accurate in determining the true native among a set of computationally predicted models.

For computing the CA component of SPECS, all the possible superpositions of two protein structures, each including the largest set of superimposable atoms are considered and for each possible superposition, distances between the aligned backbone Cα atoms are calculated. Then the distances are assigned to four different distance thresholds of 0.5, 1, 2 and 4 Å. Finally, from all the residue sets generated from different superpositions, the maximal residue set for each cutoff is selected, followed by averaging the proportion of residues in four different distance thresholds as shown below:

$$SPECS_{CA} = \frac{p_{dCA\_05} + p_{dCA\_1} + p_{dCA\_2} + p_{dCA\_4}}{4.0} \quad (3)$$

where p_{dCA_05}, p_{dCA_1}, p_{dCA_2} and p_{dCA_4} are the proportions of the maximal set of residues which belong to the 0.5, 1, 2 and 4 Å distance thresholds, respectively.

For computing the remaining four components of SPECS, we determine the optimal superposition between the backbone Cα atoms of the model and native. Now for the SC component of SPECS, the distances between the aligned side-chain reference atoms in model and native, r_{ij}, are calculated. Then, these distances, r_{ij}, are assigned to ten different bins ranging from 0.5 Å to 5.0 Å

with a step size of 0.5 Å, followed by averaging the proportion of residues in ten different bins as shown below:

$$SPECS_{SC} = \frac{2\ \sum_{i=1}^{10}(k-i+1)p_{rsc_i}}{k(k+1)} \quad (4)$$

where k = 10 is the number of bins and p_{rSCi} is the fraction of reference atoms assigned to bin i.

Now, we divide the $\theta_{ij}^{(1)}$ and $\theta_{ij}^{(2)}$ planar angles into four bins of $[0^{\circ}, 30^{\circ}]$, $(30^{\circ}, 60^{\circ}]$, $(60^{\circ}, 90^{\circ}]$ and $(90^{\circ}, 120^{\circ}]$, followed by averaging the proportion of residues in four different bins as shown below:

$$SPECS_{\theta^{(1)}} = \frac{2\ \sum_{i=1}^{4}(k-i+1)p_{\theta^{(1)}{}_i}}{k(k+1)} \quad (5)$$

where k = 4 is the number of bins and $p_{\theta}{}^{(1)}{}_i$ is the fraction of residues assigned to bin i.

$$SPECS_{\theta^{(2)}} = \frac{2\ \sum_{i=1}^{4}(k-i+1)p_{\theta^{(2)}{}_i}}{k(k+1)} \quad (6)$$

where k = 4 is the number of bins and $p_{\theta}{}^{(2)}{}_i$ is the fraction of residues assigned to bin i.

Then, we divide the Φ_{ij} dihedral angle into ten bins of $[0^{\circ}, 30^{\circ}]$, $(30^{\circ}, 60^{\circ}]$, $(60^{\circ}, 90^{\circ}]$, $(90^{\circ}, 120^{\circ}]$, $(120^{\circ}, 150^{\circ}]$, $(150^{\circ}, 180^{\circ}]$, $(180^{\circ}, 210^{\circ}]$, $(210^{\circ}, 240^{\circ}]$, $(240^{\circ}, 270^{\circ}]$ and $(270^{\circ}, 300^{\circ}]$, followed by averaging the proportion of residues in ten different bins as shown below:

$$SPECS_{\emptyset} = \frac{2\ \sum_{i=1}^{10}(k-i+1)p_{\emptyset_i}}{k(k+1)} \quad (7)$$

where k = 10 is the number of bins and $p_{\Phi i}$ is the fraction of residues assigned to bin i.

Finally, SPECS is calculated as a weighted average of the two distance based and three angle based components as shown below:

$$SPECS = \frac{4 * SPECS_{CA} + SPECS_{SC} + SPECS_{\theta^{(1)}} + SPECS_{\theta^{(2)}} + SPECS_{\emptyset}}{8.0} \quad (8)$$

4.4 RESULTS AND DISCUSSION

4.4.1 Comparison between SPECS and other model-native similarity scores on regular single domain targets

We compared SPECS on the regular single domain targets in CASP 12 [54] and 13, to determine its correlation with the existing backbone Cα based model-native similarity metrics. The CASP 12 decoy set consists of 55 single domain protein targets and the CASP 13 decoy set consists of 32 single domain protein targets. The targets were divided into template-based (TBM), free modeling (FM) and unresolved (TBM/FM) categories as defined by the assessors, to understand the density of the models in different categories. GDT-TS [2], TMScore [3] and Sphere Grinder (SPGR) score [55] were taken from the data archive of the Prediction Center (http://www.predictioncenter.org/), whereas SPECS was calculated as described in Materials and Methods. The plots displaying the relationship between SPECS and GDT-TS, TMScore, Sphere Grinder (SPGR) score are shown in Figure 4.3. From the average Pearson and Spearman correlation coefficients in the Figure 4.3, it is evident that there is a strong correlation between SPECS and superposition-based scores like GDT-TS, TMScore and local model accuracy scores like SPGR. The high Pearson's correlation coefficients show the existence of a linear relationship between SPECS and the popular backbone Cα based similarity measures. The high Spearman's correlation coefficients indicate the high level of agreement in ranking by SPECS and ranking by GDT-TS, TMScore and Sphere Grinder score. The high correlations with SPGR score show that, SPECS can also be used as reliable measure for analyzing the local accuracy of protein models [55].

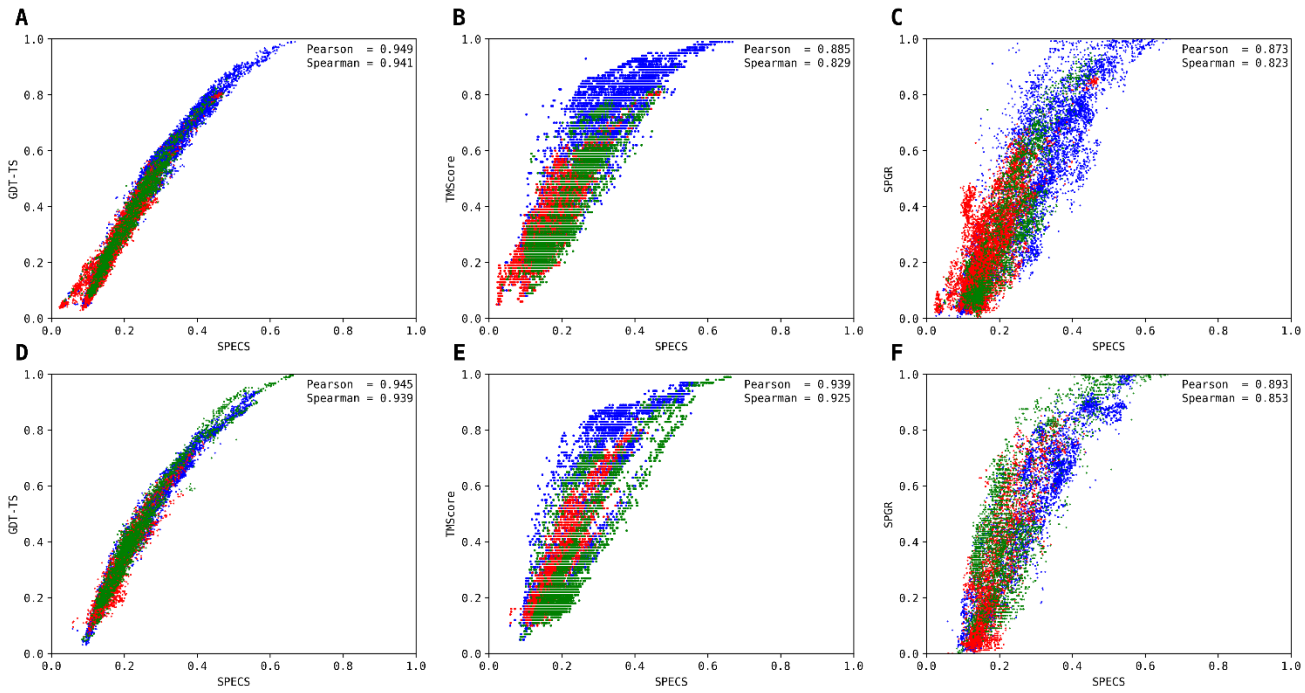

Figure 4.3. Comparisons between SPECS (horizontal axis) and existing model-native similarity metrics namely GDT-TS, TMScore and SPGR (vertical axis) using models in CASP 12 (A-C) and 13 (D-F) regular single domain targets. Average Pearson (P) and Spearman (S) correlation coefficients are shown for each plot. Blue, red, and green colors represent models assessed in template-based (TBM), free modeling (FM) and unresolved (TBM/FM) categories respectively. Higher color intensity reflects higher density of models.

4.4.2. Comparison between SPECS and other model-native similarity scores on high accuracy refinement targets

To further test the performance of SPECS on high accuracy models, we decided to compare SPECS with GDT-HA score [48], CAD-AA (all atoms) score [5], GDC-SC score [53] and LDDT score [4] on CASP 12 [54] and 13 refinement targets. On the whole in CASP 12 and 13, there are 37 refinement targets. GDT-HA score, CAD-AA score, GDC-SC score and LDDT score were taken from the data archive of the Prediction Center (http://www.predictioncenter.org/), whereas SPECS was calculated as described in Materials and Methods. The plots displaying the relationship between SPECS and GDT-HA, CAD-AA, GDC-SC, LDDT scores are shown in Figure 4.4. From the average Pearson and Spearman correlation coefficients in the Figure 4.4, it is evident that there

is a strong correlation between SPECS and superposition-based scores like GDT-HA, GDC-SC and superposition-free scores like CAD-AA and LDDT. From the results shown in Figure 4.4, it can be noted that SPECS is highly correlated with backbone Cα based similarity metric like GDT-HA, side-chain based similarity metric like GDC-SC, and all atom based similarity metrics like CAD-AA and LDDT.

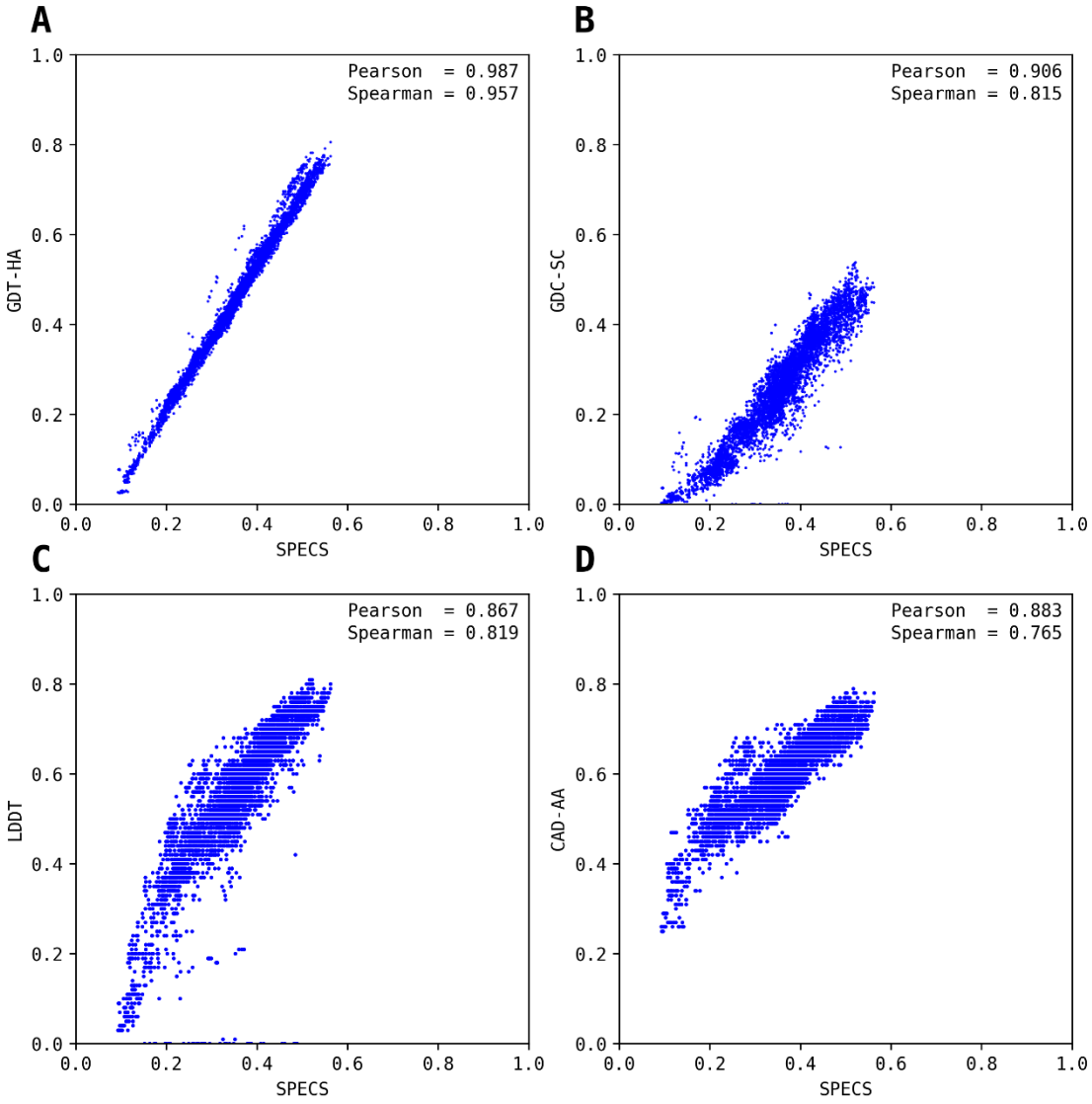

Figure 4.4. Comparisons between SPECS (horizontal axis) and existing model-native similarity metrics namely GDT-HA (A), GDC-SC (B), LDDT (C) and CAD-AA (D) (vertical axis) using models in CASP 12 and 13 refinement targets. Average Pearson (P) and Spearman (S) correlation coefficients are shown for each plot.

This shows the all-round ability of the SPECS score. The high Spearman's correlation coefficients indicate the high level of agreement in ranking between SPECS and all the other similarity scores. The all-round performance of SPECS on refinement targets and on regular single domain targets, shows that SPECS is a reliable measure for model-native similarity in case of both highly accurate refined predictions and low/moderately accurate predictions.

4.4.3. SPECS as a reliable Model Variation score

To test the effectiveness of SPECS in differentiating between models with correct and distorted stereochemical features, we benchmarked it on 3D Robot Decoy set [58]. The 3D Robot Decoy set consists of 200 protein targets, each consisting of 300 models. We further divided the 60000 protein models into three bins namely [0 – 2) Å, [2 – 6) Å and [6-12] Å based on their RMSD scores when compared against their natives. The division of the models into three bins, helped us analyze the effectiveness of SPECS as a Model Variation score on high accuracy, medium accuracy and low accuracy protein models. To understand the relationship between the SPECS score assigned to a model and its physical realism, we analyzed pairs of models for which SPECS and TMScore [3] were in conflict, for example, in a pair consisting of model_1 and model_2, SPECS assigned better score to model_1 and TMScore assigned better score to model_2. Now, in these conflicting pairs of models, we compared the consistency of the SPECS and TMScore with the physical realism of the models. We selected MolProbity score [57] to help us in assessing the variation of stereochemical features in the models. MolProbity consists of four components namely clash score, rotamer outlier score, Ramachandran outlier score and Ramachandran favored score to evaluate the correctness of protein structures. Like TMScore and GDT-HA score [48], MolProbity doesn't determine the true nativity of a protein model, instead it evaluates the structural quality of the model. From the pie charts in Figure 4.5, it is evident that there is a high level of

agreement in the ranking between SPECS and MolProbity scores. Hence, when compared to TMScore, SPECS is a reliable measure of variation in the models from all the three bins. We repeated this experiment, using SPECS and GDT-HA scores, to compare the performance of SPECS against a high accuracy and stringent metric like GDT-HA. From the pie charts in Figure 4.5, it can be observed that SPECS is a better measure of variation in the models from the high accuracy bin. Therefore, it is clear that a model's SPECS score and its physical realism are directly proportional to each other.

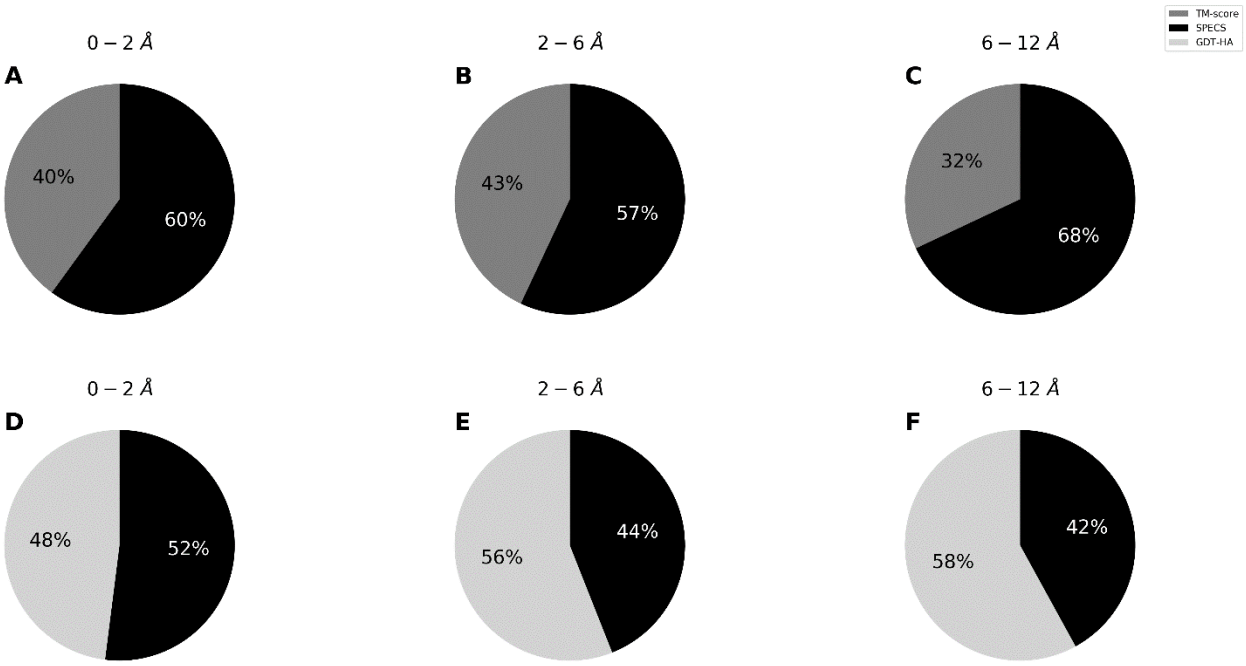

Figure 4.5. Pairs of 3DRobot models with conflicting ranking by SPECS and TMScore, SPECS and GDT-HA score. The 3DRobot models are divided to three bins based on their RMSD scores. Pie charts represent the MolProbity score agreement with rankings by SPECS and TMScore (A-C), SPECS and GDT-HA score (D-F). The ranges of the bins are indicated above each chart.

4.4.4. Evaluating Side-Chain Conformations using SPECS

To analyze the performance of SPECS as an evaluation measure for side-chain positioning, we considered the side-chain conformations of the monomeric proteins, predicted by three most widely used side-chain prediction methods. The monomeric protein dataset consists of 231 proteins and 33461 residues. The backbone Cα atoms of the proteins in the monomeric dataset are perfectly aligned with those in the natives, hence all the backbone Cα based similarity scores are 1. The side-chain conformations were predicted by RASP [62], Rosetta [63] and SCWR4L [64]. Then, the prediction accuracy of the three methods was evaluated in terms of the Angular RMSD of the χ1 side-chain torsion angle. The χ1 angle is the dihedral angle between the planes defined by the atoms N, Cα, Cβ, and Cγ [59]. Initially, the χ1 angle was calculated for each residue using the PDB module [65] of the Biopython package [66]. Then, the Angular RMSD values were calculated at the target level, from the corresponding χ1 angles [67]. The relationship between the ranking of side-chain prediction methods by Angular RMSD and the χ1 angle was determined using a boxplot as shown in Figure 4.6. From Figure 4.6, it is evident that the ranking of the three methods matches exactly with the ranking mentioned in Peterson L, Et al [59]. For evaluating the side-chain conformations predicted by the three methods using SPECS, we decided to determine its correlations with the Angular RMSD values of the methods. From Table 4.1, it can be observed that the SPECS score is moderately correlated with the angular RMSD values of the methods, when the backbone Cα based similarity scores are all perfect. Hence, it is clear that SPECS is a reliable measure for evaluating side-chain conformations.

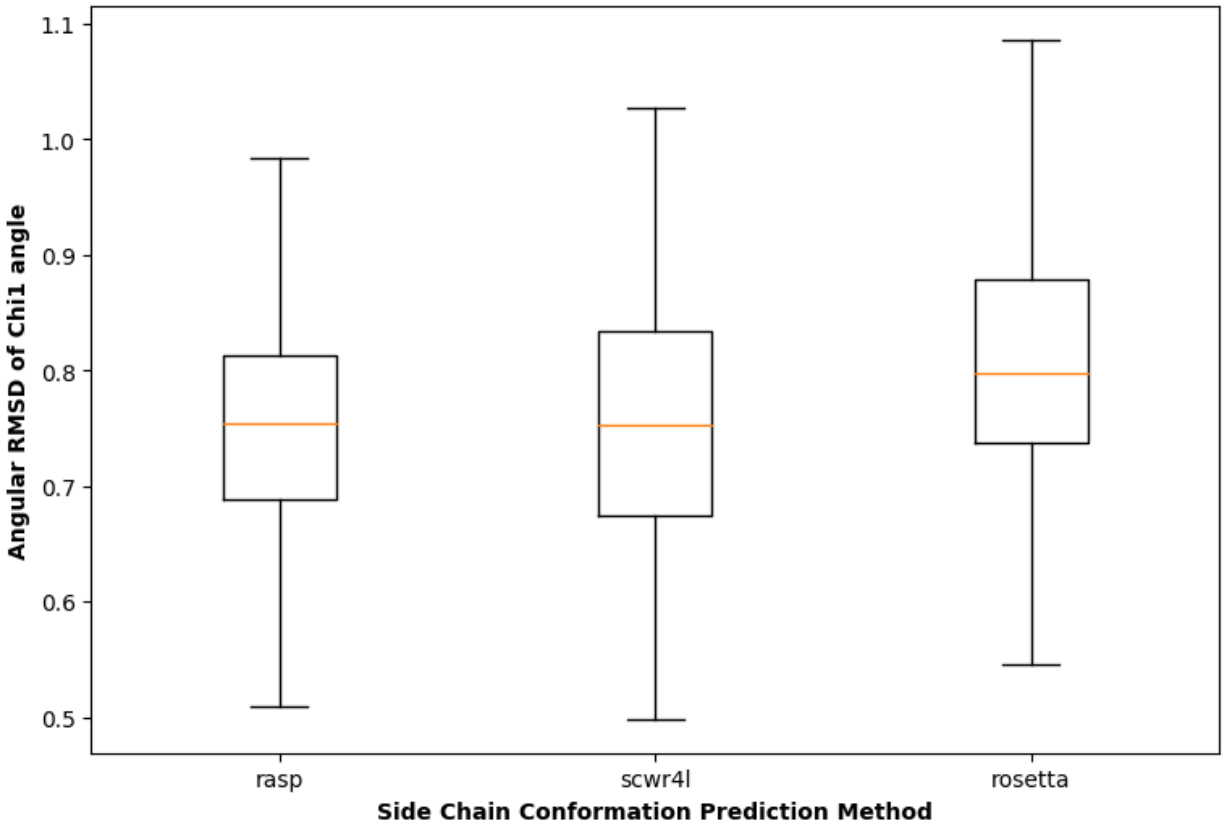

Figure 4.6. Prediction Accuracy by method using Angular RMSD. Lower and upper hinges: 1st and 3rd quartile. Whisker length: 1.5 times the interquartile range.

Table 4.1. Spearman Correlations between SPECS and the Angular RMSDs of side-chain conformation prediction methods.

Prediction Method	Spearman Correlation
RASP	0.4216
SCWR4L	0.3105
ROSETTA	0.3041

4.5. CONCLUSION

We develop a superposition-based model-native similarity score, which considers both the protein backbone and side-chain conformations for determining the true nativity of a model. SPECS is calculated as a weighted average of the two distance based components which quantify the positioning of the CA and SC atoms in the model and native and three angle based components which quantify the orientations of backbone and side-chain in the model and native. The experimental results suggest that SPECS is a reliable measure for model-native similarity in case of both highly accurate refined predictions and low/moderately accurate predictions. SPECS is highly correlated with superposition-based scores like GDT-HA, GDT-TS, GDC-SC and TMScore, superposition-free scores like CAD-AA and LDDT, and local model accuracy scores like SPGR. SPECS is also highly correlated with backbone Cα based similarity metric like GDT-HA, GDT-TS, TMScore, side-chain based similarity metric like GDC-SC, and all atom based similarity metrics like CAD-AA, LDDT and SPGR. In addition to being highly correlated with TMScore and GDT-HA, SPECS displays a stronger emphasis on the physical realism of models. Moreover, SPECS acts as a robust measure for evaluating side-chain conformations when the backbone Cα based similarity scores are all perfect. Collectively, these results demonstrate that SPECS is a valuable addition to protein structure comparison in particular and protein structure prediction in general.

REFERENCES

[1] "What are proteins and what do they do?" by U.S. National Library of Medicine is in the Public Domain.

[2] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic acids research*, vol. 31, no. 13, pp. 3370-3374, 2003.

[3] Y. Zhang, and J. Skolnick, "Scoring function for automated assessment of protein structure template quality", *Proteins*, pp. 702-710, 2004.

[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*, pp. 2722-2728, 2013.

[5] K. Olechnovic, E. Kulerkyte, and C. Venclovas, "CAD-score: A new contact area difference-based function for evaluation of protein structural models", *Proteins*, pp. 149-162, 2012.

[6] Arjun Ray, Erik Lindahl and Björn Wallner, "Improved model quality assessment using ProQ2", BMC Bioinformatics 2012, 13:224.

[7] R. Cao, B. Adhikari, D. Bhattacharya, M. Sun, J. Hou, and J. Cheng, "QAcon: single model quality assessment using protein structural and contact information with machine learning techniques", *Bioinformatics*, 2017.

[8] Alapati, R., Bhattacharya, D. "clustQ: Efficient Protein Decoy Clustering Using Superposition-free Weighted Internal Distance Comparisons", *In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018.

[9] Wang Q, Vantasin K, Xu D, Shang Y. "MUFOLD-WQA: A new selective consensus method for quality assessment in protein structure prediction", *Proteins*, 2011.

[10] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, and A.M.J.J. Bonvin, "Structure, Function, and Bioinformatics", *Proteins,* vol. 86, Issue Supplement S1, pp. 1-399, 2018.

[11] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede and A. Tramontano, "Structure, Function, and Bioinformatics", *Proteins,* vol. 84, Issue Supplement S1, pp. 1-391, 2016.

[12] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede and A. Tramontano, "Structure, Function, and Bioinformatics", *Proteins,* vol. 82, Issue Supplement S2, pp. 1-230, 2014.

[13] E. F. Blean, C. P. Stuart, and A. B. Boraston, "Structure, Function, and Bioinformatics", *Proteins,* vol. 79, Issue S10, pp. 1-207, 2011.

[14] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost and A. Tramontano, "Structure, Function, and Bioinformatics", *Proteins,* vol. 77, Issue S9, pp. 1-228, 2009.

[15] M. B. David, O. N. Brik, A. Paz, J. Prilusky, J. L. Sussman, and Y. Levy, "Assessment of CASP 8 structure predictions for template free targets", *Proteins,* pp. 50-65, 2009.

[16] A. Kryshtafovych, and B. Monastyrskyy, "Assessment of model accuracy estimations in CASP12", *Wiley,* 2018.

[17] A. Elofsson, K. Joo, C. Keasar, and J. Lee, "Methods for estimation of model accuracy in CASP12", *Wiley,* 2018.

[18] L.J. McGuffin, "The ModFOLD server for the quality assessment of protein structural models", *Bioinformatics,* 2008.

[19] Z. Wang, J. Eickholt, and J. Cheng, "APOLLO: a quality assessment service for single and multiple protein models", *Bioinformatics,* 2011.

[20] A. Herbert, and M. J. E. Sternberg, "MaxCluster: a tool for protein structure comparison and clustering", 2008.

[21] R. Cao, D. Bhattacharya, B. Adhikari, J. Li, and J. Cheng, "Large-scale model quality assessment for improving protein tertiary structure prediction", *Bioinformatics,* 2015.

[22] S. C. Li, and Y. K. Ng, "Calibur: a tool for clustering large numbers of protein decoys", *BMC Bioinformatics,* 2010.

[23] J. Zhang and Y. Zhang, "A Distance-Dependent Atomic Potential Derived from Random-Walk Ideal Chain Reference State for Protein Fold Selection and Structure Prediction", *PLoS One,* vol. 5, 2010.

[24] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. M. E. Strauss, and D. Baker, "Rosetta in CASP4: progress in ab-initio protein structure prediction", *Proteins,* vol. 45, pp. 119–126, 2001.

[25] B. John, and A. Sali, "Comparative protein structure modeling by iterative alignment, model building and model assessment", *Nucleic Acids Research,* 2003.

[26] L. J. McGuffin, and D. B. Roche, "Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments", *Bioinformatics,* 2010.

[27] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano, "Assessment of model accuracy estimations in CASP12", *Proteins,* 2017.

[28] M. J. Skwark, and A. Elofsson, "PconsD: ultra-rapid, accurate model quality assessment for protein structure prediction", *Bioinformatics,* 2013.

[29] N. Siew, A. Elofsson, and L. Rychlewski, "MaxSub: an automated measure for the assessment of protein structure prediction quality," *Bioinformatics,* vol. 16, no. 9, pp. 776-785, 2000.

[30] Y. Zhang, "I-TASSER server for protein 3D structure prediction", *BMC Bioinformatics,* pp. 9- 40, 2008.

[31] D. Xu and Y. Zhang, “Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field”, *Proteins*, pp. 1715-1735, 2012.

[32] R. Jauch, H. C. Yeo, P. R. Kolatkar, and N. D. Clarke, “Assessment of CASP7 structure predictions for template free targets”, *Proteins*, 2007.

[33] J. Moult, T. Hubbard, and K. Fidelis, “Critical assessment of methods of protein structure prediction (CASP): round III”, *Wiley*, 1999.

[34] J. Moult, K. Fidelis, and A. Zemla, “Critical assessment of methods of protein structure prediction (CASP): round IV”, *Wiley*, 2001.

[35] J. Moult, K. Fidelis, and A. Zemla, “Critical assessment of methods of protein structure prediction (CASP): round V”, *Wiley*, 2003.

[36] J. Moult, K. Fidelis, B. Rots, and T. Hubbard, “Critical assessment of methods of protein structure prediction (CASP): round VI”, *Wiley*, 2005.

[37] J. Moult, K. Fidelis, and A. Kryshtafovych, “Critical assessment of methods of protein structure prediction: Round VII”, *Wiley*, 2007.

[38] J. Moult, K. Fidelis, and A. Kryshtafovych, “Critical assessment of methods of protein structure prediction: Round VIII”, *Wiley*, 2009.

[39] A. Kryshtafovych, A. Barbato, K. Fidelis, B. Monastyrskyy, T. Schwede and A. Tramontano, “Assessment of the assessment: Evaluation of the model quality estimates in CASP10”, *Proteins*, 2014.

[40] R. Cao, D. Bhattacharya, J. Hou, and J. Cheng. “DeepQA: improving the estimation of single protein model quality with deep belief networks”, *BMC Bioinformatics*, 2016.

[41] R. Cao, D. Bhattacharya, B. Adhikari, J. Li, and J. Cheng. “Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11”, *Proteins: Structure, Function, and Bioinformatics,* 2015.

[42] D. Bhattacharya, and J. Cheng, “i3Drefine software for protein 3D structure refinement and its assessment in CASP10”, PLOS ONE, 2013.

[43] B. Adhikari, X. Deng, J. Li, D. Bhattacharya, and J. Cheng, “A Contact-Assisted Approach to Protein Structure Prediction and Its Assessment in CASP10”, *AAAI Workshop,* 2013.

[44] Zhang Y, “Protein structure prediction: when is it useful?”, *Current Opinion in Structural Biology* 2009;19(2):145–155.

[45] Holm L, Sander, “Protein structure comparison by alignment of distance matrices”, *Journal of Molecular Biology* 1993;233(1):123–138.

[46] Kufareva I, Abagyan R, “Methods of Protein Structure Comparison”, In: Orry AJW, Abagyan R, editors. *Homology Modeling.* Volume 857. Totowa, NJ: Humana Press; 2011. p 231–257.

[47] Koehl P, “Protein structure similarities”, *Current Opinion in Structural Biology* 2001;11(3):348–353.

[48] Kopp J, Bordoli L, Battey JND, Kiefer F, Schwede T, “Assessment of CASP7 predictions for template-based modeling targets”, *Proteins: Structure, Function, and Bioinformatics* 2007;69(S8):38–56.

[49] W Kabsch, “A solution for the best rotation to relate two sets of vectors”, *Acta Crystallograp Sec A* 1976;32:922–923.

[50] J Moult, T Hubbard, SH Bryant, K Fidelis, JT Pedersen, “Critical assessment of methods of protein structure prediction (CASP): round II”, *Proteins 1997*;Suppl 1:2–6.

[51] J Moult, JT Pedersen, R Judson, K Fidelis, “A Large-Scale Experiment to Assess Protein Structure Prediction Methods”, *Proteins* 1995;23:ii–v.

[52] Miao Z, Cao Y, “Quantifying side-chain conformational variations in protein structure”, *Scientific Reports* 2016;6(1).

[53] MacCallum JL, Hua L, Schnieders MJ, Pande VS, Jacobson MP, Dill KA, “Assessment of the protein-structure refinement category in CASP8”, *Proteins: Structure, Function, and Bioinformatics* 2009;77(S9):66–80.

[54] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A, “Critical assessment of methods of protein structure prediction (CASP)-Round XII”, P*roteins: Structure, Function, and Bioinformatics* 2018;86:7–15.

[55] Kryshtafovych A, Monastyrskyy B, Fidelis K, “CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL: CASP Prediction Center”, *Proteins: Structure, Function, and Bioinformatics* 2014;82:7–13.

[56] Sadreyev RI, Shi S, Baker D, Grishin NV, “Structure similarity measure with penalty for close non-equivalent residues”, *Bioinformatics* 2009;25(10):1259–1263.

[57] Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC, “MolProbity : all-atom structure validation for macromolecular crystallography”, *Acta Crystallographica Section D Biological Crystallography* 2010;66(1):12–21.

[58] Deng H, Jia Y, Zhang Y, “3DRobot: automated generation of diverse and well-packed protein structure decoys”, *Bioinformatics* 2016;32(3):378–387.

[59] Peterson LX, Kang X, Kihara D, “Assessment of protein side-chain conformation prediction methods in different residue environments: Side-Chain Conformation Prediction Accuracy”, *Proteins: Structure, Function, and Bioinformatics* 2014;82(9):1971–1984.

[60] Liwo A, O?dziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA, “A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data”, *Journal of Computational Chemistry* 1997;18(7):849–873.

[61] Bhattacharya D, Cao R, Cheng J, “UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling”, *Bioinformatics* 2016;32(18):2791–2799.

[62] Miao Z, Cao Y, Jiang T, “RASP: rapid modeling of protein side chain conformations”, *Bioinformatics* 2011;27(22):3117–3122.

[63] Kuhlman B, Baker D, “Native protein sequences are close to optimal for their structures”, *Proceedings of the National Academy of Sciences* 2000;97(19):10383–10388.

[64] Krivov GG, Shapovalov MV, Dunbrack RL, “Improved prediction of protein side-chain conformations with SCWRL4”, *Proteins: Structure, Function, and Bioinformatics* 2009;77(4):778–795.

[65] T Hamelryck, B Manderick, “PDB file parser and structure class implemented in Python”, *Bioinformatics* 2003;19(17):2308–2310.

[66] 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.

[67] Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T, “A generative, probabilistic model of local protein structure”, *Proceedings of the National Academy of Sciences* 2008;105(26):8932–8937.

Appendix A:

Software and Web-services

In this chapter, we provide a brief overview of the freely available software and web-servers developed based on the aforementioned methods for the scientific community. In particular, we developed a freely available web-server named clustQ based on the methods presented in Chapter 2. Finally, a publicly accessible web-server named SPECS based on the methods presented in Chapter 4 is made available.

A.1. clustQ

A.1.1. Overview

clustQ is a webservice for rapid protein decoy clustering. It is a consensus based QA method, that employs a WQ-score based multi-model pairwise comparison approach for model quality assessment. The goal of clustQ is to rank and select models in a time efficient manner, using pairwise comparison between all the models in a pool.

A.1.2. Availability

http://watson.cse.eng.auburn.edu/clustQ/

A.1.3. Input

The input to clustQ server must be a target protein sequence, a decoy tarball and a job name of user's choice. When the user provides the target protein sequence and decoy tarball for estimating the model accuracy estimation, the server validates the user entries. If successful, the job is queued.

Otherwise the user is a shown an error message. The user needs to ensure that the decoy tarball is either a zip or a tar.gz file.

A.1.4. Output

clustQ server automatically redirects the user about the status of the current submission. After the job is completed, the decoys in the tarball will be displayed in the descending order of their quality scores assigned by clustQ. The user can download a text file consisting of the decoy rankings and their corresponding scores. The user can bookmark the results page to view the results later. In case the user has provided an email address, the decoy ranking along with the quality scores assigned by clustQ will be emailed immediately after the job is complete.

A.1.5. Software Architecture

clustQ is developed in C++. Source code, executable and example data of clustQ for Linux are freely available to non-commercial users.

A.2. SPECS

A.2.1. Overview

SPECS is a webservice for protein structure comparison. It is a superposition based model-native similarity measure, that integrates the high accuracy version of the Global Distance Test (GDT-HA) metric, and side-chain distance and orientation in a singular framework. The goal of SPECS is to compare two protein structures by taking into consideration their main chain Cα atoms, SC atoms along with their orientation and assign a similarity score in the range of 0 to 1, higher the better.

A.2.2. Availability

http://watson.cse.eng.auburn.edu/SPECS/

A.2.3. Input

The input to SPECS server must be two protein structural files in PDB format. When the user provides the model and native files for structure comparison, the server validates the user entries. If successful, the job is queued. Otherwise the user is a shown an error message. The user needs to ensure that the protein structural files are in PDB format.

A.2.4. Output

SPECS server automatically redirects the user about the status of the current submission. After the job is completed, the results of the structural comparison consisting of SPECS score, TMScore, MaxSub score, GDT-TS and GDT-HA scores will be displayed. The user can bookmark the results page to view the results later.

A.2.5. Software Architecture

SPECS is developed in C++. Source code, executable and example data of SPECS for Linux are freely available to non-commercial users.