

Contents

List of Figures	ii
List of Tables	iii
List of Equations	iv
List of Abbreviations	v
1 Introduction	1
2 Materials & Methods	2
3 Residue contacts predicted by evolutionary covariance extend the application of <i>ab initio</i> molecular replacement to larger and more challenging protein folds	3
4 Evaluation of ROSETTA distance-restraint energy functions on contact-guided <i>ab initio</i> structure prediction	4
5 Alternative <i>ab initio</i> structure prediction algorithms for AMPLE	5
6 Decoy subselection using contact information to enhance MR search model creation	6
6.1 Introduction	7
6.2 Materials & Methods	8
6.2.1 Target selection	8
6.2.2 Computation of range-specific satisfaction scores	8
6.2.3 Decoy subselection	8
6.2.4 Molecular Replacement	9
6.3 Results	10
6.3.1 Contact pair satisfaction correlates with decoy quality	10
6.3.2 Long-range contact satisfaction metric to filter decoy sets	13
6.3.3 AMPLE's cluster-and-truncate approach with filtered decoy sets . .	16
6.3.4 MR search models by processing single decoys	20
6.3.5 Decoy subselection extends AMPLE's performance	24
6.4 Discussion	27
7 Protein fragments as search models in Molecular Replacement	28
8 Conclusion	29
A Appendix	30
Bibliography	31

List of Figures

6.1	Linear regression model between decoy TM-score and contact satisfaction .	12
6.2	Top-1 decoy TM-score and contact satisfaction analysis	13
6.3	TM-score comparison pre- and post-decoy subselection	15
6.4	Effect of decoy subselection on SPICKER clusters	18
6.5	Effect of decoy subselection on THESEUS variance	19
6.6	Selection of single decoys by long-range satisfaction	21
6.7	Difference in RMSD for individually processed decoys	22
6.8	Relationship between decoy quality and fraction of residues retained	23
6.9	Molecular Replacement summary of decoy-subselected AMPLE ensembles .	25

List of Tables

6.1	Correlation analysis between decoy TM-score and contact satisfaction . . .	11
-----	----------------------------------------------------------------------------	----

List of Equations

List of Abbreviations

CC	Correlation Coefficient
KDE	Kernel Density Estimate
MR	Molecular Replacement
PDB	Protein Data Bank
RMSD	Root-Mean-Square Deviation
TM-score	Template-Modelling score

Chapter 1

Introduction

Chapter 2

Materials & Methods

Chapter 3

Residue contacts predicted by evolutionary covariance extend the application of *ab initio* molecular replacement to larger and more challenging protein folds

Chapter 4

Evaluation of ROSETTA

distance-restraint energy functions

on contact-guided *ab initio*

structure prediction

Chapter 5

Alternative *ab initio* structure prediction algorithms for AMPLE

Chapter 6

Decoy subselection using contact information to enhance MR search model creation

6.1 Introduction

Work presented in previous chapters highlighted the much improved *ab initio* decoy quality achievable by restraining the conformational search space with residue-residue contact information. Furthermore, the data also highlighted that this improvement extends AM-
PLE’s tractability of achieving structure solution for more challenging targets. However, the data also indicated that AM-
PLE’s current protocol is not tailored towards decoy sets with overall much higher accuracy. Decoy sets with correctly predicted folds (average Template-Modelling score (TM-score) > 0.5 per 1,000 decoys) did not generate any ensemble search models, which enabled Molecular Replacement (MR) structure solution. It also became apparent that certain decoy sets contained some very high-quality decoys, nevertheless, such decoys are often lost in the process of clustering due to the lack of other similar predictions.

Furthermore, *ab initio* decoy similarity in exceptional cases approaches a near-identical fold (Root-Mean-Square Deviation (RMSD) $< 1.5\text{\AA}$) to the crystallised one. Although challenging by current means to identify these decoys, it is of great interest to structural biologists to do so since these decoys might be sufficient by themselves as MR search models. Contact information, which is typically used to restrain the folding protocol might provide enough information to drive such filtering. Indeed, Kosciolk and Jones [1] and De Oliveira et al. [2] highlighted the usefulness of long-range residue-residue contact pair satisfaction, which correlates well with decoy quality. Additionally, Adhikari and Cheng [3] use long-range contact satisfaction routinely in CONFOLD2 to exclude the worst decoys amongst the set predicted ones.

Thus, this chapter focuses on exploring alternative strategies of decoy selection in AM-
PLE. In particular, work presented here focuses on exploiting long-range contact information to drive search model generation to extend AM-
PLE’s target tractability further.

6.2 Materials & Methods

6.2.1 Target selection

The dataset for this study consisted of 113 ROSETTA decoy sets generated throughout the works outlined in previous chapters. The 113 decoy sets covered all targets in the ORIGINAL (??), PREDICTORS (??) and TRANSMEMBRANE (??) datasets. Top- L (> 5 residues sequence separation) CCMPRED [4], PCONSC2 [5], METAPSICOV STAGE 1 [6] and MEMBRAN [7] contact pairs were used in combination with the *FADE* energy function to restrain the *ab initio* structure prediction process.

6.2.2 Computation of range-specific satisfaction scores

The satisfaction of short- (> 5 residues sequence separation), medium- (> 12 residues sequence separation) and long-range contact pairs (> 23 residues sequence separation; see ??) were computed for each decoy in each set. Hereby, the contact pairs of the original set of contact pairs used to restrain the *ab initio* structure prediction protocol were extracted, matched against the contact pairs extracted from individual decoys and the contact pair range-specific satisfaction score evaluated.

6.2.3 Decoy subselection

Each set of decoys was ranked in descending order by their long-range contact pair satisfaction scores and the n decoys with the lowest scores removed from each set. The number of decoys to remove n was selected using a number of different strategies:

<i>NONE</i>	leave the original set unchanged
<i>LINEAR</i>	remove the worst 500 decoys
<i>CUTOFF</i>	remove all decoys with a score of < 0.287
<i>SCALED</i>	remove all decoys with a scaled score of < 0.5 , where the scaled score is score divided by set average

INDIVIDUAL keep the top-5 decoys only

The fixed definition in the *CUTOFF* strategy was determined by De Oliveira et al. [2]. The scaled score used by the *SCALED* strategy was computed by dividing each decoy’s long-range contact pair satisfaction by the set’s average.

The *INDIVIDUAL* subselection strategy differs substantially from the others. The top-5 decoys by long-range contact satisfaction were selected and subjected to treatment outside of AMPLE. The per-decoy treatments were the following:

default	leave the decoy unchanged
domain	remove all residues with $kde < \frac{1}{2}max_{kde}$, where kde corresponds to the Kernel Density Estimate (KDE) and max_{kde} to the maximum KDE obtained by applying the algorithm described by Sadowski [8] to the top-5 L contact map
dssp	remove all residues with DSSP-assigned secondary structure [9] not in $\{H, B, E, G, I\}$
fragment	remove all residues that do not form a consecutive fragment of at least 3 residues in the contact map extracted from the decoy, where a contact between two atoms was existent if its C β (C α in case of Gly) were within 8Å
variance	remove all residues with cluster variance computed by THESEUS [10] of more than 5Å (cluster variance determined from <i>NONE</i> strategy)

6.2.4 Molecular Replacement

To evaluate the benefits of such subselection to MR in AMPLE, a subset of 35 decoy sets (spanning 35 unique targets) were processed as described above and subjected to AMPLE v1.2.0 and CCP4 v7.0.28. Default options were chosen with few exceptions: decoys in all 10 clusters were used, subcluster radii thresholds were set to 1 and 3Å, and

side-chain treatments were set to `polyala` only. This change in protocol was shown to be advantageous in most cases by Thomas [11], and thus trialled in this context.

To allow comparability of these results to previous AMPLE runs, an additional condition was added, namely *NONE_classic*. The decoy set from the *NONE* strategy was hereby subjected to the AMPLE protocol with default settings except `-num_clusters`, which was set to sample the three largest clusters.

All individual decoys created under *INDIVIDUAL* strategy were subjected as poly-Alanine decoys to MRBUMP [12] with identical settings to those used in AMPLE.

Each MR run was assessed using the criteria defined in ??.

6.3 Results

This chapter focuses on identifying further uses of predicted residue-residue contact pairs in unconventional MR. In particular, the exclusion of *ab initio* decoys by their contact satisfaction scores is under investigation. A total of 113 decoy datasets were used to identify potential means of identifying the best or worst decoys. Furthermore, three strategies were trialled alongside two standards to test the viability of excluding the worst decoys in ensemble search model preparation in AMPLE.

6.3.1 Contact pair satisfaction correlates with decoy quality

Kosciolek and Jones [1] previously identified a correlation between the TM-score of a decoy and its fraction of satisfied contact pairs. Albeit the striking positive correlations (short-range: $\rho = 0.50$; medium-range: $\rho = 0.57$; long-range: $\rho = 0.87$) for top-1 decoys, the study by Kosciolek and Jones [1] was limited to 10 representative targets with a maximum chain length of 158 residues. Furthermore, FRAGFOLD [13] was used for *ab initio* protein structure prediction, a method with inferior performance to ROSETTA [14] when using the decoys in unconventional MR (see Chapter 5). Thus, the more diverse set of decoys generated in this study might be more representative in determining a correlation between decoy TM-scores and contact pair satisfaction.

A correlation analysis with 113 ROSETTA decoy sets representing 56 globular and transmembrane targets shows a positive linear correlations between a decoy's TM-score and short-, medium- and long-range contact satisfaction (Table 6.1). Furthermore, separating the correlation analysis of all targets by fold classification reveals that all- α , mixed α - β and transmembrane protein targets show the strongest positive correlations for long-range contact satisfaction (Table 6.1). All- β and mixed α - β decoy sets show the strongest correlations for short- and medium-range contact satisfaction, whereby the former shows a stronger positive correlation between the decoy's TM-score and its medium-range contact satisfaction than its long-range contact satisfaction (medium-range: $\rho = 0.54$; long-range: $\rho = 0.50$) (Table 6.1). Notably, the decoys of transmembrane protein targets show no correlation between TM-score and short-range contact satisfaction ($\rho = 0.08$; Table 6.1).

Table 6.1: Pearson's Correlation Coefficient (CC) analysis between a ROSETTA decoy's TM-score and short-, medium- and long-range contact satisfaction. Probability values for all ρ coefficients is < 0.01 .

Target class	Pearson's CC		
	Short-range	Medium-range	Long-range
all	0.11	0.18	0.64
all- α	0.30	0.44	0.69
all- β	0.40	0.54	0.50
mixed α - β	0.42	0.55	0.69
transmembrane	0.08	0.48	0.70

Following on from the correlation analysis, a linear regression model was fitted to individual subsets of the data used for the correlation analysis to see if a decoy's TM-score could be predicted from its contact satisfaction score. However, weak coefficients of determination indicate that only some cases show models with reasonably good fits to the data (Fig. 6.1). Nevertheless, all models further support the positive linear correlations between a decoy's TM-score and its range-dependent contact satisfaction. Interestingly, the strongest and best fits of the linear regression model to its corresponding data is for long-range contact pairs, where the linear regression models are also near identical between the different fold categories (Fig. 6.1).

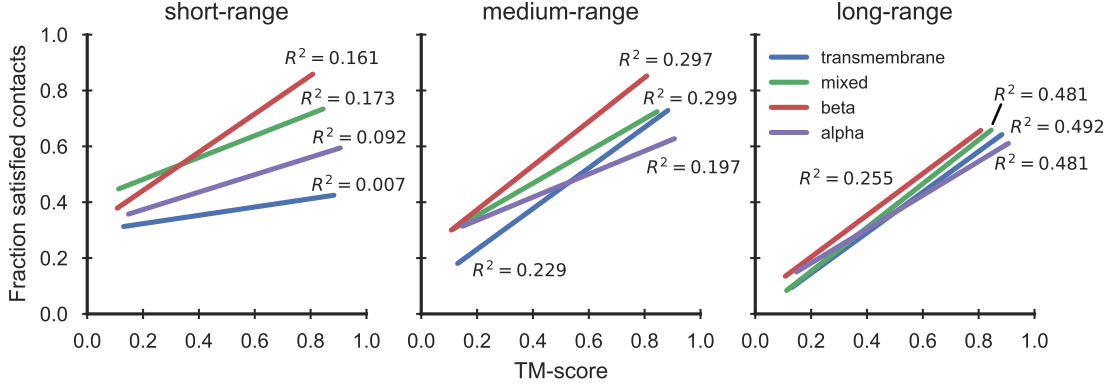


Figure 6.1: Linear regression model fitted to decoy TM-scores and corresponding fractions of satisfied, range-dependent contacts. Targets were further separated by fold classification. Coefficients of determination (R^2 -values) added alongside each regression model.

An analysis of the correlation between the TM-score and long-range contact satisfaction of individual decoy sets further highlights the potential to subselect decoy sets by their long-range contact satisfaction. One hundred and Eight decoy sets show statistically significant positive correlations between decoy TM-scores and their long-range contact satisfaction (ρ -values in range of 0.09 to 0.97 with p -value < 0.01). A singleton ROSETTA decoy set, derived for the Glycolipid transfer protein with Protein Data Bank (PDB) ID 2eum and restrained with METAPSICOV STAGE 1 contact data, shows a weak negative correlation ($\rho = -0.10$, $p < 0.01$). The remaining four decoy sets, derived for targets with PDB IDs 1chd, 1gm4, 2x6u and 3ouf and restrained with METAPSICOV STAGE 1 contact data except for 2x6u (PCONSC2), show no statistically significant correlation between the TM-score and long-range contact satisfaction of the decoy sets.

A further subdivide of the previously presented data by metapredictor highlights that no predictor outperforms the others. Decoy sets from all metapredictors result in decoy sets with stronger and weaker correlations. Similarly, target chain length and fold do not show overall stronger or weaker correlations.

So far, all analyses focused on entire sets of decoys (1,000 decoys per set); however, it is often desirable to know if we could better estimate the accuracy of the best decoy by some measure. Kosciolk and Jones [1] demonstrated strong positive correlations for short-, medium- and long-range contact satisfaction with a decoy's corresponding TM-score. In this work, these findings are confirmed albeit the strength of the correlation for

long-range contact satisfaction is much weaker than previously observed (short-range: no correlation; medium-range: $\rho = 0.52$; long-range: $\rho = 0.69$) (Fig. 6.2). The weak positive correlation for short-range contact satisfaction is statistically non-significant, and thus cannot be validated.

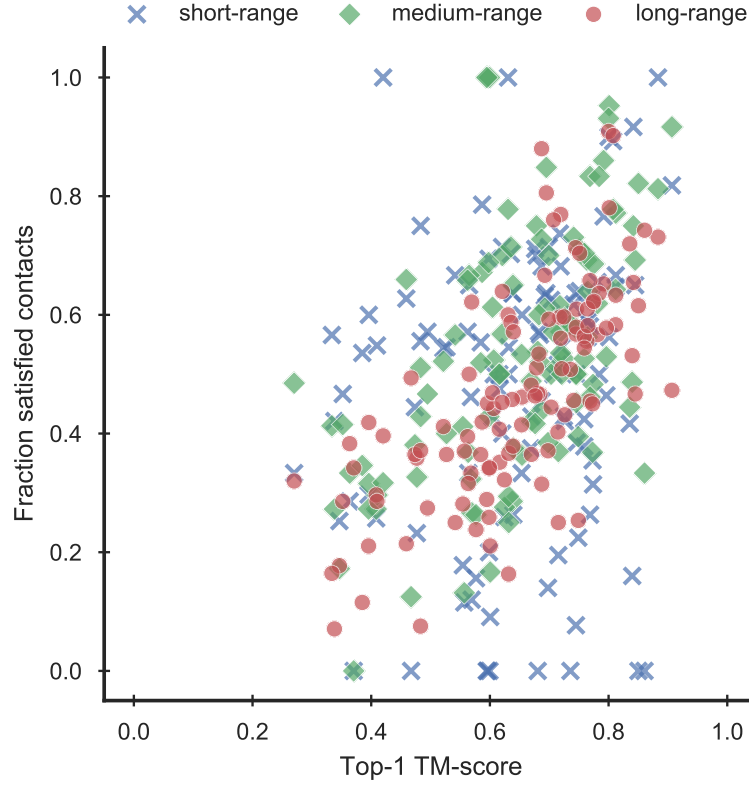


Figure 6.2: Analysis of the relationship between TM-score and contact satisfaction for the top-1 decoy (as ranked by TM-score) in each decoy set.

6.3.2 Long-range contact satisfaction metric to filter decoy sets

In the previous section, the data highlighted that decoy quality correlates positively with contact satisfaction. In particular, a strong positive correlation between long-range contact satisfaction and decoy quality could be established for almost all decoy sets in this study. A key ambition in this work is to determine if this correlation could be used to include or exclude decoys prior to submission to the AMPLE processing pipeline to enrich the quality and number of ensemble search models trialled in MR.

The difference in mean TM-score of each decoy set before and after applying a sub-selection strategy (see Section 6.2.3) is shown in Fig. 6.3. Estimating a decoy's quality

by short-range contact satisfaction results in marginal mean TM-score changes of decoy sets ($\Delta_{CUTOFF} = -0.003$; $\Delta_{LINEAR} = 0.008$; $\Delta_{CUTOFF} = 0.001$). In comparison, medium- ($\Delta_{CUTOFF} = 0.005$; $\Delta_{LINEAR} = 0.015$; $\Delta_{CUTOFF} = 0.002$) and long-range ($\Delta_{CUTOFF} = 0.025$; $\Delta_{LINEAR} = 0.032$; $\Delta_{CUTOFF} = 0.005$) contact satisfaction are better estimates to improve the mean TM-score values of each decoy set. Notably, per-decoy long-range contact satisfaction provides the best estimate for identifying and excluding the least accurate decoys independent of the subselection strategy.

A comparison of Δ TM-scores between subselection strategies shows that the *LINEAR* strategy results in the largest improvements across all contact range categories. Nevertheless, such great changes are expected compared to other strategies since the *LINEAR* strategy removes on average the most decoys from each set (*CUTOFF*=280; *LINEAR*=500; *SCALED*=66). However, the sample-dependent strategies (*CUTOFF* and *SCALED*) may remove a much greater number of decoys from a set if the corresponding satisfaction scores fall below a certain threshold.

In certain cases, some subselection strategies greatly changed the overall size and quality of the starting decoy set. The METAPSICOV STAGE 1 decoy set of the ankyrin sequence (PDB ID: 2qyj) shows overall quality improvements from 0.006 (short-range *SCALED*; $n_{models} = 958$) to 0.213 (long-range *CUTOFF*; $n_{models} = 218$). The CCM-PRED decoy set of sensory rhodopsin II sequence (PDB ID: 1gu8) shows overall changes from -0.155 (short-range *CUTOFF*; $n_{models} = 2$) to 0.06 (long-range *LINEAR*; $n_{models} = 500$).

Overall, the optimal strategy to select or exclude decoys from a starting set of structures appears to be long-range contact satisfaction given the improved similarity metric of the set compared to the reference crystal structure.

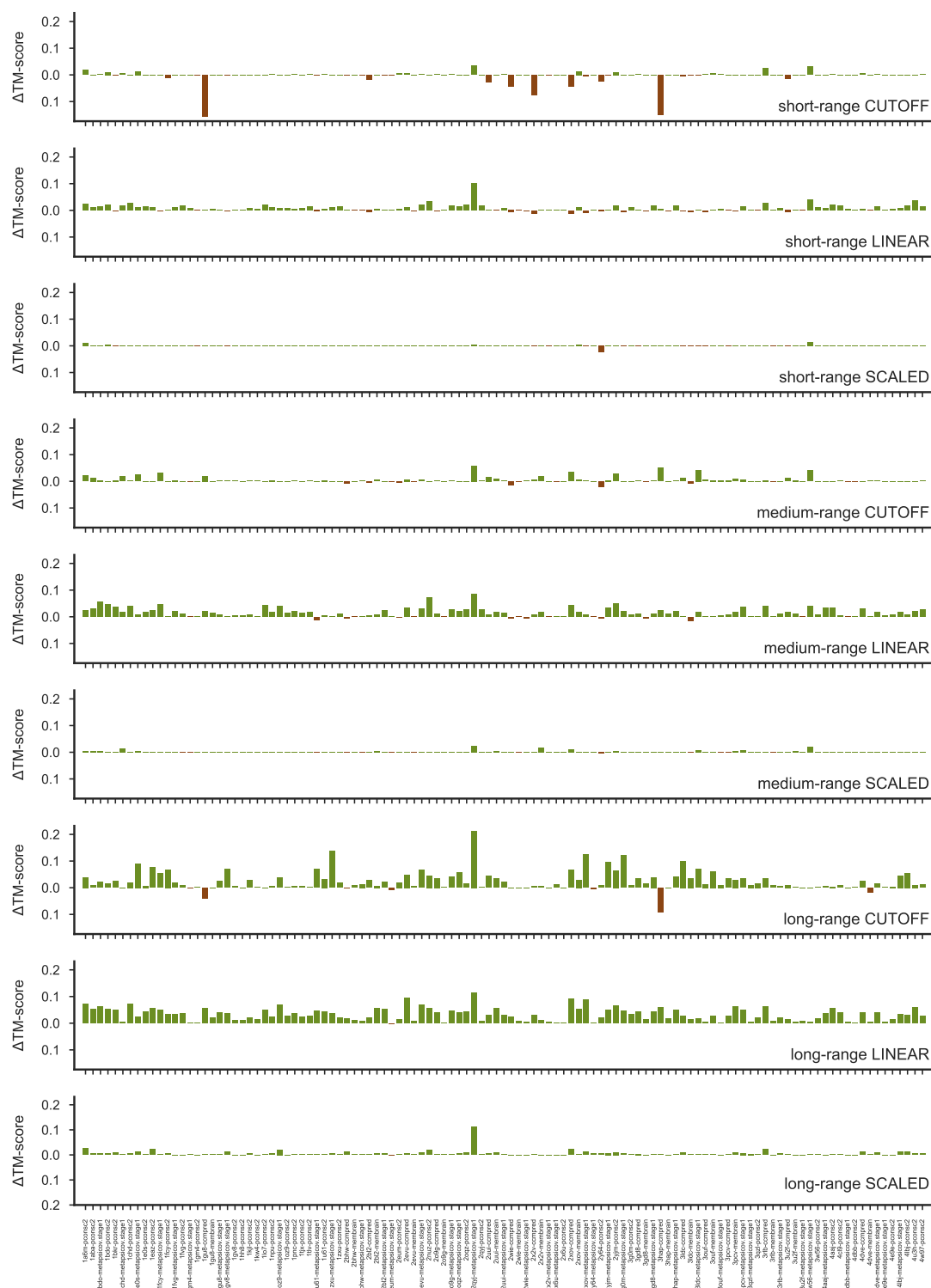


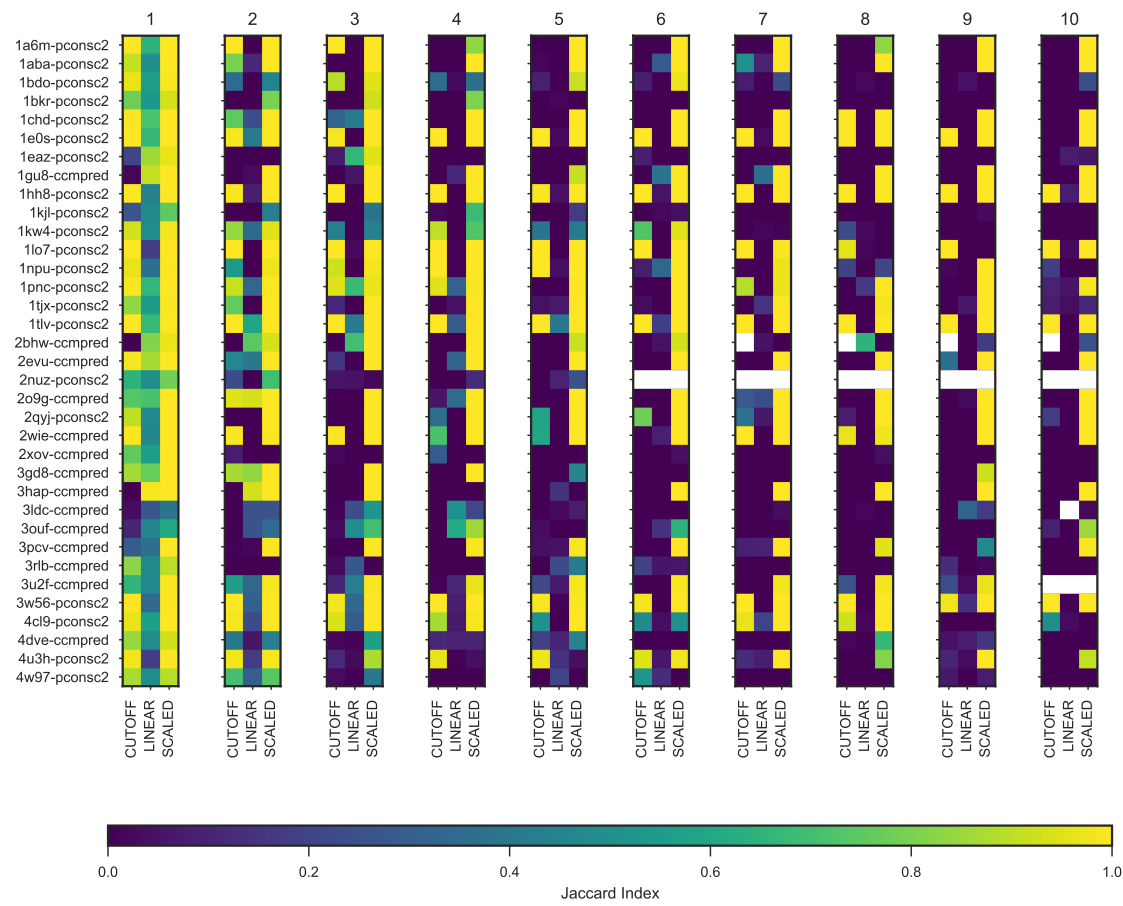
Figure 6.3: Differences in mean TM-score for decoy sets pre- and post-decoy subselection. Each subselection strategy is stated in each subplot along with the contact range used to establish decoy inclusion in the final set.

6.3.3 AMPLE’s cluster-and-truncate approach with filtered decoy sets

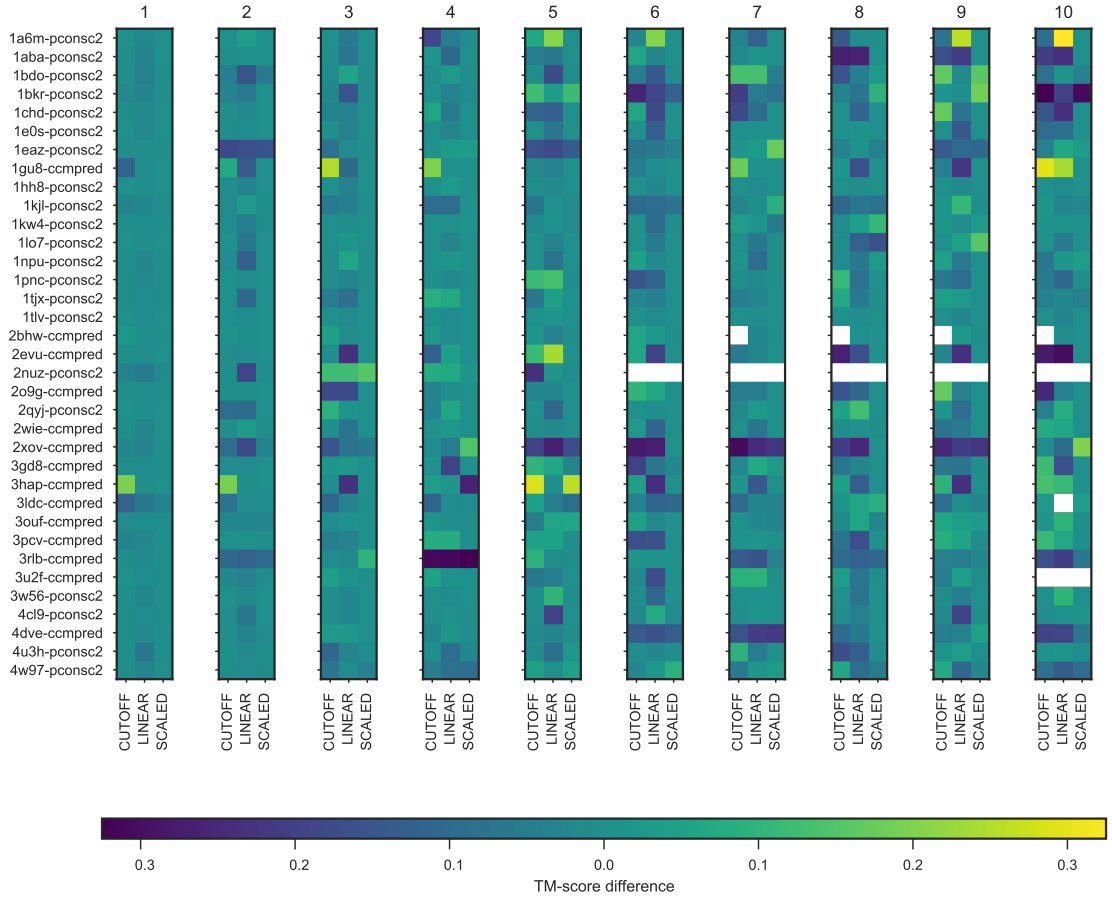
The original dataset was limited to a sub-sample of 35 decoy sets spanning 35 unique targets (21 globular and 14 transmembrane targets) for MR trials. The contact prediction algorithms generating the restraints for the *ab initio* structure predictions were PCONSC2 (globular targets) and CCMPRED (transmembrane targets). Each decoy set was subjected to the AMPLE pipeline with certain decoys removed according to one of five subselection strategies.

The initial step in the AMPLE pipeline is the clustering of decoys. A comparison of SPICKER clusters between the *NONE* default strategy and the *CUTOFF*, *LINEAR* and *SCALED* subselection strategies highlights an important difference. Larger clusters — those ranked higher — show higher similarity between a subselection strategy and the default (Fig. 6.4a). The top SPICKER cluster shows high similarities between the *NONE* strategy and all other subselection ones, whereby it has to be noted that the *LINEAR* strategy contains only 50% of the starting decoys and thus can at most show a Jaccard index of 0.5. With increasing cluster index, the overall similarity degrades and most of the decoys in cluster 10 are non-identical between each subselection strategy and the default.

Furthermore, a similar analysis to compare the overall quality of each cluster compared to the target structure revealed less difference between the default and each subselection strategy for higher-ranked SPICKER clusters (Fig. 6.4b). With decreasing SPICKER cluster index, the difference in median TM-scores starts to alternate without any particular pattern. Thus, pre-selecting decoys prior to AMPLE’s cluster-and-truncate approach most certainly preserves the top cluster for the *CUTOFF* and *SCALED* subselection strategies, whereby lower clusters show more deviation from the default.



(a)



(b)

Figure 6.4: Effect of decoy subselection on SPICKER clusters. Effect illustrated by (a) the Jaccard Index and (b) median TM-score difference. Values were calculated for clusters resulting from the full starting set of decoys and the *CUTOFF*, *LINEAR* and *SCALED* subselection strategies.

The mean of the inter-decoy variance computed by THESEUS — used in AMPLE to guide truncation of each cluster — is reduced in lower clusters compared to the *NONE* default strategy (Fig. 6.5). The clusters of decoys based on the galectin-3 domain (PDB ID: 1kjl) sequence show overall the highest reduction in mean inter-decoy variance up to -15Å compared to the default strategy. Similarly, clusters 4 and 8 of the K⁺-channel protein domain (PDB ID: 3ouf) show reductions in mean inter-decoy variance of up to -20Å. In general and also true for the aforementioned examples, clusters starting from *CUTOFF*-subselected decoys show the greatest mean inter-decoy variance reductions, followed by *LINEAR* and then *SCALED*-subselected decoys sets.

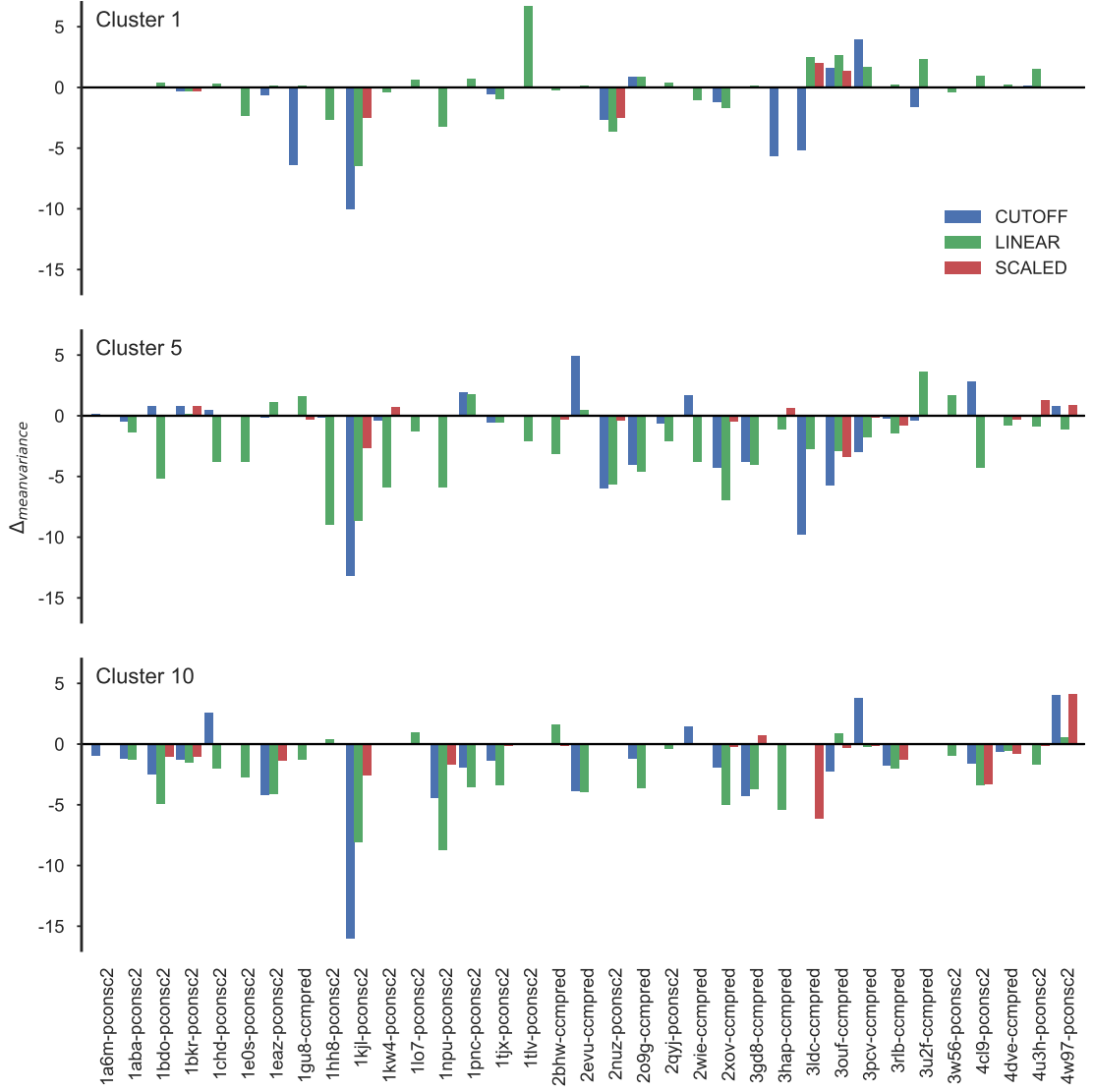


Figure 6.5: Effect of decoy subselection on mean inter-decoy THESEUS variance. Difference in mean variance calculated between the default and the three decoy subselection strategies *CUTOFF*, *LINEAR* and *SCALED*. Data for clusters 1, 5 and 10 shown as examples.

A comparisons of intermediate stages in the AMPLE pipeline resulting from differently subselected decoy sets is very difficult. Each strategy results in different starting sets, which result in different clusters. Since AMPLE’s objective truncation procedure is based on the inter-decoy variance, it might be greatly affected by differing clusters. Nevertheless, structure solution is more likely when AMPLE generates more ensemble search models because a greater number of search models relates to greater inter-cluster decoy similarity and trialling a greater number should provide a higher chance of success. A count of generated AMPLE ensemble search models reveals that the *SCALED* strategy generates

the most search models ($n = 7,611$), which is roughly 300 more than the default *NONE* strategy ($n = 7,340$). The *CUTOFF* subselection strategy generates the least ensemble search models ($n = 7,237$), whilst the *LINEAR* strategy’s count ($n = 7,401$) is very similar to the *NONE* one.

Further inspection of the number of AMPLE-generated ensemble search models by target reveal near identical numbers between the *NONE*, *LINEAR* and *SCALED* strategies (Fig. 6.9). In fact, only few outliers for each of those methods distinguish them from the others. The *CUTOFF* strategy shows greater deviation from the other three, especially for certain targets with differences up to approximately 200 ensemble search models (Fig. 6.9). If we compare all these strategies to the previous default processing in AMPLE (*NONE_classic*; further details in Section 6.2.4), we can see that the number of search models is greatly reduced (Fig. 6.9). A comparison of the previous default (*NONE_classic*) with the new one (*NONE*) shows on average 144 fewer ensemble search models per target, whilst sampling a larger range of folds through all 10 clusters.

6.3.4 MR search models by processing single decoys

This study attempts to identify single decoys of sufficient quality to be used directly as MR search model in addition to the decoy set subselection. The top-5 decoys, as judged by long-range contact satisfaction, were selected from each decoy set. Four distinct processing approaches were applied to each decoy and compared against the unmodified initial decoy.

The correlation between a decoy’s long-range contact satisfaction and its TM-score have previously been outlined and is further confirmed here (Fig. 6.6). However, the positive correlation was dependent on the target’s fold class and the overall accuracy of the decoy set. Here, an analysis of the top-5 decoys by long-range contact satisfaction in each decoy set shows that 50% of selected decoys fall in the 80th percentile or greater of TM-scores in each decoy set whilst 90% are in the at least the 40th percentile (Fig. 6.6). Although the latter is less striking, it highlights the current limits of selecting individual decoys by this approach.

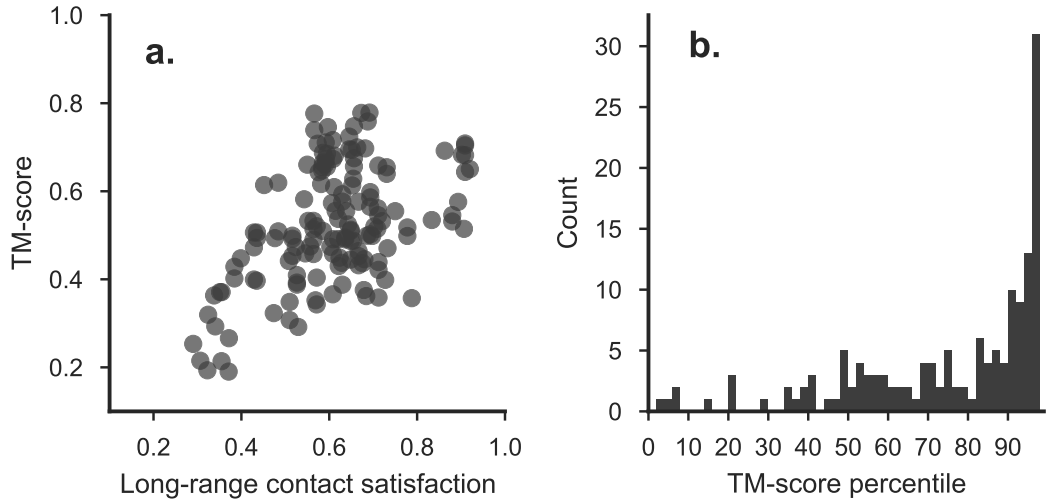


Figure 6.6: Analysis of long-range contact-satisfaction-based decoys with respect to the relationship between (a) the satisfaction and decoy quality and (b) the quality compared to the remaining, excluded decoy set.

A comparison of RMSD value changes indicated that the “fragment” and “variance” metrics provide the best approximation to identifying less-reliable regions in each decoy. The average RMSD change compared to the original decoys is just under 4.0\AA . This compares to a slightly lower RMSD change of 2.1\AA for “DSSP”-treated decoys and 1.5\AA for the “domain” treatment. Although almost all decoys were improved by either of the treatments, a small number of decoys worsen in terms of RMSD compared to its native structure. All treatments except the “fragment” one had worsened decoys in the final set, with changes up to -1.6\AA .

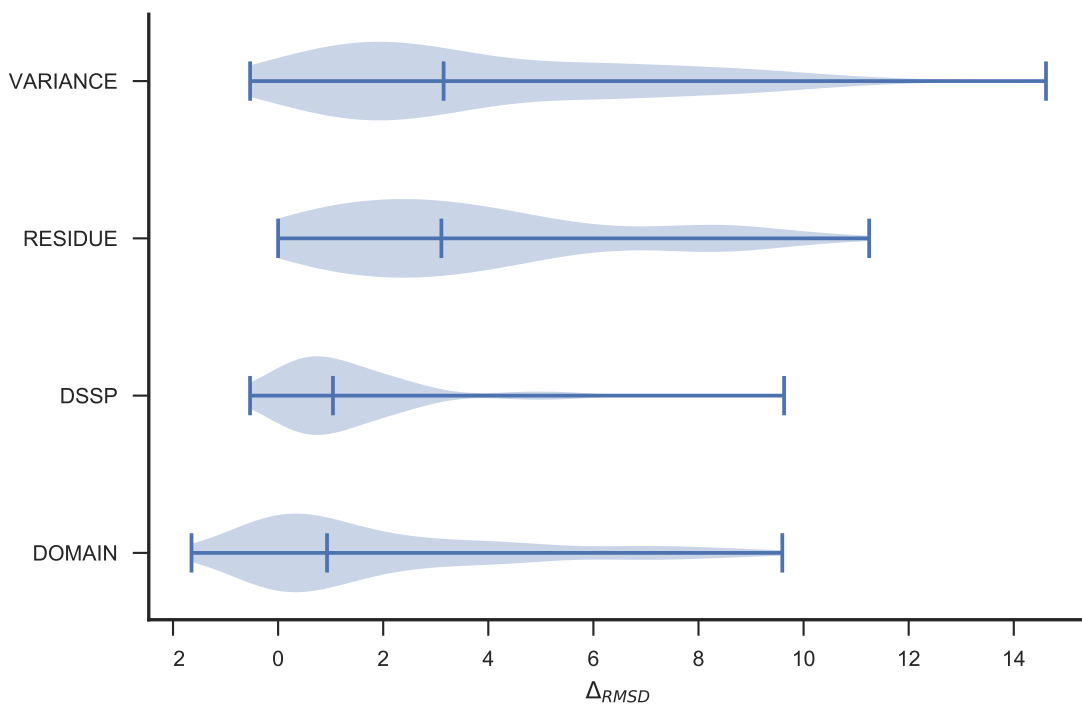


Figure 6.7: Distribution of differences in RMSD values between the initial and modified decoys under four different treatments.

A comparison of the range of RMSD values revealed much greater changes for the “fragment” and “variance” conditions (Fig. 6.7). However, these changes are not reflected in the fraction of residues retained in each decoy. Most residues were removed by the “domain” treatment ($\mu=61.6\%$), whilst the “fragment” one saw the least removal ($\mu=34.5\%$). Similarly to the range in RMSD values, the “fragment” and “variance” treatments resulted in the greatest spread of fraction of residues in the treated decoy. The values range for both treatments from retaining less than 5% of the initial decoy up to 100%.

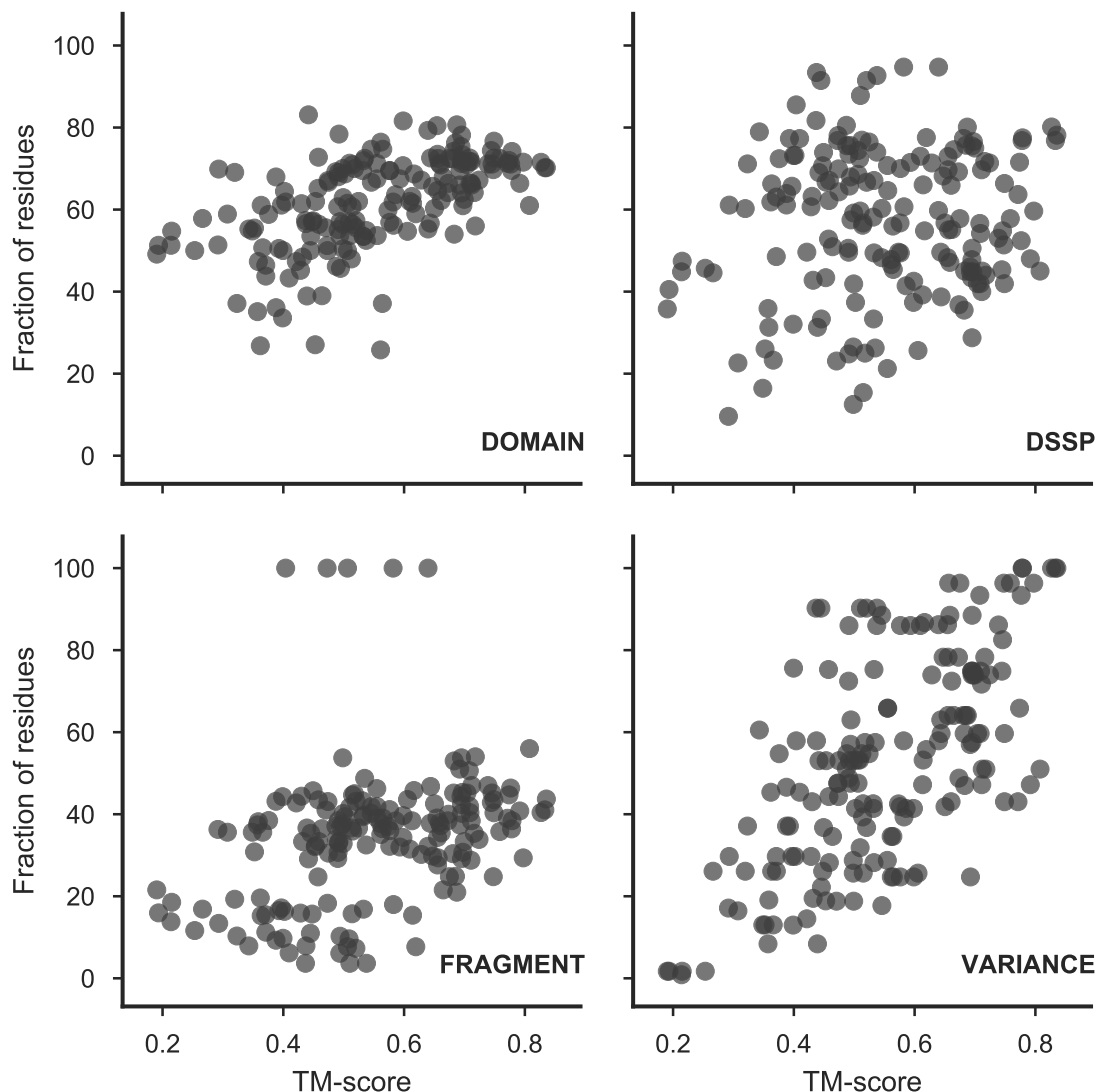


Figure 6.8: Scatter plots of initial decoy TM-score and the fraction of residues retained after one of four different residue removal treatments.

A further aspect of the decoy treatments highlights that the fraction of residues retained after decoy post-processing correlates with the cluster variance of the decoy (Fig. 6.8). Such correlation cannot be determined for any other processing step and explains at least in part why we see much greater changes in RMSD value between the initial and processed decoy for the “variance” treatment compared to the others. However, if a decoy is of particularly poor quality (TM-score < 0.3), the “variance” treatment retains as little as 0.87% and 1.7% of the initial decoy (2 and 4 residues) whilst the others retain a much larger fraction of at least 40% for equivalent decoys.

6.3.5 Decoy subselection extends AMPLE's performance

The final step in this study is the assessment of AMPLE-generate ensemble search models and single-decoy-based search models in MR. In particular, the comparison of different decoy subselection strategies and individual decoy-processing treatments is of great interest since it might allow us to extend AMPLE's performance beyond that described in previous chapters.

A comparison of the total number of targets solved by each subselection strategy shows that the *CUTOFF*-subselected decoys lead to most structure solutions (14 out of 35) (Fig. 6.9). Although slightly less successful, the *LINEAR* and *SCALED* subselection strategies lead to approximately 6% more solved targets than the *NONE* strategy. The *LINEAR* and *SCALED* strategies are on par with AMPLE's default, the *NONE_classic* strategy (Fig. 6.9). Although the *NONE_classic* strategy generates two version of each ensemble search model with poly-Alanine and all-atom side chain treatment, the former was enough to solve all targets outlined in (Fig. 6.9). Therefore, the *LINEAR* and *SCALED* subselection strategies would be the minimum processing requirement to maintain previous performance in AMPLE.

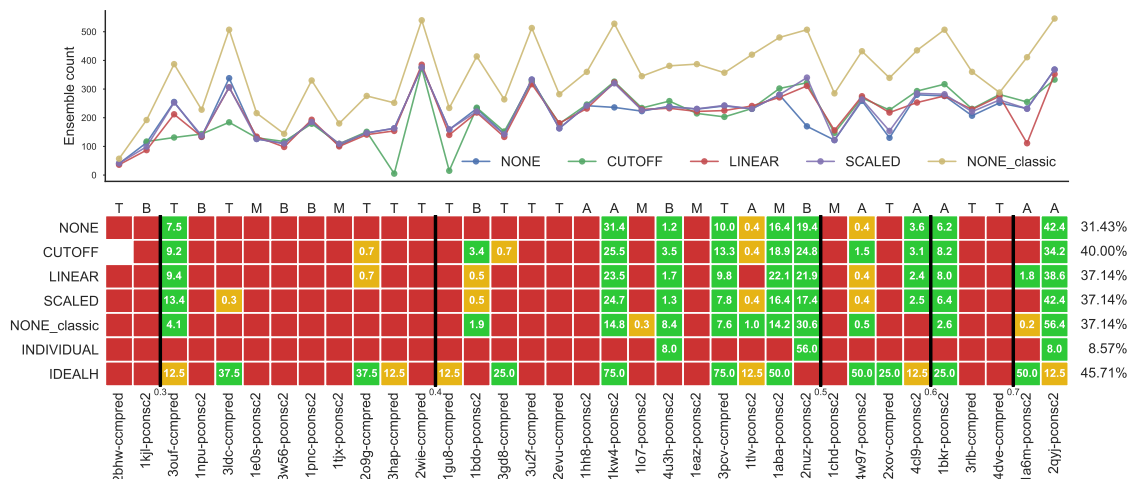


Figure 6.9: Molecular Replacement summary of decoy-subselected AMPLE ensembles. AMPLE-generated ensemble counts illustrated at the top with Molecular Replacement results in grid below: red cell equates to no solution; orange to a singleton solution; and green to multiple solutions. All *INDIVIDUAL* attempts were compressed to a single row per decoy set. The number in the orange and green cells indicates the percentage of ensemble search models leading to structure solutions. One letter code above each column indicates the target fold: “T” for transmembrane; “A” for all- α ; “B” for all- β ; “M” for mixed α - β). Percentage alongside each row indicates the number of targets with solutions by each strategy. Targets are sorted from left to right with increasing median TM-score of the starting decoy set. The black lines highlight TM-score thresholds from 0.3 to 0.7 from left to right. The subselection strategy *IDEALH* refers AMPLE’s ideal helix library.

The *CUTOFF* method yields the highest number of structure solutions based on AMPLE-generated ensemble search models whilst generating the fewest search models. In fact, this subselection strategy has generated no ensemble search models for target 2bhw. Furthermore, the *CUTOFF* method achieves amongst the best ratio of search models leading to structure solution compared to the total number generated.

In few exceptions, only a singleton AMPLE search model lead to a structure solution (orange cells in Fig. 6.9). Upon closer inspection, 71% of all singleton solutions were achieved with AMPLE ensemble search models containing at least 30% of the target sequence. Twenty-nine percent of the singleton solutions contain at least 50% of the target sequence, whilst none contain more than 70%. Three out of four search models with less than 30% of the target sequence were derived from the PCONSC2 decoy set predicted for the ketosteroid transcriptional regulator KstR2 (PDB ID: 4w97) sequence and contained one, two or three small helical fragments.

The rank order of targets by median TM-score of the initial starting decoy set in

Fig. 6.9 shows that no decoy set with median TM-score of less than 0.3 score units lead to structure solution; however, only two such cases exist in the dataset, and therefore this threshold may only serve as indication. With increasing median TM-score, i.e. increasing similarity between the decoy set and its reference target structure, the chances appear to increase to achieve structure solution. Beyond a threshold of 0.4 TM-score units, structure solutions are much more likely (over 50% of targets solved with one of the four subselection strategies), which highlights AMPLE's success in processing such accurate decoy sets appropriately.

The work in this study further explored, if individual decoys could be selected via their long-range contact satisfaction and trialled directly as MR search models. The *INDIVIDUAL* subselection strategy explored this aspect with numerous post-selection processing steps. However, structure solutions for only three targets could be obtained using this approach (Fig. 6.9). All processing steps obtained lead to structure solutions based on the PCONSC2 decoy set of the α -spectrin SH3 domain (PDB ID: 2nuz). The other two targets with solutions, PDB IDs 2qyj and 4u3h with PCONSC2 decoy sets, acquired at least one structure solutions with a single decoy and processing with the "domain", "DSSP", "fragment" and "variance" treatments. Across the three targets, only five decoys (three based on the sequence of PDB ID 2nuz) with a minimum TM-score of 0.682 resulted in the 20 structure solutions (PDB ID 2nuz: 16 solutions; PDB ID 2qyj: 2 solutions; PDB ID 4u3h: 2 solutions).

A comparison of decoy-derived search models and AMPLE's simplistic ideal helix library [15] in MR was done. Ideal helices achieved the most structure solutions solving nearly 46% of all targets (Fig. 6.9). In particular, ideal helices achieved structure solutions for more transmembrane targets. Eight out of 14 transmembrane targets were solved with at least one ideal helix, which compares to six out of 14 for all decoy-based search models combined. Ideal helices also managed to near identical results for all- α and mixed α - β targets in the set compared to decoy-derived search models. However, three structure solutions for all- β targets are intractable.

6.4 Discussion

The subselection of decoy sets by long-range contact satisfaction is a concept originally proposed by Kosciulek and Jones [1] and later confirmed and extended by De Oliveira et al. [2] and Adhikari and Cheng [3]. In this study, these findings are further confirmed by re-analysing all decoy sets previously presented in this thesis.

In this chapter, further work was conducted to explore potential benefits of applying long-range contact satisfaction to subselect decoy sets with the aim to extend AMPLE’s target tractability. In part, such extension is possible. In particular, AMPLE’s default mode has changed since the start of this study and equates in today’s version to the *NONE* subselection strategy. Given the results presented here, contact-based decoy subselection can extend the performance whilst also reducing the number of generated search models, such as in the case of the *CUTOFF* subselection strategy. The advantage of such approach would be two-fold: a typical AMPLE run would require less computation resources since the *CUTOFF* strategy rejects decoys which satisfy the contact information poorly; and structure solution is possible whilst sampling less ensemble search models in general. However, subselecting decoy sets by contact satisfaction poses a real risk of biasing all attempts to a particular fold, especially if the contact prediction is less accurate.

Beyond subselecting decoys sets, some very preliminary work in this chapter aimed to explore the possibility of identifying, processing and trialling individual *ab initio* structure predictions as MR search models. Findings supported previous challenges in the field of identifying the very best decoys reliably by long-range contact satisfaction. Although a general correlation exists for most decoy sets, the best by long-range satisfaction are not necessarily always the very best by TM-score. Thus, this approach has proven less successful in this work than AMPLE-generated ensembles. Further work is required to explore better decoy selection and processing since the latter’s attempts were relatively simplistic and better sequence- and structure-based metrics might be available. Additionally, single decoys might be further treated with CONCOORD — in an identical manner to Rigden et al. [16] — to generate AMPLE-compatible decoys.

Chapter 7

Protein fragments as search models in Molecular Replacement

Chapter 8

Conclusion

Appendix A

Appendix

Bibliography

- [1] T. Kosciolok, D. T. Jones, en, *PLoS One* **Mar. 2014**, *9*, e92197.
- [2] S. H. P. De Oliveira, J. Shi, C. M. Deane, en, *Bioinformatics* **Feb. 2017**, *33*, 373–381.
- [3] B. Adhikari, J. Cheng, en, *BMC Bioinformatics* **Jan. 2018**, *19*, 22.
- [4] S. Seemayer, M. Gruber, J. Söding, en, *Bioinformatics* **Nov. 2014**, *30*, 3128–3130.
- [5] M. J. Skwark, D. Raimondi, M. Michel, A. Elofsson, en, *PLoS Comput. Biol.* **Nov. 2014**, *10*, e1003889.
- [6] D. T. Jones, T. Singh, T. Kosciolok, S. Tetchner, en, *Bioinformatics* **Apr. 2015**, *31*, 999–1006.
- [7] J. Yang, R. Jang, Y. Zhang, H. B. Shen, en, *Bioinformatics* **Oct. 2013**, *29*, 2579–2587.
- [8] M. I. Sadowski, en, *Proteins: Struct. Funct. Bioinf.* **Feb. 2013**, *81*, 253–260.
- [9] D Frishman, P Argos, en, *Proteins* **Dec. 1995**, *23*, 566–579.
- [10] D. L. Theobald, D. S. Wuttke, en, *Bioinformatics* **Sept. 2006**, *22*, 2171–2172.
- [11] J. M. H. Thomas, PhD thesis, University of Liverpool, **Jan. 2017**.
- [12] R. M. Keegan, S. J. McNicholas, J. M. H. Thomas, A. J. Simpkin, F. Simkovic, V. Uski, C. C. Ballard, M. D. Winn, K. S. Wilson, D. J. Rigden, *Acta Crystallographica Section D Structural Biology* **Mar. 2018**, *74*, 167–182.
- [13] D. T. Jones, en, *Proteins: Structure Function and Genetics* **2001**, *Suppl 5*, 127–132.
- [14] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, en, *Methods Enzymol.* **2004**, *383*, 66–93.
- [15] J. M. H. Thomas, R. M. Keegan, J. Bibby, M. D. Winn, O. Mayans, D. J. Rigden, en, *IUCrJ* **Mar. 2015**, *2*, 198–206.
- [16] D. J. Rigden, J. M. H. Thomas, F. Simkovic, A. Simpkin, M. D. Winn, O. Mayans, R. M. Keegan, *Acta Crystallographica Section D Structural Biology* **Mar. 2018**, *74*, 183–193.