UNIVERSITY OF
LIVERPOOL

Covariation-derived residue contacts in *ab initio* modelling and Molecular Replacement

Felix Simkovic

Thesis submitted in accordance with the requirements of the
University of Liverpool
for the degree of
Doctor in Philosophy

September 2018

Institute of Integrative Biology
University of Liverpool
United Kingdom

Contents

List of Figures **iv**

List of Tables **v**

List of Equations **vi**

List of Abbreviations **vii**

1 Introduction **1**

2 Materials & Methods **3**

3 Evolutionary covariance in *ab initio* structure prediction-based Molecular Replacement **5**

3.1 Introduction 6
3.2 Materials & Methods 6
  3.2.1 Target selection 6
  3.2.2 Contact prediction 6
  3.2.3 Contact-to-restraint conversion 7
  3.2.4 *Ab initio* structure prediction 8
  3.2.5 Molecular Replacement in AMPLE 8
3.3 Results 8
  3.3.1 Residue-residue contact prediction 9
  3.3.2 Protein structure prediction 12
  3.3.3 Molecular Replacement 17
3.4 Discussion 23

4 Evaluation of ROSETTA distance-restraint energy functions on contact-guided *ab initio* structure prediction **27**

5 Alternative *ab initio* structure prediction algorithms for AMPLE **29**

6 Decoy subselection using contact information to enhance MR search model creation **31**

7 Protein fragments as search models in Molecular Replacement **33**

8 Conclusion & Outlook **35**

A Appendix **37**

Bibliography **39**

List of Figures

3.1 Alignment depth and contact precision analysis of globular and transmembrane protein targets . . . 10
3.2 Evaluation of BBCONTACTS contact pairs . . . 12
3.3 Effect of contact distance restraints on *ab initio* decoy quality . . . 13
3.4 TM-score comparison for globular targets separated by fold . . . 14
3.5 Decoy analysis of effects of BBCONTACTS contact addition . . . 15
3.6 TM-score difference between contact-assisted and simple ROSETTA decoys 17
3.7 Structure solution summary for globular targets . . . 18
3.8 Structural superposition of three search models for target 1lo7 . . . 19
3.9 Top-PHASER solutions for target 1e0s . . . 20
3.10 Effect of progressive truncation on RMSD of ensemble centroid . . . 21
3.11 Summary of AMPLE truncation ranges for structure solution . . . 22

List of Tables

3.1 Summary of raw conact prediction precision values in PCONSC2 11

List of Equations

List of Abbreviations

CC	Correlation Coefficient
LLG	Log-Likelihood Gain
MR	Molecular Replacement
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank
RIO	Residue-Independent Overlap
RMSD	Root-Mean-Square Deviation
TFZ	Translation Function Z-score
TM-score	Template-Modelling score

Chapter 1

Introduction

Chapter 2

Materials & Methods

Chapter 3

Evolutionary covariance in *ab initio* structure prediction-based Molecular Replacement

Note: *The majority of the work presented in this chapter was published in two independent pieces of work. All work relating to the globular targets was published by Simkovic et al. [1], and a great majority of work relating to the transmembrane targets by Thomas et al. [2]. As such, this chapter consists of extracts from both publications with additional information where appropriate. Text duplicated from either publication was written by Felix Simkovic, all other elements were adapted.*

3.1 Introduction

The introduction of residue-residue contacts as distance restraints in *ab initio* protein structure prediction has proven to be a highly successful approach to limiting the conformation search space thereby enabling successful fold predictions of larger and more β-rich protein structures [e.g., 3–11]. In AMPLE, such proteins have historically proven the most difficult targets [12]. Furthermore, the initial AMPLE study by Bibby et al. [12] focused solely on globular targets, whilst Thomas [13] focused only much later on transmembrane protein targets. Contact information was shown to be useful for both target classes, and thus should prove invaluable to AMPLE users.

Since the application of much more accurate *ab initio* protein structure prediction — obtained by restraining the conformational search space with residue-residue contacts — has not yet been explored, this initial study examines the impact on AMPLE performance of contact predictions. The aim is to extend the target tractability with particular focus on larger and more β-rich protein structures.

3.2 Materials & Methods

3.2.1 Target selection

In this study, targets from the ORIGINAL and TRANSMEMBRANE datasets were used. This resulted in a final set of 21 globular and 17 transmembrane protein targets. For details on how the targets were selected refer to **????**, and for details on each target refer to **????**.

3.2.2 Contact prediction

For all globular targets, one contact map was predicted with the fully automated metapredictor PCONSC2 v1.0 [14]. In summary, four Multiple Sequence Alignment (MSA)s were generated with each of JACKHMMER v3.1b2 [15] against the `uniref100` v2015-10 database and HHBLITS v2.0.15 [16] against the `uniprot20` v2013-03 database

[17] at E-value cutoffs of $10e^{-40}$, $10e^{-10}$, $10e^{-4}$ and 1. Each MSA was analysed with PSICOV v2.13b3 [18] and PLMDCA v2 [19] to produce 16 individual contact predictions. All 16 predictions and per-target PSIPRED v3 [20] secondary structure prediction, NETSURFP v1.0 [21] solvent accessibility information and HHBLITS v2.0.15 [16] sequence profile were provided to the PCONSC2 deep learning algorithm [14] to identify protein-like contact patterns. The latter produced a final contact map for each target sequence.

An additional contact map for β-structure containing targets was predicted using CCMPRED v0.3 [22] and reduced to β-sheet contact pairs using the CCMPRED-specific filtering protocol BBCONTACTS v1.0 [23]. Each MSA for CCMPRED contact prediction was obtained using HHBLITS v2.0.15 [16]. This entailed two sequence search iterations with an E-value cutoff of 10^{-3} against the `uniprot20` v2013-03 database [17] and filtering to 90% sequence identity using HHFILTER v2.0.15 [16] to reduce sequence redundancy in the MSA. Besides the contact matrix as input, BBCONTACTS requires a secondary structure prediction and an estimate of the MSA diversity. The secondary structure prediction was taken from the PCONSC2 step whilst the diversity factor was calculated using **??**.

For each transmembrane protein target, a MSA was generated using HHBLITS v2.0.16 [16] against `uniprot20` v2016-02 database [17]. Three search iterations were selected at an E-value cutoff of $1e^{-3}$ and minimum coverage with the target sequence of 60%. Contact predictions for each transmembrane target were obtained using the metapredictor METAPSICOV v1.04 [24], which in turn used the contact prediction algorithms CCMPRED v0.3.2 [22], FREECONTACT v1.0.21 [25] and PSICOV v2.1b3 [18]. Additionally, a set of contacts was also predicted using the MEMBRAIN server v2015-03-15 [26].

3.2.3 Contact-to-restraint conversion

For all targets, the predicted contact maps were converted to ROSETTA restraints to guide *ab initio* structure prediction. The FADE energy function was used to introduce a restraint in ROSETTA's folding protocol. The implementation described by Michel et al. [4] was used, which defined a contact to be formed during folding if the participating Cβ atoms (Cα in case of glycine) were within 9Å of one another. The top-L (L corresponds to the number of residues in the target sequence) contact pairs were converted to ROSETTA restraints, and if satisfied a “squared-well” bonus of -15.00 added to the energy function.

Additionally, all β-containing targets were subjected to a further conversion step in a separate condition. The approach of adding BBCONTACTS restraints to a previous prediction is outlined in **??**.

3.2.4 *Ab initio* structure prediction

Fragments for all targets were selected using the `make_fragments.pl` script shipped with ROSETTA. To ensure no closely homologous fragments were included in the fragment libraries, the `-nohoms` flag was set. This performs a PSIBLAST search to identify sequence homologs, whose corresponding Protein Data Bank (PDB) IDs are subsequently excluded from the fragment search. Each target's secondary structure prediction was provided to the fragment picker using the `-psipredfile` argument. The fragment libraries, contact restraints and secondary structure prediction were provided to the ROSETTA `AbinitioRelax` protocol [27] to predict 1,000 decoys per target. ROSETTA options were chosen according to the default protocol in AMPLE v1.0 [12]. ROSETTA v2015.05.57576 was used for globular targets and v2015.22.57859 for transmembrane ones for all ROSETTA-related protocols.

3.2.5 Molecular Replacement in AMPLE

All generated decoys were subjected to AMPLE v1.0 [12] for ensemble search model generation.

All transmembrane protein targets were processing using AMPLE's default parameters. Molecular Replacement (MR) trials were performed with software versions shipped in CCP4 v6.5.13 [28], with the exception of SHELXE v2014/14 [29] and ARP/wARP v7.5 [30].

All globular protein targets were subjected to AMPLE with two deviations from the default parameters. The `-use_scwrl` was set to subject all decoys to side-chain remodelling using SCWRL4 [31]. Furthermore, the number of clusters to trial was set increased from one to three via the `-num_clusters`parameter. All MR trials were performed with the version of software shipped with CCP4 v6.5.15 [28].

All MR solutions were assessed for success using the criteria described in ??.

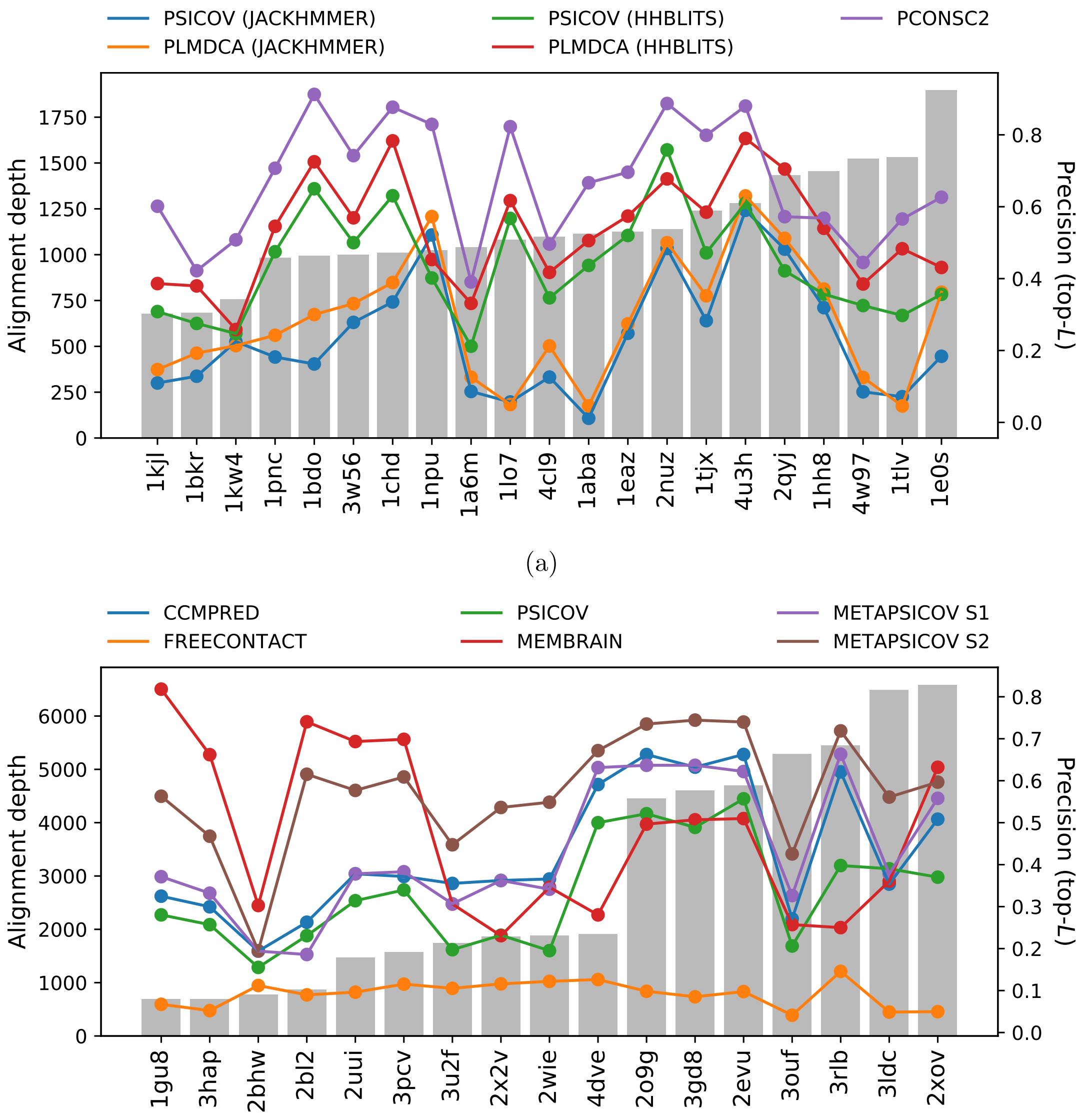
3.3 Results

In this study, the application of residue-residue contact predictions to *ab initio* protein structure prediction and subsequently MR was investigated. This proof-of-concept work is based on two datasets covering a range of globular and transmembrane protein targets. At the time of conducting this study, state-of-the-art contact prediction algorithms were applied to obtain the best possible contact predictions to see how much AMPLE performance could be improved [12].

3.3.1 Residue-residue contact prediction

Accurate coevolution-based residue-residue contact prediction is highly dependent on the availability of many divergent homologous sequences. As such, it is important to validate that the selected targets in this study satisfy such requirement.

The depth of MSAs obtained for each target sequence suggests that sufficient numbers of divergent homologous sequences are available. Across all globular targets, the minimum alignment depth is obtained for Galectin-3 domain (PDB ID: 1kjl) with 679 effective sequences and the maximum for G-protein Arf6-GDP (PDB ID: 1e0s) with 1,897 effective sequences (Fig. 3.1a). The median alignment depth for all globular targets is over 1,000, which is beyond the often suggested threshold of 200 sequences [32]. The MSAs for all transmembrane protein targets also surpass this threshold comfortably. The median alignment depth is much higher than for globular targets with 1,878 sequences (Fig. 3.1b). The minimum, which was obtained for Sensory rhodopsin II (PDB ID: 1gu8), is 692 sequences and the maximum for the sequence of Rhomboid protease GLPG (PDB ID: 2xov) is 6,583.



(a)

(b)

Figure 3.1: Alignment depth and contact precision analysis of (a) globular and (b) transmembrane protein targets. Contact predictions were obtained with several contact prediction algorithms. Precision scores were calculated for the top-L contact pairs. JACKHMMER and HHBLITS alignments for PSICOV and PLMDCA contact predictions in (a) were obtained with E-value $1e^{-4}$.

In coevolution-based contact prediction, the precision depends on alignment depth. Despite sufficient number of effective sequences across all targets, findings in this study suggest that some (meta-)predictors cannot fully utilise greater alignment depths to correct contact pairs (Fig. 3.1).

PCONSC2 — a metapredictor using eight starting alignments and two contact predictors — outperforms its individual parts for almost all globular targets (Fig. 3.1a). Although only four individual components are shown in Fig. 3.1a, the pattern translates across all 16 individual predictions per target. Such results suggest that precision greatly depends on the tool used to identify and select homologous sequences for the MSA. A closer inspection of mean precision scores resulting from HHBLITS- and JACKHMMER-based alignments shows higher precision scores for top-L contact pairs based on the former alignments (Table 3.1). Nevertheless, the Machine Learning approach in PCONSC2 to combine more and less precise individual predictions results in superior precision in the output (Table 3.1). No correlation could be established between alignment depth and precision for either individual predictors or the metapredictor PCONSC2 (Fig. 3.1a).

Table 3.1: Summary of mean PCONSC2 raw contact prediction precision based on JACKHMMER and HHBLITS alignments and PSICOV, PLMDCA and PCONSC2 coevolution-based contact prediction.

Contact prediction		**Alignment E-value cutoff**			
		$1e^{0}$	$1e^{-4}$	$1e^{-10}$	$1e^{-40}$
PSICOV	JACKHMMER	0.240	0.239	0.213	0.167
	HHBLITS	0.439	0.435	0.354	0.209
PLMDCA	JACKHMMER	0.293	0.288	0.252	0.140
	HHBLITS	0.545	0.530	0.447	0.224
PCONSC2		0.667			

Contacts for transmembrane protein targets in this study were predicted with the metapredictor METAPSICOV and the transmembrane-specific predictor MEMBRAIN. METAPSICOV STAGE 1 and STAGE 2 predictions outperform MEMBRAIN in nine and ten cases, respectively, whilst MEMBRAIN outperforms METAPSICOV for the rest (Fig. 3.1b). The METAPSICOV algorithm utilises the raw predictions by CCMPRED, FREECONTACT and PSICOV to generate its STAGE 1 and STAGE 2 predictions. METAPSICOV STAGE 1 predictions are near identical to CCMPRED, whereby 15 of 17 targets show an absolute $\Delta_{precision}$ of less than 0.05 (Fig. 3.1b). This similarity does not propagate to METAPSICOV STAGE 2 predictions with only a single target showing such similar precision values (Fig. 3.1b). Amongst the three raw predictors used by METAPSICOV, FREECONTACT performs by far the worst with a mean precision of 0.09 across all transmembrane targets. PSICOV shows similar trend to CCMPRED when assessed by target, which results in a mean absolute $\Delta_{precision}$ of 0.10.

The addition of BBCONTACTS contact pairs to improve structure prediction accuracy for β-structure containing targets is a novel aspect introduced in this study. The initial step of the addition of BBCONTACTS contact pairs included the filtering of one- and two-pair strand contacts from the original BBCONTACTS list, since those contained high numbers of false positives (Jessica Andreani, personal communication). The findings in this study confirm this for all β-structure containing targets. Precision values improved for all targets with changes ranging from 0.01 to 0.14 whilst retaining on average 80% of all contacts. Filtered BBCONTACTS predictions were combined with other contact maps, i.e. PCONSC2, to either upweight or add contact pairs. Findings in this study highlight that upweighted contact pairs are more precise than ones to be added. The minimum precision score for a set of upweighted contacts is 0.72 for 29 contact pairs and the maximum of 1.00 for up to 27 contact pairs. In comparison, novel BBCONTACTS contact pairs not present in the base contact map range in precision scores from 0.22 (nine contacts) to 0.76 (21 contacts).

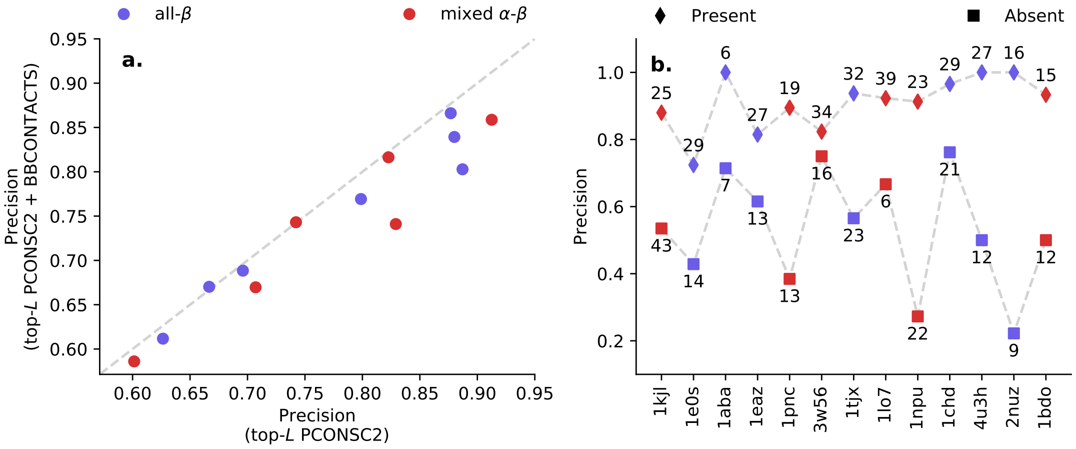


Figure 3.2: Evaluation of BBCONTACTS contact pairs. (a) Precision evaluation of PCONSC2 contact map with and without BBCONTACTS. (b) Precision evaluation of BBCONTACTS contact pairs split by status of presence or absence in the base PCONSC2 contact list. Numbers besides each marker indicate the number of contacts. The order scatter points is identical between both subplots and based on the PCONSC2 precision values (x-axis) in (a).

Despite the high precision of BBCONTACTS contact pairs, the merge of such pairs with top-L PCONSC2 contact pairs results in an expected loss in precision for the resulting contact set (Fig. 3.2). True positive contacts, which dominate the BBCONTACTS contact set, are usually also predicted by PCONSC2, and thus upweighted (Fig. 3.2). Since upweighting does not affect the precision, the value remains unaffected after this procedure. However, contact pairs unique to BBCONTACTS contain more false positives. Once added to the base PCONSC2 contact list, these contacts therefore reduce the precision value (Fig. 3.2). Either subset of BBCONTACTS contacts does not show any correlation between the number it contains and its precision. The fold of the target does not show any clear distinction between better and worse sets of contacts either (Fig. 3.2).

3.3.2 Protein structure prediction

Predicted contact information is particularly useful to limit the conformation search space in *ab initio* protein structure prediction. Since such predictions are the basis for

AMPLE studies presented in this thesis, it is important to analyse the improvement in decoy quality.

Globular protein targets benefit greatly from the addition of PCONSC2 residue contacts. All but one target see median Template-Modelling score (TM-score) improvements of at least 0.05 when comparing contact-assisted PCONSC2 decoys with simple ROSETTA decoys (Fig. 3.3). The greatest improvement over 1,000 decoys was achieved for Oxy-myoglobin (PDB ID: 1a6m) with an improvement in median TM-score of 0.42. The decoys for Ankyrin (PDB ID: 2qyj) show a minor decrease in median TM-score of 0.04; however, the median TM-score for ROSETTA decoys is 0.78, and thus a minor decrease may be negligible.

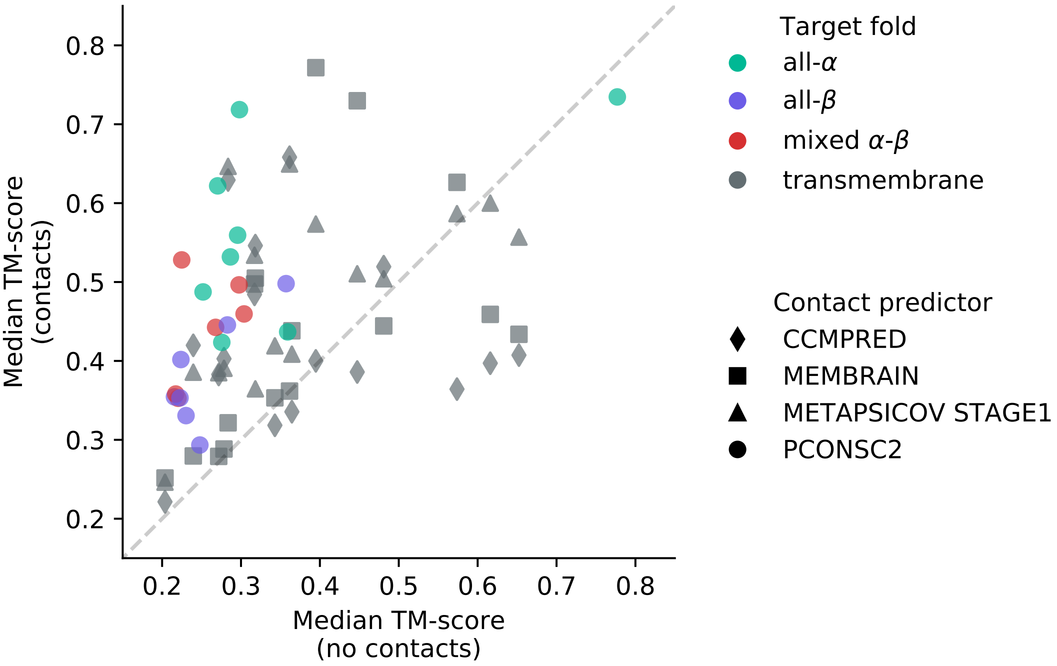


Figure 3.3: Effect of contact distance restraints on *ab initio* decoy quality by comparison of unrestrained (*no contacts*) and contact-restrained (*contacts*) median TM-scores for 1,000 decoys per target. Colours indicate the target fold and symbols the contact prediction algorithm.

Previously, *ab initio* protein structure prediction for globular targets was greatly limited by target fold and chain length. The addition of residue-residue contacts enhances decoy quality primarily for α-helical and mixed α-β protein targets (Fig. 3.4). Whilst only one all-α target has more than 50% native-like decoys in its ROSETTA decoy set, five targets surpass this threshold when PCONSC2 contact data is used to restrain the folding procedure. Similarly, the median TM-score of no mixed α-β target decoy set surpasses the TM-score threshold of 0.5 with ROSETTA decoys compared to one for PCONSC2 decoys with three further ones greater than 0.4. All-β targets also benefit from the addition of contact restraints, although decoy sets do not surpass the native-like threshold by median TM-score (Fig. 3.4). Larger targets do not benefit any more than smaller targets from the addition of residue contacts to the structure prediction protocol. The only real exception to this are the decoys for the CheB methylesterase domain (PDB ID: 1chd), for which the majority of ROSETTA decoys are almost random-like whilst PCONSC2 decoys are native-like (Fig. 3.4).

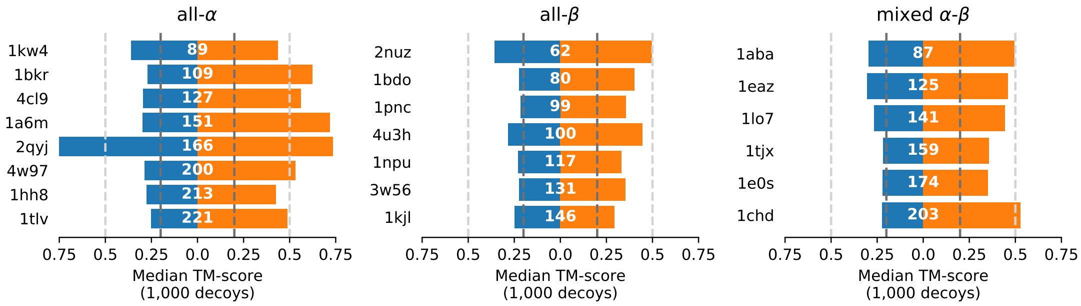
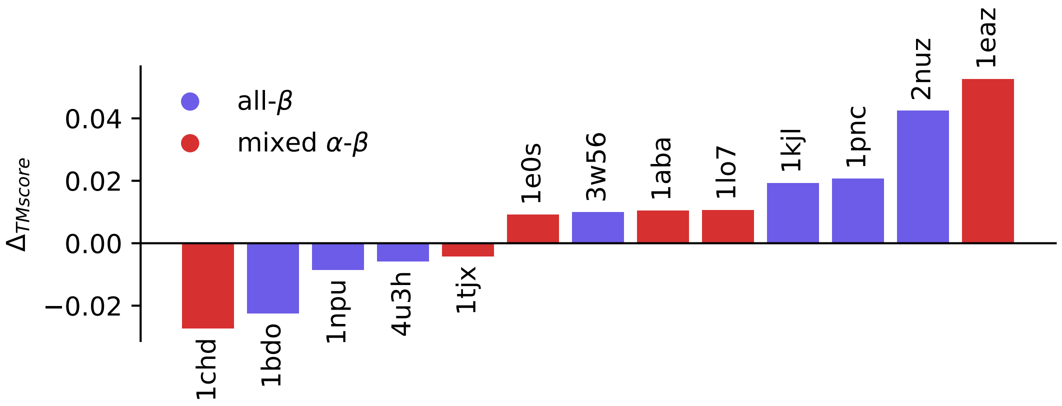
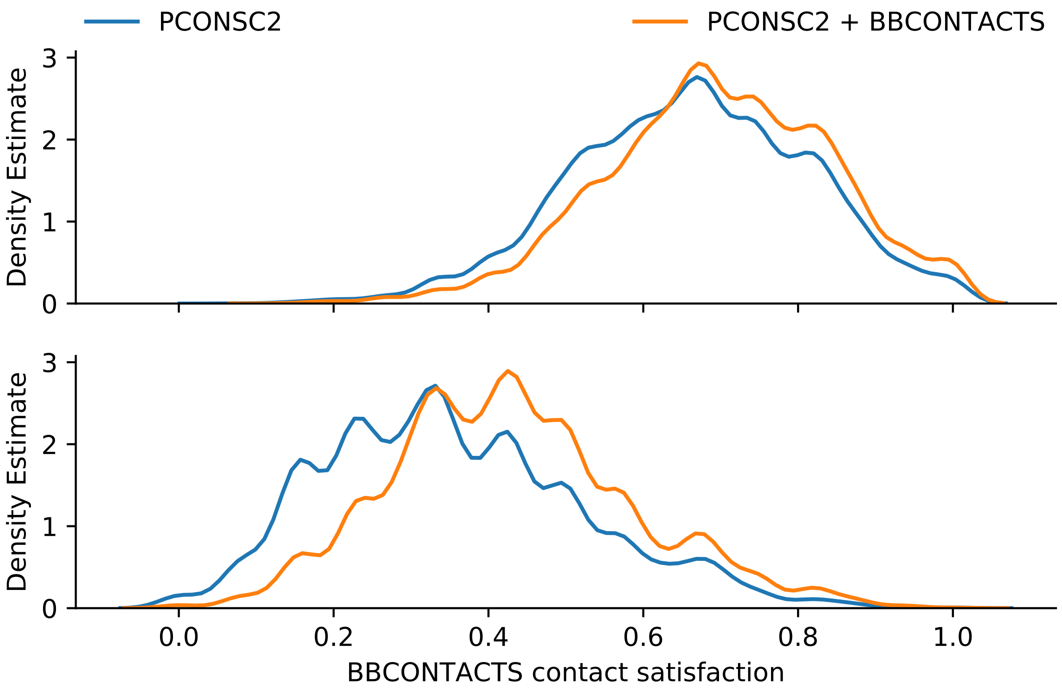


Figure 3.4: TM-score comparison for globular targets separated by fold and ordered by target chain length. Median TM-scores for 1,000 decoys generated with simple ROSETTA (orange) or contact-assisted ROSETTA (blue) runs. White numbers in each row correspond to the target chain length. Bars surpassing the dark gray line indicate that the majority of structures are better than random, whilst the light gray line indicates that the majority of structures are native-like [33].

The enhancement of β-structure specific contact pairs is an important part of this study. Previously, the precision of the added BBCONTACTS contact pairs has been demonstrated. The next essential step is to explore how the BBCONTACTS supplement enhances or degrades decoy quality after ROSETTA *ab initio* protein structure prediction. Given 13 β-structure containing targets, eight targets achieve better overall decoy quality with added BBCONTACTS (Fig. 3.5a). The smallest improvement is observed for target 1e0s with 0.01 TM-score units, whilst the largest for target 1eaz with 0.05 units. The remaining five targets — PDB IDs 1chd, 1bdo, 1npu, 4u3h and 1tjx — saw decreases in median TM-score up to 0.03 when BBCONTACTS contact pairs were added as restraints (Fig. 3.5a). No clear difference between fold classes, i.e. mixed α-β or all-β targets, could be observed, although mixed α-β targets do show slightly greater extremes (Fig. 3.5a).



(a)



(b)

Figure 3.5: (a) TM-score comparison for β-structure containing globular targets separated by fold and ordered by the difference in median TM-score between PCONSC2 and PCONSC2+BBCONTACTS decoys. Positive values indicate a better median TM-score in favour of PCONSC2+BBCONTACTS decoys, whilst negative values those for PCONSC2. PDB IDs are provided alongside each bar. (b) Satisfaction of BBCONTACTS contact pairs in decoys with added β-structure contact restraints (PCONSC2 + BBCONTACTS) and those without (PCONSC2).

An analysis of the satisfaction of BBCONTACTS contact pairs in decoys where extra β-sheet contact pairs were used as restraints compared those where they were not highlights a greater satisfaction in the former. This indicates that the added and upweighted BBCONTACTS β-structure contact restraints enhance the formation of β-sheets in the resulting decoys, which would explain the overall improved decoy quality for more than half the targets outlined previously. A separation of contact satisfaction by upweighted and added BBCONTACTS contact pairs indicates that the upweighting has less effect compared to the addition (Fig. 3.5b). Although the former shows a marginal improvement in BBCONTACTS contact satisfaction of decoys without upweighted restraints, the difference is minimal. In comparison, PCONSC2 decoys without the added BBCONTACTS restraints show less satisfaction for such contacts, indicating that they did not form as often compared to PCONSC2+BBCONTACTS. In combination with the upweighting, this explains that β-rich regions are predicted more accurately when BBCONTACTS contact pairs supplement PCONSC2 contacts.

Transmembrane protein targets were modelled using residue-residue contact predictions derived with CCMPRED, MEMBRAIN and METAPSICOV STAGE1. A ROSETTA benchmark was also run to compare contact-assisted decoys to the current norm. Findings in this study highlight the much improved decoy quality for almost all targets when contact information is used to reduce the conformational sampling space (Fig. 3.3). Across all methods, only two targets suffered from the addition of contact restraints during *ab initio* protein structure prediction, namely the domains of ATP synthase C chain (PDB ID: 2wie) and ATP synthase subunit C (PDB ID: 3u2f). In both cases, ROSETTA generates decoys with median TM-score of greater than 0.6 when no contact restraints are used. This contrasts strongly with contact-assisted decoy sets, for which only METAPSICOV STAGE1 predictions yielded overall native-like decoys, i.e. median TM-score of greater than 0.5.

A split for decoy quality comparison between no-contact and contact-assisted decoy sets by contact prediction algorithm shows that CCMPRED contacts are not sufficiently precise to always improve decoy quality. Six out of 16 targets are predicted more accurately without CCMPRED contact information (Fig. 3.6). In comparison, MEMBRAIN and METAPSICOV STAGE1 contact predictions result in enhanced decoy quality to the extend that only three and two decoy sets are worse than their no-contact counterparts (Fig. 3.6). Most notably, either of the three contact predictions per target performs better for certain targets. The most extreme example may be the decoy sets for Bacteriorhodopsin (PDB ID: 3hap) for which CCMPRED contacts result in decoy quality degradation of 0.06, METAPSICOV STAGE1 in a slight improvement of 0.06 and MEMBRAIN in an improvement of 0.28 TM-score units. This translates into absolute decoy counts with native-like fold — i.e., TM-score ≥ 0.5 — of the following: 274 for decoys without contact guidance, 289 for CCMPRED contact guidance, 538 for METAPSICOV STAGE1 contact guidance, and 996 for MEMBRAIN contact guidance. Similar examples exist (e.g., PDB IDs 1gu8, 3rlb or 4dve in Fig. 3.6) and highlight that no single method yields the best decoys in all circumstances.

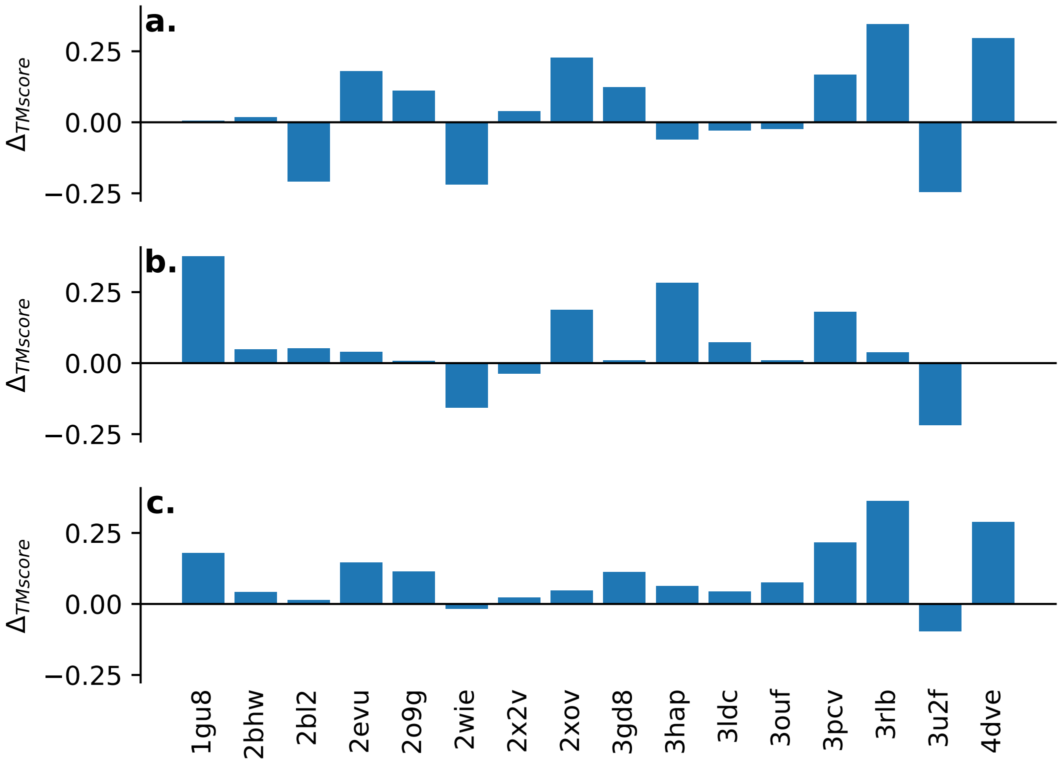


Figure 3.6: TM-score difference between contact-assisted and simple ROSETTA decoys for transmembrane protein targets. Positive $\Delta_{TMscore}$ values indicate more accurate contact-assisted decoys, whilst negative values better decoys without the addition of contacts. $\Delta_{TMscore}$ values were computed by median TM-score. Contact restraints were obtained with (a.) CCMPRED, (b.) MEMBRAIN, and (c.) METAPSICOV STAGE1.

3.3.3 Molecular Replacement

The most important aspect of this study is the impact of contact-assisted decoys in AMPLE-MR. AMPLE is primarily limited by a target's chain length and fold, which typically cannot exceed 150 residues, and performs poorly for β-rich folds [12]. Findings presented in previous sections of this chapter outlined improvements in overall decoy quality when contact information was used as distance restraints in *ab initio* protein structure prediction. However, it is yet to be seen how the improved decoy quality translates into MR structure solutions.

3.3.3.1 Globular protein targets

Structure solutions were attempted for a total of 21 globular targets. Simple ROSETTA decoys — those without contact restraints and AMPLE's current default — resulted in nine structure solutions (Fig. 3.7). The addition of PCONSC2 contact-restraints to the structure prediction procedure improved decoy quality to achieve four additional structure solutions. However, the structure of the N-terminal region of P67Phox (PDB ID: 1hh8) was not solved when PCONSC2-restrained decoys were used compared to simple ROSETTA ones. The addition of BBCONTACTS distance restraints to up-weight and supplement PCONSC2 contacts enabled a further unique solution for the Phosphoinositol (3,4)-bisphosphate PH domain (PDB ID: 1eaz) (Fig. 3.7).

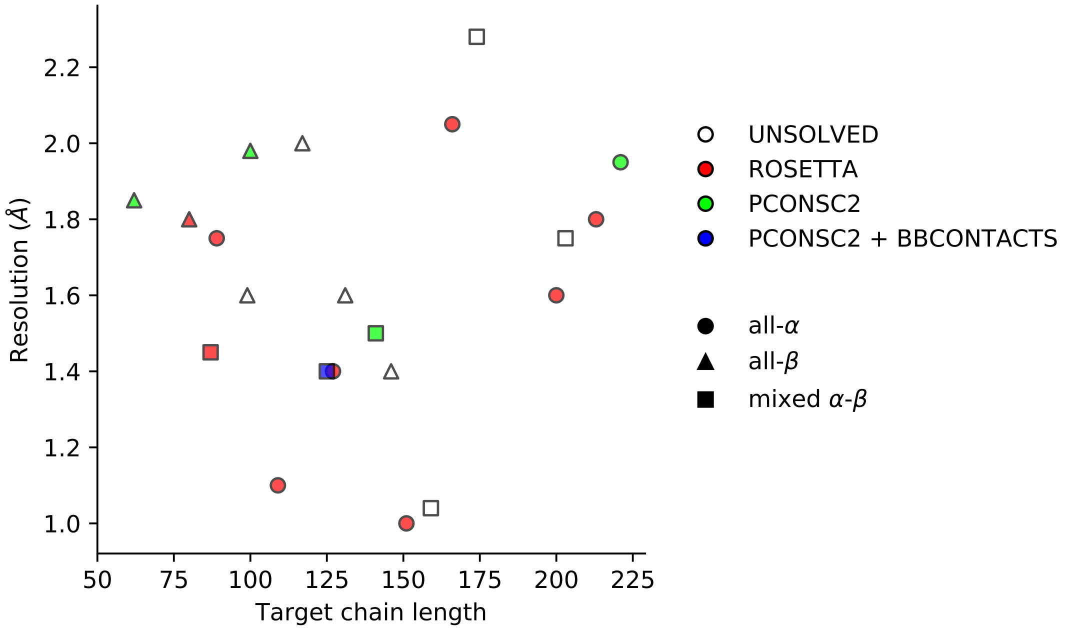
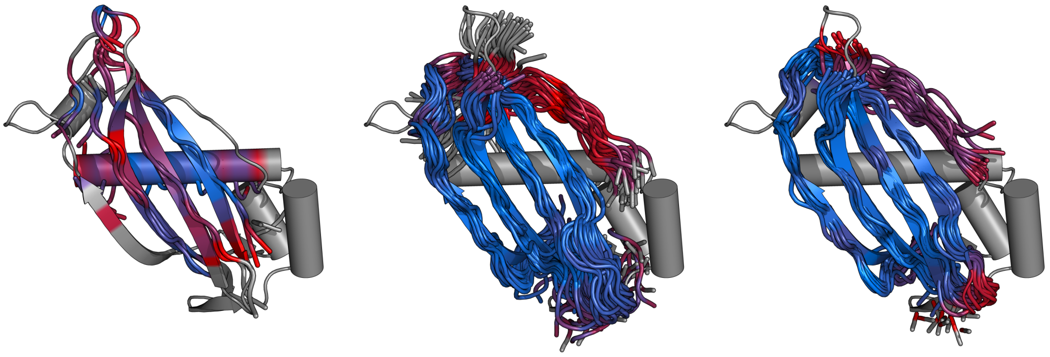


Figure 3.7: Summary of structure solutions obtained with AMPLE using no-contact ROSETTA, PCONSC2-contact-restraint-assisted ROSETTA and PCONSC2 + BBCONTACTS-assisted ROSETTA decoy sets. Empty markers indicate unsolved targets. Filled markers highlight the minimum decoy set by complexity of the prediction procedure to obtain MR structure solution. Marker shape distinguishes the fold class.

The majority of structure solutions were obtained for all-α targets in the dataset, with a total of eight structure solutions (Fig. 3.7). Seven of those eight structure solutions were achieved with unrestrained ROSETTA decoys, with target chain lengths up to 213 residues. The largest target in the globular dataset, and the only all-α target that required residue contacts, totals 221 residues in target chain length, which exceeds AMPLE's previously benchmarked limits for globular targets greatly [12]. In comparison to all-α targets, β-structure containing proteins require the contact restraints to result in sufficiently accurate decoys for MR. Across all-β and mixed α-β targets, only two structure solutions were obtained with unrestrained ROSETTA decoys. This contrasts to an additional three targets when PCONSC2 restraints were used during *ab initio* structure prediction. Furthermore, the addition of BBCONTACTS contact restraints enabled an additional structure solution, yielding much greater success for β-structure containing protein targets compared to the previous default. Structure solutions for β-containing targets were obtained for target chain lengths up to 141 residues (Fig. 3.7).

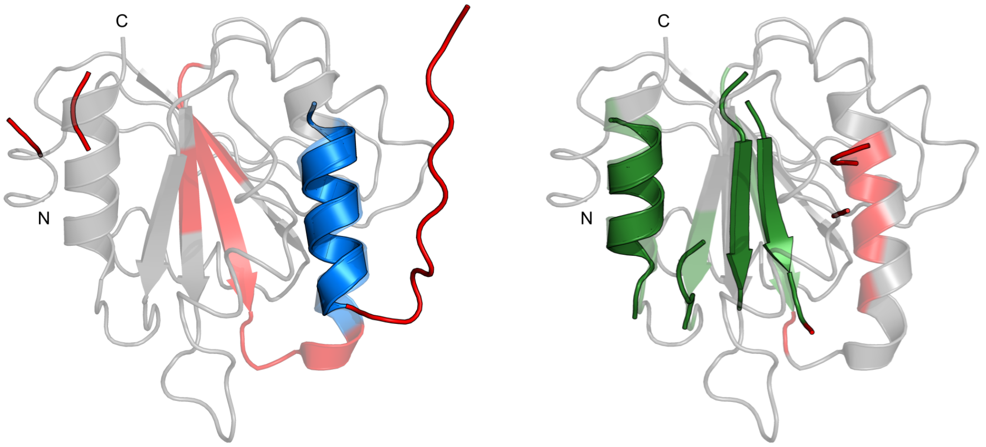
In this study, two exceptional cases exemplify the application of contact predictions and their benefit to MR. The first, the structure solutions for 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7) based on AMPLE ensemble search models derived from the PCONSC2 and PCONSC2+BBCONTACTS decoy sets. Without contact restraints, AMPLE search models do not accurately represent the target fold (Fig. 3.8a). In comparison, precise residue-residue contacts primarily restraining the large β-sheet yield decoys of sufficient quality to achieve MR structure solution with both PCONSC2 (Fig. 3.8b) and PCONSC2+BBCONTACTS (Fig. 3.8c) decoy sets.



(a) (b) (c)

Figure 3.8: Structural superposition of the (a) ROSETTA (Cα Root-Mean-Square Deviation (RMSD) 2.814 Å; ensemble contains two structures), (b) PCONSC2 (Cα RMSD 1.748 Å; 30 members) and (c) PCONSC2+BBCONTACTS (Cα RMSD 1.760 Å; 15 members) search-model ensembles for 4-hydroxybenzoyl CoA thioesterase (PDB ID: 1lo7). Examples are the highest scoring search models based on SHELXE Correlation Coefficient (CC) score, with only (b) and (c) leading to successful MR structure solutions. Search models are shown as tubes and crystal structures as cartoons. (a) and (c) are 50% of the target sequence, while (b) is 55%. The colour scale illustrates the pairwise Cα RMSD between each search-model ensemble (represented by its first member) and the crystal structure, with blue representing the minimum Cα RMSD and red the maximum. Unaligned residues are coloured grey.

The second exceptional case, PDB ID 1e0s, does not yield any MR structure solution with either decoy set according to the stringent criteria for MR success applied in this study (see ??). However, a Residue-Independent Overlap (RIO) analysis of PHASER solutions, i.e. after MR, indicates that some PCONSC2 and PCONSC2+BBCONTACTS AMPLE search models were placed partially correctly (Fig. 3.9). For the top PCONSC2 search model, 40% (12 residues) of the search model residues are correctly superimposed, albeit out of register on the target structure (PHASER Translation Function Z-score (TFZ)=4.7, PHASER Log-Likelihood Gain (LLG)=16) (Fig. 3.9a). For the top PCONSC2+BBCONTACTS search model, 77% (30 residues) of the search model were superimposed in an in-register fashion (PHASER TFZ=5.3, PHASER LLG=17) (Fig. 3.9b). For the latter, expert manual intervention might allow structure determination, but in this case the correct solution was not prominent in the list of MR placements. Nevertheless, it is clear that even when overall structure solution was not automatically achieved the PCONSC2+BBCONTACTS search model provided better results which might be recoverable as successes in the future as MR and post-MR software improves still further.



(a) (b)

Figure 3.9: Top-PHASER solutions for PDB ID 1e0s based on RIO scores for (a) PCONSC2 (RIO score 12) and (b) PCONSC2+BBCONTACTS (RIO score 30) search models. Search-model colour coding indicates useful superposition of residues by in- (green) or out-of-sequence register (blue) residues as well as misplaced (red) residues. The addition of BBCONTACTS restraints produced a more accurate model with correctly placed β-strands that was placed correctly. Both structures are shown in cartoon representation with the crystal structure shown as a transparent cartoon. Unaligned reference crystal structure residues are coloured grey.

With much improved decoy quality deriving from the use of predicted residue restraints to guide *ab initio* structure prediction, the question arises whether AMPLE's existing cluster-and-truncate approach remains the most suitable for obtaining a conserved, native-like core from the decoys found in the largest clusters. For globular targets solved using simple ROSETTA decoys, certain features throughout AMPLE's cluster-and-truncate approach typically correlated with eventual success in structure solution [12]. In general, the greater the number of decoys in the largest cluster the more likely the success was with derived search models. Truncation removed structurally variant parts leading to smaller more accurate ensemble subsets of the cluster decoys. Although successful search models were found at every truncation interval, the majority were derived with search models containing around 30 residues. Lastly, each of the potential nine search models derived at each truncation level (three subclustering radii with three side chain treatments each) can lead to non-redundant structure solutions. Similar observations, particularly with respect to the most successful search model size range were made for other target classes [2, 34] and for *ab initio* decoys made with QUARK [35].

A size comparison of the largest clusters of ROSETTA and PCONSC2+BBCONTACTS (or PCONSC2 for all-α) decoys indicated a median increase of 122 decoys per cluster in the latter. All cluster sizes increased except for target 2qyj. More accurate *ab initio* decoys are directly linked to larger cluster sizes because of the associated increase in convergence [36]. Here, as expected, the largest cluster contains better than average quality decoys but the size of the largest cluster does not link to the total number of successful search models.

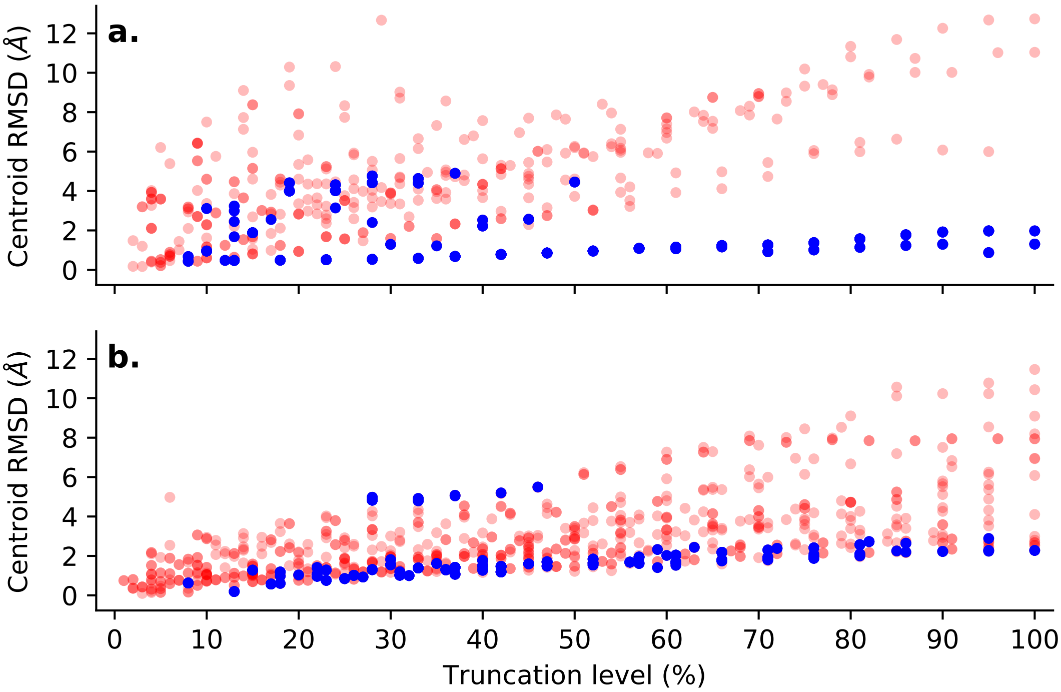


Figure 3.10: The percentage of sequence in the search model is mapped against the RMSD over all Cα atoms of the first representative of each search-model ensemble derived from the largest cluster against the native structure. Successful structure solutions of individual search models are highlighted in blue and unsuccessful solutions in red. Progressively darker shades of either colour correspond to increasing numbers of overlapping points. Progressive truncation is shown for (a) ROSETTA and (b) PCONSC2+BBCONTACTS decoys (or PCONSC2 decoys for all-α targets).

In comparison to the clustering step, the progressive truncation of decoys in the largest cluster at 20 different intervals directly affects the number of successful search models. An analysis of the progressive truncation and the effects on search model accuracy revealed that all successful search model ensembles had a Cα-RMSD better than 5.5Å compared to the native structure (Fig. 3.10). Although the latter cutoff is independent of whether contact information was provided during *ab initio* modelling, a clear difference between the ROSETTA and PCONSC2+BBCONTACTS (or PCONSC2 for all-α) ensemble search models for all targets can be observed. In total, ROSETTA decoys for all targets produced 1,314 ensemble search models based on the largest clusters. In comparison, PCONSC2+BBCONTACTS decoys generated for the same targets 2,469 search model ensembles from the largest clusters. This increase is the result of a more successful subclustering process due to the increased structural homogeneity across the decoys in the largest cluster. The most notable difference between the two sets is detected for the Small G-protein ARF6-GDP (PDB ID: 1e0s), which produced three ensemble search models based on ROSETTA decoys and 90 based on PCONSC2+BBCONTACTS decoys. Additionally, ensemble search models with structural fragments of 15-40 residues of the target sequence are more likely to succeed in MR phasing than larger or smaller search models [12]. Here we find that the same range is most successful for contact-assisted decoys (Fig. 3.11). Out of 246 successful search models for PCONSC2+BBCONTACTS decoys derived from the largest cluster (PCONSC2 for all-α), 101 successful search models contained 15-40 residues. Significantly, some cases like the PH domain of TAPP1 (PDB ID: 1eaz) and the N-terminal bromodomain of human BRD4 (PDB ID: 4cl9) only solved with truncated search models in this size range. Nevertheless, structure solutions were also achieved with larger or smaller search models. The smallest search model leading to a structure solution contained nine residues (8% of total sequence) and solved the Calponin Homology domain from human β-spectrin (PDB ID: 1bkr). In comparison, the largest successful search model in terms of residues was found for the designed full consensus ankyrin (PDB ID: 2qyj) domain with 158 residues (95% of total), and in terms of percentage of the total sequence the untruncated, 62 residue search model for α-spectrin SH3 domain (PDB ID: 2nuz) was successful. Therefore, although truncating the *ab initio* decoys at different levels remains essential for contact-assisted decoys, biasing sampling into the most successful size range may be advantageous in future runs.

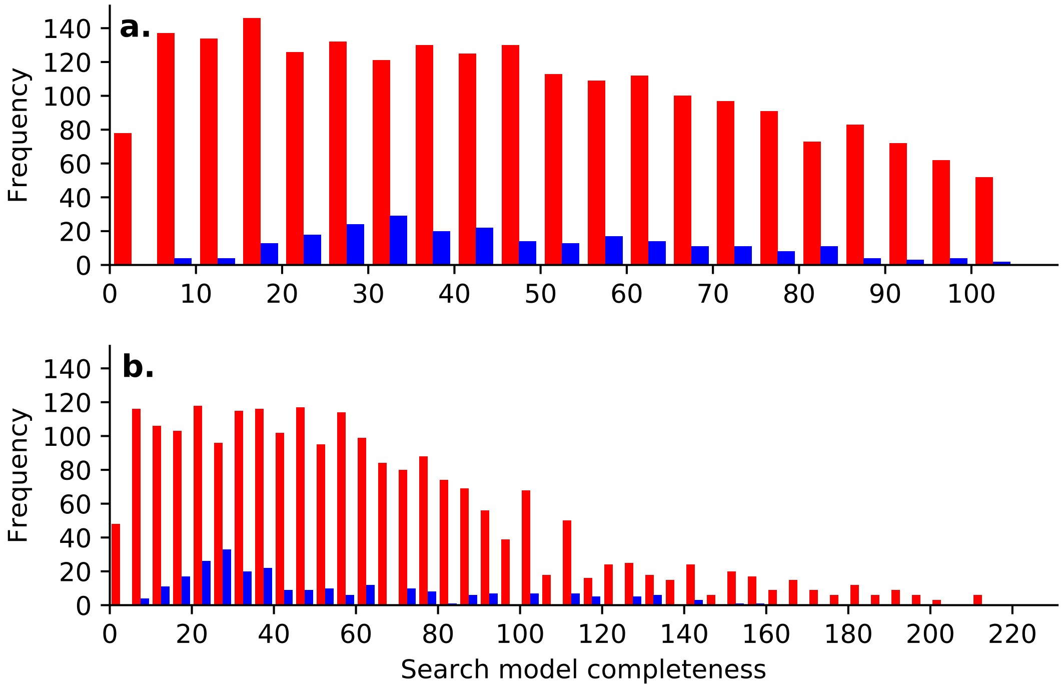


Figure 3.11: Summary of AMPLE truncation ranges for structure solution. (a) Percentage of residues and (b) number of residues per chain in search model mapped against the number of search models leading to structure solution (blue) or not (red).

The truncated decoys are further processed by subclustering at three different atomic radii, with the resulting subclusters previously found to be similarly successful [12]. Similar trends are seen here: 36% of structure solutions with ROSETTA decoys were achieved with a subclustering radius of 1Å, 36% at a radius of 2Å, and 28% at a radius of 3Å. For PCONSC2+BBCONTACTS (or PCONSC2 for all-α) decoy sets similar numbers were observed (35% at radius of 1Å; 40% at 2Å; 25% at 3Å). Nevertheless, in terms of number of targets solved all three subclustering radii were essential. Largest-cluster decoys for target 1eaz produced a total of 327 search models, but only one solved and this derived from a subclustering radius of 1Å. In comparison, contact-assisted decoys from the largest cluster for target 4u3h achieved structure solutions solely with decoys subclustered at 2Å. A single search model with subclustering radius of 3Å solved the target 4cl9 with ROSETTA decoys. The final step in search model creation is the side-chain processing of each subclustered ensemble. Similarly to the subclustering, no difference was observed between ROSETTA and PCONSC2+BBCONTACTS decoys. For both the polyalanine treatment is most successful, covering 37% of successful search models for ROSETTA decoys and 44% for PCONSC2+BBCONTACTS decoys. For almost all targets, the polyalanine side-chain treatment would be enough to obtain a structure solution. However, some cases, like the target 1eaz, only solve with either or both of the remaining treatments. Thus, relying solely on polyalanine side-chain treatment may limit the overall success rate, although trialling polyalanine ensemble search models first might lead to structure solution faster.

3.3.3.2 Transmembrane protein targets

The MR structure solution attempts given the decoy sets for transmembrane protein targets was conducted by Dr Jens Thomas and is documented in Thomas [13] and Thomas et al. [2].

In summary, MR structure solution successes with decoys restrained by either of the three contact prediction protocols — CCMPRED, MEMBRAIN and METAPSICOV STAGE1 — were mixed. CCMPRED solved three targets, MEMBRAIN solved five and METAPSICOV STAGE1 decoys solved four. Simple ROSETTA decoys resulted in four structure solutions. CCMPRED and METAPSICOV STAGE1 both solved target 4dve, which could not be solved with any other method, and METAPSICOV STAGE1 also solved target 2o9g, which had previously only been solved with the AMPLE library of ideal helices.

3.4 Discussion

The change in statistical model for residue-residue contact prediction has enabled great improvements to its precision. Today, contact information is often used to restrain the conformational search space to enable accurate *ab initio* protein structure prediction. In this study, the effect of such improved structure predictions was examined with a particular interest of their application in unconventional MR in AMPLE. The main focus of the presented work rests with the aim to extend AMPLE's target tractability, both for larger and more β-rich protein targets.

The addition of predicted residue-residue contacts unsurprisingly improved the quality of *ab initio* protein structure predictions, which is in line with numerous previous studies [e.g., 3–11]. The improved decoy quality directly translates to further structure solutions with AMPLE. Contact-unassisted decoys, i.e. the current default, achieved nine and four solutions for globular and transmembrane protein targets, respectively. In comparison, contact-assisted decoys solve a further five globular targets, whilst contact-assisted decoys solved some different targets compared to contact-unassisted decoys for transmembrane protein targets.

The initial findings in this study highlighted the successful application of contact prediction to extend the target tractability with regards to the target chain length. Bibby et al. [12] previously benchmarked *ab initio* protein structure predictions up to chain lengths of 120 residues. However, the findings indicated that larger targets should be tractable with AMPLE, especially all-α ones [12]. In this study, we confirm such extended target tractability, with contact-unassisted decoys leading to structure solutions up to 213 residues for globular targets and 223 residues for transmembrane protein targets. The addition of contacts to limit the conformational search space enabled structure solutions for the largest target in the globular target dataset with a 221-residue chain length and the transmembrane dataset with a 249-residue chain length. The fact that both of these targets are the largest in their sets is highly suggestive that contact-assisted decoys may enable solutions for much larger targets. In fact, recent research highlighted the successful *ab initio* structure prediction of globular and transmembrane protein targets with target chain lengths in excess of 300 residues [10], which further supports this claim.

AMPLE was previously also limited by the target fold [12]. Whilst the majority of all-α protein targets were comfortably tractable, mixed α-β and all-β targets were not [12]. This limitation primarily arose from upstream limitations in *ab initio* protein structure prediction but also the challenging task of tracing β-sheets in SHELXE, which was used to assess the successful structure solution. The use of contact-assisted decoys in AMPLE improved the target tractability for β-structure-containing protein targets. Structure solutions for four additional, β-structure-containing targets were obtained when contact-assisted decoys were used. A novel approach of combining β-sheet-specific contact pairings with a normal base prediction enabled the structure solution of one further target. Although no MR structure solutions were lost when BBCONTACTS contact pairs were added to a base set of contact restraints, further studies are required to support routine application in AMPLE. Furthermore, BBCONTACTS contact pairs were identified by analysis of a CCMPRED contact map, which is generally much noisier than metapredictor alternatives. Thus, further studies may explore the benefits or drawbacks of BBCONTACTS based on alternative contact maps. Lastly, since the release of BBCONTACTS, other β-strand specific contact identification protocols have been developed [37], which may need to be explored too.

Beyond the proof-of-concept study outlined in this chapter, it is very important to appreciate new limitations and unexplored areas of this work. At the time of conducting this study, PCONSC2 proved to be the state-of-the-art metapredictor. However, numerous alternatives have since been developed with more advanced Machine Learning architectures to post-process multiple individual contact predictions [e.g., 6, 24]. Furthermore, the optimal introduction of contacts as distance restraints into *ab initio* protein structure prediction protocols is not yet clearly defined, and thus leaves the choice to the user without much comparison or guidance as to which works best. Lastly, contact information was exclusively used to restrain the *ab initio* protein structure prediction procedure despite other potential applications in the AMPLE cluster-and-truncate algorithm. Therefore, it is crucial in subsequent chapters to explore different aspects of the outlined limitations of the work presented here.

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

Chapter 7

Protein fragments as search models in Molecular Replacement

Chapter 8

Conclusion & Outlook

Appendix A

Appendix

Bibliography

[1] F. Simkovic, J. M. H. H. Thomas, R. M. Keegan, M. D. Winn, O. Mayans, D. J. Rigden, en, *IUCrJ* **July 2016**, *3*, 259–270.

[2] J. M. H. Thomas, F. Simkovic, R. M. Keegan, O. Mayans, Y. Zhang, D. J. Rigden, C. Zhang, Y. Zhang, D. J. Rigden, *Acta Crystallographica Section D Structural Biology* **Dec. 2017**, *73*, 985–996.

[3] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, en, *PLoS One* **Dec. 2011**, *6*, e28766.

[4] M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, A. Elofsson, en, *Bioinformatics* **Sept. 2014**, *30*, i482–8.

[5] T. Kosciolek, D. T. Jones, en, *PLoS One* **Mar. 2014**, *9*, e92197.

[6] S. Ovchinnikov, L. Kinch, H. Park, Y. Liao, J. Pei, D. E. Kim, H. Kamisetty, N. V. Grishin, D. Baker, en, *Elife* **Sept. 2015**, *4*, e09248.

[7] S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, D. Baker, en, *Proteins* **Sept. 2016**, *84 Suppl 1*, 67–75.

[8] M. Michel, D. Menéndez Hurtado, K. Uziela, A. Elofsson, *Bioinformatics* **2017**, *33*, i23–i29.

[9] S. H. P. de Oliveira, E. C. Law, J. Shi, C. M. Deane, en, *Bioinformatics* **Nov. 2017**, DOI `10.1093/bioinformatics/btx722`.

[10] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, D. Baker, en, *Science* **Jan. 2017**, *355*, 294–298.

[11] S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, en, *PLoS Comput. Biol.* **Jan. 2017**, *13*, e1005324.

[12] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, D. J. Rigden, en, *Acta Crystallogr. D Biol. Crystallogr.* **Dec. 2012**, *68*, 1622–1631.

[13] J. M. H. Thomas, PhD thesis, University of Liverpool, **Jan. 2017**.

[14] M. J. Skwark, D. Raimondi, M. Michel, A. Elofsson, en, *PLoS Comput. Biol.* **Nov. 2014**, *10*, e1003889.

[15] L. S. Johnson, S. R. Eddy, E. Portugaly, en, *BMC Bioinformatics* **Aug. 2010**, *11*, 431.

[16] M. Remmert, A. Biegert, A. Hauser, J. Söding, en, *Nat. Methods* **Dec. 2011**, *9*, 173–175.

[17] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, J. Zhang, en, *Nucleic Acids Res.* **Jan. 2017**, *45*, D158–D169.

[18] D. T. Jones, D. W. A. Buchan, D. Cozzetto, M. Pontil, *Bioinformatics* **Jan. 2012**, *28*, 184–190.

[19] M. Ekeberg, T. Hartonen, E. Aurell, *J. Comput. Phys.* **Nov. 2014**, *276*, 341–356.

[20] D. T. Jones, en, *J. Mol. Biol.* **Sept. 1999**, *292*, 195–202.

[21] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, C. Lundegaard, en, *BMC Struct. Biol.* **July 2009**, *9*, 51.

[22] S. Seemayer, M. Gruber, J. Söding, en, *Bioinformatics* **Nov. 2014**, *30*, 3128–3130.

[23] J. Andreani, J. Söding, en, *Bioinformatics* **June 2015**, *31*, 1729–1737.

[24] D. T. Jones, T. Singh, T. Kosciolek, S. Tetchner, en, *Bioinformatics* **Apr. 2015**, *31*, 999–1006.

[25] L. Kaján, T. A. Hopf, M. Kalaš, D. S. Marks, B. Rost, en, *BMC Bioinformatics* **Mar. 2014**, *15*, 85.

[26] J. Yang, R. Jang, Y. Zhang, H. B. Shen, en, *Bioinformatics* **Oct. 2013**, *29*, 2579–2587.

[27] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, en, *Methods Enzymol.* **2004**, *383*, 66–93.

[28] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, en, *Acta Crystallogr. D Biol. Crystallogr.* **Apr. 2011**, *67*, 235–242.

[29] A. Thorn, G. M. Sheldrick, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2013**, *69*, 2251–2256.

[30] S. X. Cohen, M. Ben Jelloul, F. Long, A. Vagin, P. Knipscheer, J. Lebbink, T. K. Sixma, V. S. Lamzin, G. N. Murshudov, A. Perrakis, en, *Acta Crystallogr. D Biol. Crystallogr.* **Jan. 2007**, *64*, 49–60.

[31] G. G. Krivov, M. V. Shapovalov, R. L. Dunbrack, *Proteins: Struct. Funct. Bioinf.* **2009**, *77*, 778–795.

[32] F. Simkovic, S. Ovchinnikov, D. Baker, D. J. Rigden, *IUCrJ* **May 2017**, *4*, 291–300.

[33] J. Xu, Y. Zhang, en, *Bioinformatics* **Apr. 2010**, *26*, 889–895.

[34] J. M. H. Thomas, R. M. Keegan, J. Bibby, M. D. Winn, O. Mayans, D. J. Rigden, en, *IUCrJ* **Mar. 2015**, *2*, 198–206.

[35] R. M. Keegan, J. Bibby, J. M. H. Thomas, D. Xu, Y. Zhang, O. Mayans, M. D. Winn, D. J. Rigden, en, *Acta Crystallogr. D Biol. Crystallogr.* **Feb. 2015**, *71*, 338–343.

[36] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **1997**, *268*, 209–225.

[37] W. Mao, T. Wang, W. Zhang, H. Gong, en, *BMC Bioinformatics* **Apr. 2018**, *19*, 146.