Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

CC	Correlation Coefficient
CNS	Crystallography & NMR System
MR	Molecular Replacement
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank

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

Acknowledgement: *I would like to thank Dr Saulo de Oliveira for his contributions to this chapter. He kindly provided his time and expertise to generating SAINT2 structure predictions included in the analysis presented in this chapter.*

5.1 Introduction

To-date, the recommended *ab initio* protein structure prediction protocol for optimal AMPLE performance is ROSETTA [1–4]. This recommendation is based primarily on the superiority of the decoy quality compared to other modelling algorithms, which was recently reaffirmed by the lastest CASP12 experiments [5, 6]. However, Keegan et al. [1] demonstrated that the alternative *ab initio* structure prediction protocol QUARK provides a suitable alternative to ROSETTA. Although inferior in the total number of structure solutions, QUARK decoys are a suitable ROSETTA alternative in most cases [1]. In particular, given ROSETTA's challenging installation procedure, availability limited to POSIX operating systems, requirement for large disk space and computationally expensive algorithm, QUARK's online server has been a very suitable alternative for AMPLE users.

Whilst ROSETTA and QUARK are amongst the best *ab initio* structure prediction algorithms currently available [5], other algorithms have been developed over the last two decades [e.g., 7–12]. Although most of these algorithms utelise fragment-assembly algorithms similar to ROSETTA and QUARK, their procedure to fragment selection or assembly is substantially different [7, 8]. Furthermore, predicted contact information has recently seen a spike in accuracy. This invaluable source of information is introduced differently in each protocol, and thus might have profound effects on the resulting decoy quality. In particular, physics-based algorithms relying entirely on this information are an interesting alternative to fragment-based approaches [9, 11, 12].

CONFOLD2 [13], a distance-geometry based algorithm, utelises predicted secondary structure and contact information to rapidly generate *ab initio* decoys. Unlike other algorithms, CONFOLD2's algorithm is driven almost entirely by the contact information to explore the fold space. Different contact selection thresholds are used to not limit the search space to a pre-defined selection. CONFOLD2 generates slightly inferior decoys compared to ROSETTA, however outperforms it in speed and simplicity of installation [13, 14].

FRAGFOLD [7], a fragment-assembly based algorithm, generates decoys in a similar fashion to ROSETTA and QUARK. However, FRAGFOLD does not rely on large structural libraries for fragment extraction. Instead, it provides a relatively small library of supersecondary structural fragments and short length fragments, which were extracted from high resolution protein structures. Since the generalised fragment library is shipped with FRAGFOLD and target-specific fragments extracted based on secondary structure and a sequence-based threading score, it enables fast and easy fragment library generation compared to ROSETTA [15].

SAINT2 [16], a further fragment-assembly based algorithm is substantially different to most others. SAINT2 attempts *ab initio* structure prediction sequentially, starting from either either terminus of the target sequence [16]. Furthermore, SAINT2 uses FLIB [17] for fragment picking, an algorithm shown to outperform ROSETTA's equivalent NNMake [18] in precision with very similar coverage.

Since some of these algorithms are readily available and often easier to install without the overhead of large databases for fragment picking, the work in this study focused on exploring three alternative *ab initio* structure prediction algorithms and their value in unconventional Molecular Replacement (MR). The *ab initio* structure prediction protocols CONFOLD2 [13], FRAGFOLD [7] and SAINT2 [16], were explored given their substantially different approaches to AMPLE's current default ROSETTA [19].

5.2 Materials & Methods

5.2.1 Target selection

This study was conducted using all 27 targets from the PREDICTORS dataset (**??**).

5.2.2 Contact prediction

Residue-residue contact information for 18 out of 27 targets were predicted using METAPSICOV v1.04 [20].

Secondary structure and solvent exposure were predicted using PSIPRED v4.0 [21] and SOLVPRED (shipped with METAPSICOV v1.04), respectively. The Multiple Sequence Alignment (MSA) for coevolution-based contact prediction was generated using HHBLITS [22] against the `uniprot20_2016_02` database. CCMPRED v0.3.2 [23], FREECONTACT v1.0.21 [24] and PSICOV v2.1b3 [25] were used by METAPSICOV to generate contact predictions.

METAPSICOV STAGE 1 contact predictions were used in *ab initio* structure prediction since those result in more accurate structure predictions [20].

5.2.3 *Ab initio* structure prediction

The ROSETTA 3- and 9-residue fragment libraries for each target were generated using the ROBETTA online server (`http://robetta.bakerlab.org/`). The option to "Exclude Homologues" was selected to avoid inclusion of homologous fragments. Each target sequence and its fragments were subjected to ROSETTA v2015.22.57859 [19] and 1,000 decoys per target generated with AMPLE v1.2.0 ROSETTA default options. Top-L contact pairs were used in combination with the *FADE* ROSETTA energy function, identical to the benchmark outlined in Chapter 5.

The CONFOLD2 decoys were generated using CONFOLD2 v2.0 [13] and Crystallography & NMR System (CNS) v1.3 [26]. Default parameters were used except for the number of decoys per run, which was increased from 20 to 25 with `-mcount 25`. This resulted in 40 seperated runs differing only in the number contact pairs used, which was increased by 0.1 from $L/10$ to $4L$.

The FRAGFOLD decoys were generated using FRAGFOLD v4.80 [7] with default options. Homologous fragments were removed from the shipped library by excluding all entries with Protein Data Bank (PDB) identifiers identical to those retrieved from the ROBETTA server. All contact pairs were used according to FRAGFOLD's internal protocol.

The fragment libraries for SAINT2 were generated using FLIB [17], which generates on average 30 fragments per target position that are 6 to 20 residues long. Homologous fragments were removed from the final fragment list using the PDB identifiers obtained from the ROBETTA online server. The secondary structure prediction and solvent accessibility scores were also obtained from the ROBETTA server. SAINT2 was used for decoy generation, and 1,000 decoys generated per target. The procedure and parameters were identical to those described in Supplementary Information (p. 16) by Oliveira et al. [16].

5.2.4 Molecular Replacement

All decoy sets were subjected to AMPLE v1.2.0 and CCP4 v7.0.28. Default options were chosen with the following 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 from AMPLE's initial mode of operation [4] was shown to be advantageous in most cases by Thomas [27], and thus trialled in this context. Each MR run was assessed using the criteria defined in **??**.

5.3 Results

The purpose of this study was to investigate the usefulness of alternative *ab initio* structure prediction in AMPLE. Three promising leads widely used in the *ab initio* modelling experiments were examined and compared against AMPLE's current algorithm of choice. This led to a direct comparison of the algorithms ROSETTA [19], CONFOLD2 [13], FRAGFOLD [7] and SAINT2 [16]. All four algorithms have recently seen great improvements through the use of residue-residue contact information, which predicted using the METAPSICOV [20] algorithm.

5.3.1 Alignment depth and contact prediction precision

The first step in this study was the prediction of residue-residue contacts using the metapredictor METAPSICOV for 18 targets in the PREDICTORS dataset [20]. Since we attempted to test each of the structure prediction boundaries in extreme cases, a variety of targets with different alignment depths were chosen. The alignment depth of METAPSICOV-generated HHBLITS alignments ranges from 6 to 6,186 across all targets (Fig. 5.1). Five targets contain alignments with a depth of less than 200, a rough threshold to indicate suitability of a MSA for coevolution-based contact prediction ([28]). A further six targets contain at least 200 and less than 1000 sufficiently-diverse sequences, whilst for the remaining 16 targets the MSA depth is at least 1,000.

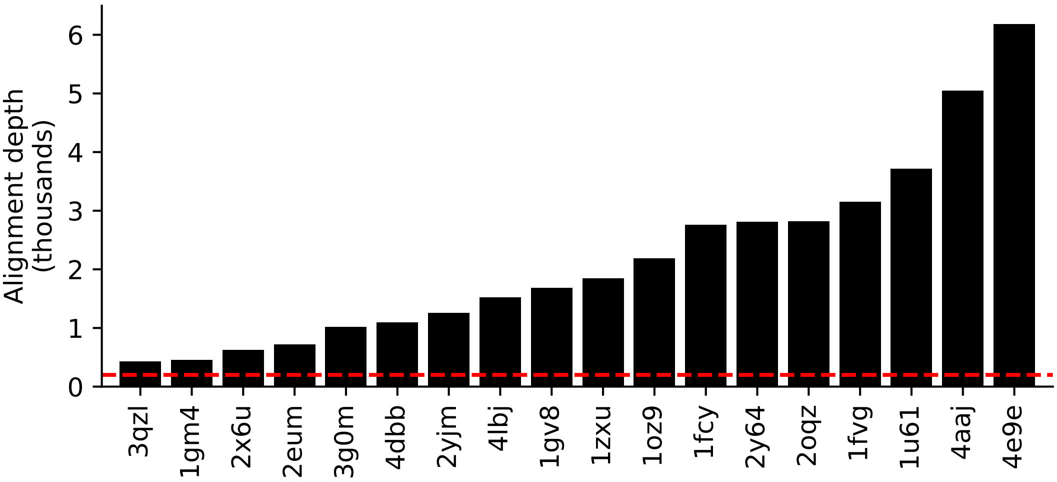

Figure 5.1: Distribution of HHBLITS alignment depth for subset of targets in the PREDICTORS dataset. Red line indicates alignment depth threshold of 200 sequences.

Coevolution-based contact predictors rely heavily on the alignment depth for accurate contact predictions. In this work, these findings are further confirmed. Sequence alignments with depths of less than 1000 sequences highlight lower precision scores across a number of cutoffs compared to those with deeper alignments (Fig. 5.2). Given the alignment depths and top-L contact predictions, a significant positive correlation between the two is found (Spearman's $\rho = 0.57$, p-value < 0.02). A moving average analysis shows that those contact predictions based on alignments with more than 1000 effective sequences yield better precision scores of at least 0.09 units up to 0.34 units. The difference between the two moving average curves highlights that the difference is greater at lower cutoff values, i.e. only the very best contacts are included in the selection. This difference declines more drastically for targets with deeper alignments (Fig. 5.2).

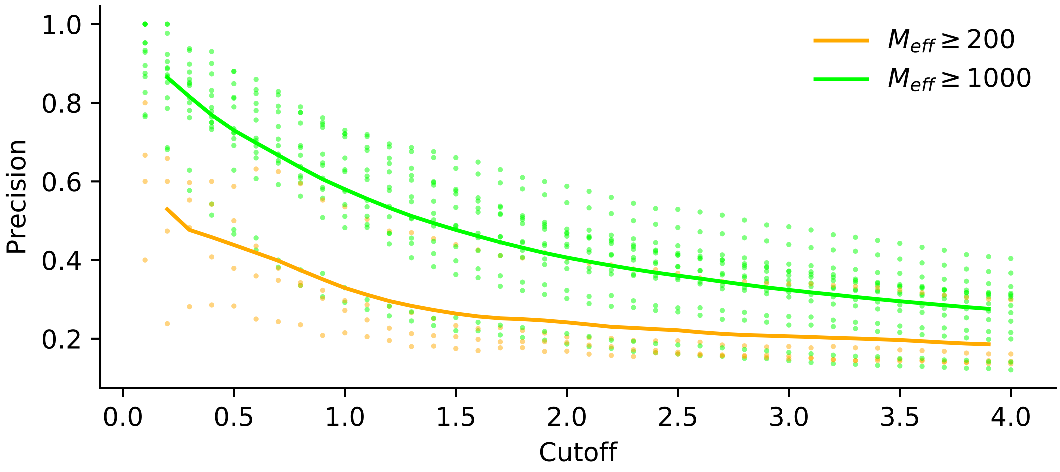

Figure 5.2: Contact precision analysis for numerous contact selection cutoffs for targets with alignment depths of more than 200 and more than 1000 sequences. Lines indicate moving averages for both categories with a window size of three residues. M_{eff} refers to the alignment depth (number of effective sequences).

5.3.2 Comparison of decoy quality

One main interest of the work presented in this chapter is the comparison of the quality of decoys predicted with four *ab initio* structure prediction algorithms. To-date, no such direct comparison exists on the same dataset, and thus might provide direct insights into the performance of each.

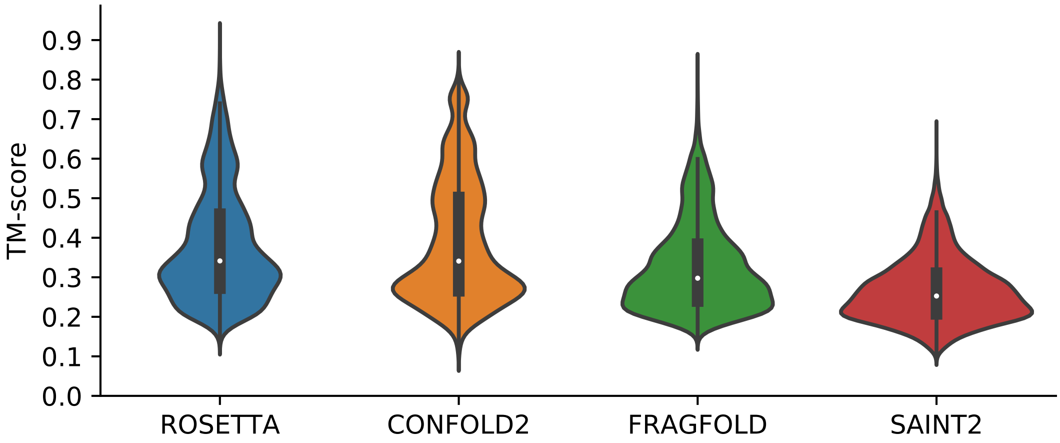

Figure 5.3: Kernel Density Estimate of decoy TM-scores for four different *ab initio* structure prediction algorithms, namely ROSETTA, CONFOLD2, FRAGFOLD and SAINT2. CONFOLD2 contains 9,000 less decoys than the remaining algorithms (for further details refer to Section 5.2.3).

An initial comparison of overall performance highlights that ROSETTA generates the highest quality decoys (Fig. 5.3). Across all modelling algorithms the distribution of TM-scores right-skewed, which indicates a higher proportion of non-native-like folds within the sets. A TM-score quantile evaluation of each decoy set by algorithm shows that ROSETTA and CONFOLD2 contain only a single set with a lower quantile of less than 0.2 TM-score units. In comparison, FRAGFOLD predicted three and SAINT2 eight decoy sets with a lower quantile of less than the aforementioned threshold. In comparison, ROSETTA, CONFOLD2 and FRAGFOLD predicted six, seven and five decoy sets with upper quantiles greater than 0.5 TM-score units, whilst SAINT2 predicted zero.

A direct comparison of the methods by median TM-score of each contact-assisted decoy set reaffirms ROSETTA's performance in predicting *ab initio* decoys accurately. Across 18 targets, ROSETTA decoy sets contain the best median TM-score for 11 targets (CONFOLD2 for remaining seven targets). This is further strengthened when comparing the top-1 decoy for which ROSETTA predicts the best in 13 cases (CONFOLD2 in three cases, FRAGFOLD and SAINT2 in one) (Fig. 5.4).

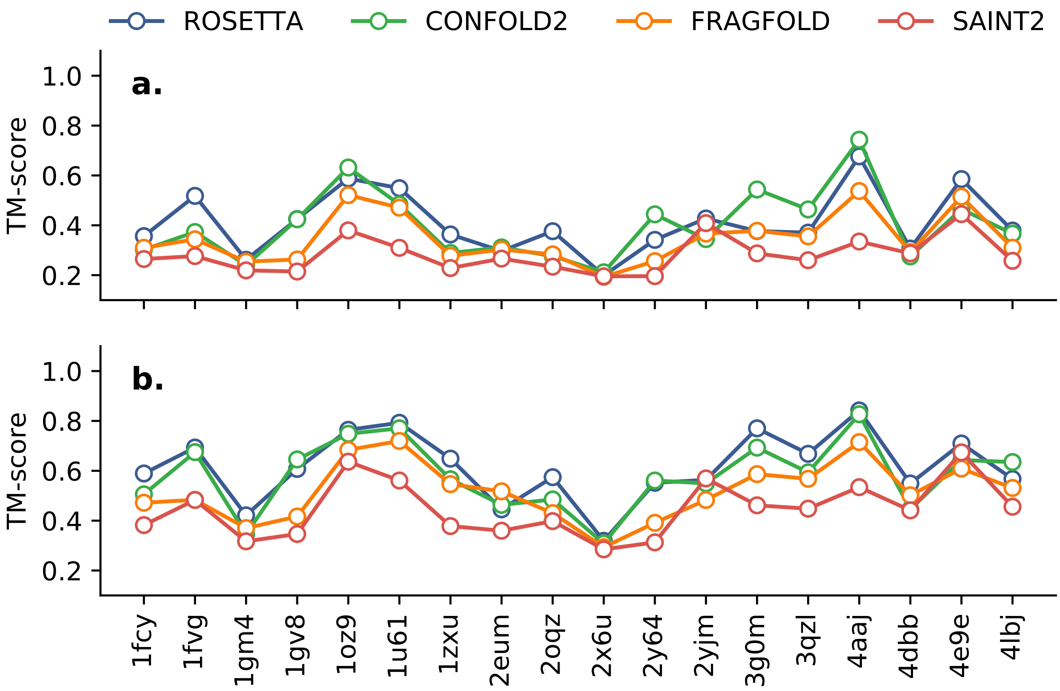

Figure 5.4: Per-target TM-score analysis for targets modelled with contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

Abriata et al. [5] recently attributed the success in the CASP12 experiments to improved accuracy of coevolution-based contact predictors and the availability of many more sequence homologs than ever before. Thus, it is of great interest to explore the structure prediction algorithms in this study with regards to their dependence on the availability of sequence homologs and precise contact predictions.

The results obtained in this study further support the conclusions made by Abriata et al. [5] but only for the ROSETTA algorithm. A Spearman's rank-order Correlation Coefficient (CC) analysis of alignment depth and median TM-score shows a significant positive correlation for ROSETTA-generated decoy sets (Spearman's $\rho = 0.68$, $p < 0.01$). This positive correlation is also found for ROSETTA-generated decoy sets with regards to their top-L precision and median TM-score (Spearman's $\rho = 0.61$, $p < 0.01$). All other modelling algorithms do not show a significant correlation, although better decoy sets are generally obtained with greater alignment depths and more precise top-L contacts (Fig. 5.5). Furthermore, the sample size for each correlation analysis was small ($n = 18$), and thus further test cases are required for a more confident inference.

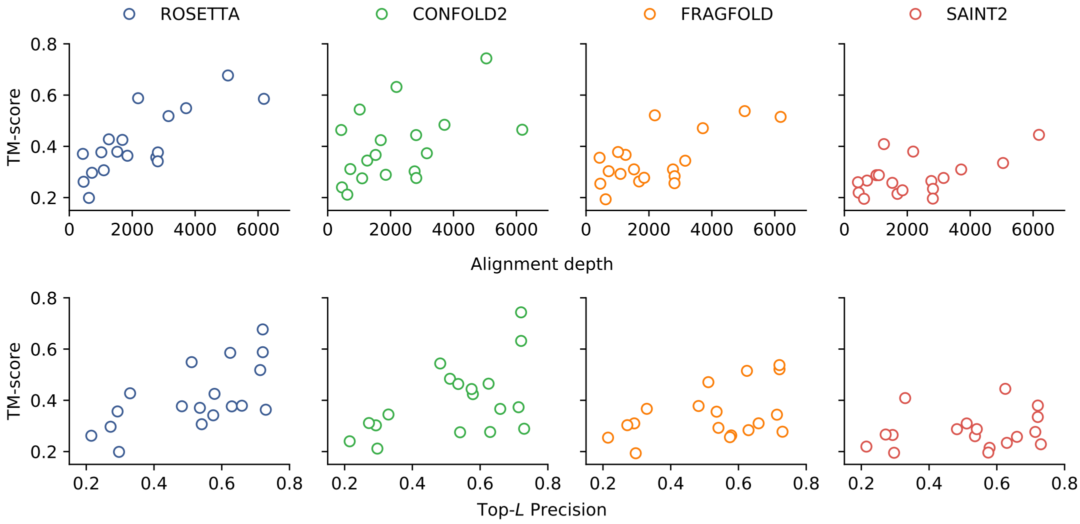

Figure 5.5: Analysis of median TM-scores of the contact-based decoy sets and their dependence on alignment depth and top-L precision.

Beyond the use of contact information, parts of this study explored the performance of ROSETTA, FRAGFOLD and SAINT2 when no contact information is provided as distance restraints in *ab initio* structure prediction (CONFOLD2 requires contact information, and thus was excluded). ROSETTA performs best for seven of the nine contact-free decoy sets based on median TM-score of the entire decoy set and the TM-score of the top-1 decoy (Fig. 5.6). However, the difference is marginal for the majority of cases. The median values for eight ROSETTA and FRAGFOLD decoy sets differ by less than 0.10 TM-score units (seven ROSETTA and SAINT2 sets by less than 0.10 units). Furthermore, the top-1 decoys for only three targets differ greatly between the modelling algorithms, whilst the rest is near identical between algorithms (Fig. 5.6).

The top decoy predicted by ROSETTA and SAINT2 based on the sequence of the FAT domain of focal adhesion kinase (PDB ID: 1k40) differs by 0.35 TM-score units. More significantly though, the top decoy predicted by ROSETTA for the outer surface protein A (PDB ID: 2ol8) is considered native-like (TM-score = 0.59), whilst the FRAGFOLD (TM-score = 0.35) and SAINT2 (TM-score = 0.24) counterparts predict incorrect folds. A near-identical scenario applies to the top decoys of the Hypothetical protein PF0907 (PDB ID: 4pgo) (ROSETTA TM-score = 0.68; FRAGFOLD TM-score = 0.27; SAINT2 TM-score = 0.39).

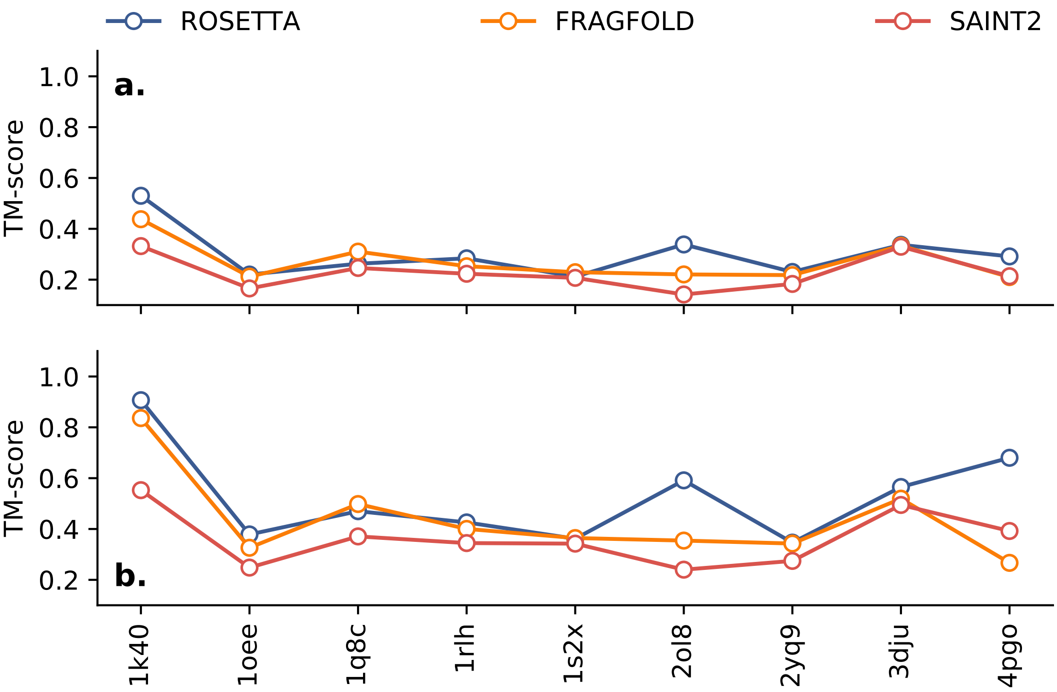

Figure 5.6: Per-target TM-score analysis for targets modelled without contact information and four separate *ab initio* structure prediction algorithms. Analysis is subdivided by (a) median TM-score of all decoys in each set and (b) TM-score of the top-1 decoy in each set.

An analysis of the modelling results by target fold shows that all-α and mixed α-β target folds are less challenging to predict than all-β targets (Fig. 5.7). The multimodal distributions of all-α and mixed α-β target decoys predicted by ROSETTA spans from 0.10 TM-score units to 0.80. In comparison, the roughly normal distribution for all-β by the same algorithm centres at 0.32 TM-score units (s.d.=0.08 TM-score units). Similarly, FRAGFOLD decoys show a more spread distribution of decoys for all-α and mixed α-β decoys compared to all-β. Although the multimodal distribution of TM-scores for all-β target decoys might indicate better performance for some targets, it is most certainly misleading since three all-β targets are missing from the dataset in comparison to the other algorithms. Lastly, the distributions of TM-scores for either fold class of SAINT2 decoys appear more similar than the others indicating less difference between the fold classes. However, similarly to the ROSETTA decoys the all-β distribution appears normal whilst the other two are right-skewed highlighting some more accurate decoys in the overall set.

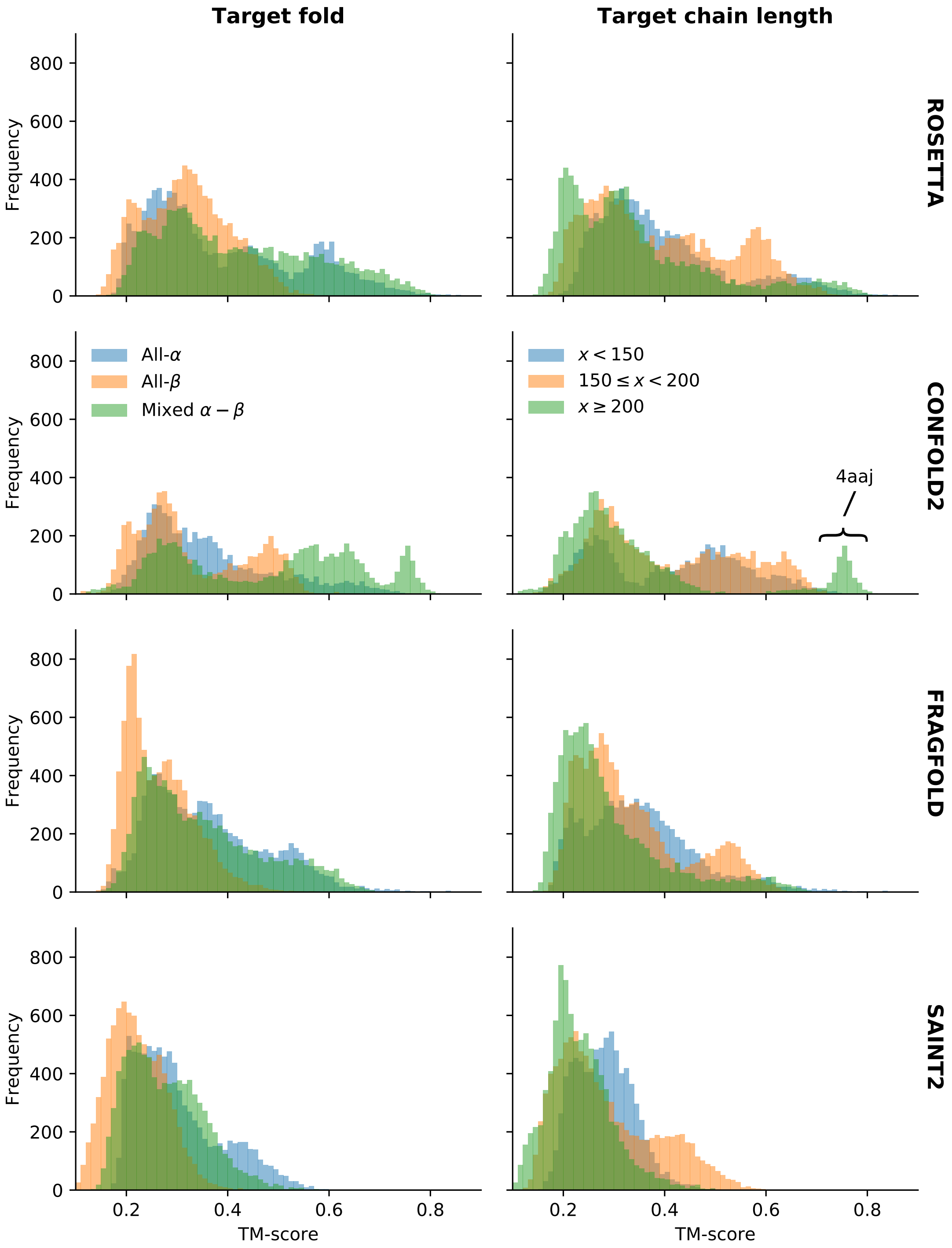

Figure 5.7: Distribution of decoy TM-scores by fold, chain length and algorithm.

A further subdivision of all target decoys is by target chain length. At the stage of target selection, three main bins were defined from which targets were randomly sampled (see **??**). These bins were defined with target chain length edges of 150 and 200 creating three bins: $0 < x < 150$ & $150 \leq x < 200$ & $x \geq 200$ (x refers to the target chain length). A grouping of the decoy TM-scores by algorithm and target chain length indicates little difference in modelling difficulty (Fig. 5.7). Each of the modelling algorithms shows the largest spread for targets with chain lengths in the bin $150 \leq x < 200$. Surprisingly, only FRAGFOLD and SAINT2 performed better for targets in the smallest bin size whilst CONFOLD2 found those targets most challenging. CONFOLD2 also generated the best decoys for one of the largest targets in the dataset. The set of CONFOLD2 decoys for N-(5-phosphoribosyl)anthranilate isomerase (PDB ID: 4aaj) have a median TM-score of 0.74 units. ROSETTA decoys show a comparable median TM-score of 0.68; however, FRAGFOLD (median TM-score=0.54) and SAINT2 (median TM-score=0.33) were unable to generate decoys of similarly high quality.

5.3.3 Molecular Replacement

Each *ab initio* structure prediction algorithm generated decoy sets, which are sufficient for structure solution (Fig. 5.8). ROSETTA and SAINT2 decoy sets led to five structure solutions each, whilst FRAGFOLD decoys to four and CONFOLD2 decoys to two. All four algorithms predicted decoys of good enough quality to solve the structures of the Hypothetical protein AQ_1354 (PDB ID: 1oz9) and Putative Ribonuclease III (PDB ID: 1u61), although SAINT2-based AMPLE search models yielded the highest ratio of successful search models in both cases (Fig. 5.8). Besides these two targets, little consensus exists amongst the targets for which structure solutions were obtained.

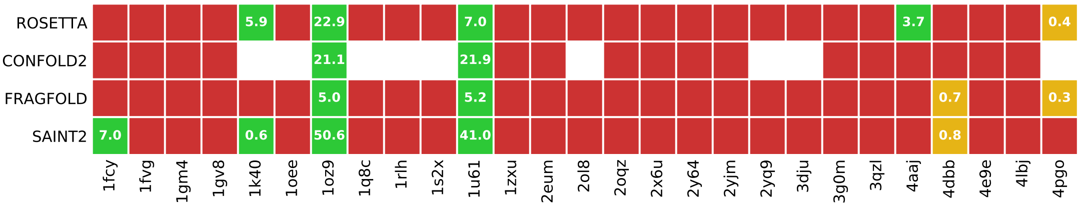

Figure 5.8: Summary of MR success with AMPLE ensemble search models. Search models are based on decoy sets generated with different *ab initio* structure prediction protocols. The colour coding indicates structure solution: no solution (red), one solution (orange), more than one solution (green). The number in cells with at least one solution states the percentage successful search models.

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

Appendix A

Appendix

Bibliography

[1] 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.

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

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

[5] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, M. Dal Peraro, en, *Proteins* **Mar. 2018**, *86 Suppl 1*, 97–112.

[6] S. Ovchinnikov, H. Park, D. E. Kim, F. Dimaio, D. Baker, en, *Proteins: Struct. Funct. Bioinf.* **Sept. 2017**, DOI `10.1002/prot.25390`.

[7] D. T. Jones, en, *Proteins: Structure Function and Genetics* **2001**, *Suppl 5*, 127–132.

[8] J. J. Ellis, F. P. E. Huard, C. M. Deane, S. Srivastava, G. R. Wood, en, *BMC Bioinformatics* **Apr. 2010**, *11*, 172.

[9] B. Adhikari, D. Bhattacharya, R. Cao, J. Cheng, en, *Proteins: Struct. Funct. Bioinf.* **Aug. 2015**, *83*, 1436–1449.

[10] D. Xu, Y. Zhang, en, *Proteins* **July 2012**, *80*, 1715–1735.

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

[12] S. Wang, W. Li, R. Zhang, S. Liu, J. Xu, en, *Nucleic Acids Res.* **July 2016**, *44*, W361–6.

[13] B. Adhikari, J. Cheng, en, *BMC Bioinformatics* **Jan. 2018**, *19*, 22.

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

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

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

[17] S. H. P. de Oliveira, J. Shi, C. M. Deane, en, *PLoS One* **Apr. 2015**, *10*, e0123998.

[18] D. Gront, D. W. Kulp, R. M. Vernon, C. E. M. Strauss, D. Baker, en, *PLoS One* **Aug. 2011**, *6*, e23294.

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

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

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

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

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

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

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

[26] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J Kuszewski, M Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T Simonson, G. L. Warren, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 1998**, *54*, 905–921.

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

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