# Contents

# List of Figures

# List of Tables

# List of Equations

# List of Abbreviations

| | |
|---|---|
| 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

## 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 struc-

tural 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.

### 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 SAINT2 decoys ... *wait for Saulo.* Saulo did not generate 1,000 decoys per target but slightly less ...

### 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

# 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**.