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

Felix Šimkovic

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

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

Felix Šimkovic

This thesis is concerned with the application of predicted residue contacts in *ab initio* protein structure prediction and Molecular Replacement.

The initial work in this thesis explored the use of predicted residue contacts to improve *ab initio* protein structure predictions, which were used to generate ensemble search models for Molecular Replacement in AMPLE. The results proved highly encouraging. Five additional targets were tractable where previous AMPLE attempts would have been unable to achieve structure solution. In particular, the improved decoy quality appeared to be the main reason for the extended target tractability.

Following on from the initial proof-of-concept study, different contact prediction algorithms and ROSETTA energy functions were trialled to identify the optimal strategy to generate the most accurate decoys for unconventional Molecular Replacement in AMPLE. The findings showed supported previous claims that METAPSICOV produces the most precise contact predictions. Furthermore, the ROSETTA `FADE` energy function outperforms the `SIGMOID` function. Nevertheless, results also demonstrate that the most accurate structure predictions do not achieve the most Molecular Replacement structure solutions.

Beyond different contact prediction algorithms and ROSETTA energy functions, many alternative fragment-based and fragment-independent protein structure prediction algorithms exist. In this chapter, results highlighted that ROSETTA remains the optimal structure prediction algorithm in combination with AMPLE to maximise structure solutions.

The most accurate protein structure predictions may not be processed optimally in AMPLE. Thus, it is important to explore alternative ensembling strategies when more accurate contact-assisted decoys are used in AMPLE. The findings in this chapter demonstrated the successful application of estimating decoy quality by the satisfaction of long-range contact predictions used initially to restrain the folding procedure. Excluding the decoys that satisfy the least long-range contacts provided further structure solutions previously intractable.

Lastly, contact-driven selection of supersecondary structure elements or subfolds during fragment picking was explored to identify suitable search models for unconventional Molecular Replacement. Preliminary results of this approach strongly hint towards a potential new approach. Two out of four protein targets were solved with

fragments extracted from sequence-independent protein targets, which crucially satisfied many predicted residue contacts.

Acknowledgements

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

MR Molecular Replacement

Chapter 1

Introduction

Chapter 2

Materials & Methods

Chapter 3

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

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

8.1 Conclusion

The successful disentanglement of direct and indirect residue contacts in contact prediction revolutionised many aspects of Structural Bioinformatics research [1]. Successful applications of contact information range from accurately defining domain boundaries [2] to identifying druggable protein-protein interfaces [3]. Although many such applications have been highlighted over the last few years [1], few concerned the topic of Molecular Replacement (MR) in X-ray crystallography. In this thesis, work was presented that made the first attempts to apply contact information to explore some of its applications in MR.

The use of contact information in *ab initio* protein structure prediction allowed researchers to predict the structure of many previously unknown protein folds based on their sequence alone [e.g., 4–12]. The major benefit of adding such information was to reduce the conformational search space, which allowed more challenging folds to be sampled correctly. Work presented in Chapters 3 to 5 further confirms such findings. More importantly, the presented results highlight that the modelling algorithm ROSETTA is very sensitive to the way contact information is introduced into the ROSETTA folding protocol. Two important examples include the up-weighting of β-strand contacts and the choice of energy function used to "reward" satisfied contacts. Furthermore, work in Chapter 5 highlights that fragment-based structure prediction algorithms may no longer be essential for accurate structure prediction. CONFOLD2, a fragment-independent algorithm, predicts the protein structure using secondary structure and contact information alone, which provided models of comparable accuracy to the state-of-the-art ROSETTA. Nevertheless, further research is required to establish the optimal routine to process CONFOLD2 decoys since AMPLE's default routine cannot generate ensemble search models sufficient for MR solutions.

Beyond the prediction of protein structures, a major focus of the presented research centred on the benefit of such improved structure predictions in unconventional MR. In line with prior expectations, better structure predictions yield more MR structure solutions. In particular, previous weaknesses of the AMPLE approach — a target's chain length and fold — can be partially overcome with contact-guided structure predictions. Some examples for which structure solutions were obtained exceed 200 residues in chain length, whilst many others contain large proportions of β-structure. Nevertheless, simply adding contact information to *ab initio* protein structure prediction is not sufficient to solve all trialled targets. In part, this limitation results from a lack of precision of contact information for some targets, since it depends significantly on the availability of divergent homologous sequences. Furthermore, further research is required to be address new limitations in AMPLE resulting from suboptimal processing of much more accurate structure predictions. One approach, outlined in Chapter 6, explored the incorporation of contact information in the AMPLE processing pipeline to address the latter issue. Contact information was used to estimate the similarity of a predicted

decoy to its native structure, by means of scoring its long-range contact satisfaction [6, 13, 14]. Exclusion of the worst decoys by this metric prior to clustering allowed more fine-grain sampling in AMPLE, which turned unsuccessful decoy sets into ones with which the native structure could by solved by MR. However, key examples presented in Chapters 3 to 5 also highlight that further developments in MR-related software are required to enable the automatic detection and subsequent processing of AMPLE ensemble search models that are correctly placed but by current metrics cannot be identified as correct structure solutions.

A further topic of research concerned the use of supersecondary structure elements or subfolds as MR search models. The default mode in AMPLE currently relies on computationally expensive *ab initio* structure predictions. Since contact predictions have reached sufficient quality for protein families with many known sequences, such information could be used to identify matching subfolds in other, unrelated protein structures. In Chapter 7, a new hybrid approach demonstrated the successful implementation of such an idea. Although imperfect at this stage, several examples highlighted the successful identification of such subfolds and subsequently successful MR structure solution. Tied to this idea may also be recent research that attempts to identify subfolds by means of matching a predicted contact map to those extracted from protein structures [11, 15].

8.2 Outlook

In this thesis the first applications of predicted contact information in MR were presented. Despite the already promising results, this area of research is still in its infancy and a great number of potentially promising routes remain unexplored [1].

Earlier studies by Rigden [16] and Sadowski [2] demonstrated the successful application of predicted residue contacts to identify domain boundaries. Although unexplored to-date, precise domain boundary predictions could be applied for better domain boundary definitions in *ab initio* structure prediction to avoid sampling of terminal loops and linkers, and thus improve protein structure prediction quality. Furthermore, contact information was used to improve the AMPLE ensemble-generation pipeline with respect to identifying poorly predicted decoys. However, the AMPLE ensemble-generation pipeline might additionally benefit from contact information to drive the truncation procedure. For example, contact data could be used to rank individual residues by their contribution to a contact network, similar to [17], and truncation driven by the rank order or a hybrid score, which also includes the structural variance. Additionally, contact prediction might be used in the context of identifying alternative conformational states [18–22], which AMPLE could exploit to identify structurally conserved residues between both states and truncate to this conserved core, or attempt remodelling after successful disentanglement of state-dependent contact pairs and try both

conformations separately as ensemble search models. Simkovic et al. [1] outlined many further such applications of predicted contact information in the field of Structural Biology. Ultimately, the precision of contact information improves daily with the increasing depth of sequence databases, thus enabling an ever-increasing number of applications with more precise outcomes. Furthermore, many more research groups start to identify the value in using contact information in their own studies, and by means of pushing the boundaries new tools and applications are most likely going to emerge.

Despite the vast space of unexplored applications, predicted residue contacts with perfect precision may never solve all current or future challenges in unconventional MR. Despite the ability to limit the conformational space search in *ab initio* protein structure prediction greatly, sampling of larger protein targets will always remain difficult unless energy functions and force fields become true representations of all properties found *in vivo*. Furthermore, computational resources need to expand to allow many more sampling steps. Additionally, many protein targets exist in multiple conformations. Energy functions in fragment-based *ab initio* protein structure prediction may always favour one such conformation over all others, which may make conventional or unconventional MR very challenging.

Beyond limitations in Bioinformatics software to facilitate the generation of search models for unconventional MR, limits are also posed in the procedure of MR itself. The most prominent limitation may be the resolution of the experimental data, and the proportion of the search model compared to the content of the crystallographic unit cell. SHELXE [23], a popular and powerful algorithm to perform density modification and main-chain autotracing, is heavily limited by a lower resolution limit of 2.5Å. Thus, MR pipelines, such as AMPLE [24] or ARCIMBOLDO [25], may not be able to detect correctly placed search models due to the current dependence on associated software metrics. Furthermore, MR is extremely challenging, if not impossible, when the scattering matter, i.e. a correctly placed search model, is particularly small in relation to the asymmetric unit content whilst the resolution of the experimental data is low.

Finally, AMPLE and similar unconventional MR software pipelines try to enable MR when one or more sufficiently similar structures are unavailable to derive the essential phase information. Despite the relative rarity of such a scenario [26], it is essential to provide routes to structure solution when conventional approaches fail since those cases may often provide novel or unexpected findings. The current toolbox for unconventional MR provides idealised fragments [25, 27, 28], supersecondary structure motifs [29], and ensemble search models extracted from a diversity of different starting structures [24, 30, 31]. The former two are usually target-independent, and thus limited by structural deviations between selected search probes and the target. In comparison, the latter depend much more on accurate and target-specific starting structures but provide a great alternative in lower resolution cases or scenarios whereby larger search

models are required. Therefore, unconventional MR requires a diversity of approaches to attempt structure solutions of the most challenging cases. AMPLE and its improvements through predicted residue contacts should therefore be considered an important tool in this set of approaches.

Appendix A

Appendix

Bibliography

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

[2] M. I. Sadowski, en, *Proteins: Struct. Funct. Bioinf.* **Feb. 2013**, *81*, 253–260.

[3] F. Bai, F. Morcos, R. R. Cheng, H. Jiang, J. N. Onuchic, en, *Proceedings of the National Academy of Sciences* **Dec. 2016**, *113*, E8051–E8058.

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

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

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

[7] 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.

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

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

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

[11] 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.

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

[13] S. H. P. De Oliveira, J. Shi, C. M. Deane, en, *Bioinformatics* **Feb. 2017**, *33*, 373–381.

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

[15] D. W. A. Buchan, D. T. Jones, *Bioinformatics* **Sept. 2017**, *33*, 2684–2690.

[16] D. J. Rigden, en, *Protein Eng.* **Feb. 2002**, *15*, 65–77.

[17] D. J. Parente, J. C. J. Ray, L. Swint-Kruse, en, *Proteins* **Dec. 2015**, *83*, 2293–2306.

[18] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, D. S. Marks, en, *Cell* **June 2012**, *149*, 1607–1621.

[19] B. Jana, F. Morcos, J. N. Onuchic, *Phys. Chem. Chem. Phys.* **2014**, *16*, 6496–6507.

[20] P. Sfriso, M. Duran-Frigola, R. Mosca, A. Emperador, P. Aloy, M. Orozco, en, *Structure* **Jan. 2016**, *24*, 116–126.

[21] F. Morcos, B. Jana, T. Hwa, J. N. Onuchic, en, *Proc. Natl. Acad. Sci. U. S. A.* **Dec. 2013**, *110*, 20533–20538.

[22] L. Sutto, S. Marsili, A. Valencia, F. L. Gervasio, en, *Proc. Natl. Acad. Sci. U. S. A.* **Nov. 2015**, *112*, 13567–13572.

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

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

[25] M. Sammito, C. Millán, D. Frieske, E. Rodríguez-Freire, R. J. Borges, I. Usón, en, *Acta Crystallogr. D Biol. Crystallogr.* **Sept. 2015**, *71*, 1921–1930.

[26] J. M. Chandonia, N. K. Fox, S. E. Brenner, en, *J. Mol. Biol.* **Feb. 2017**, *429*, 348–355.

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

[28] H. T. Jenkins, *Acta Crystallographica Section D Structural Biology* **Mar. 2018**, *74*, 205–214.

[29] M. Sammito, C. Millán, D. D. Rodríguez, I. M. De Ilarduya, K. Meindl, I. De Marino, G. Petrillo, R. M. Buey, J. M. De Pereda, K. Zeth, G. M. Sheldrick, I. Usón, en, *Nat. Methods* **Nov. 2013**, *10*, 1099–1104.

[30] 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.

[31] J. Bibby, R. M. Keegan, O. Mayans, M. D. Winn, D. J. Rigden, en, *Acta Crystallogr. D Biol. Crystallogr.* **Nov. 2013**, *69*, 2194–2201.