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

Institute of Integrative Biology
University of Liverpool
United Kingdom

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

MR	Molecular Replacement
MSA	Multiple Sequence Alignment

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

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 prediction of larger and more β-rich protein structures [e.g., 3–11]. In AMPLE, these two domains are the major limitation for a more successful approach [12]. This typically results in user success being limited to small globular and primarily α-helical folds, or time- and resource-demanding attempts most likely going to be unsuccessful for larger targets

With the advent of contact information, is has thus become essential to identify the extend to which this invaluable bit of information is going to help AMPLE users in the future.

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 transmembrain protein targets. For details in 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 [13]. In summary, four Multiple Sequence Alignment (MSA)s were generated with JACKHMMER v3.1b2 [14] against the `uniref100` v2015-10 database and HHBLITS v2.0.15 [15] against the `uniprot20` v2013-03 database [16] at E-value cutoffs of 10^{-40}, 10^{-10}, 10^{-4} and 1. Each MSA was analysed with PSICOV v2.13b3 [17] and PLMDCA v2 [18] to produce 16 individual contact predictions. All 16 predictions and per-target PSIPRED v3 [19] secondary structure prediction, NET-

SURFP v1.0 [20] solvent accessibility information and HHBLITS v2.0.15 [15] sequence profile were provided to the PCONSC2 deep learning algorithm [13] 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 [21] and reduced to β-sheet contact pairs using the CCMPRED-specific filtering protocol BBCONTACTS v1.0 [22]. Each MSA for CCMPRED contact prediction was obtained using HHBLITS v2.0.15 [15]. This entailed two sequence search iterations with an E-value cutoff of 10^{-3} against the `uniprot20` v2013-03 database [16] and filtering to 90% sequence identity using HHFILTER v2.0.15 [15] 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 [15] against `uniprot20` v2016-02 database [16]. Contact predictions for each transmembrane target were obtained using the metapredictor METAPSICOV v1.04 [23], which in turn used the contact prediction algorithms CCMPRED v0.3.2 [21], FREECONTACT v1.0.21 [24] and PSICOV v2.1b3 [17]. Additionally, a set of contacts was also generated using the MEMBRAIN server v2015-03-15 [25].

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 to above, 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 homologous fragments were included in the fragment libraries, the `-nohoms` flag was set. 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 subjected to the ROSETTA `AbinitioRelax` protocol [26] 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 [27], with the exception of SHELXE v2014/14 [28] and ARP/wARP v7.5 [29].

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 [30]. 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 [27].

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

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 [31]. Successful applications of contact information range from accurately defining domain boundaries [32] to identifying druggable protein-protein interfaces [33]. Although many such applications have been highlighted over the last few years [31], few concerned the topic of MR in X-ray crystallography. In this thesis, work was presented that made first attempts to apply contact information to explore some of its application spectrum 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., 3–11]. 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 confirm 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.

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 weakpoints of the AMPLE routine — the target chain length and fold — can successfully be addressed 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. Thus, further research, outlined in Chapter 6, explored one way of incorporating contact information in the AMPLE processing pipeline. 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. 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 from which the native structure can be eludicated.

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.

8.2 Outlook

In this thesis 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. Earlier studies by Rigden [34] and Sadowski [32] demonstrated the successful application of 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 pipeline with respect to excluding poorly predicted decoys. However, the AMPLE ensemble generation pipeline might additionally benefit from contact information to aid the driving of the truncation procedure. For example, contact data could be used to rank individual residues by their contribution to a contact network, similar to [35], and truncation driven by the rank order. Additionally, contact prediction might be used in the context of identifying alternative conformational states [36–40], which AMPLE could exploit to identify 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. Besides the application of contact data in protein structure prediction, other alternatives need to be considered too. Recently, first tools were developed to match predicted contact maps to ones extracted from protein structures [10, 41]. It might be of interest to investigate how search models, such as distant homologs, could be identified by sequence searches aided with contact map matching.

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

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

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

[16] 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.

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

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

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

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

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

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

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

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

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

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

[27] 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.

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

[29] 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.

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

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

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

[33] 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.

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

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

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

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

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

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

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

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