# 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 |
| PDB | Protein Data Bank |
| TM-score | Template-Modelling score |

# 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 **??**.

## 3.3 Results

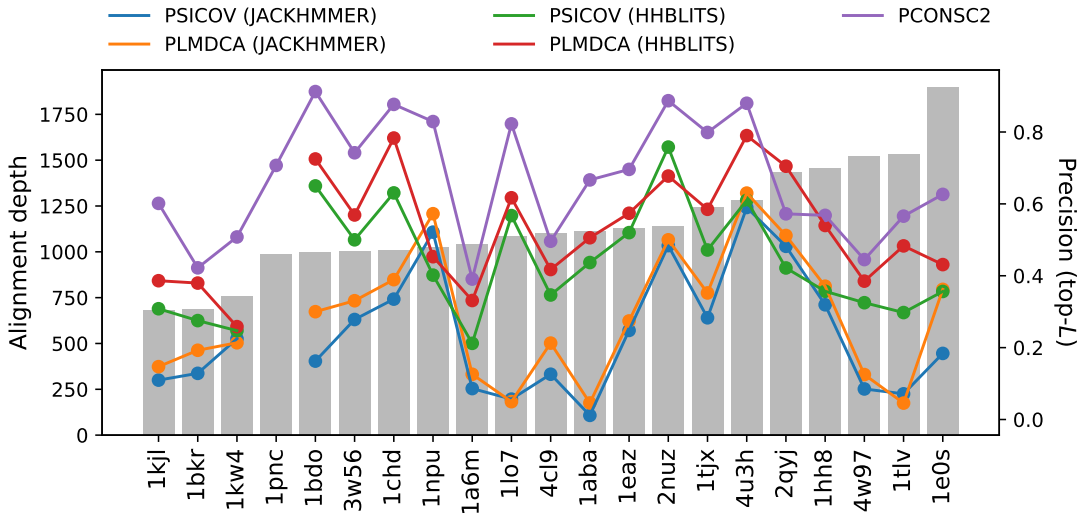
In this study, the application of residue-residue contact predictions to *ab initio* protein structure prediction and subsequently MR was investigated. This proof-of-concept work is based on two datasets covering a range of globular and transmembrane protein targets. At the time of conducting this study, state-of-the-art contact prediction algorithms were applied to obtain the best possible contact predictions to identify the extend of pushing the boundaries previously incurred in AMPLE studies [12].

### 3.3.1 Residue-residue contact prediction

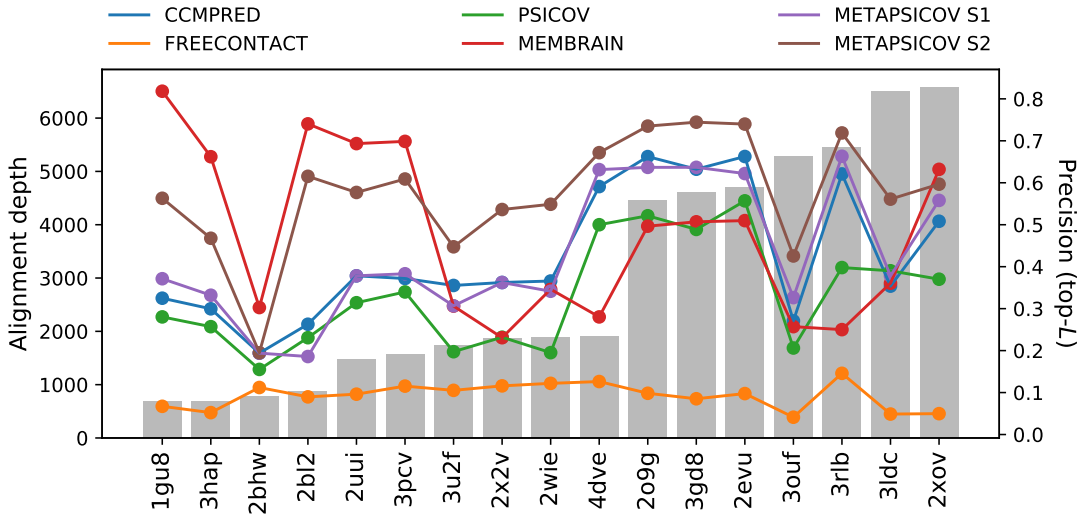
Accurate coevolution-based residue-residue contact prediction highly depends on the availability of many divergent homologous sequences. As such, it is important to validate that the selected targets in this study satisfy such requirement.

The depth of MSAs obtained for each target sequence suggests that sufficient num-

bers of divergent homologous sequences are available. Across all globular targets, the minimum alignment depth is obtained for Galectin-3 domain (Protein Data Bank (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 [31]. 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 utelise greater alignment depths to correct contact pairs (Fig. 3.1).

PCONSC2 — a metapredictor using numerous starting alignments and two contact predictors — outperforms its individual parts for almost all globular targets (Fig. 3.1a). Despite only a fraction of all generated contact predictions illustrated in Fig. 3.1a, the pattern translates across all 16 predictions per target, which suggests that precision greatly depends on the tool to generate the input alignment. 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 utelises 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 $\Delta_{TMscore}$ of less than 0.05 (Fig. 3.1b). This similarity traverses not to METAPSICOV STAGE 2 predictions with only a single target showing such similar Template-Modelling score (TM-score) values (Fig. 3.1b). Amongst the three raw predictors used by METAPSICOV, FREECONTACT performs by far the worst with a mean TM-score of 0.09 across all transmembrane targets. PSICOV shows similar trend to CCMPRED when assessed by target, which results in an mean absolute $\Delta_{TMscore}$ of 0.10.

### 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 TM-score improvements of at least 0.05 when comparing contact-guided PCONSC2 decoys with simple ROSETTA decoys (Fig. 3.2). The greatest improvement over 1,000 decoys was achieved for the Oxy-myoglobin (PDB ID: 1a6m) with an improvement in median TM-score of 0.42. The decoys for Ankyrin (PDB ID: 2qyj) shows a minor decrease of 0.04; however, the median TM-score for ROSETTA decoys is 0.78, and thus a minor decrease in decoy quality may be negligible.

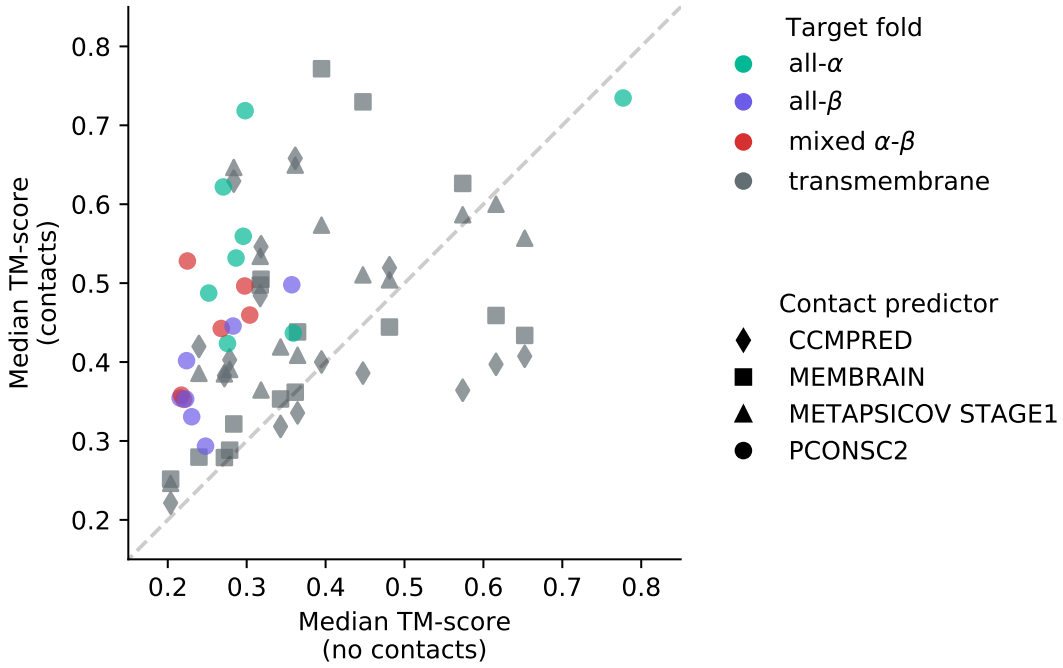



Figure 3.2: Effect of contact distance restraints on *ab initio* decoy quality by comparison of unrestrained (*no contacts*) and contact-restrained (*contacts*) median TM-scores for 1,000 decoys per target. Colours indicate the target fold and symbols the contact prediction algorithm.

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. 3.3). Whilst only one all-α target has more than 50% native-like decoys based on ROSETTA decoys, five targets surpass this threshold when PCONSC2 contact data is used to restrain the folding procedure. Similarly, 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. 3.3). 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. 3.3).

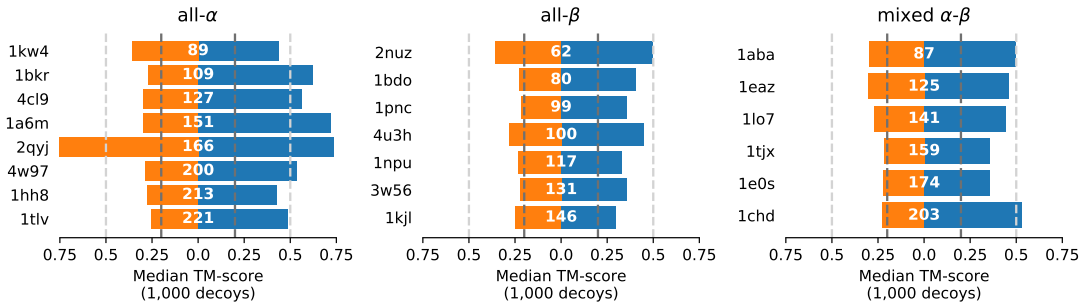



Figure 3.3: TM-score comparison for globular targets separated by fold and ordered by target chain length. Median TM-scores for 1,000 decoys generated with simple ROSETTA (orange) or contact-guided ROSETTA (blue) runs. White numbers in each row correspond to the target chain length. Bars surpassing the 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 [32].

One further important aspect in this study is the addition of BBCONTACTS contact pairs to improve structure prediction accuracy for β-structure containing targets.

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