manuscript rr5187 for review

STRUCTURAL BIOLOGY

Molecular replacement using structure predictions from databases

Daniel Rigden, Adam Simpkin, Felix Simkovic and Ronan Keegan

Category: *research papers*

Co-editor:

Professor R. Read

Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, UK

Telephone: 01223 336500

Fax: 01223 336827

Email: rjr27@cam.ac.uk

Submitting author:

Daniel Rigden

Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, L69 7ZB, United Kingdom

Telephone: 0151 795 4467

Fax: 0151 795 4467

Email: drigden@liverpool.ac.uk

Molecular replacement using structure predictions from databases

Adam J. Simpkin[a], Jens M.H. Thomas[a], Felix Simkovic[a], Ronan M. Keegan[b] and Daniel J. Rigden[a]*

[a] Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, United Kingdom,

[b] Research Complex at Harwell, STFC Rutherford Appleton Laboratory, OX11 0FA, United Kingdom.

*Correspondence e-mail drigden@liv.ac.uk.

Running Title: MR using structure predictions from databases

Abstract

Molecular Replacement (MR) is the predominant route to solution of the phase problem in macromolecular crystallography. Where the lack of a suitable homologue precludes conventional MR, one option is to predict the target structure with bioinformatics. Such modelling, in the absence of homologous templates, is called *ab initio* or *de novo* modelling. Recently the accuracy of such models has improved significantly as a result of the availability, in many cases, of residue contact predictions derived from evolutionary covariance analysis. Covariance-assisted *ab initio* models representing structurally uncharacterised Pfam families are now available on a large scale in databases, potentially representing a valuable and easily accessible supplement to the PDB as a source of search models.

Here we deploy the unconventional MR pipeline AMPLE to explore the value of structure predictions in the GREMLIN and PconsFam databases. We test whether these deposited predictions, processed in various ways, can solve the structures of PDB entries subsequently deposited. The results were encouraging: nine of 27 GREMLIN cases solved, covering target lengths of 109-355 residues and a resolution range of 1.4-2.9Å, and with target-model shared sequence identity as low as 20%. AMPLE's cluster-and-truncate approach proved essential for most successes. For the overall lower quality structure predictions in the PconsFam database, remodelling with Rosetta within the AMPLE pipeline proved to be the best approach, generating ensemble search models from single structure deposits. Finally, we show that AMPLE-obtained search models deriving from Gremlin deposits are of sufficiently high quality to be selected by the sequence-independent MR pipeline SIMBAD. Overall the results help point the way towards the optimal use of expanding *ab initio* structure prediction databases.

Introduction

Macromolecular crystallography requires a source of phasing information to supplement the measured diffraction intensities and thereby solve a structure. Although experimental methods are available, the most popular method for obtaining phase information is Molecular Replacement (MR). MR involves the positioning of a search model in the asymmetric unit, usually by sequential rotation and translation steps, thereby providing approximate phase information which, together with the measured diffraction data, allows for the calculation of initial electron density maps (Rossmann & Blow, 1962).

Conventional MR typically employs the structure of a homologue of the target protein as a search model, often after some manual or automatic editing. The editing is designed to remove loops or side chains that sequence comparison shows differ between homologue and target, or which are flexible and hence prone to adopt different conformations in the known and unknown structures (Schwarzenbacher *et al.*, 2004,Stein, 2008,Bunkoczi & Read, 2011,Lebedev *et al.*, 2008). Conventional MR becomes more difficult as the target-search model relationship becomes more distant and, consequently, the structures tend to differ more. Considerable effort is therefore applied to push the boundaries of conventional MR by non-trivial treatments of distantly homologous structures (Bunkoczi & Read, 2011,Rigden *et al.*, 2018) (Sammito *et al.*, 2014) and/or their advantageous superposition

to serve as ensemble search models (Leahy *et al.*, 1992,Adams *et al.*, 2010,Keegan et al., 2018). Ensemble search models work particularly effectively with the maximum likelihood scoring approach used by Phaser (McCoy, 2004,McCoy *et al.*, 2007). Selection of homologues to serve as search models is typically done by a sequence homology search of the Protein Data Bank (PDB; (wwPDB consortium, 2018)) but the imperfect correlation between sequence- and structural similarity (eg in protein families that can adopt multiple conformations) means that large-scale sequence-independent screens of the PDB or a derivative database are also undertaken (Hatti *et al.*, 2016,Stokes-Rees & Sliz, 2010,Simpkin *et al.*, 2018).

Beyond the boundaries of conventional MR, for very distant homologues or even novel folds, unconventional MR approaches have been developed. These exploit other sources of search models such as ideal regular secondary structure elements or motifs (Rodriguez *et al.*, 2012), recurring tertiary folding patterns (Sammito *et al.*, 2013), or *ab initio* models (Bibby *et al.*, 2012,Keegan *et al.*, 2015,Simkovic *et al.*, 2016). *Ab initio* models are structure predictions that can be obtained based on sequence alone, independent of structural information from homologues being present in the PDB. The first broadly successful approach, as used by the programs Rosetta (Shortle *et al.*, 1998,Leaver-Fay *et al.*, 2011) and Quark (Xu & Zhang, 2012), builds structures from fragments of unrelated proteins using Monte Carlo algorithms to sample search space and sophisticated search functions to recognise structures that share features of experimental protein structures. Early work on the use of *ab initio* models (Qian *et al.*, 2007,Rigden *et al.*, 2008) inspired the development of the pipeline AMPLE using Rosetta in particular for the modelling (Bibby *et al.*, 2012). However, its utility was limited by the size of protein that could be accurately modelled - up to around 120 residues at the time - and by the poorer quality in general of structures that were rich in β-structure, in comparison to α-helical proteins (Bibby *et al.*, 2012).

More recently, the availability of intra- and inter-molecular residue contact predictions, derived from evolutionary covariance analysis of deep protein sequence alignments (Morcos *et al.*, 2011), has revolutionised structural bioinformatics (de Oliveira & Deane, 2017) with many implications for structural biology (Simkovic *et al.*, 2017). It was immediately perceived that good quality contact predictions would enable the folding *ab initio* of much larger proteins (Marks *et al.*, 2011). Indeed, reasonably accurate fold predictions were soon obtained for globular proteins of >200 residues (Marks *et al.*, 2011) and transmembrane helical proteins containing more than 500 residues (Hopf *et al.*, 2012). Several groups use distance geometry structure prediction methods implemented in CNS (Brunger *et al.*, 1998,Brunger, 2007) but others continue with fragment assembly approaches, with particularly impressive results obtained by exploiting metagenomics databases to deepen the sequence alignments that can be obtained for targets and thereby obtain more accurate contact predictions (Ovchinnikov et al., 2017).

With the rapid development of contact-assisted *ab initio* modelling methods, several groups have given thought to producing structure predictions to cover protein sequence space, using Pfam (El-Gebali *et al.*, 2018) as a convenient definition of protein families. Prominent among these are the GREMLIN database (Ovchinnikov et al., 2017), containing representatives of 614 Pfam families resulting from sophisticated iterative modelling with Rosetta, and the PconsFam database (Lamb *et al.*, 2019), covering a much larger number of protein families - 13,617 - but with more rapidly obtained models. Since these models

represent Pfam families, often with thousands of members, they provide a degree of structural information for many proteins: for example, the GREMLIN authors calculate that their models with predicted TM-scores of >0.65 (where a TM-score > 0.5 is taken as a correct fold prediction (Zhang & Skolnick, 2004a,Xu & Zhang, 2010)) cover almost half a million sequences in UniRef100 (Suzek *et al.*, 2007). Thus, as models have become increasingly accurate, and especially as they are likely to become ever more readily accessible at prominent protein sequence databases in the near future, an exploration of their potential for MR is timely. Here we show that the MR pipeline AMPLE provides an effective way to prepare search models from entries in the GREMLIN and PconsFam databases. The former are clustered and truncated directly using the same protocols developed for locally produced *ab initio* models: this solves many more structures than use of deposited structure predictions more directly. The single deposited structure predictions in the PconsFam database are best dealt with by Rosetta remodelling, that can be conveniently done within the AMPLE pipeline, with clustering and truncating of the results to compose ensemble search models (Figure 1). A preliminary exploration of the use of database-derived search ensembles in the sequence-independent MR pipeline SIMBAD (Simpkin *et al.*, 2018) is also presented.

Methods

Test set Selection

Cases were chosen from the GREMLIN database (Ovchinnikov et al., 2017) which contains 30 structure predictions for each of 614 proteins, each protein representing a Pfam family (El-Gebali *et al.*, 2018) that was structurally uncharacterised (i.e. the Pfam database recorded no experimentally determined structure in the family entry) at the time of modelling. At the time of publication of the database structures had subsequently been determined for six families. 30 families that were structurally characterised post-modelling between Jan 2017 and Dec 2018 were identified by mining the Pfam database for structures related to the 614 families. This gave a total of 36 (Supplementary Table 1). Of these, 10 were eliminated as only having diffraction data to > 3A resolution (one case), or where the quality of the model was too poor (nine cases). Poor modelling was defined as resulting in models (represented by the first of the 30 structures deposited for each protein) that gave TM-scores (Zhang & Skolnick, 2004a), normalised either to the target structure or to the model, that were both <0.5: such values indicate that the overall fold has not been correctly modelled (Xu & Zhang, 2010). We asked whether the remaining 26 cases (Table 1) could have been solved using the results of the modelling deposited in the databases.

The PconsFam database (Lamb *et al.*, 2019) contains single structure predictions for 13,617 proteins, again each representing a Pfam family. As well as addressing novel folds, it contains models for families that are structurally characterised. For 22 of the 26 cases above models were available from the PconsFam database. However, only six of the 22 passed the TM-score >0.5 criterion and one of these (4xb6) was not attempted since the models were rather poor (TM-score of 0.55) and eight copies of the target protein were present in the asymmetric unit. Since the number of suitable PconsFam models was rather small, experiments were also undertaken with selected other families for which it was known that high quality models were available in the PconsFam database. These were the Ras family

(PF00071) where the model was used to try to solve the structure deposited in the PDB as 1yzq (1.78 Å resolution) and the DUF305 family (PF03713; 5ffa; 1.50 Å)

Search model generation

For the 26 GREMLIN test cases, the 30 structure predictions deposited for each were used as direct input to AMPLE 1.4.6 in CCP4 7.0.68 (Winn *et al.*, 2011) . The current default processing options were used for search model composition, namely for each of the top 10 SPICKER (Zhang & Skolnick, 2004b) clusters, truncate progressively in 20 steps from 100% (untruncated) down to around 5% remaining, subcluster (Bibby *et al.*, 2012) using 1 or 3Å radii and remove all side chains to leave polyalanine search models. Models are truncated into bins as close to the desired percentage intervals as possible, but as protein sequences are discrete entities of variable lengths they are not always evenly divisible into the desired bins. As the actual size of the truncation bins is reported, the size of the bins may vary a little from the ideal percentage values.

Two additional attempts were made for comparison: all 30 structure predictions were presented directly to Phaser as an ensemble; and entries in the separate database of single 'final models' were processed in AMPLE single structure mode (Rigden *et al.*, 2018) using VoroMQA (Olechnovic & Venclovas, 2017) to provide per-residue quality scores which drove progressive truncation over a set of 20 thresholds. Retention of side chains or editing to polyalanine were specified so that 40 search models were derived for each case.

Since the PconsFam database contains only single models per Pfam family, three approaches were tried. Firstly, truncation of the single models in AMPLE was done using its single structure mode as above using VoroMQA server protein structure quality predictions. Secondly, Rosetta remodelling was done using the PconsFam model as a basis. This approach was previously employed with NMR ensembles and proved to improve performance. Using the -nmr_remodel flag causes AMPLE to idealise the input structure, here the PconsFam model, and then remodel the result, using a provided target sequence, into a number of new structures, sampling conformational space in a fragment-dependent fashion. Fragment libraries were obtained from the Robetta server (Kim *et al.*, 2004) with the exclude homologues option selected in order that the remodelling was not influenced by any knowledge of the target structure or homologues. Here, 100 structures were derived from each PconsFam model and given to AMPLE for clustering and truncation as above. Thirdly, for selected targets, the PconsFam single structures were transformed into ensembles using CONCOORD (de Groot *et al.*, 1997) as previously (Rigden *et al.*, 2018). Briefly, CONCOORD extracts restraints from a given structure and then uses distance geometry methods to build a set of variant structures that differ from the original but which obey the derived restraints. By this procedure, less well-packed regions such as loops exhibit structural divergence in resulting derivative structures and hence, by the AMPLE algorithm, are subject to truncation.

Molecular Replacement

Within the AMPLE pipeline, MrBUMP (Keegan et al., 2018) trialled the search models using Phaser 2.8.2 (McCoy *et al.*, 2007,Read & McCoy, 2016). The default AMPLE estimated rmsd error of 0.1 Å was used but this value was adjusted internally by Phaser where inconsistent with the internal structural variability of the ensemble. Success was judged as a placement that yielded a map Correlation Coefficient (CC) of 0.25 or higher using *phenix.get_map_cc_mtz_pdb* (Adams *et al.*, 2010). All of these cases also produced a CC of >25% upon main chain tracing using SHELXE 2016 (Thorn & Sheldrick, 2013) with the single exception of 5uw2, for which diffraction data to only 2.9 Å were available, which produced a marginally lower score of 24.8. All these solutions could be refined to Rfree <0.45, using either just the BUCCANEER (Cowtan, 2006) plus REFMAC (Murshudov *et al.*, 2011) protocol built into AMPLE's default operation or, where necessary (for 5oon and 5uw2) by directly refining the PHASER placement with REFMAC (Murshudov *et al.*, 2011), or with manual model building. For comparison we attempted solution of all 27 using AMPLE's ideal helix mode with a Phaser time limit per search model of 24 hours.

SIMBAD

SIMBAD is a MR pipeline that uses the rotation function to screen large databases of structures (Simpkin *et al.*, 2018). SIMBAD has recently been modified to run Phaser's likelihood-enhanced fast rotation function (Simpkin *et al.*, 2019). This has increased the sensitivity of the pipeline and also allowed single search models to be replaced with ensembles. The MoRDa (Vagin & Lebedev, 2015) ensemble database that SIMBAD is typically run against, was modified to include AMPLE derived ensembles made from the models in the GREMLIN database. Initial experiments suggested that the rotation function was not sensitive enough to pick up these poor models, so SIMBAD was modified to also run Phaser's likelihood enhanced fast translation search (McCoy *et al.*, 2005) but only on the best orientation identified in the rotation function. In this work the top 200 solutions by translation score were taken forward for MR and refinement as opposed to the top 200 solutions by rotation score in previously published work.

Results and Discussion

Using models from the GREMLIN database

The 27 cases studied include many cases that are challenging in terms of the relatively high structural deviations between model and target and/or the complex and sometimes heterooligomeric composition of the asymmetric unit: only eight cases contained a single chain in the asymmetric unit. When the GREMLIN structure predictions, each comprising 30 models of a given protein representing a particular Pfam family, were supplied to AMPLE for its default clustering and truncation approach, nine of the 27 cases were solved (Table 1). These nine cases include four transmembrane helical proteins, one globular helical protein and four mixed fold proteins. Thus successes spanned all fold classes but the numbers are too small to suggest whether certain types of protein may be particularly (un)favourable. The ultimately successful structure predictions overall can be considered of medium quality,

sharing rmsd 1.5 to 2.8Å on Calpha atoms (TM-scores of 0.63-0.84) with the targets. The solved cases cover a range of lengths of 112-355 residues and a resolution range of 1.35-2.85Å.

In most cases, the modelled member of a given Pfam family was closely related (>90% shared sequence identity) to the member ultimately structurally characterised. However, there were three exceptions. The first was 5cuo, the crystal structure of *Rhodopseudomonas palustris* PduL which was solved with models of phosphate propanoyltransferase from *Bacillus megaterium* (Pfam: PF06130, Uniprot: D5DKA5) with which it shared only 49% sequence identity. The second was 5xj5, the structure of *Aquifex aeolicus* glycerol-3-phosphate acyltransferase where the model of *Bacillus subtilis* (Pfam: PF02660, Uniprot: Q45064), shared only 36% sequence identity with the target. Most remarkable was 5mlz, the structure of dolichyl phosphate mannose synthase, where the model of an uncharacterised GtrA-family protein from *Bacillus subtilis* (Pfam: PF04138, Uniprot: O31821), shared only 20% sequence identity with the target. When considering these successes with relatively distant homologues, it is worth remembering that the covariance signal, which strongly influences the modelling, will be strongest for features shared throughout the superfamily. This may well help produce models that serve to solve targets from across a superfamily. However, it is also true that the GREMLIN structure predictions are derived from an all-atom, fully sequence-aware protocol that would be expected to give authentically different predictions for homologous proteins. As such, it remains encouraging that structure predictions can solve quite distantly homologous targets. In the three cases mentioned here the secondary structure of the GREMLIN prediction matched that of the target quite well (Supp figs 1-3).

As expected, cases with multiple chains in the asymmetric unit solved less often, but AMPLE succeeded with 5caj (two chains) and 5uw2 (three chains). Since some of the targets contained multiple domains, the search models sometimes represented only a portion of the target. Such was the case with 5mlz where the available model was 123 residues long but solved a structure of 352 residues.

The ease of solution of the nine cases, as expressed as the proportion of search model ensembles that succeeded, varies widely. For 5edl, 132 of 170 search models (78%) succeeded while for 5caj the figures were six of 132 (4.5%). 5edl solved with search models containing 11-100% of the starting model residues while others solved over a narrower range of search model sizes - 27-41% for 5mlz, for example.The most truncated successful search model contained 7% of the starting structure (19 residues) of the 5azb model. This target is the structure of *E. coli* lipoprotein diacylglyceryl transferase, an integral membrane enzyme, 300 residues long, determined to a resolution of 1.6 Å. The 7% successful search model comprises an antiparallel pair of helices. Successful search models for a given target tended to derive from different clusters but cluster 1, containing the largest number of the input 30 models, was not always successful: 5cuo, for example, only solved with search models deriving from clusters 2 and 3. Overall, the results suggest that AMPLE's cluster-and-truncate approach, intensively sampling many non-trivial edits of ensembles deriving from the deposited models, is an appropriate strategy to deal with these structures.

The need to use AMPLE's automated processing and sampling for best performance is illustrated by the poorer performance of two simple baseline approaches. When the top

model for each protein, provided separately to the ensembles in the GREMLIN database, was used, using VoroMQA quality measurements to produce a series of truncated derivatives, only two cases were solved, 5mlz and 5edl. Secondly, when the 30 structures were presented as an ensemble to Phaser directly, only one case could solve. The successful case was PDB code 5edl, where models in the ensemble were between 1.59 Å and 2.30 Å rmsd (TM-scores 0.4-0.87) from the true structure.

The successes presented undoubtedly cover targets that could potentially have been solved alternatively using fragment-based approaches (Rodriguez *et al.*, 2009,Jenkins, 2018). Although the simple ideal helix mode in AMPLE performed relatively poorly, only solving three, more sophisticated approaches might well do better, particularly for cases with higher-resolution diffraction data, helix-rich composition and/or small asymmetric unit contents. The more challenging cases to be solved therefore include 5cuo, a largely beta structure containing two ~200 residue chains; 5uw2 with diffraction only to 2.9 Å resolution; and 5caj, where diffraction data to 1.65A resolution were available but the asymmetric unit contained 510 residues. Fig 2 illustrates that the most successful search models in these three cases are only moderately truncated down to 54, 70 or 80% of the starting structures, indicating that correct overall fold prediction is important (see also Supp Figs 4-6). In contrast, the best-performing search model for 5azb (Fig 2) contained only 12% of the starting structure, and truncations to below 33% were required for success (Supp Fig 7). This observation demonstrates the importance of AMPLE's sampling of truncations over a wide range.

Using models from the PconsFam database

Applying the same TM-score threshold of 0.5, indicating a broadly correct predicted fold (Xu & Zhang, 2010), only five of the 26 families considered above were represented by PconsFam structure predictions that were good enough to take to MR trials. PconsFam contains only single structure predictions for representative proteins of Pfam domains. Three different strategies were therefore employed: truncation of that single structure according to local model quality prediction from the VoroMQA server, generation of ensembles using the distance geometry method CONCOORD and Rosetta remodelling using the PconsFam deposit as a starting point.

The simplest approach, editing a single model according to per-residue predicted quality scores, failed to solve any of the five targets. Rosetta remodelling was successful with two of the five, 5xj5 and 5azb, each transmembrane helical proteins. 5xj5 solved with two search models out of 49, being truncated ensembles from the first cluster containing 23 or 41 residues. The SHELXE traces were automatically rebuilt using BUCCANEER within the AMPLE pipeline to final Rfree values of 28-29%. The larger search model, c1_23_r3_polyAla (where c1 means deriving from cluster 1, 23 means 23% of the initial model remains, r3 refers to a 3Å subclustering radius, and polyAla refers to the side chain treatment) contains most of the C-terminal 3-helical subdomain of the target structure which is more accurately predicted (Figure 3). 5azb was solved by a single search model from the 200 produced. It was derived from the 7th cluster and truncated until it contained 57 residues, mainly composing portions of four of the transmembrane helices. Again, automated rebuilding

produced an Rfree of 29%. Neither of these cases was solved by the simpler and somewhat less time-consuming approach of ensemble generation with CONCOORD.

In order to further explore approaches that could convert PconsFam models into successful search models, some trials were done with Ras protein (Pfam accession PF00071; PDB code 1yzq) and DUF305 (PF03713; 5ffa). For these, high-quality structure predictions were available with TM-scores of 0.85 and 0.76 respectively and both solved using Rosetta remodelling. The Ras structure was solved with 29 from 175 search model ensembles generated, deriving from clusters 1, 2 3 or 7, containing 53-170 residues (170 residues being the full size of the model), and tracing and refining to Rfree values as low as 33% within the AMPLE pipeline. The DUF305 structure solved with 18 of 175 search model ensembles. These were derived from clusters 2, 3, 6 or 7, contained between 79 and 143 residues and automatically traced and refined to Rfrees as low as 33% (Table 2).

Interestingly, CONCOORD-derived ensembles could solve the Ras structure but not the DUF305 case. In the successful run, seven search model ensembles out of a total of 400 generated were successful, deriving from clusters 5, 7 8 or 9 and containined 50-75% of the original model, corresponding to 79-119 residues. Although deriving from different clusters, the successful search models were similar in having discarded less accurately modelled loops but retaining the core fold of well-captured secondary structure elements (Figure 4).

Several factors could be contributing to the relative success of the Rosetta remodelling approach as compared to the single PconsFam model. Most obviously, remodelling the target sequence could take the structure closer to that of the target, especially in cases where the sequence identity between the target and the PconsFam deposit is low. This would combine with the use of a sophisticated energy function in Rosetta (Alford *et al.*, 2017), rather than the simpler function used by PconsFam's structure building algorithm CONFOLD (Adhikari *et al.*, 2015), to potentially allow for more accurate modelling i.e. the PconsFam structure might be 'refined' by the Rosetta step. Secondly, modelling based on covariance information guided distance geometry methods, as in PconsFam, can often lead to results in which local backbone geometry is poor. Potentially the backbone geomtry could be improved by running through Rosetta's fragment-based remodelling. Finally, as is well-established (Qian *et al.*, 2007,Rigden *et al.*, 2008), the comparison across the multiple structures resulting from remodelling, allows the inference of quality enabling truncation to more accurately modelled core regions. Supplementary Table 2 shows the overall accuracy and stereochemical quality of the PconsFam models and the Rosetta structures derived from them.

The results confirm a clear and consistent improvement in backbone geometry as measured by Ramachandran plot statistics and an overall G-factor calculated on backbone dihedrals with positive values indicating better quality. However, these suggest that Rosetta does not generally act to refine the PconsFam models: in fact, in three of the four cases the average correctness of the models, measured as TM-scores, is worse than for the PconsFam starting model. Where the starting structure is poorer quality, it seems that Rosetta fragment-based conformational exploration can effectively unfold the structure. Options to try to prevent this in the future could include the imposition of evolutionary covariance derived contact predictions or more generalised restraints to maintain the structure in the vicinity of the

starting model. Nevertheless, the AMPLE protocol, being based on clustering, is tolerant of some unfolded structures among the input set.

Overall the results suggest that simple editing of the single structure PconsFam models is unlikely to transform them into successful search models. However, where the overall fold has been correctly captured, Rosetta remodelling with subsequent clustering and truncating to generate ensembles, can be effective. This approach clearly outperforms CONCOORD for ensemble generation.

SIMBAD and search models derived from databases

SIMBAD is a sequence-independent MR pipeline that attempts to solve structures using a lattice search, a search of a curated database of known contaminant structures and/or a large-scale search of domain structures (around 120,000) from the MoRDa database. Since recent developments to SIMBAD (Simpkin *et al.*, 2019) have improved its sensitivity - by using Phaser in place of the original AMoRe and through the use of ensemble search models - we tested whether truncated search model ensembles derived from the GREMLIN database that succeeded in AMPLE could succeed in SIMBAD too.

Success in the large scale MorDa screen can arise in two ways in SIMBAD. First, if a tested search model yields a Phaser RFZ high enough (>7) to generally indicate an accurate rotation then it is immediately trialled in a full MR protocol the success of which (R-values below 0.45 and/or both LLG>120 and TFZ>8) would lead to termination of the SIMBAD without testing any remaining search models. Alternatively, if no search model reaches the RFZ threshold, then at the end of the rotation function screen of all search models the 200 that have the highest RFZ scores are trialled for full MR.

The GREMLIN structure predictions are of moderate accuracy at best and require significant processing to succeed. Therefore, we first assessed whether they would score RFZ values likely to lead to their selection in the top 200 in a full MorDa+GREMLIN run. Supp Table 3 shows the range of RFZ values obtained for the range of truncated search models produced by AMPLE for cases that successfully solved. In general the results were somewhat disappointing: no search model ensemble achieved an RFZ greater than 6.11. Although full SIMBAD runs were not done, experience suggests that these values are unlikely to place the search model ensembles, even those that ultimately succeeded in AMPLE, within the top 200. As such, they would never proceed to the full MR step.

In a bid to improve the sensitivity of the SIMBAD pipeline further, we therefore experimented with the addition of the Phaser translation function on just the top ranked orientation in the rotation search. We reasoned that placing the search model would improve the signal to noise from good search models. Preliminary results suggested that this worked well: for example, search model ensembles for 5xj5 gave LLG/TFZ scores of up to 90.35/7.68 while ensembles for 5edl gave LLG/TFZ scores as high as 147.32/13.05. These values are indicative of success.

A version of SIMBAD in which the database, in this case MoRDa supplemented by GREMLIN-derived ensembles, is screened using a rotation function in combination with the

rapid translation function was then produced. As a proof of principle this was tested on 5edl due to the high TFZ scores observed. This gave a clear success with six AMPLE ensembles being reported in the top 200 (c1_74_r3_polyala, c1_t89_r3_polyala, c1_t74_r1_polyala, c1_t79_r1_polyala, c1_t84_r3_polyala, c1_t100_r3_polyala) with the best example being shown in Figure 5.

Naturally, the additional translation function can increase the runtime of SIMBAD but this will be compensated for, to some extent, by more frequent early termination due to the improved sensitivity with which good search models can be selected.

Conclusions

Databases of protein homology models have a long history (Kiefer *et al.*, 2009,Pieper *et al.*, 2014,Guex & Peitsch, 1997), more recently under the aegis of the Protein Model Portal (Haas *et al.*, 2013), and homology models have been used for MR e.g. (Horsefield *et al.*, 2008,Jung *et al.*, 2011). Nevertheless, we are unaware of cases where a homology model, much less an *ab initio* model, downloaded from a database has been used as a search model. These new results demonstrate that recently emerged databases of *ab initio* models, representing Pfam families with structures that are very different from anything deposited in the PDB, already contain information that can solve structures of proteins from those families by MR. The success of the MR in AMPLE should be considered in the context of the quality of the models available in the GREMLIN and PconsFam databases. We could collect 36 cases representing Pfam families that were structurally characterised at the time of their GREMLIN modelling but subsequently deposited in the PDB. Of these 26 had GREMLIN models of the right fold (TM-score >0.5) while the figure was only five for the PconsFam database. This observation can be related to the more sophisticated modelling protocol behind the GREMLIN database and its exploitation of metagenomic data to improve the quality of the contact predictions driving the modelling (Ovchinnikov et al., 2017). However, within those different sets the success by MR was actually comparable - 9/26 with GREMLIN-derived search models, two from five with PconsFam. GREMLIN predictions with TM-scores as low as 0.64 could succeed while the two successful PconsFam cases in the set of five were based on structure predictions with TM-scores of 0.80 and 0.69: for the additional PconsFam cases (Ras and DUF305) these values were 0.85 and 0.76. Overall, the results suggest that models should score somewhat better than the correct fold criterion of TM-score >0.5 in order to succeed.The current advantage of the PconsFam databases is its coverage, but the simpler modelling protocol is likely to mean that its predictions are poorer quality on average than the GREMLIN contents. A user may currently estimate the likely model quality of a PconsFam model by looking at its Pcons (Lundstrom *et al.*, 2001), or ProQ3D (Uziela *et al.*, 2017) model quality scores, or the underlying alignment depth (number of effective sequences) upon which the contact prediction was done.

The requirement of the Rosetta remodelling approach for success with some PconsFam models might invite the comment that a user could simply generate her own models rather than work with those from the database. However, databases like PconsFam and GREMLIN contain models derived using state-of-the-art contact predictions and, in the latter case, complex, bespoke and iterative modelling pipelines. For a crystallographer to recapitulate

these approaches, within or without AMPLE, is certainly more demanding in computational skills and infrastructure than the comparatively rapid (around 80 min on 10 cores) remodelling approach outlined here.

In summary, these results demonstrate that *ab initio* structure predictions deposited in online databases are already of sufficient quality to form the basis of successful MR search models. Some of the targets addressed here could undoubtedly be alternatively solved with sophisticated fragment-based methods (Rodriguez *et al.*, 2009,Jenkins, 2018), but AMPLE conveniently provides a unifying framework to attempt solution of such cases (typically higher resolution, higher helical content) as well as harder cases (Fig 2) where high quality modelling is key and moderately edited search models containing almost entire folds succeed. However, the evidence currently suggests that non-trivial processing is required for optimal performance, to transform single models into ensembles and to eliminate inaccurate regions from ensembles such that better-modelled core regions remain. These *ab initio* models are calculated using covariance-driven approaches, and represent sometimes large families of structurally uncharacterised proteins. The GREMLIN database has much smaller coverage at the time of writing, but there are plans to liaise with the Pfam database (El-Gebali *et al.*, 2018) and use the latter as a means to disseminate models that cover more of protein sequence space. Such models will be periodically recalculated as and when expansion of sequence databases allowed for improved contact predictions and hence better modelling (R Finn, personal communication). These plans run alongside similar efforts to collect homology models from structural bioinformatics resources such as Genome3D (Lewis *et al.*, 2013) and make them available within the InterPro database (Mitchell *et al.*, 2019) (R Finn, personal communication). In the near future these databases will facilitate access to increasingly available and high quality models, be they *ab initio*- or homology-based. As such, they will increasingly be viewed as a valuable supplement to the PDB as sources of MR search models.

Acknowledgements

This work was supported by BBSRC grant BB/L009544/1 ‘CCP4 Grant Renewal 2014–2019: Question-driven crystallographic data collection and advanced structure solution’. FS was supported by a BBSRC DTP PhD scholarship at the time of the work.

Figure legends

Figure 1) Flowchart showing the methods used to treat search models obtained from GREMLIN and PconsFam prior to AMPLE/AMPLE single model mode. The relative success of each method is represented in green, orange or red where green represents a more successful method and red represents a less successful one.

Figure 2) a) The 30 models obtained from the GREMLIN database for PF01790 (magenta) aligned with the crystalised structure, 5azb (rainbow from blue at the N-terminus to red at the C-terminus). b) The best-performing AMPLE derived ensemble (magenta), derived by truncating cluster 1 down to 12% (33 residues), aligned with the crystalised structure, 5azb (rainbow). c) The 30 models obtained from the GREMLIN database for PF02470 (magenta) aligned with the crystal structure, 5uw2 (rainbow). d) The best-performing AMPLE derived ensemble (magenta), derived by truncating cluster 2 down to 80% (96 residues), aligned with the crystal structure, 5uw2 (rainbow). e) The 30 models obtained from the GREMLIN database for PF03883 (magenta) aligned with the crystalised structure, 5caj (rainbow). f) The best-performing AMPLE derived ensemble (magenta), derived by truncating cluster 1 down to 54% (137 residues), aligned with the crystal structure, 5caj (rainbow). g) The 30 models obtained from the GREMLIN database for PF06130 (magenta) aligned with the crystalised structure, 5cuo (rainbow). h) The best-performing AMPLE derived ensemble (magenta), derived by truncating cluster 3 down to 70% (138 residues), aligned with the crystal structure, 5cuo (rainbow).

Figure 3) a) PconsFam model for PF02660 (magenta) aligned with the crystalised structure, 5jx5 (rainbow). b) An untruncated AMPLE ensemble (magenta ribbon), following Rosetta remodelling, aligned with the crystalised structure, 5jx5 (rainbow). c) The truncated AMPLE ensemble (c1_23_r3_polyAla) obtained from the Rosetta remodelled versions of the PconsFam model for PF02660 (magenta) aligned with the crystalised structure, 5jx5 (rainbow).

Figure 4) a) PconsFam model for PF00071 (magenta) aligned with the crystalised structure, 1yzq (rainbow). b) An untruncated AMPLE ensemble (magenta ribbon), following CONCOORD, aligned with the crystalised structure, 1yzq (rainbow). c) The AMPLE ensemble obtained from the CONCOORD derivatives for PF00071 (magenta) aligned with the crystalised structure, 1yzq (rainbow).

Figure 5) Cross-eyed stereo view of the AMPLE ensemble (c1_t74_r3_polyala) which gave the best score in the SIMBAD search for PF09819 (magenta) aligned with the crystalised structure, 5edl (rainbow from blue at the N-terminus to red at the C-terminus).

References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213-221.

Adhikari, B., Bhattacharya, D., Cao, R. & Cheng, J. (2015). *Proteins.* **83,** 1436-1449.

Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L.,Jr, Das, R., Baker, D., Kuhlman, B., Kortemme, T. & Gray, J. J. (2017). *J. Chem. Theory Comput.* **13,** 3031-3048.

Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Crystallogr. D Biol. Crystallogr.* **68,** 1622-1631.

Brunger, A. T. (2007). *Nat. Protoc.* **2,** 2728-2733.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905-921.

Bunkoczi, G. & Read, R. J. (2011). *Acta Crystallogr. D Biol. Crystallogr.* **67,** 303-312.

Cowtan, K. (2006). *Acta Crystallogr. D Biol. Crystallogr.* **62,** 1002-1011.

de Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A., Vriend, G. & Berendsen, H. J. (1997). *Proteins.* **29,** 240-251.

de Oliveira, S. & Deane, C. (2017). *F1000Res.* **6,** 1224.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. & Finn, R. D. (2018). *Nucleic Acids Res.*

Guex, N. & Peitsch, M. C. (1997). *Electrophoresis.* **18,** 2714-2723.

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L. & Schwede, T. (2013). *Database (Oxford).* **2013,** bat031.

Hatti, K., Gulati, A., Srinivasan, N. & Murthy, M. R. (2016). *Acta Crystallogr. D. Struct. Biol.* **72,** 1081-1089.

Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C. & Marks, D. S. (2012). *Cell.* **149,** 1607-1621.

Horsefield, R., Norden, K., Fellert, M., Backmark, A., Tornroth-Horsefield, S., Terwisscha van Scheltinga, A. C., Kvassman, J., Kjellbom, P., Johanson, U. & Neutze, R. (2008). *Proc. Natl. Acad. Sci. U. S. A.* **105,** 13327-13332.

Jenkins, H. T. (2018). *Acta Crystallogr. D. Struct. Biol.* **74,** 205-214.

Jung, J., Kim, J. K., Yeom, S. J., Ahn, Y. J., Oh, D. K. & Kang, L. W. (2011). *Appl. Microbiol. Biotechnol.* **90,** 517-527.

Keegan, R. M., Bibby, J., Thomas, J., Xu, D., Zhang, Y., Mayans, O., Winn, M. D. & Rigden, D. J. (2015). *Acta Crystallogr. D Biol. Crystallogr.* **71,** 338-343.

Keegan, R. M., McNicholas, S. J., Thomas, J. M. H., Simpkin, A. J., Simkovic, F., Uski, V., Ballard, C., Winn, M. D., Wilson, K. S. & Rigden, D. J. (2017). *Acta Crystallogr. D Biol. Crystallogr.* **74**, 167-182 .

Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L. & Schwede, T. (2009). *Nucleic Acids Res.* **37,** D387-92.

Kim, D. E., Chivian, D. & Baker, D. (2004). *Nucleic Acids Res.* **32,** W526-31.

Lamb, J., Jarmolinska, A. I., Michel, M., Menendez-Hurtado, D., Sulkowska, J. I. & Elofsson, A. (2019). *J. Mol. Biol.*

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). *Cell.* **68,** 1145-1162.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. (2011). *Methods Enzymol.* **487,** 545-574.
Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2008). *Acta Crystallogr. D Biol. Crystallogr.* **64,** 33-39.

Lewis, T. E., Sillitoe, I., Andreeva, A., Blundell, T. L., Buchan, D. W., Chothia, C., Cuff, A., Dana, J. M., Filippis, I., Gough, J., Hunter, S., Jones, D. T., Kelley, L. A., Kleywegt, G. J., Minneci, F., Mitchell, A., Murzin, A. G., Ochoa-Montano, B., Rackham, O. J., Smith, J., Sternberg, M. J., Velankar, S., Yeats, C. & Orengo, C. (2013). *Nucleic Acids Res.* **41,** D499-507.

Lundstrom, J., Rychlewski, L., Bujnicki, J. & Elofsson, A. (2001). *Protein Sci.* **10,** 2354-2362.
Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. & Sander, C. (2011). *PLoS One.* **6,** e28766.

McCoy, A. J. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2169-2183.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Crystallogr.* **40,** 658-674.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Crystallogr. D Biol. Crystallogr.* **61,** 458-464.

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H. Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Nuka, G., Orengo, C., Pandurangan, A. P., Paysan-Lafosse, T., Pesseat, S., Potter, S. C., Qureshi, M. A., Rawlings, N. D., Redaschi, N., Richardson, L. J., Rivoire, C., Salazar, G. A., Sangrador-Vegas, A., Sigrist, C. J. A., Sillitoe, I., Sutton, G. G., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Yong, S. Y. & Finn, R. D. (2019). *Nucleic Acids Res.* **47,** D351-D360.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. & Weigt, M. (2011). *Proc. Natl. Acad. Sci. U. S. A.* **108,** E1293-301.

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Crystallogr. D Biol. Crystallogr.* **67,** 355-367.

Olechnovic, K. & Venclovas, C. (2017). *Proteins.* **85,** 1131-1145.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P. S., Pavlopoulos, G. A., Kim, D. E., Kamisetty, H., Kyrpides, N. C. & Baker, D. (2017). *Science.* **355***,* 294-298.

Pieper, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., Khuri, N., Spill, Y. G., Weinkam, P., Hammel, M., Tainer, J. A., Nilges, M. & Sali, A. (2014). *Nucleic Acids Res.* **42,** D336-46.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature.* **450,** 259-264.

Read, R. J. & McCoy, A. J. (2016). *Acta Crystallogr. D. Struct. Biol.* **72,** 375-387.

Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Crystallogr. D Biol. Crystallogr.* **64,** 1288-1291.

Rigden, D. J., Thomas, J. M. H., Simkovic, F., Simpkin, A., Winn, M. D., Mayans, O. & Keegan, R. M. (2018). *Acta Crystallogr. D. Struct. Biol.* **74,** 183-193.

Rodriguez, D., Sammito, M., Meindl, K., de Ilarduya, I. M., Potratz, M., Sheldrick, G. M. & Uson, I. (2012). *Acta Crystallogr. D Biol. Crystallogr.* **68,** 336-343.

Rodriguez, D. D., Grosse, C., Himmel, S., Gonzalez, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Uson, I. (2009). *Nat. Methods.* **6,** 651-653.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Crystallogr.* **15,** 24-31.

Sammito, M., Meindl, K., de Ilarduya, I. M., Millan, C., Artola-Recolons, C., Hermoso, J. A. & Uson, I. (2014). *FEBS J.* **281,** 4029-4045.

Sammito, M., Millan, C., Rodriguez, D. D., de Ilarduya, I. M., Meindl, K., De Marino, I., Petrillo, G., Buey, R. M., de Pereda, J. M., Zeth, K., Sheldrick, G. M. & Uson, I. (2013). *Nat. Methods.* **10,** 1099-1101.

Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60,** 1229-1236.

Shortle, D., Simons, K. T. & Baker, D. (1998). *Proc. Natl. Acad. Sci. U. S. A.* **95,** 11158-11162.

Simkovic, F., Ovchinnikov, S., Baker, D. & Rigden, D. J. (2017). *IUCrJ.* **4,** 291-300.

Simkovic, F., Thomas, J. M. H., Keegan, R. M., Winn, M. D., Mayans, O. & Rigden, D. J. (2016). *IUCr J.* **15,** 259-270.

Simpkin, A. J., Simkovic, F., Thomas, J. M. H., Savko, M., Lebedev, A., Uski, V., Ballard, C., Wojdyr, M., Shepard, W., Rigden, D. J. & Keegan, R. M. (2019). *Acta Crystallogr. D Biol. Crystallogr.*

Simpkin, A. J., Simkovic, F., Thomas, J. M. H., Savko, M., Lebedev, A., Uski, V., Ballard, C., Wojdyr, M., Wu, R., Sanishvili, R., Xu, Y., Lisa, M. N., Buschiazzo, A., Shepard, W., Rigden, D. J. & Keegan, R. M. (2018). *Acta Crystallogr. D. Struct. Biol.* **74,** 595-605.

Stein, N. (2008). *J. Appl. Cryst.* **41,** 641-643.

Stokes-Rees, I. & Sliz, P. (2010). *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21476-21481.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. (2007). *Bioinformatics.* **23,** 1282-1288.

Thorn, A. & Sheldrick, G. M. (2013). *Acta Crystallogr. D Biol. Crystallogr.* **69,** 2251-2256.

Uziela, K., Menendez Hurtado, D., Shu, N., Wallner, B. & Elofsson, A. (2017). *Bioinformatics.* **33,** 1578-1580.

Vagin, A. A. & Lebedev, A. A. (2015). *Acta Crystallogr. A.* **71,** s19.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Crystallogr. D Biol. Crystallogr.* **67,** 235-242.

wwPDB consortium. (2018). *Nucleic Acids Res.* **47**, D520-D528

Xu, D. & Zhang, Y. (2012). *Proteins.* **80,** 1715-1735.

Xu, J. & Zhang, Y. (2010). **26,** 889-895.

Zhang, Y. & Skolnick, J. (2004a). *Proteins.* **57,** 702-710.

Zhang, Y. & Skolnick, J. (2004b). *J. Comput. Chem.* **25,** 865-871.

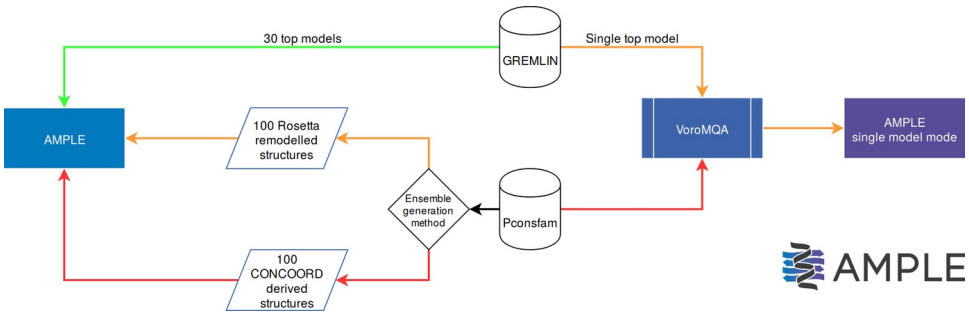
30 top models
GREMLIN
Single top model
AMPLE
100 Rosetta remodelled structures
VoroMQA
AMPLE single model mode
Ensemble generation method
Pconsfam
100 CONCOORD derived structures
AMPLE

Figure 1

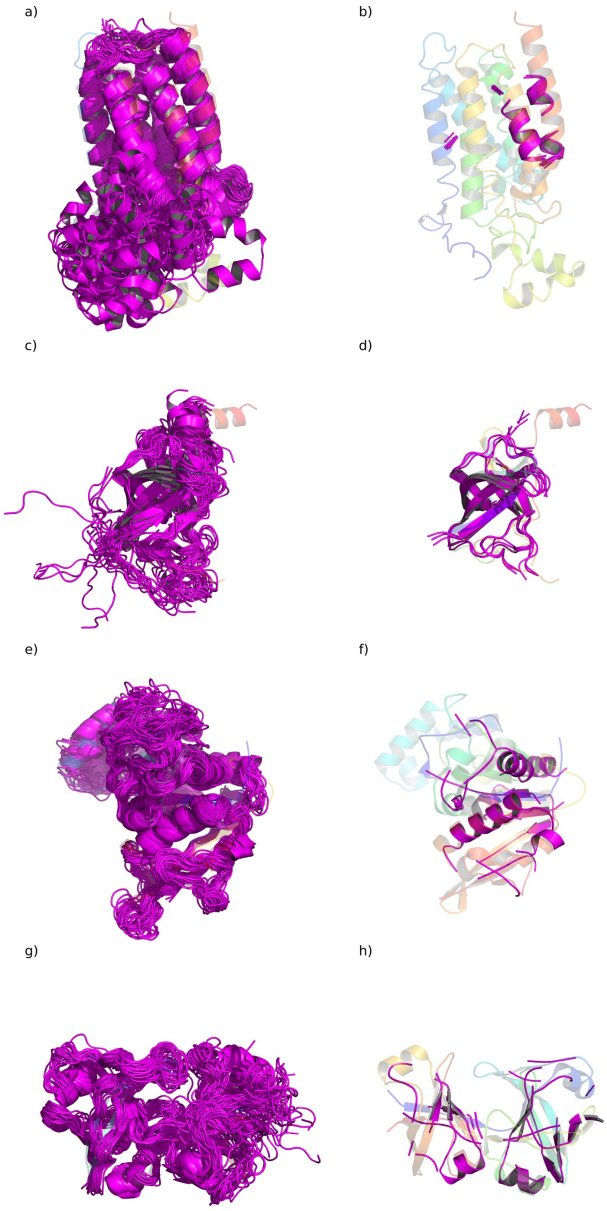

Figure 2

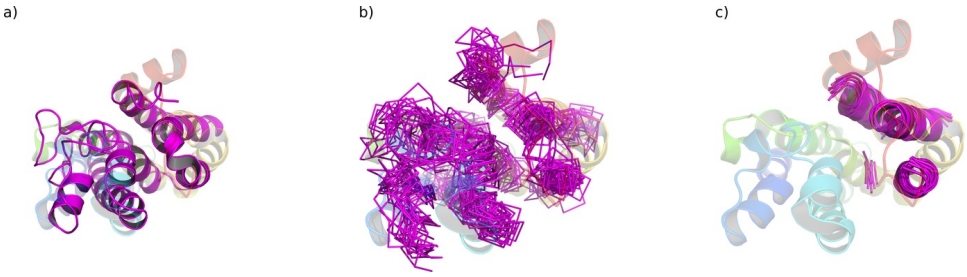

Figure 3

a) b) c)

Figure 4

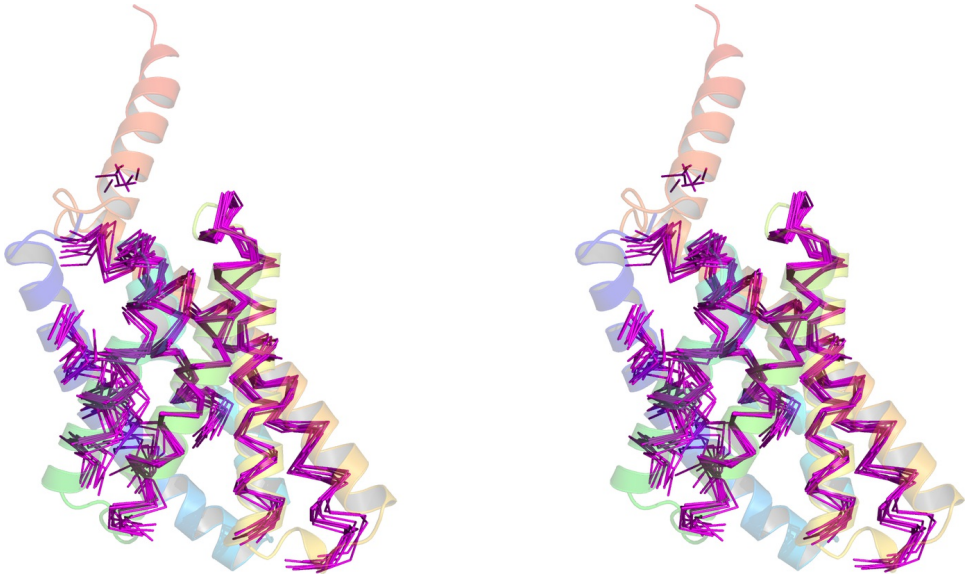

Figure 5