

## Residue contacts predicted by evolutionary covariance extend the application of *ab initio* molecular replacement to larger and more challenging protein folds

Felix Simkovic,<sup>a</sup> Jens M. H. Thomas,<sup>a</sup> Ronan M. Keegan,<sup>b</sup> Martyn D. Winn,<sup>c</sup> Olga Mayans<sup>a‡</sup> and Daniel J. Rigden<sup>a\*</sup>

Received 10 February 2016

Accepted 18 May 2016

Edited by Z.-J. Liu, Chinese Academy of Sciences, China

‡ Present address: Department of Biology, Universität Konstanz, 78457 Konstanz, Germany.

**Keywords:** molecular replacement; protein structure prediction; evolutionary covariation; predicted contacts; *ab initio* modelling.

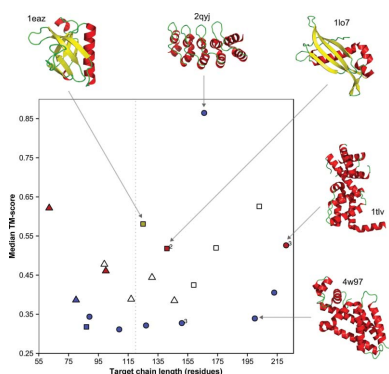
**Supporting information:** this article has supporting information at [www.iucrj.org](http://www.iucrj.org)

<sup>a</sup>Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, England, <sup>b</sup>Research Complex at Harwell, STFC Rutherford Appleton Laboratory, Didcot OX11 0FA, England, and <sup>c</sup>Science and Technology Facilities Council, Daresbury Laboratory, Warrington WA4 4AD, England. \*Correspondence e-mail: [d.rigden@liv.ac.uk](mailto:d.rigden@liv.ac.uk)

For many protein families, the deluge of new sequence information together with new statistical protocols now allow the accurate prediction of contacting residues from sequence information alone. This offers the possibility of more accurate *ab initio* (non-homology-based) structure prediction. Such models can be used in structure solution by molecular replacement (MR) where the target fold is novel or is only distantly related to known structures. Here, *AMPLE*, an MR pipeline that assembles search-model ensembles from *ab initio* structure predictions ('decoys'), is employed to assess the value of contact-assisted *ab initio* models to the crystallographer. It is demonstrated that evolutionary covariance-derived residue–residue contact predictions improve the quality of *ab initio* models and, consequently, the success rate of MR using search models derived from them. For targets containing  $\beta$ -structure, decoy quality and MR performance were further improved by the use of a  $\beta$ -strand contact-filtering protocol. Such contact-guided decoys achieved 14 structure solutions from 21 attempted protein targets, compared with nine for simple *Rosetta* decoys. Previously encountered limitations were superseded in two key respects. Firstly, much larger targets of up to 221 residues in length were solved, which is far larger than the previously benchmarked threshold of 120 residues. Secondly, contact-guided decoys significantly improved success with  $\beta$ -sheet-rich proteins. Overall, the improved performance of contact-guided decoys suggests that MR is now applicable to a significantly wider range of protein targets than were previously tractable, and points to a direct benefit to structural biology from the recent remarkable advances in sequencing.

### 1. Introduction

Molecular replacement (MR) is the most common technique for deriving the lost phase information of the unknown target structure in X-ray crystallography. MR places a structurally similar protein in the unit cell of the unknown target to best reproduce the diffraction data. The correct placement of the similar structure provides the basis for the initial phase calculation of the target. The obtained phasing information and the measured diffraction intensities allow the initial calculation of the electron-density map (Blow & Rossmann, 1961). However, homologues of the target structure do not always exist or may be too structurally distinct. On the other hand, experimental alternatives to MR such as anomalous dispersion (Matthews, 1966; Hendrickson *et al.*, 1985; Wang, 1985) or isomorphous replacement (Green *et al.*, 1954; Perutz, 1956; Blow & Rossmann, 1961) can be time-consuming or difficult to implement in certain cases. These considerations



OPEN ACCESS

have driven recent developments in computational crystallography to derive and trial search models from unconventional sources. Such sources include ideal secondary-structure elements or structural motifs (Rodríguez *et al.*, 2009), libraries of tertiary structural cores derived from mining the Protein Data Bank (PDB; Sammito *et al.*, 2013) and *ab initio* protein structure predictions ('decoys'; Qian *et al.*, 2007; Rigden *et al.*, 2008; Das & Baker, 2009; Bibby *et al.*, 2012; Keegan *et al.*, 2015; Rämisch *et al.*, 2015; Thomas *et al.*, 2015). Clearly, the success of approaches based on *ab initio* protein modelling will depend sensitively on the quality of the structure predictions available.

In recent years, a step change in the accuracy of residue–residue contact predictions (Giraud *et al.*, 1999; Miller & Eisenberg, 2008; Weigt *et al.*, 2009; Burger & van Nimwegen, 2010), based on sequence information alone, has enabled striking advances to be made in structural bioinformatics, including in *ab initio* modelling. Although evolutionary covariance analysis for contact prediction is a research area with a long history (Levitt & Warshel, 1975; Vendruscolo *et al.*, 1997), only recently has the prediction of such contacts become sufficiently accurate to guide *ab initio* structure prediction successfully (Supplementary Fig. S1; Marks *et al.*, 2011; Kosciolok & Jones, 2014; Michel *et al.*, 2014; Adhikari *et al.*, 2015; Ovchinnikov *et al.*, 2015). The basic rationale behind the prediction of residue–residue contacts rests on the existence of strong evolutionary constraints, reflected in the covariation of contacting residues, to maintain functionally important conformations. Such evolutionary constraints can be detected at a sequence level, but thousands of homologous protein sequences are required to detect them. The great challenge of separating direct and indirect residue–residue contacts (direct, A–B and B–C; indirect, A–C) was recently overcome through the use of cooperative ('global') statistical probability models. These approaches not only treat contact pairs independently, but analyse their dependence on each other, thereby increasing the signal-to-noise ratio in the predicted list of contacts or 'contact map' (Giraud *et al.*, 1999; Miller & Eisenberg, 2008; Weigt *et al.*, 2009; Marks *et al.*, 2011). Therefore, contacts with the strongest signal, indicated by the highest global statistical scores, are most likely to represent the true residue interactions in a protein conformation (Marks *et al.*, 2012).

Since the successful separation of direct and indirect contacts, various evolutionary covariance-analysis applications have been developed to increase the accuracy and speed of contact predictions (Balakrishnan *et al.*, 2011; Jeong & Kim, 2012; Jones *et al.*, 2012; Ekeberg *et al.*, 2013, 2014; Kamisetty *et al.*, 2013; Wang & Xu, 2013; Feinauer *et al.*, 2014; Kaján *et al.*, 2014; Schneider & Brock, 2014; Seemayer *et al.*, 2014; Skwark *et al.*, 2014). Broadly, these applications can be divided into three categories depending on the cooperative statistical model implemented to derive evolutionary covariance amongst multiple homologous sequences. The first category employs a pseudo-likelihood maximization model and can be found in applications such as *plmDCA* (Ekeberg *et al.*, 2013; Kamisetty *et al.*, 2013), *GREMLIN* (Kamisetty *et al.*, 2013) or

*CCMpred* (Seemayer *et al.*, 2014). The second and third categories include sparse covariance matrix inversion models such as *PSICOV* (Jones *et al.*, 2012) or mean-field direct coupling analysis models such as *EVFold-mfDCA* (Kaján *et al.*, 2014), respectively. Although these methods differ in accuracy and speed, a recent study (Jones *et al.*, 2015) revealed a high similarity of around 90% in top-ranked contacts amongst those three categories. To capture the full spectrum of top-ranked contacts and produce the best possible contact map, metapredictors such as *PconsC2* (Skwark *et al.*, 2013, 2014) or *MetaPSICOV* (Jones *et al.*, 2015) combine individual contact predictions across two or three of these categories. Additionally, an ongoing aim of prediction tools is to achieve successful detection of evolutionary covariance from homologous sequence sets of reduced diversity and number. Any decrease in the number of homologous sequences required would make covariance analysis applicable to smaller protein families.

One of the principal applications of predicted contact maps is to predict structures for large protein families (Marks *et al.*, 2011; Kosciolok & Jones, 2014; Michel *et al.*, 2014; Adhikari *et al.*, 2015; Ovchinnikov *et al.*, 2015). Without contact information, accurate homology-independent fold predictions for globular proteins in *Rosetta* are limited to chain lengths of up to ~130 residues (Kinch *et al.*, 2011; He *et al.*, 2013; Tai *et al.*, 2014). Several covariance analysis tools, such as *GREMLIN* (Ovchinnikov *et al.*, 2015), *PconsC2* (Michel *et al.*, 2014; Skwark *et al.*, 2014), *MetaPSICOV* (Jones *et al.*, 2015; Kosciolok & Jones, 2015) and *EVFold* (Marks *et al.*, 2011), reported accurate fold predictions for much larger globular proteins, illustrating how the use of contact predictions can greatly expand the capabilities of *ab initio* folding protocols.

The availability of improved tertiary-structure predictions from contact-assisted fragment-assembly *ab initio* modelling naturally enhances the prospects for their use in MR. Broadly speaking, two approaches have been reported. The first entails highly CPU-intensive modelling to produce an overall fold prediction that is sufficiently accurate to serve as a search model in the same way as a crystal structure or homology model would conventionally be deployed (Qian *et al.*, 2007; Das & Baker, 2009). Alternatively, more cheaply obtained, coarse-grained models can be clustered into search-model ensembles and, recognizing their limited accuracy, treated to truncation to attempt to isolate sufficiently accurate core regions (Rigden *et al.*, 2008; Bibby *et al.*, 2012). *AMPLE* (*ab initio* modelling of proteins for molecular replacement; Bibby *et al.*, 2012) is a pipeline that implements the latter strategy and is available in the *CCP4* software suite (Winn *et al.*, 2011). *AMPLE* overcomes the absence of suitable crystal structures or homology models through a cluster-and-truncate approach that processes computationally cheap and minimally refined *Rosetta* (Bibby *et al.*, 2012) or *QUARK* (Keegan *et al.*, 2015) *ab initio* decoys into search models. In brief, 1000 *ab initio* decoys are clustered based on their structural similarity, after which the decoys in the largest resulting cluster (containing a maximum of 200 decoys) are truncated at 20 different intervals. Truncation is rationally guided by inter-decoy structural

variance within the cluster (Qian *et al.*, 2007; Bibby *et al.*, 2012). The truncated decoys are then subclustered under three different  $C^\alpha$  r.m.s.d. radii (1, 2 and 3 Å), whereby a maximum of 30 decoys within the cluster, those closest to the cluster centroid, are selected and combined into an ensemble. Lastly, each ensemble search model undergoes three different side-chain treatments: polyalanine (all side chains are truncated at their  $C^\beta$  atom), reliable side chains (only side chains are kept that are usually well modelled; Shapovalov & Dunbrack, 2007) and all-atom (all side chains are kept). Up to 120 ensemble search models can be obtained per target through this cluster-and-truncate approach, but this number strongly depends on the structural similarity of the initial decoys and the similarity of the decoys after truncation. Each ensemble search model is processed using *MrBUMP* (Keegan & Winn, 2008), which in turn uses *Phaser* (McCoy *et al.*, 2007) and/or *MOLREP* (Vagin & Teplyakov, 1997, 2010) for MR, *SHELXE* (Sheldrick, 2010; Thorn & Sheldrick, 2013) for main-chain tracing and *ARP/wARP* (Cohen *et al.*, 2008) or *Buccaneer* (Cowtan, 2006) for automatic rebuilding of the *SHELXE* trace.

In an initial study of 295 small globular proteins with fewer than 120 residues and resolution better than 2.2 Å, 43% of the targets were solved successfully (Bibby *et al.*, 2012). However, the application of *ab initio* modelling to MR, and therefore the success of *AMPLE*, is greatly limited by the size and fold class of the protein target. These limitations arise at the initial stage during decoy prediction, where successful fold predictions of protein structures without homologues of known structure are currently limited to a chain length of ~130 residues (Kinch *et al.*, 2011; He *et al.*, 2013; Tai *et al.*, 2014). *AMPLE* has been shown to succeed with protein targets greater than 150 amino acids in size, but these cases were not comparable in the nature of the target and/or the methodology employed: the successes were achieved with either anisometric folds (*e.g.* coiled coils), ensemble search models derived from distant structural homologues or NMR structures (Bibby *et al.*, 2013; Bruhn *et al.*, 2014; Hotta *et al.*, 2014; Thomas *et al.*, 2015). Thus, the largest globular protein target previously solved with *AMPLE* using *ab initio* models is 120 amino acids in length (Bibby *et al.*, 2012), although it was noted in that work that success rates had not declined to zero at this size threshold. In addition to the issue of protein size, the success rate of *AMPLE* strongly depends on the fold architecture, as reflected in the widely varying success rates of all- $\alpha$  (80%), mixed  $\alpha$ - $\beta$  (including  $\alpha/\beta$  and  $\alpha+\beta$  folds; 37%) and all- $\beta$  protein targets (2%) in the original test set of 295 small globular proteins. In sum, both the size and the fold of the target can limit *ab initio* folding protocols and thus the success rate of *AMPLE*.

*Ab initio* modelling of proteins without exploiting information from known folds is a longstanding challenge in the field of computational structural biology and success currently strongly depends on the chain length and fold architecture of the target protein. Recent successful advances in the derivation of direct residue-residue contacts from large multiple sequence alignments have greatly increased the accuracy of *ab initio* structure predictions, especially for larger and all- $\beta$  protein targets, which are the greatest challenges for *ab initio*-

based MR approaches. Here, we set out to explore the impact of the improved contact-guided decoys on the success rate of MR. For this, we use our automated pipeline *AMPLE*, as its cluster-and-truncate approach has proven to be highly successful in the downstream processing of *ab initio* decoys for MR (Bibby *et al.*, 2012). We report that contact-guided decoys allow the successful solution of targets that were previously unsolvable using the *AMPLE* method. In addition, we report that combining independently obtained contact maps further improves decoy quality, which in turn extends the tractable MR target range to  $\beta$ -rich proteins.

## 2. Methods

### 2.1. Data set

A test set of 21 globular protein targets was used throughout. They were manually selected to include a range of chain lengths, fold architectures, X-ray diffraction data resolutions and divergent sequence counts in a multiple sequence alignment. The test set covered the three fold classes ( $\alpha$ -helical, mixed  $\alpha$ - $\beta$  and  $\beta$ -sheet) and each target was grouped based on its secondary-structure content as defined by *DSSP* (Kabsch & Sander, 1983; Joosten *et al.*, 2011; Supplementary Table S1). The chain length of the sequences ranged from 62 to 221 residues and each crystal structure contained one molecule in the asymmetric unit. The resolutions of the crystal structures ranged from 1.0 to 2.3 Å. The FASTA sequences of each target, as provided in the PDB entry (Rose *et al.*, 2015), were modelled, rather than the sequence that was visibly present in the crystallographic model. A number of divergent ('effective') sequences ( $N_{\text{eff}}$ ) available for a target of greater than 100 is considered to be the minimum requirement for accurate covariance-based contact predictions (Skwark *et al.*, 2014). The formula  $N_{\text{eff}} = \sum[(1/n_1) + (1/n_2) + \dots + (1/n_i)]$  (Jones *et al.*, 2015) defines  $N_{\text{eff}}$  as the sum of fractional weights of  $n$  sequences in  $i$  clusters in a multiple sequence alignment (MSA). To calculate this parameter for our targets, each target sequence formed the basis of an MSA which was obtained from a database search with *HHblits* v.2.0.15 (Remmert *et al.*, 2012). Two sequence-search iterations were performed with an *E*-value cutoff of  $10^{-3}$  against the nonredundant UniProt20 database v.2013.03 (The UniProt Consortium, 2015). All sequences in each resulting alignment were then clustered using *CD-HIT* v.4.6.3 (Li *et al.*, 2001, 2002; Fu *et al.*, 2012) at 62% sequence identity (Jones *et al.*, 2015) and  $N_{\text{eff}}$  was calculated.

### 2.2. Evolutionary covariance analysis

One contact map was predicted for each of the 21 targets using the fully automated metapredictor *PconsC2* (Skwark *et al.*, 2014). In summary, MSAs were generated with *Jackhammer* (Johnson *et al.*, 2010) against the UniRef100 database and with *HHblits* v.2.0.15 (Remmert *et al.*, 2012) against the non-redundant UniProt20 database v.2013.03 (The UniProt Consortium, 2015) at *E*-value cutoffs of  $10^{-40}$ ,  $10^{-10}$ ,  $10^{-4}$  and 1. Each MSA was then analysed with *PSICOV* (Jones *et al.*,

2012) and *plmDCA* (Ekeberg *et al.*, 2013, 2014) to produce 16 individual sets of contact predictions. All 16 predictions, combined with a secondary-structure prediction, solvent-accessibility information and a sequence profile, were then provided to a deep-learning algorithm (Skwark *et al.*, 2014) to identify protein-like contact patterns. The latter produced a final contact map for each target sequence.

An additional contact map for  $\beta$ -structure-containing targets was predicted using *CCMpred* (Seemayer *et al.*, 2014) and reduced to  $\beta$ -sheet contact pairs using the *CCMpred*-specific filtering protocol *bbcontacts* (Andreani & Söding, 2015). Each MSA for *CCMpred* contact predictions was obtained using *HHblits* v.2.0.15 (Remmert *et al.*, 2012). This entailed two sequence-search iterations with an *E*-value cutoff of  $10^{-3}$  against the nonredundant UniProt20 database v.2013.03 (The UniProt Consortium, 2015) and filtering to 90% sequence identity using *HHfilter* v.2.0.15 (Remmert *et al.*, 2012) to reduce sequence redundancy in the MSA. Besides the contact matrix as input, *bbcontacts* requires a secondary-structure prediction and a factor describing the range of predicted contacts in the MSA. The latter was shown to depend on the sequence count in the MSA ( $N$ ) and the target chain length ( $L$ ). Thus, the factor describing this MSA-specific diversity was calculated using the equation  $\eta = (N/L)^{1/2}$  (Andreani & Söding, 2015). The secondary structure for each sequence was predicted using the *adss.pl* (Remmert *et al.*, 2012) script distributed with *HH-suite* v.2.0.16 (Söding, 2005). Hereafter, the term *bbcontacts* will be used to describe the full process from the target sequence to the filtered  $\beta$ -strand contact map. At no point do contact-prediction algorithms use structural information from structurally characterized proteins.

### 2.3. Conversion of contact maps to contact restraints

For all targets, the predicted contact maps from *PconsC2* were converted to *Rosetta* (Rohl *et al.*, 2004) restraints to guide *ab initio* folding of the target sequences. The *FADE* energy function was used to introduce a restraint in the folding protocol of *Rosetta*. As described in *PconsFold* (Michel *et al.*, 2014), a restraint was satisfied during folding if the participating  $C^\beta$  atoms ( $C^\alpha$  in the case of glycine) were within 9 Å of one another. If a pre-defined contact restraint was satisfied, a smoothed ‘squared-well’ bonus was added to the internal energy scoring function of *Rosetta* during folding. The shape of this function therefore rewards conformations that place residues within 9 Å of each other, but has no influence on the energy outside this range. Thus, a false-positive prediction between two positions that are in fact distant in the target structure will not lead to an undesirable long-distance attraction between the two residues. As defined by Michel *et al.* (2014), the ‘squared-well’ bonus (parameter *wd* in *FADE*) was set to  $-15.00$ . Adopting the same benchmarked approach as Michel *et al.* (2014), only the top  $L$  ranked contacts (based on confidence scores, with  $L$  again representing target length) from each *PconsC2* contact map were selected and converted to *Rosetta* restraints.

For  $\beta$ -containing targets, an alternative selection of predicted contacts, hereafter called *PconsC2+bbcontacts*, was made by a novel combination of *PconsC2* and *bbcontacts* predictions, as follows. Firstly, inter-strand predictions composed of only one or two contacts were removed from the *bbcontacts* contact list owing to their high false-positive rate (Jessica Andreani, personal communication). For all present contact pairs between residues  $i$  and  $j$  and any neighbouring contacts (*i.e.*  $i, j \pm 1$ ;  $i, j \pm 2$ ;  $i \pm 1, j$ ;  $i \pm 2, j$ ) in the top- $L$  *PconsC2* contact list the ‘squared-well’ bonus was doubled from  $-15.00$  to  $-30.00$ , which proved to be the most effective after several options were tried (unpublished data). In addition, all contact pairs solely present in the filtered *bbcontacts* contact map were added to the modified *PconsC2* and *bbcontacts* contact list with a ‘squared-well’ bonus of  $-15.00$ . It is worth noting that the added *bbcontacts* contacts were also present in the full *PconsC2* contact map, although they were not within the top- $L$  cutoff. This approach allowed a strengthening of the weight on  $\beta$ -strand contacts during *ab initio* structure prediction. After uniting the two predictions in this way no further length-based cutoff was applied, so that the *PconsC2+bbcontacts* restraint list fed to *Rosetta* for  $\beta$ -containing proteins might be longer than the simple *PconsC2* list.

The final contact-prediction lists were compared with the corresponding crystal structure contacts to determine their accuracy. For this, all pairs of  $C^\beta$  atoms ( $C^\alpha$  in the case of glycine) within 9 Å of one another in the crystal structure were considered as reference contacts. Predictions were assigned as true or false positives according to whether they were in the list of reference contacts or not. The precision or positive predictive value (PPV) for the each restraint list was then determined using the formula  $PPV = (\text{true positives})/(\text{true positives} + \text{false positives})$ .

### 2.4. *Ab initio* structure prediction of decoys

Fragments were picked for unassisted *Rosetta* modelling with secondary-structure prediction from *PSIPRED* (McGuffin *et al.*, 2000) and for contact-assisted decoys with the secondary-structure prediction obtained during evolutionary covariance analysis. Homologous structures were excluded using the *nohoms* flag to make all experiments equivalent to predictions of unknown folds. Protein decoys were generated using the *AbinitioRelax* folding protocol of *Rosetta* in v.2015wk05 (Rohl *et al.*, 2004). As recommended in the *Rosetta* documentation, special modelling parameters included the helix and loop atom-refinement flags *abinitio::rsd\_wt\_helix* and *abinitio::rsd\_wt\_loop* with a reweight factor of 0.5. The two flags *abinitio::relax* and *relax::fast* were set during *ab initio* modelling to obtain an all-atom refinement using the *Rosetta* full-atom force field. For each target in the data set, the structure folds were predicted under two different restraint conditions: without any residue–residue contact restraints and with *PconsC2*-only contact restraints. Targets containing  $\beta$ -folds were additionally modelled with a third restraint condition: *PconsC2+bbcontacts* contact

**Table 1**

Restraint-guided *ab initio* modelling improves model quality.

Median template-modelling scores for *ab initio* decoys found in the largest cluster (values for all decoys are shown in parentheses) predicted for threefold classes using three different types of residue–residue constraint settings.

Fold classification	<i>Rosetta</i>	<i>PconsC2</i> -only	<i>PconsC2+bbcontacts</i>
All- $\alpha$	0.377 (0.298)	0.609 (0.531)	–
Mixed $\alpha$ - $\beta$	0.314 (0.252)	0.537 (0.433)	0.565 (0.441)
All- $\beta$	0.323 (0.247)	0.467 (0.374)	0.471 (0.381)
Mixed $\alpha$ - $\beta$ + all- $\beta$	0.320 (0.249)	0.506 (0.397)	0.522 (0.403)

restraints derived as described above. *bbcontacts*-only contact restraints were not treated as a separate condition owing to the low count of predicted contacts. A total of 1000 decoys were modelled for each target under each of the three different restraint conditions. Decoy quality was assessed based on the template-modelling score (TM-score; Zhang & Skolnick, 2005), a measure of fold similarity between two structures with identical sequences, in this case a decoy and the corresponding crystal structure. TM-scores range from 0 to 1, with a TM-score above 0.5 usually indicating a correct fold prediction.

### 2.5. Molecular replacement

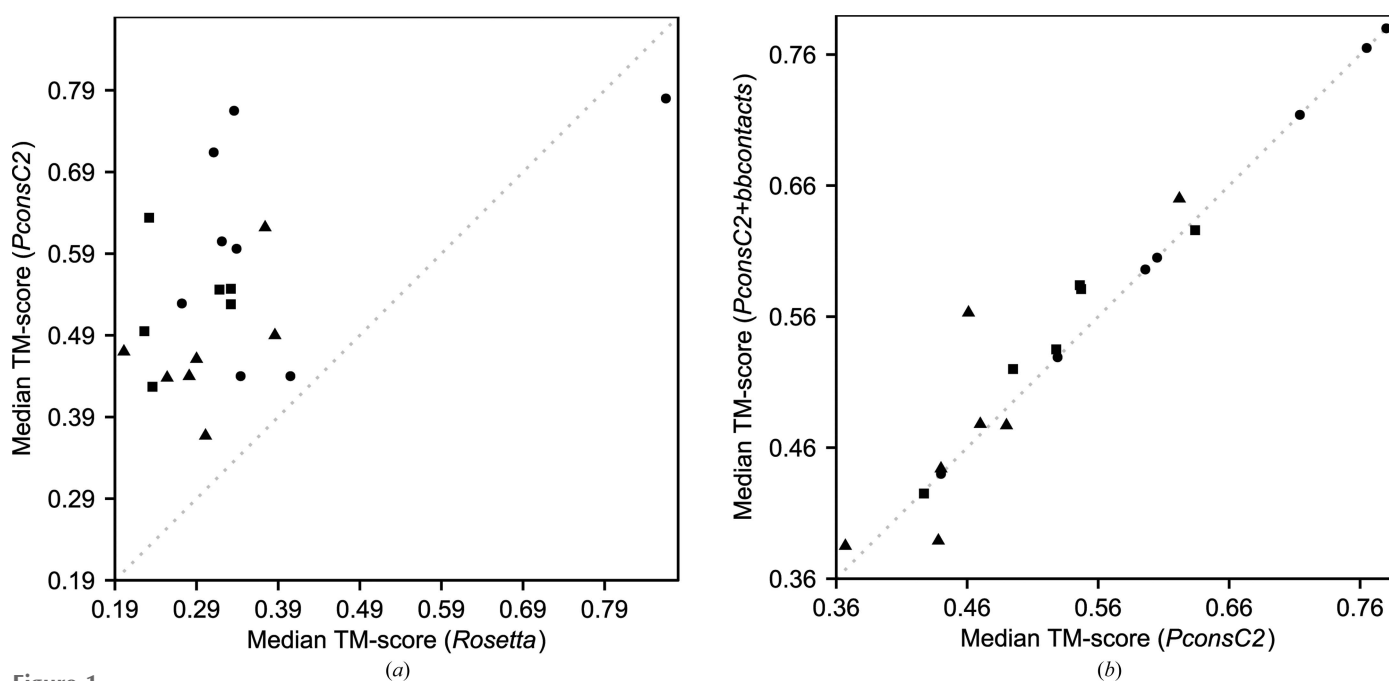
The three sets of *ab initio* decoys for each target were subjected to *SCWRL4* side-chain remodelling (Canutescu *et al.*, 2003; Krivov *et al.*, 2009). Afterwards, all three sets of decoys for each target were run in the automated MR pipeline *AMPLE* v.1.0 using default parameters, with the exception of the number of clusters to trial, which was changed from one to three (Bibby *et al.*, 2012). The associated structure-factor

amplitudes for the crystal structure of each target were retrieved from the PDB. The correct placement of search models by *Phaser* was assessed using the recently developed residue-independent overlap (RIO) score (Thomas *et al.*, 2015). In short, the RIO score assesses the in-sequence and out-of-sequence register overlap of the placed search-model residues (fragments of at least three residues) with the corresponding crystal structure. To be considered a success, MR using *AMPLE* was required to give a *SHELXE* correlation coefficient (CC) of  $\geq 25.00$  and an average chain length (ACL) of  $\geq 10.00$  (Sheldrick, 2010; Thorn & Sheldrick, 2013; Keegan *et al.*, 2015). Additionally, structure rebuilding of the *SHELXE* chain traces was attempted using both *ARP/wARP* (Cohen *et al.*, 2008) and *Buccaneer* (Cowtan, 2006) and an  $R_{\text{free}}$  value of  $\leq 0.45$  from either method was required.

## 3. Results

### 3.1. Both general and $\beta$ -strand-specific residue–residue contact maps improve *ab initio* protein structure predictions

The initial part of this study evaluates the use of contact restraints for improving *ab initio* protein structure prediction. For each protein target in the data set, 1000 decoys were predicted using *Rosetta*, either unassisted or with restraints deriving from *PconsC2* alone or from our novel fusion of *PconsC2* and *bbcontacts* predictions. Since *AMPLE* by default processes decoys from the largest cluster to create search models, structure quality was primarily assessed for these structure predictions alone. However, we also report overall improvements for all decoys (Table 1).

**Figure 1**

Residue–residue contact restraints improve *ab initio* model quality. (a) Median template-modelling scores (TM-scores) for *Rosetta* decoys plotted against median TM-scores for *PconsC2*-only coupling-guided decoys. (b) Median TM-score for *PconsC2*-only coupling-guided decoys plotted against median TM-scores for *PconsC2+bbcontacts* decoys (Skwark *et al.*, 2014; Andreani & Söding, 2015). Median TM-scores derived from decoys found in the largest cluster. The symbol shapes correspond to the three different fold classes: all- $\alpha$  (circles), all- $\beta$  (triangles) and mixed  $\alpha$ - $\beta$  (squares).

**Table 2**

Summary of *ab initio* structure-prediction results of  $\beta$ -structure-containing targets.

All data shown are for *PconsC2+bbcontacts* (*PconsC2*-only) guided decoys.

Fold classification	PDB code	No. of effective sequences	No. of contacts	PPV	TM-score	
					Top-cluster decoys	1000 decoys
Mixed $\alpha$ - $\beta$	1aba	1037	92 (87)	0.787 (0.782)	0.584 (0.546)	0.507 (0.496)
	1chd	852	222 (203)	0.924 (0.931)	0.626 (0.634)	0.501 (0.528)
	1e0s	1831	184 (174)	0.691 (0.713)	0.520 (0.495)	0.362 (0.353)
	1eaz	1060	136 (125)	0.928 (0.944)	0.581 (0.547)	0.512 (0.460)
	1lo7	1026	146 (141)	0.980 (0.986)	0.535 (0.528)	0.453 (0.443)
	1tjx	1189	178 (159)	0.857 (0.887)	0.425 (0.427)	0.354 (0.358)
	All- $\beta$	1bdo	940	91 (80)	0.913 (0.963)	0.477 (0.490)
1kjl		272	183 (146)	0.704 (0.727)	0.385 (0.367)	0.313 (0.293)
1npu		943	136 (117)	0.835 (0.940)	0.389 (0.438)	0.322 (0.331)
1pnc		887	111 (99)	0.830 (0.889)	0.478 (0.470)	0.375 (0.354)
2nuz		1048	70 (62)	0.901 (0.952)	0.650 (0.622)	0.540 (0.498)
3w56		949	146 (131)	0.896 (0.906)	0.444 (0.440)	0.363 (0.353)
4u3h		1226	109 (100)	0.911 (0.950)	0.563 (0.461)	0.440 (0.446)

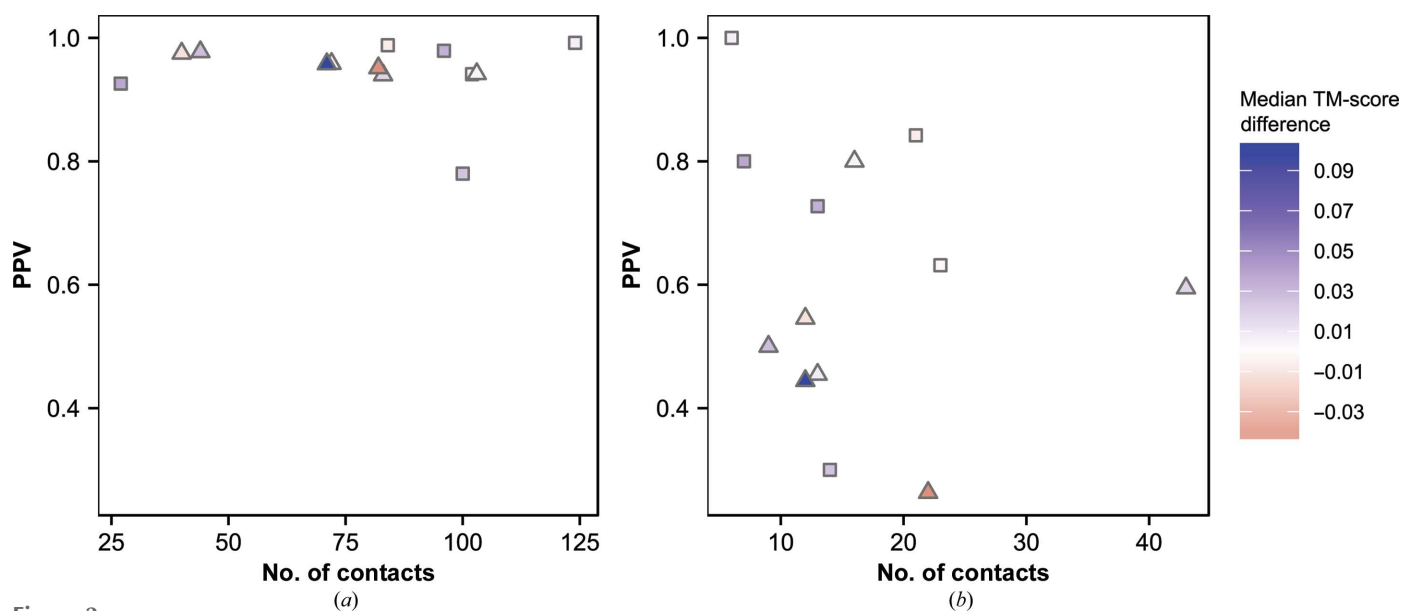
A high degree of sequence diversity is a prerequisite for the identification of residue covariance in MSAs. The range of effective sequences in the MSAs of target sequences ranged from 272 to 1831. Typically, higher numbers of effective sequences correlate with more accurate contact predictions (Jones *et al.*, 2012; Kamisetty *et al.*, 2013; Ekeberg *et al.*, 2014; Skwark *et al.*, 2014; Ma *et al.*, 2015). Here, similar results were observed, as illustrated in Supplementary Fig. S2. Considering the three fold classes, more accurate predictions were obtained for  $\beta$ -structure-containing proteins (median  $PPV_{\beta}$  = 0.940; median  $PPV_{\alpha-\beta}$  = 0.909) compared with all- $\alpha$  targets (median  $PPV_{\alpha}$  = 0.655). In plain language, for  $\beta$ -structure-containing proteins over 90% of the intramolecular contacts predicted by evolutionary covariance methods are indeed present in the crystal structure. Although untested in this

study, a higher accuracy for  $\beta$ -sheet-containing proteins is achieved owing to the regular pattern of contact pairs that is easily detectable in a contact map (Skwark *et al.*, 2014; Andreani & Söding, 2015). The deep-learning procedure used during the final step of *PconsC2* filters these contact pairs better, therefore increasing the overall accuracy of the prediction.

As expected (Michel *et al.*, 2014; Skwark *et al.*, 2014), the inclusion of *PconsC2*-predicted contact information substantially improved the quality of structure predictions. A simple *Rosetta* run without contact information yielded a largest cluster median TM-score of 0.342 for all 21 protein targets,

compared with 0.542 for *PconsC2*-only decoys (Fig. 1a). 20 of 21 targets were modelled better, with median TM-score improvements ranging from 0.035 to 0.429. A single target, PDB entry 2qyj, the unassisted *Rosetta* models for which were already of exceptionally high quality (median TM-score of largest cluster structure predictions of 0.865), was modelled slightly worse (0.780) when contact information was included.

For 13  $\beta$ -strand-containing proteins in the data set we developed a novel approach of combining the top-*L* predicted *PconsC2* (Skwark *et al.*, 2014) contacts with the filtered  $\beta$ -sheet-specific *bbcontacts* (Andreani & Söding, 2015) contacts. This procedure resulted in the upweighting of some contacts already present in the *PconsC2* list and the addition of others. The number of contacts affected in each category is shown in Fig. 2. At least 80% of upweighted contact restraints



**Figure 2**

Effects of upweighting and addition of  $\beta$ -sheet-specific contacts on contact accuracy and decoy quality. Number of (a) upweighted and (b) added *bbcontacts*  $\beta$ -sheet-specific contact restraints for 13  $\beta$ -sheet-containing (seven all- $\beta$ , triangles; six mixed  $\alpha$ - $\beta$ , squares) targets plotted against their corresponding positive predictive value (PPV). The colour fill of each point corresponds to the resulting difference in median TM-scores between the largest cluster decoys from *PconsC2*-only and *PconsC2+bbcontacts* decoy sets (positive values favour the latter).

(present in the final contact lists of *PconsC2* and *bbcontacts* predictions) proved to be true positives (Fig. 2*a*). The average PPV of upweighted contacts was 95% (Fig. 2*a*). The quality of the added contacts was generally lower (Fig. 2*b*), but nine of the 13 targets had a PPV of at least 50% (Fig. 2*b*). Thus, at the cost of the inclusion of some false-positive contacts, our approach generally provides extra valuable information for the folding process.

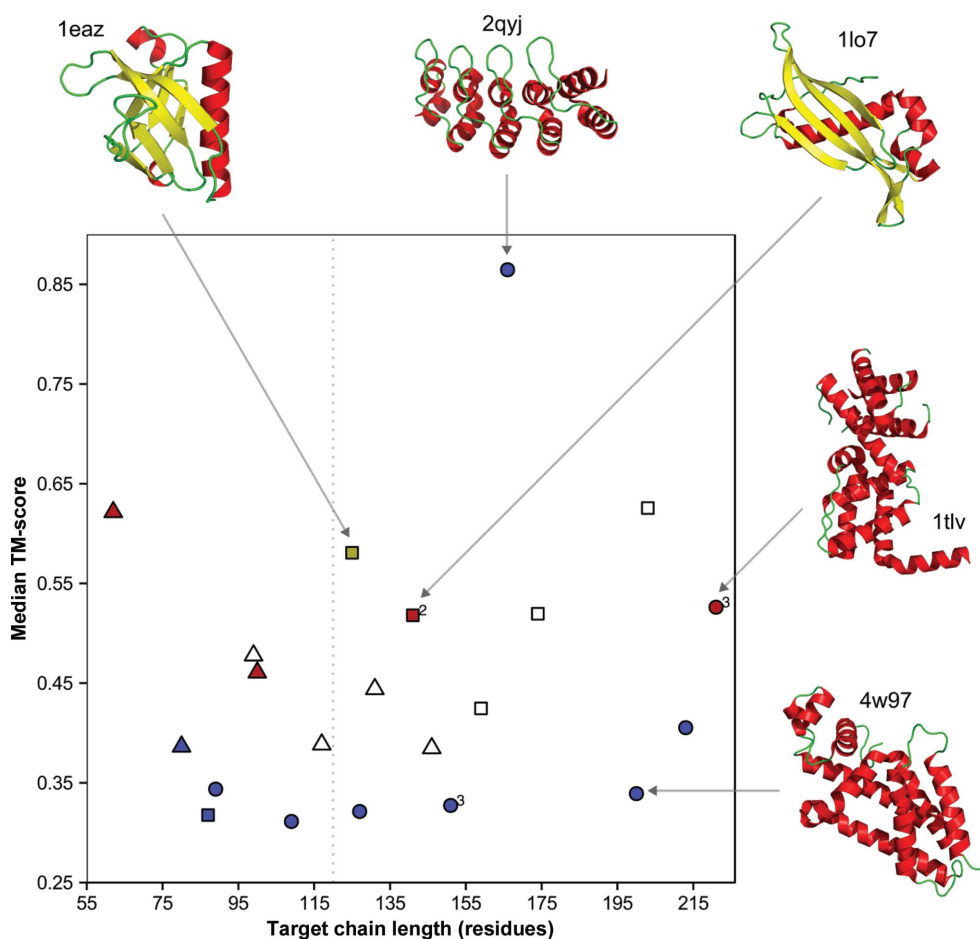
Models based on *PconsC2+bbcontacts* contacts were again somewhat improved compared with those built using the *PconsC2* contacts: the median TM-scores for the two model sets were 0.522 and 0.506, respectively (Fig. 1*b*). Model quality improved for nine targets, of which five showed improvements in median TM-score of at least 0.02 (Table 2). Model quality

deteriorated for four targets, but for three of these the difference was very small: less than 0.02 (Table 2).

### 3.2. Contact-guided *ab initio* models extend the tractable target range of *AMPLE*

With a demonstrable and significant improvement in decoy quality evident from the use of predicted contact restraints, the ability of *AMPLE* to solve the 21 protein targets using contact-guided decoys was then tested. For all targets, two sets of decoys were trialled deriving from *ab initio* structure prediction with no contact-prediction restraints or with *PconsC2*-only restraints. For  $\beta$ -strand-containing targets a third decoy set was created using *PconsC2+bbcontacts*-derived restraints. Structure solution of each target was attempted with these sets of structure predictions using default *AMPLE* methods. Successful MR was detected as previously by the ability of *SHELXE* main-chain tracing and density modification run on the MR placement to reach a correlation coefficient (CC) of  $\geq 25$  with a mean traced chain length of  $\geq 10$  residues (Sheldrick, 2010; Bibby *et al.*, 2012; Thorn & Sheldrick, 2013; Keegan *et al.*, 2015; Thomas *et al.*, 2015). As previously (Keegan *et al.*, 2015; Thomas *et al.*, 2015), we further required that an  $R_{\text{free}}$  value of  $\leq 0.45$  could be achieved after *ARP/wARP* (Cohen *et al.*, 2008) or *Buccaneer* (Cowtan, 2006) automatic rebuilding of the resulting *SHELXE* chain traces.

Based on these stringent success criteria, the default algorithm of *AMPLE* achieved eight structure solutions for decoys predicted without contact restraints (Fig. 3, blue). Six out of eight all- $\alpha$ , one mixed  $\alpha$ - $\beta$  and one all- $\beta$  target were solved with chain lengths up to 213 residues. Success using *ab initio* models has not been previously reported for such large globular protein targets (Table 3), but these findings recapitulate the fold-class preferences observed previously: *AMPLE* works well for all- $\alpha$  targets but less so for mixed  $\alpha$ - $\beta$  and particularly all- $\beta$  proteins (Bibby *et al.*, 2012). Previously,



**Figure 3**

Contact restraint-guided *ab initio* models extend the tractable target range of *AMPLE*. Molecular-replacement (MR) success mapped against target chain length and median template-modelling score (TM-score). The point shape corresponds to the fold class of the target: all- $\alpha$  (circles), all- $\beta$  (triangles) and mixed  $\alpha$ - $\beta$  (squares). The point colour indicates successful structure solutions for the contact constraints used: none (blue), *PconsC2*-only (red) and *PconsC2+bbcontacts* (gold). Points for successful solutions were considered in the order of *Rosetta*, *PconsC2*-only and *PconsC2+bbcontacts* decoys. In cases of unsuccessful molecular-replacement attempts (empty symbols), TM-scores for the largest clusters of *PconsC2+bbcontacts* decoys are shown. Median TM-scores for each point correspond to the largest decoy cluster (compared with the crystal structure), leading to a structure solution (cluster indices given next to each point for targets that were not solved with the largest cluster). The dashed grey line highlights the tested target chain-length limit of *AMPLE* (120 residues) for globular proteins (Bibby *et al.*, 2012). Cartoon representations of crystal structures of five different targets exemplify the diversity of structure solutions (PDB identifiers are provided next to each crystal structure).  $\alpha$ -Helices are shown in red,  $\beta$ -sheets in yellow and loops in green.

**Table 3**

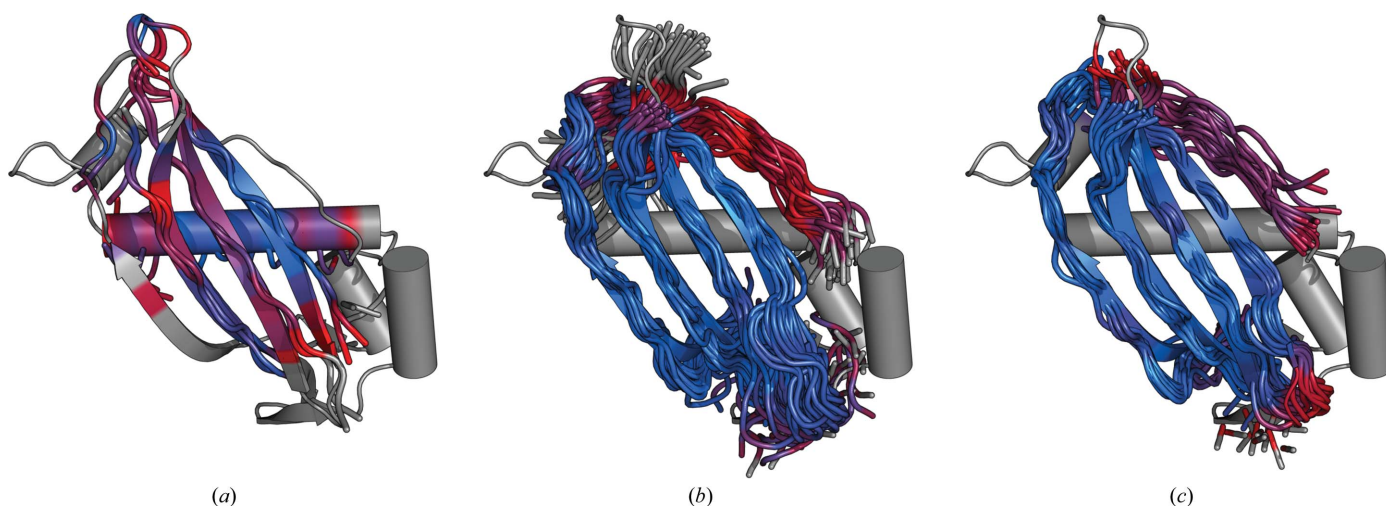
Summary of molecular-replacement solutions of 21 protein targets.

The total number of ensemble search models derived from *ab initio* decoys from the three largest clusters is provided in parentheses after the individual number of successful search models.

Fold classification	PDB code	Resolution (Å)	Target chain length	No. of successful (total) search models		
				<i>Rosetta</i>	<i>PconsC2</i> -only	<i>PconsC2+bbcontacts</i>
All- $\alpha$	1kw4	1.75	89	137 (393)	101 (468)	—
	1bkr	1.10	109	21 (105)	13 (459)	—
	4cl9	1.40	127	1 (210)	1 (408)	—
	1a6m	1.00	151	1 (102)	4 (327)	—
	2qyj	2.05	166	378 (501)	329 (453)	—
	4w97	1.60	200	6 (114)	3 (399)	—
	1hh8	1.80	213	3 (66)	0 (297)	—
	1tlv	1.95	221	0 (18)	2 (399)	—
	Mixed $\alpha$ - $\beta$	1aba	1.45	87	4 (312)	58 (429)
1eaz		1.40	125	0 (135)	0 (345)	28 (327)
1lo7		1.50	141	0 (120)	3 (327)	3 (333)
1tjx		1.04	159	0 (27)	0 (165)	0 (150)
1e0s		2.28	174	0 (15)	0 (195)	0 (207)
1chd		1.75	203	0 (12)	0 (279)	0 (225)
All- $\beta$	2nuz	1.85	62	0 (393)	76 (444)	183 (453)
	1bdo	1.80	80	27 (343)	16 (381)	19 (372)
	1pnc	1.60	99	0 (126)	0 (300)	0 (297)
	4u3h	1.98	100	0 (273)	14 (357)	1 (372)
	1npu	2.00	117	0 (111)	0 (210)	0 (180)
	3w56	1.60	131	0 (129)	0 (123)	0 (150)
	1kjl	1.40	146	0 (63)	0 (174)	0 (201)

the strong performance on all- $\alpha$  targets has been at least partly attributed to the greater accuracy of *Rosetta* modelling of those proteins (Bibby *et al.*, 2012), but these results make a second important contribution more explicit. Most all- $\alpha$  targets were solved despite the overall accuracy of their models being poor (Fig. 3). This suggests that their success in MR, nevertheless, lies with the superior ability of *SHELXE* to autotrace helices compared with other secondary structures (Sheldrick, 2010). The  $\beta$ -containing targets solved here were

*PconsC2+bbcontacts* decoys achieved all of the structure solutions using *PconsC2*-only decoys. Additionally and most notably, *PconsC2+bbcontacts* decoys led to the structure solution of the mixed  $\alpha$ + $\beta$  PH domain of the human TAPP1 protein (PDB entry 1eaz; Fig. 3, gold). The structure solution of 1eaz uniquely using *PconsC2+bbcontacts* restraints highlights the importance of the fusion of contact maps developed here. In total, the largest cluster decoys modelled with *PconsC2* and *bbcontacts* restraints solved 11 out of 21 targets.



**Figure 4**

Contact restraints improve the search-model quality of  $\beta$ -strand-containing targets. Structural superposition of the (a) *Rosetta* ( $C^\alpha$  r.m.s.d. 2.814 Å; ensemble contains two structures), (b) *PconsC2*-only ( $C^\alpha$  r.m.s.d. 1.748 Å; 30 members) and (c) *PconsC2+bbcontacts* ( $C^\alpha$  r.m.s.d. 1.760 Å; 15 members) search-model ensembles for 4-hydroxybenzoyl CoA thioesterase (PDB entry 1lo7). Examples are the highest scoring search models based on *SHELXE* CC score, with only (b) and (c) leading to successful structure solutions. Search models are shown as tubes and crystal structures as cartoons. (a) and (c) are 50% of the target sequence, while (b) is 55%. The colour scale illustrates the pairwise  $C^\alpha$  r.m.s.d. between each search-model ensemble (represented by its first member) and the crystal structure, with blue representing the minimum  $C^\alpha$  r.m.s.d. and red the maximum. Unaligned residues are coloured grey.



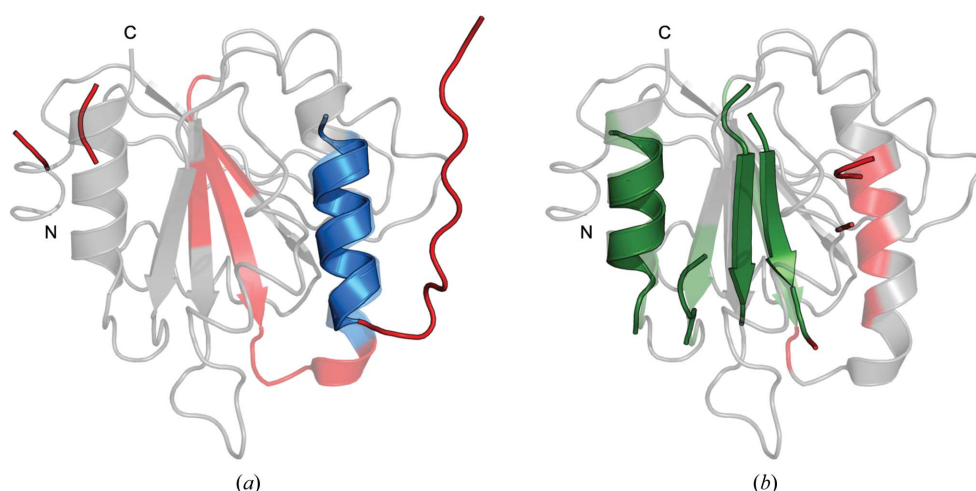
By default, the *AMPLE* algorithm processes *ab initio* models solely from the largest cluster. When trialling ensemble search models based on decoys from the three largest clusters, an additional three structure solutions were obtained (Fig. 3). The successful solution of haem-bound oxymyoglobin (PDB entry 1a6m) was achieved with *Rosetta* decoys from the third cluster. Notably, *ab initio* modelling of this target was performed without its large bound haem group, yet structure solution was achieved. The second target, the inactive LicT PRD domain from *Bacillus subtilis* (PDB entry 1tlv), was solved with *PconsC2*-only decoys from the third cluster. This all- $\alpha$  target with a chain length of 221 residues is, to our knowledge, the largest globular protein to be solved using search models derived from *ab initio* protein structure modelling. As mentioned above, all- $\alpha$  targets such as this benefit from the powerful helix tracing in *SHELXE* (Sheldrick, 2010), as do the programs of the *ARCIMBOLDO* suite (Rodríguez *et al.*, 2009), which can also solve large all- $\alpha$  protein structures (Fourati *et al.*, 2014). Lastly, *PconsC2*-only and *PconsC2+bbcontacts* decoys derived from second largest clusters yielded search models that solved the 4-hydroxybenzoyl CoA thioesterase domain structure (PDB entry 1lo7). Particularly notable about this solution is the topology of the search models. Although this mixed  $\alpha+\beta$  target contains a number of helices, the best structure solutions (based on *SHELXE* CC scores) were obtained from search models containing the accurately modelled, central, four-stranded  $\beta$ -sheet (Fig. 4). This accurate modelling, which is required for successful MR, was only achieved with the guidance of contact restraints. In total, the addition of these three structure solutions results in 14 out of 21 structure solutions for *PconsC2+bbcontacts* decoys compared with nine for simple *Rosetta* decoys.

Although the stringent criteria of MR success used here did not indicate a successful structure solution for target 1e0s, the beneficial effect of including joint *PconsC2+bbcontacts* contact predictions was evident in the search-model placement as assessed by RIO scores (Fig. 5). For the top *PconsC2*-only search model, 40% (12 residues) of the search-model residues were correctly superimposed, albeit out of register (blue) on the target structure (*Phaser* TFZ = 4.7, *Phaser* LLG = 16). For the top *PconsC2+bbcontacts* search model, 77% (30 residues) of the search model were superimposed in an in-register fashion (*Phaser* TFZ = 5.3, *Phaser* LLG = 17) (Fig. 5). For the latter, expert manual intervention might allow structure determination, but in this case the correct solution was not prominent in the list of MR placements. Nevertheless, it is clear that even when overall structure solution was not automatically achieved the *PconsC2+bbcontacts* model provided better results which might be recoverable as successes in the future as MR and post-MR software improves still further.

Within the range explored, the success of structure solution did not appear to depend significantly on the resolution of the available crystallographic data (Supplementary Fig. S3). Successful targets ranged in resolution from 1.00 to 2.05 Å (mean  $\pm$  standard deviation of  $1.62 \pm 0.32$  Å), while unsuccessful targets spanned 1.04–2.28 Å ( $1.67 \pm 0.40$  Å). The solvent content of the protein crystals appeared to have a modest impact on MR success. Targets with successful structure solutions ranged from 36.0 to 55.3% (mean  $\pm$  SD of  $46.1 \pm 5.2\%$ ) solvent content compared with 25.8–48.0% ( $39.1 \pm 8.1\%$ ) for unsuccessful targets.

Given that the inclusion of predicted contact information is a significant change to the modelling protocol, we re-examined the performance and importance of the key features of the operation of *AMPLE*. A detailed analysis of the characteristics of the successful search models is provided in the Supporting Information; only a summary is provided here.

*AMPLE* uses well established clustering of decoys (Simons *et al.*, 1997) to pick out those likely to be the most accurate. This continues to be effective here, as picking the largest clusters selects better than average decoys from the sets available (Supplementary Fig. S4) and there is a good correlation between the largest cluster size and the median TM-score of the decoys in that cluster (Supplementary Fig. S5). However, the size of the largest cluster does not correlate well with the total number of successful search models (Supplementary Fig. S6). *AMPLE* also relies on rational, variance-based truncation to trim ensembles



**Figure 5**

*PconsC2+bbcontacts* contact-derived models give a more accurate MR placement for PDB entry 1e0s which, although not solved automatically, might yield to expert manual intervention. Top *Phaser* solutions of 1e0s based on RIO scores for (a) *PconsC2*-only (RIO score 12) and (b) *PconsC2+bbcontacts* (RIO score 30) search models for target 1e0s. Search-model colour coding indicates useful superposition of residues by in-sequence (green) or out-of-sequence register (blue) residues as well as misplaced (red) residues. The addition of *bbcontacts* restraints produced a more accurate model with correctly placed  $\beta$ -strands that was placed correctly. Both structures are shown in cartoon representation with the crystal structure shown as a transparent cartoon. Unaligned reference crystal structure residues are coloured grey.

down to more accurate core structures, with the size range 15–40 residues found to be most successful (Bibby *et al.*, 2012). Here, the truncation is further validated (Fig. 6) and a similar mapping of success onto search-model size is observed (Supplementary Fig. S7). Fig. 6 further illustrates the overall positive impact of contacts on accuracy: note the larger number of low-r.m.s.d. ensembles on the right of Fig. 6(b) compared with Fig. 6(a). However, Fig. 6 also illustrates that targets that are already well modelled by simple *Rosetta* and successful in MR (blue points on the right in Fig. 6a) can be modelled slightly worse when contact information is included (somewhat raised r.m.s.d.s in Fig. 6b), presumably owing to the influence of false-positive contact predictions. Previously, we have found that sampling across three subclustering radii and three modes of side-chain treatment were both required for solution of the largest possible number of targets (Bibby *et al.*, 2012). This remains largely the case in the current exercise, as

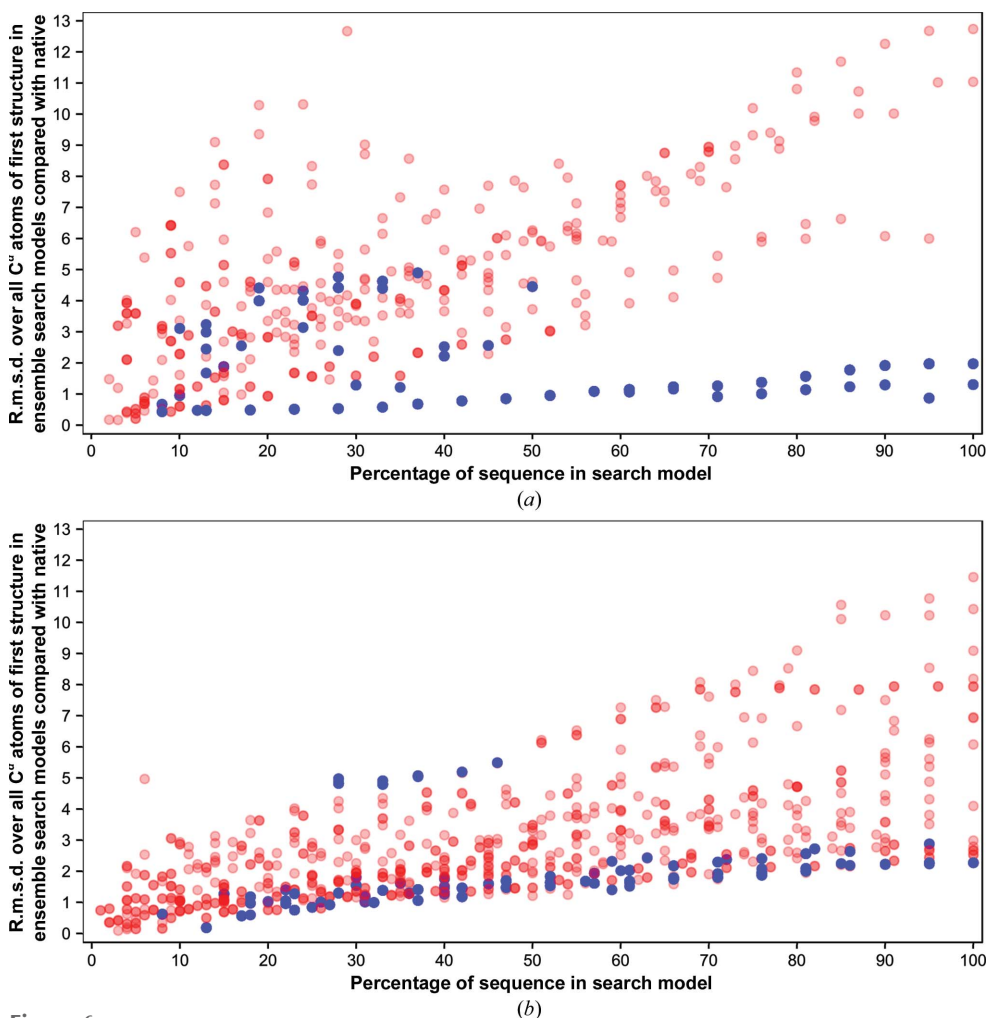
unique solutions were obtained for each of the subclustering radii (Supplementary Table S1). Polyalanine side-chain search models were the most successful, but a single target, PDB entry 1eaz, was only solved using one of the alternative treatments (Supplementary Table S1).

#### 4. Discussion

The recently emerged ability to predict contacting residues from large protein sequence alignments is one of the most exciting developments in structural bioinformatics for many years. The key statistical breakthrough allowing the disentangling of predicted direct contacts (Giraud *et al.*, 1999; Miller & Eisenberg, 2008; Weigt *et al.*, 2009; Marks *et al.*, 2011), *i.e.* neighbouring amino acids from pairs of residues whose identities covary indirectly, has been followed by a wave of papers not only dealing with the accuracy of predictions but

also considering the manifold applications of the information. Predicted contact information is of immediate benefit to crystallographers in many ways that are yet to be fully appreciated, including parsing of domains for structural analysis (Rigden, 2002; Sadowski, 2013) and interpretation of crystal structure composition (Nicoludis *et al.*, 2015). Here, we considered how the better protein *ab initio* models that can be produced by exploiting information can serve as a source of improved search models for MR. We use the MR pipeline *AMPLE* as a convenient and effective tool for the analysis.

Challenged by the lesser success of *AMPLE* with  $\beta$ -structure-containing proteins (Bibby *et al.*, 2012), and motivated by the accuracy improvements in *ab initio* fold predictions through contact restraints (Marks *et al.*, 2011; Michel *et al.*, 2014; Jones *et al.*, 2015; Ovchinnikov *et al.*, 2015), we developed a new approach for combining predicted contact-restraint lists from *PconsC2* (Skwark *et al.*, 2014) and *bbcontacts* (Andreani & Söding, 2015) to elevate the *ab initio* modelling accuracy of  $\beta$ -structure protein targets. Structure predictions guided by the resulting *PconsC2+bbcontacts* contact restraints improved the decoy quality for nine out of 13



**Figure 6**

Variance-based truncation remains an effective way to derive successful search models from higher quality contact restraint-assisted *ab initio* decoys. The percentage of sequence in the search model is mapped against the root-mean-square deviation (r.m.s.d.) over all C $\alpha$  atoms of the first representative of each search-model ensemble derived from the largest cluster against the native structure. Successful structure solutions of individual search models are highlighted in blue and unsuccessful solutions in red. Progressively darker shades of either colour correspond to increasing numbers of overlapping points. Progressive truncation is shown for (a) *Rosetta* decoys and (b) *PconsC2+bbcontacts* decoys (or *PconsC2*-only decoys for all- $\alpha$  targets).

$\beta$ -structure-containing protein targets. Our approach, which involved both selective upweighting of and addition to the *PconsC2* set, based on the specialist  $\beta$ -sheet predictions, may well be of more general use to the protein-modelling community. The value of these contact-guided *ab initio* models for structure solution by MR of targets treated as novel folds is demonstrated. Nine of the 21 targets in the data set were solved using the *AMPLE* algorithm to process unassisted *Rosetta* structure predictions. This number rose to 14 using contact-guided modelling. The 100% success rate for all- $\alpha$  targets is highly encouraging and, along with the comparable MR pipeline *ARCIMBOLDO* (Rodríguez *et al.*, 2009), graphically illustrates the power of the  $\alpha$ -helical tracing implemented in *SHELXE* (Sheldrick, 2010) and the relative tractability of  $\alpha$ -rich targets to unconventional MR. More  $\beta$ -rich,  $\alpha$ -poor targets are harder for both *AMPLE* and *ARCIMBOLDO*, so our demonstrable advances with these targets, leveraging the value of contact restraints during *ab initio* modelling, are exciting. The fusion of top-ranked *PconsC2* and *bbcontacts* contacts developed as part of this study proved to be a key part in one successful structure solution, further highlighting the importance of the approach.

The size of the targets solved is another notable feature of this work. We originally suggested (Bibby *et al.*, 2012) that all- $\alpha$  protein targets larger than the 120-residue threshold then tested could be suitable for the cluster-and-truncate approach of *AMPLE*. Here, we demonstrate this to be true, with unassisted decoys leading to the solution of a 213-residue protein and contact-assisted models leading to the successful solution of a 221-residue chain. To our knowledge, these are the largest targets to be solved with search models derived from *ab initio* structure decoys.

The availability of reliable contact restraints to aid MR with *ab initio* models clearly widens the range of targets for which *AMPLE* is a viable option for structure solution. The accuracy of contact predictions is directly related to the number of protein sequences deposited in sequence databases such as UniProt (The UniProt Consortium, 2015), and thus will benefit from the continuous growth of those databases. Notably, this manuscript focused solely on globular proteins; yet the *AMPLE* algorithm is equally well suited to coiled-coil and transmembrane proteins (Thomas *et al.*, 2015 and unpublished data). Specific contact predictors for the latter are available (Wang *et al.*, 2011; Hopf *et al.*, 2012; Yang *et al.*, 2013; Zhang *et al.*, 2016) and future research will explore their application to MR using *AMPLE*. In conclusion, the current and future broadening of the target range tractable by *AMPLE* through the use of evolutionary restraints during *ab initio* modelling highlights the value of the software as an effective alternative to experimental phasing approaches in X-ray crystallography.

In summary, we confirm here that predicted contacts can significantly improve *ab initio* model quality in a way that directly impacts on structure solution by MR. Our novel mode of uniting general and  $\beta$ -structure-specific contact predictions brings further tangible model improvements to the particularly difficult  $\beta$ -rich protein targets. All of these methodo-

logical advances have immediate benefits for crystallographers facing targets with novel or divergent folds which cannot be addressed by conventional MR. *AMPLE* proves to be an efficient framework for rendering these contact-assisted decoys into search-model ensembles, with truncation and extensive sampling remaining key to success. Future inevitable expansion of sequence databases and predictable improvements in contact-prediction software will undoubtedly extend the reach of MR with *ab initio* models still further.

## 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 and JT are supported by BBSRC DTP PhD scholarships.

## References

- Adhikari, B., Bhattacharya, D., Cao, R. & Cheng, J. (2015). *Proteins*, **83**, 1436–1449.
- Andreani, J. & Söding, J. (2015). *Bioinformatics*, **31**, 1729–1737.
- Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. (2011). *Proteins*, **79**, 1061–1078.
- Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Cryst.* **D68**, 1622–1631.
- Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2013). *Acta Cryst.* **D69**, 2194–2201.
- Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195–1202.
- Bruhn, J. F., Barnett, K. C., Bibby, J., Thomas, J. M. H., Keegan, R. M., Rigden, D. J., Bornholdt, Z. A. & Saphire, E. O. (2014). *J. Virol.* **88**, 758–762.
- Burger, L. & van Nimwegen, E. (2010). *PLoS Comput. Biol.* **6**, e1000633.
- Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L. (2003). *Protein Sci.* **12**, 2001–2014.
- Cohen, S. X., Ben Jelloul, M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J., Sixma, T. K., Lamzin, V. S., Murshudov, G. N. & Perrakis, A. (2008). *Acta Cryst.* **D64**, 49–60.
- Cowtan, K. (2006). *Acta Cryst.* **D62**, 1002–1011.
- Das, R. & Baker, D. (2009). *Acta Cryst.* **D65**, 169–175.
- Ekeberg, M., Hartonen, T. & Aurell, E. (2014). *J. Comput. Phys.* **276**, 341–356.
- Ekeberg, M., Lökvist, C., Lan, Y., Weigt, M. & Aurell, E. (2013). *Phys. Rev. E*, **87**, 012707.
- Feinauer, C., Skwark, M. J., Pagnani, A. & Aurell, E. (2014). *PLoS Comput. Biol.* **10**, e1003847.
- Fourati, Z., Roy, B., Millan, C., Coureux, P.-D., Kervestin, S., van Tilbeurgh, H., He, F., Usón, I., Jacobson, A. & Graille, M. (2014). *J. Mol. Biol.* **426**, 3689–3702.
- Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012). *Bioinformatics*, **28**, 3150–3152.
- Giraud, B. G., Heumann, J. M. & Lapedes, A. S. (1999). *Phys. Rev. E*, **59**, 4983–4991.
- Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. Lond. A Math. Phys. Sci.* **225**, 287–307.
- He, Z., Alazmi, M., Zhang, J. & Xu, D. (2013). *PLoS One*, **8**, e74006.
- Hendrickson, W. A., Smith, J. L. & Sheriff, S. (1985). *Methods Enzymol.* **115**, 41–55.
- Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C. & Marks, D. S. (2012). *Cell*, **149**, 1607–1621.
- Hotta, K., Keegan, R. M., Ranganathan, S., Fang, M., Bibby, J., Winn, M. D., Sato, M., Lian, M., Watanabe, K., Rigden, D. J. & Kim, C.-Y. (2014). *Angew. Chem. Int. Ed.* **53**, 824–828.
- Jeong, C.-S. & Kim, D. (2012). *Protein Eng. Des. Sel.* **25**, 705–713.

- Johnson, L. S., Eddy, S. R. & Portugaly, E. (2010). *BMC Bioinformatics*, **11**, 431.
- Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. (2012). *Bioinformatics*, **28**, 184–190.
- Jones, D. T., Singh, T., Kosciolk, T. & Tetchner, S. (2015). *Bioinformatics*, **31**, 999–1006.
- Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hoofst, R. W. W., Schneider, R., Sander, C. & Vriend, G. (2011). *Nucleic Acids Res.* **39**, D411–D419.
- Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
- Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S. & Rost, B. (2014). *BMC Bioinformatics*, **15**, 85.
- Kamisetty, H., Ovchinnikov, S. & Baker, D. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 15674–15679.
- Keegan, R. M., Bibby, J., Thomas, J., Xu, D., Zhang, Y., Mayans, O., Winn, M. D. & Rigden, D. J. (2015). *Acta Cryst.* **D71**, 338–343.
- Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* **D64**, 119–124.
- Kinch, L., Yong Shi, S., Cong, Q., Cheng, H., Liao, Y. & Grishin, N. V. (2011). *Proteins*, **79**, 59–73.
- Kosciolk, T. & Jones, D. T. (2014). *PLoS One*, **9**, e92197.
- Kosciolk, T. & Jones, D. T. (2015). *Proteins*, doi:10.1002/prot.24863.
- Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Jr (2009). *Proteins*, **77**, 778–795.
- Levitt, M. & Warshel, A. (1975). *Nature (London)*, **253**, 694–698.
- Li, W., Jaroszewski, L. & Godzik, A. (2001). *Bioinformatics*, **17**, 282–283.
- Li, W., Jaroszewski, L. & Godzik, A. (2002). *Bioinformatics*, **18**, 77–82.
- Ma, J., Wang, S., Wang, Z. & Xu, J. (2015). *Bioinformatics*, **31**, 3506–3513.
- Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. & Sander, C. (2011). *PLoS One*, **6**, e28766.
- Marks, D. S., Hopf, T. A. & Sander, C. (2012). *Nat. Biotechnol.* **30**, 1072–1080.
- Matthews, B. W. (1966). *Acta Cryst.* **20**, 230–239.
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
- McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). *Bioinformatics*, **16**, 404–405.
- Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S. & Elofsson, A. (2014). *Bioinformatics*, **30**, i482–i488.
- Miller, C. S. & Eisenberg, D. (2008). *Bioinformatics*, **24**, 1575–1582.
- Nicoludis, J. M., Lau, S.-Y., Schärfe, C. P. I., Marks, D. S., Weihofen, W. A. & Gaudet, R. (2015). *Structure*, **23**, 2087–2098.
- Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., Kamisetty, H., Grishin, N. V. & Baker, D. (2015). *eLife*, **4**, e09248.
- Perutz, M. F. (1956). *Acta Cryst.* **9**, 867–873.
- Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature (London)*, **450**, 259–264.
- Rämisch, S., Lizatović, R. & André, I. (2015). *Acta Cryst.* **D71**, 606–614.
- Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2012). *Nat. Methods*, **9**, 173–175.
- Rigden, D. J. (2002). *Protein Eng.* **15**, 65–77.
- Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* **D64**, 1288–1291.
- Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nat. Methods*, **6**, 651–653.
- Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. (2004). *Methods Enzymol.* **383**, 66–93.
- Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Green, R. K., Goodsell, D. S., Westbrook, J. D., Woo, J., Young, J., Zardecki, C., Berman, H. M., Bourne, P. E. & Burley, S. K. (2015). *Nucleic Acids Res.* **43**, D345–D356.
- Sadowski, M. I. (2013). *Proteins*, **81**, 253–260.
- Sammito, M., Millán, C., Rodríguez, 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. & Usón, I. (2013). *Nat. Methods*, **10**, 1099–1101.
- Schneider, M. & Brock, O. (2014). *PLoS One*, **9**, e108438.
- Seemayer, S., Gruber, M. & Söding, J. (2014). *Bioinformatics*, **30**, 3128–3130.
- Shapovalov, M. V. & Dunbrack, R. L. Jr (2007). *Proteins*, **66**, 279–303.
- Sheldrick, G. M. (2010). *Acta Cryst.* **D66**, 479–485.
- Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). *J. Mol. Biol.* **268**, 209–225.
- Skwark, M. J., Abdel-Rehim, A. & Elofsson, A. (2013). *Bioinformatics*, **29**, 1815–1816.
- Skwark, M. J., Raimondi, D., Michel, M. & Elofsson, A. (2014). *PLoS Comput. Biol.* **10**, e1003889.
- Söding, J. (2005). *Bioinformatics*, **21**, 951–960.
- Tai, C.-H., Bai, H., Taylor, T. J. & Lee, B. (2014). *Proteins*, **82**, 57–83.
- The UniProt Consortium (2015). *Nucleic Acids Res.* **43**, D204–D212.
- Thomas, J. M. H., Keegan, R. M., Bibby, J., Winn, M. D., Mayans, O. & Rigden, D. J. (2015). *IUCrJ*, **2**, 198–206.
- Thorn, A. & Sheldrick, G. M. (2013). *Acta Cryst.* **D69**, 2251–2256.
- Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
- Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* **D66**, 22–25.
- Vendruscolo, M., Kussell, E. & Domany, E. (1997). *Fold. Des.* **2**, 295–306.
- Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
- Wang, X.-F., Chen, Z., Wang, C., Yan, R.-X., Zhang, Z. & Song, J. (2011). *PLoS One*, **6**, e26767.
- Wang, Z. & Xu, J. (2013). *Bioinformatics*, **29**, i266–i273.
- Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. (2009). *Proc. Natl Acad. Sci. USA*, **106**, 67–72.
- Winn, M. D. *et al.* (2011). *Acta Cryst.* **D67**, 235–242.
- Yang, J., Jang, R., Zhang, Y. & Shen, H.-B. (2013). *Bioinformatics*, **29**, 2579–2587.
- Zhang, H., Huang, Q., Bei, Z., Wei, Y. & Floudas, C. A. (2016). *Proteins*, **84**, 332–348.
- Zhang, Y. & Skolnick, J. (2005). *Nucleic Acids Res.* **33**, 2302–2309.