Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

AMPLE: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed *ab initio* models

Jaclyn Bibby,[a] Ronan M. Keegan,[b] Olga Mayans,[a] Martyn D. Winn[c] and Daniel J. Rigden[a]*

[a]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, England, [b]Research Complex at Harwell, STFC Rutherford Appleton Laboratory, Didcot OX11 0FA, England, and [c]Science and Technology Facilities Council Daresbury Laboratory, Daresbury WA4 4AD, England

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

Received 20 July 2012
Accepted 13 September 2012

Protein *ab initio* models predicted from sequence data alone can enable the elucidation of crystal structures by molecular replacement. However, the calculation of such *ab initio* models is typically computationally expensive. Here, a computational pipeline based on the clustering and truncation of cheaply obtained *ab initio* models for the preparation of structure ensembles is described. Clustering is used to select models and to quantitatively predict their local accuracy, allowing rational truncation of predicted inaccurate regions. The resulting ensembles, with or without rapidly added side chains, solved 43% of all test cases, with an 80% success rate for all-α proteins. A program implementing this approach, *AMPLE*, is included in the *CCP*4 suite of programs. It only requires the input of a *FASTA* sequence file and a diffraction data file. It carries out the modelling using locally installed *Rosetta*, creates search ensembles and automatically performs molecular replacement and model rebuilding.

1. Introduction

Molecular replacement (MR) remains a popular and important means of solving the phase problem as it allows the rapid determination of crystal structures from native X-ray diffraction data. From the beginning of 2011 to the time of writing, it accounted for almost 80% of submissions to the Protein Data Bank (PDB). MR requires the availability of a protein structure that is sufficiently similar to the target to allow its placement, in rotational and translation terms, within the asymmetric unit. This placement produces approximate phasing information, which is often of good enough quality to allow manual or automatic refinement of the model. Traditionally, the protein structure used for the MR search will be a homologous structure or a homology model of the target. Pipelines are available that allow convenient automatic identification, preparation and processing of such search models (Keegan & Winn, 2008; Schwarzenbacher *et al.*, 2008; Long *et al.*, 2008). The success of MR depends on the latter being sufficiently similar to the target. However, in many cases no homologous structure is available or those available are evolutionarily too distant. In order to address these situations, *ab initio* (or *de novo*) models are increasingly becoming considered as search models in molecular replacement. *Ab initio* modelling attempts to predict three-dimensional structures of proteins in the absence of any guidance from known homologous structures. Current *ab initio* methods such as *Rosetta* (Shortle *et al.*, 1998; Simons *et al.*, 1997, 1999), *I-TASSER* (Lee & Skolnick, 2007; Roy *et al.*, 2010; Wu *et al.*, 2007; Zhang, 2008) and *QUARK* (Xu & Zhang, 2012) work initially with a reduced representation of the protein assembled from suitable structural fragments obtained from experimental structures deposited in the PDB and proceed to a complete all-atom representation. At the fragment-assembly stage, thousands of models (termed 'decoys') are clustered and centroid representatives of large top clusters are considered as candidate fold predictions. The appearance of a large top cluster at this stage is generally indicative of reliable modelling (Shortle *et al.*, 1998). Side chains may then be added to selected decoys and the results refined under a more realistic physics-based force field than is applied during fragment assembly (Xu & Zhang, 2012). The CPU time required to reach the initial fold stage is modest (less than 24 h for sequences up to 120 residues), but the all-atom second stage can be highly demanding depending on the sampling regime. For *Rosetta*, the package that is used here, supercomputers or distributed computing resources are typically required for the second stage (Bradley *et al.*, 2005; Das *et al.*, 2007).

Use of *ab initio* models in MR has followed two tracks. Firstly, an intensive all-atom modelling approach has been employed to produce single search models of maximum completeness and accuracy which first demonstrated the potential of *ab initio* modelling for MR (Qian *et al.*, 2007) and later solved around one third of a small test set of 30 cases (Das & Baker, 2009). However, the computational demands of this approach (around 100 CPU days per case in the latter study) place it out of reach of typical crystallography laboratories, despite attempts to reduce these demands (Shrestha *et al.*, 2011). More recently, we have demonstrated the feasibility of a more economical approach in a small-scale pilot study, which showed that polyalanine 'decoys' resulting from the early fragment-assembly step in *Rosetta* could be assembled into successful MR search models (Rigden *et al.*, 2008; Caliandro *et al.*, 2009). For this, the most accurate decoys were selected by reference to the crystal structure and bespoke processed by combinations of clustering into ensembles, with rapid side-chain addition and the truncation of inaccurate portions. Crucially, we, like others (Qian *et al.*, 2007), identified a correlation between structural diversity in the decoy set and deviation from the native structure, in principle allowing inaccurate regions to be rationally predicted and removed.

We present here a large-scale assessment of the suitability of cheaply obtained *ab initio* decoys for automatic MR of crystal structures of small proteins. Our previous *ad hoc* processing has been replaced by automatic steps for generating a set of ensembles for use as MR search models. The ensembles are automatically truncated to various degrees based on local structure diversity and processed to contain all, selected or no side chains. In contrast to the prevailing approach attempting to generate a single, complete and universally accurate search model, our method attempts to extract and combine features which are likely to have reliable phasing power from the thousands of decoys produced by *ab initio* modelling. Such features may only represent a small part of the protein target. We benchmarked the method against 295 test cases, resulting in a success rate of ~43%. To facilitate the broad adoption of our method, we have implemented it as the software *AMPLE* (**ab** *initio* **m**odelling of **p**roteins for mo**le**cular replacement). *AMPLE* constitutes an automatic pipeline for the computation of suitable *ab initio* model ensembles, their trialling in MR and subsequent rebuilding of the target structure. Thereby, it allows routine, cost-effective structure solution of the crystal structures of small proteins.

2. Materials and methods

2.1. The test set of structures

A test set of 295 proteins was selected from the PDB (Rose *et al.*, 2011) of X-ray protein structures containing 40–120 residues that were determined to 2.2 Å resolution or better (Supplementary Table S1[1]). Structures that contained bound metal ions, nucleic acids or modified residues were excluded and suitable data quality ($R \leq 0.25$, $R_{\rm free} \leq 0.35$) was required. This set was further filtered using *PISCES* (Wang & Dunbrack, 2005) to eliminate sequence redundancy to a 5% level. The sequence modelled, of 40–120 residues, was that provided as a *FASTA* sequence by the PDB, which does not necessarily exactly match the experimentally visualized protein structure.

2.2. Preparation of models

Coarse-grained decoys were rapidly generated by the *Rosetta* (Shortle *et al.*, 1998; Simons *et al.*, 1997, 1999) *ab initio* protocol followed by rapid side-chain addition and minimization by the fast-relax application. Visual inspection of some decoys suggested that they were compacted when compared with the native structure: the side-chain addition was an attempt to deal with this issue. For each sequence, fragment libraries were generated informed by secondary-structure prediction with *PSIPRED* (Jones, 1999). PDB-derived fragments with greater than 5% sequence identity were excluded in order to treat each target as a novel fold. For each target 1000 decoys were produced, with the modelling time, which was principally dependent on length, being no more than one CPU day per case (the CPUs used in this work were Intel Xeon E5540 or E5640 processors running at 2.53 or 2.67 GHz, respectively). This computation time is easily achievable for a single-core desktop available to any crystallography laboratory. In order to take advantage of the specialized side-chain addition tool *SCWRL* (Canutescu *et al.*, 2003; Krivov *et al.*, 2009), with its backbone-dependent rotamer libraries, the side chains that were added during the rapid-relax procedure of *Rosetta* were remodelled using *SCWRL*.

The 1000 decoys were clustered with *SPICKER* (Zhang & Skolnick, 2004) and the three largest resulting clusters were selected for processing. To determine whether the largest cluster generated by *SPICKER* contained the most accurate decoys, each decoy was compared with the deposited structure using Global Distance Test (GDT) total scores calculated with

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: TZ5014). Services for accessing this material are described at the back of the journal.

the *Local–Global Alignment* (*LGA*) program (Zemla, 2003). The rank percentile of the most native-like model was determined as its ordered position in the population.

In order to implement rational truncation, for each of the top three clusters the 200 decoys most similar to the cluster centroid as output by *SPICKER* were structurally aligned in *THESEUS* and the residues were ranked along the chain by structural variance. In only a few cases did the largest cluster include more than 200 decoys (Supplementary Fig. S1). Truncation was carried out by incremental removal of the most divergent residues at intervals of 5% of the sequence length. At least 20 truncated ensembles were thus generated, but more were generated for some cases as the number of residues to be trimmed is rounded down to the nearest integer. For example, a 100-residue protein would be trimmed in increments of five residues, resulting in 20 truncated ensembles (per *SPICKER* cluster) with 100, 95, 90, …, 5 residues. A 99-residue protein would be trimmed by four residues at each step, giving 25 truncated ensembles of 99, 95, 91, …, 3 residues. Overall, the number of truncated ensembles ranged from 20 to 29 for different targets.

Each truncated ensemble was then further subclustered to derive ensembles with different degrees of structural heterogeneity. An r.m.s.d. comparison between all members of each truncated ensemble was carried out using *MAXCLUSTER*. For each decoy, all other decoys that were similar within 1 Å were collected into the subcluster centred on that decoy. The largest of these subclusters, which must contain more than one decoy to proceed, was structurally aligned using *THESEUS* and proceeded to side-chain treatment. This process is repeated with 2 and 3 Å thresholds, typically generating three search models per truncated ensemble, but sometimes generating fewer where no subcluster contains more than one decoy. If there were more than 30 structures in a subcluster then it was cut down to the first 30 structures.

Three side-chain treatments were then employed for each resulting subcluster. Two of these were the ensembles proceeding as polyalanine (or glycine) or with all side chains retained. The third option, modelling only more reliable side chains, trimmed the side chains previously placed by *SCWRL* back to C^{β} for Met, Asp, Pro, Gln, Lys, Arg, Glu and Ser: these side chains are statistically harder to place reliably (Shapovalov & Dunbrack, 2007). The three different modes resulted in around 180 ensembles per *SPICKER* cluster.

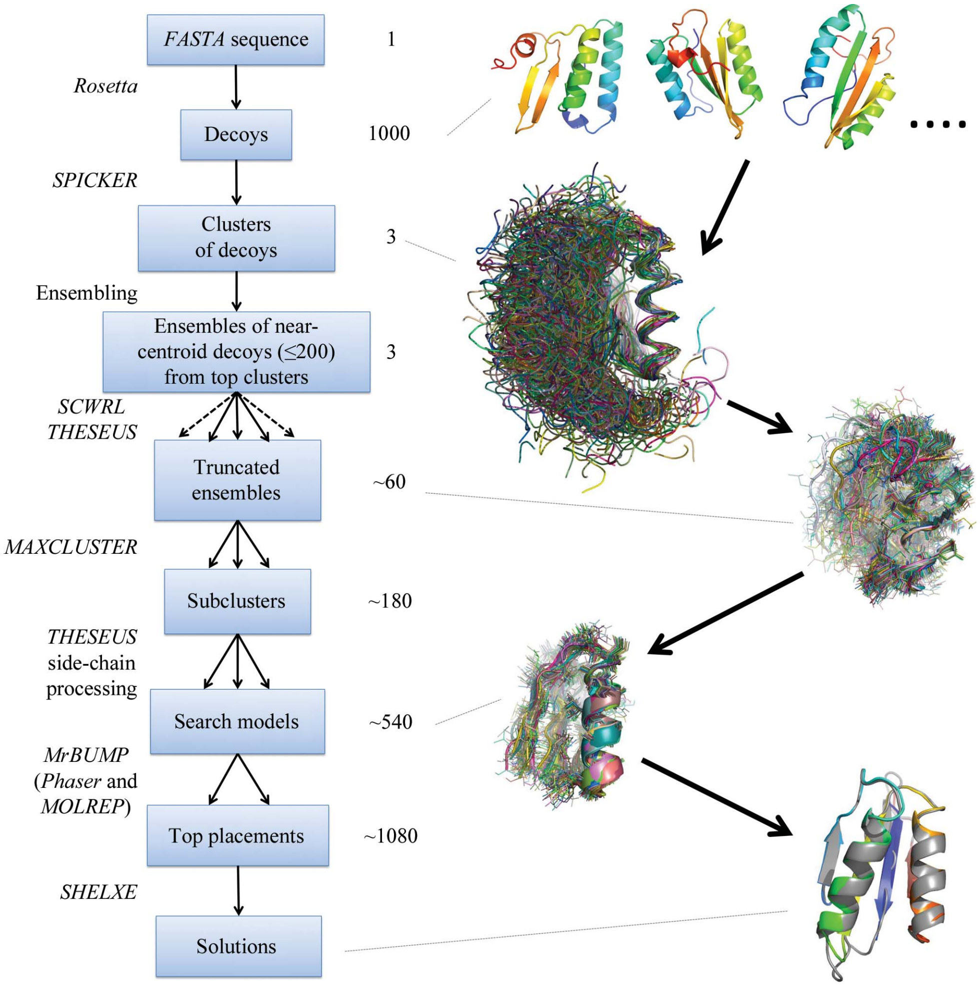

Figure 1
Flowchart illustrating the processing of input structures (*ab initio* decoys in this work) into MR search models and their subsequent trialling. Programs employed by *AMPLE* at various steps are given on the left. Steps not in italic text are carried out internally by *AMPLE*. The numbers to the right of the flowchart indicate typical numbers at different steps: each target will generate three *SPICKER* clusters, resulting in around 60 truncated ensembles and eventually the trialling of around 540 search models. On the right are snapshots during the successful solution of target 3hz7. The largest *SPICKER* cluster consists of highly structurally diverse decoys, although a helix and following strands (less visible) are shared between them. The truncated ensemble in the same orientation shows greater structural homogeneity and has side chains placed with *SCWRL*, but still contains decoys that are too diverse to use as an ensemble MR search model. The search model shown results from subclustering, in this case using an r.m.s.d. threshold of 2 Å, and side-chain treatment; here, the retention of only more reliably predicted side chains. The behaviour of the top *Phaser* placement in *SHELXE* (an increase of the CC to a value of 49%) was indicative of success; indeed, the structure output by *SHELXE* (coloured blue to red from the N-terminus to the C-terminus) is almost complete and very largely correct compared with the deposited crystal structure (grey).

2.3. Molecular replacement

All ensembles generated were fed into *MrBUMP* (Keegan & Winn, 2008) using its default protocol (a version of *MrBUMP* later than or including that distributed in *CCP*4 v.6.3.0 is required) and MR was carried out by both *Phaser* (Storoni *et al.*, 2004; McCoy *et al.*, 2005, 2007) and *MOLREP* (Vagin & Teplyakov, 2010). Default parameters were used for these methods; the inputs comprise the number of molecules in the asymmetric unit predicted by *MrBUMP*, the resulting estimated molecular weight of the asymmetric unit and estimated error values of 1 Å (*Phaser*) or SIM = 1 (*MOLREP*).

Only the top solution from each of these programs, represented by the first member of each placed ensemble, was analysed. The accuracy of its placement was measured with the *REFORIGIN* program of the *CCP*4 package (Winn *et al.*, 2011) using C$^{\alpha}$ atoms only. A stricter criterion of success was then used, namely the ability to proceed to a refined structure from the MR solution. This is reliably predicted by the behaviour of a putative solution upon rapid tracing in *SHELXE* (Usón *et al.*, 2007; Sheldrick, 2010): 15 runs of density modification each composed of 20 cycles followed by main-chain tracing. When the CC score rises and achieves levels above 25–30 the solution can generally be assumed to be correctly rebuilt (http://strucbio.biologie.uni-konstanz.de/ccp4wiki/index.php/SHELX_C/D/E; Rodríguez *et al.*, 2012). We additionally required that the average chain length after *SHELXE* should be at least ten residues (Tim Gruene, personal communication). For comparison, $R_{\rm free}$ values were measured after *SHELXE* using *REFMAC* (Murshudov *et al.*, 2011). We verified that all but one case (a partial success) could be automatically rebuilt with *ARP*/*wARP* (Langer *et al.*, 2008; Cohen *et al.*, 2008) and/or *Buccaneer* (Cowtan, 2006) to an $R_{\rm free}$ value of below 45%.

3. Results and discussion

3.1. Generation of *ab initio* search models by clustering and truncation

We have developed a method for the processing of *ab initio* generated models that applies clustering and truncation to derive a set of ensembles to be used as search models in MR. For clarity, we will hereafter refer to the *ab initio* model structures as decoys, reserving the term model for search models destined for MR attempts. Our methodology is outlined in Fig. 1 and is further described in §2. Briefly, 1000 decoys were generated for each target, built from fragments of unrelated proteins *ab initio*, and side chains were placed with *SCWRL* (Canutescu *et al.*, 2003; Krivov *et al.*, 2009). The decoys were clustered with *SPICKER* (Zhang & Skolnick, 2004) and only decoys from the largest three clusters were taken further. It is well established that the scoring function in the initial fragment-assembly step cannot reliably pick out native-like decoys; the largest cluster(s) were instead treated as broad fold predictions. The top three clusters were subsequently treated separately. The 200 decoys nearest to the centroid of each cluster were structurally aligned using a maximum-likelihood algorithm implemented in the program *THESEUS* (Theobald & Wuttke, 2006) that can effectively downweight variable regions and thereby reveal any structurally conserved core that may be present. This alignment results in a variance score along the target sequence. As mentioned above, our observations and those of others (Qian *et al.*, 2007) show that high-variance regions are often inaccurately modelled. Based on this variance score, our approach computes 20 or so derivative clusters by progressively eliminating the 5% of sequence with the highest variance values, then the next highest 5%, and so on until 95% deletion (see §2). These truncated clusters are then reclustered in ordered to generate three ensembles containing 2–30 decoys each. These values were considered to be optimal, as preliminary experiments showed that ensembles containing more decoys slowed down the MR step unduly. Each resulting

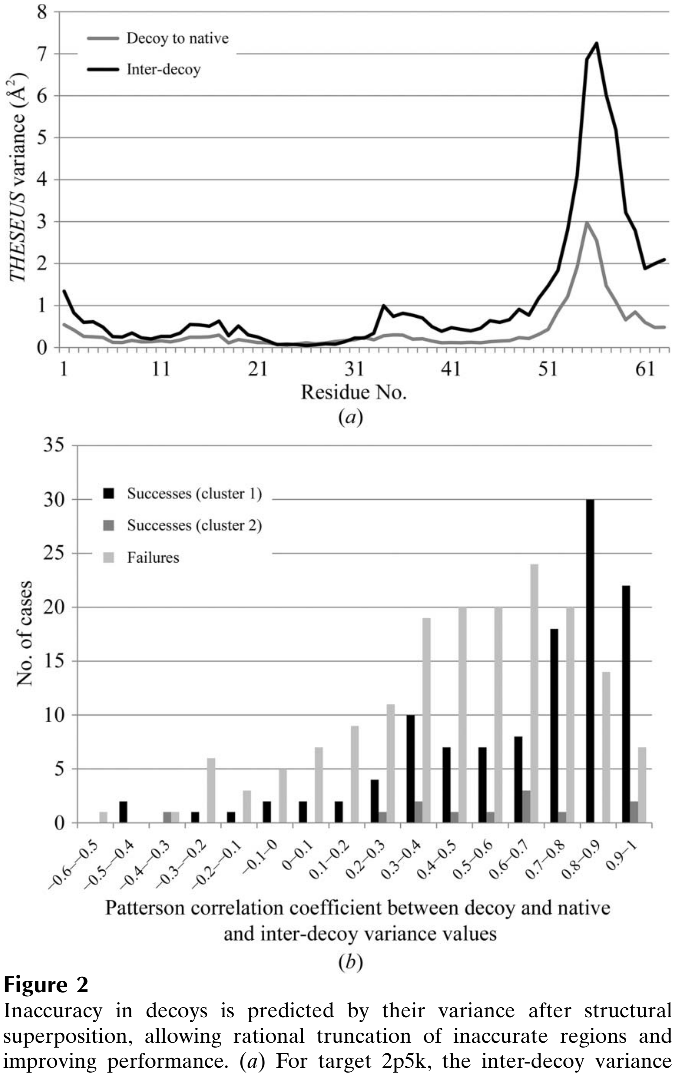

Figure 2
Inaccuracy in decoys is predicted by their variance after structural superposition, allowing rational truncation of inaccurate regions and improving performance. (*a*) For target 2p5k, the inter-decoy variance (dark trace) mirrors the mean variance between each decoy and the deposited native structure (light trace). (*b*) Successes and failures *versus* bins of Patterson correlation coefficients (PCCs) that measure the correlation between inter-decoy and native–decoy variance values. Most successes, mainly deriving from *SPICKER* cluster 1 (black bars) and a few from cluster 2 (dark grey), originate in clusters of decoys for which the PCCs are high. Failing targets are indicated by light grey bars and PCC values derived from their respective top *SPICKER* clusters: the distribution of failure is shifted leftwards, indicating that the PCCs were generally lower and local inaccuracy was thus less well predicted, impacting on the search-model processing.

ensemble is then treated in three ways with respect to side chains: (i) all side chains are trimmed back to C^{β}, (ii) all remain or (iii) only a subset that are generally most reliably predicted (Shapovalov & Dunbrack, 2007) are kept, with the others being trimmed to C^{β}. The final set of ensembles, 360–540 for a given target, is then used for MR, with the results sent to *SHELXE* (Usón *et al.*, 2007; Sheldrick, 2010) for tracing. To test the efficiency of this method, we derived a nonredundant test set of proteins of 40–120 residues from the PDB (Rose *et al.*, 2011) determined to 2.2 Å resolution or better with $R < 0.25$ and $R_{\rm free} < 0.35$ and for which diffraction data have been deposited (PDB codes for all test cases are provided in Supplementary Table S1). Complexes with nucleic acids or metal ions were eliminated since these ligands cannot presently be considered in protein *ab initio* modelling yet can strongly influence protein conformation.

For our method to work best, our selected decoys to be transformed into search models should be among the most accurate produced by *Rosetta.* To determine whether this was the case, each decoy selected from the largest three clusters (the 200 nearest the cluster centroid for each cluster) was scored and ranked against the deposited structure using the Global Distance Test (GDT) total score calculated by the *Local–Global Alignment* (*LGA*) program (Zemla, 2003). This measure is commonly used in *ab initio* modelling as a measure of model accuracy, balancing local and global comparisons. In 44% of cases the top cluster contained a model better than the 99th percentile, *i.e.* one of the best ten models in the population. The equivalent figures for the second and third largest clusters were 31 and 23%, respectively. More importantly, in 97% of cases the top cluster contained a model better than the 70th percentile. Evidently, *SPICKER* clustering serves as a computationally inexpensive way to select a subset from the initial 1000 decoys that is enriched in more accurate structures.

Also key to the success of our method is the ability to automatically identify and remove inaccurate regions. This can be achieved since the variability between decoys in each cluster correlates with their deviation from the deposited native structure (Qian *et al.*, 2007). This is exemplified here by the test case 2p5k, in which the C-terminal region could be predicted from inter-decoy variability as the least reliably modelled portion (Fig. 2*a*). Fig. 2(*b*) further shows this relationship, represented as a Patterson correlation coefficient (PCC), across all 295 targets. The PCC value varies widely, but is clearly biased towards high values, indicating the good predictive value of inter-decoy variability for inaccuracy. For example, the PCC is greater than 0.7 for 39% of the 295 cases. The effect of progressively greater truncation on two examples is shown in Supplementary Fig. S2. The importance of being able to identify inaccurate regions is suggested by the high proportion of successes that were achieved when the PCC between inter-model variability and deviation from native structure was high (Fig. 2*c*); successes only outweighed failures when PCC > 0.8. Although not the principal focus of this work, we suggest that the variable/inaccurate regions (generally loops and termini) are those for which the existence of fewer intramolecular contacts leads to a less well defined (broader, shallower) energy well in the vicinity of the native conformation. This would naturally reduce the degree of structural homogeneity in these regions as fewer decoys would sample and be retained in the near-native energy basin. In some cases it may also be true that the local structural variability within clusters faithfully reflects the lack of a single preferred conformation in the native structure, *e.g* regions that would not have well defined electron density.

3.2. Performance of processed *ab initio* model clusters in MR

We tested the performance of our cluster-and-truncate approach on a test set of 295 proteins selected from the PDB. As explained, the decoys of each target were processed into 360–540 search models. Initially, we assessed whether MR successfully placed the search model by comparison with the available crystal structure using the *CCP*4 program *REFORIGIN* (Winn *et al.*, 2011). However, a more stringent measure of success is the ability of a putative solution to be automatically rebuilt. We used the *SHELXE* CC score for a partially traced structure against the native data (Usón *et al.*, 2007; Sheldrick, 2010) for this purpose. It has been reported that solutions for which the CC score, after a rapid procedure of density modification and main-chain tracing, increased and attained levels of 25–30% could consistently be automatically rebuilt (http://strucbio.biologie.uni-konstanz.de/ccp4wiki/index.php/SHELX_C/D/E; Rodríguez *et al.*, 2012). We additionally imposed a second criterion, an average chain length of greater than ten residues after *SHELXE* tracing, which improves the selection of solutions which may be automatically refined (Tim Gruene, personal communication). Conversely, a CC score that reduced and remained below this threshold can reliably be discounted as incorrect. Our ensembles were partitioned very cleanly into two groups by *SHELXE* CC score (Supplementary Fig. S3). We classified ensembles with CC > 25% and a mean chain length of >10 residues as successes and verified that all but one case (for which the density indicated partial success) could be automatically rebuilt with *ARP*/*wARP* (Langer *et al.*, 2008; Cohen *et al.*, 2008) and/or *Buccaneer* (Cowtan, 2006) to an $R_{\rm free}$ value of below 45%. The relationship between CC and *REFORIGIN* root-mean-squared difference (r.m.s.d.) calculated on C^{α} atoms *versus* the deposited structure is shown in Supplementary Fig. S4(*a*) and the relationship with $R_{\rm free}$ is shown in Supplementary Fig. S4(*b*). The CC negatively correlates with $R_{\rm free}$ (calculated by *REFMAC* post-*SHELXE*), but many ensembles give $R_{\rm free}$ values which in themselves would not guarantee the possibility of automatic rebuilding, confirming the superiority of the CC score for this purpose. Supplementary Fig. S4(*a*) shows that a proportion of successes derive from inaccurately placed search models. These are discussed later along with successes from inaccurate models.

It is generally considered that a *Phaser* translation-function *Z*-score (TFZ) of at least 5 is required for successful solution, and even then model bias may not always allow successful rebuilding and refinement. This rule of thumb is based on experience with conventional MR and may not hold when *ab* *initio* models are used. For our unconventional truncated ensembles, a TFZ of greater than 14 was required to guarantee successful tracing (Supplementary Fig. S5). It was also interesting that TFZ scores as low as 2.3 also gave successes; in fact, 66 successful ensembles belonging to 12 test cases had TFZ scores below 3. These results show the presence of traceable solutions with statistics that would typically lead to their being discarded. Typically, poorer density would be expected for low-TFZ solutions, suggesting that *SHELXE* is a powerful tool for bootstrapping from poorer MR results.

As expected, targets for which the *ab initio* modelling led to a large top cluster of structurally similar decoys were more likely to be successfully solved than targets yielding smaller clusters (Supplementary Fig. S1). We processed decoys from the three largest *SPICKER* clusters into search models. We then assessed the contribution of each cluster to success. Since the largest cluster often contains the best model in the population, it was likely that this cluster would give the most successes, with fewer yielded by the smaller clusters 2 and 3. Indeed, search models derived from the largest *SPICKER* cluster were able to solve 115 cases. Search models from the second largest cluster solved 100 cases, but only 11 of these were uniquely solved by this cluster, giving a total of 126 solved cases. For 114 cases the third largest cluster was also sampled. Although there were 22 successful cases, all were also solved using search models from clusters 1 or 2; thus, sampling beyond the two largest clusters is unnecessary.

Using the *SHELXE* CC score as an indicator, we achieved 126 successes from the 295 cases (43%). This number includes only the solutions that could be automatically refined and excludes a few additional marginal cases in which the search model was well positioned but automatic tracing and refinement failed. Such cases might prove to be soluble after manual rebuilding. Nevertheless, the success rate of 43% was for search models deriving from 1000 computationally inexpensive *ab initio* decoys generated in at most one CPU day per target. A comparison with the only previous study (Das & Baker, 2009) is difficult since that work used a different criterion of success that was based on *Phaser* TFZ scores. As discussed later, the use of *SHELXE* in the present work allowed the successful tracing of solutions with relatively low TFZ scores (Supplementary Fig. S5). As well as the different success criteria, the test set of the earlier paper included targets that diffracted to lower resolution than our set.

3.3. Characteristics of successful targets

An evaluation of the parameters influencing success highlighted target length as well as the type and amount of secondary structure. Smaller proteins containing fewer than 100 residues are generally more likely to succeed in MR than longer proteins (Fig. 3). This is as would be expected from the general reduction in the accuracy of *ab initio* modelling with increasing target size. This trend, however, is not necessarily followed by all-α proteins, for which a consistent success rate was observed as the length increased and did not decline to zero at the upper sequence size limit that we imposed of 120 residues. Thus, it is reasonable to expect some success with even larger all-α proteins in future studies.

Secondary-structure class was another major factor that affected success. Two methods of assessing secondary structure were used: the true secondary structure of the deposited structure as assigned by *DSSP* (Kabsch & Sander, 1983) and a *PSIPRED* prediction (Jones, 1999). Our analyses refer to the latter since obviously only this would be available to the user in practice. Here, the proteins were categorized as all-α, all-β or mixed α/β. Our test set contained 76, 43 and 175 targets, respectively, in these three categories, as well as one that lacked regular secondary structure. A high proportion of all-α cases (80%) were successful, but only one all-β protein (2%) was successfully solved. Mixed α/β proteins achieved an intermediate success rate of 37%. The case (PDB entry 2qsk) that lacked regular secondary-structure elements was not successful. Such proteins are unlikely to be correctly modelled using present *ab initio* methods. As fold class is a determinant of success, its prediction from *PSIPRED* results is reported to the user by *AMPLE* as a preliminary indicator of likelihood of success when a secondary-structure prediction has been performed locally.

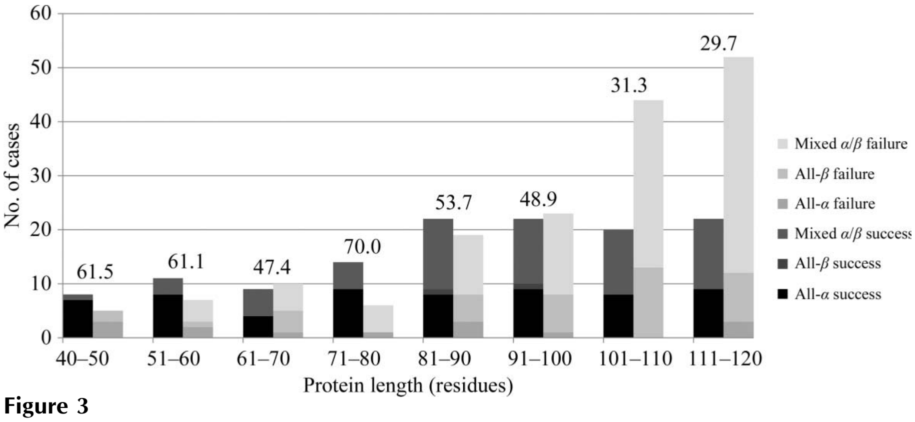

Figure 3
Characteristics of targets and their influence on MR success. In each of the size-range bins shown, the ensembles are divided into successes (left bars, dark colours) and failures (right bars, light colours). The percentage success rate in each size range is shown above the bars: larger targets are generally harder but successes are seen up to the upper size limit. Shades indicate secondary-structure classes derived from *PSIPRED* predictions and show that all-α targets are usually successful and mixed α/β targets are often successful, but that all-β targets generally fail.

The amount of regular secondary structure also influences success. No solutions were found in cases with less than 26% predicted secondary-structure content. The number of successful cases increases almost linearly with the content of helical structure, with 100% success in cases where there is 80–90% helix. This decreases at 90–100% helix content, likely owing to this group containing several particularly elongated coiled-coil-like structures which are not well modelled by *Rosetta*. Conversely, as the content of sheet increases the success rate decreases.

Since secondary-structure prediction is used to select and rank fragments during *ab initio* modelling in *Rosetta*, the accuracy of the prediction will be important for modelling accuracy. *PSIPRED* is one of the most accurate available methods although, as with most methods, helices are predicted somewhat more accurately than strands (Zhang *et al.*, 2011). Indeed, we found that the *DSSP* results and the *PSIPRED* predictions agreed very well, with few exceptions (not shown). This, and the relatively small performance deficit with strands, suggest that inaccurate secondary-structure prediction is not responsible for the lower success rate with all-β proteins. Instead, the explanation may lie in the fact that *ab initio* modelling underperforms for all-β structures (Xu *et al.*, 2011; see also Supplementary Fig. S6) and/or in specific difficulties in MR of proteins containing β-sheets. These difficulties arise from the innate diversity of β-sheets, for example in the variability of the sheet twist between related structures, which can stymie MR attempts. Specific problems may also be generated by discordant α-helices, *i.e.* observed helical regions which are strongly predicted as β-structure using secondary-structure prediction (Gendoo & Harrison, 2011). At least one protein in our test set (PDB entry 2o9u; monellin) is known to contain discordant helices (Gendoo & Harrison, 2011) and, indeed, *AMPLE* failed to solve its structure.

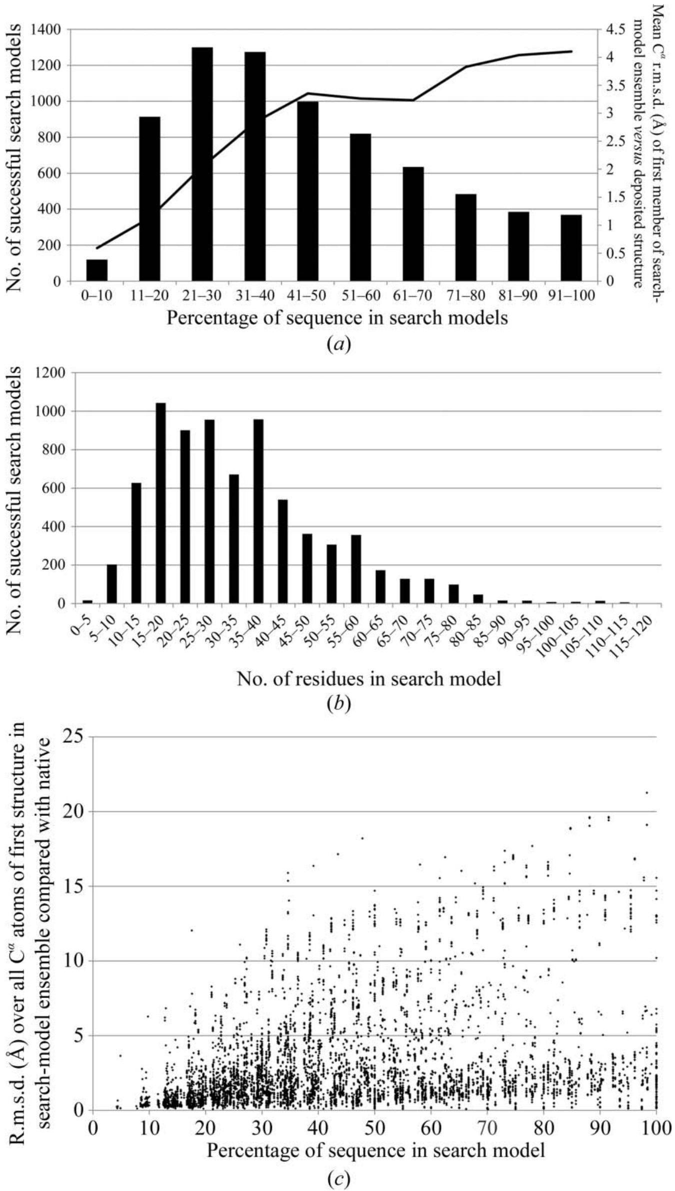

Figure 4
The influence of different modes of search-model preparation on success. (*a*) Successful search models are found covering between 4.2 and 100% of the target sequence, with the highest success rate in the range 21–40%. The trace indicates the mean coordinate error, expressed as C$^{\alpha}$ r.m.s.d. compared with the deposited structure, of ensembles in each bin. (*b*) Successful search models are found covering between three and 116 residues of the target sequence, with the highest success rate in the 15–40-residue range. (*c*) Successful search models include both longer and less accurate and short but very accurate ensembles. Accuracy is estimated from the r.m.s.d. over all C$^{\alpha}$ atoms of the first member of the search-model ensemble. This is an underestimation of the accuracy, since it does not include other models in the ensemble which may better represent certain regions of the native structure.

Finally, space group (Supplementary Fig. S7), solvent content (Supplementary Fig. S8) and the high-resolution limit of the diffraction data (at least in the range covered here; Supplementary Fig. S9) seemed to have little impact on the success rate.

3.4. Characteristics of successful search models

As expected, many successful search models were derived from parent *SPICKER* clusters in which the most accurate decoy had a GDT total score of greater than 40 (Supplementary Fig. S10), a level usually associated with a native fold (Zemla, 2003); below this level, the fold is random or non-native. However, a sizeable proportion came from clusters of less accurately modelled decoys; the best decoy in the parent cluster had a GDT total score of <40 in 43% of cases (Supplementary Fig. S10). The fact that inaccurate decoys can be transformed into successful search models is a striking illustration of the power of our cluster-and-truncate approach. Note that the best model in the parent cluster is not necessarily selected since a maximum of 200 near-centroid decoys progress to truncation (Fig. 1), but the size of the largest cluster only rarely exceeds 200 (Supplementary Fig. S1).

Successful cases were solved using at least one search model, but often several were successful. We therefore investigated which types of model preparation were the most successful. The degree of truncation of ensembles was clearly an important factor in success. While success was achieved across a full range from truncated ensembles containing as little as 4.2% of the target sequence to as much as 100% of the sequence, the highest success rates were obtained with 21–40% of the target sequence included in the ensemble, corresponding to rather severe truncation (Fig. 4*a*). Corresponding data for model size as the number of residues are shown in Fig. 4(*b*). Successes were achieved across a wide range of sizes from three to 116 residues, with the greatest number of successful models in the size range from 15 to 40 residues. There is an interplay between the completeness and accuracy of the ensembles. The truncation step improves the accuracy of the remaining structure, but simultaneously reduces the MR signal by representing a smaller portion of the contents of the asymmetric unit. Accordingly, we found that poorer models were successful if very complete, while conversely incomplete models with ~5% of the residues present can sometimes be successful if they are very accurate (Fig. 4*c*).

There is an interesting comparison to be drawn with the approach implemented in *ARCIMBOLDO* (Rodríguez *et al.*, 2009, 2012), in which the placement of a number of small ideal model fragments such as α-helices by MR is followed by density modification and rebuilding using *SHELXE*. Although the treatment of potential MR solutions is similar, our approach differs dramatically in terms of the derivation of the search models. *ARCIMBOLDO* works with short idealized elements of secondary structure, most commonly α-helices of 10–14 residues, representing as little as 5% of the target. Our successful search models, which were derived from *ab initio* models rather than idealized structures, overlap the *ARCIMBOLDO* range in both the number of residues and the percentage of target represented (Fig. 4), but extend upwards to 100% of the model and 116 residues (typical cases are shown in Supplementary Fig. S11). Most commonly, our successful search models are larger than those dealt with by *ARCIMBOLDO*, with 21–40% and 15–40 residues being the most successful categories. While a direct comparison of the two approaches is not yet available, *ARCIMBOLDO* may well succeed where the *ab initio* modelling employed here for obtaining search models fails completely, but the present method avoids the need for intensive computing that is a characteristic of *ARCIMBOLDO*.

Our approach is based on the idea that a sufficiently accurate modelled fragment ensemble can be identified in the *ab initio* model set and placed sufficiently accurately to allow successful MR. Supplementary Fig. S11 shows illustrative cases over a range of sizes where this idea works in practice. Indeed, Supplementary Fig. S4(*c*) shows that many successful search models have a low error with respect to the deposited structure and are well placed by MR. However, some successes are derived from inaccurately modelled search models (Fig. 4*b*, Supplementary Fig. S4*c*) and some result from inaccurate MR placements (Supplementary Fig. S4), even of accurate search models (Supplementary Fig. S4*c*). These more surprising successes, *e.g.* ensembles with an r.m.s.d. of >10 Å (around 5% of the total), were found to be associated preferentially with coiled-coil structures or other folds containing long helices. Supplementary Fig. S12 shows illustrative cases over a range of model errors and *REFORIGIN* placement errors. In these cases globally inaccurate models may nevertheless accurately recapitulate sufficient helical structure and broad modes of helical packing for successful structure solution. Similarly, accurate search models may be misplaced but positioned in such a way as to yield sufficient phasing power for tracing to proceed.

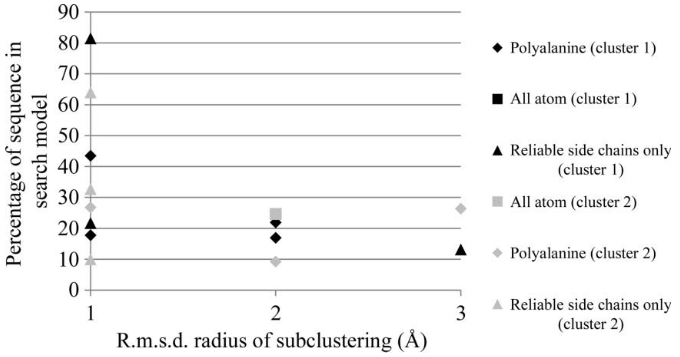

Figure 5
Processing modes of singleton solutions, *i.e.* search models uniquely capable of solving their target. The completeness of the search model is plotted against the tightness of the subclustering, with side-chain modelling and the *SPICKER* cluster indicated by the symbols shown in the legend.

Post-truncation, the decoys are subclustered by r.m.s.d. at 1, 2 and 3 Å, with each subcluster then serving as the basis for the generation of three search models differing in side-chain treatment. The success rate of ensembles derived from each class of subcluster is similar: 28% for 1 Å subcluster ensembles, 37% for 2 Å subcluster ensembles and 35% for 3 Å subcluster ensembles. The 2 Å ensembles may represent a point at which divergence within the cluster provides useful information to programs such as *Phaser* regarding reliability, but the divergence has not yet reached a degree at which additional noise outweighs this advantage.

The number of (processed) decoys in each subcluster varies according to the results of the subclustering, but we imposed a limit of 30 for reasons of speed. Since subclusters of 30 decoys produced the most successful search models, it may be that allowing greater numbers of models in the ensemble would increase the success rate at the expense of greater computational demands. Interestingly, there was a relationship between the r.m.s.d. radius of the subcluster and the number of decoys needed for success. For subclusters with a 1 Å radius that gave rise to successful search models, only 56% had the maximum of 30 models. Furthermore, successes were achieved with search models deriving from 1 Å radius subclusters containing from two to 30 decoys. In contrast, for successful subclusters with a 2 Å radius 77% have 30 decoys; for those with a radius of 3 Å this rises to 90%. Hence, it seems to be the case that as structural diversity between models in a subcluster rises, an increased number of component decoys is required for successful search models to derive from it.

The side chains of each subcluster are treated in three ways: they may be represented as polyalanine, retain all side chains previously added or keep only those side chains that are statistically more reliably predicted. There is little difference in success between the different side-chain treatments, but overall the search models with just the reliable side chains performed best. 35% of these were successful, while the figures for polyalanine and all-side-chain search models were 30 and 34%, respectively.

We observed that for some proteins most search models were successful, while for others only one was successful. At one extreme, 497 of 663 search models for target 1gvd were successful. On the other hand, 15 of the 126 successful cases were solved by only one search model. That is to say, only a certain truncation with one particular treatment of side chains could give success in these cases. Notably, such singleton successes included examples of all three modes of side-chain treatment and spanned a range of truncations containing 10–81% of the target sequence (Fig. 5) and a range of sizes from ten to 44 residues. Singleton successful search models also derived from both the top and second largest *SPICKER* clusters (not shown). Taken together, these findings justify the structure of our pipeline and demonstrate the sometimes very narrow requirements for MR success, highlighting the benefits of our automated approach that comprehensively samples ensembles and their derivatives without the requirement for user intervention.

3.5. *AMPLE* implementation

The program *AMPLE* was written in Python to control the procedure of model generation, processing, molecular replacement and rebuilding. It is available as part of the *CCP*4 package (Winn *et al.*, 2011; http://www.ccp4.ac.uk). When given a *FASTA* sequence file, *AMPLE* will call locally installed *Rosetta* to generate *ab initio* models using the sequence and the default *ab initio* modelling protocol. Fragments may be generated locally or can be accepted from the *Robetta* server (http://robetta.bakerlab.org/). *AMPLE* can also accept pre-made models in the form of a folder containing individual PDB files. This allows the input of structures from any origin, including other modelling programs and NMR ensembles. Currently, the models must each have the same sequence. *AMPLE* calls *MrBUMP* (Keegan & Winn, 2008), which acts as a wrapper for both *Phaser* (Storoni *et al.*, 2004; McCoy *et al.*, 2005, 2007) and *MOLREP* (Vagin & Teplyakov, 2010). The two programs are complementary, with some overlap, but each also has unique successes (Supplementary Fig. S13). For example, 74 cases were solved by both programs, but *Phaser* alone solved a further 50 unique cases while *MOLREP* added a further two. Currently, only the first MR placement from the results of each program is tested using *SHELXE* (Usón *et al.*, 2007; Sheldrick, 2010) tracing as a reliable indicator of a correct solution that can be rebuilt. During operation, *AMPLE* reports on two key characteristics that are somewhat predictive of success. Firstly, if a local secondary-structure prediction has been performed *AMPLE* will assess the fold class of the target (all-α, mixed α/β or all-β) and report that the chance of success is high, intermediate or low, respectively. Secondly, once the *Rosetta* modelling has finished, *AMPLE* will report on the likely success rate based on the size of the largest *SPICKER* cluster (Supplementary Fig. S1). *AMPLE* can employ multiple cores to parallelize both the modelling and MR steps on clusters and multi-core workstations. When run on a single computer (not a cluster, for technical reasons), *AMPLE* stops as soon as a successful solution is indicated by *SHELXE.*

4. Conclusion

We have presented a novel approach to processing rapidly obtained *ab initio* decoys into successful MR search models. Importantly, the computation time required is such that single-core desktop machines available to any crystallography laboratory are sufficient: no access to clusters is required. Our approach is implemented in the program *AMPLE*, which is already available as part of the *CCP*4 package. Our thorough characterization of the performance of *AMPLE* provides helpful guidelines for the crystallographer to assess the chance of their target being successfully solved. All-α proteins are particularly favourable (80% success) and mixed α/β targets were solved in 36% of cases, but all-β targets are presently unlikely to succeed. Once the user has obtained *ab initio* decoys, the size of the largest cluster is somewhat predictive of success (Supplementary Fig. S1), with larger sizes being indicative of structural convergence and more reliable predictions. Most importantly, a CC of >25% after brief rebuilding with *SHELXE* is reliably indicative of solutions which can then be automatically traced. Although the test set here was limited to a maximum of 120 residues, success with some of the largest targets suggests that some proteins longer than this may be tractable. Thus, with its simple interface to *Rosetta*, we argue that *AMPLE* brings *ab initio* modelling for MR to the crystallographer in a convenient and accessible form for the first time. Further applications of the core cluster-and-truncate methodology to cases of distant homology, to missing domains and to NMR structures are under active exploration.

We gratefully acknowledge the Biotechnology and Biological Sciences Research Council for funding (award BB/H01330X/1).

References

Bradley, P., Misura, K. M. & Baker, D. (2005). *Science*, **309**, 1868–1871.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2009). *Acta Cryst.* D**65**, 477–484.

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.* D**64**, 49–60.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Das, R. *et al.* (2007). *Proteins*, **69**, Suppl. 8, 118–128.

Das, R. & Baker, D. (2009). *Acta Cryst.* D**65**, 169–175.

Gendoo, D. M. & Harrison, P. M. (2011). *Protein Sci.* **20**, 567–579.

Jones, D. T. (1999). *J. Mol. Biol.* **292**, 195–202.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 119–124.

Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. (2009). *Proteins*, **77**, 778–795.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nature Protoc.* **3**, 1171–1179.

Lee, S. Y. & Skolnick, J. (2007). *Proteins*, **68**, 39–47.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

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.

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

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

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

Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 1288–1291.

Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nature Methods*, **6**, 651–653.

Rodríguez, D., Sammito, M., Meindl, K., de Ilarduya, I. M., Potratz, M., Sheldrick, G. M. & Usón, I. (2012). *Acta Cryst.* D**68**, 336–343.

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M. & Bourne, P. E. (2011). *Nucleic Acids Res.* **39**, D392–D401.

Roy, A., Kucukural, A. & Zhang, Y. (2010). *Nature Protoc.* **5**, 725–738.

Schwarzenbacher, R., Godzik, A. & Jaroszewski, L. (2008). *Acta Cryst.* D**64**, 133–140.

Shapovalov, M. V. & Dunbrack, R. L. (2007). *Proteins*, **66**, 279–303.

Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.

Shortle, D., Simons, K. T. & Baker, D. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.

Shrestha, R., Berenger, F. & Zhang, K. Y. J. (2011). *Acta Cryst.* D**67**, 804–812.

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). *J. Mol. Biol.* **268**, 209–225.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). *Proteins*, **34**, 82–95.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.

Theobald, D. L. & Wuttke, D. S. (2006). *Bioinformatics*, **22**, 2171–2172.

Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. (2007). *Acta Cryst.* D**63**, 1069–1074.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.

Wang, G. & Dunbrack, R. L. (2005). *Nucleic Acids Res.* **33**, W94–W98.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Wu, S, Skolnick, J. & Zhang, Y. (2007). *BMC Biol.* **5**, 17.

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

Xu, D., Zhang, J., Roy, A. & Zhang, Y. (2011). *Proteins*, **79**, Suppl. 10, 147–160.

Zemla, A. (2003). *Nucleic Acids Res.* **31**, 3370–3374.

Zhang, Y. (2008). *BMC Bioinformatics*, **9**, 40.

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

Zhang, H., Zhang, T., Chen, K., Kedarisetti, K. D., Mizianty, M. J., Bao, Q., Stach, W. & Kurgan, L. (2011). *Brief. Bioinform.* **12**, 672–688.