BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

Automated *de novo* phasing and model building of coiled-coil proteins

Sebastian Rämisch, Robert Lizatović and Ingemar André*

Department of Biochemistry and Structural Biology, Lund University, Sweden. *Correspondence e-mail: ingemar.andre@biochemistry.lu.se

Received 14 November 2014
Accepted 30 December 2014

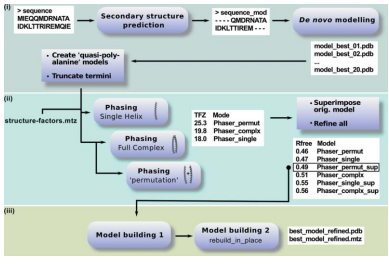
Keywords: molecular replacement; *de novo* phasing; *Fold-and-Dock*; coiled coils.

Supporting information: this article has supporting information at journals.iucr.org/d

Models generated by *de novo* structure prediction can be very useful starting points for molecular replacement for systems where suitable structural homologues cannot be readily identified. Protein–protein complexes and *de novo*-designed proteins are examples of systems that can be challenging to phase. In this study, the potential of *de novo* models of protein complexes for use as starting points for molecular replacement is investigated. The approach is demonstrated using homomeric coiled-coil proteins, which are excellent model systems for oligomeric systems. Despite the stereotypical fold of coiled coils, initial phase estimation can be difficult and many structures have to be solved with experimental phasing. A method was developed for automatic structure determination of homomeric coiled coils from X-ray diffraction data. In a benchmark set of 24 coiled coils, ranging from dimers to pentamers with resolutions down to 2.5 Å, 22 systems were automatically solved, 11 of which had previously been solved by experimental phasing. The generated models contained 71–103% of the residues present in the deposited structures, had the correct sequence and had free *R* values that deviated on average by 0.01 from those of the respective reference structures. The electron-density maps were of sufficient quality that only minor manual editing was necessary to produce final structures. The method, named *CCsolve*, combines methods for *de novo* structure prediction, initial phase estimation and automated model building into one pipeline. *CCsolve* is robust against errors in the initial models and can readily be modified to make use of alternative crystallographic software. The results demonstrate the feasibility of *de novo* phasing of protein–protein complexes, an approach that could also be employed for other small systems beyond coiled coils.

1. Introduction

Coiled coils are protein domains that are found in a wide range of proteins involved in a diverse set of biological functions. Structurally, coiled coils are α-helical bundles in which individual helices are wound around a common superhelical axis. In most cases, coiled coils act as protein-oligomerization domains, but they can also be involved in filament formation in both the extracellular matrix and cytoskeletal networks, as well as in membrane-fusion processes. In accordance with their primary role as oligomerization domains, coiled coils can exist in a variety of oligomeric states. Besides their functional importance, coiled coils also serve as important model systems to understand protein assembly and interface formation. A large body of work in the protein-engineering field has been devoted to uncovering the relationship between sequence, structure and stability in coiled-coil proteins (Harbury *et al.*, 1993; Lumb & Kim, 1995; Kammerer *et al.*, 2005; Grigoryan & Keating, 2008; Lee *et al.*, 2003; McClain *et al.*, 2001; Monera *et al.*, 1994; Ramos & Lazaridis, 2006; Tsatskis *et al.*, 2008; Yoon *et al.*, 2007).



High-resolution molecular structures obtained using X-ray crystallography or NMR are central for our understanding of the biological functions and biophysical properties of proteins. In order to extract structural information from X-ray diffraction data, the recorded intensities must be supplemented by additional phase information. This leads to the crystallographic phase problem, which can be tackled by a number of approaches. One such method is molecular replacement (MR), which is convenient from an experimental point of view because it does not require additional data collection from heavy-atom derivatives. However, MR is dependent on the availability of an accurate model of the crystallized system that can be correctly placed in the asymmetric unit to obtain initial phase estimates. Owing to the large number of solved structures, homologous proteins with highly similar backbone structures can often be identified in structural databases and used as effective search models. However, for proteins with limited sequence identity (below 30%) to proteins with known structures, MR is still very difficult.

Whereas overall the percentage of protein structures in the Protein Data Bank (PDB; Berman *et al.*, 2000) solved by MR is about 70%, for coiled-coil proteins this value is 50%. One should further note that the vast majority of these are point mutants and fusion constructs of coiled coils whose structures have previously been solved using experimental phasing methods. In a couple of cases the phase problem has been solved by MR despite the lack of structural homologues using nonstandard approaches. In one such example, initial phases were obtained by using parameterized coiled-coil models as search models for MR (Harbury *et al.*, 1993). In another, various coiled-coil structures with truncated side chains were used (Thépaut *et al.*, 2004). However, for systems without close homologues, in particular for *de novo*-designed sequences, which may deviate considerably from known structures, traditional MR approaches may become intractable. It is beneficial therefore to develop methods that do not rely on sequence homology to produce initial phase estimates.

Combinations of advanced structure modelling and MR are becoming increasingly useful for solving crystal structures where traditional methods fail. DiMaio, Terwilliger *et al.* (2011) developed an iterative procedure that combines energy- and density-guided structure rebuilding of homology models in order to lower the sequence-identity threshold where MR can be effectively applied. Further improvements of the method have resulted in successful structure solution starting from templates with down to 15% sequence identity (DiMaio, 2013). However, even with those improvements this method is still limited by the availability and the identification of suitable homologues. A method capable of surpassing this limitation, especially to find initial phases for α-rich proteins, is implemented in the program *ARCIMBOLDO* (Rodríguez *et al.*, 2009). *ARCIMBOLDO* uses fragments from ideal α-helices to obtain initial phases and uses *SHELXE* to generate an initial backbone trace. However, the data resolution needs to be better than 2 Å. Sammito *et al.* (2013) extended this method by introducing a fragment-based approach that allows *de novo* phasing of proteins with folds beyond purely α-helical.

However, for proteins without close homologues, and for cases with data of lower resolution, these improved methods may still fail in producing an interpretable map. This is where *de novo* structure modelling may become very useful. The general applicability of phasing using *de novo* models has recently been assessed in several studies (Bibby *et al.*, 2012; Das & Baker, 2009; Qian *et al.*, 2007; Rigden *et al.*, 2008; Shrestha *et al.*, 2011). Das and Baker estimated that about one sixth of small proteins could successfully be phased with *de novo* models generated using the *Rosetta* macromolecular modelling suite (Das & Baker, 2008; Leaver-Fay *et al.*, 2011; Rohl *et al.*, 2004). More recently, Bibby and coworkers developed the program *AMPLE*; using a larger benchmark set, they obtained MR solutions for about 43% of the included proteins by phasing large numbers of low-resolution *de novo* models (Bibby *et al.*, 2012). Das and Baker also found a correlation between the ease of phasing and the molecular weight of the protein: larger structures required less accurate search models than smaller ones, which was in accordance with previous observations. This implies that coiled-coil structures, given their relatively small size, will require more accurate search models for successful phasing using MR.

Despite the rapid development of algorithms for MR in the absence of close homologues, there are still areas of application of these methods where the task can be very challenging. For example, *de novo* phasing has mostly been tested on small monomeric, globular proteins. Extending these approaches to assemblies consisting of several subunits is a challenging but an important problem as crystallization of protein complexes becomes increasingly successful. Such complexes, especially when subunits assemble in an intertwined manner, are much more difficult to model, and MR can be challenging.

Currently, there is no systematic study that has demonstrated the feasibility of using models from *de novo* structure prediction to solve structures of protein oligomers using MR. The goal of this study was to develop an automated procedure tailored to *de novo* phasing of small homo-oligomeric proteins using *de novo* models. By using accurate models of complete complexes, we hoped to also enable MR for lower resolution diffraction data. A further goal was to produce fully built models in an automated manner, with fully connected residues and side chains, which would only require minor manual editing to produce finalized structures.

Coiled coils can serve as a model system for small protein oligomers. *Rosetta*'s *Fold-and-Dock* protocol was developed to predict the structures of small symmetric oligomers, and we have recently shown that the method can be used to generate accurate models of homomeric coiled coils of various oligomeric states (Rämisch *et al.*, 2015). Here, we used *de novo* models of homo-oligomeric coiled-coil structures generated by *Rosetta*'s *Fold-and-Dock* protocol and tested whether these could be used to solve structures using MR. We developed a fully automated procedure, named *CCsolve*, that produces fully refined, relatively complete models with all side chains present; it requires no prior information other than the

sequence and the oligomeric state. We benchmarked phasing and model building on 24 coiled-coil proteins with different oligomerization states. In 22 cases, *CCsolve* generated structures close to the reported crystal structures, with some of the most accurate models having relatively low data resolutions between 2.0 and 2.5 Å. The models were sufficiently accurate so that limited manual rebuilding was sufficient to produce models with similar quality to that of the corresponding published structures. *CCsolve* can be run on a single multicore workstation and uses freely available software components.

2. Results

2.1. The method

The automated pipeline starts with *de novo* structure modelling, proceeds with phasing the top-ranked models using MR, selects the most promising candidate model and finishes with iterative model building and refinement. The details of each step are described below. The method requires standard processing of crystallographic data as well as the following specific information as input.

(i) The oligomeric state of the crystallized coiled coil.

(ii) The complete sequence as it was used in crystallization trials.

(iii) A sequence file modified according to secondary-structure prediction (described below).

(iv) A file in MTZ format containing the experimental structure-factor amplitudes.

(v) The number of helices in the asymmetric unit, derived from Matthews analysis (Matthews, 1968).

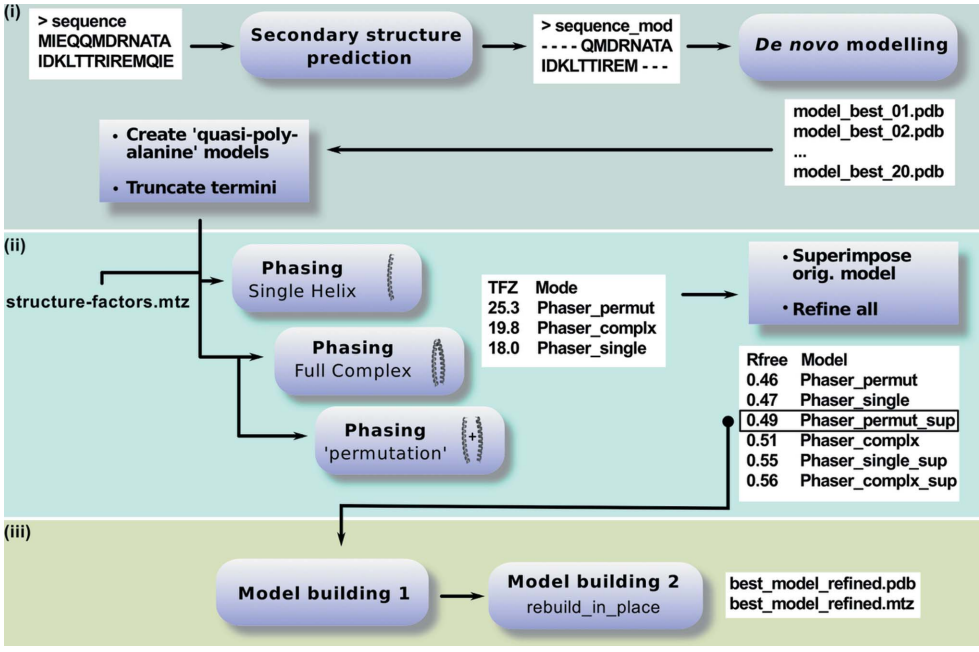


Figure 1
Schematic representation of the automated *de novo* phasing and model-building procedure. (i), (ii) and (iii) indicate the three main phases. Round boxes represent the main steps, all of which involve commonly available software, and square boxes show intermediate steps, which are partially performed by customized scripts. White boxes represent output, *i.e.* the result of a certain step that is used to proceed. In phase (ii), the *Phaser* solutions can be ranked by LLG, alternatively to R_{free} after refinement.

Once this information is available, all subsequent steps are performed in a fully automated manner. A full flowchart describing the modelling approach is shown in Fig. 1.

2.1.1. *De novo* structure modelling. Owing to its success in *de novo* structure prediction of homomeric coiled coils, we chose to use *Rosetta*'s *Fold-and-Dock* application (Das *et al.*, 2009) to generate starting models for MR. The method is particularly useful in predicting the structures of intertwined complexes or assemblies consisting of subunits that are not stable on their own. We have recently shown that the structure of homomeric coiled coils could be accurately predicted with backbone r.m.s.d. values ranging from 0.8 to 2.5 Å in a benchmark of 33 proteins (Rämisch *et al.*, 2015).

Fold-and-Dock requires only sequence information and does not rely on the availability of homologous structures. Nevertheless, the presence of long flexible stretches at the termini, which is commonly observed in structures of coiled coils, may impair *de novo* structure modelling. Accordingly, instead of using the full sequence of the peptides as used in crystallization trials, our procedure excludes regions that are predicted to form terminal loops. Secondary-structure prediction using *PSIPRED* (Jones, 1999) is carried out to identify potentially flexible stretches at both termini. Terminal segments with a low probability of helix formation were removed from the modelled sequence. This modified sequence serves as input for *Fold-and-Dock* simulations, which are carried out as described elsewhere (Das *et al.*, 2009). In brief, for trimers to pentamers, *Fold-and-Dock* was run in symmetric mode using standard fragment files. For benchmarking, we excluded backbone fragments from the structure in question as well as from all homologues. The generation of 5000 models is usually sufficient; however, anti-parallel tetramers require 20 000 prediction runs owing to the higher number of degrees of freedom inherent in the *D*2 symmetry. Dimers were predicted with an asymmetric version of *Fold-and-Dock* (Rämisch *et al.*, 2015). The introduction of structural asymmetry has been demonstrated to be important for the accurate prediction of dimers. Owing to the higher number of degrees of freedom in this simulation, a larger number of models (50 000) needs to be generated. If the Matthews analysis is ambiguous and the oligomeric state in solution is not known, *Fold-and-Dock* can be used to generate models assuming different oligomeric states to test which one produces accurate MR solutions.

2.1.2. Initial placement and ranking. The 20 lowest-energy

Table 1
Summary of results.

i.m. indicates that the identical model was selected as when using R_{free}. n.k., not known; n.d., not determined; ASU, asymmetric unit.

															Result		
									No. of residues			Deposited			TFZ=		LLG
Oligomeric state	Target PDB code	Original method†	Resolution (Å)	*Fold-and-Dock* r.m.s.d.‡ (Å)	Helices per ASU	Solvent content (%)	Space group	*Phaser* solution§	TFZ=	Crystal	a.b.¶ (%)	Final (%)	R	R_{free}	R_{free} (re-ref.)††	R	R_{free}	R_{free}
2-mer	1pl5	SIR	2.5	3.3	2	69.6	$P6_5$	S	21.5	151	79	81	0.26	0.29	0.27	0.27	0.28	0.34
2-mer	1t3j	MAD	2.5	1.5	1	68.0	$F4_132$	S	10.7	122	98	98	0.25	0.29	0.29	0.24	0.27	0.28
2-mer	1uii	MAD	2.0	2.0	2	65.3	$P2_12_12_1$	S	17.0	135	95	98	0.22	0.24	0.23	0.25	0.27	n.d.
2-mer	1uix	MAD	1.8	32 (1.7)‡‡	2	46.7	$C2$	P	11.0	74	102	103	0.22	0.24	0.25	0.22	0.25	i.m.
2-mer	1x8y	MR	2.2	1.1	1	71.2	$P6_522$	S	19.4	84	85	86	0.28	0.30	0.31	0.26	0.30	i.m.
2-mer	2oqq	SIRAS	2.0	1.0	2	47.7	$C2$	P	14.4	128	76	83	0.24	0.30	0.27	0.22	0.27§§	i.m.
2-mer	2q6q	MR	2.0	1.6	2	38.4	$P2_12_12_1$	P	15.1	139	73	79	0.22	0.30	0.28	0.28	0.32	i.m.
2-mer	2w6a	SAD	1.4	1.1	2	36.9	$P2_1$	S	14.3	167	86	95	0.16	0.20	0.20	0.21	0.23	i.m.
2-mer	3bas	MR	2.3	3.0	2	52.8	$C2$	C	12.9	62	98	95	0.25	0.29	0.32	0.29	0.38	i.m.
3-mer	1wt6	MAD	1.6	0.8	3	48.0	$P2_12_12_1$	C	14.5	195	120¶¶	120¶¶	0.23	0.27	0.27	0.27	0.32	0.25
3-mer	1zvb	MR	1.7	0.9	3	45.9	$C2$	C	18.7	101	100	101	0.20	0.24	0.23	0.22	0.24	i.m.
3-mer	2akf	MR	1.2	1.5	3	36.2	$P1$	C	13.0†††	96	100	100	0.16	0.20	0.27	0.29	0.31‡‡‡	0.29
3-mer	2pnv	MR	2.1	0.9	2	43.4	$P6_3$	S	9.8	78	71	74	0.22	0.28	0.33	0.30	0.35	0.28
3-mer	2wz7	SAD	2.5	1.9	6	40.0	$P2_12_12_1$	n.d.	n.d.	403	n.d.	n.d.	0.22	0.30	n.d	n.d.	n.d.	n.d.
3-mer	3efg	MAD	2.0	1.4	1	n.k.	$P6_3$	S	9.1	53	77	74	0.20	0.21	0.23	0.28	0.27	0.30
3-mer	3g9r	MR	2.0	1.1	6	52.1	$P2_1$	S	11.9	244	88	101	0.22	0.27	0.30	0.27	0.30	0.29
4-mer	2bni	SIR/MR	1.5	1.0	4	30.6	$P3_1$	S	20.1	128	77	84	0.24	0.28	0.27	0.29	0.32	n.d.
4-mer	2gus	MAD	1.8	1.0	1	39.9	$P42_12$	S	13.0	42	83	83	0.24	0.29	0.28	0.26	0.30	i.m.
4-mer	2ovc	MR	2.1	1.2	1	43.3	$I4$	S	13.4	30	90	90	0.20	0.22	0.25	0.22	0.27	0.25
4-mer	2r2v	MR	1.9	0.8	8	42.8	$P4_3$	C	18.5	259	71	83	0.19	0.24	0.31	0.29	0.34	0.28
4-mer	3e7k	MR	2.0	2.0	8	43.0	$C2$	C	20.6	432	n.d.	n.d.	0.19	0.25	0.24	n.d	n.d	n.d.
4-mer	3m9h	MR	2.0	1.2	6	30.1	$F222$	S	22.4	258	86	86	0.21	0.27	0.29	0.25	0.29	i.m.
5-mer	2guv	MAD/MR	1.4	1.3	5	n.k.	$P2_1$	S	14.1	280	91	95	0.20	0.25	0.26	0.25	0.28	n.d.
5-mer	3miw	MR	2.5	1.4	10	35.5	$P4_2$	C	25.2	432	92	93	0.26	0.30	0.37§§§	0.29	0.39	i.m.

† Methods used for solving the deposited structures: MAD, multiple-wavelength anomalous diffraction; SIRAS, single isomorphous replacement with anomalous signal; MR, molecular replacement; SIR, single isomorphous replacement. ‡ C^{α} r.m.s.d. of the *Fold-and-Dock* model that resulted in the lowest R_{free} after initial phasing; it includes only those chains that are present in the asymmetric unit. § Best *Phaser* solution; P, permutation; C, complete complex; S, single helices. ¶ Residues built after the first model-building step. †† Crystal structure refined after removing all nonprotein atoms using the same refinement settings as used for automated model building. ‡‡ The model had a wrong helix orientation (antiparallel); the r.m.s.d. of the extracted helix used for phasing is given in parentheses. §§ Refinement performed using *REFMAC*5. ¶¶ *AutoBuild* added several terminal residues although no density was visible; 100% of the originally resolved residues were present. ††† RFZ score. ‡‡‡ The deposited R_{free} was obtained after anisotropic refinement (not performed here). §§§ The deposited structure contains a large number of buffer molecules and the X-ray data are twinned.

models from *Fold-and-Dock* are used as starting points for MR. Before initial phasing, these models are modified in two ways. Firstly, three residues are removed from both termini of each helix because termini tend to be less accurately modelled than the core portion of a coiled coil. Secondly, most amino acids are substituted by alanine. Only Phe, Gly, His, Ile, Ser, Thr, Val, Tyr and Trp residues remain unchanged. This essentially yields polyalanine helices containing a few 'anchor residues' (subsequently referred to as quasi-polyalanine models), which are intended to prevent register shifts. After helix shortening and side-chain truncation, the structures are handed over to *Phaser* to find initial phase estimates.

If structure prediction yields a model that is very close to the crystallized structure, the full complex may be placed accurately; alternatively, single helices may be employed. Our procedure tries both MR strategies in parallel to increase the chance of optimal placement. For the latter approach, only one helix per oligomer needs to be considered, because the *Fold-and-Dock* models are perfectly symmetric. In summary, 40 different models (20 complete complexes and 20 single helices) are used as input for *Phaser* (McCoy *et al.*, 2007). If the expected number of helices within the asymmetric unit is one, only single-chain MR needs to be performed.

Asymmetric dimers can be viewed as complexes of two different molecules. For such cases, *Phaser* offers an alternative mode in which the molecules are placed consecutively. Thus, in addition to the single-chain and full-complex modes, we also run *Phaser* with dimers in 'permutation' mode, which tests both possible search orders. Because the individual chains in asymmetric dimers usually do not display large deviations in backbone geometry, we perform the single-chain-based placement using only one chain. If no satisfactory result were obtained the second chain could be tried, although this was not necessary for our test cases.

The *Phaser* results are ranked according to their translation-function Z-score equivalent (TFZ=; the TFZ score is the number of standard deviations above the mean value), except for $P1$ crystals with only one monomer in the asymmetric unit, where the rotation-function Z-score equivalent (RFZ=) is used instead. The TFZ score equivalents of runs using

complete complexes and single helices cannot be directly compared. To select which *Phaser* solution to continue with, the best solution at this point is identified by refinement of the highest-scored models from each phasing mode and comparison of their R_{free} values. Refinement of the two (or three) models from the different MR modes is performed using *phenix.refine* (Afonine *et al.*, 2005). Additionally, a non-truncated model with the native sequence is obtained by superimposing individual helices from the corresponding *Fold-and-Dock* model onto the placed 'quasi-polyalanine' complex. These models are also subjected to refinement. The resulting four (or six) refined models are then ranked by their R_{free} values and only the best one is selected for model building.

Alternatively, the phased models can be ranked according to the log-likelihood gain (LLG). Because the LLG can be used to compare models in a straightforward manner, no refinement is needed. Hence, when ranking by LLG instead of TFZ=, the refinement step is omitted. Results using this alternative selection strategy are described at the end of §3.

2.1.3. Model rebuilding. The *AutoBuild* program in the *PHENIX* software suite (Terwilliger *et al.*, 2008) is used for model rebuilding. In order to enable the building of residues that were excluded from modelling in *Fold-and-Dock*, the complete sequence as used for crystallization trials is provided. The starting point for model building is the best *Phaser* solution with all side chains and previously removed terminal amino acids added back to give the correct sequence. Firstly, a new model is built from scratch into the electron-density map, including information from the provided model. *AutoBuild*'s much more conservative `rebuild_in_place` procedure is not sufficient at this point because it can neither correct register shifts nor add or delete residues.

The rebuilt and refined model from *AutoBuild* typically has a highly modified sequence and contains various chain breaks and partially incomplete helices. We implemented an algorithm to reconnect split helices and identify the most complete helix within the complex. This selected helix is then superimposed onto all other helices of the incomplete model. If several helices have the same length, a sequence alignment is carried out using *ClustalW* (Thompson *et al.*, 1994) in order to identify the helix with highest sequence recovery. The rebuilt model is then subjected to a second rebuilding step, but this time with the `rebuild_in_place` option enabled. This second step enables subtle refinement of the structure using the correct sequence. Because our aim was to study the capability of fully automated coiled-coil *de novo* phasing and building, we did not carry out manual rebuilding subsequent to this stage, except for three examples.

The described procedure is the default protocol, but it is flexible enough that a number of modifications can easily be introduced. This is important because solving problematic crystal structures often requires some amount of trial and error. Modifications may include the use of different input structures at various stages of the protocol or the use of different crystallographic model-building software. A few examples are described below.

2.2. Benchmark results

We tested the performance of the procedure described above using 24 crystal structures of homomeric coiled coils, as previously used in our modelling benchmark, with structure-factor amplitudes deposited in the PDB. The benchmark set contains nine dimers, seven trimers, six tetramers and two pentamers. The data extend to resolutions ranging from 1.2 to 2.5 Å and the content of the asymmetric unit ranges from single chains to two copies of a complete pentamer (crystallographic data for the set of structures are summarized in Table 1). The starting point for structure determination was the complete sequence of a peptide as used in crystallization trials. The sequence served as input for secondary-structure prediction and subsequent *de novo* modelling, and the deposited structure-factor amplitudes were used for phasing and model building.

In order to avoid the modelling of potentially unstructured regions, we performed a secondary-structure prediction and removed loop regions that were predicted with high confidence. According to the *PSIPRED* results, all but four sequences were truncated at the termini. 12 of the resulting sequences for *de novo* modelling were longer and five of the sequences were shorter than those resolved in the deposited structure. Most *Fold-and-Dock* simulations yielded models with low r.m.s.d. values relative to the deposited structures (Table 1). Following the *PSIPRED* result, the pentameric 3miw was modelled with a sequence that contained more amino acids than were resolved in the crystal structure. These extra amino acids were unstructured in the *Fold-and-Dock* model, demonstrating that the structure-prediction protocol was able to recapitulate the lack of stable secondary structure.

The first critical step in structure solution is the determination of initial phases *via* MR. For all but two cases in the benchmark, initial phasing succeeded and yielded TFZ equivalents between 9.1 and 25.2 (Table 1). Placement of 2wz7 failed, even when the deposited crystal structure was used as input for *Phaser* using our chosen input parameters, whereas initial phases were found for 3e7k but subsequent model-building and refinement steps did not improve the model beyond a final R_{free} of 0.56. We considered these two cases as failures and excluded them from the remaining analysis. Two dimers have only one helix per asymmetric unit. Of the remaining seven dimers, three showed the best results when running *Phaser* in permutation mode, with three placing single chains and one placing the complete dimer. For the other coiled coils, single-helix phasing gave the best result in nine cases and full-complex phasing was more successful in five cases. The R_{free} values of the models before remodelling were between 0.29 and 0.57 (Fig. 2*a*).

The first remodelling step in *AutoBuild* reduced R and R_{free} significantly in all cases (Fig. 2*a*). In several cases we observed that initial register shifts are corrected in this step, as for 2guv, which crystallized head-to-tail; helices from the neighbouring asymmetric unit appear as exact continuations of the previous helix. Despite the almost random distribution of the placed single helices along the quasi-infinite helix density by *Phaser*,

the register shift was corrected in *AutoBuild* without any human intervention.

Although the model accuracy is improved, the resulting structures are usually fragmented and a large proportion of amino acids have non-native side chains. Furthermore, the backbones of different helices are traced to different degrees, resulting in helices of differing lengths. Additionally, the final refinement step in *AutoBuild* failed in one case (see §2.3.2). The average percentage of built residues compared with the deposited structures after this step is 88%. Substantial manual model building and refinement are required to obtain more complete helices with the correct connectivity. In order to obtain correctly connected helices with high residue completeness, we implemented a method that identifies the longest built helix, connects its fragments if necessary and superimposes it onto all other helices in the structure. This step determines the completeness of the final obtained model from the automated procedure. On average, 91% of the residues present in the deposited models are present at this stage and all residues within a helix are connected. The resulting model serves as input for final model building: a second *AutoBuild* `rebuild_in_place` run corrects the sequence and gives a final $R_{\rm free}$ for the fully automatically generated model. These values were between 0.23 and 0.39 (Fig. 2*a* and Table 1); however, they are not directly comparable to the reported $R_{\rm free}$ values in the PDB. Different software, or at least different software versions, had been used; re-refining the unmodified structures as downloaded from the PDB yields an average difference in $R_{\rm free}$ of 0.03 (s.d. = 0.04). A more informative estimation of the performance of the method can be made when comparing with the final $R_{\rm free}$ values after re-refining the deposited crystal structure using *phenix.refine* without ligands using the software employed here. The comparison reveals that for most structures the model accuracy that the described method achieves is close to that of the respective crystal structure when stripped of ligands and before more specialized refinement methods are applied. The average difference in $R_{\rm free}$ is 0.01 (s.d. = 0.03; Table 1). Only the final models of 3bas and 3miw were of very low quality, with $R_{\rm free}$ values of 0.38 and 0.39, respectively. In the case of 3bas, an N-terminal 11-residue stretch was built for only one of the helices by *AutoBuild* and the missing amino acids were added to the model by duplicating and superimposing the longer helix, resulting in this part being quite different from the real structure. For this case, omitting the second `rebuild_in_place` step and instead manually building the structure after the first *AutoBuild* run would probably prove to be more beneficial because the built helices in this model are virtually identical to the asymmetric deposited structure.

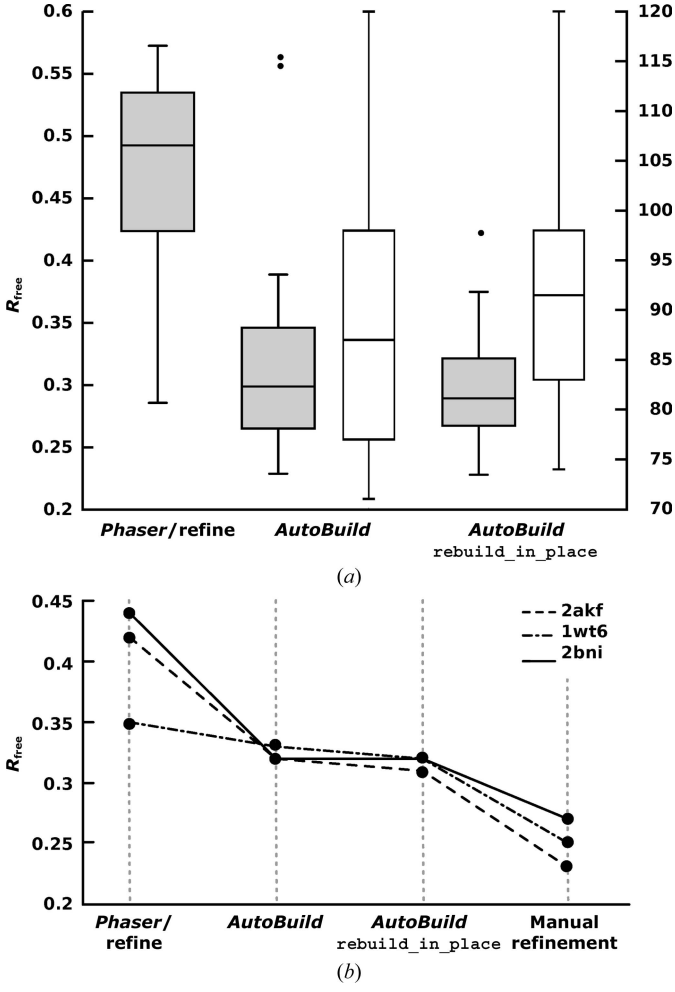


Figure 2
(*a*) Grey boxes: distributions of the 22 $R_{\rm free}$ values obtained after phasing (and refinement) as well as after the first and second rebuilding steps; white boxes, percentage of amino acids built compared with the deposited crystal structure. (*b*) $R_{\rm free}$ values for three proteins that were manually refined subsequent to the automated procedure.

In summary, all 22 systems could be placed, although some were initially out of register, and all of them were significantly improved by automatic rebuilding. Remarkably, the r.m.s.d. between the *Fold-and-Dock* model and the crystal structures of 1pl5 (dimer), 1uii (dimer), 1uix (dimer), 2q6q (dimer) and 3bas (dimer) were higher than what is commonly accepted as the upper limit for MR and yet MR solutions could be obtained with our procedure. This approach is also able to correct significant errors in side-chain conformations of the starting models. 2guv, a designed pentamer with mostly phenylalanines in the core, had all of its rotamers initially oriented wrongly in the *Rosetta* model; all of these were correct in the final model.

2.3. Examples

A description of those cases where a deviation from the default procedure was necessary follows below. Here, we report the types of problems that we observed and how the procedure may be modified if necessary.

2.3.1. 2q6q. When rebuilding this model, the refinement step in *AutoBuild* did not converge towards a low $R_{\rm free}$ value. As an alternative to refinement in *PHENIX*, we tried using *REFMAC*5 (Murshudov *et al.*, 2011) instead. For unknown reasons, this strategy resulted in a low $R_{\rm free}$ value, whereas the previous strategy did not. Therefore, we omitted the *PHENIX* refinement step in the *AutoBuild* protocol and replaced it with

refinement using *REFMAC*5 in both model-building steps for this particular case.

2.3.2. 2bni. Upon model building, the R_{free} did not decrease. We noticed that the initial refinement of the complete sequence model was significantly worse than that of the quasi-polyalanine model (R_{free} of 0.59 and 0.44, respectively). Accordingly, we tested model building using the latter (after refinement) as input. This decreased the final R_{free} from 0.50 to 0.32. The quality of the obtained electron-density maps was good enough to largely recover the correct sequence in the first round of model building. Using a subset of the benchmark proteins, we then tested whether directly using the truncated models instead of superimposing the helices from the *Fold-and-Dock* models, as performed in this example, would be a good general strategy. The resulting models had R_{free} values comparable those obtained before. However, the sequence completeness was significantly reduced. We concluded that using these models might be beneficial in individual cases but that it should not be part of the default pipeline.

2.3.3. 3miw. Here, the asymmetric unit contains two pentameric coiled coils, which makes it the largest structure in our benchmark set. It is also one of the structures with the lowest resolution of the data, which overall makes it into a difficult MR problem. The default procedure yielded an R_{free} of 0.51. Visual inspection of the *Fold-and-Dock* model revealed several nonhelical residues at the C-terminus. In *de novo* structure prediction such regions are often wrongly modelled. Thus, we deleted these positions from the starting model. Although the TFZ score decreased from 25.2 to 15.9, the R_{free} of the best superimposed and refined model decreased from 0.55 to 0.44. After the first round of rebuilding, we obtained a model without any fragmentation, with the correct sequence and with an R_{free} of 0.39. The second model-building step increased the R_{free} by 0.05; thus, we considered the result of the first building round as the final solution.

2.4. Manual model optimization: three examples

The automated model-building procedure is not perfect. Many of the final models still contained wrong rotamers. Furthermore, residues that were excluded from *Fold-and-Dock* modelling were not necessarily added by *AutoBuild*. However, the resulting electron-density maps are of sufficient quality to allow the manual building of missing residues and the correction of many wrong rotamers (Fig. 3) as well as correcting backbone traces near the termini. To test whether the automatically generated models could be improved with relatively little effort, we performed further manual rebuilding of three models that had large differences between the R_{free} of our model and the deposited model: 2akf, 1wt6 and 2bni. We were able to improve all three models substantially in a few cycles (3–10) of building and refinement using *Coot* (Emsley *et al.*, 2010) and *phenix.refine* (Fig. 2*b*). For 2akf, the high resolution allowed us to perform individual anisotropic refinement, which led to a decrease in R_{free} of 0.06 (final R_{free} of 0.23). The obtained R_{free} values of 1wt6 (0.25) and 2bni (0.27) indicated that we reached model accuracies comparable with those of the deposited structures. Except for 1wt6, where too many residues were built by *AutoBuild* and were removed during manual refinement, several missing terminal residues were visible in the electron-density maps and could be built successfully.

2.5. Model selection by LLG

As noted in the description of the method, the LLG can be used to identify successful MR solutions. We tested the performance of the pipeline when taking the highest LLG *Phaser* solution for model building and refinement; the results are summarized in Table 1. In ten cases the highest LLG solution was identical to the previously selected solution based on the R_{free} value. In six cases the final R_{free} was improved and in three of these the values were significantly better. In two cases the resulting R_{free} was worse than using the initial selection method, and for three structures that could previously be successfully built *AutoBuild* failed to produce models that could be processed automatically. These models contained highly scattered helical fragments which prevented the identification of the longest helix for subsequent superposition and model refinement. The structures of 2wz7 and 3e7k, which were not solved initially, could not be solved using the top LLG solutions either. Because significantly improved models were obtained in a few cases, we provide selection by

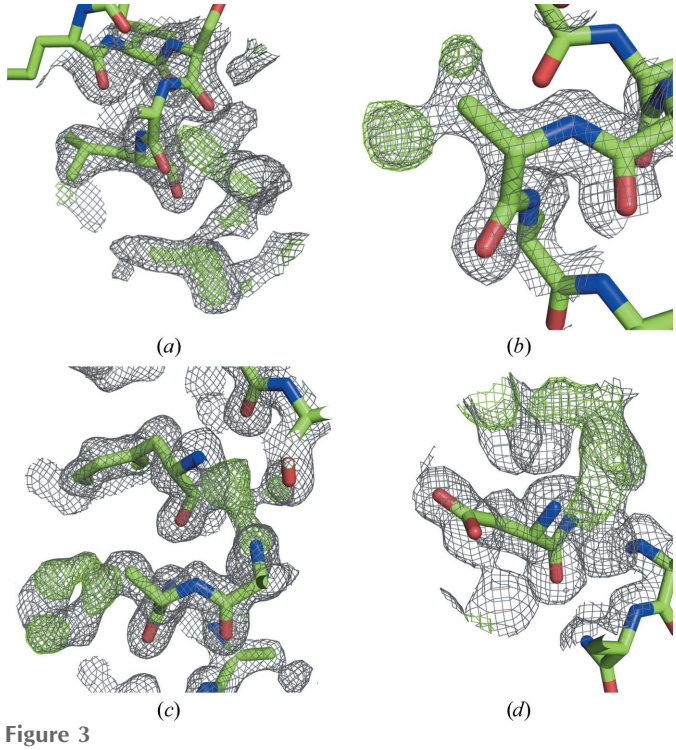


Figure 3
Examples of map quality at the end of the automatic procedure. Final electron-density maps are contoured at 1σ (grey) and positive difference density is contoured at 3σ (green). (*a*) C-terminus of the dimer 2q6q; the density shows that the chain can be extended further. (*b*) Missing threonine side chain of the pentamer 2guv. (*c*) Missing phenylalanine side chain and a missing amino acid in the pentamer 2guv. (*d*) N-terminus of the dimer 1uix; the density shows that more residues can be added.

LLG as an optional parameter to change the default behaviour of *CCsolve*.

3. Discussion

All but three models in our benchmark set yielded solutions with an accuracy close to the published crystal structures. In agreement with previous studies, our results demonstrate the feasibility of using *de novo* models for *de novo* phasing. The comparison with $R_{\rm free}$ values after re-refinement of the ligand-free crystal structures shows that the accuracy of our automatically built models is close to what can be theoretically achieved before manual introduction of ligands and special treatments such as anisotropic refinement or the application of twin laws. Notably, several structures could be solved with $R_{\rm free}$ values comparable with the deposited structure, although the data resolution is as low as 2.5 Å. However, it is not clear whether this can be extrapolated to proteins with different folds.

Some of the *Fold-and-Dock* models produced were rather inaccurate (Table 1), yet for some of these cases a reasonable solution was obtained. This surprising result demonstrates that the procedure can be quite robust against inaccuracies in the *de novo* models, even for small systems such as short dimeric coiled coils.

A potential problem in MR of coiled coils is finding the correct α-helix register. Our results show that *de novo* models can be accurate enough to be placed unambiguously with the correct register. Subsequent refinement produces maps that show clear side-chain density. Even when the molecules were placed out of register by up to four helical turns, the resulting maps contained sufficient detail to guide backbone tracing in *AutoBuild* such that the correct register was found. Hence, finding the correct register during initial placement is not a strict requirement for successful model building.

The electron-density maps obtained at the end of our procedure are of high quality and hence serve as good starting points for finalizing a structure. There is a possibility that owing to inaccurate secondary-structure prediction, too few or too many residues are considered in *Fold-and-Dock*. As we could show for three examples in which automated structure solution was complemented with manual modelling, adding and deleting residues as well as correcting wrong rotamers is relatively straightforward because the corresponding densities are clearly visible. Hence, *de novo* modelling of too large or too small a portion of a protein has little effect on the final outcome.

Because coiled coils are often relatively symmetric structures, the procedure includes phasing using single chains. We expect that in some circumstances phasing using a single chain might succeed even if the overall complex was modelled rather inaccurately, for example if the relative orientation of helices was wrong. Our results show that in some instances this method does result in better phasing than when a complete complex is used. For dimers there seems to be a higher chance of success when individual helices are placed consecutively in alternative orders (permuted), whereas for all other oligomers we could not see a general advantage of one method over the other. Which method was more successful correlated neither with size nor with data resolution.

A potential problem is the replacement of all helices by an identical structure before the last model-building step. If that one helix is not sufficiently correct, it may decrease the overall model quality. In fact, for the dimer 3bas, which is rather asymmetric, the initially traced model is very close to the crystal structure, except for the terminal three turns of one helix. Duplication of the more complete chain worsened the model and resulted in an $R_{\rm free}$ of 0.38.

All current methods for *de novo* modelling are computationally intense. This is also the case for *CCsolve*, but the procedure is computationally lean enough that it can be run on a single multicore workstation. The majority of structures could be solved within 2 days on a 16-core machine starting from the protein sequence and the amplitude data. Because the program only relies on free and popular academic software, it should be relatively straightforward to install and use.

Successful structure determination using *CCsolve* depends on correct *de novo* structure prediction and the capabilities of *PHENIX* to build correct models. Despite clearly visible electron density, *AutoBuild* did not always add missing residues. In our procedure, we did not attempt to automatically include ligands, ions or buffer components, although those can contribute substantially to the overall electron density. Hence, even for already well refined models some further manual improvement is still necessary. Furthermore, we did not investigate whether a higher number of 'anchor' residues could increase the chances of finding the correct register. There might not be a global optimum for this number; removing side chains is a commonly used strategy to increase the chance of obtaining an MR solution (Schwarzenbacher *et al.*, 2004). Increasing the number of side chains may improve the results in some cases, but it may also have negative effects.

CCsolve can be downloaded freely from http://www.cmps.lu.se/biostruct/people/ingemar-andre/ccsolve.

4. Conclusions

Here, we have tested a fully automated procedure to obtain well refined crystal structure solutions of homomeric coiled coils using *de novo* models. Thus, the method outlined here should facilitate streamlined structure determination of coiled coils by MR when no homologous proteins are available. This is of particular interest for *de novo*-designed proteins, as these are not necessarily based on existing structures. Although *Fold-and-Dock* might be particularly well suited to predict the structure of coiled coils, it has been successfully applied to the prediction of other folds. Furthermore, none of the main steps in the described procedure is specifically tailored towards coiled coils. Thus, the procedure outlined here can potentially be applied to other small oligomers.

Acknowledgements

We thank Derek Logan and Susanna Törnroth-Horsefield for their critical reading of the manuscript and helpful discussions.

This work was supported by the Swedish Research Council (Vetenskapsrådet).

References

Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsl. Protein Crystallogr.* **42**, contribution 8.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

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

Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C. H., Szyperski, T. & Baker, D. (2009). *Proc. Natl Acad. Sci. USA*, **106**, 18978–18983.

Das, R. & Baker, D. (2008). *Annu. Rev. Biochem.* **77**, 363–382.

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

DiMaio, F. (2013). *Acta Cryst.* D**69**, 2202–2208.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwai, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature (London)*, **473**, 540–543.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Grigoryan, G. & Keating, A. E. (2008). *Curr. Opin. Struct. Biol.* **18**, 477–483.

Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. (1993). *Science*, **262**, 1401–1407.

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

Kammerer, R. A., Kostrewa, D., Progias, P., Honnappa, S., Avila, D., Lustig, A., Winkler, F. K., Pieters, J. & Steinmetz, M. O. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 13891–13896.

Leaver-Fay, A. *et al.* (2011). *Methods Enzymol.* **487**, 545–574.

Lee, D. L., Ivaninskii, S., Burkhard, P. & Hodges, R. S. (2003). *Protein Sci.* **12**, 1395–1405.

Lumb, K. J. & Kim, P. S. (1995). *Biochemistry*, **34**, 8642–8648.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

McClain, D. L., Binfet, J. P. & Oakley, M. G. (2001). *J. Mol. Biol.* **313**, 371–383.

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.

Monera, O. D., Kay, C. M. & Hodges, R. S. (1994). *Biochemistry*, **33**, 3862–3871.

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.

Rämisch, S., Lizatović, R. & André, I. (2015). *Proteins*, **83**, 235–247.

Ramos, J. & Lazaridis, T. (2006). *J. Am. Chem. Soc.* **128**, 15499–15510.

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.

Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). *Methods Enzymol.* **383**, 66–93.

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). *Nature Methods*, **10**, 1099–1101.

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

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

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 61–69.

Thépaut, M., Maiorano, D., Guichou, J.-F., Augé, M.-T., Dumas, C., Méchali, M. & Padilla, A. (2004). *J. Mol. Biol.* **342**, 275–287.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.

Tsatskis, Y., Kwok, S. C., Becker, E., Gill, C., Smith, M. N., Keates, R. A. B., Hodges, R. S. & Wood, J. M. (2008). *Biochemistry*, **47**, 60–72.

Yoon, M.-K., Kim, H.-M., Choi, G., Lee, J.-O. & Choi, B.-S. (2007). *J. Biol. Chem.* **282**, 12989–13002.