CCFold: rapid and accurate prediction of coiled-coil structures and application to modelling intermediate filaments

Dmytro Guzenko and Sergei V. Strelkov *

Department of Pharmaceutical and Pharmacological Sciences

KU Leuven, Leuven 3000, Belgium.

*To whom correspondence should be addressed.

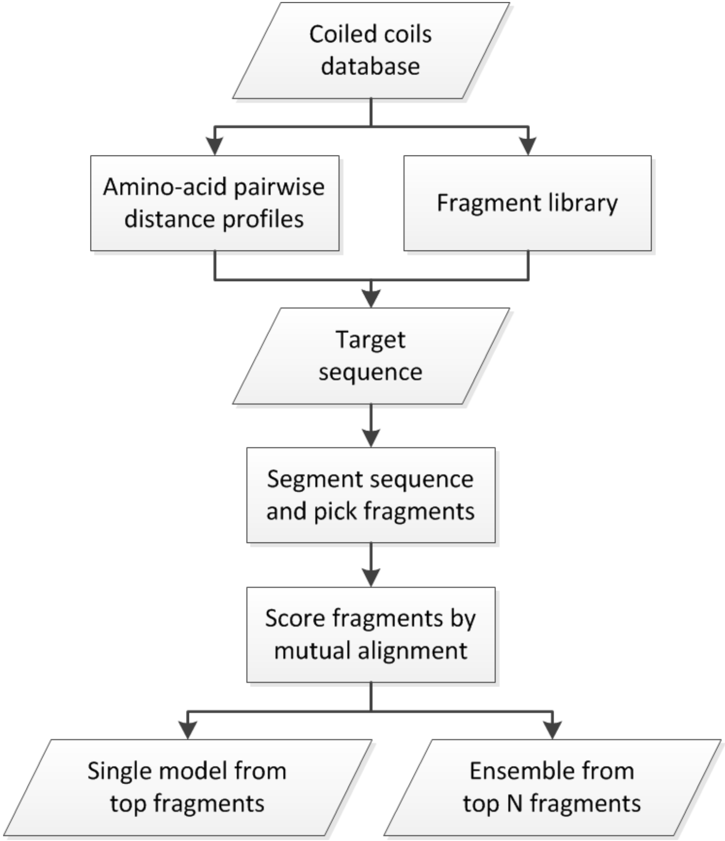
Abstract

Accurate molecular structure of the protein dimer representing the elementary building block of intermediate filaments (IFs) is essential towards the understanding of the filament assembly, rationalizing their mechanical properties and explaining the effect of disease-related IF mutations. The dimer contains a ∼300-residue long α-helical coiled coil which is not assessable to either direct experimental structure determination or modelling using standard approaches. At the same time, coiled coils are well-represented in structural databases. Here we present CCFold, a generally applicable threading-based algorithm which produces coiled-coil models from protein sequence only. The algorithm is based on a statistical analysis of experimentally determined structures and can handle any hydrophobic repeat patterns in addition to the most common heptads. We demonstrate that CCFold outperforms general-purpose computational folding in terms of accuracy, while being faster by orders of magnitude. By combining the CCFold algorithm and Rosetta folding we generate representative dimer models for all IF protein classes. The source code is freely available at https://github.com/biocryst/IF

1 Introduction

Intermediate filaments (IFs) are an important example of a protein assembly based on α-helical coiled coils (CCs). IFs together with microtubules and actin filaments are key elements of the cytoskeleton in vertebrate cells. The human body contains over 70 different IF proteins, belonging to five major classes by sequence similarity and including both cytoplasmic and nuclear IFs. Inherited and sporadic mutations in IF genes were linked to numerous diseases, including muscle, heart, skin and neurological disorders (Omary, 2009), all of which are incurable at present. Currently our structural understanding of IFs is scarce, particularly when compared to that of actin and microtubules. This problem is linked to both their complexity and partial disorder. The elementary building block of all IFs is an elongated dimer with a length of ∼45 nm and a diameter of 2-3 nm. The overall structure of the dimer is defined by the formation of a CC by its conserved central 'rod' domain. This domain contains three CC segments denoted coil1A, coil1B and coil2 which

Figure 1: Flowchart of the CCFold algorithm.


Coiled coils database
Amino-acid pairwise distance profiles
Fragment library
Target sequence
Segment sequence and pick fragments
Score fragments by mutual alignment
Single model from top fragments
Ensemble from top N fragments

are interconnected by short linkers L1 and L12. The rod is flanked by the intrinsically disordered, highly variable head and tail domains (Chernyatina *et al.*, 2015). Following our divide-and-conquer strategy (Strelkov *et al.*, 2001) multiple short fragments of the IF rod could be resolved at atomic detail using X-ray crystallography (Guzenko *et al.*, 2017), while a full-length IF dimer is clearly unsuitable for crystallisation (Chernyatina *et al.*, 2016).

In general, CC is a widespread structural motif in proteins involved in a multitude of functions (Lupas and Bassler, 2016). Its idealised geometry can be described with just a few parameters (Crick, 1953). The Crick parametrisation is routinely used to analyse existing CC structures (Strelkov and Burkhard, 2002) or to create theoretical models thereof (Offer *et al.*, 2002; Wood *et al.*, 2014). The driving force of CC formation is a regular pattern of hydrophobic amino acids that allows two or more α-helices to associate together, forming a common hydrophobic core. The most widespread pattern is a heptad (7-residue motif) HxxHxxx, where H indicates a hydrophobic amino acid and x stands for any amino acid. This pattern, with residue positions traditionally labelled *abcdefg*, results in a 'canonical' left-handed CC. Another possibility is a hendecad (11-residue repeat) HxxHxxxHxxx, which corresponds to an addition of a four-residue block (stutter) to a heptad (Lupas and Gruber, 2005). Such a pattern supports parallel packing of the chains. Two consecutive stutters after a heptad define a quindecad (15-residue repeat), promoting right-handed supercoiling. Further variation of the CC geometry is possible by insertions of three-residue blocks (stammers). In general, stammers demand a more pronounced left-handed supercoil, while stutters cause its unwinding and eventually a switch to a right-handed geometry (Lupas and Gruber, 2005).

Sequences featuring long regions with heptad patterns result in structures with a more or less uniform left-handed supercoiling and as such are the most straightforward to model (Grigoryan and DeGrado, 2011). However, sequences of many naturally occurring CCs, including those found in the IF rod, contain intermixed patterns. Experimental data reveal that such transitions can often be accommodated within a continuously α-helical structure, causing adaptation of the local CC parameters, so that the hydrophobic core packing is preserved (Strelkov and Burkhard, 2002). For such less regular CC structures, accurate modelling remains a challenging task. The use of Rosetta, a popular general-purpose protein folding algorithm (Leaver-Fay *et al.*, 2011), for CC structures was recently described (Rämisch *et al.*, 2015). Upon composing the initial fragment library from α-helices found in CC structures and employing an asymmetric 'fold-and-dock' protocol, accurate models of CCs could be obtained, including those with deviations from the heptad pattern. However, the use of Rosetta required substantial computational time (weeks) even for short CCs. For the ~300-residue rod domain of the elementary IF dimer this is computationally prohibitive.

At the same time, the relative simplicity and the 'linear' nature of a parallel CC (meaning that the residues located further away in the sequence do not interact with each other) suggest that it may be amenable to reliable structure prediction. A logical option is to employ the 'threading' approach which can be efficient towards the *in silico* prediction of protein structure (Pieper *et al.*, 2014). Indeed, there are ample experimental data for CC proteins, which can be conveniently assessed through the CC+ database (Testa *et al.*, 2009). At the time of writing it contains nearly 15 thousand CCs of 11 residues or longer, 450 of which include non-canonical (*i.e.*, other than heptad) patterns.

Here we describe a novel threading-based algorithm, CCFold, specifically designed for the prediction of CC structure. It is based on picking multiple CC fragments for short overlapping segments of the input sequence. For each fragment, the probability of the correct match between the sequence and structural features, such as the presence of certain residues in the core positions of the CC, is evaluated with the help of statistics obtained from the analysis of the CC+ database. The best-scoring fragments along the protein sequence are merged together through an optimization procedure. The algorithm was implemented and optimised for the parallel dimeric CCs, such as found within the IF dimer, but can also be applied to other orientations and oligomeric states. We demonstrate that CCFold clearly outperforms the existing approaches in terms of both accuracy and speed. Further, by combining the CCFold for the CC domains and general Rosetta folding we produce representative models of the elementary dimers for all main IF protein classes. Here, the availability of X-ray structures of multiple IF rod fragments allows us to further evaluate the performance of the algorithm.

2 System and Methods

Flowchart of the CCFold algorithm is shown in Fig. 1.

2.1 Amino-acid profile of pairwise distances in parallel dimeric CCs

All parallel in-register dimeric CCs were extracted from the CC+ database and truncated to polyalanine. Chain sequences were clustered at 80% sequence identity, which resulted

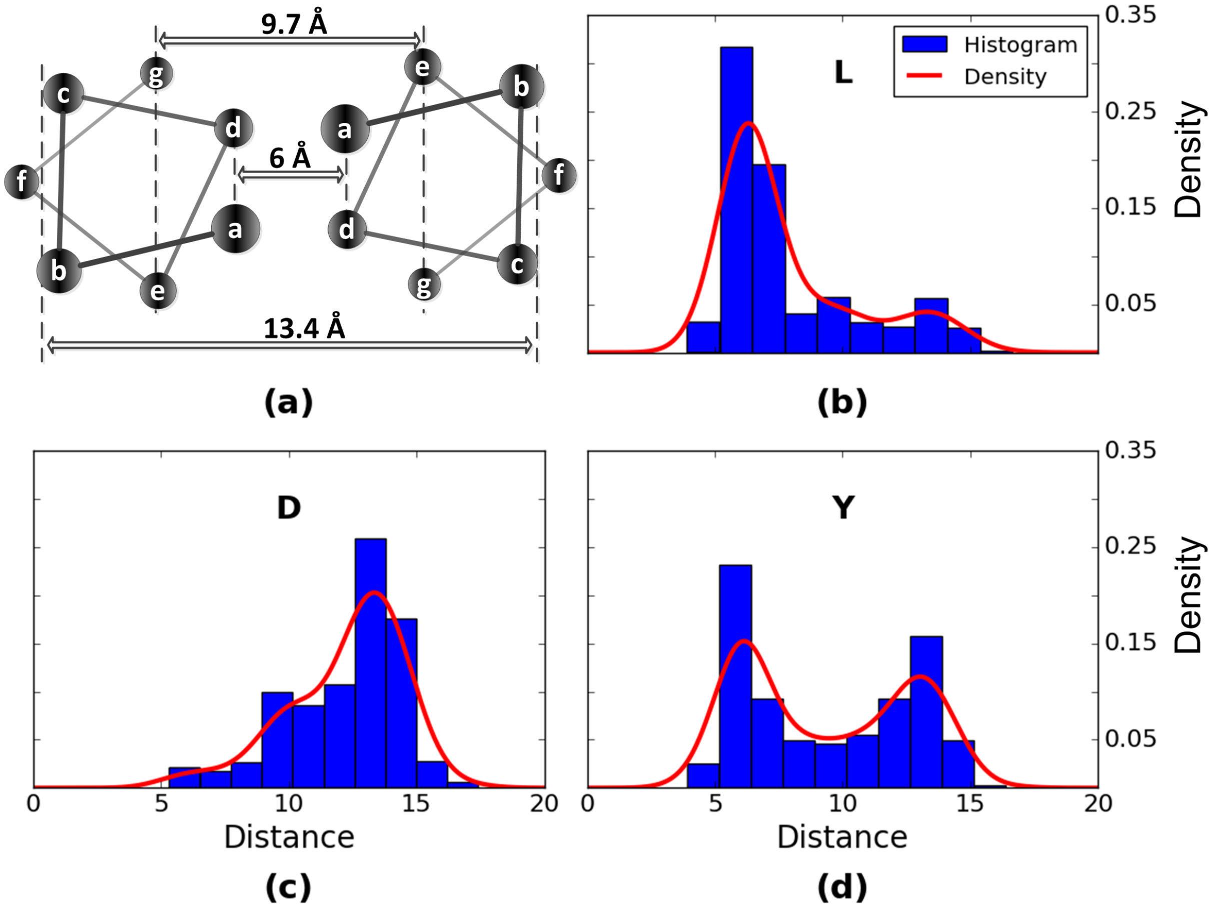


Figure 2: (a) Schematic diagram of an idealised CC dimer showing pairwise distances between Cα atoms of the corresponding residues in both chains. (b,c,d) Distributions of such distances obtained from the CC+ database for valine, aspartic acid and tyrosine residues, respectively. Kernel density approximation is plotted over the original histogram.

in 336 non-redundant CCs containing from 15 to 148 residues per chain. Statistics was collected on the distance between the Cα-atoms of aligned residues in both chains (such as the equivalent residues in the case of parallel in-register homodimers). The number of distances measured for a particular amino acid ranged from 21 (proline) to 1899 (leucine), making a total of 11847 measurements. As seen in Fig. 2 and S1, the distributions readily reveal three major maxima, which structurally correspond to residues situated in heptad positions *a* or *d* (shortest distance), *e* or *f* (intermediate distance), and *b*, *c* or *f* (longest distance), respectively. At the same time, the distance distributions vary greatly across the 20 amino acids, which ultimately appears the main reason towards the high predictive power of our algorithm.

2.2 CC fragment library

Next, we wanted to capture the structural diversity of experimentally determined CCs, regardless of the particular amino-acid sequences that produced such structures. To this end, a set of all possible 2x15-residue fragments of parallel two-stranded CCs was extracted from the CC+ database, truncated to polyalanine. After all-*vs.*-all coordinate root-mean-square deviation (RMSD) comparison, we selected a non-redundant subset

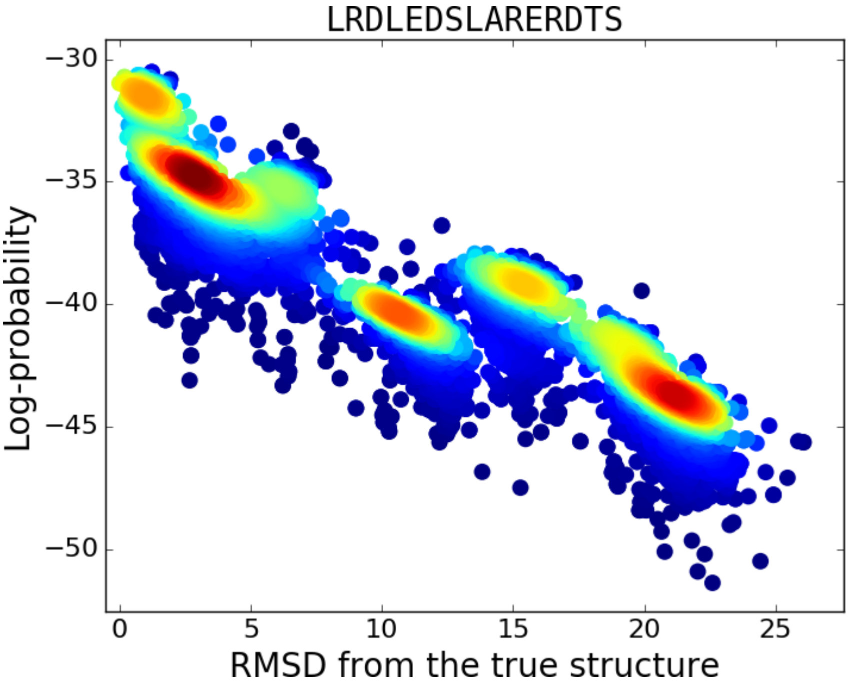


Figure 3: Scatter plot comparing log-probability (eq. 1) of each of 14175 database fragments for a 15-residue segment of human lamin A (residues 320 to 334) and their Cα RMSD with respect to the corresponding part of the crystal structure (PDB entry 1X8Y). Colour indicates the data point density.

having a pairwise RMSD between any two fragments of at least 0.2Å. This resulted in a library of 14175 dimeric CC fragments void of sequence information.

2.3 Delineation of CC domains

The region of the target protein sequence that is expected to form a CC structure can be predicted using standard tools (Li *et al.*, 2015). In addition, the expected multiplicity of the CC (dimer, trimer, *etc.*) and the orientation of the helices (parallel or anti-parallel) must be provided. Several methods to predict the CC multiplicity exist (Vincent *et al.*, 2013; Trigg *et al.*, 2011). The current implementation of CCFold is focused on parallel dimeric CCs, with limited support of the anti-parallel orientation for benchmarking purposes.

3 The CCFold algorithm

Starting input is the amino-acid sequence to be assembled into a parallel in-register dimeric CC. The main steps of the algorithm include segmenting this sequence into a series of overlapping sequence windows, picking a pool of fragments for each window, selecting a set of fragments that optimally agree with each other in the overlapping regions, and finally merging them together.

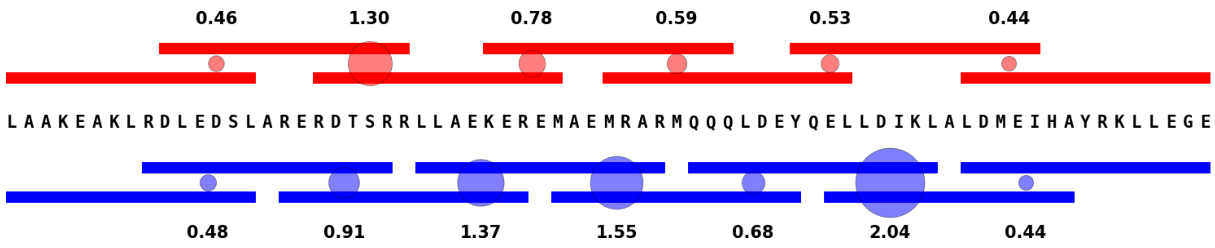

0.46
1.30
0.78
0.59
0.53
0.44
L A A K E A K L R D L E D S L A R E R D T S R R L L A E K E R E M A E M R A R M Q Q Q L D E Y Q E L L D I K L A L D M E I H A Y R K L L E G E
0.48
0.91
1.37
1.55
0.68
2.04
0.44

Figure 4: An example of sequence segmentation. Optimized set of overlapping 15-residue windows for a lamin A fragment (residues 313 to 383), shown on top. For comparison, a simple set of windows with constant overlap is shown at bottom. Average overlap RMSD (in Å) between the fragments in the windows is illustrated by circles.

3.1 Sequence segmentation

For a target sequence of length N, initially a full set of $N - 14$ possible overlapping sequence windows of length 15 is considered. For every window we evaluate the probability that it would form each of the fragments from our library, assuming independence between the amino acid pairs:

$$P(f|A) = \prod_{i=1}^{15} D(d(f_i)|a_i), \tag{1}$$

where a_i is the i-th amino acid of the sequence window A, $d(f_i)$ is the distance between the aligned Cα atoms of the i-th residue within the fragment f, and $D(d(f_i)|a_i)$ is the kernel density function for amino acid a (Fig. 2). Importantly, the function (1) correlates well with the likeness to the true structure in terms of minimal coordinate RMSD even for non-canonical CCs. Fig. 3 demonstrates this for an experimentally determined coil2 fragment of lamin A containing a stutter (Strelkov *et al.*, 2004).

At this point, a particular set of overlapping 15-residue sequence windows covering the entire target sequence is selected. To this end, 100 top structural fragments are picked by probability (1) for every possible window. The optimal segmentation of the target sequence is given by a two-pass dynamic programming as detailed below.

Let $\mathbf{F}_i$ be a set of fragments for the i-th sequence window and let o_i be a set of overlapping residues between fragment sets $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$, number of which is bounded by reasonable values $o_{min} \leq |o_i| \leq o_{max}$. The segmentation of the target sequence is defined as $\mathcal{O} = \langle o_1, o_2, ..., o_{K_\mathcal{O}-1} \rangle$, with number of resulting windows $K_\mathcal{O}$ varying for different segmentations. Let $\mathcal{R}(f_i, f_{i+1}|o_i)$ denote the RMSD of two fragments f_i and f_{i+1} from consecutive fragment sets $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$ superposed by the residues in o_i. Further, let $\mathcal{R}(\mathbf{F}_i, \mathbf{F}_{i+1}|o_i)$ denote the average RMSD of all fragments from $\mathbf{F}_i$ and all fragments from $\mathbf{F}_{i+1}$. The first pass of dynamic programming is employed to minimize the maximal average overlap RMSD:

$$R_{mm} = \min_{\mathcal{O} \in \mathbf{O}} \max_{1 \leq i < K_\mathcal{O}} \mathcal{R}(\mathbf{F}_i, \mathbf{F}_{i+1}|o_i), \tag{2}$$

where $\mathbf{O}$ is a set of all possible overlaps. Afterwards the second pass minimizes the total alignment RMSD under condition that no single one surpasses the value determined at the first step:

$$\mathcal{O} = \operatorname*{argmin}_{\mathcal{O} \in \mathbf{O}} \left\{ \sum_{1 \leq i < K_\mathcal{O}} \mathcal{R}(\mathbf{F}_i, \mathbf{F}_{i+1}|o_i) : \max_{1 \leq i < K} \mathcal{R}(\mathbf{F}_i, \mathbf{F}_{i+1}|o_i) \leq R_{mm} \right\} \tag{3}$$

An example of a resulting sequence segmentation is shown on Fig. 4. We have found that such an approach yields more accurate models compared to just using a fixed set of overlapping fragments with constant overlap, especially in the cases when non-canonical repeats are present or the hydrophobic core assignment is ambiguous.

3.2 Optimal set of overlapping structural fragments

Next, for each sequence window from the selected overlapping set we choose a large pool (200-1000) of the most probable structural fragments. Logically, the majority of such fragments in the majority of windows should represent the correct fold for a given sequence. Based on this assumption, we re-score the fragments in the pools according to how well they align with fragments selected in other windows, using sum-product belief propagation (Barber, 2012). The factor graph required for this procedure is constructed by representing each sequence window by a variable node, and every overlap between two windows by a factor node that specifies alignment RMSD between the overlapped fragments.

3.3 Output model generation

As the first option, the final model is constructed by simply merging the top-scoring fragments in each window. To this end the fragments are superimposed upon RMSD minimisation. Thereafter the atomic coordinates in the overlapping parts are averaged with weights decreasing towards the ends of the merged fragments. The resulting single model should be the most accurate one locally, but its overall shape (such as the bending of the CC axis in particular) is not controlled.

Alternatively, an ensemble model can be output. The advantage of such a model is that it collectively reveals the uncertainty of the structural prediction. In addition, ensemble models may be desired for some applications such as for phasing X-ray data using molecular replacement (MR). The procedure to obtain an ensemble model is as follows.

Let each sequence window be represented by any of 20 to 100 of highest-scoring fragments. This enables the solution of a 'shortest path' problem, *i.e.*, choosing fragments that result in the lowest sum of overlap RMSDs throughout the whole structure. Let us define a *conditional* shortest path, the condition being that the path is restricted to a particular fragment f in a certain window $\mathbf{F}_j$. With $\mathbf{F}\backslash\mathbf{F}_j = \mathbf{F}_1\times\ldots\times\mathbf{F}_{j-1}\times\mathbf{F}_{j+1}\times\ldots\times\mathbf{F}_K$ we define:

$$\mathrm{SP}(\mathcal{O},\mathbf{F}|\mathbf{F}_j = f) = \underset{\substack{f_1,\ldots,f_K\in\mathbf{F}\backslash\mathbf{F}_j \\ f_j=f}}{\operatorname{argmin}} \{\sum_{1\leq i<K} \mathcal{R}(f_i, f_{i+1}|o_i)\} \tag{4}$$

Next, we condition eq. (4) on all fragments in all windows to produce an ensemble model as follows:

$$\mathrm{SP}_{\mathrm{ens}}(\mathcal{O},\mathbf{F}) = \bigcup_{j=1..K} \bigcup_{f\in F_j} \mathrm{SP}(\mathcal{O},\mathbf{F}|F_j = f) \tag{5}$$

Each individual model resulting from the procedure (5) will differ from the rest in at least one 15-residue fragment. Moreover, choosing a different fragment in one window may cause selection of alternative fragments in other windows to satisfy the RMSD minimisation criterion.

Finally, the procedure (5) can be used to produce a model with a nearly straight CC axis at a small cost to the local accuracy. Let the sequence $S = \langle f_1, f_2, ..., f_K \rangle$ of chosen fragments resulting from eq. (4) be characterised by the total fragment overlap RMSD $\mathcal{R}(S|\mathcal{O}) = \sum_{1 \leq i < K} \mathcal{R}(f_i, f_{i+1}|o_i)$, which naturally reflects overall quality of the produced model. As an indicator of straightness, we use superposition RMSD of the two chains of the full-length model $\mathcal{R}_{sym}(S)$. To combine these two measures, models with higher-than-median overlap RMSD are discarded from the ensemble, and the model with minimal inter-chain RMSD is selected among the rest of the ensemble:

$$\begin{aligned} \mathrm{SP}_{\mathrm{sym}}(\mathcal{O}, \mathbf{F}) &= \underset{S \in \mathrm{SP}_{\mathrm{ens}}(\mathcal{O}, \mathbf{F})}{\operatorname{argmin}} \{\mathcal{R}_{sym}(S) : \mathcal{R}(S|\mathcal{O}) \leq M\} \\ M &= \underset{S \in \mathrm{SP}_{\mathrm{ens}}}{\operatorname{median}} \{\mathcal{R}(S|\mathcal{O})\} \end{aligned} \tag{6}$$

An example of output models depending on the options chosen is given in Fig. S2.

4 Implementation

The algorithm is implemented as a Python script. Biopython (Cock *et al.*, 2009) is used to process the Protein Data Bank (PDB) files (Berman *et al.*, 2000). Structure superpositions and RMSD calculations are done with PyRMSD (Gil and Guallar, 2013). The smooth density functions are estimated using the Gaussian kernel available in Scikit-learn (Pedregosa *et al.*, 2011), with the bandwidth set to 1.

Program parameters are as follows:

- target=one,termini,all. Whether to produce one model, a collection of models with alternative terminal fragments, or a collection with alternative fragments in all positions.
- straighten=true,false. Whether to produce models with a straight axis.
- segmentation˙pool (default 100). Number of fragments used to determine the optimal segmentation of the sequence.
- belief˙propagation˙pool (default 350). Number of top fragments in each sequence window selected for the belief propagation procedure.
- shortest˙path˙pool (default 50). Number of top fragments in each window selected for producing the shortest paths given the segments. Not applicable if target=one.

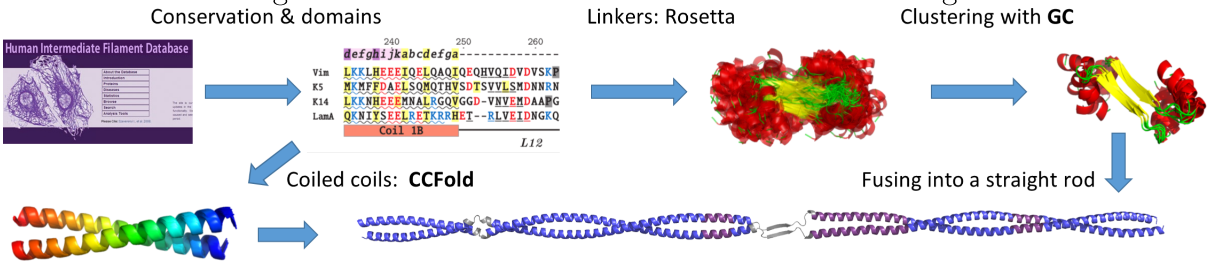
5 Modelling of the IF rod domain

The overall flow of the modelling is shown in Fig. 5.

5.1 CC segments of the IF rod

IF protein sequences were retrieved from the Human Intermediate Filament Database (Szeverenyi *et al.*, 2008). The CC segments (coil1A, coil1B and coil2) of each protein were defined as in Chernyatina *et al.* (2015) and fed into the CCFold algorithm. In case of heterodimers, such as for type I/II keratins, these were pairs of aligned sequences. Ensembles including 100 models with alternative N- and C-terminal fragments were produced. Straight models were preferred.

Figure 5: Flowchart of the IF rod domain modelling.


Conservation & domains
Human Intermediate Filament Database
240
250
260
defghijkabcdefga----------------
Vim LKKLHEEEIQELQAQIQEQHVQIDVDVSKP
K5 MKMFFDAELSQMQTHVSDTSVVLSMDNNRN
K14 LKKNHEEEMNALRGQVGGD-VNVEMDAAPG
LamA QKNIYSEELRETKRRHET--RLVEIDNGKQ
Coil 1B
L12
Linkers: Rosetta
Clustering with GC
Coiled coils: CCFold
Fusing into a straight rod

5.2 Linkers L1 and L12

Rosetta modelling suite (Leaver-Fay *et al.*, 2011) was employed as general structure prediction method for the short linkers that interconnect the CC segments. The asymmetric fold-and-dock protocol (Rämisch *et al.*, 2015) was used. This was necessary for the modelling of heterodimers such as type I/II keratins. In addition, for linker L12 a parallel β-hairpin structure with a two-residue offset of the two chains could be stably produced.

To this end, the sequences of the linkers were first ’capped’ with eight-residue stretches of an α-helix at either end. ’AtomPair’ distance constraints of 6Å were placed on the Cα-atoms of hydrophobic core residues of these ’CC caps’, to simulate the context of a dimeric rod. 5000 decoys were produced for each linker. 500 top-scoring decoys were analysed with our GC algorithm (Guzenko and Strelkov, 2017), and those forming the largest cluster were taken. Such an asymmetric procedure implied that the CC caps at either end of the linker were generally not aligned along the same axis. While the linkers are widely assumed to serve as points of flexibility of the rod domain, a ’straight’ model is preferable as the default conformation. Accordingly, we have implemented a genetic algorithm that modifies backbone torsion angles of the linker residues which lack secondary structure in order to bring the CC domains to the same axis, as illustrated by Fig. S3. Finally, the modelled linkers were merged with the flanking CC domains using the caps.

5.3 Assembly of the complete rod domain

The complete rods were constructed from the obtained pools of models for the CC segments and the two linkers by superposing the overlapping parts. The side chains were then added with SCWRL4 (Krivov *et al.*, 2009). Finally, the complete models were energy-minimized using REFMAC5 (Murshudov *et al.*, 2011).

6 Results

6.1 CCFold algorithm validation

The CCFold algorithm enables modelling of dimeric CC structures with just the amino-acid sequence on input. The procedure is fast, allowing complete modelling of a ~100-residue CC within a few seconds when using a regular contemporary PC. In particular, we have modelled 10 dimeric CC structures previously used as a benchmarking set for the asymmetric fold-and-dock (AFnD) protocol (Rämisch *et al.*, 2015). Anti-parallel dimers were modelled with a modified version of the CCFold algorithm utilising a database of anti-parallel fragments and corresponding pairwise distance distributions. To prevent

Table 1: Performance comparison of CCFold and AFnD. Better values are in bold.

PDB ID	lDDT score	—	TM score	
	CCFold	AFnD	CCFold	AFnD
1hf9	**0.809**	0.750	**0.855**	0.668
1pl5	**0.967**	0.931	**0.935**	0.915
1r48	**0.845**	0.810	**0.781**	0.728
1t3j	**0.920**	0.881	**0.884**	0.866
1uii	0.867	**0.910**	**0.866**	0.850
1x8y	**0.893**	0.811	**0.936**	0.857
2oqq	**0.942**	0.941	0.924	**0.930**
2q6q	**0.981**	0.963	**0.961**	0.923
2w6a	**0.965**	0.943	**0.933**	0.925
3bas	**0.946**	0.910	**0.917**	0.868

any bias from the known experimental structures, homologues of the target structure at PSIBLAST E-value of 0.05 were excluded from the database.

The models output by CCFold and the best decoys obtained by the AFnD protocol were both compared to the experimentally determined structures using two orthogonal criteria, namely the lDDT score which evaluates the local model quality (Mariani *et al.*, 2013), and the TM score which assesses the similarity of the global fold (Zhang and Skolnick, 2004). Only Cα atoms were considered. As seen from Table 1, 8 out of 10 resulting models produced by our method are superiour by both local and global similarity measures. The remaining two targets show a better result by one of the measures and worse by the other. At the same time, our algorithm is radically faster than the Rosetta-based AFnD procedure.

6.2 IF dimer structure

We have used the CCFold algorithm to produce representative rod domain structures for each of the main sequence homology classes of the IF family (Fig. 6). The resulting models were superposed with the available crystal structures of the rod fragments. The structural agreement is quite good, as the observed deviations are largely confined to the bending of the CC axis of the experimental structures, which are likely due to the crystal contacts they are involved in.

Most structural features of the rod are preserved across various IF classes, which is in line with a considerable sequence conservation (Guzenko *et al.*, 2017). Coil1A always appears as a relatively short CC segment with a regular left-handed structure. Coil1B also consistently reveals a left-handed structure for the most part (whereby coil1B of lamins contains an insert of 42 residues or 6 heptads compared to cytoplasmatic IF proteins), except for a single hendecad and therefore a local unwinding near its C-terminus. Finally, coil2 domain is absolutely conserved in both length and hydrophobic repeat pattern throughout all IF classes, and contains a parallel α-helical bundle (3 hendecad repeats) at its N-terminus, plus a single extra hendecad repeat (equivalent to a stutter) approximately two thirds into its length (Fig. 6).

In line with previous analyses (Kapinos *et al.*, 2010), the linker L1 of nuclear lamins could be modelled as fully α-helical, so that a continuous left-handed CC is formed

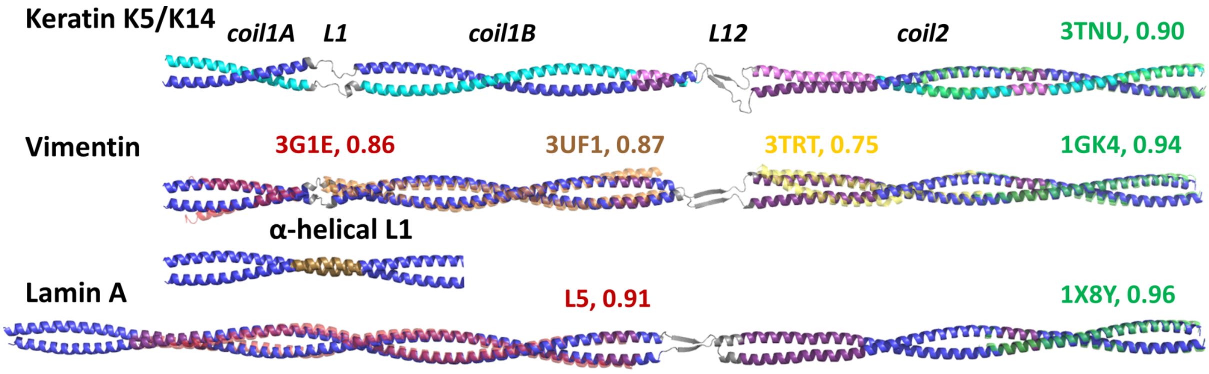


Figure 6: IF rod domain models. Regions with hendecad repeats are highlighted in violet. Linker regions are shown in grey. A possible fully α-helical conformation of linker L1 in vimentin is also shown, with 19-residue repeat highlighted in brown. The model is superposed with the available crystal structures of fragments, with PDB codes and TM scores of the superposition indicated.

running from coil1A into coil1B, with a local unwinding (2 hendecads) at the linker. In cytoplasmic IF proteins, this linker was traditionally considered as a non-helical and flexible (Smith *et al.*, 2002), although recent experimental data suggested a rigid structure (Aziz *et al.*, 2012). Interestingly, by feeding the complete coil1A-L1-coil1B sequence of vimentin into the CCFold algorithm we could obtain a continuous CC with a single 19-residue motif at positions 135-153. While experimental verification of the latter possibility is still necessary, in Fig. 6 we present two alternative conformations for vimentin linker L1. Finally, in line with earlier sequence-based predictions (Parry and Steinert, 1999), our modelling of linker L12 yielded a short stretch of a β-structure in all IF types. Interestingly, the predicted parallel β-hairpin often reveals a two-residue shift of one chain relative to another.

We have further assessed the quality of the predicted IF rod structure by analysing its CC geometry using the program Twister (Strelkov and Burkhard, 2002). Indeed, comparison of the geometrical parameters such as CC pitch and radius for the *in silico* model and crystal structures is an efficient means to evaluate the local modelling accuracy. Fig. S4a shows the inverse of the CC pitch angle which reveals the local CC geometry (left-handed or right-handed supercoiling and its pitch), along the length of our vimentin rod model and in experimental fragments. The local unwinding of the left-handed CC at the positions of hendecad repeats, as manifested by a pronounced drop in the inverse CC pitch value, is consistently observed in both our model and crystal structures. At the same time, our model has a practically constant CC pitch (∼147 Å) and CC radius (∼4.9 Å) throughout the regions with regular heptad repeats, apparently corresponding to the average values in the fragment database, while the experimental structures reveal more local deviations (Fig. S4b). Although some of the latter may be related to both crystal contact artefacts and truncation effects, there can also be structural reasons, in particular the size of the side chains in core positions, towards 'real' variations of CC radius and pitch along the length of the structure. Indeed, for CC fragments crystallized as several copies in the asymmetric unit, similarities in the CC pitch profile were observed (Strelkov and Burkhard (2002), Fig. 5). At this moment, the CCFold algorithm is not able to simulate these fine details. Moreover, we note that even a small variation of the CC pitch value will naturally result in a major relative rotation of the ends of a sufficiently long CC. These global differences can be decisive in some applications, such as crystallographic phasing by MR. Particularly in the latter case we recommend to use the CCFold to generate an ensemble of models, which could ultimately be crucial for a successful phasing.

In addition, using the CCFold algorithm, we have produced dimer models of isomin and crescentin, which had previously been proposed as IF-like proteins in insects (Mencarelli *et al.*, 2011) and bacteria (Ausmees *et al.*, 2003), respectively. In line with previous analyses (Herrmann and Strelkov, 2011) pointing to conservation of several sequence features, the isomin dimer model revealed a CC structure resembling that of the nuclear lamins (Fig. S5a). This included a short linker corresponding to L1 as well as a parallel α-helical bundle near the beginning of coil2. However, our new modelling suggested a possibility of a fully helical structure for the linker L12 in isomin. Also the coil2 of isomin contains an insertion of 2 residues in the place of a traditional stutter in IF proteins, resulting in an additional short linker. In contrast, our modelling of crescentin yielded a continuous CC structure without any linkers, predominantly based on regular heptads with only two instances of stutters/hendecads (Fig. S5b). We conclude that crescentin dimer bears little resemblance to the segmented IF rod domain.

6.3 Application to molecular replacement

By serendipity, further evidence on the quality of CC models obtained by the CCFold algorithm could be obtained. In the past, we have collected X-ray diffraction data for crystals of a lamin A fragment (residues 65-222 corresponding to linker L1 and coil1B; Chernyatina *et al.*, in preparation). All previous attempts to phase the data by MR failed, despite a large number of search models tested. Indeed, the MR procedure for CC structures is known to be challenging in general and highly sensitive to the quality of the search model (Guzenko *et al.*, 2017). However, an MR search using a model of this fragment produced by the CCFold algorithm was successful. The refined crystal structure (fragment L5 on Fig. 6) shows only minor deviations from the modelled structure and in particular supports a fully α-helical conformation for the linker L1.

7 Discussion

Existing methods to predict CC domains almost exclusively employ heptad repeat as the main paradigm to define a CC (Li *et al.*, 2015). Correspondingly, such algorithms tend to assign lower CC scores to regions with other repeat patterns, which are only seen as 'discontinuities' in the heptad periodicity. With time, however, decads, hendecads, quindecads, *etc.* were determined experimentally to be consistent with a continuous CC structure (Lupas and Gruber, 2005). With this in mind, here we do not attempt to route the modelling of a particular protein sequence towards this or that type of CC periodicity, but pose a more general question: can this sequence be folded into a continuous CC with a plausible hydrophobic core? To answer this question we base ourselves on threading through all available experimental CC structures.

By threading fragments of CCs rather than single helices we eliminate the necessity of docking the helices together, as in the fold-and-dock approach (Rämisch *et al.*, 2015). Moreover, α-helices forming a CC exhibit less structural variation than a combination of independent helices (Grigoryan and DeGrado, 2011). Together with fragment picking according to the amino-acid sequence and re-scoring based on mutual alignment of the fragments, this approach drastically reduces the conformational search space.

It should be noted that the starting hypothesis of a CCFold run on a given sequence is that this sequence indeed forms a dimeric parallel CC. Thus the algorithm would attempt to construct a continuous CC even in cases when this hypothesis is wrong. In order to see how the CCFold algorithm deals with this situation, we tried to deliberately feed in amino-acid sequences not forming a CC. The models output in these cases include clear structural abnormalities, for instance a deviation of the α-helices from the optimal geometry (3.6 residues per turn and the canonical hydrogen bonding pattern of main-chain atoms). Thus, beyond its initial purpose as a *de novo* structural modelling tool, our algorithm can be useful as a predictor of CC domains. At the same time, in borderline cases the CC structure produced by our algorithm may still need experimental verification, such as for the alternative model of vimentin rod with fully α-helical linker L1 (Fig. 6). The predictive power of CCFold directly relates to the quality of the underlying structural database and inclusion of various possible types of CC geometry. Thus the accuracy of our predictions should further increase with time, as more CC structures are determined experimentally.

The CCFold algorithm brings the modelling of the IF dimer structure to a new level. Past attempts included choosing some starting conformation, based either on the available crystal structures of protein fragments or some rather crude structural assumptions, followed by molecular dynamics (MD) (Chou and Buehler, 2012; Bray *et al.*, 2015). The problem here is that, given the size of the IF dimer and current computational capacities, such MD simulations could never guarantee a convergence to a correct fully energy-minimised structure. Typically, only a locally optimized structure could be obtained, which was still principally defined by the starting conformation.

Here we have presented an accurate molecular structure of the conserved rod domain, the 'signature' feature of the IF protein class. This directly provides for the modelling of the full IF dimer, since the variable, intrinsically disordered N- and C-terminal domains can readily be simulated using standard Rosetta tools (Chernyatina *et al.*, 2015). The knowledge of the dimer structure is indispensable towards the understanding of the process of IF assembly. Indeed, experimental constraints for the higher-level association of dimers are available, including chemical crosslinks in particular, as well as a more detailed information on the association of dimers into tetramers obtained through crystallography and site-directed spin labelling (Chernyatina *et al.*, 2016). New improvements in cryoEM technique, as recently applied to nuclear lamins (Turgay *et al.*, 2017), also suggest that locating individual dimers in tomographic reconstructions is within reach. Altogether, these advances suggest that building up a complete 3D atomic model of an IF may be achieved in the near future.

References

Ausmees, N., *et al.* (2003). The bacterial cytoskeleton: an intermediate filament-like function in cell shape. *Cell*, **115**(6), 705–713.

Aziz, A., *et al.* (2012). The structure of vimentin linker 1 and rod 1b domains characterized by site-directed spin-labeling electron paramagnetic resonance (sdsl-epr) and x-ray crystallography. *Journal of Biological Chemistry*, **287**(34), 28349–28361.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Berman, H. M., *et al.* (2000). The protein data bank. *Nucleic acids research*, **28**(1), 235–242.

Bray, D. J., *et al.* (2015). Complete structure of an epithelial keratin dimer: implications for intermediate filament assembly. *PloS one*, **10**(7), e0132706.

Chernyatina, A. A., *et al.* (2015). Intermediate filament structure: the bottom-up approach. *Current opinion in cell biology*, **32**, 65–72.

Chernyatina, A. A., *et al.* (2016). How to study intermediate filaments in atomic detail. *Methods in enzymology*, **568**, 3–33.

Chou, C.-C. and Buehler, M. J. (2012). Structure and mechanical properties of human trichocyte keratin intermediate filament protein. *Biomacromolecules*, **13**(11), 3522–3532.

Cock, P. J., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Crick, F. H. (1953). The Fourier transform of a coiled-coil. *Acta crystallographica*, **6**(8-9), 685–689.

Gil, V. A. and Guallar, V. (2013). pyRMSD: a Python package for efficient pairwise rmsd matrix calculation and handling. *Bioinformatics*, **29**(18), 2363–2364.

Grigoryan, G. and DeGrado, W. F. (2011). Probing designability via a generalized model of helical bundle geometry. *Journal of molecular biology*, **405**(4), 1079–1100.

Guzenko, D. and Strelkov, S. V. (2017). Granular clustering of de novo protein models. *Bioinformatics*, **33**(3), 390–396.

Guzenko, D., *et al.* (2017). Crystallographic studies of intermediate filament proteins. In D. A. Parry and J. M. Squire, editors, *Fibrous Proteins: Structures and Mechanisms*, pages 151–170. Springer International Publishing.

Herrmann, H. and Strelkov, S. V. (2011). History and phylogeny of intermediate filaments: now in insects. *BMC biology*, **9**(1), 16.

Kapinos, L. E., *et al.* (2010). Characterization of the head-to-tail overlap complexes formed by human lamin a, b1 and b2 ”half-minilamin” dimers. *Journal of molecular biology*, **396**(3), 719–731.

Krivov, G. G., *et al.* (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, **77**(4), 778–795.

Leaver-Fay, A., *et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, **487**, 545.

Li, C., *et al.* (2015). Critical evaluation of in silico methods for prediction of coiled-coil domains in proteins. *Briefings in bioinformatics*, page bbv047.

Lupas, A. N. and Bassler, J. (2016). Coiled coils–a model system for the 21st century. *Trends in Biochemical Sciences*.

Lupas, A. N. and Gruber, M. (2005). The structure of α-helical coiled coils. *Advances in protein chemistry*, **70**, 37–38.

Mariani, V., *et al.* (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**(21), 2722–2728.

Mencarelli, C., *et al.* (2011). Isomin: a novel cytoplasmic intermediate filament protein from an arthropod species. *BMC biology*, **9**(1), 17.

Murshudov, G. N., *et al.* (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, **67**(4), 355–367.

Offer, G., *et al.* (2002). Generalized crick equations for modeling noncanonical coiled coils. *Journal of structural biology*, **137**(1-2), 41–53.

Omary, M. B. (2009). "IF-pathies": a broad spectrum of intermediate filament–associated diseases. *The Journal of clinical investigation*, **119**(7), 1756–1762.

Parry, D. A. and Steinert, P. M. (1999). Intermediate filaments: molecular architecture, assembly, dynamics and polymorphism. *Quarterly reviews of biophysics*, **32**(02), 99–187.

Pedregosa, F., *et al.* (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pieper, U., *et al.* (2014). Modbase, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, **42**(D1), D336–D346.

Rämisch, S., *et al.* (2015). Exploring alternate states and oligomerization preferences of coiled-coils by de novo structure modeling. *Proteins: Structure, Function, and Bioinformatics*, **83**(2), 235–247.

Smith, T. A., *et al.* (2002). Sequence comparisons of intermediate filament chains: evidence of a unique functional/structural role for coiled-coil segment 1a and linker l1. *Journal of structural biology*, **137**(1-2), 128–145.

Strelkov, S. V. and Burkhard, P. (2002). Analysis of α-helical coiled coils with the program TWISTER reveals a structural mechanism for stutter compensation. *Journal of structural biology*, **137**(1), 54–64.

Strelkov, S. V., *et al.* (2001). Divide-and-conquer crystallographic approach towards an atomic structure of intermediate filaments. *Journal of molecular biology*, **306**(4), 773–781.

Strelkov, S. V., *et al.* (2004). Crystal structure of the human lamin a coil 2b dimer: implications for the head-to-tail association of nuclear lamins. *Journal of molecular biology*, **343**(4), 1067–1080.

Szeverenyi, I., *et al.* (2008). The human intermediate filament database: comprehensive information on a gene family involved in many human diseases. *Human mutation*, **29**(3), 351–360.

Testa, O. D., *et al.* (2009). CC+: a relational database of coiled-coil structures. *Nucleic acids research*, **37**(suppl 1), D315–D322.

Trigg, J., *et al.* (2011). Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS One*, **6**(8), e23519.

Turgay, Y., *et al.* (2017). The molecular architecture of lamins in somatic cells. *Nature*, **543**(7644), 261–264.

Vincent, T. L., *et al.* (2013). LOGICOIL – multi-state prediction of coiled-coil oligomeric state. *Bioinformatics*, **29**(1), 69–76.

Wood, C. W., *et al.* (2014). CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics*, page btu502.

Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, **57**(4), 702–710.