Article:

Jenkins, Huw Thomas orcid.org/0000-0002-3302-6966 (2018) Fragon: rapid high-resolution structure determination from ideal protein fragments. Acta crystallographica. Section D, Structural biology. pp. 205-214. ISSN 2059-7983

https://doi.org/10.1107/S2059798318002292

Fragon: Rapid high-resolution structure determination from ideal protein fragments

Huw T. Jenkins[a]*

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington,
York, YO10 5DD, UK

Correspondence email: huw.jenkins@york.ac.uk

Synopsis A new pipeline to solve structures by molecular replacement with ideal protein fragments is described and benchmarked against two test sets of mixed α/β and all-β folds at relatively high resolution.

Abstract Correctly positioning ideal protein fragments by molecular replacement presents an attractive method for obtaining preliminary phases when no template structure for molecular replacement is available. This has been exploited in several existing pipelines. This paper presents a new pipeline, named *Fragon*, in which fragments (ideal α-helices or β-strands) are placed using *Phaser*, then the phases calculated from these coordinates are improved by the density modification methods provided by *ACORN*. The reliable scoring algorithm provided by *ACORN* identifies success. In these cases the resulting phases are usually of sufficient quality to enable automated model building of the entire structure. *Fragon* was evaluated against two test sets comprising mixed α/β folds and all-β folds at resolutions between 1.0 and 1.7 Å. Success rates of 61% for the mixed α/β test set and 30% for the all-β test set were achieved. In almost 70% of successful runs fragment placement and density modification took fewer than 30 minutes on relatively modest 4-core desktop computers. In all successful runs the best set of phases enabled automated model building with *ARP/wARP* to complete the structure.

Keywords: Molecular replacement; Density Modification; Fragments

1. Introduction

The molecular replacement (MR) approach (Rossmann & Blow, 1962), combining experimentally derived structure factor amplitudes with phases generated from a partially correct model, is the most common method used to solve macromolecular crystal structures accounting for almost 80% of the X-ray structures deposited in the protein data bank (PDB) in the last five years. When the differences between the search model and the new structure are small, it is usually straightforward to place the model and refinement quickly improves the phases. In challenging cases where the only template structures available are from distantly related homologues more sophisticated tools are required. Molecular modelling (DiMaio *et al.*, 2011; Qian *et al.*, 2007; Wang *et al.*, 2016) may be able to modify the template to generate a better model or, in some cases, *de novo* predictions may be sufficiently accurate. The application of maximum likelihood approaches in the molecular replacement search algorithms provided by *Phaser* improves the chance of positioning even very partial models correctly (McCoy *et al.*, 2007; Read, 2001). Then once a potential solution is found, iterative cycles of phase improvement via density modification and automatic chain tracing (Cowtan, 2006; Sammito *et al.*, 2014; Terwilliger, 2003; Terwilliger *et al.*, 2008; Thorn & Sheldrick, 2013) have increased the chance of success. However, it is still often impossible to solve the structure with these techniques: either no reasonable template can be identified, or it cannot be correctly placed, or the initial phases calculated from even a correctly placed template are too poor to guide model improvement.

As most proteins contain secondary structure elements (α-helices and β-strands), standardised fragments with ideal secondary structure geometry can provide alternative MR search models. These fragments are likely to be highly similar to some regions of the unknown structure but only represent a tiny fraction of the structure. *Phaser* (McCoy *et al.*, 2007) can position such fragments correctly, but the signal to noise ratio is very low and it is difficult to detect correct matches from false. The challenge of this approach is then to reliably identify which of the many phase sets calculated from potential solutions are good enough to trigger successful current phase improvement procedures and to generate an interpretable map.

Several pipelines have been developed to build upon the power of *Phaser* to position small fragments such as *ARCIMBOLDO* (Rodríguez *et al.*, 2009), *AMPLE* (Bibby *et al.*, 2012) and *FRAP* (Shrestha & Zhang, 2015). *ARCIMBOLDO* and *AMPLE* in particular provide a variety of ways for selecting fragments, and for assessing MR success (Bibby *et al.*, 2013; Sammito *et al.*, 2013; Sammito *et al.*, 2015; Keegan *et al.*, 2015; Thomas *et al.*, 2015). Both use *SHELXE* (Sheldrick, 2002; Sheldrick, 2010) to probe whether any of the placed fragments provide sufficient phase information to lead to a complete model through iteration of density modification and chain tracing to build more atoms into the fragment. The use of chain tracing enables atoms additional to those in the initial fragments to be used for calculation of phases in subsequent cycles and thus if these are correctly placed chain tracing contributes to the phase improvement. In addition, as *SHELXE* reports the correlation coefficient (CC) between experimentally derived normalised structure factor amplitudes (E_{obs}) and those calculated from the trace (E_{calc}), this allows progress to be monitored. When the resolution of the diffraction data extends to better than 2.5 Å resolution, a value above 25% appears to be a reliable indication of success (Thorn & Sheldrick, 2013).

The *Fragon* pipeline uses the “dynamic density modification” (DDM) method coded in the program *ACORN* (Foadi *et al.*, 2000; Yao *et al.*, 2005) to test if phases calculated from a starting fragment can be improved to generate an interpretable map. This approach has been shown to be successful with atomic resolution data: the structure of triclinic lysozyme at 1.0 Å resolution was solved from initial phases calculated from a single 10-residue ideal poly-alanine α-helix (Foadi *et al.*, 2000) and other novel structures at atomic resolution were determined using this approach (Chavali *et al.*, 2005; Dubrava *et al.*, 2008).

The premise is that if the initial phases are not completely wrong (i.e. the fragments are correctly placed) maps calculated with these phases will show correct new atomic positions, albeit at a low level. Previous work indicated this was possible providing the measured data extend to sufficient resolution to give atomic peaks (Foadi *et al.*, 2000). In the DDM procedure firstly $\sigma(\rho)$, the standard deviation of the map density, is calculated then the density modification proceeds as follows: (i) negative density is replaced by zero. (ii) Positive density is replaced by $\rho\tanh\{0.2[\rho/\sigma(\rho)]^{3/2}\}$. (iii) The modified density is truncated to $kn\sigma(\rho)$ where k is by default 3 and n is the lower of the cycle number and 5 (Foadi *et al.*, 2000). The minimum and maximum values for truncation are $0.1\rho_{\max}$ and $0.8\rho_{\max}$, respectively where $\rho_{\max}$ is the maximum of the map. The truncation reduces the bias from the fragment but is progressively decreased over the first few cycles of DDM so as not to remove density that appears outside the starting fragment (Yao *et al.*, 2006). The biggest difference between DDM and density modification methods used in other programs such as the ‘sphere of influence’ method employed in *SHELXE* (Sheldrick, 2002) is that in DDM the map is modified solely according to the ratio $\rho/\sigma(\rho)$ i.e. the modification applied to the density at any grid point in each cycle is not determined by whether the grid point is assigned to the protein or solvent region (Yao, 2002) and therefore, the performance of density modification cannot be influenced by the value entered for the solvent content.

ACORN modifies an atomic resolution E_{obs} map calculated using only the larger E values (in this work $E > 0.8$). It has been shown that also using “reflections” beyond the measured resolution limit approximated as $E = 1$ enhances the map atomicity (Yao *et al.*, 2005). Previous work demonstrated that with measured data extending to 1.5 Å, and starting phases calculated from fragments representing as little as 3.2% of the final model, these could be sufficiently improved to enable automated model building of the entire structure (Yao *et al.*, 2005). The current work extends this to starting with phases calculated from placed fragments with ideal secondary structure geometry at similar resolutions. The basis of *Fragon* is that given data extending to sufficiently high resolution the density modification procedure within *ACORN* is powerful enough to reliably screen out incorrectly placed fragments (Fig. 1) and generate phases suitable for automated model building without the requirement for extensive cycles of chain-tracing. Removing the requirement for chain-tracing allows many more potential solutions to be tested with modest computing requirements. This makes it feasible to attempt challenging cases where only one of 100s of potential solutions is correct on a desktop machine. Consideration of the ratio of the height of a peak at an atomic centre relative to the standard deviation of the electron density map at 1.7 Å resolution with a phase error of 75° suggests that DDM is unlikely to be effective below this resolution (Yao *et al*., 2005; Yao *et al*., 2006).

2. Methods

2.1. Implementation

Fragon (Fig. 2) is essentially a Python wrapper with all the underlying crystallographic calculations performed by existing software. *Fragon* calls *Phaser* through its Python interface and *ACORN* via shell scripts. Manipulation of reflection data and coordinate files is handled by functionality provided by the *Clipper Python* (Cowtan, 2003; McNicholas *et. al.,* 2018) and *cctbx* (Grosse-Kunstelve *et al.*, 2002) libraries.

Fragment location simply runs *Phaser* in its MR_AUTO mode with many options set to their default values. One important exception is that all solutions with clashing fragments are rejected (to speed up the run time). The default criteria for purging the list of solutions after the rotation and translation functions are used but in searches for only one copy of a fragment the purge of solutions after rigid-body refinement is removed to output more potential solutions to test. In multi-copy searches this final purge governs the number of rotation functions performed in the subsequent fragment search and so if the default cut-off (removal of solutions with log-likelihood gain (LLG) lower than 75% of the difference between the mean and top LLG) retains more than 100 partial solutions only the top 100 are kept.

2.2. Selecting test sets

Previous work showed that only a tiny fraction of the structure was required to generate starting phases that could be improved by *ACORN* (Yao *et al*., 2005) and this suggested that *Fragon* might be able to solve structures when only one or two small fragments were placed. This would be particularly powerful for determining structures where only a small part of the fold could be represented by an ideal α-helix. Previous approaches are able to produce success rates of over 80% for all-α test cases (Bibby *et al.*, 2012; Keegan *et al*., 2015; Sammito *et al.* 2015), therefore, in this work no all-α test cases were used and instead two test sets were generated: the first containing mixed α/β folds with limited α-helical content and the second containing all-β folds.

2.2.1. Mixed α/β test set

A test set of mixed α/β folds was selected from the PDB. The criteria applied were: (i) Mixed α/β folds with the ratio of α to β content of less than or equal to 1. (ii) Data resolution between 1.0 and 1.7

Å. (iii) A single chain of 80-200 residues in the asymmetric unit. This meant that for the largest structures a model fragment of 7-10 residues would represent about 3-5% of the asymmetric unit. (iii) As the presence of heavy atoms facilitates density modification, structures containing elements heavier than chlorine were removed. (iv) The test set was further filtered with *PISCES* (Wang & Dunbrack, 2003) to remove structures with detectable sequence identity. The final test set contained 103 structures (Supporting information).

2.2.2. All-β test set

MR with fragments of β-sheets (either extracted from structures in the PDB (Sammito *et al.,* 2013) or generated by truncation of *ab initio* models (Bibby *et al.*, 2012; Keegan *et al.*, 2015)) is much more challenging than using ideal α-helical fragments as the varied geometry of β-strands in β-sheets usually requires large libraries of β-sheets to be sampled in order to identify one with a similar r. m. s. deviation to a region of β-sheet in the target as an ideal α-helix has to many sections of α-helix in proteins. To test structure determination using idealised β-strands, a test set of all-β structures was selected from the PDB. The criteria applied were: (i) No residues assigned as α-helical by DSSP (Kabsch & Sander, 1983). (ii) Data resolution between 1.0 and 1.7 Å. (iii) A single chain of 80-200 residues in the asymmetric unit. This test set was further filtered as described in §2.2.1 to leave 74 structures (Supporting information).

2.2.3. Deposited data

For both test sets the high resolution limit reported in the PDB deposition was used to select and filter structures, however, for test cases where the deposited data extend to higher resolution (12 in total) all deposited data were used. For PDB entry 2pnd in the all-β test set the deposited data extend to 0.97Å (but are only 22.8% complete in the range 1.0-0.97 Å). No analysis was attempted to detect anisotropy in the data sets.

As *Phaser* now uses a log-likelihood-gain target based on intensities and their associated experimental error estimates (Read & McCoy, 2016) these were used when available in preference to structure factor amplitudes.

2.3. Testing approach

2.3.1. Success criteria

Successful runs were identified on the basis of the value of the CC between the smaller E_{obs} not used in the map calculation and their calculated values generated by back-transforming the modified map reported by *ACORN*, referred to as CC_s. For all test cases in this work CC_s was calculated for reflections with E values between 0.1 and 0.8. If the best CC_s was above 0.2 the run was deemed successful. In more marginal cases, where the best CC_s was between 0.09 and 0.2, if either the difference between the best and worse CC_s was greater than 75% of the best CC_s or the number of solutions was less than or equal to 10% of the maximum allowed (so as not to reject runs where there were multiple correct solutions and no incorrect solutions) the run was deemed successful. All successful runs were verified by using the improved phases from the *ACORN* run with the highest value of CC_s for automated model building with *ARP/wARP* (Langer *et al*., 2008). For all successful runs except for those from PDB entry 4gu2, *ARP/wARP* was able to build a model with R_{free} below 0.3. For 4gu2 the best models after refinement with anisotropic ADPs had R_{free} in the range 0.33 – 0.36 but since the deposited structure (1.35 Å resolution) and the re-built and re-refined model from *PDB_REDO* (Joosten *et al*., 2014), have R_{free} of 0.269 and 0.273, respectively, this case was also deemed successful.

2.3.2. Mixed α/β folds with ideal α-helices

Bias towards the known structure was avoided by making no attempts to tailor the search fragment. For each test case 8 separate runs searching for one copy of an ideal (φ = -57.8°, ψ = -47°) poly-alanine α-helix of between 7 and 14 residues were performed. Up to 100 potential solutions for each helix length were tested by density modification with *ACORN*. For test cases where no run was deemed successful by the criteria defined in §2.3.1, a further 8 runs searching for two copies of an ideal poly-alanine α-helix of between 7 and 14 residues were performed.

2.3.3. All-β folds with ideal β-strands

Ideal poly-alanine β-strands comprising between 3 and 5 residues with identical φ/ψ angles of -120°/115°, -125°/120°, -130°/130°, -135°/135° and -140°/135° were generated with *Coot* (Emsley *et al.*, 2010). The β-strands were arranged in pairs with parallel and anti-parallel orientations and tilt angles between the β-strands of 0° - 30° in 5° increments. Ensembles (containing five models) of individual β-strands and pairs of β-strands were produced with *phaser.ensembler* and sorted so the closest-to-mean structure was the first model in the ensemble. For the 20 atomic ($d_{min} \leq 1.2$ Å) resolution structures in the test set searches for one copy of an ensemble of single β-strands with lengths of 3, 4 and 5 residues were performed. If these were unsuccessful, and also for the remaining 54 test cases, searches for one copy of each ensemble of pairs of 5 residue β-strands (14 ensembles in total) were performed. In these runs the pair of β-strands is refined by *Phaser* as a single rigid-body and the model output is the first structure in the ensemble (as no *a priori* information about which model in the ensemble has the lowest r. m. s. deviation to the target is available). In an attempt to increase the success rate, the *Phaser* option to refine individual chains as separate rigid bodies can be employed to optionally further refine either each of the two β-strands as individual rigid bodies or split the strands about the central Cα atom and refine each half of the β-strand as a rigid body.

2.4. Use of the deposited model to identify correctly placed fragments

In these non-blind test cases it is possible to determine whether each *Phaser* solution is correctly located by using the model from the PDB as a reference. In order to do this rapidly the following procedure was developed. Firstly the allowed origin shift that optimally superimposed (F_{calc} ,φ_{calc}) maps from each placed fragment and the deposited structure was calculated with *RESOLVE* (Terwilliger, 2000). This offset was applied to the solution and then this was superimposed on the deposited structure with *csymmatch*. A small *Clipper*-based utility was written to calculate the correlation coefficient between (F_{calc} ,φ_{calc}) maps calculated from both the fragment and the deposited structure in the region encompassed by the fragment. To maximise the discrimination between correctly and incorrectly located fragments only backbone atoms were included in the calculation, the ADPs of both the fragment and deposited structure were set to a constant value and only grid points containing density for the fragment were included in the calculation. To avoid confusion with more commonly quoted CCs between (F_{obs} ,φ_{calc}) maps, this measure is termed ‘placement score’. The use of (F_{calc} ,φ_{calc}) maps eliminates the need for atom-matching which would be required to determine coordinate r. m. s. deviation between the search fragment and the deposited structure.

2.5. Benchmarking

Benchmarks were performed on desktop computers with a single Intel Core i7-6700 (8 MB L3 Cache, 3.4-4.0 Ghz) or Core i7-4790 (8 MB L3 Cache, 3.6-4.0 Ghz) processor and 16 GB RAM running Scientific Linux release 7.3. Hyper-threading is enabled on these processors but a maximum of 4 simultaneous threads (i.e. one per physical core) were used.

3. Results

3.1. Overall performance

As there was no overlap between the fragments used as search models in the tests performed in this work - i.e. the idealised β-strands were only used for the all-β test set and the ideal α-helices used only for the mixed α/β test set the overall performance of *Fragon* against the two test sets is described separately. The overall success rate (i.e. at least one run deemed successful by the criteria in §2.3.1) for *Fragon* against the mixed α/β test set of 103 structures was 61%. The overall success rate against the all-β test set of 74 structures was 30% (Table 1, Supplementary Fig. 1 and 2, Supporting information). Success was achieved with ideal fragments accounting for under 3% of the total scattering for 22 runs in the mixed α/β test set and 13 runs in the all-β test set - some examples are shown in Fig. 3. The success rate is correlated with resolution: in the mixed α/β test set it reached 89% for test cases at resolutions between 1.0 and 1.2 Å, falling to 68% for those with resolution between 1.21 and 1.49 Å and further to 36% for those with resolution between 1.5 and 1.7 Å. While the success rate of *Fragon* with ideal β-strands against the all-β test set is low it is still encouraging as a small library of ideal strands (17 in total) was capable of solving 22 test cases. The options to further refine the β-strands placed as a pair as individual strands or to split each strand in half each resulted in success for 2 test cases. Success rate in the all-β test set was also better for structures at high resolution with 65% of the test cases with resolution between 1.0 and 1.2 Å solved. This includes six test cases solved by searching for one copy of an ensemble of single β-strands.

3.2. Fragment placement by *Phaser*

The high redundancy of the testing performed here (multiple correctly placed fragments over all eight runs for some test cases (Supporting information), combined with the (F_{calc} ,φ_{calc}) map correlation based placement scoring (§2.4) enabled a huge number (nearly 170,000) of *Phaser* solutions (henceforth the term 'solution' refers to a potential fragment placement) to be evaluated (Fig. 4). If a placement score of 0.3 (the edge of the cluster of low scoring solutions in Fig. 4 (*a*) and (*b*)) is taken as the lower bound for a correctly placed fragment, 2505 (2.85 %) of the solutions in the mixed α/β test set and 850 (1.04 %) of the solutions in the all-β test set are above this threshold. Such a small fraction of correctly placed fragments illustrates the scale of the challenge. Moreover, in the case of ideal α-helices there are often multiple correctly placed solutions from a single run, corresponding to one residue shifts of the short search fragment along a helix in the target structure. It is important to note that while low values (i.e. less than 0.3) of the placement score clearly indicate an incorrectly placed solution and placement scores in the range 0.8-1.0 indicate very accurately located fragments, intermediate scores are harder to interpret. For example, the placement score does not differentiate solutions where some atoms are very accurately placed but some extend into solvent from those where all atoms are somewhat inaccurately placed. The phases calculated from the former situation are likely to be more correct than the latter.

3.3. Analysis of eLLG for runs

Before a MR calculation is performed the expected value of the LLG for a correctly placed model can be estimated. The eLLG is the total expected LLG summed over all reflections. (McCoy *et al*., 2017). The eLLG values vs the best placement score (§2.4) for the 824 runs in the mixed α/β test set are shown in Fig. 5 (a). As expected for runs with low eLLG values in many cases none of the *Phaser* solutions are correctly placed (placement scores clustered around 0.2) but as the eLLG increases an increasing proportion of runs contain (at least) one correctly placed solution. The proportion of successful runs increases with eLLG which is as expected as the eLLG increases with the fraction of scattering accounted for by the search model and the number of reflections (McCoy *et al*., 2017). The 28 unsuccessful runs in the one-helix searches where searching for two copies of the helix was successful have eLLG values towards the lower end of the range. The same plot for the 1012 runs in the all-β test set (Fig. 5 (b)) is less informative. This is mainly because for each test case the 14 ensembles of pairs of β-strands have similar eLLG values but it is clear from Table 1 and the Supporting information that for a successful test case only one or a few of the ensembles resulted in success. The range of eLLG values is much smaller than for the helices in the mixed α/β test set as the eLLG increases quadratically with the fraction of scattering (McCoy *et al.*, 2017) and the largest fragment used in the all-β test set is 10 residues compared to 14 residues in the mixed α/β test set. Interestingly Fig. 5 (b) shows that in 4 runs success came from fragments with low placement scores. In 3 of the 4 cases the low placement scores reflect that one of the ends of one or both β-strands in the fragment is placed into solvent but in one case the strand lies across three strands in a β-sheet in the target structure. By chance some atoms are located at atomic positions in the true structure so the mean phase error for the strongest E values (E > 1.6, 1999 reflections) is 80.8° and with data extending to 1.2 Å resolution *ACORN* was able to improve the initial phases calculated from this rather inaccurately placed model.

3.4. Identification of correctly placed fragments by density modification with *ACORN*.

Fig. 4 illustrates that the CC between the smaller E_{obs} not used in the map calculation and E_{calc} from the density modified map (CC$_s$) can reliably identify correctly located fragments. As would be expected, the discriminatory power of CC$_s$ increases with higher resolution and, therefore, fewer of the E_{obs} used in map calculation approximated by $E = 1$. Fig. 4 (*a*) and (*b*) reveals that there are cases where density modification fails to improve the phases from accurately placed fragments. In cases where these fragments are from runs that were ultimately successful this has no impact on the overall success rate. However, there are fragments with high placement scores in runs that were unsuccessful suggesting that further tuning of the parameters for density modification or placement of additional fragments may lead to increased success rates. Comparison of Fig. 4 (*a*) and (*b*) suggests that the low success rate in the all-β test set result is due to searches with β-strands resulting in many fewer correctly placed fragments than searches with ideal α-helices in the mixed α/β test set.

3.5. Analysis of unsuccessful test cases

All runs were unsuccessful for 40 test cases from the mixed α/β test set. Of these in 20 cases the best placement score from all eight runs searching for one helix was in the range 0.15 - 0.36 indicating that no solution contained a correctly placed fragment. As up to 100 solutions were tested in the one helix runs and the number of partial solutions kept in the searches for two copies of an ideal helix was limited to 100, for these 20 test cases two helix searches would not be able to successfully place two copies of the search fragment. For the remaining 20 test cases the best placement score from all eight runs searching for one helix was in the range 0.62 - 0.96. Of these six of the test cases only contained one α-helix so ideal helices with length 7-14 residues could only represent one part of the structure. Accordingly, for these test cases in the runs searching for two helices in five cases no runs produced any solution where both helices had a placement score > 0.3. The exception was 1y9l where the run searching for two copies of an 8-residue ideal α-helix produced one solution where the helices were arranged with both fragments corresponding to parts of the 18 residue α-helix in this structure (placement scores of 0.88 and 0.59 for the first and second helix respectively). For nine of the remaining 14 test cases none of the runs searching for two helices produced any solution where both helices had a placement score > 0.3. Therefore, of the 40 unsuccessful test cases in only six were two copies of a helix correctly placed but density modification with *ACORN* was unable to improve the phases.

Of the 52 unsuccessful test cases in the all-β test set, the distribution of the best placement score from all runs for each test case is less informative. For 23 test cases it is < 0.30, for 20 it is in the range 0.31-0.59 and for the remaining 9 it is in the range 0.62-0.8. For the 29 test cases with best placement score > 0.30 the eLLG ranges from 6.3 to 23.7 for the corresponding runs indicating that *Phaser* is unlikely to find a solution. For 26 of these runs the LLG of the top solution ranges from 26.0 to 65.1 but for 3 runs the LLGs of the top solutions are 123.4, 134.1 and 281.4. However, for these three test cases (4ld1, 4rlc, 4gei), 308 of 564, 27 of 141 and 144 of 144 solutions have LLG > 120, indicating that in these cases high LLG does not definitively identify a correct solution. As Fig. 4 (b) and 5 (b) illustrate that fragments with placement scores in the range 0.3-0.8 can lead to success, for many of the unsuccessful test cases in the all-β test set failure cannot be due to *Phaser* failing to correctly place fragments but instead this must be because the fragments do not accurately enough match the corresponding region of β-sheet in the target structure for *ACORN* to be able to improve the phases.

3.6. Timing

The highly redundant testing, together with the placement scores for all solutions (Fig. 4 (*a*), (*b*)), allowed definition of criteria based on CC_s after density modification with *ACORN* that indicate a definitive solution has been found and no further solutions should be tested. For atomic resolution data ($d_{min} \leq 1.2$ Å) this is simply that CC_s is greater than 0.3. For data with resolution between 1.2 and 1.7 Å this is once the difference between the highest and lowest CC_s for the solutions tested exceeds 0.15. As many solutions can be tested in parallel and running *ACORN* processes are not terminated once the first definitive solution is identified, several definitive solutions may be produced before the run finishes. Applying these criteria to the 382 successful runs in the mixed α/β test set and 51 successful runs in the all-β test set allowed evaluation of the run times to be performed (Fig. 6). These benchmarks were performed on reasonably low specification desktop hardware (§2.6). As no attempt was made to modify parameters based on the results of previous testing the results are identical to those presented in Table 1.

3.6.1. Fragment location with *Phaser*

Fig. 6 illustrates how the time taken to place the fragment(s) dominates the run time in many runs and shows the greatest variability. This is not unexpected as the MR_AUTO mode of *Phaser* has been carefully optimised so that the signal in the rotation function determines how many potential solutions are tested in the translation function. When the signal is low this can result in a long run time in which

1000s of potential solutions are tested. However, if the translation function results in high scoring solutions the many low-scoring potential solutions are discarded. For 10 of the 15 successful runs in the mixed α/β test set with *Phaser* run times over 90 minutes this is the case and thus the translation function dominates the *Phaser* run time. All of these runs were from test cases in one of the 11 pairs of enantiomorphic space groups for which two translation functions are required to test both possibilities. Limiting the number of rotations tested in the translation function would speed up the run time but since in two of the 10 cases the highest scoring rotation was over 1000 places down the list this would have to be balanced against the risk of missing solutions. These 10 runs are from four test cases and for each of these there were runs searching with alternative length ideal α-helices in which the shortest *Phaser* run time was between 2.9% and 25% that of the longest. In the other five runs with *Phaser* run times over 90 minutes the translation function failed to produce high scoring solutions and in two of these runs the rigid-body refinement dominates the run time (for the other three the translation function still required more time than the rigid-body refinement). It should be noted that in unsuccessful runs thousands of low-scoring potential solutions are retained throughout the run and the consecutive rigid-body refinement of these solutions accounts for most of the run time.

3.6.2. Density modification with *ACORN*

The median time for density modification with *ACORN* to identify a definitive solution in the mixed α/β test set was 3.48 minutes. The outliers with times longer than 40 minutes predominately reflect runs where either the relatively conservative criteria for early termination were not triggered (three runs for test case 4xh7) or around 50% of the solutions were tested before the first correct placement was found. The only exception is for PDB entry 1sxv where density modification with *ACORN* was particularly slow requiring 50 minutes to test 9 solutions. In the all-β test set the median time for density modification with *ACORN* to identify a definitive solution was 9.2 minutes and the two outlier times in Fig. 6 correspond to runs where the first correct solution was number 67 and 92 respectively (shortest to longest time).

3.6.3. Overall run times

The shortest time required for density modification with *ACORN* in the mixed α/β test set was under 10 seconds and in 43 runs the time required was under one minute. When the time for fragment location with *Phaser* was also very short this results in extremely short overall run times with the fastest being under 40 seconds to solve the 1.15 Å resolution structure of Monellin (PDB entry 2o9u) with a single 7-residue α-helix. Of the 382 successful runs in the mixed α/β test set 55 took fewer than five minutes, 141 fewer than 10 minutes and 265 (69%) fewer than 30 minutes. The run times for the 51 successful runs in the all-β test set ranged from under two minutes to nearly 1.5 hours, however, all but five runs took under 1 hour and 34 runs (67%) finished in fewer than 30 minutes.

4. Discussion

It is clear that when sufficiently high resolution data are available, placement of one or two secondary structure elements such as an ideal α-helix or a β-strand followed by improvement of the phases calculated from the placed fragment by density modification can result in phases of sufficient quality to enable automatic model building to complete the structure. The challenge lies in testing a sufficient number of potential solutions to identify one that is correctly placed without relying on massive computational resources.

Fragon was implemented to address this challenge by enabling rapid testing of potential solutions. The run times for successful *Fragon* runs is primarily governed by how easily *Phaser* is able to place the fragment as this affects both the time for fragment location and how far down the solution list the first correct solution lies. For many of the test cases presented here *Fragon* requires fewer than 10 minutes on a relatively modest 4-core desktop computer to solve the structure. Moreover, on the same hardware nearly 70% of the successful runs are finished in under 30 minutes.

This speed does not come at the expense of performance - the overall success rates for the mixed α/β test cases with ideal α-helices was 61%. An alternative approach is used in *ARCIMBOLDO_LITE* (Sammito *et al.*, 2015) where improved ranking and filtering of potential solutions enables vastly fewer potential solutions to be tested than in the original *ARCIMBOLDO* (Rodríguez *et al.*, 2009) approach and thus, for easier target structures, the computational demands required for success to be vastly reduced. The 30% success rate for the all-β test set is encouraging as the same set of ensembles of β-strands (17 in total) was able to solve 22 structures suggesting that the use of large libraries of β-sheets extracted from the PDB, as employed in *ARCIMBOLDO_BORGES* (Sammito *et al.*, 2013) is not always required. Comparison is difficult because the fragments used in *Fragon* represent idealised secondary structure elements, i.e. the atomic positions are not derived from any template structure in the PDB and so the same fragments can be used for all test cases.

Powerful automated systems capable of generating vast numbers of results require well-designed interfaces to cater to the needs of users with varying levels of expertise. In order to guide the user in choosing sensible options and most importantly to clearly present the results from 100s of potential solutions, an interface to *Fragon* has been added to the *CCP4i2* graphical user interface (Potterton *et. al.*, 2018). Fig. 7 illustrates the use of this interface to solve the structure of the soluble domain of FlaF (PDB entry 4zbh, Banerjee *et al.*, 2015) from ideal β-strands. The structure of this 146 residue all-β fold at 1.5 Å resolution was originally solved by single isomorphous replacement with anomalous scattering from a platinum derivative. The user needs only provide the reflection data and expected composition of the asymmetric unit either explicitly by sequence or as an estimate of the solvent content. The interface allows easy selection of helix length or ensembles of β-strands and the detailed documentation helps inform suitable choice of search fragment. The results are presented in a table (Fig. 7) and phases from *ACORN* for the best solution suitable for subsequent automated model building pipelines and map coefficients for viewing in *Coot* (Emsley *et al.*, 2010) are provided.

5. Availability

Fragon will be submitted for distribution with the CCP4 suite and currently runs on Linux and macOS operating systems. A graphical interface within *CCP4i2* (Fig. 7) is available. The default parameters are those used for the benchmarks in this study.

Acknowledgements The work presented here was performed whilst the author was a postdoctoral associate in the group of Fred Antson. I thank Randy Read, Airlie McCoy, Eleanor Dodson, Jon Agirre and Johan Turkenburg for helpful discussions and Eleanor Dodson and Fred Antson for critical comments on the manuscript.

References

Banerjee, A., Tsai, C.L., Chaudhury, P., Tripp, P., Arvai, A. S., Ishida, J. P., Tainer, J. A., & Albers S.V. (2015). *Structure*, **23**, 863-872.

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

Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2013). *Acta Cryst.* D**69**, 2194–2201.

Chavali, G. B., Ekblad, C. M., Basu, B. P., Brissett, N. C., Veprintsev, D., Hughes-Davies, L., Kouzarides, T., Itzhaki, L. S. & Doherty, A. J. (2005). *J Mol Biol.* **350**, 964-973.

Cowtan, K. D. (2003). *IUCr Comput. Commun. Newsl.* **2**, 4–9.

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

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., Iwaï, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature (London),* **473**, 540–543.

Dubrava, M. S., Ingram, W. M., Roberts, S. A., Weichsel, A., Montfort, W. R. & Cordes, M. H. (2008) *Protein Sci.* **17**, 803-812.

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

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Yao, J.-X. & Chao-de, Z. (2000). *Acta Cryst*. D**56**, 1137–1147.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.

Joosten R. P., Long F., Murshudov G. N. & Perrakis A. (2014). *IUCrJ.* **1**, 213-220.

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

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

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

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

McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R., Tryggvason, K., Lohkamp, B. & Read, R. J. (2017). *Proc. Natl Acad. Sci. USA,* **114**, 3637-3641

McNicholas, S., Croll, T., Burnley, T., Palmer, C. M., Hoh, S. W., Jenkins, H. T., Dodson, E., Cowtan, K., & Agirre, J. (2018) *Protein Sci.* **27**, 207-216.

Potterton, L., Agirre, J., Ballard, C., Cowtan K., Dodson, E. J., Evans, P. R., Jenkins, H. T., Keegan, R. M., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S. J., Nicholls, R. A., Noble, M. E., Pannu, N. S., Roth, C., Sheldrick, G.M., Skubak, P., Turkenburg, J. P., Uski, V., von Delft, F., Waterman, D. G., Wilson, K. S., Winn, M & Wojdyr, M. (2018). *Acta Cryst.* D**78**, 68-84.

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

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Read, R. J. & McCoy, A. J. (2016) *Acta Cryst.* D**72**, 375–387.

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.

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

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.

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

Sammito, M., Millán, C., Frieske, D., Rodríguez-Freire, E., Borges, R. J. & Usón, I. (2015). *Acta Cryst.* D**71**, 1921–1930.

Shrestha, R. & Zhang, K. Y. J. (2015). *Acta Cryst.* **D**71, 304–312.

Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644-650.

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

Terwilliger, T. C. (2000). *Acta Cryst*. **D**56, 965–972.

Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 1174-1182.

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.

Thomas, J. M., Keegan, R. M., Bibby, J., Winn, M. D., Mayans, O., & Rigden, D. J. (2015) *IUCrJ*, **2**, 198-206.

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

Wang, G. & Dunbrack, R. L. Jr. (2003) *Bioinformatics,* **19**, 1589-1591.

Wang, Y., Virtanen, J., Xue, Z., Tesmer, J. J. G, & Zhang, Y. (2016). *Acta Cryst.* D**72**, 616-628.

Yao J.-X. (2002). *Acta Cryst.* **D**58, 1941-1947.

Yao J.-X., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2005). *Acta Cryst*. D**61**, 1465–1475.

Yao J.-X., Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006) *Acta Cryst.* **D**62, 901-908.

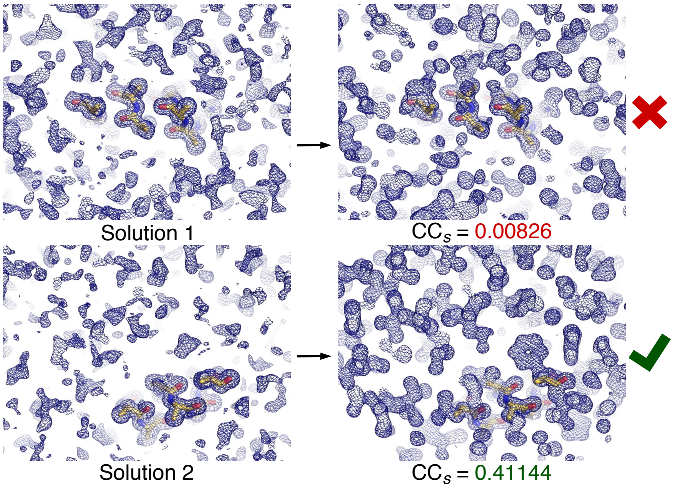


Figure 1 Phase improvement by density modification with *ACORN* illustrated for test case 1sxv (1.3 Å resolution) with phases calculated from a 10 residue ideal α-helix. Solutions are tested until the CC between the $E_{\rm obs}$ not used in the map calculation and their calculated values generated by back-transforming the modified map (CC_s) indicates phases from a solution have been sufficiently improved to enable automated model building.

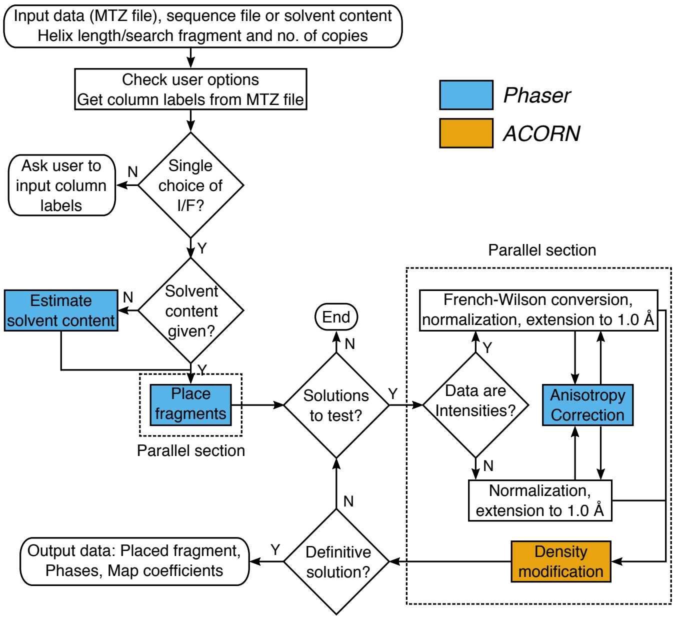
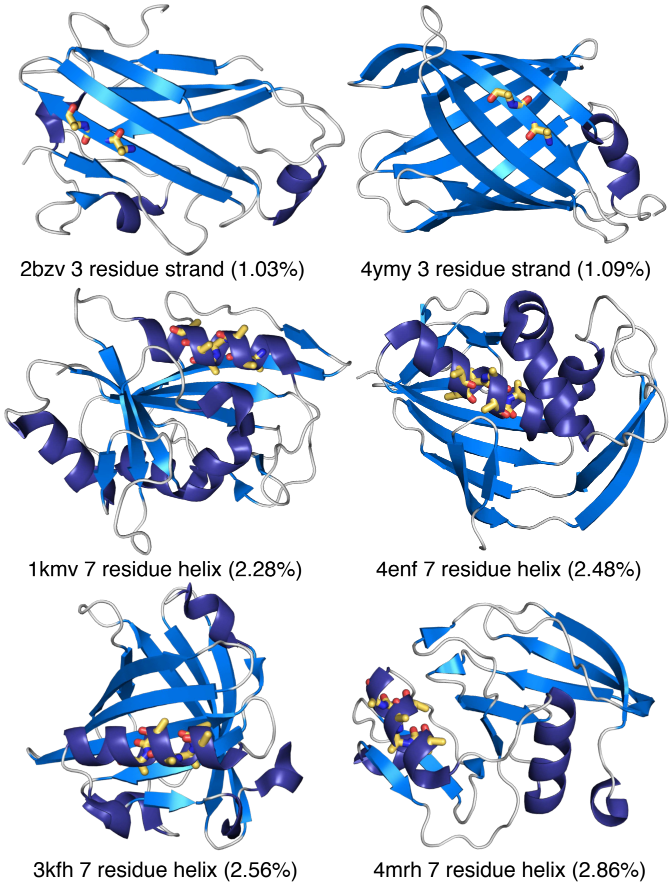

Input data (MTZ file), sequence file or solvent content
Helix length/search fragment and no. of copies
Check user options
Get column labels from MTZ file
Single choice of I/F?
N
Ask user to input column labels
Y
Solvent content given?
N
Estimate solvent content
Y
Place fragments
Parallel section
Solutions to test?
N
End
Y
Parallel section
French-Wilson conversion, normalization, extension to 1.0 Å
Data are Intensities?
Y
N
Anisotropy Correction
Normalization, extension to 1.0 Å
Density modification
Definitive solution?
N
Y
Output data: Placed fragment, Phases, Map coefficients
Phaser
ACORN

Figure 2 Flow diagram of the *Fragon* pipeline.



2bzv 3 residue strand (1.03%)

4ymy 3 residue strand (1.09%)

1kmv 7 residue helix (2.28%)

4enf 7 residue helix (2.48%)

3kfh 7 residue helix (2.56%)

4mrh 7 residue helix (2.86%)

Figure 3 Selection of test cases solved by *Fragon* from fragments of ideal secondary structure accounting for under 3% of the total scattering. The fragment placed by *Phaser* is shown as yellow sticks and the deposited structure shown as ribbons. In each case the PDB id of the test case, the size of the search fragment and the percentage of the scattering (as reported by *Phaser*) this represents is shown.

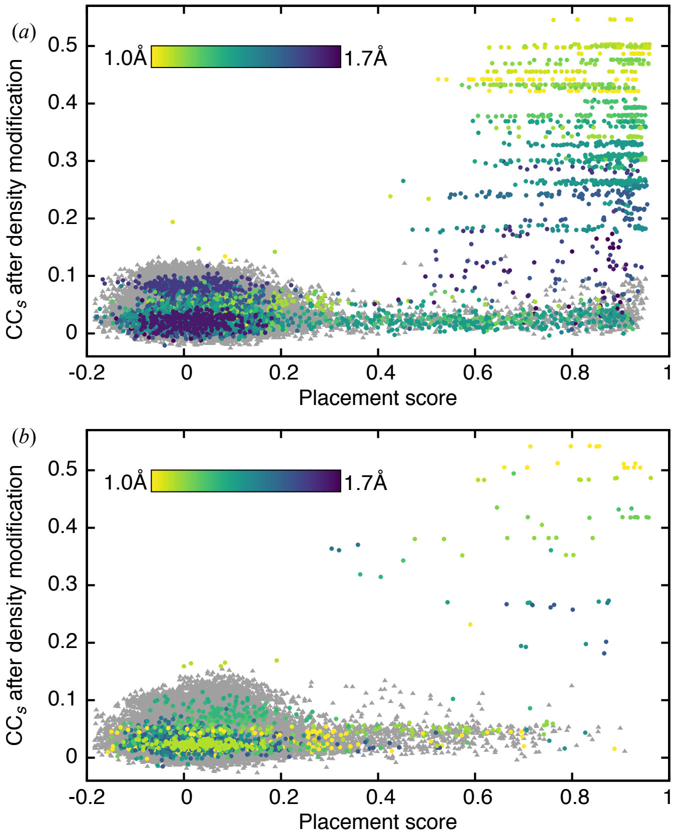


Figure 4 CC$_s$ after *ACORN* density modification vs 'Placement score' (§2.4). (*a*) Mixed α/β test set, (*b*) all-β test set. In both cases *Phaser* solutions from successful runs are indicated by circles coloured by resolution and *Phaser* solutions from unsuccessful runs are shown as grey triangles. For runs in which two ideal α-helices were placed the placement score for the second α-helix is shown as these runs were only carried out when all runs with a single α-helix were unsuccessful.

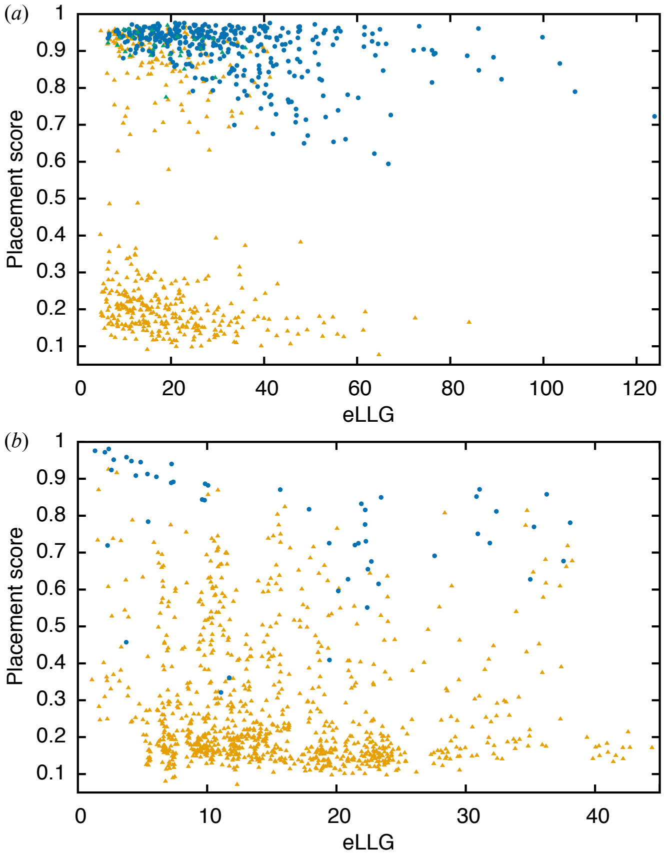


Figure 5 Placement score of the best placed solution against eLLG for searches performed with one copy of an ideal α-helix or one copy of an ensemble of ideal β-strands or pairs of ideal β-strands. (a) Runs searching for one copy of an ideal α-helix in the mixed α/β test set (824 runs). Successful runs are plotted as filled blue circles, unsuccessful runs as orange triangles and unsuccessful runs for which runs searching for two copies of the same α-helix were successful as green triangles. (b) Runs searching for one copy of an ensemble of ideal β-strands or an ensemble of pairs of ideal β-strands in the all-β test set (1012 runs). Successful runs are plotted as filled blue circles, unsuccessful runs as orange triangles.

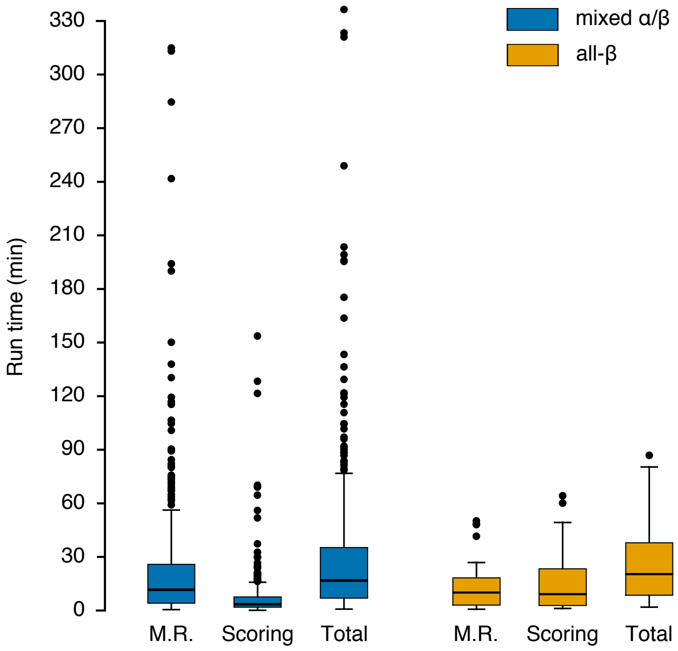
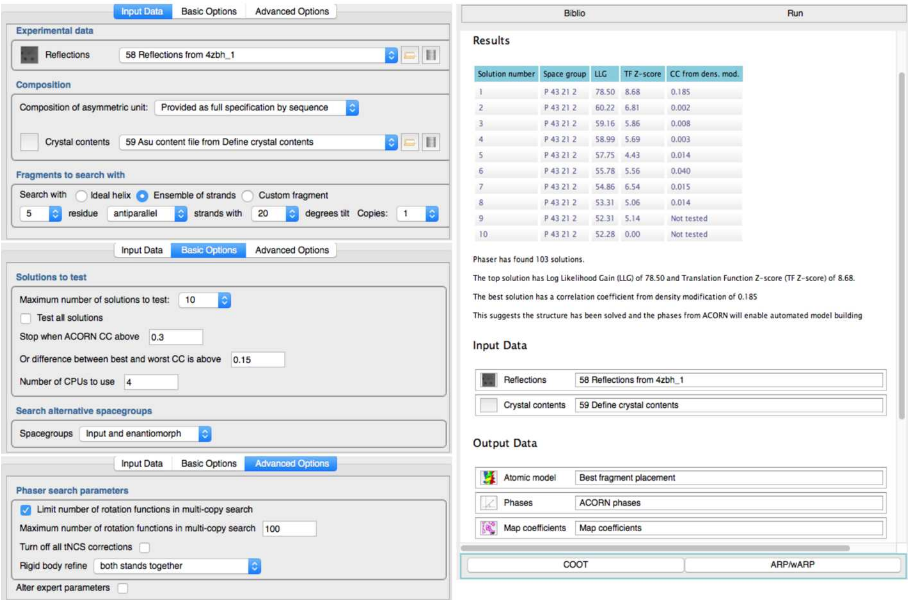


Figure 6 Box and whisker plots summarising run times for successful test cases. Whisker length is 1.5x inter-quartile range. 'M.R.' is time for fragment location with *Phaser,* 'Scoring' is time for density modification with *ACORN* which includes all time for reflection file manipulation.



Experimental data

Reflections: 58 Reflections from 4zbh_1

Composition

Composition of asymmetric unit: Provided as full specification by sequence

Crystal contents: 59 Asu content file from Define crystal contents

Fragments to search with

Search with ○ Ideal helix ● Ensemble of strands ○ Custom fragment

5 residue antiparallel strands with 20 degrees tilt Copies: 1

Solutions to test

Maximum number of solutions to test: 10

□ Test all solutions

Stop when ACORN CC above 0.3

Or difference between best and worst CC is above 0.15

Number of CPUs to use 4

Search alternative spacegroups

Spacegroups: Input and enantiomorph

Phaser search parameters

☑ Limit number of rotation functions in multi-copy search

Maximum number of rotation functions in multi-copy search 100

Turn off all tNCS corrections □

Rigid body refine: both stands together

Alter expert parameters □

Results

Solution number	Space group	LLG	TF Z-score	CC from dens. mod.
1	P 43 21 2	78.50	8.68	0.185
2	P 43 21 2	60.22	6.81	0.002
3	P 43 21 2	59.16	5.86	0.008
4	P 43 21 2	58.99	5.69	0.003
5	P 43 21 2	57.75	4.43	0.014
6	P 43 21 2	55.78	5.56	0.040
7	P 43 21 2	54.86	6.54	0.015
8	P 43 21 2	53.31	5.06	0.014
9	P 43 21 2	52.31	5.14	Not tested
10	P 43 21 2	52.28	0.00	Not tested

Phaser has found 103 solutions.

The top solution has Log Likelihood Gain (LLG) of 78.50 and Translation Function Z-score (TF Z-score) of 8.68.

The best solution has a correlation coefficient from density modification of 0.185

This suggests the structure has been solved and the phases from ACORN will enable automated model building

Input Data

Reflections	58 Reflections from 4zbh_1
Crystal contents	59 Define crystal contents

Output Data

Atomic model	Best fragment placement
Phases	ACORN phases
Map coefficients	Map coefficients

Figure 7 Solution of 4zbh using ideal β-strands with *Fragon* in *CCP4i2*. The input parameters are shown on the left and the report and output data on the right.

Table 1 Performance of *Fragon* against two test sets.

Test set	Resolution range (Å)	No. of structures	Search fragment	Copies	Runs	Solutions	Successful runs	No. solved
Mixed α/β	1.00 – 1.20	27	α-helix	1	216	10603	168	24
	1.21 – 1.49	37			296	20675	138	22
	1.50 – 1.70	39			312	23335	48	9
	1.00 – 1.20	3[†]		2	24	2269	0	0
	1.21 – 1.49	15[†]			120	9993	12	3
	1.50 – 1.70	30[†]			240	21123	16	5
Total		103			1208	87998	382	63
All-β	0.97[‡] – 1.20	20	β-strand	1	60	4781	15	6
	1.00 – 1.20	14[§]	2 β-strands	1	196	16213	17	7
	1.21 – 1.49	28			392	31946	14	7
	1.50 – 1.70	26			364	28760	5	2
Total		74			1012	81700	51	22

[†] Test cases solved searching for one copy of an ideal α-helix were not tested

[‡] 1 test case has a reported resolution of 1.0Å but deposited data extend to 0.97 Å

[§] Test cases solved searching with a single β-strand were not tested

Supporting information

Detailed results for the performance of *Fragon* against both test sets.

Figures 1 and 2