**MrParse**: finding homologues in the PDB and the EBI AlphaFold database for molecular replacement and more

Adam J. Simpkin, a Jens M. H. Thomas, a Ronan M. Keeganb and Daniel J. Rigdena*

*Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom, and bUKRI–STFC, Rutherford Appleton Laboratory, Research Complex at Harwell, Didcot OX11 0FA, United Kingdom.

*Correspondence e-mail: drigden@liverpool.ac.uk

Crystallographers have an array of search-model options for structure solution by molecular replacement (MR). The well established options of homologous experimental structures and regular secondary-structure elements or motifs are increasingly supplemented by computational modelling. Such modelling may be carried out locally or may use pre-calculated predictions retrieved from databases such as the EBI AlphaFold database. *MrParse* is a new pipeline to help to streamline the decision process in MR by consolidating bioinformatic predictions in one place. When reflection data are provided, *MrParse* can rank any experimental homologues found using eLLG, which indicates the likelihood that a given search model will work in MR. Inbuilt displays of predicted secondary structure, coiled-coil and transmembrane regions further inform the choice of MR protocol. *MrParse* can also identify and rank homologues in the EBI AlphaFold database, a function that will also interest other structural biologists and bioinformaticians.

1. Introduction

The dominant approach to solving the phase problem in crystallography is molecular replacement (MR). At the time of writing, 86% of crystal structures deposited in the Protein Data Bank (PDB; Burley et al., 2021) in 2021 were solved by this method. In MR, initial phase estimates are derived from the placement of a search model in the asymmetric unit, typically by successive rotation and translation steps (Scapin, 2013). Successful placement requires that the search model bear a sufficiently close structural resemblance to (part of) the target structure. Conventional MR typically deploys experimental PDB structures that are inferred to be homologous to the target structure (or one of its chains or domains). The inference of homology, from a significant result in a sequence-based database search with the target as a query, allows a reasonable supposition of structural similarity of the target and the PDB deposition, although this assumption can break down where a protein family can adopt distinct conformations. Furthermore, with distant homologues the degree of structural similarity between the target and the search model may be too low for successful placement, even with advanced maximum-likelihood-based methods (McCoy, 2004; McCoy et al., 2007; Read, 2001) and the use of methods to maximize their value (Rigden et al., 2018; Sammito et al., 2014).

Unconventional MR generally uses bioinformatic predictions to suggest or construct search models. Thus, a detailed consideration of the sequence properties of the target can help direct the structure-solution strategy (Pereira & Alva, 2021).
For example, a secondary-structure prediction can point to simple regular structural elements such as $\alpha$-helices (Rodríguez et al., 2012) or recurring tertiary packing features composed of several such elements (Sammito et al., 2013) as potential search models. Novel and divergent folds can also be explicitly predicted using $ab$ $initio$ modelling (also known as de novo, free or template-independent modelling). The first broadly successful algorithms in the field (Leaver-Fay et al., 2011; Xu & Zhang, 2012) used fragment-assembly approaches, limiting their application to relatively small targets. Limited accuracy also meant that their results often needed sampling across a range of ensembles and rational edits in order to succeed in MR (Rigden et al., 2008; Bibby et al., 2012). However, $ab$ $initio$ modelling methods have advanced with remarkable speed, first by exploiting the residue-contact information available from sequence alignments (see, for example, Marks et al., 2011) and then, dramatically, using bespoke deep neural networks (Senior et al., 2020; Jumper et al., 2021). CASP14 saw the stunning performance of AlphaFold2 (AF2), which in many cases produced predictions that resembled the target as closely as a different crystal form typically would (Pereira et al., 2021). The value of predictions from AF2 and the AF2-inspired RoseTTAFold (Baek et al., 2021) as search models was quickly demonstrated, although some cases still required domain splitting or other editing (Millán et al., 2021; Baek et al., 2021; McCoy et al., 2022; Pereira et al., 2021).

As $ab$ $initio$ modelling methods have advanced, so have the corresponding databases of structure predictions. Earlier efforts typically sampled uncharacterized fold space using Pfam domain definitions (Mistry et al., 2021) as a convenient foundation (Ovchinnikov et al., 2017; Lamb et al., 2019; Wang et al., 2019). Although Pfam domain boundaries inferred from sequence alignment alone are not always accurately defined, the entries in these databases could, especially with ensembling, succeed as search models (Simpkin et al., 2019). More recently, AF2 has been used to model complete sequences of 21 whole proteomes, including the human proteome (Tunyasuvunakool et al., 2021), and the results have been made available in the EBI AlphaFold database (AFDB; https://alphafold.ebi.ac.uk/). The often high accuracy of the predictions (and they are accompanied by high-quality residue-level error estimates) makes the database a very significant new source of search models for MR.

Here, we present MrParse, which addresses a number of issues in MR. It will find and rank search models from both the PDB and the AFDB, providing convenient visualization of the results. It also guides choices in unconventional MR through secondary-structure prediction and predictions of regions that are relevant to MR strategy such as coiled coils (Thomas et al., 2015, 2020; Caballero et al., 2018) and transmembrane helices. When MrParse is provided with diffraction data information it can flag the crystal pathologies that can hinder successful MR (Sevvana et al., 2019; Caballero et al., 2021) and rank experimental homologues from the PDB according to eLLG (Oeffner et al., 2018), which is a good predictor of their suitability as search models.

2. Methods

2.1. Reflection data classification

If a reflection file is provided, MrParse creates a table providing information from the reflection file (resolution and space group) and information about the crystal pathology calculated with CTRUNCATE (Evans, 2011) (noncrystallographic symmetry, twinning and anisotropy).

2.2. PDB search

MrParse uses phmmer (Eddy, 2011) to search either the full PDB or a 95% sequence identity redundancy-reduced version of it, as provided by MrBUMP (Keegan et al., 2018). Phmmer also provides information about the regions in the target protein that the hits correspond to. This is used to create a visualization of the search results using Pfam Domain Graphics (Mistry et al., 2021), which allows easy interpretation of how much of the target the search model covers. If a reflection file is provided, Phaser (Oeffner et al., 2018) is used to calculate the eLLG for each of the hits identified by phmmer. It has been shown that eLLG is a better indicator of whether a search model will succeed in MR than sequence identity (Oeffner et al., 2018). Therefore, when a reflection file is provided the search results are ranked by eLLG. Any hits are downloaded from the PDB and trimmed according to their match to the target sequence.

2.3. Protein classification

MrParse performs protein classification analysis on the input sequence to predict secondary structure, transmembrane regions and coiled-coil regions. Secondary structure is predicted using the JPred4 (Drozdetskiy et al., 2015) RESTful Application Programming Interface (API), transmembrane regions are predicted by TMHMM (Krogh et al., 2001) and coiled-coil regions are predicted by DeepCoil (Ludwiczak et al., 2019). Currently, coiled-coil and transmembrane predictions require local installations of TMHMM and DeepCoil.

2.4. EBI AlphaFold database search

MrParse uses phmmer to search the sequence database provided by the EBI AlphaFold database (https://alphafold.ebi.ac.uk/). As in the PDB search, information from phmmer is used to create a visualization of the search results using Pfam Domain Graphics. For the EBI AlphaFold database, these visualizations are coloured by Predicted Local Distance Difference Test (pLDDT) on an orange to blue scale, where orange indicates very low confidence in the model and blue indicates very high confidence in the model. Additional information is provided about the quality of the AF2 models, including the average pLDDT and a new measure of structural quality called the H-score.

The H-score can be calculated with the following equation, where $N$ represents a list of pLDDT scores and $\|N\|$ represents the number of elements in $N$.  

\[ H = \frac{\sum_{i=1}^{\|N\|} (1 - \frac{1}{\sqrt{N[i]}}))}{\|N\|} \]
degree of sequence (and hence structural) divergence. The target and search model, likely with a relationship defined by a combination of intrinsic modelling error and the error resulting from the use of a search model outside of the target sequence. These are downloaded into the homologous subdirectory in the EBI AlphaFold database' panel (Fig. 1, blue) provides information about AF2 predictions identified by phmmer in the EFD database. In this example, we can see a large number of AF2 hits. These hits are largely of very high quality, with an average pLDDT score of >80 for all of the hits. The visualization on the right-hand side shows the regions of the models that correspond to and provides information about predicted model reliability at a residue level. For example, the few models that match the C-terminal region of the target structure (P24942, P43003 and D7RVS0) all have lower predicted reliability in this region. Any matches are downloaded from the AFDB and trimmed to match the target sequence and undergo a pLDDT to estimated B-factor conversion to improve their performance in MR. These are downloaded into the models subdirectory in the MrParse run directory.

3.2. Use of an AFDB entry for MR when a PDB search model is lacking

PDB entry 7dry is a crystal structure of Aspergillus oryzae Rib2 deaminase experimentally determined by Zn-SAD (Chen et al., 2021). A phmmer search of the PDB only identified a single hit (PDB entry 2cvj) that only covers a 71-residue region of the target protein with 31% sequence identity (Figs. 2a and 2b). This homologue was insufficiently similar to the target protein to succeed in MR. A search of the EBI AlphaFold2 database identified a number of models that covered a larger region of the target protein and with a higher sequence identity. MR with the model of Q12362, the best hit ranked by H-score (Figs. 2a and 2c), was successfully placed by Phaser (LLG =173, TFZ = 15.4) and rebuilt with Buccaneer (Cowtan, 2006; R factor = 0.23, Rfree = 0.25).

4. Discussion

A crystallographer attempting to solve a macromolecular crystal structure by MR should be aware of the existence of any crystal pathologies and has an increasing range of search-model options to choose from. MrParse is designed to bring together a range of relevant information in a single place and present it with useful visualizations and sortable tables. For most effective use, it expects both diffraction data and a target sequence, but it can run without the former. Conventional MR using homologous structures identified in the PDB is supported by the presentation of potential search models.
Figure 1
Highlighted sections of an MrParse report page. In red is information on the input reflection file, including resolution, space group and crystal pathology. In teal is information about the PDB entries identified by phmmer and visualizations of the matches. In purple is the protein classification report; this includes a secondary-structure prediction, a coiled-coil prediction and a transmembrane prediction. Finally, in blue is information about the AlphaFold models identified by phmmer and visualizations of the matches coloured by pLDDT on an orange to blue scale, where orange indicates very low confidence in the model and blue indicates very high confidence in the model.
discovered by phmmer, with graphics that illustrate their extent relative to the target and numerical data that illustrate their size and characteristics. In the future, more sensitive HHpred (Söding, 2005) sequence searching will be supported. With diffraction data supplied, search models are ordered by default by eLLG as a predictor of their relative utility in MR. At present, PDB files are available locally and through the CCP4i2 GUI (Potterton et al., 2018) and online through the CCP4 Cloud setting (Krissinel et al., 2018). In the future, options for inline composition of ensembles will be implemented. The PDB files, which are trimmed according to their match to the target sequence and modified to convert the predicted residue error into a $B$ factor (Claudia Millán; https://gitlab.developers.cam.ac.uk/scm/haematology/readgroup/phaser_voyager/-/blob/master/src/Voyager/MDSLibraries/pdb_structure.py), can be fed directly to programs such as Phaser (McCoy et al., 2007) or MOLREP (Vagin & Teplyakov, 2010) or may, in more difficult cases, require special treatment.

Figure 2
(a) MrParse results page; components are as seen previously except for a coiled-coil prediction (labelled CC) under the Sequence Based Predictions heading. (b) The closest match in the PDB (PDB entry 2civ, blue) aligned with the crystal structure (PDB entry 7dry, grey). (c) The closest match in the EBI AlphaFold database (Q12362, coral) aligned with the crystal structure (PDB entry 7dry, grey).
Perhaps the most exciting and forward-facing aspect of MrParse is its discovery of structure predictions, especially those generated by ab initio (also known as de novo or template-independent) methods. The potential of these methods for MR of targets with novel or divergent folds has been recognized for some time (Rigden et al., 2008; Bibby et al., 2012; Qian et al., 2007). Nevertheless, their (until recently) considerable CPU demands and specialist software have undoubtedly proved offputting to structural biologists, despite the convenience offered by some servers (Keegan et al., 2015). In addition, the accuracy of ab initio methods has historically not always been sufficient for MR and only smaller proteins were tractable using the earliest methods. This picture has changed rapidly in recent years with first AlphaFold (Senior et al., 2020) and then AlphaFold2 (Jumper et al., 2021), each providing a step-change in model accuracy. These developments have been mirrored in online databases of ab initio structure predictions. Databases based on earlier methods such as GREMLIN (Ovchinnikov et al., 2017), PconsFam (Lamb et al., 2019) and C-QUARK (Wang et al., 2019) typically modelled single representatives of Pfam families. These provided useful sampling of uncharacterized protein fold space, sometimes being suitable for MR (Simpkin et al., 2019), but were limited by the fact that the domain boundaries of Pfam entries are not always, in the absence of some kind of structural information, accurately determined from sequence analysis (Bateman et al., 2010). The AFDB, in contrast, includes full-length models from 21 essentially complete proteomes, with the ambition to cover UniRef90 (Suzek et al., 2015), so that no protein of interest will be less than 90% identical to an entry in the database, by end of 2021. Models in the AFDB are likely to be much more accurate than models available elsewhere, and are accompanied by accurate residue-level error estimates. Their availability therefore has profound implications for the choice of crystallographic phasing method (Kryshtafovych et al., 2021; McCoy et al., 2022) and the already very high market share of MR will only increase further.

MrParse currently provides a second graphical panel devoted solely to matches in the AFDB. These can be ranked by clicking on column headings for two measures of model quality: the novel H-score described here or the percentage sequence identity between the protein of interest and the model. While the experience of the CASP structure-prediction experiment suggests that many models serve, unaltered, as successful search models, downstream editing of models after retrieval via MrParse will sometimes be necessary (McCoy et al., 2022; Millán et al., 2021; Pereira et al., 2021). This can eliminate regions with low predicted accuracy (McCoy et al., 2022) or sample a variety of truncated versions (Pereira et al., 2021), or excise domains from multi-domain models, recognizing that inter-domain packing remains a challenge for AF. Future work will undoubtedly address the automatic identification or ranking of AFDB-derived search models, for example recognizing that small but very accurate substructures may be suitable search models where high-resolution diffraction data are available (McCoy et al., 2017). Furthermore, a systematic exploration of the characteristics of AFDB entries and their ability to predict coordinate error with respect to a given target, as performed with PDB entries (Hatti et al., 2020), will also be highly valuable.

Presently, hits are found by a phmmer (Eddy, 2011) search against a local database containing the sequences of entries in the AFDB. With the ambitious plans to expand the AFDB, this arrangement becomes increasingly awkward as ever-larger databases would need to be distributed with CCP4. Happily, the 3D-Beacons initiative (Orengo et al., 2020) will shortly be launching an API for sequence-based retrieval of models not only from the AFDB but also from a variety of other resources containing protein structure predictions. Thus, we envisage that the importance of MrParse in facilitating access to a wide range of potential MR search models, both experimental structures and predictions, will only grow in the future. In addition, its ability to search AFDB in particular and conveniently visualize the results is likely to prove useful to bioinformaticians and cryo-EM researchers (Kryshtafovych et al., 2021; Simpkin et al., 2021) as well as to crystallographers.

Acknowledgements
The authors declare no conflicts of interest.

Funding information
This work was supported by the Biotechnology and Biological Sciences Research Council (BB/S007105/1) and by CCP4 grants to support AJS and JMHT.

References
Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathina swamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. (2021). Science, 373, 871–876.
Bateman, A., Coggill, P. & Finn, R. D. (2010). Acta Cryst. F66, 1148–1152.
Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). Acta Cryst. D68, 1622–1631.
Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., Dutta, S., Feng, Z., Ganesan, S., Goodsell, D. S., Ghosh, S., Green, R. K., Guranirovic, V., Guzenko, D., Hudson, B. P., Lawson, C. L., Liang, Y., Lowe, R., Namkoong, H., Pfeirash, E., Persiko, I., Randle, C., Rose, A., Rose, Y., Sali, A., Segura, J., Sekharan, M., Shao, C., Tao, Y.-P., Voigt, M., Westbrook, J. D., Young, J. Y., Zardecki, C. & Zhuravleva, M. (2021). Nucleic Acids Res. 49, D437–D451.
