

Applications of residue contact predictions in structural biology

Filomeno Sanchez Rodriguez^{1,2}, Ronan M. Keegan³, Melanie Vollmar², Gwyndaf Evans²,
Daniel J. Rigden¹

¹University of Liverpool, Liverpool, United Kingdom, ²Life Science, Diamond Light Source, Harwell Science and Innovation Campus, Didcot, Oxfordshire, United Kingdom, ³STFC, Rutherford Appleton Laboratory, Harwell Oxford, United Kingdom
drigden@liverpool.ac.uk

Recent developments in the field of evolutionary covariance and machine learning have enabled the precise prediction of residue-residue contacts and increasingly accurate inter-residue distance predictions. Access to accurate covariance information has played a pivotal role in the recent advances observed in the field of protein bioinformatics, particularly the improvement of prediction of protein folds by ab initio protein modelling. As this work seeks to showcase, this data is of equal value in the field of X-ray crystallography, with several practical applications in MR, model validation and map interpretation.

The most prevalent technique for the solution of the phase problem in macromolecular crystallography is molecular replacement (MR). In most cases, the availability and detection of a suitable search model, typically a solved structure homologous to the target of interest, is the key limitation of conventional MR. In those cases where no such structure is available, unconventional MR approaches are used. Recent results suggest that even in those cases where no homologous structures are found for a given target, it may still be possible to find suitable search models among unrelated structures, in the form of regions that share high, albeit local, structural similarity with the target. The challenge then becomes the accurate identification of such search models among the vast number of available solved structures. Here we present *SWAMP*, a novel pipeline for the solution of structures of transmembrane proteins, which exploits the latest advances in residue contact predictions for the detection of fragments later to be used as search models. *SWAMP* includes a library of ensembles built by clustering commonly observed packings of transmembrane helical pairs in close contact, mined from the available databases. Residue contact predictions are used in the process of search model selection: the contact maximum overlap between the target's predicted contacts and the observed contacts of each member of the library is used to estimate the likelihood of the helical pair being a successful search model. Preliminary results show that *SWAMP* is capable of detecting valid search models originating from unrelated solved structures solely exploiting this contact information. This enables the solution of new and challenging structures without the use of experimental phasing techniques, and opens a whole new avenue of research in which predicted contact information is used to extend the reach of unconventional MR.

The outcome of X-ray crystallographic experiments is the determination of the structure of interest, which requires building a model that satisfies the experimental observations. However, experimental limitations can lead to the presence of unavoidable uncertainties during model building resulting in regions that require validation and potentially further refinement. Many metrics are available for model validation but are mostly limited to the consideration of the physico-chemical aspects of the model or its match to the map. We present new metrics based on the availability of accurate inter-residue distance predictions, which are then compared with the distances observed in the emerging model. Early results suggest that these metrics are capable of detection of register and other errors, even in challenging cases where conventional metrics may struggle.

Residue contact and inter-residue distance predictions are usually represented respectively as two-dimensional binary matrices called contact maps and distograms. These typically omit contacts between sequential near neighbours resulting in a blank space on and near the diagonal axis of the matrix. A multitude of properties can be predicted by other sequence-based methods and researchers often need to consider diverse sources of information in order to form a complete and integrated picture for the inference of structural features that can facilitate the structure solution. Here we present ConPlot, a web-based application which uses the typically empty space near the contact map or distogram diagonal to display multiple-coloured tracks representing other sequence-based predictions. These predictions can be uploaded in various popular file formats. This web application is currently available online at www.conplot.org, along with documentation and examples.

Keywords: Molecular replacement; X-ray crystallography; Residue contact predictions; Model validation