Choosing the best enzyme complex structure made easy

Sayoni Das[1] and Christine Orengo[1]

[1] Institute of Structural and Molecular Biology, University College London, London, UK

In this issue of Structure, Tyzack et al (2018) present a study of enzyme-ligand complexes in the Protein Data Bank (PDB) and show that the molecular similarity of bound and cognate ligands can be used to choose the biologically most appropriate complex structure for analysis when multiple structures are available.

Main Text

Enzymes, comprising ~50% of all known proteins, are biological catalysts that catalyze most biochemical reactions within a cell. They carry out their function by binding to ligands *in vivo* (cognate ligand) in a specific manner, which often results in a conformational change in the enzyme. Therefore, characterization of the molecular mechanisms of ligand-binding and recognition is key to gain an understanding of molecular functions *in vivo*. Moreover, characterization of enzyme-ligand interactions has huge pharmaceutical and biotechnological implications. The availability of large numbers of protein structures in the PDB and the use of structure-based approaches during various stages of drug discovery and development has been a significant step towards this.

One of the main challenges of any protein structure-based study is the selection of the most appropriate target or template structure that affects to a great extent the outcome of drug discovery, structural modelling, molecular dynamics, protein design and engineering (Śledź and Caflisch 2018). The criteria generally used for choosing structures can vary with the biological question being asked but is often based on one or more of the following − resolution, completeness and/or conformation of the structure, presence/absence of amino acid mutations, bound ligands or cofactors, and absence of any additives or artifacts of crystallization. In some applications, such as structural modelling, there are advantages of using multiple structures. However, for most drug development applications such as pharmacophore identification, virtual screening and drug design, where characterization of enzyme-ligand interactions play a major role, the analyses benefit from using the biologically most relevant structure.

A large number of enzyme-ligand complexes contain molecules that are non-cognate (i.e. non-biological ligands), such as non-biological inhibitors or substrate analogues (Bashton, Nobeli, and Thornton 2006). Therefore, the availability of enzyme structures bound with the cognate ligand and/or, the identification of the closest representative, when multiple structures of the enzyme are available bound to many different ligands, presents new challenges to analyses that aim to study the characteristics of the biological ligand. There are a few studies or resources, such as Binding MOAD (Ahmed et al. 2015) and BioLip (Yang, Roy, and Zhang 2012), which can help in differentiating ligands and crystallographic additives/artifacts or metals, however, none of them help in distinguishing between cognate and non-cognate ligands.

Tyzack et al. (2018) present a study of all the PDB structures that can be mapped to entries in the Enzyme Commission (EC) numbers and Kyoto Encyclopedia of Genes and Genomes (KEGG) reactions with molecular structures for the cognate ligands (reactants and/or products). This comprises a dataset of 56,994 enzyme-ligand complex structures, 16.9% of

which do not have any bound ligands. For the remaining structures that have a bound ligand, they calculate the molecular similarity of the bound ligand with known cognate ligand(s) for each of them. The molecular similarity scores are calculated using a new method called PARITY (Proportion of Atoms Residing in Identical Topology) that identifies the proportion of atoms that are of the same type and lie in identical topological positions in both the bound (which could be cognate or non-cognate) and cognate ligands. For each structure, they also calculated a combined similarity score for all cognate reactants and products in a reaction. The similarity scores range from 0-1, where higher scores indicate higher bound-cognate ligand similarities. The fact that the authors find more than 31% of the structures have similarity scores of ≤0.3 and only 26% (14,839) of the structures have scores ≥0.7, it gives us an estimate of the challenges researchers face while choosing suitable structures. As part of their study, they rank related enzyme-ligand structures in the PDB on the basis of their similarity of the bound non-cognate ligands from that of the cognate ligand for each EC and KEGG reaction. This information will be valuable in assisting researchers in making an informed choice for structure selection (see example shown in Figure 1).

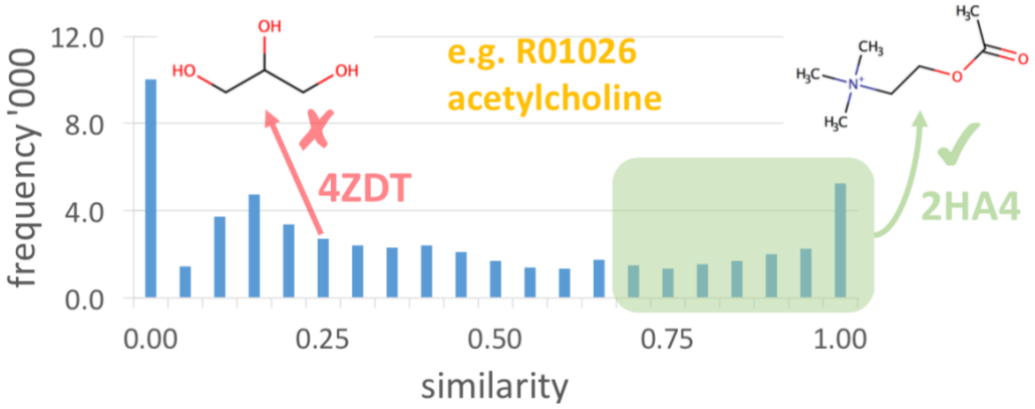

Figure 1: The graph shows the distribution of the binned similarity scores (measured using the PARITY method) of the bound ligands from PDB structures and cognate ligands from KEGG reactions for the most-similar PDB-KEGG match for each PDB structure. Only about 26% of the PDB structures (highlighted by green box) were found to have bound-cognate similarity scores of greater than or equal to 0.7. Adapted from Tyzack et al. (2018).

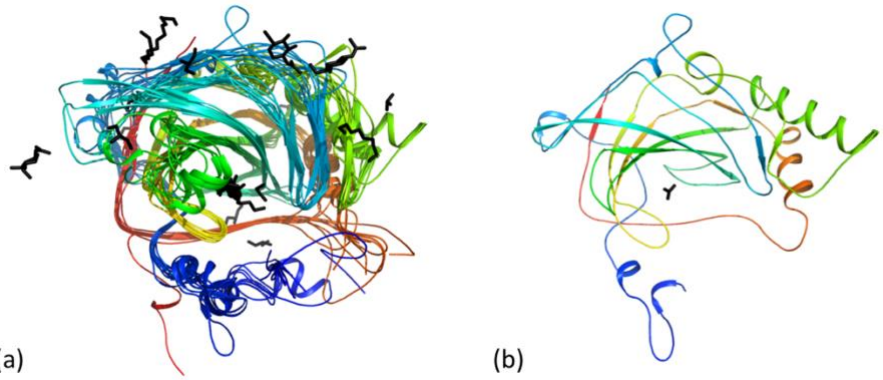

Figure 2: (a) Structural superposition of members of the carbonic anhydrase superfamily in CATH along with their bound ligands. All the members shown are carbonic anhydrases (EC 4.2.1.1). The protein structures are shown in ribbon representation using the rainbow colour scheme and the ligands are shown as black sticks. The structure superposition figures has been generated using cath-superpose (Dawson et al. 2017), https://github.com/UCLOrengoGroup/cath-tools). (b) The structure of carbonic anhydrase (PDB 1Y7W), selected using the ranked lists provided by Tyzack et. al. (2018), that has a ligand ACY (shown as black sticks) bound, that is similar (similarity score=0.75) to the cognate ligand of the enzyme.

In fact, the advantages of this study of enzyme-ligand complexes is manifold. It will not only assist researchers to choose the most relevant enzyme-ligand complex for their analyses in structure-based drug design applications but will also help the structural bioinformatics community in selecting suitable structural representatives for protein family-based studies on ligand diversity (Najmanovich 2017), function evolution (Das, Dawson, and Orengo 2015), structural modelling (Lam et al. 2017) among others. For example, the carbonic anhydrase superfamily in CATH (Dawson et al. 2017) contains many carbonic anhydrase enzyme structures with different ligands bound in different parts of the structure (Figure 2a). Using the ranking of PDB structures for each EC by Tyzack et al. (2018), it is possible to easily find the enzyme structure with the cognate ligand bound (exact match in this case, Figure 2b). Another research area in which this study will contribute towards massively is the structure-guided interpretation of how genetic variants may impact the structure and function of proteins.

In summary, the study presents a very valuable strategy for selecting structures of enzyme-cognate ligand complexes for structural biologists, structural bioinformaticians, and biomedical researchers, alike, with helpful data on PDB structures, provided for these users.

References

Ahmed, Aqeel, Richard D. Smith, Jordan J. Clark, James B. Dunbar Jr, and Heather A. Carlson. 2015. "Recent Improvements to Binding MOAD: A Resource for Protein-Ligand Binding Affinities and Structures." *Nucleic Acids Research* 43 (Database issue): D465–69.

Bashton, Matthew, Irene Nobeli, and Janet M. Thornton. 2006. "Cognate Ligand Domain Mapping for Enzymes." *Journal of Molecular Biology* 364 (4): 836–52.

Das, Sayoni, Natalie L. Dawson, and Christine A. Orengo. 2015. "Diversity in Protein Domain Superfamilies." *Current Opinion in Genetics & Development* 35 (December): 40–49.

Dawson, Natalie L., Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. 2017. "CATH: An Expanded Resource to Predict Protein Function through Structure and Sequence." *Nucleic Acids Research* 45 (D1): D289–95.

Lam, Su Datt, Sayoni Das, Ian Sillitoe, and Christine Orengo. 2017. "An Overview of Comparative Modelling and Resources Dedicated to Large-Scale Modelling of Genome Sequences." *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 8): 628–40.

Najmanovich, Rafael J. 2017. "Evolutionary Studies of Ligand Binding Sites in Proteins." *Current Opinion in Structural Biology* 45 (August): 85–90.

Śledź, Paweł, and Amedeo Caflisch. 2018. "Protein Structure-Based Drug Design: From Docking to Molecular Dynamics." *Current Opinion in Structural Biology* 48 (February): 93–102.

Tyzack, Jonathan D., Laurent Fernando, Antonio J. M. Ribeiro, Neera Borkakoti, Janet M Thornton. 2018. "Ranking enzyme structures in the PDB by bound ligand similarity to biological substrates." *Structure.*

Yang, Jianyi, Ambrish Roy, and Yang Zhang. 2012. "BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand–protein Interactions." *Nucleic Acids Research* 41 (D1): D1096–1103.