3 November 2023

Australasian Structural Phylogenetics Meeting 2023: Brainstorm Session

On October 24-25, 2023, a group of researchers convened in B09-303 and B11-303, Science Building, University of Auckland to present and discuss novel research in the emerging field of structural phylogenetics. On the afternoon of October 25, the participants below formed groups of three to four people. No group contained more than one person from the same research group. Each group was given 45 minutes to discuss the following statement, and then share their ideas with the full room:

“Looking back on this meeting, what are the major challenges faced in structural phylogenetics? What methodological advances would you like to see in the future?”

Session chairs: Caroline Puente-Lelievre and Jordan Douglas
Group 1: Ashar Malik, Chandra Marie Rodriguez, Matt Baker, Peter Wills
Group 2: Desiree Langer, Remco Bouckaert, Tristan de Rond
Group 3: Daniel Body, Jamiema Sara Philip, Nick Matzke
Group 4: Jane Allison, Martin Steinegger, Matthew Fulmer, Pietro Ridone

Summary

For context, the discussion was largely focused on ‘3Di’ characters, which enable one dimensional representations of three-dimensional protein structures. These conversations followed from the keynote presentation by Martin Steinegger “Structure analysis in the era of next-generation structure prediction”. In this talk, Martin discussed his new method FoldSeek, which uses 3Di characters to identify structural homologs. For more information, please see

van Kempen, Michel, et al. "Fast and accurate protein structure search with Foldseek." Nature Biotechnology (2023): 1-4.

Use of 3Di characters in phylogenetics

- [ ] Sequences of 3Di characters could be useful for making phylogenetic inference, using standard methods. As structure is more conserved than sequence, this approach has the potential to infer much deeper phylogenies than using amino acid sequence alone. However, it is unclear whether this is a valid approach for phylogenetic inference, especially given that phylogenetic methods make the mathematical assumption of independence between positions, which is most certainly violated in the case of 3Di characters.
- [ ] To validate this approach, and other potential methods for inferring protein phylogenies, we would need:
    - [ ] An accurate forward-time simulator of protein structure evolution, conditional on a phylogenetic tree. By inferring phylogenies from large volumes of simulated data, one could robustly validate such methods and establish their statistical properties and biases.
    - [ ] An empirical dataset of protein structure evolution where we know the ancestral sequences or structures. Ideally, the sequences should have evolved so much that they are <10% similar (i.e., past the twilight zone). This dataset could be used to validate methods for inferring deep phylogenies. It is unclear whether such a dataset already exists.

Protein and RNA structural models

- [ ] Although AlphaFold2 has made accurate protein structure prediction feasible, accurate RNA structure prediction remains challenging.
- [ ] Protein conformational rearrangements are hard to capture and they present difficulties when inferring structural similarity and classifying or aligning protein structures.
- [ ] It is not always clear what regions of a protein structure to include, or what protein structures to sample, when performing phylogenetic inference or searching for homologs.
- [ ] Disordered proteins are difficult to work with both experimentally and computationally. An ideal computational approach to studying disorder would sample an ensemble of disordered structures with probabilities assigned to each state. It is unclear how 3Di characters would work on disordered proteins.
- [ ] Difficulties modelling interactions between monomers and understanding how 3Di reacts to interfaces between subunits

Innovative ways to visualise structural and evolutionary data

- [ ] There is a need for a visualiser that combines phylogenetic and structural information, for example a hybrid between PyMOL and FigTree.

- [ ] The use of a standardised set of taxonomic ‘emojis’ to help communicate phylogenies. For example, there should be a canonical gorilla icon to denote its species or genus, which can be displayed at the leaves of a tree to denote gorilla genes/proteins. Similar to PhyloPic.
- [ ] Visualising large phylogenetic trees is challenging. Circular trees are often employed in such a case, however the room was divided on whether they are easy to interpret or if they just add confusion.
- [ ] Ancestral reconstructions should become the norm with any phylogenetic inference. Users could scroll through the time axis to see the ancestral sequences and structures fade in and out of view.
- [ ] Visualising 3Di alignments is challenging because the default amino acid colouring schemes (e.g. ClustalX) are inappropriate for 3Di characters. Following this feedback, Martin Steinegger’s colleagues in South Korea promptly released a tentative colour scheme for use in programs like JalView.

Integrating sequence and structural data with other types of data

- [ ] Biochemical data could be integrated into phylogenetic analyses, for example by considering catalytic functions from a database like BRENDA.
- [ ] Ligand binding could be incorporated into such analyses, for example by describing the molecular interactions between a protein and its ligand(s) using a 3Di-like notation.