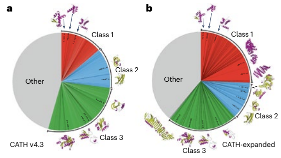
Research highlights

Computational biology

Predicted protein structures expand the CATH database



Structural distribution of CATH v4.3 (a) and the expanded CATH database with AlphaFold2 predicted structures (b). Adapted with permission from Bordin, N. et al. *Commun. Biol*. 6, 160, (2023), CCBY 4.0.

The number of high-quality predicted protein structures has exploded following the recent advances in deep learning-based protein structure prediction methods. AlphaFold2 has already released predicted structures for more than 200 million proteins. These developments have set off reevaluation or revamping of the traditional protein structure databases, including the CATH Protein Structure Classification database. CATH provides four levels of manually curated protein structures classification: the protein class (C), architecture (A), topology (T) and homologous superfamily (H).

Christine Orengo from University College London, one of the original developers, maintains the CATH database. Orengo and colleagues now describe CATH-Assign, a set of automated methods for assigning protein structural domains, to handle the sudden expansion in protein structural data. "CATH-Assign includes profile HMM-based methods for identifying domains in UniProt proteins (CATH-Resolve-Hits); a deep learning method for assigning homology to known families in CATH (CATHe); and fast structure comparison methods (FoldSeek), developed by the Martin Steinegger group, for verifying these relationships through determination of structure similarity to known relatives. Methods for assessing structure quality are also applied," says Orengo.

CATHe makes use of sequence embeddings generated by Prot-BERT-T5, a large language model developed by the Hannes Rost group, and is highly sensitive. It enabled the assignment of even remote homologues with less than 20% sequence identity to CATH superfamilies. The researchers used this approach to analyze the AlphaFold2-generated models for the proteomes of 21 model organisms. About half of these were of high enough quality for CATH classification, from which 92% could be assigned to existing CATH superfamilies. The researchers "manually analyzed a subset of unclassified structure clusters containing at least one human protein that could not be assigned to CATH superfamilies, and identified 24 novel superfamilies. Novel architectures were found, one of which, the 'heart' domain, adopts alternative conformations in solution," says Orengo.

Although CATH-Assign is faster and more sensitive than other approaches for structural classification, processing the full AlphaFold Protein Structure Database is still prohibitive. The researchers are working to extend CATH-Assign for more accurate domain detection and a complete end-to-end computational workflow. "Preliminary trials suggests that 100,000 models can be processed in less than 10 hours on a single CPU node," says Orengo. Full analysis is expected to further expand the CATH superfamily classification.

Arunima Singh
Nature Methods

Original reference: *Commun. Biol*. **6**, 160 (2023).

Developmental biology

Bat–virus entanglements

Bats are one of the most fascinating groups of mammals in the animal kingdom. Along with remarkable traits such as echolocation and an extremely long lifespan, bats are also tolerant of myriad viral infections, suggesting a long history of coevolution with viruses.

Previous studies have suggested that, upon viral infection, bats rapidly modulate their innate immune response to develop a state of tolerance. To further investigate the host–virus relationship in bats, researchers led by Thomas Zwaka at the Icahn School of Medicine at Mount Sinai have generated the first bat induced pluripotent stem (iPS) cells.

In a study reported in *Cell*, the team initially discovered that the protocol established for reprogramming human or mouse cells into iPS cells using the Yamanaka factors failed for bat embryonic fibroblasts, leading instead to the formation of non-proliferating primitive stem cell colonies. Following optimization, the team identified a combination of the Yamanaka factors and additional growth factors that led to successfully reprogrammed bat iPS cells. These cells expressed the pluripotency marker Oct4, as do other mammalian iPS cells, and had a proliferation capacity comparable to that of human iPS cells. Interestingly, the bat iPS cell nucleus has one or two large nucleoli filled with tiny vesicles, a feature not observed in other mammalian cells.

Transcriptomic and epigenetic analyses confirmed the pluripotent state of the bat iPS cells, but the researchers noticed that these cells simultaneously expressed key markers for both naive and primed iPS cells, suggesting a distinct stage of pluripotency. The team then functionally confirmed pluripotency by differentiating the iPS cell into the three germ layers and by demonstrating the formation of embryoid bodies, teratomas and blastoids.

Next, the researchers collected transcriptomic profiles from iPS cells from five divergent mammalian species — mouse, human, marmoset, dog and pig — and compared them to the bat transcriptome. Principal component analysis revealed that the bat iPS cells were the furthest away from all the other species, indicating a transcriptomic state unique to bats. This bat-specific gene signature was highly enriched for pathways associated with viral infection, implying that antiviral transcriptional pathways were probably triggered by the presence of endogenous viral sequences. Of note, the team discovered that the most enriched pathway in the bat transcriptome was related to coronavirus disease.

Indeed, metagenomic analyses confirmed the presence of a vast diversity of integrated and expressed viral elements in bat iPS cells, including those belonging to endogenous retroviruses and coronaviruses as well as sequences with homologies to viruses like monkeypox, squirrelpox and others.

This method to efficiently and reproducibly generate bat iPS cells is a starting point for studies into mechanisms of viral tolerance and reservoir establishment, for insights into bat biology and for efforts towards pandemic preparedness.

Madhura Mukhopadhyay
Nature Methods

Original reference: *Cell* **186**, 957–974.e28 (2023).