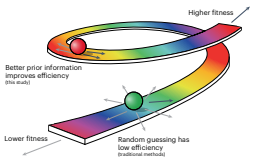
Molecular engineering

Protein language models guide directed antibody evolution



Better prior information on evolutionary plausibility helps improve protein engineering efficiency.

Directed protein evolution for generating antibodies with improved binding affinity or stability requires exploration of a vast space of possible mutations. Experimental high-throughput antibody engineering methods screen thousands to millions of variants using techniques like phage display or cell surface display. This imposes a heavy experimental burden. Computational methods provide a structure-guided rationale for selecting mutations, typically within the complementarity-determining regions, but still require experimental testing of many mutants.

A team of researchers led by Peter Kim at Stanford University has performed guided protein evolution using protein language models that were trained on millions of natural protein sequences. The models thereby learn amino acid patterns that are likely to be seen in nature. "Because the models are trained on millions of protein sequences produced by natural evolution, they are also helpful in suggesting mutations that are likely to have a functional impact when conducting directed evolution in the laboratory," says Brian Hie, the lead author of the paper. "Unlike other methods for machine-learning-guided directed evolution, our method also requires no initial task-specific training data and recommends mutations directly from the wild-type sequence alone", Hie adds.

Using these models, the team evolved seven human immunoglobulin G antibodies that bind to antigens from coronavirus, ebolavirus and influenza A virus. Screening 20 or fewer variants of each antibody across only two rounds of laboratory evolution allowed then to improve the affinity of all antibodies – an impressive feat. "We were able to improve the neutralization potency of an FDA-approved antibody against an Ebola pseudovirus and showed that for weak binders we can improve the affinity up to two orders of magnitude using very low-throughput experimentation," says Kim.

While the results show that general language models outperform antibody-specific language models, because the training is done on general sequences, there is no guarantee that the language-model-recommended mutations will improve binding affinity in every instance. "Another open question is whether these models could be applicable to evolving more unnatural, de novo designed proteins", says Hie. The researchers plan to leverage data beyond protein sequence information including protein structure information and binding affinity data to further improve the outcome. We look forward to developments in the area.

Arunima Singh
Nature Methods

Original reference: *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01763-2 (2023)

Immunology

Tracing lymphocyte development

The function of responding T or B cells is strongly influenced by their cell surface lymphocyte receptor. Although methods for multiomic analysis have enabled the study of lymphocytes at unprecedented resolution, there still remain some gaps in our interpretation of transcriptomic data in the context of the adaptive immune receptor repertoire (AIRR). To address this, researchers at the Wellcome Sanger Institute and University of Cambridge developed Dandelion, a computational framework for paired analysis of single-cell RNA sequencing (scRNA-seq) and single-cell VDJ sequencing (scVDJ-seq) data.

"We wanted to bring some of the bulk BCR-seq analysis capabilities to single cells, but we realized there are problems unique to single cells which other methods haven't tackled," says Kelvin Tuong, now at the University of Queensland and one of the lead authors of the study, published in *Nature Biotechnology*.

The Dandelion workflow acts as a bridge between AIRR sequencing tools and scRNA-seq data. Using scVDJ-seq data, Dandelion facilitates high-confidence contig annotation and B cell receptor (BCR) mutation calling. Unlike most contig annotation workflows, Dandelion does not filter out unproductive contigs. The researchers delved into the large proportion of T cell receptor (TCR) and BCR data from single-cell human fetal tissues that were unproductive, with absent V (variable region) genes. They found that multiple J (junction) genes could be mapped on the same mRNA contig, a phenomenon they called multi-J mapping, which is a unique functionality enabled by Dandelion.

Analysis of the developmental trajectory of immune cells poses a unique challenge since not only are developmental processes interrupted by proliferation and transcriptomic convergence but cell fate is often governed by the receptor. To allow refined analysis of combined gene expression and AIRR data, Dandelion creates a pseudobulked VDJ feature space. Here, cells can be grouped by neighborhoods or pseudobulks based on metadata for differential analysis or trajectory inference.

The team used AIRR data as 'timekeeper' and found that, by sampling cell neighborhoods with developing T cells, they could create a trajectory of development and associated gene expression patterns, mapping T cells from the initial double-positive stage to $CD4^+$ or $CD8^+$ lineage commitment. They also noticed a high proportion of non-productive TCR-β contigs in pre and pro mature B1 cells, suggesting alternate developmental routes for B1 cells and conventional B cells, as is consistent with findings from mouse studies. They also found unproductive TCR-β, TCR-δ and TCR-γ in innate lymphoid cells and the nature killer cell lineage, implying that these cells veer away from the T cell path between the double-negative and quiescent double-positive T cell stages.

The Dandelion workflow is an enabling method for studying the trajectories of immune cells in various in vivo and in vitro settings and has the potential to improve our understanding of dynamic immune processes.

Madhura Mukhopadhyay
Nature Methods

Original reference: *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01734-7 (2023)