Biological Data Scientist Technical Questions

Timothy Daley

Questions

Suppose we have an assay that measures some functional activity of a protein. We want to build a machine learning model based on the output of screening experiments using this assay, each screen containing thousand of protein molecules derived from a large number of species. Please describe how you would design the experiments, the model training process, and the model evaluation process to ensure that the model is generalizable.

I want the candidate to describe in some vague detail about the model training, and setting up the experiments to ensure that the model built will not overfit on technical noise or batch effects.

- Mentioning batch effects and the need to control them is what we're looking for. A simple train-test split will not work here because of the batch effects.
- They should also recognize that models trained on one species might not generalize to others.
- An ideal split is either a species holdout, or an experimental hold out, or both.

Suppose a wet lab scientist comes to you for advice about an experiment. They want to identify the mechanism of a particular drug treatment. How would you suggest they design the experiment and what questions would you ask?

The key here is that they identify questions for the wet lab scientist. e.g.

- What is currently known about the drug?
- What work has previously been done?
- What is this drug used for?
- What is important outcome of the experiment? What do we want out of the experiment?