Problem Statement

Computationally predict CD spectra from primary amino acid sequences.

Context

The ability to accurately predict protein properties from primary amino acid sequence has long been an area of interest for academia and industry alike. From an academic perspective, our ability to do so reports directly on the field's understanding of the first principles driving these properties. For companies, *in silico* discovery and developability programs offer the possibility of screening vastly higher numbers of potential sequences and lowering failure rates for costly later steps in the development pipeline. The influx of novel protein design algorithms over the last decade has added additional nuances to the problem; these sequences lack evolutionary context and constraints and thus often behave in a surprising manner, suggesting that the algorithms trained on natural proteins are incorporating information beyond the physics necessitated by the primary sequence.

Circular dichroism (CD) is a spectroscopic technique that relies on differential absorption of right- and left-circularly polarized light. Readouts in the far-UV range (190-250 nm) report primarily on the angles adopted by the peptide backbone and therefore protein secondary structure. Each main type of secondary structure has characteristic wavelengths for spectrum maxima and minima. The technique is sufficiently sensitive that it is traditionally used to determine thermodynamic stability by monitoring change in signal at a single wavelength (usually 222 nm) in response to perturbations like temperature or chemical denaturant or as a "fingerprint" technique to make structural observations about changes due to mutation or ligand binding by observing where the resultant spectra differ.

Criteria for success

Deliverable: Model that takes amino acid sequence as input and produces a CD spectrum as output. This should have sufficient documentation and be packaged in such a way as to be easily accessible to potential users (GitHub Repo with report/slide deck/other user guidance).

Considerations for validation:

1) Predicted spectra accurately reproduce experimentally determined spectra.
2) Able to predict spectra for proteins without experimentally determined spectra; spectra are consistent with other available structural knowledge.

Scope of solution space

CD spectra that have been entered into PCDDB. Amino acid sequences can be pulled from UniProt using UniProt IDs. Validation can be done with small subset of individually identified spectra/sequence combinations from the literature.

Some other potential questions to explore, depending on time & data: How do these predictions compare with predictions based on structure? If put through a deconvolution algorithm, how do the result compare with secondary structure prediction algorithms? Is there something interesting to be said about areas where these do not align? Is this prediction algorithm sufficiently nuanced to predict spectral changes for protein variants?

Constraints

CD spectra for the same protein can look different under different conditions (e.g. buffer conditions, temperature, etc.). Developing the model should take this into account and think through how to deal with it both for the training dataset and for the prediction presentations.

As alluded to in the context, it is likely that all or most of the proteins in the PCDDB are natural proteins. Therefore, how these predictions will hold up for artificially designed sequences is an open question.

Raw CD data is generally in units of millidegrees; this is then converted to mean residue ellipticity (MRE) or equivalent to allow for greater comparability to other experiments. How to deal with these is something to be considered, as the conversion required a very precise measure of concentration (a potential source of error/difference between labs and techniques) and pathlength (not an issue for traditional instruments, but problematic for high-throughput instruments where pathlength is dictated by the volume per well).

Stakeholders

Springboard

Scientific community, esp.

- Those using CD and/or secondary structure prediction tools
- Those interested in *in silico* construct screening based on biophysical parameters

Data sources

Protein Circular Dichroism Data Bank (PCDDB) https://pcddb.cryst.bbk.ac.uk/

Protein sequences can be scraped from:

- UniProt https://www.uniprot.org/help/api_downloading
- NCBI https://www.ncbi.nlm.nih.gov/guide/data-software/

CD spectra prediction from structures: DichroCalc https://comp.chem.nottingham.ac.uk/dichrocalc/