protein language models: summary

A protein language model turns each amino acid sequence into an embedding, a row of numbers reflecting patterns the model learned from vast numbers of protein sequences. Proteins with similar embeddings tend to share evolutionary, structural or functional features, and because a transformer's attention mechanism relates every amino acid to every other, the embedding can reflect residues that are far apart in the sequence but close in the folded protein. Unlike a multiple sequence alignment, an embedding needs no homologous sequences and is fast to compute once the model is trained. The workflow has three steps: retrieve sequences, embed them, and run a PCA on the embedding columns to see whether proteins with different functions separate.

Step	Function	What it is for	Arguments and settings used in this chapter
1	`searchNCBI()`	retrieve protein sequences from NCBI as an AAStringSet	`search_term`, `retmax`
1	(subsetting)	keep sequences short enough to embed whole, since esm2 truncates long sequences (the chapter keeps those under 1024 amino acids)	`all_sequences[all_sequences@ranges@width < 1024]`
2	`embedAminoAcids()`	embed each sequence with a pre-trained protein language model	`amino_acid_stringset`, `platform` (`"nvidia"` runs esm2-650m, bioLM offers several models), `nvidia_api_key` or `biolm_api_key`
3	`runMatrixAnalysis()`	PCA on the embedding columns to see which proteins group together	`analysis = "pca"`, `columns_w_values_for_single_analyte` (the embedding columns), `columns_w_sample_ID_info`