GROUP FINAL PROJECT

This project was done individually since all the other groups were full and one of the members backed out :((

Goal: Increase the structural stability of the MS2 L-protein

The MS2 bacteriophage L-protein (75 residues) is intrinsically disordered in its soluble N-terminal domain and depends on the E. coli chaperone DnaJ for proper folding. A key resistance mechanism bacteria use is a point mutation in DnaJ that prevents it from interacting with the L-protein, abolishing lysis. Engineering a more intrinsically stable L-protein that can fold without DnaJ would directly overcome this resistance mechanism.

Chosen sub-problem: Increase intrinsic stability of the soluble domain (residues 1–40)

This is the easiest of the three goals because:

(1) stability is the most tractable property for computational protein design tools

(2) the ESM language model scores directly reflect evolutionary likelihood which correlates with fold stability

(3) We don't need to model complex protein-protein interactions or membrane physics to make progress.

Proposed pipeline:

1. Run an ESM2 deep mutational scan on the wild-type L-protein to get per-position log-likelihood ratio (LLR) scores for all possible single-point substitutions. Positions with consistently positive LLR scores are predicted to be tolerated or stabilizing by the language model.

2. Cross-reference the LLR scores against the experimental mutational analysis dataset (Chamakura et al., 2017) to filter out any high-scoring mutations that are known to abolish lysis in practice. This step ensures computational predictions are grounded in experimental reality.

3. Check ClustalOmega alignments of L-protein homologs (from pBLAST results) to identify conserved positions. Avoid mutating fully conserved residues (conservation is a strong proxy for functional or structural essentiality).

4. Combine the top 4–6 compatible substitutions in the soluble domain into candidate mutant sequences. Prioritize positions with positive LLR, positive experimental outcome,

and low conservation.

5. Fold each candidate mutant using ESMFold and compare predicted pLDDT (confidence) scores to the wild-type. Higher pLDDT in the soluble domain would suggest the mutant adopts a more defined, stable fold without chaperone assistance.

6. Run AF2-Multimer to co-fold each promising mutant with DnaJ, and separately as an 8-mer oligomer, to confirm that: (a) the DnaJ interaction is disrupted or reduced (DnaJ independence), and (b) the capacity to oligomerize into a membrane-disrupting pore is preserved.

Why these tools are appropriate:

ESM2 was trained on hundreds of millions of natural protein sequences, learning the statistical rules of what makes proteins stable and functional. Its LLR scores serve as an unsupervised proxy for fitness residues that the model predicts as unlikely in a given context tend to be destabilizing. ESMFold then translates sequence changes into predicted structural confidence, giving a direct readout of whether a mutation improves structural definition. AF2-Multimer adds a check on the biological output (DnaJ disruption, oligomerization) to confirm the stability-improving mutations don't destroy function.

Potential pitfalls:

- ESM was trained on evolutionarily diverse proteins but may have limited coverage of short phage proteins like the L-protein. Its LLR scores may be less reliable here than for well-conserved globular proteins, leading to false positives. This is why cross-checking against experimental data is essential.
- The L-protein's soluble domain is intrinsically disordered. ESMFold and even AlphaFold are less accurate at modeling disordered regions, so low pLDDT scores may reflect genuine disorder rather than poor design. This makes it hard to distinguish "good mutant, still disordered" from "bad mutant, misfolded."