Improving PPI Prediction with ESM-derived Features

戴佳文 齐奕婷 王安瑞

1. Problem and Goal

Understanding protein interactions (PPIs) is vital for biology and medicine.

Problem: Improve computational PPI prediction accuracy and understanding.

Goal: Use features from a new protein AI model, ESM, to make better and more understandable PPI predictions. We think ESM's rich protein understanding will give us better features.

2. Current Methods

Existing computational methods have issues:

1. **Using biological annotations:** Relies on existing data, doesn't work well for less-known proteins.
2. **Using only protein sequences:** Can be good but struggles with new proteins, and hard to apply across different species without more context.

3. Our New Approach & What We'll Do

1. Use the advanced **ESM model** to create powerful new features(high-dimensional embeddings) for individual proteins.
2. Use these features to predict protein pairs that interact.

What we will contribute:

1. **Creating New Features from a cutting-edge AI (ESM):**
    - We'll extract high-dimensional embeddings from ESM for single proteins.
    - We'll figure out how to combine these single-protein features to represent a *pair* of proteins, ready for interaction prediction.
    - **Why this is new:** ESM features understand sequence, structure, and function together, unlike older methods.

2. **Trying Different ML Models & Combining Features:**
    - We'll test our new ESM features with various ML models (like MLPs, SVMs, CNNs).
    - **Crucially, we'll develop ways to mix our new ESM features with older types of features** (like those from biological annotations).
    - **Why this is important:** Combining the strengths of new AI features and existing biological knowledge should lead to better, more robust predictions.

3. **Using Unsupervised Learning:**
   - We'll explore if analyzing our ESM features without using known PPI labels can help.
   - This might involve grouping proteins by their features to find better feature subsets or estimate how likely proteins are to interact based on their feature similarity.
   - **Why this helps:** This could reveal hidden patterns in ESM's understanding of proteins that improve or complement our prediction task.