Paper Review

ProCyon

Paper Review

Background and Motivation

- 20% of human proteins lack known function
- 95% of research focus on 5000 well-studied proteins
- Majority of human proteome is unexplored… can we automate it?

Research Gap

- Many relies on pre-defined functional categories
- Most of protein language models (PLMs) are text-based
  - Sequences cannot capture all possible phenotypes
  - e.g. it cannot capture convergent evolution

Paper Review

Research Gap

- Phenotype is confusing -> not one-to-one relationship

ProCyon aims to solve all of the gaps

- Allow for open-ended QA (thanks to 11B LLaMA3 parameters)
  - Can determine free-text phenotype beyond what is seen in training
- Multimodal (not just text-based)
  - Protein sequence, protein structure, natural language
- Can model many-to-many relationships between proteins and phenotypes

Paper Review

ProCyon Functions

1. Retrieval: phenotype description => matched proteins
2. QA: phenotype + protein => match / no match
3. Free-text phenotype generation: protein + context => predicted phenotype descriptions
    - e.g. Given a protein and information about a drug, ProCyon can describe the mechanism of the drug with the proteins

Paper Review

Results

- F_{max} = 0.743 compared to 0.618 for the next best model
- QA accuracy of 72.7% compared to 67.8% of the next best model
- Able to determine the correct drug-binding protein domains 65.8% of the time (zero-shot)

Paper Review

ProCyon-BIND

- To show the capability of ProCyon in zero-shot task transfer
- Goal: finetune ProCyon to identify peptides that bind to target proteins

- ProCyon is finetuned with data from PDBBind to get ProCyon-BIND
- Experimentally validated by experimentally validated data
    - Between ACE2 receptor and 58 protein binders + 5072 protein non-binders

Paper Review

ProCyon-BIND

- ProCyon-BIND yields statistically significant separation between binders and non-binders
  - AUROC 0.6480
  - Compared to ESM-2 with 0.3923 AUROC
- Not the best, but there is no clear peptide sequence motif that experimentally identified to drive binding
  - So distinguishing is pretty hard
- Illustrated the versatility of ProCyon as a base model

Paper Review

Strengths

- Represents a shift from modeling protein sequences to modeling protein phenotypes
- Zero shot task transfer capabilities are quite cool
- Natural language abilities democratizes access to unexplored proteins phenotyping

Paper Review

Limitations

- Evaluating novel phenotypes is challenging
  - Due to scarce ground truth data
- Static nature of ProCyon limits its ability to adapt to evolving biological research
- No conversational abilities yet
- No uncertainty metrics generated together with predictions