S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure

Qing Shao
qshao@uky.edu
Department of Chemical and Materials Engineering
University of Kentucky
https://www.biorxiv.org/content/10.1101/2023.08.06.52203v1.full
Acknowledgment

• Dr. Duolin Wang (U. Missouri)
• Dr. Dong Xu (U. Missouri)
• Dr. Jin Chen (Uky, currently U. Alabama)
• Usman Abbas (UKy)

• Funding Support:
  • University of Kentucky Start-up Funds
  • AI for Medicine Alliance Pilot
Why need protein language models?

• Proteins are biomolecules composed of twenty natural amino acids
• Proteins play a central role in human health

Chemical structure of 20 amino acids

https://www.compoundchem.com/2014/09/16/aminoacids/
Why need protein language models?

• The protein universe is huge.
• (One cell may have ~10 K different types of proteins)
• We need tools to provide reliable information about proteins in a quick way
Why need protein language models?

- Proteins structures determine their properties

https://www.mun.ca/biology/scarr/iGen3_06-04.html
Why need protein language models?

• It is quite natural to find similarities between the protein primary structure and a sentence

Protein sequence

A sentence

I will present at the CCS 2023 summit.

Thus, protein language models can be the tool we are looking for.

A typical protein language model

The development of a protein language model is similar to the development of a language model. One of the most valuable products is the embedding of the language model. We could use the embedding to develop downstream tasks for protein design and property prediction.

What is embedded in the embedding is the key

https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1.full
The current protein language models learn the primary structure

- Proteins structures determine their properties
We must let the protein language model learn the other structures

https://www.mun.ca/biology/scarr/iGen3_06-04.html
Multiview contrastive learning enables the fusion of information from different sources.
Multiview contrastive learning enables the fusion of information from different sources
The architecture of our model

We adapt simCLR loss function

What is contact map?

Contact map is a 2D matrix. The element (i,j) is the distance between amino acid i and j. The distance is determined based on the Cα atom.
Where to get the protein structure?

- We cannot do our work without the great AlphaFold2

https://alphafold.ebi.ac.uk/
Our data

- Swiss-Prot Database (~ 580 K protein sequences)
- Gain the 3D structure from AlphaFold2 predictions
- Gain the contact map using self-developed Python code.

https://alphafold.ebi.ac.uk/
Some other useful detail

- 540 K proteins for training
- 40 K proteins for validation
- Play with hyperparameters
- Trained on our LCC V100
- Code developed based Pytorch
Some preliminary results

Contrastive learning enhances the alignment between sequence and structure embedding. This enhancement implies that the embedding knows structure better.

https://www.biorxiv.org/content/10.1101/2023.08.06.552203v1.full
CATH superfamily task

Our sequence embedding performs better on structure task than the original ESM2 model
Some preliminary results

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Metrics</th>
<th>S-PLM</th>
<th>PEER paper ESM-1b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betalactamase ($\beta - lac$)</td>
<td>Spearmanr</td>
<td>0.90 (0.002)</td>
<td>0.84 (0.053)</td>
</tr>
<tr>
<td>Solubility (Sol)</td>
<td>Accuracy</td>
<td>72.09 (0.002)</td>
<td>70.23 (0.75)</td>
</tr>
<tr>
<td>Subcellular localization (Sub)</td>
<td>Accuracy</td>
<td>79.84* (0.001)</td>
<td>79.82* (0.18)</td>
</tr>
<tr>
<td>Secondary structure (SSP)</td>
<td>Accuracy</td>
<td>86.88* (0.001)</td>
<td>83.14* (0.10)</td>
</tr>
</tbody>
</table>

* Used as a feature extractor with the pre-trained PLM weights frozen. The task names used in the PEER paper (Table 3 [18] ) are indicated in parentheses.

https://arxiv.org/abs/2206.02096
Summary

• We have successfully implemented structure information into sequence embedding
• The developed structure-aware protein language models perform better in some downstream tasks