2024-07-10

TO-DO

- [x] ~~Terminar de leer Survey~~
- [x] ~~Métricas~~
- [x] ~~Posibles áreas de oportunidad~~
- [ ] Estudiar alphafold
- [x] ~~Estado del arte~~

Survey

1. Backgrounds

Evolution based and single sequence based PSP

Ideas importantes:

- polypeptide is flexible and can fold into a staggering number of different shapes
- computational methods for predicting 3D protein structures from protein sequences have traditionally taken two parallel paths, focusing on either physical interactions or evolutionary principles
- Protein databases support to get multiple sequence alignment (MSA) of homologous proteins, which significantly benefits the development of evolutionary methods.
- The introductions of MSA input and pLMs have led to the vast success of AlphaFold2 (AF2)

- we outline several works for single sequence PSP, structure prediction of antibodies, protein complexes, protein-ligands, protein-RNA, and protein conformational ensembles
- https://github.com/bozhenhhu/A-Review-of-pLMs-and-Methods-for-Protein-Structure-Prediction

2 y 3.Notions and Terms and Protein and Language

- Compared with algorithms in NLP, protein tokenization methods still remain at a low level without a well-defined and biologically meaningful protein token algorithm (Vu et al., 2022). This may be a direction for unlocking the secrets of proteins.

4.Language Models

- **RNN**
- **LSTM**
- **Attention Mechanism and Transformer**
- **Pre-trained Language Models**

Table 1: Representative pre-trained LMs in general domains

Model	Network	Objective	#Params.	Comments
ELMo (Peters et al., 2018)	LSTM	Bidirectional LM	93.6M	the first deep contextualized word representation
GPT (Radford et al., 2018a)	Transformer	Autoregressive LM	110M	a pre-trained LM for predicting the next word
BERT (Devlin et al., 2018)	Transformer	Masked LM	340M	the most commonly-used LM for predicting masked tokens
Transformer-XL (Dai et al., 2019)	Transformer	Autoregressive LM	257M	enabling learn dependencies beyond a fixed length
XLM (Lample and Conneau, 2019)	Transformer	Multi-task	665M	cross-lingual pre-training
Udify (Kondratyuk and Straka, 2019)	BERT	Multi-task	~340M	leveraging a multilingual BERT self-attention model
GPT-2 (Radford et al., 2022)	Transformer	Autoregressive LM	1.5B	larger model, more training data
Grover (Zellers et al., 2019)	GPT2	Autoregressive LM	1.5B	defending against the general neural fake news
XLNet (Yang et al., 2019b)	BERT	Autoregressive LM	~340M	more training data, integrates ideas from Transformer-XL
RoBERTa (Liu et al., 2019d)	BERT	Masked LM	355M	more training data, dynamic masking
CTRL (Keskar et al., 2019)	Transformer	Autoregressive LM	1.63B	trained to control particular aspects of the generated text
Megatron-LM (Shoeybi et al., 2019)	Transformer	Autoregressive LM	8.3B	a large transformer model
ALBERT (Lan et al., 2019)	BERT	Masked LM	223M	a lite BERT
DistillBERT (Sanh et al., 2019)	BERT	Masked LM	65M	a distilled version of BERT
SpanBERT (Joshi et al., 2019)	BERT	Masked LM	~340M	presenting and predicting the masked span
MASS (Song et al., 2019)	Transformer	Seq2Seq LM	~307M	masked Seq2Seq pre-training
MT-DNN (Liu et al., 2019c, 2020)	BERT	Multi-task	~340M	for multiple natural language understanding tasks
MT − DNN_{KD} (Liu et al., 2019b)	MT-DNN	Multi-task	~340M	incorporating knowledge distillation
ERNIE (Zhang et al., 2019)	BERT	Masked LM	~114M	incorporating knowledge graphs
KnowBERT (Peters et al., 2019)	BERT	Masked LM	~110M	incorporating knowledge bases into BERT
KEPLER (Wang et al., 2019)	RoBERTa	Masked LM	~125M	incorporating knowledge embedding
VideoBERT (Sun et al., 2019a)	BERT	Multimodal model	~340M	modelling between the visual and linguistic domain
VisualBERT (Li et al., 2019)	BERT	Multimodal model	~110M	modelling a broad range of vision and language tasks
ERNIE (Sun et al., 2019b)	BERT	Masked LM	~110M	a new masking strategy
PEGASUS (Zhang et al., 2020c)	Transformer	Masked & Seq2Seq LM	568M	for abstractive text summarization
Unicoder-VL (Li et al., 2020b)	BERT	Multimodal model	~110M	cross-modal learning
UNILM (Bao et al., 2020)	BERT	Bidirectional & Seq2Seq LM	~110M	for natural language understanding and generation tasks
Turing-NLG (Rasley et al., 2020)	Transformer	Autoregressive LM	17B	a hugely large LM
ELECTRA (Clark et al., 2020)	BERT	Generator & Discriminator	335M	token detection
GPT-3 (Brown et al., 2020)	GPT-2	Autoregressive LM	175B	extending the model size
T5 (Raffel et al., 2020)	Transformer	Seq2Seq LM	11B	producing new text as output
Switch Transformer (Fedus et al., 2021)	Transformer	Masked LM	1.6T	increasing the pre-training speed
BEIT (Bao et al., 2021)	Transformer	Masked image model	307M	a vision Transformer
MT-NLG (Smith et al., 2022)	Transformer	Autoregressive LM	530B	the largest publicly monolithic transformer

All examples report the largest model of their public series. Network displays high-level backbone models preferentially if they are used to initialize parameters. #Param. means the number of parameters; M, millions; B, billions; T, trillions; & , and; ~ means estimated data. Related terminologies are listed in Section 2.

5.Protein Language Models

LSTM models

desarrollados para predecir varias características estructurales de las proteínas a partir de sus secuencias de aminoácidos.

Importante: Rao et al. (2019) (TAPE): Evaluaron un grupo de modelos de proteínas en un panel de tareas y concluyeron que existen oportunidades para diseñar innovaciones específicas en los modelos de proteínas y métodos de entrenamiento para LSTM y Transformers estándar

-Combinación con Redes Neuronales Convolucionales

-Enfoque en Secuencias Únicas

-Pre-entrenamiento y Embeddings para Tareas Downstream

-Representaciones Ricas sin Datos Estructurales o Evolutivos

-Supervisión Estructural Adicional

-Maximización de la Información Mutua

Transformers models

ventaja: aprenden la gramática de las proteínas incluso sin usar información evolutiva

-Entrenamiento a Gran Escala

-Estrategia de Enmascaramiento Compleja

-Captura de Correlaciones de Coevolución

-Mejora del Conocimiento Multisource de las Representaciones de Proteínas (MSA, funciones, estructuras y conocimientos biológicos previosembeddings)

IMPORTANTE: PEER benchmarks (Xu et al., 2022) were built for protein sequence understanding, including protein function and structure prediction, protein-protein interaction prediction, protein-ligand interaction prediction tasks, etc.

- **Primeros pLMs**: predecir características estructurales (estructura secundaria, SOLVENT, ángulos de torsión, homología remota y mapas de contacto.)
- **Actuales**: predecir la predicción de la estructura de proteínas de manera integral.

Evoformer (Jumper et al., 2021) ALPHAFOLD 2

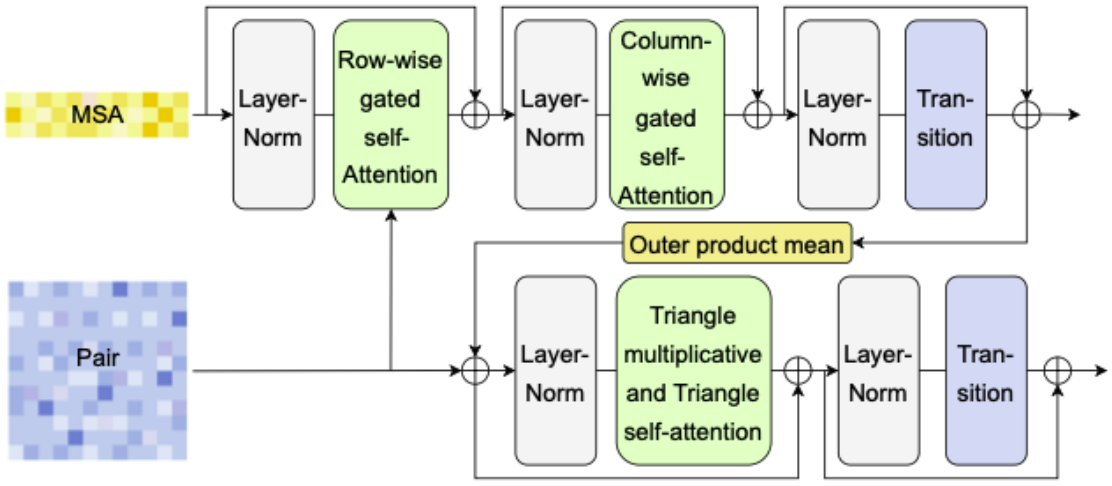

Figure 6: Evoformer block.

6. Methods of Protein Structure Prediction (PSP)

Trans

Las características estructurales incluyen características 1D (estructura secundaria , solvente , ángulos de torsión, densidad de contacto, etc.) y características 2D (mapa de contactos y mapa de distancias), las cuales son útiles para predecir estructuras de proteínas. En las primeras etapas de la predicción de estructuras de proteínas, en lugar de predecir directamente las coordenadas de los átomos, se predecían características estructurales.

(...)

7. Discussion: Limitations and Future Trends

Problemas

- AF2 no predice bien el impacto de mutaciones en proteínas.

- Falta de interpretabilidad en los modelos basados en pLMs

propuestas:

- Combinar información adicional, incluyendo priors biológicos y físicos, información evolutiva, diferentes niveles de estructuras, GO, etc., puede reducir el tamaño de los modelos y mejorar su rendimiento.
- Aprendizaje Multi-tarea o Multi-moda
- Desarrollar métodos de tokenización de proteínas de alta calidad o combinarlos con otros métodos de aprendizaje

8. **Databases**

Table 4: Information of Protein Databases

Dataset	#Proteins	Disk Space	Description	Link
UniProtKB/Swiss-Prot	500K	0.59GB	knowledgebase	https://www.uniprot.org/uniprotkb?query=*
UniProtKB/TrEMBL	229M	146GB	knowledgebase	https://www.uniprot.org/uniprotkb?query=*
UniRef100	314M	76.9GB	clustered sets of sequences	https://www.uniprot.org/uniref?query=*
UniRef90	150M	34GB	90% identity	https://www.uniprot.org/uniref?query=*
UniRef50	53M	10.3GB	50% identity	https://www.uniprot.org/uniref?query=*
UniParc	528M	106GB	Sequence	https://www.uniprot.org/uniparc?query=*
PDB	190K	50GB	3D structure	https://www.wwpdb.org/ftp/pdb-ftp-sites
CATH4.3	N/A	1073MB	hierarchical classification	https://www.cathdb.info/
BFD	2500M	272GB	sequence profile	https://bfd.mmseqs.com/
Pfam	47M	14.1GB	protein families	https://www.ebi.ac.uk/interpro/entry/pfam/
AlphaFoldDB	214M	23 TB	predicted 3D structures	https://alphafold.ebi.ac.uk/
ProteinKG25	5.6M	147MB	a KG dataset with GO	https://drive.google.com/file/d/1iTC2-zbvYZCDhWM_wxRufCvV6vvPk8HR
Uniclust30	N/A	6.6GB	clustered protein sequences	https://uniclust.mmseqs.com/
SCOP	N/A	N/A	structural classification	http://scop.mrc-lmb.cam.ac.uk/
SCOPe	N/A	86MB	extented version of SCOP	http://scop.berkeley.edu

K, thousand; M, million, disk space is in GB or TB (compressed storage as text), which is estimated data influenced by the compressed format.

- **UNIPROT**
  - **UniProtKB**:

  Base de datos central de proteínas que proporciona información detallada sobre las secuencias y anotaciones de proteínas.

  Swiss-Prot:

  Subconjunto revisado manualmente que proporciona anotaciones de alta calidad y baja redundancia.

  TrEMBL:

  Descripción: Subconjunto que contiene entradas anotadas automáticamente y no revisadas manualmente

  - **UniRef**:

  Agrupa secuencias de proteínas similares para reducir la redundancia

  - **UniParc (The UniProt Archive)**:

Base de datos no redundante de secuencias de proteínas que rastrea la historia de cada secuencia a través de múltiples bases de datos.

- **Protein Data Bank (PDB)**:

Base de datos que recopila estructuras tridimensionales de proteínas y ácidos nucleicos obtenidas mediante técnicas experimentales como la cristalografía de rayos X y la resonancia magnética nuclear (NMR).

- **AlphaFold Protein Structure Database**:

Creada por DeepMind y EMBL-EBI que contiene estructuras de proteínas predichas por AlphaFold.

- **Big Fantastic Database (BFD)**:

Compendio de secuencias de proteínas de varias bases de datos, utilizado para entrenar modelos de lenguaje de proteínas como ProtTrans.

- **Pfam**:

Base de datos de familias de proteínas que proporciona anotaciones y alineamientos de secuencias para familias de proteínas.

- **InterPro**:

Base de datos que integra diversas bases de datos de familias y dominios de proteínas para proporcionar anotaciones comprensivas.

- **Ensembl**:

Base de datos que proporciona anotaciones de genes y secuencias genómicas.

- **RefSeq**:

Base de datos de secuencias de referencia anotadas para genomas, genes y proteínas.

- **GenBank**:

Base de datos de secuencias nucleotídicas y proteicas proporcionada por el NCBI.

- **DDBJ**:

Secuencias nucleotídicas mantenida por el DNA Data Bank of Japan.

- **MGnify**:

Base de datos de secuencias metagenómicas que proporciona análisis y anotaciones de datos metagenómicos.

Métricas

Benchmarks

TAPE

1. Predicción de Funciones de Proteínas

1. **Predicción de Fluorescencia**

- **Fuente de Datos**: Dataset de Sarkisyan
- **Tamaño**: 21,446 (Entrenamiento), 5,362 (Validación), 27,217 (Prueba)
- **Métrica**: Coeficiente de correlación de Spearman

2. **Predicción de Estabilidad**

- **Fuente de Datos**: Dataset de Rocklin
- **Tamaño**: 53,571 (Entrenamiento), 2,512 (Validación), 12,851 (Prueba)
- **Métrica**: Coeficiente de correlación de Spearman

3. **Predicción de Actividad de β-lactamasa**

- **Fuente de Datos**: Envision
- **Tamaño**: 4,158 (Entrenamiento), 520 (Validación), 520 (Prueba)
- **Métrica**: Coeficiente de correlación de Spearman

4. **Predicción de Solubilidad**

- **Fuente de Datos**: DeepSol
- **Tamaño**: 62,478 (Entrenamiento), 6,942 (Validación), 1,999 (Prueba)
- **Métrica**: Precisión

2. Predicción de Localización de Proteínas

1. **Predicción de Localización Subcelular**

- **Fuente de Datos**: DeepLoc
- **Tamaño**: 8,945 (Entrenamiento), 2,248 (Validación), 2,768 (Prueba)

- **Métrica**: Precisión

2. **Predicción Binaria de Localización**

- **Fuente de Datos**: DeepLoc
- **Tamaño**: 5,161 (Entrenamiento), 1,727 (Validación), 1,746 (Prueba)
- **Métrica**: Precisión

3. Predicción de Estructura de Proteínas

1. **Predicción de Contactos**

- **Fuente de Datos**: ProteinNet
- **Tamaño**: 25,299 (Entrenamiento), 224 (Validación), 40 (Prueba)
- **Métrica**: Precisión en L/5

2. **Clasificación de Plegamientos**

- **Fuente de Datos**: DeepSF
- **Tamaño**: 12,312 (Entrenamiento), 736 (Validación), 718 (Prueba)
- **Métrica**: Precisión

3. **Predicción de Estructura Secundaria**

- **Fuente de Datos**: NetSurfP-2.0
- **Tamaño**: 8,678 (Entrenamiento), 2,170 (Validación), 513 (Prueba)
- **Métrica**: Precisión

4. Predicción de Interacciones Proteína-Proteína

1. **Predicción de PPI en Levadura**

- **Fuente de Datos**: Dataset de Guo
- **Tamaño**: 1,668 (Entrenamiento), 131 (Validación), 373 (Prueba)
- **Métrica**: Precisión

2. **Predicción de PPI en Humanos**

- **Fuente de Datos**: Dataset de Pan
- **Tamaño**: 6,844 (Entrenamiento), 277 (Validación), 227 (Prueba)
- **Métrica**: Precisión

3. **Predicción de Afinidad de PPI**

- **Fuente de Datos**: SKEMPI

- **Tamaño**: 2,127 (Entrenamiento), 212 (Validación), 343 (Prueba)
- **Métrica**: RMSE (Error cuadrático medio)

5. Predicción de Interacciones Proteína-Ligando

1. **Predicción de Afinidad en PDBbind**

- **Fuente de Datos**: PDBbind-2019
- **Tamaño**: 16,436 (Entrenamiento), 937 (Validación), 285 (Prueba)
- **Métrica**: RMSE

2. **Predicción de Afinidad en BindingDB**

- **Fuente de Datos**: BindingDB
- **Tamaño**: 7,900 (Entrenamiento), 878 (Validación), 5,230 (Prueba)
- **Métrica**: RMSE

PEER

1. Predicción de Funciones de Proteínas

1. **Predicción de Fluorescencia**

- **Fuente de Datos**: Dataset de Sarkisyan
- **Tamaño**: 21,446 (Entrenamiento), 5,362 (Validación), 27,217 (Prueba)
- **Métrica**: Coeficiente de correlación de Spearman

2. **Predicción de Estabilidad**

- **Fuente de Datos**: Dataset de Rocklin
- **Tamaño**: 53,571 (Entrenamiento), 2,512 (Validación), 12,851 (Prueba)
- **Métrica**: Coeficiente de correlación de Spearman

3. **Predicción de Actividad de β-lactamasa**

- **Fuente de Datos**: Envision
- **Tamaño**: 4,158 (Entrenamiento), 520 (Validación), 520 (Prueba)
- **Métrica**: Coeficiente de correlación de Spearman

4. **Predicción de Solubilidad**

- **Fuente de Datos**: DeepSol
- **Tamaño**: 62,478 (Entrenamiento), 6,942 (Validación), 1,999 (Prueba)
- **Métrica**: Precisión

2. Predicción de Localización de Proteínas

1. **Predicción de Localización Subcelular**

- **Fuente de Datos**: DeepLoc
- **Tamaño**: 8,945 (Entrenamiento), 2,248 (Validación), 2,768 (Prueba)
- **Métrica**: Precisión

2. **Predicción Binaria de Localización**

- **Fuente de Datos**: DeepLoc
- **Tamaño**: 5,161 (Entrenamiento), 1,727 (Validación), 1,746 (Prueba)
- **Métrica**: Precisión

3. Predicción de Estructura de Proteínas

1. **Predicción de Contactos**

- **Fuente de Datos**: ProteinNet
- **Tamaño**: 25,299 (Entrenamiento), 224 (Validación), 40 (Prueba)
- **Métrica**: Precisión en L/5

2. **Clasificación de Plegamientos**

- **Fuente de Datos**: DeepSF
- **Tamaño**: 12,312 (Entrenamiento), 736 (Validación), 718 (Prueba)
- **Métrica**: Precisión

3. **Predicción de Estructura Secundaria**

- **Fuente de Datos**: NetSurfP-2.0
- **Tamaño**: 8,678 (Entrenamiento), 2,170 (Validación), 513 (Prueba)
- **Métrica**: Precisión

4. Predicción de Interacciones Proteína-Proteína

1. **Predicción de PPI en Levadura**

- **Fuente de Datos**: Dataset de Guo

- **Tamaño**: 1,668 (Entrenamiento), 131 (Validación), 373 (Prueba)
- **Métrica**: Precisión

2. **Predicción de PPI en Humanos**

- **Fuente de Datos**: Dataset de Pan
- **Tamaño**: 6,844 (Entrenamiento), 277 (Validación), 227 (Prueba)
- **Métrica**: Precisión

3. **Predicción de Afinidad de PPI**

- **Fuente de Datos**: SKEMPI
- **Tamaño**: 2,127 (Entrenamiento), 212 (Validación), 343 (Prueba)
- **Métrica**: RMSE (Error cuadrático medio)

5. Predicción de Interacciones Proteína-Ligand

1. **Predicción de Afinidad en PDBbind**

- **Fuente de Datos**: PDBbind-2019
- **Tamaño**: 16,436 (Entrenamiento), 937 (Validación), 285 (Prueba)
- **Métrica**: RMSE

2. **Predicción de Afinidad en BindingDB**

- **Fuente de Datos**: BindingDB
- **Tamaño**: 7,900 (Entrenamiento), 878 (Validación), 5,230 (Prueba)
- **Métrica**: RMSE

Estado del Arte

CASP15

Modelado de Proteínas Individuales y Dominios:

Evaluación de la precisión de las proteínas individuales y dominios, con énfasis en la precisión fina de motivos de la cadena principal y cadenas laterales

Ensamblaje:

Modelado de interacciones proteína-proteína

Modelos de Estructuras y Complejos de RNA:

Experimento piloto para evaluar la precisión del modelado de estructuras de RNA y complejos proteína-RNA.

Ensamblajes Conformacionales de Proteínas:

Evaluación de métodos para predecir ensamblajes estructurales, desde conformaciones desordenadas hasta dinámicas locales de proteínas.

Resultados

AlphaFold2 (AF2):

AF2 sigue dominando la competencia, aunque DeepMind no participó directamente.

- **Nuevos Desafíos**: predicción de interacciones proteína-ligand y las diferentes conformaciones que pueden adoptar algunas proteínas.

Mejoras y Nuevas Tecnologías

Modelos Basados en Energía:

- **Atom Transformer** y **GraphEBM**: Modelos que predicen conformaciones de cadenas laterales basados en energía.

Refinamiento de Estructuras:

- **EquiFold** y **ATOMRefine**: para el refinamiento de estructuras de proteínas.

3. **Modelos Generativos**:

- **DCGAN**: Uso de redes generativas adversariales para generar mapas de distancia por pares y recuperar estructuras 3D robustas.
- **Modelos Basados en Difusión**: Generación de estructuras de proteínas de alta calidad mediante modelos generativos basados en difusión.

Areas de oportunidad

1. Predicción de la Función de Proteínas

AlphaFold se enfoca principalmente en la predicción de estructuras tridimensionales. Sin embargo, la función de una proteína no siempre se deduce directamente de su estructura.

*Integrar información estructural y secuencial para predecir funciones específicas de proteínas, como actividad enzimática, interacción con ligandos o capacidad de unión a otras moléculas.

*aprovechar grandes bases de datos anotadas con funciones conocidas para mejorar las predicciones funcionales a partir de secuencias de proteínas.

2. Predicción de Efectos de Mutaciones

AF3 tiene limitaciones en la predicción del impacto de mutaciones sobre la estabilidad y la función de proteínas.

* Modelar cómo las mutaciones específicas afectan la estructura y función de las proteínas, incluyendo efectos sobre la estabilidad, la interacción con otras proteínas y la actividad enzimática.

3. Interacciones Proteína-Proteína y Complejos Multiméros

AF 3 ha mostrado avances en la predicción de complejos proteína-proteína, pero aún existen desafíos en la predicción de interacciones en grandes complejos multiméros.

5. Predicción de Estructuras en Condiciones No Nativas

pH extremos, altas concentraciones de sal, etc.

*Predecir conformaciones de proteínas en una variedad de condiciones ambientales y de estrés.

6. Interpretabilidad y Explicabilidad de Modelos