Tarea 1

Y. Sarahi García González

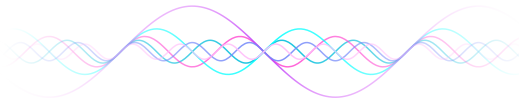
CIMAT

- Los investigadores que has encontrado que trabajan en temas cercanos y sus papers sobre el tema.

    1. Carlos Francisco Méndez-Cruz Centro de ciencias genómicas UNAM
       cmendezc@ccg.unam.mx
       Liga unam: `https://www.ccg.unam.mx/carlos-francisco-mendez-cruz/`
       Luga GS: `https://scholar.google.com/citations?user=nyQNW0gAAAAJ&hl=es`
       Papers relevantes: Fine-tuning BERT models to extract transcriptional regulatory interactions of bacteria from biomedical literature,
    2. Fabien Plisson Cinvestav Irapuaro
       fabien.plisson@cinvestav.mx
       -Sesgo de algoritmo en el diseño de péptidos guiado por aprendizaje automático.
       Liga Cinvestav: `https://portal.cinvestav.mx/ira/investigacion/directorio-de-investigacio dr-fabien-gerard-christian-plisson`
       Liga GS: `https://scholar.google.com.au/citations?user=5Shvy8wAAAAJ&hl=en`
       Papers relavantes: Machine learning-guided discovery and design of non-hemolytic peptides

- Los datasets que has encontrado y los papers donde los presenten o usen, en caso de que ya tengas una idea la conformación del dataset incluye una descripción.

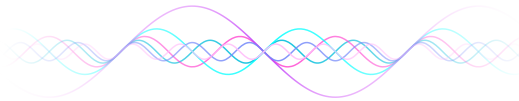
    1. Protein Data Bank: `https://www.rcsb.org`
       Recopila y proporciona acceso a estructuras tridimensionales de biomoléculas, principalmente proteínas y ácidos nucleicos.
       Entradas: Cada entrada en el PDB incluye información detallada sobre la estructura tridimensional de una molécula obtenida mediante técnicas como la cristalografía de rayos X, la resonancia magnética nuclear, y la microscopía electrónica.
       Datos Asociados: Además de las coordenadas atómicas, las entradas pueden incluir información sobre las condiciones experimentales, los métodos de determinación de la estructura, y enlaces a publicaciones científicas relevantes.
       Año:1971 y continua en crecimiento,
       Número de muestras: 220 472 Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive
       Importancia en NLP: Marco para validar y refinar las predicciones de secuencias generadas por estos modelos.
       Disponibilidad: Disponible públicamente
       Improtancia NLP: Para entrenarse en una amplia gama de secuencias



2. UniParc (Universal Protein Resource Archive) `https://www.uniprot.org/help/uniparc`
   A comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Cada secuencia recibe un Identificador de Proteína Universal (UPI) que es estable y único. Este identificador nunca se elimina, cambia ni reasignao.
   Objetivo: El objetivo principal de UniParc es rastrear la historia de cada secuencia de proteína, incluyendo las modificaciones, fusiones y fragmentaciones a lo largo del tiempo.
   Conformación: secuencias de proteínas de una variedad de bases de datos de secuencias biológicas, como UniProtKB, Ensembl, RefSeq, GenBank, DDBJ, y PDB, entre otras.
   Disponibilidad: Disponible públicamente. Cada entrada en UniParc está vinculada a sus bases de datos fuente. Buscar en UniParc es equivalente a buscar en muchas bases de datos simultáneamente.
   Etiquetas: No. Y no contiene anotaciones biológicas detalladas, para obtener información detallada sobre las funciones, localizaciones celulares y otros aspectos biológicos de una proteína, hay que referirse a las bases de datos originales mediante estas referencias cruzadas.
   Numero de muestras: 441,169,278 entradas de secuencias de proteínas. Incluye secuencias de proteínas de una amplia variedad de organismos, desde bacterias hasta plantas y animales.
3. UNIREF
   UniRef100:combina secuencias idénticas (100 %idénticas) y sus fragmentos exactos en una sola entrada con la secuencia más larga como representante.
   UniRef90:agrupa secuencias de proteínas que tienen al menos un 90 %de identidad secuencial y una longitud de alineación de al menos 80 %con la secuencia representativa más larga. Reduce la redundancia sin perder la diversidad funcional de las proteínas, siendo útil para análisis comparativos donde es importante la diversidad pero se busca evitar redundancias excesivas.
   UniRef50: UniRef50 agrupa secuencias de proteínas que tienen al menos un 50 %de identidad secuencial y una longitud de alineación de al menos 80 %con la secuencia representativa más larga. Este nivel es beneficioso para estudios que requieren un análisis de datos a gran escala y una representación no redundante de secuencias de proteínas.

- Si tienes otras referencias relevantes incluye estos papers y describe por que los ves relevantes.

1. ProteinBert: Otro tipo de Transformers. Utilizan atención pero con arquitectura distinta. Se preentrenó utilizando el conjunto de datos UniProtKB y UniRef90, que contiene aproximadamente 106 millones de secuencias de proteínas. UniRef90 es una base de datos no redundante. Esto asegura que el preentrenamiento cubra una amplia diversidad de secuencias proteicas. Se ajustó a estructura de proteínas, homología remota, modificaciones postraduccionales y propiedades biofísicas.
2. ProtTrans: Varios LLM pre-entrenados en grandes corpus de texto
3. AlphaFold 2 y 3: Estado del arte en predicción de proteinas, utilizan propiedades físicas relacionadas con los estados de menor energía. Codigo en `https://github.com/google-deepmind/alphafold`
   Datasets: `https://alphafold.ebi.ac.uk`



4. Promises of large langage models: pendiente
5. Peptide Properties with Recurrent Neural Networks: Propiedades particulares de péptidos utilizando RNN

- Paper survey