**"OligoPredict Polimeromics Interactive Dashboard"**

Creación de una herramienta accesible y práctica para investigadorescon enfoque en Bioinformática. Combinaré machine learning y visualización interactiva para elevar la utilidad del proyecto,

Enfoque en:

* **9606.protein.links.detailed.v12.0.txt.gz**: Contiene interacciones detalladas con subscores.
* **9606.protein.physical.links.detailed.v12.0.txt.gz**: Ofrece interacciones físicas detalladas.

9606.protein.physical.links.detailed.v12.0.txt.gz

**BIOGRID-ORCS-ALL-homo\_sapiens-1.1.16.screens.tar.gz** (657.96 MB), contiene datos relacionados con interacciones y características de proteínas humanas (**Homo sapiens**)

**Plan con el Archivo Actual:**

1. **Explorar y Limpiar los Datos:**
   * Revisar la estructura del archivo para seleccionar las columnas relevantes.
   * Identificar y tratar valores faltantes o inconsistentes.
2. **Integrar con los Datos Iniciales:**
   * Relacionar la información de interacciones físicas con las características estructurales del dataset original.
3. **Avanzar en el Modelo:**
   * Construir y evaluar un modelo predictivo basado en las características disponibles.

**Ventajas del formato para el proyecto:**

1. **Estructura Tabular**: Es fácil de importar y analizar con herramientas como Python (pandas), R (data.frames), o incluso SQL.
2. **Información Clave**:
   * OFFICIAL\_SYMBOL y ALIASES pueden ser usados para mapear genes.
   * SCORE.X permite análisis cuantitativos (importante para modelos predictivos).
   * HIT (YES/NO) puede servir como variable objetivo para clasificación.
3. **Especificidad del Organismo**:
   * ORGANISM\_ID y ORGANISM\_OFFICIAL son útiles si se necesita filtrar por especie.

**Posible Uso en el Proyecto:**

1. **Modelado Predictivo**:
   * SCORES y otras características podrían ser utilizadas para predecir si un gen será considerado un "HIT" (YES).
   * Análisis exploratorio de las puntuaciones para entender su relación con el HIT.
2. **Análisis de Tendencias**:
   * Ver cómo se distribuyen las puntuaciones o qué genes aparecen consistentemente como HIT.
3. **Integración con Otros Datos**:
   * Si tenemos datos adicionales (como genes específicos involucrados en defectos), podríamos unir estas tablas con claves como IDENTIFIER\_ID.

**Pasos a Seguir:**

1. **Preparación**: Revisar si los archivos tienen la misma estructura en todas las columnas.
2. **Procesamiento**:
   * Consolidar los archivos en un único dataset para análisis conjunto.
   * Limpiar valores faltantes (por ejemplo, en columnas de SCORE.X).
3. **Análisis**:
   * Identificar correlaciones entre SCORE.X y HIT.
   * Evaluar si estas columnas aportan información suficiente para modelos predictivos.
4. **Modelos**:
   * Crear modelos como Random Forest o XGBoost para clasificar o predecir HIT.