Propuesta de impartición del Módulo 3 – Retos

Presentada por: Pastor E Pérez Estigarribia,
Diego Pinto & Jose Colbes

Febrero 2023

Contenido

Inteligencia Artificial para clasificación de especies a partir de secuencias de ADN

Descripción del reto

Relevancia, novedad del reto y los desafíos involucrados

En las últimas décadas la secuenciación genética ha revolucionado la forma en la que se estudia la vida. Tan solo la versión 250.0 de GenBank, publicada en junio de 2022, contenía más de 17 billones de bases de nucleótidos en más de 2450 millones de secuencias[1]. Además, esta plataforma sigue creciendo a un ritmo explosivo a partir de envíos directos de laboratorios individuales, así como de envíos masivos de centros de secuenciación a gran escala. Hoy por hoy, extraer conocimientos de toda esta información sobrepasa la capacidad en números de científicos que se dedican a analizar estos datos.

[1]GenBank release notes (Release 250) https://ftp.ncbi.nih.gov/genbank/gbrel.txt

Por ejemplo, esta revolución genética ha impactado de forma drástica sobre el descubrimiento de nuevas especies. Sin embargo, aún ni siquiera somos capaces de responder una simple pregunta ¿Cuántas especies hay? Como resumió Robert May en un artículo publicado en Science[2]

Si alguna versión extraterrestre de Starship Enterprise visitara la Tierra, ¿cuál podría ser la primera pregunta de los visitantes? Creo que sería: "¿Cuántas formas de vida distintas, especies, tiene su planeta?" Vergonzosamente, nuestra mejor respuesta aproximada estaría en el rango de 5 a 10 millones de eucariotas (sin importar los virus y las bacterias), pero podríamos defender números superiores a 100 millones o tan bajos como 3 millones.

Tan solo responder a esta pregunta sigue siendo un desafío en la biología, pero por fortuna la generación de secuencias de ADN está ayudando en esta tarea.

Como es de esperarse los algoritmos de inteligencia artificial pueden brindar apoyo en la automatización de esta difícil tarea[3]. El desafío que te planteamos aquí es: enseñarles cómo.

Formulación del problema

Ciertos fragmentos de ADN (e.g. secuencias mitocondriales, nucleares y plástidas) se han definido como códigos de barras utilizados como marcadores para identificar y clasificar especies. Estos marcadores pueden entenderse como secuencias de cuatro letras A, C, G o T que varían entre individuos de una misma especie y entre especies. Estas secuencias en algunos casos pueden ser codificantes y por lo tanto traducibles de forma automática a secuencia de aminoácidos (secuencias de alfabetos de aproximadamente 21 letras) los cuales guardan información estructural o de propiedades fisicoquímicas.

El problema de clasificación trata sobre asignar un espécimen desconocido a una especie conocida mediante el análisis de su código de barras. Con base a esto se formula el siguiente problema:

[2] OWD: How many species are there? https://ourworldindata.org/how-many-species-are-there
[3] Supervised DNA Barcodes species classification: analysis, comparisons and results. https://doi.org/10.1186/1756-0381-7-4

Dado un conjunto de secuencias de ADN ¿Qué método de aprendizaje automático supervisado muestra mayor eficacia para clasificar especies a partir de las secuencias de código de barras de ADN?

Subdesafíos requeridos y sugeridos

1. Alinear las múltiples secuencias de ADN para establecer sitios homólogos.
2. Definir sitios homólogos en las secuencias como atributos y las especies como etiqueta de clase.
    2.1. A partir de una misma matriz de datos de secuencias alineadas se puede definir cada sitio homólogo como un atributo.
    2.2. Sin embargo, es posible definir sitios según longitudes homogéneas, por ejemplo, cada dos sitios, o cada tres sitios, o cada n-sitios como un atributo distinto.
    2.3. También es posible hacer particiones heterogéneas aleatorias reagrupando los sitios como atributos.

    Note que cuando mayor sea la longitud de palabra en sitios homólogos para un atributo definido, mayor será la cardinalidad en valores posibles (en lenguaje bioinformático: motivos) para esa variable. Esta etapa de preprocesamiento de los datos puede comprometer o mejorar el rendimiento final de los algoritmos de clasificación.
3. Buscar el mejor modelo de clasificación supervisada.
4. Seleccionar atributos y optimizar el o los modelos.
5. Evaluar y reportar el rendimiento del modelo con conjuntos de entrenamiento y prueba.
6. Evaluar el modelo inferido con un conjunto de datos de prueba final del desafío (El conjunto X).
7. Informar la clase asignada a cada muestra del conjunto X según ID preasignado.

Métrica de evaluación de resultados

La métrica de evaluación entre modelos será `cohen_kappa_score`.

```
from sklearn.metrics import cohen_kappa_score
>>> cohen_kappa_score(y_test, y_pred)
```

La mejor propuesta será aquella que consiga el mejor rendimiento de clasificación. En caso de empate en las métricas se tendrá en cuenta otros aspectos de la solución propuesta. Como por ejemplo, novedad en la propuesta, simplicidad del modelo, entre otros. Para juzgar el proyecto los participantes deben proporcionar un video en youtube explicando la implementación.

Calendario tentativo para la impartición del módulo

- 26 de junio - Presentación del problema, descripción de los datos y las métricas de evaluación. Presentación de un sistema base para la solución del reto.
- 27 y 28 de junio - Asesorías
- 29 - Fecha límite para entrega de predicciones.
- 30 de junio - Presentación de soluciones de alumnos.

Comité organizador del reto

Nombre completo y afiliación de los organizadores, con una descripción sucinta de sus intereses de investigación/académicos, áreas de especialización y experiencia en la organización de eventos similares.

1. **Pastor Enmanuel Pérez Estigarribia** *(Facultad Politécnica de la Universidad Nacional de Asunción - FPUNA, Paraguay)*: es Licenciado en Biología de la Universidad Nacional de Asunción (UNA, Paraguay), Mágister con mención en Zoología de la Universidad de Concepción (Chile) y Doctor en Ciencias de la Computación de la FPUNA. Es docente investigador de la FPUNA y sus intereses de investigación son: Bioinformática, Biología Matemática, Bioestadística, Zoología y Epidemiología Matemática.

2. **Diego Pedro Pinto Roa** *(FPUNA)*: es Ingeniero en Electrónica de la Universidad Católica de Asunción (Paraguay) y Doctor en Ciencias de la Computación de la

FPUNA. Es docente investigador y director de investigación de la FPUNA; sus áreas de interés son: Investigación de Operaciones, Optimización Combinatoria y Multiobjetivo, Algoritmos Metaheurísticos, Aprendizaje de Máquina, Redes de Telecomunicaciones.

3. **José Domingo Colbes Sanabria**: (*FPUNA*): es Ingeniero en Electrónica de la UNA, Maestro en Ciencias de la Computación del Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE, México) y Doctor en Ciencias de la Computación del CICESE. Es docente investigador de la FPUNA y sus intereses de investigación son: optimización combinatoria, predicción de estructuras de proteínas, diseño de proteínas y aprendizaje de máquina.

Requerimientos técnicos

Equipo de cómputo:

- Computadora personal o portátil.
- Conexión a Internet estable.
- Python
- Jupyter Notebook
- Bibliotecas de Python

Persona de contacto

Pastor E Pérez Estigarribia, peperez.estigarribia@pol.una.py , Cel: +595984439677