Identification of biological properties in organisms using Machine Learning techniques on whole genome sequences.

Identificación de propiedades biológicas en organismos utilizando técnicas de Machine Learning sobre secuencias de genoma completo.

Authors: Ferella, Nicolás[1,2], Pizio, Pablo Román[1,3]
Thesis director: Pons, Claudia[1,4]

[1]Facultad de Informatica
[2]Jefatura de gabinete de ministros de la Republica Argentina
[3]Nutanix Inc
[4]CAETI, Facultad de Tecnología Informática - Universidad Abierta Interamericana
nicoferella@gmail.com
piziopablo@gmail.com
cpons@lifia.info.unlp.edu.ar

Abstract

The advance in technology and genome sequencing processes in the recent decades have made large volumes of biological data available to researchers from all over the world, which, due to the large scales, are difficult to analyze in their entirety. Therefore, it is intuitive to think of Artificial Intelligence to work with such information.

In order to reduce the existing gap between the researchers and the Artificial Intelligence tools, a software was developed that allows the creation of a workspace for biological organisms, the processing of its corresponding genomes, and the creation and training of models of Machine Learning, everything using a simple (yet powerful) graphical interface.

The trained models are then analyzed to find which patterns determine the result of the property that is being investigated on the biological organism, finding in the process the genes with the greatest impact on the model's predictions, allowing the researcher to subsequently analyze the desired genes in the laboratory, saving time and resources in the process.

keywords: Artificial Intelligence, Genetics, Big Data, DNA, Machine Learning.

Identificación de propiedades biológicas en organismos utilizando técnicas de Machine Learning sobre secuencias de genoma completo.

Resumen

El avance de la tecnología y los procesos de secuenciación de genomas de las últimas décadas ha logrado poner al alcance de investigadores de todo el mundo grandes volúmenes de datos biológicos, que debido a su gran escala, los mismos resultan difíciles de analizar en su totalidad, por lo cual es intuitivo pensar en Inteligencia Artificial para trabajar con dicha información.

Con el objetivo de disminuir la brecha existente entre el investigador y las herramientas de Inteligencia Artificial, se desarrolló un software que permite crear un espacio de trabajo para un organismo biológico, realizar el procesamiento de los genomas correspondientes y permitir la creación y entrenamiento de modelos de Machine Learning desde una interfaz gráfica.

Los modelos entrenados luego se analizan para buscar qué patrones determinan el resultado de la propiedad biólogica a investigar sobre el organismo biológico en cuestión, y así encontrar los genes de mayor impacto en las predicciones del modelo, permitiendo al investigador el posterior análisis en laboratorio de un gen deseado.

Palabras claves: Inteligencia Artificial, Genética, Big Data, ADN, M.

Introducción

La salud y la enfermedad son parte integral de la vida, del proceso biológico y de las interacciones medio ambientales y sociales. Se entiende a la enfermedad como la pérdida de la salud, cuyo efecto negativo es consecuencia de una alteración estructural o funcional de un órgano a cualquier nivel. Muchas enfermedades son causadas por organismos externos, como virus y bacterias.

Para comprender la epidemiología en términos de cómo se diseminan, las características que poseen, la importancia de ciertos genes y la gravedad de las enfermedades que pueden producir, los investigadores deben realizar análisis de la estructura biológica de los organismos estudiados. Esto se logra por medio de la recolección de muestras en sujetos infectados, para su posterior aislamiento y cultivo en laboratorio, permitiendo luego realizar la secuenciación del genoma de los mismos.

Los avances de la tecnología han traído maneras más eficientes en cuanto a costos y tiempos de secuenciación de genomas; procesos que costaban miles de dólares para realizar sobre una bacteria ahora cuestan menos de cien.

A su vez, los tiempos de secuenciación han disminuido considerablemente y su precisión ha aumentado, permitiendo pasar en las últimas décadas de un reducido numeros de muestras secuenciadas a cientos de miles.

Organizaciones como GenBank, la cual es una base de datos de acceso público de secuencias de nucleótidos de más de 100.000 organismos diferentes, han puesto al alcance de investigadores alrededor del mundo grandes y variados volúmenes de datos biológicos que de otro modo les serían imposibles de recolectar, abriendo las puertas a nuevos tipos de estudios.

Es dentro de este marco que se gesta la tesina: la realización de un software para investigadores que ayude en el análisis de propiedades biológicas sobre cientos o miles de secuencias de genomas completos de un organismo mediante técnicas de Machine Learning, permitiendo realizar predicciones y encontrar los genes de mayor impacto, que en caso de ser genes no clasificados hasta la fecha, resultan de interés para su posterior análisis en laboratorio.

El desarrollo de este trabajo fue en conjunto con investigadores del Instituto Nacional de Enfermedades Infecciosas (INEI) y el Centro Nacional de Genómica y Bioinformática (CNGB) de la Administración Nacional de Laboratorios e Institutos de Salud “Dr. Carlos Malbrán” (ANLIS - Malbrán), realizando una primera investigación y caso de uso de la herramienta sobre la invasividad de los genes de la bacteria *S. pyogenes*, contenidos en 1638 genomas.

Biología

Organismos: definición biológica

En biología, un organismo es cualquier entidad individual que contiene toda propiedad necesaria para considerarse una forma de vida. Todo organismo posee la capacidad de reproducirse, crecer y desarrollarse, mantenerse, y además

poseen algun grado de respuesta a un estímulo externo. Los organismos pueden dividirse en 3 reinos: El reino de las eucariotas (organismos cuyas celulas poseen un núcleo), el reino de las arqueas y el reino de las bacterias.

Bacteria

Las bacterias son un tipo de célula biológica. De un tamaño de apenas unos micrones de largo, las bacterias están dentro de las primeras formas de vida de la Tierra. Están presentes dentro de la mayoría de los hábitats que se encuentran en la Tierra.

Varias bacterias, que van desde el Streptococcus del grupo A, *Clostridium perfringens*, *E. coli* y *S. aureus*, pueden causar una infección de tejidos blandos grave llamada fascitis necrotizante (a veces llamada bacteria carnívora), la cual afecta los tejidos que rodean los músculos, los nervios, la grasa y los vasos sanguíneos; es tratable, siempre y cuando se detecte temprano.

ADN, ARN, genoma y gen

El ácido desoxirribonucleico (ADN) es una molécula compuesta por dos cadenas de polinucleótidos que se entrelazan entre sí, formando una doble hélice que contienen las instrucciones genéticas para el desarrollo, funcionamiento, crecimiento y reproducción de todos los organismos vivos conocidos, y como también de numerosos virus. El ácido ribonucleico (ARN), es una molécula polimérica, la cual es esencial en varios roles biológicos para la codificación, decodificación, regulación y expresión de genes. Tanto el ADN como el ARN son conocidos como ácidos nucleicos los cuales, junto a las proteínas, los lípidos y los carbohidratos complejos, conforman los 4 tipos principales de macromoléculas esenciales para todo tipo de vida.

Cada molécula de ADN está compuesta por miles de copias de 4 bases específicas ricas en nitrógeno:

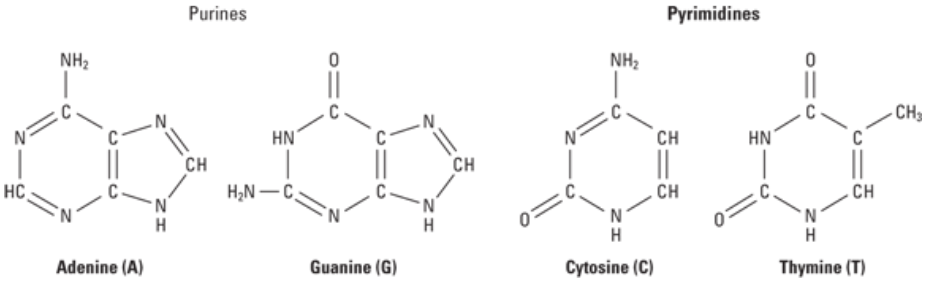

Figura 1: Las cuatro bases del ADN

De la misma manera que una secuencia de caracteres puede proveer información en forma de palabras o sentencias, la secuencia de estas bases es la que contiene el mensaje del ADN, el cual provee la información correspondiente para producir las proteínas necesarias. La secuencia total de ADN de un organismo

se denomina genoma. Un gen es una secuencia de ADN dentro del genoma que codifica un producto, el cual puede ser proteína (como es el caso de la mayoría de los genes) o puede ser ARN.

Representación Computacional del ADN

Ya comprendiendo las diferentes partes que componen el ADN, se puede observar que el mismo puede interpretarse como una cadena de bases, las cuales pueden representarse con 4 letras, una para cada posible base (A=Adenina, G=Guanina, C=Citosina, T=Timina), permitiendo que el mismo pueda ser descripto como una cadena de caracteres.

Redes Neuronales Artificiales

Las RNA constituyen uno de los tipos de modelos posibles dentro de *Machine Learning*, una rama de la *inteligencia artificial* dedicada al estudio de técnicas que permitan que algoritmos computacionales aprendan mediante la experiencia.[1, 2]

Definición

Las RNA son un modelo computacional inspiradas en el funcionamiento de las redes neuronales del cerebro humano. Su propósito es generalizar comportamientos a partir de un conjunto de datos de entrenamiento, y así obtener un modelo que permita hacer predicciones para un conjunto de datos más amplio.

Al igual que una red neuronal biológica, las RNA están formadas por un conjunto de neuronas, las cuales se conectan entre sí y se agrupan por capas. Las neuronas de cada capa reciben información de las neuronas de la capa anterior y envian información a las neuronas de la capa siguiente, formando una estructura de grafo dirigido.

Componentes

Capas

Las RNA se componen por un conjunto de capas, donde cada capa posee una o más neuronas. Se pueden distinguir tres tipos de capas: capa de entrada, capa de salida y capa oculta.

Neuronas y conexiones

Una neurona artificial es la unidad básica de procesamiento dentro de una RNA. Al igual que una neurona biológica recibe una serie de impulsos nerviosos, una neurona artificial recibe valores de entrada con los cuales realiza un cálculo y genera un valor de salida. Una neurona artificial es en realidad una función matemática.

Aprendizaje

Una RNA debe tener los pesos de las conexiones con ciertos valores para que resuelva el problema para el cual es modelada. La parte interesante de estos algoritmos es que encuentren por si solos cuáles son los valores adecuados para estos parámetros, este proceso es lo que se denomina *aprendizaje*.

Proceso de aprendizaje

La técnica de aprendizaje utilizada en este trabajo es el aprendizaje supervisado. Este proceso consiste en encontrar cuáles son los valores que deben tener los pesos de las conexiones de la red para que la misma tenga el comportamiento deseado.

Conjunto de datos

Para realizar la técnica de aprendizaje supervisado es necesario disponer de un conjunto de datos donde cada dato esté etiquetado con el resultado esperado. El objetivo es que la RNA encuentre patrones para cada posible etiqueta del conjunto de datos, esto requiere que el conjunto sea grande, ya que si son pocos los datos sobre los cuales el modelo realiza el entrenamiento, se dificulta la tarea de identificar patrones. Que un conjunto tenga pocos datos o la cantidad necesaria depende de la complejidad del problema a resolver.

Caso de uso

El *Streptococcus Pyogenes* ó estreptococo del grupo A (Group A beta-hemolytic Streptococcus, GAS) es una bacteria causante de numerosas enfermedades, entre ellas, la faringoamigdalitis, la escarlatina y el impétigo. El estudio global de enfermedades realizado en el 2010 por The Lancet en conjunto con la Organizacion Mundial de la Salud (WHO) estimó que hay 140.495.000 casos de impétigo en todo el mundo cada año. Esta carga infecciosa lo coloca entre las 50 enfermedades más comunes en todo el mundo. Las infecciones cutáneas por *S. pyogenes* es una de las causas más importantes de morbilidad en entornos con recursos limitados[3].

El GAS también puede provocar enfermedades invasivas graves, como la fascitis necrotizante y el síndrome de shock tóxico estreptocócico, las cuales presentan elevadas tasas de mortalidad, con un estimado de 517.000 muertes anuales a nivel global[4].

Para comprender completamente su epidemiología en términos de cómo se disemina, las características de la cepa, su importancia para la transmisión y la gravedad de la enfermedad que puede producir, los investigadores deben realizar análisis de la estructura biológica de estas bacterias. Esto se logra por medio de la recolección de muestras de personas infectadas para el posterior aislamiento y cultivo en laboratorio, permitiendo luego realizar una secuenciación de la estructura genómica de la cepa.

Descarga y preparación de datos

Para la creación del dataset se comenzó con la búsqueda, identificación y descarga de genomas en formato FASTQ dentro numerosos estudios publicados en la base internacional de datos de secuencias de nucleótidos, llegando a mas de 1600 descargas con un total de 300Gbs en el transcurso de 3 meses. La búsqueda e identificación se realizó en conjunto con investigadores del Centro Nacional de Genómica y Bioinformática (CNGB).

Los archivos FASTQ pueden contener hasta millones de entradas y pueden tener desde varios megabytes, hasta gigabytes de tamaño, haciéndolos demasiado pesados para ser trabajados en grandes sets, por lo que se consideran archivos de salida intermedios, los cuales se utilizan como entrada para herramientas que realizan análisis posteriores. Estos archivos están compuestos por múltiples lecturas de células cultivadas de la muestra biológica a analizar, por lo que el mismo contendrá "trozos" de estas lecturas. Esto permite tener redundancia en las lecturas, la cual es una manera de asegurar o puntuar la calidad de los datos obtenidos, al costo de sumar complejidad a la hora de analizarlo como un tipo de dato homogéneo y continuo. Otro desafío que presentan este tipo de archivos es que su estructura variará dependiendo del sistema y los parámetros que han sido utilizados para realizar la secuenciación de las muestras. Esto nos llevó a buscar realizar un proceso de estandarización de los archivos, para lo cual utilizamos diferentes programas con la supervisión de los científicos del CNGB, para convertir los archivos a formatos con menor redundancia de datos y mayor contenido de metadata.

Se seleccionaron 1638 muestras de *S. pyogenes*, 819 consideradas invasivas debido a los síntomas generados en cada uno de los sujetos de los cuales se extrajo la muestra, y 819 no invasivas. Estas muestras fueron procesadas con los programas descriptos anteriormente para obtener su correspondiente archivo GBF.

Parseo de datos y creación de DB

Ya teniendo un archivo GBF de cada muestra se procedió a analizar la información de los genes de cada una de ellas.

La cantidad de genes encontrados en cada uno de los genomas analizados fue entre 2500 y 3800. Una vez obtenidos los genes de cada muestra, se procedió a realizar un pangenoma[1] de las muestras del estudio. Se obtuvieron más de 158.000 genes diferentes.

[1] El pangenoma (o supragenoma) en biología molecular describe la colección de todos los genes en una especie (aplicado típicamente a bacterias y arqueas, que pueden presentar una gran variación de contenido genético entre cepas estrechamente relacionadas). Es un superconjunto de todos los genes de todas las cepas de una especie

Red neuronal: Creación y entrenamiento

Creación de la red neuronal

Utilizando el pangenoma generado de todas las muestras del estudio, se procedió a crear una matriz de pertenencia de genes para cada una de las muestras del estudio el cual, junto a la propiedad de invasividad de cada una de ellas, fueron utilizadas en el entrenamiento de una red neuronal.

Se comenzaron las primeras pruebas de entrenamiento utilizando una taza de aprendizaje de 0.005 y una cantidad de 5 a 10 capas secuenciales compuestas por una reducción escalonada de neuronas desde 8072 hasta 1.

Debido a la cantidad de datos de entrada (arreglos de más de 158.579 elementos para cada una de las 1638 muestras), los tiempos de ejecución para el aprendizaje del modelo resultaban muy elevados, haciéndolos pocos eficientes tanto en tiempo como en recursos. Para mejorar los tiempos de aprendizaje, se buscó utilizar el concepto de alelo[2] en las muestras de entradas. Para ello, utilizando el concepto de *extracción de características* descripto en el capítulo 3, se procedió a identificar los genes que producían la misma proteína como resultado, generando con estos un nuevo arreglo de pertenencia para cada una de las muestras, reduciendo asi los datos de entradas de cada una de ellas de 158.579 a 82.520.

A su vez, como un segundo enfoque, y para buscar mejorar aun más los tiempos de aprendizaje, se utilizó el concepto de selección de características descripto en el capítulo 3. Para ello, se examinaron los datos de entrada, analizando la distribución de genes dentro de las muestras. Durante este proceso, se detectó que muchos de estos alelos podían encontrarse en igual cantidad de muestras invasivas y no invasivas, haciendo que los mismos no tengan un aporte significativo al aprendizaje. Definiendo un nuevo parámetro de configuración, se utilizó un porcentaje de tolerancia máximo en la diferencia de pertenencia entre muestras invasivas y no invasivas. Utilizando un 1 % de tolerancia (diferencia de 16 muestras en nuestro caso de estudio), se redujo el tamaño de cada dato de entrada de 82.520 a 5.218 genes.

En cuanto a la función de activación para las diferentes capas, se obtuvieron los mismos porcentajes de precisión tanto en el uso de las funciones sigmoid como RelU, siendo esta última más eficiente en terminos de tiempo de ejecución.

Modelos

A continuación se describen los resultados obtenidos de 5 modelos diferentes en el análisis de las muestras de nuestro caso de uso. La cantidad de muestras utilizadas para todos los modelos fue de 1638 genomas, realizando un total de 500 épocas en cada caso.

[2]Un alelo es una variación de la misma secuencia de nucleótidos que codifica la síntesis de un producto génico en el mismo lugar de una molécula de ADN. La mayoría de los alelos producen poco o ningún cambio en la función del producto génico que codifica.

Model\Acc	**100 épocas**	**250 épocas**	**+500 épocas**	**tiempo de entr. +500 épocas**
Modelo 1	0.7488	0.7668	0.7668	297:57 min.
Modelo 2	0.7578	0.7764	0.7795	147:47min.
Modelo 3	0.6708	0.7236	0.7240	51:18 min.
Modelo 4	0.7298	0.7453	0.7578	197:43 min.
Modelo 5	0.7391	0.7547	0.7871	123:17 min.

Con las pruebas realizadas no se observó impacto significativo en el porcentaje de aciertos entre genes y alelos, viendo una reducción significativa en los tiempos de entrenamiento en estos ultimos (entre un 25 % y 50 % más rápidos). Por el ajuste de diferencia de muestras, un porcentaje de 1 % demostró tener los resultados más prometedores, mientras que valores mayores (Modelo 3, 2 %) devolvió los resultados de aciertos más bajos, aunque su entrenamiento resultó en tiempos de ejecuciones menores a los demás modelos.

Desarrollo de la herramienta

Habiendo creado y entrenado un modelo de inteligencia artificial que puede evaluar si el genoma de una muestra de una bacteria de *S. Pyogenes* es invasiva o no, se puede notar que los datos que recibe como entrada el modelo es una matriz de números 1 y 0, con un tamaño igual a la cantidad de genes que se encuentran en las muestras utilizadas en el entrenamiento, donde cada elemento de la matriz se vincula con un gen y tiene como valor 1 si la muestra tiene ese gen y 0 en caso contrario. El valor de salida del modelo es un valor binario, 1 si la muestra es invasiva y 0 si no lo es.

Si nos abstraemos de la parte biológica, se puede ver que los patrones que busca el modelo para que el resultado sea 1 o 0, se determinan según el valor que tiene cada elemento de una matriz de unos y ceros.

Entonces surge la pregunta, se puede utilizar el modelo para problemas con la misma estructura? Es decir problemas que respeten la estructura de los datos de entrada y los de salida. No importa la bacteria o el virus que se quiera tratar, mientras se pueda generar para cada muestra una matriz de pertenencia de los genes que se encuentran en todas las muestras, y no importa la pregunta binaria que se quiera preguntar, mientras que el valor de las etiquetas para cada muestra sea 1 o 0.

Resulta interesante desarrollar una herramienta que permita realizar un flujo como el que se explicó en el capítulo anterior, desde el procesamiento de los archivos utilizando programas de bioinformática ya existentes hasta el entrenamiento de un modelo y la obtención de las estadísticas del impacto de los genes en ese modelo, para cualquier bacteria o virus y cualquier pregunta binaria que se quiera evaluar, por medio de una interfaz gráfica amigable para no informáticos.

Arquitectura y tecnologías elegidas

Arquitectura

La herramienta se plantea como un sistema web. Una interfaz gráfica web que interactúa con una API REST, la cual dispone de dos bases de datos, una en PostgreSQL para el manejo de usuarios y sesiones y la otra no relacional en ElasticSearch para el almacenamiento de la gran cantidad de datos que implican los genomas. En total son cuatro servicios, y la arquitectura es la siguiente:

- Interfaz gráfica desarrollada con ReactJS.
- API RESTful desarrollada con Django.
- Base de datos en PostgreSQL.
- Base de datos no relacional en ElasticSearch.

Tecnologías

- ReactJS es una librería para JavaScript para desarrollar SPA (single page applications), y la elección de la misma se debe a que es una librería simple de usar, eficiente y muy usada a nivel global. Las SPA son una buena opción cuando en la arquitectura de un sistema existe una API RESTful, ya que el usuario carga toda la interfaz con una sola petición al servidor web, y luego el resto de la interacciones se realizan con la API RESTful para el intercambio de datos y el flujo de la aplicación. ReactJS además se puede integrar con Material Design, una librería creada por Google para diseñar interfaces gráficas siguiendo una línea de buenas prácticas de diseño.
- Django es un framework para Python. Más precisamente en este sistema se utiliza DRF (Django Rest Framework), un framework para desarrollar API's RESTful. También es simple de usar y uno de los frameworks más utilizados. Django cuenta con un ORM (object–relational mapping) para el manejo de bases de datos a alto nivel, con la opción de elegir entre varios motores de bases de datos, en este caso se eligió PostgreSQL para gestionar los usuarios y las sesiones.
- Como se observa en el capítulo anterior, el objetivo de este trabajo implica procesar y almacenar una gran cantidad de información, por eso resulta conveniente usar una base de datos no relacional como ElasticSearch, un motor de base de datos adecuado para realizar consultas sobre volúmenes de información con un tamaño grande como ocurre en este caso.
- Para la creación, entrenamiento y evaluación de las redes neuronales se utiliza la misma librería que en el capítulo anterior, Tensorflow para Python.
- Otra tecnología que se decide incorporar es Docker, la cual permite empaquetar una aplicación y sus dependencias en un contenedor. De esta

manera, la herramienta se puede ejecutar en cualquier servidor que tenga instalado Docker.

Desarrollo

Proyectos

Como se mencionó previamente, un proyecto es el espacio de trabajo que se asocia a un organismo biológico, el cual se crea para procesar los archivos FASTQ, FASTA y GBF vinculados al mismo, y también crear los estudios deseados.

Estados del proyecto

El proyecto tiene 5 estados posibles:

- Inicial: es el estado en el cual se crea el proyecto, y es el estado al que vuelve el proyecto luego de procesar archivos si el estado previo al procesamiento era **Inicial**.
- Procesando FASTQ: es el estado al que pasa el proyecto cuando se procesan archivos FASTQ, para poder procesar archivos FASTQ el estado del proyecto debe ser **Inicial** o **Listo para estudios**. Al terminar el procesamiento y convertir los archivos FASTQ a FASTA, vuelve al estado previo del procesamiento.
- Procesando FASTA: es el estado al que pasa el proyecto cuando se procesan archivos FASTA, para poder procesar archivos FASTA el estado del proyecto debe ser **Inicial** o **Listo para estudios**. Al terminar el procesamiento y convertir los archivos FASTA a GBF, vuelve al estado previo del procesamiento.
- Procesando GBF: es el estado al que pasa el proyecto cuando se procesan archivos GBF, para poder procesar archivos GBF el estado del proyecto debe ser **Inicial** o **Listo para estudios**. Al terminar el procesamiento de los archivos GBF y crear las matrices de pertenencia de los genomas, pasa al estado **Listo para estudios**.
- Listo para estudios: es el estado al que pasa el proyecto luego de procesar archivos GBF, y es el estado que habilita acciones sobre los estudios del proyecto.

Procesamiento de archivos

Esta herramienta requiere el procesamiento de tres tipos de archivo: el de archivos FASTQ, que se realiza con el programa *spades*, quien convierte los archivos a FASTA. El procesamiento de archivos FASTA, que se hace con *prokka*, convirtiendo los mismos a GBF. Y por último, el procesamiento de archivos GBF, que funciona de la siguiente manera:

1. Por cada archivo GBF que se encuentre en la carpeta del proyecto:
    a. Se crea un JSON con el nombre de la muestra y una lista vacía para ir agregando los genes.
    b. Por cada gen que se encuentra en el archivo se lo agrega a la lista guardando: su nombre, el producto, locus_tag y translation (cadena de aminoácidos del gen).
    c. Cuando se termina de procesar el archivo, se guarda el JSON en la base de datos ElasticSearch, en el índice de genomas del proyecto.
2. Luego de insertar los nuevos genomas, se obtienen todos los genomas del proyecto (los recién insertados y los ya existentes) para obtener todos los genes y armar dos conjuntos de genes:
    a. Un conjunto eligiendo como atributo único la cadena de aminoácidos del gen, cuyos genes se insertan en el índice de genes del proyecto. A cada gen se le agrega un atributo id numérico que se utiliza para identificar el gen dentro del índice.
    b. Otro conjunto eligiendo como atributo único el nombre del gen, para agrupar en este caso las distintas cadenas de aminoácidos posibles (alelos) para cada gen, cuyos genes se insertan en el índice de alelos del proyecto. A cada gen se le agrega un atributo id numérico que se utiliza para identificar el gen dentro del índice.
3. Una vez creados los índices de genes y de alelos, se pueden generar las matrices de pertenencia de los genomas, donde para cada genoma se crean dos matrices: la matriz de pertenencia de genes y la matriz de pertenencia de alelos. El objetivo es agregar a estas matrices los identificadores de los genes que contenga el genoma. Cada matriz se corresponde con un índice distinto, en la matriz de genes se agregan los identificadores de los genes del índice de genes y en la matriz de alelos los identificadores del índice de alelos.

Estudios

Como se mencionó previamente, un estudio es la pregunta binaria que se quiere trabajar sobre un organismo, por ejemplo, aplicando el mismo caso descripto en el capítulo anterior: si el objetivo es estudiar la invasividad de la bacteria *S. pyogenes*, se crea un proyecto para dicha bacteria y dentro del mismo se crea el estudio **Invasividad**. Al momento de crear un estudio se debe elegir si se desea trabajar con genes o alelos.

El propósito de crear un estudio es el de entrenar una red neuronal con las muestras obtenidas de la bacteria a tratar, para que detecte los patrones que determinan el resultado de la pregunta binaria que se quiere investigar.

La carpeta **/keras_model** se utiliza para guardar el modelo y toda la estructura de archivos y carpetas que requiere Keras para entrenar la red neuronal.

En **/training_logs** se almacenan todos los logs de los entrenamientos y optimizaciones que se realizan sobre el modelo, además de los análisis que se hagan sobre el modelo con la librería Shap.

La carpeta **/genomes_to_evaluate** es donde se agregan los archivos FASTQ, FASTA o GBF que se quieran evaluar con la red neuronal, y dentro de la misma se encuentran los subdirectorios necesarios para el procesamiento de los archivos de la misma manera que ocurre en el proyecto. En la carpeta **/results** se guardan los resultados de las evaluaciones de los genomas por parte de la red neuronal.

Por último, **/multifasta_muscle_tree** se usa al crear archivos multifasta, para alinear archivos multifasta y para generar archivos IQTree.

Entrenamiento de redes neuronales y evaluación

Con el estudio creado, lo que se necesita para poder crear la red neuronal y empezar con el entrenamiento es etiquetar los genomas del proyecto al que pertenece el estudio.

Para etiquetar las muestras, la herramienta permite exportar un CSV que cuenta con dos columnas: una columna con el nombre de cada muestra perteneciente al proyecto y la otra es una columna vacía para completar con el valor de cada etiqueta. Luego este CSV se puede cargar para que las muestras sean etiquetadas.

Lo que implica esta carga de etiquetas es agregar un atributo con el nombre del estudio y el valor de la etiqueta, en cada genoma indicado en el CSV en el índice de genomas del proyecto. Entonces, el JSON que representa a un genoma en ElasticSearch, luego de agregar la etiqueta para el estudio **Invasividad** con etiqueta **1**, queda asi:

```
{
    "name": "ERR227094"
    "genes" : [
        {
            "product" : "Maltodextrin phosphorylase",
            "gene" : "malP",
            "translation" : "MTRFTEYVETKLGKSLTQAS...",
            "locus_tag" : "JJOEMFFN_00006"
        },
        {
            "product" : "hypothetical protein",
            "gene" : "",
            "translation" : "MTKKHLLTLLLISFFTSFLV...",
            "locus_tag" : "JJOEMFFN_00007"
        },
        ...
    ],
    "gen_pertenence" : [258, 1762, 1763, ...],
    "alleles_pertenence" : [258, 1659, 1660, ...],
    "invasividad": 1
}
```

Extracto 1: Ejemplo de JSON para un genoma etiquetado

Etiquetar las muestras habilita el entrenamiento de la red neuronal. El entrenamiento admite varios parámetros para realizar distintas pruebas y llegar a la combinación que genere los mejores resultados de tasa de acierto y error. Con estos parámetros definidos, la herramienta tiene todo definido para realizar el entrenamiento.

Una vez que termina el primer entrenamiento de la red de un estudio, se puede seguir optimizando el modelo, es decir, realizar más pasadas con los mismos valores que se habían definido para los parámetros en el entrenamiento, para que la misma mejore su tasa de acierto y reduzca el error.

Por otro lado, también es posible en un estudio volver a realizar un entrenamiento con otros valores en los parámetros, pero esto implica crear una nueva red y perder la anterior.

Estadísticas pos entrenamiento e impacto de genes

La última sección que tiene un estudio es el análisis del modelo entrenado. Algo interesante de tener una red neuronal que reconoce qué patrones hacen que un genoma dé un resultado u otro, que para el ejemplo en cuestión sería invasiva o no, es poder ver cuáles son esos patrones, es decir, cuáles son las neuronas (genes) o combinaciones de neuronas que tienen mayor impacto en el resultado.

Luego de realizar un primer entrenamiento para un modelo de un estudio, se habilita la acción para analizar dicho modelo. Este análisis es un proceso que se lleva a cabo utilizando la librería **SHAP (SHapley Additive exPlanations)**, una librería para Python que se usa para explicar las predicciones de modelos

de *Machine Learning*. La misma contiene funciones integradas que hacen uso del método **Shapley value**[5], el cual pretende determinar cuál es el aporte que brinda cada participante de un grupo a un resultado final.

El funcionamiento de este método consiste en obtener el Shapley value para cada feature de los datos de entrada de un modelo (en este caso, para cada gen de los datos de entrada) y así determinar qué impacto tiene cada gen sobre las predicciones del modelo.

En resumen, el Shapley value de un gen es el aporte promedio que el gen brinda al modelo a través de todas las posibles combinaciones de genes.

Luego de obtener el Shapley value de cada gen, el análisis se queda con los 50 genes con mayor impacto y de esa forma genera en la interfaz:

- Un gráfico donde se aprecia el impacto de cada uno de los 50 genes.
- Un listado para detallar de cada gen su id, su nombre, su product, la cantidad de genomas en cuales está presente, cuántos de esos genomas están etiquetados con SI y cuántos con NO, y por último el conjunto de translations (alelos) obtenidos para ese gen. Además por cada gen del listado se permiten cuatro acciones.

Las cuatro acciones que se pueden ejecutar son las siguientes:

- Generar multifasta: genera un archivo que contiene todos los translations (alelos) del gen en un formato FASTA. Esto quiere decir, que por cada translation crea dos líneas: la primera como cabecera, la cual empieza con '>', proporciona un nombre/identificador único a la secuencia, y en la segunda el translation.
- Alinear multifasta: sirve para alinear un archivo multifasta, o sea todos los translations que contiene un multifasta, para comparar las diferencias entre los aminoácidos de cada cadena.
- Generar IQtree: genera un archivo para ser visualizado por el programa IQtree, este mismo se genera a partir del alineamiento del multifasta. Con el IQtree se puede observar un árbol que explica los distintos conjuntos en los cuales se clasifican los translations de un gen.
- Detalle del gen en XLSX: genera un archivo Excel con dos hojas:
    1. Un listado de los translations del gen, donde por cada uno detalla en cuántos genomas está presente y cuántos de ellos fueron etiquetados con SI y cuántos con NO.
    2. Un listado de los genomas al cual pertenece al gen, donde por cada uno se detalla la etiqueta de ese genoma y el translation del gen que contiene dicho genoma.

Con estas cuatro acciones disponibles para los 50 genes de mayor impacto en las predicciones de una red neuronal, los usuarios pueden obtener información más detallada y estadísticas sobre genes que resultan de mayor interés sobre el resto de los genes, y que si lo requieren, lleguen a un análisis en laboratorio.

Conclusiones y trabajo futuro

Se logró realizar un software para investigadores que ayuda en el análisis de propiedades biológicas sobre miles de secuencias de genomas completos de un organismo mediante técnicas de Machine Learning, el cual permite realizar predicciones y encontrar los genes de mayor impacto, que en caso de ser genes no clasificados hasta la fecha, resultan de interés para su posterior análisis en laboratorio.

La herramienta reduce la brecha entre las dos ciencias, al permitir analizar grandes volúmenes de información mediante una interfaz amigable a biólogos con poca experiencia en IA.

A su vez, gracias al trabajo en conjunto con investigadores del Centro Nacional de Genómica y Bioinformática (CNGB), se extendió la funcionalidad del software para agilizar el trabajo del día a día de los científicos, como son la generación de archivos multifasta, el alineamiento de los mismos y la generación de árboles IQTREE sobre los genes de mayor impacto en un estudio.

Se pudo además encontrar soluciones informáticas a problemas técnicos reales, partiendo del entendimiento del dominio de la biología con ayuda de especialistas en el tema, para poder procesar información cruda de gran tamaño y complejidad y así estructurarla teniendo presente la resolución del objetivo de la tesina.

Se consiguió desarrollar un software simple de distribuir, pudiendo ser ejecutado en cualquier servidor que disponga de Docker. Esto brinda portabilidad y pocos pasos para lograr su instalación y uso. El desarrollo fue realizado enteramente en Gitlab, permitiendo el uso y extensión del mismo por parte de otros usuarios o desarrolladores que así lo deseen.

Se realizó la instalación de la herramienta en el CNGB para el uso por parte de los científicos, con el objetivo de seguir trabajando en conjunto y así mejorar y utilizar el software en nuevos tipos de organismos biológicos.

Trabajo a futuro

- Extender la plataforma para poder indicar una lista de genomas hosteados en la nube y así poder utilizar datos públicos que se encuentren en diferentes portales (ej. NCBI).
- Extender el funcionamiento aprovechando los beneficios de Docker para distribuir el entrenamiento a nivel hardware, para reducir los tiempos de ejecución considerablemente.

- Extender el comportamiento de las redes neuronales que se crean en los estudios para permitir predicciones sobre preguntas no binarias, es decir, más de dos etiquetas posibles para las muestras, dando lugar a la investigación de propiedades biológicas de organismos que requieren una clasificación más compleja.

Bibliografía

[1] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. `http://www.deeplearningbook.org`.

[2] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 ed., 2010.

[3] A. S. Sanyahumbi, S. Colquhoun, R. Wyber, and J. Carapetis, "Global disease burden of group a streptococcus," *National Center for Biotechnology Information*, 2 2016.

[4] J. Carapetis, A. Steer, E. Mulholland, and M. Weber, "The global burden of group a streptococcal diseases," *The Lancet infectious diseases*, vol. 5, pp. 685–94, 12 2005.

[5] R. J. Aumann and L. S. Shapley, *Values of non-atomic games*. Princeton Legacy Library, Princeton, NJ: Princeton University Press, Apr. 2016.