Módulo 3 Retos - Bionformática

Presentado por: André Borges Farias, Maria Carolina del Valle Sisco y Edgardo Galán Vásquez

Junio 2025

Clasificación de sitios de unión a factores de transcripción de bacterias

La expresión genética es el proceso mediante el cual la información almacenada en el ADN es decodificada a un producto funcional, como son las proteínas. Este proceso permite determinar el conjunto de genes que se expresan en una condición específica, por lo cual es altamente complejo y regulado.

Esta regulación está principalmente mediada por proteínas reguladoras llamadas factores de transcripción. Los cuales desempeñan un papel fundamental en la regulación del inicio de la transcripción de los genes de ADN a ARN mensajero, activando o inhibiendo la expresión de sus genes regulados. Esta regulación se da por medio de la interacción física entre el factor de transcripción y una región específica del ADN denominada sitio de unión del factor de transcripción (TFBS por sus siglas en inglés), la cual se encuentra en la región promotora de los genes.

La identificación de estas TFBS es un problema abierto debido a que tienden a ser secuencias de AND cortas que van desde los 10 a 20 pares de bases y que tienden a variar en su contenido de nucleótidos. Sin embargo, la identificación de estos sitios es crucial para comprender los mecanismos de regulación celular, que son un paso crítico en el desarrollo de estrategías para comprender cómo los organismos responden a estímulos intra- y extracelularmente.

La identificación de estos sitios por medio de matrices de peso, que es una de la estrategías más utilizadas, tiende a tener un gran número de falsos positivos. Por lo que, el uso de datos públicos combinado con técnicas de inteligencia artificial surge como un enfoque prometedor para predecir estos sitios de unión. De manera, que este curso tiene como objetivo capacitar a los estudiantes para trabajar con un conjunto de datos de sitios de unión a factores de transcripción reales, que implica el preprocesamiento, construcción y validación de modelos predictivos aplicados a la identificación de sitios de unión factores de transcripción.

Objetivo

1. Presentar los principales conceptos relacionados con la regulación genética mediada por interacciones entre proteínas y ADN.
2. Estimular el pensamiento crítico con respecto a los pasos de preprocesamiento y curación de conjuntos de datos biológicos.

3. Alentar a los estudiantes a explorar diferentes arquitecturas de modelos de inteligencia artificial, incluidos enfoques interpretativos que permitan comprender los mecanismos reguladores.
4. Desarrollar habilidades prácticas en la construcción, entrenamiento y validación de modelos basados en inteligencia artificial aplicados a la biología molecular.

Desafío

Se proporcionará un conjunto de datos biológicos compuesto por secuencias de nucleótidos obtenidos de varias bases de datos públicas, como RegulonDB, CollecTF y JASPAR, los cuales han sido previamente curadas por el grupo del taller. Las secuencias están organizadas en formato FASTA, separadas por familias de factores de transcripción y de longitudes diferentes.

Los estudiantes deben organizar estas secuencias de manera adecuada, así como implementar estrategías para manejar valores faltantes, convertir las secuencias de nucleótidos en representaciones numéricas adecuadas para el entrenamiento de los modelos. Otro aspecto importante será la definición de una estrategía para construir el conjunto de datos negativo, es decir, secuencias que no representan sitios de unión de factores de transcripción (no-TFBS).

Al finalizar el desafío, se espera que los estudiantes puedan:

- Analizar críticamente los pasos de preprocesamiento de datos y justificar la elección del enfoque utilizando para entrenar modelos de inteligencia artificial.
- Implementar y representar los algoritmos utilizados para convertir secuencias de nucleótidos en variables numéricas.
- Presentar e interpretar las métricas de validación obtenidas tanto en el conjunto de pruebas como en la validación externa; comprobando los resultados con modelos previamente publicados en la literatura.
- Realizar un estudio de caso basado en la predicción de nuevas secuencias proporcionadas por el equipo del curso.
- Elaborar un informe técnico que describa la metodología utilizada y los resultados obtenidos en la predicción de sitios de unión de factores de transcripción.

Estudio de caso

Como parte del desafío final, a los estudiantes se les proporcionará un conjunto adicional de aproximadamente 30 secuencias de nucleótidos, seleccionadas para simular un escenario realista de predicción y clasificación de sitios de unión de factores de transcripción (TFBS). Este conjunto incluirá:

- Secuencias de TFBS de organismos bacterianos.
- Secuencias de humanos, que representan posibles contaminaciones biológicas o ruido experimental.

- Secuencias aleatorias, no asociadas con ningún factor de transcripción conocido.

El estudio de caso tiene dos propósitos:

- Clasificación. Los estudiantes deben aplicar los modelos desarrollados a lo largo del curso para predecir si cada secuencia representa o no un sitio de unión de un factor de transcripción.
- Caracterización. Para las secuencias identificadas como TFBS, los estudiantes deberían indicar, siempre que sea posible, la familia de factores de transcripción correspondiente, basándo se en los datos utilizados para el entrenamiento y/o herramientas de análisis complementarias.

El estudio de caso representa una oportunidad para que los participantes prueben el modelo construido en una situación práctica.

Evaluación

Dado el conjunto de secuencias de ADN, el modelo debe predecir si una secuencia es o no un sitio de unión a un factor de transcripción (TFBS) y el tipo de familia de factor de transcripción al que pertenece.

- Métricas de evaluación
    - Precisión, Recall y F1-score
    - Área bajo la curva ROC
- Interpretabilidad del modelo
- Justificación del diseño del modelo
- Reporte

Público objetivo

Este curso está dirigido a estudiantes de pregrado y posgrado, así como a profesionales e investigadores en las áreas biológicas, bioinformáticas, ciencia de datos y áreas afines, que estén interesadas en aplicar técnicas de inteligencia artificial al análisis de datos biológicos. La propuesta es especialmente relevante para aquellos que desean explorar los mecanismos de regulación genética basados en la identificación de sitios de unión de factores de transcripción, combinando conocimientos de biología molecular y modelado computacional.

Prerrequisitos

Idealmente los participantes deberían haber tomado el módulo 2 de Bioinformática impartido en el MeIA.

Recursos necesarios

Para el seguimiento y desarrollo de las actividades propuestas, los participantes deberán disponer de un ordenador con conexión a internet. Se recomienda utilizar un entorno de programación de Python, con acceso a las principales librería de código abierto, como:

- Pandas
- Seaborn
- Scikit-learn
- Tensorflow o Keras

No es necesario contar con equipos de alto rendimiento, pero se sugiere que las computadoras tengan capacidades adecuadas para ejecutar scripts de procesamiento y entrenamiento del modelo en un tiempo razonable.

Listado de integrantes

En caso de seleccionar la modalidad por equipo, indicando quién es el representante del equipo.

○ Nombre completo y Apellidos: Edgardo Galán Vásquez
○ Universidad de adscripción: Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la Universidad Nacional Autónoma de México
○ Grado académico: Doctorado
○ Área de especialidad: Bioinformática y Biología de sistemas
○ Correo electrónico: edgardo.galan@iimas.unam.mx

○ Nombre completo y apellidos: André Borges Farias
○ Universidad de adscripción: Laboratório de Bioinformática - Laboratório Nacional de Computação Científica
○ Grado académico: PhD
○ Área de especialidad: Modelado Molecular de sistemas biológicos

○ Nombre completo y apellidos: Maria Carolina del Valle Sisco Zerpa

- Universidad de adscripción: Laboratório de Bioinformática - Laboratório Nacional de Computação Científica
- Grado académico. PhD
- Área de especialidad. Microbiología y Bioinformática