Clasificación de Péptidos Antimicrobianos
Universidad Nacional de Colombia

En las últimas décadas, la resistencia bacteriana se ha convertido en una de las problemáticas más graves a nivel mundial. Un agravante de esta problemática es que existe un declive en la búsqueda y desarrollo de nuevas moléculas antimicrobianas que puedan ser usadas contra bacterias resistentes a los antibióticos actuales. En parte, la disminución del interés en la investigación y desarrollo de nuevos antibióticos se debe tanto al costo y el tiempo que requieren las pruebas en laboratorio necesarias en la generación de nuevas moléculas antimicrobianas, como al escaso retorno de la inversión que esta representa para las empresas farmacéuticas.

Como parte de solución a la problemática, los péptidos antimicrobianos (o AMPs, de su sigla en inglés *Antimicrobial Peptides*) han tomado importancia en el desarrollo de nuevos antibióticos por su rol como agente inhibidor, no solo de bacterias sino también de virus, hongos y parásitos, entre otros (Porto et al., 2017; Yoshida et al., 2018). Los AMPs son parte esencial de todos los organismos vivos y configuran la primera línea de defensa contra bacterias, microbios y parásitos. Este tipo de péptidos causan la muerte de las bacterias y los microbios, bien sea interfiriendo las funcionalidades de la membrana celular o interrumpiendo sus funciones intracelulares (Brogden, 2005; Yeaman & Yount, 2003). De ahí la importancia que tienen el determinar la capacidad antimicrobiana de un péptido.

Si bien la identificación de péptidos antimicrobianos se ha desarrollado de manera manual, en los últimos años se ha venido realizando importantes avances en lo que respecta al uso de los algoritmos de inteligencia artificial. Esto puede tener un impacto positivo en el contexto farmacéutico dado que el uso de algoritmos de aprendizaje automático puede reducir los costos y el tiempo empleados en el proceso de búsqueda y diseño de nuevas biomoléculas sintéticas para la producción de antibióticos.

Lo anterior, ha motivado a que los grupos de investigación en Biología Funcional e Investigación y Desarrollo en Inteligencia Artificial (GIDIA), ambos de Universidad Nacional, sede Medellín, y el grupo de investigación en Automática Electrónica y Ciencias Computacionales (AEyCC) del ITM, unan sus esfuerzos en la búsqueda y construcción de soluciones computacionales que apoyen la investigación y desarrollo de nuevas moléculas antimicrobianas.

Como resultado parcial se ha recopilado, de las diferentes bases de datos de péptidos disponibles, péptidos que se han identificado que pertenecen a una de dos clases: antimicrobianos y NO-antimicrobianos. Para dicho conjunto de péptidos se ha calculado aproximadamente 1700 descriptores a partir de propiedades como la carga eléctrica, la hidrofobicidad, el momento hidrofóbico, el punto isoeléctrico, la estructura primaria y otras características fisicoquímicas de los péptidos.

Para el ejercicio se entregan 3 archivos:

- Base de datos positiva, la cual contiene los péptidos antimicrobianos.
- Base de datos negativa, la cual contiene los péptidos NO antimicrobianos.
- Un conjunto de validación para determinar el desempeño de los algoritmos seleccionados como mejores y con base en la cual se determinará el orden de la competencia.

La competencia tiene dos partes. El informe de ser entregado en formato pdf y la implementación debe ser entregada en un solo Notebook de Jupyter.

Parte 1: Clasificación sin selección

1. Cargue los archivos de los conjuntos de datos y aplique normalización.

2. Divida el conjunto de datos en 2 subconjuntos: un conjunto con el 80% de las muestras para el entrenamiento (training) y otro con el 20% restante como conjunto de pruebas (testing).
3. Entrene mínimo 3 técnicas de clasificación diferentes, haciendo ajuste a los hyperparámetros.
4. El clasificador debe ser entrenado con los datos de entrenamiento de la partición. Se debe usar validación cruzada para la medición del desempeño.
5. Use el conjunto de prueba para evaluar los clasificadores. Las métricas a usar son: matriz de confusión, precision, recall y F1-score.

Parte2: Clasificación con selección:

1. Sobre el conjunto de entrenamiento, aplique un método de selección y un método de extracción (o transformación) para seleccionar las mejores características del conjunto de datos (se aplican de manera independiente sobre el conjunto de datos original normalizado).
2. Utilizar los clasificadores de la Parte1 con el nuevo conjunto de características. Pueden realizar hypertunning.
3. Use el conjunto de prueba para evaluar los clasificadores. Las métricas a usar son: matriz de confusión, precision, recall y F1-score.
4. Compare estos resultados con los clasificadores de la parte anterior. ¿Qué hallazgos hay?

Entregables.

1. **Informe**. Se debe presentar un informe, no mayor a 6 páginas, con la metodología usada para la solución del problema, además de una descripción, análisis y discusión de los resultados obtenidos y las conclusiones generales del proyecto.

   Nota: se espera que en el informe se describa cada una las pruebas realizadas. ¿Qué técnicas de machine learning exploró? ¿Cuál conjunto de hyperparámetros ajustó? Analizar los resultados por cada métrica de cada una de las técnicas exploradas. Mostrar los tiempos de entrenamiento para cada una de las técnicas usadas. ¿Qué pasos fueron los de mayor complejidad y retadores?

2. **Desarrollo**. Notebook (*.ipynb) con la solución completa, desde la carga de los datos hasta la visualización de resultados. El notebook debe estar correctamente comentado y con descripciones amplias en las celdas de mayor relevancia para la solución.

Los entregables deben ser enviados al correo **srobles@unal.edu.co** hasta el 14 de marzo de 2020 hasta las 11:59pm.

Asunto del correo: CRP – G3 - Taller 4 Competencia

Nota: Aquellas entregas que no cumplan con el asunto del correo, **no** se tendrán en cuenta.

Nota: En la revisión de los informes, es posible que se llamen a los grupos para sustentar el trabajo realizado.