# Predicción 'in-silico' de sitios de escisión reconocidos por la proteasa del HIV-1

## Introducción

En esta PEC vamos a realizar un informe que analiza un caso basado en los datos del artículo:

**State of the art prediction of HIV-1 protease cleavage sites. Rögnvaldsson et al. Bioinformatics, 2015, 1-7. doi: 10.1093/bioinformatics/btu810**

Los datos se pueden obtener directamente de la revista Bioinformatics o del dataset "HIV-1 protease cleavage data" desde la *UCI Machine Learning Repository*, disponible en http://archive.ics.uci.edu/ml.

En dicho artículo se investiga la predicción in-silico de los sitios en las proteinas que son reconocidos para la escisión (cleavage) por la proteasa del HIV-1.

*There are two different approaches to predicting cleavage by HIV-1 protease: molecular modeling and sequence analysis. It has been argued that the HIV-1 protease recognizes shape rather than a specific amino acid sequence (Prabu-Jeybalan et al., 2002), which supports aiming for the molecular modeling approach. However, the method is cumbersome and no large scale study has been done on the accuracy of molecular modeling approaches so it is very unclear if the approach is, or will be, competitive with the sequence based approach. This article demonstrates the current state-of-theart prediction, which uses the sequence-based approach.*

En particular, en artículo de Rögnvaldsson et al. se centra en la tarea de predecir si dado un octamero (secuencia de 8 aminoacidos), éste será o no reconocido por la proteasa.

*The HIV-1 cleavage problem is described in detail in (Rögnvaldsson et al., 2007) together with discussions on different encoding schemes. Only a concise description is given here. The classification task is to tell whether a given octamer (sequence of eight amino acids) will be cleaved or not between the fourth and the fifth position.*

La manera elegida para representar los datos es un paso crucial en los algoritmos de clasificación. En el caso que nos ocupa, análisis basados en secuencias, usaremos el mismo tipo de representación que los autores emplearon en el su estudio.

*The octamer is represented using an orthogonal encoding where each amino acid is represented by a 20-bit vector with 19 bits set to zero and one bit set to one (other encodings have been suggested, see later). This maps each octamer to an 8 by 20 binary matrix that is transformed into a 160-dimensional vector.*

En la PEC se implementará un algoritmo `knn` para predecir aquellos octameros que son sustrato para de la proteasa del HIV-1.

## Enunciado

ampliar?

1. Escribir en el informe una sección con el título "Algoritmo k-NN" en el que se haga una breve explicación de su funcionamiento y sus caracterisiticas. Además, se presente una tabla de sus fortaleza y debilidades.
2. Lectura del artículo (especialmente las secciones de Introducción y Métodos)
3. Desarrollar una función en R que implemente una codificación ortogonal (*orthogonal encoding*) de los octameros (ver la sección de Métodos).
4. Desarrollar un script en R que implemente un clasificador `knn`. El script debe realizar los siguientes apartados:

(a) Leer los datos `impensData.txt` y `schillingData.txt`. Crear un nuevo conjunto de datos que sea la unión de ambos y hacer una breve descripción de los datos. Incluir en esta descripción el patrón de cada clase de octamero mediante la representación de su secuencia logo (https://en.wikipedia.org/wiki/Sequence_logo). Para realizar esta representación se puede usar el paquete `ggseqlogo` descargable desde github.

(b) Utilizar la función de codificación ortogonal para representar los octameros. **NOTA:** En caso de no poder hacer la función, se puede descargar el fichero `ort_enc.csv` con los octameros ya transformados.

(c) Utilizando la semilla aleatoria `123`, separar los datos en dos partes, una parte para training (67%) y una parte para test (33%).

(d) Utilizar un knn (k = 3, 5, 7, 11) basado en el training para predecir que octameros del test tienen o no *cleavage site*. Además, realizar una curva ROC para cada $k$ y calcular su área bajo la curva (AUC).

(e) Comentar los resultados de la clasificación en función del AUC, número de falsos positivos, falsos negativos y error de clasificación obtenidos para los diferentes valores de $k$. La clase que será asignada como positiva es la `1`.

## Informe de la PEC

Las soluciones se presentarán mediante un informe dinámico R markdown con la estructura habitual de los ejercicios no evaluables realizados hasta ahora. En primer lugar, el informe tendra un titulo (igual que el de la PEC), el autor, la fecha de creación y el indice de apartados de la PEC. En segundo lugar, se crea una sección con el título "Algoritmo k-NN" donde se haga una breve explicación de su funcionamiento y sus caracteristicas. Además, se presenta la tabla de sus fortaleza y debilidades. En tercer lugar se realizan los diferentes apartados de la PEC pero con la estructura de Step1 hasta Step5. Fijaros que que la codificación ortogonal es una transformación de los datos originales y corresponde a un Step.

Un caracteristica que se valorará es hasta que punto es el informe "dinámico". En el sentido de adaptarse el informe a cambios en los datos.

Se entregaran dos ficheros:

1. Fichero ejecutable (.Rmd) que incluya un texto explicativo que detalle los pasos implementados en el script y el código de los análisis.
2. Informe (pdf) resultado de la ejecución del fichero Rmd anterior.

## Puntuacions de los apartados

Apartado 1 (5%), Apartado 3 (25%), Apartado 4 (60%), Calidad del informe dinámico (10%).