# Empleando representaciones numéricas y espacios métricos para correlacionar secuencias con propiedades

Este documento busca servir de base para guiar el trabajo de tesis de la utilización de técnicas de representación numérica y codificación de secuencias para analizar respuestas o propiedades como enfoque o guía para identificar vectores de interés o espacios latentes reducidos relacionados con respuestas favorables para guiar el diseño de nuevas variantes mutacionales. Teniendo esto en mente se buscan responder los siguientes puntos:

1. Cuáles son las mejores estrategias para codificar o representar secuencias de proteínas
2. Cómo se pueden emplear las métricas de distancia para elaborar estrategias de comparación de secuencias
3. Cómo podemos aplicar las representaciones para identificar subgrupos o definir subespacios latentes para incrementar la probabilidad de mejora de las propiedades deseables
4. Determinar si existe una correlación entre distancias y cambios evolutivos.

De manera general, este documento tiene el objetivo de guiar los pasos a desarrollar y los resultados que se deben obtener con respecto a la tesis, contemplando las estrategias de codificación, representación numérica, cálculo de distancias y aplicaciones de clustering. Sin embargo, es solo una guía, el alumno es quien debe realizar todos los cálculos y estimaciones necesarias, así como también la automatización de los códigos de ser necesario y la ejecución en los diferentes casos de estudio.

## Casos de estudio

Los casos de estudio a analizar corresponden a conjuntos de secuencias trabajadas mediante técnicas de evolución dirigida, las cuales se detalla en el paper del siguiente link: <https://academic.oup.com/bioinformatics/article/34/15/2642/4951834>

## Codificación de secuencias

Las secuencias en formato texto, deben ser codificadas empleando estrategias de codificación, para ello, se emplearon dos puntos de vista:

1. Propiedades fisicoquímicas
2. Transformadas de Fourier sobre propiedades fisicoquímicas

Ver el script testing\_encoders.py del source code para más detalles. En total, se pueden generar 8 codificaciones empleando tópicos de propiedades y a los cuales se les puede aplicar FFT para formar un total de 16 posibles estrategias de codificaciones para un mismo conjunto de datos.

## Representación numérica de secuencias

Empleando lenguajes de proteínas o Protein Language Model, se pueden generar representaciones desde los textos mediante modelos pre entrenados usando técnicas como text mining, autoencoders, y NLP. En este trabajo, se emplean de manera inicial los siguientes modelos pre entrenados:

1. Bepler
2. ESM
3. Fasttext
4. PlusRNN
5. Prottrans

La representación numérica generada es un vector de iguales dimensiones independiente sea el tamaño de la secuencia de entrada, solo dependiendo del método empleado. Para codificar esto, se trabajó con Diego Fernández, quien, en su tesis, está evaluando el uso de PLM para el entrenamiento de modelos predictivos.

## Cálculo de distancia

La distancia se emplea para comparar dos secuencias, para ello, se hace un cálculo de distancia de todos contra todos y se genera una matriz de adyacencia. Es necesario explorar diferentes estrategias de cálculos de distancia, siendo habilitadas las siguientes:

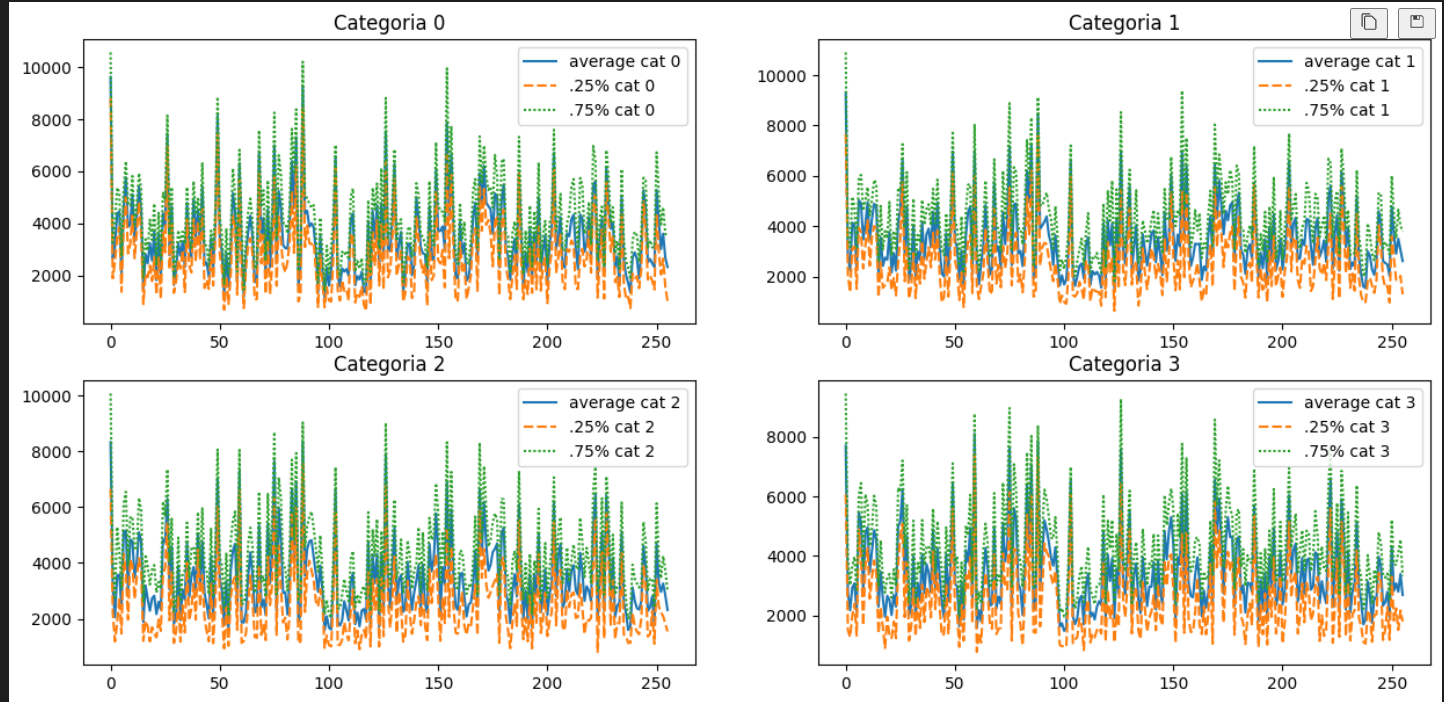
1. Euclidean
2. Braycurtis
3. Canberra
4. Chebyshev
5. Cityblock
6. Correlation
7. Cosine
8. Minkowski
9. Hamming

La idea es obtener las matrices de adyacencia empleando las diferentes estrategias de distancia, para cada una de las codificaciones o métodos de representación evaluadas.

Ver los scripts de la carpeta distance\_estimator del source code y el notebook distance\_estimator de la carpeta notebooks.

Una vez las distancias son estimadas, es posible analizarlas evaluando las distribuciones de distancias, respuestas, categorizando, etc. Para ello es necesario mirar lo implementando en el notebook exploring\_distances de la carpeta notebooks.

A su vez, también pueden analizarse los conjuntos de datos, ploteando la información por categorías:



Chart

Description automatically generated

Obteniendo imágenes tal como se pueden observar. En ellas, se observa que no existe una clara diferencia entre las categorías por cuartil definidas, siendo necesario un método más robusto para analizar diferencias o pudiendo denotar que en el caso de análisis de termoestabilidad, empleando las propiedades de FFT, no existen diferencias significativas de manera general. Sin embargo, en espacios específicos o en ventanas, si se logran identificar diferencias que pueden estar relacionadas con los cambios de funciones o mejoras en las actividades.

Chart, line chart, histogram

Description automatically generated

Se recomienda analizar las codificaciones generadas para “mirar” los conjuntos obtenidos y poder ir sacando conclusiones visuales relacionadas a posibles correlaciones entre patrón de curva y función.

## Evaluación de clustering mediante distancias

Se propone emplear clustering por louvain para detectar comunidades en estructuras de grafo para identificar proteínas relacionadas con baja distancia entre ellas, lo cual indicaría posibles relaciones de forma y evaluación de correlaciones entre la respuesta de interés.

Para ello, se implementa un script simple (ver notebook clustering\_using\_distances) para poder identificar grupos mediante matrices de adyacencia. Este script realiza filtros relacionados a remover las altas distancias y quedarse con las bajas simplemente, ya que ellas son las que importan para este análisis. Como resultado, este script genera labels o etiquetas que indican la pertenencia de un nodo a una comunidad. Las ejecuciones deben estar mediadas por la variación de los criterios de corte.

## Evaluación de clustering mediante aprendizaje no supervisado

A partir de las secuencias codificadas, se busca determinar la factibilidad de armar cluster de secuencias empleando funciones de espacios métricos y algoritmos de aprendizaje no supervisado. Para ello, el script clustering\_using\_unsupervised\_learning de la carpeta notebooks puede facilitar la aplicación de estas técnicas. En él, se encuentra implementado la exploración del algoritmo k-means. No obstante, se requiere la exploración de variados algoritmos. La idea es armar etiquetas de grupos con similitud en las secuencias en el espacio métrico empleado.

## Generación de perfiles

Los perfiles se generan a partir de la combinación de las aplicaciones de los métodos de clustering explorados para armar intersecciones y espacios latentes reducidos. Para ello, se implementa el script merge\_responses\_and\_labels en la carpeta notebooks, el cual combina los grupos obtenidos por distancia (grafos) y los grupos obtenidos por clustering, generando perfiles.

## Análisis de perfiles

Los perfiles generados se deben analizar desde un punto de vista descriptivo:

1. Qué tienen en común las secuencias
2. Cuáles son las propiedades de interés
3. Cómo se relacionan la forma o los puntos clave con las respuestas de interés
4. Cómo mejora la actividad el perfil con respecto a la secuencia nativa, existen cambios relacionados a la secuencia o a las propiedades, etc.
5. Cómo se relaciona los PLM o los codificadores con la respuesta, y en que puntos del espacio existe una mayor variación para poder generar nuevos vectores.
6. Se puede definir una probabilidad para que las nuevas secuencias que pertenezcan a un perfil en específico adquieran una mejora en su actividad
7. Cómo se pueden emplear los subgrupos definidos.

Para esto, se debe relacionar los IDS de las secuencias codificadas con los IDS de las secuencias existentes en el archivo unique\_sequences.csv de la carpeta others.

Elementos por explorar

1. Explorar todos los sets de datos, siendo un total de 4 conjuntos de datos a trabajar de manera inicial.
2. Explorar todas las estrategias de representación numérica o codificación, esto hace un total de 21 posibles representaciones.
3. Evaluar las diferentes estrategias de distancia, siendo un total de 9
4. De esta forma, la totalidad de experimentos a hacer es 4x21x9: 756
5. Para los 756 se deben hacer los cálculos de clustering y análisis de grupos y obtener resultados de todos, revisarlos y sacar algunas discusiones.