# Anotación de secuencias con *Machine Learning* aplicado al análisis de coevolución en Fascina 1

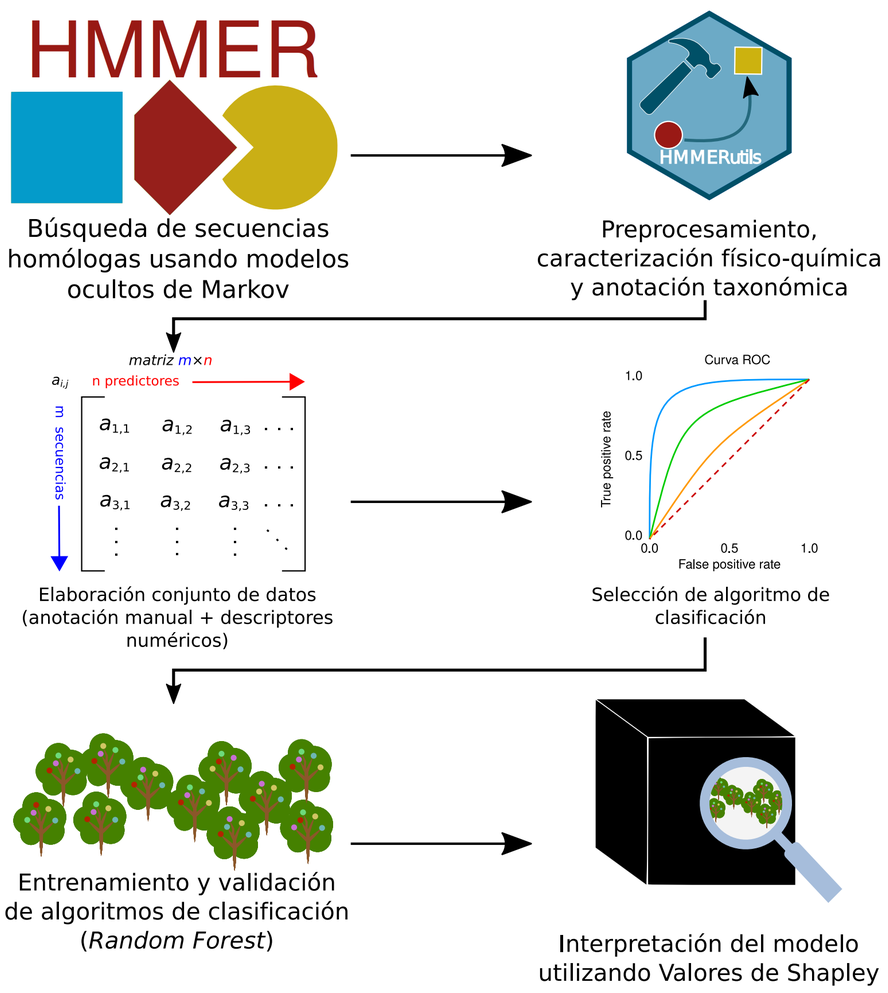
Fco. de Borja Campuzano Jiménez[1], Irene Luque Fernández[2], Coral del Val Muñoz[3]

## Introducción

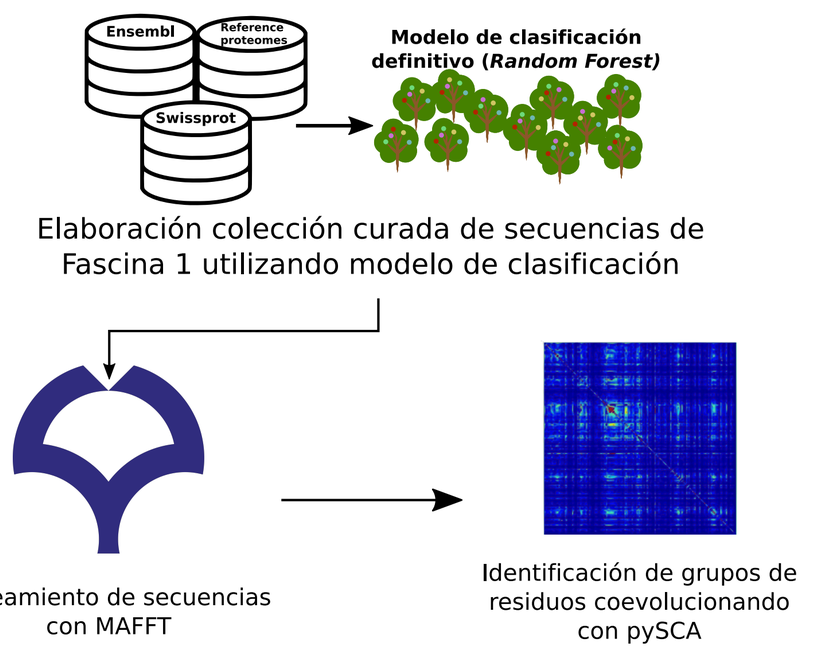
Fascina 1 es una proteína globular multidominio de unión a actina. Es una diana prometedora para el desarrollo de fármacos alostéricos antimetastásicos debido a su baja expresión en tejidos adultos normales su implicación en la diseminación y posterior supervivencia de células cancerosas, y su elevada flexibilidad conformacional. Un abordaje para la identificación de sitios alostéricos es el análisis de coevolución de residuos. No obstante, este análisis requiere de alineamientos extensos y el estado actual de la anotación en la familia de proteínas fascina es inadecuado para llevarlo a cabo. Para solventarlo, en este trabajos proponemos la utilización de técnicas de aprendizaje supervisado para desarrollar una herramienta de anotación automática a partir de propiedades físico-químicas con la que anotar las secuencias homólogas utilizadas para el posterior análisis de coevolución de residuos.

## Metodología

### 1) Algoritmo de clasificación de secuencias a partir de sus propiedades físico-químicas




### 2) Análisis de coevolución de residuos




## Resultados

**Herramienta de anotación automática**

Elaboramos un conjunto de datos formado por 425 secuencias representativas de las 5 clases de interés (Fascina, Fascina 1, Fascina 2, Fascina 3 y no Fascina) y equilibradas en número; y 669 descriptores ampliamente usados en quimiogenómica. Realizamos distintos ensayos para determinar el algoritmo que ofrecía mejores resultados para el problema de multiclasificación.

*Tabla 1. Métricas de evaluación para los 6 algoritmos de clasificación estudiados. Valores obtenidos mediante validación cruzada de diez particiones utilizando la partición de entrenamiento (298 secuencias).*

| | *Random Forest* | SVM | XGBOOST | KNN | Árbol de decisión | Regresión multinomial |
|---|---|---|---|---|---|---|
| Sensibilidad | 0.998 | 0.994 | 0.982 | 0.971 | 0.919 | 0.200 |
| Especificidad | 0.999 | 0.998 | 0.995 | 0.992 | 0.980 | 0.800 |
| PPV | 0.997 | 0.994 | 0.981 | 0.967 | 0.920 | 0.200 |
| NPV | 0.999 | 0.998 | 0.995 | 0.993 | 0.980 | 0.800 |
| Precisión | 0.998 | 0.993 | 0.979 | 0.965 | 0.919 | 0.175 |
| Valor-F1 | 0.998 | 0.993 | 0.979 | 0.964 | 0.912 | 0.296 |
| Exactitud equilibrada | 0.995 | 0.998 | 0.988 | 0.981 | 0.950 | 0.500 |

El algoritmo escogido fue Caret, una implementación de *Random Forest*. Este modelo fue validado con 127 nuevas secuencias (partición de validación), cometiendo únicamente un error. Por tanto, el modelo definitivo es capaz de predecir correctamente la anotación de nuevas secuencias y es capaz de identificar errores en la anotación automática (por ejemplo, en A0A068WNP9, A0A2Y9L644 y A0A2B4RKU0).

**Interpretación del modelo de Machine Learning**

Utilizamos los valores de Shapley para la interpretación del modelo. De este modo, identificamos que las variables a las que, en promedio, el modelo "prestaba más atención" corresponían 3 tríadas conjuntas. Las tríadas conjuntas representan la frecuencia con que aparecen ciertas combinaciones de tipos de aminoácidos (clasificados según volumen y momento dipolar).

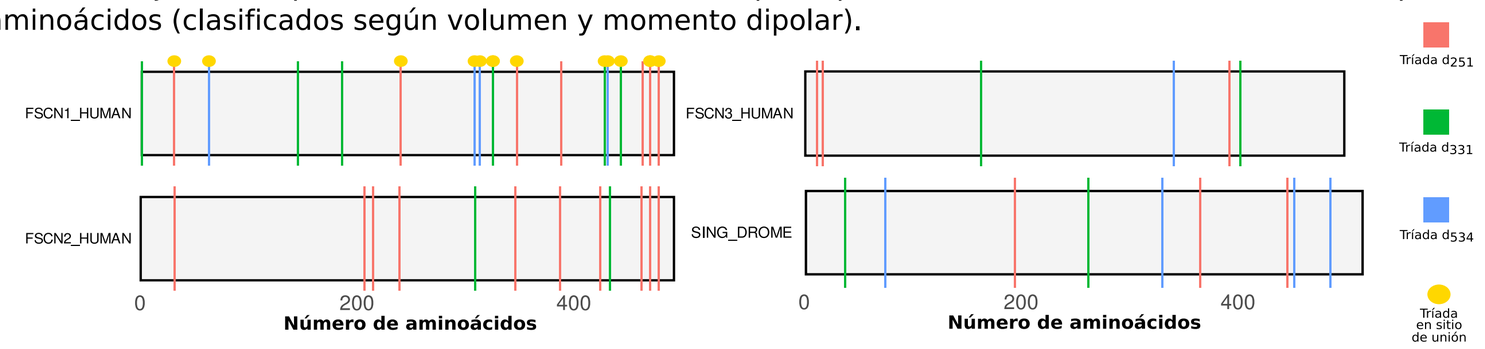


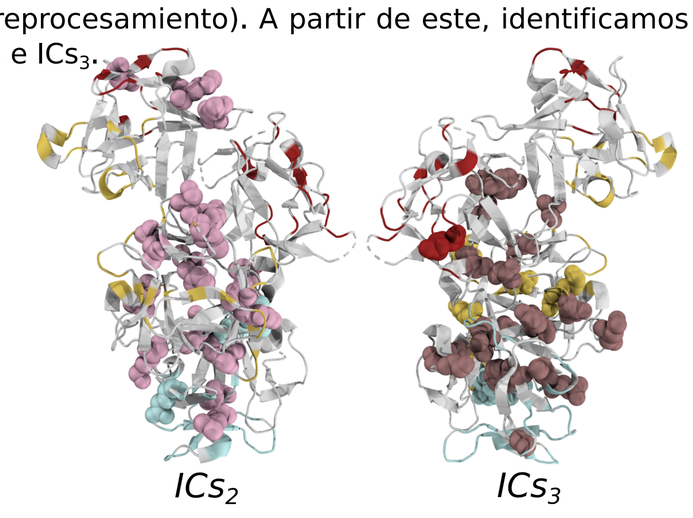

*Figura 1. Distribución de las tríadas $d_{251}$, $d_{331}$ y $d_{534}$ a lo largo de las secuencias de referencia para Fascina 1, Fascina 2, Fascina 3 y Fascina (secuencias que no evolucionaron de la duplicación que tuvo lugar en vertebrados). Estos motivos se distribuyen de manera diferente en los distintos tipos de secuencias y, además, forman parte, en su mayoría, de los sitios de unión en Fascina 1 humana. El modelo, por tanto, no solo anota nuevas secuencias correctamente sino que además lo hace utilizando variables con significado biológico.*

**Análisis de coevolución de residuos en Fascina 1**

Ayudándonos del modelo de anotación automática, construimos un alineamiento curado de Fascina 1 (dimensiones 469 secuencias x 489 posiciones tras el preprocesamiento). A partir de este, identificamos 2 grupos de residuos que han coevolucionado entre sí, $ICs_2$ e $ICs_3$.

*Figura 2. Representación de los dos grupos de residuos que han coevolucionado sobre Fascina 1 humana (código: 3LLP). Estos dos grupos forman una red de residuos próximos en el espacio y conectan los tres sitios de unión. Podrían ser responsables del alosterismo de Fascina 1.*

*Los residuos que han coevolucionado se muestran como esferas sobre la proteína cartoon. Los sitios de unión 1, 2 y 3 se muestran coloreados en amarillo, rojo y azul, respectivamente. Residuos de los grupos $ICs_2$ e $ICs_3$ sin anotación previa se muestran en rosa y marrón, respectivamente.*




## Conclusiones

- El desarrollo de herramientas de anotación automática a partir de propiedades físico-químicas inferidas a partir de la secuencia puede ser utilizado para anotar nuevas secuencias y para caracterizar e identificar propiedades de interés biológico.
- El análisis de coevolución de residuos se ve altamente influenciado por la composición del alineamiento de partida y, gracias al modelo de anotación utilizado, hemos podido identificar dos grupos de residuos que han coevolucionado entre sí y podrían formar parte de los sitios alostéricos de Fascina 1.

## Afiliaciones

*Fco. de Borja Campuza Jiménez[1]: Graduado en Biotecnología, estudiante en prácticas en Instituto Andaluz de Investigación en Data Science and Computational Intelligence (DaSCI), Universidad de Granada, Granada, España*
*campuzanocurro@gmail.com*

*Irene Luque Fernández[2]: Departamento de Química Física e Instituto de Biotecnología, Facultad de Ciencias, Universidad de Granada, 18071 Granada, España*
*iluque@ugr.es*

*Coral del Val Muñoz[3]: Departamento de Ciencias de la Computación e Inteligencia Artificial, Instituto Andaluz de Investigación en Data Science and Computational Intelligence (DaSCI), Universidad de Granada, Granada, España*
*delval@ugr.es*

UNIVERSIDAD DE GRANADA

Facultad de Ciencias

CEEBI - CONGRESO ESTATAL DE ESTUDIANTES DE BIOTECNOLOGÍA