

Evaluación del rendimiento de metodologías univariantes, multivariantes y de aprendizaje automático en el análisis de variaciones genómicas

Objetivo del proyecto

El objetivo general de este trabajo consiste en el desarrollo de una nueva metodología basada en *machine learning* que sea capaz de determinar qué SNPs pertenecientes a un pathway dado son más relevantes a la hora de diferenciar en un estudio de genoma amplio entre casos de cáncer y controles. Más concretamente, el método que se propone se basa en el uso de un tipo de algoritmos evolutivos denominado algoritmos genéticos, pero que nada tiene que ver más allá de su nombre con la genética biológica, así como de las máquinas de vectores de soporte, metodología propia del *machine learning* muy adecuada para los problemas de clasificación.

Líder del proyecto

El grupo GIGAS (Grupo de Investigación en Interacciones Gen-Ambiente-Salud) de la Universidad de León, www.unileon.es está dirigido por el Dr. Vicente Martín, y lleva a cabo investigación epidemiológica en la que combina las exposiciones ambientales y hábitos de vida, con los datos genómicos de los pacientes para desarrollar modelos que permitan valorar su asociación con eventos de salud tanto crónicos como cáncer o enfermedades cardiovasculares, como enfermedades transmisibles como el COVID-19 o la gripe.

Es Unidad de Investigación Consolidada de la Junta de Castilla y León (UIC-109) que forma parte de un grupo CIBERESP.

Mantiene colaboraciones con el grupo de Modelización Matemática Aplicada, que ha sido con el que ha desarrollado estos modelos de *machine learning* para los que se ha empleado Caléndula, en el marco de una tesis doctoral conjunta.

Participantes del proyecto

Universidad de Oviedo, www.uniovi.es

SCAYLE, Supercomputación Castilla y León, www.scayle.es

Ejecución: Del 2018 al 2022.

Financiación del proyecto

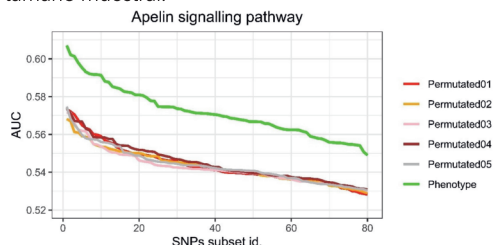
Sin financiación.

Funciones de SCAYLE

Se aplicaron metodologías de *machine learning* y optimización que requieren de cálculo intensivo.

Justificación del proyecto

El proyecto busca desarrollar métodos de evaluación y estudio del genoma mediante técnicas de *machine learning* que mejoren la capacidad para discriminar la relación entre la variabilidad genética y la aparición de patologías, en un marco con una gran cantidad de variables en comparación con el tamaño muestral.



Valores del área bajo la curva ROC de las 80 iteraciones realizadas con el pathway apelin signalling en la base de datos de casos controles y en 5 repeticiones con las etiquetas de casos y controles permutadas.