Congreso Colombiano de Bioinformática y Biología Computacional
2022

Integración de datos genómicos para las interacciones proteína-proteína

Vanessa Orjuela Lagos[1,a], Liliana López Kleine[1,b], Yesid Cuesta Astroz[2,c]

[1]Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia

[2]Instituto Colombiano de Medicina Tropical - Universidad CES, Medellín, Colombia

Resumen

Las etapas críticas en la biología de un patógeno están mediadas principalmente por interacciones proteína-proteína entre el hospedero y el patógeno. Para poder conocer, entender y proponer alternativas terapéuticas, es necesario identificar las interacciones a nivel molecular; de manera que es pertinente analizar predecir con base en datos genómicos disponibles dicha interacción entre proteínas. Para poder realizar estas predicciones el primer paso es la obtención e integración de datos con información genómica disponible de alta calidad y seleccionada con base en criterios biológicos y estadísticos. El objetivo de esta propuesta es construir una base de datos multi-ómica para cada

[a]Autora. E-mail: vorjuelal@unal.edu.co
[b]Autora presentadora. E-mail: llopezk@unal.edu.co
[c]Autor. E-mail: ycuesta@ces.edu.co

organismo de un sistema hospedero-patógeno, que debido a su naturaleza heterogénea, requieren de una transformación adecuada para ser comparables.

El segundo paso es la representación de dichos datos para lo cual se ha seleccionado la representación en matrices de similitud tipo kernel que permiten transformar los datos mediante una función kernel asociada, obteniendo como finalidad medidas de similitud entre proteínas o genes para cada tipo de variable. Y que además pueden ser comparables, pues estas se encuentran en un mismo espacio y permiten encontrar relaciones no lineales.

El tercer paso es el análisis de los datos para extraer las características comunes de los datos genómicos multivariados. Se usa el método de análisis de correlación canónica del kernel (KCCA), empleando un kernel integrado por pesos; buscando, además de encontrar el espacio de mayor correlación, poder sintetizar la información entre los datos conocidos y desconocidos. El proceso de predicción de las interacciones proteína a proteína se hace bajo un enfoque supervisado, donde el entrenamiento se basa en el método KCCA usado un kernel de referencia de las relaciones conocidas en la red de proteínas (las proteínas que se sabe que interactúan físicamente gracias a la validación experimental) y el proceso de prueba se realiza proyectando el kernel asociado a las variables en el espacio de características de KCCA. Con esto se busca proyectar la función conocida con la desconocida y realizar predicciones.

Se presentarán resultados preliminares sobre la calidad de los datos, la representación en matrices de similitud kernel, así como la calidad de los datos de referencia.