Desarrollo de un clasificador para identificar variantes patogénicas en regiones no codificantes del genoma humano.

Ben Omega Petrazzini

Dra. Lucia Spangenberg, Institut Pasteur de Montevideo
Unidad de Bioinformática

Msc. Fernando López-Bello, Quanam
Líder de BigData y Computación Cognitiva

2019

Introducción

La *Organización Mundial de la Salud* estima que alrededor de 400 millones de personas sufren por enfermedades raras a nivel mundial[1]. Estas patologías se caracterizan por ser de difícil diagnóstico, con síntomas comprometedores que pueden surgir en edades muy tempranas. Existen entre 5000 y 8000 diferentes tipos[1], la mayoría de las cuales tienen bases genéticas, siendo esta en muchos casos una sola mutación en un único ácido nucleico del ADN (polimorfismo de nucleótido simple o SNP) con baja frecuencia a nivel poblacional. Por esta razón, los estudios genómicos han permitido grandes avances en el diagnóstico de dichas patologías. En el marco del proyecto del genoma de los uruguayos (URUGENOMES) se ha secuenciado la información nucleotídica completa de 80 individuos, 30 de ellos padecen hoy una de estas enfermedades.

La secuenciación de un genoma humano puede tener como resultado cerca de 3 millones de variantes, de las cuales solo la minoría resultan ser patogénicas. La automatización de procesos permite filtrar grandes números de variantes según diferentes parámetros. De esta manera se pueden obtener mutaciones potencialmente causales de la enfermedad. Dicho proceso de filtrado, si bien complejo, puede ser aplicado de forma sistemática para variantes codificantes. Sin embargo, para variantes no codificantes la complejidad aumenta sustancialmente. Dos características de las mismas hacen que el conjunto de variantes no codificantes sea demasiado grande y no tenga suficiente información. En primer lugar, la longitud de dichas regiones nos lleva a trabajar con un número mucho más grande de datos, se estima que el 95% del genoma humano no codifica para proteínas, estas son intrones, secuencias UTR, fragmentos intergénicos, pseudogenes y demás. En segundo lugar, las dinámicas de los procesos evolutivos hacen que los cambios nucleotídicos en las mismas sean mucho más frecuentes, por lo que obtenemos una variabilidad mayor en estas regiones. Además, la mayoría de los scores que permiten predecir la patogenicidad de una variante se limitan a regiones codificantes. Esto se debe a que dichos evaluadores se basan en las propiedades fisicoquímicas de los aminoácidos, ausentes en regiones no codificantes. Por ello es importante encontrar un mecanismo aplicable a regiones no codificantes que nos permita discernir entre las mutaciones que pueden tener relevancia clínica y las que no.

El proceso de filtrado de variantes exónicas no arrojo resultados concluyentes para un gran número de los 30 pacientes analizados como parte del proyecto URUGENOMES, resta así examinar el otro 95% del genoma en regiones no codificantes. Para esto debemos encontrar una solución al problema planteado en el párrafo anterior. Así surgió la posibilidad de desarrollar un clasificador basado en Machine Learning que permita priorizar SNPs no codificantes según su significancia biológica. Este permitiría discernir variantes *patogénicas* / *potencialmente patogénicas* de variantes *benignas* / *potencialmente benignas* / *significado incierto*, reduciendo el set de datos a analizar.

Desde el surgimiento de la *Next Generation Sequencing (NGS)*, la obtención masiva de datos genómicos ha permitido integrar técnicas de Machine Learning a los estudios bioinformáticos. Trabajos como *Zhou J., et al. Nature Methods. 2015* han podido aplicar algoritmos de Deep Learning en la resolución de problemas biológicos[2]. En este caso han predicho exitosamente los efectos a nivel de la cromatina de variantes en regiones no codificantes, particularmente en secuencias reguladoras del genoma humano.

Los primeros pasos de nuestro trabajo consisten en pulir y entender los datos disponibles, generando una tabla en la que los diferentes abordajes de los scores de predicción nos permitan asignarle a cada variante un perfil característico. Luego del curado manual de la tabla se procede a entrenar diferentes algoritmos de clasificación, esto nos permitirá distinguir las propiedades (columnas) con mayor peso en la clasificación de las variantes. Utilizando parámetros que indiquen la calidad de la predicción (ej: cross-validation) se seleccionará uno o varios modelo/s óptimo/s para aplicar a nuestro objeto de estudio. Por lo que el último paso del proyecto consiste en emplear el clasificador para priorizar variantes no codificantes de interés, posiblemente asociadas con la patología de los pacientes. Obtendríamos así un subconjunto de datos a partir del cual el genetista clínico podrá tomar una decisión informada sobre el diagnóstico del paciente.

De esta manera esperamos contribuir a la solución del problema presentado al principio. Dado que el estudio de las variantes exónicas no arrojó en muchos casos resultados concluyentes, confiamos en que este abordaje nos permitirá indagar en las amplias regiones no codificantes de sus genomas. Una vez finalizado el algoritmo podremos clasificar mutaciones que puedan afectar la regulación génica, compactación / estructura del ADN, replicabilidad, entre otras. Esperamos que estos resultados nos ayuden a identificar bases genéticas causantes de las enfermedades raras de algunos de los pacientes del proyecto URUGENOMES, pudiendo contribuir a su diagnóstico lo antes posible.

Objetivos

A mediados de abril se dio inicio a la primer etapa del proyecto de tesis. En esta se busca pulir la información disponible en las bases de datos de variantes nucleotídicas para poder utilizarla. Si bien hemos visto que la base de datos ClinVar tiene anotaciones bastante confusas, este proceso no debería tomar más de seis o siete semanas. Una vez curados los datos, se anotarán las variantes de ClinVar con el software ANNOVAR para recabar la mayor cantidad de información disponible de cada polimorfismo.

De aquí se pasa al segundo paso del proyecto, el cual consiste en analizar cada característica (columna) de la tabla evaluando su balance y posibles sesgos de información. Esto nos permitirá descartar atributos que puedan comprometer la fiabilidad del algoritmo, previniendo así outliers y overfitting.

El siguiente paso consiste en una revisión bibliográfica para entender el funcionamiento de los clasificadores que vamos a utilizar. El objetivo aquí es ser capaz de analizar los resultados de la predicción, entendiendo como maneja los datos el algoritmo de clasificación.
El cuarto paso del proyecto es poner a prueba los diferentes clasificadores, este se realizará de manera paralela a la búsqueda bibliográfica. El testeo recursivo de los clasificadores nos permitirá comparar las eficiencia de cada uno, para poder elegir el/los que mejor se adapte/n a nuestro set de datos.
Una vez seleccionado/s el/los algoritmo/s de clasificación más óptimo/s se procede a la última etapa del proyecto. Esta consiste en aplicar los algoritmos entrenados a un conjunto de pacientes del proyecto URUGENOMES. Esperamos obtener un set reducido de variantes a analizar, con una predicción confiable por parte del clasificador.

Materiales y Métodos

Los análisis estadísticos y el manejo de las tablas se realizaran con el software de acceso libre R. Las anotaciones en cada variante se realizaran con el programa de acceso libre ANNOVAR. Las bases de datos utilizadas y sus scores correspondientes son de acceso publico, algunas de ellas son Clinvar, dbNSFP, InterVar, gnomAd, Kaviar, fatHMM y gwava.
Los procesos computacionalmente demandantes serán corridos a través de un servidor disponible en el Institut Pasteur Montevideo. El mismo cuenta con 256 Gb de RAM y 32 procesadores.
Para el primer paso de curado de la información disponible vamos a descargar la base de datos Clinvar y analizar manualmente la significancia biológica asignada a cada variante. Estas etiquetas requieren un filtrado particularmente especifico y es un paso importante para los próximos objetivos del proyecto ya que las predicciones del mismo van a depender del distintivo que se le haya asignado a cada mutación. Una vez curados los datos, se anotarán las variantes de ClinVar con el software ANNOVAR para recabar la mayor cantidad de información disponible de cada polimorfismo.
En el segundo paso del proyecto se utilizara el software R para analizar en detalle cada característica (columna) de la tabla. Para estudiar el desbalance y la presencia de outliers se verá la proporción de variantes que tienen una anotación en cada base de datos, corroborando que las significancias biológicas de las mismas sean equilibradas.
Para el tercer paso simplemente se analizara bibliografía presente en textos como *Bishop C., Pattern Recognition and Machine Learning*[3] y *Goodfellow I. et al, Deep Learning*[4]. Además de bibliografía disponible en la Unidad de Bioinformática del Institut Pasteur de Montevideo y en la biblioteca de la Facultad de Ciencias, UdelaR.
Para el cuarto paso del proyecto utilizaremos algoritmos de aprendizaje supervisado altamente informativos, como pueden ser Knn, SVM y decision trees. Esto nos permitirá distinguir las propiedades

(columnas) con mayor peso en la clasificación de las variantes. Si bien estos métodos son fácilmente interpretables, en algunos casos no proporcionan suficiente potencial de predicción. Por lo que progresivamente se irán testeando algoritmos con mejor tasa de predicción, pero posiblemente menos interpretables. Utilizando parámetros que indiquen la calidad de la predicción (ej: cross-validation) se seleccionará uno o varios modelo/s óptimo/s para aplicar a nuestro objeto de estudio.

En la última etapa del proyecto se genera una tabla para cada individuo con las mismas características que la utilizada para entrenar el programa. Una vez generados estos archivos se aplica el/los clasificador/es a un conjunto de pacientes del proyecto URUGENOMES.

Bibliografía

[1] Van Welly S. y Leufkens H.G.M. (2004), Priority Medicines for Europe and the World "A Public Health Approach to Innovation", 6.19 Rare Diseases. World Health Organization.
Disponible en: https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf

[2] Zhou, J. y Troyanskaya, O. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10), pp.931-934.

[3] BISHOP, C. (2016). *PATTERN RECOGNITION AND MACHINE LEARNING*. [Place of publication not identified]: SPRINGER-VERLAG NEW YORK.

[4] Goodfellow, I., Bengio, Y. y Courville, A. (2016). *Deep learning*. Cambridge (EE. UU.): MIT Press.