Metodología de la investigación avanzada: introducción al estudio de los sistemas complejos y sus aplicaciones. Parte VII: Estudios de interacción de genes (epistasis y aprendizaje automático)

Lucas Costa y Pablo Argibay

RESUMEN

En medicina, pocas son las enfermedades de base genética que pueden ser explicadas por la presencia de una alteración monogénica; antes bien, es en la compleja red de interacciones de genes y factores extragénicos donde se pueden encontrar los fundamentos moleculares de algunas enfermedades. La epistasis o epistasia es el fenómeno de interacción de genes, cuyo resultado es un determinado carácter fenotípico, comportamiento o proceso molecular relacionado, entre otras cosas con diversas enfermedades. Es un fenómeno complejo y en la actualidad forma parte de los análisis en los que se pretende relacionar la interacción y combinatoria de genes con determinada enfermedad. Un método de aprendizaje de máquinas (*machine learning*), llamado reducción de dimensionalidad multifactorial (MDR), se presenta como una interesante alternativa a los métodos paramétricos tradicionales para la detección y caracterización de interacciones no lineales entre genes.

Palabras clave: interacción gen-gen, epistasis, aprendizaje de máquinas, minería de datos, reducción de dimensionalidad multifactorial.

ADVANCED RESEARCH METHODOLOGY: INTRODUCTION TO THE STUDY OF COMPLEX SYSTEMS AND ITS APPLICATIONS. PART VII: STUDIES ON GENE INTERACTION (EPISTASIS AND MACHINE LEARNING)

ABSTRACT

In medicine, there are few genetic diseases that can be explained by the presence of a monogenic disorder, rather it is in the complex interactions of genes and extragenic factors where we can find the molecular underpinnings of some diseases. Epistasis is the phenomenon of interaction of genes, resulting in a particular phenotypic trait, behavior or molecular process related, inter alia with various diseases. It is a complex phenomenon and is currently part of the analysis that aims to link the combinatorial interaction of genes with a particular disease. A machine learning method, called multifactor dimensionality reduction (MDR) is presented as an interesting alternative to traditional parametric methods for detecting and characterizing nonlinear interactions among genes.

Key words: mendelian genetics, epistasis, genotype-phenotype relationship, data mining, machine learning.

Rev. Hosp. Ital. B.Aires 2014; 34(1): 00-00.

RELACIONANDO GENES CON ENFERMEDADES

Un objetivo central de la genética es identificar variaciones de la secuencia de ADN, conocidas como polimorfismos, que confieren un riesgo aumentado para desarrollar enfermedades particulares. Conocer la relación de correspondencia entre variaciones de la secuencia de ADN y la susceptibilidad a la enfermedad abre el camino para estudiar posibles mejoras en el diagnóstico, prevención y tratamiento.

En el caso de trastornos de genes individuales, tales como la anemia de células falciformes o la fibrosis quística, la relación genotipo-fenotipo es aparentemente simple, ya que el genotipo mutante es explícitamente responsable de la enfermedad. Sin embargo, esta situación es la excepción y, en la mayoría de las enfermedades, dicha relación es extremadamente difícil de caracterizar debido a que la enfermedad es el resultado de una interacción compleja de factores genéticos y ambientales.

La multicausalidad genético-ambiental y las interacciones no lineales hacen necesario el empleo de diferentes estrategias de investigación más allá del uso de herramientas estadísticas convencionales de correlación o covarianza entre variables. Por otra parte, no es eficaz el uso de inferencias basadas en la herencia de caracteres mendelianos estrictos. La mayoría de las enfermedades son complejas y tienen patrones de herencia ambiguos, generalmente formados por la combinatoria de genes correspondientes

Entregado 20/02/14 Aceptado 17/04/14

Laboratorio de Aprendizaje Biológico y Artificial (LBAL). Instituto de Ciencias Básicas y Medicina Experimental. Hospital Italiano de Buenos Aires.
Correspondencia: lucas.costa@hospitalitaliano.org.ar

a muchos loci.[1] Un tipo de interacción genética interesante es la interacción gen-gen o epistasis, fenómeno bajo el cual se engloba el grado de no linealidad en el mapeo entre genotipo y fenotipo. La forma más simple de explicar el concepto de epistasis es la siguiente: un alelo en un locus enmascara la expresión de un alelo en otro locus. Por ejemplo, en un nivel fenotípico claro: el gen que causa el albinismo ocultaría al gen que controla el color del pelo.
En una red compleja de eventos moleculares relativamente simples como la expresión génica o la presencia de polimorfismos, la epistasis es una manera de conceptualizar la interacción entre genes (Fig. 1). Existen dos formas conceptuales para caracterizar la epistasis, una de ellas es reconocerla como desviaciones de los patrones de herencia simples observados por Mendel y la otra es entenderla como desviaciones de la aditividad en un modelo estadístico lineal.

APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE INTERACCIONES GEN-GEN

Aunque los enfoques estadísticos tradicionales de búsqueda de genes candidatos, como los análisis de asociación y vinculación, han tenido mucho éxito en el descubrimiento de los genes responsables de enfermedades causadas por un único gen, son relativamente ineficaces en el análisis de las redes génicas causales de enfermedades multifactoriales complejas. Esto se debe en gran parte al hecho de que, ya que muchos genes interactúan para causar enfermedades complejas, los efectos individuales son tan insignificantes que resultan difíciles de detectar mediante los métodos tradicionales que no fueron diseñados para tener en cuenta las interacciones, sino más bien para detectar efectos individuales fuertes. A medida que el énfasis en la genética humana se ha alejado del estudio de los trastornos monogénicos, extremadamente raros, se ha puesto énfasis en el estudio de interacciones complejas, propias de enfermedades más prevalentes como el cáncer, las enfermedades cardiovasculares, los trastornos metabólicos y las enfermedades neurológicas y psiquiátricas. Esto ha hecho necesario el desarrollo de métodos estadístico-computacionales dirigidos específicamente a la detección de epistasis.
Las interacciones gen-gen son difíciles de detectar y caracterizar mediante el uso de métodos estadísticos paramétricos tradicionales como la regresión logística, debido a la escasez de datos que generalmente se presentan en espacios de altas dimensiones. Es decir, cuando se consideran las interacciones entre múltiples variables, hay muchas celdas de la tabla de contingencia que tienen un conteo de datos bajo o nulo. Esta situación se conoce como la "maldición de la dimensionalidad" y puede dar lugar a estimaciones de los parámetros que tienen errores estándar muy grandes, ocasionando un aumento del error de tipo I. Los modelos lineales desempeñan un papel muy importante en la genética moderna y la epidemiología, ya que tienen una sólida base teórica, son fáciles de poner en práctica mediante el uso de una amplia gama de programas estadísticos, y fáciles de interpretar. Sin embargo, hay que destacar que tienen grandes limitaciones para la detección de patrones no lineales de interacción. Dichas limitaciones del modelo lineal y de otros métodos estadísticos paramétricos han motivado el desarrollo de nuevos métodos computacionales, tales como los de aprendizaje automático[2] y minería de datos[3], que hacen menos suposiciones sobre la forma funcional del modelo y los efectos que están modelando, además de controlar de una manera mucho más adecuada el sobreajuste[4] del modelo. Es por esto que un método de

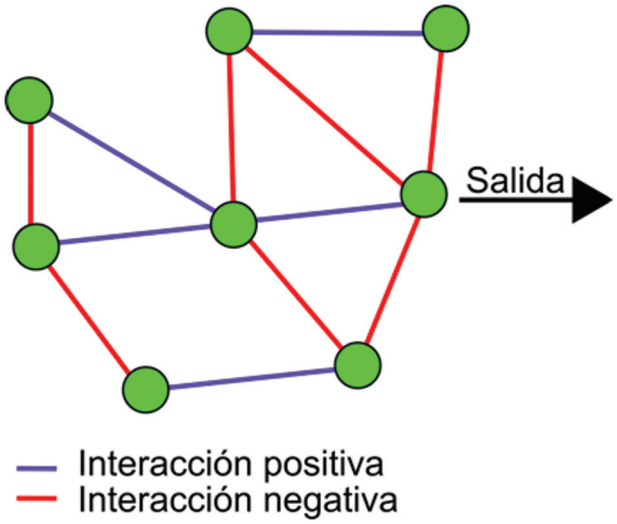

Figura 1. Hipotética red génica de una enfermedad en la cual se muestran las interacciones entre genes (representados por círculos verdes). Algunos pares de genes tienen interacciones positivas (líneas azules), mientras que los otros pares de genes tienen interacciones negativas (líneas rojas). En conjunto, estas interacciones dan como resultado un fenotipo de salida. El fenómeno de epistasis, generado por la presencia de interacciones no lineales entre genes, es el causante de aumentar la complejidad de la red; descubrirlo y entenderlo permite comprender determinados aspectos clave de la enfermedad.

1. Locus: se refiere en los cromosomas homólogos, a la posición específica de un determinado gen o uno de sus alelos (secuencias de bases diferentes que ocurren en el mismo gen de determinado locus en un cromosoma homólogo). *Oxford Dictionary of Biochemistry and Molecular Biology*. Edición revisada, año 2000.

2. El aprendizaje automático o aprendizaje de máquinas (*machine learning*) es la ciencia de conseguir el aprendizaje por computadoras sin la utilización de programas explícitos. En la última década, el aprendizaje automático ha llevado al desarrollo de vehículos autoconducidos, el reconocimiento práctico de voz, la búsqueda de páginas web eficaces y una comprensión muy mejorada del genoma humano.

3. La minería de datos es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos. Véase: exa.unne.edu.ar/depar/areas/informatica/.../Mineria_Datos_Vallejos.pdf

4. Sobreajuste (*overfitting*) se refiere a la situación estadística en la cual el polinomio se ajusta forzadamente a los datos de una muestra, no prediciendo adecuadamente lo que ocurre en la población.

aprendizaje de máquinas llamado reducción de dimensionalidad multifactorial (MDR por sus siglas en inglés) se ha convertido en un importante y novedoso método para la detección y caracterización de modelos de epistasis estadística en los estudios de asociación genética, complementando el paradigma de modelado lineal.

REDUCCIÓN DE DIMENSIONALIDAD MULTIFACTORIAL (EL MÉTODO MDR)

El MDR fue desarrollado como una estrategia no paramétrica (ningún parámetro es estimado) y libre de modelo genético (ningún modelo genético es preasumido) de aprendizaje automático, para la identificación de combinaciones de factores genéticos y ambientales específicos que sean predictores de una situación clínica determinada. Es un método ampliamente utilizado en estudios epidemiológicos para detectar e interpretar efectos epistáticos (interacciones gen-gen no lineales), cuando no existen efectos principales significativos.

En estudios de enfermedades multifactoriales, el área original de aplicación del MDR, la principal fortaleza del algoritmo es que facilita la detección y caracterización simultánea de múltiples loci, asociados con determinados rasgo clínicos, mediante la reducción de la dimensionalidad de los datos. En un principio existen tantas dimensiones como variables independientes (loci) se incluyan en el software; de esta manera se puede ejemplificar que dos polimorfismos con tres genotipos cada uno forman nueve combinaciones de genotipos de dos loci; cada una de esas combinaciones posee su propio conteo de sujetos casos y controles. La razón entre caso y controles de cada genotipo multilocus es comparada con un valor umbral definido por el usuario; de esta manera se clasifican los genotipos multiloci en dos grupos, grupo de alto riesgo (si supera el umbral) o grupo de bajo riesgo (si no supera el umbral). El MDR produce así una reducción de dimensionalidad en un conjunto de datos, llevando un problema *n*-dimensional a uno de una única dimensión (una variable con categorías bajo y alto riesgo). Esta nueva variable unidimensional se utiliza para entrenar y testear el algoritmo de *machine learning*. Por tratarse de una situación de aprendizaje supervisado,[5] la fracción ponderada de casos y controles correctamente etiquetados por el algoritmo, llamada exactitud equilibrada (*balanced accuracy*), se utiliza como un indicador para seleccionar el mejor modelo, el cual presenta un mínimo en el error de predicción. Por otro lado, la aplicación de la técnica de validación cruzada genera aleatoriamente M divisiones (típicamente M es igual a 10) del *set* de datos; cada división aparta un 90% de individuos para la etapa de entrenamiento y un 10% para la etapa de validación. De esta manera, la consistencia de validación cruzada (*CV consistency*) muestra cuántas de las M veces que se corrió el algoritmo, un determinado modelo fue elegido como sobresaliente, ayudando así a identificar el mejor modelo que presente la máxima coherencia. Luego, para evaluar si el mejor modelo hallado es estadísticamente significativo, se aplica una prueba de permutación, basada en simulación de datos, para obtener un valor de p.

Resumiendo de una manera más técnica el método, se debe entender que el MDR identifica las interacciones entre variables discretas que influyen en un resultado binario. El MDR se utiliza para determinar el modelo óptimo de orden k-ésimo (la interacción del conjunto de k variables que mejor predicen la clase) entre las *n* posibles variables, los pasos del algoritmo MDR pueden verse en la Fig. 2.

Ningún método de búsqueda de solución se incorpora explícitamente en el algoritmo; en su lugar, una búsqueda

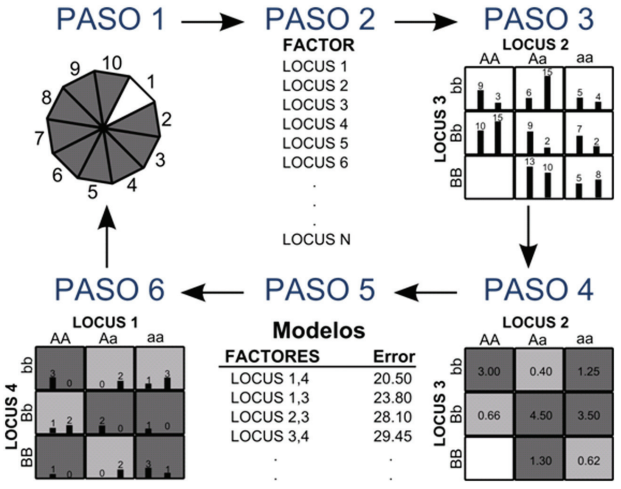

Figura 2. Resumen de los pasos generales involucrados en la aplicación del método MDR. *Paso 1:* el *set* de datos es dividido en un *set* de entrenamiento (9/10 de los datos) y un *set* de testeo (1/10 de los datos). *Paso 2:* se elige un determinado número (*n*) de factores (loci); esta elección determina el orden del modelo para construir. *Paso 3:* los *n*-loci elegidos y sus correspondientes genotipos forman un espacio *n*-dimensional. Se realiza el conteo de casos y controles en cada celda. *Paso 4:* cada celda del espacio *n*-dimensional es etiquetada como alto riesgo (celdas gris oscuro) o bajo riesgo (celdas gris claro) dependiendo de si su razón –casos controles– supera o no un determinado umbral T (p. ej., T=1). De esta manera se reduce el modelo *n*-dimensional a una sola dimensión. *Paso 5:* todas las posibles combinaciones de *n* factores son evaluadas según su capacidad de clasificar correctamente los casos y controles utilizando el *set* de entrenamiento. De esta evaluación surge el mejor modelo de orden *n*. *Paso 6:* el modelo elegido en el paso 5 es evaluado para clasificar los datos del *set* de testeo y así se obtiene el error del predicción del modelo. Para cumplir con la validación cruzada, los pasos 1 al 6 se repiten 10 veces con el *set* de datos dividido en 10 diferentes *sets* de entrenamiento y testeo.

5. Aprendizaje supervisado se refiere a la situación en la cual se conocen los valores de las variables independientes (entradas) así como también los de la variable dependiente (salida) de los individuos que conforman un set de datos. En este caso, el algoritmo de aprendizaje se encarga de inferir la relación entre la salida y las entradas; esto es aprender a utilizar las entradas para predecir los valores de la salida.

exhaustiva se utiliza en la aplicación; esto hace que la intensidad computacional del MDR sea notable. La intensidad computacional no está afectada notablemente por el tamaño de la muestra, pero sí depende ampliamente del número de atributos, N, y el orden del modelo, k. Debido a esto, la complejidad computacional se vuelve monumental cuando se desea trabajar con más de 10 loci. Como se puede advertir, el método MDR no escapa completamente a la "maldición de la dimensionalidad"; sin embargo, son muchos los ajustes y variantes aplicados al algoritmo para solucionar este efecto indeseado.

CONCLUSIONES

La epistasis es una importante fuente de complejidad en el mapa genotipo-fenotipo y requiere métodos computacionales especiales para su análisis. En este artículo se ha presentado brevemente un método de gran alcance para el análisis de atributos llamado reducción de dimensionalidad multifactorial, que se destaca para detectar interacciones no lineales en estudios genéticos de enfermedades humanas comunes.

La genética humana y la epidemiología se encuentran en la era de la genómica (acceso a toda la información en el genoma), lo cual aumenta cada vez más su dependencia de la informática (bioinformática), para lograr un análisis pertinente y una adecuada interpretación de las enormes bases de datos.

La minería de datos y el descubrimiento de conocimiento en bases de datos desempeñarán un papel cada vez más importante en la genética humana. A medida que esta disciplina se aleje del enfoque de estudios de asociación de gen candidato hacia el enfoque de estudios de asociación de genoma completo, el conocimiento biológico experto será cada vez más importante para el desempeño exitoso de la minería de datos y es por esto que los algoritmos de aprendizaje deberán adaptarse para explotar esta valiosa información.

Agradecimientos: los autores agradecemos la colaboración del Ingeniero Nicolás Quiroz por su invalorable ayuda en la realización de los gráficos de la presente monografía.

Conflictos de interés: Los autores declaran no tener conflictos de interés.

BIBLIOGRAFÍA

- Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003;19(3):376-82.
- Mei H, Cuccaro ML, Martin ER. Multifactor dimensionality reduction-phenomics: a novel method to capture genetic heterogeneity with use of phenotypic variables. Am J Hum Genet. 2007;81(6):1251-61.
- Moore JH. Detecting, characterizing, and interpreting nonlinear gene-gene interactions using multifactor dimensionality reduction. Adv Genet. 2010;72:101-16.
- Moore JH, Gilbert JC, Tsai CT, et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol. 2006;241(2):252-61.
- Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am J Hum Genet. 2009;85(3):309-20.
- Mustavich L. The use of multifactor dimensionality reduction to detect epistasis among potential causal genes of alcoholism. [Internet]. Disponible en: http://www.gersteinlab.org/courses/545/07-spr/proj/proj_rpt.Laura.pdf. [Consulta: 16/05/2014].