Análisis de asociaciones en escenarios de datos masivos

Ingrid Teich, Ana María Planchuelo y Mónica Balzarini
Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, CONICET, Córdoba, Argentina

Resumen

La cantidad de datos en estudios biológicos se ha incrementado dramáticamente en los últimos años debido a la incorporación de nuevas fuentes y capacidades de procesamiento. Los secuenciadores automáticos del ADN y los microarreglos han permitido obtener grandes volúmenes de información a nivel molecular. Asimismo, los sensores remotos y los Sistemas de Información Geográfica proveen muchos datos a escala ecosistémica. Para estudiar a los organismos en forma integrada es necesario asociar los datos producidos por diferentes tecnologías. En biología de sistemas es esencial integrar información genómica, transcriptómica, proteómica, fenómica y ambiental. Modelar estadísticamente dichas interacciones es dificultoso desde el punto de vista computacional y biológico, sin embargo, los métodos algorítmicos filtran las principales señales de la información, permitiendo estudiar asociaciones, y el posterior modelado estadístico. En este trabajo se presentan distintas técnicas multivariadas que pueden ser usadas para comparar el nivel de covariación de distintos sets de datos entre situaciones o tratamientos y encontrar asociaciones entre ellos. Se describen e ilustran con datos de especies de importancia las técnicas de Análisis de Correlaciones Canónicas, Procrustes, Regresión por Cuadrados Mínimos y mapas Autorganizativos.

Abstract

The amount of available information in biological studies has increased dramatically in the last years. Technologies as DNA sequencing and microarrays have swamped data bases. Additionally, remote sensing and Geografical Information Systems provide a constant influx of environmental data difficult to handle. Concomitantly, the need to study organisms as a whole and find associations between data sets of different nature has increased. In System Biology, it is essential to analyze genotype, phenotypes, environment and the interactions among them. Statistical models of such complex interactions are difficult for both computational and biological interpretation aspects. On the contrary, algorithmic methods can be used to filter the main signals of genetic data and to study the association between sets of genotypic, phenotypic and environmental data. These methods are more straightforward and provide meaningful insight for later statistical modeling. In this work we present different multivariate techniques that can be used to compare the level of covariation of different data sets between situations or treatments and to find associations. In this work, we propose the use Cannonical Correlation Analysis, Procrustes, Partial Least Squares and Self Organizing Maps for the association of different data types and and illustrate their application in agronomical important species.

Palabras Clave

Correlaciones Canónicas, Procrustes, PLS, Mapas Auto-organizativos.

Introducción

El avance en el desarrollo de diferentes tecnologías e instrumental para la obtención automatizada de datos sobre distintas características de los organismos vivos ha colapsado las capacidades de almacenamiento y procesamiento para la obtención de información de la cual se pueda derivar conocimientos socialmente relevantes. Los datos generados por la biotecnologías "ómicas", como la genómica, transcriptómica, metabolómica, proteómica y fenómica, se caracterizan por su masividad y heterogeneidad en tipo o naturaleza de variables, generando nuevos desafíos no sólo computacionales sino también metodológicos para su análisis [1].

Las investigaciones que realizan interacciones o asociaciones entre diferentes datos ómicos mediante modelos estadísticos, son de difícil operatividad tanto desde el punto de vista computacional como desde el biológico. Por el contrario, los métodos basados en algoritmos computacionales, más que en la teoría probabilística clásica, permiten filtrar las principales señales contenidas en estas bases de datos multidimensionales. Los estudios de asociación entre grupos de variables son de particular interés; por ejemplo la asociación entre datos fenómicos y datos genómicos permite identificar regiones genómicas responsables de la variabilidad observada [2], así como, la asociación entre perfiles metabolómicos y transcriptómicos permite identificar rutas metabólicas involucradas en las respuestas de los seres vivos [3]. En este contexto, los algoritmos computacionales con alta capacidad de visualización de los agrupamientos de datos multivariados, permiten identificar relaciones importantes, que son confirmadas en posteriores etapas de modelación estocástica.

Las técnicas de reducción de dimensión (TRD) permiten explorar las relaciones existentes entre las observaciones multidimensionales mediante ordenaciones de las mismas sobre planos que, bajo distintos criterios de representación, son óptimos para ordenar las observaciones y analizar interdependencias. En esencia, los métodos de ordenación extraen sucesivos componentes desde una matriz de similitudes (o distancias) entre las observaciones o casos en estudio calculada a partir de múltiples variables. Esos componentes son usados como ejes para la representación gráfica de los objetos. En la ordenación, cada unidad de estudio es ubicada sobre uno o más ejes tal que su posición geométrica relativa refleja las similitudes y/o distancias entre ellos [4]. Generalmente las TRD son usadas con fines exploratorios y no requieren de supuestos sobre distribuciones probabilísticas. No obstante, existen TRD específicamente diseñadas para evaluar las correlaciones entre ordenaciones o cuantificar la magnitud de las asociaciones entre grupos heterogéneos de variables. Los métodos de agrupamiento multivariados también permiten identificar estados potencialmente relacionados mediante la generación de grupos. Los algoritmos de clasificación permiten analizar las clases existentes y explorar asociaciones entre distintos grupos de variables, que pueden servir de base para interpretaciones de relaciones taxonómicas y filogenéticas de especies vegetales como las que han llevando a cabo en el género *Lupinus* [5,6]. Más allá de los algoritmos de conglomerados jerárquicos y no jerárquicos, las ciencias ómicas se nutren de métodos bioinformáticos basados en redes neuronales e inteligencia artificial. Estos métodos han probado ser adecuados para manejar grandes dimensiones como las que se generan a partir de un alto volumen de datos y evidenciar, al mismo tiempo, patrones de relaciones ocultas en los mismos [7]. En éste trabajo se ilustran distintas aproximaciones metodológicas útiles para conducir estudios de asociación en escenarios de masiva cantidad de datos biológicos de distinta naturaleza.

Elementos del Trabajo y Metodología

El principal concepto que subyace a la aplicación de métodos bioestadísticos y bioinformáticos con distintos grupos de datos multidimensionales es que el consenso entre clasificaciones realizadas sobre las entidades biológicas en estudio, bajo los distintos tipos de datos, es una forma de cuantificar correlaciones en sentido multivariado. Entre los principales algoritmos computacionales desarrollados para el análisis de asociaciones multivariadas, aparecen frecuentemente en el análisis de datos ómicos, TRD como el análisis de correlaciones canónicas (ACC), la regresión por mínimos cuadrados parciales (PLS), el análisis de procrustes generalizado (GPA) y más recientemente, métodos de

inteligencia computacional como los Mapas auto-organizativos (SOM). Varios algoritmos de reducción de dimensión pueden ser aplicados en Info-Gen [8]. Entre los software de uso libre, el R (www.r-project.org) es uno de los principales desarrollos para implementar variantes específicas de tales métodos multivariados y bioinformáticos.

Análisis de Correlaciones Canónicas (ACC). La técnica de ACC es probablemente una de las primeras desarrolladas para evaluar la asociación entre grupos de variables en hiperespacios [9]. Se utiliza para determinar la correlación lineal entre dos grupos de variables métricas. El ACC permite identificar y cuantificar la asociación entre conceptos que no son medidos directamente sino a través de múltiples descriptores, como por ejemplo el metaboloma y el genoma, donde cada uno de ellos es explorado por un número, que a pesar de ser grandes (ej: marcadores, transcriptos o metabolitos), no describen individualmente el concepto. El ACC provee una medida de correlación entre una combinación lineal de las variables en un conjunto, con una combinación lineal de las variables en el otro conjunto. En un primer paso del análisis, se determina el par de combinaciones lineales con máxima correlación. En un segundo paso, se identifica el par con máxima correlación entre todos los pares no correlacionados con el par de combinaciones seleccionadas en el primer paso y así sucesivamente. Las combinaciones lineales de un par son llamadas variables canónicas y la correlación entre ellas, es llamada correlación canónica. La primera correlación canónica nunca es menor que la mayor de las correlaciones múltiples entre cualquier variable y otra del grupo opuesto. Podría pasar que la primera correlación canónica sea muy alta mientras que todas las correlaciones múltiples para predecir una variable desde el conjunto opuesto sean pequeñas. El ACC asume correlación del tipo lineal, otras correlaciones pueden pasar desapercibidas y distorsionar el análisis. La incorporación y eliminación de variables puede modificar sustancialmente el análisis al igual que la presencia de puntos influyentes. Técnicas de diagnóstico comunes en el análisis de regresión pueden ser utilizadas para la identificación de puntos influyentes. El número de correlaciones canónicas que puede ser extraído desde estas descomposiciones es igual al mínimo de los números p y q (cardinalidad de cada uno de los conjuntos de variables que se desean correlacionar). Los coeficientes de correlación canónica al cuadrado representan la proporción de la varianza total explicada por cada variable canónica. Usualmente se reporta bajo el nombre de "estructura canónica total" a las correlaciones simples entre las variables respuestas y las variables canónicas.

Las correlaciones canónicas han sido utilizadas para asociar datos de variabilidad genética en especies vegetales con datos ambientales y morfológicos. Sork et al. [10] realizaron estudios de asociación entre datos genómicos (microsatélites) y climáticos en bosques de en el hemisferio norte encontrando una asociación entre los alelos obtenidos y variables climáticas. Souto & Smouse [11] realizaron asociaciones entre datos genéticos (isoenzimas), ambientales y morfológicos mediante ACC para poblaciones de *Embothrium coccineum* de los bosques templados de la Patagonia, encontrando que la variación genética y morfológica se relaciona con la variabilidad climática y el aislamiento geográfico. En éste trabajo se ilustran los resultados obtenidos por Teich et al. [12], quienes utilizando datos provenientes de sensores remotos, encontraron que sitios más inestables albergan mayores niveles de diversidad genética.

Análisis de Procrustes Generalizado (GPA). Cuando las observaciones son caracterizadas mediante $k \geq 2$ conjuntos de variables (ej: descriptores moleculares y fenotípicos) puede ser de interés obtener una ordenación para cada conjunto, donde las variables pueden ser de

igual o diferente naturaleza, y luego, consensuar las ordenaciones obtenidas para lograr una única configuración (la ordenación en el espacio del consenso). La cuantificación del consenso mediante análisis procrustes generalizado (APG) provee información acerca de la armonización o adecuación de las configuraciones producidas por cada conjunto de variables. El porcentaje de consenso entre ambas ordenaciones es una medida univariada de la asociación entre ambos grupos de variables. El APG se basa en rotaciones, traslaciones y escalamientos de las ordenaciones individuales para su representación en un mismo espacio (espacio de consenso). Dado que las características de mayor interés para el estudio de las especies son fenotípicas más que genotípicas, la asociación genotipo-fenotipo es objetivo principal de numerosas estudios de asociación en especies de importancia agronómica. Bramardi et al. [13] utilizaron GPA para determinar las relaciones entre genotipos via el uso simultáneo de caracteres agronómicos y datos moleculares de variedades de pepino (*Cucumis sativus* L.), pudiendo discriminar las variedades mediante la ordenación de consenso.

Regresión por cuadrados mínimos cuadrados parciales (PLS del inglés, Partial Least Squares). PLS es un método estadístico multivariado que generaliza y combina el Análisis de Componentes Principales (ACP) y el análisis de Regresión Lineal [14]. Es particularmente útil cuando se desea predecir un conjunto de variables dependientes (Y) desde un conjunto (relativamente grande y correlacionadas) de variables predictoras (X). El objetivo del método PLS es describir Y a partir de X y su estructura de variación común. Cuando hay más observaciones que variables predictoras y no existe multicolinealidad, la predicción de Y en función de X puede realizarse eficientemente con un análisis de regresión lineal múltiple. No obstante, PLS permite realizar el análisis aún cuando existe correlación entre las variables predictoras y/o existen más predictoras que observaciones. El problema de la estimación en estos casos se resuelve combinando linealmente las predictoras con un ACP y luego regresionando Y con un número reducido de componentes principales. Hay que recordar que las CP explican variación en X y nada nos dicen sobre la relación de Y con X. Por el contrario la técnica PLS busca una solución óptima o de compromiso entre el objetivo de explicar la máxima variación en X y encontrar las correlaciones de éstas con Y. La idea en PLS es obtener un vector de coeficientes (A_j), uno para cada variable en X, y un vector de coeficientes (B_j), uno para cada variable en Y, tal que el producto AB^T aproxime bien a la matriz de relaciones entre variables en el sentido mínimo cuadrático. Podría decirse que estos coeficientes permiten combinar las variables de cada conjunto para explicar la variabilidad debida a la relación o correlación entre ambos bloques de variables. Una aplicación clásica de PLS es extender la regresión múltiple cuando existe correlación entre las predictoras o como se indicó anteriormente, cuando el número de observaciones es pequeño en relación al número de regresoras. Vargas et al. [15] usa PLS para relacionar las asociaciones entre datos genómicos expresados por cientos de marcadores moleculares con datos fenotípicos (rendimientos y sus componentes) en distintos ambientes en maíz. Crossa et al. [16] compara el método junto a otros algoritmos bioinformáticos en su performance para predecir valores genéticos en caracteres cuantitativos en mejoramiento vegetal. La diferencia entre numerosos algoritmos corrientemente usados para este fin a pesar de grande a su naturaleza es pequeña en cuanto a los resultados obtenidos.

Mapas auto-organizativos (SOM). Los SOM son un modelo de red neuronal desarrollado por Kohonen [17]. Este procedimiento procesa una base de datos o casos multidimensionales, resultando en un mapa (usualmente bidimensional) donde casos similares se "mapean" en regiones cercanas de la red neuronal. De esta manera "vecindad" significa "similaridad". La red se estructura como una capa usualmente bidimensional de nodos no conectados entre sí. Todos los nodos se asocian con un dato de entrada, se inicializan los pesos de cada nodo, se busca el nodo ganador respecto a su similitud con el caso de entrada, y se actualizan los pesos del nodo ganador y de sus vecinos, reiterando los pasos hasta que se satisface un criterio de detección impuesto previamente. Los mapas auto-organizativos constituyen un método de conglomeración similar a los métodos no jerárquicos donde los grupos que se conforman son ubicados espacialmente sobre la estructura de una red predefinida. Especialmente en cuanto a la integración de datos de diferentes tipos, Hirai et al. [18] propuso un modelo de SOM para el estudio integral del metaboloma y transcriptoma de *Arabidopsis thaliana* (especie modelo) y recientemente se ha propuesto una aplicación para la agrupación y visualización de asociaciones entre transcriptos y metabolitos en tomate [19]. Asimismo se ha desarrollado un software [20] que permite utilizar este modelo facilitando la obtención de clusters (o neuronas) y a su vez permite visualizar los agrupamientos de manera de extraer la máxima información.

Resultados

En las figuras de la 1 a la 4 se presentan las salidas obtenidas de los procesos de análisis aplicando las metodologías descriptas previamente, para conjuntos de datos ómicos de distintas naturalezas. Estos resultados sirven para ilustrar el tipo de información y conocimiento que se persiguieron en el análisis de las correlaciones multivariadas.

En la tabla y figura 1 se presenta una aplicación del ACC en un estudio de variabilidad genética espacial de bosques de una especie nativa (*Polylepis australis* Bitt.) y su asociación con características ambientales a través del rango distribucional de la especie en Argentina. Este estudio surge por el interés creciente de estudiar y conservar la biodiversidad de los bosques y comprender los factores biofísicos asociados, especialmente frente a los cambios producidos en el ambiente por el hombre y al cambio climático. En este contexto, la relación entre estabilidad ambiental y biodiversidad es de particular interés. Teich et al. [12] examinaron la asociación entre estabilidad ambiental y diversidad genética en 18 poblaciones de *P. australis*, a lo largo de su rango de distribución. Para caracterizar la estabilidad ambiental de los distintos sitios se utilizaron series temporales de 12 años (432 valores) del índice de vegetación diferencial normalizado (NDVI), el cual es un indicador de la respuesta de la vegetación a la variabilidad ambiental. Cada año fue caracterizado por distintos estadísticos (máximo, mínimo, media, y desvío estándar) de NDVI. La estabilidad ambiental a largo plazo fue caracterizada por el coeficiente de variación de dichos estadísticos entre años. Aquellos sitios con bajos coeficientes de variación interanual corresponden a zonas más estables y sitios con altos coeficientes de variación caracterizan zonas de mayor inestabilidad. La diversidad genética se calculó para cada población a partir de los perfiles moleculares de 208 árboles obtenidos de 244 bandas de polimorfismos de la longitud de fragmentos amplificados (AFLP). Resultaron significativas las dos primeras correlaciones entre combinaciones lineales de indicadores de variabilidad genética obtenida de datos genómicos (244 marcadores AFLP) e indicadores de la inestabilidad ambiental de los sitios en donde se desarrollan estos bosques, obtenidos

a partir de sensores remotos (series temporales de 432 valores de NDVI pertenecientes a un lapso de 12 años). La estructura canónica total sugiere que uno de los principales contribuyentes en explicar la variabilidad en la diversidad genética es el indicador ambiental denominado CV_{seas} el cual mide la variación a largo plazo de la estacionalidad expresada a través del Índice de vegetación normalizado (NDVI).

Tabla 1: Resumen del Análisis de Correlaciones Canónicas entre estimadores de diversidad genética y de inestabilidad ambiental

(A) Coeficientes Canónicos y valores p

Eje Canónico	Correlación canónica	valor p
1	0.66	0.0031
2	0.42	0.0349
3	0.12	0.6362

(B) Coeficientes de correlación de Pearson entre los ejes canónicos de inestabilidad ambiental y cada componente

	Eje 1_{IA}	**Eje 2_{IA}**
CV_{seas}	0.92	0.18
CV_{ave}	0.62	-0.30
Cv_{max}	0.50	0.18
CV_{min}	0.51	-0.73

(C) Coeficientes de correlación de Pearson entre los ejes canónicos de diversidad genética y cada componente

	Eje 1_{DG}	**Eje 2_{DG}**
Loci polimórficos (P)	0.98	-0.15
Shannon Weaver (SW)	0.80	0.53
Diversidad Génica (Pn)	0.96	0.26

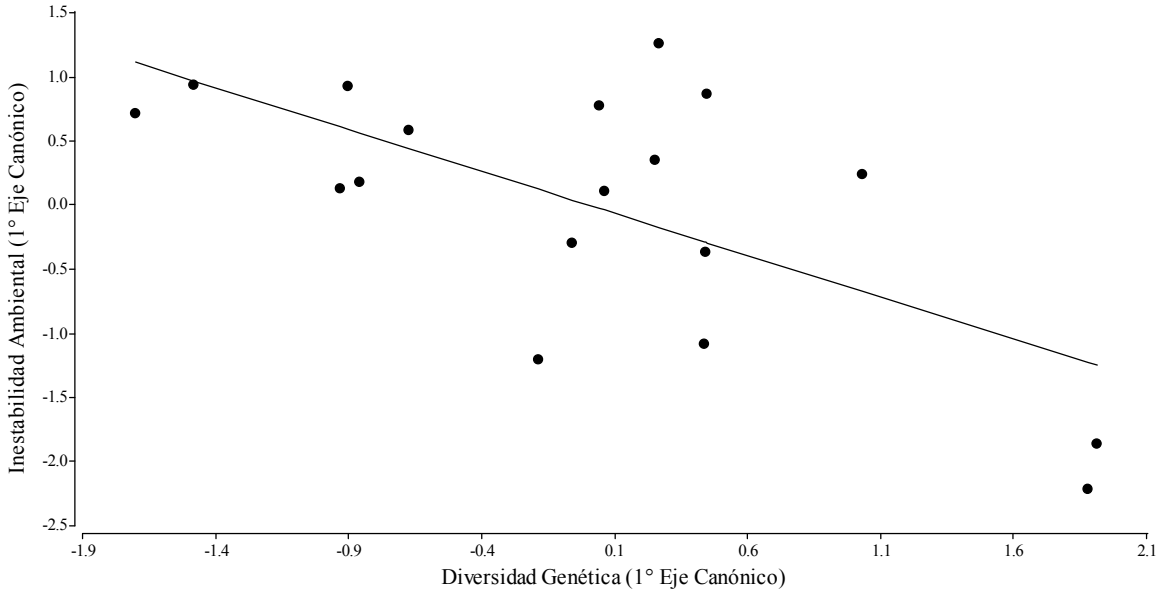

Figura 1: Correlación Canónica entre Inestabilidad Ambiental y Diversidad Genética para 18 poblaciones de Polylepis australis

En la figura 2 se esquematiza la ordenación de 7 entidades de estudio caracterizadas cada una multidimensionalmente en dos espacios según dos conjuntos distintos de variables. Usualmente, el fenotipo agronómico se evalúa a través de 3 a 5 caracteres de importancia económica como son el rendimiento y sus componentes en 5 a 10 ambientes, mientras que la dimensionalidad de los datos moleculares es aún mayor (cientos a miles de marcadores). Para ilustrar la capacidad del GPA para generar ordenaciones que consensúan configuraciones alternativas de las mismas entidades en la figura 2 mostramos la ordenación de 20 entidades por dos conjuntos de variables y la ordenación de consenso. Los datos utilizados pertencen a tres grupos, primero fueron ordenados a una TRD basada en datos genómicos y luego via otra TRD basada en datos fenotípicos. Aún cuando se conoce que existen tres grupos, ambas ordenaciones realizadas independientemente ponen en evidencia dos grupos. Sólo cuando las observaciones se ordenan consensuando las dos ordenaciones previas, emergen los tres grupos.

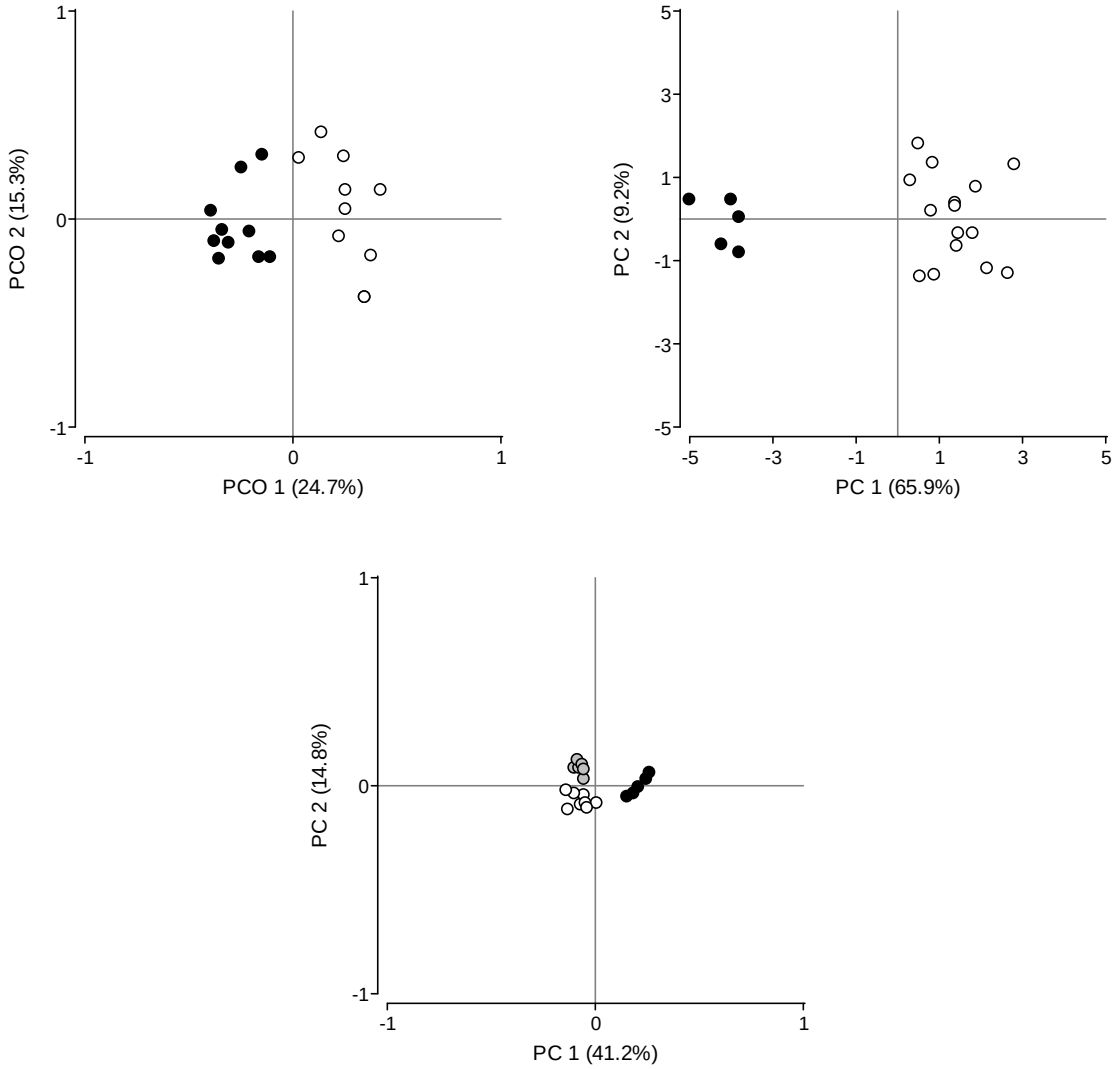

Figura 2. Ordenación en dos dimensiones de 20 entidades: A) Análisis de Coordenadas principales para datos cualitativos (genómicos) y B) Análisis de Componentes Principales para datos cuantitativos (morfológicos) C) Configuración de Consenso del Análisis de Procrustes.

En la figura 3 se reproducen los resultados obtenidos por Cirilo et al. [21] en un ensayo experimental multiambiental donde se analizó el rendimiento de grano de maíz en respuesta a la disponibilidad de nitrógeno. La técnica de PLS fue utilizada para modelar la relación entre la interacción genotipo x ambiente (G x A) y los caracteres morfo-fisiológicos de los individuos de maíz. Como variables dependientes se utilizaron estimaciones de G x A, mientras que los caracteres morfofisiológicos constituyeron el grupo de variables independientes. PLS permitió detectar aquellos ideotipos con caracteres morfofisiológicos asociados a interacciones positivas en ambientes pobres en nitrógeno. Mediante esta técnica también sería posible identificar las variables ambientales asociadas a interacciones G x A positivas para los genotipos analizados. En la figura se observan las entidades biológicas (genotipos de maíz), los distintos ambientes y las variables morfofisiológicas. Por ejemplo,

se observa que los híbridos H3 y H1 (círculos negros) presentan altas interacciones positivas en el ambiente codificado como ON (zona de Oliveros con fertilizante nitrogenado), asociados principalmente a la variable índice de cosecha (NHI), calculada como la relación entre el contenido de nitrógeno en los granos y el contenido de nitrógeno acumulado en la madurez fisiológica.

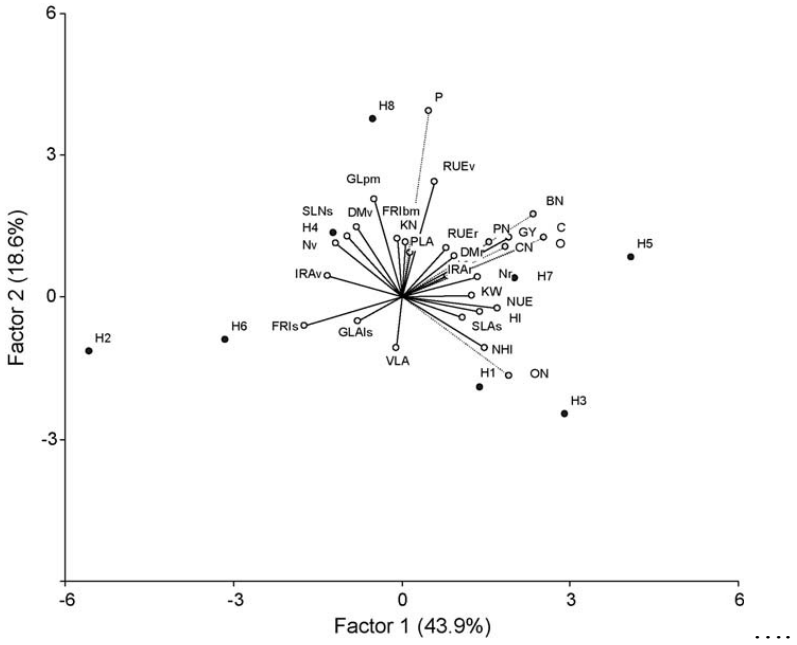

Figura 3. Resultados obtenidos por Cirilo et al. [21] en un estudio de asociación entre la interacción genotipo x ambiente (8 híbridos x 7 ambientes) con caracteres morfo-fisiológicos usados como variables predictoras en un PLS. En negro se observan los híbridos, en gris los ambientes y en blanco los caracteres morfo-fisiológicos.

En la figura 4 se reproducen los resultados obtenidos por Stegmayer et al. [19] que utilizan un SOM para integrar los perfiles transcripcionales y metabólicos obtenidos a partir de frutos de tomate de 21 líneas de introgresión (IL). Los perfiles metabólicos se obtuvieron mediante la técnica de GC-MS. Los datos de expresión de genes se obtuvieron utilizando la plataforma de TOM2 que consiste en microarreglos de 12860 clones EST que representan ~8,500 loci independientes de tomate. Mediante la técnica de SOM pudieron encontrar grupos con transcriptos y metabolitos asociados, que permitieron estudiar las rutas metabólicas en tomate.

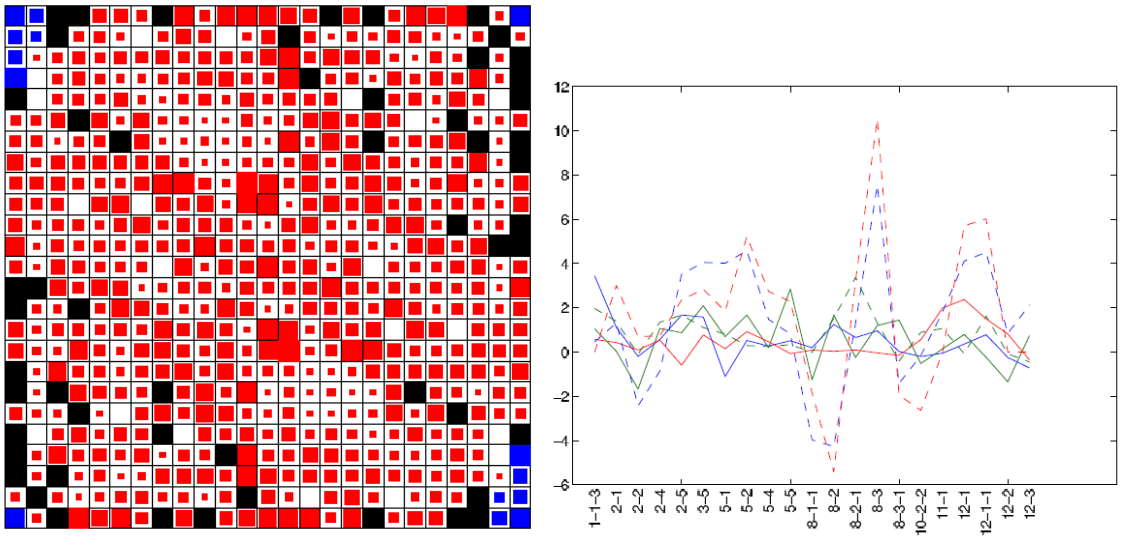

Figura 4: Resultados obtenidos por Stegmayer et al. [19] (a) Mapa SOM obtenido con los perfiles transcripcionales y metabólicos de 21 líneas de tomate utilizando el programa omeSOM. Los recuadros negros indican clusters con meabolitos y transcriptos, los rojos con sólo transcriptos y los azules con metabolitos y transcriptos. Los blancos son vacíos. El tamaño del recuadro del color indica la cantidad de componentes agrupados. A mayor tamaño mayor cantidad de componentes. (B) patrones de tres metabolitos (líneas sólidas) y tres transcriptos (líneas punteadas) clasificados en un mismo nodo (nodo 625).

Discusión

Los análisis multivariados proveen herramientas útiles para estudiar asociaciones entre bases de datos de distinta naturaleza y presentan ventajas respecto a otros procedimientos estadísticos. En primer lugar, no requieren muchos supuestos, incluso se benefician de estructuras de correlación fuertes entre los variables. Esta es la situación en la mayor cantidad de estudios biológicos y agronómicos, en los que las variables suelen no ser independientes. En segundo lugar, las técnicas de reducción de la dimensión, ofrecen la posibilidad de resumir información de muchos rasgos en pocas variables sintéticas. La claridad que emerge de esta propiedad, es invaluable en estudios de asociación. Sin embargo, para proveer información confiable, es necesario aplicarlas adecuadamente. La pertinencia de cada método dependerá de los objetivos planteados como también en la naturaleza de los datos y la existencia de otras covariables.

La masiva cantidad datos que proveen las nuevas biotecnologías provee constantemente conjuntos de datos disponibles para obtener información relevante. La secuenciación del ADN, mediante el desarrollo de secuenciadores automáticos, ha permitido la difusión de más de 1000 genomas completos de diversos organismos y ha generado cientos de billones de datos genómicos (pares de bases) depositados en GenBank y accesibles via Internet. Los microarreglos han revolucionado la caracterización de los perfiles transcriptómicos permitiendo conocer la regulación y expresión del genoma bajo distintas condiciones y estados. La proteómica y la metabolómica exploran la consecuente funcionalidad y expresión en los organismos generando a la vez, grandes volúmenes de datos disponibles. La fenómica, como nivel último de expresión, sintetiza la variabilidad de

las regulaciones que operan a escalas más finas, permitiendo estudiar el impacto de cambios genómicos o ambientales sobre el fenotipo. Asociar los datos provistos por dichas tecnologías ómicas es una tarea multidisciplinaria a la cual la estadística multivariada aporta valiosas herramientas.

Referencias

[1] Balzarini M., Bruno C., Fernandez E. (2011) Multivariate Analysis in Phytopathology: Options and opportunities in data mining to face new molecular information. Phytopathology in the omics era. Editorial Singpost Research , México.

[2] Houle D., Govindaraju DR , Omhol S. (2010) Phenomics: the next challenge. Nature Reviews Genetics. 11 (855-865).

[3] Carrari F., Baxter C., Usadel B., Urbanczyk-Wochniak E., Zanor M.I., Nunes-Nesi A., Nikiforova V., Centeno D., Ratzka A., Pauly M., Sweetlove L., Fernie A.R. (2006) Integrated Analysis of Metabolite and Transcript Levels Reveals the Metabolic Shifts that Underlie Tomato Fruit Development and Highlight Regulatory Aspects of Metabolic Network Behavior. Plant Physiology 142(4):1380-1396.

[4] Gabriel K. (1971) The biplot graphic display of matrices with application to principal component analysis. Biometrika 58:453-467.

[5] Merino E. F., Maestri D., Planchuelo A.M. (1999) Chemotaxonomic evaluation of leaf alkanes inspecies of *Lupinus* (Leguminosae). Bioch. Sist. & Ecology 27 (3): 297-301.

[6] Planchuelo A.M., Fuentes E. (2001) Taxonomic evaluation and new combinations in *Lupinus gibertianus-L. linearis* complex (Fabaceae). NOVON 11:442-450.

[7] Fernandez E., Balzarini M. (2007) Improving cluster visualization in Self-Organizing Maps: Application in Gene Expression Data Analysis. Computers in Biology and Medicine 37(3):1677-1689.

[8] Balzarini M., Di Rienzo J. (2004) InfoGen. Statistical software for genetic data. Universidad Nacional de Córdoba, Argentina.

[9] Hotelling H. (1936) Relations Between Two Sets of Variables," Biometrika 28, 321-377.

[10] Sork V.L., Davis F.W, Westfall R., Flint A., Ikegami M., Wang H.F., Grivet D. (2010) Gene movement and genetic association with regional climate gradients in California valley oak (Quercus lobata Née) in the face of climate change. Molecular Ecology 19: 3806–3823.

[11] Souto C., Smouse P. (2011) Patrones concordantes de variación genética y morfológica a lo largo del paisaje en poblaciones de *Embothrium coccineum* de los bosques templados de la Patagonia. III Jornadas Argentinas de Ecología del Paisaje, Bariloche.

[12] Teich I., García C., Swinnen E., Tote C., Hensen I., Balzarini M. Genetic diversity in a changing world: stability matters. Enviado a Molecular Ecology.

[13] Bramardi S., Bernet G., Asíns M., Carbonell E. (2005) Simultaneous Agronomic and Molecular Characterization of Genotypes via the Generalized Procrustes Analysis: An Application to Cucumber. *Crop Sci* 45(4):1603-1609.

[14] Abdi H. (2003) Partial least squares regression (PLS-regression). In M. Lewis-Beck, A. Bryman, T. Futing (Eds): Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage. pp. 792-795.

[15] Vargas M., van Eeuwijk F.A., Crosa J., Ribaut J.M (2006) Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods TAG. Theoretical and Applied Genetics 112 (6): 1009-1023.

[16] Crossa J., de los Campos G., Pérez P., Gianola D., Burgueño J., Araus J.L., Makumbi D., Singh R.P., Dreisigacker S., Yan J., Arief V., Banziger M., Braun H.J. (2010) Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186:713-724.

[17] Kohonen T. (1997) Self-Organization Maps, second ed., Springer, Berlin

[18] Hirai M.Y., Yano M., Goodenowe D.B., Kanaya S., Kimura T., Awazuhara M., Arita M., Fujiwara T., Saito K. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 101: 10205–10210.

[19] Stegmayer G., Milone D., Kamenetzky L., López M., Carrari F. (2009) Neural network model for integration and visualization of introgressed genome and metabolite data. Proceedings of the IJCNN'09.

[20] Milone D., Stegmayer G., Kamenetzky L., López M., Lee J.M., Giavanonni J.J., Carrari F. (2010) omeSOM: a software for clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. Bioinformatics 11:438.

[21] Cirilo A.G., Dardanelli J., Balzarini M., Andrade F.H., Cantarero M., Luque S,. Pedrol H.M. (2009) Morpho-physiological traits associated with maize crop adaptations to environments differing in nitrogen availability. Field Crops Research 113:(2)116-124.

Datos de Contacto

Ingrid Teich. Centro de Relevamiento y Evalución de Recursos Agrícolas y Naturales, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba.. Av. Valparaíso s/n CC. ingridteich@gmail.com.