Análisis de conglomerados en la identificación de estructura genética a partir de datos de marcadores moleculares

Cluster analysis for identification of genetic structure from molecular marker data

Peña Malavera, Andrea;[I,II] Bruno, Cecilia;[I,III] Teich, Ingrid;[I,III]
Fernández, Elmer;[III,IV] Balzarini, Mónica.[I,III]

Resumen. En el contexto de abundante información genómica, como la producida a partir de marcadores moleculares basados en ADN, es de interés identificar la estructura genética subyacente en un conjunto de individuos, previo al análisis de asociación entre expresión de marcadores y fenotipo. Cuando existen subgrupos de individuos que difieren sistemáticamente en las frecuencias alélicas de sus marcadores, se origina una estructura genética que, de no ser considerada, incrementa el riesgo de detectar asociaciones espurias entre marcadores y fenotipo. Diversos métodos estadísticos son utilizados para determinar la agrupación de individuos desde datos de marcadores moleculares que producen información discreta multidimensional, entre ellos métodos basados en algoritmos de conglomerados jerárquicos (UPGMA), conglomerados no jerárquicos (K-means), redes neuronales como los mapas auto-organizativos (SOM) y métodos de conglomerados bayesianos. En este trabajo comparamos la capacidad de tales algoritmos para detectar subpoblaciones (conglomerados genéticos) bajo dos escenarios biológicos de estructura poblacional: modelo de islas y modelo de contacto. Los algoritmos de conglomerado fueron evaluados simultáneamente usando conjuntos de datos de marcadores moleculares de expresión binaria simulados bajo ambos modelos biológicos. El método de conglomeración bayesiano fue el que mejor identificó, entre los evaluados, las subpoblaciones simuladas bajo el modelo de migración de islas. Para el modelo de contacto la identificación de subgrupos fue difícil con cualquiera de los cuatro algoritmos de conglomeración evaluados.

Palabras clave: conglomerados jerárquicos, conglomerados no jerárquicos, mapas auto-organizativos, conglomerado bayesiano, modelos de migración.

Abstract. Prior to association studies, and in the context of abundant genomic information provided by molecular markers, it is of interest to identify the underlying genetic structure of individuals. Genetic structure arises when markers´ allele frequen-

I Estadística y Biometría-FCA, UNC
II FONCyT
III CONICET Córdoba- Argentina
IV Universidad Católica de Córdoba. Correo electrónico: andreapema@gmail.com

cies differ systematically between subgroups, and if it is not considered in association analysis, it increases the risk of detecting spurious associations between molecular markers and the phenotype of interest.

A variety of statistical methods are used to determine groups of individuals from molecular markers that produce multidimensional discrete data, such as methods based on hierarchical (UPGMA) and non-hierarchical clustering algorithms (K-means), neural networks (SOM), and Bayesian clustering. In this study, we compared the capacity of these algorithms to detect genetic clusters under two different biological scenarios: the island model and the contact model. The clustering algorithms were simultaneously evaluated using binary molecular marker data simulated under both biological scenarios. Bayesian clustering was the best model to identify subpopulations under the island migration model. However, in the contact model the identification of subgroups was difficult with all algorithms.

Key words: Hierarchical clusters, Non-hierarchical clusters, self-organizing maps, Bayesian clustering, Migration models.

1. INTRODUCCIÓN

En genética humana, como en genética de animales y plantas, es de interés identificar loci del ADN que rigen caracteres complejos (generalmente de expresión continua). El estudio de asociación se realiza en un contexto de abundante información genómica como es la producida por distintos tipos de marcadores moleculares del ADN usando individuos seleccionados al azar de una población con distinto nivel de relación de parentesco y estructuración genética. El procedimiento para el análisis simultáneo de información genómica y fenotípica con el propósito de identificar asociaciones no aleatorias entre loci de marcadores moleculares y loci del carácter fenotípico en estudio, es conocido como mapeo de asociación o mapeo por desequilibrio de ligamiento. Para este procedimiento se han utilizado con éxito distintos modelos estadísticos, los cuales deben contemplar si existen estructuras genéticas subyacentes en la población de estudio, es decir la posible existencia de subpoblaciones con distinto nivel de relación genómica. Diversos métodos analíticos computacionales de naturaleza exploratoria multivariada e índices de diversidad genética entre y dentro de grupos pueden usarse para detectar estructura genética (Hedrick, 2005).

Los estudios de asociación, suponen que si una mutación o cambio en el estado alélico del marcador incrementa la expresión de una característica en una población de individuos, entonces se puede esperar que el o los alelos asociados a tal característica sean más frecuentes entre los individuos que la comparten que entre el resto de los

individuos. Una población estructurada genéticamente es aquella en la que coexisten subgrupos o conglomerados de individuos que difieren sistemáticamente en sus frecuencias alélicas para los diferentes loci. Cuando se lleva a cabo un análisis de asociación en estas poblaciones sin considerar los efectos de la subestructura poblacional subyacente, se aumenta el riesgo de detectar asociaciones espurias entre marcadores y la característica fenotípica de interés.

El análisis estadístico de la información molecular a través de algoritmos de clasificación no supervisada, se utiliza para revelar tales estructuras de subpoblaciones que luego son usadas como un factor de clasificación dentro de los modelos de asociación. Diversos algoritmos estadístico-computacionales para la agrupación de entidades en espacios multidimensionales, tales como los métodos de conglomerados jerárquicos (UPGMA), conglomerados no jerárquicos (K-means), métodos de conglomeración bayesianos basados en cadenas de Markov Monte Carlo (MCMC), y redes neuronales que producen mapas auto-organizativos (SOM), pueden aplicarse a datos multialelos-multilocus para identificar subpoblaciones o conglomerados genéticos. El objetivo de este trabajo es ilustrar el desempeño relativo de estos métodos bajo distintos escenarios biológicos de presencia de subpoblaciones genéticas caracterizados por modelos de migración comunes en estudios de ecología y genética de poblaciones.

2. MATERIALES Y MÉTODOS

2.1 Datos

Se simularon datos genéticos multivariados del tipo binario con una estructura poblacional conocida a priori bajo dos escenarios biológicos caracterizados por diferentes modelos de migración: (A) modelo de islas y (B) modelo de contacto. El programa utilizado fue EASYPOP versión 2.0.1 (Balloux, 2001). Para el modelo de islas se consideró una estructura poblacional compuesta por 5 grupos, mientras que para el modelo de contacto se simuló una estructura poblacional de 2 grupos con 5 subpoblaciones cada uno (figura 1). El modelo de islas es considerado básico porque resume la estructura poblacional fundamental (Wright, 1931), y es utilizado en muchos estudios de genética de poblaciones para predecir procesos evolutivos causados por deriva génica, mutación, selección y migración (Chiappero et al., 2010). Este modelo asume que los migrantes tienen una frecuencia génica igual a la del conjunto de las subpoblaciones, y es considerado matemáticamente simple. El modelo de contacto se considera opuesto al de islas en el continuo de estructuras poblacionales. En él la

población no está dividida en subunidades donantes o receptoras de migrantes, ni es una unidad panmíctica. Los cruzamientos al azar están limitados por la distancia, de modo que los individuos tendrán una mayor probabilidad de aparearse con vecinos que con individuos más lejanos. De este modo se pueden agrupar a los individuos en «vecindarios», áreas definidas por «individuos centrales» cuyos progenitores se pueden tratar como extraídos al azar. La representación más sencilla del modelo de contacto es un hábitat lineal a lo largo del cual existe una distribución normal de las distancias entre los lugares de nacimiento de los padres y la progenie.

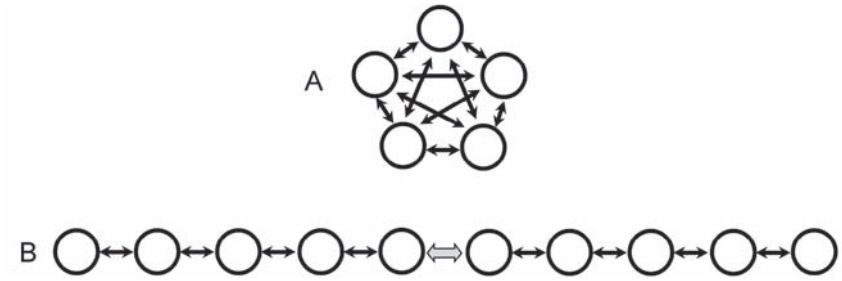

Figura 1. Representación esquemática de los dos modelos biológicos de migración simulados. A) Modelo de islas. B) Modelo de zona de contacto.

Los parámetros de simulación fueron fijados en este estudio con la finalidad de lograr datos moleculares que caractericen un conjunto de líneas o individuos genéticamente estructurados, y obtener escenarios comparables a la mayor parte de los estudios que se realizan. Se definieron los siguientes parámetros: nivel de ploidía (diploide), proporción de recombinación (0.01), número de individuos por población (100), número de loci (150), tasa de mutación (0.01), modelo de mutación (modelo de mutación paso a paso -SSM-), número de posibles estados alélicos (2), variabilidad de la población inicial (mínima), tasa de migración dentro de poblaciones (0.01), tasa de migración entre poblaciones (0.001). Todas las simulaciones fueron realizadas para 4000 generaciones.

El hecho de considerar una tasa de migración pequeña entre y dentro de poblaciones, significa que en cada generación existirá una baja proporción de la población que será sustituida por inmigrantes, tanto dentro de una misma población como entre poblaciones. Se espera que este modelo de migración provoque una alta endogamia local y elevada homocigosis entre individuos de una misma población, y por tanto mayor diferencia entre poblaciones, es decir que las poblaciones estén estructuradas genéticamente. Cada escenario o realización del modelo genético para los parámetros

expuestos anteriormente se simuló 100 veces y de cada conjunto de datos simulado se seleccionaron por muestreo aleatorio simple para retener en el análisis entre 120 y 180 individuos con igual proporción de representación de cada una de las subpoblaciones simuladas, en cada muestra emulada. El valor de la cantidad de líneas o individuos seleccionados fue elegido en función de los tamaños de muestras que se utilizan en la práctica en estudios de mapeo de asociación.

2.2 Procedimientos evaluados

Sobre cada conjunto de datos simulado se aplicaron simultáneamente 4 algoritmos de conglomerados: 1. Método jerárquico (UPGMA), 2. Método no jerárquico K-means, 3. Método basado en redes neuronales (SOM) y 4. Método bayesiano basado en MCMC.

En el método de conglomerados jerárquico UPGMA (Sokal y Michener, 1958), como en otros métodos de conglomerados de esta naturaleza, se parte de una matriz de distancias conteniendo todas las distancias entre pares de objetos, y se los comienza a agrupar teniendo en cuenta la mínima distancia. Luego de agrupar el primer par de individuos el proceso continua recalculando la matriz de distancias de manera tal que el conglomerado recientemente formado se trata como un nuevo objeto. Por esto es necesario definir una métrica de distancia entre objetos simples y conglomerados o entre conglomerados. Ésta se obtiene promediando todas las distancias entre pares de objetos, donde un miembro del par pertenece a uno de los conglomerados y el otro miembro al segundo conglomerado. Este es un método simple que se ha encontrado exitoso en numerosas aplicaciones de clasificación. La expresión para calcular la distancia promedio entre conglomerados es:

donde es la distancia entre el objeto, que pertenece al conglomerado AB y el objetoque pertenece al conglomerado C, siendo la sumatoria sobre todos los posibles pares de objetos entre dos conglomerados y donde y son los números de objetos en los conglomerados AB y C respectivamente. El método tiende a producir grupos de igual varianza (Milligan, 1980). La historia de formación de conglomerados en los distintos pasos del algoritmo de agrupamiento suele representarse gráficamente a través de un dendrograma.

El método de conglomerados no-jerárquico (K-means) agrupa objetos en k grupos haciendo máxima la variación entre conglomerados y minimizando la variación dentro de cada conglomerado. Este método comienza con un agrupamiento inicial o

con un grupo de puntos semilla (centroides) que formarán los centros de los grupos (partición inicial del grupo de objetos en k ítems), prosigue asignando cada objeto al grupo que tiene el centroide (media) más cercano. La partición lograda es aquella tal que la suma de la suma de las distancias al cuadrado de los miembros del grupo respecto a su centroide es mínima. El método se basa así en el principio de los k mejores centroides, éstos son modificados cada vez que un objeto se transfiere de un grupo al otro. El algoritmo K-means es óptimo en cada paso. Los resultados finales dependen de la configuración inicial, de la secuencia en que son considerados los objetos a agrupar y claramente del número de grupos. A los fines de alcanzar un óptimo global, es recomendable usar varias particiones iniciales y seleccionar aquella partición final con mínimo valor de la función objetivo.

Si bien estos métodos basados en conglomerados jerárquicos y no-jerárquicos se encuentran implementados en numerosos softwares, en este trabajo usamos la implementación del software Info-Gen (Balzarini y Di Rienzo, 2004). La función objetivo que se minimiza en Info-Gen es la suma de distancias al cuadrado entre los objetos de un mismo conglomerado. El gráfico que muestra la evolución de esta función objetivo para 2, 3,…, k conglomerados es usado para identificar el número de grupos subyacentes en la población de objetos, i.e. cantidad de grupos donde la función decrece a una menor tasa.

Los Mapas Auto-Organizativos o Self-Organizing Maps (SOM) son un modelo de red neuronal desarrollado por Kohonen (Kohonen, 1997). 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" (Fernández y Balzarini, 2007). 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. Si bien originalmente fue implementado en Sompack (Kohonen, 1997) actualmente se puede implementar en los softwares Matlab, SAS y R, con distintos métodos de representación gráfica de los resultados. En el presente trabajo se usa el método de representación "U-matrix" disponible en Sompack.

El método de conglomeración bayesiano, basado en Cadenas de Markov de Monte Carlo propuesto por Pritchard y sus compañeros (2000) es frecuentemente usado para estudios de estructura genética poblacional. Se estima la distribución a posteriori de coeficientes asociados a cada individuo que se corresponden a los distintos subgrupos en los que éste puede clasificarse. El valor esperado de la distribución a posteriori provee una estimación de la proporción que el genoma de un individuo tiene o comparte con los distintos subgrupos. El algoritmo se encuentra implementado en el software STRUCTURE 2.2.3 (Pritchard et al., 2000).

2.3 Evaluación del desempeño de los procedimientos de conglomerados

Bajo cada escenario de modelo biológico de migración usado para simular estructura genética se contabilizó el porcentaje de corridas donde el algoritmo de conglomerado usado detectó correctamente el número de grupos o subpoblaciones simuladas (K). Adicionalmente se estimó para el conglomerado jerárquico UPGMA, no jerárquico K-means y el método bayesiano, el porcentaje de entidades bien clasificadas a través de la comparación de la subpoblación a la cual pertenece cada entidad (según simulación) y el conglomerado asignado. Para los mapas SOM se contabilizó el porcentaje de nodos en zonas de la red no claramente definidas como un conglomerado (transición). Estos porcentajes fueron calculados en aquellos escenarios donde se detectó con éxito el número de grupos de subpoblaciones (5 y 2 para el modelo de islas y de contacto, respectivamente). En las figuras 2 y 3 presentamos los gráficos obtenidos con cada técnica de conglomerados para un caso simulado bajo cada modelo biológico.

3. RESULTADOS

3.1 Desempeño de los algoritmos de agrupamiento a partir del modelo de Islas

En la figura 2 se representan los resultados obtenidos a partir de los 4 métodos de conglomerados. Cuando se utilizó el algoritmo de conglomerados jerárquico UPGMA, el dendrograma resultante permitió de manera relativamente fácil visualizar el agrupamiento de las líneas en 5 subgrupos con un umbral de corte para el eje de las abscisas (distancias entre conglomerados) posicionado aproximadamente en el percentil 95 de las distancias calculadas. En la gráfica obtenida luego de implementar un análisis de conglomerados no jerárquicos (K-means), se observó que la tasa de disminución del criterio que define la función objetivo (suma de cuadrados o heterogeneidad dentro de grupos) decae a partir de 5 conglomerados. La representación

de la red del mapa auto-organizativo debería permitir la visualización de 5 zonas con igual nivel de grises; sin embargo, en este ejemplo la observación de las mismas resultó difícil. Los puntos blancos representan nodos de la red que no han podido ser agrupados claramente en un conglomerado. En el método bayesiano se logró un gráfico en donde los 5 grupos fueron bien separados, sin embargo, en la clasificación de dicho algoritmo se observó la existencia de algunos individuos que perteneciendo a un grupo en particular, contienen características similares a las de otros conglomerados, lo que ocasiona en el gráfico líneas de individuos con más de un color.

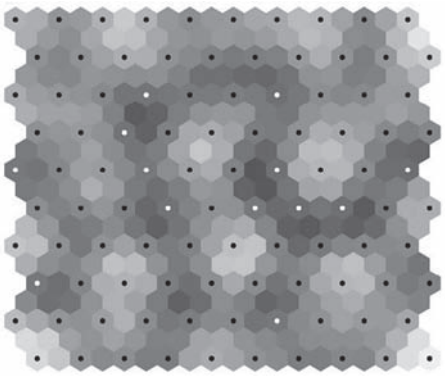

Figura 2. Representaciones gráficas de los resultados logrados por los algoritmos de agrupamiento (de izquierda a derecha, arriba) método jerárquico UPGMA, método no-jerárquico K-means, (abajo) redes neuronales basados en mapas auto-organizativos (SOM) y método bayesiano basados en MCMC, respectivamente. Modelo de islas con una tasa de migración dentro de poblaciones =0.01 y K=5.

3.2 Desempeño de los algoritmos de agrupamiento a partir del modelo de contacto

Las representaciones del modelo de contacto se presentan en la figura 3. La detección del número de grupos (K=2) fue relativamente fácil a través de los dendrogramas, al igual que en el modelo de islas, mientras que la detección del agrupamiento de las líneas en 10 subgrupos (subpoblaciones) resultó difícil. En la representación gráfica obtenida a partir del análisis de conglomerados no jerárquicos se observa que la tasa de disminución del criterio que define la función objetivo no decae sustancialmente. La representación de la red del mapa auto-organizativo debería permitir la visualización de 10 zonas; en este ejemplo la observación de las mismas resultó difícil. Con el método bayesiano el gráfico resultante no permitió la identificación de los subgrupos (subpoblaciones) que conforman la estructura poblacional.

Figura 3. Representación gráfica de los resultados logrados por 4 algoritmos de agrupamiento (de izquierda a derecha, arriba) método jerárquico UPGMA, métodos no-jerárquico K-means, (abajo) redes neuronales basados en mapas auto-organizativos (SOM) y métodos bayesianos basados en MCMC, respectivamente. Modelo de contacto con una tasa migración dentro=0.01, tasa migración entre=0.001 y 2 grupos con 5 subpobaciones; total 10 subpoblaciones.

En la tabla 1 se presenta, para cada método de agrupamiento y para cada modelo de migración, la cantidad de veces que cada uno clasificó de manera correcta a cada individuo en el grupo o subpoblación que le correspondía según fuera simulado. Puede observarse que el método bayesiano fue el que presentó el mayor porcentaje en la detección del número de grupo simulados (90%), seguido del método no jerárquico K-means que clasificó correctamente cada grupo el 83% de las veces para el modelo de islas. El método UPGMA y el SOM obtuvieron el número correcto de conglomerados el 56 y 23% de las veces, respectivamente. Esto último sugiere que tanto el método jerárquico como el basado en redes (SOM) son sensibles a modelos biológicos con estructuras más complejas. Con respecto a la clasificación correcta de individuos en los grupos bien identificados, los resultados estuvieron por encima del 90% para el modelo de islas en todos los métodos, excepto para el método SOM, que clasificó correctamente el 80% de los nodos. Con el método bayesiano el 99% de los individuos fue correctamente asignado en los grupos identificados. Los métodos UPGMA y K-means asignaron individuos correctamente en un 93% de las veces. Bajo el modelo de contacto, los 4 algoritmos evaluados presentaron una baja proporción de éxito en la detección del número de grupos (menor al 50%). El mayor porcentaje de éxito fue obtenido, a diferencia del modelo de islas, con el algoritmo de conglomerados jerárquicos, UPGMA.

Luego, a partir de los porcentajes de éxito en la detección del número de grupos o poblaciones (K), se contabilizó la proporción de individuos bien clasificados para cada grupo. Para los algoritmos jerárquico, bayesiano y SOM los porcentajes de entidades bien clasificadas fueron mayores al 80%. El método de conglomerados no

jerárquico K-means tuvo en esta simulación bajo porcentaje de clasificación correcta de individuos (10%) cuando la estructura de población se obtuvo bajo un modelo biológico de contacto.

Migración	Éxito en la detección							
	K-poblaciones+				Individuos @			Nodos*
	UPGMA	Kmeans	SOM	Bayesiano	UPGMA	Kmeans	Bayesiano	SOM
Islas	56	83	23	90	93	93	99	80
Contacto	50	36	7	23	87	10	97	82

Tabla 1. Comparación de métodos conglomerados respecto a la identificación de subestructuras genéticas bajo distintos modelos de migración que generan estructura genética de poblaciones.

+ Porcentaje de veces en que el K fue detectado, @ Porcentaje de individuos bien clasificados, * Porcentaje de nodos bien clasificados en la red. UPGMA (encadenamiento promedio) algoritmo jerárquico, K-means algoritmo no-jerárquico, SOM (self-organizing maps) algoritmo basado en redes neuronales y Bayesiano (Pritchard et al., 2000).

4. DISCUSIÓN

Los algoritmos de conglomerados confrontados en este trabajo funcionaron de forma diferencial en cada modelo biológico, para detectar los grupos o poblaciones, debido a la estructura genética explícita en cada uno. El método jerárquico para el modelo de islas identificó los 5 grupos simulados. Si bien Sagnard y sus compañeros (2002) sugieren usar un método de análisis discriminante (AD) en vez de algoritmos UPGMA para estudiar la estructura de diversidad genética y poder agrupar poblaciones geográficas o ecológicas, el AD requiere conocer a priori los grupos, mientras que el análisis de conglomerados jerárquicos se usa en la situación contraria, i.e. cuando no se conocen a priori la cantidad de grupos. El algoritmo K-means tuvo un buen desempeño en la detección de los grupos simulados bajo el modelo de islas. Fernández y sus compañeros (2007) lo citan como uno de los algoritmos divisivos más simples de calcular. Jombart y sus compañeros (2010) se refieren a k-means como el algoritmo natural a utilizar para definir el número de grupos cuando éste no se conoce a priori, debido a que utiliza el mismo modelo que el análisis discriminante y una medida similar en la diferenciación de grupos.

El algoritmo basado en redes SOM es una alternativa recientemente usada en estudios donde se desea encontrar la estructura genética en datos obtenidos a partir de

marcadores moleculares. Fernández y Balzarini (2007) propusieron el uso de SOM junto a una aplicación desarrollada como herramienta para mejorar la visualización de los resultados en el contexto de abundante información genómica. Las adaptaciones específicas usadas en ese trabajo permitieron una mejor identificación de grupos de genes que el algoritmo K-means e incluso el algoritmo UPGMA. Sin embargo en nuestro estudio, con una cantidad sustancialmente menor de marcadores y sin el uso de la herramienta gráfica, el método SOM no fue el algoritmo que mejor identificó los grupos y las subestructuras emuladas. El método bayesiano basado en cadenas de Markov de Monte Carlo demostró un buen desempeño en el reconocimiento de la estructura simulada bajo el modelo de islas, pero no para el modelo de contacto, reduciendo su porcentaje de detección desde el 90% al 20%, respectivamente. Además, en la representación gráfica no se visualizan de manera clara los grupos, ni las subpoblaciones. Evanno y sus compañeros (2005) evaluaron el método bayesiano para detectar estructura genética sobre datos simulados bajo tres modelos de migración, concluyendo también que este algoritmo resulta eficaz no sólo para un modelo de isla, sino para esquemas más complejos de migración jerárquica.

5. COMENTARIOS FINALES

Para el modelo de islas el algoritmo que mejor identificó los grupos de subpoblaciones simulados fue el bayesiano. Para el modelo de contacto fue mejor el conglomerado jerárquico UPGMA, aunque la identificación de la estructura genética bajo este modelo fue pobre bajo todos los algoritmos. Los algoritmos SOM y bayesiano distinguieron mejor subpoblaciones que grupos de subpoblaciones. Los resultados sugieren que el desempeño de los algoritmos depende del modelo biológico subyacente para la estructuración genética de poblaciones, y que existe menos error de clasificación con el modelo de islas que con el de contacto. Nuevos desarrollos son necesarios para mejorar la identificación de la subestructura genética subyacente en situaciones biológicas

BIBLIOGRAFÍA

Balloux, F. EASYPOP versión 1.7. (2001) A computer program for the simulation of population genetics. J. Heredity 92: 301-302.

Balzarini, M. y Di Rienzo, J. (2004) Info-Gen, Software para análisis estadístico genómico. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba. http://web.info-gen.com.ar. Versión actualizada 2010.

Chiappero, M., Panzetta-Dutari, G., Gómez, D., Castillo, E., Polop, J. y Gardenal, C.

(2010) Contrasting genetic structure of urban and rural populations of the wild rodent Calomys musculinus (Cricetidae, Sigmodontinae). Mammalian Biology. doi:10.1016/j.mambio.2010.02.003.

Evanno, G., Regnaut, S. y Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14: 2611-2620.

Fernández, E. y Balzarini, M. (2007) Improving cluster visualization in self-organizing maps: Application in gene expression data analysis. Computers in Biology and Medicine 37:1677– 1689.

Fernández, E., Alvarez, M., Podhajcer, O., Stolovitzky, G. (2007) Genomica funcional: En busca de la función de los genes. National Academy of Science, Argentine. XIII 63-76.

Hedrick, P. (2005) Large variance in reproductive success and the Ne/N ratio. Evolution 59:1596-1599.

INFOSTAT. Manual del usuario. (2008) Grupo InfoStat, FCA, Universidad Nacional de Córdoba. Primera Edición, Editorial Brujas. Córdoba, Argentina.

Jombart, T., Devillard, S. y Balloux, F. (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11:94.

Kohonen, T. (1997) Self-Organizing Maps. Second Ed. Springer. Berlin.

Milligan, G. (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45:325-342.

Pritchard, J., Stephens, M. y Donnelly, P. (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Sagnard, F., Barberot, C. y Fady, B. (2000) Structure of genetic diversity in Abies alba Mill. from southwestern Alps: multivariate analysis of adaptive and non.adaptive traits for conservation in France. Forest Ecology and Management 157:175-189.

Sokal, R. y Michener, C. (1958) A statistical methods for evaluating systematic relationships. University of Kansas Science Bulletin 38:1409-1438.

Wright, S. (1931) Evolution in mendelian populations. Genetics 16:97-159.

Referencia	Fecha de recepción	Fecha de aprobación
Peña Malavera Andrea Natalia, Bruno Cecilia, Teich Ingrid y Balzarini Mónica. Clasificación en la identificación de estructura genética de poblaciones a partir de datos moleculares. Revista *Tumbaga* (2010), 5, 225 - 236	Día/mes/año 30/08/2010	Día/mes/año 1/09/2010