Rev. FCA UNCuyo. Tomo XLI. N° 3. Año 2009. 171-182.

Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos

Genetic distances between molecular profiles obtained from multilocus-multiallele markers

Cecilia Bruno
Mónica Balzarini

Originales: Recepción: 11/07/2009 - Aceptación: 14/09/2009

RESUMEN

Para expresar la magnitud de la identidad genética (similaridad) o su complemento (distancia) entre dos individuos caracterizados molecularmente a través de marcadores del tipo microsatélites (SSR), que son multilocus-multialélicos, es necesario elegir una métrica acorde con la naturaleza multivariada de los datos. Comúnmente, las métricas de distancias genéticas son diseñadas para expresar, en un único número, la diferencia genética entre dos poblaciones y son expresadas como función de la frecuencia alélica poblacional. Dichas métricas pueden también ser utilizadas para calcular la distancia entre perfiles individuales, pero las frecuencias alélicas no son continuas en este caso. Alternativamente, se pueden usar distancias geométricas obtenidas como el complemento del índice de similaridad para datos binarios que indican la presencia/ausencia de cada alelo en un individuo. El objetivo de este trabajo fue evaluar simultáneamente el desempeño de ambos tipos de métricas para ordenar y clasificar individuos en una base de datos generadas a partir de loci de marcadores microsatélites SSR. Se calcularon 11 métricas de distancias a partir de 17 loci SSR obtenidos desde 17 introducciones

ABSTRACT

In order to express the magnitude of the genetic identity (similarity) or its complement (distance) between individuals genotyped with microsatellites (SSR), which are multilocus-multiallele markers, is necessary to choose a metric in agreement with the multivariate nature of the marker data. Most of the metrics of genetic distances were designed to express, as a single quantity, the genetic difference between two populations and they are expressed as function of population allele frequencies. Such metrics can also be used to calculate distances between individual profiles, but the allele frequencies are not longer continuous. On the other hand, geometric distances obtained as complement of similarity indexes for binary data indicating allele presence/absence in each individual, are commonly used for pairwise individual comparisons. However, they do not take into account the nested allele within locus structure of SSR data. The objective of this work was to simultaneously evaluate the performance of both metric types to order and classify individuals in a multivariate basis generated by the use of SSR loci. We applied 11 different distance metrics to a dataset involving 17 SSR loci obtained from

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Estadística y Biometría. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba. Av. Valparaíso s/n. Ciudad Universitaria. C. C. 509. (5000) Córdoba. Argentina. cebruno@agro.unc.edu.ar

de un banco de germoplasma de soja [*Glycine max* (L.) Merr.]. Se evaluó el consenso de los resultados obtenidos para la clasificación de los 17 perfiles moleculares desde varias métricas. Los resultados sugieren que los diferentes tipos de métricas producen información similar para comparar individuos. No obstante, se realizó una clasificación de las métricas que responden a diferencias entre los núcleos de las expresiones de cálculo.

Palabras clave

similaridad genética • microsatélite • clasificación

17 entries of a soya [*Glycine max* (L.) Merr.] germoplasm, and evaluated the consensus in the results obtained from the classification of the 17 molecular profiles from severl metrics. The results suggest that most of the evaluated metrics yield similar information about marker profiles in the context of pairwise individual comparisons. We provide a kernel-based metric classification.

Keywords

genetic similarity • microsatellite • classification

INTRODUCCIÓN

Las regiones genómicas que contienen secuencias simples repetidas (SSR) amplificadas por PCR (*Polymerase Chain Reaction*) constituyen marcadores de ADN altamente polimórficos, conocidos como microsatélites. Cada marcador microsatélite, independientemente del elemento repetido, representa un locus genético multialélico. Los locus SSR son somáticamente estables y poseen expresión co-dominante, es decir, es posible diferenciar los distintos alelos de un locus. Los microsatélites proveen abundante información para calcular distancias entre poblaciones y también entre individuos ya que permiten distinguir los estados de los alelos de cada locus de marcador (datos de alelos por locus).

Para expresar la magnitud de la identidad genética (similaridad) o su complemento (distancia) usando datos de marcadores multilocus-multialélicos, es necesario elegir una métrica acorde a la naturaleza multivariada de la información. El problema de concebir una métrica eficiente de similaridad/distancia entre poblaciones de individuos ha sido bien tratado por Nei (13) y por Hedrick (7). Las métricas de distancias genéticas fueron diseñadas para expresar, como un único número, la diferencia entre dos poblaciones. Así, las distancias constituyen un recurso para reducir la dimensión de una matriz de datos multivariados. Generalmente, cuando no hay diferencias entre los objetos de estudio, la distancia es 0, mientras que si éstos no tienen alelos en común para ningún locus (máxima diferencia), la distancia es 1.

La selección de la métrica más apropiada depende del tipo de marcador que provee los datos e incluso de la codificación usada para representar los eventos de amplificación. Si bien las métricas descriptas por Nei son aplicables a datos genotípicos obtenidos a partir de marcadores SSR, sus propiedades han sido evaluadas desde la perspectiva de estudios poblacionales (14) y no respecto de su desempeño en la comparación de individuos (perfiles moleculares individuales). Cuando estas métricas de distancia genética se calculan entre pares de individuos, las frecuencias alélicas son de naturaleza discreta.

En la comparación de poblaciones se ha usado como *kernel* de varias métricas el numerador de la medida de similitud de Nei, conocido también como índice de Sneath (I_s), que representa un coeficiente de parentesco entre las poblaciones. Dicho índice se basa en las frecuencias alélicas de los componentes de un locus de interés en cada población y se calcula a partir del producto de las frecuencias de alelos de un mismo locus en cada población. Sin embargo, puede suceder que dos poblaciones sean idénticas en las frecuencias alélicas observadas pero que no tengan la máxima similaridad (I_s). Para evitar este inconveniente se trabaja con el I_s normalizado (Nei estándar, I_N) cuyo logaritmo natural es una métrica de uso frecuente para comparar poblaciones. También existen otras métricas, basadas en el índice de Sneath, que han sido desarrolladas para controlar otros tipos de problemas que surgen cuando toma valores extremos, como la métrica Nei mínimo, o cuando los tamaños de muestras, extraídos de cada población son pequeños, como la métrica Nei insesgado (13). Para el caso multilocus, Sokal y Sneath (16) propusieron calcular I_s para cada locus y luego obtener la identidad media promediando a través de todos los loci. Se considera que las medidas de distancia genética mencionadas anteriormente tienen una base biológica y como fueron concebidas en el contexto de la genética de poblaciones, rara vez son aplicadas en la comparación de perfiles individuales.

Alternativamente, la comparación de perfiles individuales puede realizarse con otras métricas que no involucran en sus fórmulas conceptos biológicos y que son conocidas como distancias geométricas, tales como las distancias de Roger (18), de Cavalli-Sforza (3) y de Prevosti (18) a partir de datos codificados en una tabla de dos vías donde cada celda contiene la información molecular para un genotipo (filas de la tabla) y un locus (columnas de la tabla). Estas distancias también se basan en frecuencias alélicas. Smouse y Peakall (15) definieron otra distancia geométrica que, al cuadrado, es un medio de la distancia Euclídea (9) entre vectores que contienen como elementos las frecuencias alélicas de cada locus y que puede ser usada para medir distancia entre individuos dentro de una población. La distancia multivariada de Smouse y Peakall (15) es obtenida sumando las distancias entre individuos para cada locus, a través de todos los loci.

Los datos de microsatélites también se pueden codificar en tablas a dos vías clasificación de individuos por alelos, *i.e.*, tantas filas como individuos y tantas columnas como alelos diferentes haya a través de todos los loci. En esta codificación no se puede identificar cuáles son los alelos provenientes de un mismo locus. La presencia o no de un alelo para un individuo es contabilizada mediante una variable indicadora (*e.g.* presencia=1, ausencia=0). Para esta codificación es común usar distancias geométricas definidas como el complemento a uno, o una función de dicho complemento, de algún índice de similitud para variables binarias. Aun cuando no contemplan la estructura de anidamiento de los alelos dentro de los locus, las métricas basadas en datos binarios se encuentran ampliamente difundidas como herramientas para la comparación de perfiles individuales obtenidos por marcadores SSR (4, 11, 17).

Los datos de marcadores SSR se usan no sólo para obtener distancias genéticas y realizar estudios de variabilidad genética sino también para ordenar y clasificar individuos. Para ello, es común el uso de métodos de clasificación o *cluster*. Muchos

de los algoritmos de clasificación tienen como *input* matrices de distancia entre individuos. Un problema común en la práctica del análisis de este tipo de datos es que las diferentes métricas pueden producir agrupamientos de individuos de mayor o menor calidad respecto de la verdadera relación que existe entre los individuos, en el espacio multidimensional.

Las técnicas de conglomerados jerárquicos, como UPGMA (*unweigthed pair-group arithmetic average method*) o Ward generan como *output* un dendrograma donde la longitud de las ramas que conectan individuos indican la magnitud de la distancia entre ellos; estas longitudes pueden variar con la métrica de distancia usada. Cuando se usan métodos no-jerárquicos, como k-*means*, el impacto de las métricas de distancia no se evalúa tan directamente ya que si bien se parte de una matriz de distancias, la clasificación final de individuos depende también de la relación entre sumas de cuadrados entre y dentro de un número determinado de grupos que se hipotetiza *a priori* caracteriza la estructura subyacente de agregamiento.

Menos impactados por la selección de la métrica son los algoritmos de clasificación no supervisados, basados en redes neuronales como los mapas auto-organizativos (SOM) (10) que trabajan directamente sobre la base de datos original, es decir, sin cálculos de distancia previa. En este trabajo se usó k-*means* y SOM para evaluar los resultados que algoritmos jerárquicos (UPGMA y Ward) producen bajo diferentes métricas de distancia entre pares de individuos.

Objetivo

Evaluar el desempeño de distintas métricas cuando son usadas con loci SSR para calcular distancias entre perfiles moleculares individuales con fines de ordenamiento y clasificación de los individuos cuyo genoma fue caracterizado molecularmente.

MATERIALES Y MÉTODOS

Datos

Se trabajó con un conjunto de datos ilustrativos compuesto por 17 loci de marcadores SSR usados para caracterizar el genoma de 17 introducciones de un banco de germoplasma de soja [*Glycine max* (L.) Merr.]. Mediante la amplificación PCR de los SSR se pudieron visualizar un total de 55 bandas (alelos) para el conjunto de las 17 introducciones. De las 55 bandas, 52 (94,5%) fueron polimórficas. El porcentaje de amplificación, en todo el experimento, fue de 34,3%. No hubo, en el conjunto de datos usados, muestras duplicadas.

En la tabla (pág. 175) se presentan medidas resúmenes que describen el polimorfismo observado para el conjunto de loci SSR usados para evaluar el desempeño de distintas métricas en el cálculo de distancias entre perfiles moleculares individuales con fines de ordenamiento y clasificación de los individuos.

Tabla 1. Estadística descriptiva del polimorfismo observado mediante 17 loci de marcadores SSR usados para caracterizar genotítpicamente 17 introducciones de soja [*Glycine max* (L.) Merr.].

Table 1. Descriptive statistics of polymorph observed in 17 SSR loci used for genotype 17 entries of soya [*Glycine max* (L.) Merr.].

Locus	AP	AM	PMF(95)	PIC	AMP	PDICMA
L1	3	0	1,00	0,31	37,25	$2,5\times10^{-14}$
L2	3	0	1,00	0,34	33,33	$1,9\times10^{-18}$
L3	4	0	1,00	0,27	33,82	$1,0\times10^{-11}$
L4	3	0	1,00	0,25	39,22	$3,9\times10^{-09}$
L5	2	0	1,00	0,31	52,94	$3,6\times10^{-12}$
L6	3	0	1,00	0,27	37,25	$1,7\times10^{-12}$
L7	2	0	1,00	0,32	55,88	$5,3\times10^{-13}$
L8	4	0	1,00	0,21	27,94	$2,8\times10^{-08}$
L9	3	1	0,75	0,26	32,35	$5,2\times10^{-08}$
L10	4	1	0,80	0,26	22,35	$1,5\times10^{-09}$
L11	0	1	0,00	0,00	100,00	$1,0\times10^{-00}$
L12	3	0	1,00	0,13	33,33	$1,8\times10^{-04}$
L13	2	0	1,00	0,37	50,00	$4,3\times10^{-17}$
L14	4	0	1,00	0,31	32,35	$4,5\times10^{-15}$
L15	4	0	1,00	0,26	26,47	$1,9\times10^{-13}$
L16	4	0	1,00	0,28	26,47	$3,8\times10^{-14}$
L17	4	0	1,00	0,25	25,00	$5,6\times10^{-13}$

AP: cantidad de alelos polimórficos; AM: cantidad de alelos monomórficos; PMF(95): proporción de marcadores polimórficos; PIC: contenidos de información polimórfica; AMP: porcentaje de amplificación; PDICMA: probabilidad de que dos individuos compartan el mismo alelo por azar (5).

AP: number of polymorphic alleles; AM: number of monomorphic alleles; PMF(95): proportion of polymorphic markers; PIC: polymorphism information content; AMP: amplification rate; PDICMA: probability that two individuals share the same allele by chance (5).

Distancias evaluadas

Se calcularon 9 medidas de distancias expresadas en términos de frecuencias alélicas por locus: Nei Estándar, Nei Insesgado, Nei Mínimo, Nei, Roger-Modificada, Prevosti, Smouse y Peakall, distancia de la cuerda (3) y distancia del arco (3). Estas métricas fueron obtenidas desde tablas a dos vías de clasificación (individuos x locus) donde cada celda es el genotipo de un individuo para un locus, *i.e.* la combinación de alelos de cada locus. Dado que los objetos a ordenar son individuos diploides, las frecuencias de cada alelo en un locus asumen los valores 0; 0,5 ó 1. Se incluyeron en el análisis dos medidas de distancia geométrica expresadas desde datos binarios que sugieren presencia/ausencia de cada alelo en cada locus de un perfil individual. Las distancias se expresaron como $(1\text{-}S)^{1/2}$ donde S representa el índice de similitud para datos binarios de Jaccard (8) o, alternativamente, el índice Emparejamiento Simple (*Simple Matching*).

Procedimientos de evaluación

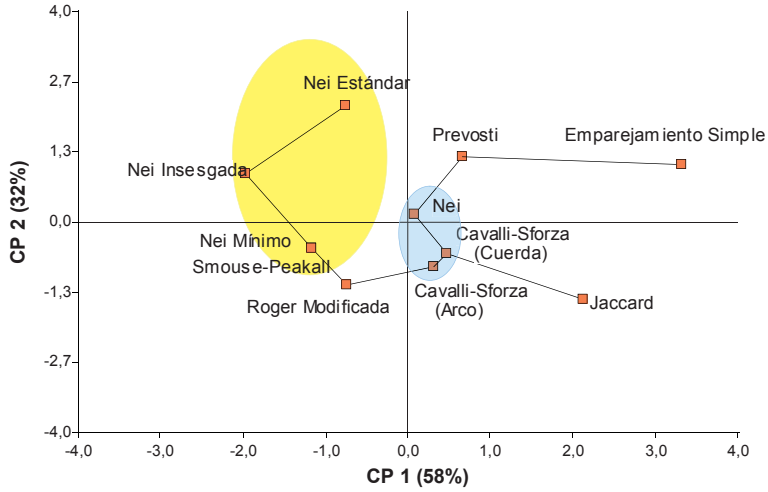
Se calcularon las 11 métricas de distancia listadas anteriormente para todos los pares de accesiones conformados a partir de las 17 variedades (introducciones) de soja. Las distancias fueron obtenidas con el software para datos genéticos Info-Gen (2) totalizando 136 valores de distancias para cada métrica. Las distancias entre pares de individuos fueron usadas para clasificar las métricas según la similitud de su desempeño.

Para ello se construyó un ordenamiento de las métricas de distancias en un plano factorial, mediante análisis de coordenadas principales (PCO) de una matriz de dimensión 136 x *k* (136 pares de individuos y *k*=11métricas). La matriz de entrada del PCO contiene en la columna *k* las distancias entre pares de individuos para la *k*-ésima métrica. También se calcularon medidas de correlación entre todas las matrices de distancia obtenidas para cada métrica, mediante el estadístico Z de Mantel (12). Finalmente, se clasificaron las 17 variedades, usando cada una de estas matrices de distancia, mediante dos técnicas de conglomeración jerárquica: UPGMA y Ward. Los métodos de agrupamiento k-*means* y SOM se usaron para evaluar los resultados que los algoritmos jerárquicos (UPGMA y Ward) produjeron bajo diferentes métricas de distancia. Los métodos k-*means* y SOM (10) se aplicaron sobre la matriz de datos originales. El algoritmo k-*means* se usó en dos sentidos: 1) sobre cada una de las 11 matrices de distancias originadas por las distintas métricas evaluadas y 2) para agrupar los resultados de la clasificación obtenida bajo cada métrica.

RESULTADOS Y DISCUSIÓN

Se encontró correlación estadísticamente significativa entre todos los pares de matrices de distancias que se compararon (Z de Mantel, $p<0,0001$). La figura 1 (pág. 177) muestra el ordenamiento de las métricas obtenido por PCO y mediante un árbol de mínimo recorrido (ARM) (1) se puede visualizar la mayor o menor cercanía de las métricas según su similitud/diferencia en la cuantificación de distancias entre pares de variedades. Las métricas más cercanas, en el plano construido por las dos primeras coordenadas principales, deben ser interpretadas como métricas altamente congruentes respecto del conjunto ordenado de distancias entre individuos que producen. Mientras mayor parecido entre el desempeño de dos métricas respecto de las distancias que asignan entre las 17 variedades, menor es la distancia del segmento que las conecta en el ARM.

La correlación entre las métricas Nei-Mínimo y Smouse-Peakall fue igual a 1. Estas métricas también se mostraron altamente correlacionadas con la distancia de Roger-Modificada (r_{Mantel}=0,9971). Las métricas Nei-Estándar, Nei-Insesgada, Nei-Mínimo, Smouse-Peakall y Roger-Modificada se ordenaron juntas según valores negativos de la primera componente principal (CP1). Éstas conforman un grupo de métricas que producen ordenamientos muy parecidos a los perfiles moleculares. Es de destacar que todas ellas tienen como núcleo de su expresión matemática el producto de las frecuencias alélicas de cada locus, entre pares de individuos. Este producto representa la correlación entre los genotipos moleculares del locus. Además todas ellas integran esta información a través de todos los loci.



Los segmentos de líneas que unen los puntos representan el árbol de recorrido mínimo (ARM) que conecta cada métrica con aquella de desempeño más parecido.

The segments of lines joining the points represent the minimum spanning tree (MST) that connects each metric performance with that of most similar.

Figura 1. Ordenamiento de métricas según su desempeño en la clasificación de individuos caracterizados genotípicamente por marcadores SSR.
Figure 1. Arrangement according to performance metrics in the classification of individuals genotyped by SSR markers.

La de mayor diferencia es la métrica de Nei-Estándar que expresa la distancia entre perfiles en la escala logarítmica de la correlación de los vectores que representan los individuos en el espacio de las frecuencias alélicas. Dicha escala se usa para evitar que la distancia genética sea mayor que el complemento a uno de la similitud cuando la correlación se acerque a cero. Al igual que Nei-Insesgada, considera una corrección por sesgo ocasionado al trabajar con frecuencias alélicas estimadas a partir de muestras pequeñas, mientras que Nei-Mínimo establece una cota para la distancia genética entre dos individuos que no comparten ningún alelo. Para tal caso las distancias anteriores no están definidas o son extremadamente grandes. Este hecho justifica el ordenamiento de las métricas de Nei en distintos cuadrantes según la Componente Principal 2 (CP2). Aun cuando los valores de la distancia de Smouse-Peakall fueron mayores en nivel medio, éstos produjeron idéntica ordenación entre individuos que la distancia Nei-Mínimo (figura 1), por tanto esta distancia no presenta un beneficio extra.

La distancia de Roger-Modificada, que es expresada directamente en términos de distancia Euclídea promedio a través de los loci, mostró un desempeño muy similar a las distancias de Nei-Mínimo y Smouse-Peakall. La alta correlación es justificada por la asociación directa entre correlación y distancia Euclídea en el espacio de las frecuencias alélicas.

Según se muestra en la figura 1 (pág. 177), un segundo grupo de métricas es el conformado por las distancias de la Cuerda y el Arco de Cavalli-Sforza, que al igual que la distancia de Nei, se generan usando la raíz cuadrada de las frecuencias alélicas como coordenadas del espacio en el que se representan los individuos.

Estas distancias asumen valores proporcionales a la correlación entre los vectores de frecuencias alélicas de ambos individuos en el espacio de sus raíces cuadradas, mientras que la distancia de Prevosti es definida a partir del valor absoluto de la diferencia de los vectores de frecuencias alélicas directamente y, por tanto, se relacionó más en desempeño con la distancia obtenida a partir del índice de Emparejamiento Simple, que también trabajó en la escala de las diferencias absolutas.

Por el contrario, la distancia basada en la transformación $(1\text{-}S)^{1/2}$ del Índice de Jaccard, donde S es el índice de similitud de Jaccard, produjo ordenamientos más parecidos a las distancias de Cavalli-Sforza. La mayor ponderación que recibe la co-presencia en relación con la co-ausencia de alelos en el índice de Jaccard, podría hacer más conservadora la inferencia sobre similitudes entre dos individuos, *i.e*. aumenta la similaridad entre individuos.

El índice de similitud de Jaccard produjo valores de distancia mayores en términos medios y de menor variabilidad que las distancias obtenidas a partir del índice Emparejamiento Simple ($\bar{d}_J = 0.8,\ CV_J = 8$ *vs.* $\bar{d}_{ES} = 0.6,\ CV_{ES}$=*13*). La variabilidad de las distancias obtenidas entre pares de individuos bajo métricas basadas en información binaria fue menor que la variabilidad observada al usar métricas basadas en frecuencias alélicas. Los coeficientes de variación de las métricas basadas en presencia/ausencia de alelos alcanzaron valores de 13% (para Emparejamiento Simple) mientras que un CV=37% se adjudicó a la métrica Nei-Insesgado.

En función de la ordenación y el agrupamiento de métricas sugerido por la figura 1 (pág. 177), la clasificación de las 17 variedades, mediante UPGMA y Ward se realizó usando como *input* matrices de distancia calculadas a partir de las métricas de Roger-Modificada, Prevosti y de los índices de similitud Jaccard y Emparejamiento Simple.

Los resultados obtenidos sugieren que, en general, existe una alta congruencia en el agrupamiento de las 17 variedades, independientemente de las métricas de distancia usada. Sin embargo, algunas variedades (K, L y G, figura 2 -pág. 179-) fueron clasificadas por UPGMA en forma diferente según la métrica.

Los dendrogramas de la figura 2 muestran que Jaccard ordena, distinto que Emparejamiento Simple, a la variedad K. Tal como sugirió la ordenación de las métricas por CP (figura 1, pág. 177), la clasificación de las variedades obtenidas vía la distancia calculada desde el índice Emparejamiento Simple fue más parecida a la clasificación obtenida bajo la métrica de Prevosti (figura 2, pág. 179). Los dendrogramas obtenidos por el método de Ward arrojaron resultados equivalentes.

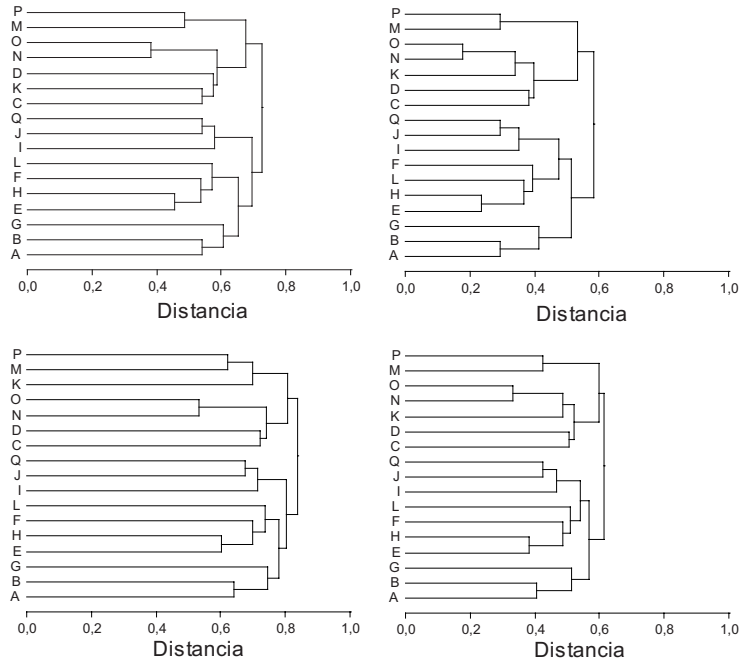
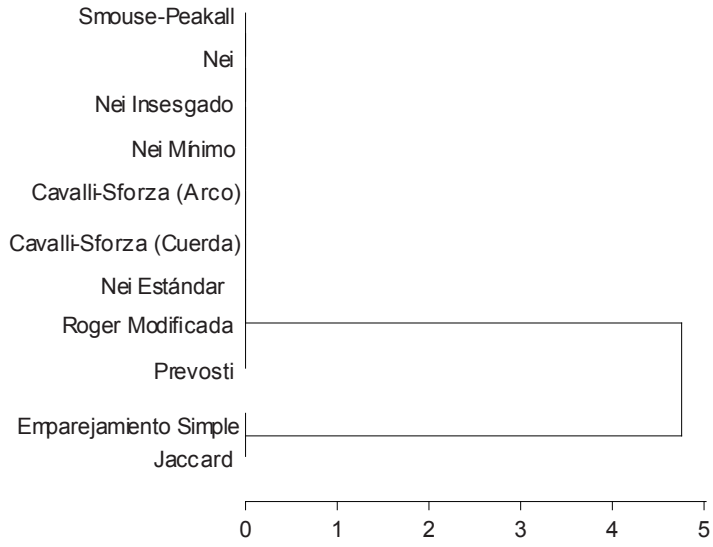


Figura 2. Clasificación de variedades en función del algoritmo jerárquico UPGMA para las métricas de distancias Roger-Modificado (panel izquierdo arriba), Prevosti (panel derecho arriba), $(1\text{-Jaccard})^{1/2}$ (panel izquierdo abajo) y $(1\text{-Emparejamiento Simple})^{1/2}$ (panel derecho abajo).

Figure 2. Classification of varieties depending on the UPGMA hierarchical algorithm for metric distances Roger-Modified (left panel above), Prevosti (right panel above), $(1\text{-Jaccard})^{1/2}$ (bottom left panel) and $(1\text{-Simple Pairing})^{1/2}$ (bottom right panel).

Al clasificar las variedades mediante el algoritmo que produce mapas auto-organizativos de Kohonem (SOM) donde no es necesario realizar un cálculo de distancia previo, se formaron los siguientes grupos de variedades: Grupo I: A-B, Grupo II: E-F-H, Grupo III: C-D-N-O-Q-J-I, Grupo IV: K-L-M-P, mientras que la variedad G no fue ubicada en ningún grupo, ya que constituyó un nodo de transición en la conformación de la red (6). Es importante notar que la mayoría de estos grupos fueron sugeridos por la clasificación jerárquica independientemente de la métrica de distancia usada. La unión o no de G con A-B dependió más del algoritmo que de la métrica. Los cambios en el agrupamiento de la variedad K mostraron que la distancia calculada a partir del índice de Emparejamiento Simple produjo clasificaciones menos parecidas que las obtenidas mediante el índice de Jaccard cuando se usó UPGMA que cuando se usó Ward.

La figura 3 muestra el dendrograma obtenido por UPGMA sobre una matriz conteniendo el grupo al que pertenece cada variedad según k-*means* desde cada métrica de distancia. Se puede visualizar que las métricas de distancias geométricas, obtenidas bajo uno u otro índice de similitud, se desempeñaron de manera similar, mientras que las métricas de distancia genética de Smouse-Pekall, Arco y Cuerda de Cavallis-Sforza, Nei, Nei-Estándar, Nei-Insesgado, Nei-Mínimo, Roger-Modificada y Prevosti conformaron otro grupo.



Emparejamiento Simple y Jaccard representan en esta figura la raíz cuadrada del complemento a uno del índice de similitud, respectivamente.

Simple Matching and Jaccard in this figure represent the square root of one's complement of the similarity index, respectively.

Figura 3. Dendrograma obtenido por UPGMA, con distancia Euclídea, sobre la clasificación de variedades sugeridas por k-*means* bajo 11 métricas de distancias entre pares de individuos.

Figure 3. Dendrogram obtained by UPGMA with Euclidean distance on the classification of varieties suggested by k-means under 11 metric distances between pairs of individuals.

CONCLUSIÓN

Las diferentes posibilidades de estimar distancias entre individuos, aun habiendo sido concebidas desde conceptos tanto biológicos como geométricos, producen resultados altamente congruentes cuando son utilizadas para clasificar perfiles individuales de datos construidos a partir de marcadores SSR. Sin embargo, los resultados sugieren una mayor similitud en las clasificaciones obtenidas por los distintos tipos de métricas (basada en frecuencias alélicas o en la presencia/ausencia de cada alelo por locus).

Las métricas Nei-Insesgada, Nei-Mínimo, Smouse-Peakall y Roger-Modificada que se basan en la correlación de frecuencias alélicas, conforman un grupo de medidas de distancias que junto a las métricas que se establecen a partir de la raíz cuadrada de frecuencias alélicas (distancias de la Cuerda y el Arco) producen clasificaciones de individuos más parecidas y ligeramente diferentes a las que se podrían obtener por Nei-Estándar o por distancias basadas en índices de similitud. La distancia de Prevosti produce resultados que se asemejan a los obtenidos utilizando la métrica basada en el índice de similitud Emparejamiento Simple cuando es usada en este contexto de comparación de individuos. Incluso cuando las distancias basadas en los índices de similitud Jaccard y Emparejamiento Simple son métricas que no tienen en cuenta la estructura de anidamiento de los alelos en los locus de los marcadores SSR, éstas producen clasificaciones de los genotipos congruentes a las que son obtenidas con algoritmos no basados en distancias como los algoritmos de redes neuronales SOM.

REFERENCIAS

1. Arroyo, A.; Balzarini, M.; Bruno, C.; Di Rienzo, J. 2005. Árboles de expansión mínimos: ayudas para una mejor interpretación de ordenaciones en bancos de germoplasma. Interciencia. Venezuela. 30(9): 550-554.
2. Balzarini, M.; Di Rienzo, J. 2003: Info-Gen: Software para análisis estadístico de datos genéticos. Universidad Nacional de Córdoba. Córdoba. Argentina. Disponible en http://www.info-gen.com.ar
3. Cavalli-Sforza, L. L; Edwards, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 19: 233-257.
4. Chandra, A.; Pandey, K. C. 2007. Identification of Wwvil (*Hypera postica* Gyll.) Resistance sources in genus Medicago and their analysis employing SSR markers. In: Li, Z. K.; Zhang, Q. F. (eds.) From genomics to plant improvement. Proceedings of the 2nd International Conference on Plan Molecular Breeding, 23-27 March, Sanya,China, p. 83-84.
5. Fernández, E. A.; Balzarini, M. 2007. Improving cluster visualization in Self-Organizing Maps: Application in Gene Expression Data Analysis. M. Computers in Biology and Medicine 37: 1677-1689.
6. Gower, J. C. 1985. Measures or similarity, dissimilarity and distance. p. 397-405. In: Kotz, S.; N. L. Jonhson (eds). Encyclopedia of statistical science. Vol. 5. Wiley, New York.
7. Hedrick, P. W. 1975. Genetic Similarity and Distance: Comments and Comparisons. Evolution 29(2): 362-366.
8. Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull. Soc. Vaudoise Sci. Nat. 37: 547-579.(1908). Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44: 223-270.
9. Johnson, R. A.;Wichern, D. W. 2007. Applied Multivariate Statistical Analysis (6ª ed.). Prentice Hall. 800 p.
10. Kohonen, T. 1997. Self-Organization Maps. 2nd. Springer, Berlin. 362 p.
11. Luan, F. S.; Sheng, Y. Y.; Wang, Y. H.; Staub, J. E. 2007. Genetic characteristics among parents and derived Melon hybrids. In: Li, Z. K.; Zhang, Q. F. (eds.) From genomics to plant improvement, Proceedings of the 2nd International Conference on Plan Molecular Breeding, 23-27 March, Sanya,China. 76 p.
12. Mantel, N. A. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209-220.

13. Nei, M. 1972. Genetic distance between populations. Am. Nat. 106: 283-291.
14. _____. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York. 512 p.
15. Smouse, P.; Peakall, R. 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82: 561-573.
16. Sokal, R. R.; Sneath, P. H. 1963. Principles of Numerical Taxonomy. San Francisco and London, W. H. Freeman. 359 p.
17. Wang, L. F.; Exbrayat-Vinson, F.; Hao, C. Y.; Roussel, V.; Zhang, X. Y.; Balfourier, F. 2007. Comparison of genetic diversity level between European and Asian Wheat germplasm, using SSR markers. In: Li, Z. K.; Zhang, Q. F. (eds.) From genomics to plant improvement, Proceedings of the 2nd International Conference on Plan Molecular Breeding, 23-27 March, Sanya, China. 74 p.
18. Wright, S. 1978. Evolution and the Genetics of Populations. Vol. 4. Chicago, IL: Variability within and among natural populations. University of Chicago Press. 628 p.

Agradecimientos

A la Universidad Nacional de Córdoba (Argentina) y al Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET).