ACTAS DEL XV CONGRESO
DR. ANTONIO A. R. MONTEIRO (2019)

RELACIONES ENTRE COEFICIENTES DE SIMILITUD PARA EL ANÁLISIS DE MARCADORES MOLECULARES EN DIFERENTES ESTRUCTURAS POBLACIONALES

ANTONIO F. GARAYALDE

RESUMEN. En este trabajo se comparan los resultados de las distancias basadas en coeficientes de similitud más utilizadas en la genética de poblaciones, en dos materiales genéticos contrastantes en su estructura poblacional. Se establecen las relaciones funcionales entre los índices y la influencia de la estructura poblacional en las relaciones entre individuos obtenidas con los diferentes coeficientes. Se concluye que en materiales con baja estructura poblacional es recomendable el uso de la distancia basada en Simple Matching dado que incluye las dobles ausencias, es euclidiana y es proporcional a la distancia euclídea entre los datos originales, por lo que permite aplicar un análisis de componentes principales, pudiéndose representar de manera conjunta individuos y variables. En el caso de materiales con estructura poblacional marcada, el uso de cualquier coeficiente de similitud llevaría a resultados similares.

1. INTRODUCCIÓN

Los marcadores moleculares son ampliamente utilizados para el estudio de las relaciones genéticas entre individuos, permitiendo detectar variaciones en la secuencia del ADN entre individuos y caracterizar sectores específicos del ADN. Los marcadores ISSR (secuencias entre repeticiones de secuencias simples) son de característica dominante, por lo que en un individuo diploide la presencia de la banda en un gel de electroforesis se corresponde con los genotipos AA, Aa y aA, mientras que la ausencia de la banda se corresponde al genotipo aa, donde A es el alelo dominante y a el alelo recesivo de un determinado locus. El patrón de bandas obtenido es entonces de característica binaria, codificándose como 1 o 0 a la presencia o ausencia de la banda, respectivamente. A partir de esta codificación se utilizan índices de similitud S, y para evaluar las diferencias entre pares de individuos se establecen índices de disimilitud definidos como $1-S$. Los coeficientes de similitud entre individuos usados habitualmente son:

- Jaccard (***J***; [8]): $J = \dfrac{a}{a+b+c}$
- Sørensen–Dice (***SD***; [4, 18]): $SD = \dfrac{2a}{2a+b+c}$
- Ochiai (***O***; [12]): $O = \dfrac{a}{\sqrt{(a+b)(a+c)}}$
- Anderberg (***A***; [1]): $A = \dfrac{a}{a+2(b+c)}$
- Simple Matching (***SM***; [17]): $SM = \dfrac{a+d}{a+b+c+d}$
- Rogers–Tanimoto (***RT***; [15]): $RT = \dfrac{a+d}{a+d+2(b+c)}$,

2020 *Mathematics Subject Classification.* Primary 92D10; Secondary 92D20.

donde a es el número de bandas presentes (o compartidas) en los individuos i y j, b es el número de bandas presentes en el individuo i pero ausentes en el j, c es el número de bandas ausentes en el individuo i pero presentes en el j, y d es la frecuencia de bandas ausentes en ambos individuos.

Asimismo, es ampliamente utilizada en el ámbito de la genética molecular la distancia genética de Huff (***GD***; [7]):

$$GD = n\left[1 - \frac{2n_{xy}}{2n}\right],$$

donde n es el número de bandas consideradas y n_{xy} el de bandas compartidas.

A partir de los algoritmos de los coeficientes y de la distancia de Huff es posible establecer las siguientes relaciones funcionales: los pares de coeficientes ***A/J***, ***J/SD***, ***RT/SM*** satisfacen $y = \frac{x}{2-x}$, donde y y x son respectivamente el primer y segundo coeficiente de cada par. Además ***GD*** es igual a $1 - $ ***SM*** multiplicada por el número de bandas consideradas, n.

La elección del coeficiente de similitud debe basarse en ciertos criterios. En primer lugar es necesario que las propiedades de la distancia sean adecuadas para los diferentes análisis a realizar. Por ejemplo, con frecuencia se utilizan análisis multivariados de coordenadas principales, por lo que se requiere euclidianidad de la distancia obtenida mediante el índice. En segundo lugar hay consideraciones sobre la inclusión o exclusión de las co-ocurrencias negativas o dobles ausencias. Dado que en los marcadores moleculares dominantes la ausencia de la banda puede deberse a diferentes mutaciones, utilizar las dobles ausencias (d) en el algoritmo del índice no es aconsejable, recomendándose generalmente el uso del índice de Jaccard [9]. Sin embargo, que la ausencia de una determinada banda en diferentes individuos se deba a diferentes mutaciones es menos frecuente en individuos estrechamente relacionados, pertenecientes a grupos donde se puedan producir cruzamientos genéticos entre ellos. En esos casos es esperable que la ausencia de la banda se deba al mismo motivo mutacional, por lo que incluir las dobles ausencias como información resulta valioso. Por el contrario, en individuos que presentan aislamiento reproductivo, en donde la estructura poblacional es más marcada, la probabilidad de que la falta de una determinada banda se deba a diferentes mutaciones es más elevada, por lo que se deberían utilizar índices de similitud que no incluyan las dobles ausencias en su algoritmo.

Se plantea la hipótesis de que las relaciones entre individuos y poblaciones serán afectadas por la medida de distancia molecular seleccionada según su estructura poblacional. El objetivo del trabajo es analizar el efecto del uso de diferentes medidas de distancias en las relaciones obtenidas entre individuos de poblaciones relacionadas o aisladas.

2. MATERIALES Y MÉTODOS

A partir de 54 marcadores obtenidos con un marcador molecular dominante ISSR en 100 individuos de girasol silvestre pertenecientes a 10 poblaciones [6] se calcularon las medidas de distancia basadas en ***J***, ***SD***, ***O***, ***A***, ***SM***, ***RT*** y ***GD***. Estos individuos pertenecen a poblaciones estrechamente relacionadas con baja estructura poblacional. Por otro lado, se calcularon esas mismas medidas a partir de 144 marcadores ISSR en 80 individuos pertenecientes a ocho accesiones de mijo perenne [2]. Esos individuos presentan una marcada estructura poblacional.

Las matrices de distancias obtenidas con cada índice fueron comparadas mediante el test de correlación de Mantel [10, 16], y las diferencias en las relaciones entre individuos mediante la comparación visual de los dendrogramas generados mediante un ligamiento completo. Los análisis se realizaron usando Infostat [3] y GenAlEx6 (Genetic Analysis in Excel) [13, 14].

3. RESULTADOS Y DISCUSIÓN

Las correlaciones de Mantel entre matrices de distancia en girasol silvestre se muestran en la Tabla 1. Las correlaciones entre todos los pares de coeficientes resultaron altas, mayores a 0.961. Se encontraron dos grupos de coeficientes con las correlaciones más elevadas. El primer grupo con los coeficientes de ***J***, ***SD***, ***O*** y ***A***, con valores que oscilaron entre 0.987 y 0.998. Estos índices no consideran las dobles ausencias en su algoritmo. El segundo grupo integrado por los coeficientes de ***SM*** y ***RT***, con una correlación de 0.997. Ambos índices sí incluyen a las dobles ausencias en su cálculo. Las correlaciones entre los coeficientes que no incluyen las dobles ausencias y los que sí la incluyen resultaron elevadas (mayores a 0.961) aunque son menores a las observadas entre coeficientes en cada uno de los grupos mencionados. Otros autores encontraron similares relaciones entre los índices trabajando con otros tipos de marcadores moleculares dominantes, como por ejemplo [11] trabajando con AFLP (polimorfismos en la longitud de fragmentos amplificados) y [5] con RAPD (amplificación aleatoria de polimorfismos del ADN). Además de la inclusión o no de las dobles ausencias, es esperable una alta correlación entre las distancias obtenidas de los coeficientes que presentan relaciones funcionales no lineales entre sí, tales como los pares de coeficientes ***A***/***J***, ***J***/***SD***, ***RT***/***SM*** que satisfacen $y = \frac{x}{2-x}$, donde y y x son respectivamente el primer y segundo coeficiente de cada par. Además ***GD*** es igual a $1 - $ ***SM*** multiplicada por el número n de bandas consideradas, de ahí que su correlación es de 1.

A pesar de que las correlaciones entre los coeficientes que no incluyen las dobles ausencias y los que sí la incluyen resultaron elevadas (mayores a 0.961), la inspección visual de los conglomerados obtenidos con las matrices de distancia mostraron diferencias en las relaciones entre los individuos generadas. En la Figura 1 se muestran los cambios en la posición relativa de ciertos individuos de girasol silvestre en el dendrograma generado a partir de Jaccard y el generado a partir de Simple Matching.

Coeficientes	***J***	***SD***	***O***	***A***	***SM***	***RT***	***GD***
J	–						
SD	0.997	–					
O	0.996	0.998	–				
A	0.996	0.988	0.987	–			
SM	0.969	0.968	0.965	0.965	–		
RT	0.969	0.963	0.961	0.971	0.997	–	
GD	0.969	0.968	0.965	0.965	1.000	0.997	–

TABLA 1. Correlaciones de Mantel entre las matrices de distancia basadas en $1 -$ *coeficiente de similitud* obtenidas a partir de 54 marcadores ISSR en 100 individuos de girasol silvestre. ***J***: Jaccard, ***SD***: Sørensen–Dice, ***O***: Ochiai, ***A***: Anderberg, ***SM***: Simple Matching, ***RT***: Rogers–Tanimoto y ***GD***: distancia genética de Huff. Todos los coeficientes fueron significativamente diferentes de cero ($p < 0{,}01$).

En el caso de mijo perenne, las correlaciones entre las matrices de distancias resultaron más elevadas que en girasol silvestre, siendo mayores a 0.938. No se observó que los grupos de índices que no consideran las dobles ausencias y los que las consideran tuvieran mayores correlaciones entre sí que las que se observan entre ambos grupos, por lo que la inclusión

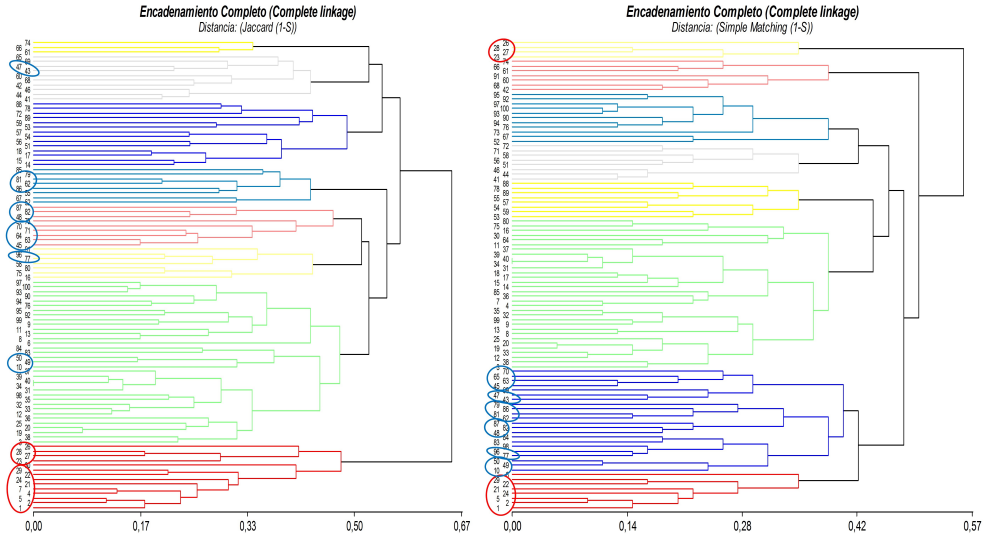

FIGURA 1. Dendrogramas obtenidos en girasol silvestre utilizando ***J*** y ***SM***, mediante un ligamiento completo.

de las dobles ausencias no parece generar cambios significativos en las distancias generadas. Mediante la inspección visual de los dendrogramas generados con Jaccard y Simple Matching en mijo perenne (Figura 2) no se observan cambios relevantes en las posiciones relativas de los individuos, manteniéndose dentro de los grupos definidos con ambos tipos de distancias.

Coeficientes	***J***	***SD***	***O***	***A***	***SM***	***RT***	***GD***
J	–						
SD	0.989	–					
O	0.989	1.000	–				
A	0.985	0.948	0.948	–			
SM	0.980	0.994	0.993	0.938	–		
RT	0.995	0.991	0.991	0.971	0.993	–	
GD	0.980	0.994	0.993	0.938	1.000	0.993	–

TABLA 2. Correlaciones de Mantel entre las matrices de distancia basadas en $1 -$ *coeficiente de similitud* obtenidas a partir de 144 marcadores ISSR en 80 individuos de mijo perenne. ***J***: Jaccard, ***SD***: Sørensen–Dice, ***O***: Ochiai, ***A***: Anderberg, ***SM***: Simple Matching, ***RT***: Rogers–Tanimoto y ***GD***: distancia genética de Huff. Todos los coeficientes fueron significativamente diferentes de cero ($p < 0{,}01$).

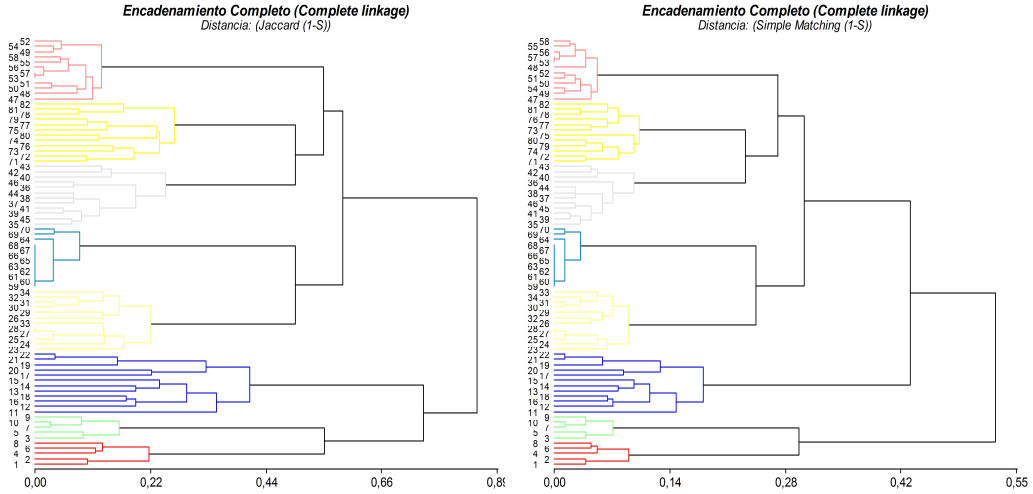

FIGURA 2. Dendrogramas obtenidos en mijo perenne utilizando ***J*** y ***SM***, mediante un ligamiento completo

4. CONCLUSIONES

La falta de banda en marcadores moleculares podría deberse a diferentes mutaciones en el ADN especialmente en genotipos poco emparentados, aunque si los individuos pertenecen a poblaciones emparentadas en donde la posibilidad de cruzamientos es alta la doble ausencia se debería principalmente al mismo motivo mutacional. Esto indica que se deberían incluir las dobles ausencias en el cálculo de la similitud (o distancia) genética. Dado que los individuos de girasol analizados en este trabajo pertenecen a poblaciones estrechamente relacionadas, con baja estructura poblacional, se deberían incluir las dobles ausencias en el índice a utilizar. Esto es un punto a tener en cuenta dado que se pudo observar que en girasol silvestre la utilización de coeficientes que no incluyen las dobles ausencias generó relaciones entre individuos que resultaron diferentes a las obtenidas con un coeficiente que sí las incluye. En estos casos es recomendable la utilización de la distancia basada en Simple Matching, dado que incluye las dobles ausencias, es euclidiana y es proporcional a la distancia euclídea entre los datos originales, por lo que permite aplicar un análisis de componentes principales, pudiéndose representar de manera conjunta individuos y variables.

En el caso de poblaciones con elevada estructura poblacional o aislamiento reproductivo que generen cambios significativos entre los grupos como en mijo perenne, se debería optar por la utilización de índices que no incluyan las dobles ausencias, ya que éstas podrían deberse a diferentes motivos mutacionales, tornándose dudoso si la falta de una determinada banda es un indicio de similitud entre individuos. Sería entonces el índice de Jaccard una apropiada elección. Sin embargo, en este trabajo se encontró que las relaciones entre individuos obtenidas con índices que incluyen o no las dobles ausencias resultaron similares. La fuerte estructura poblacional en este grupo de individuos, en donde las accesiones se diferenciaron marcadamente, generó que el uso de cualquiera de estos índices muestre los mismos agrupamientos.

Estos resultados parecen indicar que en poblaciones con baja estructura poblacional la inclusión o no de las dobles ausencias en el algoritmo del índice de similitud a utilizar genera cambios en las relaciones entre individuos obtenidas, mientras que si la estructura poblacional es fuerte el uso de cualquier índice daría resultados similares.

AGRADECIMIENTOS

A la Dra. Nélida Winzer y el Lic. Ricardo Camina por establecer las relaciones funcionales entre los índices y a las Dras. Lorena Armando y Alicia Carrera por sus valiosos comentarios.

REFERENCIAS

[1] Anderberg MR (1973) Clustering analysis for applications. London, Academic Press. MR 0326934.
[2] Armando LV, Tomas AC, Garayalde AF, Carrera A (2015) Assessing the genetic diversity of *Panicum coloratum* var. *makarikariense* using agro-morphological traits and microsatellite-based markers. *Ann. Appl. Biol.* 167, no. 3, 373–386. https://doi.org/10.1111/aab.12234
[3] Di Rienzo JA, Casanoves F, Balzarini MG, Gonzalez L, Tablada M, Robledo CW (2008) InfoStat. Versión 2008. Grupo InfoStat. FCA, Universidad Nacional de Córdoba, Argentina. http://www.infostat.com.ar
[4] Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology* 26, no. 3, 297–302. https://doi.org/10.2307/1932409
[5] Duarte MC, Santos JB, Melo LC (1999) Comparison of similarity coefficients based on RAPD markers in the common bean. *Genet. Mol. Biol.* 22, no. 3, 427–432. https://doi.org/10.1590/S1415-47571999000300024
[6] Garayalde AF, Poverene M, Cantamutto M, Carrera A (2011) Wild sunflower diversity in Argentina revealed by ISSR and SSR markers: an approach for conservation and breeding programs. *Ann. Appl. Biol.* 158, no. 3, 305–317. https://doi.org/10.1111/j.1744-7348.2011.00465.x
[7] Huff DR, Peakall R, Smouse PE (1993) RAPD variation within and among natural populations of outcrossing buffalograss [*Buchloë dactyloides* (Nutt.) Engelm]. *Theor. Appl. Genet.* 86, 927–934. https://doi.org/10.1007/BF00211043
[8] Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Voud. Sc. Nat.* 37, no. 142, 547–579. https://doi.org/10.5169/seals-266450
[9] Kosman E, Leonard KJ (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Mol. Ecol.* 14, no. 2, 415–424. https://doi.org/10.1111/j.1365-294X.2005.02416.x
[10] Mantel NA (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27, no. 2, 209–220. https://pubmed.ncbi.nlm.nih.gov/6018555/
[11] Meyer AD, Garcia AAF, de Souza AP, de Souza CL (2004) Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L). *Genet. Mol. Biol.* 27, no. 1, 83–91. https://doi.org/10.1590/S1415-47572004000100014
[12] Ochiai A (1957) Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions. *Bull. Jap. Soc. Sci. Fish.* 22, no. 9, 526–530. https://doi.org/10.2331/suisan.22.526
[13] Peakall R, Smouse PE (2006) GenAlEx 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, no. 1, 288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x
[14] Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinform.* 28, no. 19, 2537–2539. https://doi.org/10.1093/bioinformatics/bts460
[15] Rogers JS, Tanimoto TT (1960) A computer program for classifying plants. *Science* 132, 1115–1118. https://doi.org/10.1126/science.132.3434.1115
[16] Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Zool.* 35, no. 4, 627–632. https://doi.org/10.2307/2413122
[17] Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* 38, 1409–1438. https://www.biodiversitylibrary.org/page/3711319#page/387/mode/1up
[18] Sørensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. K. Danske Vidensk. Selsk. Biol. Skr. 5, no. 4. https://www.royalacademy.dk/Publications/Low/295_S%C3%B8rensen,%20Thorvald.pdf

(A. F. Garayalde) UNIVERSIDAD NACIONAL DEL SUR, DEPARTAMENTO DE MATEMÁTICA, AVENIDA ALEM 1253, BAHÍA BLANCA, ARGENTINA
Email address: agarayalde@criba.edu.ar