Macat, Paula[1,2*]
Kovalevski, Leandro[2*]
Quaglino, Marta[2]
Pratta, Guillermo[1,3]

[1] *Consejo Nacional de Investigaciones Científicas y Técnicas*

[2] *Instituto de Investigaciones Teóricos y Aplicados, Escuela de Estadística, Facultad de Ciencias Económicas y Estadística UNR, Rosario, Argentina*

[3] *Cátedra de Genética, Facultad de Ciencias Agrarias UNR, Zavalla, Argentina*

**Autores con igual contribución en esta Investigación*

BIOINFORMÁTICA EN EL ESTUDIO DE LA MADUREZ DE FRUTOS DE TOMATE

Resumen

La maduración del fruto de tomate es un proceso biológico afectado por diferentes fuentes multidimensionales de variación: el estado de madurez, el genotipo y la expresión de proteínas. El análisis de correspondencias (AC) es una técnica de escalamiento multidimensional que permite una rápida visualización de las asociaciones entre las diferentes fuentes de variación evaluadas mediante datos dicotómicos. El objetivo de este trabajo es visualizar el proceso de maduración del fruto de tomate mediante un AC que permite medir la contribución relativa de los diferentes genotipos, estados de madurez y bandas de polipéptidos (constituyentes de las proteínas) a la variación total observada durante todo el proceso, en una aplicación bioinformática a nivel individual de la organización de los sistemas biológicos.

Frutos de 15 genotipos (padres e híbridos de un diseño de cruzamientos dialélicos) fueron seleccionados por los perfiles polipeptídicos en SDS-PAGE, en los que se observaron 25 bandas en 4 estados de madurez: Verde Maduro (VM), Rojo Pintón (RP), Rojo Maduro en planta (RMp) y Rojo Maduro en estantería (RMe) de acuerdo a Marchionni Basté et al. (2014). Se realizó un análisis descriptivo univariado sobre los datos para evaluar la presencia de cada banda (total y por estado de madurez) y en segundo lugar un Análisis multivariado de Correspondencias (AC) en cada estado. Por último, se hizo un AC integrador a la base de datos completa.

Para la mayoría de las bandas de polipéptidos, su presencia varía a través de los diferentes estados de madurez. Una mayor variación entre los genotipos para la expresión de proteínas fue encontrada en los estados VM y RMp mediante el AC. Las dos primeras dimensiones explican el 35% de la variación total en el estado VM, que fue la etapa de madurez más variable para los perfiles polipeptídicos analizados. Los genotipos más divergentes y sus correspondientes polipéptidos asociados fueron variando de acuerdo al estado de madurez. Finalmente, el AC integrador identificó un híbrido como el individuo más variable a lo largo de la maduración, y siete bandas de polipéptidos altamente asociadas a su comportamiento discrepante en relación con los otros genotipos del cruzamiento dialélico.

Palabras claves: Análisis de Correspondencias; Bioinformática; Proteómica de la madurez del fruto de tomate.

1. Introducción

El tomate (*Solanum lycopersicum*) es un fruto climatérico cuya maduración se caracteriza por cambios secuenciales en la expresión de proteínas, dando como resultado diferentes perfiles de bandas polipeptídicas en cada estado de madurez (Giovannonni, 2004). Sin embargo, frutos de diversos genotipos de tomate varían en su madurez (Rodriguez et al., 2008). Por lo tanto la maduración del fruto de tomate es un proceso biológico afectado por fuentes multidimensionales de variación: el estado de madurez, el genotipo y la expresión de proteínas.

El análisis de correspondencias (AC) es una técnica de escalamiento multidimensional que permite una rápida visualización de las asociaciones entre las diferentes fuentes de variación evaluadas mediante datos dicotómicos (Lebart et al., 1984). AC se aplicó en estudios de microarrays (Fellenberg et al., 2001) y de proteínas funcionales (Chang et al., 2013). El objetivo de este trabajo es visualizar el proceso de maduración del fruto de tomate mediante un AC que permite medir la contribución relativa de los diferentes genotipos, estados de madurez y bandas de polipéptidos a la variación total observada durante todo el proceso, en una aplicación bioinformática a nivel individual de la organización de los agentes biológicos.

2. Materiales y Métodos

Frutos de 15 genotipos (cinco líneas puras recombinantes -RIL- y sus diez Híbridos de Segundo Ciclo -SCH- obtenidos según cruzamientos dialélicos) fueron seleccionados por sus perfiles polipeptídicos en SDS-PAGE, compuestos por 25 bandas en 4 estados de madurez: Verde Maduro (VM), Rojo Pintón (RP), Rojo Maduro en planta (RMp) y Rojo Maduro en estantería (RMe) de acuerdo a Marchionni Basté et al. (2014). Una base de datos de dimensión 15 x 25 x 4 se analizó en primer lugar, mediante un análisis descriptivo univariado para evaluar la presencia de cada banda (total y por estado) y en segundo lugar, mediante el Análisis multivariado de Correspondencias (AC) en cada estado de madurez. Por último, se hizo un AC integrador a la base de datos completa.

Análisis de Correspondencias

El Análisis de Correspondencias es una técnica con la cual es posible encontrar una representación multidimensional de la dependencia entre las filas y columnas de una tabla de contingencia bidimensional. El objetivo del análisis es poder interpretar, resumir los datos, dar sentido a los ejes, detectar los puntos singulares, determinar las relaciones causa-efecto y las proximidades significativas.

2.1. Tabla de contingencia y notaciones preliminares

Se considera la tabla de contingencia bidimensional F con elementos f_{ij}, formada por las variables X (filas) e Y (columnas) con "I" y "J" categorías respectivamente definidas sobre una población o una muestra. Se define con "n" al número de elementos de la población o una muestra. Se define con "n" al número de elementos de la población clasificados según las categorías de X e Y. Un subíndice se reemplaza por "+" cuando se suma a través de la correspondiente variable, obteniéndose las siguientes relaciones:

$f_{i+} = \sum_j f_{ij}$

$f_{+j} = \sum_i f_{ij}$

$f_{++} = \sum_i f_{i+} = \sum_j f_{+j} = \sum_i \sum_j f_{ij} = n$, para todo i = 1,…,I para todo j = 1,…,J

donde f_{i+} representa la frecuencia de la categoría i, f_{+j} representa la frecuencia de la categoría j y f_{++} coincide con el valor de "n" igual al número total de elementos.

Se llama perfil de un elemento i de las I filas a los valores $f_{j/i} = f_{ij} / f_{i+}$ siendo $f_{i+} \neq 0$. Se llama perfil de un elemento j de las J columnas a los valores $f_{i/j} = f_{ij} / f_{+j}$ siendo $f_{+j} \neq 0$.

Luego se tiene que $\sum_j f_{j/i} = 1$ para todo i fijo, con i = 1,…,I y $\sum_i f_{i/j} = 1$ para todo j fijo, con j = 1,…,J.

Se observa que estas notaciones son similares a las usuales en el cálculo de probabilidades.

2.2. Hipótesis de Independencia

Se dice que hay independencia entre dos variables aleatorias categóricas X e Y, si para todo i = 1,…,I y para todo j = 1,…,J se tiene $p_{ij} = p_{i+} . p_{+j}$, simbolizando con p_{ij} a la probabilidad de aparición conjunta de la categoría i de X (i = 1,…,I) y j de Y (j = 1,…,J), con p_{i+} a la probabilidad marginal de la categoría i de X y con p_{+j} a la probabilidad marginal de la categoría j de Y.

En términos de las frecuencias empíricas, esta relación de independencia es la siguiente:

$f_{ij} = f_{i+} . f_{+j} / n$.

El clásico test Chi-cuadrado de Pearson para las tablas de contingencia permite apreciar la diferencia entre las probabilidades empíricas dadas por f_{ij} y $f_{i+} . f_{+j} / n$. Bajo la hipótesis de independencia, resulta que, para cualquier j: $f_{ij} / f_{i+} = f_{+j} / n$.

Si todos los perfiles de filas son iguales entre sí y por consecuencia idénticos al perfil medio correspondiente, existe independencia entre X e Y. Lo mismo ocurre para los perfiles columnas donde cualquiera sea i: $f_{ij} / f_{+j} = f_{i+} / n$.

De esta manera, examinar las proximidades entre cada perfil y su perfil medio, permitirá estudiar la relación entre dos variables nominales, es decir, la discrepancia de la independencia.

2.3. Representación gráfica

La tabla F de datos originales no presenta interés en este análisis, sino la tabla de perfiles filas y la de perfiles columnas, es decir, las distribuciones condicionales de una fila o de una columna, permitiendo comparar las categorías de una misma variable.

Geométricamente, un elemento i de las I filas está representado por un vector de R^J, vector cuyas componentes son los perfiles $\{f_{ij} / f_{i+}$, j = 1,…,J$\}$. Es lo mismo para un elemento j de las J columnas que está representado por un vector de R^I, vector cuyas componentes son los perfiles $\{f_{ij} / f_{+j}$, i = 1,…,I$\}$. Dos elementos i e i' de las I filas serán idénticos sí y sólo sí sus perfiles correspondientes son idénticos, es decir si dos filas i e i' de la tabla F son proporcionales.

Conjunto de puntos filas: El conjunto de perfiles fila forma una nube de I puntos en el espacio de las J columnas notada N(I). Cada punto i tiene por coordenada en R^J a $\{f_{ij} / f_{i+}$, j = 1,…,J$\}$ y está afectado por la masa f_{i+} / n que es su frecuencia relativa. Como $\sum_j f_{ij} / f_{i+} = 1$, los I puntos de la nube están situados en un subespacio de (J-1) dimensiones. El centro de gravedad de esta nube es la media de perfiles filas afectados por sus masas y corresponde al perfil medio. Su componente j-ésima será: $\sum_i f_{i+} / n . f_{ij} / f_{i+} = f_{+j} / n$, la cual es la frecuencia marginal de las columnas.

Conjunto de puntos columnas: De la misma manera, el conjunto de los J perfiles columnas

constituye una nube de J puntos en el espacio de las I filas notada N(J). Las coordenadas en R^I del punto j son dadas por: $\{f_{ij} / f_{+j}$, i = 1,…,I$\}$. Cada punto está afectado por una masa f_{+j} / n. Los J puntos de la nube están situados en un subespacio de (I-1) dimensiones ya que $\sum_i f_{ij} / f_{+j} = 1$. El centro de gravedad de la nube de los puntos columnas es el perfil medio. La componente i-ésima será: $\sum_j f_{+j} / n \; . \; f_{ij} / f_{+j} = f_{i+} / n$, la cual es la frecuencia marginal de las filas.

Elegir a los perfiles como coordenadas para graficar las nubes en R^J y R^I conduciría a dar a todas las categorías de X e Y la misma importancia. Sin embargo esta importancia es modificada a través de la masa que afecta a cada punto (proporcional a su frecuencia) a fin de no privilegiar las categorías de totales pequeños y de reflejar la distribución real de la población. El punto i de R^I, tiene por lo tanto un peso de f_{i+} / n, y el peso del punto j de R^I es f_{+j} / n. Esta masa interviene por una parte en el cálculo de las coordenadas del centro de gravedad de la nube y por otra parte en el criterio de ajuste, el cual se definirá más adelante.

2.4. Elección de las distancias

La distancia euclídea en el espacio de los perfiles sobre J (para el conjunto N(I)), cuyo cuadrado se define como:

$$\partial^2 (i,i') = \sum_j (f_{ij} / f_{i+} - f_{i'j} / f_{i'+})^2 \qquad \text{para todo } j = 1,...,J$$

Favorece a las columnas que tienen una masa f_{+j} / n importante.

Otra métrica utilizada, diferente a la distancia euclídea común, es la llamada “distancia Chi-cuadrado”, que resulta de ponderar cada desviación por la inversa de la masa de la columna. Esta distancia al cuadrado entre los perfiles filas se definen como:

$$\partial^2 (i,i') = \sum_j [1 / (f_{+j} / n)] \; . \; (f_{ij} / f_{i+} - f_{i'j} / f_{i'+})^2 \qquad \text{para todo } j = 1,...,J$$

Cuando i e i’ tienen el mismo perfil ($f_{ij} / f_{i+} = f_{i'j} / f_{i'+}$ para todo j) el resultado es que $\partial^2 (i,i') = 0$.

De la misma manera, la distancia al cuadrado entre los perfiles columnas, está dada por:

$$\partial^2 (j,j') = \sum_i [1 / (f_{i+} / n)] \; . \; (f_{ij} / f_{+j} – f_{ij'} / f_{+j'})^2 \qquad \text{para todo } i = 1,...,I$$

La inercia total de las nubes de puntos de puntos filas (o de puntos columnas) calculada con esta distancia es proporcional a la clásica Chi-cuadrado de Pearson utilizada para probar la independencia de las filas y las columnas de una tabla de contingencia. Esta distancia ponderada Chi-cuadrado, así como el rol simétrico jugado por las filas y las columnas de la tabla de contingencia que particularizan al análisis de correspondencias, asegura dos propiedades remarcables: la equivalencia distribucional y las relaciones de transición.

2.5. Equivalencia distribucional

La propiedad de equivalencia distribucional permite sumar dos categorías de una misma variable (por ejemplo, los perfiles filas i’ e i’’ en R^J), en una nueva categoría (perfil fila i), afectada por la suma de las masas de i’ e i’’. Los dos puntos i’ e i’’ estarán fusionados y las distancias entre las categorías de la variable X y las distancias entre las categorías de la variable Y permanecerán invariantes. Ocurre lo mismo para los perfiles columnas en R^I.

Esta propiedad es fundamental, ya que garantiza la invariancia de los resultados respecto a cómo las categorías fueron originalmente codificadas. Por lo tanto, no hay pérdida de información cuando se juntan categorías y viceversa, no hay nada para ganar subdividiéndolas.

2.6. Reducción de la dimensionalidad

Los puntos usualmente están situados en espacios de una alta dimensionalidad en los cuales se hace imposible poder observarlos. Sin embargo, podría ocurrir que los puntos no "llenen" el espacio de dimensión total de igual manera en todas las direcciones, sino que se sitúen aproximadamente en un subespacio de menor dimensión. Si se podría identificar este subespacio de menor dimensión, (preferentemente no más de dos o tres dimensiones) ubicado cerca de todos los puntos, entonces se podrían proyectar los puntos en este subespacio y considerar las posiciones proyectadas (de los puntos en este subespacio) como una aproximación a las posiciones verdaderas en el espacio original.

A partir de este proceso de reducción de dimensionalidad se pierde el conocimiento de qué tan lejos y en qué dirección los puntos se sitúan "fuera" de este subespacio, pero se gana visión de los puntos que de otra manera sería imposible de ver.

Para esto, se utiliza la inercia como una medida de la variación total o dispersión geométrica. Las distancias entre los puntos proyectados son distancias chi-cuadrado aproximadas, dado que las distancias chi-cuadrado exactas se dan sólo en el espacio de dimensión completa. Las distancias en el subespacio son siempre iguales o menores a las correspondientes en el espacio de dimensión completa.

El siguiente paso consiste en definir la "cercanía" o el "mejor ajuste" de los puntos al subespacio, es decir, el criterio de ajuste.

2.7. Reglas de interpretación

La interpretación de las nubes de puntos está basada en las distancias chi-cuadrado: cuando dos puntos filas (o dos puntos columnas) están cerca, sus perfiles son similares. Cuando los perfiles difieren considerablemente, la distancia entre los puntos es grande.

También es legítimo interpretar las posiciones relativas de un punto de un conjunto con respecto a todos los puntos del otro conjunto: cada punto es un "promedio ponderado" de todos los puntos columnas y viceversa. Si un punto fila y un punto columna están próximos, significa que esa combinación se da con alta frecuencia.

El centro de gravedad, ubicado en el origen de los ejes, corresponde a los perfiles promedios de ambos conjuntos de puntos. Por lo tanto, un perfil cercano al origen, no difiere mucho del perfil promedio y tiene una distribución indiferenciada entre las categorías de la otra variable.

A cada eje le corresponde un porcentaje que es la porción de variancia explicada por el mismo. Este porcentaje da una idea conservadora de la porción de información explicada por los ejes principales. Es conservadora porque es sólo una forma parcial de medir la información; puede ocurrir que, a pesar de porcentajes pequeños, los ejes correspondientes restauren gran parte de la información contenida en el conjunto de datos.

Estas reglas de interpretación se mantienen perfectamente en un espacio de dimensión completa. Sin embargo, es posible evaluar hasta qué punto unas pocas primeras dimensiones proveen una buena representación de ese espacio. Esta evaluación se realiza a través de la inspección de la secuencia de autovalores y porcentajes de inercia, según se detalla a continuación.

2.7.1. Inercia y test de independencia

El valor de la inercia global está ligado al test clásico de chi-cuadrado. La inercia total I de la nube de puntos en relación al centro de gravedad se escribe por definición:

$I = \sum_i (f_{i+}/n) \cdot \partial^2 (f_{ij} / f_{i+} , f_{+j} / n) = \sum_j f_{+j} / n \cdot \partial^2 (f_{ij} / f_{+j} , f_{i+} / n) = \sum_i \sum_j 1 / (f_{i+} \cdot f_{+j}) \cdot (f_{ij} - f_{i+} \cdot f_{+j} / n)^2$

es decir, es el promedio ponderado de las distancias chi-cuadrado entre los perfiles fila (o perfiles columna) y su perfil promedio. Siendo el total de individuos igual a "n", se reconoce en nI la estadística que asintóticamente se distribuye siguiendo la distribución chi-cuadrado con (I-1)(J-1) grados de libertad. Bajo las hipótesis de independencia: $X^2 = nI$. La inercia se expresa igual a: $I = tr\ \Lambda = \sum_\alpha \lambda_\alpha$, para todo $\alpha = 1,\ldots,p$. Es decir, la suma de los autovalores no triviales de un análisis de correspondencias tiene una interpretación estadística simple. Se puede rechazar la hipótesis nula de independencia de las variables en filas y en columnas si el valor observado X^2 es mayor que un valor X_0^2 que antiacumula en la distribución una probabilidad igual o inferior a un nivel de significación fijado previamente.

El valor de la inercia global es un indicador de la dispersión de la nube y mide la relación entre las dos variables. Sin embargo, no resulta interesante sólo la dispersión de la nube, sino la existencia de direcciones privilegiadas en esa nube.

Por lo tanto, conviene observar las inercias (autovalores) como así también los porcentajes de inercia (o partición de X^2) de cada eje en relación a los otros. Si la forma de la nube es esférica significa que no hay direcciones privilegiadas. Cuando existen relaciones privilegiadas, el análisis de correspondencias intervendrá con utilidad para describir esta dependencia entre filas y columnas.

De manera general, dos variables son independientes si los perfiles de sus categorías son similares a los perfiles medios: la inercia total es débil y no existe dirección privilegiada, es decir, todos los puntos están concentrados alrededor del centro de gravedad de la nube siguiendo una forma esférica.

La inercia de un factor no puede ser superior a 1. Un autovalor cercano a 1 indica una dicotomía al nivel de los datos; obteniendo para cada variable dos grupos de categorías que separan la nube de puntos en dos sub-nubes. Cuando dos autovalores son cercanos a 1, se obtienen tres sub-nubes y las categorías de las variables se descomponen en tres grupos. Si todos los auntovalores son cercanos a 1, cada categoría de una variable está en correspondencia casi exclusiva con una sola categoría de la otra variable.

Inercia de fila (o columna): en términos geométricos, es el producto de la masa de la fila (o columna) por su distancia chi-cuadrado al perfil promedio. Como cada celda realiza una contribución positiva a la inercia total, entonces cada contribución puede expresarse relativa a su total.

Si todas las filas (columnas) tuvieran el mismo perfil en una cierta categoría de columna (fila), entonces el punto de esta categoría de columna (fila) se ubicará en el centro del gráfico, mientras que las categorías con grandes variaciones a través de las filas (columnas) estarán lejos del centro del gráfico.

2.7.2. Contribuciones absolutas y relativas

Ya se ha visto como la inercia puede descomponerse entre los ejes factoriales, entre las filas y entre las columnas. En esta sección, se utilizará además la descomposición de la inercia en componentes de fila y en componentes de columna. Estas contribuciones de inercia se expresan mejor en cifras relativas y son las siguientes:

Contribuciones absolutas: Indican la proporción de variancia explicada por cada categoría de la variable en relación a cada eje principal: es decir, determinan cuánto contribuye cada categoría de la variable a la formación de cada factor. Para los perfiles filas estas contribuciones están definidas por el cociente:

$$Cr_\alpha (i) = (f_{i+} /n) \cdot r_{i\alpha}^2 / \lambda_\alpha$$

el cual permite conocer en qué proporción un punto i contribuye a la inercia λ_α de la nube proyectada sobre el eje α, en otras palabras, permite diagnosticar cuáles puntos i tuvieron un rol principal en la orientación del eje α. Se observa que para todo eje α:

$$Cr_\alpha (j) = (f_{+j} / n) \, . \, c_{j\alpha}{}^2 / \lambda_\alpha \qquad \text{con} \ \sum_j Cr_\alpha (j) = 1$$

Contribuciones relativas: También llamadas cosenos cuadrados, expresan la proporción de la variancia de una cierta categoría, explicada por un factor particular. Es decir, determinan qué tan bien una categoría de la variable es representada por un factor. El cuadrado de la distancia de un punto al centro de gravedad se descompone en suma de cuadrados de las coordenadas sobre los ejes.

Para un punto i de R^J, se tiene:

$$\partial^2 (i \, , (f_{+j} / n)^{1/2}) = \sum_j [\, f_{j/i} \, (f_{+j} / n)^{-1/2} \,]^2 \qquad \text{con } i = f_{j/i} \, (f_{+j} / n)^{-1/2}$$

la cual se anula cuando el punto es igual al perfil medio.

El cuadrado de la proyección del punto i sobre el eje α vale: $\partial_\alpha{}^2(i, (f_{+j} / n)^{1/2}) = r_{i\alpha}{}^2$

con: $\sum_\alpha \partial_\alpha{}^2(i, (f_{+j} / n)^{1/2}) = \partial^2(i, (f_{+j} / n)^{1/2})$

La calidad de representación de un punto I sobre el eje α puede evaluarse por los cosenos del ángulo entre el eje y el vector que une al centro de gravedad de la nube con el punto i:

$$Cos_\alpha{}^2(i) = \partial_\alpha{}^2(i, (f_{+j} / n)^{1/2}) / \partial^2(i, (f_{+j} / n)^{1/2}) = r_{i\alpha}{}^2 / \partial^2(i, (f_{+j} / n)^{1/2})$$

Cuanto mayor es este coseno cuadrado, la posición del punto observado proyectado estará más cerca de la posición real del punto en el espacio, por lo tanto un valor pequeño de este coseno indicaría una pobre representación de dicho punto en ese eje. En tal caso, su posición deberá interpretarse con mayor cautela. La calidad de representación de un punto en un plano se obtiene de la suma de los cosenos cuadrados sobre los ejes estudiados, dado que para todo i, es: $\sum_\alpha Cos_\alpha{}^2(i) = 1$.

De la misma manera, se puede medir la contribución relativa del factor α a la posición del punto j por los cosenos cuadrados de j:

$$Cos_\alpha{}^2(j) = \partial_\alpha{}^2(j, (f_{i+} / n)^{1/2}) / \partial^2(j, (f_{i+} / n)^{1/2}) = c_{j\alpha}{}^2 / \partial^2(j, (f_{i+} / n)^{1/2}) \qquad \text{con } j = f_{i/j} \, (f_{i+} / n)^{-1/2}$$

Cumpliéndose también que para todo j: $\sum_\alpha Cos_\alpha{}^2(j) = 1$.

Relación entre las contribuciones absolutas y relativas:

- Se espera que un punto que haya tenido una gran contribución a un eje, se encuentre ubicado bastante cerca al eje y generalmente es lo que ocurre.
- No sólo los puntos con una gran contribución son los que tienen altas correlaciones con los ejes; un punto con una contribución pequeña a la orientación del eje podría ubicarse en la dirección de este eje y estar bien explicado por el mismo.

Observación: para las contribuciones absolutas y las relativas, no hay niveles de significación a partir de los cuales se pueda decir que un valor es “fuerte” o “débil”. Las apreciaciones se hacen empíricamente, en función del conjunto de valores calculados y varían según los datos. Es usual multiplicar por 100 a las contribuciones absolutas, expresando en porcentaje la parte que le corresponde a cada punto.

3. Resultados

La presencia global de todas las bandas de polipéptidos (definida como proporción de polipéptidos presentes) en los 4 estados de madurez para los 15 genotipos fue de 0.52, con valores de 0.46 en VM, 0.55 en el RP, 0.53 en RMp y 0.54 en RMe. La presencia global mínima y máxima de cada banda varió entre 0.05 (casi ausente) y 1 (presencia completa) para dos polipéptidos dados. Para la mayoría de las bandas de polipéptidos, su presencia varía a través de los diferentes estados de madurez. Algunos polipéptidos fueron más frecuentes en las etapas de madurez más avanzadas, mientras que otros estaban sólo presentes en las primeras etapas.

Una mayor variación entre los genotipos para la expresión de proteínas fue encontrada en los estados VM y RMp mediante el AC (Tabla 1), apoyando la hipótesis de que es de esperar una diversidad genética más amplia para los rasgos del fruto que están menos expuestos a las presiones de la selección natural (Marchionni Basté et al., 2014).

Tabla 1. Resultados del AC para cada estado de madurez del fruto de tomate.

Estado de madurez	Inercia Principal Total	Chi-Cuadrado
VM	0.873	150.986
RP	0.617	127.165
RMp	0.656	130.581
RMe	0.529	106.976

Las dos primeras dimensiones explican el 35% de la variación total en el estado VM, que fue la etapa de madurez más variable para los perfiles polipeptídicos analizados. Dos RIL y dos SCH se diferencian claramente del resto de los genotipos en esta etapa, la mayoría de las bandas polipeptídicas asociadas a cada uno de estos cuatro genotipos son completamente opuestas en su presencia (Figura 1).

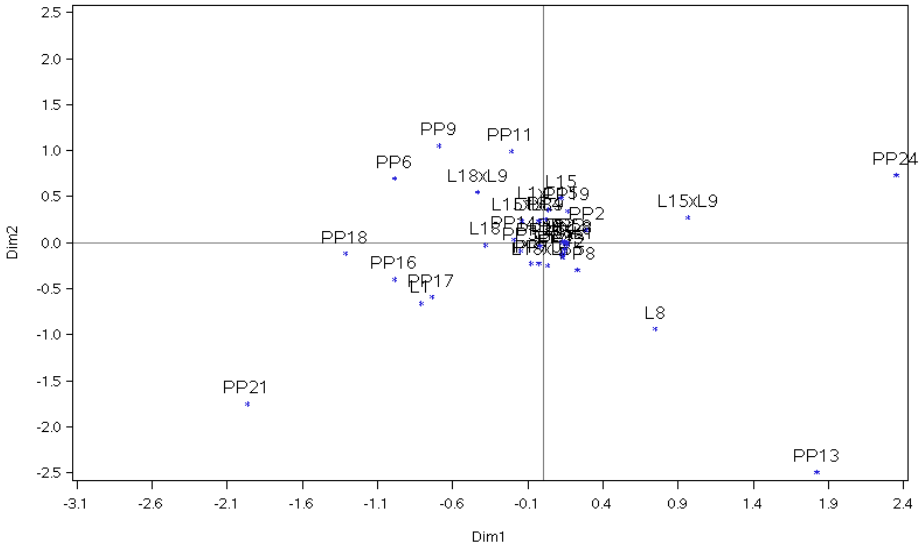

Figura 1. Representación de las 25 bandas polipeptídicas (PP) y los 15 genotipos (RILs indicados como LN y SCH indicados como $LN_X x LN_Y$, siendo N el número asignado a cada RIL obtenido por los criadores de tomates) de acuerdo al AC en el estado de madurez VM.

Respecto a los demás estados de madurez, las dos primeras dimensiones explican el 37% de la variación total en el RP, el 53% en el RMp y el 48% en el RMe. Los genotipos más divergentes y sus correspondientes polipéptidos asociados fueron variando de acuerdo a la etapa de madurez, verificando que el proceso de maduración se ve afectado en forma conjunta por las tres fuentes de variación consideradas en este trabajo, es decir, es un proceso biológico multidimensional.

Finalmente, el AC integrador identificó un híbrido como el individuo más variable a lo largo de la maduración, y siete bandas de polipéptidos altamente asociadas a su rendimiento discrepante en relación con los otros genotipos del cruzamiento dialélico (Figura 2).

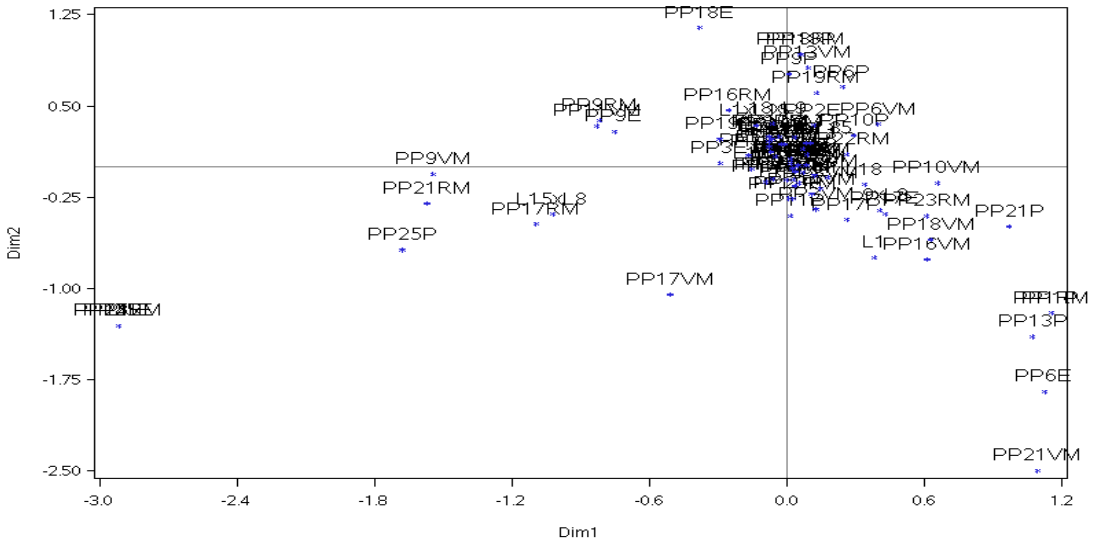

Figura 2. Representación de las 25 bandas polipeptídicas (PP) y los 60 genotipos (RILs indicados como LN y SCH indicados como LN_XxLN_Y, siendo N el número asignado a cada RIL obtenido por los criadores de tomates) de acuerdo al AC integrador de los cuatro estados de madurez.

4. Conclusiones

La visualización del proceso de maduración del fruto de tomate en cuatro estados de madurez permite medir la contribución relativa de la diversidad genética y la proteómica de este proceso biológico multidimensional. La aplicación bioinformática a nivel individual de la organización fue eficiente para la identificación de los genotipos más variables y sus bandas de polipéptidos asociados a cada etapa de madurez diferente y a lo largo de la madurez completa.

REFERENCIAS BIBLIOGRÁFICAS

Chang, J. M.; Taly, J. F.; Erb, I; Sung, T. Y.; Hsu, W. L.; Tang, C. Y.; Notredame, C. and Su ECY (2013). Efficient and interpretable prediction of protein functional classes by Correspondence Analysis and Compact Set Relations. PLOS 2013, 8: e75542. doi:10.1371/journal.pone.0075542.

Fellenberg, K.; Hauser, N. C.; Brors, B.; Neutzner, A.; Hoheisel, J. D. and Vingron, M. (2001). Correspondence analysis applied to microarray data. PNAS, 98, 10781-86.

Giovannonni, J. J. (2004). Genetic regulation of fruit development and ripening. The Plant Cell, 16, S160-76.

Greenacre, M.; Blasius, J. (1994). Correspondence Analysis in the Social Sciences. Harcourt brace & Company, Publishers.

Lebart, L.; Morineau, A. and Warwick, K. M. (1984). Multivariate Descriptive Statistical Analysis. New York, J. John Wiley & Sons, Inc.

Macat, P.; Kovalevski, L.; Quaglino, M. and Pratta, G. (2014). Visualization of genetic and proteomic biodiversity in four maturity stages of tomato fruit ripening.

Marchionni Basté, E.; Pereira da Costa, J. H.; Rodríguez, G. R.; Zorzoli, R. and Pratta, G. R. (2014). Genetic analysis of tomato fruit ripening at polypeptide profiles level through quantitative and multivariate approaches. American Journal of Plant Sciences, 5, 1926-35.

Rodriguez, G. R.; Sequin, L.; Pratta, G. R.; Zorzoli, R. and Picardi, L. A. (2008). Protein profiling in F1 and F2 generations of two tomato genotypes differing in ripening time. Biología Plantarum, 52, 548-52.