Inteligencia Artificial 44(2009), 56-66
doi: 10.4114/ia.v13i44.1046

INTELIGENCIA ARTIFICIAL

http://erevista.aepia.org/

Métodos de agrupamiento no supervisado para la integración de datos genómicos y metabólicos de múltiples líneas de introgresión

D. Milone[1], G. Stegmayer[2], M. Gerard[1,2], L. Kamenetzky[3], M. López[3] y F. Carrari[3]
[1]SINC-FICH-UNL, CONICET, Ciudad Universitaria UNL - Santa Fe (Argentina)
[2]CIDISI-UTN-FRSF, CONICET, Lavaise 610 - Santa Fe (Argentina)
[3]IB-INTA, CONICET, Castelar - (Argentina)
d.milone@ieee.org

Abstract Las numerosas aplicaciones de la inteligencia artificial a la biología de sistemas han dado lugar a nuevos algoritmos, además de la adaptación y reutilización de los existentes. En tareas de minería de datos se han aplicado diversos métodos estándar, como por ejemplo el bien conocido k-medias. Sin embargo, las capacidades de estos métodos son limitadas en relación a otros algoritmos más recientes, tanto en su desempeño para el agrupamiento de patrones como para la representación e interpretación de los resultados obtenidos. En este trabajo se compara el desempeño de tres métodos de agrupamiento no supervisado en la tarea de integración y descubrimiento de relaciones entre variaciones en los contenidos de metabolitos y la expresipon de genes de frutos de tomate. Los métodos considerados son el k-medias, el agrupamiento jerárquico y un método recientemente propuesto que se basa en mapas auto-organizativos. Se presentan los resultados obtenidos del análisis objetivo de la calidad de los agrupamientos y su significancia biológica. El modelo auto-organizado ha mostrado las más altas tasas de desempeño en las medidas de cohesión y separación, brindando además la máxima coherencia de las agrupaciones obtenidas desde el punto de vista del significado biológico.

Keywords: Métodos de agrupamientos no supervisado, integración de datos genómicos y metabólicos, líneas de introgresión.

1. Introducción

El procesamiento y descubrimiento de relaciones en la enorme cantidad de datos que deben analizarse en ciertas áreas de la bioinformática representan actualmente grandes desafíos. Desde el punto de vista de la aplicación biológica, la integración de estos datos podría poner de manifiesto relaciones ocultas que permitan inferir nuevos conocimientos acerca de los procesos biológicos que los involucran. Sin embargo, descubrir patrones ocultos en datos de expresión génica y perfiles metabólicos de plantas de interés económico para la agrobiotecnología es actualmente un reto ya que, además del gran volumen de datos, el empleo de algún tipo de algoritmo para reconocimiento de patrones se ve entorpecido por la llamada maldición de la dimensionalidad. Esto pone en evidencia la necesidad de desarrollar nuevas técnicas tendientes a superar las limitaciones de las existentes, por lo cual muchos métodos tradicionales de agrupación fueron adaptados para tales fines en los inicios de la bioinformática [12]. Entre los métodos aplicables en el área, en el caso del descubrimiento de clases se exploran los datos desde el punto de vista de la existencia o no de relaciones y mecanismos desconocidos para formular hipótesis que expliquen estos

ISSN: 1988-3064(on-line)

mecanismos. Por ejemplo, el algoritmo de agrupación jerárquica es un método determinista que ha sido aplicado para el descubrimiento de relaciones en datos genómicos y otras tareas similares [16]. En este algoritmo se establecen pequeños grupos de genes/condiciones que tienen un patrón de expresión común y posteriormente construye un dendrograma de forma secuencial. Este algoritmo, sobre la base de una matriz de distancia, permite inferir un árbol para los compuestos que luego es podado y a partir de las ramas de este árbol se pueden detectar grupos con características comunes y definir clases que identifiquen a estos grupos [2]. En cuanto a los algoritmos de tipo no-jerárquicos, generalmente se comienzan a calcular las distancias a partir de un número predefinido de grupos y se van colocando de forma iterativa los genes en los diferentes grupos hasta minimizar la dispersión interna de cada uno. El algoritmo más representativo de este tipo de agrupación es k-medias [3].

Un problema de actual interés consiste en detectar la presencia de genes y metabolitos relacionados por los mismos mecanismos regulatorios. En particular la integración de datos transcriptómicos y metabolómicos de plantas, relacionando perfiles de transcripción con variaciones en los perfiles de un gran número de moléculas no proteicas, puede ser usado para identificar cambios fenotípicos silenciosos asociados a vías metabólicas[1] [1]. La determinación de los enlaces entre genes, proteínas y reacciones es una tarea no trivial y de especial interés para la reconstrucción de una red metabólica, la cual podría intervenir en la obtención de un producto final con una determinada característica [9].

Existen trabajos recientes tendientes a mejorar el desempeño de los algoritmos mencionados, en los que se proponen el uso de técnicas de la inteligencia computacional [7], y en particular las redes neuronales artificiales del tipo mapas auto-organizativos (SOM del inglés Self-Organizing Maps) [8] para manejar grandes dimensiones y evidenciar, al mismo tiempo, patrones de relaciones ocultas en datos metabólicos [10]. En [5] se usa un modelo SOM para el análisis integrado y temporal de datos del metaboloma y transcriptoma de la planta *Arabidopsis thaliana*. Un trabajo relacionado [19] muestra que un modelo SOM para agrupar este tipos de datos ha sido de ayuda para explicar un mecanismo metabólico en respuesta a una deficienca de ácido sulfúrico. Los resultados obtenidos muestran que genes relacionados se agrupan en las mismas neuronas o en neuronas vecinas entre sí. El examen manual de esos agrupamientos fue de ayuda para la deducción de funciones de genes involucrados en la biosíntesis de un determinado compuesto. Sin embargo, el experimento y el modelo fueron específicamente diseñados para seguir la evolución temporal de una condición pre-establecida (deficiencia de ácido sulfúrico y nitrógeno), y por lo tanto el modelo fue más bien diseñado para corroborar una hipótesis y no para descubrir nuevas relaciones entre los datos. Sin embargo, en la mayoría de los casos, los grupos no se conocen *a-priori* y el interés se centra, justamente, en encontrarlos sin la ayuda de una variable de respuesta.

Además, en muchos experimentos biológicos no se pretende estudiar la evolución temporal de una condición particular, sino que el interés se centra en el estudio de las diferencias entre los genomas de varias plantas. Por ejemplo, puede querer estudiarse el genoma original de una planta que ha sido modificado a través de líneas de introgresión (ILs, del inglés Introgression Lines). Una línea de introgresión se define como un genotipo que contiene material genético derivado de una especie similar, como una especie silvestre. Esto puede ser de utilidad para estudiar nuevos genotipos introduciendo rasgos exóticos, para domesticación de cultivos o para identificar puntos biológicamente significativos (marcadores) que están ocultos dentro de una gran cantidad de mediciones analíticas de, por ejemplo, acumulación de metabolitos.

Para estas tareas, varias herramientas de software han aparecido recientemente. MarVis [6] realiza minería de datos solamente en perfiles de intensidad de metabolitos usando un mapa auto-organizativo en una dimensión. KaPPA-view [17] es una herramientas basada en la web para la representación cuantitativa de datos de transcriptos o metabolitos individualmente sobre vías metabólicas de plantas.

Específicamente en cuanto a la integración de datos de diferentes tipos y para experimentos que en lugar de centrarse en evolución en el tiempo se enfocan en poder detectar similitudes o diferencias entre compuestos, recientemente se ha propuesto un modelo para la agrupación y visualización de transcriptos y metabolitos en frutos de diferentes ILs de tomate [14]. Este modelo, denominado IL-SOM, se basa en la premisa de que genes y metabolitos que se comporten de forma similar pueden ser parte de redes de regulación comunes. Este principio se denomina "guilt-by-association" [13] y postula que un conjunto de genes involucrados en un proceso biológico están co-regulados –y por lo tanto co-expresados– bajo el control de una misma vía.

[1] Vía metabólica: colección de objetos (metabolitos, reacciones bioquímicas, enzimas o genes) y sus relaciones.

La motivación de este trabajo es estudiar con mayor profundidad, y de forma comparativa, el desempeño de los métodos de agrupamiento no-supervisados arriba mencionados y del modelo IL-SOM, para la tarea de integración y descubrimiento de relaciones en datos biológicos de distinto tipo. Para esto se utiliza en primer lugar un conjunto de medidas objetivas para cuantificar la calidad de los agrupamientos obtenidos por cada método (más allá de su significado biológico). Además, para verificar la consistencia de los agrupamientos desde el punto de vista biológico, se analiza en qué medida los diferentes transcriptos y metabolitos agrupados participan en vías metabólicas conocidas en los frutos de la especie domesticada de tomate (*Solanum lycopersicum*) a partir de alteraciones en el metabolismo producidas por introgresiones de alelos silvestres provenientes de la especie *Solanum pennellii*.

La organización de este trabajo es la siguiente. La Sección 2 describe las etapas de pre-procesamiento, selección, integración y análisis de datos biológicos a través del modelo IL-SOM. La Sección 3 presenta las medidas objetivas para la evaluación del desempeño de los métodos de agrupación. En la Sección 4 se presentan los resultados experimentales con la correspondiente discusión de las capacidades de agrupación y un análisis de la relevancia biológica. Finalmente las conclusiones se presentan en la Sección 5.

2. Procesamiento de datos y agrupamiento con IL-SOM

En el preprocesamiento de los perfiles metabólicos obtenidos para cada IL y sus réplicas de control, se detectan y se marcan metabolitos con menos de dos repeticiones válidas ($> 0,001$). Metabolitos marcados en todas las IL son eliminados. En cada IL ℓ de cada metabolito m, se calcula el logaritmo del cociente de las réplicas válidas ($logR_\ell^m$). Los niveles de transcripción se obtuvieron a partir de microarreglos de ADN. Los puntos del arreglo de baja calidad o sin señal fueron filtrados. Se detectaron aquellos que no mostraban expresión del gen en una réplica de una IL respecto del control, de acuerdo a un umbral de expresión típicamente usado en el área [2].

La siguiente etapa de procesamiento involucra el control de réplicas invertidas, en la cual se verifica que las réplicas de los genes seleccionados sean consistentes en cuanto a la sobre/sub expresión del gen en el experimento de una cierta IL contra el control. Se aplicó también el control de falsos positivos [15], que consiste en calcular la tasa esperada de predicciones falsas en relación al conjunto total de predicciones de cambios de un gen respecto al control (en cada IL). Los transcriptos t que pasaron las etapas anteriores fueron incluidos en el análisis usando el logaritmo del cociente del promedio de las réplicas válidas ($logR_\ell^t$). Los logaritmos de los cocientes resultantes de las etapas de preprocesamiento y selección fueron normalizados (en el caso de los transcriptos se aplicó la normalización LOWESS [2])y combinados antes de alimentar el modelo IL-SOM. Para cada patrón, la suma del cuadrado del logaritmo de los cocientes fue normalizada a 1.

Para la etapa de entrenamiento, los datos normalizados fueron organizados como una matriz conteniendo tantas filas como patrones y tantas columnas como ILs (dimensiones) se analizaran. Para encontrar todas las posibles relaciones entre los datos, el conjunto de entrenamiento incluyó también una copia de los patrones con su signo invertido. Esto permite ver al mismo tiempo, por cada IL, relaciones directas (genes sobre-expresados y metabolitos incrementados, genes sub-expresados y metabolitos disminuidos) e inversas (genes sub-expresados agrupados junto con metabolitos incrementados y genes sobre-expresados agrupados con metabolitos disminuídos) en los datos. Esta clase de análisis puede ser de ayuda para la inferencia de vías metabólicas desconocidas que involucren los datos agrupados. Antes de alimentar el IL-SOM, cada columna/dimensión-IL es normalizada en el rango $[0, 1]$ de acuerdo a una ecualización del histograma de los valores de los patrones. Denominaremos a estos patrones $\mathbf{R}^*$. Para más detalle véase [14].

Si bien se pueden utilizar diferentes topologías y estrategias de inicialización para el mapa, para el IL-SOM se han utilizado mapas de tipo grilla cuadrada $N = n \times n$, siendo N cantidad total de neuronas del mapa. Los pesos iniciales se obtienen mediante análisis de componentes principales, lo que permite que el resultado del proceso de aprendizaje sea reproducible y se vuelva independiente del orden en que se ingresan los patrones de entrenamiento. El método de aprendizaje usado es el algoritmo estándar de entrenamiento por lotes [2], con la distancia euclídea estándar y una función de vecindad de tipo gaussiana. Para la visualización del mapa de características resultante se colorean las neuronas de acuerdo al tipo

[2]Para más detalles: `http://www.cis.hut.fi/projects/somtoolbox/`.

de datos que agrupan, permitiendo una rápida identificación de tipos de datos combinados: negro para agrupamiento de metabolitos y transcriptos, azul para sólo metabolitos y rojo para sólo transcriptos. Se puede definir además una *vecindad de visualización* para la evaluación de las neuronas que integran los dos tipos de patrones mixtos (metabolito/transcripto). El uso de varios radios posibles en la vecindad de visualización de una neurona es útil para la identificación de grupos, permitiendo el análisis dinámico de su formación sin necesidad de volver a entrenar el IL-SOM. Si el set de datos incluye los datos originales y éstos mismos cambiados de signo, el mapa resultante mostrará una configuración "triangular" simétrica, en la cual las esquinas superior izquierda e inferior derecha agruparán exactamente los mismos datos pero con signos opuestos.

3. Medidas objetivas para la comparación de métodos de agrupamiento

Denominaremos nodo a cada uno de los k elementos que conforman la estructura del método de agrupación. Los métodos que se compararon son el agrupamiento jerárquico (HC_k, del inglés *hierarchical clustering*), k-medias y el modelo IL-SOM_k. En el caso del IL-SOM un nodo es equivalente a una neurona, en el caso de HC cada rama conforma un nodo y en el caso del método de k-medias los nodos corresponden a las k partes en que se dividen los datos. En los tres casos identificaremos con el índice j al nodo, con $\mathbf{w}_j$ a su centroide y con A_j al conjunto de patrones que quedaron agrupados en él. Se usará el término "nodo integrador" para hacer referencia a los nodos que contengan agrupamientos de patrones de diferente tipo (metabolitos/transcriptos). Para las comparaciones se definirán medidas objetivas y criterios de validación biológica. Las medidas objetivas miden la calidad de los agrupamientos encontrados con cada técnica, sin considerar su significado biológico. La calidad de los nodos encontrados se puede evaluar usando tres tipos de medidas: I) medidas de cohesión, II) medidas de separación y III) medidas combinadas [4].

Para medir la cohesión entre los patrones de cada nodo, se utilizó

$$\overline{C}_j = \frac{1}{|A_j|} \sum_{\forall \mathbf{R}^i \in A_j} \|\mathbf{R}^i - \mathbf{w}_j\|_2, \tag{1}$$

donde $|A_j|$ es el número de patrones en el nodo j. Como medida global de cohesión simplemente se utiliza el promedio sobre todos los nodos $\overline{C} = \frac{1}{k}\sum_j \overline{C}_j$. Valores de $\overline{C}$ cercanos a 0 indican nodos más compactos.

La separación entre nodos se evaluó midiendo los valores medios, mínimo y máximo de la distancia euclídea entre centroides, según

$$\overline{S} = \frac{2}{k^2 - k} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \|\mathbf{w}_i - \mathbf{w}_j\|_2, \tag{2}$$

$$S_m = \min_{0<i\neq j\leq k} \{\|\mathbf{w}_i - \mathbf{w}_j\|_2\}, \quad S_M = \max_{0<i\neq j\leq k} \{\|\mathbf{w}_i - \mathbf{w}_j\|_2\}. \tag{3}$$

donde $\overline{S}$ cercano a cero indica cercanía entre nodos.

Las primera de las medidas combinadas que se utilizaron fue el índice de Davies-Bouldin, definido como [18]

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j\neq i} \left(\frac{\overline{C}_i + \overline{C}_j}{\|\mathbf{w}_i - \mathbf{w}_j\|_2} \right). \tag{4}$$

Este índice es una medida del solapamiento entre nodos, por cual valores de DB cercanos a cero indican poco solapamiento. La otra medida combinada que se utilizó fue la tasa de dispersión intranodo, que se define como [11]

$$\Upsilon = 1 - \left(\frac{\sum_{j=1}^{k} \left\| \mathbf{w}_j - \left(\frac{1}{N} \sum_{\ell=1}^{N} \mathbf{R}^\ell \right) \right\|_2}{\sum_{i=1}^{N} \left\| \mathbf{R}^i - \left(\frac{1}{N} \sum_{\ell=1}^{N} \mathbf{R}^\ell \right) \right\|_2} \right), \tag{5}$$

donde N es el número total de patrones analizados. El numerador en (5) corresponde a la suma de las distancias entre los centroides y el centro del conjunto de los datos; el denominador es la suma de las distancias entre cada patrón y el centro del conjunto de datos. Cuanto menor sea el valor de Υ, menor será la dispersión intranodo.

4. Resultados experimentales y discusión

Los datos con los cuales se ha trabajado fueron perfiles metabólicos y transcripcionales de frutos de tomate cultivados en condiciones controladas de campo y cosechados en etapa de maduración. Las muestras de metabolitos se analizaron por cuadruplicado. En el caso de los transcriptos, se realizaron seis u ocho réplicas por cada medición en microarreglo. Luego de la etapa de preprocesamiento y selección explicadas en la Sección 2, quedaron seleccionados 71 metabolitos y 1385 transcriptos con niveles de detección y expresión respectivamente para 21 ILs.

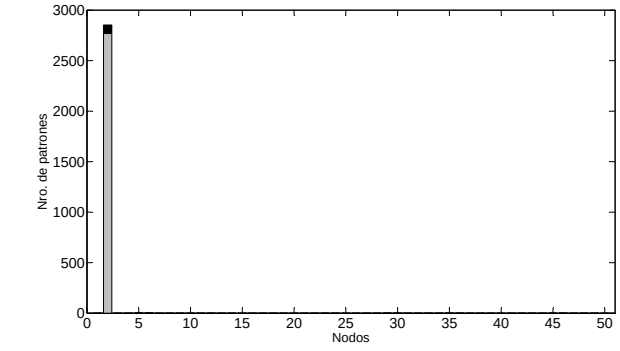
4.1. Análisis comparativo basado en medidas objetivas

La Figura 1 muestra los histogramas resultantes para cada método con $k = 50$. Para hacer una comparativa equivalente y dado que para el IL-SOM el hecho de tener cada patrón y su invertido genera un mapa simétrico como fue discutido en la Sección 2, el histograma para esta técnica fue generado únicamente con las neuronas de una de las mitades del mapa. Como se puede apreciar, el HC agrupa la gran mayoría de los patrones en una misma rama (Figura 1.a). Esto pone en gran desventaja a la técnica, tanto desde la perspectiva de sus capacidades como método de agrupación en este tipo de datos como desde la perspectiva de la información acerca de los procesos biológicos que se puedan inferir a partir de tal agrupación. Es importante destacar que independientemente de la profundidad en la que se corte la ramificación jerárquica, el método tiende a agrupar siempre la gran mayoría de los patrones en unos pocos nodos. Otra inconsistencia importante que se detectó en este caso es que, como se detallará más adelante, los patrones originales junto con su versión invertida han sido agrupados en muchos casos en el mismo nodo.

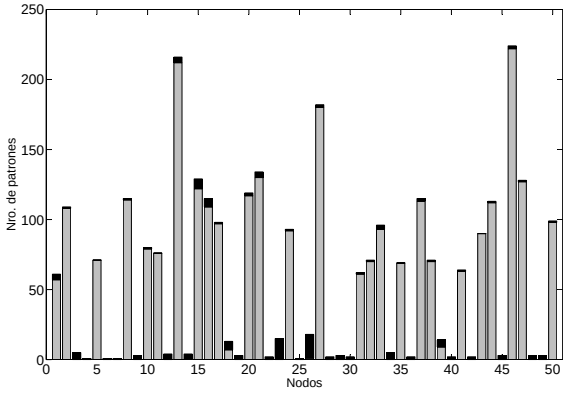
Al comparar los histogramas de k-medias (Figura 1.b) y IL-SOM (Figura 1.c) se puede observar que la distribución en el caso del IL-SOM es mucho más uniforme. Es decir, mientras k-medias posee varios nodos con muy pocos patrones y algunos pocos nodos con muchos patrones, la distribución de los patrones en el IL-SOM es más equilibrada a lo largo de los nodos. Esto se debe principalmente a la influencia de las vecindades durante el proceso de entrenamiento de un mapa auto-organizativo. Mientras para k-medias cada nodo se entrena independientemente de los demás, en el caso del IL-SOM se realiza una actualización de los nodos cercanos a la neurona ganadora, lo que permite que los centroides no se alejen tanto y los patrones puedan distribuirse más uniformemente en regiones de neuronas con centroides similares. La distribución más equilibrada de los patrones y la posibilidad de analizar a las neuronas individuales y extender el análisis a aquellas situadas en las cercanías para diferentes radios de vecindad, es una clara ventaja del IL-SOM en relación a k-medias.

La Tabla 1 muestra los resultados obtenidos en la comparación de los tres, para diferentes cantidades de nodos. El método de agrupamiento jerárquico concentró más del 85 % de los patrones en un mismo nodo, e incluyó en ellos datos directos e invertidos, por lo que no resultaría un método válido para detectar cambios coordinados en los patrones. Los nodos más compactos y con menor separación internodo ($\overline{C}$ y $\overline{S}$) fueron obtenidos con IL-SOM. En cuanto a la separación, se pudo observar que el HC y k-medias tienden a ubicar un centroide en cada uno de los patrones más distantes del conjunto de datos analizado. Los mapas auto-organizativos son más robustos, manteniendo las distancias entre los centroides de las neuronas gracias a la actualización por vecindades durante las primeras etapas de entrenamiento. Claramente, al aumentar la cantidad de neuronas en el mapa se amplía el grado de libertad y los centroides más externos del mapa pueden acercarse a los patrones más alejados del conjunto. Se puede notar también que la separación mínima se reduce en el IL-SOM en relación a los otros métodos, lo que permite recorrer el mapa con una mayor confianza en que los cambios entre nodos cercanos son graduales y pueden conformar agrupaciones de mayor interés biológico.

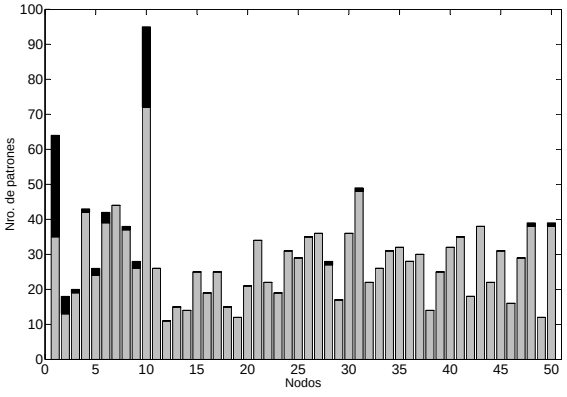
Con respecto a las medidas de tipo III, hay que considerar que la dispersión intranodo siempre es mayor para el IL-SOM, independientemente de la cantidad de nodos. Esto significa que la distancia



a)



b)



c)

Figura 1: Histogramas de distribución de patrones (gris para transcriptos, negro para metabolitos) a lo largo de los nodos de cada método de agrupación. a) agrupación jerárquica; b) k-medias; c) IL-SOM.

Tabla 1: Medidas de calidad para los diferentes métodos de agrupamiento estudiados. Subrayados se destacan los dos mejores valores para cada medida y cantidad de nodos.

Tipo	Medida	HC		k-medias		IL-SOM	
		50	200	50	200	50	200
I	$\overline{C}$	4.728	3.511	4.608	3.167	<u>3.433</u>	<u>3.294</u>
II	$\overline{S}$	21.51	13.57	12.29	9.577	<u>1.344</u>	<u>2.118</u>
II	S_m	8.635	4.159	1.892	1.369	<u>0.239</u>	<u>0.196</u>
II	S_M	45.92	45.92	45.92	45.92	<u>3.483</u>	<u>7.363</u>
III	DB	<u>2.280</u>	<u>2.901</u>	5.764	3.630	24.42	20.67
III	Υ	0.936	0.842	0.967	0.895	<u>0.995</u>	<u>0.971</u>

Tabla 2: Medidas de calidad para los diferentes métodos de agrupamiento estudiados, considerando solamente nodos integradores. Subrayados se destacan los dos mejores valores para cada medida y cantidad de nodos.

Tipo	Medida	HC		k-medias		IL-SOM	
		50	200	50	200	50	200
	N	1	13	26	35	13	21
I	$\overline{C}$	3.787	<u>3.356</u>	<u>3.638</u>	3.588	4.104	4.660
II	$\overline{S}$	–	7.493	4.214	6.591	<u>1.609</u>	<u>2.913</u>
II	S_m	–	4.372	1.892	2.733	<u>0.302</u>	<u>0.371</u>
II	S_M	–	12.51	11.02	15.14	<u>3.483</u>	<u>6.118</u>
III	DB	–	<u>2.516</u>	<u>2.472</u>	3.274	18.13	11.28
III	Υ	–	0.994	0.993	0.986	<u>0.998</u>	<u>0.995</u>

media entre los centroides del IL-SOM y el centro global de los patrones es menor (en forma relativa a la dispersión total de los patrones) que la distancia media de los centroides de los otros métodos. Dado que HC encuentra un gran nodo con la mayoría de los patrones y k-medias forma muchos nodos dispersos con unos pocos patrones lejanos, todas las distancias entre los nodos dispersos (y sus centroides) son grandes. Debido a esto, el índice DB es el menor para el caso de HC. Esto no pasa en IL-SOM porque las distancias entre centroides son siempre más pequeñas, dado que están mejor distribuidos y no se asocian centroides a patrones distantes y aislados, con lo que tienen más centroides para repartir y no se ve forzado a concentrar muchos patrones en pocos centroides. Además, dado que los patrones alejados (probablemente *outliers*) tienen que asociarse a algún centroide, las cohesiones también bajan y por lo tanto la medida de DB es la más alta.

La Tabla 2 muestra los resultados obtenidos en la comparación de los tres métodos, considerando en este caso únicamente nodos integradores. El interés de este análisis en particular radica en que los patrones agrupados en estos nodos podrían ser partes componentes de una misma vía metabólica. Como se puede observar, se ha agregado una fila más a la tabla que da cuenta del número de nodos integradores encontrados por cada técnica. Como es de esperar, al agregar mayor grado de libertad a las técnicas se encuentra mayor cantidad de este tipo de nodos. k-medias es el método que mayor cantidad de nodos encuentra y con mayor cohesión. Sin embargo, el detalle de las agrupaciones en esos nodos indica que se han agrupado patrones sin invertir que en algunos casos en su versión invertida no quedaron en un mismo grupo (lo que es incoherente). En el otro extremo, HC encuentra la menor cantidad de nodos, pero con los problemas anteriormente destacados en cuanto a la agrupación en un mismo nodo de patrones en su versión directa e invertida, y su correspondiente invalidez biológica. El IL-SOM en cambio siempre encuentra nodos que agrupan coherentemente los datos. En cuanto a las medidas de tipo II, el SOM obtiene los mejores índices en cuanto a la separación mínima, máxima y media de los nodos integradores encontrados. La técnica de HC con $k = 50$ produjo un único nodo integrador con el 98 % de los datos. Las medidas de tipo II no pueden calcularse para este caso ya que no hay separación entre nodos (hay un único nodo integrador). En las medidas combinadas, la dispersión internodo y el índice DB tampoco pueden calcularse para este caso ya que no tiene sentido medir el solapamiento de un único nodo integrador. La dispersión intranodo es la mejor para el IL-SOM, aunque es muy cercana a 1.0 en todos los otros métodos

también. Por los mismos motivos que se analizaron anteriormente, en el IL-SOM se sigue observando el mayor índice de solapamiento de agrupaciones y la menor separación entre nodos. El índice DB privilegia agrupamientos compactos y bien separados entre sí, lo cual, como ya se ha dicho, no es que caracteriza a los mapas auto-organizativos. Adicionalmente, desde el punto de vista biológico, no sería útil tener agrupaciones con un índice DB alto, porque hay patrones que deben estar cerca de muchos otros patrones, si pensamos que las agrupaciones reflejan componentes de vías metabólicas comunes y hay patrones que pueden participar en varias vías al mismo tiempo.

4.2. Agrupaciones y vías metabólicas

Para poder evaluar la significancia de las agrupaciones descriptas previamente desde el punto de vista de la aplicación biológica de los resultados encontrados se propone verificar la pertenecia a alguna vía de regulación conocida de los agrupamientos encontrados. Se analizaron los nodos integradores formados por los tres métodos para $k = 50$ buscando metabolitos y transcriptos involucrados en rutas metabólicas conocidas. En este análisis se consideraron vías metabólicas básicas[3] relacionadas con la producción de energía (glucólisis y ciclo de Krebs) y algunas reacciones asociadas, debido a su importancia en la subsistencia de todos los organismos y a la gran cantidad de información disponible sobre ellas. Además, excepto en algunos casos puntuales, la elección de procesos biológicos comunes a la gran mayoría de los organismos es un punto de partida importante para la comparación, ya que cualquier método de agrupamiento que se utilice para analizar estos datos debería poder encontrar relaciones tan básicas. Esta comparación se realizó en base a la búsqueda de los metabolitos y transcriptos agrupados en los nodos integradores, que a su vez estuvieran relacionados en las vías metabólicas, evaluando la cantidad de agrupaciones válidas encontradas en cada caso. Para la comparación se descartó diretamente el método de agrupamiento jerárquico dado que agrupa la gran mayoría de los compuestos en un único nodo, incluso con inconsistencias importantes, como se detalló más arriba.

En la Figura 2 se muestra un esquema simplificado de las vías metabólicas usadas. En el caso de los transcriptos, se usaron los códigos "EC" correspondientes a la nomenclatura estándar para enzimas. Los metabolitos que estaban presentes en los datos de entrenamiento han sido destacados con un círculo. Los restantes compuestos (en cursiva) no se tendrán en cuenta en el presente análisis dado que no han sido medidos. En esta figura se ha destacado el número de nodo integrador en el que cada compuesto fue agrupado, distinguiendo a la derecha el nodo correspondiente al método IL-SOM y a la izquierda el correspondiente a k-medias. En el caso de enzimas que son codificadas por más de un gen, se indican los nodos en los cuales quedó agrupado cada gen. Para simplificar la notación en el análisis a continuación se presentan entre corchetes $[\cdots]$ los compuestos que fueron agrupados en un mismo nodo integrador. El método k-medias encontró relaciones coherentes pero más dispersas a lo largo de diferentes nodos integradores. Por ejemplo, se pueden observar [β-D-glucosa y D-fructosa-6-fosfato], [succinato y fumarato], [glicina, L-serina y 4-aminobutirato], [EC 4.2.1.2 y 1 gen de EC 4.1.1.31], [malato y 1 gen de EC 1.1.1.1] y [L-ascorbato y EC 1.1.1.29]. Sin embargo, las asociaciones no siempre reflejaron las relaciones opuestas, tales como en el caso de [maltosa y D-glucosa], [L-glutamato y EC 1.1.1.29], [fumarato y EC 4.2.1.2] entre otros, donde para una configuración de signos se agruparon en el mismo nodo y para la configuración opuesta lo hicieron en nodos diferentes. Inclusive, en el caso de [sacarosa y 1 gen de EC 1.1.1.1], quedaron agrupados cuando uno de los dos aparece con su signo invertido y no cuando ambos tienen sus valores directos. Estas inconsistencias, si bien no tan significativas como las encontradas en el agrupamiento jerárquico, limitan al método y arrojan dudas en cuanto a su aplicabilidad en la búsqueda de relaciones en este tipo de datos.

Si bien el IL-SOM generó la mitad de nodos integradores que k-medias (como se pudo ver en la Tabla 2), a diferencia de éste la asociación de patrones con un dado signo en un nodo particular se mantuvo de forma consistente para el mismo conjunto de datos con su signo invertido y las agrupaciones claramente relacionaron más compuestos de la misma vía en menos nodos integradores. Este fue el caso de [maltosa, D-glucosa, fructosa, D-fructosa-6-fosfato, L-alanina, glicina, 3-fosfoglicerato, EC 1.1.1.27, EC 4.2.1.2 y 1 gen de EC 4.1.1.31] y de [citrato, L-glutamato, succinato, malato y sacarosa]. El modelo IL-SOM permite además analizar relaciones con distintos radios de vecindad en el mapa [14], lo que ofrece un nivel más de análisis en relación a los otros métodos. Si se consideran los primeros vecinos de cada neurona (es decir,

[3]LycoCyc: `http://solcyc.sgn.cornell.edu/LYCO/server.html`.

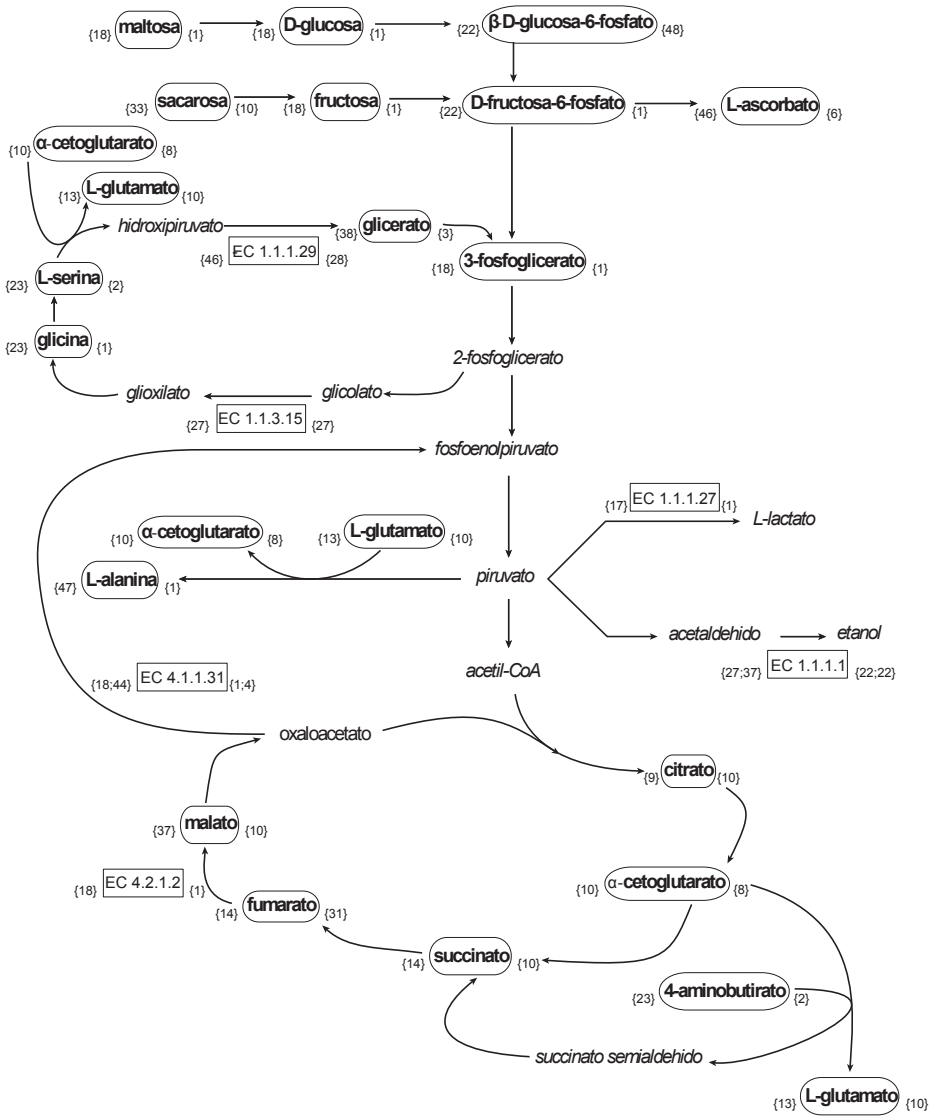


Figura 2: Esquema simplificado de la glucólisis, ciclo de Krebs y reacciones asociadas.

radio de vecindad 1), se pueden encontrar otras relaciones de interés como [glicina, L-serina, glicerato, 4-aminobutirato y EC 4.1.1.31], [EC 1.1.1.29 y EC 1.1.3.15] y [fumarato y 2 genes de EC 1.1.1.1]. En el primero de estos grupos se pueden encontrar compuestos que habían sido agrupados por k-medias pero no por IL-SOM con radio de vecindad 0. Adicionalmente, los dos genes que codifican para la enzima EC 1.1.1.1 también quedaron agrupados en el mismo nodo.

5. Conclusiones

En este trabajo se ha presentado una comparación entre métodos no supervisados para agrupamiento de datos biológicos, en particular, metabolitos y transcriptos de frutos de tomate. Se compararon los métodos de agrupamiento jerárquico, k-medias y el modelo IL-SOM basado en un mapa auto-organizativo para la integración de datos genómicos y metabólicos de múltiples líneas de introgresión. Se definieron medidas objetivas para el análisis de la calidad de los agrupamientos y se propuso una forma de medir su significancia biológica, la cual fue abordada desde la perspectiva de la utilidad de las agrupaciones para identificar aquellos patrones que cambian coordinadamente y por lo tanto pertenecen a vías comunes de regulación metabólica. El agrupamiento jerárquico concentró la mayoría de los patrones en un mismo nodo, y en varios casos incluyó en ellos al patrón original con su copia de signo invertido, por lo que no resultaría un método válido para detectar cambios coordinados en metabolitos y transcriptos. El método de k-medias, si bien fue el que encontró mayor cantidad de nodos agrupando datos de ambos tipos, el detalle de los patrones encontrados indica que se han agrupado patrones con signo directo en algunos casos y en el caso inverso no se han agrupado coherentemente esos mismos patrones, lo cual representa una clara limitación del método para su aplicabilidad en la búsqueda de relaciones en este tipo de datos.

En cambio, el modelo IL-SOM ha mostrado altas tasas de desempeño en la mayoría de las medidas objetivas de calidad, además de la máxima coherencia desde el punto de vista del significado biológico de los agrupamientos entre metabolitos y transcriptos obtenidos. Una de sus ventajas principales radicó en la posibilidad de usar el radio de vecindad para localizar otros patrones vecinos a los agrupados y que también pudiesen pertenecer a la vía regulatoria bajo estudio. En conjunto, estos resultados permiten predecir la consistencia del método IL-SOM para el análisis de agrupamientos de metabolitos y transcriptos.

Referencias

[1] Fernando Carrari and et al. Integrated analysis of metabolite and transcript levels reveals the metabolic shifts that underlie tomato fruit development and highlight regulatory aspects of metabolic network behavior. *Plant Physiol.*, 142:1380–1396, 2006.

[2] H.C. Causton, J. Quackenbush, and A. Brazma. *Microarray Gene Expression Data Analysis: A Beginner's Guide*. Blackwell Publishers, 2003.

[3] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 2003.

[4] Julia Handl, Joshua Knowles, and Dourglas B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[5] Masami Hirai, Kenjiro Sugiyama, Yuji Sawada, Takayuki Tohge, Takeshi Obayashi, Akane Suzuki, Ryoichi Araki, Nozomu Sakurai, Hideyuki Suzuki, Koh Aoki, Hideki Goda, Osamu Ishizaki Nishizawa, Daisuke Shibata, and Kazuki Saito. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, 101:10205–10, 2004.

[6] Alexander Kaever, Thomas Lingner, Kirstin Feussner, Cornelia Gbel, Ivo Feussner, and Peter Meinicke. MarVis: a tool for clustering and visualization of metabolic biomarkers. *BMC Bioinformatics*, 10:92+, March 2009.

[7] Arpad Kelemen, Ajith Abraham, and Yuehui Chen. *Computational Intelligence in Bioinformatics*. Springer, 2008.

[8] T. Kohonen, M. R. Schroeder, and T. S. Huang. *Self-Organizing Maps*. Springer-Verlag New York, Inc., 2001.

[9] Vincent Lacroix and et al. An introduction to metabolic networks and their structural analysis. *IEEE Transactions on computational biology and bioinformatics*, 5(4):594–617, 2008.

[10] John C. Lindon, Jeremy K. Nicholson, and Elaine Holmes, editors. *The Handbook of Metabonomics and Metabolomics*. Elsevier, 2007.

[11] Sueli A. Mingoti and Joab O. Lima. Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174(3):1742–1759, November 2006.

[12] Andrzej Polanski and Marek Kimmel. *Bioinformatics*. Springer-Verlag, NY, 2007.

[13] Kazuki Saito, M.Y. Hirai, and K. Yonekura-Sakakibara. Decoding genes with coexpression networks and metabolomics - majority report by precogs. *Trends in Plant Science*, 13:36–43, 2008.

[14] Georgina Stegmayer, Diego Milone, Laura Kamenetzky, Mariana Lopez, and Fernando Carrari. Neural network model for integration and visualization of introgressed genome and metabolite data. In *International Joint Conference on Neural Networks*, pages 2983–2989, 2009.

[15] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[16] D.K. Tasoulis, V.P. Plagianakos, and M.N. Vrahatis. *Computational Intelligence in Bioinformatics*, volume 94 of *Studies in Computational Intelligence*. Springer, 2008.

[17] T. Tokimatsu, N. Sakurai, H. Suzuki, H. Ohta, K. Nishitani, T. Koyama, T. Umezawa, N. Misawa, K. Saito, and D. Shibata. KaPPA-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiology*, 138(3):1289–1300, 2005.

[18] Rui Xu and II Donald C. Wunsch. *Clustering*. Wiley and IEEE Press, 2009.

[19] Mitsuru Yano, Shigehiko Kanaya, Md. Altaf-UI-Amin, Ken Kurokawa, Masami Yokota Hirai, and Kazuki Saito. Integrated data mining of transcriptome and metabolome based on bl-som. *Journal of Computer Aided Chemistry*, 7:125–136, 2006.