LAYOUT DE GRAFOS INTERACTIVOS PARA MATRICES DE EXPRESIÓN GÉNICA DE GRAN VOLUMEN

Raquel Guardia Villalba

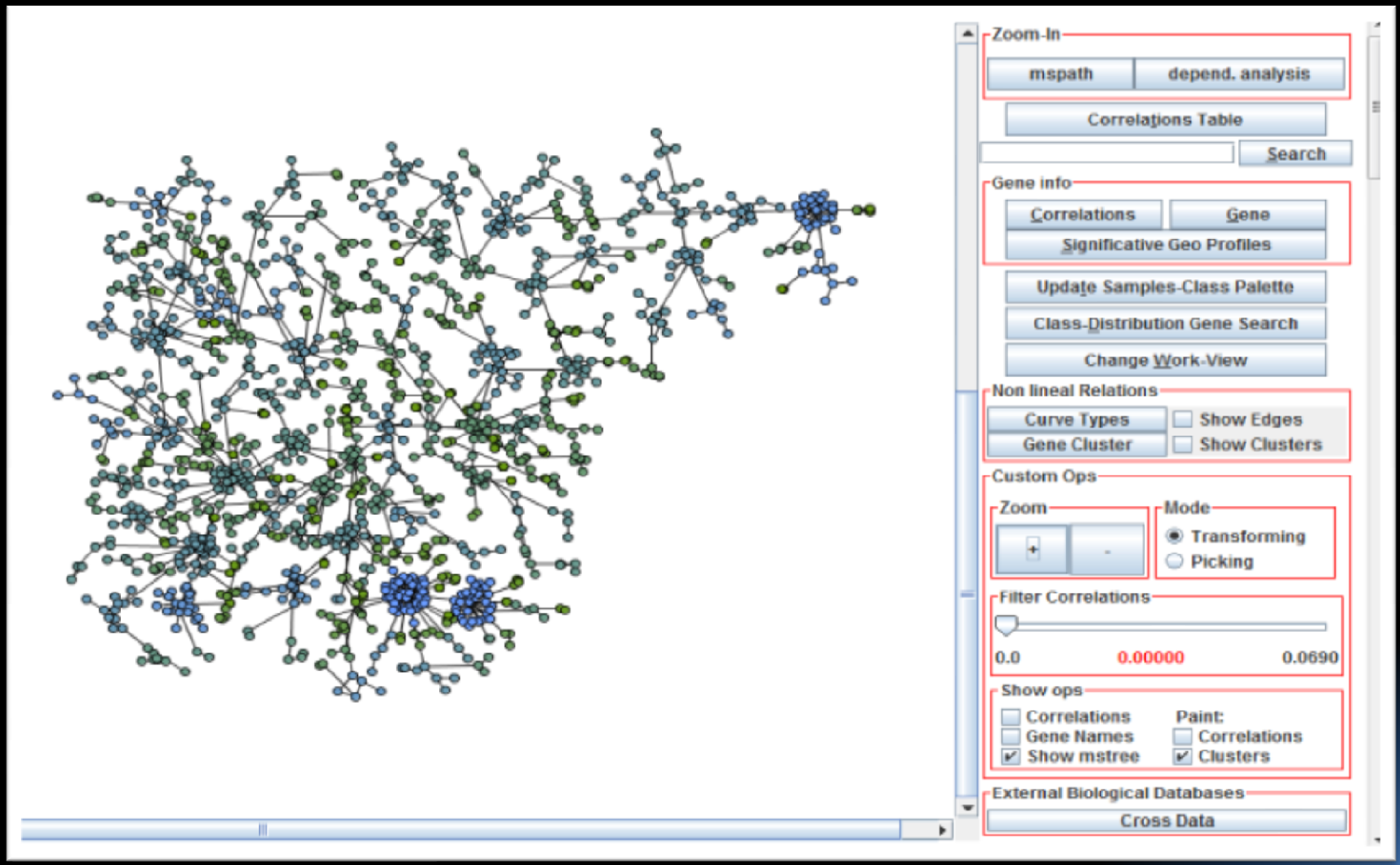
Índice de contenidos

1. Introducción

Los genes al expresarse, sintetizan las diferentes proteínas las cuales son encargadas de llevar a cabo las diferentes funciones de la célula. De esta forma, cuando los genes se expresan determinan el estado celular y modificando su expresión, provocan un cambio en la célula que puede llevar de un estado sano a uno patológico o viceversa.

ANÁLISIS DE MICROARRAYS

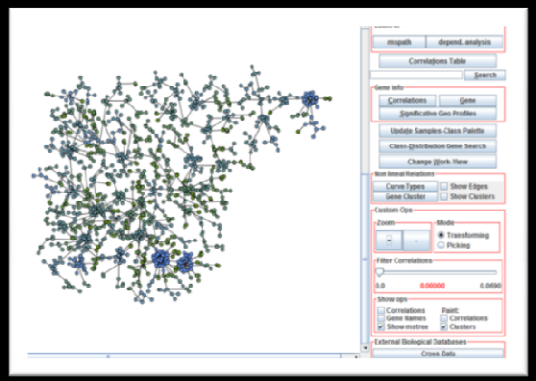
1. Introducción



1. Introducción

PCOPGene-Net es una aplicación web creada por el IBB pensada para facilitar el estudio de las relaciones entre las expresiones génicas bajo las condiciones de las microarrays que se analicen.

Problema: Solo opera con microarrays pequeñas.



Urge encontrar la forma de visualizar y trabajar con grafos interactivos de gran magnitud.

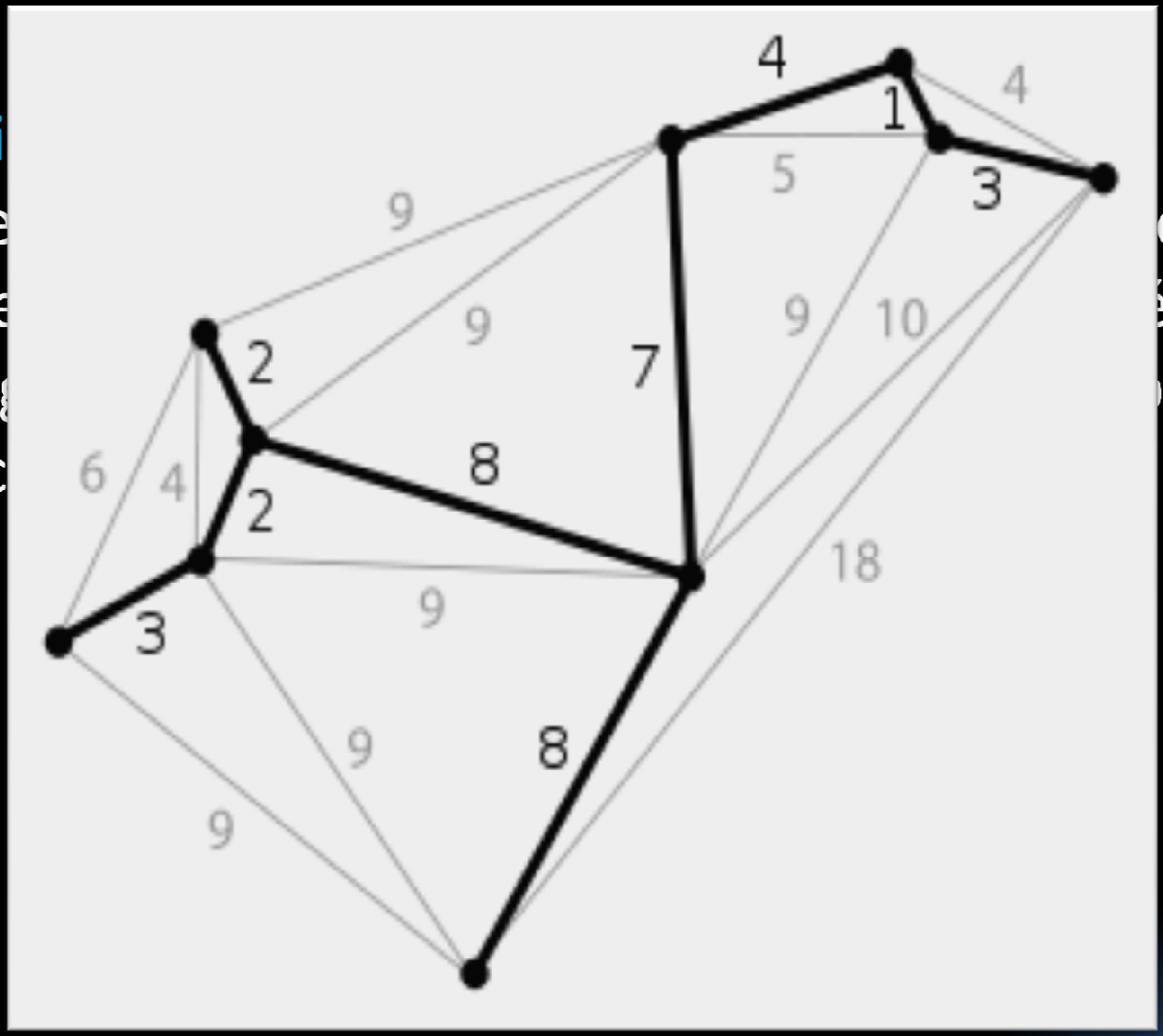
2. Fundamentos teóricos

- Microarrays: Matrices de genes frente a diversas condiciones muestrales. Cada uno de los valores de la matriz representa el nivel de expresión de un determinado gen bajo una cierta condición muestral.

- Clustering: Su objetivo es reducir la gran cantidad de datos caracterizándolos en grupos (clusters) más pequeños de individuos similares.

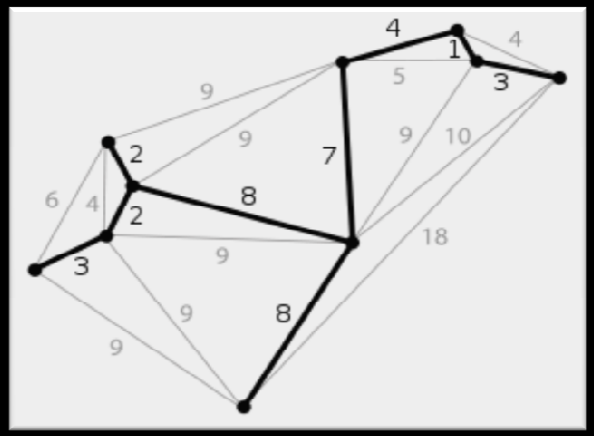
2. Fundamentos teóricos

- Mini grafo
  cone o que
  tiene rtices
  del y se
  busc



2. Fundamentos teóricos

- Minimum Spanning Tree (MST): Dado un grafo conexo, un MST de ese grafo es un subgrafo que tiene que ser un árbol y contener todos los vértices del grafo inicial. Cada arista tiene un peso y se busca que la suma de éstos sea mínima.



3 Objetivos

El aplicativo web abrirá simultáneamente los diferentes applets que muestran las particiones que conforman el total de genes de la microarray analizada.

las diferentes particiones de la microarray.

Objetivos

- Conseguir la máxima funcionalidad, entendibilidad y operatividad para todo tipo de microarrays.
- Diseño de nuevas fórmulas para cribar las relaciones de expresión no lineales entre genes.
- Diseño e implementación de un algoritmo para la división en clusters, hyperclusters y hyperclusters de segundo orden.
- Adaptación del layout.
- Partición de los datos necesarios para el applet.
- Diseño e implementación de un algoritmo para un último filtrado de relaciones de expresión no lineales por tipología.

Objetivos

- Tratamiento diferenciado de las microarrays pequeñas y de gran tamaño.
- Para microarrays de gran tamaño, trabajo con particiones de la microarray.
- Coordinación con las aplicaciones externas al applet y coordinación entre los distintos applets que contienen las diferentes particiones de la microarray.

Objetivos

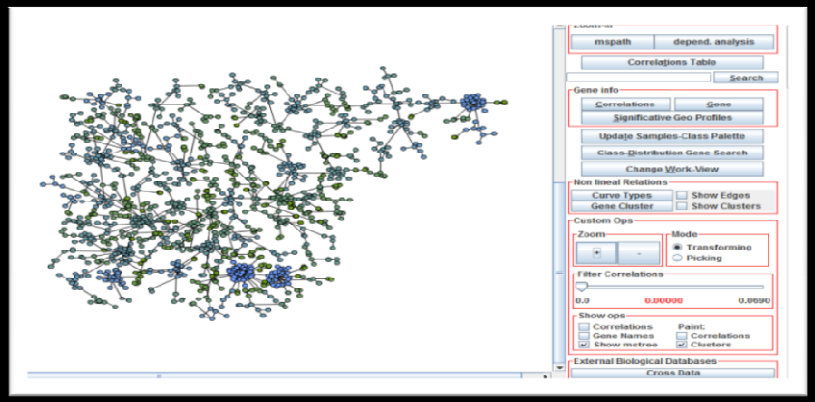
El aplicativo web abrirá simultáneamente los diferentes applets que muestran las particiones que conforman el total de genes de la microarray analizada.

4. Fases y resultados

1. Conocimientos previos en el ámbito de la bioinformática y del proyecto.
2. Mejora del preproceso para analizar los datos de microarrays pequeñas.
3. Tratamiento de microarrays de gran tamaño.
4. Adaptación del applet .
5. Filtrado de relaciones de expresión no lineales.
6. Adaptación del aplicativo web.

3.1 Conocimientos previos en el ámbito de la bioinformática y del proyecto

- Adquirir conocimientos sobre la bioinformática.
- Familiarizarme con el aplicativo PCOPGene.



- Familiarizarme con el preproceso para analizar los datos de microarrays pequeñas.

Correlaciones entre todos los genes y relaciones de expresión no lineales

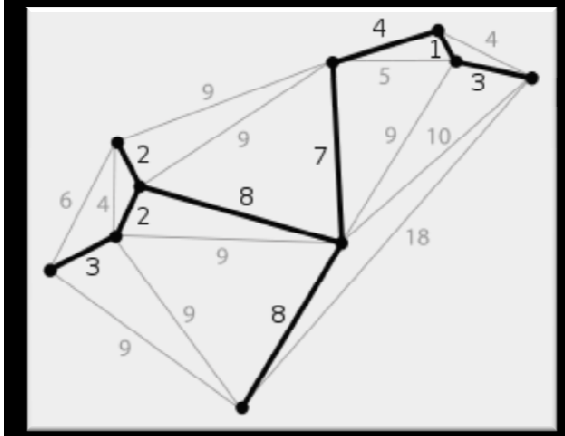
1 2 0.012246
1 3 0.019694
1 4 0.005982
⋮
1 n 0.059434
2 3 0.035903
2 4 0.036961
⋮
n-1 n 0.25541

Búsqueda de los genes mejor correlacionados

Búsqueda del gen mejor correlacionado con cada gen

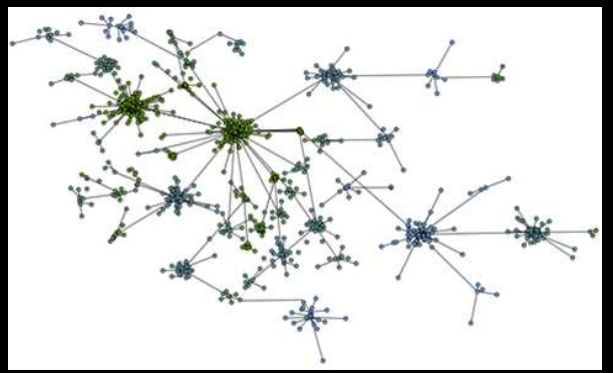
1 934 0.012246
2 3 0.019694
3 2 0.019694
4 1388 0.019932
5 4 0.036034
⋮
1414 588 0.059434
1415 962 0.035903

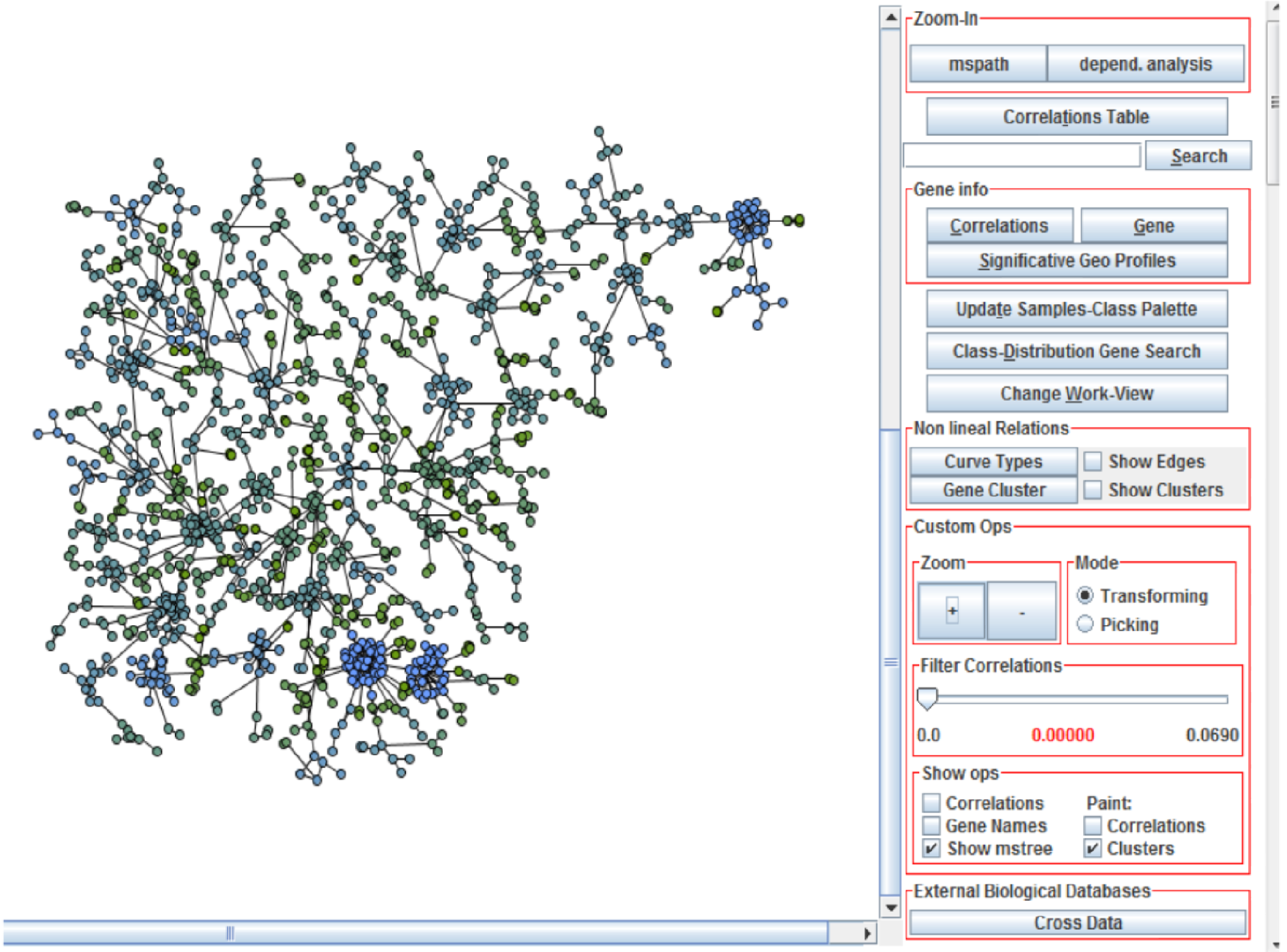
MST



Clustering

Layout




Zoom-In
mspath
depend. analysis
Correlations Table
Search
Gene info
Correlations
Gene
Significative Geo Profiles
Update Samples-Class Palette
Class-Distribution Gene Search
Change Work-View
Non lineal Relations
Curve Types
Show Edges
Gene Cluster
Show Clusters
Custom Ops
Zoom
+
-
Mode
Transforming
Picking
Filter Correlations
0.0
0.00000
0.0690
Show ops
Correlations
Gene Names
Show mstree
Paint:
Correlations
Clusters
External Biological Databases
Cross Data

1. Cálculo de las correlaciones entre genes y determinación de las relaciones de expresión no lineales entre genes.
2. Búsqueda de los genes mejor correlacionados
3. Búsqueda del gen mejor correlacionado con cada gen
4. Cálculo del mínimum spanning tree entre los genes de la microarray
5. Proceso de clustering de genes por la correlación entre sus expresiones
6. Cálculo del Layout de la microarray

4.2 Optimización del preproceso

1. Optimización del cálculo de correlaciones entre genes
2. Adaptaciones en la búsqueda de los genes mejor correlacionados
3. Mejoras en la búsqueda del gen mejor correlacionado con cada gen
4. Adaptaciones en el cálculo del MST
5. Proceso de clustering
6. Optimizaciones en el cálculo del layout

4.2 Optimización del preproceso

1. Optimización del cálculo de correlaciones entre genes

En caso que a la microarray le faltase la respuesta de algún gen a la última condición muestral, éste proceso omitía dicho gen y el siguiente y reenumeraba los genes restantes.

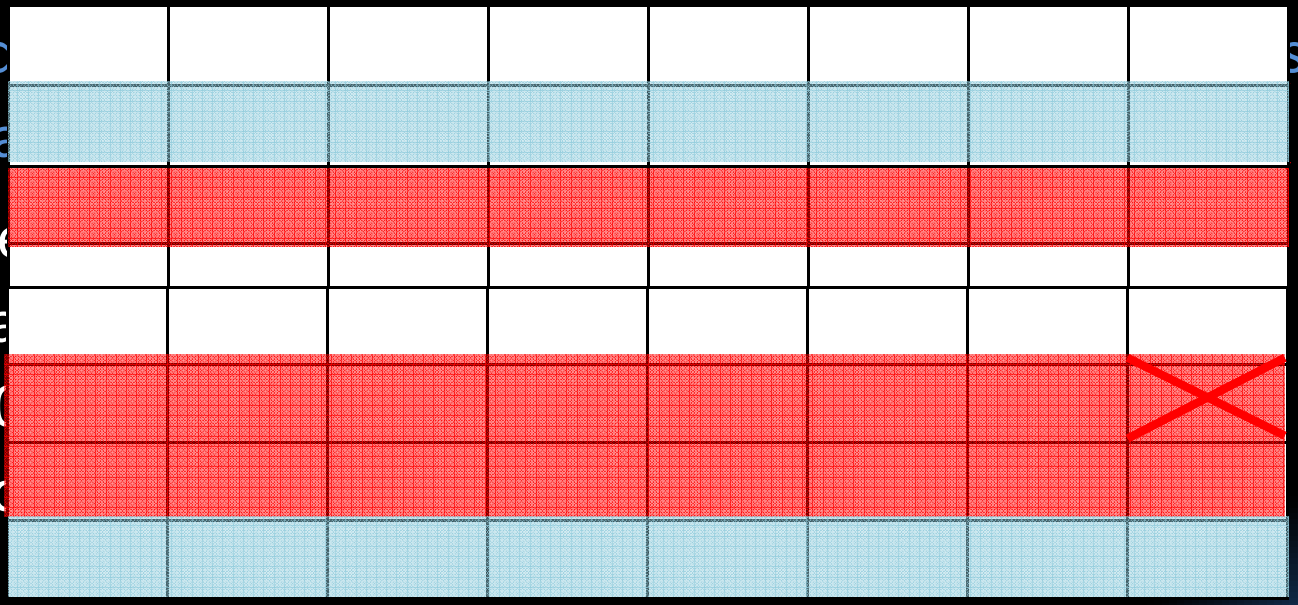
2. Ac s mejor

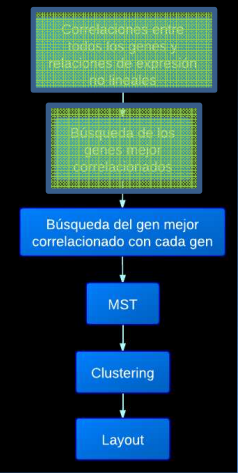
En este

microa

los 50

unto





1. Optimización del cálculo de correlaciones entre genes

En caso que a la microarray le faltase la respuesta de algún gen a la última condición muestral, éste proceso omitía dicho gen y el siguiente y reenumeraba los genes restantes.

2. Adaptaciones en la búsqueda de los genes mejor correlacionados

En este proceso se crea un fichero para cada gen de la microarray en el que figuran ordenados por correlación los 500 genes mejor correlacionados con el primero junto con las correlaciones que mantienen.

4.2 Optimización del preproceso

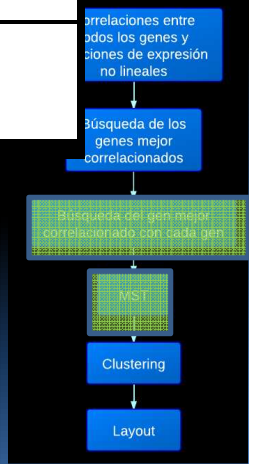
3. Mejoras en la búsqueda del gen mejor correlacionado con cada gen

El p… …ía tar… …na mi…

Número de genes	Proceso anterior	Nuevo proceso
2.998	5.971 s	2 s
5.000	28.856 s	6 s
7.702	101.248 s = 1,17 días	19 s
14.012	889.056 s = 10,29 días	47 s

4. A… en

Este proceso es el encargado de crear el minimun spanning tree (MST).



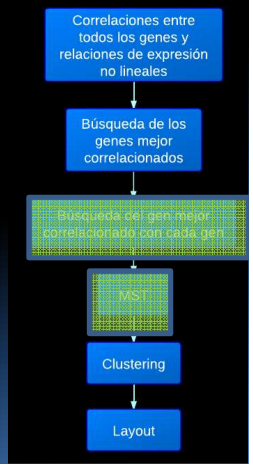
3.2 Optimización del preproceso

3. Mejoras en la búsqueda del gen mejor correlacionado con cada gen

El problema de este proceso es que estaba mal diseñado; podía tardar varios minutos en obtener los resultados para una microarray de 1.400 genes.

4. Adaptaciones en el cálculo del mínimum spanning tree entre los genes de la microarray

Este proceso es el encargado de crear el minimun spanning tree (MST).



4.2 Optimización del preproceso

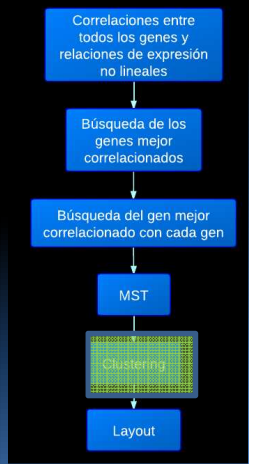
5. Proceso de clustering de genes por la correlación entre sus expresiones

Para hallar los clusters de genes se siguen estos pasos:

1. Obtener una tabla en la que figuren todos los genes junto con el gen con el que mantienen una mayor correlación.

2. Recorrer la tabla anterior y estudiar en cada caso el gen asociado.
   - 2.1 Si el gen asociado se encuentra ya en un cluster se añade el gen inicial al mismo cluster.
   - 2.2 En caso contrario se crea un nuevo cluster con los 2 genes.

3. Tanto en el caso 2.1 como en el 2.2 es necesario mirar si el gen inicial se encuentra ya en un cluster y, en este caso, si se encuentra en el mismo cluster que el gen asociado. En caso contrario los dos clusters serán fusionados.



Proceso de clustering

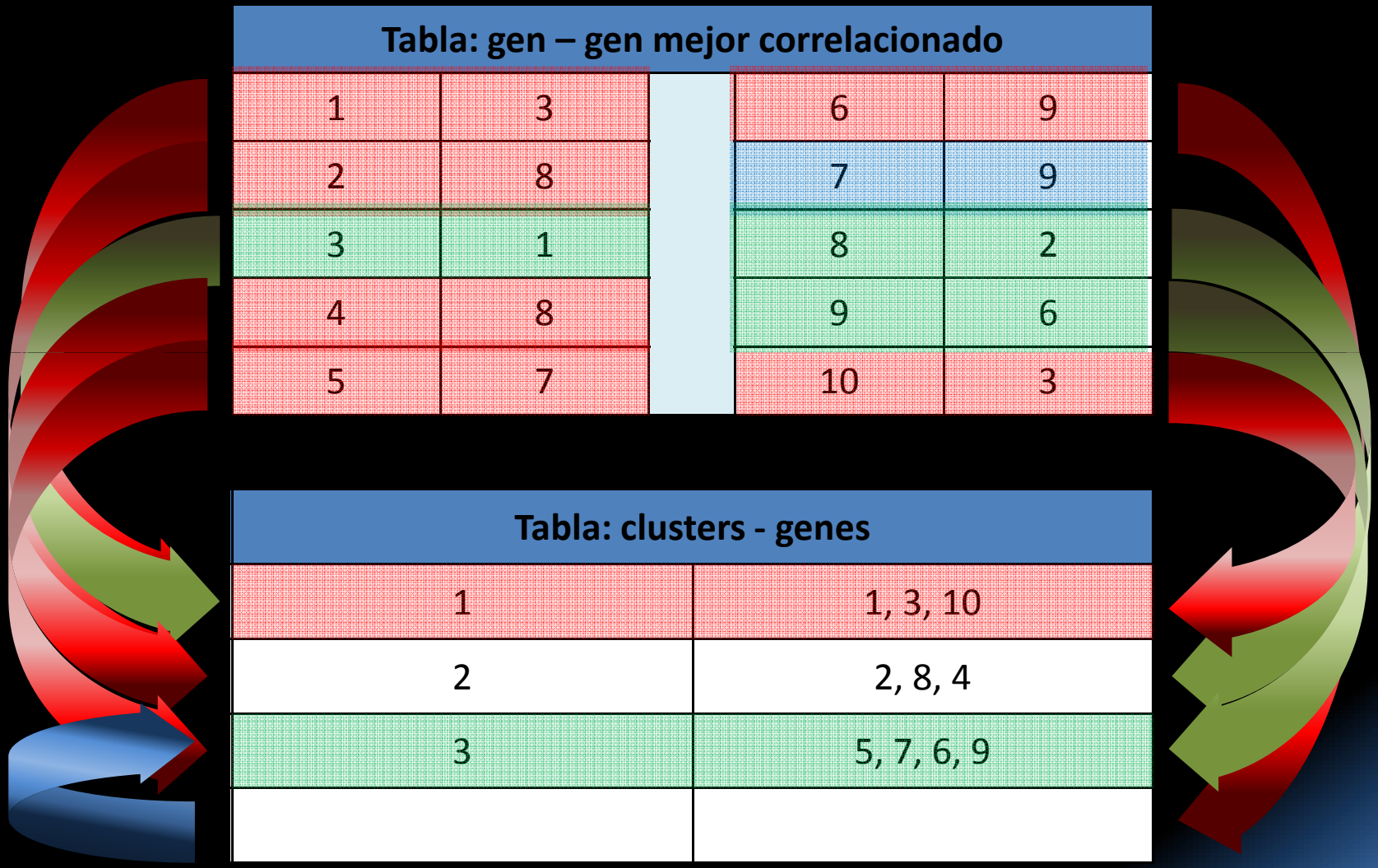


Tabla: gen – gen mejor correlacionado			
1	3	6	9
2	8	7	9
3	1	8	2
4	8	9	6
5	7	10	3

Tabla: clusters - genes	
1	1, 3, 10
2	2, 8, 4
3	5, 7, 6, 9

Tabla: clusters - genes	
1	1, 3
2	2, 8

Tabla: clusters - genes	
1	1, 3, 10
2	2, 8, 4
3	5, 7, 6, 9

4.2 Optimización del preproceso

6. Optimizaciones en el cálculo del layout

El programa que realiza el layout tiene como objetivo generar las coordenadas de cada gen en función de la correlación entre los genes de la microarray.

- Layout Local
- Layout Global

Problemas:

- Existencia de casos que conducían a error.

4.3 Tratamiento de microarrays de gran tamaño

1. Comprobación del grado de correlación entre los genes
2. Proceso de clustering de genes por la correlación entre sus expresiones
3. Proceso de partición de la microarray
4. Separación de los ficheros que necesita el applet para las diversas particiones
5. Generación del layout para cada partición concreta

4.3 Tratamiento de microarrays de gran tamaño

1. Comprobación del grado de correlación entre los genes

Problema: Existencia de correlaciones menores a $1 \cdot 10^{-6}$.

ERRORES

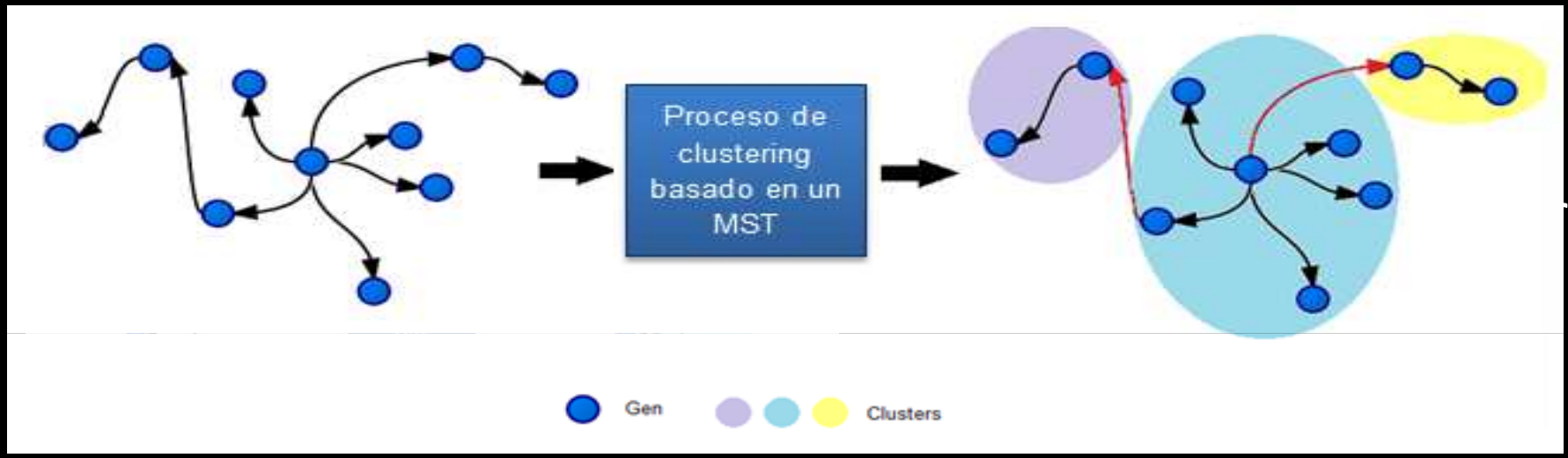
Solución: Detección y Corrección.

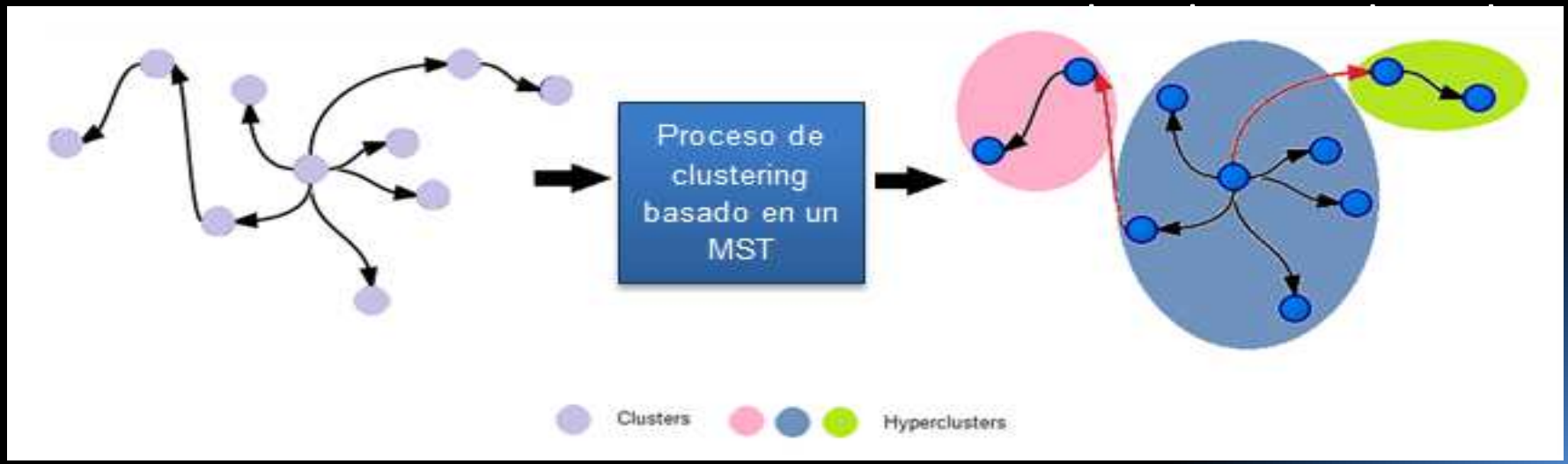
Detección: Mc

x – 1 – 0.000077
x – 2 – 0.000035
x – 3 – 0.000001
x – 4 – 0.000063
x – 5 – 0.000001

por 0.000001

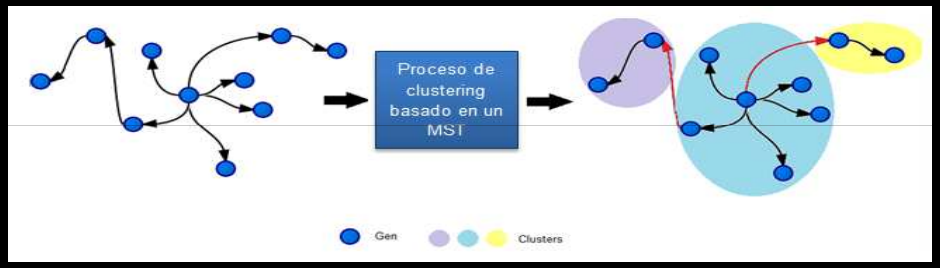
4.3 Tratamiento de microarrays de gran tamaño

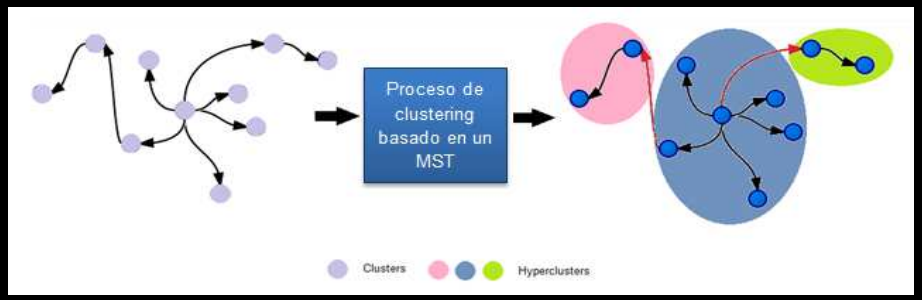




4.3 Tratamiento de microarrays de gran tamaño

2. Proceso de clustering de genes





Objetivo: Encontrar los clusters de nivel n que formarán las particiones

- Busca los clusters de todos los niveles necesarios
- Informa de la cantidad de clusters de cada nivel

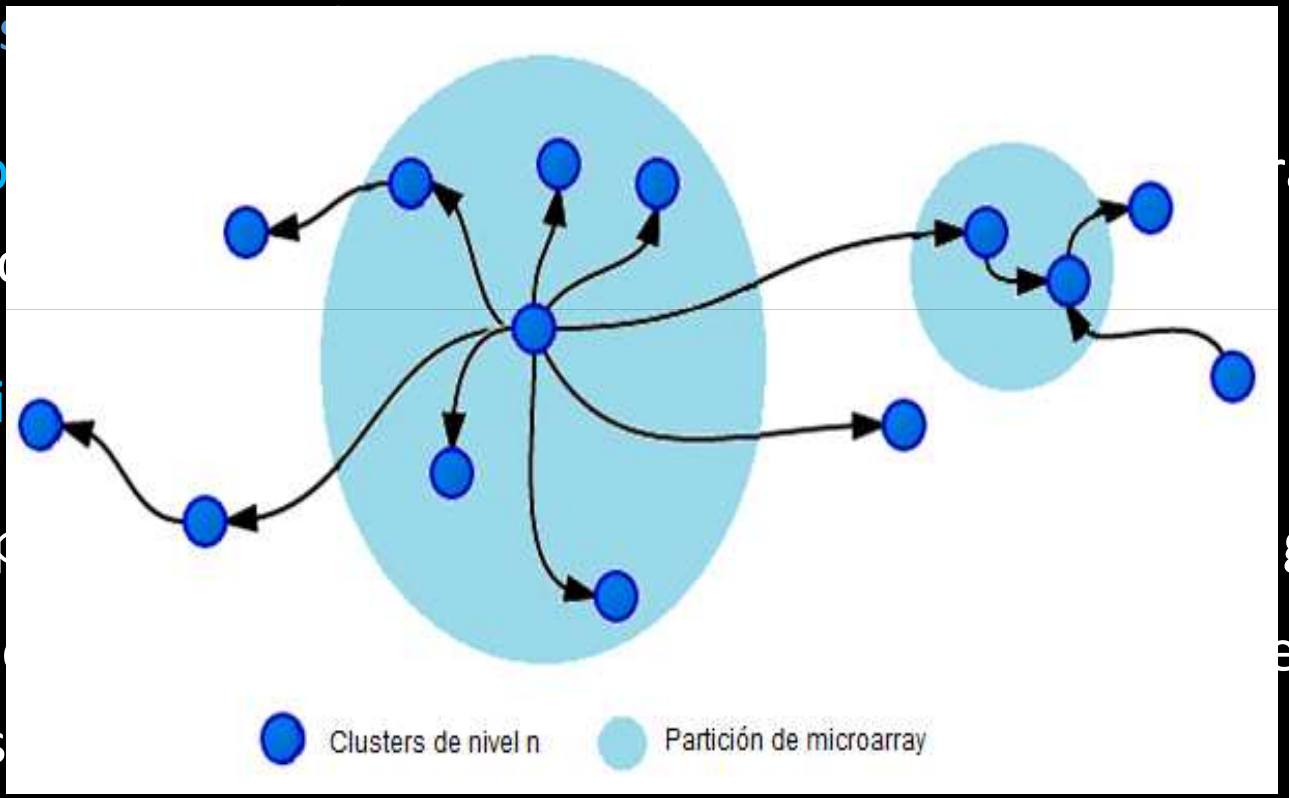
4.3 Tratamiento de microarrays de gran tamaño

3. Proces

Objetivo ra crear las parti

Restricci

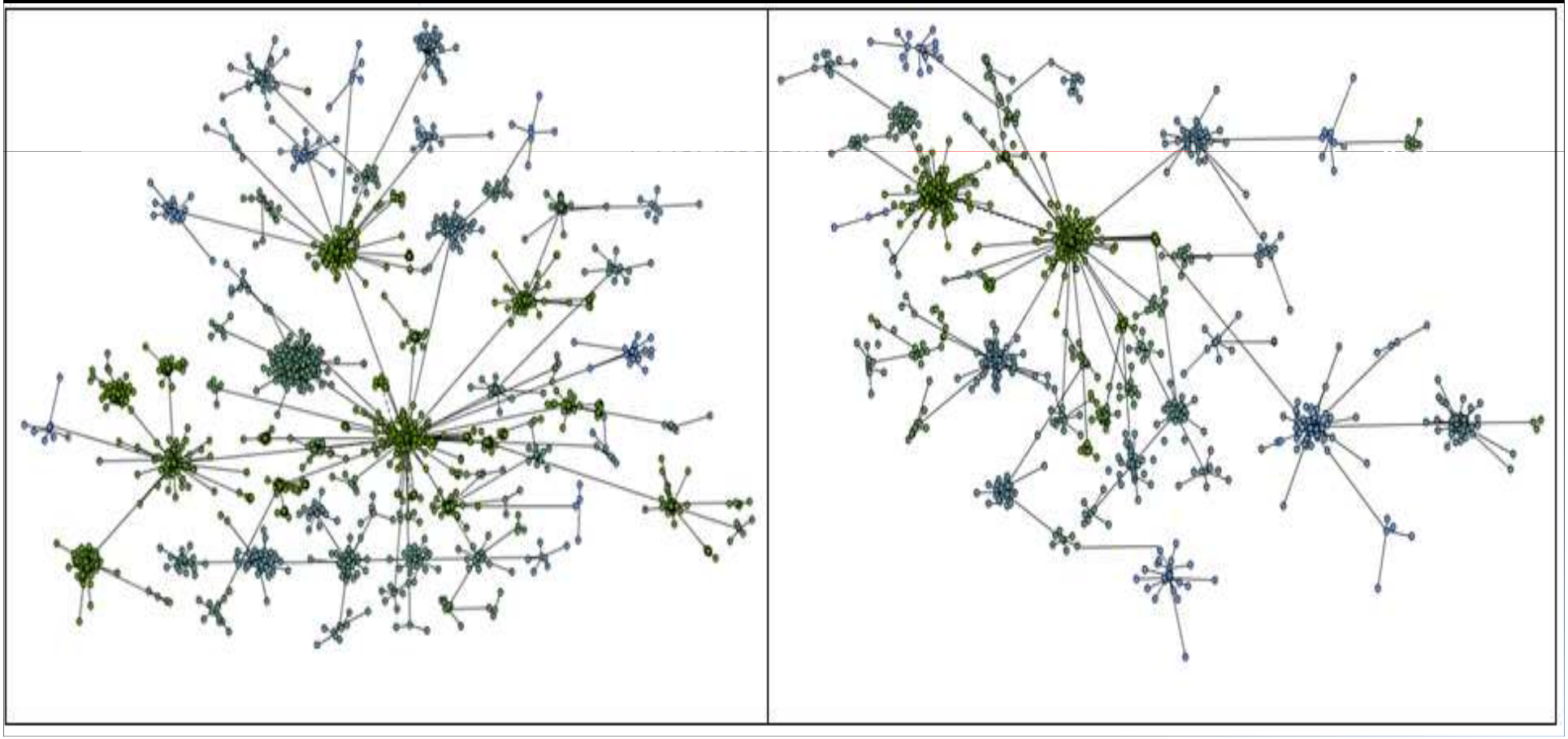
- Los ap genes.
- Los g e entre todos
- Tratamiento de clusters huérfanos.



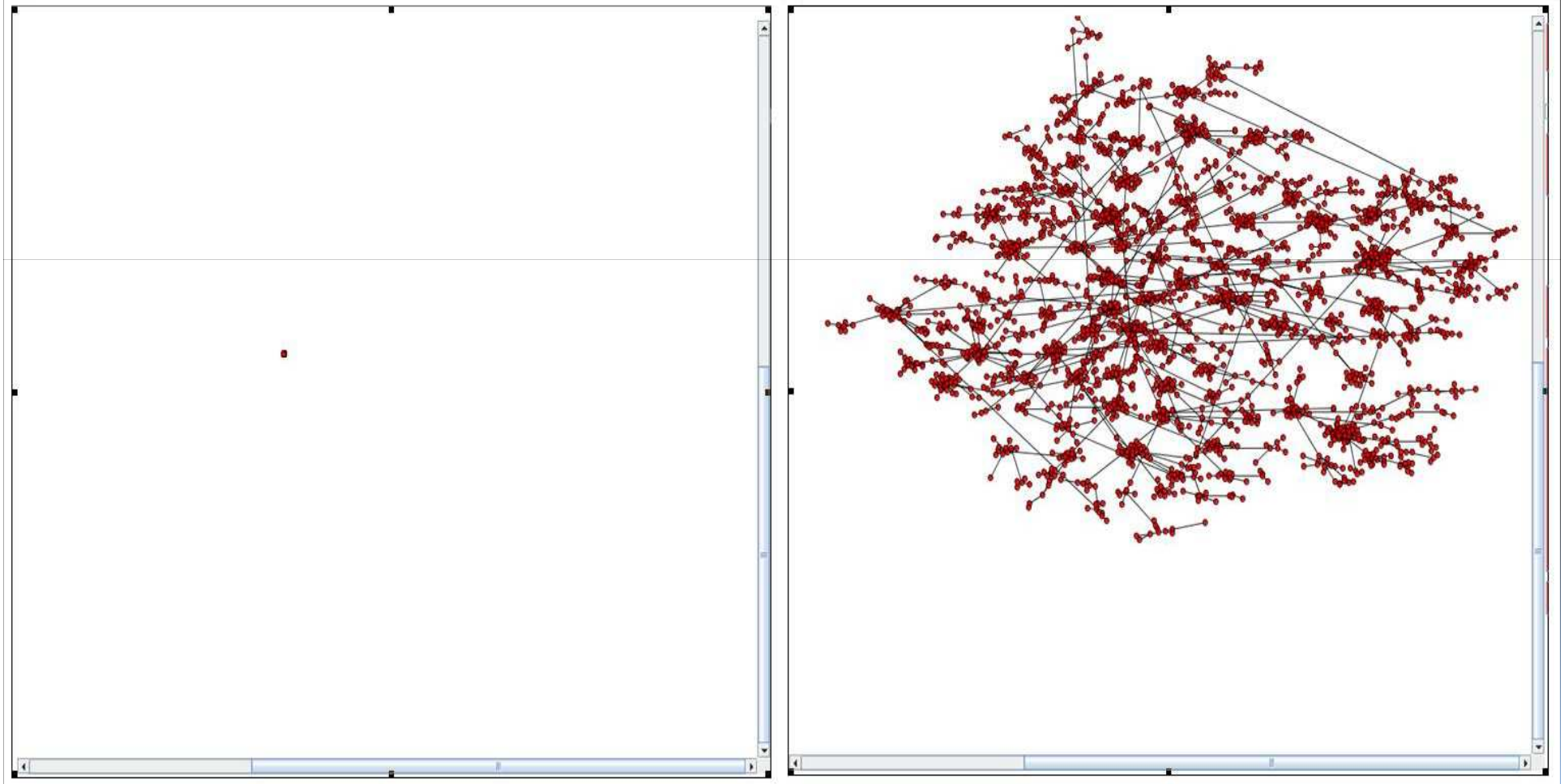
4.3 Tratamiento de microarrays de gran tamaño

3. Proceso de partición de la microarray

4.716



4.3 Tratamiento de microarrays de gran tamaño



4.3 Tratamiento de microarrays de gran tamaño

4. Separación de los ficheros que necesita el applet para las diversas particiones

Objetivo: Separar todos los ficheros previos para las diversas particiones modificando los identificadores de microarray.

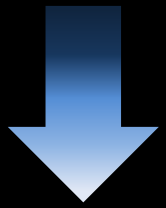
5. Generación del layout para cada partición concreta

Grandes microarrays → Pequeñas correlaciones

Distancias muy pequeñas

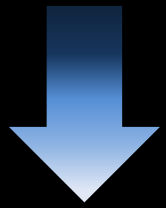
4.4 Adaptación del applet

Ficheros



Diferentes nombres de ficheros según se trabaje con particiones o con microarrays.

Genes



Conversión de identificadores si se trabaja con microarrays.

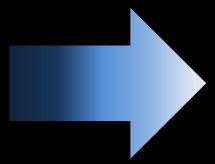
4.5 Filtrado de relaciones de expresión no lineales

- Durante la detección

$$\text{Correlación} < 0.08 \cdot \left(\frac{1.500}{\textit{número de genes}}\right)^{0.6}$$

$$\text{Correlación} < 0.12 \cdot \left(\frac{1.600}{\textit{número de genes}}\right) - \left(\frac{\textit{número de genes}}{40.000}\right)^{18}$$

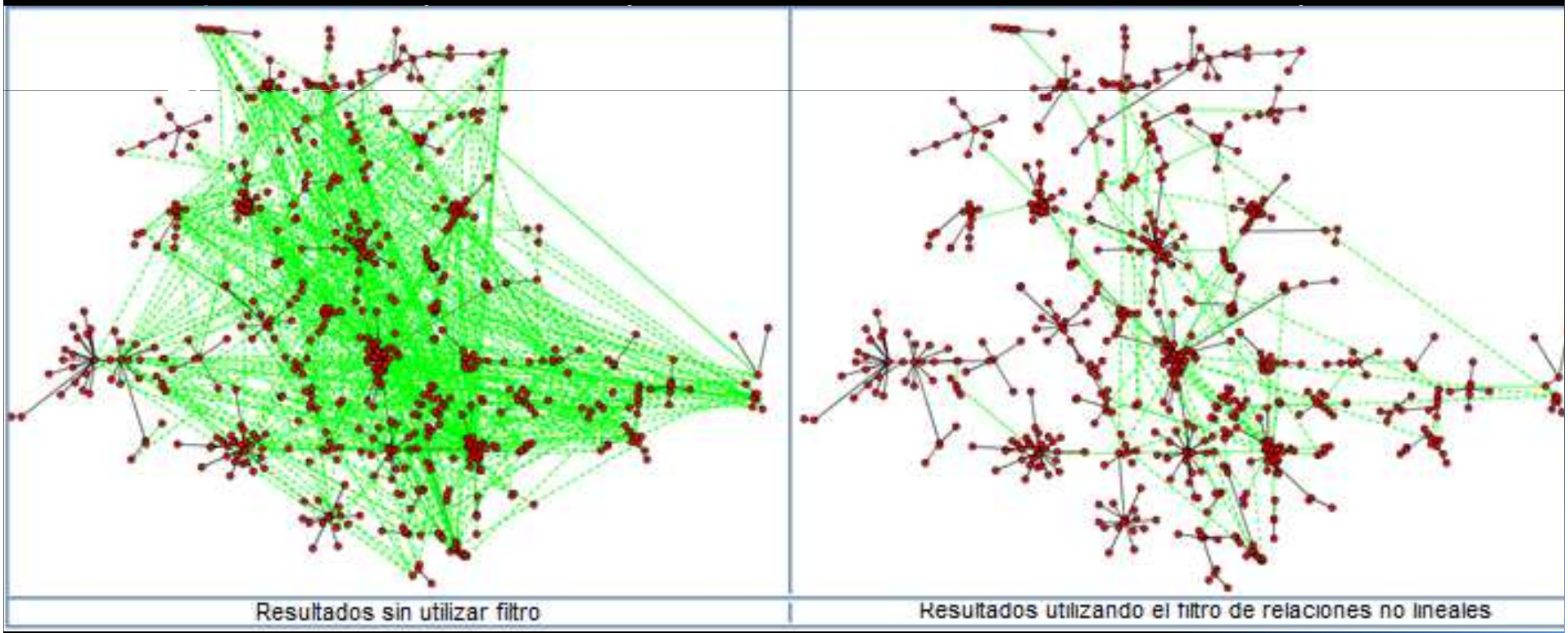
Número de genes	Valor del filtro
673	0,12940043
1.600	0,07696136
14.000	0,02094440
16.000	0,01933182
20.000	0,01690934
27.500	0,01396835
30.000	0,01325782



Número de genes	Valor del filtro
673	0,28528975
1.600	0,12000000
14.000	0,01371428
16.000	0,01199993
20.000	0,00959619
27.500	0,00580446
30.000	0,00076229

4.5 Filtrado de relaciones de expresión no lineales

- Durante la detección
- Mostradas en el applet



4.5 Filtrado de relaciones de expresión no lineales

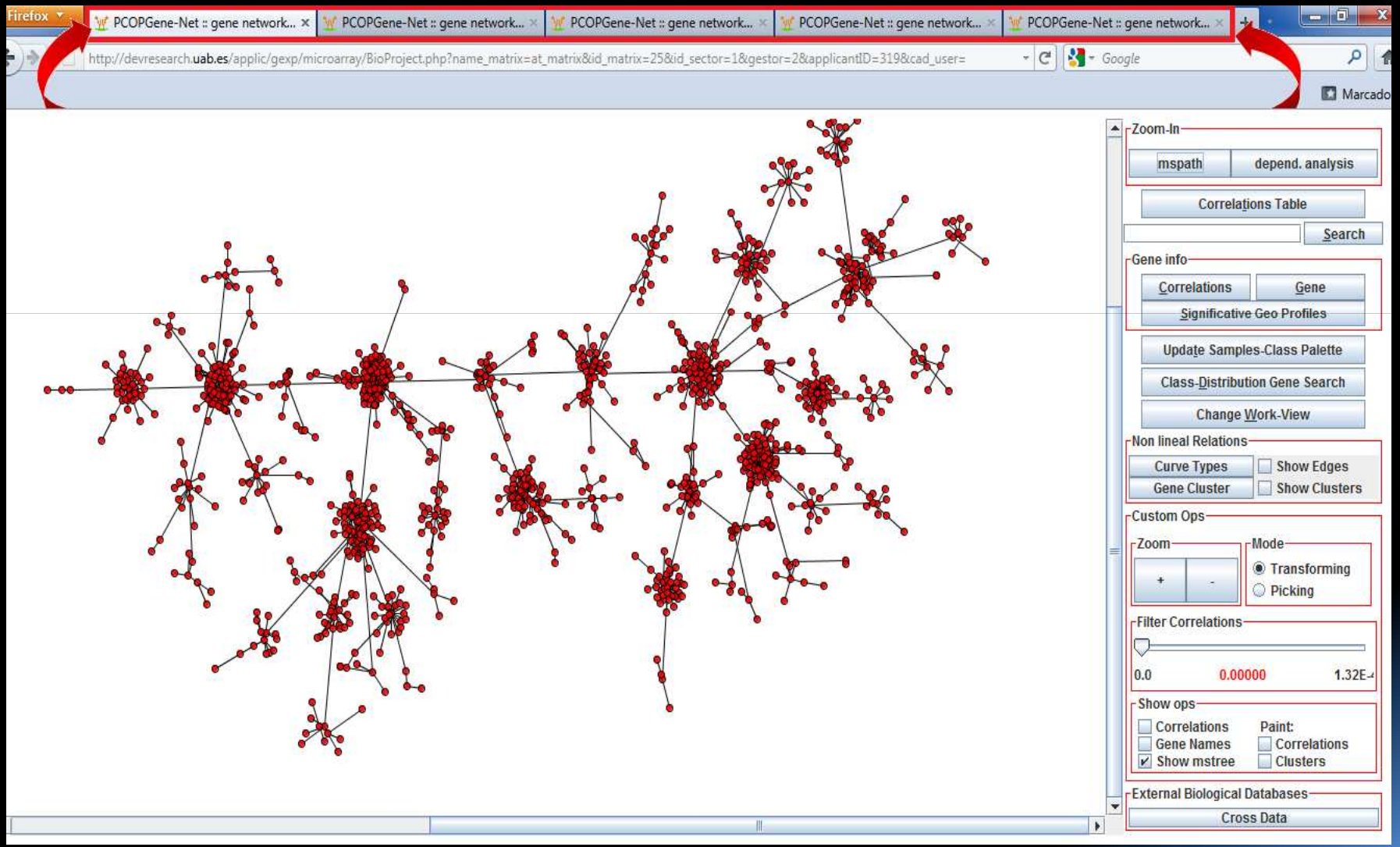
- Durante la detección
- Mostradas en el applet

Objetivo: Seleccionar las mejores curvas para mostrarlas en el applet.

Ventajas:

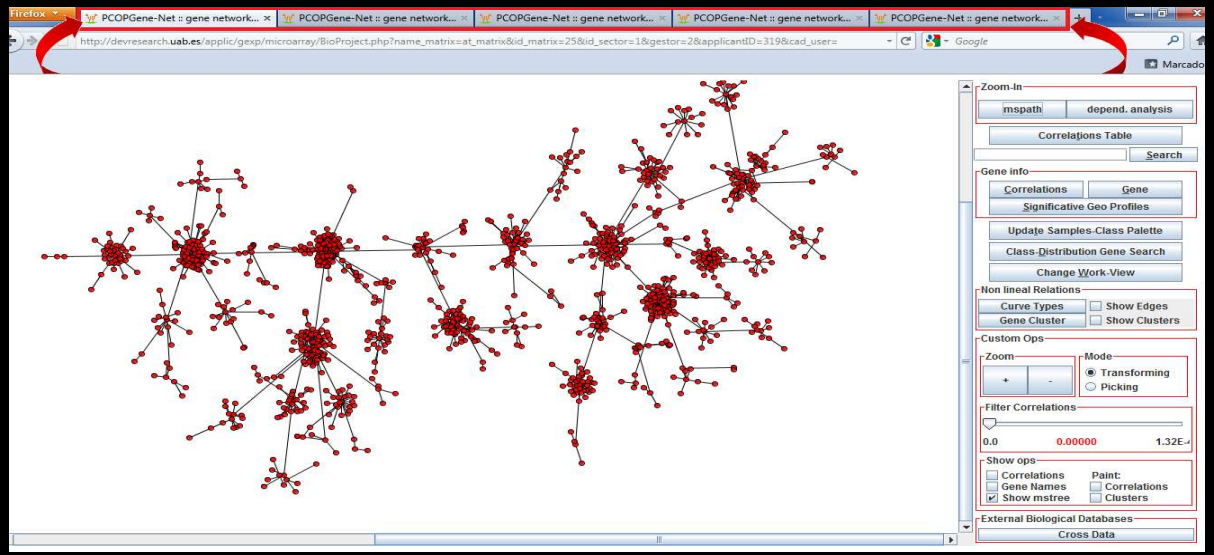
- Visualización más nítida.
- Se evitan problemas derivados del exceso de carga de datos.
- El applet funciona de una manera más rápida.

4.6 Adaptación del aplicativo web



4.6 Adaptación del aplicativo web

Al seleccionar una microarray de gran tamaño se han de abrir todas las particiones que la conforman hasta un máximo de siete.



5 Conclusiones

Los objetivos marcados para la realización del proyecto han sido alcanzados con creces.

Como resultado de mi trabajo ahora se ofrece una nueva herramienta muy útil para los investigadores en el campo de la biología molecular y totalmente adaptada al crecimiento en el volumen de datos que dicha ciencia genera.

6 Bibliografía

- Delicado, P.(2001) Another look at principal curves and surfaces. Journal of Multivariate Analysis, 77, 84-116.
- Delicado, P. and Huerta, M. (2003): 'Principal Curves of Oriented Points: Theoretical and computational improvements'. Computational Statistics 18, 293-315.
- Cedano J, Huerta M, Estrada I, Ballllosera F, Conchillo O, Delicado P, Querol E. (2007) A web server for automatic analysis and extraction of relevant biological knowledge. Comput Biol Med. 37:1672-1675.
- Huerta M, Cedano J, Querol E. (2008) Analysis of nonlinear relations between expression profiles by the principal curves of oriented-points approach. J Bioinform Comput Biol. 6:367-386.
- Cedano J, Huerta M, Querol E. (2008) NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships Advances in Bioinformatics, vol. 2008.
- Huerta M, Cedano J, Peña D, Rodriguez A, Querol E. (2009) PCOPGene-Net: holistic characterisation of cellular states from microarray data base on continuous and non-continuos analysis og gene-expression relationships. BMC Bioinformatics 2009 May 9;10:138.
- Huerta M, Fernández-Márquez J, Cabello JL, Medrano A, Querol1 A, Cedano J (2011) Studying glucocorticoids' Dual Behaviour and Other Tumour-Progression Paradoxes by means of Exhaustive Analysis of Phenotypic Interdependences, Nature Oncogene [Accepted]

GRACIAS