Manual Toolbox para la clusterización de genes en análisis de datos de microarrays de series temporales

Índice

A continuación se describe el uso de la interfaz gráfica de usuario.

Para llevar a cabo una ejecución completa de uno de los algoritmos, se deben realizar los siguientes pasos:

1. Carga de datos de entrada: como se indica posteriormente, existen distintas opciones para efectuar esta operación.
2. Preprocesado: opcionalmente se puede llevar a cabo una selección de características a través de los filtros proporcionados, así como una normalización.
3. Selección de un algoritmo, configuración de sus umbrales y aplicación del mismo.
4. Visualización de los resultados.

Cuando se ejecuta la aplicación por primera vez aparece la siguiente pantalla inicial:

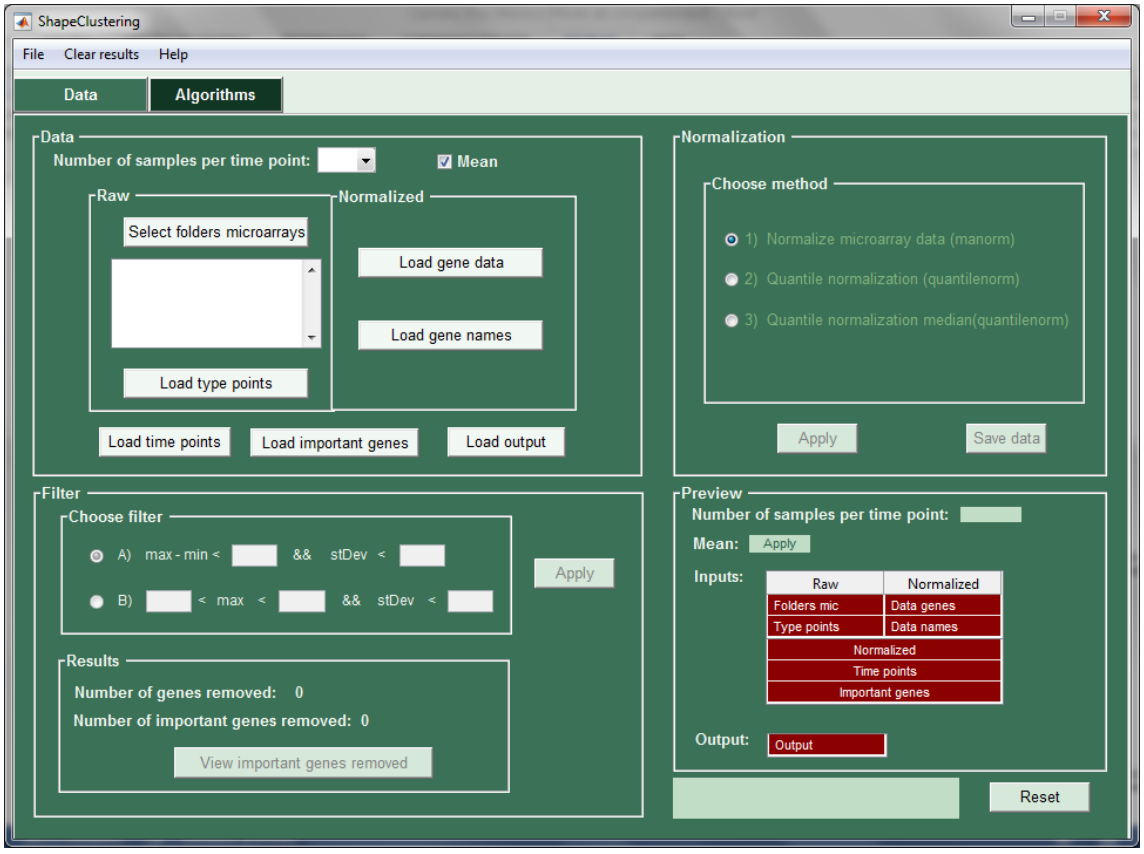


Ilustración 1 Ventana principal de la aplicación

1.1. Barra de menús

Menú File

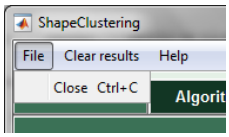


Ilustración 2 Barra de menús – File

- **Close:** cierra la aplicación.

Menú Clear results

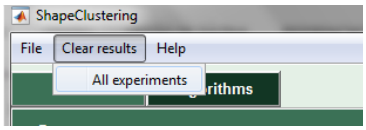


Ilustración 3 Barra de menús - Clear results

- **All experiments:** borra todos los experimentos generados de la aplicación, reinicia la aplicación, pero no elimina los resultados que se han ido guardando en la carpeta de resultados.

Menú Help

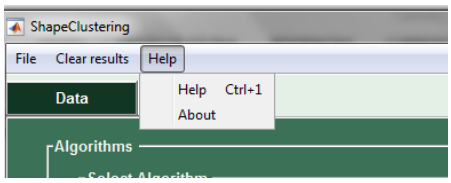


Ilustración 4 Barra de menús - Help

- **Help:** muestra la ayuda en línea de la aplicación, para que el usuario pueda consultar cualquier duda de funcionamiento sin tener que emplear este manual.
- **About:** muestra una pantalla con la información sobre el autor de la aplicación.

1.2. Paneles

En la aplicación existen dos paneles que se visualizan a través de su pestaña correspondiente:

Pestaña *Data*

Panel que se muestra por defecto al inicializar la aplicación. A través de este panel se deben cargar y preparar los diferentes datos de entrada al algoritmo.

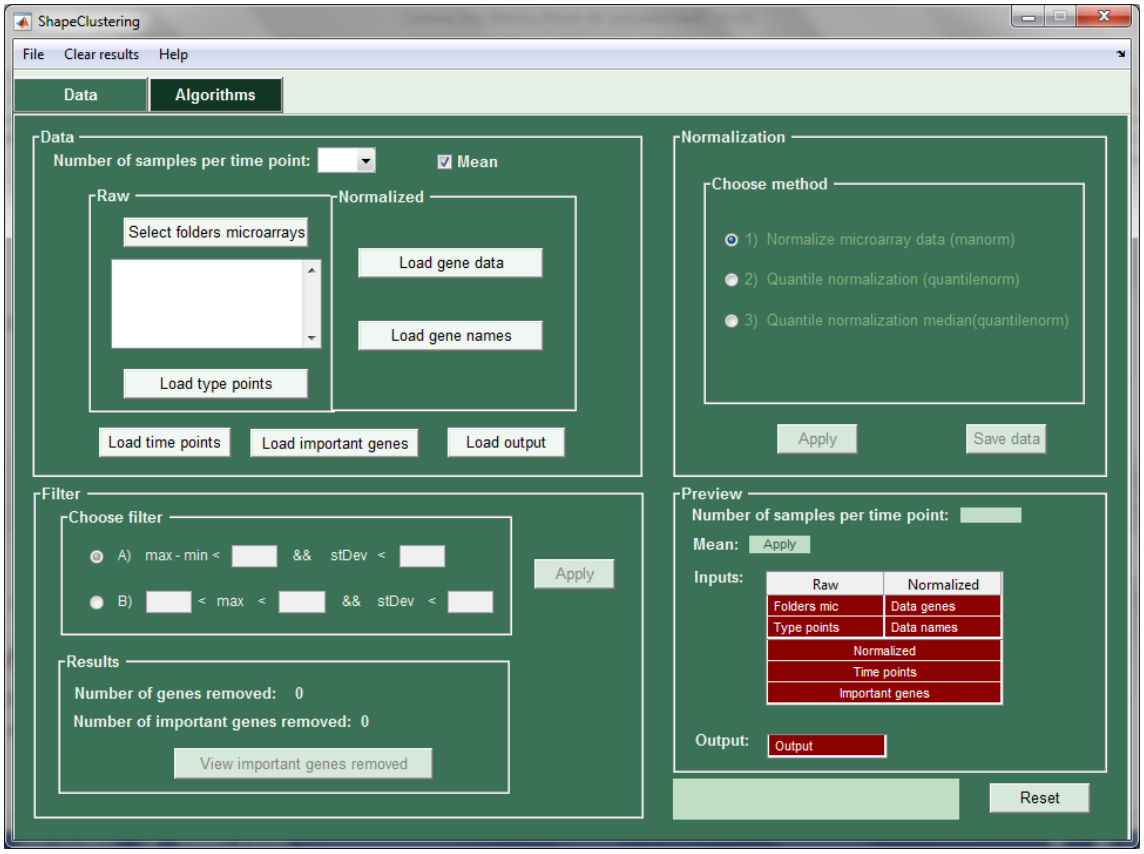


Ilustración 5 Pestaña 'Data' – Por defecto

Este panel está dividido en 4 subpaneles:

Preview

Este subpanel se encuentra en la parte inferior derecha. Por defecto cuando se abre la aplicación muestra el siguiente aspecto:

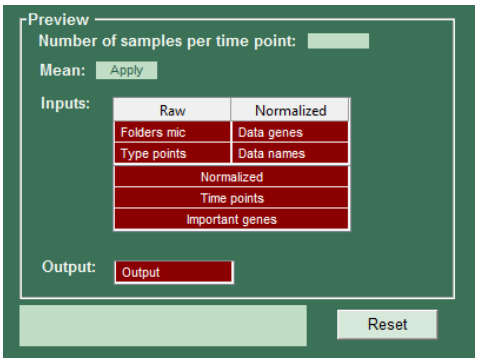


Ilustración 6 Pestaña 'Data' - Preview

A medida que se van cargando los datos del problema este panel se irá actualizando, indicando el número de muestras por instante de tiempo, si se quiere o no aplicar la media sobre los datos cargados, las entradas y la salida y en la parte inferior se va mostrando un log con los mensajes que genera la propia aplicación para informar al usuario.

Se colorean en rojo los elementos (Inputs y Output) que no se han cargado satisfactoriamente todavía. De ahí que aparecen inicialmente todos coloreados. Cabe destacar que, no es requisito que todos ellos se carguen para la ejecución de los algoritmos ya que los datos de entrada que estos necesitan varían de unos a otros.

Data

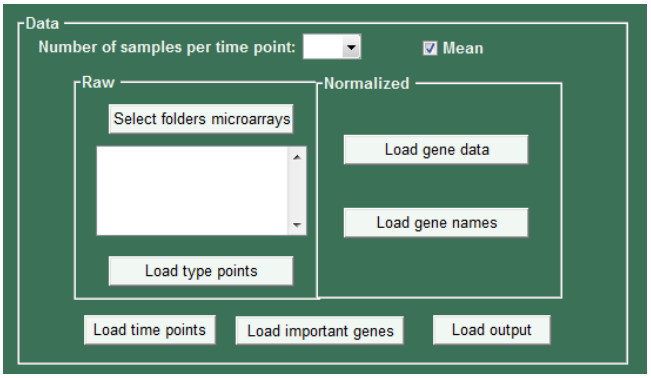


Ilustración 7 Pestaña 'Data' - Data

En este panel en la parte superior izquierda, a través del desplegable *'Number of samples per time point',* se permite indicar el número de muestras que se han tomado de un mismo gen en cada instante de tiempo. Además permite indicar si en los experimentos a realizar se quieren llevar a cabo sobre la media de los datos tomados o no.

Este panel permite la introducción de los datos del problema de dos maneras, en el CD que acompaña a esta documentación se proporcionan los ficheros con los datos asociados a la experimentación llevada a cabo en este estudio:

1. En **bruto** (en la parte izquierda *Raw*) si se tienen los datos sin normalizar: desde el botón ***'Select folders microarrays'*** se seleccionan las carpetas que contienen los microarrays de datos, cada carpeta contendrá un fichero por cada instante de tiempo en que se han tomado las muestras y todas las carpetas tienen que contener el mismo número de ficheros. El número de carpetas introducidas y el número de muestras por instante de tiempo seleccionadas deben coincidir.

Desde el botón ***'Load type points'*** se carga el fichero que indica los tipos de puntos necesario para llevar a cabo la normalización de los datos.
En el CD los ficheros que contienen estos datos se encuentran dentro de la carpeta '*distribuciones_datos_Ejemplo\microarrays*', las carpetas *serie1, serie2* y *serie3* contienen los microarrays de datos y el fichero *typePoints.txt* contiene los tipos de puntos.

A continuación se muestra el estado del panel 'Data' después de cargar los datos en bruto:

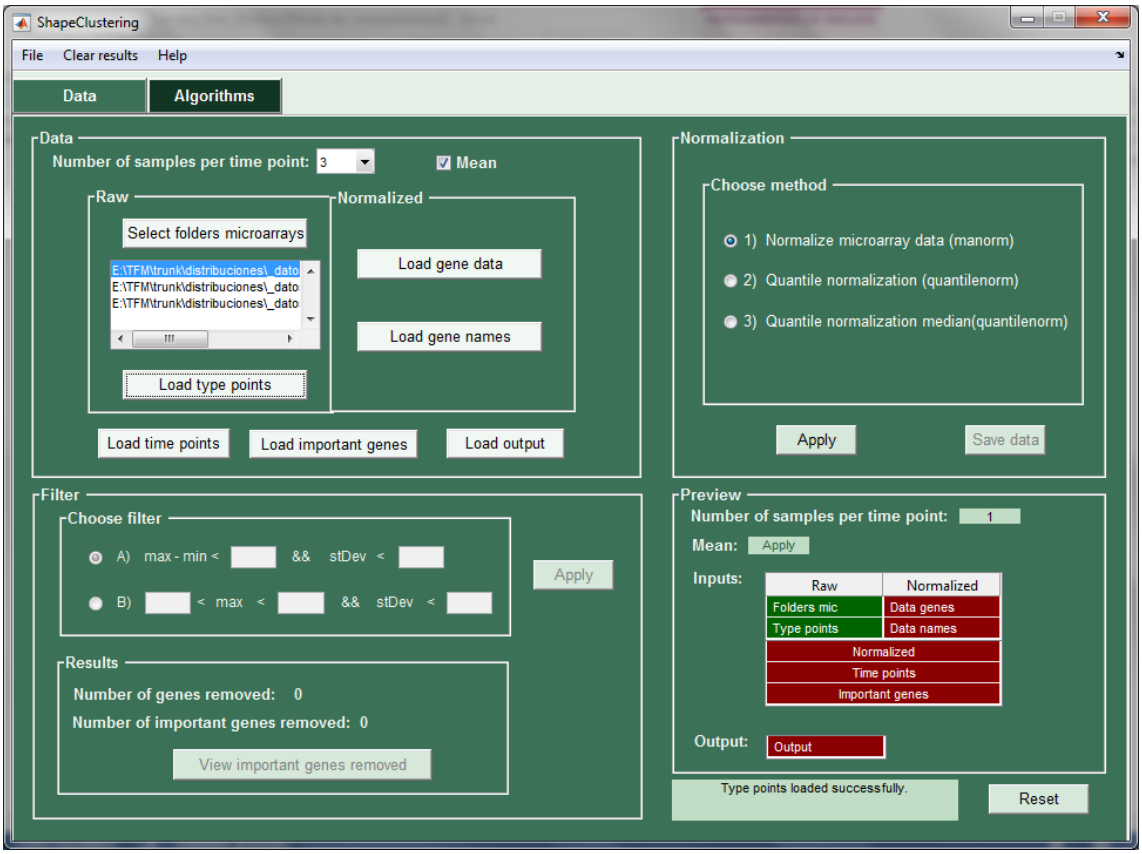


Ilustración 8 Pestaña 'Data' - Datos en bruto cargados

2. **Normalizados** (en la parte derecha *Normalized*), si ya se dispone de los datos normalizados desde el botón *'Load gene data'* se cargan los ficheros de los datos de los genes; el número de ficheros cargados debe coincidir con el número de muestras por instante de tiempo seleccionadas, además el número de genes en cada fichero debe ser el mismo, es decir el tamaño de las matrices de cada fichero ha de ser idéntico. Desde el botón ***'Load gene names'*** se deberá cargar el fichero que contiene el nombre identificativo de cada gen, el

número de nombres de este fichero debe coincidir con el número de genes de los ficheros cargados a través de ***'Load gene data'***.
En el CD los ficheros que contienen estos datos se encuentran dentro de la carpeta '*distribuciones_datos_Ejemplo\data'*, los ficheros *genesDataSerie1.txt, genesDataSerie2.txt* y *genesDataSerie3.txt* contienen los datos de 8848 genes en 12 instantes de tiempo y el fichero *genesName.txt* contiene el nombre de los 8848 genes.

A continuación se muestra el estado del panel 'Data' después de cargar los datos normalizados:

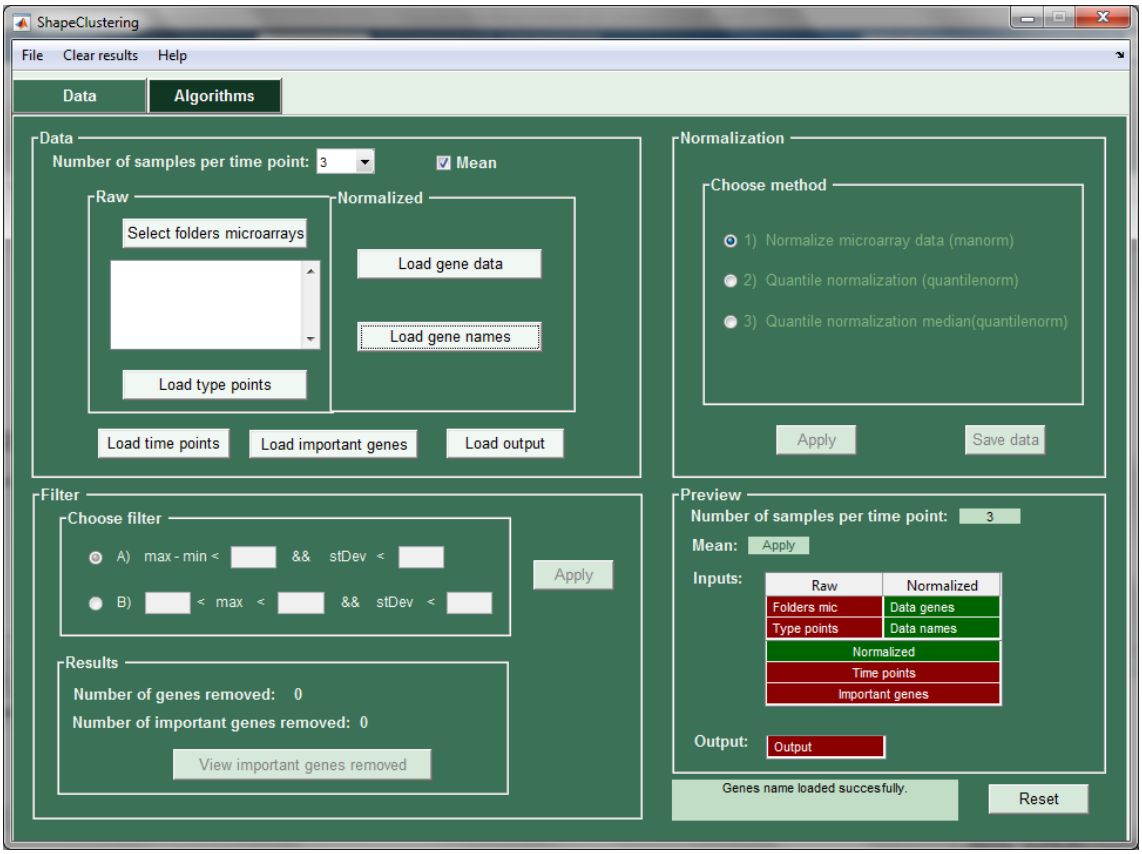


Ilustración 9 Pestaña 'Data' - Datos normalizados cargados

Además, desde este mismo subpanel se deben cargar los siguientes ficheros:

- ***'Load time points'**:* fichero que contienen los **instantes de tiempo** en que se han tomado las muestras. El número de instantes de tiempo de este fichero tiene que coincidir con el número de muestras que se tienen de cada gen en los ficheros cargados a través de *'Load gene data'* o con el número de ficheros de microarrays que contiene cada carpeta seleccionada por *'Select folders microarrays'*.
  En el CD el fichero a cargar desde este botón se encuentra en '*distribuciones_datos_Ejemplo\data\timePoints.txt'.*

Panel ‘Data’ después de cargar los instantes de tiempo:

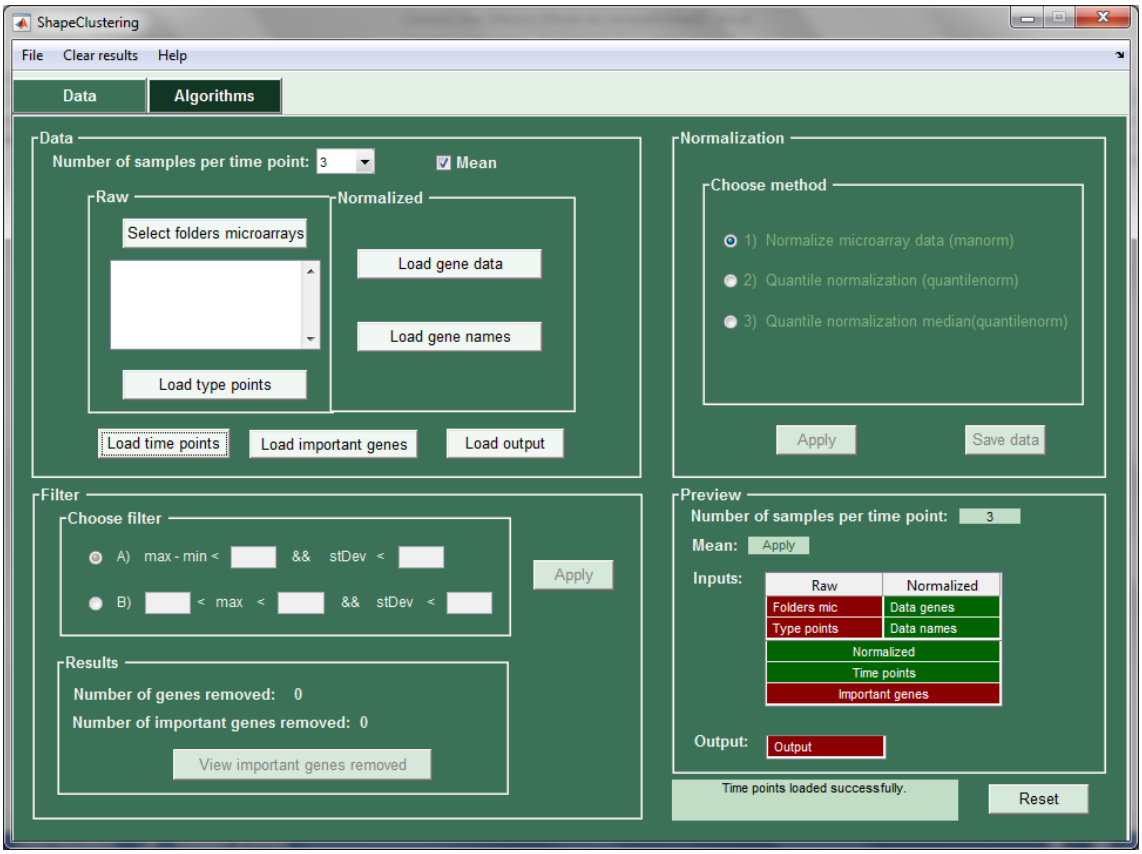


Ilustración 10 Pestaña 'Data' - Instantes de tiempo cargados

- ***‘Load important genes’:*** fichero que identifica los genes más importantes. Los nombres que se encuentren dentro de este fichero han de estar contenidos dentro del fichero cargado desde el botón *‘Load gene names’.*
  En el CD el fichero a cargar desde este botón se encuentra en ‘*distribuciones_datos_Ejemplo\data\nameImportantGenes.txt’.*

Panel ‘Data’ después de cargar los genes importantes:

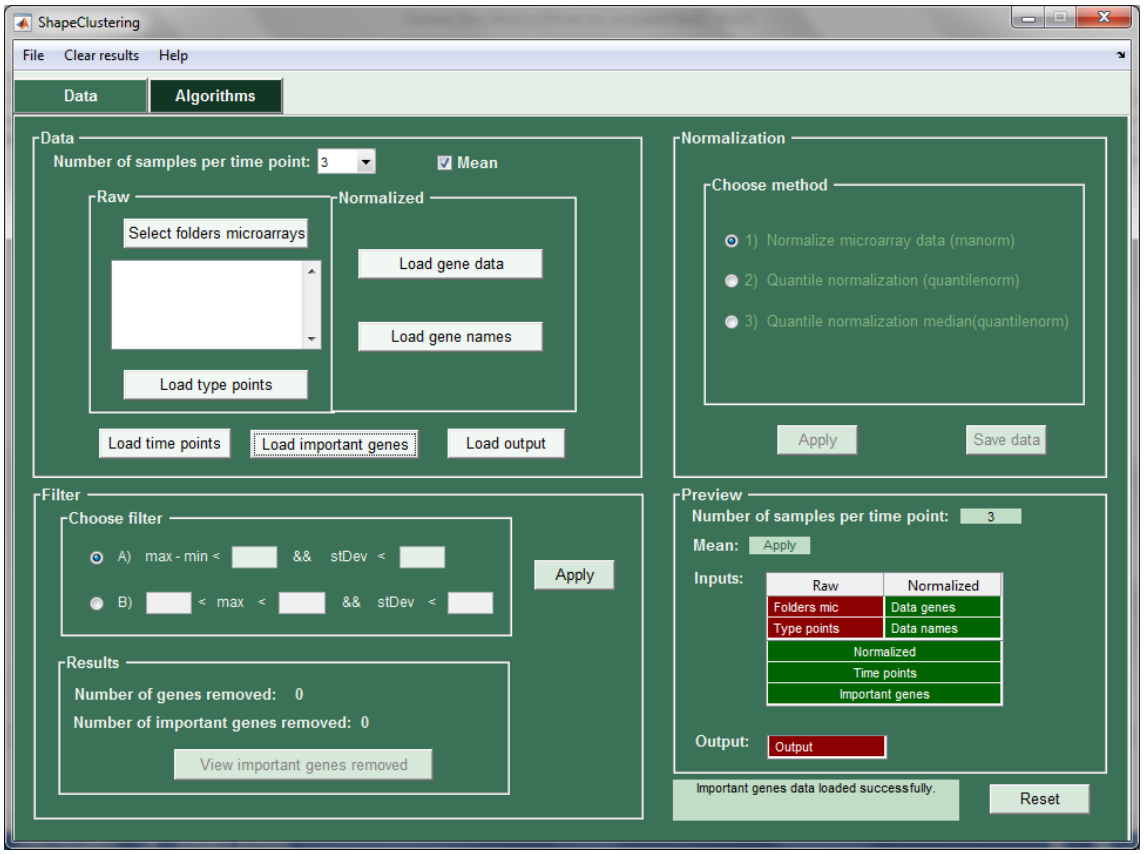


Ilustración 11 Pestaña 'Data' - Genes importantes cargados

- ***'Load output'**: f*ichero/s que contienen los datos de salida. Dos de los cinco algoritmos implementados tienen en cuenta la correlación de cada gen con la salida, para ejecutar esos algoritmos se han de cargar los ficheros necesarios. El número de ficheros cargados debe coincidir con el número de muestras por instante de tiempo seleccionadas.

  En el CD estos ficheros se encuentran dentro de la carpeta *'distribuciones_datos_Ejemplo\data'*: *growthSerie1.txt, growthSerie2.txt* y *growthSerie3.txt.*

  Panel 'Data' después de cargar los datos de salida:

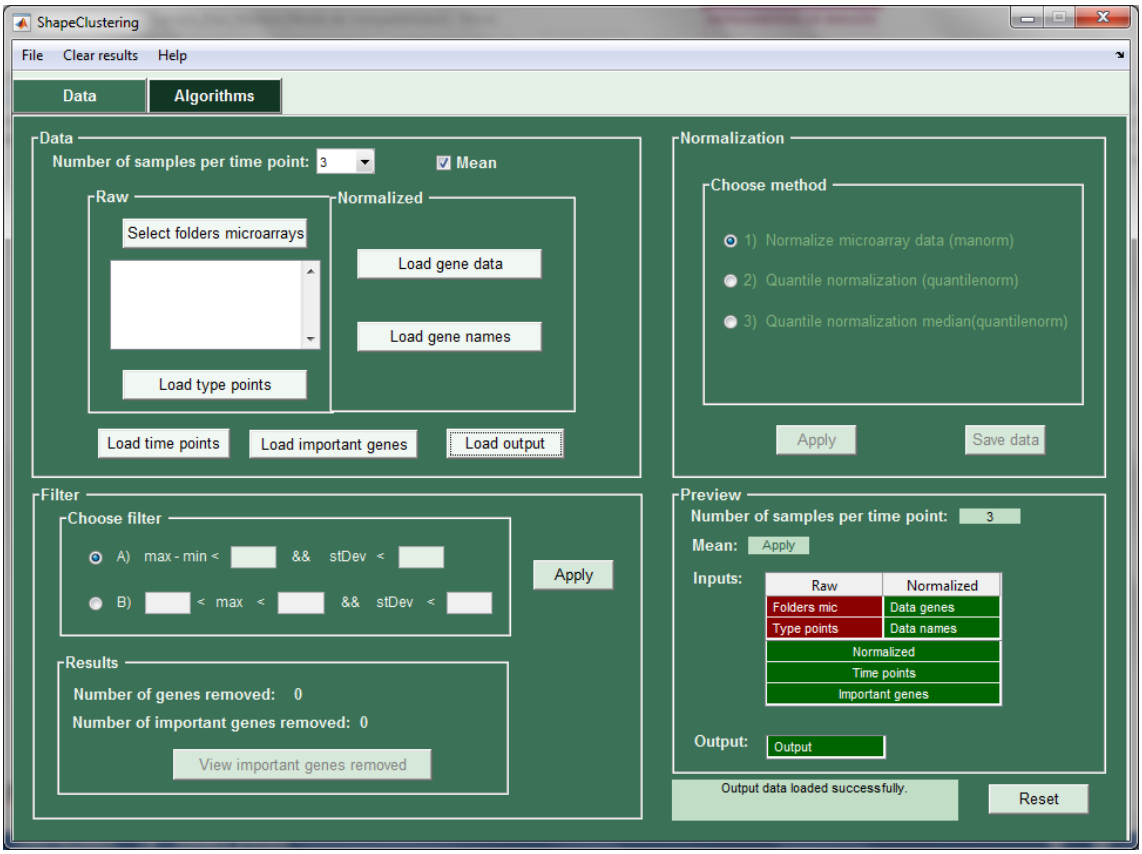


Ilustración 12 Pestaña 'Data' - Datos de salida cargados

Normalization

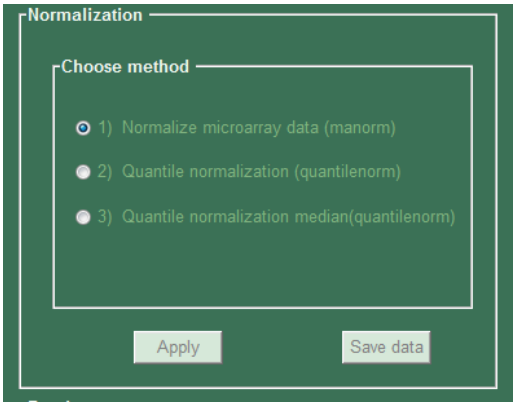


Ilustración 13 Pestaña 'Data' - Normalization

Este subpanel por defecto se encuentra deshabilitado. Si en el subpanel anterior se cargan los datos en bruto (primer caso del apartado anterior) este panel se habilitará automáticamente para poder llevar a cabo la normalización de los datos cargados, de acuerdo con 3 opciones disponibles, tal y como se muestra en la siguiente imagen.

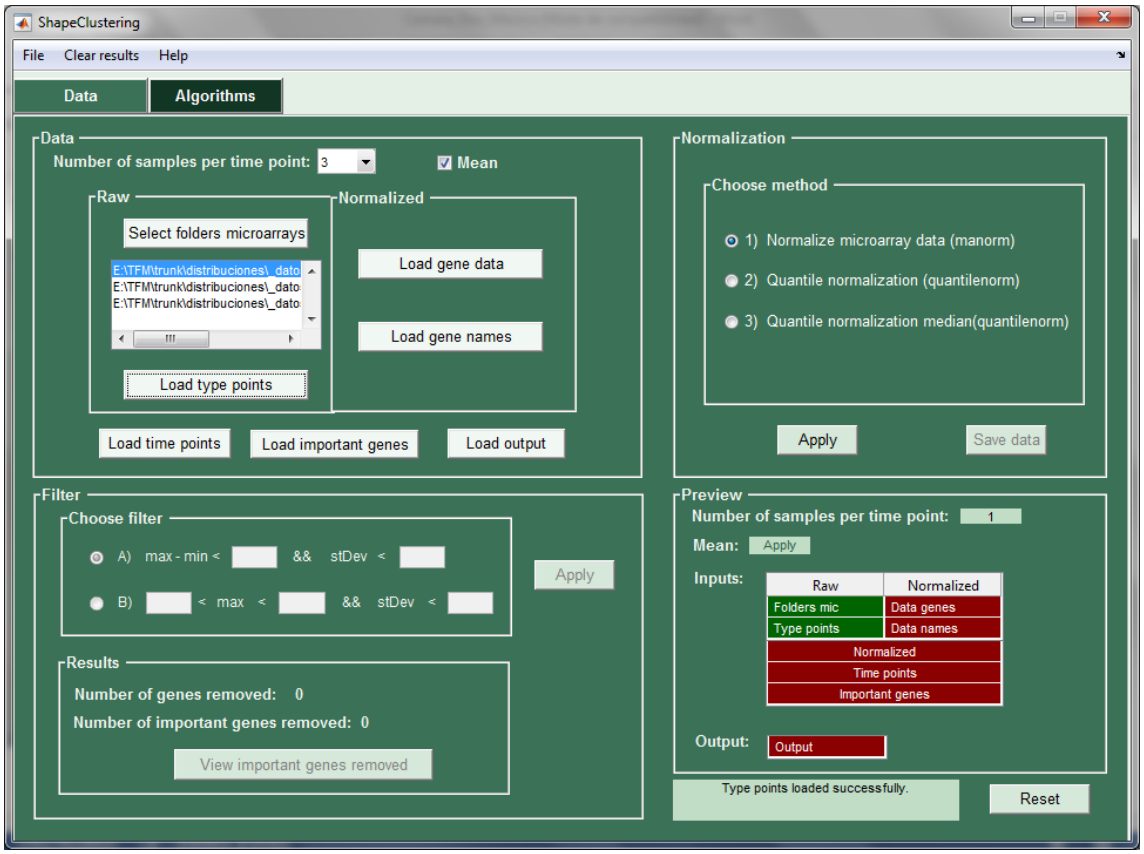


Ilustración 14 Pestaña 'Data' - Datos en bruto cargados y subpanel Normalization activo

Este subpanel permite la opción de tres formas de normalización:

- **Normalize microarray data** **[**¡Error! No se encuentra el origen de la referencia.**]:** función de Matlab que normaliza datos de mircroarrays. *XNorm = manorm(X)* escala los valores de cada columna de X dividiendo por la intensidad media de la columna.
- **Quantile normalization [**¡Error! No se encuentra el origen de la referencia.**]:** función de Matlab que normaliza XXX. *NormData = quantilenorm(Data)* normaliza las distribuciones de los valores en cada columna de Data.
- **Quantile normalization median [**¡Error! No se encuentra el origen de la referencia.**]:** *NormData = quantilenorm(...,'MEDIAN', true)*, misma función que la anterior pero toma la mediana de los valores clasificados en lugar de la media.

Mientras se está llevando a cabo la normalización en el subpanel 'Preview' se muestra la siguiente información:

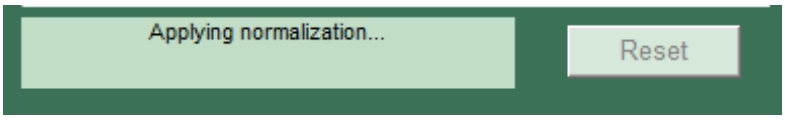


Ilustración 15 Pestaña 'Data' - Aplicando normalización

A continuación se muestra el estado del panel 'Data' después de aplicar una de las normalizaciones sobre los datos. Como se puede observar en la esquina inferior derecha, aparece en verde el elemento 'Normalized' y se indica en la ventana de log:

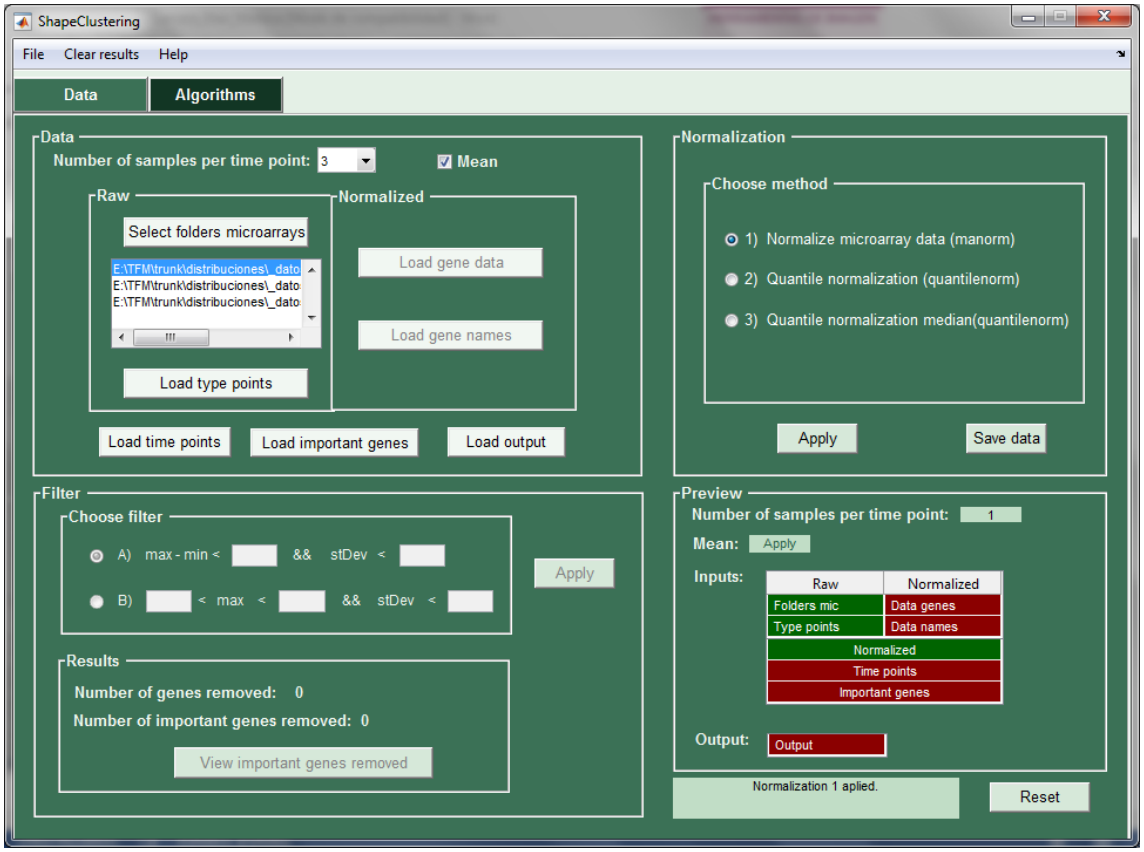


Ilustración 16 Pestaña 'Data' - Después de aplicar una normalización sobre los datos

A través del botón 'Save data' del mismo subpanel se permite guardar los datos que se acaban de normalizar. Cuando se pulsa este botón la aplicación pregunta al usuario por la carpeta donde quiere guardar los datos e indicar un nombre:

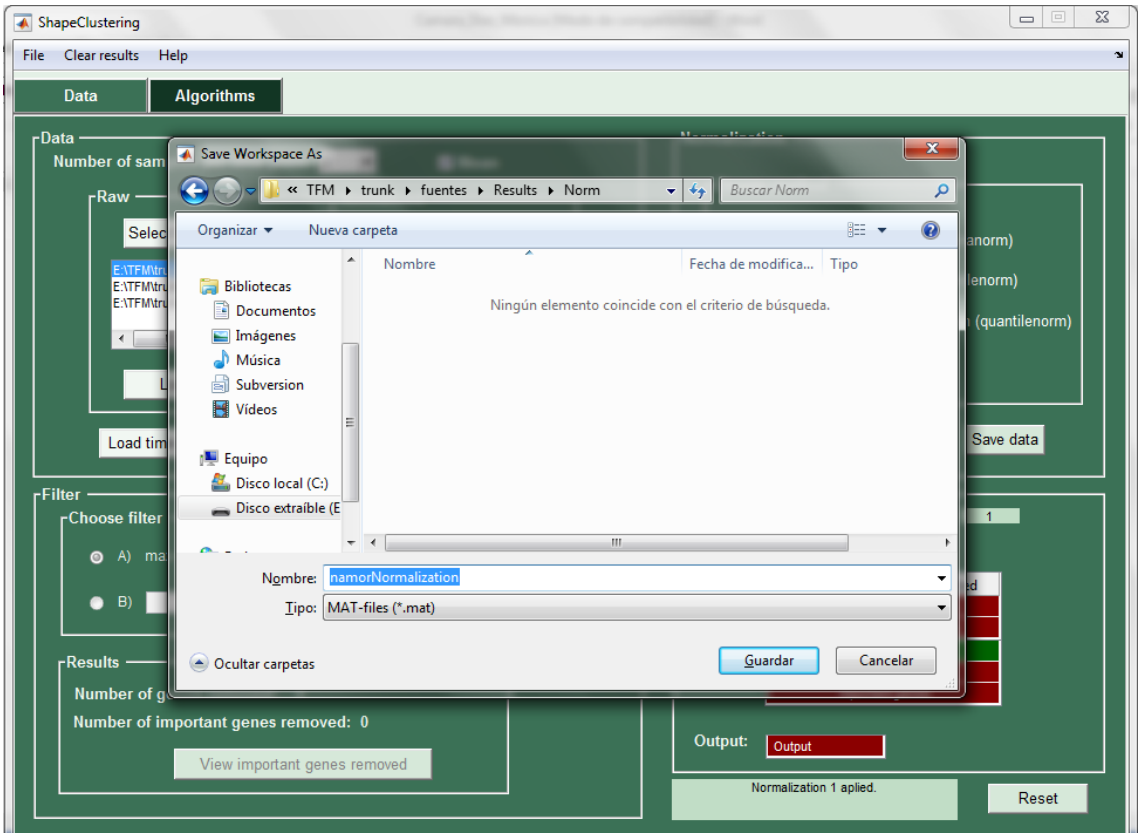


Ilustración 17 Pestaña 'Data' - Guardar datos normalizados

La aplicación guarda un fichero .mat con el resultado de la normalización, un fichero *x_genesName.txt* con el nombre de los genes y un fichero *x_genesData.txt* por cada muestra con los datos de los genes, como se muestra en la siguiente imagen.

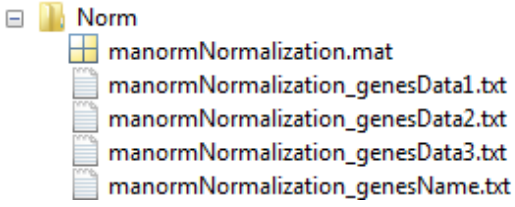


Ilustración 18 Pestaña 'Data' - Ficheros guardados para normalización

Los ficheros .txt almacenados de esta manera están preparados para poder ser usados en la propia aplicación.

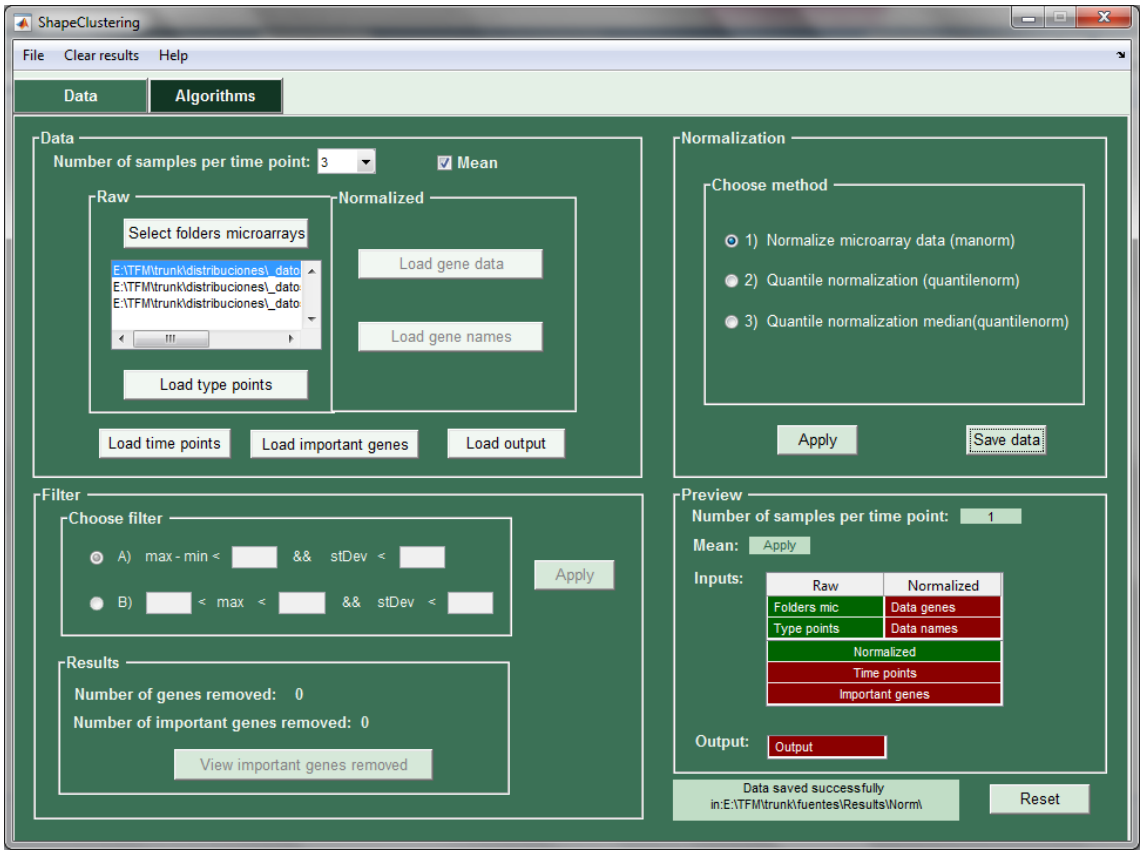


Ilustración 19 Pestaña 'Data' - Después de guardar los datos normalizados

Filter

Este subpanel por defecto se encuentra deshabilitado, habilitándose una vez que haya datos normalizados disponibles. Permite realizar una selección de características seleccionando una de las dos posibles opciones de los filtros, así como especificar valores para sus parámetros.

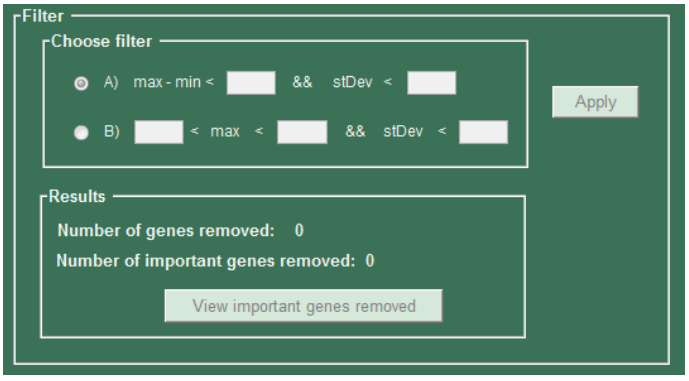


Ilustración 20 Pestaña 'Data' - Filter por defecto

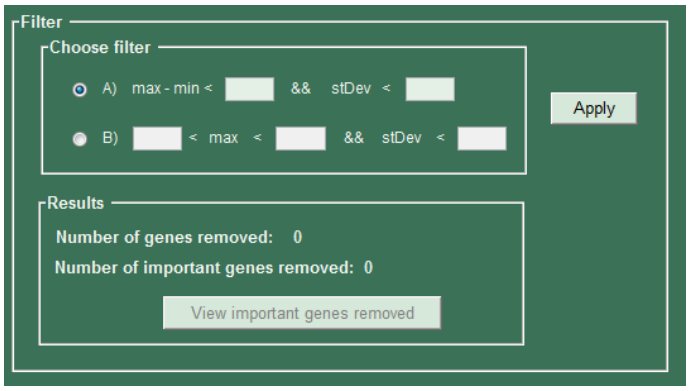


Ilustración 21 Pestaña 'Data' - Filter cuando se dispone de datos

La aplicación permite aplicar diferentes filtros sobre el conjunto de datos inicial con el fin de eliminar ruido.

Se puede elegir entre dos tipos de filtros, estos filtros incluyen condiciones sobre el valor máximo y/o mínimo y la desviación estándar sobre el nivel de expresión genética.

Una vez elegido el filtro e indicados los parámetros que se quieren aplicar, se ha de pulsar el botón *'Apply'* y en el cuadro de *Results* se mostrarán los resultados obtenidos. En este cuadro se informa acerca de:

- Número de genes eliminados a través del filtro.
- De esos genes eliminados cuántos de ellos son importantes. En el caso de no haber ninguno el botón *'View important genes removed'* seguirá deshabilitado y en el caso de encontrarse genes importantes entre los eliminados este botón se activará para poder visualizar cuales son esos genes importantes.

En la Ilustración 22 se muestra el estado del subpanel *Filter* después de aplicar el primer filtro con los parámetros 0.75 y 0.25. Se indica que el número de genes eliminados es de 566 y ninguno de ellos es importante.

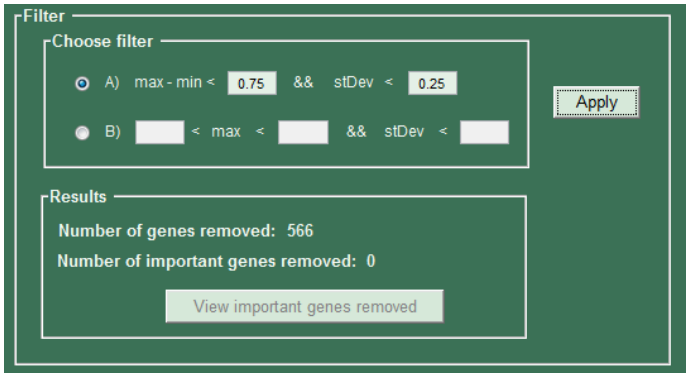


Ilustración 22 Pestaña 'Data' - Filter después de aplicar el primer filtro

En la Ilustración 23 se muestra el estado del subpanel Filter después de aplicar el primer filtro con los siguientes valores para sus parámetros: 1 y 0,25. Se indica que el número de genes eliminados es de 897 y uno de ellos es importante.

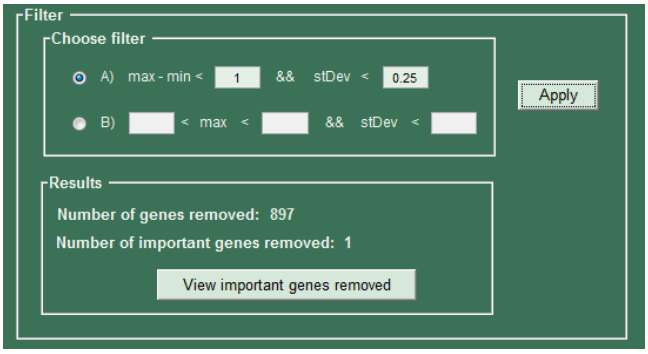


Ilustración 23 Pestaña 'Data' - Filter después de aplicar el primer filtro con genes importantes eliminados

Pulsando sobre el botón 'View importante genes removed' se abre la imagen mostrada en la Ilustración 24, ofreciendo información sobre el gen importante que se ha eliminado como consecuencia del filtrado:

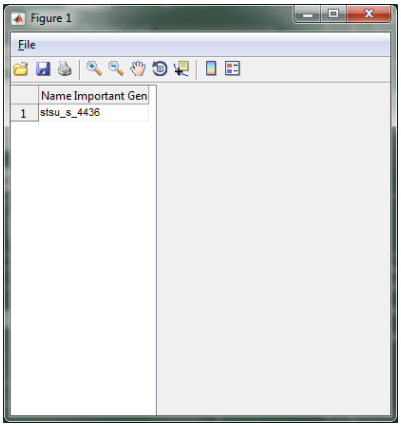


Ilustración 24 Genes importantes eliminados a través de la aplicación de un filtro

Pestaña Algorithms

Este panel al iniciar la aplicación se encuentra deshabilitado, como se muestra en la siguiente imagen:

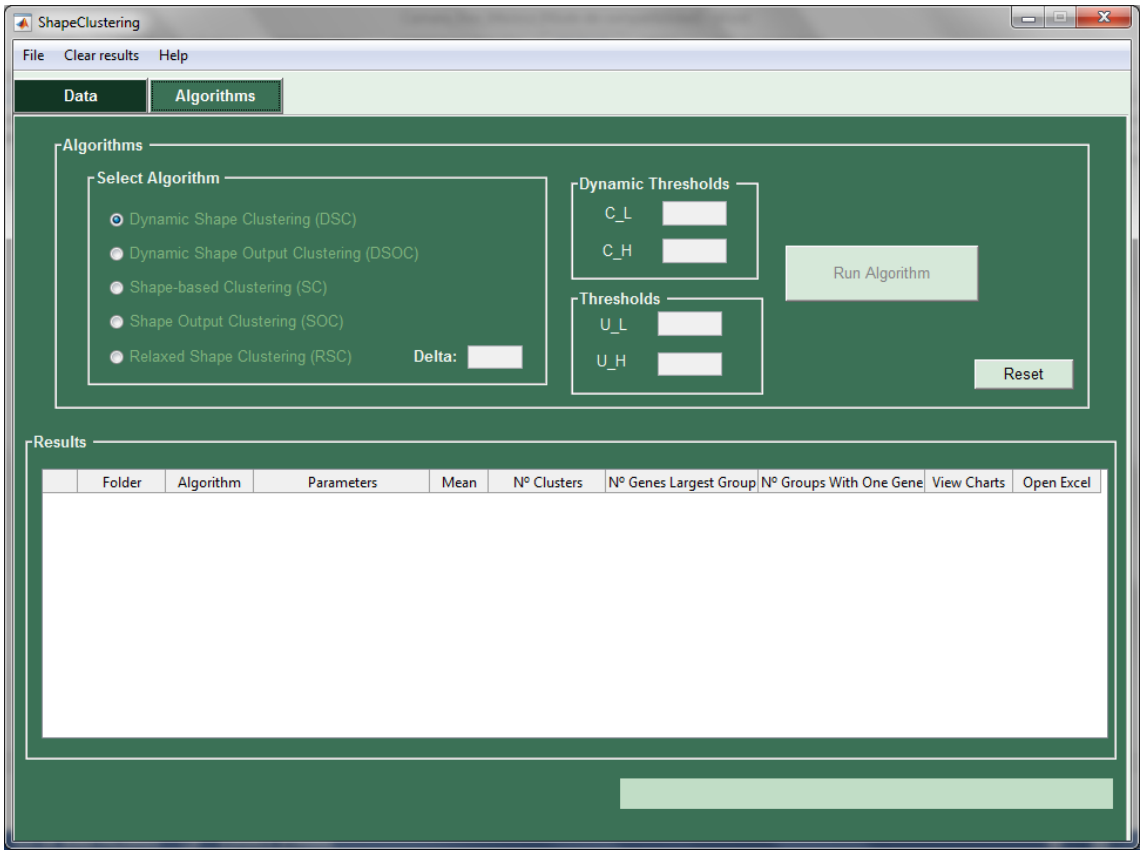


Ilustración 25 Pestaña 'Algorithm' – Por defecto

Una vez que se han cargado datos a través del panel *Data* se habilitan las diferentes opciones de este, como se puede observar a continuación:

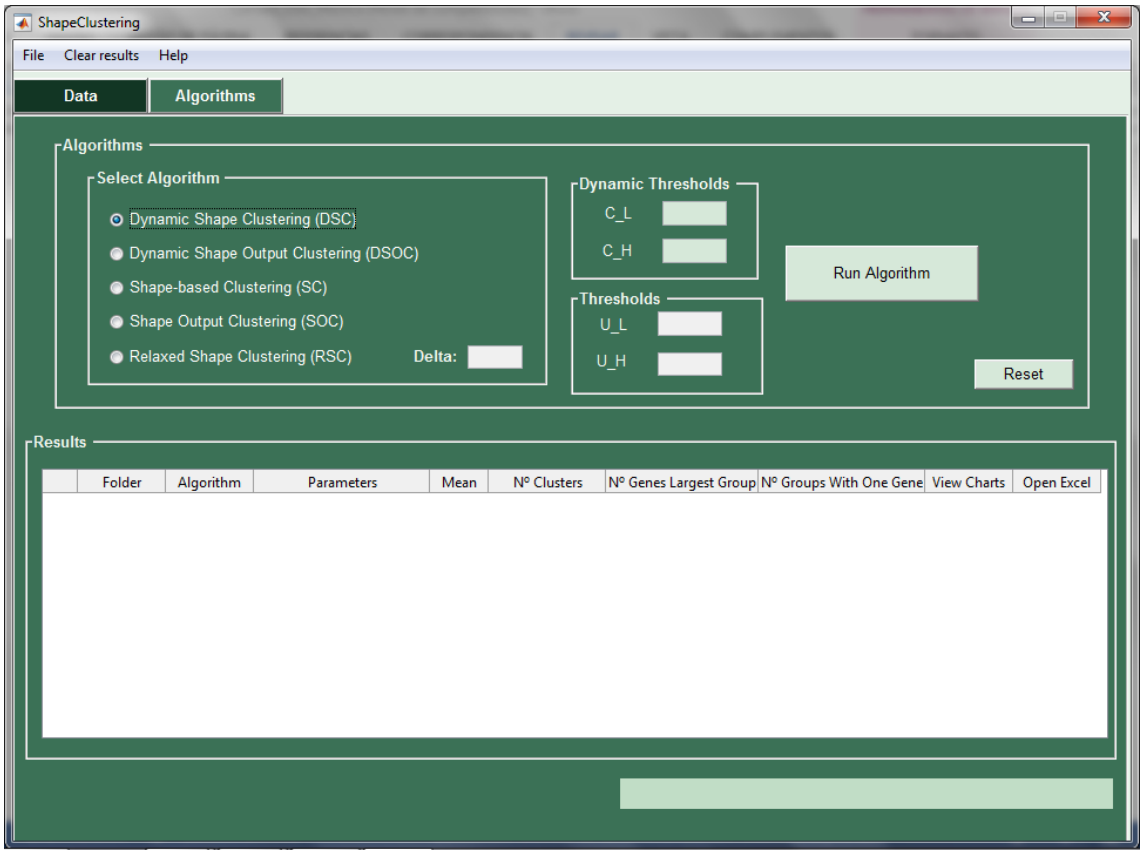


Ilustración 26 Pestaña 'Algorithm' - Habilitado

El panel permite seleccionar uno de entre los cinco algoritmos disponibles. Si se selecciona el algoritmo SC o SOC se han de introducir valores para los parámetros φ_L y φ_H a través de los campos denominados como U_L y U_H respectivamente:

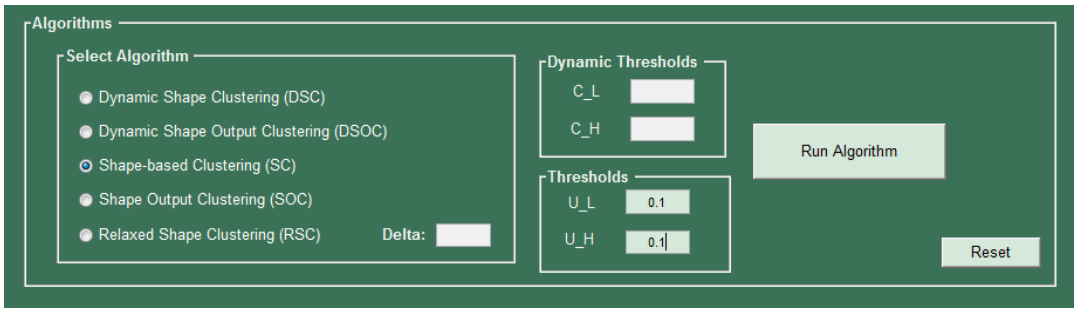


Ilustración 27 Pestaña 'Algorithm' – Seleccionado algoritmo SC

Si se selecciona el algoritmo RSC además de indicar los parámetros anteriores se ha de indicar también el parámetro δ en el campo indicado como Delta.

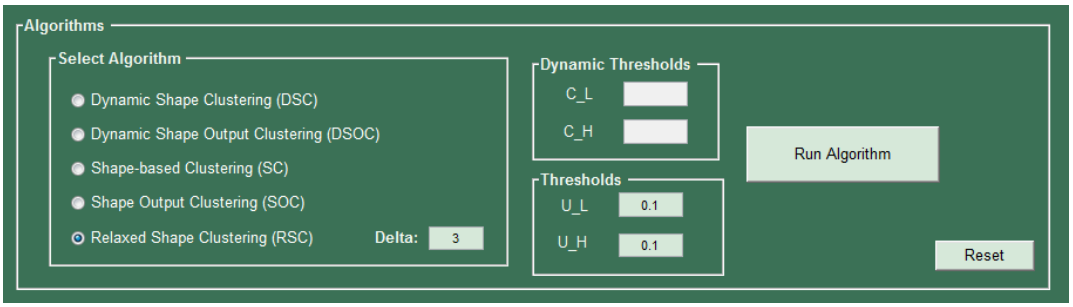


Ilustración 28 Pestaña 'Algorithm' – Seleccionado algoritmo RSC

Si las opciones elegidas son DSC o DSOC los parámetros cuyos valores se deben introducir son C_L y C_H a través de los campos C_L y C_H:

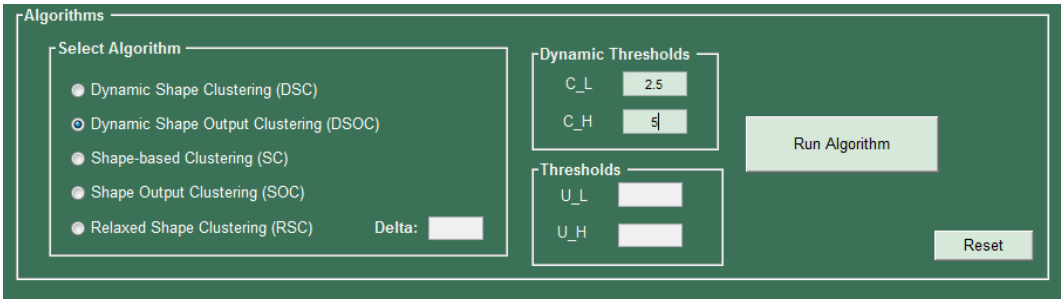


Ilustración 29 Pestaña 'Algorithm' – Seleccionado algoritmo DSOC

Una vez seleccionado el algoritmo e introducidos los parámetros se ha de pulsar el botón:

Ilustración 30 Pestaña 'Algorithm' – Botón para ejecutar un algoritmo

Durante la ejecución del algoritmo en la parte inferior de la ventana se va indicando al usuario los pasos que se están llevando a cabo:

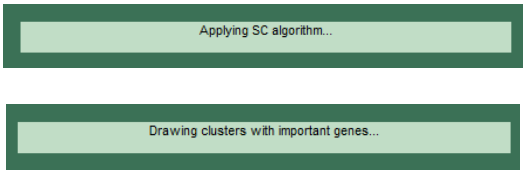


Ilustración 31 Pestaña 'Algorithm' – Mensajes durante la ejecución del algoritmo

Al lanzar el algoritmo se genera una nueva carpeta dentro de la carpeta */Results* con la nomenclatura 'YYYYMMDDhhmm', de acuerdo con el año (YYYY), mes (MM), día (DD), hora (hh) y minutos (mm) en que se ha llevado a cabo la ejecución. Dentro de esta carpeta se genera una nueva carpeta con el nombre del algoritmo ejecutado y los

valores asignados a sus parámetros. En la Ilustración 32 se puede ver el resultado de la carpeta */Results* después de lanzar el algoritmo de la Ilustración 29:

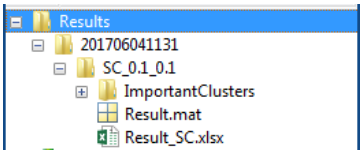


Ilustración 32 Carpeta Results después de la ejecución de un algortimo

Una vez que el algoritmo ha finalizado, automáticamente guarda los resultados obtenidos en formato Matlab en un archivo .mat, genera un fichero .xlsx y además genera y guarda las gráficas que contienen genes importantes. El archivo *Result.mat* contiene tres estructuras de datos: *fil*, *obj* y *Analysis*. En la Ilustración 33 se puede observar que el filtro aplicado es el *FilterA* y el algoritmo que se ha ejecutado es el *SC*:

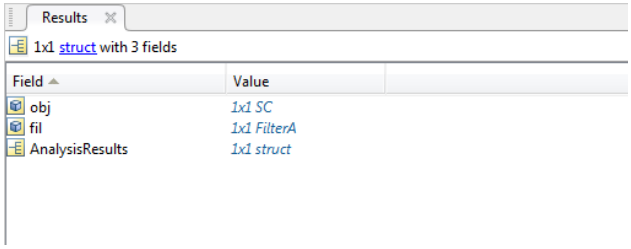


Ilustración 33 Contenido del archivo Results.mat

La estructura *fil,* Ilustración 34*,* guarda los datos originales antes de la aplicación del filtro (*genesData, genesName* y *nameImportantGenes*), y el resultado obtenido después de aplicar el filtro sobre ellos. Como resultado se guardan los parámetros con que se ha aplicado el filtro (*maxmin* y *stdev*), el identificador de los genes eliminados (*geneIdRemove*), el nombre de los genes eliminados (*geneNameRemove*), el nombre los genes importantes que se han eliminado como consecuencia de aplicar el filtro (importantGeneNameRemove) y el número de genes importantes eliminado (nImportantGeneRemove):

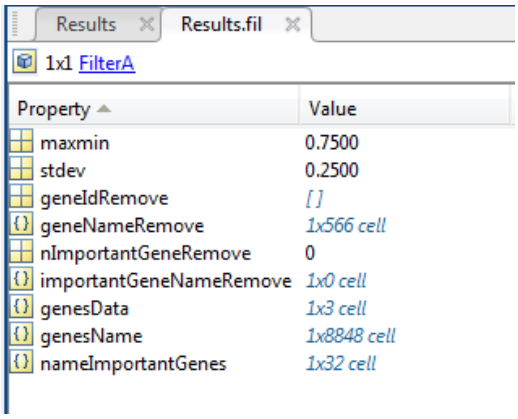


Ilustración 34 Archivo Result.mat resultados del filtro

La estructura *obj,* Ilustración 35*,* guarda el resultado después de la aplicación del algoritmo. Se guardan los parámetros con que se ha ejecutado el algoritmo (*threshlodL* y *threshlodH*), los datos que se han analizado (*dataAnalysis*) que en el caso de que se haya aplicado la media será la media de los datos que se encuentran en *genesData* y en el caso de no aplicarse la media estos datos coincidirán con los de *genesData*, los grupos formados por el algoritmo (*clusters*), el valor de *shapeIndex* para cada grupo (*gindex*), si se ha aplicado la media (*mean*), el número de muestras tomadas por instante de tiempo (*nSampleEachTime*) y el contenido de los ficheros cargados con los instantes de tiempo (*timePoints*) y la salida (output):

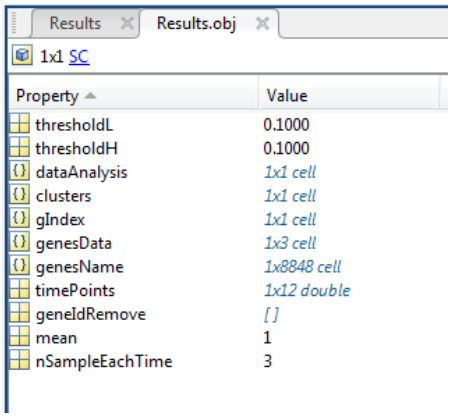


Ilustración 35 Archivo Result.mat resultados del algoritmo

Por último la estructura *Analysis* contiene el resultado de aplicar el análisis sobre los datos obtenidos después de la ejecución del algoritmo. Eta estructura lo que guarda es el tamaño del cluster más grande formado y el número de clusters que contienen un solo gen.

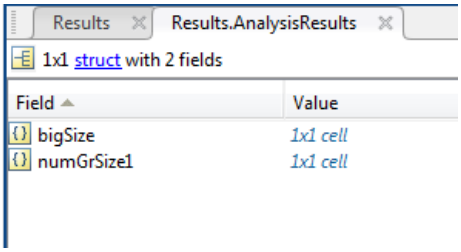


Ilustración 36 Archivo Result.mat resultados del análisis

En la aplicación (dentro de la tabla en el subpanel 'Results') se muestra un resumen de los resultados, donde se indica:

- Identificador de la carpeta donde se han guardado los resultados que coincide con la fecha y hora de cuando se ha ejecutado el algoritmo.
- Algoritmo ejecutado.
- Valores proporcionados a los parámetros del algoritmo.

- Si se ha solicitado aplicar la media o no sobre los datos cargados.
- Número de grupos que ha dado como resultado el algoritmo ejecutado.
- El número de genes que contiene el grupo más grande formado.
- EL número de grupos que contienen un solo gen.
- Visualizador de las gráficas generadas para los grupos con genes importantes.
- Visualizador del Excel generado con los resultados.

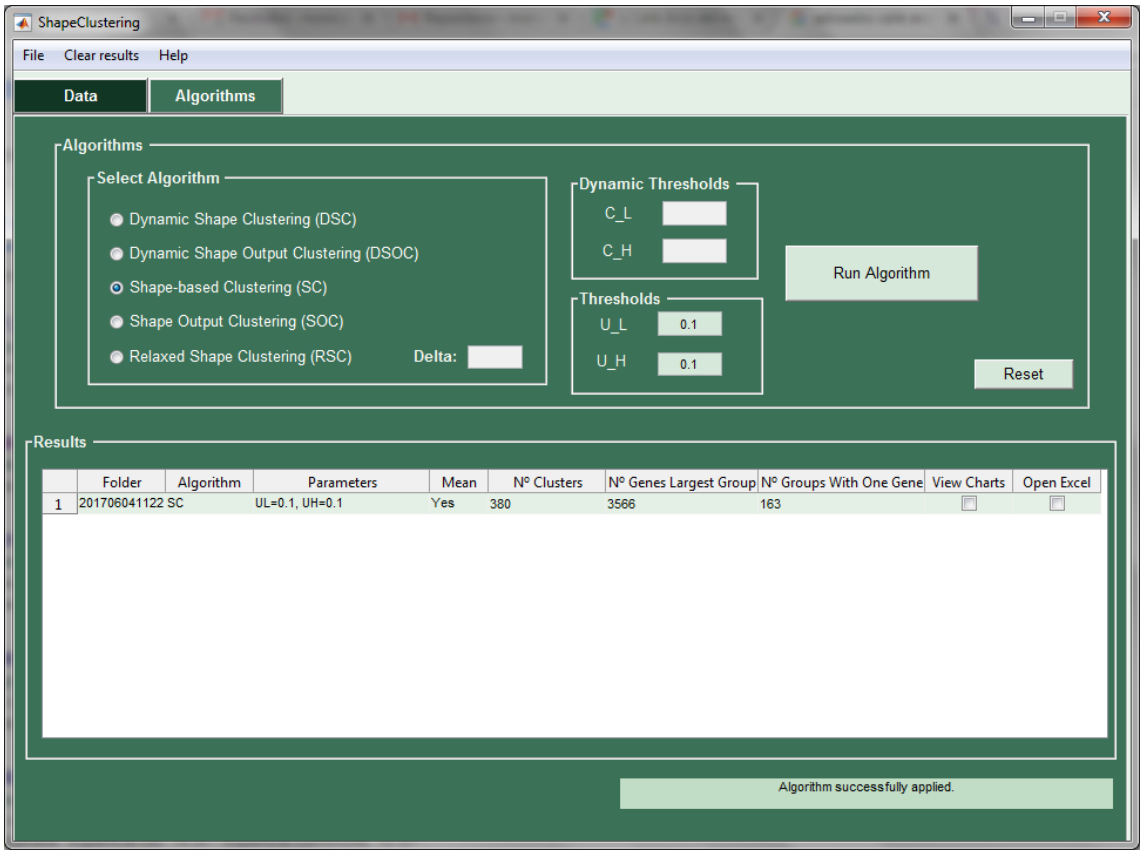


Ilustración 37 Pestaña 'Algorithm' – Resultados para algoritmo SC con parámetros 0.1 y 0.1

A través de la columna 'View Charts' de esta tabla se pueden ver las gráficas generadas para los grupos que contienen genes importantes, en cada una de las ejecuciones. Un ejemplo de estas se muestra en la Ilustración 38:

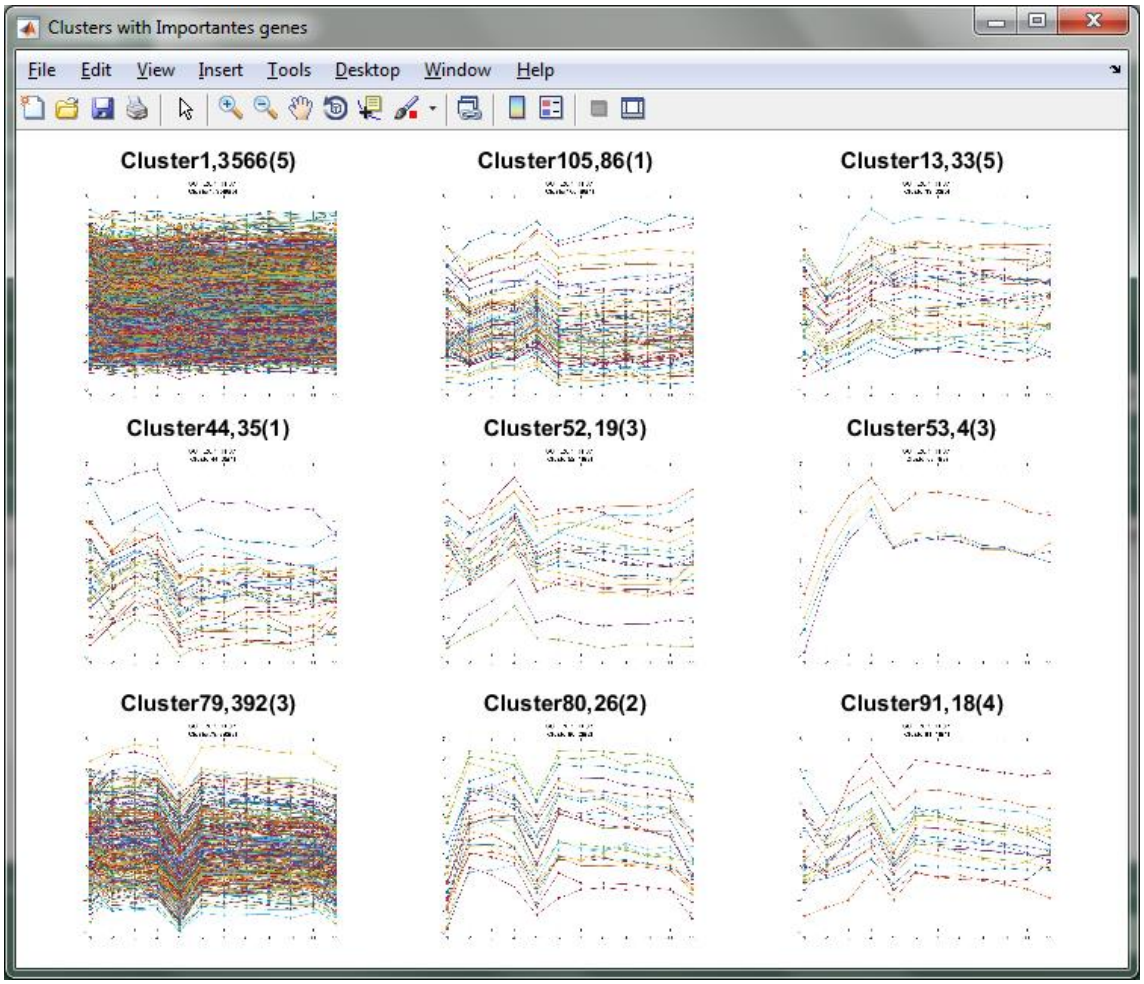
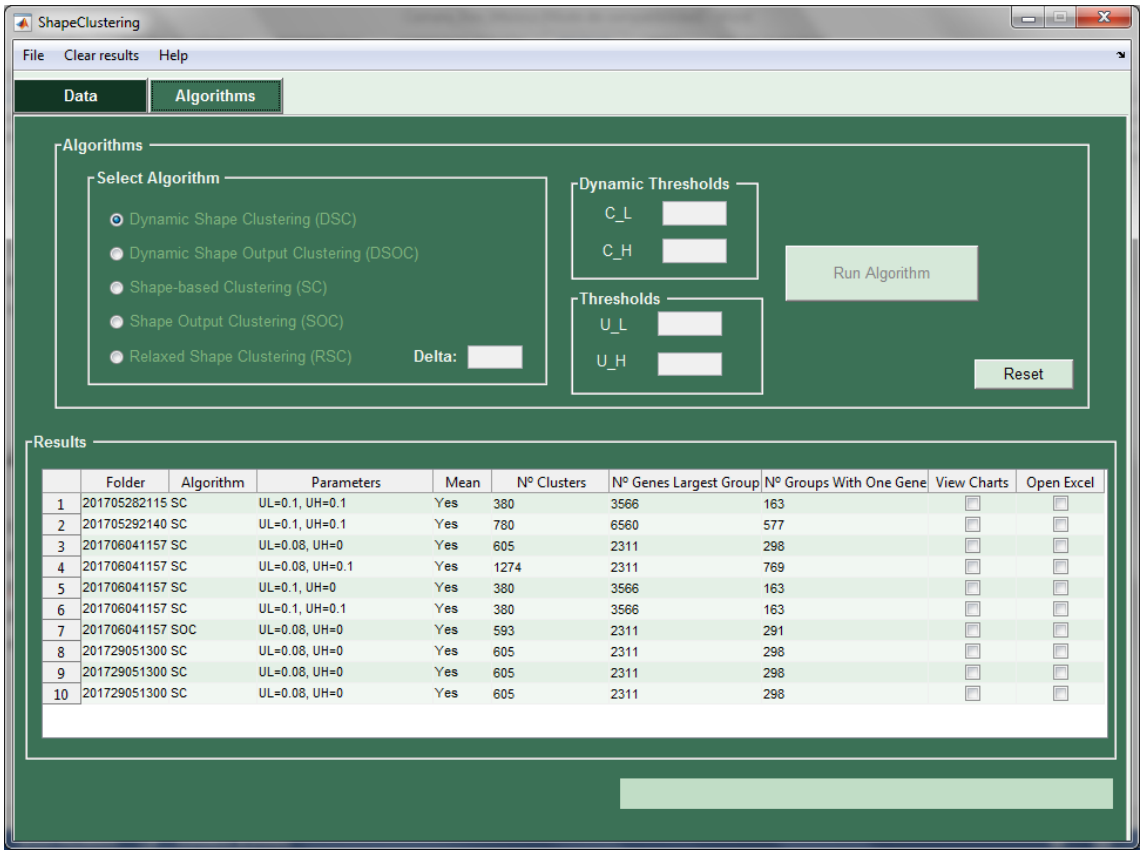


Ilustración 38 Gráficas - Grupos con genes importantes

A través de la columna 'Open Excel' se puede abrir el ficheros .xlsx guardado con los resultados de la ejecución del algoritmo.

Si la carpeta que alberga los resultados ya dispone de resultados de alguna experimentación previa, estos se mostrarán. De esta manera, al abrir la aplicación se visualizan, en la tabla inferior, los últimos diez resultados generados si es que existen:



	Folder	Algorithm	Parameters	Mean	Nº Clusters	Nº Genes Largest Group	Nº Groups With One Gene	View Charts	Open Excel
1	201705282115	SC	UL=0.1, UH=0.1	Yes	380	3566	163		
2	201705292140	SC	UL=0.1, UH=0.1	Yes	780	6560	577		
3	201706041157	SC	UL=0.08, UH=0	Yes	605	2311	298		
4	201706041157	SC	UL=0.08, UH=0.1	Yes	1274	2311	769		
5	201706041157	SC	UL=0.1, UH=0	Yes	380	3566	163		
6	201706041157	SC	UL=0.1, UH=0.1	Yes	380	3566	163		
7	201706041157	SOC	UL=0.08, UH=0	Yes	593	2311	291		
8	201729051300	SC	UL=0.08, UH=0	Yes	605	2311	298		
9	201729051300	SC	UL=0.08, UH=0	Yes	605	2311	298		
10	201729051300	SC	UL=0.08, UH=0	Yes	605	2311	298		

Ilustración 39 Pestaña 'Algorithm' – Cargados por defecto los últimos resultados generados previamente