



Instituto Tecnológico de Castilla y León

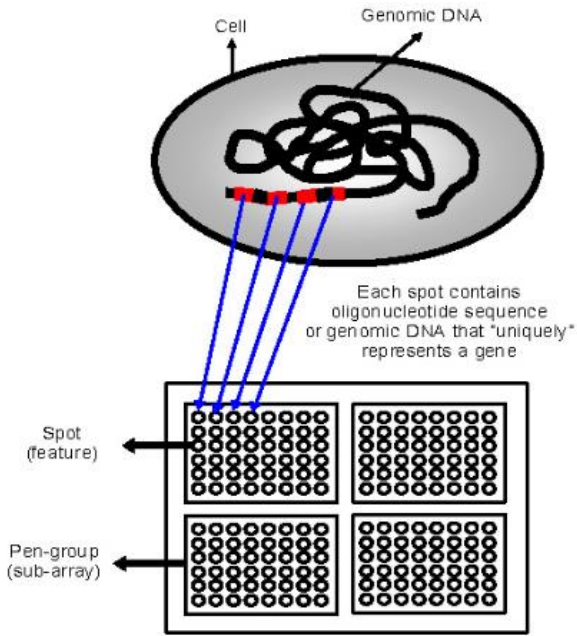
**ALGORITMOS Y TOOLBOX PARA
LA CLUSTERIZACIÓN DE GENES
EN ANÁLISIS DE DATOS DE
MICROARRAYS DE SERIES
TEMPORALES**

**Dr. Javier Sedano
Mónica Cámara**



1. OBJETIVOS
2. MICROARRAYS
3. ALGORITMOS
3. ALGORITMO GENERAL
4. TOOLBOX
5. RESULTADOS
6. PUBLICACIONES RELACIONADAS





- Instantánea de miles de niveles de expresión génica medidos en diferentes momentos y en diferentes condiciones.
- Utilidades:
 - Diagnóstico de enfermedades
 - Guía de tratamientos
 - Medicamentos
 - ...

Microarray. Figure from Grant, Richard P. *Computational Genomics: Theory And Application*. Horizon Bioscience, 2004.

El análisis de datos de microarrays de series temporales requiere técnicas eficientes para extraer y visualizar información genética relevante. Encontrar los genes más importantes en un espacio de entrada de una alta cardinalidad y dimensionalidad es uno de los mayores problemas de la bioinformática y de la biología computacional.



Shape-based Clustering (SC)

Shape Output Clustering (SOC)

Dynamic Shape Clustering (DSC)

Dynamic Shape Output Clustering (DSOC)

Relaxed Shape Clustering (RSC)

- ❖ Diseño e implementación de varios **algoritmos propios de agrupamiento** basados en formas para el análisis de datos específicos de microarrays de series temporales.
- ❖ Algunos de los algoritmos propuestos solamente tienen en cuenta los valores de expresión en el tiempo de cada gen, mientras que otros analizan la correlación entre el cambio de expresión genética y un valor de salida, como puede ser, la producción de un fármaco o el crecimiento celular.
- ❖ Los algoritmos diseñados han sido probados en un conjunto de datos de microarrays reales con una interpretación desde una perspectiva biológica.

SC

- La variante más simple del algoritmo es SC que agrupa directamente los genes sin tener en cuenta la correlación de cada gen con la salida. El objetivo es crear grupos de genes basados exclusivamente en su patrón de expresión en el tiempo.

SOC

- Incorpora conocimiento sobre la forma de los niveles de expresión genética y su correlación con los valores de la salida.

DSC y DSOC

- Son una versión dinámica de los métodos SC y SOC en el sentido de que el punto de referencia utilizado para la determinación de g_class , así como los valores de los umbrales, se ajustan dinámicamente para cada gen basándose en los valores de expresión genética particulares.

RSC

- El algoritmo RSC es una extensión del algoritmo SC relajando las condiciones en que se forman los grupos. En todos los algoritmos anteriores, se permite que dos genes estén en el mismo grupo si tiene exactamente el mismo valor g_index . En RSC se agrupan los genes que tienen un índice perteneciente a un intervalo $[g_index-\delta, g_index+\delta]$.

El **algoritmo general** propuesto está basado en la forma en la que se agrupan los datos de microarrays de series temporales.

Descripción:

1. Calcular ***g_step*** para cada intervalo de tiempo. El valor de *g_step* se utiliza para decidir como de significativo es el cambio en el nivel de expresión genética entre un instante de tiempo y el inmediatamente siguiente:

$$g_step(t_i, t_{i+1}) = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$$

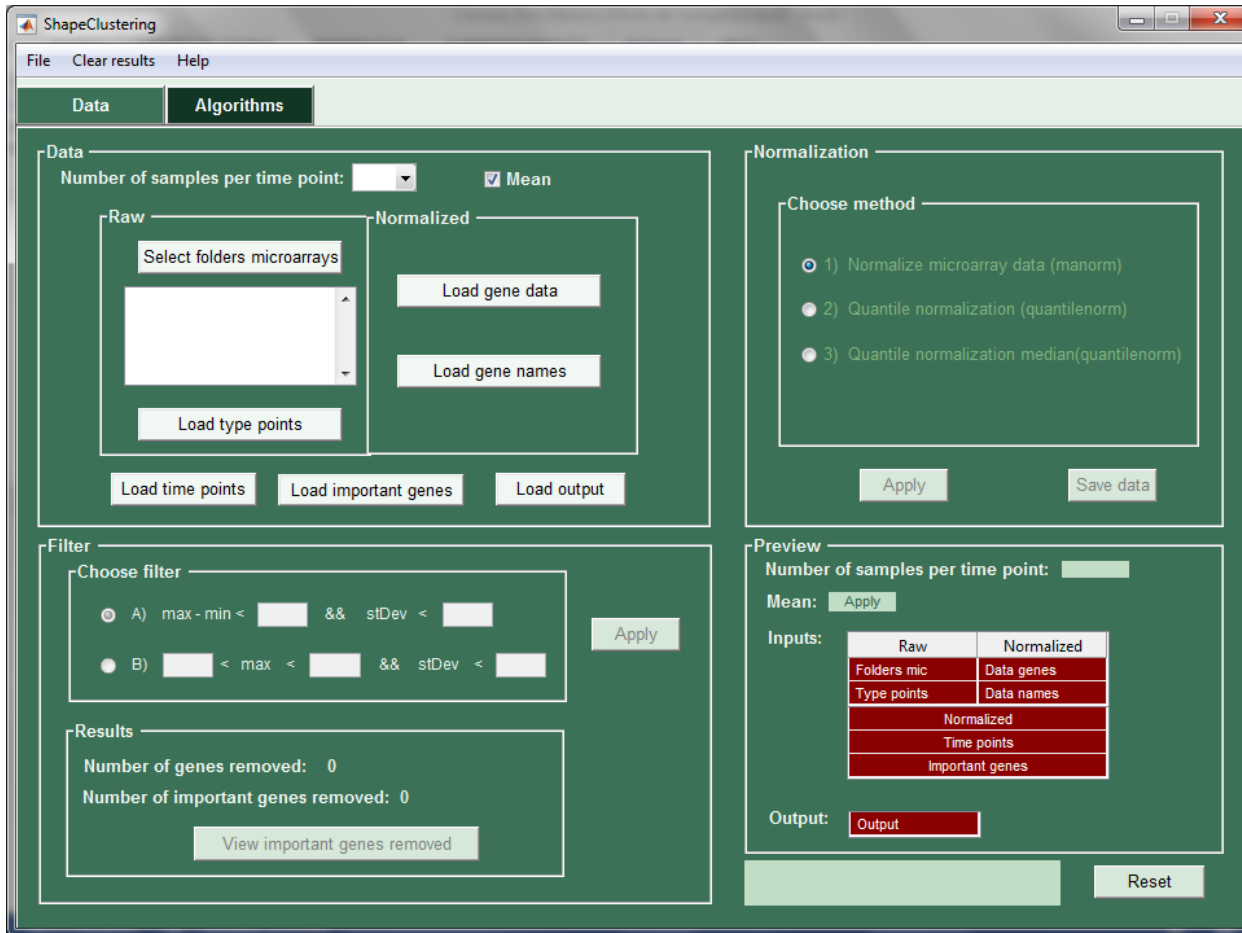
2. Asignar un nivel de ***g_class*** a cada intervalo de tiempo mediante el uso de umbrales. El umbral indica el nivel de diferencia aceptable entre dos niveles de expresión genética consecutivos.

Categorías (l) de *g_class*:

- Si *g_step* es negativo y por debajo del umbral negativo -> **decremento**
- Si *g_step* se encuentra entre los dos umbrales -> **estable**
- Si *g_step* es positivo y por encima del umbral considerado -> **incremento**

3. El array de valores de *g_class* se utiliza para calcular un valor de *g_index* para cada gen. Este valor de ***g_index*** se calcula de manera diferente para cada algoritmo y es utilizado para formar los clusters agrupando los genes que tienen el mismo valor.

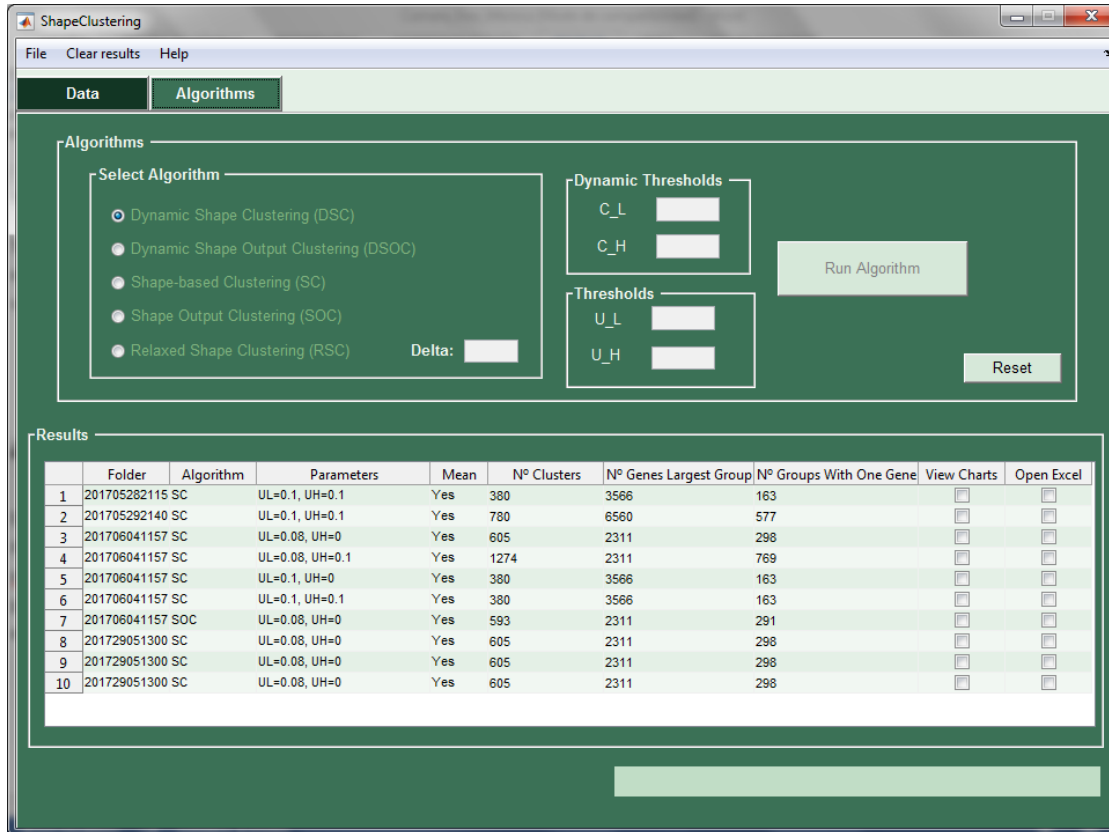
$$g_index^{SC} = \sum_{i=1}^{T-1} l^i * g_class(t_i, t_{i+1}) \quad g_index^{SOC} = \sum_{i=1}^{T-1} l^i * g_class(t_i, t_{i+1}) * y_class(t_i, t_{i+1})$$



Partes:

- 1. Carga de los datos:** permite cargar los datos en bruto o ya normalizados.
- 2. Normalización:** aplicar una normalización a datos en bruto.
- 3. Selección de características:** se permite elegir entre dos filtros para eliminar genes irrelevantes.
- 4. Visualización de los pasos llevados a cabo.**

1. Selección, configuración y ejecución de un algoritmo.
2. Visualización de los resultados.

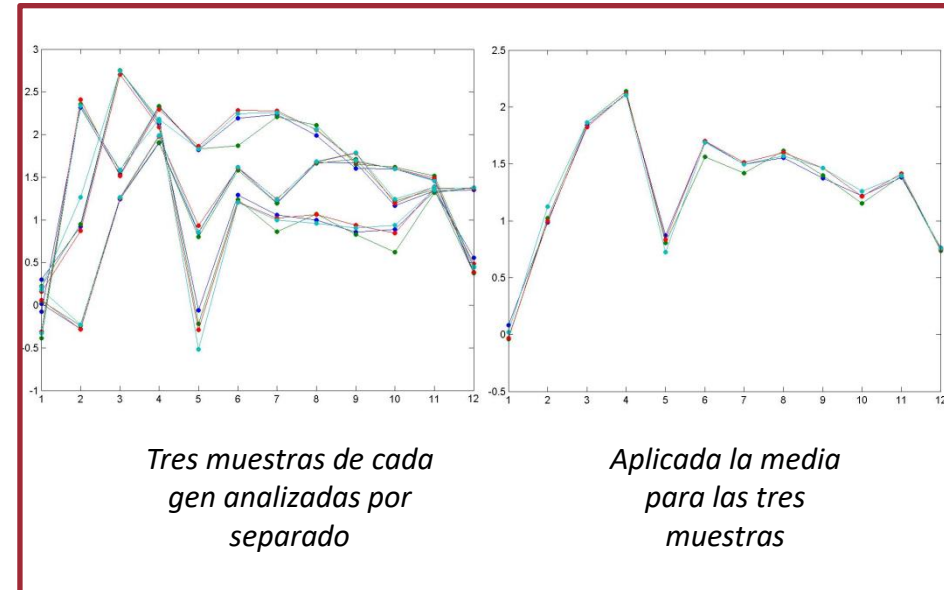
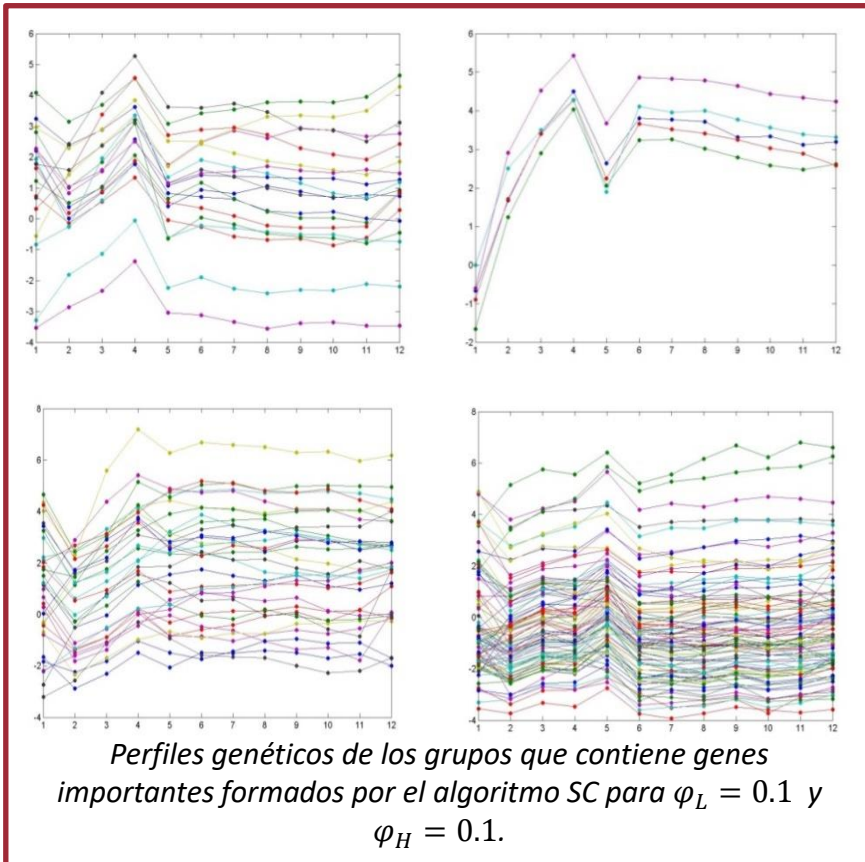


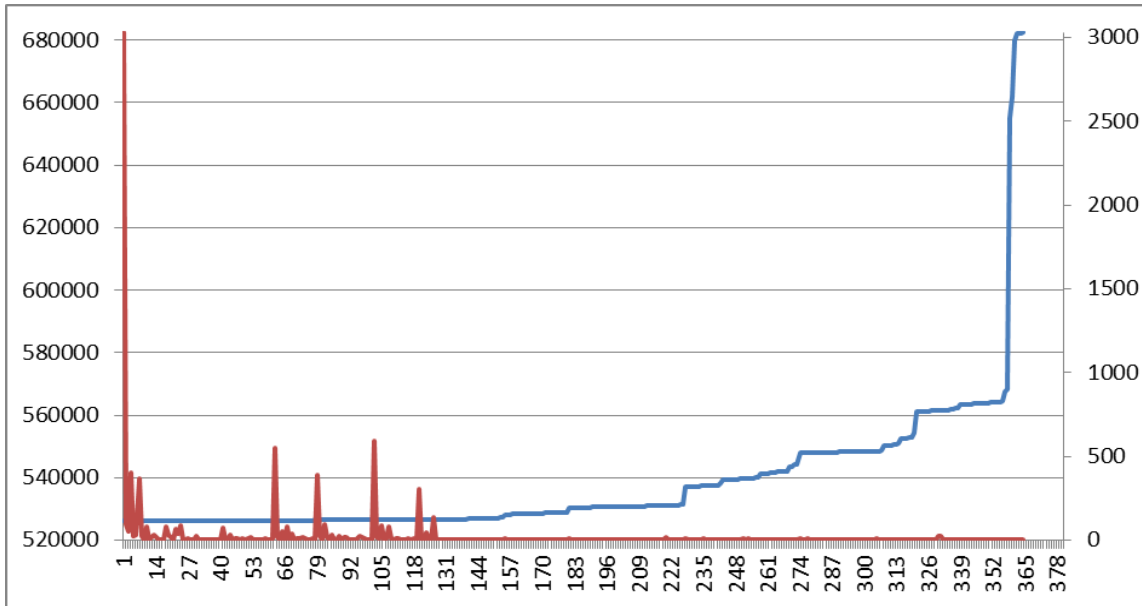
The screenshot shows the ShapeClustering application window. It has a menu bar with 'File', 'Clear results', and 'Help'. Below the menu bar are two tabs: 'Data' and 'Algorithms'. The 'Algorithms' tab is active, showing a 'Select Algorithm' section with five radio buttons: 'Dynamic Shape Clustering (DSC)', 'Dynamic Shape Output Clustering (DSOC)', 'Shape-based Clustering (SC)', 'Shape Output Clustering (SOC)', and 'Relaxed Shape Clustering (RSC)'. A 'Delta:' input field is next to the RSC option. To the right, there are 'Dynamic Thresholds' (C_L, C_H) and 'Thresholds' (U_L, U_H) input fields. A 'Run Algorithm' button is positioned to the right of these fields, and a 'Reset' button is at the bottom right. Below the configuration area is a 'Results' section containing a table with 10 rows of data.

	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	<input type="checkbox"/>	<input type="checkbox"/>
2	201705292140	SC	UL=0.1, UH=0.1	Yes	780	6560	577	<input type="checkbox"/>	<input type="checkbox"/>
3	201706041157	SC	UL=0.08, UH=0	Yes	605	2311	298	<input type="checkbox"/>	<input type="checkbox"/>
4	201706041157	SC	UL=0.08, UH=0.1	Yes	1274	2311	769	<input type="checkbox"/>	<input type="checkbox"/>
5	201706041157	SC	UL=0.1, UH=0	Yes	380	3566	163	<input type="checkbox"/>	<input type="checkbox"/>
6	201706041157	SC	UL=0.1, UH=0.1	Yes	380	3566	163	<input type="checkbox"/>	<input type="checkbox"/>
7	201706041157	SOC	UL=0.08, UH=0	Yes	593	2311	291	<input type="checkbox"/>	<input type="checkbox"/>
8	201729051300	SC	UL=0.08, UH=0	Yes	605	2311	298	<input type="checkbox"/>	<input type="checkbox"/>
9	201729051300	SC	UL=0.08, UH=0	Yes	605	2311	298	<input type="checkbox"/>	<input type="checkbox"/>
10	201729051300	SC	UL=0.08, UH=0	Yes	605	2311	298	<input type="checkbox"/>	<input type="checkbox"/>

* Para cualquier duda con la aplicación contactar con: monica.camara@itcl.es

Clusters de genes, con perfiles de expresión similar, para la optimización de la producción de inmunodepresores (*streptomyces tsukubaensis*), para la obtención de tacrolimús.





Azul: valor de g_index (en este caso hay 380 valores diferentes, correspondiente a 380 clusters) ordenados de menor a mayor.

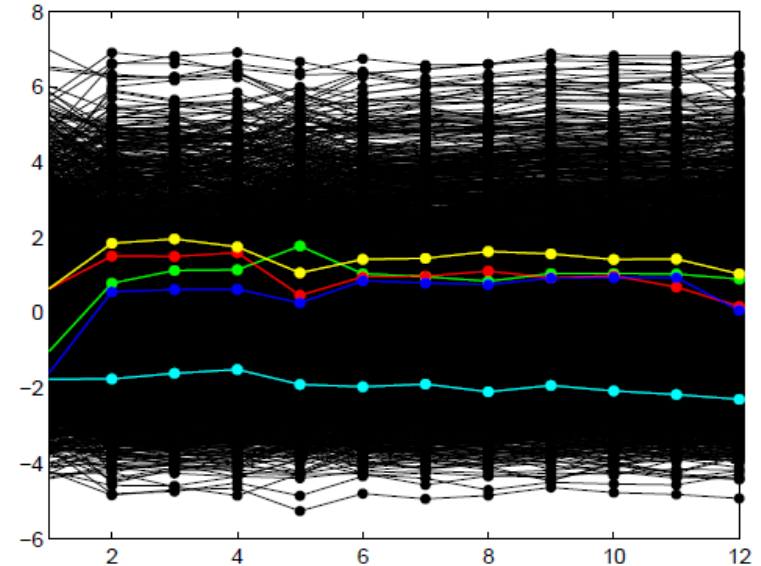
Rojo: número de genes en cada cluster.

Representación del tamaño de los grupos obtenidos mediante el algoritmo SC para $\varphi_L = 0.1$ y $\varphi_H = 0.1$.

La mayoría de los genes están agrupados por un g_index de corto alcance (ya que los grupos más grandes se forman en la primera mitad del intervalo), mientras que para los valores más grandes de g_index el tamaño de los grupos es mucho menor. Sin embargo, estos grupos de menor tamaño se caracterizan por una mayor diversidad entre ellos, ya que la brecha del índice genético entre grupos consecutivos es mucho mayor.

El análisis de los resultados incluye la **identificación de los genes importantes** dentro de los grupos formados para permitir una investigación más profunda de estos grupos. Se sabe que los genes funcionalmente relacionados tienden a tener valores de expresión similares, y por tanto, la posibilidad de obtener grupos con un perfil de expresión común es de gran interés ya que aumenta la importancia biológica.

Para la explicación biológica de los resultados se tuvieron en cuenta 32 genes que participan en el proceso de producción.



El grupo más grande para SC para $\varphi_L = 0.1$ y $\varphi_H = 0.1$ consta de 3029 genes y contiene 5 genes importantes que se pueden ver representados en la figura.

- Barbero I., Chira C., Sedano J., Prieto C. and Villar J.R. et al. (2012) “Merge Method for Shape-Based Clustering in Time Series Microarray Analysis Lecture Notes in Computer Science,” 2012, Volume 7435, Intelligent Data Engineering and Automated Learning – IDEAL 2012, Pages 834-841.
- Camelia Chira, Javier Sedano, José Ramón Villar, Carlos Prieto, Emilio Corchado: Gene Clustering in Time Series Microarray Analysis. SOCO-CISIS-ICEUTE 2013: 289-298.
- Camelia Chira, Javier Sedano, Mónica Cámara, Carlos Prieto, José Ramón Villar, and Emilio Corchado. A cluster merging method for time series microarray with production values. International Journal of Neural System, 24(6), 2014.
- Camelia Chira, Javier Sedano, José Ramón Villar, Monica Camara, Carlos Prieto: Shape-Output Gene Clustering for Time Series Microarrays. SOCO 2015: 241-250.
- Camelia Chira, Javier Sedano, José Ramón Villar, Monica Camara, and Carlos Prieto: Gene clustering for time-series microarray with production outputs. *Soft Comput.*, 20(11):4301–4312, 2016.
- Registro de Software: BU-54-17. Agrupaciones de formas genéticas (shape clustering genes).

Quedamos a su disposición para cualquier duda en:



- Javier Sedano: javier.sedano@itcl.es
- Mónica Cámara: monica.camara@itcl.es



C/ López Bravo 70
P.I. Villalonquejar, 09001 Burgos



947 298 471
947 298 008



info@itcl.es
www.itcl.es