Universitat Autònoma de Barcelona

escola d'enginyeria

2684-1

Identificar los genes que promueven los cambios fenotípicos

Memoria del Proyecto Fin de Carrera

de Igeniería en Informática

realizado por

Carles Hernández Ferrer,

dirigido por

Jordi González Sabaté

y codirigido por

Mario Huerta Casada

Bellaterra, 12 de septiembre de 2011

Universitat Autònoma de Barcelona

escola d'enginyeria

El sotasignat, Jordi González Sabaté,

Professor/a de l'Escola Tècnica Superior d'Enginyeria de la UAB,

CERTIFICA:

Que el treball a què correspon aquesta memòria ha estat realitzat sota la seva direcció per en Carles Hernández Ferrer

I per tal que consti firma la present.

Signat:

Bellaterra, 12 de Setembre de 2011

Universitat Autònoma de Barcelona

escola d'enginyeria

El sotasignat, Mario Huerta Casada,

de l'empresa, Insitut de Biotecnología i Biomedicina de la UAB

CERTIFICA:

Que el treball a què correspon aquesta memòria ha estat realitzat a l'empresa sota la seva supervisió mitjançant conveni amb la Universitat Autònoma de Barcelona.

Així mateix, l'empresa en té coneixement i dona el vist-i-plau al contingut que es detalla en aquesta memòria.

Signat:

Bellaterra, 12 de Setembre de 2011

Índice

1. Introducción

1.1. Motivación personal

Mi proyecto de fin de carrera ha representado una excelente oportunidad para poner a prueba los conocimientos y habilidades que he adquirido durante todo el trascurso de mi formación universitaria. Es importante recalcar que poder realizar el proyecto enmarcado en un ambiente de investigación interdisciplinar que relaciona la informática con la biología y la medicina ha sido el elemento de mayor motivación, junto a poder participar de todo aquello que conlleva trabajar en investigación.

De esta forma el proyecto me ofrecía la oportunidad de participar en un proyecto real y aplicable, además de formarme en unos ámbitos ajenos a la ingeniería informática como son el análisis de expresiones entre genes, el análisis de microarrays, la interacción entre proteínas, las interacciones entre biomoléculas y proteínas, etc.

Poder realizar un proyecto de este calibre ha supuesto una gran fuente de motivación, además de saber que los resultados obtenidos, tanto a nivel de datos resultantes como a nivel de herramientas desarrolladas, van a servir a la comunidad científica que trabaja estudiando la expresión génica.

El Instituto de Biotecnología y Biomedicina (IBB – UAB) [1], donde he desarrollado el proyecto, es un centro puntero en el desarrollo de herramientas para la investigación y análisis de datos en distintos campos de la bioinformática, biología celular, inmunología, microbiología y proteómica. Esto ha propiciado que el proyecto cumpliera todas mis expectativas.

1.2. Estado del arte

La bioinformática es una disciplina científica que utiliza las tecnologías de la información para organizar, analizar y distribuir información biológica con la finalidad de responder preguntas complejas en biología, biotecnología y biomedicina. La bioinformática es, por tanto, un área de investigación multidisciplinar que puede ser definida como la relación entre diversas disciplinas biológicas (como la biología, la biomedicina,...) y las tecnologías que nos permiten aprovechar el potencial computacional actual (algoritmia, paralelización, tecnologías de la información, data mining...).

La bioinformática permite el análisis de forma masiva y automática de una gran cantidad de datos de entrada. Uno de los puntos de entrada de datos experimentales en bioinformática es la tecnología de microarrays. La tecnología de microarrays genera una gran cantidad de datos sobre expresión génica, ya que obtiene los niveles de expresión de una gran cantidad de genes (entre 10^3 y 10^4) para una gran cantidad de condiciones muestrales (entre 10^2 y 10^3).

Como resultado se genera una gran matriz de datos donde cada fila representa un gen, cada columna una condición muestral y cada celda (intersección entre fila y columna) el nivel de expresión de ese gen para esa condición muestral.

Así pues, la información que se obtiene de una microarray depende de las condiciones muestrales aplicadas (falta de oxígeno, falta de agua, aplicación de drogas,...). Si las condiciones muestrales de la microarray son para el estudio del impacto del tabaco en relación al cáncer de pulmón, la microarray nos proporcionará los niveles de expresión de los genes en el tejido pulmonar bajo estas circunstancias. De ahí que las condiciones muestrales nos proporcionen la repuesta génica a distintos fármacos, a diferentes fases de una patología, o a diferentes características de diferentes pacientes, entre otras.

Un procedimiento habitual en el análisis de datos de microarrays es agrupar las condiciones muestrales en clusters. Este proceso estadístico agrupa determinadas condiciones muestrales en un cluster por tener un efecto similar sobre la expresión de los genes y tener un efecto diferente a las condiciones muestrales agrupadas en el resto de clusters. Este proceso se realiza haciendo uso de diferentes métodos de clustering, como pueden ser el Hierarchical Clustering o el Point Accepted Mutation.

Otros métodos para el análisis de microarrays son los métodos basados en la extracción de componentes principales. La detección de componentes principales nos permite obtener los patrones de las relaciones de expresión entre los genes. Este método genera una recta como aproximación de la nube de datos (la nube de datos serían las condiciones muestrales y cada gen a relacionar una de las variables). Esta recta, cuando es obtenida de la relación de expresión entre dos o más genes, pasará a ser el patrón de la relación de expresión entre los genes analizados. En el análisis de relaciones de expresión génica, que el patrón sea una recta es un problema, porque gran parte de las relaciones de expresión no son lineales. Hacer uso del método de extracción de curvas principales de puntos orientados (PCOP) [2][3] soluciona el problema, ya que nos permite estudiar las relaciones no lineales entre expresiones de genes [4]. En resumen, las PCOPs son curvas que representan la relación de expresión entres dos o más genes, incluso cuando esta relación no es lineal.

Un fenotipo son las características observables de un organismo, célula o tejido, como su morfología, sus propiedades bioquímicas, fisiológicas o de comportamiento .

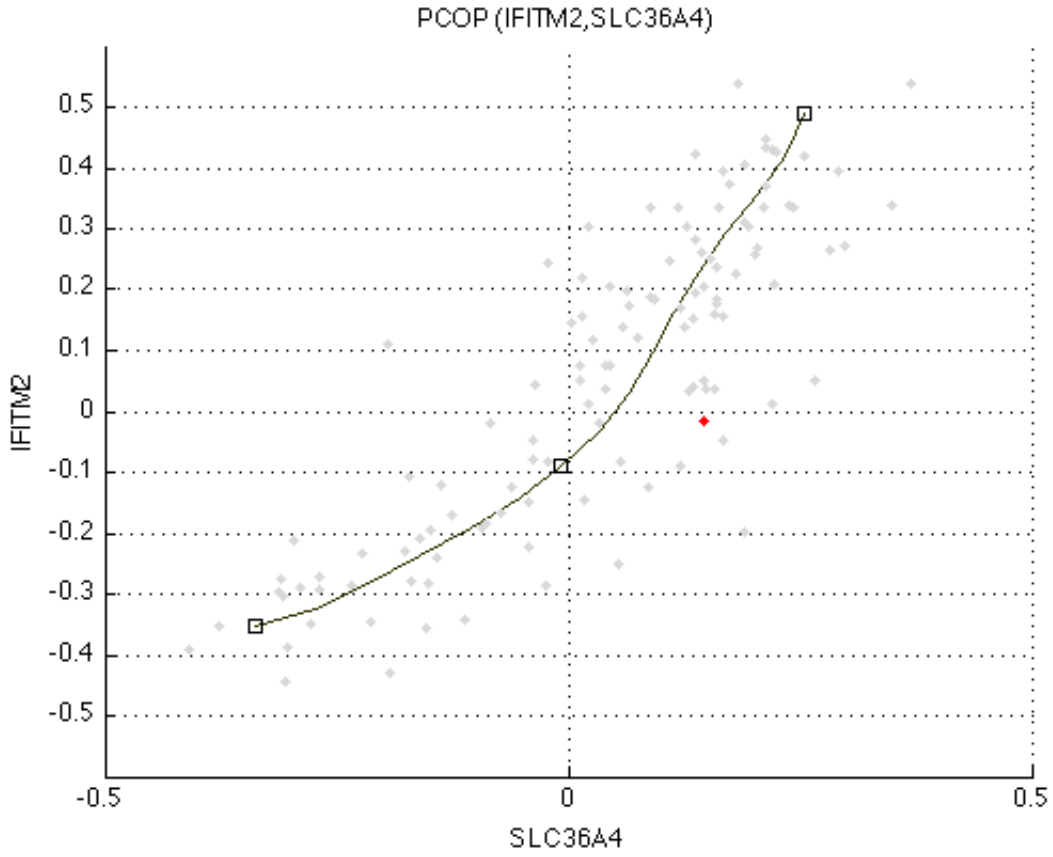


Ilustración 1.1: Los ejes del gráfico representan los niveles de expresión de los genes SLC36A4 (eje X) y IFITM2 (eje Y). Cada uno de los puntos del gráfico, coloreados en gris, corresponde a una condición muestral de la micrarray, mostrándonos el nivel de expresión de cada gen para esa condición muestral. El punto rojo corresponde a la condición muestral de aplicar la droga: 67574-TU-TU-Tu-:-:-:-Vincristine-sulfate (S06757408MDQ), que induce unos valores de expresión: 0,1470 para el gen SLC36A4 y -0,0170 para el gen IFITM2. Mediante el método de extracción de curvas principales(PCOP) a la nube de puntos obtenemos una curva que representa la relación de expresión entre los dos genes. La PCOP aparece dibujada en negro en el gráfico. Los cuadrados negros identifican los puntos de curvatura de la PCOP, dividiendo la curva en dos segmentos.

Un cambio fenotípico en un organismo es un cambio observable provocado por la influencia de factores ambientales o factores internos, como los ritmos circadianos o diferentes procesos secuenciales. El estudio de los cambios fenotípicos en biomedicina es muy importante porque pueden significar el paso de un fenotipo sano a uno patológico o viceversa.

Un ejemplo de diferentes fenotipos para un mismo genotipo lo encontramos en los insectos sociales: Colonias que dividen sus miembros en castas distinguidas (obreros, soldados o reina). Los individuos de las distintas castas son claramente diferentes entre ellos, tanto físicamente como en su comportamiento. No obstante, estas diferencias no son genéticas sino que surgen durante el desarrollo del individuo. Dependen de cómo se tratan los huevos, manipulando factores como la dieta del embrión, la temperatura de incubación, etc. El genoma de cada individuo contiene todas las instrucciones para desarrollarse en cualquiera de las formas, pero los factores ambientales promoverán la expresión de una parte concreta de los genes del programa de desarrollo y a un nivel de expresión concreto. Esto determinará el fenotipo concreto del individuo.

Otro ejemplo de cambio fenotípico lo encontramos en el cáncer de cuello uterino del ser humano, donde la transición epitelio-mesenquimal es el cambio que transforma una célula benigna en maligna. El nuevo citoesqueleto pseudomesenquimal otorga a la célula epitelial original, una vez transformada, las propiedades de migración, invasión y diseminación. Este fenotipo maligno puede ser reversible a través de la corrección de las claves que facilita el microambiente tumoral.

Los métodos de clustering que hemos visto anteriormente nos ayudan a encontrar las condiciones muestrales que representan los diferentes fenotipos contenidos en la microarray. Cada cluster generado es, en definitiva, un fenotipo, ya que todas sus condiciones muestrales están produciendo un efecto similar en la expresión de los genes. Esta expresión concreta de los genes es la que determina el fenotipo que se está dando. Sin embargo, los métodos de clustering, aunque nos sirven para detectar los diferentes fenotipos contenidos en la microarray, no nos proporcionan información sobre los cambios fenotípicos (paso de un fenotipo a otro).

Las relaciones de expresión no lineales (descritas por las PCOPs) nos sirven para detectar cambios fenotípicos ya que si se produce un cambio brusco y repentino de pendiente en la relación de expresión entre dos genes podemos asegurar, sin lugar a dudas, que estamos cambiando a un nuevo fenotipo. El punto de curvatura donde se produce ese cambio brusco de pendiente es el punto en el que se inicia el cambio fenotípico, un cambio fenotípico lo suficientemente fuerte como para cambiar la tendencia de la relación de expresión entre ambos genes (por ejemplo, si la relación entre un gen y otro pasa de ser 1:1 a 1:5, sin lugar a duda, esto nos lleva a un nuevo fenotipo).

De esta forma, los puntos de curvatura nos marcan los cambios fenotípicos y los intervalos entre estos puntos de curvatura nos indican los fenotipos entre los que se transita. Sin embargo, estos cambios fenotípicos descritos en las PCOPs, sólo atañen a los genes estudiados en la relación, es decir, no tienen porque ser unos cambios fenotípicos que afecten al nivel de expresión del total de genes de la microarray.

El Instituto de Biotecnología y de Biomedicina (IBB – UAB) dispone de una línea de investigación para el análisis de microarrays [2][3][4][5][6][7]. Dentro de esta línea de investigación se ha desarrollado un servidor de aplicaciones para el análisis de datos de microarray: http://revolutionresearch.uab.es [8].

El servidor de aplicaciones para el análisis de datos de microarray del IBB dispone de una herramienta para el análisis de los datos de microarray, subidos al servidor, mediante diferentes métodos de clustering especialmente adecuados para el análisis de datos de

microarray. Estos métodos de clustering son:

- Hierarchical Clustering.
- Point Accepted Mutation.
- Self-Organizing Map.
- Self-Organizing Tree Algorithm.

El análisis de los datos de miacroarray a través de estos métodos de clustering nos permite detectar los diferentes fenotipos ocultos en la matriz de expresión génica. Es decir, apreciar qué condiciones muestrales de la microarray representan, realmente, fenotipos diferentes.

Otra herramienta disponible en el servidor de aplicaciones es PhenoSamples-cl, que permite detectar todas las relaciones de expresión no lineales entre pares de genes de la microarray. También clasifica estas relaciones según su tipología e identifica los puntos de curvatura. Todo esto se obtiene mediante el citado cálculo de las PCOPs [2][5].

Como ya hemos visto, los métodos de clustering nos determinan los fenotipos contenidos en una microarray. El problema es determinar las transiciones entre fenotipos, es decir, encontrar si se pasa de un fenotipo a otro, de qué fenotipo a qué fenotipo, y los genes que promueven dichos cambios de fenotipo.

Los puntos de curvatura de las PCOPs determinan la transición entre dos fenotipos pero a nivel local, es decir, afectando a 2 o más genes pero no a todos los genes de la microarray. En otras palabras, analizando solamente la relación entre 2 genes no sabemos si ese cambio fenotípico afectará a la expresión del resto de genes.

Como hemos comentado antes, el estudio de cambios fenotípicos es clave para el avance de la ciencia biomédica y es un reto que se afronta desde todos los frentes posibles, no únicamente desde la bionformática. Sin embargo, hoy en día, tanto en bioinformática como en biomedicina, no existe una herramienta que facilite el estudio en detalle de los cambios fenotípicos. Una de las mejores aproximaciones es la que proporciona el análisis de microarrays, porque sí que nos permite estudiar todos los genes afectados por uno u otro fenotipo mediante los métodos de clustering. Sin embargo, actualmente, la búsqueda de cambios fenotípicos en los datos de microarray se limita a la búsqueda de genes que se sobreexpresan en un fenotipo y se infraexpresan en otro. No existe hoy en día una herramienta que permita relacionar estos fenotipos, identificar los genes responsables de esos cambios y el rol de los genes en la transición.

1.3. Objetivos

Nuestro objetivo es desarrollar una herramientas sólida para la búsqueda, a partir de los datos de microarray, de los genes que promueven los cambios fenotípicos y el rol de dichos genes en la transición.

Para cumplir con el objetivo descrito, se desarrollará un proceso que cruce:

1. Los clusters de condiciones muestrales obtenidos del análisis de los datos de microarray, mediante los métodos de clustering actualmente disponibles en el servidor http://revolutionresearch.uab.es [8].

2. Las PCOPs obtenidas del análisis de las relaciones de expresión entre los genes de la microarray, mediante el aplicativo PhenoSampels-cl del servidor http://revolutionresearch.uab.es [8].

El cruce de ambos tipos de datos nos permitirá detectar aquellos genes que promueven los cambios fenotípicos entre los fenotipos obtenidos por los métodos de clustering. Es decir la transición entre fenotipos que afectan al total de genes de la microarray (o, al menos, al mayor número de genes).

De esta forma, con el proceso de cruce, encontramos aquellas relaciones de expresión que nos permiten conocer no sólo cuándo sucede un cambio fenotípico, sino también los genes involucrados en la transición entre fenotipos y si estos fenotipos son fenotipos globales, es decir, fenotipos que afectan a la expresión del total de genes de la microarray y no sólo a la expresión de los genes relacionados. Esto es posible porque cada uno de los puntos de curvatura de una PCOP se corresponde con un cambio fenotípico, así que, mediante el cruce con los clusters de condiciones muestrales, se podrá detectar si estos clusters de condiciones muestrales quedan ubicados a uno u otro lado del punto de curvatura, es decir, a uno u otro lado del cambio fenotípico.

Para cumplir los objetivos del proyecto se propone el desarrollo de la aplicación web CrossingClusters, que estará, también, albergada en el servidor http://revolutionresearch.uab.es [8]. Este aplicativo web constará de:

- El preproceso encargado de realizar el cruce entre los clusters de condiciones muestrales y las PCOPs.
    - Cruzar Cada distribución de clusters obtenida por un método de clustering con todas la PCOPs.

- Filtrar la aplicación de distribuciones de clusters sobre las PCOP exigiendo un número de clusters para cada intervalo entre dos puntos de curvatura consecutivos. El objetivo es encontrar PCOPs donde realmente se distribuyan los clusters por los diferentes intervalos.

- Agrupar las PCOPs resultantes del filtro anterior según si los mismos clusters aparecen juntos en diferentes intervalos de diferentes PCOPs, si además aparecen en el mismo orden en el intervalo, y si además mantienen el mismo grado de intersección entre ellos.

- Generar los ficheros de salida con un formato fácil de parsear por la aplicación web.

- Programa para generar el gráfico de una PCOP sobre la que se aplica una distribución de clusters. Este programa debe:

  - Generar un gráfico cuya estructura sea similar a la de un diagrama Box-Plot que contenga:

    - Los puntos de curvatura de la PCOP.

    - El punto de mayor densidad de muestras de la PCOP.

    - La distribución de los clusters a lo largo de la PCOP.

    - El punto de mayor densidad de muestras de cada uno de los clusters representados, variando su intensidad de color según la cantidad de muestras del cluster.

  - Ser un programa independiente del proceso de cruce de distribuciones de clusters con PCOPs con el objetivo de facilitar su re-utilización.

  - No hacer uso de ficheros de entrada sino usar argumentos de programa para que se puedan realizar múltiples llamadas simultaneas al programa.

- Aplicación web para visualizar los resultados del cruce entre los clusters de condiciones muestrales y las PCOPs. Esta aplicación debe:

  - Tener una interfaz general que debe:

    - Permitir comparar de forma visual los clusters de condiciones muestrales que los diferentes métodos de clustering han obtenido.

- Permitir escoger qué método de clustering deseamos estudiar para comparar las distribuciones de clusters obtenidas para los diferentes parámetros de entrada del método.

- Permitir comparar una selección de las mejores distribuciones de clusters proporcionadas por los diferentes métodos de clustering. Las mejores distribuciones de clusters serán aquellas que posean un mayor grado de integridad.

- Tener una interfaz de detalle por cada distribución de clusters que debe:

  - Listar las relaciones de expresión que nos mostrarán los genes que han llevado a cabo una transición entre los fenotipos descritos por la distribución de clusters y cómo. La tipología de la PCOP nos mostrará el rol del cada uno de los genes en dicha transición.

  - Mostrar estas relaciones entre genes agrupadas por participar en la transición entre unos mismos fenotipos. Es decir, aquellas PCOPs en las que diferentes clusters de la distribución aparecen juntos en un mismo intervalo entre dos puntos de curvatura, incluso en el mismo orden dentro del intervalo, incluso con el mismo grado de intersección con los clusters vecinos. Estas PCOPs estarán describiendo la misma transición entre fenotipos porque son los mismos clusters los que aparecen a un lado u otro del punto de curvatura.

  - Permitir visualizar la información de cada una de las PCOPs.

  - Mostrar en el listado el gráfico que ilustra la PCOP sobre la que se ha aplicado una distribución de clusters.

1.4. Organización de la memoria

Para alcanzar los objetivos descritos anteriormente se ha seguido la planificación que se expone a continuación.

El trabajo realizado ha tenido una duración de 11 meses. Se inició en octubre de 2010 y finalizó en agosto de 2011. Este trabajo se divide en tres procesos bien diferenciados: adquisición de conocimientos, creación del preproceso y creación de la aplicación web.

En la siguiente sección de la memoria, expondremos los Fundamentos teóricos

necesarios para comprender los conceptos con los que hemos trabajado a lo largo del proyecto.

Acto seguido detallaremos las fases que forman el diseño y la implementación del preproceso y la aplicación web. Primero nos centraremos en el algoritmo diseñado para realizar el cruce de datos expuesto en la sección objetivos y los filtros que se aplican a los resultados. En segundo termino trataremos la aplicación web desarrollada, explicando cada uno de sus elementos.

Durante cada fase del trabajo se han realizado modificaciones y adaptaciones para mejorar la funcionalidad, operatividad y usablidad del conjunto de aplicaciones bajo la supervisión y asesoramiento del codirector del proyecto Sr. Mario Huerta.

La estructura seguida por la memoria y su contenido es el siguiente:

- Fundamentos teóricos: Donde exponemos los conocimientos biológicos y técnicos necesarios para comprender el proyecto.
- Fases: Donde se describe el trabajo desarrollado para alcanzar los objetivos propuestos, los problemas encontrados y las soluciones adoptadas.
- Análisis de resultados: Donde comentamos los resultados obtenidos y justificamos las decisiones tomadas.
- Informe técnico: Donde se describen los programas implementados y la estructura de directorios del servidor web.
- Conclusiones.
- Bibliografía.
- Resumen.

2. Fundamentos teóricos

2.1. Bioinformática

La bioinformática es un área de investigación multidisciplinar donde se reúnen las dos ciencias que le dan nombre: Biología y Computación. El objetivo último de esta disciplina científica es mejorar la condición y calidad de la vida humana a través de la investigación genómica.

Entre sus funciones, la bioinformática, comprende la solución de problemas complejos haciendo uso de herramientas de sistemas y herramientas de computación. Además de gestionar, organizar y almacenar la gran cantidad de información biológica existente.

Dentro de la bioinformática existen otras subdisciplinas importantes:

- Desarrollo e implementación de herramientas que permitan el acceso, uso y manejo de diversos tipos de información.

- Desarrollo de nuevos algoritmos (matemáticos y estadísticos) con los que se puedan relacionar partes de un enorme conjunto de datos, como predecir estructuras o funciones de proteínas.

No obstante, el crecimiento exponencial de la biología molecular hace que la bioinformática no sea un caso más de la aplicación de las ciencias de la computación ya que el número y la complejidad de los datos con los que debe tratar aumenta a un ritmo más alto que la capacidad de procesamiento de los ordenadores actuales.

Junto a las técnicas de la bitotecnologia y la biomedicina, la bioinformática es capaz de identificar los genes que intervienen en una patología y las proteínas que intervienen en los mecanismos de resistencia a antibióticos. Esto es posible al reto fundamental de la bioinformática, que es crear, adaptar y utilizar técnicas computacionales existentes para poder extraer información útil de los datos bioémdicos y biológicos (*data mining*).

2.2. Microarray

En el estudio de la expresión génica se trabaja haciendo uso de conjuntos de datos, como son las microarrays, que contienen valores observables de k características para n individuos. Es lo que llamamos datos multivariantes.

La tecnología de microarrays se basa en someter una gran cantidad de genes a muchas condiciones. Para cada una de estas condiciones, a las que llamaremos condiciones muestrales, obtendremos el nivel de expresión de cada gen. Es decir, en la tecnología de microarrays, cada

una de las *k* características va a corresponder a cada una de las condiciones muestrales a las que se someten cada uno de los *n* genes (individuos) [Ilustración 2.1].

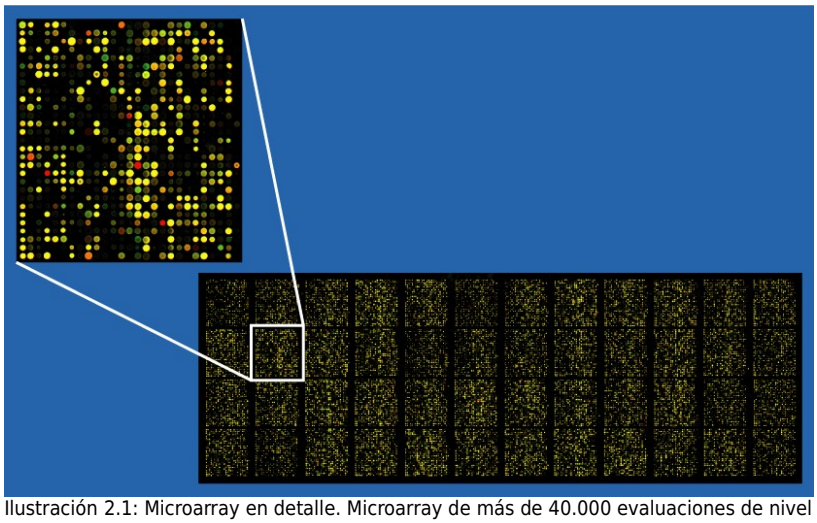


Ilustración 2.1: Microarray en detalle. Microarray de más de 40.000 evaluaciones de nivel de expresión génica (166 genes evaluados en 348 condiciones muestrales). La luminiscencia de cada uno de los puntos nos indica el nivel de expresión de un gen en una condición muestral.

La tecnología actual de microarrays nos permite evaluar entre 10^3 y 10^4 genes bajo un número de condiciones muestrales que se encuentra entre 10^2 y 10^3. En el caso de la microarray 17 del servidor del IBB, se contemplan 1416 genes (individuos) sometidos a 118 condiciones muestrales (características). Es decir, obtenemos un total de 167.088 evaluaciones de nivel de expresión génica.

Los materiales usados para construir las superficies de las microarrays son muy variados (vidrio, silicio,...). Sobre esta superficie se depositan los componentes usados para forzar la expresión del gen dada una condición muestral. A través de técnicas de análisis de imagen se evalúa la microarray, generando una matriz de valores numéricos. Cada celda de esta matriz corresponde a cada uno de los genes sometidos a cada una de las condiciones muestrales.

2.3. Métodos de clustering

Los métodos de clustering se basan en métodos estadísticos para agrupar individuos en clusters de forma que los clusters resultantes sean homogéneos o distantes entre sí.

Aplicando los métodos de clustering sobre los datos obtenidos por la tecnología de microarray obtenemos, en vez del nivel de expresión individual de cada gen en cada condición muestral, aquellos clusters de condiciones muestrales que entendemos provocan el mismo

nivel de expresión sobre el conjunto de genes (pero diferente al nivel de expresión que representan el resto de clusters).

Además de los métodos de clustering estadísticos también existen métodos de clustering que se basan en criterios biomédicos para generar los clusters en lo que se agrupan los individuos.

2.4. Principal Curves of Oriented Points (PCOPs)

Las PCOPs son curvas que nos permiten estudiar relaciones no lineales entre dos o más variables. La PCOP se define generalizando a nivel local las propiedades de la varianza en el cálculo de las componentes principales. Esta generalización de la varianza de las componentes principales realizada a nivel local permite generar los Principal Oriented Points (POPs). Cada POP se corresponde con un análisis a nivel local del espacio muestral por componentes principales. Estos POPs generados son lo que constituyen la PCOP o patrón interno de la nube de muestras que describe la relación entre las variables, en nuestro caso genes. El primer componente de cada POP será el vector director de la curva y el segundo componente será ortogonal a la curva. Sobre este segundo componente se cumplirá la propiedad antes citada: Al proyectar el subespacio muestral, sobre el que se han calculado los componentes principales, sobre este segundo componente la varianza será mínima.

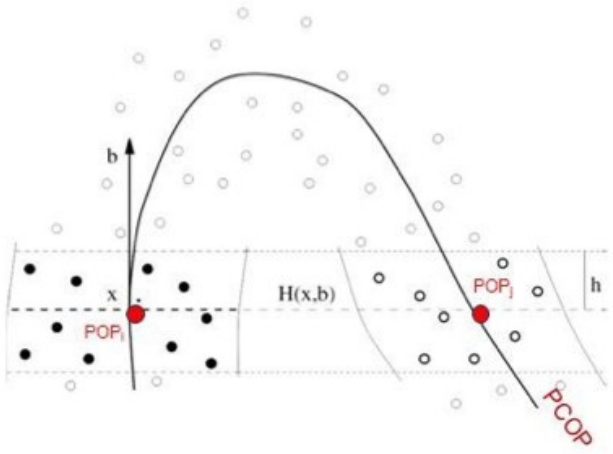


Ilustración 2.2: En la imagen [2] vemos la nube de muestras que forman el espacio muestral, además de la PCOP (línea negra) y dos de los POPs que la forman (puntos rojos). Para cada uno de los POPs vemos el subespacio de muestras que representa: en negro las muestras que generan el POPi y en blanco las que generan el POPj.

3. Fases

3.1. Planificación

A continuación vemos el diagrama de Gantt con la planificación inicial del proyecto:

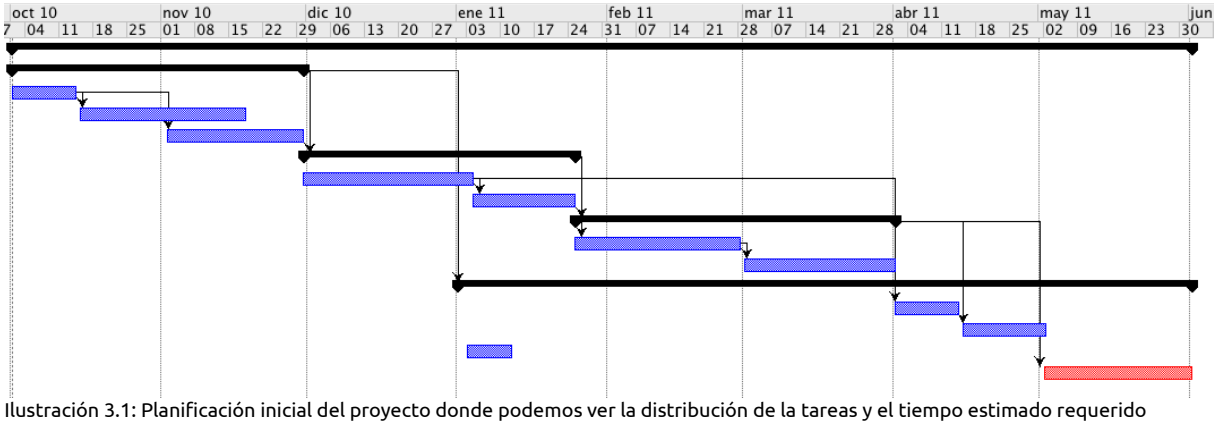


Ilustración 3.1: Planificación inicial del proyecto donde podemos ver la distribución de la tareas y el tiempo estimado requerido

Detallamos, en la siguiente tabla, el concepto de cada tarea, su durada, fecha de inicio y fecha estimada de finalización.

Nombre	Duración	Inicio	Final
Fase 1	61 días	01/10/2010	30/11/2010
Conocimientos	14 días	01/10/2010	14/10/2010
PCOP	35 días	15/10/2010	18/11/2010
Métodos de clustering	29 días	02/11/2010	30/11/2010
Fase 2	57 días	30/11/2010	25/01/2011
Cruce de Datos	35 días	30/11/2010	04/01/2011
Generación de Imágenes	22 días	04/01/2011	25/01/2011
Fase 3	67 días	25/01/2011	01/04/2011
Aplicación web	35 días	25/01/2011	28/02/2011
Pruebas y optimización	32 días	01/03/2011	01/04/2011
Fase 4	152 días	01/01/2011	01/06/2011
Informe Previo	12 días	01/01/2011	12/01/2011
Memoria	32 días	01/05/2011	01/06/2011

Ahora exponemos el diagrama de Gantt con la planificación real del proyecto:

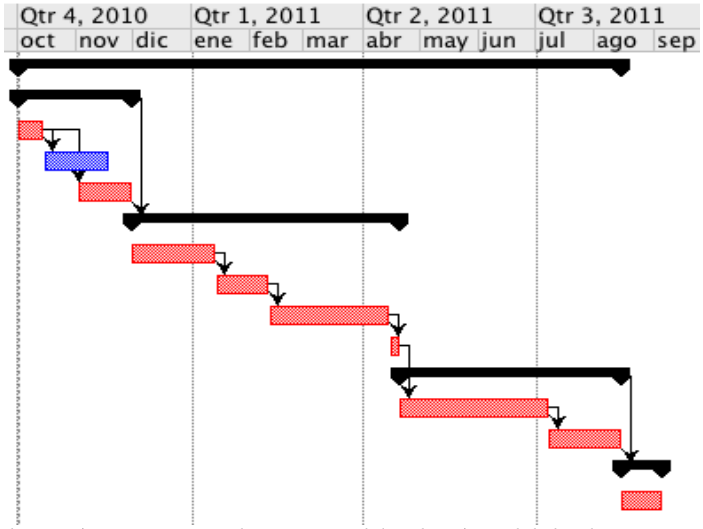


Ilustración 3.2: Diagrama de Gantt con al distribución real de las distintas tareas del proyecto.

A continuación detallamos el concepto de cada tarea, su durada, fecha de inicio y fecha estimada de finalización.

Nombre	Duración	Inicio	Final
Fase 1	61 días	01/10/2010	30/11/2010
Conocimientos	14 días	01/10/2010	14/10/2010
PCOP	35 días	15/10/2010	18/11/2010
Métodos de clustering	29 días	02/11/2010	30/11/2010
Fase 2	141 días	30/11/2010	20/04/2011
Cruce Datos	49 días	30/11/2010	18/01/2011
Generación de Imágenes en phyton	27 días	19/01/2011	15/02/2011
Generación de Imágenes en C	58 días	15/02/2011	14/04/2011
Lanzadora	5 días	15/04/2011	20/04/2011
Fase 3	117 días	20/04/2011	15/08/2011
Aplicación Web	78 días	20/04/2011	07/07/2011
Pruebas y optimización.	39 días	08/07/2011	15/08/2011
Fase 4	217 días	01/01/2011	05/09/2011
Informe Previo	12 días	01/01/2011	12/01/2011
Memoria	20 días	15/08/2011	05/09/2011

En la Fase 1 encontramos las etapas *Conocimientos*, *PCOP* y *Métodos de clustering*. Se

trata de tres etapas diferenciadas de adquisición de conocimientos: Conocimiento general sobre biología, nociones acerca de las PCOPs y nociones de métodos de clustering.

La Fase 2 está compuesta por las etapas *Cruce de Datos*, *Generación de Imágenes en python*, *Generación de Imágenes en C* y *Lanzadora*. En comparación a la planificación inicial, podemos ver que se han añadido las etapas *Generación de Imágenes en C* y *Lanzadora*. La etapa *Cruce de Datos* corresponde al tiempo dedicado a implementar el preproceso para el cruce entre los clusters de condiciones muestrales y las PCOPs. Las etapas *Generación de Imágenes en python* y *Generación de Imágenes en C* corresponden a la implementación del programa para generar el gráfico de una PCOP sobre la que se aplica una distribución de clusters. La etapa *Lanzadora* corresponde a la modificación del programa encargado de ejecutar todos los procesos de cálculo y agrupación de cada microarray para integrar el preproceso generado. Esta etapa estaba incluida, originalmente, dentro de *Cruce Datos* pero se decidió terminar primero con la etapa *Generación de Imágenes en python*, que derivó en una segunda etapa (*Generación de Imágenes en C*).

La etapa *Generación de Imágenes en C* apareció debido a un problema con la ejecución de código python en los servidores del IBB. Este problema fue resuelto con éxito realizando un nuevo programa en código C/C++ con las mismas funcionalidades que el programa original en python.

La Fase 3, compuesta por *Aplicación Web* y *Pruebas y optimización*, es la fase donde se desarrolló y testeó el aplicativo web. En la planificación inicial se planteó usar la interfaz web del aplicativo PhenoSamples-cl ya que es muy útil para presentar los datos que se generan con los análisis desarrollados en el proyecto. Sin embargo, la interfaz web de PhenoSamples-cl no está preparada para tratar datos con un gran número de clusters, así que se tuvo que reprogramar gran parte de la interfaz.

3.2. Fase 1: Adquisición de conocimientos

Para realizar correctamente el trabajo que conlleva el proyecto es necesario aprender los conceptos biológicos fundamentales para poder entender la plenitud del proyecto que se ha escogido.

Los conceptos adquiridos han sido referentes al análisis de microarrays, marco en el que se encuentra el proyecto y la herramienta implementada.

Tal y como hemos visto, los métodos de clustering son unos métodos estadísticos que nos permiten analizar las microarrays agrupando aquellas condiciones muestrales que producen un efecto similar sobre la expresión de un grupo de genes y que tienen un efecto diferente a las condiciones muestrales agrupadas en el resto de clusters. Es decir, que nos permite detectar los diferentes fenotipos escondidos en la matriz de expresión génica.

Las PCOPs, que nos describen las relaciones de expresión no lineales entre genes, nos sirven para encontrar cambios fenotípicos porque cuando los genes cambian súbitamente la pendiente de su relación de expresión es señal inequívoca de un cambio entre fenotipos. Así pues, los puntos de curvatura de las PCOPs determinan las transiciones entre dos fenotipos pero lo hacen a nivel local (sólo afectando a 2 o más genes, pero no a la totalidad de genes de la microarray).

3.3. Fase 2: Cruce de datos entre clusters de condiciones muestrales y PCOPs

3.3.1. Proceso para la realización del cruce de datos

El cruce de datos entre las distribuciones de clusters generadas por los métodos de clustering y las PCOPs genera los listados de las PCOPs que superan los filtros que determinan si realmente describen una transición entre los clusters de la distribución de clusters. Las PCOPs del listado aparecerán agrupadas en una agrupación jerárquica según si describen la transición entre unos mismos clusters y de una misma forma o bien transiciones diferentes entre clusters diferentes de la misma distribución de clusters.

Para ello se sigue el siguiente algoritmo, expresado en el siguiente diagrama de flujo [Ilustración 3.3] que se encuentra en la siguiente página.

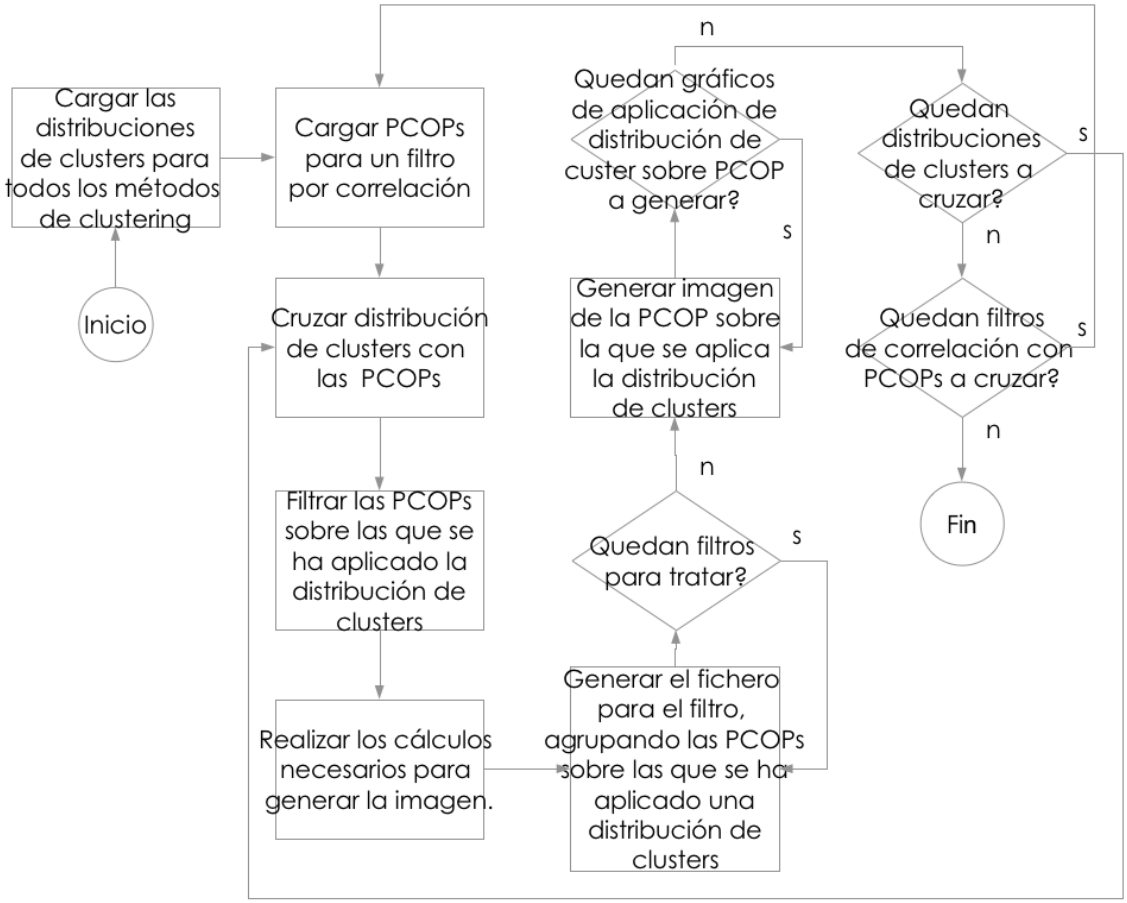


Ilustración 3.3: En el proceso *Cargar las distribuciones de clusters para todos los métodos de clustering* cargamos, en memoria, todas las distribuciones de clusters que aplicaremos sobre las PCOPs. En el siguiente proceso *Cargar PCOPs para un filtro por correlación* cargamos las PCOPs para un filtro por correlación que serán cruzadas con las distribuciones de clusters. El proceso *Cruzar distribución de clusters con las PCOPs* cruza una distribución de clusters con todas las PCOPs cargadas. *Filtrar las PCOPs sobre las que se ha aplicado la distribución de clusters* es la parte del proceso que descarta aquellas PCOPs sobre las que se ha aplicado la distribución de clusters que no son útiles exigiendo un número de clusters para cada intervalo entre dos puntos de curvatura consecutivos. A continuación se procede a obtener la información necesaria para generar la imagen que las representa en *Realizar los cálculos necesarios para generar la imagen*. Se *Genera el fichero para el filtro, agrupando las PCOP sobre las que se ha aplicado una distribución de clusters* según si los mismos clusters aparecen juntos en diferentes intervalos de diferentes PCOPs, si además aparecen en el mismo orden en el intervalo, y si además mantienen el mismo grado de intersección entre ellos. Para aquellas PCOPs sobre las que se ha aplicado la distribución de clusters que han sido agrupadas se procede a generar el gráfico que representa la PCOP sobre la que se ha aplicado una distribución de clusters en *Generar gráfico de la aplicación de la distribución de clusters sobre la PCOP*.

El algoritmo está pensado para facilitar su posterior paralelización y por ello empieza cargando los datos correspondientes a las distribuciones de clusters, en el proceso llamado *Cargar las distribuciones de clusters para todos los métodos de clustering* [Ilustración 3.3].

Las distribuciones de clusters, generadas a través de la ejecución de los métodos de clustering, las encontramos clasificadas en dos tipos. Las que han dado grandes niveles de integridad y las que no. Las distribuciones de clusters con alta integridad son aquellas que se consideran óptimas para diferenciar fenotipos, es decir, que detectan aquellos clusters más distanciados en los datos de microarray. En el proceso *Cargar las distribuciones de clusters para todos los métodos de clustering* [Ilustración 3.3] cargamos todas las distribuciones de clusters, independientemente que las consideremos de alta integridad o baja integridad.

Con las distribuciones de clusters cargadas, el siguiente paso es cargar las PCOPs que se van a usar en el proceso de cruce. Las PCOPs se encuentran clasificadas por correlación. Así los procesos que siguen repetirán su ejecución tantas veces como clasificaciones de PCOPs por correlación dispongamos. El proceso de carga de PCOPs se lleva a cabo en el proceso *Cargar PCOPs para un filtro por correlación* [Ilustración 3.3].

El proceso de cruce se realiza en: *Cruzar distribución de clusters con las PCOPs* [Ilustración 3.3]. En este proceso se toma una única distribución de clusters y se cruza con todas las PCOPs cargadas para un filtro por correlación concreto. Es decir, que el proceso que comprende desde el proceso *Cruzar distribución de clusters con las PCOPs* hasta el proceso *Generar imagen de la PCOP sobre la que se aplica la distribución de clusters* [Ilustración 3.3] se realiza una vez para cada una de las distribuciones de clusters con todas las PCOPs de una correlación concreta.

El proceso de aplicar la distribución de clusters sobre la PCOP esta formado por dos etapas:

1. Generar la tabla con la distribución de los clusters a lo largo de la PCOP. Cada fila de la tabla es un cluster de la distribución de clusters y cada columna un POP de la PCOP (tal y cómo se ha explicado detalladamente en Fundamentos teóricos, los POPs son los puntos que forman la PCOP). En cada una de las celdas de la tabla encontramos el número de muestras del cluster (fila) que tienen como POP más cercano, en el espacio de 2 dimensiones, a determinado POP (columna) [Ilustración 3.4]. Con esta tabla podremos conocer cómo se distribuyen los clusters de la distribución de clusters sobre una PCOP concreta.

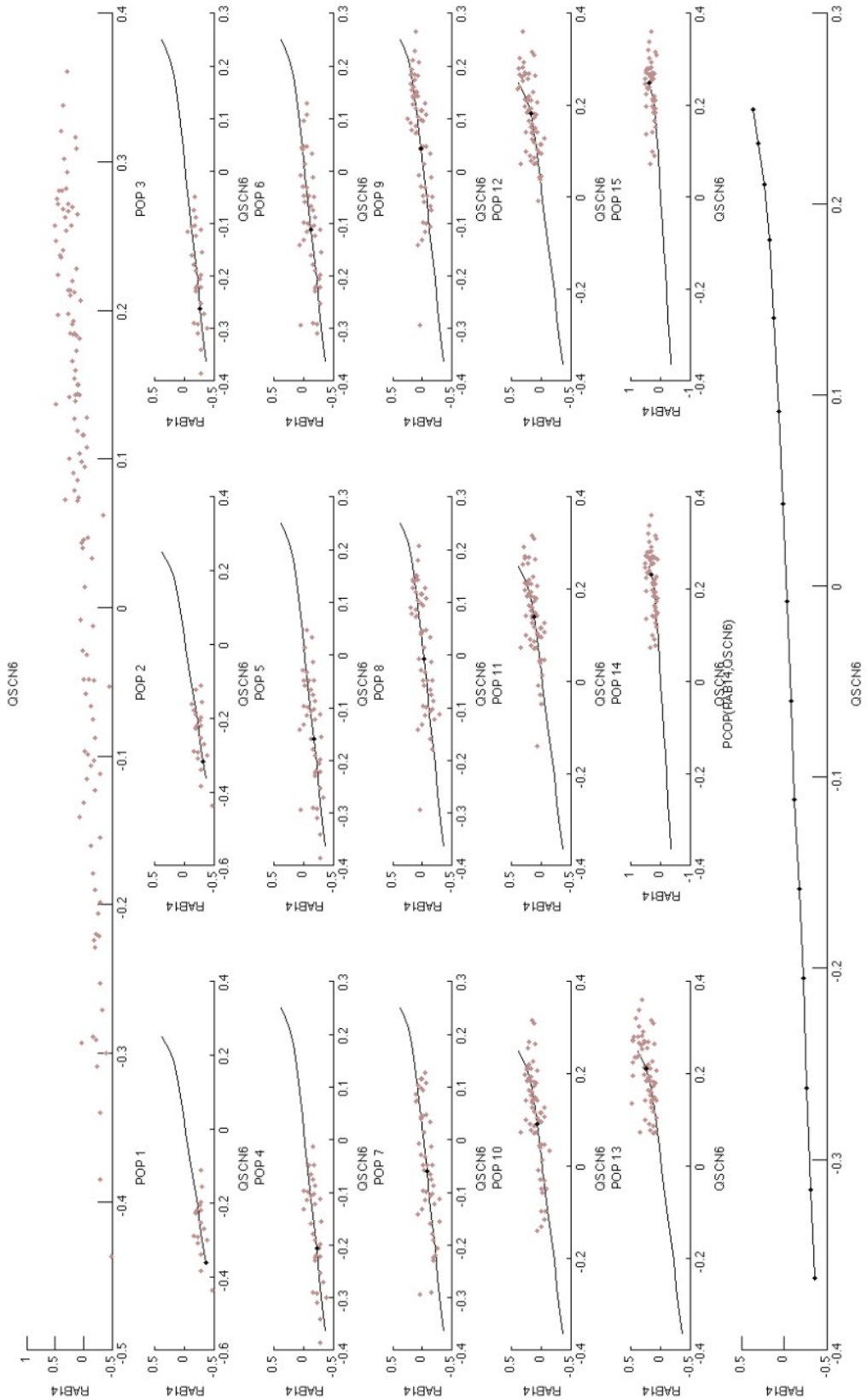


Ilustración 3.4: Los ejes de los gráficos representan los niveles de expresión de los genes QSCN6 (eje X)y RAB14 (eje Y). El primero de los gráficos nos muestra como se dispersan las condiciones muestrales de la microarray que nos muestran el nivel de expresión de cada uno de estos genes. En los siguientes 15 gráficos vemos, en negro, la PCOP(RAB14,QSCN6). En cada uno de ellos vemos uno de los 15 POPs, puntos en negro, que forman la PCOP, junto a las muestras, en gris, que les han sido vinculadas en el proceso que genera la PCOP. El gráfico inferior muestra la PCOP(RAB14,QSCN6) con todos sus POPs.

2. Filtrar la tabla anterior descartando aquellas celdas que representen menos del 5% de las muestras que forman el total del cluster (fila). Así, descartamos las muestras alejadas del resto de muestras del cluster (outlayers) y nos quedamos con los POPs con una mayor densidad de muestras de cada cluster. Esto nos servirá para determinar si un cluster aparece íntegro entre dos puntos de curvatura.

(1) Tabla de número de muestras del cluster

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	19	21	23	20	19	11	6	5	4							
2	5	7	8	11	14	13	11	10	7	5	2	1	1			
3		1	2	3	5	11	14	18	22	26	25	23	18	15	7	5
4						3	5	12	16	23	21	39	35	35	28	21
5		1	2	3	4	4	9	9	9	7	6	5	2			

(2) Tabla de porcentaje de muestras del cluster

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	14,84%	16,41%	17,97%	15,63%	14,84%	8,59%	4,69%	3,91%	3,13%							
2	5,26%	7,37%	8,42%	11,58%	14,74%	13,68%	11,58%	10,53%	7,37%	5,26%	2,11%	1,05%	1,05%			
3		0,51%	1,03%	1,54%	2,56%	5,64%	7,18%	9,23%	11,28%	13,33%	12,82%	11,79%	9,23%	7,69%	3,59%	2,56%
4						1,26%	2,10%	5,04%	6,72%	9,66%	8,82%	16,39%	14,71%	14,71%	11,76%	8,82%
5		1,64%	3,28%	4,92%	6,56%	6,56%	14,75%	14,75%	14,75%	11,48%	9,84%	8,20%	3,28%			

(3) Tabla de número de muestras del cluster sin outlayers

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	19	21	23	20	19	11										
2	5	7	8	11	14	13	11	10	7	5						
3						11	14	18	22	26	25	23	18	15		
4								12	16	23	21	39	35	35	28	21
5					4	4	9	9	9	7	6	5				

Ilustración 3.5: Representación del proceso de cruce de datos. La primera tabla (número de muestras del cluster) nos muestra cómo se relacionan los clusters de la distribución de clusters generada por el método Hierarchical dunn (k=7) sobre la PCOP(IFITM2,SLC36A4). La segunda tabla (tabla de porcentaje de muestras del cluster) nos muestra el porcentaje de muestras del cluster vinculado con cada POP (celda) respecto al total de muestras del cluster (fila). La tercera tabla (tabla de número de muestras del cluster sin outlayers) nos muestra el resultado tras desechar aquellas celdas que corresponden a menos del 5% de las muestras totales del cluster.

En la imagen superior [Ilustración 3.5] se aprecian tres tablas: la tabla de número de muestras del cluster, la tabla de porcentaje de muestras del cluster, la tabla de número de muestras del cluster sin outlayers. Cada una de las columnas de las tablas es un POP de la PCOP sobre la que se aplica la distribución de clusters. Cada una de las filas de las tablas es uno de los clusters de la distribución de clusters.

La tabla de número de muestras del cluster nos muestra cómo se relacionan los clusters de la distribución de clusters generada por el método Hierarchical dunn (k=7) sobre la PCOP(IFITM2,SLC36A4). Cada una de las celdas indica la cantidad de muestras de los datos de microarray que han sido vinculados, a al vez, a un cluster (fila) y a un POP (columna).

La tabla de porcentaje de muestras del cluster nos muestra el porcentaje de muestras del cluster vinculado con cada POP (celda) respecto al total de muestras del cluster (fila).

La tabla de número de muestras del cluster sin outlayers nos muestra el resultado tras

desechar aquellas celdas que corresponden a menos del 5% de las muestras totales del cluster. Esta es la tabla resultante del proceso de cruce de datos. Tan sólo nos indica cómo se relacionan los POPs de la PCOP con los clusters de la distribución de clusters, sin tener en cuenta los puntos de curvatura de la PCOP.

Esta tercera tabla (tabla de número de muestras del cluster sin outlayers) que se ha generado nos sirve para conocer como se relacionan los POPs de una PCOP con los clusters de una distribución de clusters. Los puntos de curvatura de la PCOP serán aplicados sobre esta tabla y, por lo tanto, podremos saber si la PCOP analizada muestra una transición entre los clusters de la distribución de clusters. Es decir, si la PCOP ubica los clusters de la distribución de clusters a un lado u otro de los puntos de curvatura.

El proceso de aplicar los puntos de curvatura de la PCOP sobre la tabla que se ha generado se realiza en el proceso *Filtrar las PCOPs sobre las que se ha aplicado la distribución de clusters* [Ilustración 3.3]. De esta manera, los puntos de curvatura de la PCOP nos sirven para evaluar la PCOP tras aplicar la distribución de clusters sobre ella.

Para filtrar las PCOPs sobre las que se ha aplicado la distribución de clusters usamos tres filtros:

1. Un primer filtro(filtro de cluster por número de POPs) que determina si un cluster de la distribución de clusters habita entre dos puntos de curvatura consecutivos de la PCOP . De esta forma sabemos qué cluster se distribuyen a un lado u otro de los puntos de curvatura de la PCOP.

2. Un segundo filtro(filtro de PCOP por número de clusters validos) que determina si un mínimo de clusters de la distribución de clusters habita entre dos puntos de curvatura consecutivos de la PCOP . En función del porcentaje de clusters de toda la distribución de clusters que superen el primer filtro, este segundo filtro aceptará la PCOP sobre la que se ha aplicado la distribución de clusters o la descartará. Este segundo filtro dispone de tres niveles de exigencia que generarán tres listados con las PCOPs que han superado cada uno de los niveles de exigencia.

3. Un tercer filtro(filtro de PCOP por clusters en diferentes intervalos) que evalúa la capacidad de la PCOP sobre la que se le ha aplicado la distribución de clusters para separar fenotipos exigiendo que los clusters habiten en intervalos diferentes. Este filtro se aplica a cada una de las PCOPs de los tres listados del filtro anterior (filtro de PCOP por número de clusters válidos).

Según el número de PCOPs que superen el proceso de filtrado, compuesto por los tres

filtros arriba explicados, podemos determinar el tipo de fenotípos que se detallan en la distribución de clusters aplicada sobre las PCOPs. Si un gran número de PCOPs superan el proceso de filtrado significa que los fenotipos encontrados por la distribución de clusters son de tipo sistémico (sucederán grandes cambios en la célula que involucran muchos genes por ejemplo transformarse en una célula cancerígena) mientras que si son pocas las PCOPs que superan el proceso de filtrado significa que los fenotipos encontrados por la distribución de clusters son de tipo transversal (que involucran pocos genes).

Una vez el total de PCOPs sobre las que se ha aplicado la distribución de clusters ha sido filtrada, se procede a realizar los cálculos necesarios para obtener los datos que permitan representar gráficamente la PCOP sobre la que se ha aplicado una distribución de clusters. Este proceso se realiza para cada una de las PCOPs sobre las que se ha aplicado la distribución de clusters que se encuentre en uno de los listados obtenidos en el proceso de filtrado anterior.

Todo este proceso se lleva a cabo en el proceso *Realizar los cálculos necesarios para generar la imagen* [Ilustración 3.3] y los cálculos que se realizan son:

1. Generar una tabla de POPs comunes entre los clusters de la distribución de clusters para relacionar los clusters entre sí dada una PCOP.

2. Generar el Layout, que nos indica cómo se relacionan entre sí los clusters de la distribución de clusters según la PCOP a la que se aplica la distribución de clusters. Esta relación se clasifica como Concatenación, Intersección o Unión. Consideramos que si dos clusters no tienen ningún POP en común son Concatenación. Si dos clusters tienen menos del 40% de POPs en común son considerados Intersección. Y si dos clusters tienen más de un 40% de POPs en común son considerados Unión.

3. Calcular los puntos medios de cada cluster de la distribución de clusters para los genes comparados en la PCOP [Ilustración 3.6]. Para encontrar el punto medio de un cluster, sumamos las coordenadas en el espacio 2D de todos las muestras del cluster y dividimos por el número de muestras. Este espacio 2D es el formado por los genes cuya relación de expresión es descrita por la PCOP, y las coordenadas de cada muestra serán el nivel de expresión de cada gen para esa muestra.

4. Buscar el POP más próximo al punto medio de cada cluster de la distribución de clusters. De esta forma conocemos el orden de aparición, sobre la PCOP, de los clusters de la distribución de clusters.

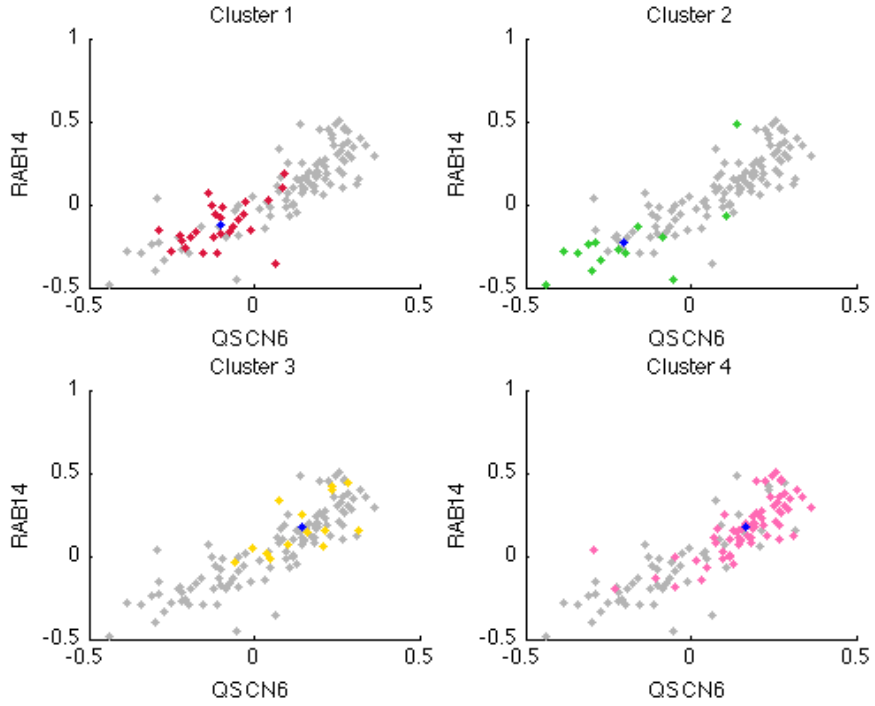


Ilustración 3.6: En cada una de las gráficas se ve, en gris, todas las condiciones muestrales de la microarray que nos indican el nivel de expresión de los genes QSCN6 y RAB14. En cada una de las 4 gráficas se han coloreado las muestras que forman los clusters que ha generado el método de clustering SOTA (k=4). En azul se ha coloreado el punto medio de cada cluster.

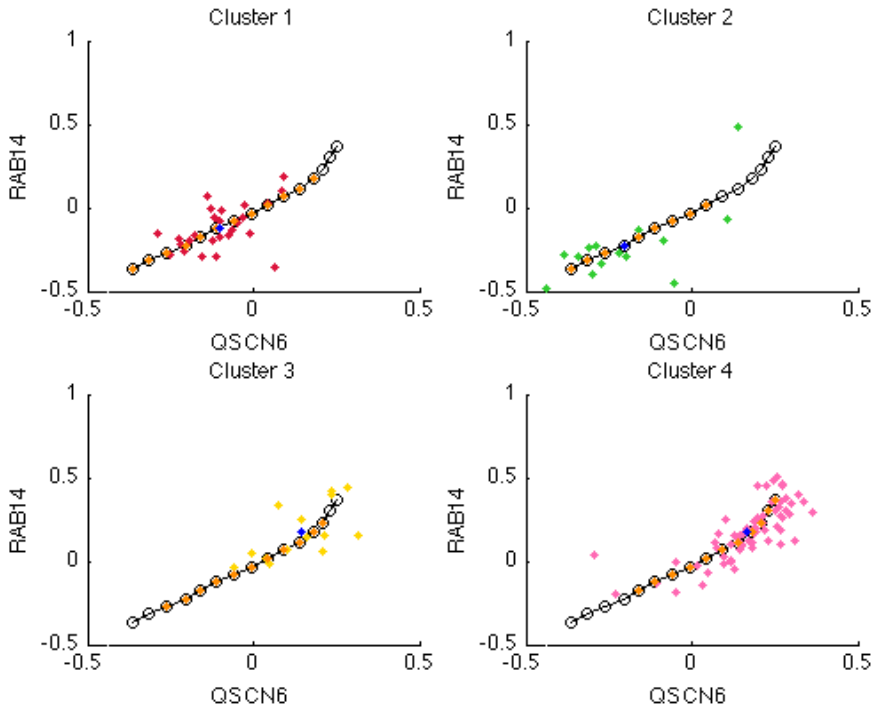


Ilustración 3.7: Vemos, en cada gráfico, las muestras que forman los 4 clusters resultantes del método de clustering SOTA (k=4). También se ha representado, en negro, la PCOP(QSCN6,RAB14) y los POPs que la componen. En cada gráfico se han coloreado en azul el punto medio de cada cluster y en naranja los POPs que cada cluster tiene vinculados.

Obtenidos ya todos los datos se procede a generar un fichero de salida por cada listado obtenido en el proceso de filtrado. Obtenemos así tres ficheros para cada distribución de clusters que se haya cargado uno por cada nivel de exigencia del proceso de filtrado. Este proceso: *Generar el fichero para el filtro, agrupando las PCOPs sobre las que se aplica la distribución de clusters* [Ilustración 3.3], tiene en cuenta, para generar los ficheros de salida, el momento en que aparecen los clusters de la distribución de clusters a lo largo de la PCOP. Encontrado el orden de aparición de los clusters de la distribución de clusters a lo largo de la PCOP, se busca en que porción de la curva entre dos puntos de curvatura consecutivos se encuentran. De esta forma se pasan a agrupar aquellas PCOPs sobre las que se ha aplicado la distribución de clusters en las que aparecen los mismos clusters juntos en una porción de la curva entre dos puntos de curvatura consecutivos de la PCOP involucrada. A cada una de estas agrupaciones le llamamos AGRUPACIÓN. Cada uno de los grupos de PCOPs que se ha formado (AGRUPACIÓN) se agrupa en ORDENACIONES. Cada ORDENACIÓN es un grupo de PCOPs sobre las que se aplica la distribución de clusters que, no solo tienen los mismos clusters juntos en una porción de curva entre dos puntos de curvatura consecutivos, sino que estos clusters aparecen, además, en el mismo orden sobre la curva. El último paso consiste en agrupar las PCOPs de cada ORDENACIÓN por tener, no sólo los clusters en un mismo orden dentro de la misma porción de la curva, sino porque estos clusters mantienen una misma relación con los clusters vecinos de la porción de la curva entre dos puntos de curvatura consecutivos (intersección, unión o concatenación), es decir, que tengan el mismo LAYOUT.

Finalmente sólo queda generar cada gráfico que representa cada PCOP sobre la que se aplica la distribución de clusters. O sea, el proceso *Generar imagen de la PCOP sobre la que se aplica la distribución de clusters* [Ilustración 2.3] se ejecuta por cada una de las PCOPs sobre las que se ha aplicado la distribución de cluster que ha superado el proceso de filtrado y ha sido agrupada en el proceso anterior.

3.3.1.1. Filtro de cluster por número de POPs

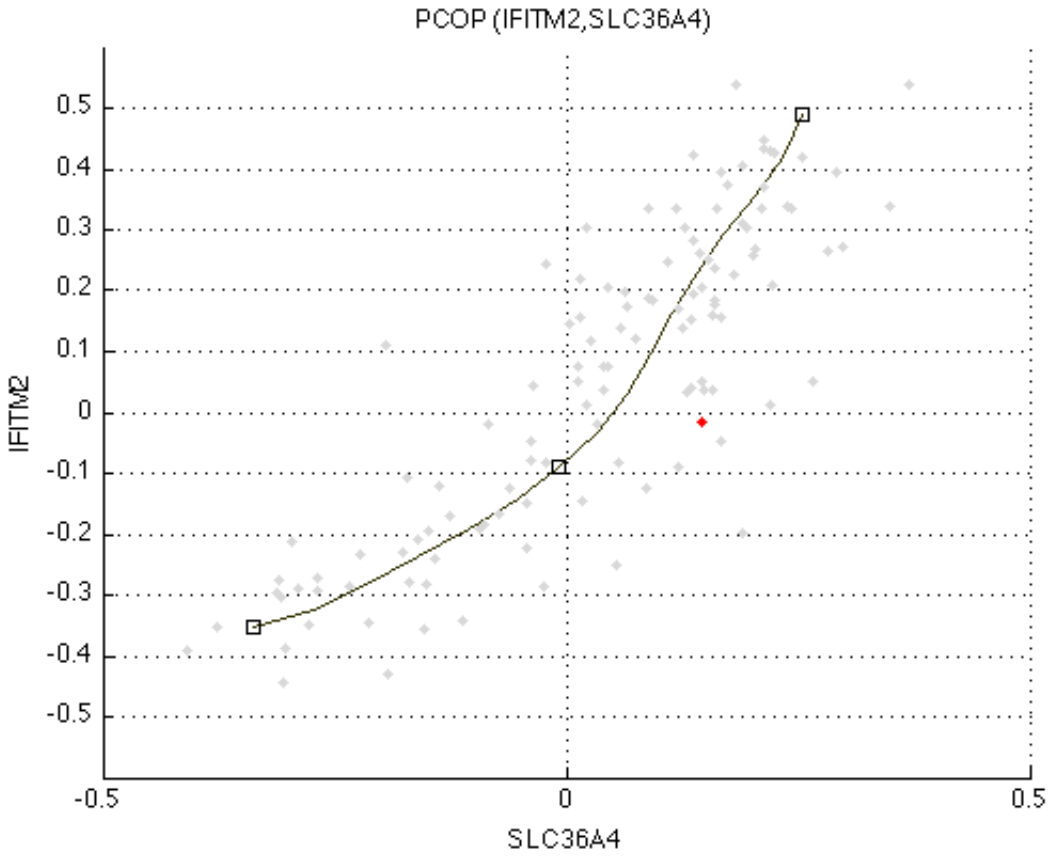


Ilustración 3.8: Los ejes del gráfico representan los niveles de expresión de los genes SLC36A4 (eje X) y IFITM2 (eje Y). Cada uno de los puntos del gráfico, coloreados en gris, corresponde a una condición muestral de la micrarray, mostrándonos el nivel de expresión de cada gen para esa condición muestral. El punto rojo corresponde a la condición muestral de aplicar la droga: 67574-TU-TU-Tu-:-:-:-Vincristine-sulfate (S06757408MDQ), que induce unos valores de expresión: 0,1470 para el gen SLC36A4 y -0,0170 para el gen IFITM2. Mediante el método de extracción de curvas principales(PCOP) a la nube de puntos obtenemos una curva que representa la relación de expresión entre los dos genes. La PCOP aparece dibujada en negro en el gráfico. Los cuadrados negros identifican los puntos de curvatura de la PCOP, dividiendo la curva en dos segmentos.

El objetivo de este filtro es determinar si un cluster de la distribución de clusters que se aplica sobre la PCOP habita entre dos puntos de curvatura consecutivos de la PCOP involucrada.

Para considerar que un cluster habita entre dos puntos de curvatura, un porcentaje mínimo del total de POPs vinculados al cluster, deben estar entre esos dos puntos de curvatura consecutivos de la PCOP sobre la que se aplica la distribución de clusters.

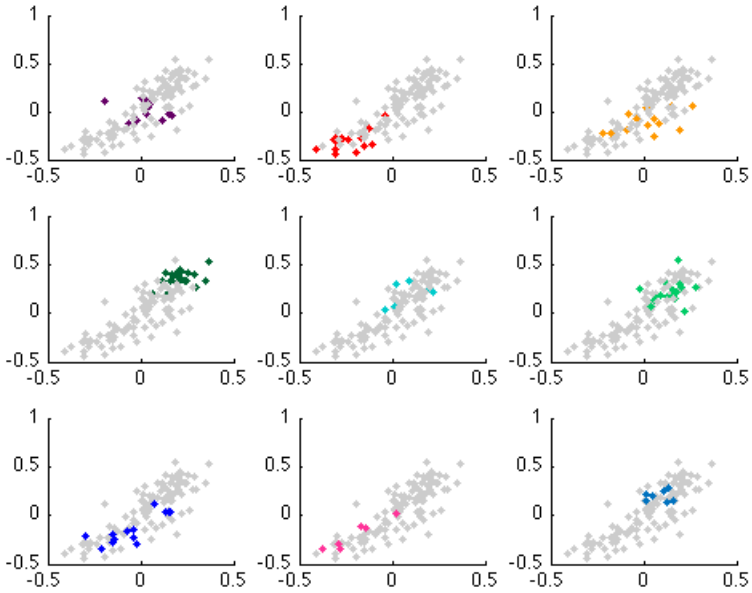
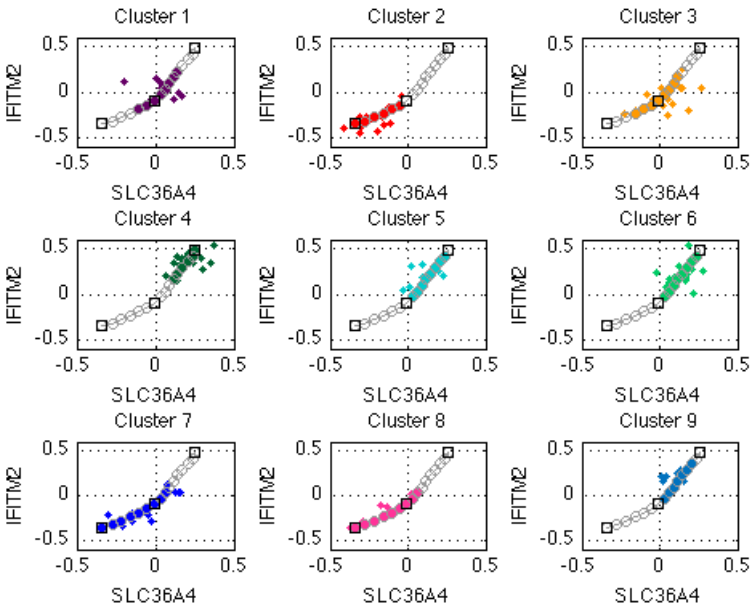


Ilustración 3.9: Los ejes del gráfico representan los niveles de expresión de los genes SLC36A4 (eje X) y IFITM2 (eje Y). Cada uno de los puntos coloreados en los gráfico, en gris, corresponde a una condición muestral de la micrarray, mostrándonos el nivel de expresión de cada gen para esa condición muestral. En cada uno de los gráficos se han coloreado de un color diferente las condiciones muetras que, aplicando el método de clustering Pc Hierarchical descarte (k=11), forman un cluster de la distribución de clusters resultante.



lustración 3.10: En cada uno de los 9 gráficos podemos observar la ubicación de uno de los clusters del método Pc Hierarchical descarte (k=11) sobre la PCOP (IFITM2,SLC36A4). La PCOP (IFITM2,SLC36A4) está representada por una línea gris, a la que se le superponen una serie de círculos grises vacíos representando los POPs que forman la PCOP. En cada gráfico se han coloreado las condiciones muestrales que forman un cluster, pintando del mismo color aquellos POPs vinculados al cluster. Además, vemos los puntos de curvatura de la PCOP marcados con un cuadrado negro. Podemos observar que el Cluster 2 (rojo) tiene la totalidad de sus POPs entre el primer y el segundo punto de curvatura de la PCOP, con lo que superará el filtro que determina si un cluster habita entre dos puntos de curvatura de la PCOP sobre la que se aplica. En cambio, el Cluster 1 (morado) tiene sus POPs repartidos alrededor del segundo punto de curvatura de la PCOP. Concretamente tiene el 37,5% de los POPs en el intervalo formado por el primer y segundo punto de curvatura de la PCOP y el 62,5% en el intervalo formado por el segundo y tercer punto de curvatura de a PCOP. Con lo que no superará el *filtro de cluster por número de POPs*.

El *filtro de cluster por número de POPs* se lanza, por defecto, con una cota mínima del 80%. Es decir, se precisa que el 80% de los POPs asignados a un cluster se encuentren entre los dos mismos puntos consecutivos de curvatura de la PCOP. Si se supera el filtro se considerará que el cluster habita en el intervalo formado por estos dos puntos de curvatura.

3.3.1.2. Filtro de PCOP por número de clusters válidos

Tal y como ya se ha indicado, el objetivo de este filtro es encontrar aquellas PCOPs que al aplicarles los clusters de la distribución de clusters, un mínimo de clusters se encuentren entre dos puntos de curvatura de la PCOP. Es decir, comprobar si las PCOPs a las que se ha aplicado una distribución de clusters son capaces de diferenciar los fenotipos determinados por la distribución de clusters.

Se determina si un mínimo de clusters de la distribución de clusters habita entre dos puntos consecutivos de curvatura de la PCOP sobre la que se aplica la distribución de clusters aplicando el primer filtro (*filtro de cluster por número de POPs*) a cada uno de los clusters de la distribución de clusters. Con ello se sabe el porcentaje de clusters de la aplicación de distribución de clusters que habitan entre los puntos de curvatura de la PCOP sobre la que se aplica.

Según el porcentaje de clusters de la distribución de clusters aplicada sobre una PCOP que superen el primer filtro *(filtro de cluster por número de POPs)* descartaremos la PCOP sobre la que se ha aplicado la distribución de clusters o la aceptaremos. Este filtro se aplica 3 veces a la relación total de PCOPs sobre las que se ha aplicado la distribución de clusters con 3 porcentajes de criba distintos. Por defecto: 40% (F1), 65% (F2) y 80% (F3). Es decir, que si una PCOP sobre la que se ha aplicado la distribución de clusters tiene el 50% de los clusters de la distribución de clusters habitando entre dos de sus puntos de curvatura consecutivos superará la primera criba del filtro (F1: 40%) pero no superará la segunda (F2: 65%) ni la tercera (F3: 80%). En contraposición a una PCOP sobre la que se ha aplicado la distribución de clusters que, teniendo el 90% de los clusters habitando entre dos de sus puntos de curvatura consecutivos, superará las 3 cribas.

A continuación vemos un ejemplo de los resultados obtenidos tras la aplicación de cada uno de los tres niveles de criba:

	Total (MC-PCOP	FI 1	FI 2	FI 3
Número de relacione	53.361	42.729	16.354	2.152
Porcentaje	100,00%	80,08%	30,65%	4,03%

Con estas cribas descartamos aquellas PCOPs sobre las que se ha aplicado la distribución de clusters cuyos clusters no habiten (en un 40%, 65% o 80%) entre dos puntos consecutivos de curvatura de la PCOP sobre la que se aplican. Los resultados obtenidos de estas cribas aseguran poder diferenciar fenotipos, pero no aseguran poder diferenciarlos en intervalos formados por distintos puntos de curvatura consecutivos de la PCOP involucrada.

3.3.1.3. Filtro de PCOP por clusters en intervalos

De los 3 filtros este es el más determinante. Su objetivo es descartar aquellas PCOPs que no tengan, por lo menos, dos clusters de la distribución de clusters habitando en intervalos diferentes. Intervalos formados por puntos de curvatura consecutivos de la PCOP sobre la que se aplica la distribución de clusters.

Este filtro evalúa la capacidad de una PCOP para distribuir los fenotipos descritos por una distribución de clusters a lo largo de la relación de expresión entre los genes.

A nivel de ejemplo podemos ver, en las siguientes gráficas [Ilustración 3.11], la distribución de cada uno de los clusters de la distribución de clusters generada por el método de clustering Pc Hierarchical (k=11) a lo largo de la PCOP (IFITM2 y SLC36A4).

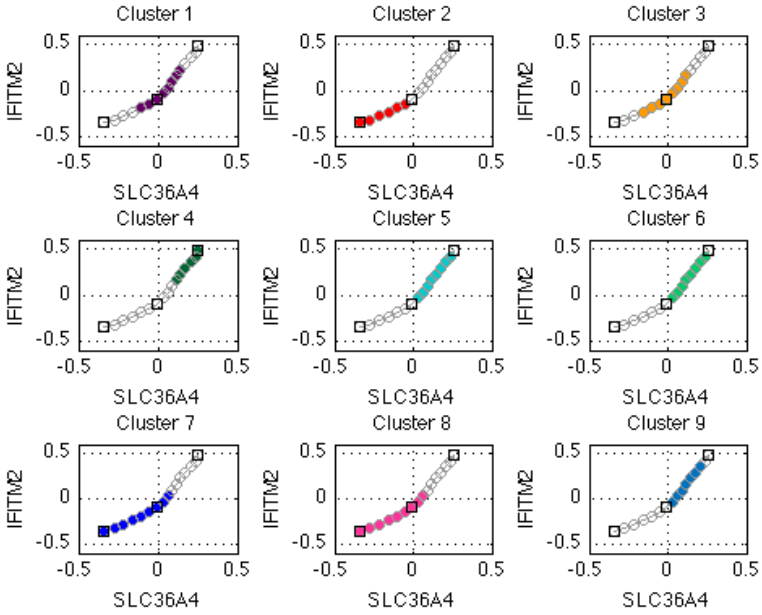


Ilustración 3.11: Podemos observar que de los 9 clusters del método de clustering, el Cluster 1 (morado) y el Cluster 3 (naranja) quedan distribuidos alrededor de un punto de curvatura, con lo que no superan el filtro que determina si un cluster habita entre dos puntos de curvatura consecutivos de la PCOP. El cluster Clusters azul (7) y el Cluster 8 (rosa) tienen 2 de sus 9 POPs fuera del intervalo formado por los dos primeros puntos de curvatura de la PCOP. Es decir, que habitan en el intervalo formado por el primer y segundo puntos de curvatura de la PCOP en un 77,7% y en el intervalo formado por el segundo y tercer punto de curvatura de la PCOP en un 22,2%. Tampoco superan el *filtro de cluster por numero de POPs*. Los demás clusters (Cluster 2 (rojo), Cluster 4 (verde oscuro), Cluster 5 (cyan), Cluster 6 (esmeralda) y Cluster 9 (azul claro)), superando el *filtro de cluster por número de POPs*. Teniendo el Clusters 1 (rojo) habitando en el intervalo formado por el primer y segundo punto de curvatura de la PCOP y los clusters 4 (verde oscuro), 5 (cyan), 6 (esmeralda) y 9 (azul claro) habitando en el intervalo formado por el segundo y tercer punto de curvatura de la PCOP, se cumple el filtro que determina si por los menos dos clusters de la distribución de clusters habitan en dos intervalos diferentes de la PCOP sobre la que se aplica la distribución de clusters.

Podemos observar [Ilustración 3.11] que de los 9 clusters de la distribución de clusters, los clusters 1 (morado) y 3 (naranja) quedan distribuidos entorno a 1 punto de curvatura, con lo que no superan el *filtro de cluster por número de POPs*. Los clusters 7 (azul) y 8 (rosa) tienen 2 de sus 9 POPs fuera del primer intervalo formado por los dos primeros puntos consecutivos de curvatura de la PCOP sobre la que se aplican. Es decir, habitan en un 77,7% en el primer intervalo formado por el primer y segundo punto de curvatura de la PCOP y en un 22,2% en el intervalo formado por el segundo y tercer punto de curvatura de la PCOP. Tampoco superan el *filtro de cluster por número de POPs*. Los demás clusters (2, 4, 5, 6 y 9), superando el *filtro de cluster por número de POPs*, habitan en el intervalo formado por el primer y segundo punto de curvatura de la PCOP o en el intervalo formado por el segundo y tercer punto de curvatura de la PCOP.

Con los clusters 2 (rojo), 4 (verde oscuro), 5 (cyan), 6 (esmeralda) y 9 (azul claro), que superan el *filtro de cluster por número de POPs*, obtenemos que 5 de los 9 clusters de la distribución de clusters que se aplica sobre la PCOP habitan en intervalos formados por puntos de curvatura consecutivos de la PCOP. Es decir, que el 55,5% de los clusters de la aplicación de distribución de clusters sobre la PCOP habitan en intervalos formados por puntos de curvatura consecutivos de la PCOP y por lo tanto encontraremos esta PCOP a la que se le ha aplicado la distribución en las PCOPs aceptadas por la criba, del *filtro de PCOP por número de clusters válido*, del 40%.

Teniendo el cluster 2 (rojo) habitando en el intervalo formado por el primer y segundo punto de curvatura de la PCOP y los clusters 4 (verde oscuro), 5 (cyan), 6 (esmeralda) y 9 (azul claro) habitando en el intervalo formado por el segundo y tercer punto de curvatura de la PCOP, se supera el filtro de PCOP por cluster en intervalos.

A continuación vemos la misma tabla de la sección anterior con los resultados de los tres niveles de criba que determina si un cluster de la distribución de clusters habitan entre dos puntos consecutivos de curvatura de la PCOP sobre la que se aplican a la que le hemos añadido los resultados obtenidos por el *filtro de PCOP por clusters en intervalos*.

	Total (MC-PCOP	FI 1	FI 1 + Fe	FI 2	FI 2 + Fe	FI 3	FI 3 +Fe
Número de relacione	53.361	42.729	7.947	16.354	2.139	2.152	97
Porcentaje	100,00%	80,08%	14,89%	30,65%	4,01%	4,03%	0,18%

3.3.1.4. Agrupación de las PCOPs sobre las que se aplica una distribución de clusters por AGRUPACIÓN, ORDENACIÓN y LAYOUT

Como ya se ha comentado con anterioridad, con las PCOPs que han superado el proceso de filtrado anterior para una distribución de clusters dada, se procede a su agrupación bajo diferentes criterios para ser mostradas de una forma visual, entendible y altamente operable en la aplicación web. Tras el proceso de filtrado hemos encontrado qué clusters aparecen juntos en la misma porción de la curva entre dos puntos de curvatura consecutivos de la PCOP involucrada. Además del orden de aparición de los clusters de la distribución de clusters a lo largo de la PCOP, y si los clusters vecinos comparten POPs a los que se vincula (es decir si el mismo POP está vinculado a diferentes clusters). Estos POPs compartidos nos determinan si hay una unión entre clusters cuando el 40% o más de todos los POPs vinculados a un cluster también lo están a otro cluster. Hay una intersección cuando menos del 40% de los POPs vinculados a un cluster también lo están a otro cluster. Hay una concatenación cuando ninguno de los POPs vinculados a un cluster lo está a otro cluster.

Una vez sabemos todo esto para la aplicación de una distribución de clusters sobre una PCOP, averiguamos si se cumplen los mismos criterios sobre los mismos clusters al aplicar la misma distribución de clusters sobre una PCOP diferente. Es decir, si los mismos clusters aparecen juntos en una porción de la curva entre dos puntos de curvatura consecutivos de la nueva PCOP, si aparecen en el mismo orden en la nueva PCOP que en la antigua PCOP y si mantienen la misma relación de unión, intersección y concatenación al ser aplicados sobre la nueva PCOP que cuando lo eran en la PCOP antigua. Una vez se ha comparado la aplicación de la distribución de clusters sobre todas las PCOPs que han pasado los filtros descritos en las anteriores secciones, pasamos a la agrupación de las PCOPs por AGRUPACIÓN, ORDENACIÓN y LAYOUT de los clusters.

Las PCOPs que agrupen los mismos clusters de la distribución de clusters en sus intervalos, se agruparán por AGRUPACIÓN. Cada uno de los grupos de PCOPs que se ha formado (AGRUPACIÓN) se agrupa de nuevo por ORDENACIÓN. Cada ORDENACIÓN es un grupo formado por aquellas PCOPs sobre las que se ha aplicado la distribución de clusters que, además de tener los mismos clusters juntos en una porción de la curva entre dos puntos de curvatura consecutivos, estos aparecen en el mismo orden sobre la curva. El último paso consiste en agrupar las ORDENACIONES agrupando las PCOPs sobre las que se ha aplicado la distribución de clusters donde los clusters, además de un mismo orden, mantienen una misma relación con los clusters vecinos (concatenación, intersección o unión), es decir, que tengan el mismo LAYOUT.

Como resultado tenemos las PCOPs que describen los cambios fenotípicos para una distribución de clusters dada, ordenados en una jerarquía de grupos y subgrupos. Esta jerarquía de subgrupos nos indicará aquellas PCOPs que lleven a cabo la misma transición entre los fenotipos, o sea las PCOPs que aparecen juntas en el mismo tercer nivel del árbol (LAYOUT); y aquellas PCOPs que lleven a cabo transiciones muy diferentes entre los fenotipos, o sea las PCOPs que aparecen separadas en el primer nivel del árbol (AGRUPACIÓN).

3.3.2. Programa para generar el gráfico de una PCOP sobre la que se aplica una distribución de clusters

Para la representación gráfica de una PCOP sobre la que se aplica una distribución de clusters se ha buscado una manera de hacerlo cuyo gráfico resultante fuese semejante a la forma de un diagrama Box-Plot.

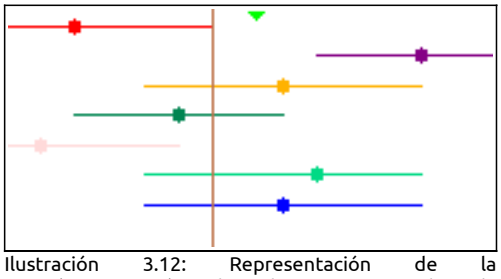


Ilustración 3.12: Representación de la PCOP(GNAS,ASNS) sobre la que se aplica la distribución de clusters generada por el método de clustering Hierarchical (k=6). La anchura de la imagen representa la longitud de la PCOP. Los clusters aparecen representados como líneas horizontales que determina su inicio y su fin en la PCOP. El punto que vemos en cada cluster representa su punto de mayor densidad, y la intensidad representa la cantidad de muestras del cluster respecto a la cantidad total de muestras. Las líneas marrones verticales son la representación de los puntos de curvatura de la PCOP. Además se sitúa un triangulo verde en la parte superior de la imagen que indica el punto de mayor densidad de muestras de la PCOP.

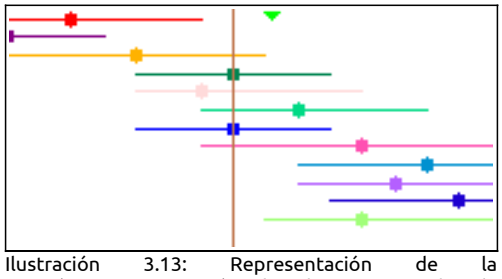


Ilustración 3.13: Representación de la PCOP(MGC14376,ITGA3) sobre la que se aplica la distribución de clusters generada por el método de clustering Som Descarte (k=12). La anchura de la imagen representa la longitud de la PCOP. Los clusters aparecen representados como líneas horizontales que determina su inicio y su fin en la PCOP. El punto que vemos en cada cluster representa su punto de mayor densidad, y la intensidad representa la cantidad de muestras del cluster respecto a la cantidad total de muestras. Las líneas marrones verticales son la representación de los puntos de curvatura de la PCOP. Además se sitúa un triangulo verde en la parte superior de la imagen que indica el punto de mayor densidad de muestras de la PCOP.

La anchura de la imagen [Ilustración 3.12 y 3.13] representa la longitud de la PCOP. La vertical de la imagen la repartimos entre la cantidad de clusters a representar.

Los clusters de la distribución de clusters son representados como una sucesión vertical de líneas horizontales (cada una de un color distinto), mientras que los puntos de curvatura de la PCOP son una sucesión horizontal de líneas marrones verticales. Además se sitúa un triangulo verde en la parte superior de la imagen que indica el punto de mayor densidad de muestras de la PCOP.

Los puntos de inicio y de final de cada uno de los clusters han sido calculados para que estos se sitúen en la posición del POP que les ha sido asignado. Además, para cada cluster, se muestra un punto indicando su centro de densidad. La intensidad de este punto representará la

cantidad de muestras del cluster respecto a la cantidad total de muestras.

Del programa para generar el gráfico de una PCOP sobre la que se aplica una distribución de clusters se generaron dos versiones distintas: Una primera versión en python y una segunda en C/C++.

La primera versión fue programada en python por la facilidad de uso e instalación de las librerías de tratamiento de imágenes PIL. Sin embargo, en el momento de realizar las pruebas de lanzamiento a través de un programa situado en el servidor web del IBB, detectamos que no se ejecutaban los lanzamientos programados. Se intentó solventar el problema llevando a cabo el siguiente el proceso:

- Revisar si el módulo de python para el servidor Apache (mod_python).
- Comprobar si los permisos del módulo (mod_python) eran los mismos que los permisos de otros módulos que sí funcionaban (mod_perl).
- Comprobar los permisos del script de python (propietario, grupo y ejecución).
- Comprobar el archivo de configuración del servidor (/etc/httpd/conf/httpd.conf) y comprobar que la extensión *.py* estuviera habilitada.
- Forzar la ruta del interprete de python en la llamada al script.

Al ver que ninguna de las comprobaciones no solventaban el problema con el lanzamiento de scripts de python en el servidor y puesto que el programa para generar el gráfico de una PCOP sobre la que se aplica una distribución de clusters debía ser usado en otros proyectos, decidí implementar una nueva versión del programa en C/C++ haciendo uso de la librería PNGwritter.

3.4. Fase 3: Aplicación web CrossingClusters para el estudio de los genes responsables de la transición entre clusters de condiciones muestrales

En la planificación inicial se contemplaba hacer uso de la aplicación web PhenoSamples-cl para presentar y poder analizar los datos generados a través del proceso de cruce.

La idea era modificar las funciones que cargan los datos presentados en la aplicación web PhenoSamples para cargar, por un lado, la totalidad de las distribuciones de clusters usadas en el proceso de cruce de datos y, por otro lado, los ficheros con las PCOPs sobre las que se han aplicado las distribuciones de clusters mencionadas y que han superado el proceso de filtrado. Con estas dos modificaciones realizadas sólo quedaba realizar la vista detalle para ver el listado de PCOPs que han superado el proceso de filtrado para cada distribución de

clusters. Esta vista detalle nos mostraría una agrupación jerárquica de las PCOPs según estén vinculadas, o no, en las mismas transiciones entre los mismos clusters de la distribución de clusters, toda la información pertinente de cada PCOP, con la oportunidad de analizar la relación de expresión en detalle, y la imagen generada de la PCOP sobre la que se ha aplicado la distribución de clusters.

No obstante, al modificar la primera función de carga de datos de PhenoSamples-gl, cuyo objetivo era cargar la totalidad de las distribuciones de clusters, se detectó que la implementación encargada de la ordenación de las condiciones muestrales de la microarray según la distribución de clusters (mostrando juntas las condiciones muestrales que pertenecen a un mismo cluster) no funcionaba correctamente.

La implementación encargada de ordenar las condiciones muestrales según la distribución de clusters [Ilustración 3.14] debe ordenar las condiciones muestrales según el orden en que aparecen los clusters en las distribuciones de clusters seleccionadas:

1. Primera distribución de clusters: Las condiciones muestrales se muestran como una sucesión de clusters ordenados.

2. Otras distribuciones de clusters: Por cada cluster de la distribución de clusters anterior, reordenamos las condiciones muestrales por los clusters de la actual distribución de clusters, sin romper las agrupaciones en clusters de la distribución anterior.

Ilustración 3.14: Ejemplo del proceso de ordenación de las condiciones muestrales de la microarray. Se ordenan las condiciones muestrales de la microarray según el cluster al que pertenecen en la distribución de clusters actual, respetando las agrupaciones en cluster de la distribución de clusters anterior.

A raíz de este problema se reprogramó gran parte de la aplicación web PhenoSampels-cl, hasta extremos de usar, únicamente, su interfaz gráfica. Es importante indicar que se intentó minimizar los cambios en la interfaz gráfica de PhenoSamples-cl para adecuarla a los datos que necesitábamos representar. El objetivo era facilitar al usuario el manejo de ambas aplicaciones al mostrar ambas una misma interfaz con la que operar.

3.4.1. Interfaz general donde se muestran las diferentes distribuciones de clusters para cada método de clustering

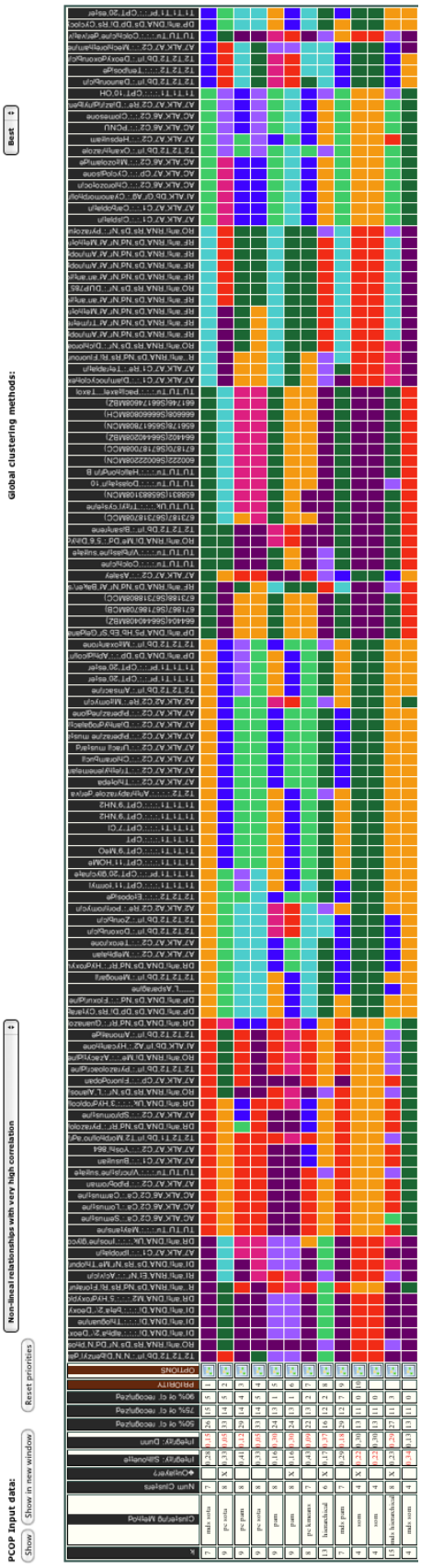


Ilustración 3.15

Pie de imagen de la Ilustración 3.15:

En la parte superior izquierda vemos una lista desplegable que nos permite seleccionar el filtro por correlación en que se clasifican las PCOPs. En la parte superior derecha vemos una lista desplegable con el listado de método de clustering que han generado las distribuciones de clusters a estudiar. Justo debajo de las listas desplegables tenemos tres botones: *Show,* actualiza la información mostrada según el filtro por correlación de PCOPs y el método de clustering seleccionados. *Show in new window*, abre una nueva interfaz general mostrando la información según el filtro por correlación de PCOPs y el método de clustering seleccionados. *Reset priorities*, resetea el orden en el que se ordenan las distribuciones de clusters listadas. La fila de etiquetas negras y marrones nos informan de lo que la columna representa: de la 1ª a la 11ª nos dan información acerca de cada distribución de clusters: k,metodo de clustering que la ha generado, número de clusters alcanzados, si el método ha aplicado inclusión u outlayer, valor de integridad aplicando el método *silhouette*, valor de integridad aplicando el parámetro *dunn, n*úmero de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 40%, número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 65%, número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 80% y su prioridad en la ordenación y opciones [Ilustración 3.27]. Las siguientes columnas (118 en la Ilustración 3.15) son las condiciones muestrales de la microarray que aparecen coloreadas según el cluster de cada distribución de clusters que se lista.

k	Clustering Method	Num Clusters	Outlayer?	Integrity: Silhouette	Integrity: Dunn	40% de reacción	75% de reacción	90% de reacción	PRIORITY	OPTIONS
13	sota	9		0,12	0,26	23	13	0	1	
11	sota	9		0,12	0,26	23	13	0	2	
12	sota	9		0,12	0,26	23	13	0	3	
15	sota	9		0,12	0,26	23	13	0	4	
16	sota	9		0,12	0,26	23	13	0	5	
14	sota	9		0,12	0,26	23	13	0	6	
13	sota	8	X	0,00	0,28	22	8	1	7	

Ilustración 3.27: Captura de las primeras 11 columnas de la interfaz general, donde vemos las primeras 11 columnas descritas con las 7 primeras distribuciones de clusters generadas por el método de clustering SOTA.

En la parte superior de la *intefaz general* nos encontramos el menú principal [Ilustración 3.15] . Las listas desplegables que encontramos nos permiten seleccionar el filtro por correlación de las PCOPs y el método de clustering que deseamos analizar.

Las primeras columnas de la izquierda muestran los distintos parámetros de cada una de las distribuciones de clusters generada por el método de clustering:

1. Valor del parámetro *k* aplicado al método de clustering .

2. Nombre del método de clustering.

3. Número alcanzado de clusters en la distribución de clusters.

4. Indicativo de si el método de clustering ha aplicado descarte o inclusión de outlayers.

5. Valor de integridad aplicando el método *silhouette* a la distribución de clusters.

6. Valor de integridad aplicando el método *dunn* a la distribución de clusters.

7. Número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 40% de clusters entre dos puntos de curvatura.

8. Número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 65% de clusters entre dos puntos de curvatura.

9. Número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 80% de clusters entre dos puntos de curvatura.

Las distribuciones de clusters listadas pueden ser ordenadas por cualquiera de los parámetros que hemos detallado hasta ahora. Para hacerlo basta con hacer clic en la etiqueta del parámetro por la que se quiera ordenar.

10. Prioridad: Para la ordenación de las condiciones muestrales (columnas) según la distribución de clusters

11. Opciones: Permite acceder a un cuadro de diálogo [Ilustración 3.16] que nos permite realizar las siguientes acciones:

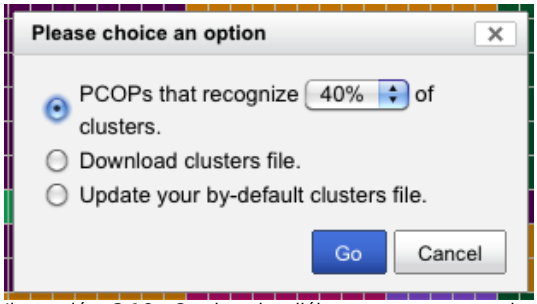


Ilustración 3.16: Cuadro de diálogo que nos permite acceder a la interfaz de detalle donde se muestran las relaciones entre genes responsables de la transición entre los clusters de una determinada distribución de clusters según el porcentaje de criba. Descargar un fichero con la distribución de clusters o usar esa distribución de clusters como distribución de clusters por defecto en el servidor de aplicaciones.

   - Usar la distribución de clusters seleccionada como distribución de clusters por defecto para el usuario en el servidor de aplicaciones.

- Descargar el fichero con la distribución de clusters seleccionada.

- Seleccionado un valor de porcentaje de criba, acceder al listado de PCOPs que han superado el porcentaje de criba y acceder a la interfaz de detalle para poder analizar los resultados.

El resto de columnas son las condiciones muestrales sobre las que se ha aplicado el método de clustering con los parámetros detallados en las primeras 11 columnas. Cada una de las condiciones muestrales aparece coloreada según al cluster al que ha sido vinculada. De manera que vemos de una forma visual y rápida la distribución de clusters que se ha obtenido gracias al método de ordenación descrito anteriormente.

La columna *prioridad* (10) nos indica el orden en el que se ordenan las condiciones muestrales (columnas) según la distribución de clusters. Al hacer clic en una distribución de clusters, esta, pasa formar parte de la ordenación, ocupando la primera posición y relegando al resto una posición. Esta ordenación afecta a la manera cómo se mostraran las condiciones muestrales de la microarray ya que añadir una nueva distribución de clusters implica ordenar, nuevamente, todas las muestras de la microarray.

3.4.2. Interfaz de detalle donde se muestran las relaciones entre genes responsables de la transición entre los clusters de una determinada distribución de clusters

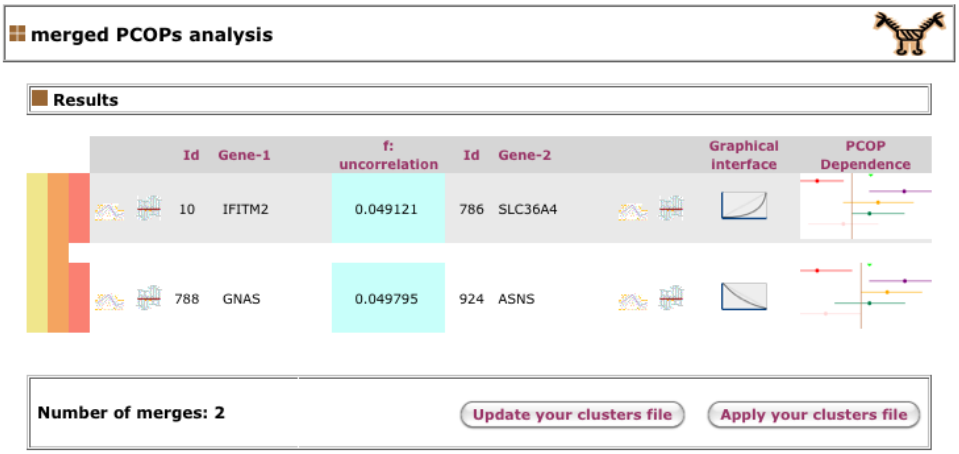


Ilustración 3.17: Captura de la vista con una lista de PCOPs. Vemos, a la izquierda 3 columnas que nos indican la AGRUPACIÓN, la ORDENACIÓN y el LAYOUT de cada una de las PCOPs. En cada entrada, que corresponde a una PCOP, tenemos los genes involucrados en la PCOP con vínculo a la información que les corresponde. Un vínculo a la representación de la PCOP y la imagen que ilustra la PCOP sobre la que se ha aplicado la distribución de clusters.

La interfaz de detalle [Ilustración 3.17] nos permite ver las PCOPs que han superado el proceso de filtrado con un cierto nivel de exigencia, seleccionado en el cuadro de diálogo que nos ha permitido acceder a esta interfaz de detalle, donde se muestran las relaciones entre genes responsables de la transición entre los clusters de una determinada distribución de

clusters.

Se presenta en formato lista todas las PCOPs sobre las que se ha aplicado la distribución de clusters y que han superado el nivel de exigencia seleccionado. De cada una de las entradas de la lista, obtenemos la información que concierne a la PCOP (vínculos a la información de cada uno de los genes relacionados por la PCOP y su nivel de correlación), un vínculo a la interfaz gráfica interactiva que nos permite estudiar la PCOP en detalle y la imagen generada que representa la PCOP sobre la que se aplica la distribución de clusters [Ilustración 3.18].

Identificar los genes que promueven los cambios fenotípicos

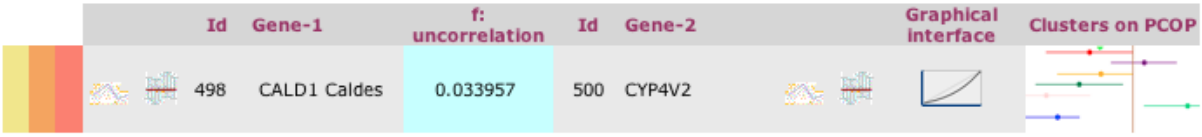


Ilustración 3.18: En la parte superior vemos la cabecera de la relación de PCOPs que podemos estudiar en la interfaz de detalle. En el listado de PCOPs, los elementos que se muestran (de izquierda a derecha) son: Los vínculos y la información del primer gen, el grado de correlación entre los genes relacionados, la información y los vínculos del segundo gen, un icono que nos muestra la tipología de la PCOP y que abre la interfaz gráfica interactiva que nos permite estudiar la PCOP en detalles y la imagen que representa la aplicación de los clusters sobre la PCOP.

La información que obtenemos de los genes cuya relación es descrita por la PCOP son su identificador numérico (*Id*) y su nombre *(Gene*). Los vínculos asociados a cada gen nos permiten acceder, por un lado, a un listado con los genes más correlacionados con cada gen y, por otro lado, a la base de datos del NCBI (National Center for Biotechnology Information) que indica si el gen es un gen marcador.

La siguiente ilustración [Ilustración 3.19] es una captura de la interfaz de detalle mostrando parte de la lista de PCOPs sobre las que se ha aplicado la distribución de clusters generada por el método de clustering SOTA (k=14) que han superado la criba del 40% de clusters entre dos puntos de curvatura.

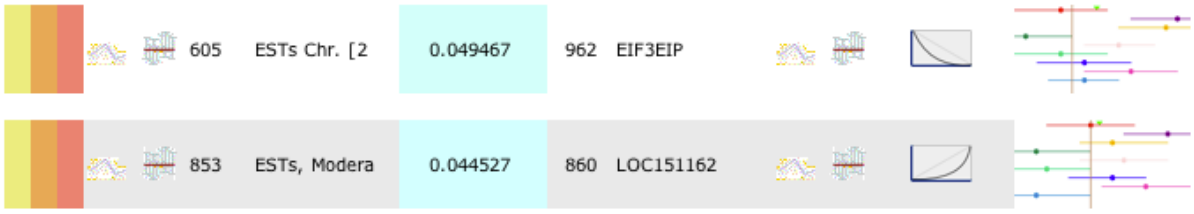


Ilustración 3.19: Captura de la interfaz de detalle de la lista de PCOPs que han superado la criba del 40% de clusters entre dos puntos de curvatura. En ella vemos 2 PCOPs: PCOP(ESTs Chr.,EIF3EIP) y PCOP(R12025, LOC151162) .

La imagen que vemos en la columna *Graphical interface* es un icono que describe la tipología de la PCOP. Además se trata de un vínculo que nos permite acceder a la interfaz gráfica interactiva para estudiar la PCOP en detalle:

Identificar los genes que promueven los cambios fenotípicos

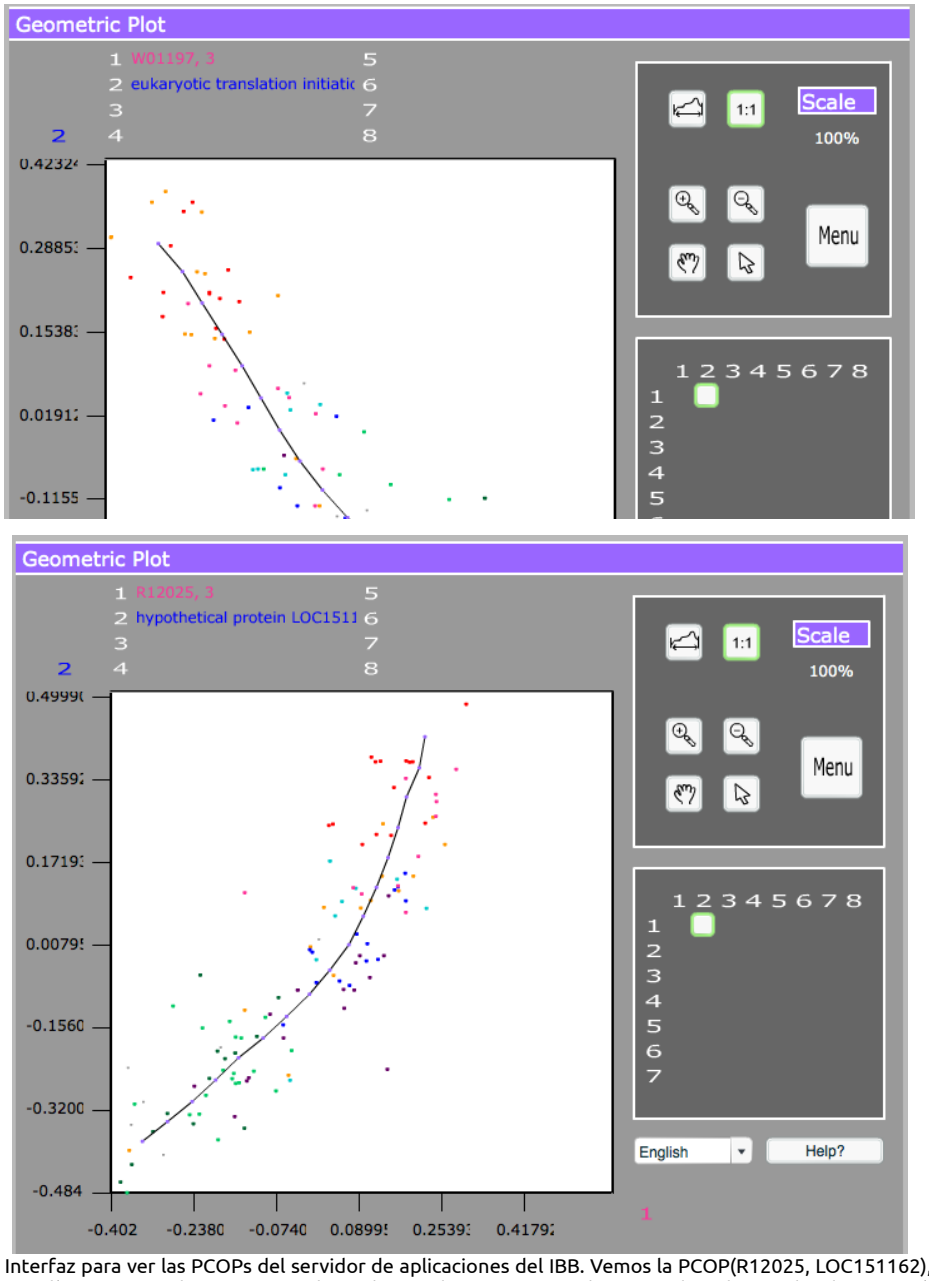


Ilustración 3.22: Interfaz para ver las PCOPs del servidor de aplicaciones del IBB. Vemos la PCOP(R12025, LOC151162), la segunda en la Ilustración 3.19. La línea negra es la PCOP, envuelta en las condiciones muestrales que indican los niveles de expresión de los genes, cuyos nombres encontramos en la parte superior. Las condiciones muestrales aparecen coloreadas según el cluster al que han sido vinculadas (morado, rojo, verde, naranja, cyan, esmeralda, azul y rosa) según la distribución de clusters generada por el método de clustering SOTA (k=14).

Los botones que encontramos al final de la interfaz [Ilustración 3.23] de detalle sirven para modificar la plantilla de colores con la que serán coloreadas las muestras en la interfaz que nos permite ver la representación de una PCOP [Ilustración 3.21, Ilustración 3.22].

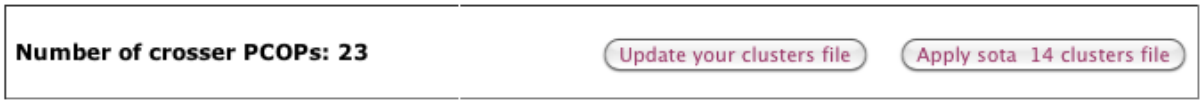


Ilustración 3.23: Pie de la interfaz de detalle donde encontramos dos botones: El primero nos permite asignar a nuestro usuario la distribución de clusters que estamos estudiando como distribución de clustrers por defecto y, el segundo, nos mostrará o la distribución que el usuario tenga por defecto o la distribucón de clusters que estamos estudiando al abrir la interfaz gráfica interactiva para el estudio de PCOPs.

El primero de los dos botones (*Update your cluster file*) nos permite asignar a nuestro usuario del servidor de aplicaciones del IBB la distribución de clusters que queramos estudiar como distribución de clusters por defecto [Ilustración 3.23]

El segundo botón (*Apply sota 14 cluster file/Apply your cluster file*) permite escoger qué distribución de clusters deseamos ver representada en la interfaz gráfica interactiva para el estudio de PCOPs. Las distribuciones de clusters entre las que podemos escoger son: La distribución de clusters vinculada al usuario por defecto o la distribución de clusters que estamos estudiando en este momento [Ilustración 3.23. 3.24 y 3.25].

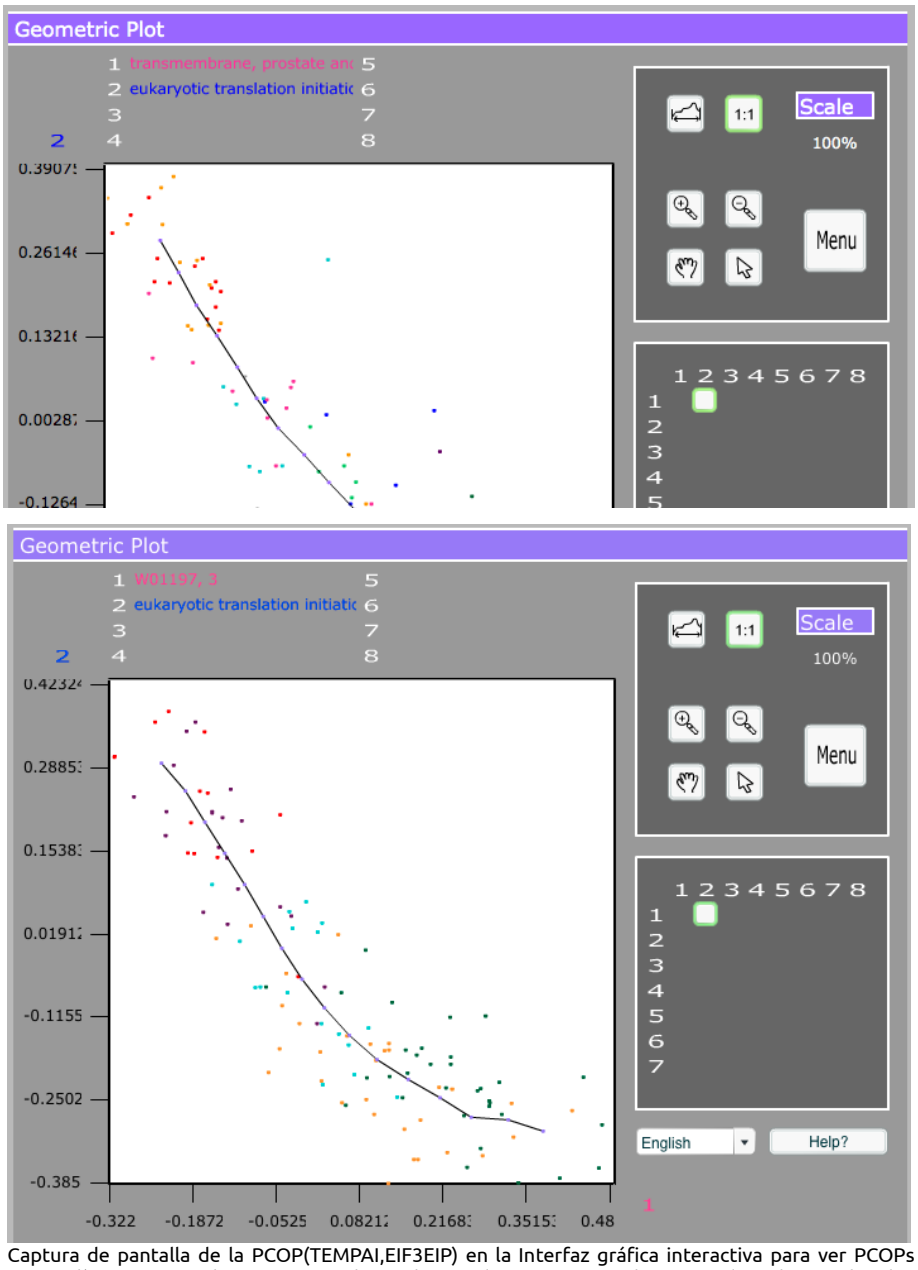


Ilustración 3.25: Captura de pantalla de la PCOP(TEMPAI,EIF3EIP) en la Interfaz gráfica interactiva para ver PCOPs del servidor de aplicaciones del IBB. La línea negra es la PCOP, envuelta en las condiciones muestrales que indican los niveles de expresión de los genes, cuyos nombres encontramos en la parte superior. En este caso, las condiciones muestrales aparecen coloradas según al cluster al qué han sido vinculadas por la distribución de clusters por defecto del usuario (Mds SOTA (k=5)).

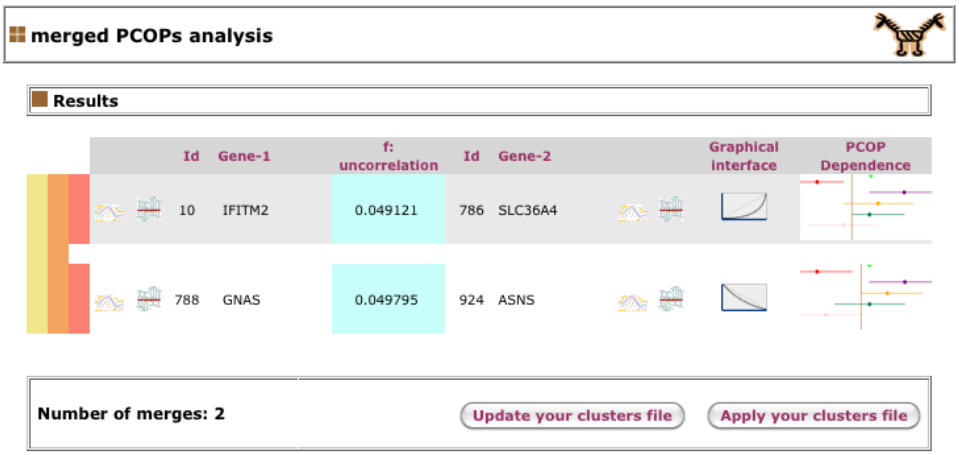


Ilustración 3.26: Captura de la vista con una lista de PCOPs. Vemos, a la izquierda 3 columnas que nos indican la AGRUPACIÓN (amarillo), la ORDENACIÓN (naranja) y el LAYOUT (rojo) de cada una de las PCOPs. En la figura podemos ver como estas dos PCOPs pertenecen a la misma AGRUPACIÓN y a la misma ORDENACIÓN pero a un LAYOUT distinto.

En el margen izquierdo del listado de PCOPs [Ilustración 3.26] encontramos tres columnas coloreadas. La primera columna (amarilla) nos aglutina aquellas PCOPs que tienen la misma AGRUPACIÓN. La segunda columna (naranja) nos aglutina aquellas PCOPs que, dentro de la AGRUPACIÓN, tienen la misma ORDENACIÓN. Finalmente, la tercera columna (roja) nos aglutina aquellas PCOPs que, dentro de la misma AGRUPACIÓN y la misma ORDENACIÓN, tienen el mismo LAYOUT.

Una AGRUPACIÓN es el grupo de PCOPs que tienen, en la misma porción de curva entre dos puntos de curvatura consecutivos de la PCOP, los mismos clusters de la distribución de clusters, sin importar el orden de aparición de los mismos.

Cada AGRUPACIÓN se organiza en ORDENACIONES que agrupan aquellas PCOPs que tienen, no solamente los mismos clusters en la misma porción de curva entre dos puntos de curvatura consecutivos de las PCOPs, sino que los tienen en el mismo orden.

Asimismo cada ORDENACIÓN se organiza en LAYOUT que agrupan aquellas PCOPs cuyos clusters de la distribución de clusters presentan las misma relación con sus vecinos, es decir, que tienen el mismo layout.

4. Análisis de resultados

4.1. Análisis de la concordancia entre la agrupación jerárquica de PCOPs por AGRUPACIÓN, ORDENACIÓN y LAYOUT y la imagen que muestra la distribución de clusters sobre la PCOP.

En la siguiente imagen [Ilustración 4.12] vemos una lista de las PCOPs que han superado la criba del 40% de clusters entre dos puntos de curvatura sobre las que se ha aplicado la distribución de clusters generada por el método de clustering SOTA (k=6) que ha realizado inclusión de outayers.

Ilustración 4.12: Captura de la vista con una lista de PCOPs. Vemos, a la izquierda 3 columnas que nos indican la AGRUPACIÓN(amarillo), la ORDENACIÓN (naranja) y el LAYOUT (rojo) de cada una de las PCOPs. En la figura podemos ver como las dos PCOPs pertenecen a la misma AGRUPACIÓN y a la misma ORDENACIÓN pero a un LAYOUT distinto. La cuarta y la quinta PCOP además pertenecen a una misa AGRUPACIÓN, a la misma ORDENACIÓN y al mismo LAYOUT pero la tercera,a pesar de pertenecer a la misma AGRUPACIÓN y a la mismo ORDENACIÓN pertenece a un LAYOUT distinto.

En la imagen [Ilustración 4.12] vemos 6 AGRUPACIONES diferentes de clusters en cada una de las porciones de la curva formadas por dos puntos de curvatura consecutivos de la

PCOP.

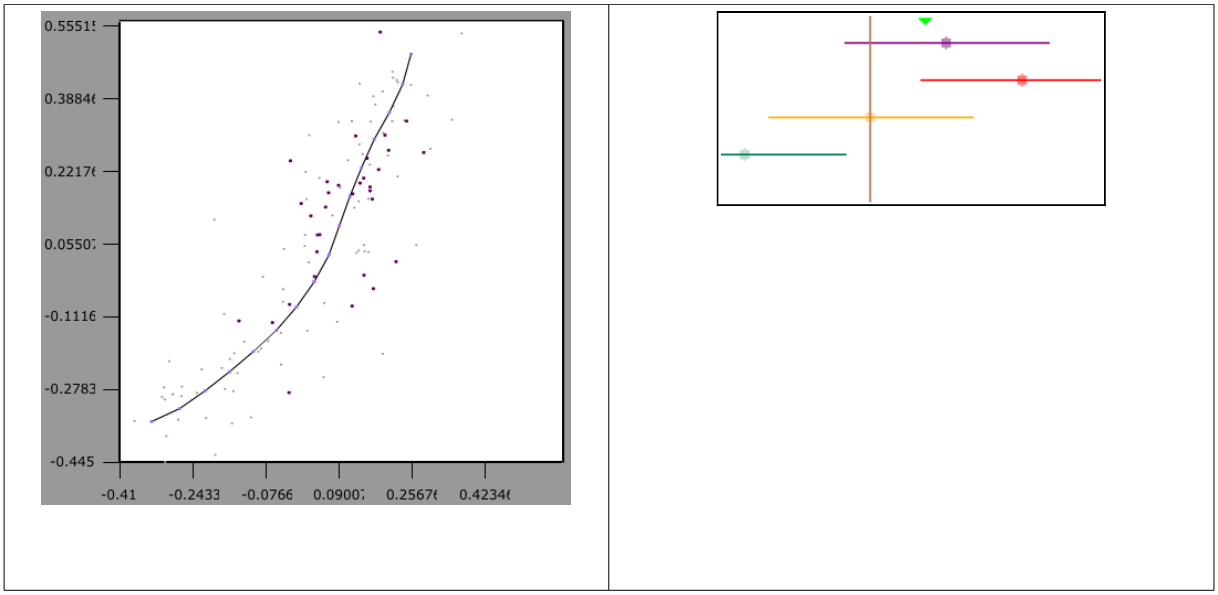
Vemos que la primera AGRUPACIÓN coloca, en la primera porción de curva, el cluster morado y el cluster verde. En la segunda porción de curva coloca los clusters rojo, naranja y esmeralda. Esta AGRUPACIÓN sólo contiene una ORDENACIÓN que agrupa, a su vez, dos tipos de LAYOUT distintos. El primer tipo de LAYOUT contiene una PCOP (PCOP(IFITM2,SLC36A4)). El segundo LAYOUT contiene una PCOP (PCOP(GNAS,ASNS)).

La segunda AGRUPACIÓN coloca, en la primer porción de curva, los clusters morado, verde y esmeralda. En la segunda porción de curva coloca el cluster rojo y el cluster naranja. Esta AGRUPACIÓN sólo contiene una ORDENACIÓN, que agrupa dos tipos de LAYOUT distintos. El primer tipo de LAYOUT reúne una única PCOP (PCOP(715,TPM1)) mientras que el segundo tipo de LAYOUT reúne dos PCOPs (PCOP(BCAR3,LEPR) y PCOP(MGC1437,715)).

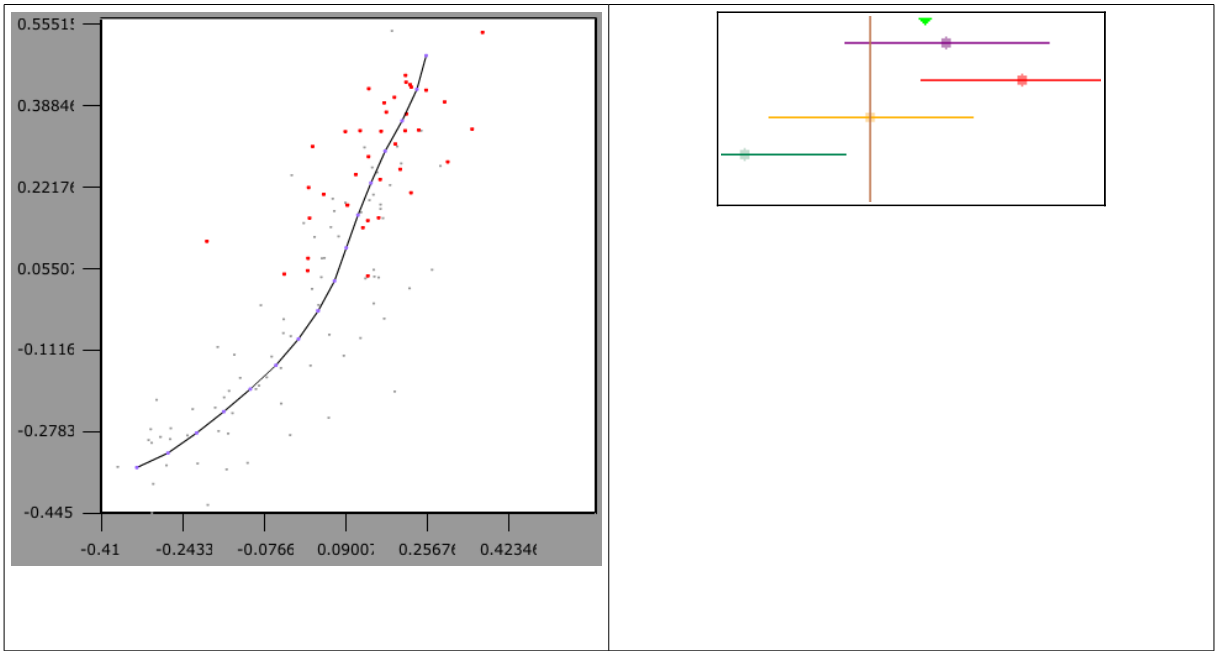
Las tres siguientes AGRUPACIONES contienen una única ORDENACIÓN, que a su vez agrupa un único tipo de LAYOUT. En cada uno de estos LAYOUTs sólo se encuentra una PCOP.

4.2. Análisis de la concordancia entre la imagen que muestra la PCOP sobre la que se ha aplicado una distribución de clusters y la interfaz gráfica interactiva para el estudio de PCOPs, para la misma distribución de clusters y la misma PCOP

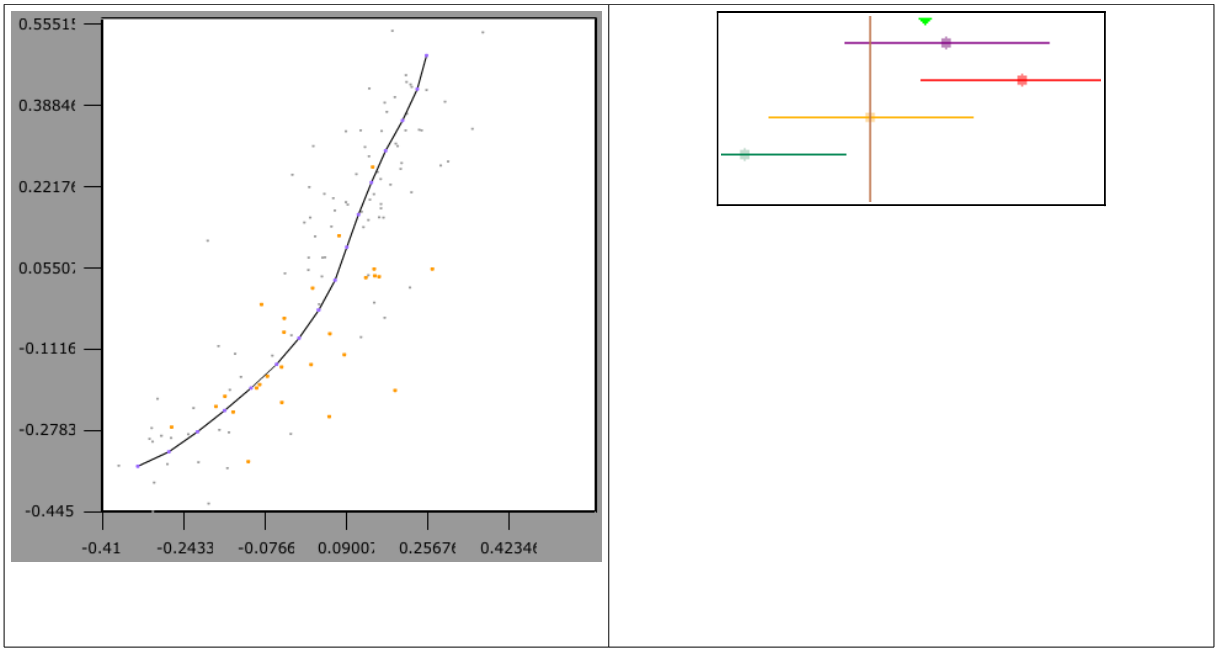
En los cuatro grupos de imágenes que siguen podemos observar la concordancia entre la distribución de clusters que podemos ver en la interfaz gráfica interactiva para el estudio de PCOPs y el gráfico generado de una PCOP sobre la que se aplica una distribución de clusters.



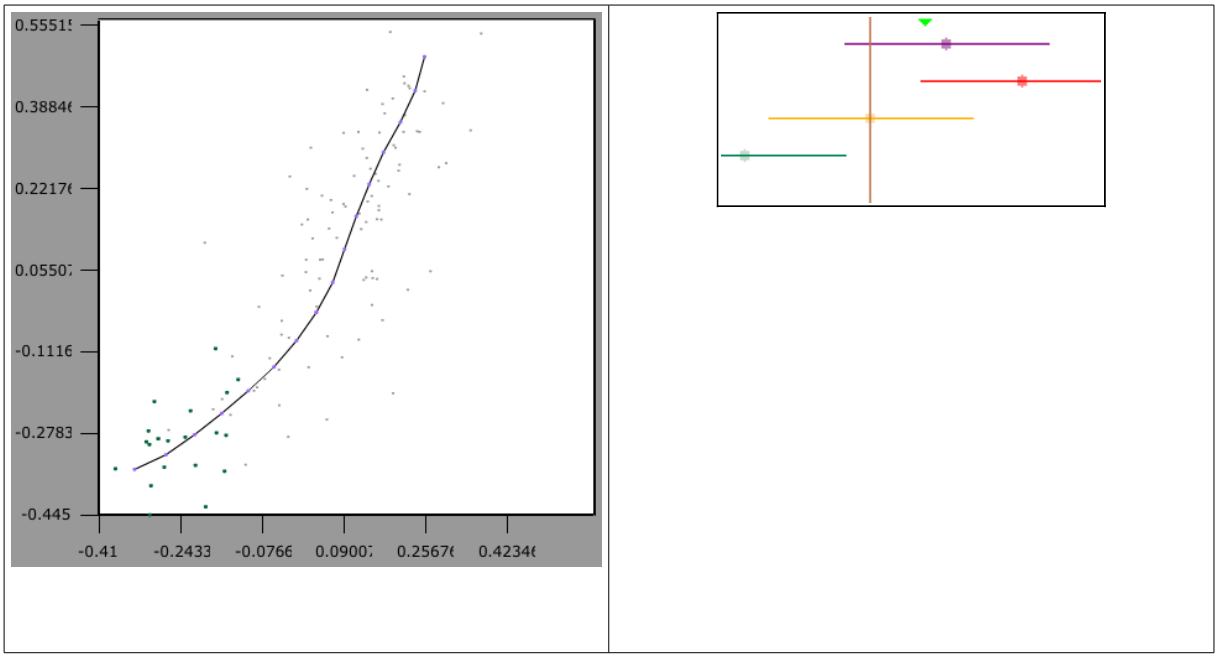
En la imagen de la izquierda vemos la PCOP(SLC36A4,IFITM2) en la que se ha coloreado el primer cluster (morado) de la distribución de clusters generada por el método de clustering Mds Som (k=4). En la imagen de la derecha, la línea morada representa la ubicación del primer cluster de la distribución de clusters sobre la PCOP.



En la imagen de la izquierda vemos la PCOP(SLC36A4,IFITM2) en la que se ha coloreado el segundo cluster (rojo) de la distribución de clusters generada por el método de clustering Mds Som (k=4). En la imagen de la derecha, la línea roja representa la ubicación del segundo cluster de la distribución de clusters sobre la PCOP.



En la imagen de la izquierda vemos la PCOP(SLC36A4,IFITM2) en la que se ha coloreado el tercer cluster (naranja) de la distribución de clusters generada por el método de clustering Mds Som (k=4). En la imagen de la derecha, la línea naranja representa la ubicación del tercer cluster de la distribución de clusters sobre la PCOP.



En la imagen de la izquierda vemos la PCOP(SLC36A4,IFITM2) en la que se ha coloreado el cuarto cluster (verde) de la distribución de clusters generada por el método de clustering Mds Som (k=4). En la imagen de la derecha, la línea verde representa la ubicación del cuarto cluster de la distribución de clusters sobre la PCOP.

4.3. Análisis cualitativo de las cotas de los filtros de cluster y de PCOP

Llamaremos *filtro de cluster por número de POPs* (FCpP) al filtro que determina si un cluster de la distribución de clusters habita entre dos puntos de curvatura consecutivos de la PCOP en base al número de POPs del cluster entre dichos puntos de curvatura.

Llamaremos *filtro de PCOP por número de clusters válidos* al filtro que determina si un mínimo de clusters de la distribución de clusters habita entre dos puntos de curvatura consecutivos de la PCOP. Diferenciaremos entre los tres niveles de exigencia que se aplican en este filtro (F1, F2, F3).

Partimos de las siguientes cotas para los filtros:

- F1: 50 % de los clusters
- F2: 90% de los clusters
- F3: 75% de los clusters
- FCpP: 80% de los clusters

Para expresar los valores de los filtros se usará la expresión F1-F2-F3/FCpP, sustituyendo las abreviaturas por su valor. El caso de partida es 50-75-90/80.

Veamos la PCOP(BFAR,LOC151162) sobre la que se ha aplicado la distribución de clusters generada por Mds Pam Descarte (k=7) obtenida con el juego de parámetros 50-75-90/80:

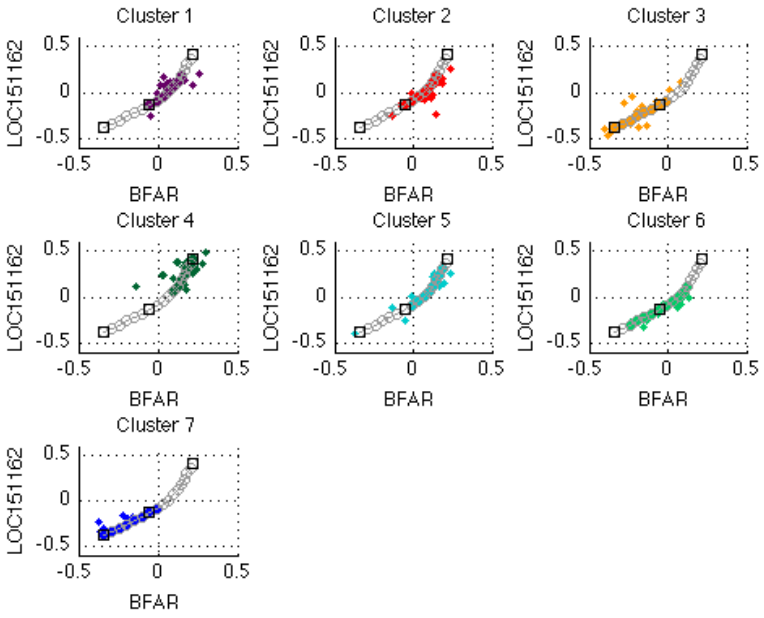


Ilustración 4.1: PCOP(BFAR,LOC151162) sobre la que se ha aplicado la distribución de clusters generado por el método de clustering Mds Pam Descarte (k=7). Cada uno de los 7 gráficos muestra, en gris, la PCOP y sus POPs. Además, se muestran los POPs vinculados a cada cluster de la distribución de clusters y las muestras de la microarray que conforman estos clusters (ambos coloreados en el mismo color). Los cuadrados negros representan la ubicación de los puntos de curvatura de la PCOP.

La representación generada por el programa que generar el gráfico de una PCOP sobre la que se ha aplicado una distribución de clusters es:

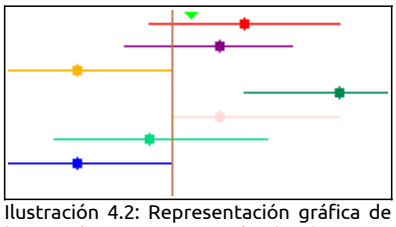


Ilustración 4.2: Representación gráfica de la PCOP(BFAR,LOC151162) sobre la que se ha aplicado la distribución de clusters generado por el método de clustering Mds Pam Descarte (k=7).

Es una relación con un 57% de clusters habitando en un intervalo de curvatura formado por dos puntos de curvatura consecutivos de la PCOP. Es decir, que 4 de sus 7 clusters habitan entre dos puntos de curvatura y los otros 3 no.

Las siguientes relaciones han sido generadas con el juego de parámetros 45-65-80/80. Ninguna de ellas tiene un mínimo del 50% de clusters habitando en un intervalo de curvatura formado por dos puntos de curvatura consecutivos de la PCOP:

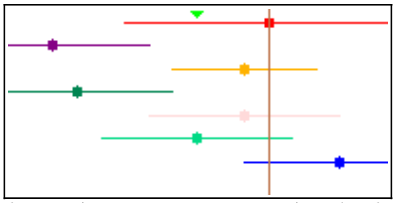


Ilustración 4.3: Representación de la PCOP(BCAR3,LEPR) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Mds Kmeans Dunn (k=7).

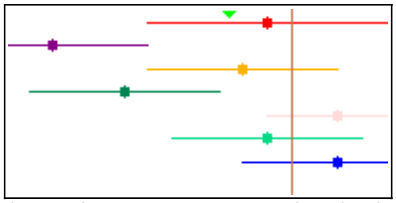


Ilustración 4.4: Representación de la PCOP(NYD-SP21,PLK2) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Mds Kmeans Dunn (k=7).

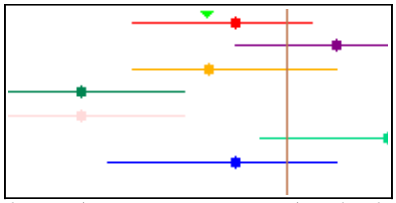
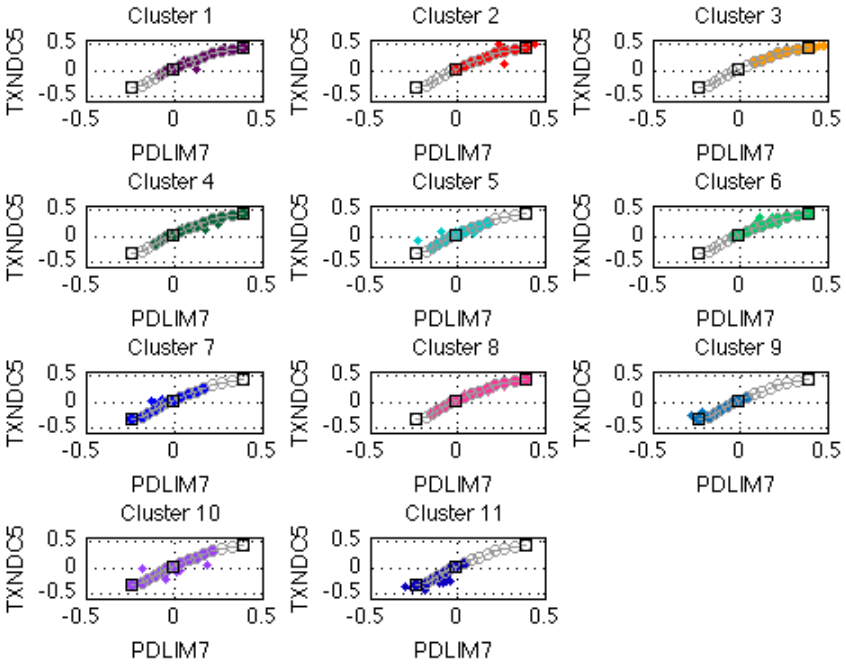


Ilustración 4.5: Representación de la PCOP(STMN4,TXNDC5) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Mds Pam Descarte Dunn (k=7).

Comparando estas tres representaciones de PCOPs [Ilustración 4.3, 4.4 y 4.5] con la representación de la PCOP(BFAR,LOC151162) [Ilustración 4.2], vemos que la calidad de estas 3 últimas, respecto a la PCOP(BFAR,LOC151162), es muy similar. Sin embargo, estas tres ultimas PCOPs, teniendo el 42% de los clusters habitando en un intervalo formado por dos puntos de curvatura consecutivos de la PCOP, son descartadas.

Una situación similar sucede con las distribuciones de clusters consideradas de baja integridad. Viendo la PCOP(RHOC,TXNDC5) sobre la que se ha aplicado la distribución de los clusters generada por el método de clustering Som Descarte (k=14):

Identificar los genes que promueven los cambios fenotípicos



Ilustracion 4.6: PCOP(PDLIM7,TXNDC5) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Som Descarte (k=15). En cada uno de los 11 gráficos vemos, representada en negro, la PCOP(PDLIM7,TXNDC5) y los POPs que la forman. Además, se muestran los POPs vinculados a cada cluster de la distribución de clusters y las muestras de la microarray que conforman estos clusters. Los cuadrados negros representan los puntos de curvatura de la PCOP.

Y la representación generada por el programa que representa gráficamente una PCOP sobre la que se ha aplicado una distribución de clusters es:

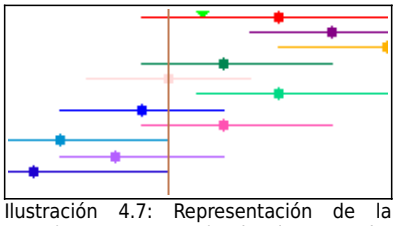


Ilustración 4.7: Representación de la PCOP(PDLIM7,TXNDC5) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Som Descarte (k=15).

Descartadas la PCOP(PDLIM7,TXNDC5) [Ilustración 4.8] y la PCOP(TLHS,TPM1) [Ilustración 4.9] sobre las que se ha aplicado la distribución de clusters generada por el método de clusterng Sota Descarte (k=10) son, a nivel cualitativo, muy similares a la anterior [Ilustración 4.7]:

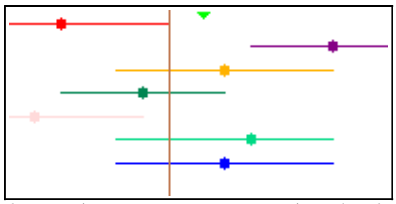


Ilustración 4.8: Representación de la PCOP(PDLIM7,TXNDC5) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Sota Descarte (k=10).

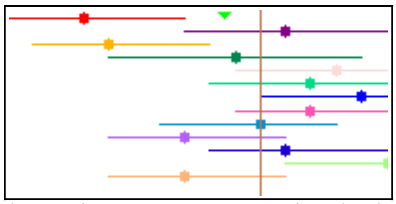


Ilustración 4.9: Representación de la PCOP(TLHS,TPM1) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Som (k=14).

Con este análisis, a nivel cualitativo, considerar reducir un 10% el nivel de las tres versiones del *filtro de PCOP por número de clusters validos* (F1, F2, F3) es una buena opción ya que conseguimos obtener nuevas PCOPs válidas.

Si en vez de considerar reducir el nivel de los *filtros de PCOP por número de clusters válidos* considerásemos reducir el nivel del *filtro de cluster por número de POPs* un 10%, el patrón de las nuevas PCOPs sobre las que se ha aplicado una distribución de clusters que aceptaríamos (con el conjunto de parámetros 50-75-90/60) sería:

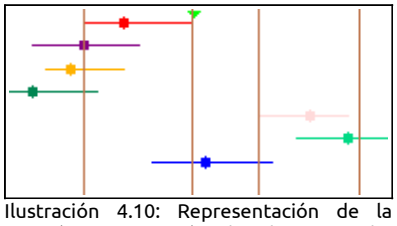


Ilustración 4.10: Representación de la PCOP(GALNT2,POLE) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Mds Kmeans Dunn (k=7).

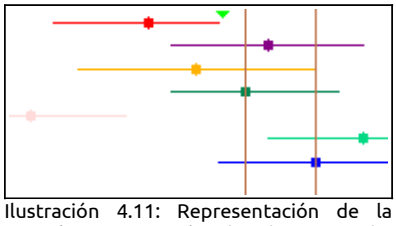


Ilustración 4.11: Representación de la PCOP(785,FLJ20232) sobre la que se ha aplicado la distribución de clusters generada por el método de clustering Sota Descarte (k=10).

Al ver cómo se distribuyen los clusters de las distribuciones de clusters sobre las dos PCOPs [Ilustración 4.10, 4.11] observamos la pérdida de calidad para diferenciar fenotipos. Por este motivo, descartamos reducir el nivel del *filtro de cluster por número de POPs* y mantenerlo al 80%.

4.4. Análisis cuantitativo de las cotas de los filtros

Para el análisis cuantitativo se han realizado cuatro lanzamientos distintos del preproceso con cuatro conjuntos de parámetro diferentes. Siguiendo la misma notación que en

el apartado anterior (F1-F2-F3/FCpP), el número de PCOPs sobre las que se ha aplicado una distribución de clusters que han superado cada uno de los lanzamientos son:

	F1	F2	F3
50-75-90/80	8.856	2.460	101
40-65-80/80	23.441	18.860	9.921
50-75-90/60	9.581	4.779	1.386
40-65-80/60	23.906	21.083	16.954

De un total de 61.152 PCOPs sobre las que se ha aplicado una distribución de clusters.

El porcentaje que corresponde a cada uno de los conjuntos de parámetros en relación al total (61.152) de las PCOPs sobre las que se ha aplicado una distribución de clusters es:

	F1	F2	F3
50-75-90/80	0,00%	0,00%	0,00%
40-65-80/80	164,69%	666,67%	9722,77%
50-75-90/60	8,19%	94,27%	1272,28%
40-65-80/60	169,94%	757,03%	16686,14%

Cambiar la cota del filtro de cluster por número de POPs del 80% inicial al 60% no es justificable ya que vemos en F1 que con esta reducción no aumentamos significativamente el número de PCOPs aceptadas. Eso quiere decir que rebajando al 60% la cota del filtro de cluster por número de POPs para el filtro de PCOP por número de clusters validos más permisivo (F1) estamos aceptando las mismas PCOPs válidas que usando el 80%, con lo que no es conveniente bajar la cota del filtro de cluster por número de POPs del 80%.

5. Informe técnico

5.1. Entorno de trabajo

El desarrollo del proyecto se ha realizado haciendo uso de herramientas open-source. Su desarrollo se ha llevado a cabo en un MacBook Pro a nivel local, probado en un entorno gnu/linux (Mandriva 2006) y desplegado en un segundo entorno gnu/linux (CentOs 5).

Para el entorno local de desarrollo se han usado las siguientes herramientas:

- Entorno de desarrollo Komodo Edit.
- Paquete de desarrollo web MAMP que incluye servidor Apache configurado para la interpretación de código PHP.
- Compilador de C++ v 4.2.1 para Mac OS X Snow Leopard.

Para la compilación del código fuente se han usado los siguientes recursos:

- Compilador de C++ v 4.0.1 para el entorno de pruebas.
- Librería PNGwriter 0.5.4 en el entorno de prueba.
- Compilador de C++ v 4.1.2 para el entorno de producción.
- Librería PNGwriter 0.5.4 en el entorno de producción.

5.2. Estructura del servidor

A continuación vamos a describir de que forma está configurado el servidor, la arquitectura de directorios y los ficheros de datos que se usan.

5.2.1. Estructura de directorios

La estructura de directorios definida en el servidor unifica los criterios de acceso a los datos por parte de todos los procesos que se puedan ejecutar. De este modo, todos y cada uno de los procesos que se ejecutan saben dónde se encuentran los datos que deben recoger y dónde guardar los datos generados.

La arquitectura de directorios del servidor tiene una configuración fija. Esto nos permite hacer uso de direcciones relativas en vez de direcciones absolutas para acceder a los datos. Poder usar direcciones relativas nos brinda la posibilidad de mover la aplicación de directorio o de servidor sin que se deban reconfigurar los distintos procesos que la componen.

En el directorio raíz se instalan todos los paquetes de software, directorios de procesos y de datos relativos a aquel directorio. Por lo tanto los procesos acceden a los datos a través de una ruta relativa al directorio raíz para facilitar la movilidad de la aplicación.

Vemos a continuación la representación de la arquitectura de directorios:

```
/ ( directorio raíz )
|
|__ microarray ( directorio de datos )
|           |
|           |__ mXX ( directorio raíz de datos de la microarray XX )
|           |           |
|           |           |
|           |           |__ nonlineal ( directorio raíz de ficheros de las relaciones no lineales )
|           |           |           |
|           |           |           |__ normal
|           |           |           |
|           |           |           |__ HeighF
|           |           |           |
|           |           |           |__ HeighFfiltered
|           |           |           |
|           |           |           |__ RH_F10filtered
|           |           |           |
|           |           |           |__ glcl-rel
|           |           |           |
|           |           |           |__ glcl-gif
|           |           |           |
|           |           |           |__ ...
|           |           |
|           |           |__ RClustering_Samples
|           |
|           |__ mYY ( directorio raíz de datos de la microarray YY )
|           |           |
|           |           |__ ...
|           |
|           |__ ...
|
|__ fullcorrelations ( directorio del programa lanzadora y de los ejecutables del servidor )
                        |
                        |__ compile ( directorio donde se guardan los códigos fuente )
                                    |
                                    |__ cldistribution2gif
                                    |
                                    |__ PCOPSamples-Glcl
```

Los directorios marcados en negrita (*glcl-rel* y *glcl-gif*) son los directorios creados por el preproceso y donde guarda los datos generados.

Dentro del directorio de salida *glcl-rel* se crea una carpeta con el nombre Best. Además se crearán tantas carpetas cómo métodos de clustering se analicen. Dentro de la carpeta Best se crearan tantos directorios como distribuciones de clusters de alta integridad se analicen.

Dentro del directorio de una distribución de clusters de alta integridad se guardarán los tres ficheros que corresponden a los tres niveles de exigencia del proceso de filtrado. Dentro de la carpeta de un método de clustering se guardan tres ficheros correspondientes a los tres

niveles de exigencia del proceso de filtrado por cada una de sus distribuciones de clusters consideradas de baja integridad.

5.2.2. Preproceso para la realización del cruce de datos

5.2.2.1. Ficheros de entrada de distribuciones de clusters para el proceso de cruce

Las distribuciones de clusters generadas por los métodos de clustering se encuentran descritas en los ficheros con extensión *.colors*, almacenadas en el directorio *Rclustering_Samples*.

La primera columna es el identificador del cluster al que se ha vinculado la condición muestral de la microarray que encontramos en la segunda columna.

```
[...]
1 109
1 110
1 117
1 118
2 24
2 25
2 26
[...]
```

Ejemplo de fichero .colors. Este fragmento corresponde al fichero 17_sota_4.colors.

La nomenclatura de estos ficheros se rige por:

<id microarray>_<nombre del método de clustering>_[descarte]_<k>.colors

Donde *id microarray* es el número de microarray que se analiza en el servidor. El *nombre del método de clustering* es el nombre del método de clustering con el que se analiza la microarray. La palabra, opcional, *descarte* aparecerá si el método de clustering aplica descarte de outlayers en vez de inclusión de outlayers. Finalmente, *k* será el valor del parámetro *k* aplicado al método de clustering.

5.2.2.2. Ficheros de entrada de PCOPs para el programa de cruce

Usamos tres tipos distintos de ficheros para obtener toda la información necesaria sobre las PCOPs:

- Los ficheros *.cluster* son los homólogos de los ficheros *.colors* de las distribuciones de clusters. En ellos se describe cómo se relacionan las condiciones muestrales de la microarray con los POPs, siguiendo la estructura:

```
1;
8 11 12 13 14 19 25 28 29 30 31 33 34 35 36 [...]
2;
8 11 12 13 14 17 18 19 21 25 28 29 30 31 33 [...]
3;
1 7 8 11 12 13 14 17 18 19 21 24 25 28 29 30 [...]
[...]
```

Ejemplo de fichero .cluster. Este fragmento corresponde al fichero g325g768h0.75d0.3.cluster

- Los ficheros *.ldom* son los que describen el comportamiento de la curva, subministrando los puntos de curvatura y la distribución de las muestras de la microarray en los intervalos formados por estos puntos.

```
[...]
        popFiCorbes1 0 11 14

segmentsCorbes1

        SEGMENT: 1 - 10 (97 mostres)
        DOMINI: 1 2 3 4 5 6 7 8 9 10 [...]

        SEGMENT: 12 - 13 (64 mostres)
        DOMINI: 2 3 39 41 42 43 46 47 [...]
```

Ejemplo de fichero .ldom. Este fragmento corresponde al fichero g325g768h0.75d0.3.ldom

- Los ficheros *.output* son los que describen los POPs de la PCOP. Se trata de una relación de los diferentes atributos de los POPs de los cuales sólo hacemos uso de la posición x e y (en espacio 2d).

```
[...]
0 -0.160770 1.817211 0.389830 0.005739 -0.059686 -0.080715 0.821323 0.570464
0 -0.090311 1.819229 0.423729 0.004624 -0.007564 -0.033304 0.821323 0.570464
0 -0.020469 1.978163 0.500000 0.003625 0.043390 0.014462 0.728729 0.684802
0 0.049392 2.212299 0.525424 0.003511 0.091806 0.064824 0.694250 0.719734
[...]
```

Ejemplo de fichero .output. Este fragmento corresponde al fichero g325g768h0.75d0.3.output.

Todos estos ficheros los encontramos agrupados en cuatro directorios dentro del directorio *nonlienal*. Estos cuatro directorios nos agrupan las PCOPs según su correlación y su curvatura. Los directorios son:

- Normal
- HeightF
- HeighFfiltered
- RH_F10filtered

5.2.2.3. Ficheros de salida del programa de cruce de las distribuciones de clusters con las PCOPs

Los ficheros de salida que genera el preproceso representan la organización AGRUPACIÓN-ORDENACIÓN-LAYOUT descrita anteriormente.

La primera línea (tabulación 0) nos indica el número de AGRUPACIONES que encontramos en el fichero.

La segunda línea (tabulación 1) nos indica el número de ORDENACIONES contenidas en esta AGRUPACIÓN.

La tercera línea (tabulación 2) nos indica el número de LAYOUTS dentro de la ORDENACIÓN.

La cuarta línea (tabulación 3) nos indica el número de PCOPs que se han agrupado en esta ARUPACIÓN-ORDENACIÓN-LAYOUT, y las encontramos en las siguientes líneas.

```
13 # Num. Agrupaciones
        3 # Num. Ordenaciones
                        1 # Num. Layouts
                                1 # Num. Items
                                        10-786
                        1 # Num. Layouts
                                1 # Num. Items
                                        648-732
                        1 # Num. Layouts
                                1 # Num. Items
                                        700-732
                        1 # Num. Ordenaciones
                                1 # Num. Layouts
                                        [...]
```

Ejemplo de fichero de salida. Este fragmento corresponde al fichero glcl-17_sota_descarte_dunn_12-f1.txt.

La nomenclatura de los ficheros de salida del preproceso sigue la estructura:

glcl-<id microarray>_<nombre del método de clustering>_[descarte]_<k>-[f1|f1|f3].txt

Donde *id microarray* es el número de microarray que se analiza en el servidor. El *nombre del método de clustering* es el nombre del método de clustering con el que se ha analizado la microarray. La palabra, opcional, *descarte* aparecerá si el método de clustering ha aplicado descarte en vez de inclusión. Donde *k* tomará el valor del parámetro *k* del método de clustering. Finalmente encontraremos *f1*, *f2* o *f3* según el nivel de exigencia del proceso de

filtrado.

5.2.2.4. Estructura de directorios

El preproceso genera una estructura de directorios en los directorios de salida *glcl-rel* y *glcl-gif*. Esta estructura de directorios consiste en diferenciar los resultados obtenidos de las distribuciones de cluters de alta integridad y las de baja integridad.

```
/glcl-rel
|
|_ HeighF
|
|_ HeighFfiltered
|
|_ normal
|           |
|           .
|           |
|           |_ 17_pc_som
|           |
|           |_ 17_pc_som_descarte
|           |
|           |_ 17_sota
|           |           |
|           |           |_ glcl-17_sota_10-f1.txt
|           |           |
|           |           |_ glcl-17_sota_10-f2.txt
|           |           |
|           |           |_ glcl-17_sota_10-f3.txt
|           |           |
|           |           |_ glcl-17_sota_11-f1.txt
|           |           |
|           |           .
|           |
|           |_ 17_sota_descarte
|           |
|           |_ Best
|
|_ RH_F10filtered
```

Para las distribuciones de clusters de baja integridad, se crea una carpeta (en la carpeta de filtro por correlación de las PCOPs) con el nombre del método de clustering que las ha generado y se guarda en ella todos los ficheros de salida.

Para las distribuciones de clusters de alta integridad se sigue una estrategia similar. Primero se crea una carpeta Best en el directorio de filtro por correlación de las PCOPs. Para cada distribución de clusters de alta integridad se creará, dentro de la carpeta Best, una carpeta con su nombre y en ella se guardaran los ficheros de salida.

5.2.2.5. Argumentos del programa de cruce de las distribuciones de cluster con las PCOPs

El preproceso que se ha generado, cuyo nombre de ejecutable es PCOPSamples-Glcl, necesita 3 parámetros para su correcta ejecución:

1. El identificador de la microarray.
2. El número de genes que se analizan en la microarray.
3. El número de condiciones muestrales que se analizan en la microarray.

Con estos tres parámetros el programa PCOPSamples-Glcl es capaz de recopilar los datos para realizar el cruce y guardar los resultados en la ubicación adecuada.

El conjunto completo de parámetros del programa PCOPSamples-Glcl es:

Modificador	Descripción
-a	Permite modificar el límite del filtro de PCOP por número de clusters 1.
-b	Permite modificar el límite del filtro de PCOP por número de clusters 2.
-c	Permite modificar el límite del filtro de PCOP por número de clusters 3.
-d	Permite modificar el límite del filtro de cluster por número de POPs.
-e	Permite modificar el numero de clusters (min) que deben encontrarse en intervalos de curvatura diferentes (filtro de PCOP por clusters en intervalos).
-f	Con este parámetro se generan los gif de todas las relaciones evaluadas.
-g	Con este parámetro no se genera ningún gif.
-h	Número de genes en la microarray.
-i	Permite modificar el límite para considerar una relación entre clusters Intersección o Unión.
-l	Permite modificar el límite del filtro de la tabla de relación (5%).
-m	Identificador de la microarray.
-o	Genera un fichero de estadísticas y un fichero de tablas.
-v	Activa el modo verbose.
-w	Número de casos muestrales en la microarray.
-x	Ruta donde guardar los ficheros de salida.

5.2.2.6. Programa lanzadora

El programa lanzadora es el encargado de ejecutar ordenadamente todos los procesos

de cálculo y agrupación de cada microarray. Este programa se ejecuta bajo demanda desde la web cada vez que se carga una microarray al sistema. Debido a la gran envergadura del proyecto tanto en procesos como en datos, este programa es el encargado de mantener el control y el orden de ejecución de los diferentes procesos. También es el encargado de crear los directorios de cada proceso para los cálculos y los resultados y de realizar la limpieza de todos aquellos directorios y ficheros temporales que se creen durante su ejecución.

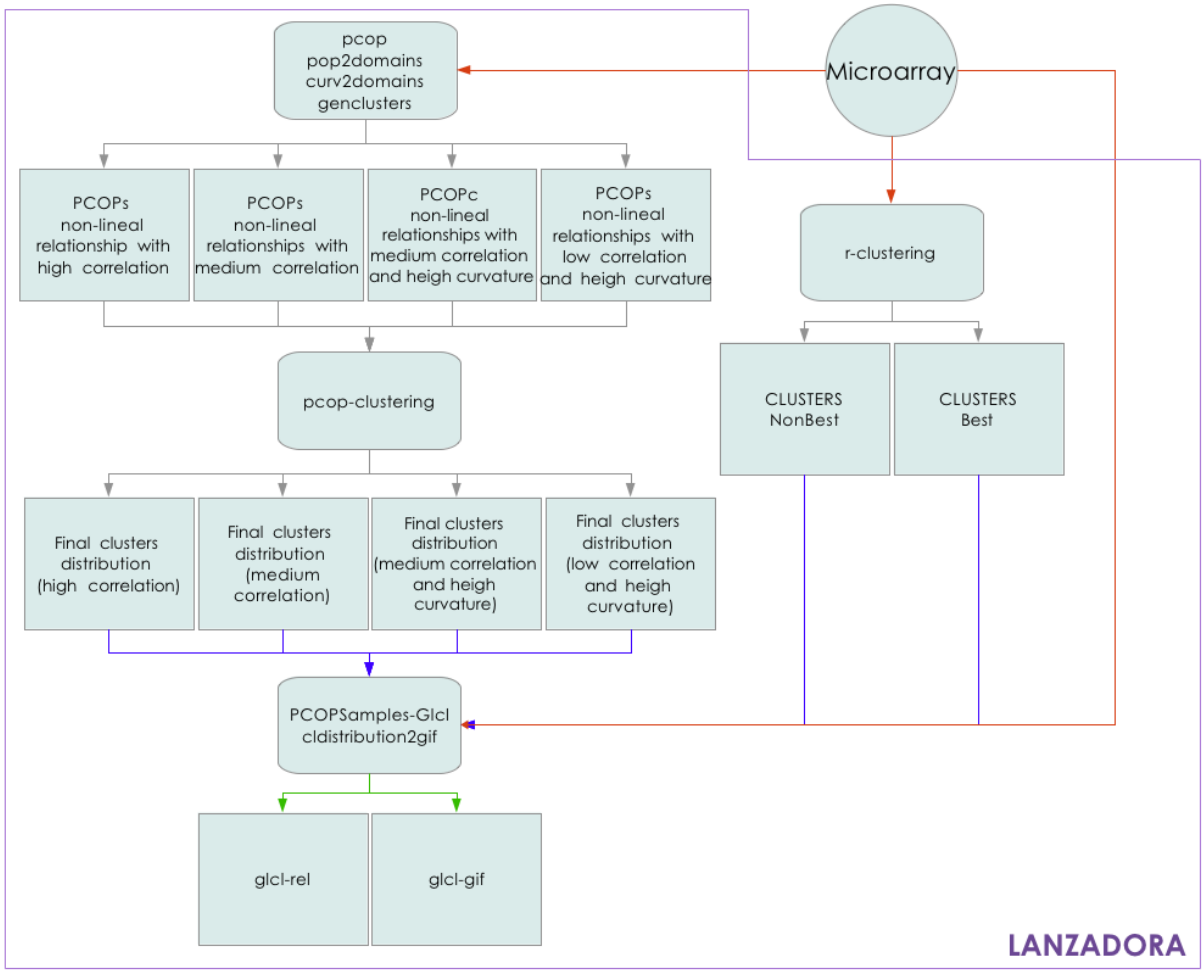


Ilustración 4.1: Esquema de ejecución del programa lanzadora para la aplicación de las distribuciones de clusters generadas por los métodos de clustering sobre las PCOPs dada una microarray.

5.2.3. Programa para generar el gráfico de una PCOP sobre la que se ha aplicado una distribución de clusters

El programa cldsitribution2gif genera un gráfico de una PCOP sobre la que se aplica una distribución de clusters. Esta distribución de clusters sobre una PCOP no es más que una lista de puntos de inicio y final. El conjunto completo de parámetros del programa cldistribution2gif es:

Modificador	Descripción
-i	Lista de datos a dibujar.
-c	Lista de los puntos de curvatura a incluir en la imagen.
-o	Nombre de la imagen de salida.
-s	Dimensiones de la imagen de salida.
-b	Lista de la intensidad de los centros de los datos de entrada.

-l	Color de fondo de la imagen.
-m	Tendencia de la distribución.

Pasamos a explicar, de forma más detallada, cada uno de los parámetros:

-i		**No opcional**
• Lista que contiene la información de las agrupaciones (clusters) a dibujar (separados por coma (',')). • Las diferentes agrupaciones están separadas por el símbolo más ('+'). • Las agrupaciones pueden tener longitud 2 ó 3. ◦ Longitud 2: inicio, fin ◦ Longitud 3: inicio, central, fin		
	Ejemplos	
• Formato de la agrupación: -0.1,0.2,0.3 • Formato de lista de agrupaciones: -0.1,0.2,0.3+0.3,0.5,0.6+0.8,0.94,0.97		

-o		**No opcional**
• Nombre del archivo de salida. • Sin extensión		
	Ejemplos	
• Formato del parámetro: /home/usuario/img		

-s		**No opcional**
• Dimensión de la imagen. • En formato *string*. • Si no se especifica, se toma por defecto 200x100.		
	Ejemplos	
• Formato del parámetro: 500x200		

-t		Opcional
• Punto de mayor densidad.		

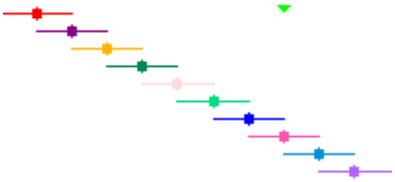
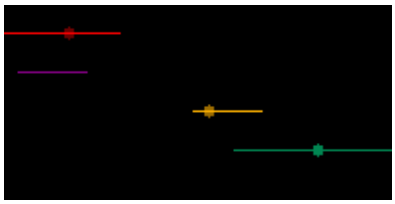
-c		Opcional
• Lista de puntos de curvatura. • Separados por coma (',').		
	Ejemplos	

• Formato del parámetro: 2,5,7		
-r		Opcional
• Lista de porcentajes de opacidad de los puntos centrales de las agrupaciones. • Sólo se aplica si las agrupaciones son de longitud 3. • Separados por coma (',').		
	Ejemplos	
• Formato del parámetro: 0.15,0.15,0.50,1.0		

-b		Opcional
• Cambia el color de fondo, por defecto blanco, a negro.		

-p		Opcional
• Cambia el formato en el que se guarda la imagen, por defecto gif, a png.		

Ejemplos y resultados de lanzamiento del programa cldistribution2gif:

Llamada	./cldistribution2gif -i 1,2,3+2,3,4+3,4,5+4,5,6+5,6,7+6,7,8+7,8,9+8,9,10+9,10,11+10,11,12 -o sample01 -t 9	./cldistribution2gif -i 0.1,0.22,0.31+0.13,0.19,0.25+0.44,0.467,0.561+0.51,0.66,0.789 -r 0.15,0.15,0.50,1.0 -o sample05 -b -p
Resultado		

5.2.4. Aplicación web CrossingClusters

5.2.4.1. Ordenación dinámica en la interfaz general donde se muestran las diferentes distribuciones de clusters para cada método de clustering

Las distribuciones de clusters que aparecen listadas en la aplicación web CrossingCusters, se ordenan, por defecto, según el valor del campo que indica el número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 65% entre dos puntos de curvatura (columna 8 de la tabla). Pero pueden ser ordenados por cualquier de los siguientes campos:

- Valor del parámetro *k* aplicado al método de clustering .

- Nombre del método de clustering.
- Número alcanzado de clusters en la distribución de clusters.
- Indicativo de si el método de clustering ha aplicado descarte o inclusión de outlayers.
- Valor de integridad aplicando el método *silhouette* a la distribución de clusters.
- Valor de integridad aplicando el método *dunn* a la distribución de clusters.
- Número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 40% de clusters entre dos puntos de curvatura.
- Número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 65% de clusters entre dos puntos de curvatura.
- Número de PCOPs sobre las que se ha aplicado la distribución de clusters que han superado la criba del 80% de clusters entre dos puntos de curvatura.

5.2.4.2. Estructura de directorios de la aplicación web CrossingClusters

En el servidor de aplicaciones del IBB tenemos por un lado los programas ejecutables que se encargan del análisis de la microarray y por otro las aplicaciones web para el estudio de los datos generados.

En /var/www/cgi-bin/pcop encontramos todos los datos que hacen referencia a las microarrays (dentro del direcotio microarray) y todos los ejecutables usados en su análisis (en fullcorrelation). También encontramos el código fuente de los ejecutables (fullcorrelation/compile).

En /var/www/html/aplic encontramos la aplicación web que nos permite acceder a las herramientas web del servidor de aplicaciones del IBB. En /var/www/html/aplic/gexp es donde se han añadido los ficheros .php y el directorio de recursos glcl necesarios para el correcto funcionamiento de la aplicación web CrossingClusters.

También en /var/www/html/aplic/gexp tenemos un enlace (microarray) al directorio microarray que se encuentra en /var/www/cgi-bin/pcop de manera que accediendo de forma relativa a este enlace nos encontramos en disposición de obtener todos los datos que sean necesarios.

El preproceso que se ha confeccionado a lo largo del proyecto guarda, en el directorio

nonlineal de la microarray, los datos que genera, creando dos ubicaciones (*glcl-rel* y *glcl-gif*). La aplicación web CrossingClusters accede a ellos mediante el enlace microarray descrito en el párrafo anterior. La aplicación web del IBB suministra a CrossingClusters el identificador de la microarray que se esta analizando, así como el identificador del usuario que realiza el estudio. Esta trasferencia se realiza a través de parámetros GET de php.

Con la estructura del servidor y accediendo de forma relativa a los datos permitimos que la aplicación web se pueda mover de directorio o de servidor sin afectar a su funcionamiento.

6. Conclusiones

Después de haber realizado este proyecto, podemos afirmar que se han cumplido todos los objetivos que se habían propuesto inicialmente.

Los objetivos referentes al cruce de datos se han cumplido satisfactoriamente ya que se ha generado el preproceso que lleva a cabo el cruce entre las distribuciones de clusteres proporcionadas por los métodos de clustering y las PCOPs. De todo el cúmulo de aplicaciones de distribuciones de clusters sobre las PCOPs se filtran aquellas que nos aportan información útil para la búsqueda de cambios fenotípicos. Para cada aplicación de una distribución de clusters que ha superado el proceso de filtrado, se llama al programa cldistribution2gif para generar el gráfico de la aplicación de sus clusters sobre la PCOP. Además, el listado de aplicaciones de distribuciones de clusters que supera el proceso de filtrado queda agrupado por: si los clusteres de la distribución aparecen juntos entre dos puntos de curvatura de diferentes PCOPs, si aparecen en el mismo orden, y si aparecen con el mismo grado de intersección entre ellos (separación, intersección, unión).

Respecto al programa cldistribution2gif, podemos decir que se cumplen todas las expectativas, ya que el gráfico generado con la aplicación de una distribución de clusters sobre una PCOP, en formato gif, muestra la ubicación y extensión sobre la PCOP de cada uno de los clusters, muestra el punto de la PCOP con mayor densidad de muestras de cada cluster y este punto se dibuja con una intensidad de color que representa el número de muestras del cluster. También se sitúan sobre la imagen, los puntos de curvatura de la PCOP, además del punto de la PCOP con mayor densidad de muestras.

Finalmente, la aplicación web creada, CrossingClusters, nos ofrece escoger qué método de clustering deseamos analizar, generando un listado de las distintas ejecuciones del mismo (con distintos valores de k) con las distribuciones de clusters obtenidas para cada una. Para cada una de las ejecuciones podemos acceder a una lista de detalle donde encontramos las relaciones de expresión que han llevado a cabo una transición entre los fenotipos descritos por la distribución de clusters. Estas relaciones entre genes se muestran, además, agrupadas por participar en la transición entre unos mismos fenotipos. En cada una de las relaciones se muestra la información que concierne a la PCOP implicada y la imagen que ilustra la aplicación de la distribución de clusters sobre la PCOP.

Este proyecto me ha resultado muy interesante porque he podido aplicar gran parte de los conocimientos adquiridos a lo largo de la carrera en un proyecto real, además me ha permitido adentrarme en la biomedicina, un campo aparentemente ajeno a la informática. La parte que me ha resultado más interesante ha sido la de desarrollar el preproceso para el cruce

de datos explicado en las secciones anteriores, por obligarme a adquirir nuevos conocimientos, tanto estadísticos como biológicos. Por otro lado, la parte que me ha resultado más tediosa ha sido la de diseño web, ya que es un campo de la informática en el que ya había trabajado previamente.

7. Trabajo futuro

A nivel biomédico, después de ver los resultados analizados en el capítulo Análisis de resultados, considerar realizar un estudio que valide el aumento de la cota del *filtro de cluster por número de POPs* del 80% al 90%.

Actualmente, las cotas de los *filtro de PCOP por número de clusters válidos* (F1, F2 y F3) son impuestos de manera estática e independientes de la cantidad y del tipo de los datos de microarray que se analicen. Una mejora plausible sería calcular las cotas de esos tres lanzamientos del filtro de forma dinámica, de tal manera que se tuviese en cuenta la naturaleza de los datos de microarray que se estén procesando.

Desde un punto de vista computacional, el algoritmo aplicado podría ser fácilmente paralelizado de tres formas distintas:

1. Paralelizar el preproceso de manera que se lanzasen, a la vez, tantas instancias del preproceso cómo métodos de correlación de PCOPs existan, de manera que no se procesaran secuencialmente.

2. Paralelizar el proceso de cruce que se realiza dado un método de clustering para poder cruzar más de un método de clustering a la vez.

3. Combinación de la primera y la segunda.

8. Bibliografía

[1] Página web oficial del Instituto de Biotecnoloía y Biomedicina de la Universidad Autónoma de Barcelona, http://www.ibb.uab.es

[2] Delicado, P. and Huerta, M. (2003): 'Principal Curves of Oriented Points: Theoretical and computational improvements'. Computational Statistics 18, 293-315.

[3] 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.

[4] Huerta, M. Cedano, J. Querol, E. (2008) Analysis of nonlineal relations between expressions profiles by the principal curves of oriented.points approach. J. Bioinform Comput Biol 6: 367-386.

[5] Delicado, P.(2001) Another look at principal curves and surfaces. Journal of Multivariate Analysis, 77, 84-116.

[6] 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. Navigation trhough non-continuous gene-expression relationships

[7] 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., 9;10:138

[8] Web server for on-line microarray analysis suported by the Institute of Biotechnology and Biomedicine of the Autonomous University of Barcelona (IBB-UAB), http://revolutionresearch.uab.es

Identificar los genes que promueven los cambios fenotípicos

Carles Hernández Ferrer

_____

Bellaterra, 12 de setiembre de 2011

Identificar los genes que promueven los cambios fenotípicos

Identificar los genes que promueven los cambios fenotípicos

Resum

El projecte que es presenta en aquesta memòria està composat per dos executables y una aplicació web. L'objectiu del treball realitzat és cercar i trobar aquells gens que promouen els canvis fenotípics, el rol que porten a terme aquests gens en la transisió i entre quins fenotips s'està canviant. A més de disenyar l'aplicació web que permetrà l'estudi dels resultats obtinguts.

Resumen

El proyecto que se presenta en esta memoria está compuesto por dos ejecutables y una aplicación web. El objetivo del trabajo realizado es buscar y encontrar aquellos genes que promueven los cambios fenotípicos, el rol que llevan a cabo estos genes en la transición y entre qué fenotipos se está cambiando. Además de diseñar la aplicación web que permitirá el estudio de los resultados obtenidos.

Summary

The project introduced in this report is composed by two executables and a web application. The aim is to detect and find the genes who promote the phenotypic changes, the role of those genes in the transition and between which phenotypes they are changing. Another aim of the project is to design and build the web application that allows the study of the results.