Universitat Autònoma de Barcelona

escola d'enginyeria

INGENIERÍA INFORMÁTICA

MEMÓRIA PREVIA DEL PROYECTO:

3899 BIOINFORMÁTICA:

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

Firma del estudiante	Firma del director/a o director/es
Nombre: Raquel Guardia Villalba Fecha: 12/01/2011	Nombre/s: Jordi Gonzàlez y Mario Huerta Dpt: Ciencias de la computación Fecha: 12/01/2011

1. Objetivo/s del proyecto

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

La tecnología de microarrays permite obtener el nivel de expresión de un gran número de genes bajo un gran número de circunstancias diferentes.

El servidor web del que se dispone para el análisis de éste tipo de datos opera actualmente con microarrays del orden los 3.000 genes. Para estas microarrays genera grafos interactivos vía web que muestran las relaciones entre los genes y permiten operar con ellos. Para ello opera con un layout de dos niveles, uno para los genes pertenecientes a cada cluster y otro para los clusters de genes. Esta aproximación por niveles permite una visión global y también en detalle del grafo. Sin embargo, la aplicación no está preparada actualmente para mostrar grafos de 10.000 o 30.000 genes.

El objetivo de este proyecto consiste entonces en modificar la aplicación actual para poder trabajar con matrices de expresión génica de mayor orden. Esto implicará también adaptar el cálculo del layout para la microarray dada, así como la interfaz web que muestra los resultados y permite operar con ellos.

Al tener un gran número de genes será necesario aumentar el nivel del layout, es decir, pasar de trabajar con un layout de dos niveles a trabajar con uno de tres. Tendré entonces por una parte los genes, por otra los clusters de genes e incorporé los hiperclusters, que no son más que clusters de clusters.

2. Breve introducción al estado del arte del tema propuesto

La bioinformática es la ciencia dedicada al estudio de los fenómenos biológicos desde un punto de vista computacional con el objetivo de ofrecer métodos robustos para la comprensión, simulación y predicción de comportamientos biológicos observados en los seres vivos. De esta manera los principales esfuerzos de investigación en este campo incluyen el alineamiento de secuencias, la predicción de genes, el alineamiento estructural de proteínas, la predicción de estructura de proteínas, el estudio de la expresión génica o las interacciones proteína-proteína entre otros.

Una constante en proyectos de este tipo es el uso de herramientas matemáticas para extraer información útil de datos producidos por técnicas experimentales de alta productividad, como la tecnología de microarrays.

En este sentido, una de las técnicas utilizadas para determinar el significado biológico de los niveles de expresión de una determinada secuencia de DNA es el análisis de microarrays. Como se ha citado anteriormente, las microarrays o matrices de expresión génica son, como su propio nombre indica, matrices en las que encontramos diferentes genes frente a diversas condiciones muestrales. De esta manera se analizan los niveles de expresión de los genes y se puede determinar en cada momento cuál es el conjunto de genes que se está expresando.

El gran problema de las microarrays es que los datos que proporciona son tantos y tan complejos que es difícil abstraer el significado biológico, es decir, analizarlos. Para conseguir este objetivo se han diseñado diferentes aplicaciones (GEO [1], BIOREL [2], ArrayExpress [3], MicroGen [4], etc.) pero ninguna de ellas conduce a una visión holística de lo que sucede en la célula.

PCOPGene-Net [5] es una aplicación web creada por el IBB (Instituto de Biotecnología y Biomedicina) pensada para facilitar el estudio de las relaciones entre las expresiones génicas bajo las condiciones de las microarrays que se analicen. Por medio de la visión global que esta facilita se muestra la red de relaciones entre los genes y mediante la visión en detalle se muestran las variaciones de la relación de expresión de los genes en detalle. Esta aplicación se alberga en el servidor de aplicaciones web para el análisis de microarrays del IBB: http://revolutionresearch.uab.es [13]

Las librerías JUNG de Java para la visualización de datos en grafos vía web [6] son utilizadas para montar el gráfico interactivo de la visión global. En el proceso de *layout*, los genes se colocan en el espacio 2D en función del grado de correlación de cada gen con sus vecinos, agrupando los genes en clusters y facilitando la exhibición del *minimum-spanning path* entre cualquier par de genes de la microarray. Esta disposición de los datos facilita al investigador la navegación a través de la nube de genes de la microarray y le proporciona la información de su interés en cada momento.

En la actualidad, estas herramientas son utilizadas para el estudio de la progresión de tumores desde un punto de vista holístico [5][7][8][9][10][11].

3. Estudio de viabilidad del proyecto

Como he comentado anteriormente, dispongo de un servidor web que opera actualmente con microarrays del orden los 3.000 genes. Para estas microarrays genera grafos interactivos que muestran las relaciones entre los genes y permiten operar con ellos. Para ello trabaja con un layout de dos niveles, uno para los genes pertenecientes a cada cluster y otro para los clusters de genes.

Dado que la nueva aplicación trabajará con un layout de tres niveles, será necesario modificar el algoritmo actual. Ahora tendré que agrupar los genes no solo por clusters sino por hiperclusters.

Por otra parte, el hecho de tener un layout de tercer nivel implicará muy probablemente cambiar la interfaz web que muestra los resultados. Hasta ahora, en un mismo applet se mostraban los genes y los clusters a los que pertenecían. En la figura 1 pueden observarse los genes de cada cluster pintados con un color diferente.

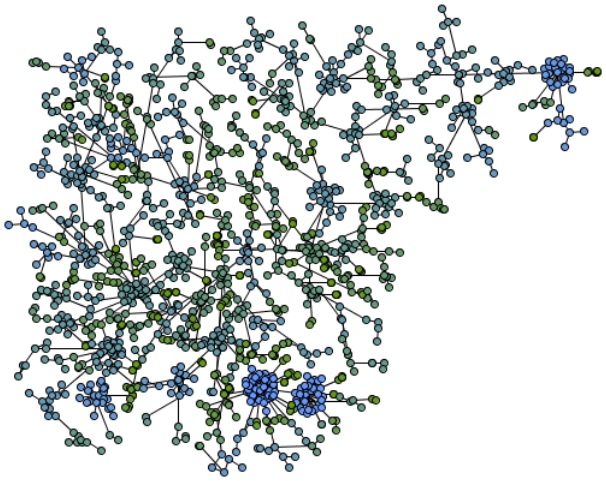


Figura 1: Diferenciación de clusters por colores.

Ahora, al tener un número tan grande de genes es posible que no puedan mostrarse todos en un mismo applet. Por ello tendré inicialmente un applet en el que se mostrarán los hiperclusters y al seleccionar uno de ellos se abrirá una nueva ventana que mostrará los clusters y los genes que pertenecen al hipercluster seleccionado.

Dado que la interfaz web sufrirá estas modificaciones, también será necesario modificar las operaciones que la misma presenta. Esto es debido a que el applet principal trabajará con hiperclusters por lo que, al seleccionar cualquier opción, tendremos que tener en cuenta que cada operación a realizar en el applet de los hyperclusters se tendrá que mostrar también a nivel de clusters y genes, y que las operaciones en el applet de los clusters de un hypercluster pueden afectar a los applets con los clusters del resto de hyperclusters. Dicho de otra manera, todos nuestros applets se tendrán que comunicar.

Finalmente comentar que cuento con todos los recursos necesarios para la realización de este proyecto, motivo por el cual considero que es viable llevarlo a cabo en el tiempo estimado.

4. Planificación temporal del trabajo

Dividiré este proyecto en seis fases:

Fase 1: Introducción. En la primera etapa del proyecto me centraré en la introducción al campo de la bioinformática así como en la primera toma de contacto con el método y código de la aplicación de partida. De esta forma, los objetivos se basan principalmente en la comprensión de los términos biológicos que envuelven el proyecto y en familiarizarse con el algoritmo y código sobre el que se trabajará, concretamente con el algoritmo a dos niveles.

Fase 2: Pruebas de código. En esta fase probaré y corregiré el código de partida para calcular el layout. De esta forma me aseguraré de que el proyecto parte de un código correcto. A continuación compararé la salida de dicho código para una microarray pequeña y otra grande. Con esto lograré ver hasta que punto soporta el método original las grandes microarrays.

Fase 3: Modificación del algoritmo de layout. Una vez tenga todo listo, empezaré con la modificación de la aplicación. En esta fase realizaré el algoritmo del layout a 3 niveles. A continuación comprobaré si el applet actual soporta el nuevo grafo o es necesaria la modificación de la interfaz.

Fase 4: Interfaz web. En esta fase me centraré en la modificación e implementación de las opciones que ofrece la interfaz web. Para ello seguiré las siguientes subfases:

Fase 4.1: Applet de hiperclusters. En este punto me centraré en la creación del applet principal, es decir, el applet de hiperclusters.

Fase 4.2: Adaptación de applets. Una vez creado el applet de hiperclusters tendré que asegurarme de que cada grupo de clusters que forma un hipercluster se abre en su applet correspondiente.

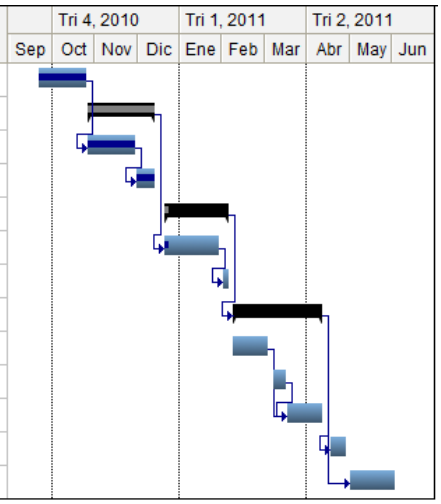
Fase 4.3: Comunicación entre applets. Cuando ya se lancen los applets correspondientes tendré que implementar la comunicación entre ellos.

Fase 5: Optimización. En esta fase optimizaré el trabajo realizado. Siempre que sea posible incluiré mejoras e intentaré que el proyecto sea lo más eficiente posible.

Fase 6: Memoria del proyecto. Finalmente, una vez acabada la implementación del proyecto, realizaré la memoria del mismo.

A continuación muestro un esquema cronológico de las fases que componen el proyecto.

Nombre	Duración	Inicio	Fin
1.- Introducción	25d?	21/09/2010	25/10/2010
2.- Pruebas de código	**35d?**	**26/10/2010**	**13/12/2010**
2.1.- Corrección del código de partida	25d?	26/10/2010	29/11/2010
2.2.- Comparación de microarrays	10d?	30/11/2010	13/12/2010
3.- Modificación del algoritmo del layout	**35d?**	**20/12/2010**	**04/02/2011**
3.1.- Realización del layout a 3 niveles	30d?	20/12/2010	28/01/2011
3.2.- Comprobación	5d?	31/01/2011	04/02/2011
4.- Interfaz web	**47d?**	**07/02/2011**	**12/04/2011**
4.1.- Creación del applet de hiperclusters	20d?	07/02/2011	04/03/2011
4.2.-Adaptación de applets	8d?	08/03/2011	17/03/2011
4.3.- Comunicación entre applets	18d?	18/03/2011	12/04/2011
5.- Optimización	10d?	18/04/2011	29/04/2011
6.- Redacción de la memória	25d?	02/05/2011	03/06/2011



5. Bibliografía

1. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles – database and tools.** *Nucleic Acids Res* 2005, 33:D562-566.

2. Antonov AV, Tetko IV, Mewes HW: **A systematic approach to infer biological relevance and biases of gene network structures.** *Nucleic Acids Research* 2006, 34(1):e6.

3. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, *et al*.: **ArrayExpress – a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res* 2005, 33:D553-555.

4. Burgarella S, Cattaneo D, Pinciroli F, Masseroli M: **MicroGen: a MIAME compliant web system for microarray experiment information and workflow management.** *BMC Bioinformatics* 2005, 6(Suppl 4)**:**S6.

5. Huerta M., Cerdano J., Peña D., Rodriguez A. y Querol E: **PCOPGene-Net: Holistic Characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships**. *BCM Bioinformatics, 2009.*

6. O'Madadhain J, Fisher D, Smyth P, White S, Boey YB: **Analysis and visualization of network data using JUNG.** *Journal of Statistical Software* 2005, VV**:**1-35.

7. Cedano J, Huerta M, Querol E. **NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships** *Advances in Bioinformatics,* vol. 2008

8. Delicado, P. **Another look at principal curves and surfaces.** *Journal of Multivariate Analysis, 77, 84-116*, 2001.

9. Delicado, P. and Huerta, M**.: 'Principal Curves of Oriented Points: Theoretical and computational improvements'.** *Computational Statistics* 18, 293-315, 2003.

10. Huerta M, Cedano J, Querol E: **Analysis of nonlinear relations between expression profiles by the principal curves of oriented-points approach**. *J Bioinform Comput Biol.* 6:367-386. 2008.

11. Cedano J, Huerta M, Estrada I, Ballllosera F, Conchillo O, Delicado P, Querol E**. A web server for automatic analysis and extraction of relevant biological knowledge.** *Comput Biol Med. 37:1672-1675.*2007.

12. Instituto de Biotecnología y de Biomedicina (IBB) de la Universidad Autónoma de Barcelona. http://ibb.uab.es/ibb

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