Ingenieria Informática

Informe Previo del Proyecto

2684
Identificar los genes que promueven los cambios fenotípicos

Bellaterra, 14 de Gener de 2010

Sig. del estudiante	Sig. del director	Sig. del coordinador
Carles Hernández	Mario Huerta IBB	Jordi Gonzàlez dCC

Índice

1. Objetivos del proyecto

El objetivo primario del proyecto es el de desarrollar una aplicación de servidor para el análisis de datos de microarray.

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

Sobre estas condiciones muestrales se pueden aplicar métodos de clustering. De este modo, toda una serie de condiciones muestrales pueden ser consideradas igual por causar el mismo efecto en la expresión de los genes.

Para desarrollar la aplicación seguiré dos líneas de actuación:

- Objetivo 1: Diseñar e implementar un método para encontrar los pares de genes que llevan a cabo los cambios fenotítpicos correspondientes a los diferentes grupos de condiciones muestrales obtenidos por los métodos de clústering más comunmente aplicados a los datos de microarrays.

- Objetivo 2: Desarrollar una aplicación web que pueda ser usada para realizar el anterior cálculo sobre la microarray (subida por el usuario) y mostrar los resultados obtenidos .

Para el primer objetivo se elaborará un método para corroborar la hipótesis que nos induce a pensar que los subespacios muestrales obtenidos a partir del análisis de las dependencias de expresión no lineales entre pares de genes se corresponden con diferentes estados celulares. Una vez corroborada esta hipótesis, el mismo algoritmo se encargará de buscar cuales son los pares de genes que dominan cada uno de los cambios fenotípicos.

Para el segundo objetivo se desarrollará la interfaz web que permita visualizar y navegar de manera gráfica por los resultados obtenidos.

Ambos objetivos se llevarán a cabo haciendo uso de herramientas de software libre.

2. Breve introducción al estado del arte

La bioinformatica es una disciplina científica que utiliza tecnología de la información para organizar, analizar y distribuir información biológica con la finalidad de responder preguntas complejas en biología. Ésta es un área de investigación multidisciplinar, la cual puede ser definida como la comunión entre diversas disciplinas (biología, computación, tecnologías de la información,...).

Uno de los puntos de entrada de datos en la bioinformatica es la tecnología de microarrays. Esta tecnología permite obtener el nivel de expresión de un gran número de genes (entre 10^3 y 10^4) bajo un gran número de condiciones muestrales diferentes (entre 102 y 103). Dadas las características de

la tecnología de microarrays, son el instrumento idóneo para el estudio de la expresión génica bajo diferentes estados celulares.

Como punto de partida en el desarrollo de la aplicación de servidor y la interfaz web, disponemos de las salidas de diversos programas y módulos existentes en el servidor para el análisis de microarrays, desarrollados en el IBB [3][4][5].

Estos módulos encuentran las relaciones no lineales entre los pares de genes mediante el cálculo de las PCOP (Principle Curves of Oriented Points). Las PCOPs, formuladas por Delicado[1][2], nos permiten el estudio de las complejas relaciones de expresión no lineales entre pares de genes.

Otros módulos aplican los métodos de clustering más comúnmente usados para agrupar las condiciones muestrales de una microarray. Los métodos utilizados son:

- Herarchical Clustering
- Self Organizing Map
- Self Organizing Tree Algorithm
- Partitioning Around Medoids

Al cruzar adecuadamente los datos obtenidos por el análisis de las PCOPs con los obtenidos por los métodos de clustering podemos encontrar los pares de genes que guían los cambios fenotípicos representados por estos clusters.

3. Estudio de viabilidad del proyecto

3.1. Viabilidad técnica

Se ha diseñado una estructura de ejecución centralizada para lanzar todos los procesos que deben ser ejecutados cuando llegan nuevos datos de microarray al servidor de aplicaciones web.

Esta estructura de ejecución se centraliza en un único ejecutable de nombre `lanzadera` (`lanzadera.cc`). Este ejecutable se encarga de realizar la llamada a los procesos a realizar, la administración de directorios y la gestión de archivos (mover y renombrar ficheros de salida, borrar ficheros intermedios,...) de forma que los resultados sean accesibles para las interficies web.

El conjunto de procesos que se lanzan desde la *lanzadera* contra las nuevas microarrays a analizar es el *preproceso* del aplicativo, mientras que la interfaz web sería la parte on-line. Diremos, por lo tanto, que el programa lanzadera es el encargado de guiar el *preproceso* de los datos.

Teniendo en cuenta la organización centralizada del *programa lanzadera*, los programas que se desarrollen tienen que ser diseñados para aceptar datos de cualquier tipo de microarray. Esto significa que todo proceso que sea iniciado por la *lanzadera* debe esperar que esta le entregue todos los parámetros que necesite (id de la microarray, ubicación de los ficheros,...).

3.2. Viabilidad Web

El diseño de aplicaciones web se caracterizara por el uso de herramientas libres:

- Sistema Operativo GNU/Linux
- Servidor Apache 2.0
- Interprete de PHP 5
- Compilador de YUI (Yahoo User Interface)

Todas estas ya se encuentran instaladas y operativas en los servidores que el IBB pone a nuestra disposición, descritos en la siguiente sección.

3.3. Viabilidad operativa

Para la realización del proyecto se dispone de dos servidores:

1. Servidor público: (revolutionresearch.uab.es) Este es el servidor público donde se encuentran la versión final del *preproceso* y de las interfaces web. Es donde se suben los nuevos datos, se realiza automáticamente su análisis y se permite navegar por los resultados generados usando las diferentes interfaces web.
2. Servidor de pruebas: Este servidor es usado para el desarrollo de las interficies web.

En ambos servidores existe la misma arquitectura de directorios (podemos verla en la Fig. 1): Encontramos replicada en el servidor de pruebas la arquitectura del servidor público.

Figura 1: Esquema de la arquitectura del servidor público

4. Planificación temporal

El plan de trabajo está formado por cuatro grandes fases consecutivas expuestas a continuación:

1. **Fase 1**: Adquisición de conocimientos previos.
   - *a*) Adquisición de **conocimientos** biológicos relacionados con el ámbito del proyecto.
   - *b*) Análisis de los ficheros de entrada de descripción de **PCOP**s.
   - *c*) Análisis de los ficheros de entrada de descripción de **PCOPClusters** y clústeres globales.
2. **Fase 2**: Algorítmia:
   - *a*) Diseño e implementación delmétodo de **cruce** de **datos**.
   - *b*) Diseño e implementación del programa para la **generación** de **imágenes**.
3. **Fase 3**: Diseño Web.
   - *a*) **Diseño** e implementación de la interfaz **web**.
   - *b*) Periodo de **pruebas** y **optimización** de la aplicación desarrollada.
4. **Fase 4**: Documentación:
   - *a*) **Documentación** técnica sobre el ***preproceso*** (cruce de datos y generador de imágenes).
   - *b*) **Documentación** técnica sobre la aplicación **web**.
   - *c*) **Informe Previo**.
   - *d*) **Memoria**.

Las palabras en negrita en las tareas corresponden con las tareas en el diagrama de la Fig. 2, un diagrama de Gantt que nos permite ver la disposición temporal de las tareas citadas.

Referencias

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

[2] Pedro Delicado, Mario Huerta (2002). Principal Curves of Oriented Points: theoretical and computational improvements. Universitat Politècnica de Calaunya.

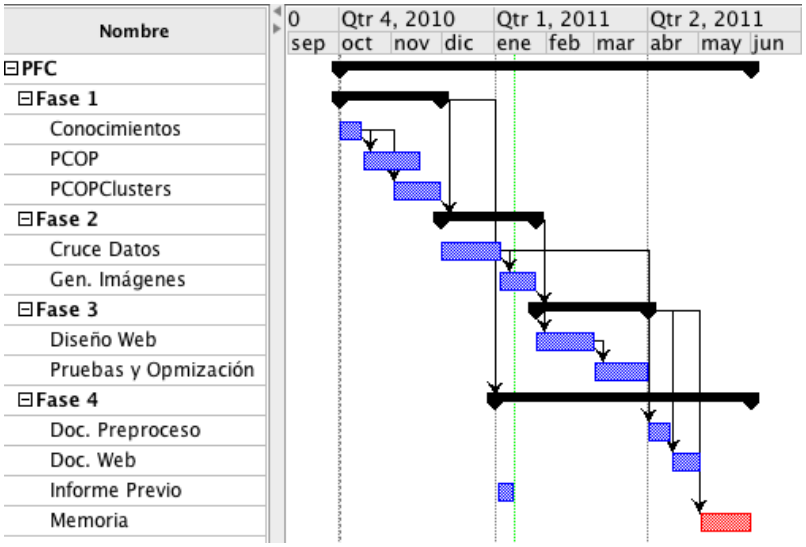

Figura 2: Diagrama de Gantt

[3] Mario Huerta, Juan Antonio Cedano, Enrique Querol (2008) Analysis of nonlinear relations between expression profiles by the principal curves of oriented-points approach. Universitat Autònoma de Barcelona.

[4] Mario Huerta, Juan Antonio Cedano, Enrique Querol (2008) NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships. Universitat Autònoma de Barcelona.

[5] Mario Huerta, Juan Antonio Cedano, Dario Peña, Antonio Rodriguez, Enrique Querol (2009) PCOPGene-Net: Holistic Characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships. Universitat Autònoma de Barcelona