Universitat Autònoma de Barcelona

escola d'enginyeria

INGENIERÍA INFORMÁTICA

MEMORIA PREVIA DEL PROYECTO

2680 Bioinformática:

Búsqueda de genes en el servidor local usando información Biomédica de Bases de Datos Remotas.

Firma del estudiante	Firma del director/a o directores/as
Nombre: David Expósito Pérez Data: 19/1/2011	Nombre/s: Jordi Gonzàlez i Sabatè / Mario Huerta Dpt: CVC IBB Data: 19/1/2011

Objetivos del proyecto

El proyecto que nos abarca se centra en el cruce de los resultados proporcionados por el servidor del IBB destinado al análisis de microarrays con información biomédica procedente de bases de datos remotas.

Las datos de microarrays son matrices de gran tamaño que contienen la expresión de un gran número de genes referente a unas circunstancias determinadas.

En dicho servidor para el análisis de microarrays existe una primera aplicación que realiza dichos cruces y que se compone de tres partes bien diferenciadas:

- La base de datos[3], que recopila toda la información biomédica conocida sobre los genes conocidos.
- El applet java[3], que nos permite escoger los parámetros de búsqueda en la base de datos y lanzar el cruce entre las base de datos de los genes y la microarray que está analizando el usuario.
- La aplicación php[3], que es la encargada de realizar los cruces y mostrar los resultados.

El proyecto consistirá entonces en alcanzar tres objetivos:

Primer objetivo: Optimizar la aplicación php[3] para que sea más rápida y eficiente, proporcionando una salida altamente intuitiva.

Segundo objetivo: Incluir el cruce de información entre un gen que no esté en la microarray con genes sí existentes la microarray. Actualmente el cruce únicamente se realiza entre genes de la misma microarray.

Esto implicará:

- Incluir en la base de datos[3] los diferentes alias del gen y de la proteína de cada gen.
- Modificar el applet java[3] para poder introducir el nombre del gen o proteína (entre otros) en las consultas.

Tercer y último objetivo: Diseñar la nueva base de datos local. Descargar del National Center for Biotechnology Information (NCBI)[2] la información necesaria para cumplimentar los nuevos campos. Diseñar e implementar la actualización mensual de la base de datos local. Este proceso también implicará eliminar de la base de datos la información innecesaria, así como actualizar los diferentes ficheros del servidor local que dependan de las últimas actualizaciones en el NCBI.

Introducción al estado del arte

El National Center for Biotechnology Information (NCBI) es parte de la Biblioteca Nacional de Medicina de Estados Unidos, una rama de los Institutos Nacionales de Salud de Estados Unidos. Está localizado en Bethesda, Maryland y fue fundado el 4 de noviembre de 1988 con la misión de ser una importante fuente de información de biología molecular[2]. Almacena y actualiza constantemente entre otras: la información referente a secuencias genómicas en GenBank, un índice de artículos científicos referentes a biomedicina, biotecnología, bioquímica, genética y genómica en PubMed, una recopilación de enfermedades genéticas humanas en OMIM, información referente a los genes para la identificación de los mismos en Gene y UniGene, y por último la base de datos de GO donde encontramos la ontología de un gen dado, así cómo aquellos genes que disponen de misma ontología que él,

KEGG (Kyoto Encyclopedia of Genes and Genomes), una base de datos de Japón, es especialmente relevante de rutas metabólicas KEGG PATHWAYS. Estos pathways son una colección de mapas dibujados manualmente en los que se representan interacciones moleculares y redes de reacción para:

- Metabolismos
- Información de Procesos Genéticos
- Información de Procesos del Entorno del gen
- Procesos celulares

Todas las bases de datos del NCBI y Kegg están disponibles en línea de manera gratuita. Las bases de datos del NCBI pueden consultadas usando el buscador Entrez y sus herramientas Eutils. Éstas nos permiten obtener la información buscada en ficheros .xml. KEGG dispone también de sus correspondientes herramientas de consulta.

El IBB es un centro de investigación que forma parte de la Universidad Autónoma de Barcelona (UAB). Actualmente disponemos en el IBB de una serie de aplicaciones web para el análisis de microarrays[1][3][4]. El resultado de estos análisis se cruza con las bases de datos biomédicas del NCBI y KEGG[3].

Con los resultados obtenidos de los cruces con las bases de datos biomédicas, los investigadores obtienen información sobre las relaciones entre los genes de la microarray más allá de lo que son las meras relaciones a nivel de expresión génica, que son las que proporciona una microarray. Sin embargo, la aplicación actual solo cruza entre sí los genes de la microarray analizada pero no permite conocer si alguno de estos genes está relacionado con la fase de alguna enfermedad, proteína o gen ausente en la microarray pero de interés para el investigador.

Estudio de viabilidad del proyecto

Puesto que el proyecto consiste en incorporar nuevas funcionalidades y herramientas a las aplicaciones ya existentes en el servidor, este resulta viable ya que las herramientas de trabajo que se necesitan ya están instaladas.

Todo lo podemos realizar con software libre, por lo tanto no tenemos repercusiones de costes. Las herramientas que necesitan los usuarios de la aplicación son herramientas estandarizadas instaladas en cualquier máquina para una correcta navegación por Internet.

Servidor	Cliente
• Sistema de Gestión de Bases de Datos (mySQL) • Java • Compiladores de Perl y PHP • Óptima conexión a Internet • Sistema operativo Linux • Espacio disponible en disco para la posterior ampliación de la base de datos	• Acceso a Internet • Máquina virtual de java • Explorador de Internet • Java

Tendré que estudiar la estructura que tiene el NCBI en el momento de descargar la información mediante FTP, necesario para ampliar nuestra base de datos.

A fecha de hoy he encontrado las posibles optimizaciones que puedo incorporar a la aplicación php existente para el cruce de las bases de datos, como obtener una navegación más intuitiva por los resultados generados y como permitir que cualquier proteína, gen o información biomédica de interés para el investigador, aunque no esté presente en la microarray analizada, pueda ser relacionada con los genes de la misma.

Planificación del proyecto

Fase 1: Conocimientos previos de la bioinformática y del proyecto. Optimización del cruce de Base de datos y exposición más intuitiva.

- Adquirir **conocimientos** sobre la bioinformática y el proyecto que nos abarca.
- Familiarizarme con el aplicativo **PCOPGene** y estudiar la forma en que muestra **resultados** la aplicación web.
- Entender la **estructura** de los archivos **.php** que tienen que ser modificados.
- Estudiar la **base de datos local** de http://revolutionresearch.uab.es/.
- Construcción de **sentencias SQL** para hacer consulta a la base de datos.
- **Modificar** archivos **.php** para incluir las mejoras.

Fase 2: Ampliación de la base de datos y modificación del applet java.

- Obtener la información necesaria del **FTP del NCBI.**
- **Modificar la base de datos** para introducir la información descargada.
- Modificar la **aplicación java** para que acepte el nuevo tipo de consultas.

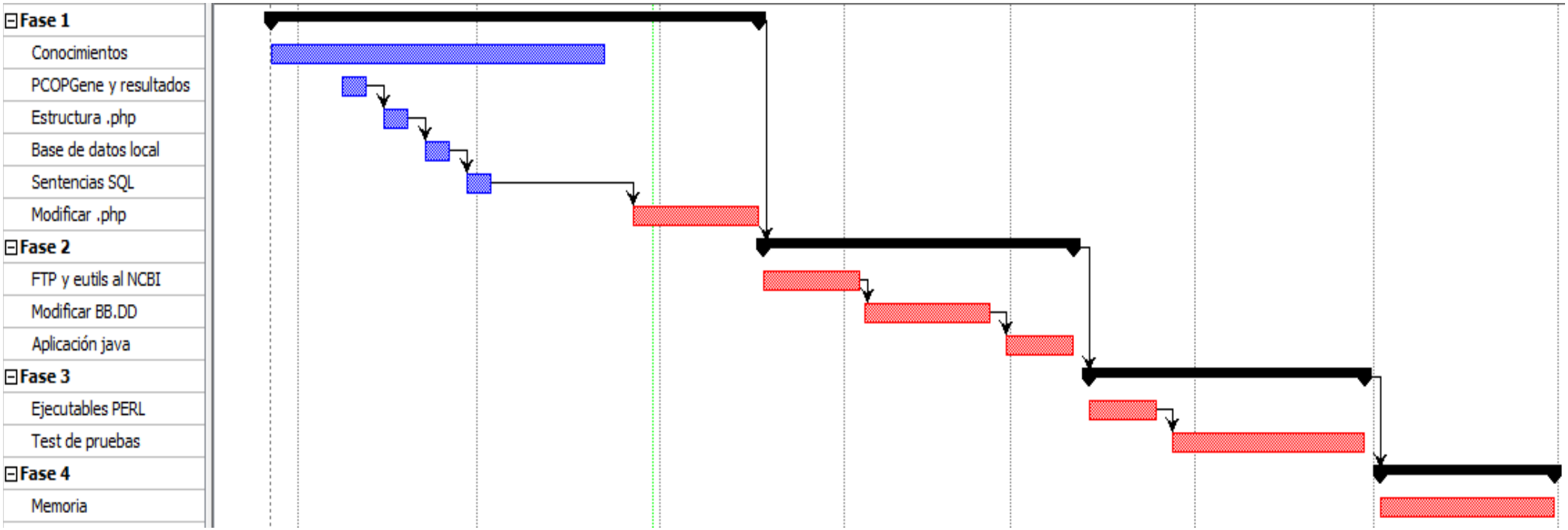
Fase 3: Actualización automática de la base de datos.

- Programar los **ejecutables PERL** que actualicen la base de datos mensualmente.
- **Test de pruebas** de todo el aplicativo.

Fase 4: Documentación

- Memoria

*Las palabras marcadas en negrita hacen referencia a las palabras clave en el diagrama de Gantt


Fase 1
Conocimientos
PCOPGene y resultados
Estructura .php
Base de datos local
Sentencias SQL
Modificar .php
Fase 2
FTP y eutils al NCBI
Modificar BB.DD
Aplicación java
Fase 3
Ejecutables PERL
Test de pruebas
Fase 4
Memoria

NOMBRE	DURACIÓN	INICIO	TERMINADO
Fase 1	62,5 days	27/10/10 12:00	21/01/11 17:00
Conocimientos	40,5 days	27/10/10 12:00	22/12/10 17:00
PCOPGene y resultados	4,875 days	08/11/10 09:00	12/11/10 17:00
Estructura .php	4,875 days	15/11/10 09:00	19/11/10 17:00
Base de datos local	4,875 days	22/11/10 09:00	26/11/10 17:00
Sentencias SQL	4,875 days	29/11/10 09:00	03/12/10 17:00
Modificar .php	15,875 days	27/12/10 09:00	17/01/11 17:00
Fase 2	35 days	18/01/11 08:00	11/03/11 17:00
FTP del NCBI	13 days	18/01/11 08:00	03/02/11 17:00
Modificar BB.DD	10 days	04/02/11 08:00	25/02/11 17:00
Aplicación java	10 days	28/02/11 08:00	11/03/11 17:00
Fase 3	35 days	14/03/11 08:00	29/04/11 17:00
Ejecutables PERL	10 days	14/03/11 08:00	25/03/11 17:00
Test de pruebas	25 days	28/03/11 08:00	29/04/11 17:00
Fase 4	22 days	02/05/11 08:00	31/05/11 17:00
Memoria	22 days	02/05/11 08:00	31/05/11 17:00

Referencias

[1] 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

[2] La Web oficial del National Center for Biotechnology Information (NCBI), un WebServer que ofrece de forma gratuita sus bases de datos.
http://www.ncbi.nlm.nih.gov/

[3] Huerta M, Cedano J, Peña D, Rodriguez A, Querol E. (2009) PCOPGene-Net: holistic characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships. *BMC Bioinformatics.*

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