HealthManagement
Promoting Management and Leadership

Volume 1 / Issue 2 Summer 2006 - Country Focus:Italy

Bio-informatics Application Service Providers - Politecnico di Milano

Authors

Dr. Marco Masseroli, *Ph.D.*

Organisation: Politecnico di Milano, Italy

Email: masseroli@biomed.polimi.it

Web: www.biomed.polimi.it/BioIntro/english/people/research/Masseroli.htm

Francesco Pinciroli

Organisation: Politecnico di Milano, Italy

Email: francesco.pinciroli@polimi.it

Web: www.biomed.polimi.it

For a copy of the references contained in this article, please contact k.ruocco.me@eahitm.org.

The Management and Interpretation of Biomolecular Data

As molecular medicine continues to gain relevance, the availability of the complete sequence of the human genome and the new nanotechnology approaches in molecular biology are permitting the quick study of thousands of genes simultaneously. With increasing biomolecular and bioinformatics advancements, many healthcare sites are offering several genetics tests at relatively low costs. Although such tests can now be easily and routinely performed, the management and interpretation of the produced data cause issues that need to be resolved. These tests produce a great amount of data that needs to be efficiently stored and statistically analysed in order to identify significant genes and proteins studied in the tested conditions. Moreover, to correctly interpret test results, structural and functional information about the identified genes and protein products requires further analysis.

Such information is increasingly available within numerous distributed databanks and easily accessible through web interfaces.

However, the spreading of required information among many heterogeneous databanks and the way most databanks provide such information (i.e. within unstructured HTML pages, one page for each gene or protein entry with all information in the databank about the entry) is not functional to its effective use for the simultaneous analysis of the relevant genes and proteins identified in each genetic test. The resulting genetic test data and interpretation of results also need to be organised, together with other clinical patient data, within clinical repositories in order to easily and effectively query them.

In order to resolve these issues, new data management and analysis approaches are being developed, and specific databases and software tools are being created. Among these is a set of bioinformatics application services developed at the BioMedical Informatics Laboratory of Politecnico di Milano.

They respectively enable:

1) collection of information regarding microarray experiments according to experiment workflow and storage in accordance to the Minimum Information About Microarray Experiments (MIAME) standard specifications;

2) statistical analysis of microarray data in order to identify significant expression patterns of relevant genes involved in the

examined conditions;

3) effective use of biomedical information publicly available in several different genomic databanks to enrich lists of identified genes with related structural and functional information;

4) statistical analysis and data mining with the aim of unveiling information patterns of co-regulated genes and highlighting newbiomedical knowledge.

Available BioInformatics Application Services

Some original bioinformatics application services are provided at http://www.bioinformatics.polimi.it/. Descriptions of system requirements and instructions for their use are available on each service web site. Among them, the GFINDer system is continuously being developed. Currently, it is one the few systems available for the analysis of genomic functional annotations of genes and their gene products, and the only one that provides analysis of human inherited disorder phenotypes.

MicroGen: A Webserver for Microarray Experiments

www.bioinformatics. polimi.it/MicroGen MicroGen consists of a core multi-database system able to store information and data

completely characterising different spotted microarray experiments according to the MIAME standard. Based on a temporal experiment workflow, MicroGen has a web interface able to support the collaborative work required among multidisciplinary actors and roles. As a result, three different actors can cooperate in the same experiment and share information about its production:

+ The Generic Public User, who can get information about MicroGen services by accessing all public sections of the system, including a presentation of MicroGen system facilities and services, a tutorial on its use, and an example of a generated experiment.

+ The Subscribed User, who can fully use the facilities provided by the system for all areas of specialisation he / she has access to.

+ The Web Master, who can use the functionalities offered by MicroGen to manage the whole system and check the work performed within it.

MicroGen additionally supports four types of subscribed user roles: the researcher who designs and requests the experiment; the spotting operator; the hybridisation operator, and the image processing operator. Composed of Active Server Page files, it uses a relational database created in MS-Access. As a result, in order to run MicroGen, an Internet Information Server (IIS) web server and MS-Access must be present. Labeling files containing information about the clones spotted on each array are generated as MSExcel files.

GAAS: Gene Array Analyser Software

www. bioinformatics.polimi.it/GAAS GAAS is an integrated software framework for the management, analysis and visualisation of large amounts of gene expression data across replicated experiments. Comprised of management, analysis and visualisation sections that work with several gene expression dataset formats, it permits custom differential expression data analyses, suitable visualisation, and storage of results.

GAAS is designed for a multi-user environment and is composed of two types of software: Gene Array Assembler Software and Gene Array Analyser Software. The Assembler performs pre-processing of gene expression data, transforming any input data structure in MS-Excel format into a built-in database structure in MS-Access format. The Analyser uses a built-in database gene expression data structure to perform fast differential gene expression analyses across multiple replica experiments. It is structured in the following sections:

+ Management: the management framework is based on the relational MasterDB system accessed and administered through software tools integrated in GAAS. MasterDB is composed of several tables available in the MasterDB management window of the Gene Array Analyser Software.

+ Analysis: the analysis framework enables management and customisation of all implemented data processing procedures subdivided in background, normalisation and gene differential expression analysis steps.

+ Visualisation: the visualisation framework enables visual navigation, both in tabular and graphical format, of data analysis results.

GAAS is developed with MS-Visual C++ and interconnected to a relational database system (MasterDB) developed with MSAccess 2000. GAAS can therefore be run on MS-Windows 98/NT/2000/XP platforms, or on Macintosh running Virtual-PC software. GAAS capabilities are compatible for single PC and local network installations in an MSWindows environment.

MyWEST:MyWeb Extraction Software Tool

http://www.bioinformatics.polimi.it/MyWEST MyWEST is a Java software package for data mining in web-interfaced biomolecular databanks. It provides an intuitive visual interface for building templates that define which information should be extracted from HTML pages in web databanks, then uses the created templates to mine information from multiple web pages of different databanks, stores and aggregates extracted data in a common database, and allows articulated queries to be performed on the aggregated data.

A template configuration module enables data mining of HTML pages in web-interfaced databanks of interest and the creation of extraction templates.

It also supports the definition of access parameters of web-accessible databanks of interest and a relational database for storing extracted data. In the data extraction module, users can provide identification codes of nucleotide or amino acid sequences of interest and use the created templates to automatically mine, in batch mode, the available annotations of interest. The resulting data is stored in Excel file format in a

relational database. Once in the database, extracted information is aggregated and structured for performing articulated queries. A specifically designed updating software agent enables the automatic updating of all information contained inside the database of the mined data.

Name	Start Date	N°. of Accesses	Distinct IPs	Downloaded Copies
MicroGen	July 2005	nearly 900	more than 150	nearly 10
GAAS	April 2003	more than 32,000	nearly 5,000	nearly 380
MyWEST	August 2003	nearly 29,400	nearly 5,000	more than 240
GFINDer	July 2004	more than 61,000	more than 3,100	(Web use only)

Table 1. Bioinformatics application services provided at the MedInfoPoli Web site (http://www.bioinformatics.polimi.it/) and their usage since their opening.

MyWEST stores data extracted from databank web pages both in single tab-delimited ASCII text files, and aggregated in relational databases connected to MyWEST.

Therefore, MyWEST can run on any operating system platform with an adequate Java Virtual Machine installed. To use database functionalities implemented in MyWEST, a suitable Data Base Management System must also be available.

GFINDer: Genome Function INtegrated Discoverer

http://www.bioinformatics.polimi.it/GFINDer GFINDer is a web tool that performs statistical analyses and data mining of functional and phenotypic annotations of gene sets identified in high-throughput biomolecular experiments. It automatically provides largescale lists of user-classified genes with functional profiles biologically characterising different gene classes. GFINDer automatically retrieves annotations of several functional categories from different sources, identifies the categories enriched in each class of a user-classified gene list and calculates statistical significance values for each category. It also enables gene classification according to functional categories and the statistical analysis of obtained results. GFINDer therefore permits a better understanding of microarray experiment results and mining hidden biomedical knowledge by examining user sequence ID lists, or gene lists, and applying clustering and statistical analysis methods to their currently available genomic annotations retrieved from several databanks. The annotation data considered in GFINDer is taken from many different databanks and includes: Gene Ontology (i.e. Biological Process, Cellular Component, and Molecular Function categories), KEGG (i.e. Biochemical Pathways), and PFAM (i.e. Protein Domains). GFINDer also considers clinical and phenotypic information provided by the OMIM databank, which describes the Phenotypes and Phenotype Locations associated with inherited disorders or genetic loci5,6.

GFINDer is organised as a flow scheme of analysis steps in distinct modules, as follows:

+ Upload: uploads user-classified gene lists to be analysed.

+ Annotation: enriches uploaded gene lists with several gene annotation categories, including structural, functional, and phenotypic annotations.

+ Exploration: studies the distribution of different classes of genes among different annotation categories.

+ Statistics: statistically estimates the relevance of each annotation category of the classes of genes considered in uploaded gene lists.

GFINDer use is open to registered and nonregistered users. Non-registered users can test the efficacy of GFINDer's main functionalities by uploading only one sequence ID list at a time that can include only a limited number of sequence IDs.

Registered users can fully access all GFINDer functionalities, upload and store in the system multiple sequence ID lists without any limitation on the number of sequence IDs in each list, save results of GFINDer analyses, and compare results obtained for different sequence ID lists.

Published on : Sun, 16 Jul 2006