TELEBIOINFORMATICS SERVICES OF POLITECNICO DI MILANO FOR POST-GENOMIC BIOMEDICAL APPLICATIONS

ABSTRACT
Molecular medicine is increasingly gaining relevance thanks to availability of the complete sequence of human genome and of new nanotechnology approaches in molecular biology that allows quickly studying thousands of genes simultaneously. Healthcare site are starting to provide several genetic tests, including genetic screenings of numerous genes simultaneously. Although such high-throughput tests can now be automatically or semi-automatically performed, efficient management and correct interpretation of produced data still present some issues. Here we present a set of telebioinformatics application services we developed at the BioMedical Informatics and Telemedicine Laboratory of Politecnico di Milano. They allow orderly collecting and managing test information and data, statistically analyzing produced data, easily gathering known information from several distributed genomic databanks about the relevant genes identified, and statistically evaluating such information with the aim to highlight information patterns that facilitate understanding fundamental biological processes and complex cellular mechanisms underlying patho-physiological phenotypes. All these telebioinformatics services are publicly available at http://www.bioinformatics.polimi.it/, where they are daily accessed by numerous users from all countries. In particular our *GFINDer* Web tool, which is currently the only tool available for performing genomic statistical analyses and data mining of phenotypic, besides functional, information of gene sets identified by high-throughput biomolecular tests.

KEY WORDS
Telegenomics, Web application, genomic databank, management and analysis of biomedical knowledge, Web-based telehealth, database and information systems

1. Introduction

In our current postgenomic era, molecular medicine is increasingly gaining relevance. Availability of the complete sequence of the human genome and of new nanotechnology approaches in molecular biology allows quickly studying thousands of genes simultaneously. With the increasing biomolecular and bioinformatics advancements, many healthcare sites are increasingly offering several genetic tests at relatively low costs. Although such tests can now be easily and routinely performed thanks to the automatic or semi-automatic procedures developed, management and interpretation of the produced data still present some issues. In fact, these tests produce a great amount of data that need to be efficiently stored and statistically analyzed in order to identify, among all genes or proteins studied in each test, those significant in the tested conditions. Moreover, to correctly interpret such test results, the known structural and functional information about the identified genes and protein products need to be further analyzed. Such information - which includes presence of specific sequence characteristics and protein domains; cytogenetic localization; expression in different cellular tissues and organs; and involvement in particular biological processes, molecular functions, biochemical pathways, genetic diseases or phenotypes - is increasingly available within numerous distributed databanks, generally easily accessible on the Internet, also through Web interfaces. However, spreading of the 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 several hundreds of relevant genes and proteins identified in each genetic test. Finally, the produced genetic test data and their interpretation results need to be orderly stored, together with other clinical patient data, within clinical repositories in order to easily and effectively query them.

With the aim to tackle the mentioned issues, new data management and analysis approaches are been developing, and specific databases and software tools are been creating. Here, we present a set of telebioinformatics application services we developed at the BioMedical Informatics and Telemedicine Laboratory of Politecnico di Milano. They respectively allow: 1) to collect all test

information regarding microarray experiment accordingly to experiment workflow, and to properly store it in accordance to the Minimum Information About Microarray Experiments (MIAME) standard specifications; 2) to statistically analyze test microarray data in order to identify lists of genes with significant expression patterns that indicate their relevant involvement in the examined conditions; 3) to effectively use the biomedical information publicly available in several different genomic databanks to enrich lists of identified genes with the structural and functional information known about each selected gene; 4) to statistically analyze and mine such available information with the aim to unveil information patterns of co-regulated genes and highlighting new biomedical knowledge.

2. Provided BioInformatics Application Services

Some original bioinformatics application services we developed in the last years (named *MicroGen*, *GAAS*, *MyWEST*, and *GFINDer*) are currently publicly provided at http://www.bioinformatics.polimi.it/, where they are daily accessed and exploited by numerous users (Table 1), also from well-known research centers and clinical institutions. Technical details of each of these services are discussed in previous scientific publications [1-6]; descriptions of system requirements and instructions for their running are available at each service Web site. Here below we illustrate the main features of each service. Among them, the *GFINDer* system is in continuous active development. It is currently one of 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. Near future enhancements will make available broader analyses of protein domains and biochemical pathways, analysis of gene expression in different tissues, and additional statistical tests to be applied to each annotation type provided.

2.1 *MicroGen*: a Web Server for Microarray Experiments

MicroGen (http://www.bioinformatics.polimi.it/MicroGen/) [1] is a MIAME compliant Web-based information system for managing all the information completely characterizing spotted microarray experiments and the produced data. Based on experiment workflow, it supports distributed collaborative work in the production pipeline of spotted microarray experiments.

MicroGen is constituted of a core multi-database system able to store all information and data completely characterizing different spotted microarray experiments according to the MIAME standard (http://www.mged.org/Workgroups/MIAME/miame.html) following a defined temporal experiment workflow.

MicroGen has an intuitive and user-friendly Web interface able to support the collaborative work required among multidisciplinary actors and roles involved in spotted microarray experiment production. Three different actors, which may also be located in a distance, can co-operate on the same experiment and share the information about its production thanks to the powerful and flexible Internet database technologies used in *MicroGen* implementation. They are:

- 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 MIAME description.
- The *subscribed user*, who can fully use all facilities provided by the system for all areas of specialization he/she has been enabled.
- The *Web master*, who can use the functionalities offered by *MicroGen* to manage the whole system and check the work performed within it.

MicroGen supports four types of subscribed user roles: the *researcher* who designs and requests the experiment, the *spotting operator*, the *hybridization operator*, and the *image processing operator*.

An *example installation* of the *MicroGen* system can be freely accessed at http://www.bioinformatics.polimi.it/MicroGen_test/.

MicroGen is composed of Active Server Page files, and uses a relational database created in MS-Access. Thus, in order to run *MicroGen*, an Internet Information Server (IIS) Web server and also a MS-Access program must be present. Labeling files containing information about the clones spotted on each array are generated as .xls files, therefore MS-Excel is recommended to be installed.

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

Name	Start Date	N° of Accesses	Distinct IPs	Downloaded Copies
MicroGen	July 2005	nearly 900	more than 150	more than 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)

2.2 GAAS: Gene Array Analyzer Software

GAAS (http://www.bioinformatics.polimi.it/GAAS/) [2] is an integrated software framework for efficient management, analysis and visualization of large amounts of gene expression data across replicated experiments. It is structured in management, analysis and visualization sections that allow dealing with several gene expression dataset formats, custom differential expression data analyses, suitable visualization, and storage of results.
The *management section* is based on a relational database system, allowing handling and analyzing gene expression data generated by different high-throughput array technologies, independently from storage formats. Besides, it ensures management and exportability of result data through custom templates defining formats of output databases where storing analysis results.
In the *analysis section* several sequential processing steps are performed: background and spot quality evaluation; background correction and data normalization; evaluation of differential gene expression in a single experiment (i.e. test vs. control condition); and determination of gene regulation (i.e. significant differential gene expression) across multiple replica experiments.
In the *visualization section*, a graphical user interface enables to interactively navigate within numerical results of gene differential expressions and their graphical plots.
GAAS is designed for a multi-user environment, enabling each user to store its own parameter values used to perform the analyses, and define data visualization schema and format of the output data.
The *GAAS* package is composed of two software: *Gene Array Assembler Software* and *Gene Array Analyzer Software*. The *Assembler* performs pre-processing of gene expression data transforming any input data structure in MS-Excel format into a built-in database-based data structure in MS-Access format. The *Analyzer* uses a built-in database-based gene expression data structure to perform fast differential gene expression analyses across multiple replica experiments. It is structured in the following sections:

- *Management section*: the management framework is based on the relational MasterDB system database accessed and administered through software tools integrated in GAAS. MasterDB is composed of several tables. All tables can be accessed and managed through the MasterDB management window of the *Gene Array Analyzer Software*.
- *Analysis section*: the analysis framework enables management and customization of all implemented data processing procedures subdivided in *background*, *normalization* and *gene differential expression* analysis steps.
- *Visualization section*: the visualization framework enables navigating visually, both in tabular and graphical format, the produced data analysis results.

GAAS is implemented in MS-Visual C++ programming language and interconnected to a relational database system (MasterDB) developed by using MS-Access 2000. Therefore *GAAS* can be run on MS-Windows 98/NT/2000/XP platforms, or Macintosh running the Virtual-PC software. *GAAS* capabilities are ensured for single PC and local network installations in MS-Windows environment.

2.3 *MyWEST*: My Web Extraction Software Tool

MyWEST (http://www.bioinformatics.polimi.it/MyWEST/) [3] is a Java software package for effective mining of Web interfaced biomolecular databanks. It provides an intuitive visual interface for building templates that define which information should be extracted from HTML pages of Web databanks, then uses the created templates to mine information from multiple Web pages of different databanks, stores and aggregates in a common database the extracted data, and allows performing articulated queries on the aggregated data for identification of hidden significant biological information.
A *template configuration* module enables the visual definition of the information to mine on selected reference HTML pages of Web interfaced databanks of interest, and the creation of extraction templates. Furthermore, it allows definition of access parameters both to Web accessible databanks of interest and to a relational database for storing all extracted data.
In a *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 from different Web interfaced databanks at once, the available annotations of interest. The mined information is stored in text excel file format for easy and immediate use, and in a relational database. In the database all extracted data are aggregated and structured to allow performing articulated queries for further comprehensive mining.
A specifically designed *updating software agent* enables automatically updating of all information contained inside the database of the mined data.

MyWEST stores data extracted from databank Web pages both in single tab-delimited ASCII text files, and aggregated in any relational database connected to *MyWEST* that has a schema such as that described at http://www.bioinformatics.polimi.it/mywest/MyWEST-DBschema.asp. 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 be available.

2.4 *GFINDer*: Genome Function INtegrated Discoverer

GFINDer (http://www.bioinformatics.polimi.it/GFINDer/) [4-6] is a Web tool for the effective use of genomic data available in many heterogeneous databanks accessible via Internet, which allows performing statistical analyses and

data mining of functional and phenotypic annotations of gene sets identified by high-throughput biomolecular experiments. It automatically provides large-scale lists of user-classified genes with functional profiles biologically characterizing the different gene classes in the list. *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. Moreover, it enables the functional classification of genes according to mined functional categories and the statistical analysis of the classifications obtained. Thus, *GFINDer* allows to better understand microarray experiment results and mine hidden biomedical knowledge by examining user sequence ID's 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* are taken from many different databanks and include several controlled vocabularies, such as those from the 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 OMIM databank, which we normalized and structured in order to obtain two controlled vocabularies, suitable for computational purposes, that describe the Phenotypes and Phenotype Locations associated with inherited disorders or genetic loci [5,6].

GFINDer is organized in distinct modules, each one providing specific functionalities, organized as a flow scheme of analysis steps as follows.

- *Upload*: to upload user classified gene lists to be analyzed.
- *Annotation*: to enrich uploaded gene lists with several gene annotation categories, including structural, functional, and phenotypic annotations.
- *Exploration*: to study the distribution of different classes of genes among different annotation categories.
- *Statistics*: to statistically estimate the relevance of each annotation category in the classes of genes considered in the input uploaded gene lists.

GFINDer use is open to registered and non-registered users. *Non-registered users* can test efficacy of *GFINDer* main functionalities by uploading only one sequence ID's 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 IDs' 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 IDs' lists.

3. Conclusion

The illustrated telebioinformatics application services developed at Biomedical Informatics and Telemedicine Laboratory of Politecnico di Milano and publicly available at http://www.bioinformatics.polimi.it/ allow order collection, efficient management, and effective analysis and interpretation of data and information from genetic test screenings. In particular, *GFINDer* Web tool performs genomic statistical analyses and data mining of functional and phenotypic information of gene sets identified by high-throughput biomolecular screenings that facilitate understanding fundamental biological processes and complex cellular mechanisms underlying patho-physiological phenotypes.

References

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

[2] M. Masseroli, P. Cerveri, P.G. Pelicci, & M. Alcalay, GAAS: Gene Array Analyzer Software for management, analysis and visualization of gene expression data, *Bioinformatics, 19*(6), 2003, 774-775.

[3] M. Masseroli, A. Stella, N. Meani, M. Alcalay, & F. Pinciroli, MyWEST: My Web Extraction Software Tool for effective mining of annotations from web-based databanks, *Bioinformatics, 20*(18), 2004, 3326-3335.

[4] M. Masseroli, D. Martucci, & F. Pinciroli, GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining, *Nucleic Acids Research, 32*, 2004, W293-W300.

[5] M. Masseroli, O. Galati, & F. Pinciroli, GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists, *Nucleic Acids Research, 33*, 2005, W717-W723.

[6] M. Masseroli, O. Galati, M. Manzotti, K. Gibert, & F. Pinciroli, Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists, *BMC Bioinformatics, 6*(Suppl 4), 2005, S18, 8 pp.