Phenotype Miner: an integrated IT system for supporting genetic studies

ID - 129

Nuzzo Angelo[1], Segagni Daniele[1], Milani Giuseppe[1], Larizza Cristiana[1], Bellazzi Riccardo[1]
[1]Department of computer Science and Systems, University of Pavia, Pavia

Motivation

A specific characteristic of the post-genomic era will be the correlation of genotypic and phenotypic information. In this context, the studies aimed at the so-called genetic dissection of complex traits represent a first crucial benchmark for Biomedical Informatics. This kind of studies is based on different types of data, i.e. clinical, genetic and genealogical data. The definition of an Information Technology infrastructure is crucial to support both phenotype discovering and genotypic traits mapping. We developed a Web-based data management system to combine such different sources of information in an integrated framework, in order to make data investigation more efficient and easier for the final user and improve the knowledge discovery process.

Methods

The overall strategy of geneticists analysis is made of 3 main steps: i) discovering the phenotype or clinical condition to be investigated, given clinical data of the population, ii) searching relationships between individuals showing the same phenotype (if any, genetic causes may be supposed), iii) choose appropriate loci to be genotyped to identify genotype-phenotype association. Thus the final purpose of the system that we have developed is to support each of these steps.

- Dynamic Query tool A first crucial task that hampers the development of automated IT solutions in genetic studies is an appropriate definition and identification of the phenotypes that geneticists want to investigate. Clinicians and biologists usually define a phenotype by a set of variables and the values they may take. In order to select (and then to analyze) the individuals satisfying that set of rules, it is necessary to write a suitable SQL statement to run a query and get them. However, as the users may have no expertise in the use of a query scripting language, we provide a tool that automatically generate the proper SQL script to select individuals with the defined phenotype using a graphical user interface. This tool is based on a phenotype formalization that corresponds to a logical tree construction, in which the nodes are the conditions, the AND operator is used to go from the top to the bottom (specialization) and the OR operator is used to add an upper node from the bottom to the top (generalization).
- OLAP engine Dealing with clinical data to analyze phenotypic information implies to take into account their heterogeneity, thus a browsing interface that allows an easy investigation of such data is needed. This means that it is required a tool for performing a multidimensional inspection of the dataset. The technique of multidimensional analysis is implemented with software tools called Online Analytical Processing (OLAP) engines.

In our system we use an OLAP engine written in Java programming language, called Mondrian (http://mondrian.pentaho.org/), which executes queries written in the MDX language (that has actually become a standard for data warehouse applications), reads data from a relational database, and presents the results in a multidimensional format through a Java API, so that the final user may choose the presentation layer. We developed a web application based on JSP pages to integrate it with the Phenotype Editor deployed as Java Web Start applications.

- Pedigree Analysis and Visualization Finally, it is necessary to analyze the relationships between them in order to make hypotheses on the possible genetic origin of that phenotype. This corresponds to the search of common shared ancestors between those individuals and analyzing the shared pedigree. To perform this task, we have developed a software engine (written in Java) that allows the use of a free pedigree analysis software, called Pedhunter (http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/pedhunter.html), directly from the Web application. Output files are provided in the standard Linkage-postMakePed format, so that any available tool can be used for visualization. We integrated our specific tool for pedigree visualization as Java Web Start application.

Results

The system described above is still under development, but it has been just used for several tests on a real dataset, the clinical database of the Val Borbera isolated population study. Geneticists can dynamically compose queries on the dataset using the graphical Phenotype Editor; then they can explore the extracted data at different levels of detail using the OLAP engine, in order to find which phenotypes could be of particular interest for the population. Several phenotypes have been identified (regarding hypertension, thyroid diseases and diabetes) analyzing the clinical database of the project, which actually contains more

than 4000 individuals and about one hundred clinical measures. Finally, the automatic mapping of the selected phenotype on the pedigree allows the geneticists to make hypotheses on the genetic origin of that phenotype and make suitable choices for the following genotyping and genetic analysis.
Availability: http://bioinfo.unipv.it/PhenotypeMiner
Email: angelo.nuzzo@unipv.it