Logistic regression of controlled functional annotations of classified genes

Masseroli M, Bellistri E, Pinciroli F

Laboratorio di Informatica BioMedica, Dipartimento di Bioingegneria, Politecnico di Milano
piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

When the value of a dichotomous variable, for example indicating presence or absence of a characteristic in a subject, depends on one or more other continuous or discrete variables, logistic regression can be used to predict the proportion of individuals who have the characteristics, or to estimate the probability that an individual will have the characteristic. In medicine logistic regression is generally used to estimate the probability that an individual, who has some characteristics or symptoms that influence the onset of a disease, will show such disease. In this case it is also used to calculate which of such characteristics or symptoms influence more significantly, and to which extent, the onset of the disease. In biomolecular medicine the same approach could be used to identify which functional characteristics better explain the binary classification of a set of genes, for instance obtained through statistical and clustering analyses of gene expression results from microarray experiments. In order to test effectiveness of this approach, we implemented a software module that allows performing logistic regression analysis of functional signature annotations of classified gene protein products.

Methods

As source of updated functional information we used the GFINDer genomic knowledge base. GFINDer (http://www.bioinformatics.polimi.it/GFINDer/) is a Web system we previously developed to gather controlled functional and phenotypic genomic annotations sparsely available in numerous different databanks accessible via Internet, and perform their comprehensive statistical analysis and mining. GFINDer is implemented in a three-tier architecture based on a multi-database structure that constitutes its genomic knowledge base. In the first tier, the data tier, a MySQL DBMS manages the knowledge base, which is kept updated by automatic procedures that automatically retrieve gene and protein annotations from several on-line public databanks as soon as new releases of them become available.

For the logistic regression we considered the following usual non linear equation:

$$\ln [p / (1 - p)] = b0 + b1\ x1 + b2\ x2 + \ldots + bi\ xi + \ldots + bn\ xn,$$

where p is the proportion of considered classified genes in two evaluated classes; xi are the proportions of considered genes in the two evaluated classes that present the i characteristics; and bi are the regression coefficients for the i characteristics. The absolute value of each of these bi coefficients indicates the importance of the corresponding i characteristic in contributing to the considered gene classification.

In order to solve the non linear equation, within GFINDer processing tier we used a straightforward Active Server Page and Javascript implementation of a standard iterative method to minimize the Log Likelihood Function, which is defined as the sum of the logarithms of the predicted probabilities of belonging to the first of the two evaluated classes for those considered genes belonging to that class, and the logarithms of the predicted probabilities of belonging to the second of the two evaluated classes for those considered genes belonging to that second class. The Null Model was used as starting guess for the iterations, i.e. all bi coefficients are zero and the b0 intercept is the logarithm of the ratio of the number of considered genes belonging to the first of the two evaluated classes to the number of considered genes belonging to the second class. Minimization is by Newton's method, with an elimination algorithm to invert and solve the simultaneous equations. No special convergence-acceleration techniques were used.

To interact with the DBMS server containing the genomic knowledge base on the data tier, we used Microsoft ActiveX Data Object technology and Standard Query Language, whereas we employed Hyper Text Markup Language and Javascript to implement a Web graphic interface for the user tier,

which is composed of any client computer connected to the Web server on the processing tier through an Internet/intranet communication network.

Results

In GFINDer Web system we implemented a Logistic Regression module that exploits controlled functional information contained within the GFINDer genomic knowledge base to allow executing logistic regression analyses of functional signature annotations of protein products of user-uploaded classified gene lists. Initial results are promising and show that the implemented logistic regression analysis helps in identifying which protein functional characteristics better explain the considered classification of a set of genes. Thus, it could support better interpretation of gene classes defined through statistical and clustering analyses of gene expression results from microarray experiments, and it could contribute to unveil new of biological knowledge about the considered genes.

Availability: http://www.bioinformatics.polimi.it/GFINDer/

Contact email: masseroli@biomed.polimi.it