Improved Singular Value Decomposition Method for Biomolecular Annotation Prediction

Chicco D, Masseroli M, Tagliasacchi M
Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Motivation
Recent biotechnologies and computer science resources have allowed the scientific community to quickly and simultaneously study thousands of genes and proteins. At the same time, advancements in information technologies and biomedical informatics are providing tools to manage the large amount of biomedical and biomolecular information produced, as well as many methods for their analysis. Biomedical domain experts are increasingly annotating bio-entities, mainly genes and their protein products, with controlled ontologies describing their structural, functional and phenotypic biological features. Currently, several controlled vocabularies are routinely used to annotate genes and proteins. Some of them are part of ontologies, like the Gene Ontology (GO). Annotation databases contain the biological information that has been gathered over the years and provide such valuable data as public repositories. Despite their relevance, there are important issues that afflict annotation databases; first, the available annotations are not exhaustive; second, they may be incorrect. In this context, it is fundamental the importance of computational tools that are able to analyze the data stored in annotation databases and infer new biomolecular annotations. Many algorithms have been proposed to predict biomolecular annotations, such as the work by Khatri et Al., based on Singular Value Decomposition (SVD) of the gene-to-term annotation matrix. We propose a novel method (SIM) which extends that SVD-based algorithm by incorporating biomolecular entity (gene or gene product) clustering based on entity functional similarity computed by means of biomolecular annotations.

Methods
SVD. Let the matrix \( A(i, j) \), with \( m \) rows corresponding to biomolecular entities and \( n \) columns corresponding to annotation terms, represent all annotations of a specific controlled vocabulary for a given organism. The entry \( A(i, j) \) assumes value 1 if entity \( i \) is annotated to term \( j \) (or to any descendant of \( j \) if the controlled vocabulary has an ontological structure), or 0 otherwise. The SVD-based annotation prediction is performed by computing a reduced rank approximation \( A_k \) of the matrix \( A \) by means of the singular value decomposition. \( A_k \) contains real valued entries related to the likelihood that gene (or gene product) \( i \) shall be annotated to term \( j \). For a defined threshold \( t \), if \( A_k(i, j) > t \), gene (or gene product) \( i \) is predicted to be annotated to term \( j \). SIM. In the SVD method,
a global term-to-term correlation matrix $T = A'A$ is estimated from the whole corpus of available annotations. As a variation, we propose an adaptive approach named SIM (Semantic IMprovement), which clusters biomolecular entities based on their original annotation profile and estimates a set of distinct correlation matrices $T_c$ for each cluster. To estimate the correlation matrices $T_c$, we cluster gene entities based on their functional similarity, expressed through their annotations, by exploiting the SVD of the matrix $A$. Thus, each gene (gene product) might belong to more than one cluster with different degrees of membership. To calculate $T_c$, for each cluster, first we generate a modified gene-to-term matrix $A_c$, in which the $i$-th row of $A$ is weighted by the membership score of the corresponding gene (or gene product) to the $c$-cluster. Then, we compute $T_c = A_c'A_c$. To obtain a more accurate clustering, we also incorporate the functional similarity between ontological terms, computed by using the Lin's similarity metrics.

**Results**

In order to evaluate the performance of our SIM method, we took as input all the GO annotations of Homo Sapiens genes available from Entrez Gene databank (ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz) on 2010-07-30. We excluded those annotations having evidence IEA (Inferred from Electronic Annotations) or ND (No Data available). We compared the new annotations predicted by our SIM method to the GO annotations of Homo Sapiens genes available from the same databank twenty months later (on 2011-03-31). Based on a threshold $t$, if $A(i, j) \leq 0$ and $A_k(i, j) > t$, a new annotation is suggested; this case is denoted as an annotation predicted (AP). Conversely, if $A(i, j) > 0$ and $A_k(i, j) \leq t$, an existing annotation is suggested to be semantically inconsistent with the available data; this case is denoted as an annotation to be reviewed (AR). From 2,418,333 input annotations, setting the threshold $t = 0.5$, our SIM method predicted 674 (annotations as APs. Out of these APs, 46 (6.83%) turned out to be present among the newer GO annotations considered, which included 31 annotations (67.3%) with evidence different from IEA or ND. Determined results demonstrate that our SIM method provides better annotation prediction compared to the SVD method previously described in literature. These better predictions can be a valuable support for scientists in the biological interpretation of biomolecular experiment results and in unveiling new biomolecular knowledge.

Contact email
davide.chicco@elet.polimi.it