Gene functional clustering for improved prediction of Gene Ontology annotations

Masseroli M, Tagliasacchi M

Motivation

To annotate biomolecular entities, several controlled vocabularies and ontologies, including the Gene Ontology (GO), are available and routinely used. This provides a computable and shareable description of the increasing knowledge of structural, functional and phenotypic features of genes in different organisms. Availability of such controlled annotations is crucial to support interpretation of experimental results and derive new biomedical knowledge. Unfortunately, only a subset of genes of sequenced organisms has been annotated, mainly through automatic annotation procedures. Indeed, considerable effort and time are required to obtain reliable curated annotations. In this context, the curation of annotation data is supported by the use of computational tools, e.g. in the assessment of the relevance of inferred annotations or in the prediction of missed annotations with high reliability. Some algorithms have been proposed to predict GO annotations. Among them, the work by Khatri et al., based on singular value decomposition (SVD) of the gene-to-term annotation matrix, seems to outperform other methods. We propose a novel method which extends that algorithm by incorporating gene clustering based on gene functional similarity computed by means of Gene Ontology annotations.

Methods

Let the matrix A(i,j), with m rows corresponding to genes and n columns corresponding to GO terms, represent all annotations of a specific GO ontology for a given organism. The entry A(i,j) assumes value 1 if gene i is annotated to term j or to any descendant of j in the GO structure, or 0 otherwise. The SVD-based annotation prediction is performed by computing a reduced rank approximation Ak of the matrix A by means of the singular value decomposition. Ak contains real valued entries related to the likelihood that gene i shall be annotated to GO term j. For a defined threshold t, if Ak(i,j) > t, gene i is predicted to be annotated to term j. The SVD method implicitly adopts a global term-to-term correlation matrix T = A'A, estimated from the whole corpus of available annotations. Instead, we propose an adaptive approach, named SIM method, which clusters genes based on their original annotation profile and estimates a set of distinct correlation matrices Tc. For each matrix Tc, a predicted annotation profile for the gene i is computed. The

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

selected predicted annotation profile for the gene i is the one that minimizes the variation, measured by the ell-2 norm, with respect to the original annotation profile of the gene. To estimate the correlation matrices Tc, we cluster genes based on their functional similarity, expressed through their annotations, by exploiting the SVD of the matrix A. Thus, each gene might belong to more than one cluster with different degrees of membership. To estimate Tc, for each cluster, first we generate a modified gene-to-term matrix Ac, in which the i-th row of A is weighted by the membership score of the corresponding gene to the c-cluster. Then, we compute Tc = Ac'Ac. To obtain a more accurate clustering, we also incorporate the functional similarity between GO terms, computed by using the Lin's similarity metrics. To assess the performance of the SVD and SIM methods, we considered the GO annotations of different organisms, including Saccharomyces cerevisiae and Drosophila melanogaster, excluding annotations with evidence code IEA (inferred electronic annotations), since they have not been checked by a manual curator. We performed k-fold cross-validation, confining our analysis to GO terms used to annotate (directly or indirectly) at least 3 or 10 genes in order to obtain more reliable predictions, and heuristically setting a fixed number of 5 clusters for all ontologies.

Results

For each possible gene-term pair, our method produces a ranking score indicating the likelihood of gene i being annotated to term j based on the whole corpus of available annotations. Evaluation results demonstrate that our SIM method generally outperforms the SVD method for all GO ontologies, showing that clustering based on the functional similarity between terms might be beneficial. Nevertheless, most of the performance gain between SIM and SVD stems from the adaptive nature of SIM, regardless on how clustering is actually performed. In fact, the SVD method, which computes similarities between clusters in terms of frequency of co-annotation, is bound to be biased towards the larger clusters, since it is unnormalized. The SIM method counterbalances such a bias with its adaptive approach of clustering genes according to their original annotation profile. The more likely predicted annotations provided by our method can help boosting the performance of data analyses that rely on existing annotations, such as the annotation enrichment analysis. Furthermore, although we considered only GO annotations, our framework can be extended to handle different and also multiple ontologies, as well as to provide predictions based on multiple data sources.

Contact e-mail

masseroli@elet.polimi.it