Improved Gene Ontology annotation predictions through Bayesian network post-processing

Tagliasacchi M, Masseroli M

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

Motivation

Several controlled vocabularies and ontologies, including in particular the Gene Ontology (GO), are currently available and used to annotate biomolecular entities with the aim of describing in a computable and shareable form the increasing knowledge of structural, functional and phenotypic features of genes in different organisms. Availability of such controlled annotations is paramount to support interpretation of experimental results and derive new biomedical knowledge. Unfortunately, only a subset of genes of sequenced organisms has been annotated so far and mainly through automatic annotation procedures, since the more reliable curated annotations require considerable effort and time. In this context, the contributions of computational tools able to analyze annotation data and assess the relevance of inferred annotations, or predict missed annotations with high reliability, are manifold. In the last years, some algorithms to predict GO annotations have been proposed. Among them, that from Khatri et al., which is based on singular value decomposition (SVD) of the gene-to-term annotation matrix, seams to outperform other methods. However, its predicted annotations generally include anomalous predictions, i.e. a gene predicted to be annotated to a GO term but, at the same time, not to some of the term ancestors, as required by the GO hierarchical structure. We propose a post-processing method that can be applied to the output of such SVD method to avoid anomalous predictions and provide more reliable annotation predictions.

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 entries i,j of A assume values 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 values entries related to the likelihood that gene i shall be annotated to the 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 might predict that gene i should be annotated to term j but, at the same time, that it should not be annotated to some of the ancestors of term j. This is an inconsistency with respect to the GO hierarchical structure, which requires that when a gene is annotated to a GO term, it is implicitly annotated also to the more generic terms for that term (i.e. all its ancestors) in the GO structure. To eliminate such anomalous predicted annotations, we constructed a Bayesian network based on the GO topology and used the output of the SVD method as prior evidence. In our Bayesian network each node Tj corresponds to the j-th term in the GO and, for a given gene i, it can assume values in the [0,1] range indicating the a-posteriori probability that term j is used to annotate gene i. For each node Tj, we created an evidence node Ej, which, for a specific gene i, assumes the real valued output of the SVD method, and an edge from Tj to Ej. For each node Tj and gene i, we specified the probability Pi of Tj being either 1 or 0 (i.e., used to annotate gene i or not), conditioned on the values assumed by its children. Finally, we specified the conditional probabilities relating nodes Ej to Tj by modeling them as a Gaussian mixture model. To evaluate the anomaly correction performed by our Bayesian network method, we considered the GO annotations of different organisms, including Saccharomyces cerevisiae and FlyBase, confining our analysis to GO terms used to annotate (directly or indirectly) at least 10 genes and excluding annotations with evidence code IEA (inferred electronic annotations), since they have not been checked by a manual curator.

Results

For each possible gene-term pair, our method produces a ranking score indicating the likelihood of gene i being annotated to term j, given the evidence and the conditional probability constraints imposed by the Bayesian network, i.e. the ontology structure. Evaluation results demonstrate that our post-processing eliminates anomalous annotations, suggesting that the annotations predicted by our method are more likely correct than those predicted by the SVD me-

thod. By providing a prioritized ranked list of more likely predicted annotations to be checked by a manual curator, our method can drive the discovery of previously unknown annotations, as well as the detection of inconsistencies in the existing annotations. Furthermore, the provided annotation profiles can help boosting the performance of data analysis methods that rely upon existing annotations, such as clustering genes based on their annotation profile. 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 : masseroli@elet.polimi.it