Truncated SVD best rank choice through ROC curves for genomic annotation prediction

D. Chicco✉, **M. Masseroli**

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy

Motivations

Correct interpretation of many biological experiments is currently based on consistency of biomolecular annotation databases. Such databases are very widespread and very useful for the scientific community, but, unfortunately, incomplete by definition. To support and quicken their time consuming curation procedure, and to improve their consistence, computational methods that supply a ranked list of predicted annotations are hence extremely useful. We depart from a previous work on the prediction of Gene Ontology (GO) annotations, based on the truncated Singular Value Decomposition (SVD) of the annotation matrix, where the truncation level k of the input matrix is a keypoint in obtaining both best biomolecular annotation predictions and best performance. Here we propose a method that chooses this truncation level by computing and evaluating the Area Under the Curve (AUC) of different Receiver Operating Characteristic (ROC) curves.

Methods

Let the matrix A(i,j), with m rows (genes) and n columns (annotation terms), represent all annotations of a specific controlled vocabulary for a given organism. The entry A(i,j)=1 if gene i is annotated to term j (or descendant), 0 otherwise. The annotation prediction is performed by computing a reduced rank approximation Ak of the matrix A, by means of the SVD. Ak contains real value entries related to the likelihood that gene i shall be annotated to term j. For a defined threshold t, if Ak(i,j)>t, gene i is predicted to be annotated to term j and, if A(i,j)<=0, a new annotation is suggested (AP). Conversely, if A(i,j)>0 & Ak(i,j)<=t, an existing annotation is suggested as semantically inconsistent with the available data (AR).

The method core is the truncation level k, which defines the size of the submatrix used by the algorithm to compute the SVD. For any considered truncation value, our greedy algorithm generates a ROC curve drawing the AR rate (1.0 - Sensitivity) vs the AP rate (1.0 - Specificity), and computes the ROC AUC. If p is the maximum rank of A, where p=min(m,n), and r<=p is the number of non-zero singular values along the diagonal of Sigma matrix, the best truncation value is in the [1;r] interval. To avoid performing the SVD and ROC analysis for every integer value in [1;r] we sample within this interval q values that could be used as adequate truncation values. To obtain the best sampling, we study the distribution of the AUC values for different truncation levels, for a sample dataset (i.e. organism Gallus gallus, GO Biological Process). First, we exclude fist and last 10% values, to avoid taking levels that, during SVD reconstruction, would consider too few or too many non-zero singular values of A. By analyzing gradient variations in AUCs distribution function, we sample q truncation values, inside the above range. We consider every qi as a new SVD truncation value, and compute the AUCqi of the corresponding ROCqi curve. Finally, we take min(AUCqi) as the best qi truncation value.

Results

For evaluation, we use old GO annotations of Gallus gallus and Bos taurus genes available on July 2009 in an old version of GO Annotation databases (http://geneontology.org). By analyzing Gallus gallus annotations between genes and Biological process (BP) (8,731 annotations; 275 genes; 610 BP terms), the algorithm suggests k=77 as the best truncation level for SVD. This level led to a ROC curve having AUC=40.27%, while the 2nd best value, 59, led to AUC=40.46%. From the 8,731 input annotations, with t=0.4, the SVD method with value 77 predicted 44 AP annotations. Out of these, 28 (63.63%) turned out to be present among the 27 month newer GO annotations in a more recent GO database version (Oct-2011); these 28 APs included 14 annotations (50%) with GO evidence different from IEA-ND. On the other hand, the 2nd best value, 59, led to worst results: same number of 44 APs, but just 14 of these (31.81%) were present among the newer GO annotations considered. Other truncation values, related to higher AUC values, led to even worst prediction results.

Erratum.

This is modified version replaced on 22 May 2012. The Editor guarantes the scientific integrety of the abstract content.