Non-negative matrix factorization for prediction of gene annotations

L. Di Persia, G. Leale, G. Stegmayer, D.H. Milone
Institute for Signals, Systems and Computational Intelligence (sinc(i)), FICH-UNL, CONICET, Argentina.

Background:
The accurate prediction of gene annotations is currently an important issue in modern computational biology. A list of putative terms/labels can be provided by the Gene Ontology (GO) and used to design targeted biological experiments in order to generate novel and validated knowledge. However, the handmade curation process of novel annotations is very time-consuming and costly. Thus novel computational tools are needed to reliably predict likely annotations and quicken the discovery of new gene functions. The proximity between GO terms (semantic similarity) can be measured through any of existing semantic measures available, in order to build a distance matrix of GO annotations (dGO) between a group of genes of interest. However, for the case of novel or non-annotated genes, this matrix will have many empty positions. Thus their similarity to annotated genes in order to infer semantically closed annotations could not be calculated. We will show how it is possible to fully reconstruct dGO by using other available information source for the genes (such as expression levels), and afterwards infer their GO labels.

Results:
We present a novel data fusion approach that allows completing the semantic similarity matrix for non-annotated genes by using the available gene data, with non-negative matrix factorization (NNMF). Suppose there are a set of genes of interest with corresponding expression data. An expression distance matrix dE among all of them can be effectively obtained (with Euclidean distance or Pearson correlation). Some of those genes are annotated in GO and some of them are not. Thus, a pairwise semantic distance matrix dGO will have many empty elements. By using NNMF it is possible to complete and reconstruct dGO by using the information in dE (see Fig 1). After that, both matrices (dE and dGO) could be combined, for example to obtain high-quality clusters of genes. For each unlabeled gene, GO annotations can be inferred from the ones belonging to its neighbors genes, according to its reconstructed semantic distance and the frequency of occurrence of GO terms within its cluster. Data from *Saccharomyces cerevisiae* was used to test the proposal (see Fig 1). The approach is sufficiently general for application to any other species. As case study, the “DOT5” gene of *S. cerevisiae* has been used for validation, supposing it was completely unlabelled and using its 64 true labels as gold standard. If label assignation is based on its real dGO alone (considering it as the best possible result), the sensitivity and precision would be 82.81% and 77.94%, respectively. For the proposed NNMF, a sensitivity of 79.68% and a precision of 75% are achieved, which are very close to the best possible case of using the real dGO matrix for annotation inference.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. Di Persia, G. Leale, G. Stegmayer & D. H. Milone; "Non-negative matrix factorization for prediction of gene annotations"
4th ISCB-LA Bioinformatics Conference, 2016.

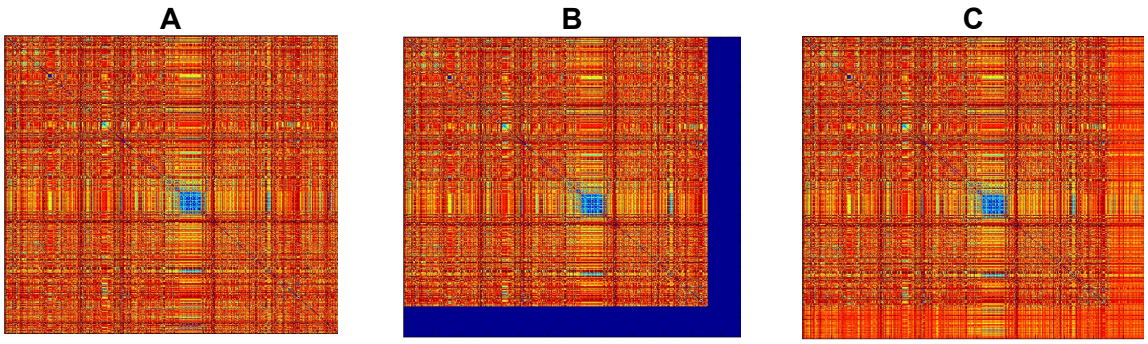

Figure 1. Reconstruction of the semantic similarity matrix using expression data with NNMF data fusion. (A) Original dGO among 587 genes of *S. cerevisiae*. (B) dGO with 10% of semantic information removed. (C) NNMF reconstructed dGO matrix.

Conclusions:

We have shown a novel approach to the of prediction of gene annotations based on NNMF fusion of the semantic and expression distances among genes, that uses the fusion to complete the unknown part of the semantic distance matrix. The reconstructed semantic matrix can be then used to infer candidates terms for the unknown genes. This approach can yield a sensitivity and precision comparable and extremely close to the one obtained by using the real semantic distance information, which shows that the NMF fusion approach was successful in capturing the information structure of the dGO matrix.

Supported by: CONICET, MinCyT, UNL

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. Di Persia, G. Leale, G. Stegmayer & D. H. Milone; "Non-negative matrix factorization for prediction of gene annotations"
4th ISCB-LA Bioinformatics Conference, 2016.