**File S4. Performance comparison between six expression profile-based GO prediction methods**

**A. The procedures of GBA strategy for expression profile-based GO prediction**

In the guilty-by-association (GBA) strategy, we select the template genes which have the highest similarity with query gene in terms of expression profiles, and then use the GO terms of templates to annotate the query, as follows.

**Training stage**

In a training dataset, the expression profiles of all genes can be represented as a matrix , where the -th row of is the expression profile for the -th gene and denoted as , is the total number of training genes, is the number of experimental samples in microarray technology [1], and is the expression value of the -th gene on the -th sample. We orderly execute z-score normalization [2] and principal component analysis (PCA) [3] on expression profile matrix to obtain a normalized matrix , where the-th row of , denoted as , is the normalized expression profile vector for the -th training gene.

**Prediction stage**

For a query gene, its expression profile can be represented as a vector . First, the z-score normalization and PCA are orderly executed on the expression profile vector to obtain a normalized vector . Then, for each training gene , we calculate its similarity score with query based on the normalized vector and . Next, we rank training genes based on the similarity scores in descending order. Finally, we select the top training genes as templates to annotate the GO terms of query. Specifically, the confidence score that the query is associated with GO term can be calculated as follows:

(S1)

(S2)

where is the weight for the -th template, and is the rank of the -th template; , if the -th template is associated with in the experimental annotation; otherwise, .

In this work, the similarity score of expression profiles between two genes are measured by four unsupervised methods, including Pearson correlation coefficient (PCC) [4], Spearman rank correlation (SRC) [5], mutual rank (MR) [6], and Euclidean distance (ED) [7], and a recently proposed supervised method, *i.e.*, metric learning for co-expression (MLC) [8].

The PCC between the-th training gene and query gene is calculated as follows:

(S3)

where and are mean values for and , respectively.

The SRC between the-th training gene and query gene is calculated as follows:

(S4)

where is rank of in the elements of in ascending order, is the rank of in the elements of in ascending order.

Due to the long computation time of MR values, we directly download MR values of genes from COXPRESdb [9] and ATTED-II databases [6]. In a species with genes, the MR value between gene and gene is calculated as follows. First, we calculate the PCC values between gene and the remaining genes based on the corresponding expression profile vectors, and rank the genes based on the PCC values in descending order. Similarly, we calculate the PCC values between gene and the remaining genes, and rank the genes in descending order based on PCC values. Then, the MR value between genes and can be calculated:

(S5)

where is the rank of gene in genes for gene , and is the rank of gene in genes for gene .

The ED between the-th training gene and query gene is calculated as follows:

(S6)

In MLC, the similarity between the-th training gene and query gene is measured by weight inner product (WIP) as follows:

(S7)

where is a diagonal matrix and can be optimized by the Broyden-Fletcher-Goldfarb-Shanno method [10].

The higher values of PCC, SRC, and WIP indicate the higher similarity, while the lower values of MR and ED mean the higher similarity.

**B. The performances of six expression profile-based GO prediction methods for each individual species**

For each of 8 species, we will evaluate the performances of six expression profile-based GO prediction methods on the corresponding test dataset. For each method, we execute it 10 times and then use the average of all prediction results as the final result.

Figure S2 show the values of Fmax and AUPR for 8 species via six expression profile-based methods. Table S5 summarizes the *P* values of Fmax and AUPR values between TNP and other five methods in student’s t-test [11] for 8 species. In comparison between TNP and MLC, we use two samples t-test [12] to calculate *P* value due to that the prediction results in 10 times are different for MLC/TNP. In comparison between TNP and PCC, MR, SRC, ED, we use single samples t-test [13] to calculate *P* value, because the prediction results in 10 times are same for PCC/MR/SRC/ED. From Figure S2 and Table S5, we can observe that TNP achieves the highest values of Fmax and AUPR among six methods for each GO aspect in each species. For example, in human species, the improvements of Fmax values between TNP and MR are 12.7%, 8.2%, 3.8%, respectively, with *P* values of 1.29×10-04, 7.83×10-09, and 5.07×10-07 for MF, BP, and CC aspects. As another example, the average improvement of AUPR values of three GO aspects between TNP and the second best performer is 8.6% with *P* values<0.05 for Arabidopsis species.

Figure S3 plots the precision-recall (PR) curves of six expression profile-based methods for three GO aspects in 8 species. For each GO aspect in each species, we can find that TNP has the highest precision values among six expression profile-based methods at all different recall rates.

[1] Heller MJ. DNA microarray technology: devices, systems, and applications. Annu Rev Biomed Eng 2002;4:129–53.

[2] Patro S, Sahu KK. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462 2015:1–5.

[3] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab Syst 1987;2:37–52.

[4] Adler J, Parmryd I. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. Cytometry A 2010;77:733–42.

[5] Zar JH. Significance testing of the Spearman rank correlation coefficient. J Am Stat Assoc 1972;67:578–80.

[6] Obayashi T, Aoki Y, Tadaka S, Kagaya Y, Kinoshita K. ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. Plant Cell Physiol 2018;59:e3–4.

[7] Wang L, Zhang Y, Feng J. On the Euclidean distance of images. IEEE Trans Pattern Anal Mach Intell 2005;27:1334–9.

[8] Makrodimitris S, Reinders MJ, van Ham RC. Metric learning on expression data for gene function prediction. Bioinformatics 2020;36:1182-90.

[9] Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Res 2019;47:D55–62.

[10] Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput 1995;16:1190-208.

[11] Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behav Ecol 2006;17:688–90.

[12] Heeren T, D'Agostino R. Robustness of the two independent samples t‐test when applied to ordinal scaled data. Stat Med 1987;6:79-90.

[13] Crawford J, Howell DC, Garthwaite PH. Payne and Jones revisited: estimating the abnormality of test score differences using a modified paired samples t test. J Clin Exp Neuropsychol 1998;20:898-905.