Journal Pre-proofs

Method

Integrating transcript expression profiles with protein homology inferences for gene function prediction

Yi-Heng Zhu, Chengxin Zhang, Yan Liu, Gilbert S. Omenn, Peter L. Freddolino, Dong-Jun Yu, Yang Zhang

PII: S1672-0229(22)00041-9
DOI: https://doi.org/10.1016/j.gpb.2022.03.001
Reference: GPB 623

To appear in: *Genomics, Proteomics & Bioinformatics*

Received Date: 20 August 2021
Revised Date: 2 March 2022
Accepted Date: 16 April 2022

Please cite this article as: Y-H. Zhu, C. Zhang, Y. Liu, G.S. Omenn, P.L. Freddolino, D-J. Yu, Y. Zhang, Integrating transcript expression profiles with protein homology inferences for gene function prediction, *Genomics, Proteomics & Bioinformatics* (2022), doi: https://doi.org/10.1016/j.gpb.2022.03.001

Integrating Transcript Expression Profiles with Protein Homology Inferences for Gene Function Prediction

Yi-Heng Zhu[1,2], Chengxin Zhang[2], Yan Liu[1], Gilbert S. Omenn[2,3], Peter L. Freddolino[2,4], Dong-Jun Yu[1,*], Yang Zhang[2,4,*]

[1] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*

[2] *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*

[3] *Departments of Internal Medicine and Human Genetics, and School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA*

[4] *Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA*

[*] Corresponding authors.

E-mail: njyudj@njust.edu.cn (Yu D), zhng@umich.edu (Zhang Y).

Running title: *Zhu Y et al / TripletGO*: *Gene Function Annotation*

Total number of references: 44

Total number of figures: 7

Total number of tables: 2

Total number of supplementary figures: 14

Total number of supplementary tables: 14

Total number of supplementary files: 8

Abstract

Gene Ontology (GO) has been widely used to annotate functions of genes and gene products. We proposed a new method, TripletGO, to deduce GO terms of protein-coding and non-coding genes, through the integration of four complementary pipelines built on transcript expression profile, genetic sequence alignment, protein sequence alignment, and naïve probability. TripletGO was tested on a large set of 5754 genes from 8 species (human, mouse, Arabidopsis, rat, fly, budding yeast, fission yeast, and nematoda) and 2433 proteins with available expression data from the third Critical Assessment of Protein Function Annotation (CAFA) challenge. Experimental results showed that TripletGO achieved function annotation accuracy significantly beyond the current state-of-the-art approaches. Detailed analyses show that the major advantage of TripletGO lies in the coupling of a new triplet network-based profiling method with the feature space mapping technique, which can accurately recognize function patterns from transcript expression profile. Meanwhile, the combination of multiple complementary models, especially those from transcript expression and protein-level alignments, improves the coverage and accuracy of the final GO annotation results. The standalone package and an online server of TripletGO are freely available at https://zhanglab.ccmb.med.umich.edu/TripletGO/.

KEYWORDS: Gene function annotation; Gene Ontology; Transcript expression profile; Triplet network; Protein-level alignment

Introduction

In the post-genome sequencing era, a major challenge is to annotate the biological functions of all genes and gene products, which are grouped, in the context of the widely used Gene Ontology (GO), into three aspects, *i.e.*, molecular function (MF), biological process (BP), and cellular component (CC) [1]. Accurate annotation of gene function provides essential knowledge to disease mechanisms and drug design [2,3]. Direct determination of the functions of genes via biochemical or genetic experiments is typically time-consuming and laborious, and often incomplete [4]. As a result, a large number of genes in the sequenced genomes have no available function annotation to date. For example, according to official statistics in the neXtProt platform [5], nearly 2000 human protein-coding genes have yet no known function; for many other organisms of biomedical or industrial importance, annotation rates are substantially lower. To fill the gap between sequence and function, it is urgent to develop efficient computational algorithms for function prediction [6,7].

Function annotations can be performed at either the protein- or gene-level. In the former case, the function of the query gene is determined by that of its encoded protein, which can be deduced from the protein sequence, structure, or family information [8–14]. However, protein-coding genes account for only ~ 2% of a typical multicellular eukaryote genome such as that of humans [15]. There are also many genes for non-coding RNAs as well as genes whose coding potential is unknown or ambiguous.

Most gene-level annotation methods deduce GO terms for queries by using a guilt-by-association (GBA) strategy, which is typically based on the similarity of expression profiles between the gene of interest and template genes with known GO annotations [16–18]. The rationale of GBA is reasonable as genes with the same functions often show similar expression profiles. This was supported by the third Critical Assessment of Protein Function Annotation (CAFA) challenge, which showed that expression profile has a great potential to improve prediction performance [19]. Despite there are some achievements of current expression profile-based methods, however, challenges remain.

First, it is tricky to define an effective similarity measure of expression profiles as the substitute for functional similarity. In previous work, several unsupervised methods (*e.g.*, Pearson correlation coefficient [20] and mutual rank [21]) and supervised methods (*e.g.*, metric learning for co-expression [17]) have been developed to measure the expression profile similarity in gene function prediction. Unfortunately, these methods cannot achieve optimal performance, because these expression similarity metrics may have no close correlation with functional similarity. Part of the reason is that these methods define the expression similarity in the original space, in which the expression data show a high dimensionality across multiple tissues and complicated distributions; as a result, the measured expression similarity is hardly associated with functional similarity and thus a higher expression similarity (by these metrics) often does not indicate a higher functional similarity. To address this issue, a promising approach is to change the data distribution via feature space mapping [22], in which the expression profiles are mapped from the original feature space to a new embedding space by non-linear functions, and the expression similarity is then associated with functional similarity in this embedding space. The second challenge is that the functional similarity of genes is often difficult to completely capture by one similarity measure. This necessitates the combinations of multiple similarity measures from different biological datasets which may help improve both accuracy and coverage of function predictions [9].

In this work, we proposed and tested a new approach, TripletGO, to integrate multi-source information from both genes and proteins for protein-coding and non-coding gene annotations. First, we extended a supervised triplet network method [23] to assess expression profile similarities in function prediction. In this extended triplet-network pipeline (TNP), the expression profiles are mapped from the original feature space to an embedding feature space via deep neural network learning, where a triplet loss function is designed to enhance the correlation between expression profile and gene function. Second, considering that most protein-coding gene functions are performed through proteins and that protein sequence alignments, which are based on 20 amino

acids, often provide more specific function associations than nucleotide sequence alignments, we proposed a protein-level method for GO prediction using protein sequence similarity. Finally, a composite model was derived by integrating the output of four complementary GO prediction pipelines, built on the TNP-based expression profile, genetic sequence alignment, naïve probability, and protein sequence alignment, through an optimal neural network training. TripletGO has been systematically tested on a large set of non-redundant genes collected from eight species, where the results demonstrated the significant advantage on accurate GO term prediction over the current state-of-the-art in the field. The standalone package and an online server of TripletGO are freely available through URL: https://zhanglab.ccmb.med.umich.edu/TripletGO/.

Method

Overview of TripletGO

TripletGO is a hierarchical approach for gene function annotation with respect to GO terms, as shown in **Figure 1**A. In TripletGO, the final GO model is a combination of the outputs of four complementary pipelines, including expression profile-based GO prediction (EPGP) by TNP, genetic sequence alignment-based GO prediction (GSAGP), protein sequence alignment-based GO prediction (PSAGP), and Naïve-based GO prediction (NGP). Here, the input is a genetic sequence with Entrez gene ID, and the output is the confidence score for the predicted GO term. First, we extract the expression profile and coding protein sequence for query gene from COXPRESdb [24] (or ATTED-II [21]) database and UniProt [25] database, respectively, using Entrez ID. Then, the expression profile, genetic sequence, and protein sequence are respectively used as the inputs of EPGP, GSAGP, and PSAGP methods to output the confidence scores of GO terms. Moreover, NGP method is also used to calculate confidence scores. Finally, for a GO term Q_j, its confidence scores by the four methods are serially combined as a vector, which used as the input of fully connected neural network to output the consensus score.

EPGP by TNP

In EPGP, a triplet network [23] is used to measure the similarity of expression profiles, as shown in Figure 1B. The input is a triplet variable (anc, pos, neg), where anc is an anchor (baseline) gene, pos is a positive gene with the same function of anc, and neg is a negative gene with the different function of anc. First, the expression profile of each gene is mapped from the original feature space to an embedding space using the same deep neural network. Next, the expression dissimilarity between two genes in embedding space is measured by Euclidean distance (ED) [26] of the mapped expression profiles. Finally, the triplet loss function is designed to associate expression similarity with functional similarity:

$$Tripletloss = max(d(anc, pos) + margin - d(anc, neg), 0) \quad (1)$$

where $d(anc, pos)$ is the ED between anchor and positive genes in embedding space, $d(anc, neg)$ is the distance between anchor and negative genes, and $margin$ is a pre-set positive value. Here, the minimization of the triplet loss requests for the maximization of $d(anc, neg) - d(anc, pos)$. In the ideal case, $tripet\ loss = 0$ when $d(anc, neg) \geq d(anc, pos) + margin$, which indicates substantially higher similarity (lower distance) of the anchor genes to the positive genes than to the negative genes.

It has been demonstrated that cross-entropy loss [27] helps to improve the performance of triplet network [28,29]. Therefore, we further combine the triplet loss with the cross-entropy loss in the TNP to predict gene function from expression profiles. The overall workflow of TNP is depicted in Figure 1C, which contains two stages.

Training stage of TNP

Procedure I: *expression profile normalization*

In a training dataset, the expression profiles of all m genes are represented as a matrix $\boldsymbol{E} = (e_{ij})_{m \times l}$, where l is the number of experimental samples in microarray technology [30], and e_{ij} is the expression value of the i-th gene on the j-th sample. Each row of $\boldsymbol{E}$ can be viewed as the expression profile of a gene. To reduce noise and

computing cost, the matrix $\boldsymbol{E}$ is transformed into a normalized matrix $\boldsymbol{E}^n = (e_{ij}^n)_{m\times h}$ ($h < l$) by performing z-score normalization [31] and principal component analysis (PCA) [32].

Procedure II: *expression profile mapping using a neural network*

The normalized expression profiles are mapped from the original feature space to an embedding space using a neural network. Specifically, the normalized matrix $\boldsymbol{E}^n$ is fed to a deep fully connected block (DFCB) with N layers to output an embedding matrix $\boldsymbol{U} = (u_{ij})_{m\times d_N}$, where d_N is number of neurons in the N-th layer. Then, L2-normaliztion is executed on $\boldsymbol{U}$ to obtain a normalized matrix $\boldsymbol{U}^n = (u_{ij}^n)_{m\times d_N}$, where $u_{ij}^n = u_{ij}/(\sum_{j=1}^{d_N} {u_{ij}}^2)^{1/2}$. Each row of $\boldsymbol{U}^n$ can be viewed as the expression profile of a gene in the embedding space.

At the same time, an output layer L_O with sigmoid activation function [33] is fully connected with DFCB to output a score matrix $\boldsymbol{S} = (s_{ij})_{m\times r}$, where r is the number of GO terms in the training dataset, and s_{ij} is the confidence score that the i-th training gene is associated with the j-th GO term. Then, we calculate triplet loss and cross-entropy loss based on matrix $\boldsymbol{U}^n$ and score matrix $\boldsymbol{S}$, respectively.

Procedure III: *loss function calculation and network optimization*

We use the "batch on hard" strategy [34,35] to calculate the triplet loss:

$$Tripletloss_{hard} = \sum_{i=1}^{m} max(d(i,pos)_{max} + margin - d(i,neg)_{min},\ 0)/m \quad (2)$$

where $d(i,pos)_{max}$ (or $d(i,neg)_{min}$) is the maximum (or minimum) value of distances between the i-th gene and all positive (or negative) genes with same (or different) function of the i-th gene in embedding space. The distance between the two genes (i,j) is measured by $d(i,j) = \sum_{k=1}^{d_N}\left(u_{ik}^n - u_{jk}^n\right)^2/4$, where the division factor of 4 is introduced to normalize $d(i,j)$ into the range of [0,1], *i.e.*, $0 \le d(i,j) \le \sum_{k=1}^{d_N}(2(u_{ik}^n)^2 + 2(u_{jk}^n)^2)/4 = 1$. Moreover, two genes are considered to have the

same function if their functional similarity is larger than a cut-off value c_f. The functional similarity of two genes is measured by the F1-score between their GO terms, as shown in File S1A.

The cross-entropy loss is calculated as:

$$Loss_{cross-entropy} = -\sum_{i=1}^{m}(\sum_{j=1}^{r} y_{ij} \cdot log\,(s_{ij}) + (1-y_{ij})\log(1-s_{ij}))/(r \cdot m) \quad (3)$$

where $y_{ij} = 1$ if the i-th gene is associated with the j-th GO term in the experimental function annotation; otherwise, $y_{ij} = 0$.

The final training loss in TNP is the combination of triplet loss and cross-entropy loss:

$$Trainingloss = Triletloss_{hard} + \alpha \cdot Loss_{cross-entropy} \quad (4)$$

where α is a trade-off value. Finally, we minimize training loss to optimize neural network by Adam optimization algorithm [36].

Prediction stage of TNP

The input is a query gene with expression profile vector $\boldsymbol{e}^q$, and the output is a confidence score vector $\boldsymbol{s}$, including the confidence scores of r GO terms for query. First, z-score normalization and PCA are orderly executed on $\boldsymbol{e}^q$ to obtain a normalized vector $\boldsymbol{e}_q^n$, which was used as the input of DFCB. Then, we execute L2-normalization on the output of DFCB to obtain a normalized embedding vector $\boldsymbol{u}^q$. Next, a distance rank-based strategy (see details in File S1B) is executed on the normalized embedding matrix of training genes ($\boldsymbol{U}^n$) and $\boldsymbol{u}^q$ to generate a confidence score vector $\boldsymbol{s}^t$. At the same time, the output layer L_O outputs another score vector $\boldsymbol{s}^c$ by sigmoid function mapping. The final score vector $\boldsymbol{s}$ is the combination of two vectors:

$$\boldsymbol{s} = w \cdot \boldsymbol{s}^t + (1-w) \cdot \boldsymbol{s}^c \quad (5)$$

where w is a trade-off value and ranges from 0 to 1.

GSAGP

In GSAGP, we search the template genes, which have the similar sequences with query

gene, from a genetic sequence database with GO annotation (named GSD-GOA, see "Datasets") for functional annotation.

For a query, we extract its RNA sequence from National Center for Biotechnology Information (NCBI) [37]. Then, Blastn [38] is used to search the templates of query with an e-value cutoff of 0.1 against GSD-GOA. To remove homology contamination, we exclude all homologous templates which have more than t_1 sequence identity with the query. Finally, the remaining templates are used to annotate the query. Specifically, the confidence score that the query is associated with GO term Q_i is calculated as:

$$S(Q_i)_{GSAGP} = \frac{\sum_{k=1}^{n} b_k \cdot I_k(Q_i)}{\sum_{k=1}^{n} b_k} \quad (6)$$

where n is the number of template genes, b_k is the bit-score of k-th template by Blastn; $I_k(Q_i) = 1$, if the k-th template is associated with Q_i in the experimental function annotation; otherwise, $I_k(Q_i) = 0$.

PSAGP

In PSAGP, we select the template genes, whose coding proteins have similar sequence with that of the query, for GO functional annotation.

For a query gene, we map it as the corresponding coding protein sequence P in the UniProt database [25]. Then, Blastp [38] is used to search the template proteins of P with a e-value cutoff of 0.1 against a protein sequence database (*i.e.*, PSD, see "Datasets"), where homologous templates with a sequence identity above t_2 to P are removed. Finally, the remaining templates are mapped back to the genes in a gene-level GO annotation database (named Gene-GOA, see "Datasets") to annotate the query. The confidence score is calculated using the same scoring function as in GSAGP (*i.e.*, Equation 6), where b_k is the bit-score of the k-th template by Blastp.

NGP

In NGP, the confidence score that a query is associated with GO term Q_i can be calculated by the frequency of Q_i in Gene-GOA:

$$S(Q_i)_{NGP} = N(Q_i)/N_{GO} \quad (7)$$

where $N(Q_i)$ is the number of genes associated with Q_i, and N_{GO} is the number of genes with at least one annotation for the same GO aspect as Q_i. This predictor can be thought of as a prior arising from the overall abundance of a particular annotation in Gene-GOA.

Consideration of hierarchical relation for evaluation of GO annotation

The GO annotation is hierarchical [19]. Specifically, for both the ground truth and the prediction, if a protein (gene) is annotated with a GO term Q_i, it should be annotated with the direct parent and all ancestors of Q_i. To enforce this hierarchical relation, we follow CAFA's rule and use a common post-processing procedure [10] for the confidence score of term Q_i in all GO prediction methods as follows:

$$S(Q_i)_{post} = \max\left(S(Q_i), S(QC_i^1)_{post}, S(QC_i^2)_{post}, \ldots, S(QC_i^n)_{post}\right) \quad (8)$$

where $S(Q_i)$ and $S(Q_i)_{post}$ are the confidence scores of Q_i before and after post-processing, $S(QC_i^1)_{post}, S(QC_i^2)_{post}, \ldots, S(QC_i^n)_{post}$ are the confidence scores of all direct children terms of Q_i after post-processing. This post-processing procedure enforces that the confidence score of Q_i is larger than or equal to the scores of all children.

Datasets

We collected all 78,170 genes with GO annotation via experimental determination from NCBI [37] to construct a gene-level GO annotation database (*i.e.*, Gene-GOA, see File S2A). The genes in Gene-GOA were used to construct the template databases (*i.e.*, GSD-GOA and PSD, see File S2B and C) in GSAGP and PSAGP, and calculate the prior probability in NGP.

To evaluate the proposed methods, we collected 57,584 genes from 8 species by the following procedures: (1) we downloaded all of 300,977 genes with expression profiles determined by microarray [30] for 20 species from COXPRESdb [24] and ATTED-II [21] databases. For each species, the total number of genes with functional annotation in Gene-GOA is shown in Table S1. Then, we selected the 8 species with the most genes

with GO annotation among the 20 species, to construct benchmark datasets. (2) For each species, we randomly selected 85% of genes with GO annotation as the training dataset, and 5% genes as the validation dataset, which were separately used to construct machine learning-based models and optimize the parameters of models. The remaining 10% genes were used as the test dataset to assess the performance of models. As a result, there are 48,954, 2876, and 5754 genes in training, validation, and test datasets, respectively, for the 8 species in total, as summarized in Table S2.

Implementation and parameter settings of TripletGO

In EPGP, DFCB consists of two fully connected layers, each including 1024 neurons with rectified linear unit (RELU) activation function [39]. The remaining parameters of EPGP are listed in Table S3. In PSAGP, we used $t_2 = 30\%$ sequence identity as the cut-off to remove homologous protein templates, following previous studies [9]. To determine the homology cutoff for nucleotide sequences, we used three different machine learning models to fit the relationship between protein sequence identity and gene sequence identity, and it was found that a 30% protein sequence identity roughly corresponds to 60% genetic sequence identity, as shown in File S3 and Figure S1. Therefore, we used $t_1 = 60\%$ sequence identity as cut-off to remove homologous templates in the GSAGP.

Evaluation metrices

Maximum F1-score (Fmax) and area under the precision-recall curve (AUPRC) are used to evaluate the performance of proposed methods. Fmax is one of the most important evaluation metrics in CAFA [19,40] and is defined as:

$$\mathrm{Fmax} = \max_{0 \le t \le 1} \left[\frac{2 \cdot pr(t) \cdot rc(t)}{pr(t) + rc(t)}\right] \tag{9}$$

where t is a cut-off value of confidence score; $pr(t)$ and $rc(t)$ are precision and recall, respectively, with confidence score $\geq t$:

$$\begin{cases} pr(t) = \frac{tp(t)}{tp(t) + fp(t)} \\ rc(t) = \frac{tp(t)}{tp(t) + fn(t)} \end{cases} \tag{10}$$

where $tp(t)$ is the number of correctly predicted GO terms, $tp(t) + fp(t)$ is the number of all predicted GO terms, and $tp(t) + fn(t)$ is the number of GO terms in experimental function annotation. AUPRC is a critical measure in multi-label prediction task [41] and ranges from 0 to 1.

The average performance of a method on multiple species is measured by weighted average Fmax (WAFmax) and weighted average AUPRC (WAAUPRC):

$$\begin{cases} \mathrm{WAFmax} = \frac{\sum_{i=1}^{M} \mathrm{Fmax}_i \cdot N_i}{\sum_{i=1}^{M} N_i} \\ \mathrm{WAAUPRC} = \frac{\sum_{i=1}^{M} \mathrm{AUPRC}_i \cdot N_i}{\sum_{i=1}^{M} N_i} \end{cases} \quad (11)$$

where M is the number of species, Fmax_i and AUPRC_i are the Fmax and AUPRC values, respectively, on the test dataset of the i-th species, and N_i is the number of test genes for the i-th species.

Results

TNP improves EPGP

TripletGO is a hierarchical approach that takes the gene sequence, the protein sequence, and the transcript expression profile data as input. Gene ontologies are then created by a set of four complementary pipelines, built on transcript expression profile, gene sequence alignment, protein sequence alignment, and naïve prior statistical calculation separately, where the final GO models are obtained by a neural network combination (Figure 1).

As TripletGO is centralized by the transcript expression profile through TNP (Figure 1C), we first compare the TNP with five existing methods in EPGP. These include four unsupervised scores: Pearson correlation coefficient (PCC) [20], Spearman rank correlation (SRC) [42], mutual rank (MR) [21], and ED [26]; and a supervised method: metric learning for co-expression (MLC) [17]. Each method is combined with the GBA strategy to predict GO terms, as described in File S4A. **Figure 2**A and B list the WAFmax and WAAUPRC values on the test dataset of 5754 genes from 8 species (human, mouse, Arabidopsis, rat, fly, budding yeast, fission yeast, and nematoda) for

the six methods, where the *P* values between TNP and the other five methods in Student's *t*-test [43] for WAFmax and WAAUPRC are summarized in Table S4. In addition, the performances of the six methods for each individual species are summarized in Figures S2 and S3, Table S5, and discussed in File S4B.

From Figure 2A and B, and Table S4, it can be observed that the accuracy of TNP is significantly higher than the other five methods in deducing function from gene expression data. Specifically, the improvements of WAFmax values between TNP and the second-best performer, MR, are 12.4% (= $(0.317 - 0.282)/0.317 \times 100\%$), 6.2%, and 4.2% for MF, BP, and CC, respectively, all with *P* values significantly below 0.05. At the same time, TNP achieves an increase of WAAUPRC values by 11.7%, 6.6%, and 7.9% compared to MR for the three GO aspects. Moreover, TNP and ED separately show the best and worst performances among six methods, although they use the similar metric functions (TNP uses the square of ED). These data suggest that the GO recognition accuracy can be improved via feature space mapping when coupled with triplet network learning. In addition, our result shows that MR obtains higher values of WAFmax and WAAUPRC than PCC for each GO aspect; this is consistent with the fact that MR has replaced PCC as the new co-expression measure in current co-expression databases [24].

In Figure 2C and D, we further compare the performances of the six expression profile-based methods on a subset of 98 non-coding genes in the test datasets of the 8 species, where TNP outperforms again all other five methods. Specifically, based on Fmax, TNP achieves 8.6%, 4.2%, and 8.4% improvements compared to the second-best performer for MF, BP, and CC, respectively. At the same time, the corresponding increases of AUPRC values are 7.7%, 6.5%, and 4.4%, respectively, for the three GO aspects.

In addition, we used the human data to examine the influence of the different characteristics of gene-expression data on the prediction performance of TNP. First, as illustrated in Table S6 and Figure S4 and discussed in File S5, the number of expression samples (*i.e.*, the dimension of expression profile vector) is not critical to the TNP

performance. In fact, the Fmax values of TNP trained on 30% of the expression samples are only slightly (*i.e.*, by 1.3%, 1.5%, and 0.7%) lower than those trained on all the data for MF, BP, and CC, respectively. This is probably due to the inherent redundancy among different samples for the same species. To illustrate this point, we further compared TNP with and without its PCA [32] procedure, which was used to reduce redundant information of expression samples. The result showed that skipping PCA leads to a clear and consistent drop of the performance (Figure S4), which confirms the negative impact of data redundancy on the performance. Finally, we found that there is no strong correlation between the function prediction accuracy of each gene (in terms of gene-level F1-score) and its mean expression level, with a neglectable PCC value ranging from -0.026 to 0.173 for all GO aspects (Figure S5).

Expression similarity has closer correlation with functional similarity in the embedding feature space than the original feature space

One important component of TNP is feature space mapping, in which the expression score calculations are transferred from the original feature space to the embedding feature space (Figure 1B and C). To examine the impact of the feature space mapping on GO prediction, we deigned and executed the following test.

For a query gene, we first rank all genes in a training dataset in descending order of the expression similarity between training gene and query, and select the top K ($K = 100$) genes as templates. In the original feature space, the expression similarity is measured by MR, PCC, MLC, SRC, and ED, respectively. In the embedding space, the expression similarity for TNP is calculated by the square of the ED. Then, the weighted functional similarity (WFS), between templates and query, can be calculated as:

$$\mathrm{WFS} = \frac{\sum_{i=1}^{K} w_i \cdot FS_i}{\sum_{i=1}^{K} w_i}, \quad w_i = 1 - (r_i - 1)/K \tag{12}$$

where w_i and r_i are the weight and rank, respectively, for i-th template, and FS_i is the functional similarity between the i-th template and query measured by F1-score between their experimental GO terms (see File S1A). Finally, the average weighted functional similarity (AVG_WFS) for all test genes is calculated by

$$\text{AVG\_WFS} = \sum_{i=1}^{N_M} \text{WFS}_i / N_M \quad (13)$$

where N_M is the total number of test genes in the M species used here. A higher value of AVG_WFS indicates a closer correlation between expression similarity and functional similarity.

Figure 3 shows the AVG_WFS values of six measures for three GO aspects in the 8 species. For each GO aspect, we found that TNP achieves the highest AVG_WFS among six measures. More specifically, the AVG_WFS values of TNP are 27.4%, 11.1%, and 7.9% higher than those of the second-best performer, MR, for MF, BP, and CC, respectively. Moreover, the AVG_WFS values of six measures in each individual species are listed in Figure S6, where TNP outperforms again other measures in all GO aspects.

As an illustrative, we listed in Figure S7 a scattering plot of F1-score versus weight for a non-coding gene *MIRLET7C* (Entrez ID: 406885) in the test dataset of human species. Here, we used three measures, *i.e.*, TNP, MR, and PCC, to select 100 templates with the highest expression similarity to the query. The expression similarity for different measures can be normalized as weight (w_i in Equation 12). The functional similarity is assessed by F1-score of experimental GO terms between two genes. It can be seen that TNP achieves a higher WFS value than both MR and PCC for each GO aspect, because it selects more templates which have a higher expression similarity (or weight) and functional similarity (F1-score) with the query than the two control measures. Since the data from TNP are directly taken from the embedding space after triplet-network training, these results suggest that the expression similarity for TNP in the embedding feature space has closer correlation with functional similarity compared to the other measures in the original space.

Protein homology inference and triplet network-based expression make most important contributions on TripletGO prediction

To examine the contributions of four component methods in TripletGO, we compared the performances of four individual methods, including EPGP by TNP, GSAGP, PSAGP,

and NGP, and five combination methods, including GSAGP + PSAGP + NGP (GPN), EPGP + PSAGP + NGP (EPN), EPGP + GSAGP + NGP (EGN), EPGP + GSAGP + PSAGP (EGP), and EPGP + GSAGP + PSAGP + NGP (EGPN = TripletGO). To be fair, we optimized the confidence scores of the combination methods using the same network in Figure 1A. **Figure 4**A and B list the WAFmax and WAAUPRC values of all nine methods on the test datasets of 8 species, where the *P* values between EGPN and the other eight methods in Student's *t*-test for WAFmax and WAAUPRC are listed in Table S7. In addition, the performances of all nine methods for each individual species are summarized in Figures S8–S10, Table S8, and discussed in File S6.

From the data in Figure 4A and B, and Table S7, we can conclude that each individual method contributes to improving the TripletGO prediction performance. Specifically, the WAFmax and WAAUPRC values of EGPN are much higher than the corresponding values by each of four individual methods. Importantly, the performance of EGPN is also significantly better than that of the other four combination methods. In terms of WAFmax, for example, EGPN gains 6.7%, 4.0%, 9.5%, and 1.1% average improvements for three GO aspects in comparison with GPN, EPN, EGN, and EPG, respectively. At the same time, the corresponding average increases of WAAUPRC are 12.3%, 6.3%, 14.2%, and 1.4%. The first and second largest increases are caused by adding PSAGP to EGN and adding EPGP to GPN, respectively, in BP and CC aspects. In addition, among the four individual methods, PSAGP and EPGP achieve the best performance for MF and BP/CC, respectively. These data demonstrate the importance of the protein-level homology inference and triplet network-based expression, separately, to the TripletGO prediction.

We further investigated the contributions of proposed methods for non-coding genes. Since non-coding genes have no available prediction results in PSAGP, we compared the performances of three individual gene-level methods and four combination methods, including GSAGP + NPG (GN), EPGP + NPG (EN), EPGP + GSAGP (EG), and EPGP + GSAGP + NGP (EGN = TripletGO). Figure 4C and D illustrate the Fmax and AUPRC values of seven GO prediction methods for 98 non-coding genes. The *P* values between

EGN and other six methods in Student's *t*-test for Fmax and AUPRC are shown in Table S9. Again, we can see that each of three gene-level methods helps to improve accuracy of GO prediction for non-coding genes. On the basis of Fmax, for example, EGN achieves the best performance among seven methods. Specifically, EGN gains 7.7%, 1.5%, and 4.2% improvements for MF, BP, and CC, respectively, compared to the second-best performer. With respect to AUPRC, although EGN shows a slightly lower value than EPGP and EN in BP, it achieves the best performance for MF and CC. In addition, EPGP significantly outperforms other two individual methods through all score metrices, which highlights again the importance of the transcript expression component to the TripletGO prediction.

Comparison of TripletGO with existing gene function prediction methods

We compared TripletGO with two most recently developed gene function prediction approaches, *i.e.*, GENETICA [7] and GeneNetwork [44], which were both based on expression profiles. Different from our work, these two approaches are designed at the term-centric level. Specifically, for a GO term Q_i, each gene is labeled as "1" or "0", where "1" means this gene is associated with Q_i in the experimental annotation. Then, each gene is assigned with a confidence score for Q_i using leave-one-out strategy. Finally, the area under receiver operating characteristic curve (AUROC) is used to evaluate the prediction performance of Q_i by combining the confidence scores and labels for all genes. In light of this, our models are compared with GENETICA and GeneNetwork by term-centric evaluation.

Between our work and GENETICA, there are 287 MF terms, 1340 BP terms, and 186 CC terms in common for human genes. As for mouse genes, there are 149, 1230, and 128 common terms for MF, BP, and CC, respectively (see File S7A). **Figure 5**A and B plot the distributions of AUROC values by GENETICA, TNP, and TripletGO on three GO aspects in human and mouse, respectively. Figure S11A and B shows the mean and median AUROC values for three methods. While TNP and GENETICA are both expression profile-based models, the former shows a better performance than the latter.

In human genes, TNP achieves 16.5%, 10.8%, 14.4% increases of the mean AUROC for MF, BP, and CC, respectively, compared to GENETICA. At the same time, the corresponding increases of median AUROC are 23.0%, 8.5%, and 12.7%. As for mouse, although TNP gains a slightly lower median AUROC than GENETICA in CC, it achieves significant improvements of the corresponding measures in MF and BP. In addition, TripletGO shows a significantly better performance than both TNP and GENETICA, mainly because it integrates complementary information from sources other than expression profiles.

There are 165, 522, and 182 common terms for MF, BP, and CC, respectively, in human genes between our work and GeneNetwork (see File S7B). Figure 5C shows the AUROC distributions of three GO aspects for GeneNetwork, TNP, and TripletGO, separately, where Figure S11C illustrates the mean and median AUROC values of three models. For each GO aspect, TNP shows higher mean and median AUROC values in comparison with GeneNetwork, while TripletGO outperforms both due to the integration of additional information from sequence homology alignments and prior statistics of Gene-GOA databases.

In File S7C and Figure S12, we made a further comparison of our methods with GENETICA and GeneNetwork in the gene-center level based on Fmax and AUPRC, where a similar trend (*i.e.*, TripletGO and TNP outperform the control methods) but with more significant distinctions between the methods can be seen as the term-centric comparisons.

Testing on the third CAFA challenge (CAFA3) targets

We further tested our methods on the dataset of CAFA3. The entire CAFA3 dataset consists of 66,841 training and 3328 test proteins [19] from 23 species. Since some targets have no available gene expression data, we only benchmarked our methods on the 2433 CAFA3 test proteins whose coding genes are originated from 7 species (human, mouse, Arabidopsis, rat, fly, budding yeast, and fission yeast) and have available expression profiles in COXPRESdb [24] or ATTED-II [21] databases. It should be

noted that we did not find any test proteins with expression data from nematoda species. The details of training and test datasets for the 7 species are summarized in Table S10. For each species, we randomly selected 90% training samples to re-train the TNP model and the remaining training samples were used to optimize the parameters of the model. Moreover, for GSAGP, PSAGP, and NGP, the entire CAFA3 training dataset was used to construct the corresponding template databases and prior probabilities of GO terms.

Figure 6A and B summarize the performance of six transcript EPGP methods on the 2433 test proteins, where the *P* values between TNP and the other five methods in Student's t-test [43] for Fmax and AUPRC are listed in Table S11. It can be found that TNP shows better performance than other five methods for all GO aspects. Compared to the second-best performer (MR), TNP achieves 10.7% and 11.2% average improvements on three GO aspects for Fmax and AUPRC, respectively, where the *P* value is statistically significant for all the comparisons except for AUPRC on MF ($1.41\mathrm{E}-01$) and CC ($8.01\mathrm{E}-01$). The performances of the six methods for each individual species are summarized in Figure S13 and Table S12, where TNP achieves the highest values of Fmax and AUPRC among six methods for each GO aspect in most species (see discussion in File S8).

We further benchmarked five proposed GO prediction methods, including four individual methods (*i.e.*, EPGP, GSAGP, PSAGP, and NGP) and their combination (*i.e.*, TripletGO), on the 2433 CAFA3 test proteins. In addition, we included two third-party protein function prediction methods (DeepGO [10] and FunFams [11]) which are the only methods that have downloadable programs and allow us to test independently on our selected CAFA dataset. Meanwhile, they represent two typical types of protocols: DeepGO is a machine learning-based method through combining convolutional neural network with protein sequence encoding, while FunFams is a template-based method and searches the functional templates using protein family information. Among them, FunFams is one of the top-performing methods and ranked at 2/4/9 position in MF/BP/CC aspects with respect to Fmax in the CAFA3 experiment [19]. To make a fair comparison between template-based and non-template-based methods, we used the pre-

set cutoff (*i.e.*, $t_1 = 60\,\%$ and $t_2 = 30\%$, see "Method") to exclude close homologies when running GSAGP and PSAGP; however, we did not exclude any homologies for the third-party programs and ran them under the default setting.

Figure 6C and D summarize the performance of seven GO prediction methods, where the *P* values between TripletGO and the other six methods in Student's t-test [43] are listed in Table S13. We found that the composite GO prediction method, *i.e.*, TripletGO, achieves a significantly better performance than other six GO prediction methods in all GO aspects, including both the third-party (FunFams and DeepGO) and the component methods of TripletGO, demonstrating again the advantage of integrating gene expression and sequence profile-based approach to function predictions.

Case studies for different GO prediction methods

As illustrations, we selected two genes from the human genome: *GALNT4* (protein-coding gene, Entrez ID: 8693) and *MIRLET7C* (non-coding gene, Entrez ID: 406885), to examine the effects of different GO prediction methods. Here, each gene is associated with 12 GO terms for CC aspect from experimental annotations. **Table 1** summarizes the numbers of correctly predicted GO terms (*i.e.*, true positives) and mistakenly predicted terms (*i.e.*, false positives) in CC aspect for the two genes by ten different methods, including six individual gene-expression methods (MR, PCC, MLC, SRC, ED, and TNP), a gene sequence alignment method (GSAGP), a protein sequence alignment method (PSAGP), a naïve-based approach (NGP), and a composite approach (TripletGO). **Figure 7** plots the directed acyclic graph of GO terms in native annotation and the correctly predicted GO terms of ten methods for the two genes. Moreover, the incorrectly predicted GO terms (*i.e.*, false positives) of each method are listed in **Table 2**. It should be noted that the predicted GO terms of each method are determined by its own cut-off value to achieve the highest Fmax value.

Several interesting observations can be made from the data. First, among six gene expression-based methods, TNP can correctly recognize the most GO terms with the least false positives for each gene. Moreover, all true positives for other five methods

can be effectively identified by TNP. More importantly, for gene *MIRLET7C*, TNP correctly recognizes 1 additional GO term, *i.e.*, GO:0005634, which are missed by the other five methods. This observation shows that TNP can predict gene function in a more precise level, because it successfully identifies some children GO terms, in which other expression-based methods fail.

Second, the combination of complementary methods increases both coverage and accuracy of the TripletGO models. In gene *GALNT4* (see Figure 7A), the four component methods (TNP, GSAGP, PSAGP, and NGP) hit 10 true positives in total, which is more than that by each individual method, indicating that the component methods derive complementary information from different sources. By taking the combination, TripletGO achieves the highest coverage with TP = 10 and the lowest false positive rate (FP = 0). Sometimes, one component method can cover all true positives predicted by other methods. For example, for gene *MIRLET7C* (see Figure 7B), all true positives of TNP and NGP are covered by GSAGP. Even in this case, the final TripletGO accuracy is not degraded by the inclusion of a less accurate method, where TripletGO shares the same performance with the best individual method by GSAGP.

In addition, to further explain what is considered as a positive prediction regarding the hierarchy, we choose *GALNT4* as an illustrative example and list the confidence scores of all the candidate GO terms by TripletGO in Table S14, where the 10 GO terms whose confidence scores are higher than the cut-off value (0.35) have been predicted as positives. In Figure S14, we plot the directed acyclic graph of the 10 predicted GO terms with corresponding confidence scores. It can be found the confidence scores of the parent terms are higher than the scores of their children, following the post-processing Equation 8.

Conclusions

We developed a new method, TripletGO, to predict the function of both protein-coding and non-coding genes by the integration of four gene-expression and protein homology

inference pipelines. The large-scale benchmark tests on 5754 non-redundant genes from a set of 8 species demonstrated that TripletGO consistently achieved significant improvements in comparison with other state-of-the-art gene function prediction methods. Detailed analysis showed that the major advantage of TripletGO stems from two aspects. First, the new triplet network-based algorithm, when coupled with feature space mapping, efficiently recognizes functional patterns from transcript expression profiles. Second, the combination of multiple complementary pipelines, especially those with protein-level homology inference and transcript expression profile, significantly improves the coverage and accuracy of the gene function annotations.

Despite the encouraging performance, there is still considerable room for further improvements. First, the TNP needs large amounts of gene expression data with GO annotation to train the prediction model. For some species, such as dog and chicken as listed in Table S1, the number of genes with GO annotation is very limited, and for many other species of interest, no such data are available. As a result, we cannot train prediction models using the TNP from expression profiles for these species. Therefore, an extended TNP model by normalizing expression profiles across different species may help solve the issue, as well as further improve the overall accuracy of the current approach. Second, the confidence scores of the four individual methods are integrated as a consensus score by a simple one-layer neural network, where an advanced machine-learning approach may help better integrate confidence scores. Meanwhile, new GO prediction methods considering other biological aspects, such as protein–protein interactions and protein/nucleic acid structures, will help improve both the coverage and accuracy of the current gene function annotation algorithms. Studies along these lines are under progress.

Code availability

The source code has been submitted to BioCode and is available at https://ngdc.cncb.ac.cn/biocode/tools/7277.

Data availability

The online server, standalone package, and all benchmark datasets and libraries are available at https://zhanglab.ccmb.med.umich.edu/TripletGO/.

CRediT author statement

Yi-Heng Zhu: Methodology, Validation, Writing - original draft, Data curation, Software. **Chengxin Zhang:** Software, Writing - review & editing. **Yan Liu:** Writing - review & editing. **Gilbert S. Omenn:** Writing - review & editing. **Peter L. Freddolino:** Writing - review & editing. **Dong-Jun Yu:** Writing - review & editing, Supervision. **Yang Zhang:** Conceptualization, Methodology, Writing - original draft, Supervision. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interest.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant Nos. 62072243 and 61772273 to Dong-Jun Yu), the Natural Science Foundation of Jiangsu (Grant No. BK20201304 to Dong-Jun Yu), the Foundation of National Defense Key Laboratory of Science and Technology (Grant No. JZX7Y202001SY000901 to Dong-Jun Yu), China Scholarship Council (Grant No. 201906840041 to Yi-Heng Zhu), the National Institute of Environmental Health Sciences (Grant No. P30ES017885 to Gilbert S. Omenn), the National Cancer Institute (Grant No. U24CA210967 to Gilbert S. Omenn), the National Institute of General Medical Sciences (Grant Nos. GM136422 and S10OD026825 to Yang Zhang), the National Institute of Allergy and Infectious Diseases (Grant No. AI134678 to Peter L. Freddolino and Yang Zhang), and the National Science Foundation (Grant Nos.

IIS1901191, DBI2030790, and MTM2025426 to Yang Zhang). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (Grant No. ACI1548562). The work was done when Yi-Heng Zhu visited the University of Michigan.

ORCID

0000-0002-3857-1533 (Yi-Heng Zhu)

0000-0001-7290-1324 (Chengxin Zhang)

0000-0002-5331-3655 (Yan Liu)

0000-0002-8976-6074 (Gilbert S. Omenn)

0000-0002-5821-4226 (Peter L. Freddolino)

0000-0002-6786-8053 (Dong-Jun Yu)

0000-0002-2739-1916 (Yang Zhang)

References

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.

[2] Murali TM, Wu C-J, Kasif S. The art of gene function prediction. Nat Biotechnol 2006;24:1474–5.

[3] Zhang J, Zhang Z, Chen Z, Deng L. Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. IEEE/ACM Trans Comput Biol Bioinform 2019;16:396–406.

[4] Peng J, Xue H, Wei Z, Tuncali I, Hao J, Shang X. Integrating multi-network topology for gene function prediction using deep neural networks. Brief Bioinform 2021;22:2096–105.

[5] Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, et al. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 2012;40:D76–83.

[6] Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, et al. GeneMANIA update 2018. Nucleic Acids Res 2018;46:W60–4.

[7] Urzúa-Traslaviña CG, Leeuwenburgh VC, Bhattacharya A, Loipfinger S, van Vugt MATM, de Vries EGE, et al. Improving gene function predictions using independent transcriptional components. Nat Commun 2021;12:1464.

[8] Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Res 2017;45:W291–9.

[9] Zhang C, Zheng W, Freddolino PL, Zhang Y. MetaGO: predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. J Mol Biol 2018;430:2256–65.

[10] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics 2018;34:660–8.

[11] Das S, Lee D, Sillitoe I, Dawson NL, Lees JG, Orengo CA. Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics 2015;31:3460–7.

[12] Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, et al. Structure-based protein function prediction using graph convolutional networks. Nat Commun 2021;12:3168.

[13] Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nat Mach Intell 2020;2:540–50.

[14] Smaili FZ, Tian S, Roy A, Alazmi M, Arold ST, Mukherjee S, et al. QAUST: protein function prediction using structure similarity search, protein interaction and functional sequence motifs. Genomics Proteomics Bioinformatics 2021; https://doi.org/10.1016/j.gpb.2021.02.001.

[15] Esteller M. Non-coding RNAs in human disease. Nat Rev Genet 2011;12:861–74.

[16] Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. Proc Natl Acad Sci U S A 2019;116:27151–8.

[17] Makrodimitris S, Reinders MJT, van Ham RCHJ. Metric learning on expression data for gene function prediction. Bioinformatics 2020;36:1182–90.

[18] Ray SS, Misra S. Genetic algorithm for assigning weights to gene expressions using functional annotations. Comput Biol Med 2019;104:149–62.

[19] Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 2019;20:244.

[20] 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.

[21] 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.

[22] Girolami M. Mercer kernel-based clustering in feature space. IEEE Trans Neural Netw 2002;13:780–4.

[23] Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering. Proc 28th IEEE Conf CVPR. 2015:815–23.

[24] 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.

[25] UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2015;43:D204–12.

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

[27] Zhang Z, Sabuncu MR. Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv 2018; https://doi.org/10.48550/arXiv.1805.07836.

[28] Taha A, Chen Y-T, Misu T, Shrivastava A, Davis L. Boosting standard classification architectures through a ranking regularizer. arXiv 2019; https://doi.org/10.48550/arXiv.1901.08616.

[29] Zhou Q, Zhong B, Lan X, Sun G, Zhang Y, Zhang B, et al. Fine-grained spatial alignment model for person re-identification with focal triplet loss. IEEE Trans Image Process 2020;29:7578–89.

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

[31] Patro S, Sahu KK. Normalization: a preprocessing stage. arXiv 2015; https://doi.org/10.48550/arXiv.1503.06462.

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

[33] Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: Mira J, Sandoval F, editors. From natural to artificial neural computation. Berlin: Springer; 1995, p.195–201.

[34] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv 2017; https://doi.org/10.48550/arXiv.1703.07737.

[35] Hoffer E, Ailon N. Deep metric learning using triplet network. In: Feragen A, Pelillo M, Loog M, editors. Similarity-based pattern recognition. Berlin: Springer; 2015, p.84–92.

[36] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2014; https://doi.org/10.48550/arXiv.1412.6980.

[37] Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2011;39:D38–51.

[38] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.

[39] Eckle K, Schmidt-Hieber J. A comparison of deep networks with ReLU activation function and linear spline-type methods. Neural Netw 2019;110:232–42.

[40] Gillis J, Pavlidis P. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). BMC Bioinformatics 2013;14 Suppl 3:S15.

[41] Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, editots. Machine learning and knowledge discovery in databases. Berlin: Springer; 2013, p.451–66.

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

[43] 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.

[44] Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat Commun 2019;10:2837.

Figure legends

Figure 1 The procedures of TripletGO

A. The flowchart of TripletGO to integrate four complementary pipelines for GO prediction. **B.** The design of a triplet-network for assessing expression profile similarity. **C.** TNP for EPGP. GO, gene ontology; EPGP, expression profile-based GO prediction; GSAGP, genetic sequence alignment-based GO prediction; PSAGP, protein sequence alignment-based GO prediction; NGP, naïve-based GO prediction; ID, identity document; TNP, triplet network-based pipeline.

Figure 2 Comparison of six transcript EPGP methods

A. The WAFmax values on the test datasets of 8 species. **B.** The WAAUPRC values on the test datasets of 8 species. **C.** The Fmax values on 98 non-coding test genes. **D.** The AUPRC values on 98 non-coding test genes. MR, mutual rank; PCC, Pearson correlation coefficient; MLC, metric learning for co-expression; SRC, Spearman rank correlation; ED, Euclidean distance; Fmax, maximum F1-score; AUPRC, area under the precision-recall curve; WAFmax, weighted average Fmax; WAAUPRC: weighted average AUPRC; MF, molecular function; BP, biological process; CC, cellular component.

Figure 3 Comparison of the AVG_WFS values by six transcript EPGP methods on 8 test species

AVG_WFS, average weighted functional similarity.

Figure 4 Comparison of GO prediction results using different methods

A. The WAFmax values on the test datasets of 8 species for nine GO prediction methods. **B.** The WAAUPRC values on the test datasets of 8 species for nine GO prediction methods. **C.** The Fmax values on 98 non-coding genes for seven GO prediction methods. **D.** The AUPRC values on 98 non-coding genes for seven GO prediction methods. EGPN = EPGP + GSAGP + PSAGP + NGP; GPN = GSAGP + PSAGP + NGP; EPN = EPGP + PSAGP + NGP; EGN = EPGP + GSAGP + NGP; EGP = EPGP + GSAGP +

PSAGP; GN = GSAGP + NPG; EN = EPGP + NPG; EG = EPGP + GSAGP.

Figure 5 Comparison of AUROC values of three GO aspects by different methods on the common dataset

A. GENETICA, TNP, and TripletGO on human genes. **B.** GENETICA, TNP, and TripletGO on mouse genes. **C.** GeneNetwork, TNP, and TripletGO on human genes. In each box, the median line and triangle represent the median and mean AUROC values, respectively. AUROC, the area under receiver operating characteristic curve.

Figure 6 Performance comparison on 2433 test proteins of 7 species from CAFA3 benchmark dataset

A. The Fmax values for six EPGP methods. **B.** The AUPRC values for six EPGP methods. **C.** The Fmax values for five proposed GO prediction methods and two existing GO prediction methods. **D.** The AUPRC values for seven GO prediction methods.

Figure 7 The directed acyclic graphs of 12 GO terms in the experimental annotation on two illustrative genes

A. The directed acyclic graphs for gene *GALNT4*. **B.** The directed acyclic graphs for gene *MIRLET7C*. The circles above each GO term refer to the prediction methods, where a circle filled with "X" on GO term "Y" indicates that method "X" can correctly predict term "Y".

Table 1 The modeling results of ten GO prediction methods on two illustrative genes

Table 2 The incorrectly predicted GO terms for ten GO prediction methods on two illustrative genes

Supplementary material

File S1 The additional definitions in TNP

A. The functional similarity for genes. **B.** Distance rank-based strategy

File S2 The construction procedures of databases in sequence alignment-based GO prediction methods

A. The construction procedures of Gene-GOA. **B.** The construction procedures of genetic sequence database with GO annotation. **C.** The construction procedures of protein sequence database.

File S3 The relationship between protein sequence identity and genetic sequence identity

File S4 Performance comparison between six EPGP methods

A. The procedures of GBA strategy for EPGP. **B.** The performances of six EPGP methods for each individual species

File S5 Exploring the influence of the characteristics of expression data on prediction performance for human species

File S6 The performances of nine GO prediction methods for each individual species

File S7 Performance comparison between TripletGO and existing gene function prediction models

A. Finding common genes and GO terms between our datasets and GENETICA's datasets. **B.** Finding common genes and GO terms between our datasets and GeneNetwork's datasets. **C.** Comparison with the existing gene function prediction models in gene-center level

File S8 The performances of six EPGP methods for each individual species on CAFA3 test dataset

Figure S1 The distribution of sequence identities for 10,000 gene–gene pairs and

10,000 mapped protein–protein pairs

Figure S2 The performance of six expression profile-based methods on the test datasets for 8 individual species

A. The Fmax values of six methods for 8 species. **B.** The AUPRC values of six methods for 8 species.

Figure S3 The PRCs of six expression profile-based methods on the test datasets for 8 individual species

PRCs, the precision-recall curves.

Figure S4 Variation curves of Fmax values of TNP and NON-PCA-TNP on the test dataset of human species versus the sampling ratios in expression data

NON-PCA-TNP, triplet network-based pipeline without principal component analysis.

Figure S5 The scattering plots of mean expression level versus F1-scores for 1470 human test genes by TNP

Figure S6 The AVG_WFS values of six measures for three GO aspects in 8 individual species

Figure S7 The scattering plots of weights versus F1-scores of 100 templates for the gene *MIRLET7C* over TNP, MR, and PCC

Figure S8 The Fmax values of nine GO prediction methods on the test datasets for 8 individual species

Figure S9 The AUPRC values of nine GO prediction methods on the test datasets for 8 individual species

Figure S10 The PRCs of five GO prediction methods on the test datasets for 8 individual species

Figure S11 Comparison of mean and median AUROC values of three GO aspects by different methods on the common dataset

A. GENETICA, TNP, and TripletGO on human genes. **B.** GENETICA, TNP, and TripletGO on mouse genes. **C.** GeneNetwork, TNP, and TripletGO on human genes.

Figure S12 Comparison of Fmax and AUPRC values of three GO aspects by different methods on the common dataset

A. GENETICA, TNP, and TripletGO on human genes. **B.** GENETICA, TNP, and TripletGO on mouse genes. **C.** GeneNetwork, TNP, and TripletGO on human genes.

Figure S13 The performance of six expression profile-based methods for 7 species on CAFA3 test dataset

A. The Fmax values of six methods. **B.** The AUPRC values of six methods.

Figure S14 The directed acyclic graph of predicted GO terms with corresponding confidence scores for gene *GALNT4* by TripletGO

Table S1 The number of genes with GO annotation of three aspects for 20 species

Table S2 The details of 8 benchmark datasets constructed in our work

Table S3 The values of α, h, $margin$, and c_f on the benchmark datasets for 8 species

Table S4 The *P* values between TNP and other five expression profile-based methods for WAFmax and WAAUPRC

Table S5 The *P* values between TNP and other five expression profile-based methods for Fmax and AUPRC on 8 species

Table S6 The Fmax values of TNP and NON-PCA-TNP on the test dataset of human species for different sampling ratios in expression data

Table S7 The *P* values between EGPN and other eight GO prediction methods for WAFmax and WAAUPRC on the test datasets of 8 species

Table S8 The *P* values between EGPN and other eight GO prediction methods for Fmax and AUPRC on 8 species

Table S9 The *P* values between EGN and other six GO prediction methods for Fmax and AUPRC on 98 non-coding genes

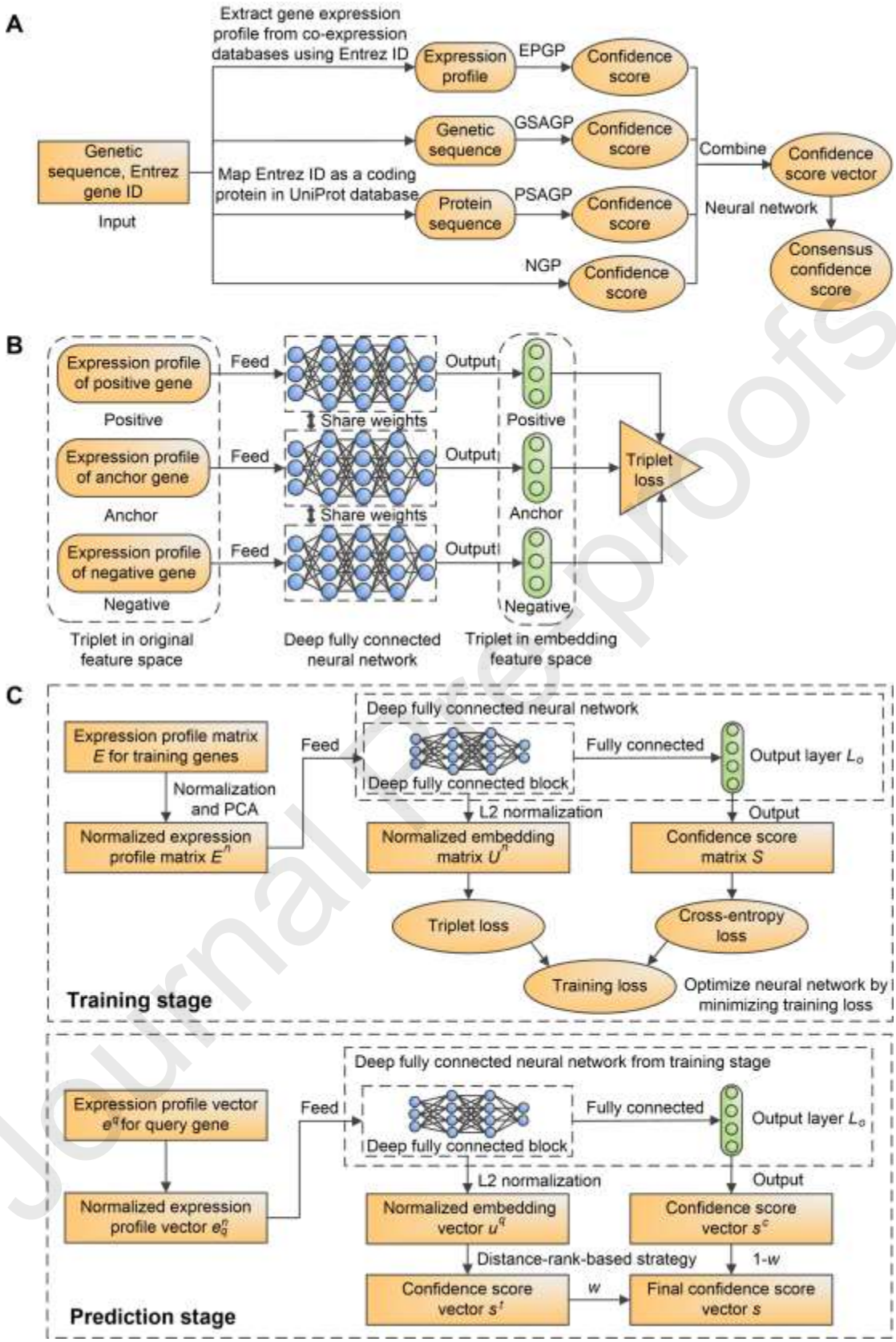
Table S10 The details of training and test datasets for 7 species in CAFA3 dataset

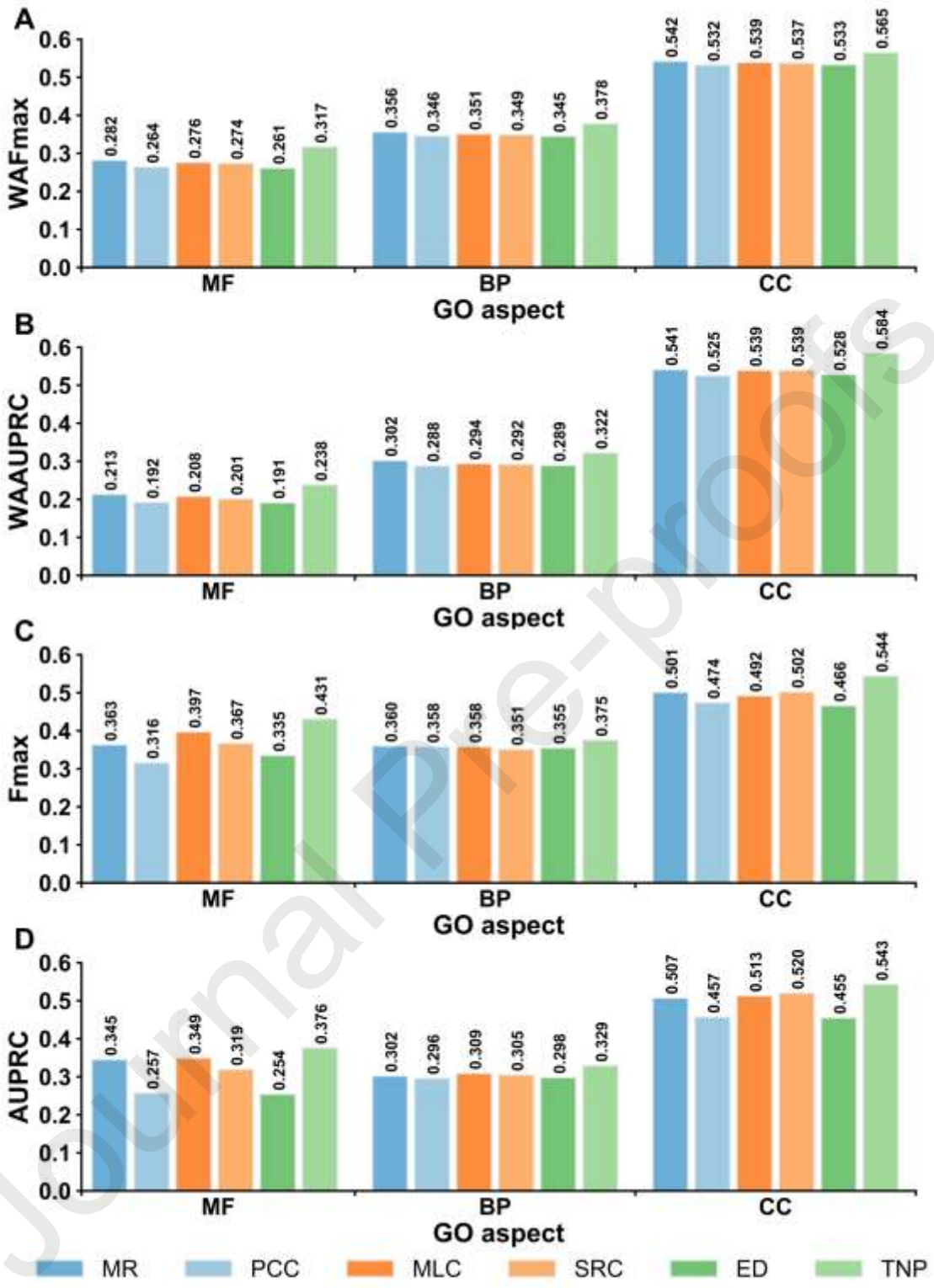
Table S11 The *P* values between TNP and other five expression profile-based methods for Fmax and AUPRC on 2433 proteins of 7 species from CAFA3 test dataset

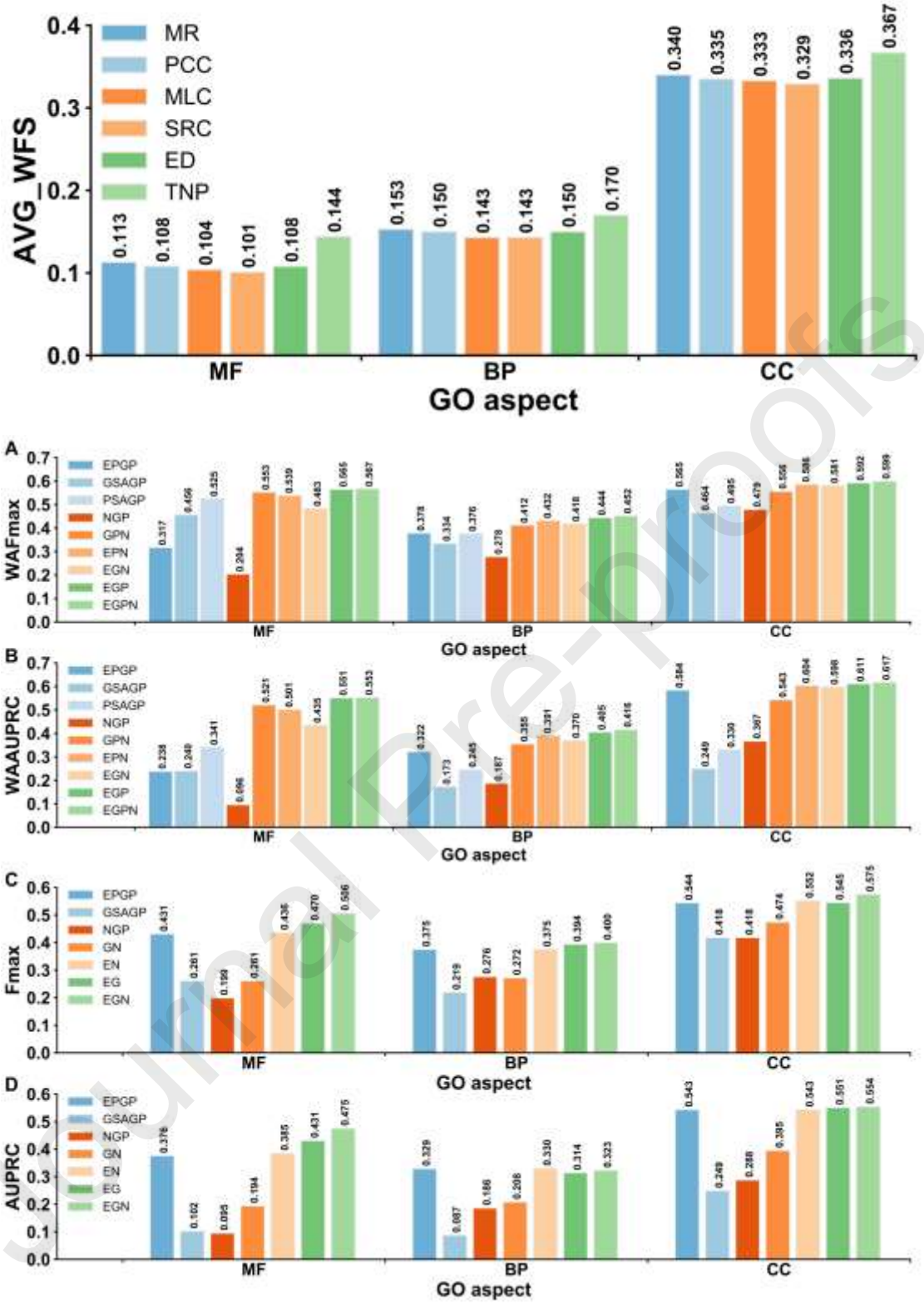
Table S12 The *P* values between TNP and other five expression profile-based methods for Fmax and AUPRC on CAFA3 test dataset for each of 7 species

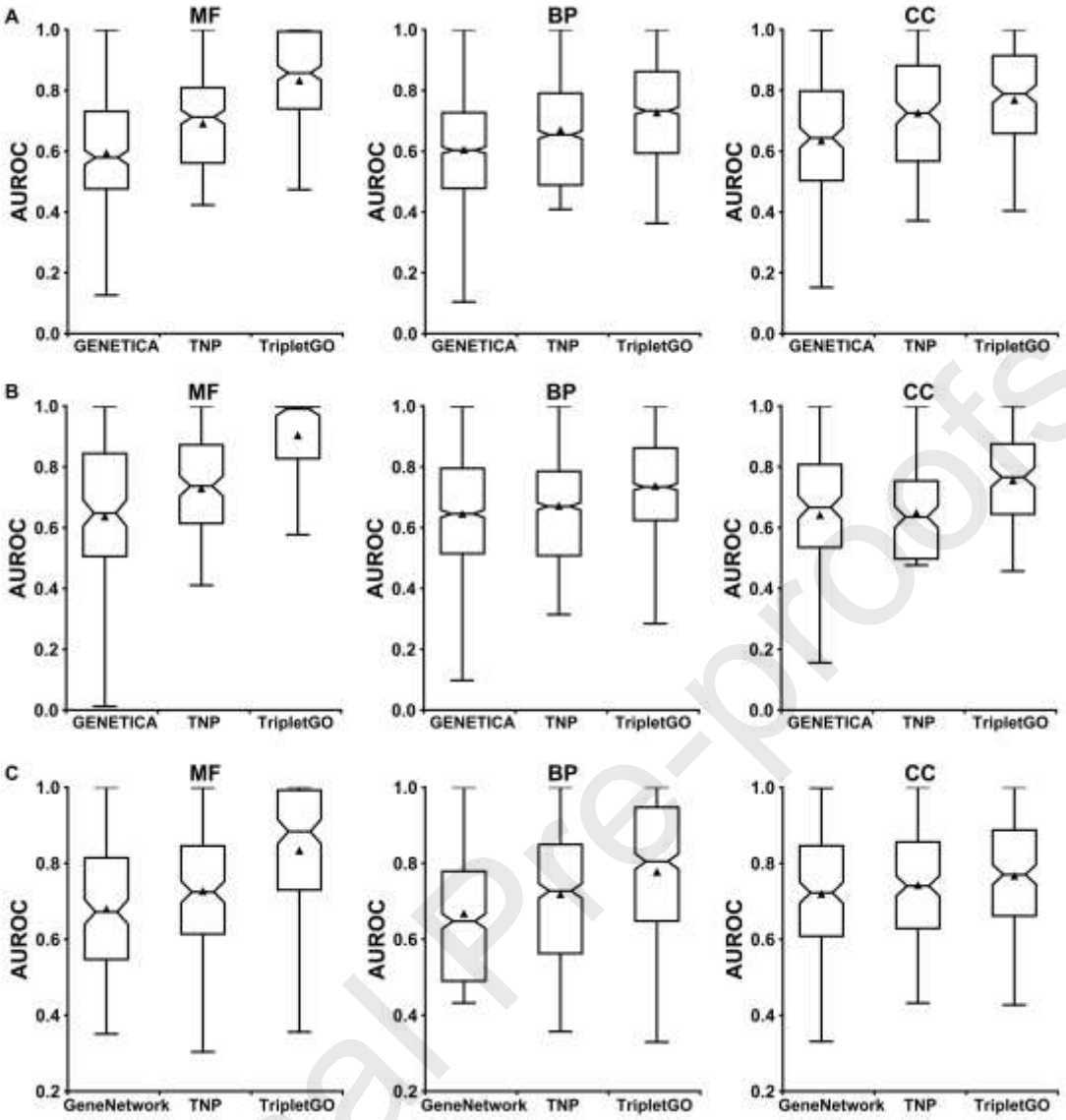
Table S13 The *P* values between TripletGO and other six GO prediction methods for Fmax and AUPRC on 2433 proteins of 7 species from CAFA3 test dataset

Table S14 The confidence scores of the candidate GO terms for gene *GALNT4* by TripletGO









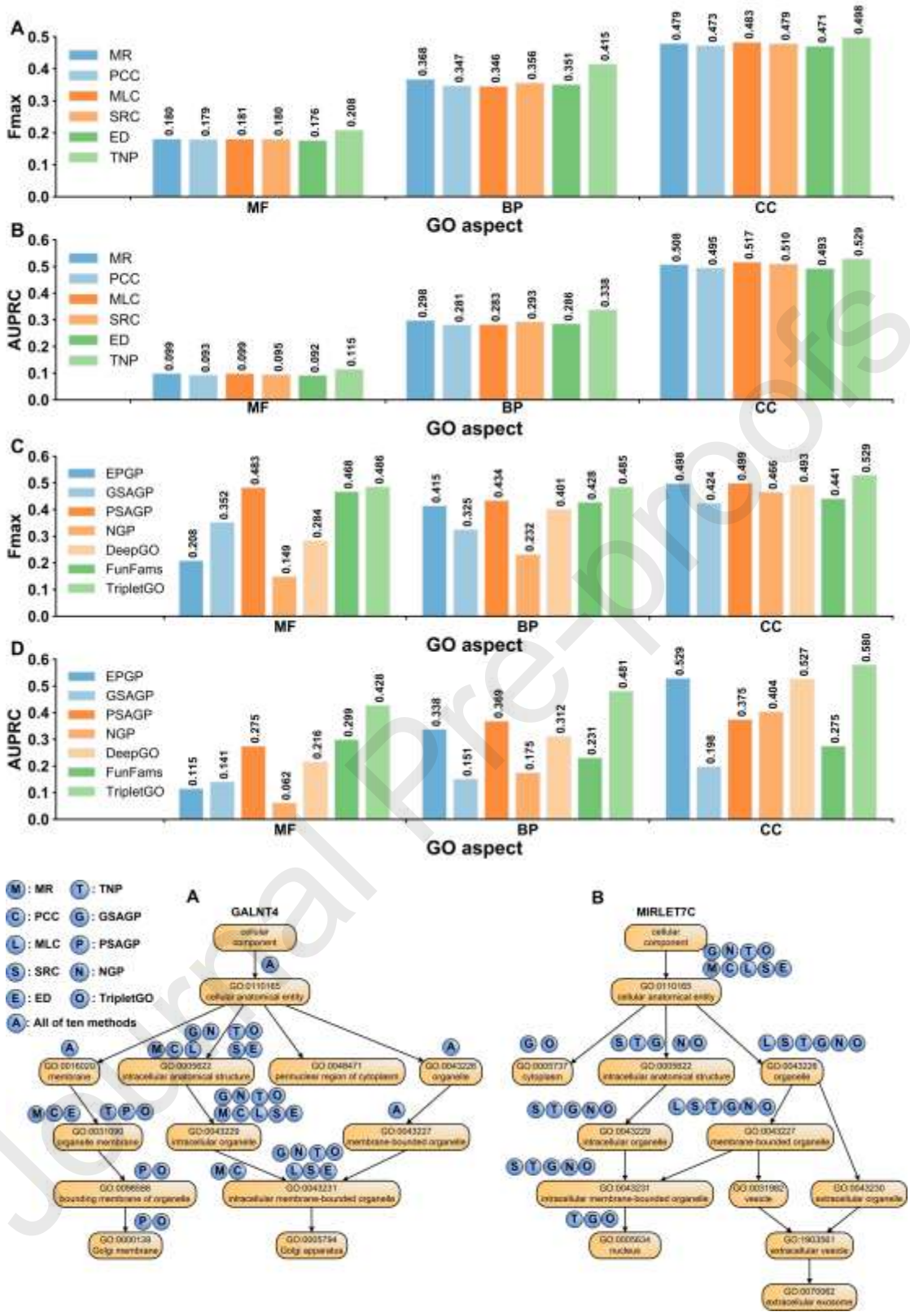


Table 1 The modeling results of ten GO prediction methods on two illustrative genes

Gene	Measure	MR	PCC	MLC	SRC	ED	TNP	GSAGP	PSAGP	NGP	TripletGO

GALNT4	NTP	8	8	7	7	8	8	7	7	7	**10**
	NFP	3	1	2	3	1	**0**	3	4	4	**0**
MIRLET7C	NTP	1	1	3	6	1	7	**8**	0	6	**8**
	NFP	2	2	2	2	2	1	**0**	**0**	2	**0**

Note: GO, Gene Ontology; MR, mutual rank; PCC, Pearson correlation coefficient; MLC, metric learning for co-expression; SRC, Spearman rank correlation; ED, Euclidean distance; TNP, triplet network-based pipeline; GSAGP, genetic sequence alignment-based GO prediction; PSAGP, protein sequence alignment-based GO prediction; NGP, naïve-based GO prediction; NTP, the number of true positives; NFP, the number of false positives. Best performers are highlighted in bold fonts in each category.

Table 2 The incorrectly predicted GO terms for ten GO prediction methods on two illustrative genes

Method	***GALNT4***	***MIRLET7C***
MR	GO:0005654 GO:0005829 GO:0032991	GO:0016020 GO:0005886
PCC	GO:0005829	GO:0016020 GO:0005886
MLC	GO:0005829 GO:0032991	GO:0016020 GO:0005886
SRC	GO:0005654 GO:0005829 GO:0005886	GO:0016020 GO:0005886
ED	GO:0005829	GO:0016020 GO:0005886
TNP		GO:0005654
GSAGP	GO:0005654 GO:0005829 GO:0005886	
PSAGP	GO:0031410 GO:0030133 GO:0097708 GO:0031982	
NGP	GO:0005829 GO:0032991 GO:0005634 GO:0005886	GO:0016020 GO:0032991
TripletGO		

Note: incorrectly predicted GO terms are false positives.