TANA: efficient approach for predicting protein functions by transferring annotation via alignment networks

CURRENT STATUS: UNDER REVIEW

BMC Bioinformatics BMC Series

Warith Eddine DJEDDI
Universite de Tunis El Manar

✉ warithsa5@gmail.com *Corresponding Author*
ORCID: <https://orcid.org/0000-0001-5751-2371>

Sadok BEN YAHIA
Tallinn University of technology

Engelbert MEPHU NGUIFO
Universite Clermont Auvergne - Campus des Cezeaux

DOI:

10.21203/rs.2.21071/v1

SUBJECT AREAS

Bioinformatics

KEYWORDS

PPI, networks, alignment, function prediction, neighborhood, gene ontology, computational assessment of function annotation

Abstract

Background: One of the challenges of the post-genomic era is to provide accurate function annotations for orphan and unannotated protein sequences. With the recent availability of huge protein-protein interactions networks for many model species, the computational methods revealed a great requirement to elucidate protein function based on many strategies. In this respect, most computational approaches integrate diverse kinds of functional interactions to unveil protein functions by transferring annotations across different species by relying on similar sequence, structure 2D/3D, amino acid motifs or phylogenetic profiles.

Results: In this work, we introduce a new approach called TANA for inferring protein functions. The main originality of the introduced approach stands on the function prediction for the unannotated protein by transferring annotation via a network alignment as well as from the direct interaction neighborhood within their PPI networks. Doing so, we are able to discover the functions of proteins that could not to be easily described by sequence homology. We assess the performance of our method using the standard metrics established by the CAFA and highlight a sharp significant improvement over other competitive methods, in particular for predicting molecular functions.

Conclusions: This research is one of the first attempts that combine sequence and networks-multiple-alignment-based function prediction approaches. We have been able to assess the accuracy of the prediction using pairwise and multiple alignment of the PPI networks for the compared species. Therefore, we recommend using different strategies (i.e pairwise, multiple, with/without neighborhood networks) especially in situations where the functions of the protein are not known in advance.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

Figure 1

The ROC curves of TANA and other predictors over BPO, MFO and CCO in full evaluation mode.

Figure 2

The precision-recall plot of TANA and other predictors over BPO, MFO and CCO in full evaluation mode.

Figure 3

The different steps of TANA for scoring the prediction of GO terms for each unannotated protein.

Figure 4

Functional annotation yield by TANA. This shows clusters of proteins composed from unannotated proteins as well as annotated ones

Strategy 1: Transferring annotation from the "Direct neighborhood"

Strategy 2: Transferring annotation from the GNA

Figure 5

Two different strategies of protein function prediction are yield by TANA. In the PPI network on the left, proteins A, E, and F are annotated with "green " function, proteins C and D are annotated with "blue" function. The node H is annotated with the "yellow" function and G and B are unannotated proteins (as denoted by white node color). With the "direct neighborhood" approach (the middle panel), node B gets annotated with "green" function, since the majority of its neighbors (E and F) have the "green" function. Similarly, G gets annotated with "blue" function, since all of its neighbors (C and D) have the "blue" function. Adding to that, by using the other strategy based on transferring annotation from the alignment of PPI network, the protein B is also annotated with two shared functions (i.e, orange and black colors) coming from the "Cluster 1". Moreover, protein G also gets annotated with one shared function (i.e., Black color) from the shared function of the

"Cluster 2". It is worth mentioning that: a) the proteins with a shape different from the circle node are a proteins coming from different species; b) The proteins B and G are the proteins with limited Knowledge which had been experimentally annotated in one or two GO ontologies (BP, MF or CC).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

[additional_file1.4.pdf](#)

[additional_file1.3.pdf](#)

[additional_file1.2.pdf](#)

[additional_file1.1.pdf](#)