Determining Rewiring Functional Effects of Alternative Splicing Variants on Protein-Protein Interactions

Nathan T. Johnson[1, $], Oleksandr Narykov [1, $], Dmitry Korkin[1*]
[1] Worcester Polytechnic Institute, Bioinformatics and Computational Biology, 01605, USA
[$] These authors contributed equally to the project
*To whom correspondence should be addressed: dkorkin@wpi.edu

1. MOTIVATION

The success or failure of proteins to interact with each other or DNA and RNA molecules drives almost any complex biological system. Therefore, it is crucial to determine not only the mere presence of a protein-mediated interaction, but also whether any genetic or transcriptional changes in the protein can affect this interaction. Alternative splicing is one of the leading factors in driving protein complexity. However no computational methods currently exist that can accurately predict whether two different isoforms of the same protein retain the same interaction. Here, we present the first ALTernative splicing driven protein-protein INteraction prediction approach (ALT-IN Tool). Our machine learning approach addresses a very important challenge: how to maximize the accuracy of the prediction by including limited structural information about some of the protein-protein interactions, while also maximizing the coverage of this approach and requiring only sequence information for the new predictions.

2. BACKGROUND

Protein-protein interactions (PPIs) underlie the cell's basic functioning. Recent studies of disease networks have linked many genetic variations (e.g., single nucleotide variations, SNVs) and posttranscriptional variations (e.g., AS isoforms) with PPIs. Understanding how the genetic and structural mutations and AS variants can rewire the PPI network mediated by proteins associated with the disease is a critical step in studying complex genetic disorders. Recently, an alternative splicing specific yeast-two-hybrid assay was used to assess how the differentially spliced products affect interaction with their partners (1). This result confirms the high estimate of functional changes due to alternative splicing. However, the interaction landscape determined by the alternative splicing variants of the disease-associated genes is far from being fully reconstructed.

3. METHOD SUMMARY

ALT-IN Tool is designed to address the following problem: Given a PPI that involves a "reference" alternative splicing isoform of a protein, will another isoform of the same protein retain this PPI or destroy it? We formulate this question as a classification problem and develop a classifier by leveraging the LUPI (Learning Under Privileged Information) paradigm instead of a traditional supervised learning method (2). The recently developed LUPI paradigm allows for a classifier to train using enriched, or "privileged", information about the training set, that is absent for a new set of examples to be classified. Here we use the limited structural information of PPIs to derive the privileged features, while the sequence-based information is available for regular features. Both, sequence- and structure-based features are extracted from the training set and used in the training of LUPI classifier, while only sequence-based features are generated to predict the rewiring effect for any new pair of isoforms. One of the unique features is a developed novel statistical potential that makes use of the frequency on protein domains interacting from our database on structural interactions of macromolecules, DOMMINO (3).

4. RESULTS

The method is trained using 2501 interactions from 638 genes with 881 alternative spliced isoforms whose rewiring properties were experimentally determined from a yeast-two-hybrid assay (1). After training the model using LUPI, stratified 10 fold cross-validation is used to test the model. The baseline accuracy of using a supervised learning method by utilizing only sequenced based features is 83% with 0.89 f-score. The LUPI trained approach on the same dataset was able to improve the supervised learning prediction accuracy and f-score to 87% and 0.92, respectively. ALT-IN tool was applied to several case studies of disease-associated AS variants to study their effects on the previously known PPIs providing new mechanistic insights into their roles in the diseases.

5. CONCLUSION

This works presents the first computational method to predict the perturbation of protein-protein interactions by the differentially spliced protein isoforms.

6. REFERENCES

1. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., ... & Spirohn, K. 2016. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164(4):805-817.

2. Vapnik, V., & Vashist, A. 2009. A new learning paradigm: Learning using privileged information. *Neural networks* 22(5):544-557.

3. Kuang, X., Han, J. G., Zhao, N., Pang, B., Shyu, C. R., & Korkin, D. 2012. DOMMINO: a database of macromolecular interactions. *Nucleic acids research* 40(D1):D501-D506.