

Identifying the topology of protein complexes from affinity purification assays

Supplementary Material

Caroline C. Friedel and Ralf Zimmer
Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstraße 17, 80333 München, Germany

1 SUPPLEMENTARY RESULTS

1.1 Effect of measurement errors in reference networks

Since the Y2H system is prone to measurement errors, the reference networks used in this study likely contain both false positive and false negative interactions. This is in particular true for the non-core interactions from the study of Ito et al. (2001) and the predicted interactions from Y2H networks for other species or domain-domain interactions. In the following, we show that the correct ranking of prediction methods can still be obtained based on ratios of true positive rate (TPR) to false positive rate (FPR) on the reference networks given certain requirements.

By measuring the TPR/FPR ratio on a reference network \( R \) for an algorithm \( A_i \), we estimate the ratio

\[
 r'_i = \frac{P_i(S|R)}{P_i(S|C \setminus R)}. \tag{1}
\]

Here, \( P_i(S|R) \) is the probability that an interaction contained in the reference network (within complexes) is predicted as positive and included in the scaffold network. \( P_i(S|C \setminus R) \) is the probability that an interaction within complexes but not included in the reference network is predicted as positive. To compare the accuracy of the prediction methods, we need the true TPR/FPR ratio

\[
 r_i = \frac{P_i(S|P)}{P_i(S|N)}. \tag{2}
\]

Here, \( P_i(S|P) \) is the probability that a true physical interaction is included in the scaffold network and \( P_i(S|N) \) the probability that a negative interaction is wrongly predicted as a physical interaction in the scaffold network.

In the following, we show that \( r'_i \) for one algorithm \( A_i \) is larger than \( r'_j \) for another algorithm \( A_j \) if and only if \( r_i \) is larger than \( r_j \). For this purpose, let \( \delta_{TP} \) be the fraction of true physical interactions among the reference interactions. Furthermore, let \( \delta_{TN} \) be the fraction of interactions not in the reference network for which the corresponding proteins truly do not interact.

**Lemma 1.** If \( \delta_{TP}\delta_{TN} > (1 - \delta_{TP})(1 - \delta_{TN}) \), we have for any two Algorithms \( A_i \) and \( A_j \) that

\[
 r'_i > r'_j \iff r_i > r_j. \tag{3}
\]

**Proof.** Algorithm \( A_i \) predicts a true physical interaction for the scaffold network with probability \( P_i(S|P) \). Furthermore, a true negative interaction is included in the scaffold network with probability \( P_i(S|N) \). Thus, the probability that a reference interaction is predicted for the scaffold is

\[
 P_i(S|R) = \delta_{TP}P_i(S|P) + (1 - \delta_{TP})P_i(S|N) \tag{4}
\]

and the probability that an interaction not contained in the reference network is predicted for the scaffold is

\[
 P_i(S|C \setminus R) = (1 - \delta_{TN})P_i(S|P) + \delta_{TN}P_i(S|N). \tag{5}
\]

This results in the following equation for \( r'_i \):

\[
 r'_i = \frac{\delta_{TP}P_i(S|P) + (1 - \delta_{TP})P_i(S|N)}{(1 - \delta_{TN})P_i(S|P) + \delta_{TN}P_i(S|N)} = \frac{(1 - \delta_{TN})P_i(S|P) + \delta_{TN}P_i(S|N)}{(1 - \delta_{TN})r_i + \delta_{TN}} \tag{6}
\]

As a consequence, we have that

\[
 r'_i > r'_j \iff \frac{\delta_{TP}r_i + (1 - \delta_{TP})}{(1 - \delta_{TN})r_i + \delta_{TN}} > \frac{\delta_{TP}r_j + (1 - \delta_{TP})}{(1 - \delta_{TN})r_j + \delta_{TN}} \]

\[
 \iff (\delta_{TP}r_i + (1 - \delta_{TP}))((1 - \delta_{TN})r_j + \delta_{TN}) > (\delta_{TP}r_j + (1 - \delta_{TP}))((1 - \delta_{TN})r_i + \delta_{TN}) \]

\[
 \iff (\delta_{TP}\delta_{TN} - (1 - \delta_{TP})(1 - \delta_{TN}))(r_i - r_j) > 0 \]

\[
 \iff r_i > r_j. \tag{7}
\]

The last equation follows from the assumption of the lemma.

The assumption of the lemma is equivalent to \( \delta_{TP} + \delta_{TN} > 1 \). Basically, this means that the fraction of true physical interactions in the reference network has to be larger than the fraction of true physical interactions among the interactions not in the reference network. In this case, the correct ranking of the algorithms can be confidently obtained from the reference networks despite false positives and false negatives. Of
accuracy of the Y2H system. The lowest confirmation rates within complexes is considerably higher than the overall complexes. Consequently, the accuracy of reference interactions compared to the possible number of interactions within subunits of different complexes. This can be explained by the significantly larger possible number of such interactions not confirmed in at least one other study. Confirmation rates were calculated separately for the reference interactions within complexes (gray background) or all reference interactions.

Table 1. Estimates for the fraction of true positive (δTP) and negative (δTN) interactions for the reference networks. These values were determined as the fraction of positive interactions confirmed in at least one other Y2H experiment and the fraction of negative interactions not confirmed in at least one other study. Confirmation rates were calculated separately for the reference interactions within complexes (gray background) or all reference interactions.

|                      | δTP ± δTN ± δTP + δTN within complexes | δTP + δTN all interactions |
|----------------------|----------------------------------------|-----------------------------|
| predicted Y2H interactions (Y2H pred) | 0.59 ± 0.939 ± 1.529 | 0.074 ± 0.999 ± 1.073 |
| domain-domain interactions (DD)        | 0.241 ± 0.951 ± 1.192 | 0.0095 ± 0.999 ± 1.0085 |
| Ito et al. (2001) (core)                | 0.647 ± 0.943 ± 1.59 | 0.389 ± 0.998 ± 1.387 |
| Ito et al. (2001) (non-core)            | 0.528 ± 0.939 ± 1.467 | 0.046 ± 0.999 ± 1.045 |

Supplementary Table 1. Estimates for the fraction of true positive (δTP) and negative (δTN) interactions for the reference networks. These values were determined as the fraction of positive interactions confirmed in at least one other Y2H experiment and the fraction of negative interactions not confirmed in at least one other study. Confirmation rates were calculated separately for the reference interactions within complexes (gray background) or all reference interactions.

In addition to evaluating TPR/FPR ratios, we compared the predictive accuracy of the MST, extended MST, connected protein complexes were used as benchmark set, we conclude that the relative performance and ranking of the different prediction methods for interactions within complexes can be accurately determined by evaluating TPR/FPR rates on the reference networks.
Supplementary Figure 1. ROC curve for the interactions predicted by the complete network within complexes (COMP) and the connected (CON), MST and extended MST (eMST, $\alpha = 1$) approach compared to all Y2H interactions in yeast within the BT-409 complexes. False positive rate (FPR) is plotted on the x-axis and true positive rate (TPR) on the y-axis. The curves for the complete and connected approach are almost identical at the beginning as 90% of the top scoring interactions of the complete network are also contained in the connected network. The networks differ mostly in the low scoring interactions contained additionally in the complete network.

and complete network using ROC curves (Fawcett, 2006). For this purpose, we plotted true positive rates against false positive rates with decreasing thresholds for predicting an interaction. Supplementary Figure 1 shows the ROC curve for the predictions compared against the complete set of Y2H interactions in yeast within complexes. Similar results can be observed for all reference sets. As can clearly be seen, significant improvements in predictive accuracy can be obtained with the MST approach. At a maximum true positive rate of 49.1%, only 13.6% false positives are predicted. At the same true positive rate, about 22% false positives are predicted by the connected and complete networks. The higher specificity of the MST approach results in a significantly lower sensitivity which can be increased by extending the MSTs. Although the false positive rate consequently increases as well, the overall performance of the extended MSTs is nevertheless significantly better than observed for the baseline predictions.

REFERENCES
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
Ito, T. *et al.* (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 98, 4569–74.