An Integrative Approach for Identifying Protein Complexes from Heterogeneous Sources

Min Wu, Zhipeng Xie, Xiaoli Li, Chee-Keong Kwoh and Jie Zheng

1 Importance of data sources

In the main manuscript, we mentioned that FS scores are supposed to differentiate well the positive and negative protein-protein interactions; however, they achieve a low AUC. The main reason for this contradiction is that the FS scores are calculated from DIP data while the positive interactions are from BioGrid data and thus 4,986 positive interactions cannot be assigned with FS scores. After removing these 4,986 positive interactions, FS scores can achieve an AUC 0.997 for the rest 21,862 interactions. For fair comparison, we also remove 4,986 positive interactions with lowest scores for other data sources and then calculate their AUC as shown in the third column (AUC-R) in Table S1. We observe that FS achieves the highest AUC after removing those positive interactions with lowest scores.

Table S1. The importance of data sources measured by various AUC. AUC in the second column is collected by ranking 26,828 positive and negative interactions on each individual data source. AUC-R in the third column is collected by removing 4,986 positive interactions with lowest scores. AUC-LOO in the fourth column is the AUC of SVM leaving one feature out. DIFF in the last column is the overall AUC of SVM minus the LOO AUC.

Data sources	AUC	AUC-R	AUC-LOO	DIFF
FS	0.627	0.997	0.827	0.052
GE	0.624	0.815	0.877	0.002
GO	0.795	0.942	0.839	0.04
TAP	0.677	0.887	0.872	0.007

As introduced in our manuscript, we considered the affinity scores from 4 data sources as 4 features for interactions. The AUC for the linear ranking SVM is 0.879. Given a feature, we can also compute the AUC of SVM leaving this feature out (denoted as leave-one-out AUC, AUC-LOO for short in the fourth column in Table S1) can thus be utilized to measure the importance of this feature [1]. Generally, if a feature has a high AUC-LOO, it is thus not important. In other words, given a feature, the overall AUC (i.e., 0.879) minus its AUC-LOO (denoted as DIFF in the fifth column in Table S1) can thus be utilized to demonstrate its importance. It is obvious that the importance measured by the AUC-R in the third column and DIFF in the last column are consistent, e.g., FS has the highest importance while GE has the lowest.

2 Data integration benefits

With 4 features (data sources) available, we have 15 feature combinations. Figure S1 shows the performance of InteHC using all these 15 feature combinations. It is obvious that the following combinations, FS+GO, FS+GO+TAP, FS+GE+GO and FS+GE+GO+TAP, achieves higher overall performance than individual features. This indicates that data integration indeed improve the performance for predicting protein complexes.

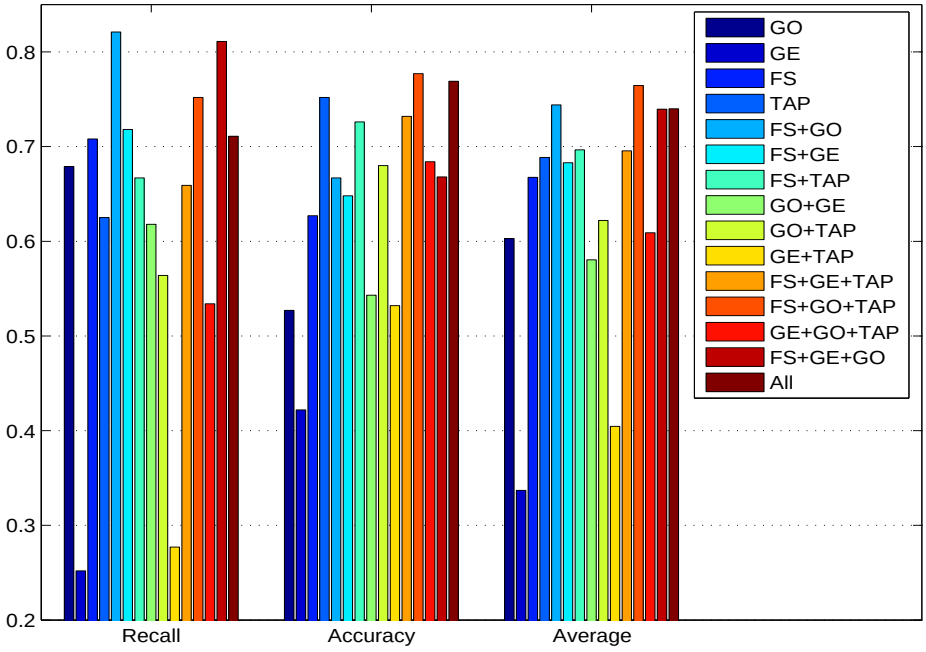

Fig. S1. The Recall and Accuracy of the complexes predicted by InteHC using various feature combinations.

We also observe that the feature combination FS+GO+TAP even achieves higher overall performance than FS+GE+GO+TAP. Although GE has the lowest importance, it is supposed to enhance the performance for protein complex prediction. Therefore, it is still highly motivated for us to investigate how to make better use of gene expression profiles for protein complex prediction in the future.

3 Case studies

Here, we will show two more example protein complexes that are more accurately detected by our InteHC. In Figure S2(A), our predicted complex with 13 proteins managed to cover all the 12 proteins in the benchmark nuclear exosome complex (GO:0000176) [3]. Meanwhile, an additional novel protein YOR076C in blue is predicted to be a member of this complex. However, it belongs to cytoplasmic exosome complex (GO:0000177). As we know, both the nuclear exosome complex and cytoplasmic exosome complex are parts of the exosome (GO:0000178). Therefore, YOR076C has similar cellular locations and molecular functions with proteins in nuclear exosome complex and it is thus clustered by InteHC into the predicted nuclear exosome complex. In Figure S2(B), the predicted complex with 10 proteins can exactly match with the TRAPP complex [2], which acts prior to SNARE complex assembly to mediate vesicle docking and fusion.

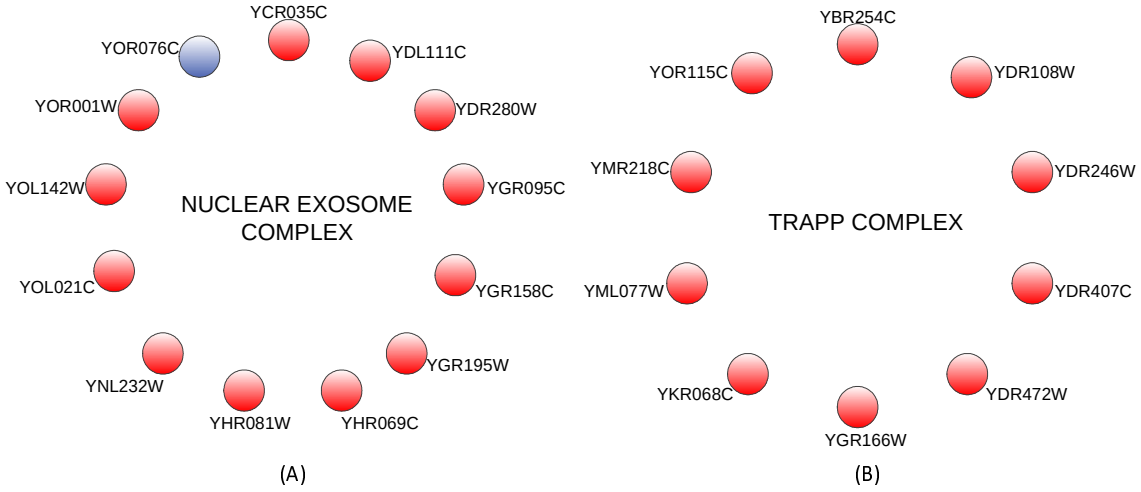

Fig. S2. Two protein complexes predicted by InteHC, (A) Nuclear exosome complex and (B) TRAPP complex.

Given a benchmark complex, its mapped complex is defined as the predicted complex that has the highest similarity (i.e., neighborhood affinity score) to this benchmark complex. Table S2 shows the mapped complexes for nuclear exosome complex and TRAPP complex predicted by various methods. As shown in Table S2, the mapped complex for nuclear exosome complex predicted by InteHC has a NA score (0.923) higher than that predicted by CACHET (0.857), C2S (0.857), InteHC-TAP (0.857), COACH (0.8), and BT (0.8). Meanwhile, the mapped complex for TRAPP complex predicted by InteHC achieves the optimal NA score 1.0 by exactly matching with the benchmark complex.

Table S2. The mapped complexes for Nuclear exosome and TRAPP complex predicted by various methods.

	Nuclear exosome (12 proteins)			TRAPP complex (10 proteins)		
Methods	NA score	Predicted size	overlap	NA score	Predicted size	overlap
InteHC	0.923	13	12	1	10	10
COACH	0.8	15	12	0.4	4	4
MCODE	0.164	73	12	1	10	10
MCL	0.167	72	12	0.667	15	10
DPClus	0.694	12	10	0.4	4	4
DECAFF	0.667	8	8	0.4	4	4
IPCA	0.675	10	9	0.5	5	5
HC-PIN	0.0309	388	12	0.833	12	10
HC-wPIN	0.632	19	12	0.833	12	10
ProRank	0.15	3	3	0.5	5	5
C2S	0.857	14	12	0.909	11	10
CACHET	0.857	14	12	0.909	11	10
BT	0.8	15	12	1	10	10
PU	0.75	16	12	1	10	10
Hart	0.6	20	12	0.833	12	10
InteHC-FS	0.75	9	9	0.6	6	6
InteHC-GO	0.25	3	3	0.2	2	2
InteHC-GE	0.042	2	1	0.05	2	1
InteHC-TAP	0.857	14	12	0.909	11	10

References

1. Y. Chang and C. Lin. Feature ranking using linear svm. *Journal of Machine Learning Research - Proceedings Track*, 3:53–64, 2008.
2. M. Sacher, Y. Jiang, J. Barrowman, A. Scarpa, J. Burston, L. Zhang, D. Schieltz, J. R. Yates, H. Abeliovich, and S. Ferro-Novick. Trapp, a highly conserved novel complex on the cis-golgi that mediates vesicle docking and fusion. *EMBO J*, 17(9):2494–2503, 1998.
3. S. A. Synowsky, M. van Wijk, R. Raijmakers, and A. J. Heck. Comparative multiplexed mass spectrometric analyses of endogenously expressed yeast nuclear and cytoplasmic exosomes. *J Mol Biol.*, 385(4):1300–13, 2009.