Protein subcomplex identification from co-purification data

Supplementary File

1 SOME PROBLEMS ASSOCIATED WITH HIGH THROUGHPUT AP DATA

A brief review of the raw data can be useful as a motivation for the data pre-processing steps and as an introduction to the main problems of the field. As an example, let us review the first purification reported by Gavin *et al.*, this is, Aac3 pulls Rpl27b, Rpl4b, Rpp0, Ssa2, Ssb1, Tef2 and Tub1 out.

False positives: Aac3 only appears in this purification and is not reported as prey of any other protein. Meanwhile, its prey proteins appear in many more purification experiments: Rpl27b (103 times), Rpl4b (216), Rpp0 (254), Ssa2 (1025), Ssb1 (779), Tef2 (399) and Tub1 (256). The heat shock protein Ssa2 is involved in protein folding and vacuolar import of proteins, while the heat shock protein Ssb1 is reported to aid in the passage of the polypeptide chains through the ribosome channel into the cytosol. Aac3, on the other hand, is an ADP/ATP translocase. This undermines the credibility of these interactions. In fact, after the pre-processing pipeline, Ssa2 will appear in no complex generated by SA or PE scores together with link communities, while appears in four overlapping complexes generated using the Dice score.

Coverage: From these eight proteins, only Aac3 and Tef2 were used as baits in this work. The other six proteins only appear in the data as preys of different experiments.

Mutual pull-out: When Tef2 is co-purified, the pulled out proteins are: Cam1, Ded1, Efb1, Eft2, Kar2, Pab1, Psa1, Rpn1, Rps24b, Rps5, Sbp1, Ssa2, Ssb1, Sse1, Tef1, Tef2, Tef4, Ura2 and Vma2. This means that no Aac3 appears, and only two of the ones that copurified before: Ssa2 and Ssb1. This raises doubts regarding the Aac3-Tef2 interaction, but the case is not better for Tef2-Ssa2 and Tef2-Ssb1. Tef2 and Ssa2 co-occured as preys in 241 purifications, while their combined number of appearances as preys is 1183 purifications, this is, a Jaccard index of 0.20. At the same time, Tef2 and Ssb1 co-occur in 216 out of 962 purifications, i.e., a Jaccard index of 0.22.

Generated complexes and bad scores: Due to the above-mentioned problems, after the scoring step, there will be no reliable interactions containing Aac3, with any of the scoring systems used. Therefore, after clustering, there will be no complexes containing Aac3, even though it was a bait pulling seven proteins out.

Generated complexes and good-enough scores: The situation is different for Tef2. Using the SA score and linkcomm, Tef2 will appear in three overlapping complexes: "CAM1, EFB1, TEF2, TEF4", "CAM1, EFB1, TEF2, SBP1", and "CAM1, EFB1, TEF2". Using the Dice score and link communities, will appear in five overlapping complexes, including "URA2, PSA1, SSA1, SSB1, TEF1, TEF2, VMA2" and four more. Using PE and link communities, will appear in no complexes. The following examples will help us understand why some interactions are good according to some methods. First, SA considers Cam1 and Efb1 are reliable interactors. In fact, Tef2 pulls Cam1 out, and Cam1 pulls Tef2 out. Even though, they only copurify as preys in 2 experiments out of 399 combined experiments, this is, a Jaccard index of 0.005. The Dice scored results consider Ura2 and Tub1 as reliable. In this case, Ura2 is not even used as a bait and, therefore, there is no chance to observe if they pull Tef2 out. However, Tef2 and Ura2 copurify as preys in 175 out of 660 experiments, this is, a Jaccard index of 0.26. Something similar occurs with Dice scores and Tub1: While Tub1 is not used as a bait, it appears with Tef2 in 112 out of 543 experiments, this is, a Jaccard index of 0.206.

Generated complexes and analysis of best scores: Using SA score, the best scored purification is Leu4-Leu9. In fact, Leu4 is a bait and only detects Leu9, while Leu9 is a bait, and only detects Leu4. Besides that, they are both pulled out by one additional protein: Rsa3. Therefore, they participate in three purifications and co-purify in all of them, to give a Jaccard index of 1. Using PE score, the best purification would be Pap1-Pfs2. Pap1 is a bait, used three times, and all 3 times finds Pfs2, while Pfs2 is a bait, used 2 times, and both times finds Pap1. As preys, they co-purify in 22 out of 28 possible experiments, this is, a Jaccard index of 0.786. Finally, using Dice score, the best purification is Pep3-Pep5. Pep3 is a bait and it finds Pep5, while Pep5 is a bait and finds Pep3. As preys, they co-purify in 7 out of 7 possible experiments, this is, a Jaccard index of 1.

The effect of clustering: Finally, some of the proteins in a highly-scored interaction not necessarily will end up in the same complex, as the clustering criterion (density of neighbors, f.ex.) may end up changing this. For example, using SA, Tef2 has reliable interactions with Rps24b, Tef4, Efb1, Hir1, Sbp1 and Cam1. From this group, Rps24b and Hir1 will not end up in any of the three predicted complexes. Using PE, Tef2 has reliable interactions with Rps24b, Rps5, Ilv1, Vps1, Mdn1, Kip1, Gcn20, Kar2, Tef4, Efb1, Ypt7, Acc1, Rfs1, Sbp1, Cam1, Hir2, YER156C, Mtq2, Ubp1, Tvp38 and Pom152. However, none of them will be a part of a complex, even though some of them (Cam1 and Efb1) do it with SA. Finally, using the Dice score, Tef2 will have Tub1, Rpl27b, Rrp5, Rpl3, Rpl20b, Gcn1, Imd3, Tub2, Cdc19 and Gfa1. Some proteins, such as Rrp5, Rpl3 and Rpl20b, will not make it to a complex here. However, some proteins that do not have reliable interactions with Tef2 here, are brought to the complex by the clustering algorithm, including: Kap123, Ura2, Psa1, Ssb1, Tef1, Vma2, Ssa2, Rps13, Rps4a, Rpl4a, Rpl7a, Rpp0, Rpl20a, Rpl30, Sam1, Pab1, Rps3, Rps4b, Rpl10, YHR020W, Adh1, Rpl7b, Rpl4b and Rps0a. Note that some of these added proteins come from the Aac3 purification, which was rejected by scoring and brought back by clustering.

The Leu4-Leu9 interaction deserves an additional consideration: While PE leads to no complexes containing Leu4, SA leads to three complexes: "ARG1, LEU4, LEU9, NOP8, RSA3, URB2", "ARG1, LEU4, LEU9, NOP8, URB2, URB1, PRC1" and "ARG1, LEU4, LEU9, NOP8, URB2, PRC1, DBP6", and Dice leads to 2 complexes: "ARG1, LEU4, LEU9, NOP8, RSA3, URB2" and "ARG1, LEU4, LEU9, NOP8, URB2, DBP6, PRC1". A Leu4-Leu9 complex does not appear in the results with any method and, in fact, Leu4 and Leu9 are the subunits of the alphaisopropylmalate synthase. This shows how a highly scored copurification, which happens to be a complex, can get filtered out by clustering.

2 EVALUATION OF ALL SCORING+CLUSTERING METHODS USING THE HYPER-GEOMETRIC INDEX

The performance of the TRIBAL method was assessed using the same procedures above explained. The method was applied to 11 different cutoff values of the reliability score, ranging from 0.0001 to 0.2, in order to select the best cutoff value. This range was selected due to the size of the resulting PIN: Below 0.001, the size of the PINs do not increase, while, after 0.2, PINs contain a small amount of edges (less than 1000). Supplementary Table 1 shows the TRIBAL results using the link communities complexes as a template, while Supplementary Table 2 shows the sensitivity, PPV and accuracy values.

Supplementary Table 1. Precision and recall analysis using link communities as a template.

Cutoff value	#Comp2 Ref	Tot_Pred_ Comp	%Comp2 Ref	#Ref2 Comp	Tot_Ref_ Comp	%Ref2 Comp
0.0001	14	14	1.0	10	214	0.05
0.001	14	14	1.0	10	214	0.05
0.0125	14	14	1.0	10	214	0.05
0.025	14	14	1.0	10	214	0.05
0.05	14	14	1.0	10	214	0.05
0.075	14	14	1.0	10	214	0.05
0.1	14	14	1.0	10	214	0.05
0.125	13	13	1.0	10	214	0.05
0.15	10	10	1.0	9	214	0.04
0.175	9	9	1.0	9	214	0.04
0.2	8	8	1.0	8	214	0.04

Supplementary Table 2. Accuracy analysis using link communities as a template.

Cutoff value	Sensitivity	PPV	Accuracy
0.0001	0.79	0.39	0.55
0.001	0.79	0.39	0.55
0.0125	0.79	0.39	0.55
0.025	0.79	0.39	0.55
0.05	0.79	0.39	0.55
0.075	0.79	0.39	0.55
0.1	0.79	0.39	0.55
0.125	0.76	0.40	0.55
0.15	0.73	0.41	0.54
0.175	0.74	0.44	0.57
0.2	0.61	0.55	0.58

Therefore, the effect of choosing a cutoff of 0.1 would be that TRIBAL produces 13 nested pairs and 5 nested groups, with 100% subcomplexes mapped to MIPS, 4.7% MIPS mapped to TRIBAL predictions, 79.3% sensitivity and 55.5% accuracy.

The results using OCG as a template can be observed in Supplementary Table 3 and 4.

Supplementary Table 3. Precision and recall analysis using OCG as template.

Cutoff value	#Comp2 Ref	Tot_Pred_ Comp	%Comp2 Ref	#Ref2 Comp	Tot_Ref_ Comp	%Ref2 Comp
0.0001	20	22	0.91	16	214	0.07
0.001	20	22	0.91	16	214	0.07
0.0125	20	22	0.91	16	214	0.07
0.025	20	22	0.91	16	214	0.07
0.05	20	22	0.91	16	214	0.07
0.075	19	21	0.90	16	214	0.07
0.1	19	21	0.90	16	214	0.07
0.125	18	20	0.90	16	214	0.07
0.15	17	19	0.89	16	214	0.07
0.175	16	16	1.00	14	214	0.06
0.2	13	13	1.00	12	214	0.06

Supplementary Table 4. Accuracy analysis using OCG as template.

Cutoff value	Sensitivity	PPV	Accuracy
0.0001	0.70	0.23	0.40
0.001	0.70	0.23	0.40
0.0125	0.70	0.23	0.40
0.025	0.70	0.23	0.40
0.05	0.71	0.23	0.41
0.075	0.72	0.24	0.41
0.1	0.72	0.24	0.41
0.125	0.70	0.25	0.41
0.15	0.70	0.26	0.42
0.175	0.67	0.30	0.45
0.2	0.61	0.36	0.47

Therefore, the effect of choosing a cutoff of 0.1 would be that TRIBAL produces 29 nested pairs and 8 nested groups, with 90% of subcomplexes mapped to MIPS, 7.5% MIPS mapped to TRIBAL predictions, 72% sensitivity and 41.3% accuracy.

Supplementary Table 5 compares the results of TRIBAL to the existing methods. TRIBAL and PE-OCG give the best precision while Dice-lcomm and PE-lcomm give the best recall.

Supplementary Table 5. Summary of precision and recall analyses.

Methods	#Comp2 Ref	Tot_Pred_ Comp	%Comp2 Ref	#Ref2 Comp	Tot_Ref_ Comp	%Ref2 Comp
Dice-lcomm	198	332	0.60	92	214	0.43
PE-lcomm	68	88	0.77	41	214	0.19
Dice-OCG	37	45	0.82	19	214	0.09
PE-OCG	52	52	1.00	25	214	0.12
TRIBAL –lcomm-PE	14	14	1.00	10	214	0.05
TRIBAL –OCG-PE	19	21	0.90	16	214	0.07

TRIBAL and PE-OCG give the best precision while Dice-lcomm and PE-lcomm give the best recall.