Supplementary Material: A novel graph clustering method with greedily heuristic search algorithm for mining protein complexes from dynamic and static PPI networks

Rongquan Wang[a,b], Caixia Wang[c], Guixia Liu[a,b,*]

[a] *College of Computer Science and Technology, Jilin University, Changchun, 130012, Jilin, China*
[b] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, Jilin, China*
[c] *School of International Economics, China Foreign Affairs University, Beijing, 100037, Beijing, China*

1. Biological relevance metrics

Here, we use a software called ProCope to calculate Co-localization score and Gosim score for the purpose of estimating discovered protein complexes. ProCope [1] is available from the website: https://www.bio.ifi.lmu.de/software/procope/index.html. Functional enrichment analysis is calculated by using data statistics method, i.e, p-value.

1.1 Co-localization score

The first biological metric is Co-localization score. Because proteins in the same protein complex have same or similar common functions [2] and they tend to have the same localization [3]. Generally, the higher Co-localization score is, the more functional similarity among proteins in the same protein complex is. For more comprehensive assessing all the discovered protein complexes, the average of Co-localization score is computed as follows:

$$Co-localizationscore = \frac{\sum_{j=1}^{m} \max_{i=1}^{n} l_{i,j}}{\sum_{j=1}^{m} N_j} \quad (1)$$

*Corresponding author
Email addresses: wangrongquanjlu@163.com (Rongquan Wang), wangcaixia@cfau.edu.cn (Caixia Wang), liugx@jlu.edu.cn (Guixia Liu)

In this Eq.(30), $l_{i,j}$ is the number of proteins in the discovered complex j allocated to the localization group i; N_j is the number of proteins in the discovered complex j; m and n are the number of discovered complexes and localization groups, respectively.

1.2 GO semantic similarity

Due to a protein complex generally performs a specific function, proteins in the same protein complex tend to have same functionality. According to GO annotations, in 2006, Schlicker et al. [4] introduce a novel algorithm for calculating the semantic similarity. For a discovered protein complex, its GO semantic similarity score is the average of semantic similarity score of all interacting proteins within the discovered protein complex. Here, for all discovered protein complexes, and their GO semantic similarity score is denoted as the average of semantic similarity score of three ontologies including cellular Component, biological process and molecular function ontology, respectively. Therefore, the higher GO semantic similarity score of discovered protein complexes is, the better the performance of discovered protein complexes is [1].

1.3 Functional enrichment analysis

The GO term enrichment is calculated by p-value based on the hypergeometric distribution to evaluate the statistical significance of the discovered protein complexes. Generally, a discovered protein complex with a higher p-value is more significant biologically. In additional, we also use the function enrichment test to measure the biological significances of discovered protein complexes by different algorithms. In this paper, we use LAGO [5] to accomplish the function enrichment test with different threshold. Note that, LAGO is a fast tool which finds significant GO terms among a list of gene names or proteins, and it computes the significance (p-value) via the hypergeometric distribution, and applies (by default) Bonferroni correction. For the details of calculating p-value, please refer to literature [5]. The p-value is denoted as follows:

$$p - value = 1 - \sum_{i=0}^{k-1} \frac{\binom{F}{i}\binom{N-F}{C-i}}{\binom{N}{C}}, \tag{2}$$

where k is the number of proteins of functional group in the protein complex, and N is the number of proteins in the PPI network. F is the size of the functional group in the PPI network, we assume that a discovered protein complex which contains C proteins. Generally, the lower the p-value is, the stronger biological significance the protein complex has. The discovered protein complex with less than 0.01 is considered to be meaningful.

2. Comparison with other methods based on several biological metrics

2.1 Comparison based on co-localization score

Based on 1.1 section, we calculate the co-localization score of discovered protein complexes by different methods from all tested PPI networks. For measuring the consistency of different discovered complexes localization, we use two yeast localization datasets consist of Kumar et al. [6] and Huh et al. [7] to calculate the colocalization scores of protein complex discovered by different algorithms. The comparison results of the Co-localization score for MPC-C with other compared methods is listed in Figure 1.

From Figure 1, on Kumar localization dataset, ClusterONE achieves the highest colocalization scores, which is better than all the other methods. On Huh localization dataset, MPC-C obtains the best colocalization score in all datasets except for String. For String dataset, IPCA is 0.6752 and the best in terms of Huh localization dataset and MPC-C is 0.6495 and the second highest. In general, the high Co-localization scores show that the protein complexes discovered by identification methods have better localization consistency. Therefore, we can see that a good proportion of proteins in the protein complexes discovered by MPC-C are locating at the same location and are therefore similar in function. Furthermore, in order to synthetically observe the GO semantic similarity scores of comparative results. In Figure 1, the y-axis represents different algorithms and the x-axis is the sum of Co-localization scores. We can see MPC-C achieves the highest composition score. Therefore, MPC-C is better

performance than the other methods for two localization datasets in terms of Co-localization score.

2.2 Comparison based on GO semantic similarity

As is known to us proteins belong to the same protein complex generally tend to achieve a similar function, thus those GO semantic similarity scores is higher than proteins get together by chance. Therefore, in order to measure that the possibility of discovered protein complex sets are the real protein complexes, we calculate separately the average of GO semantic similarity score of all discovered protein complexes for the BP, CC and MF ontologies to evaluate each the functional similarity of a group of proteins in a discovered complex based on GO terms which is used for annotating these proteins. All experimental results are listed in Table 1. Note that we use the arithmetic mean similarity of the three ontology scores as the final GO semantic similarity score of discovered protein complexes to evaluate the performance of different algorithms in the last column Table 1. According to Table 1, we can see that the GO semantic similarity score by MPC-C is highest in most tested PPI networks, which shows obviously advantage than other methods. Overall, MPC-C has the better functional similarity over the other algorithms.

Meanwhile, we also use Figure 2 to synthetically observe the GO semantic similarity scores of comparative results by all algorithms. As shown in Figure 2, MPC-C obtains the highest composition score. Therefore, our method outperforms other compared algorithms for all five datasets with respect to GO semantic similarity.

2.3 Functional enrichment analysis of discovered protein complexes

Finally, we perform GO enrichment analysis for all protein complexes discovered by different algorithms and list the number and percentage of the discovered protein complexes that are significant enrichment with the lowest p-value of GO term including biological process, molecular function and cellular component domains in Tables 2 and 3. Due to discovered protein complexes with

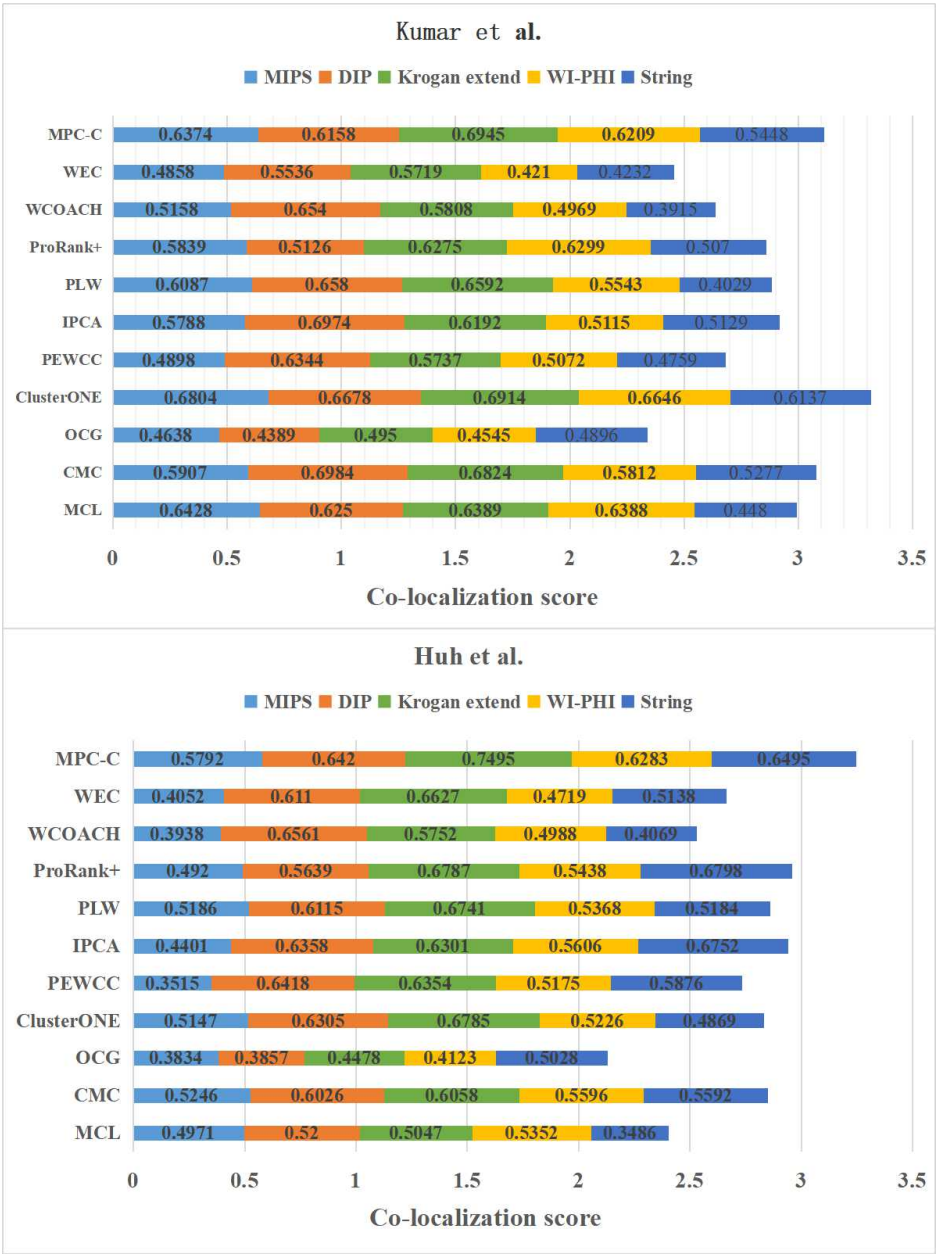

Figure 1: Comparison of all competing algorithms in terms of the co-localization scores. Shades of the different color indicate different PPI networks. The total height of each bar is the sum of Co-localization scores of five datasets for a tested algorithm on (a) Kumar et al. and (b) Huh et al. localization datasets. Larger Co-localization scores are better.

Table 1: GO semantic similarity scores by different algorithms based on BP, CC and MF ontologies in different PPI networks.

	MCL	CMC	OCG	ClusterONE	IPCA	PLW	ProRank+	WCOACH	WEC	MPC-C
MIPS										
BP	0.5868	0.7821	0.5615	0.6874	0.7327	0.7632	0.6707	0.6960	0.6763	**0.8959**
CC	0.7616	0.8523	0.7506	0.7972	0.8343	0.8419	0.8147	0.8065	0.7859	**0.9227**
MF	0.5467	0.6433	0.4812	0.6233	0.5656	0.6459	0.5973	0.5330	0.5332	**0.7369**
Average	0.6317	0.7592	0.5978	0.7026	0.7109	0.7503	0.6942	0.6785	0.6651	**0.8518**
DIP										
BP	0.6058	0.7426	0.569	0.8071	0.7973	0.8319	0.7228	0.7862	**0.8877**	0.8721
CC	0.7843	0.8816	0.7681	0.8759	0.9479	0.9036	0.8504	0.9189	0.9531	**0.9564**
MF	0.5695	0.6385	0.4948	0.7227	0.6538	0.7166	0.6444	0.6556	0.7302	**0.7448**
Average	0.6532	0.7542	0.6106	0.8019	0.7997	0.8174	0.7392	0.7869	**0.8570**	0.8577
Krogan extend										
BP	0.5757	0.7343	0.6036	0.8002	0.8297	0.8474	0.8195	0.7571	0.8705	**0.9343**
CC	0.7646	0.8761	0.7896	0.8805	0.9123	0.918	0.9007	0.8874	0.9454	**0.9690**
MF	0.555	0.6353	0.5346	0.7646	0.7504	0.7941	0.7761	0.6331	0.8226	**0.8503**
Average	0.6318	0.7486	0.6426	0.8151	0.8308	0.8532	0.8321	0.7592	0.8795	**0.9178**
WI-PHI										
BP	0.5876	0.6929	0.5953	0.5665	0.7786	0.724	0.5873	0.7162	0.7245	**0.9000**
CC	0.7550	0.8379	0.78	0.7425	0.8839	0.8487	0.8075	0.8686	0.8536	**0.9353**
MF	0.5893	0.6106	0.5238	0.5613	0.6781	0.6499	0.5967	0.6253	0.6308	**0.7512**
Average	0.6440	0.7138	0.6330	0.6234	0.7802	0.7409	0.6638	0.7367	0.7363	**0.8621**
String										
BP	0.4841	0.6986	0.6991	0.6120	0.8385	0.7247	0.8760	0.6624	0.7607	**0.8941**
CC	0.6950	0.8578	0.8453	0.7794	0.9235	0.8704	0.9473	0.8580	0.8848	**0.9528**
MF	0.5188	0.604	0.6394	0.5895	0.8163	0.6312	**0.8490**	0.6019	0.6870	0.7696
Average	0.5660	0.7201	0.7279	0.6603	0.8594	0.7421	0.8908	0.7074	0.7775	**0.8721**

biological significance have lower p-values, the higher values indicate these proteins in discovered protein complexes don't random combination together and they probably consist of a real protein complex. Furthermore, some of them may be the new protein complexes that have not been discovered. As shown in Tables 2 and 3, for all five yeast PPI networks, we find that MPC-C achieves the best in terms of the number of discovered protein complexes that are significant in DIP and WI-PHI PPI networks. IPCA obtains the best numbers for MIPS, Krogan extend and String PPI networks but with the smaller percentage of discovered protein complexes that are functionally significant than MPC-C. Meanwhile, regarding the performance in terms of the percentages of discovered

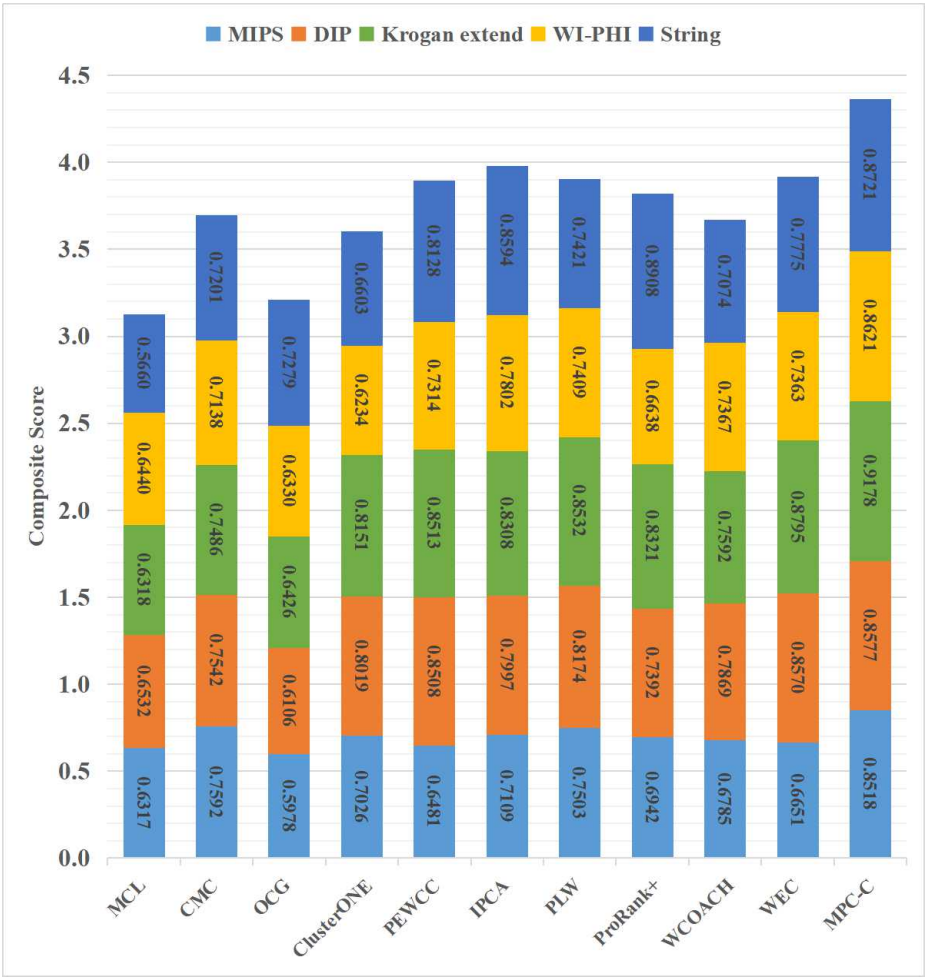

Figure 2: Results comparison of the ten algorithms in five datasets with respect to the GO semantic similarity scores. The y-axis represents the Composite Score of GO semantic similarity score different algorithms in all tested datasets while the x-axis corresponds to different algorithms. Various colors of the same column denote the Go semantic similarity score of different PPI networks. The total height of each column is the composite score of each algorithm in all tested datasets. Large scores show clustering results is better.

protein complexes that are functionally significant, WCOACH is best score, and MPC-C is slightly lower than the best WCOACH. There are two reasons to explain this results. One is the fact that MPC-C discovers many more protein

Table 2: Function enrichment analysis of the protein complexes discovered by MPC-C and other algorithms on different datasets.

Algorithms	PC	<E-15	<E-10	<E-5	significant
MIPS					
MCL	594	16(2.69%)	55(9.26%)	236(39.73%)	298(50.17%)
CMC	408	34(8.33%)	93(22.79%)	260(63.72%)	308(75.48%)
OCG	520	29(5.58%)	94(18.08%)	305(58.66%)	359(69.04%)
ClusterONE	690	35(57%)	143(20.72%)	407(58.98%)	496(71.88%)
IPCA	907	148(16.32%)	416(45.87%)	755(83.25%)	802(88.43%)
PLW	341	39(11.44%)	141(41.35%)	275(80.65%)	303(88.86%)
ProRank+	233	38(16.31%)	74(31.76%)	168(72.1%)	197(84.55%)
WCOACH	580	453(78.1%)	538(92.76%)	580(100.0%)	580(100.0%)
WEC	223	64(28.7%)	113(50.67%)	183(826%)	200(89.68%)
MPC-C	790	100(12.66%)	323(40.89%)	737(93.3%)	766(96.97%)
DIP					
MCL	628	35(5.57%)	78(12.42%)	246(39.17%)	317(50.48%)
CMC	1192	65(5.45%)	154(12.92%)	511(42.87%)	689(57.8%)
OCG	393	60(15.27%)	119(30.28%)	290(73.79%)	319(81.17%)
ClusterONE	341	41(12.02%)	85(24.92%)	217(63.63%)	238(69.79%)
IPCA	1242	348(28.02%)	620(49.92%)	1011(81.4%)	1088(87.6%)
PLW	577	93(16.12%)	204(35.36%)	439(76.09%)	481(83.37%)
ProRank+	167	27(16.17%)	61(36.53%)	144(86.23%)	150(89.82%)
WCOACH	967	800(82.73%)	916(94.73%)	966(99.9%)	967(100.0%)
WEC	253	153(60.47%)	182(71.93%)	232(91.69%)	239(94.46%)
MPC-C	1382	431(31.19%)	808(58.47%)	1319(95.45%)	1360(98.42%)
Krogan extend					
MCL	515	28(5.44%)	58(11.27%)	179(34.77%)	217(42.15%)
CMC	650	43(6.62%)	73(11.24%)	205(31.55%)	286(44.01%)
OCG	432	76(17.59%)	134(31.02%)	307(71.07%)	348(80.56%)
ClusterONE	240	38(15.83%)	74(30.83%)	154(64.16%)	171(71.24%)
IPCA	990	278(28.08%)	387(39.09%)	745(75.25%)	825(83.33%)
PLW	422	136(32.23%)	228(54.03%)	331(78.44%)	356(84.36%)
ProRank+	184	48(26.09%)	79(42.94%)	149(80.98%)	164(89.13%)
WCOACH	788	609(77.28%)	728(92.38%)	785(99.61%)	785(99.61%)
WEC	362	222(61.33%)	272(75.14%)	330(91.16%)	339(93.65%)
MPC-C	751	296(39.41%)	463(61.65%)	719(95.74%)	737(98.14%)

complexes than WCOACH in all five tested PPI networks. The other reason is that WCOACH predicts protein complex is vary large size, but standard protein complexes and discovered protein complexes by MPC-C is smaller size. For example, based on MIPS PPI network, WCOACH identifies the average of

Table 3: Function enrichment analysis of the protein complexes discovered by different algorithms on different PPI networks. (Continued)

Algorithms	PC	<E-15	<E-10	<E-5	significant
WI-PHI					
MCL	772	42(5.44%)	83(10.75%)	242(31.35%)	297(38.47%)
CMC	1971	213(10.81%)	353(17.91%)	872(44.24%)	1069(54.23%)
OCG	412	163(39.56%)	237(57.52%)	357(86.65%)	371(90.05%)
ClusterONE	1313	51(3.88%)	109(8.3%)	304(23.15%)	412(31.38%)
IPCA	2182	594(27.22%)	824(37.76%)	1292(59.21%)	1371(62.83%)
PLW	692	312(45.09%)	406(58.67%)	520(75.14%)	540(78.03%)
ProRank+	255	63(24.71%)	89(34.91%)	157(61.58%)	170(66.68%)
WCOACH	2146	1851(86.25%)	2022(94.22%)	2124(98.97%)	2128(99.16%)
WEC	729	351(48.15%)	396(54.32%)	510(69.96%)	522(71.61%)
MPC-C	2569	934(36.36%)	1566(60.96%)	2292(89.22%)	2361(91.91%)
String					
MCL	120	5(4.17%)	15(12.5%)	41(34.17%)	59(49.17%)
CMC	3045	177(5.81%)	276(9.06%)	987(32.41%)	1477(48.5%)
OCG	624	296(47.44%)	383(61.38%)	529(84.78%)	556(89.11%)
ClusterONE	848	95(11.2%)	184(21.7%)	515(60.73%)	594(70.05%)
IPCA	4661	4041(86.7%)	4248(91.14%)	4497(96.48%)	4557(97.77%)
PLW	872	793(90.94%)	824(94.5%)	857(98.28%)	861(98.74%)
ProRank+	780	663(85.0%)	711(91.15%)	764(97.94%)	772(98.97%)
WCOACH	2647	2563(96.83%)	2611(98.64%)	2643(99.85%)	2644(99.89%)
WEC	1108	942(85.02%)	975(88.0%)	1036(93.51%)	1052(94.95%)
MPC-C	2720	1601(58.86%)	2105(77.39%)	2659(97.76%)	2697(99.16%)

NOTE: The table lists the number and percentage of protein complexes detected by different algorithms in the different PPI networks. "PC" denotes the number of discovered protein complexes. As for "A(B%)", where A presents the number of discovered protein complexes whose p-value fall within different value ranges. B defines the percentage of discovered protein complexes whose p-value fall within different value ranges.

size of discovered protein complexes is 48.30 and the average of size of reference complexes 2 is 8.97 and discovered protein complexes by MPC-C is 6.24, respectively. In summary, considering the number and percentage of the discovered protein complexes that are significant in analyzing the functional enrichment of discovered protein complexes. Therefore, MPC-C could discover significantly many more protein complexes than other competing algorithms by functional enrichment test based on calculating p-value.

References

[1] J. Krumsiek, C. C. Friedel, R. Zimmer, Procopeprotein complex prediction and evaluation, Bioinformatics 24 (18) (2008) 2115–2116.

[2] M. Wu, X. Li, C.-K. Kwoh, S.-K. Ng, A core-attachment based method to detect protein complexes in ppi networks, BMC bioinformatics 10 (1) (2009) 169.

[3] S. Wang, F. Wu, Detecting overlapping protein complexes in ppi networks based on robustness, Proteome science 11 (1) (2013) S18.

[4] A. Schlicker, F. S. Domingues, J. Rahnenführer, T. Lengauer, A new measure for functional similarity of gene products based on gene ontology, BMC bioinformatics 7 (1) (2006) 302.

[5] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, G. Sherlock, Go:: Termfinderopen source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes, Bioinformatics 20 (18) (2004) 3710–3715.

[6] A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, et al., Subcellular localization of the yeast proteome, Genes & development 16 (6) (2002) 707–719.

[7] W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, E. K. O'shea, Global analysis of protein localization in budding yeast, Nature 425 (6959) (2003) 686.