Supplementary document for “Protein Complexes Identification with FWER Control”

Zengyou He*(Corresponding Author)
School of Software,
Dalian University of Technology, Dalian, China.
Key Laboratory for Ubiquitous Network and Service Software of
Liaoning Province, Dalian, China.
*Email: zyhe@dlut.edu.cn

Can Zhao
School of Software,
Dalian University of Technology, Dalian, China.
Email: can.zhao1114@hotmail.com

Hao Liang
School of Software,
Dalian University of Technology, Dalian, China.
Email: 18642883268@163.com

Quan Zou*(Corresponding Author)
School of Computer Science and Technology,
Tianjin University, Tianjin, China.
*Email: zouquan@nclab.net

1 The calculation of p-value when the vertex belongs to the subgraph

Let $G = (V, E)$ be an undirected graph with a set of vertices V and a set of edges E. For a given subgraph S, if one vertex $i \in S$ has k_i^{in} neighbors in subgraph $S \setminus \{i\}$ and has k_i^{out} neighbors in $G \setminus \{S\}$, then $k_i = k_i^{in} + k_i^{out}$ is the degree of vertex i. In addition, D_S is used to denote the degree of subgraph S, $D_{\hat{S}}$ is used to represent the total degree of the rest of vertices in $G \setminus \{S\}$, and D is the total degree of all the vertices in G.

When the vertex i is included in subgraph S, we consider the subgraph that excludes i, $S \setminus \{i\}$, as the current subgraph. Hence the vertices in G can be divided into two groups: $S \setminus \{i\}$ and $G \setminus \{S\}$, and we introduce two binary variables $C(u, S)$ and $B(i, u)$ whose definitions are the same as that in the main manuscript. That is, $C(u, S) = 1$ if vertex u is included in the subgraph $S \setminus \{i\}$ and $C(u, S) = 0$ if $u \in G \setminus \{S\}$, and $B(i, u) = 1$ if vertex i has an edge with vertex u and $B(i, u) = 0$ otherwise. Therefore, when the vertex i belongs to the given subgraph S, we can construct the following contingency table as shown in Table 1.

Table 1: The contingency table for a vertex (protein) i when it is included in the given subgraph S.

	$B(i,u) = 1$	$B(i,u) = 0$	Row totals
$C(u,S) = 1$	k_i^{in}	$D_S - k_i - k_i^{in}$	$D_S - k_i$
$C(u,S) = 0$	k_i^{out}	$D_{\hat{S}} + k_i - k_i^{out}$	$D_{\hat{S}} + k_i$
Col totals	k_i	$D - 2k_i$	$D - k_i$

2 Family-Wise Error Rate (FWER) and False Discovery Rate (F-DR)

FWER and FDR are widely used for measuring the rate of type I errors in multiple hypothesis testing. FWER is the probability of making one or more type I error when performing multiple hypotheses tests. FDR is defined as the expected proportion of false "discoveries". Suppose that there are m null hypotheses, denoted by $H_1, H_2, \cdots, H_m$, and $p_1, p_2, \cdots, p_m$ represent their corresponding p-values. We sort these p-values in ascending order, which are denoted by $p_{(1)}, p_{(2)}, \cdots, p_{(m)}$. In each hypothesis test, we will either accept the alternative hypothesis or retain the null hypothesis. Summing up the outcomes from m hypothesis tests will yield the following information in Table 2. In this table, m_0 is the number of true null hypotheses, R is the number of rejected hypotheses, V is the number of Type I errors (false positives) and T is the number of Type II errors (false negatives).

Table 2: The contingency table for m hypothesis tests.

	# true null hypotheses	# false null hypotheses	Total
# Significant	V	$R - V$	R
# Non-significant	$m_0 - V$	T	$m - R$
Total	m_0	$m - m_0$	m

The definition of FWER is given in the following formula:

$$\text{FWER} = \Pr(V \geq 1). \tag{1}$$

Thus, by making FWER $\leq \alpha$, the probability of making at least one type I error in m hypothesis tests is controlled at the significance level α. The Bonferroni procedure is a popular strategy to control the FWER, in which we will reject a null hypothesis H_i if $p_i \leq \alpha/m$.

FDR is defined as follows:

$$\text{FDR} = \text{E}[\frac{V}{R}|R > 0] \cdot \Pr(R > 0). \tag{2}$$

The Benjamini-Hochberg procedure (BH step-up procedure) [1] is widely used for controlling the FDR at the significance level α. It works as follows: (1) find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$; (2) reject each $H_{(i)}$, where $i = 1, \cdots, k$.

3 Supplementary experimental results

In the main manuscript, we adopt the overlap score to judge whether a set A is matched to a set B. The overlap score between two complexes A and B is defined as follows [2]:

$$w(A, B) = \frac{|A \cap B|^2}{|A||B|}. \tag{3}$$

3.1 The performance comparison of the complex size distribution

We compare the protein complexes predicted by each method based on the following summary statistics: the size of protein complexes (minimal size, maximal size, average size), the overlap among identified complexes (the average degree of vertices in a protein complex and the average number of complexes that non-background vertices belong to), the number of background vertices found. More precisely, the detailed results are presented in Table 3, where $|C|_{min}$, $|C|_{max}$ and $|\bar{C}|$ denote the minimal size, maximal size and average size of predicted complexes, respectively. In addition, $|\bar{D}|_{com}$ denotes the average degree of the vertices in a protein complex, and $|\bar{D}|_{bkg}$ is the average degree of the background vertices. $\bar{n_C}$ is the average number of complexes to which non-background vertices belong, and P_{bkg} is the proportion of background vertices.

We could observe that the average degree of the vertices in a protein complex $|\bar{D}|_{com}$ is higher than the average degree of the background vertices $|\bar{D}|_{bkg}$, which means that the true protein complex is more dense than the background vertices. The P_{bkg} value of SSF is larger than that of other methods, which indicates that SSF could filter more background vertices than other three methods.

Table 3: The comparison of complex size distribution of different methods.

	Algorithm	Predicted	$\|C\|_{min}$	$\|C\|_{max}$	$\|\bar{C}\|$	$\|\bar{D}\|_{com}$	$\|\bar{D}\|_{bkg}$	$\bar{n_C}$	P_{bkg}
Collins	SSF	127	3	113	9.9448	7.3046	1.7600	1.1684	0.3335
	MCL	158	3	158	8.5063	5.2027	1.1727	1	0.1714
	ClusterOne	203	3	103	7.4400	5.7200	4.2900	1.1686	0.2028
	MDS	333	3	102	24.3934	17.0378	1	5.8355	0.1418
Gavin	SSF	167	3	63	8.3053	7.2888	2.9300	1.1195	0.3321
	MCL	220	3	75	7.5682	5.4688	3.6053	1	0.1024
	ClusterOne	294	3	40	6.9300	5.5300	5.9300	1.2555	0.1245
	MDS	586	3	50	11.3703	10.0080	1	3.6833	0.0248
KroganC	SSF	91	3	38	7.1868	7.0966	3.7800	1.0365	0.7670
	MCL	374	3	46	5.8930	4.3231	3.5020	1	0.1861
	ClusterOne	242	3	23	5.2400	4.9000	4.5800	1.1806	0.6034
	MDS	1671	3	27	6.0934	17.1273	1	3.8833	0.0318
KroganE	SSF	77	3	40	8.7700	11.5918	6.2300	1.0696	0.8279
	MCL	515	3	48	6.0816	6.7434	8.7074	1	0.1471
	ClusterOne	239	3	29	5.5700	6.8990	7.0400	1.1948	0.6966
	MDS	2776	3	31	7.3246	37.6049	1	5.5646	0.0049
BioGRID	SSF	128	3	219	21.3280	31.2200	11.2000	1.3324	0.6367
	MCL	91	3	4404	59.8352	13.5251	15.6256	1	0.0346
	ClusterOne	473	3	87	7.5700	16.8600	17.9700	1.3872	0.5426
	MDS	3759	3	101	29.7574	317.5739	*NA*	19.8330	0.0000

3.2 The analysis of predicted complexes which are not detected by SSF

We analyze those complexes that are not detected by SSF but found by other methods. ClusterONE\SSF denotes the set of complexes that are reported by ClusterONE but are not detected by SSF and not included in gold standard reference set(CYC2008,MIPS and SGD). Similarly, MDS\SSF (MCL\SSF) denotes the difference set between MDS (MCL) and SSF. The result is shown in Table 4 in which Num denotes the size of difference set of complexes and P_o is the fraction of complexes in the difference set which are detected by other three methods. We could observe that the value of P_o has a negative correlation with the value of Num. In other words, the larger the size of difference set is, the smaller P_o is. This means that other methods may report more additional valid complexes that are not contained in the reference sets at the cost of generating more false positives.

Table 4: The analysis of complexes that are not detected by SSF but found by other methods.

	Collins		Gavin		KroganC		KroganE		BioGRID	
	Num	P_o	Num	P_o	Num	P_o	Num	P_o	Num	P_o
ClusterONE\SSF	48	41.7%	115	40.9%	102	73.5%	121	57.0%	272	9.9%
MDS\SSF	46	50.0%	171	36.3%	1151	10.3%	2093	5.1%	2461	1.2%
MCL\SSF	31	58.1%	72	54.2%	250	28.4%	419	16.2%	179	14.0%

3.3 The performance comparison of SSF, MCL, ClusterONE and MDS when MIPS and SGD are used as the reference set

Table 5: The performance comparison of SSF, MCL, ClusterONE and MDS when MIPS is used as the reference set.

	Algorithm	Matched	Predicted	NMI	ACC	Frac	MMR	Composite	Precision	Recall	F1 score
Collins	SSF	82	127	**0.3168**	0.5362	0.6891	0.3440	1.5693	0.4803	0.6891	**0.5661**
	MCL	88	158	0.3161	0.5490	0.7395	0.3596	1.6481	0.3987	0.7395	0.5181
	ClusterONE	89	203	0.2854	0.5421	0.7479	0.3963	**1.6863**	0.3596	0.7479	0.4857
	MDS	82	333	0.1350	0.5609	0.6891	0.3654	1.6154	0.4384	0.6891	0.5359
Gavin	SSF	69	167	**0.2510**	0.5005	0.6000	0.3020	1.4025	0.3114	0.6000	**0.4100**
	MCL	69	220	0.1836	0.5089	0.6000	0.2720	1.3809	0.2273	0.6000	0.3297
	ClusterONE	74	294	0.1542	0.4944	0.6435	0.3115	**1.4494**	0.2041	0.6435	0.3099
	MDS	74	586	0.0865	0.4800	0.6435	0.3003	1.4238	0.2321	0.6435	0.3411
KroganC	SSF	56	91	0.1707	0.3986	0.4118	0.1769	0.9873	0.3956	0.4118	**0.4035**
	MCL	74	374	0.1048	0.4337	0.5441	0.2291	1.2069	0.1471	0.5441	0.2315
	ClusterONE	67	242	**0.1919**	0.3919	0.4926	0.2406	1.1251	0.2438	0.4926	0.3262
	MDS	92	1671	0.0997	0.4505	0.6765	0.3571	1.4841	0.1855	0.6765	0.2912
KroganE	SSF	55	77	0.1517	0.3763	0.3503	0.1373	0.8639	0.4675	0.3503	**0.4005**
	MCL	61	515	0.0405	0.3778	0.3885	0.1512	0.9175	0.0893	0.3885	0.1452
	ClusterONE	60	239	**0.1728**	0.3777	0.3822	0.1873	0.9472	0.2427	0.3822	0.2969
	MDS	89	2776	0.0711	0.4082	0.5669	0.2823	**1.2574**	0.1549	0.5669	0.2433
BioGRID	SSF	59	128	**0.1124**	0.4456	0.3122	0.1083	0.8660	0.3516	0.3122	**0.3307**
	MCL	7	91	0.0171	0.2442	0.0370	0.0138	0.2950	0.0659	0.0370	0.0474
	ClusterONE	88	473	0.0954	0.4401	0.4656	0.1864	1.0921	0.1734	0.4656	0.2527
	MDS	89	3759	0.0159	0.5002	0.4709	0.1952	**1.1663**	0.2370	0.4709	0.3153

Table 6: The performance comparison of SSF, MCL, ClusterONE and MDS when SGD is used as the reference set.

	Algorithm	Matched	Predicted	NMI	ACC	Frac	MMR	Composite	Precision	Recall	F1 score
Collins	SSF	99	127	**0.4017**	0.6849	0.7388	0.4303	1.8540	0.5827	0.7388	**0.6515**
	MCL	106	158	0.3670	0.7066	0.7910	0.4580	1.9556	0.5127	0.7910	0.6221
	ClusterONE	108	203	0.3598	0.7228	0.8060	0.5254	**2.0542**	0.4532	0.8060	0.5802
	MDS	102	333	0.1042	0.6488	0.7612	0.4739	1.8839	0.3423	0.7612	0.4723
Gavin	SSF	83	167	**0.3028**	0.7006	0.6484	0.3837	1.7327	0.4012	0.6484	**0.4957**
	MCL	85	220	0.2364	0.7117	0.6641	0.3383	1.7141	0.2955	0.6641	0.4090
	ClusterONE	93	294	0.1975	0.6930	0.7266	0.3953	**1.8149**	0.2653	0.7266	0.3887
	MDS	93	586	0.1108	0.6464	0.7266	0.3704	1.7434	0.3038	0.7266	0.4284
KroganC	SSF	75	91	0.3144	0.5382	0.4545	0.2527	1.2455	0.6374	0.4545	**0.5306**
	MCL	99	374	0.1615	0.6181	0.6000	0.3102	1.5283	0.2299	0.6000	0.3325
	ClusterONE	93	242	**0.3210**	0.5776	0.5636	0.3486	1.4898	0.3884	0.5636	0.4599
	MDS	129	1671	0.1179	0.5840	0.7818	0.4567	**1.8225**	0.2053	0.7818	0.3252
KroganE	SSF	70	77	**0.2882**	0.5152	0.3743	0.2133	1.1029	0.7273	0.3743	**0.4943**
	MCL	74	515	0.0742	0.5441	0.3957	0.1884	1.1282	0.1243	0.3957	0.1891
	ClusterONE	80	239	0.2770	0.5319	0.4278	0.2472	1.2069	0.3682	0.4278	0.3958
	MDS	119	2776	0.0824	0.5484	0.6364	0.3530	**1.5378**	0.1726	0.6364	0.2715
BioGRID	SSF	76	128	0.1498	0.5239	0.3262	0.1421	0.9922	0.4766	0.3262	**0.3873**
	MCL	7	91	0.0171	0.2442	0.0370	0.0138	0.2950	0.0659	0.0370	0.0474
	ClusterONE	131	473	**0.1633**	0.6279	0.5622	0.2713	**1.4614**	0.2770	0.5622	0.3711
	MDS	129	3759	0.0125	0.4835	0.5536	0.2345	1.2716	0.0944	0.5536	0.1614

3.4 The performance comparison of SSF, ESSC and OSLOM

To further verify the performance of SSF, we carry out some additional experiments where ESSC and OSLOM are selected as the baseline algorithms. The details of comparison results are shown in Supplementary Table 6 - Table 8. In general, there are no algorithms that can always achieve the best performance over all assessment measures. Furthermore, we can observe that SSF could achieve better performance in most cases in terms of NMI. As to composite value and F1 score, we could not get an unified conclusion to claim which algorithm is better than others. Overall, SSF is competitive with the-state-of-art methods in the detection of statistically significant subgraphs.

Table 7: The performance comparison of SSF, ESSC and OSLOM when CYC2008 is used as the reference set.

	Algorithm	Matched	Predicted	NMI	ACC	Frac	MMR	Composite	Precision	Recall	F1 score
Collins	SSF	111	127	**0.5122**	0.7448	0.7708	0.4729	1.9885	0.6929	0.7708	**0.7298**
	ESSC	111	165	0.3805	0.6938	0.7708	0.5125	1.9771	0.6667	0.7708	0.7150
	OSLOM	111	402	0.3805	0.7729	0.7708	0.4976	**2.0413**	0.2363	0.7708	0.3617
Gavin	SSF	97	167	**0.3731**	0.7069	0.7029	0.4155	1.8253	0.4731	0.7029	0.5655
	ESSC	100	234	0.2476	0.6754	0.7246	0.4508	**1.8508**	0.5171	0.7246	**0.6035**
	OSLOM	81	192	0.2504	0.7351	0.5870	0.3073	1.6294	0.3385	0.5870	0.4294
KroganC	SSF	84	91	0.3822	0.6364	0.5122	0.3057	**1.4543**	0.7582	0.5122	**0.6114**
	ESSC	77	99	**0.4572**	0.6169	0.4695	0.3135	1.3999	0.7778	0.4695	0.5856
	OSLOM	58	231	0.1107	0.6738	0.3537	0.1667	1.1942	0.2294	0.3537	0.2783
KroganE	SSF	75	77	**0.3519**	0.6150	0.4144	0.2434	**1.2727**	0.8182	0.4144	**0.5501**
	ESSC	60	66	0.3165	0.5214	0.3315	0.1899	1.0428	0.8636	0.3315	0.4791
	OSLOM	32	109	0.0400	0.5847	0.1768	0.0625	0.8240	0.2844	0.1768	0.2180
BioGRID	SSF	80	128	**0.1730**	0.5887	0.3390	0.1584	**1.0860**	0.4844	0.3390	**0.3988**
	ESSC	50	80	0.0928	0.5031	0.2119	0.1060	0.8210	0.5000	0.2119	0.2976
	OSLOM	55	151	0.0625	0.6505	0.2331	0.0980	0.9816	0.3245	0.2331	0.2713

Table 8: The performance comparison of SSF, ESSC and OSLOM when MIPS is used as the reference set.

	Algorithm	Matched	Predicted	NMI	ACC	Frac	MMR	Composite	Precision	Recall	F1 score
Collins	SSF	82	127	**0.3168**	0.5362	0.6891	0.3440	1.5693	0.4803	0.6891	0.5661
	ESSC	82	165	0.2378	0.5089	0.6891	0.3736	**1.5716**	0.4909	0.6891	**0.5734**
	OSLOM	77	402	0.2311	0.5426	0.6471	0.3268	1.5165	0.1493	0.6471	0.2426
Gavin	SSF	69	167	**0.2510**	0.5005	0.6000	0.3020	1.4025	0.3114	0.6000	0.4100
	ESSC	76	234	0.1645	0.4745	0.6609	0.3365	**1.4719**	0.3462	0.6609	**0.4543**
	OSLOM	61	192	0.1531	0.5003	0.5304	0.2289	1.2596	0.2344	0.5304	0.3251
KroganC	SSF	56	91	0.1707	0.3986	0.4118	0.1769	0.9873	0.3956	0.4118	0.4035
	ESSC	57	99	**0.2116**	0.4051	0.4191	0.1949	**1.0191**	0.4444	0.4191	**0.4314**
	OSLOM	27	231	0.0472	0.4054	0.1985	0.0848	0.6887	0.1169	0.1985	0.1471
KroganE	SSF	55	77	**0.1517**	0.3763	0.3503	0.1373	**0.8639**	0.4675	0.3503	**0.4005**
	ESSC	46	66	0.1391	0.3578	0.2930	0.1154	0.7662	0.5455	0.2930	0.3812
	OSLOM	17	109	0.0160	0.3463	0.1083	0.0322	0.4868	0.1193	0.1083	0.1135
BioGRID	SSF	59	128	**0.1124**	0.4456	0.3122	0.1083	**0.8660**	0.3516	0.3122	**0.3307**
	ESSC	35	80	0.0504	0.4213	0.1852	0.0600	0.6665	0.3625	0.1852	0.2451
	OSLOM	39	151	0.0352	0.4429	0.2063	0.0657	0.7149	0.1987	0.2063	0.2024

Table 9: The performance comparison of SSF, ESSC and OSLOM when SGD is used as the reference set.

	Algorithm	Matched	Predicted	NMI	ACC	Frac	MMR	Composite	Precision	Recall	F1 score
Collins	SSF	99	127	**0.4017**	0.6849	0.7388	0.4303	1.8540	0.5827	0.7388	0.6515
	ESSC	97	165	0.3156	0.6305	0.7239	0.4894	1.8438	0.5939	0.7239	**0.6525**
	OSLOM	100	402	0.2879	0.7164	0.7463	0.4518	**1.9145**	0.2164	0.7463	0.3355
Gavin	SSF	83	167	**0.3028**	0.7006	0.6484	0.3837	1.7327	0.4012	0.6484	0.4957
	ESSC	89	234	0.2127	0.6392	0.6953	0.4351	**1.7696**	0.4444	0.6953	**0.5423**
	OSLOM	72	192	0.1744	0.6853	0.5625	0.2685	1.5163	0.2917	0.5625	0.3841
KroganC	SSF	75	91	0.3144	0.5382	0.4545	0.2527	1.2455	0.6374	0.4545	0.5306
	ESSC	73	99	**0.3802**	0.5386	0.4424	0.2780	**1.2590**	0.7071	0.4424	**0.5443**
	OSLOM	51	231	0.0815	0.5568	0.3091	0.1462	1.0121	0.1991	0.3091	0.2422
KroganE	SSF	70	77	**0.2882**	0.5152	0.3743	0.2133	**1.1029**	0.7273	0.3743	**0.4943**
	ESSC	59	66	0.2582	0.4384	0.3155	0.1658	0.9197	0.7727	0.3155	0.4481
	OSLOM	28	108	0.0212	0.4640	0.1497	0.0503	0.6640	0.2477	0.1497	0.1866
BioGRID	SSF	76	128	**0.1498**	0.5239	0.3262	0.1421	**0.9922**	0.4766	0.3262	**0.3873**
	ESSC	47	80	0.0953	0.4673	0.2017	0.0936	0.7626	0.4625	0.2017	0.2809
	OSLOM	42	151	0.0429	0.5647	0.1803	0.0724	0.8174	0.2517	0.1803	0.2101

3.5 The performance comparison on BioPlex 2.0

To test the performance of different algorithms on large-scale PPI network, we choose BioPlex 2.0[3] in our experiment.Firstly,we compare different methods with respect to the size distribution of predicted complexes in Table 10, where $|C|_{min}$, $|C|_{max}$ and $|\bar{C}|$ denote the minimal size, maximal size and average size of predicted complexes, respectively. In addition, $|\bar{D}|_{com}$ denotes the average degree of the vertices in a protein complex, and $|\bar{D}|_{bkg}$ is the average degree of the background vertices. $\bar{n_C}$ is the average number of complexes to which non-background vertices belong, and P_{bkg} is the proportion of background vertices. We could observe that SSF reports the least number of protein complexes, which indicates that our method is conservative. Meanwhile, the average size of protein complexes of SSF is much larger than that of other methods.

Since BioPlex 2.0 is the largest human PPI network so far, we may find many novel valid protein complexes that are still not contained in the current reference sets. Anyway, to compare the performance of different methods with some widely used performance indicators such as NMI, we use the Corum database[4] as the reference set. Obviously, such a comparison may not fully reflect the merits of different methods due to the incompleteness of reference set, it at least can reveal some underlying features of different algorithms to some extend. As shown in Table 11, SSF can achieve the highest precision and F1-score on BioPlex 2.0. Meanwhile, it is the second best performer with respect to NMI.

Table 10: The comparison on the complex size distribution of different methods on BioPlex 2.0

	Predicted	$\lvert C\rvert_{min}$	$\lvert C\rvert_{max}$	$\lvert\bar{C}\rvert$	$\lvert\bar{D}\rvert_{com}$	$\lvert\bar{D}\rvert_{bkg}$	P_{bkg}	$\bar{n_C}$
SSF	391	3	279	16.8747	16.5339	6.5987	0.6641	1.8047
MCL	1332	3	300	6.4752	8.4786	12.7586	0.2075	1
ClusterONE	785	3	47	5.4803	9.3446	9.9765	0.6808	1.2383
MDS	8147	3	38	6.2270	47.8120	1.0	0.0040	4.6804

Table 11: The performance comparison of different algorithms on BioPlex 2.0 when using Corum as the reference set.

Algorithm	Matched	Predicted	NMI	ACC	Frac	MMR	Composite	Precision	Recall	F1 score
SSF	23	391	0.0137	0.4476	0.1494	0.0687	0.6657	0.0639	0.1494	**0.0895**
MCL	33	1332	0.0136	0.5209	0.2143	0.1010	0.8362	0.0248	0.2143	0.0444
ClusterONE	33	785	**0.0212**	0.4695	0.2143	0.1018	0.7856	0.0484	0.2143	0.0790
MDS	50	8147	0.0032	0.4780	0.3247	0.1360	**0.9387**	0.0228	0.3247	0.0427

3.6 FWER vs. FDR

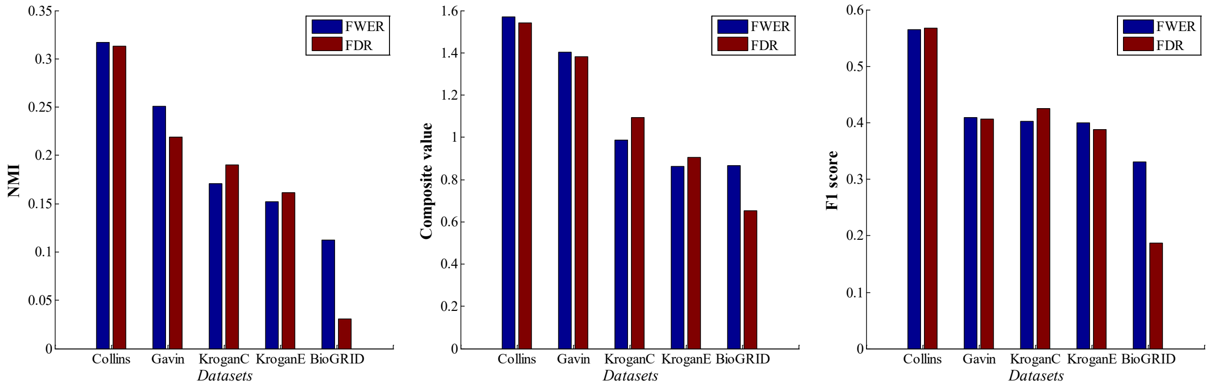

Figure 1: The performance comparison of two variants of SSF that are equipped with FWER and FDR. Here MIPS is used as the reference set.

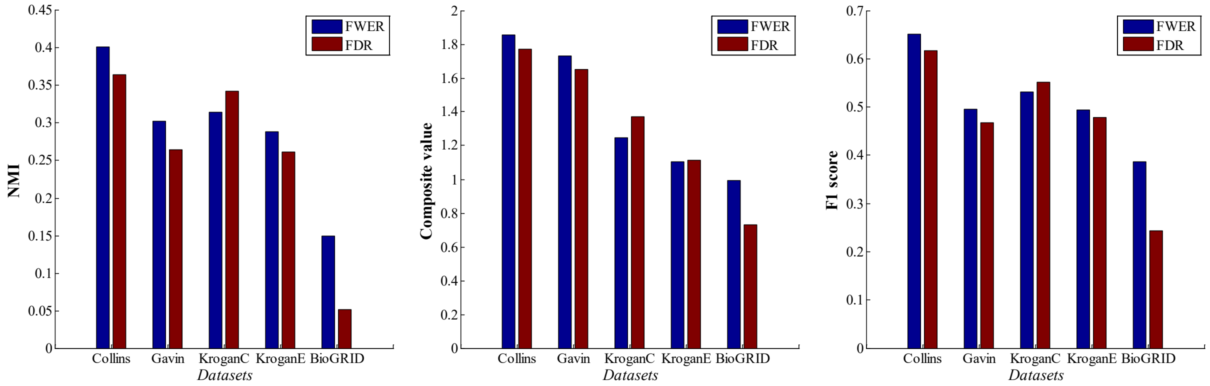

Figure 2: The performance comparison of two variants of SSF that are equipped with FWER and FDR. Here SGD is used as the reference set.

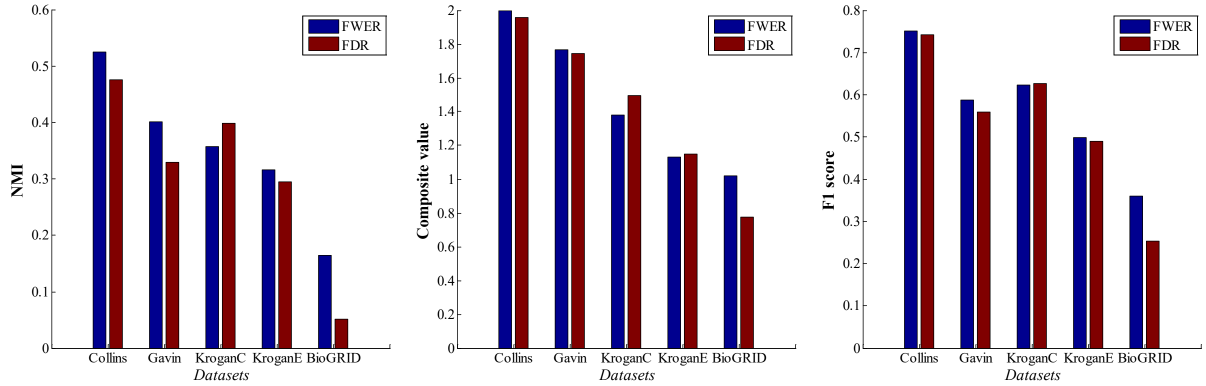

Figure 3: The performance comparison of two variants of SSF that are equipped with FWER and FDR, where a binomial distribution is used to calculate the p-values and CYC2008 is used as the reference set.

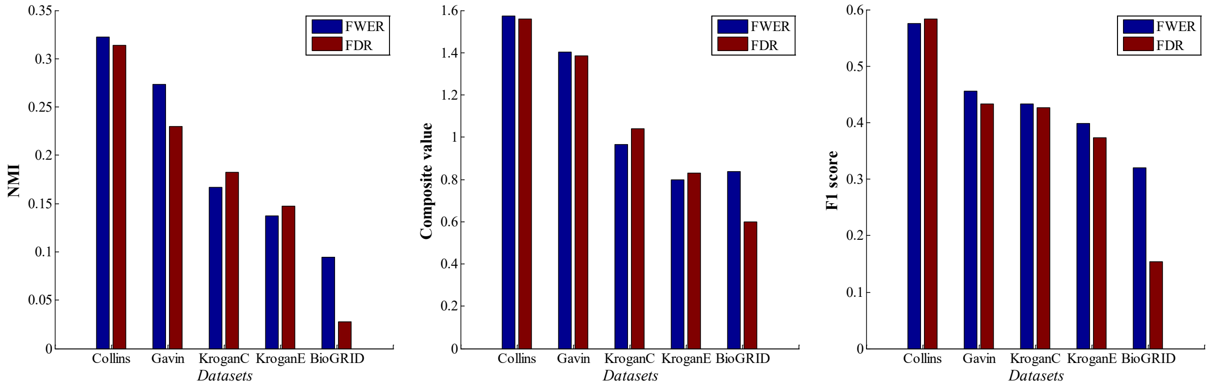

Figure 4: The performance comparison of two variants of SSF that are equipped with FWER and FDR, where a binomial distribution is used to calculate the p-values and MIPS is used as the reference set.

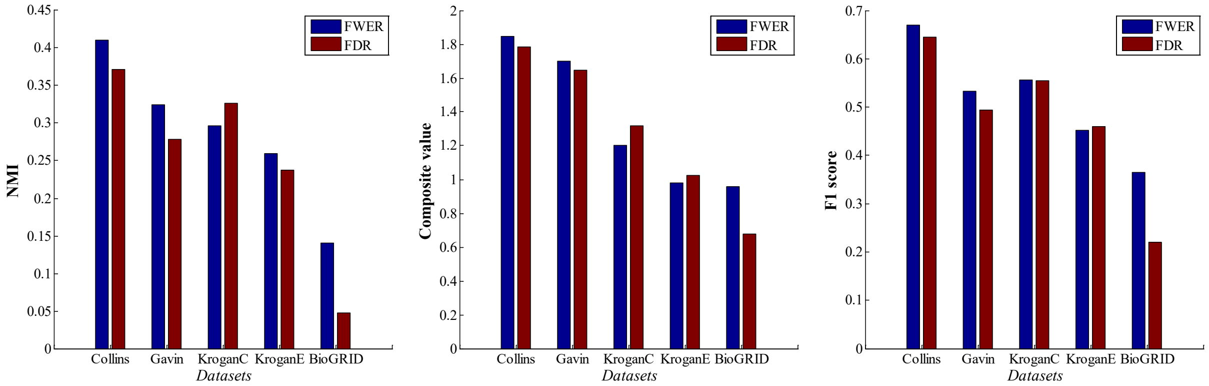

Figure 5: The performance comparison of two variants of SSF that are equipped with FWER and FDR, where a binomial distribution is used to calculate the p-values and SGD is used as the reference set.

3.7 Hypergeometric distribution vs. Binomial distribution

In order to test the effect of using different p-value calculation methods, we compare the identification performance between our method based on Fisher's exact test and the method based on binomial distribution in ESSC. The detailed results are presented in Supplementary Fig.6 – Supplementary Fig. 8, where CYC2008, MIPS and SGD are used as the reference set, respectively. In these figures, we adopt hypergeometric distribution (Fisher) and binomial distribution (Binomial) as the probability density function to calculate the p-value. We can observe that SSF equipped with hypergeometric distribution could achieve better performance in most cases. This indicates that the proposed p-value calculation method in this paper is more plausible in the context of protein complexes identification.

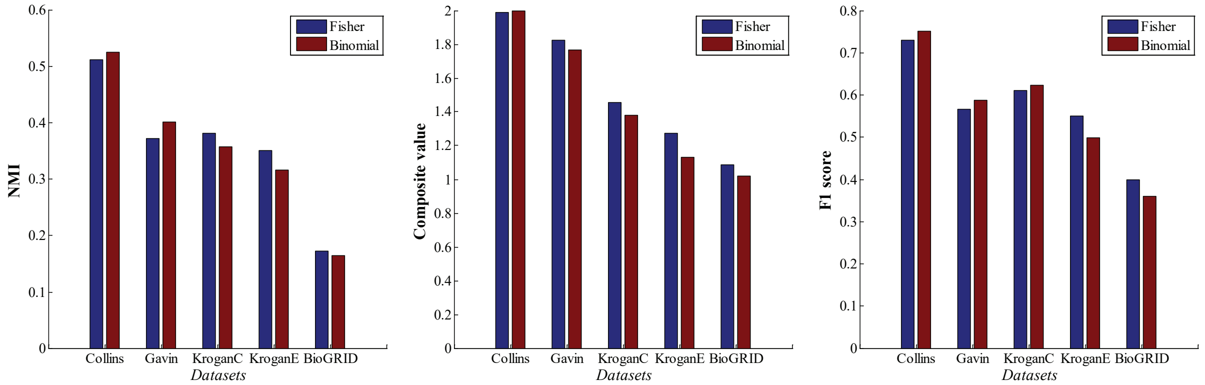

Figure 6: The performance comparison between two p-value calculation methods that are based on hypergeometric distribution and binomial distribution when CYC2008 is used as the reference set.

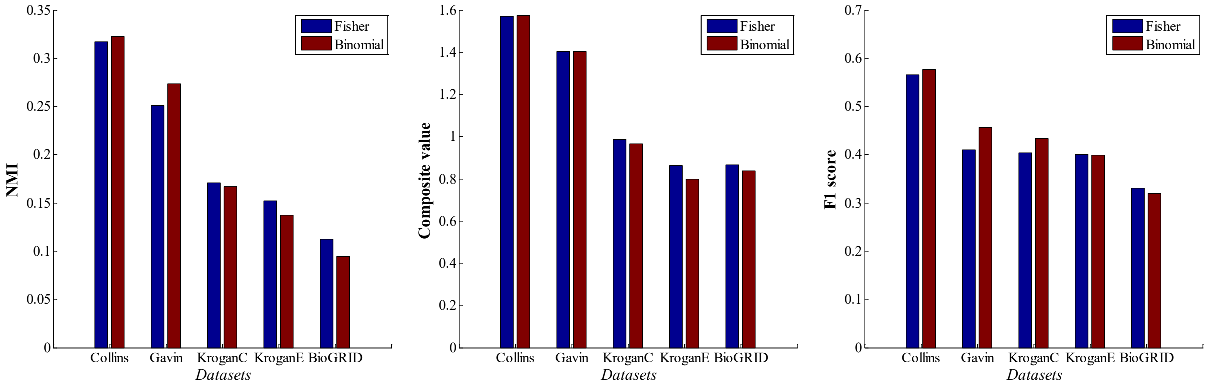

Figure 7: The performance comparison between two p-value calculation methods that are based on hypergeometric distribution and binomial distribution when MIPS is used as the reference set.

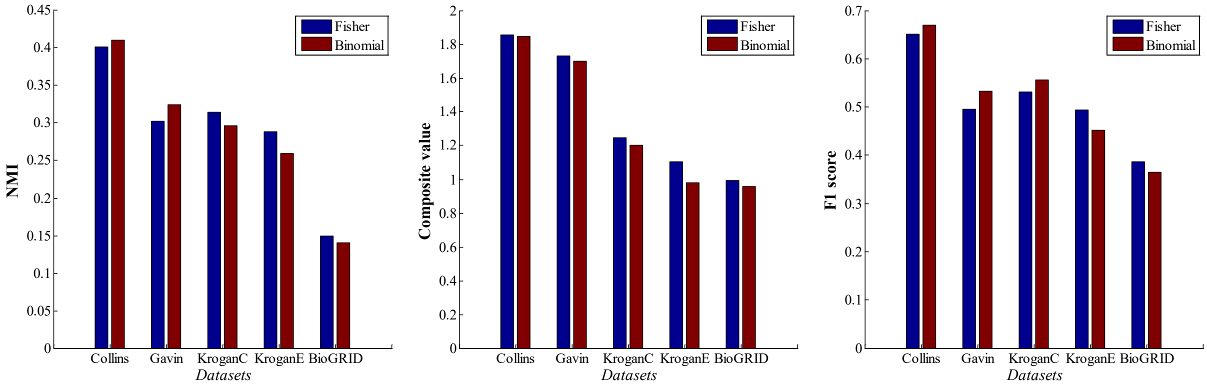

Figure 8: The performance comparison between two p-value calculation methods that are based on hypergeometric distribution and binomial distribution when SGD is used as the reference set.

3.8 Parameter sensitivity

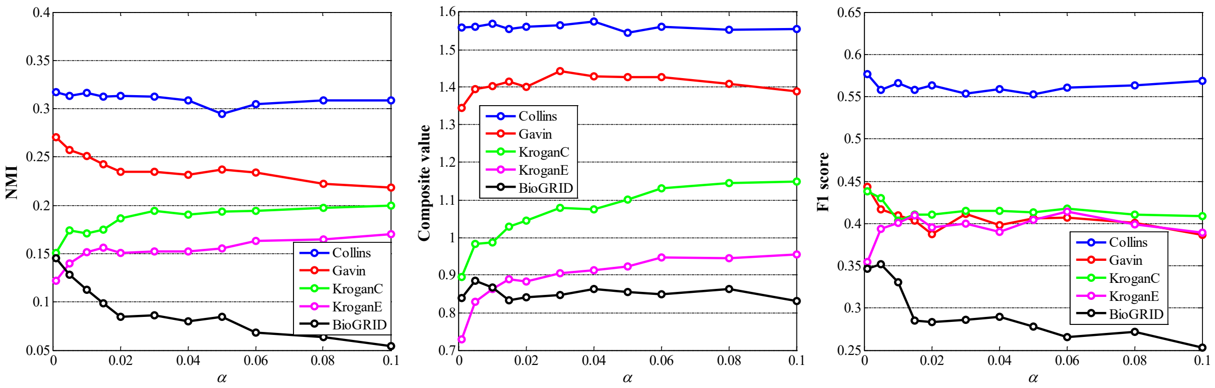

Figure 9: The effect of significance level α on the identification performance of SSF when MIPS is used as the reference set.

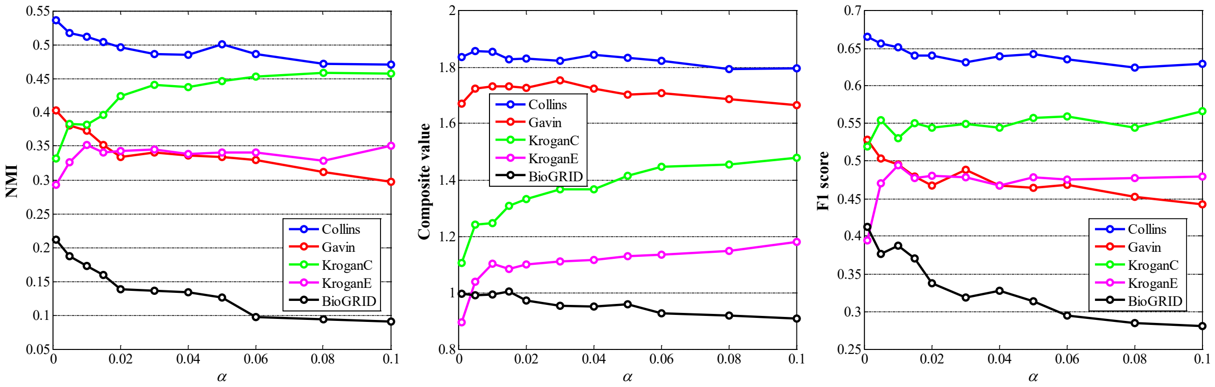

Figure 10: The effect of significance level α on the identification performance of SSF when SGD is used as the reference set.

In order to show that identification result is stable with respect to α in most data sets in a quantitative manner, we compute some statistics in Table 12, Table 13, and Table 14. In these tables, μ, D, σ, R respectively stands for the mean, the variance, the standard deviation, and the range of the measures when α ranges from 0.01 to 0.1. We can find that the variance, the standard deviation and the range is relatively small (compared to the mean) in most cases, which indicates that α has no significant effect on the identification result within [0.01,0.1].

Table 12: The stability of the identification performance of SSF with respect to α when CYC2008 is used as the reference set.

	Measures	μ	D	σ	R
Collins	NMI	0.4966	0.0004	0.0199	0.0659
	Composite	1.9722	0.0001	0.0115	0.0344
	F1 score	0.7186	0.0000	0.0065	0.0192
Gavin	NMI	0.3446	0.0009	0.0307	0.1056
	Composite	1.7954	0.0004	0.0194	0.0588
	F1 score	0.5428	0.0008	0.0289	0.1005
KroganC	NMI	0.4190	0.0017	0.0408	0.1267
	Composite	1.5425	0.0177	0.1329	0.4434
	F1 score	0.6272	0.0006	0.0245	0.0905
KrogenE	NMI	0.3361	0.0003	0.0163	0.0587
	Composite	1.2638	0.0075	0.0865	0.3076
	F1 score	0.5306	0.0009	0.0304	0.1064
BioGRID	NMI	0.1406	0.0015	0.0393	0.1216
	Composite	1.0480	0.0008	0.0281	0.0811
	F1 score	0.0368	0.0014	0.0368	0.0982

Table 13: The stability of the identification performance of SSF with respect to α when MIPS is used as the reference set.

	Measures	μ	D	σ	R
Collins	NMI	0.3100	0.0000	0.0063	0.0226
	Composite	1.5601	0.0001	0.0081	0.0297
	F1 score	0.5617	0.0000	0.0069	0.0236
Gavin	NMI	0.2395	0.0002	0.0152	0.0525
	Composite	1.4076	0.0007	0.0261	0.0965
	F1 score	0.4063	0.0002	0.0156	0.0577
KroganC	NMI	0.1842	0.0002	0.0148	0.0489
	Composite	1.0561	0.0062	0.0785	0.2528
	F1 score	0.4158	0.0001	0.0100	0.0347
KrogenE	NMI	0.1526	0.0002	0.0131	0.0486
	Composite	0.8891	0.0042	0.0651	0.2254
	F1 score	0.2947	0.0011	0.0330	0.0985
BioGRID	NMI	0.0913	0.0008	0.0277	0.0914
	Composite	0.8520	0.0003	0.0164	0.0541
	F1 score	0.2947	0.0011	0.0330	0.0985

Table 14: The stability of the identification performance of SSF with respect to α when SGD is used as the reference set.

	Measures	μ	D	σ	R
Collins	NMI	0.3877	0.0004	0.0203	0.0698
	Composite	1.8289	0.0004	0.0207	0.0641
	F1 score	0.6416	0.0002	0.0122	0.0414
Gavin	NMI	0.2851	0.0005	0.0214	0.0741
	Composite	1.7116	0.0008	0.0277	0.0870
	F1 score	0.4780	0.0006	0.0247	0.0862
KroganC	NMI	0.3406	0.0007	0.0267	0.0835
	Composite	1.3418	0.0126	0.1122	0.3738
	F1 score	0.5474	0.0002	0.0132	0.0471
KrogenE	NMI	0.2664	0.0001	0.0122	0.0495
	Composite	1.0945	0.0057	0.0758	0.2862
	F1 score	0.4702	0.0007	0.0261	0.1000
BioGRID	NMI	0.1301	0.0013	0.0364	0.1199
	Composite	0.9610	0.0011	0.0331	0.0967
	F1 score	0.3366	0.0020	0.0442	0.1321

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[2] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2, 2003.

[3] Edward L Huttlin, Raphael J Bruckner, Joao A Paulo, Joe R Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P Gygi, Hannah Parzen, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 2017.

[4] Pierre C Havugimana, G Traver Hart, Tamás Nepusz, Haixuan Yang, Andrei L Turinsky, Zhihua Li, Peggy I Wang, Daniel R Boutz, Vincent Fong, Sadhna Phanse, et al. A census of human soluble protein complexes. *Cell*, 150(5):1068–1081, 2012.