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

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 (FDR)

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 SSF, ClusterONE and MDS when MIPS and SGD are used as the reference set

Table 3: The performance comparison of SSF, 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
	ClusterONE	89	203	0.2854	0.5421	0.7479	0.3963	**1.6863**	0.3596	0.7479	0.4857
	MDS	84	307	0.1333	0.5631	0.7059	0.3638	1.6328	0.4886	0.7059	**0.5775**
Gavin	SSF	69	167	**0.2510**	0.5005	0.6000	0.3020	1.4025	0.3114	0.6000	**0.4100**
	ClusterONE	74	294	0.1542	0.4944	0.6435	0.3115	**1.4494**	0.2041	0.6435	0.3099
	MDS	70	554	0.0806	0.4796	0.6087	0.2848	1.3731	0.2509	0.6087	0.3553
KroganC	SSF	56	91	0.1707	0.3986	0.4118	0.1769	0.9873	0.3956	0.4118	**0.4035**
	ClusterONE	67	242	**0.1919**	0.3919	0.4926	0.2406	1.1251	0.2438	0.4926	0.3262
	MDS	89	1663	0.1016	0.4524	0.6544	0.3552	**1.4620**	0.1930	0.6544	0.2981
KroganE	SSF	55	77	0.1517	0.3763	0.3503	0.1373	0.8639	0.4675	0.3503	**0.4005**
	ClusterONE	60	239	**0.1728**	0.3777	0.3822	0.1873	0.9472	0.2427	0.3822	0.2969
	MDS	88	2772	0.0778	0.4133	0.5605	0.2825	**1.2563**	0.1840	0.5605	0.2770
BioGRID	SSF	59	128	**0.1124**	0.4456	0.3122	0.1083	0.8660	0.3516	0.3122	**0.3307**
	ClusterONE	88	473	0.0954	0.4401	0.4656	0.1864	1.0921	0.1734	0.4656	0.2527
	MDS	90	3759	0.0167	0.5003	0.4762	0.1951	**1.1716**	0.2519	0.4762	0.3295

Table 4: The performance comparison of SSF, 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**
	ClusterONE	108	203	0.3598	0.7228	0.8060	0.5254	**2.0542**	0.4532	0.8060	0.5802
	MDS	101	307	0.1026	0.6433	0.7537	0.4629	1.8599	0.3811	0.7537	0.5062
Gavin	SSF	83	167	**0.3028**	0.7006	0.6484	0.3837	1.7327	0.4012	0.6484	**0.4957**
	ClusterONE	93	294	0.1975	0.6930	0.7266	0.3953	**1.8149**	0.2653	0.7266	0.3887
	MDS	90	554	0.1054	0.6460	0.7031	0.3568	1.7059	0.3339	0.7031	0.4528
KroganC	SSF	75	91	0.3144	0.5382	0.4545	0.2527	1.2455	0.6374	0.4545	**0.5306**
	ClusterONE	93	242	**0.3210**	0.5776	0.5636	0.3486	1.4898	0.3884	0.5636	0.4599
	MDS	127	1663	0.1173	0.5827	0.7697	0.4504	**1.8028**	0.2177	0.7697	0.3394
KroganE	SSF	70	77	**0.2882**	0.5152	0.3743	0.2133	1.1029	0.7273	0.3743	**0.4943**
	ClusterONE	80	239	0.2770	0.5319	0.4278	0.2472	1.2069	0.3682	0.4278	0.3958
	MDS	118	2772	0.0850	0.5441	0.6310	0.3484	**1.5235**	0.2006	0.6310	0.3044
BioGRID	SSF	76	128	0.1498	0.5239	0.3262	0.1421	0.9922	0.4766	0.3262	**0.3873**
	ClusterONE	131	473	**0.1633**	0.6279	0.5622	0.2713	**1.4614**	0.2770	0.5622	0.3711
	MDS	130	3759	0.0128	0.4853	0.5579	0.2303	1.2735	0.1051	0.5579	0.1769

3.2 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 5 - Table 7. 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 5: 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 6: 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 7: 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

3.3 FWER vs. FDR

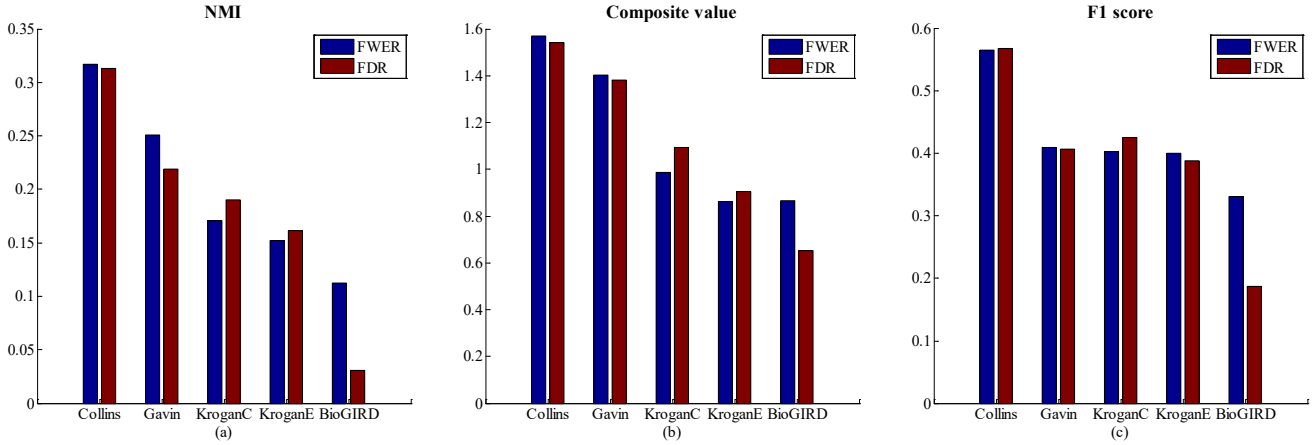

Figure 1: The performance comparison of two variants of SSF that are equipped with FWER and FDR. Here MIPS is used as the reference set.

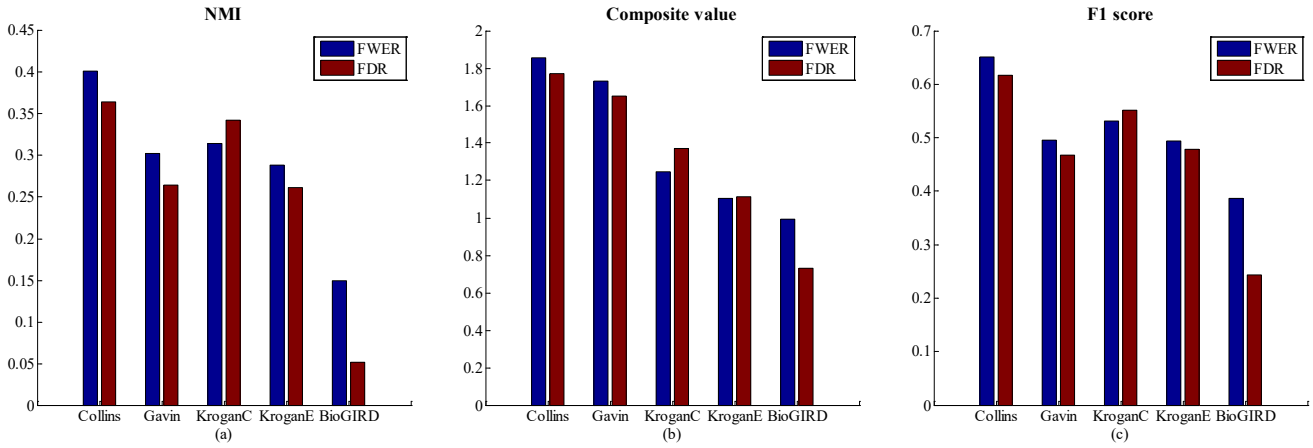

Figure 2: The performance comparison of two variants of SSF that are equipped with FWER and FDR. Here SGD is used as the reference set.

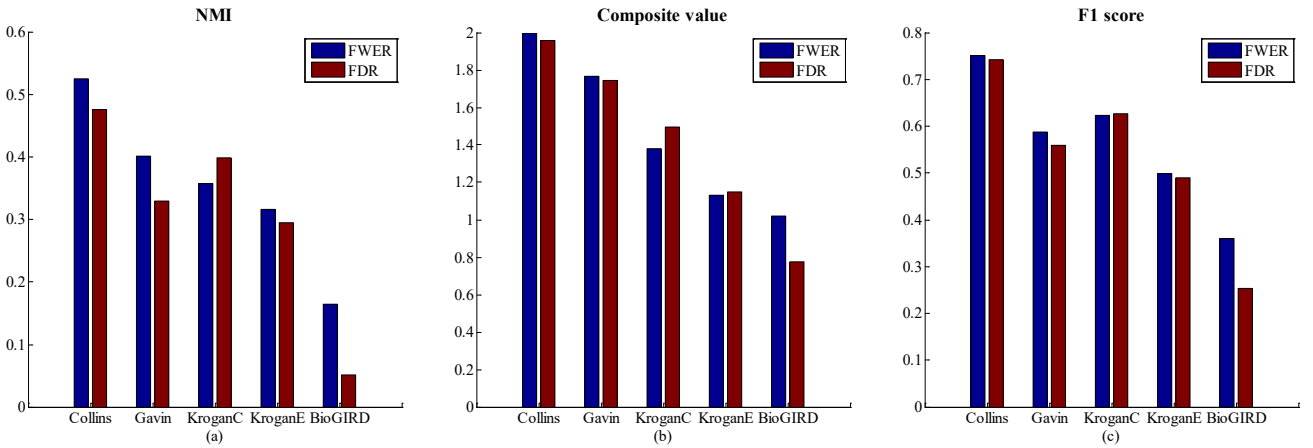

Figure 3: The performance comparison of two variants of SSF that are equipped with FWER and FDR, where a binomial distribution is used to calculate the p-values and SGD is used as the reference set.

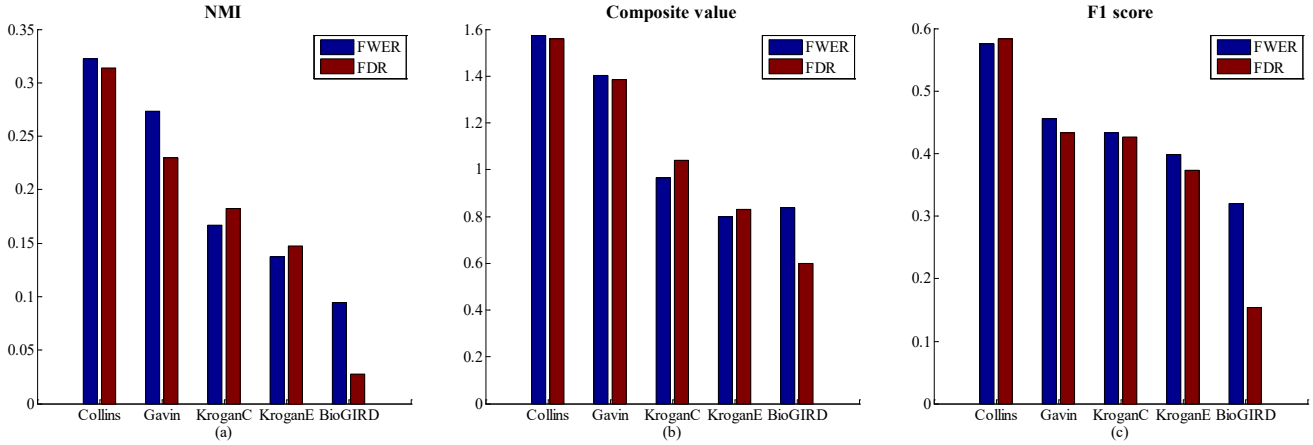

Figure 4: The performance comparison of two variants of SSF that are equipped with FWER and FDR, where a binomial distribution is used to calculate the p-values and MIPS is used as the reference set.

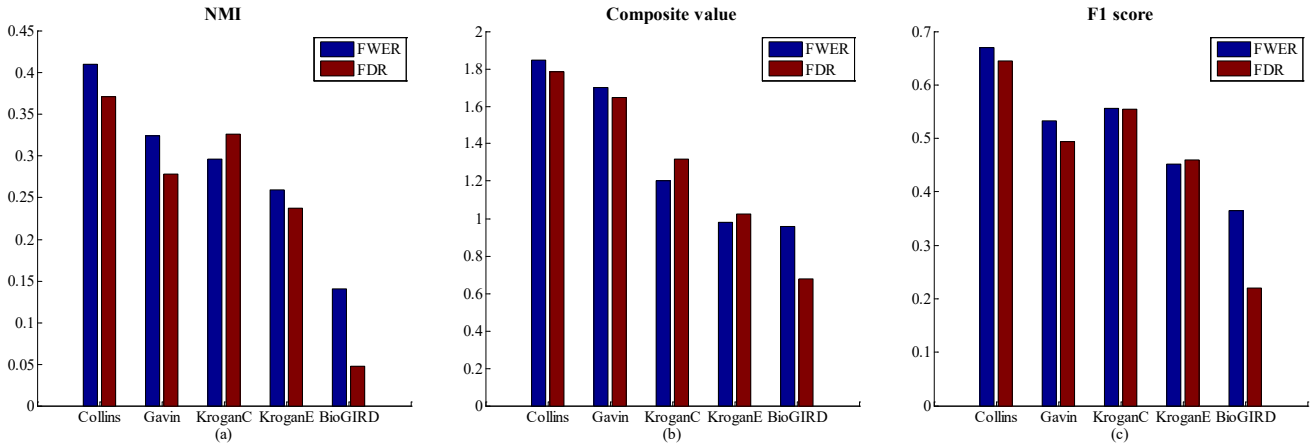

Figure 5: The performance comparison of two variants of SSF that are equipped with FWER and FDR, where a binomial distribution is used to calculate the p-values and SGD is used as the reference set.

3.4 Hypergeometric distribution vs. Binomial distribution

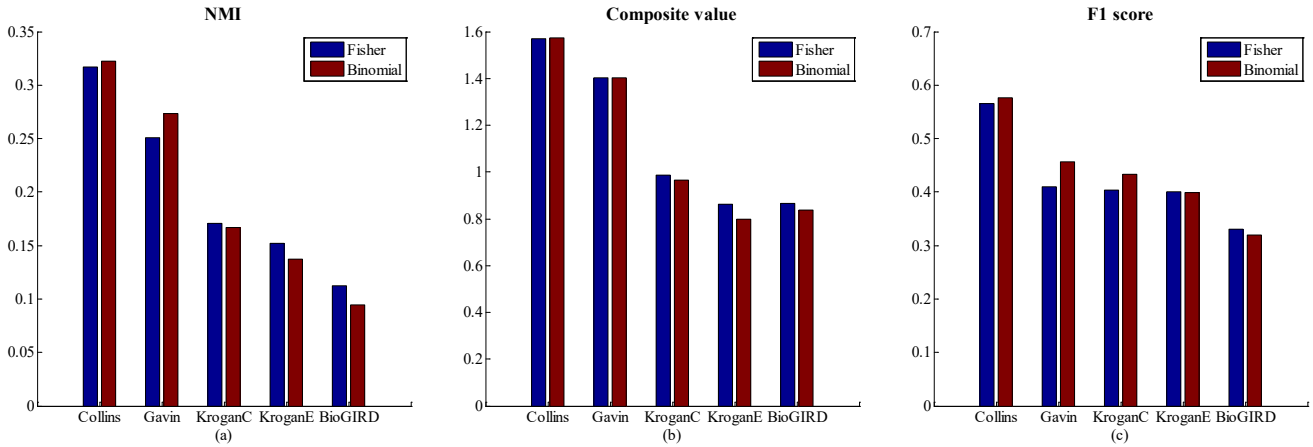

Figure 6: The performance comparison between two p-value calculation methods that are based on hypergeometric distribution and binomial distribution when MIPS is used as the reference set.

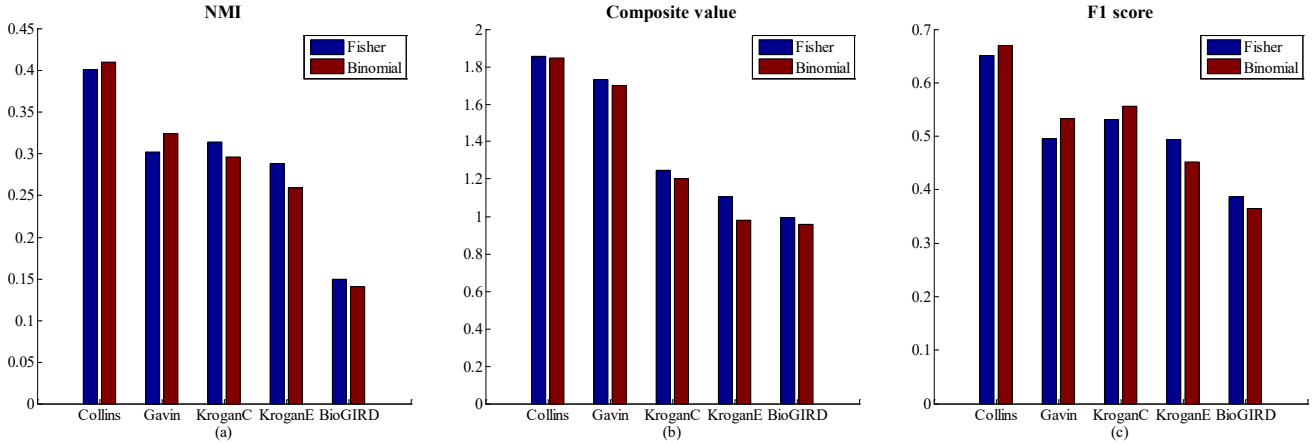

Figure 7: The performance comparison between two p-value calculation methods that are based on hypergeometric distribution and binomial distribution when SGD is used as the reference set.

3.5 Parameter sensitivity

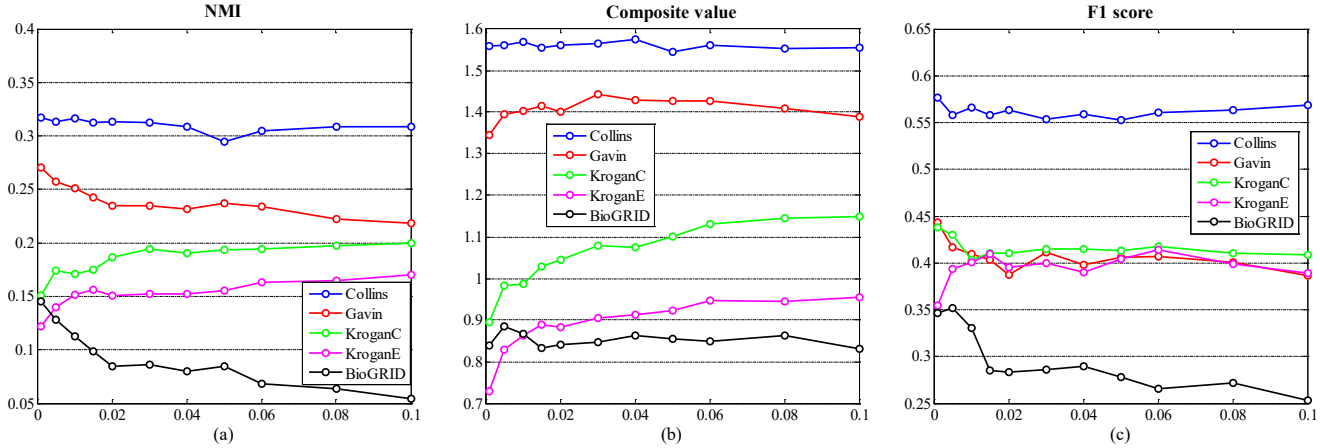

Figure 8: The effect of significance level α on the identification performance of SSF when MIPS is used as the reference set.

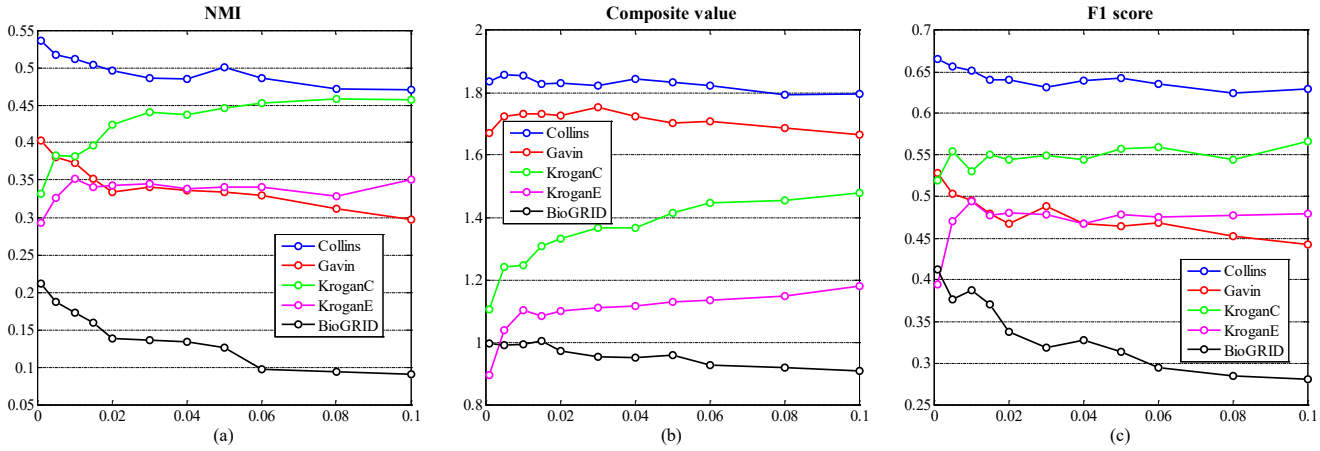

Figure 9: The effect of significance level α on the identification performance of SSF when SGD is used as the reference set.

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.