Supplementary Information

Predicting essential genes based on network and sequence analysis

Yih-Chii Hwang, Chen-Ching Lin, Jen-Yun Chang, Hirotada Mori,
Hsueh-Fen Juan, Hsuan-Cheng Huang

December 15, 2008

S1 Correlation between network and sequence features

The correlation coefficients between different network and sequence features used for the SVM classifiers are listed in Table S1 (*S. cerevisiae*) and Table S2 (*E. coli*). Most characteristics have low correlation with each other. Figure S1 shows the scattered plots of essential index (EI) versus degree (K) in *S. cerevisiae*.

Table S1: Correlation coefficients between different features in *S. cerevisiae*

	K	CFK	BC	CC	CCo	NID	KL	EI	PR	ORFL
K	1.00	0.95	0.85	0.58	0.06	0.61	0.55	0.04	0.24	0.08
CFK		1.00	0.71	0.60	0.12	0.66	0.66	0.09	0.32	0.11
BC			1.00	0.36	-0.04	0.28	0.24	-0.01	0.11	0.03
CC				1.00	0.09	0.34	0.55	0.18	0.24	0.12
CCo					1.00	0.18	0.59	0.14	0.14	0.06
NID						1.00	0.55	0.06	0.18	0.03
KL							1.00	0.17	0.35	0.15
EI								1.00	0.15	0.08
PR									1.00	0.38
ORFL										1.00

Table S2: Correlation coefficients between different features in *E. coli*

	K	CFK	BC	CC	CCo	NID	KL	EI	PR	ORFL
K	1.00	0.95	0.93	0.30	-0.02	0.82	0.30	0.01	0.18	0.04
CFK		1.00	0.86	0.27	-0.01	0.88	0.29	0.02	0.25	0.05
BC			1.00	0.18	-0.02	0.70	0.13	-0.01	0.10	0.02
CC				1.00	0.29	0.22	0.66	0.23	0.12	0.04
CCo					1.00	0.02	0.51	0.10	0.01	-0.01
NID						1.00	0.28	0.03	0.17	0.00
KL							1.00	0.22	0.16	0.02
EI								1.00	0.07	-0.06
PR									1.00	0.08
ORFL										1.00

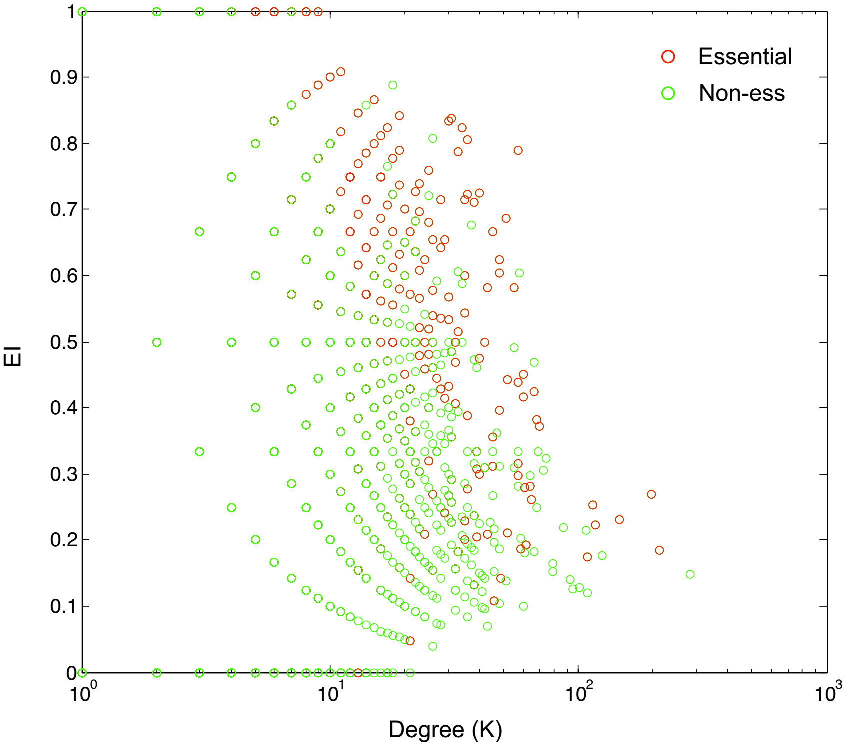

Figure S1: Scattered plots of essential index (EI) versus degree (K) for essential (red) and nonessential genes (green) in *S. cerevisiae*.

S2 Dependence of clustering coefficients on possible false positive interactions in *E. coli*

As described in the main text of this paper, we did not observe any difference of the clustering coefficients between essential and nonessential genes in *E. coli*, while the essential genes of yeast have significantly higher clustering coefficients in average than nonessential. One possible reason for such inconsistency might be due to the noisy systematic measurements of protein-protein interaction in *E. coli*, which were performed using pull-down assays. The false positive interactions might dilute the difference of clustering coefficients. In our previous work of systematic measurements of protein-protein interactions [1], we performed control experiments ($n = 16$) using the pCA24N vector to identify proteins with an intrinsic affinity for the Ni^{2+}-NTA resin (that is, no bait proteins). The frequently found proteins (in at least 10 of 16 experiments without bait proteins) were removed in order to reduce false positive interactions. To investigate whether false-positive interactions do affect the clustering coefficients and dilute the difference between essential proteins and nonessential, we reanalyzed the protein-protein interaction data with various levels of possible false-positive contaminant. Based on the results of control experiments, we kept all the prey proteins or removed part of them ($n \geq 10$, $n \geq 3$, $n \geq 2$, $n \geq 1$) according to the frequency (n) in which the prey protein was found in the 16 control experiments. For each different set of data, we calculated the clustering coefficients of essential proteins and nonessential, respectively, and also the p-values of the two categories. By removing all the possible false positive interactions ($n \geq 1$), the averaged clustering coefficient of essential proteins is $\sim 20\%$ higher than nonessential with statistical significance ($p < 10^{-4}$, Table S3). As shown in Fig. S2, there seems to be a trend that discrimination between essential proteins and nonessential becomes better by removing more possible false positive interactions.

Table S3: The numbers of proteins (Node), protein-protein interactions (Edge), essential proteins (Ess.), the averaged clustering coefficients of essential proteins (Ess. CCo) and nonessential (non-ess CCo), and the corresponding p-values of protein-protein interaction data with various levels of possible false positive interactions in *E. coli*.

	Node	Edge	Ess.	Ess. CCo	Non-ess CCo	p-value*
Original	3,035	11,475	260	0.056	0.065	0.13
≥ 10	3,004	10,872	255	0.048	0.059	0.07
≥ 3	2,973	9,929	245	0.044	0.055	0.10
≥ 2	2,941	9,590	238	0.044	0.054	0.07
≥ 1	2,734	7,561	211	0.031	0.026	8.8×10^{-5}

* p-values were calculated using Wilcoxon rank sum test.

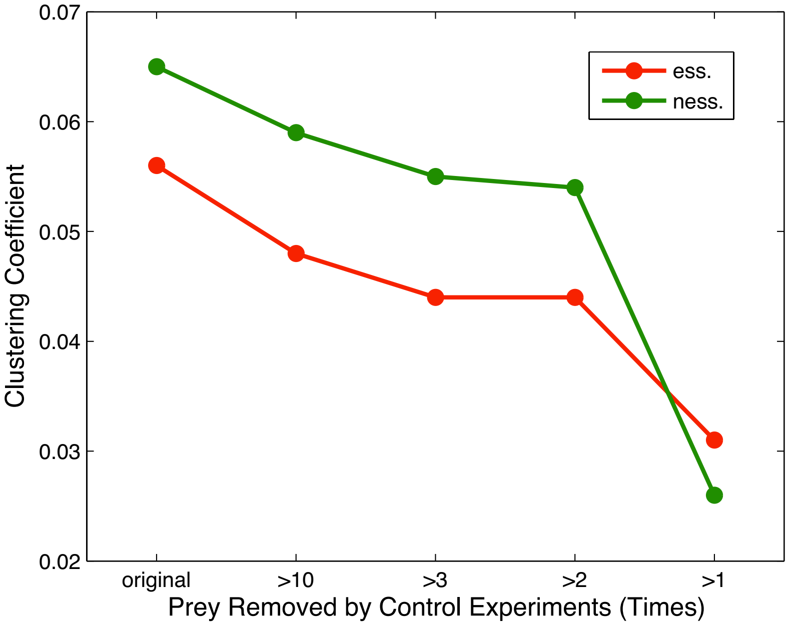

Figure S2: Averaged clustering coefficients of essential proteins and nonessential of protein-protein interaction data with various levels of possible false positive interactions in *E. coli*.

References

[1] Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, Hirai A, Tsuzuki K, Nakamura S, Altaf-Ul-Amin M, Oshima T, Baba T, Yamamoto N, Kawamura T, Ioka-Nakamichi T, Kitagawa M, Tomita M, Kanaya S, Wada C, Mori H: **Large-scale identification of protein-protein interaction of Escherichia coli K-12**. *Genome Res* 2006, **16**(5):686–91.