Finding friends and enemies in an enemies-only network

Supplementary Methods

Yan Qi Yasir Suhail Yu-yi Lin Jef D. Boeke

Joel S. Bader*

*To whom correspondence should be addressed, joel.bader@jhu.edu

September 16, 2008

Convergence

We derive the conditions for **G** to be a kernel.

$$\begin{aligned}\mathbf{G} &= [\gamma\mathbf{I}-(\mathbf{A}-\mathbf{S})]^{-1} \\ &= (\gamma\mathbf{I}-\mathbf{H})^{-1} \\ &= \gamma^{-1}(\mathbf{I}-\gamma^{-1}\mathbf{H})^{-1} \\ &= \gamma^{-1}[\mathbf{I}+\gamma^{-1}\mathbf{H}+\gamma^{-2}\mathbf{H}^2+\cdots]. \qquad (1)\end{aligned}$$

The matrix $\mathbf{H}=\mathbf{A}-\mathbf{S}$, the negative of the graph Laplacian, is symmetric and negative semi-definite [Kondor and Lafferty 200 Let $\{\lambda_i;\vec{\phi}_i\}, i=1,\cdots,n$ be the eigenvalue-eigenvector pairs of **H**. From the definitions of eigenvalues and eigenvectors and symmetric property of **H**, we have

$$\begin{aligned}\mathbf{H}\vec{\phi}_i &= \lambda_i\vec{\phi}_i, \qquad (2) \\ \vec{\phi}_i^T\mathbf{H}^T &= \vec{\phi}_i^T\mathbf{H}=\vec{\phi}_i^T\lambda_i. \qquad (3)\end{aligned}$$

Any n by 1 vector $\vec{c}$ can be written as a linear combination of the orthonormal eigenvectors of **H**,

$$\vec{c}=\sum_{i=1}^{n}a_i\vec{\phi}_i. \qquad (4)$$

Combining Eqs. (1), (2), (4), we have for any vector $\vec{c}$,

$$\begin{aligned}\vec{c}^T\mathbf{G}\vec{c} &= \sum_i a_i\vec{\phi}_i^T[\gamma^{-1}(\mathbf{I}+\gamma^{-1}\mathbf{H}+\gamma^{-2}\mathbf{H}^2+\cdots)]\sum_j a_j\vec{\phi}_j^T \\ &= \sum_i a_i^2\gamma^{-1}\vec{\phi}_i^T(I+\gamma^{-1}\mathbf{H}+\gamma^{-2}\mathbf{H}^2+\cdots)\vec{\phi}_i \\ &= \sum_i a_i^2\gamma^{-1}(1+\lambda_i/\gamma+\lambda_i^2/\gamma^2+\cdots) \qquad (5)\end{aligned}$$

In order for the RHS of Eq. (5) to converge to the LHS, a sufficient condition is $\lambda_i/\gamma<1$ for all i. When this is true, Eq. (5) reduces to

$$\vec{c}^T\mathbf{G}\vec{c}=\sum_i\frac{a_i^2}{\gamma-\lambda_i} \qquad (6)$$

From Eq. (6) it is clear that under the same condition for convergence, $\mathbf{G}$ is also positive semi-definite. Hence a sufficient conditions for $\mathbf{G}$ to converge to a kernel is that γ is larger than the maximum eigenvalue of $\mathbf{H}$. Since $\mathbf{H}$ is negative semidefinite, this implies that $\gamma > 0$ will always converge. In this application, convergence was always achieved with the allowed relaxation of $\mathbf{G}$ to a pseudo-inverse.

Kernel scores conditioned on SFL and co-complex status

Diffusion kernels for known SFL pairs from BioGRID [Stark et al. 2006] were calculated using $\gamma = 1$. SFL edges were randomly assigned to one of five cross-validation groups, and kernels were computed separately for each four-fifths of the data. For known SFL pairs, kernel scores were taken from the single calculation that excluded that edge; for all other pairs, kernel scores were averaged across the five folds. Known co-complex pairs were obtained from the MIPS protein complex catalog [Mewes et al. 2004]. Only the human-curated complexes were used, not the high-throughput complexes. For complexes with hierarchical structure, only pairs present in a sub-complex at the deepest level were taken as known PPI positives. Negative examples were pairs that were not co-complexed at any level in the hierarchy.

The histogram of $\mathbf{G}^{+}$ for known positive co-complex members is bimodal. The peak at higher kernel score is from pairs where each gene was used as a query in a high-throughput SFL screen. The lower peak is from pairs where at least one gene was not used as a query.

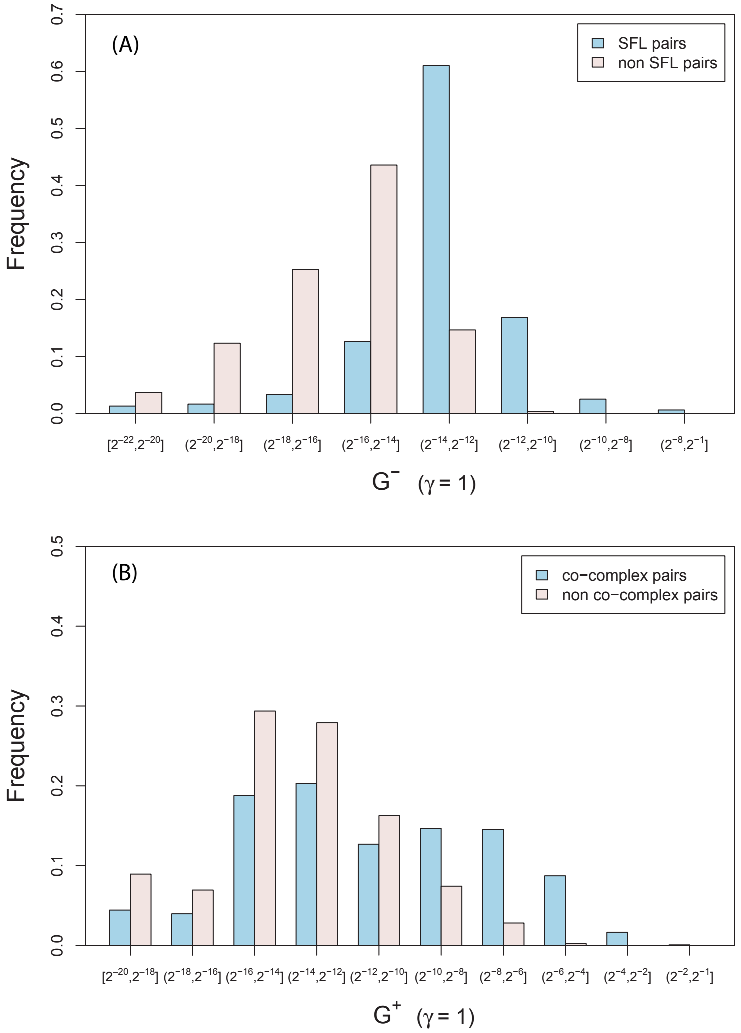

Figure 1: Diffusion kernels $\mathbf{G}^{-}$ and $\mathbf{G}^{+}$ were calculated from the SFL network with symmetric normalization and $\gamma = 1$. (A) Histograms of $\mathbf{G}^{-}$ for SFL gene pairs and non-SFL gene pairs. Scores for known SFL pairs are shifted to higher values. (B) Histograms of $\mathbf{G}^{+}$ for co-complexed and non-co-complexed gene products. Scores for known co-complexed pairs are shifted to higher values.

SFL prediction

The performance of the diffusion kernels depends on the diffusion parameter γ. Large γ corresponds to short diffusive paths while small γ allows the kernel to explore more of the network via longer paths. We evaluated the kernels under 14 different γ values ranging from 0.01 to 256 in order to find the optimal parameter for each kernel and compare the performances of the kernels. In addition to the three diffusion kernels, we also examined the performance of three methods based on the number of length-3 paths, denoted by $\mathbf{A}^3$, $N(\mathbf{A}^3)$ and $[N(\mathbf{A})]^3$ (see Main Text, Materials and Methods). The performances of the three counting methods do not depend on γ.

The performances of SFL prediction by the above mentioned six methods (three diffusion kernels and three counting methods) are assessed by the maximal F-score and the AUC of the ROC curve (Materials and Methods). Both the AUC and the F-score metrics show that the odd-parity kernel $\mathbf{G}^-$ is a better predictor for SFLs than the even-parity kernel $\mathbf{G}^+$ and the full kernel $\mathbf{G}$, with $\mathbf{G}^+$ being the worst among the three at all γ values (Supplemental Fig. 2). The AUC of $\mathbf{G}^+$ decreases abruptly when γ increases from 16 to 32. This is a numerical artifact due to machine precision: kernel scores decrease exponentially with γ, and the smallest scores in $\mathbf{G}^+$ fell below the machine precision for $\gamma \geq 32$. These scores account for the tail of the ROC curve at large false positive rate but have negligible influence on the PR curve. The $\mathbf{G}$ kernel achieves the best performance at an intermediate value of $\gamma = 1$, according to the F-score metric. In cross-validation, the direct edge between pairs to be predicted have been removed from the training data and hence the dominant term in $\mathbf{G}^-$ and $\mathbf{G}^+$ are length-3 and length-2 paths respectively. As γ increases, though the performance of $\mathbf{G}^-$ improves, $\mathbf{G}$ is increasingly dominated by the length-2 path term in $\mathbf{G}^+$, resulting in worse performance of $\mathbf{G}$.

The optimization procedure shows that $\mathbf{G}^-$ achieves its best performance on the BioGRID dataset at large γ values around $\gamma \geq 32$. At $\gamma \geq 32$, only length-3 paths contribute to the $\mathbf{G}^-$ kernel. Indeed, one of the counting method, $[N(\mathbf{A})]^3$, works as well as the optimized $\mathbf{G}^-$ kernel. This method which pre-normalize the adjacency matrix significantly out-performs the raw count $\mathbf{A}^3$, and $N[(\mathbf{A})^3]$ which post-normalizes the raw count matrix.

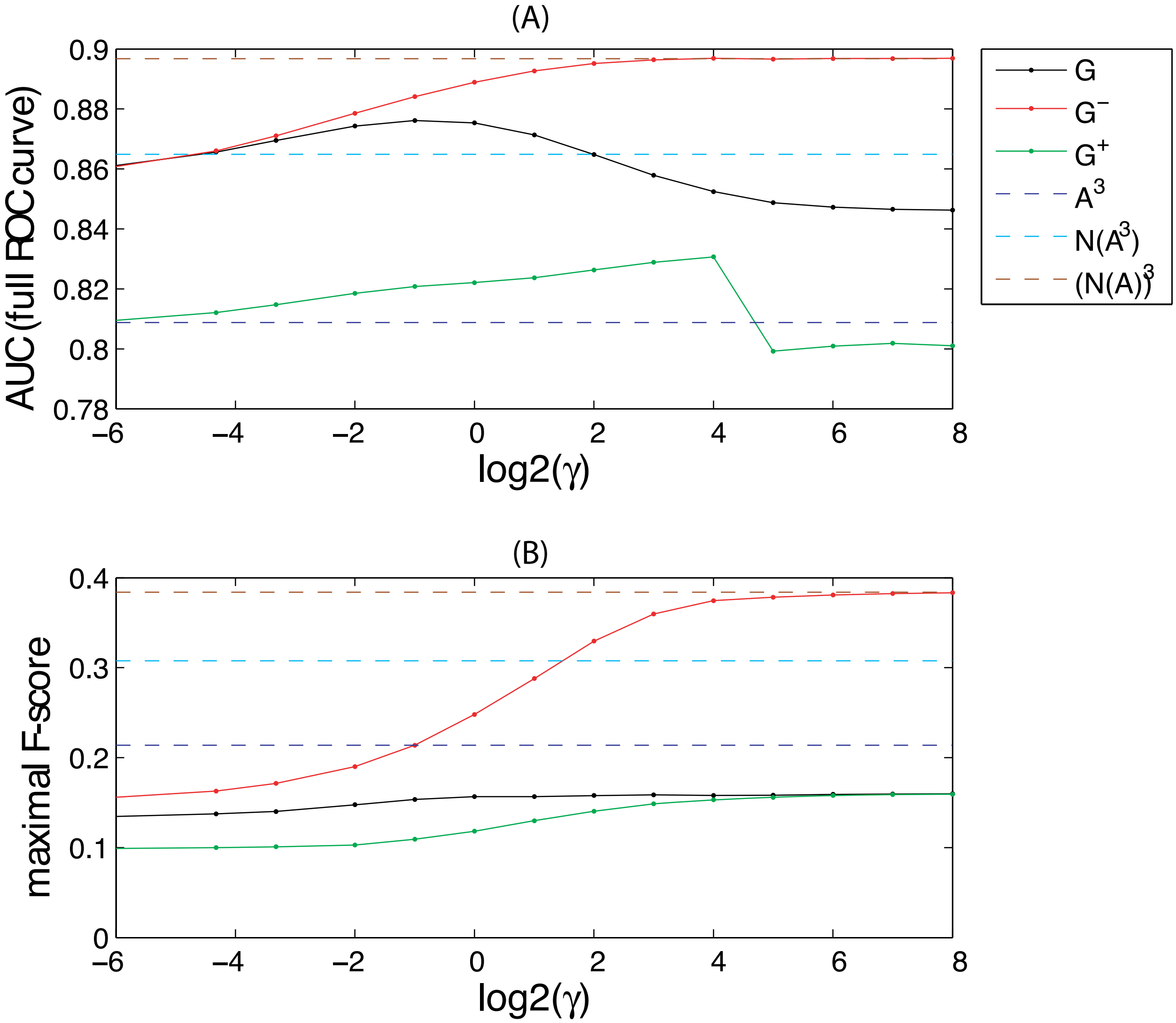

Figure 2: Parameter optimization for predicting genetic interactions from BioGRID by diffusion kernels and performance comparison with three counting methods. The $N(\cdot)$ operation represents the symmetrical normalization in Eq. (6). $\mathbf{A}^3$: raw counts of length-3 paths; $N(\mathbf{A}^3)$: symmetrically normalized $\mathbf{A}^3$; $[N(\mathbf{A})]^3$: counts of length-3 paths adjusted for node degrees. See text for details about $\mathbf{A}^3$, $N(\mathbf{A}^3)$ and $[N(\mathbf{A})]^3$. The odd-parity kernel significantly outperforms the other two diffusion kernels and has the same performance as the normalized counting method $[N(\mathbf{A})]^3$ at large γ values. (A) Area under the curve (AUC) of the full ROC curve as a function of the diffusion parameter γ. (B) Maximal F-score. F-score is defined as 2*Precision*Recall/(Precision + Recall). Maximal F-score is the maximized F-score across the entire Precision-Recall curve.

The PR curve for $\mathbf{G}^{-}$ has a distinctive shape, falling rapidly but then maintaining a plateau of about 45% precision as the recall increases. Known positives and negatives used for testing were separated into query-query and query-target pairs based on knowledge of the 179 genes used as queries in high-throughput studies. For query-query pairs, $\mathbf{G}^{-}$ and the raw path-3 count $\mathbf{A}^{3}$ perform much better than $\mathbf{G}$ or $\mathbf{G}^{+}$ (Supplemental Fig. 3). There are many more query-target pairs, however, and for these $\mathbf{G}^{-}$ performs better than any of the other predictors (Supplemental Fig. 4).

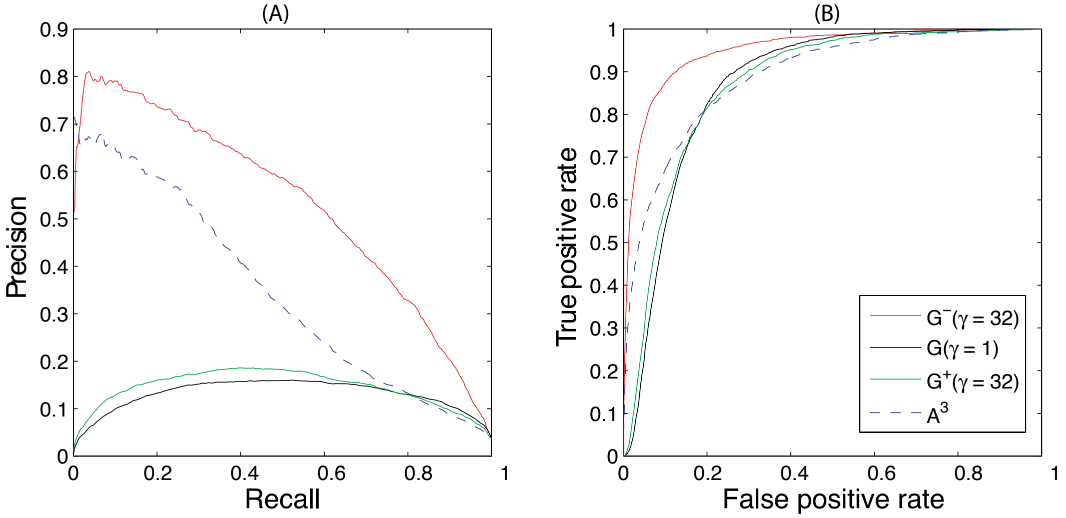

Figure 3: Performance of genetic interaction prediction where both genes in each gene pair in the test set are known query genes. The odd-parity kernel $\mathbf{G}^{-}$ and the raw count of length-3 paths $\mathbf{A}^3$ have high prediction accuracy while the diffusion kernels $\mathbf{G}^{+}$ and $\mathbf{G}$ perform poorly. (A) Precision recall curves. (B) Receiver operator characteristic curves.

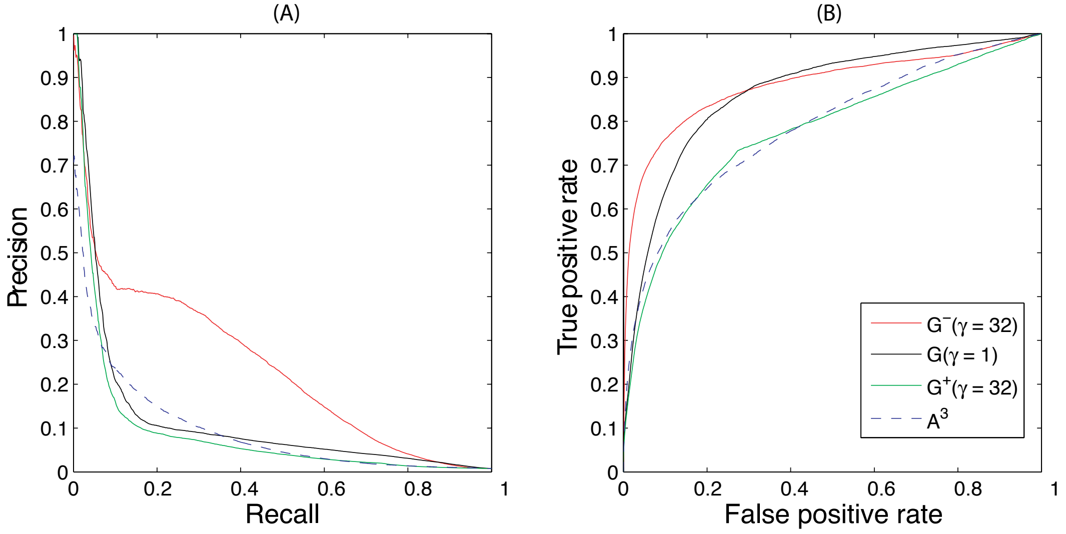

Figure 4: Performance of genetic interaction prediction where each gene pair in the test set contains one known query gene and one target gene. The odd-parity kernel $\mathbf{G}^{-}$ significantly outperforms the other three methods. (A) Precision recall curves. (B) Receiver operator characteristic curves.

Comparison with previous SFL predictions on an earlier data set

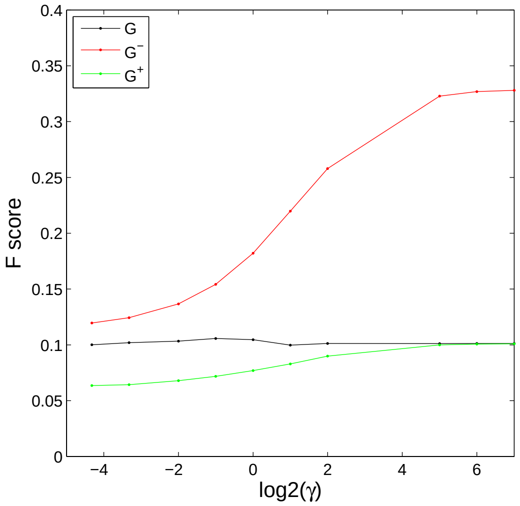

Figure 5: Performance of SSL prediction by three diffusion kernels on a smaller dataset. Five fold cross-validation is carried out for a dataset obtained from the supporting website of (Kelley and Ideker 2005). The maximal F-score is plotted as a function of γ. The $\mathbf{G}^{-}$ kernel achieves the best performance at $\gamma \geq 64$.

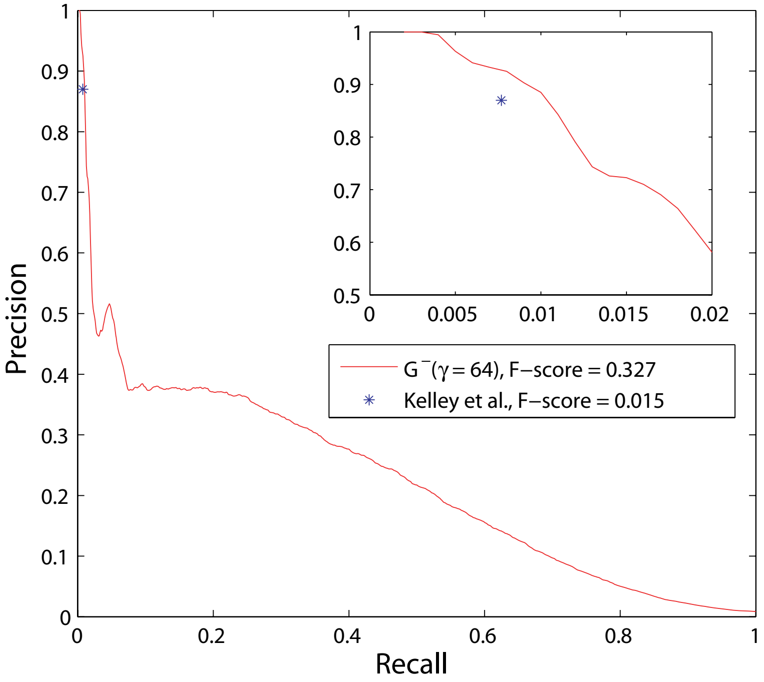

Figure 6: The performance of SFL prediction by $\mathbf{G}^{-}(\gamma = 64)$ is compared to previous results for a smaller dataset [Kelley and Ideker 2005]. Previously reported results were 87% precision (37 true positives) at 0.77% recall. At 87% precision, $\mathbf{G}^{-}$ has 1.1% recall, and at 0.77% recall it has 92% precision. The F-score for $\mathbf{G}-$ is calculated from its best overall performance of 28% precision at 40% recall.

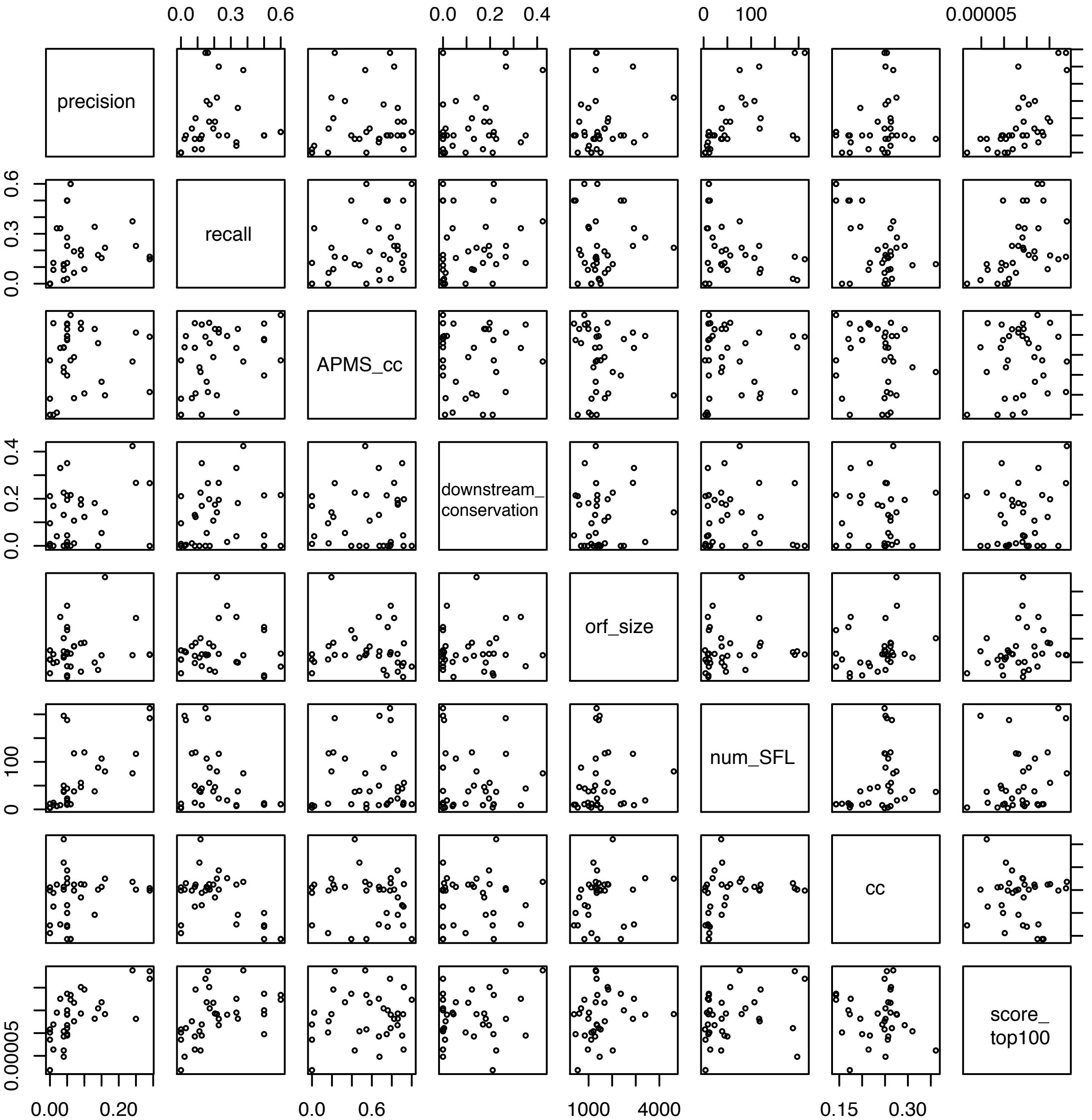

Figure 7: Scatter plot matrix of precision, recall of the top 100 predictions for each query and six biological and topological features that are significantly associated with SFL prediction quality. APMS_cc: clustering coefficient of the query protein in the protein-protein interaction network obtained by affinity purification and mass spectrometry; cc: average clustering coefficient of known SFL partner in the SFL network; score_top100: average $\mathbf{G}^{-}$ score for top 100 predictions; num_SFL: number of SFL partners from (Lin et al. 2008). See Supplementary data 2 for details of the regression model with these significant factors.

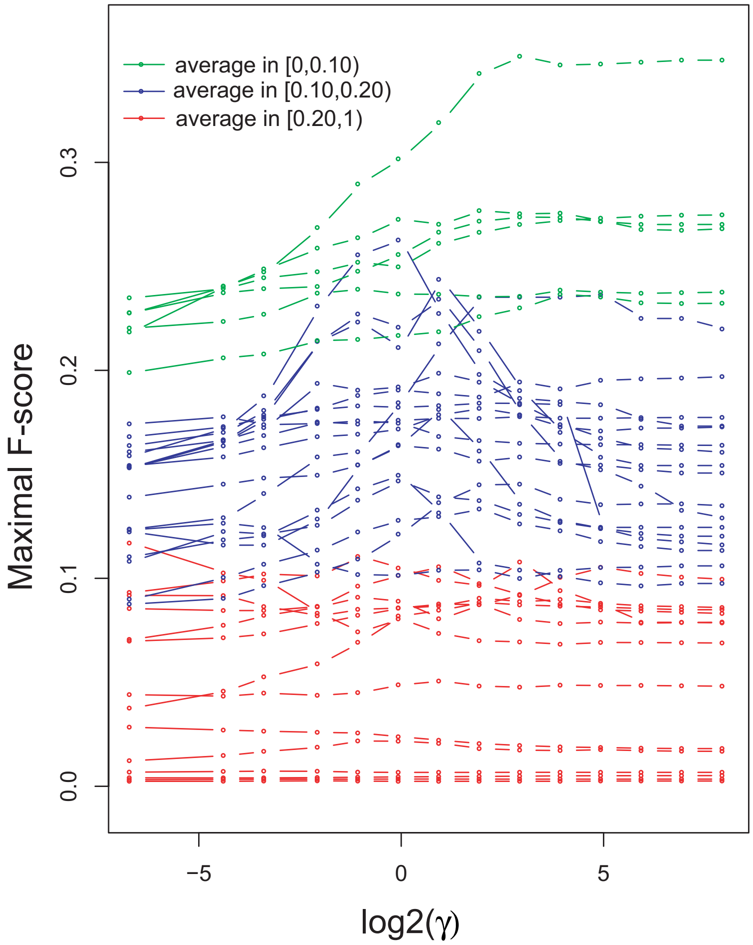

Figure 8: Performance of genome-scale SFL target prediction by the $\mathbf{G}^{-}$ kernel for 37 query genes. The maximal F-score is used as the performance metric and plotted as a function of γ. Each curve represents the prediction performance for one query. Queries are color coded according to the maximal F-score averaged across different γ values tested.

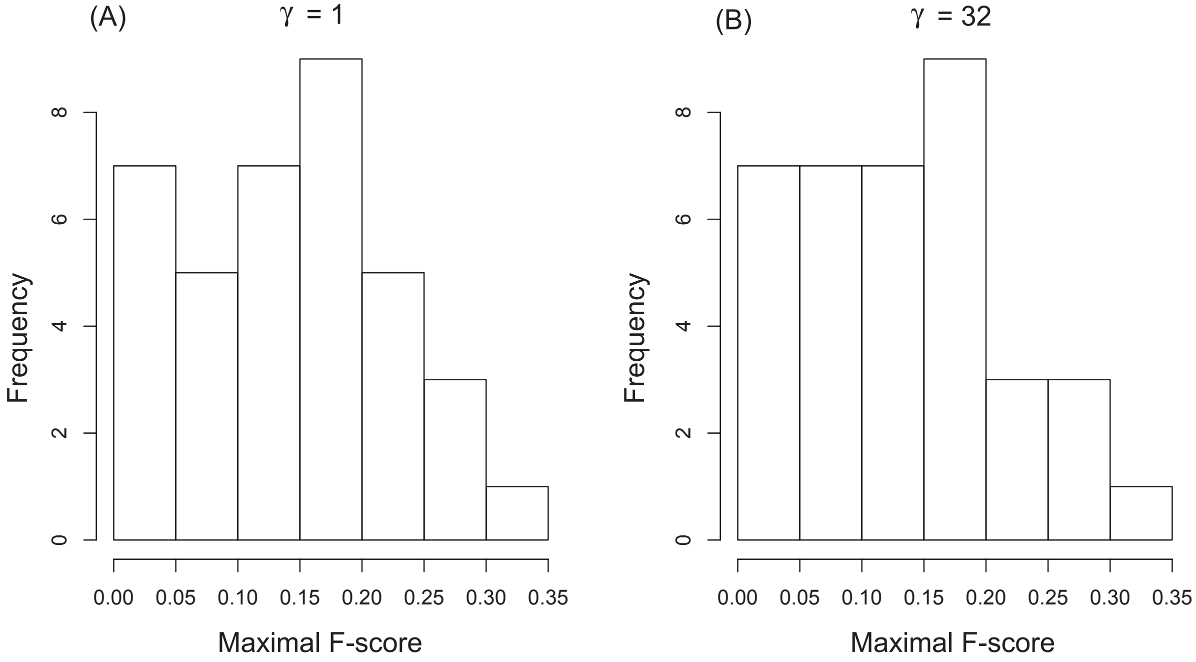

Figure 9: Histograms of the maximal F-score for 37 query genes at $\gamma = 1$ vs. $\gamma = 32$ by the $\mathbf{G}^{-}$ kernel. Eighteen queries have a maximal F-score greater than 0.15 at $\gamma = 1$ vs. sixteen at $\gamma = 32$.

Co-complex/pathway membership prediction

We optimized the diffusion parameter for the kernels in predicting co-complex membership by evaluating the kernels at 14 different γ values. We compared the performances of the diffusion kernels to those of four methods based on length-2 paths. One such method is the congruence score, which has been the best algorithm for inferring co-complex/pathway membership from only genetic interactions (Materials and Methods). The other three methods, $\mathbf{A}^2$, $N(\mathbf{A}^2)$, and $[N(\mathbf{A})]^2$, are (normalized) counts of length-2 paths (see Main Text, Materials and Methods).

According to the two performance metric, the maximal F-score and the AUC of the ROC curve, the general trend is that the $\mathbf{G}^+$ kernel is a better predictor for co-complex membership than $\mathbf{G}^-$ at all γ values examined (Supplemental Fig. 10). This is consistent with our hypothesis that an excess of even-length paths increases the likelihood of co-complex membership. There is a clear trade-off between precision and coverage. Although γ on the small extreme have larger AUC of the ROC curve, the F-score metric suggests the optimal γ values are at 0.1, 0.25 and 0.05 for $\mathbf{G}$, $\mathbf{G}^+$ and $\mathbf{G}^-$, respectively. While inferior to the $\mathbf{G}^+$ kernel, the odd-parity kernel predicts a non-negligible fraction of co-complex proteins. As a result, the best kernel overall is $\mathbf{G}$ at $\gamma = 0.1$ while the $\mathbf{G}^+$ kernel is similar or slightly better than $\mathbf{G}$ for $\gamma \geq 32$. At $\gamma = 0.25$, the even-length paths that contribute to $\mathbf{G}^+$ include those much longer than length-2.

According to the F-score metric, all three diffusion kernels are significantly better than the congruence score. The raw count of length-2 paths, $\mathbf{A}^2$ is the worst predictor among all. The congruence score is better than $\mathbf{A}^2$ but not as good as the two normalized counting methods. Of course, it is possible that the true performance of the congruence score is better than presented here, given perfect knowledge of query or target status. But the requirement of knowing query or target status can itself be a limitation, which the graph diffusion kernels do not have. Interestingly, the simple predictor $[N(\mathbf{A})]^2$ performed very well. The optimized $\mathbf{G}^+$ kernel at $\gamma = 0.25$ outperforms $[N(\mathbf{A})]^2$ in terms of the AUC of the ROC curve but only slightly better in terms of the maximal F-score. This analysis shows that the improvement of $\mathbf{G}^+$ over congruence score and $\mathbf{A}^2$ comes mainly from the normalization of the adjacency matrix which accounts for the degrees of all the nodes on the length-2 paths that connect a pair of gene i and j, while the congruence score only considers the degrees of i and j themselves and $\mathbf{A}^2$ completely ignores node degree information.

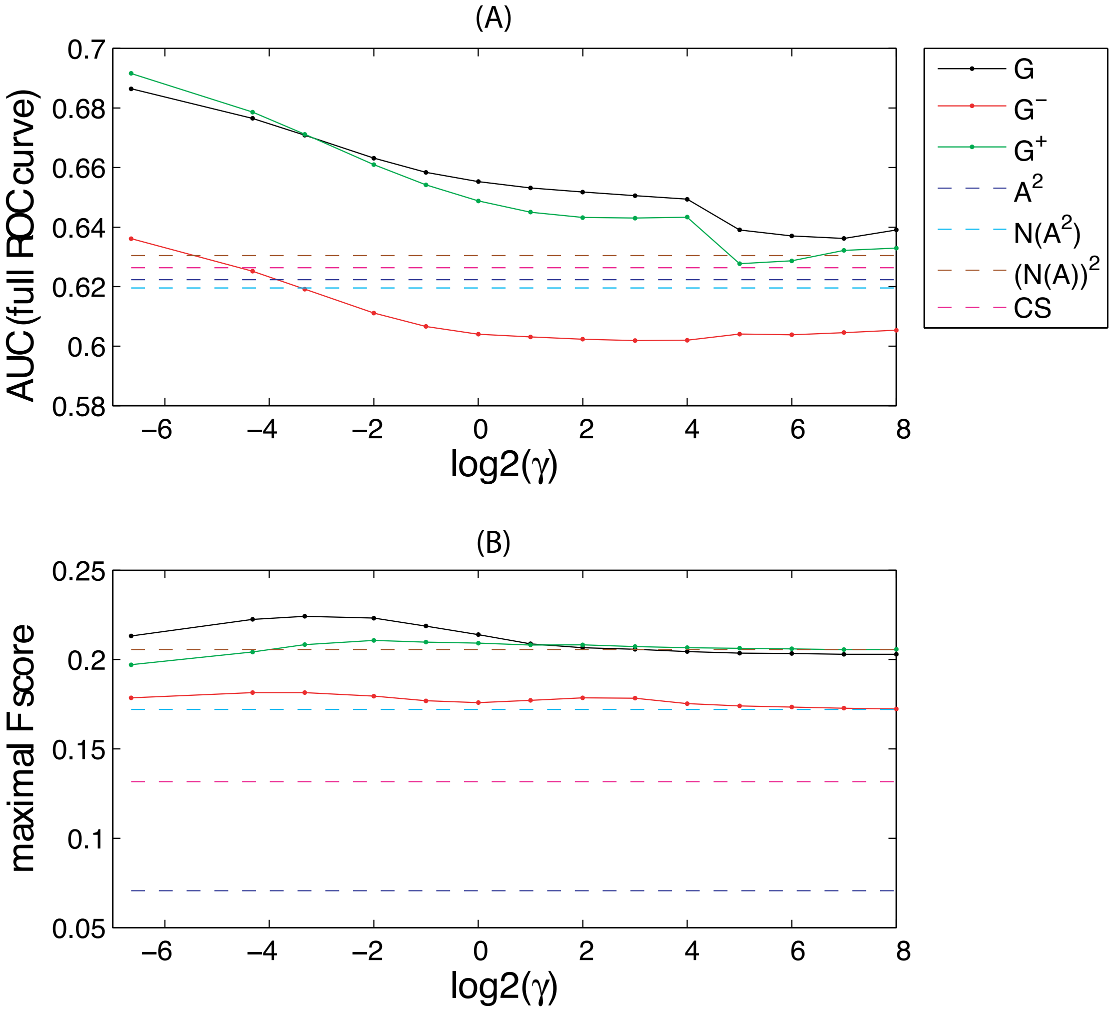

Figure 10: Parameter optimization for predicting co-complex/pathway membership by diffusion kernels and performance comparison with the congruence score and three counting methods. CS: congruence score. The $N(\cdot)$ operation represents the symmetrical normalization in Eq. (6). $\mathbf{A}^2$: raw counts of length-2 paths; $N(\mathbf{A}^2)$: symmetrically normalized $\mathbf{A}^2$; $[N(\mathbf{A})]^2$: counts of length-2 paths adjusted for node degrees. See text for details about $\mathbf{A}^2$, $N(\mathbf{A}^2)$ and $[N(\mathbf{A})]^2$. The three best methods are the full kernel $\mathbf{G}$ at $\gamma = 0.1$, the even-parity kernel $\mathbf{G}^+$ at $\gamma = 0.25$ and the counting method $[N(\mathbf{A})]^2$.(A) Area under the curve (AUC) of the full ROC curve as a function of the diffusion parameter γ. (B) Maximal F-score. F-score is defined as 2*Precision*Recall/(Precision + Recall).

Searches seeded by MIPS complexes

Protein complexes were obtained from all levels of the MIPS catalog of curated complexes. A compound query was built from the genes in complex. All genes in the SFL network, including the known members of a complex, were ranked according to the score

$$s_i = \frac{n}{n-\delta_i} \sum_{j \in \text{complex}, j \neq i} \mathbf{G}_{ij}. \tag{7}$$

The condition $j \neq i$ indicates that the self-terms $\mathbf{G}_{ii}$ are omitted when ranking known members of a complex. The term δ_i is defined as 1 for $i \in$ complex and 0 otherwise. The normalization prefactor $n/(n-\delta_i)$ corrects the score for the excluded self-term. This procedure provides a natural calibration of precision and recall with respect to known complex members.

In many cases, genes not annotated as complex members are interspersed with known members. All complexes with at least 10% precision at 80% recall were analyzed to identify why non-members were ranked higher than known members. As discussed in the main text, a common explanation was the identification of proteins that are co-complexed at higher levels in the hierarchy and in linked biological processes. A performance summary for all complex indicates good performance for complexes with at least 5 to 8 known members (Supplemental Fig. 11). Complexes with 4 or fewer known members have worse performance, with F-scores typically below 0.1.

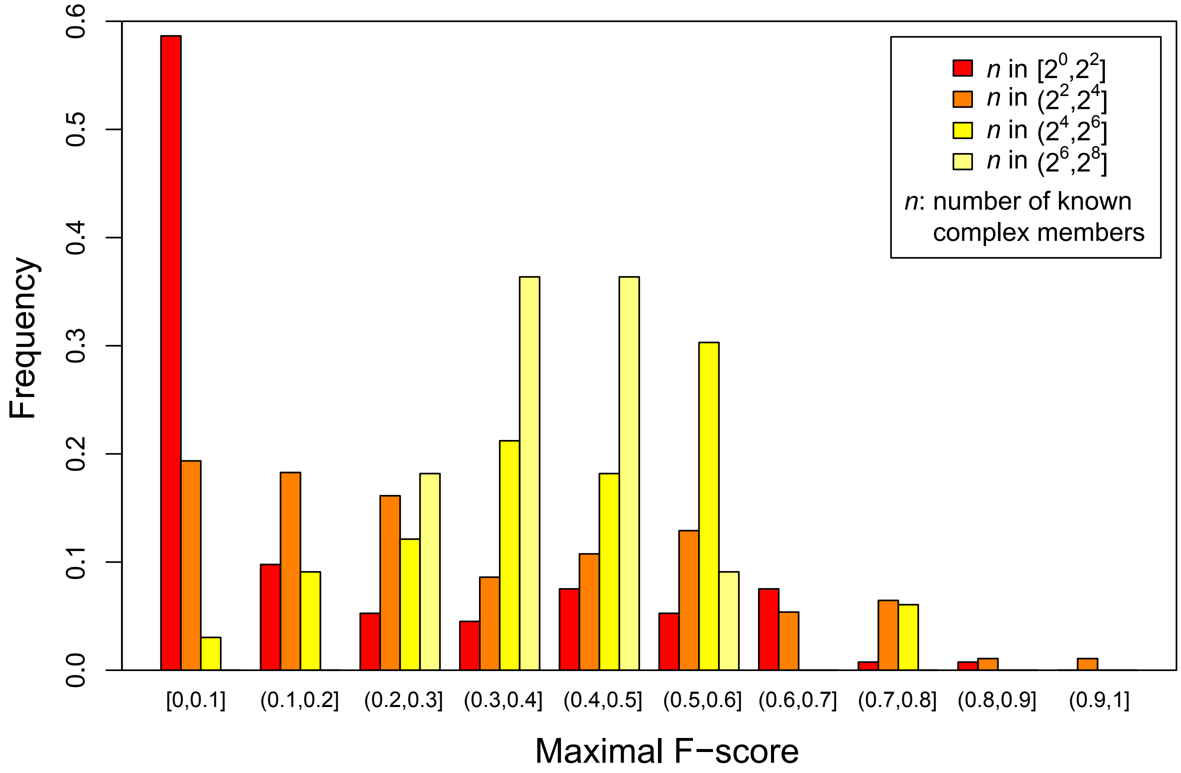

Figure 11: Performance in recovering MIPS complexes is displayed according to F-score and the number of known complex members.

The exponential kernel

While the main text describes a steady-state diffusion kernel, several other kernels have been described for graphs. Perhaps most widely used is the exponential kernel, which can be interpreted as a transient response [Kondor and Lafferty 2002]. The solution to the heat equation $(d/d\beta)\mathbf{K}_\beta = -\mathbf{L}\mathbf{K}_\beta$ is

$$\mathbf{K} = \exp(-\beta\mathbf{L}) \tag{8}$$

where the matrix exponentiation transforms local structure of the graph captured by $\mathbf{L}$ into global structure of the graph characterized by $\mathbf{K}$. When $\mathbf{L}$ is symmetric, $\mathbf{K}$ is symmetric and positive semi-definite and hence a kernel. The kernels described by Eq. (8) form an exponential family with generator $\mathbf{L}$ and bandwidth parameter β. In particular, a generator for an undirected, unweighted graph is $\mathbf{L} = \mathbf{S} - \mathbf{A}$, where $\mathbf{A}$ is the graph adjacency matrix and $\mathbf{S}$ is a diagonal matrix containing the node degrees. The bandwidth parameter β can be interpreted as the time delay before measuring the transient reponse. Small β corresponds to shallow diffusion (analogous to large γ for the steady-state kernel) while large β corresponds to long diffusion (analogous to small γ). Calculations for the exponential kernel used the unnormalized adjacency matrix.

It is possible to define a parity for the exponential kernel,

$$\begin{aligned}
\mathbf{K}^{\pm} &= \mathbf{B}^{\pm}(\beta) + \int_0^{\beta} d\beta' \mathbf{K}_2(\beta - \beta')\mathbf{K}^{\pm}(\beta') \\
\mathbf{B}^{+}(\beta) &= \exp(-\beta\mathbf{S}) \\
\mathbf{B}^{-}(\beta) &= \int_0^{\beta} d\beta' \exp[-(\beta - \beta')\mathbf{S}]\mathbf{A}\exp[-\beta'\mathbf{S}] \\
\mathbf{K}_2(\beta) &= \int_0^{\beta} d\beta' \exp[-(\beta - \beta')\mathbf{S}]\mathbf{A}\mathbf{B}^{-}(\beta').
\end{aligned} \tag{9}$$

Whereas calculating the full exponential kernel requires only repeated matrix multiplications, calculating parity-specific exponential kernels requires either quadrature over a grid in β-space or conversion of convolutions over parameter β into products of Laplace transforms, followed by inverse Laplace transform. Under some normalization schemes, the computation is easier because the diagonal matrix $\mathbf{S}$ is proportional to the identity matrix, removing the necessity of integrals over β'. We did not pursue these approaches.

Because only the full kernel was calculated, only performance in recovering PPIs was investigated. The

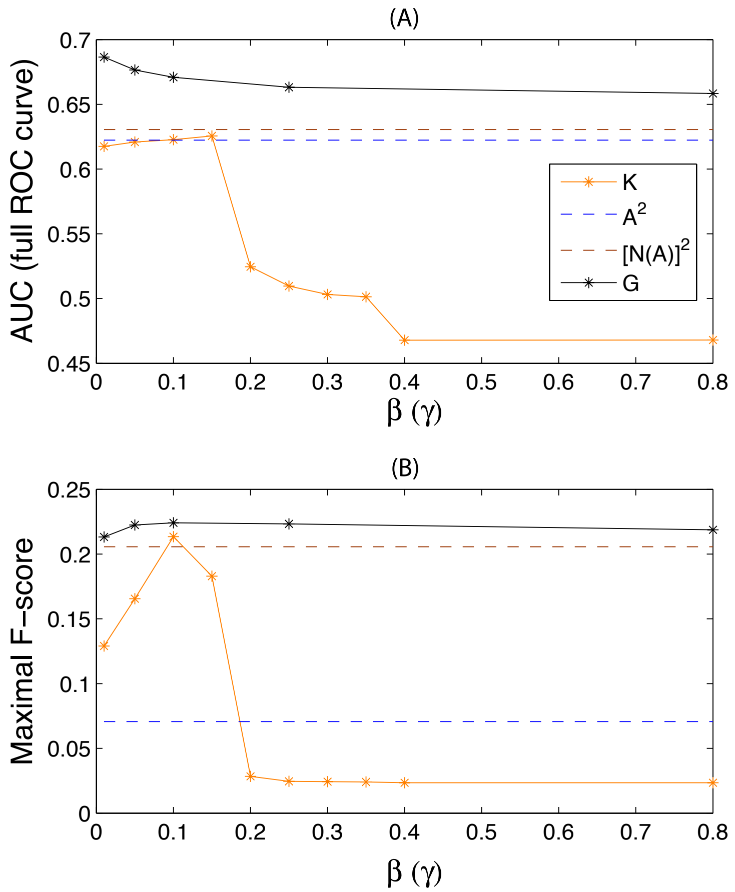

Figure 12: Performance of the exponential kernel **K** in predicting co-complex/pathway membership. The performance of the exponential kernel is plotted as a function of the bandwidth parameter β. Large β corresponds to long diffusion and small β corresponds to short diffusion. (A) AUC of the full ROC curve as a function of β. (B) Maximal F-score as a function of β. The optimal β is 0.1.

full kernel did only slightly better than the count of normalized length-2 paths (Supp. Fig. 12), and not as well as the steady-state kernel. It is possible that normalization of the adjacency matrix prior to kernel calculation would improve the performance of the exponential kernel. But as our objective is investigating parity-specific kernels, we did not pursue this point.

SVM performance

	SVM (SFL only)	SVM (no SFL)	SVM (ALL)
Precision	0.857	0.639	0.864
Recall (True pos. rate)	0.794	0.403	0.807
False pos. rate	0.132	0.228	0.127
Accuracy	0.831	0.588	0.840
F-score	0.824	0.494	0.835

Table 1: Results of SFL prediction by three SVM classifiers. Each statistic is the average across five test sets and within test sets averaged over 5-fold cross-valiation. Accuracy is defined as (# true pos. + # true neg.)/(# known pos. + # known neg.).

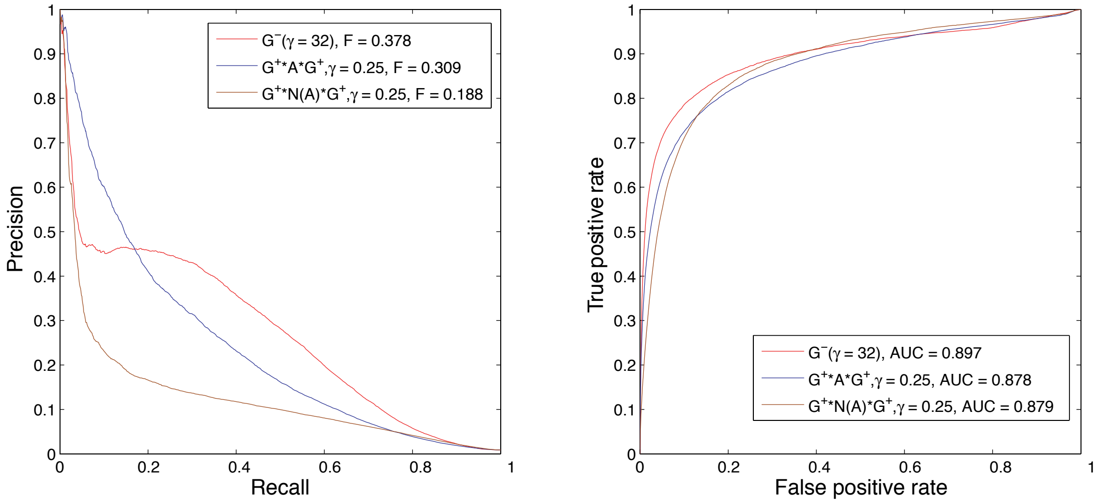

Figure 13: Performance of the three diffusion kernels $\mathbf{G}^{-}$, *GAG*1 and *GAG*2, in predicting genetic interactions. *GAG*1, termed *GAG* in the main text, is $\mathbf{G}^{+}\mathbf{A}\mathbf{G}^{+}$; $GAG2 = \mathbf{G}^{+}N(\mathbf{A})\mathbf{G}^{+}$. $\mathbf{G}^{-}$ is the best kernel overall but the *GAG*1 kernel has a superior performance at low recall region.

References

[Kelley and Ideker 2005] Kelley, R. and Ideker, T. 2005. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* **23**: 561–566.

[Kondor and Lafferty 2002] Kondor, R. and Lafferty, J. 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[Mewes et al. 2004] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., et al. 2004. Mips: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**: D41–D44.

[Stark et al. 2006] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. 2006. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res* **34**: D535–D539.