# Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network

**Manuel Middendorf[†], Etay Ziv[‡], and Chris H. Wiggins[§¶‖]**

[†]Department of Physics, [‡]College of Physicians and Surgeons, [§]Department of Applied Physics and Applied Mathematics, and [¶]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10027

Communicated by Barry H. Honig, Columbia University, New York, NY, December 20, 2004 (received for review September 7, 2004)

**Naturally occurring networks exhibit quantitative features revealing underlying growth mechanisms. Numerous network mechanisms have recently been proposed to reproduce specific properties such as degree distributions or clustering coefficients. We present a method for inferring the mechanism most accurately capturing a given network topology, exploiting discriminative tools from machine learning. The *Drosophila melanogaster* protein network is confidently and robustly (to noise and training data subsampling) classified as a duplication–mutation–complementation network over preferential attachment, small-world, and a duplication–mutation mechanism without complementation. Systematic classification, rather than statistical study of specific properties, provides a discriminative approach to understand the design of complex networks.**

machine learning | systems biology | motifs | classification | evolution

Recent research activity in biological networks has often focused on understanding the emergence of specific features such as scale-free degree distributions (1–3), short mean geodesic lengths, or clustering coefficients (4). The insights gained into the topological patterns have motivated various network growth and evolution models to determine what simple mechanisms can reproduce the features observed. Among these are the preferential attachment model (3, 5), exhibiting scale-free degree distributions, and the small-world model (4), exhibiting high clustering coefficients despite short mean geodesics. Additionally, various duplication–mutation mechanisms have been proposed to describe biological networks (6–11) and the World Wide Web (12). However, in most cases model parameters can be tuned such that multiple models of widely varying mechanisms perfectly fit the motivating real network in terms of single selected features such as the scale-free exponent and the clustering coefficient (compare Fig. 1). Because networks with several thousands of vertices and edges are highly complex, it is also clear that these statistics can capture only limited structural information.

Here, we make use of *discriminative classification* techniques recently developed in machine learning (13, 14) to classify a given real network as one of many proposed network mechanisms by enumerating local substructures. Determining what simple mechanism is responsible for a natural network's architecture (*i*) facilitates the development of correct priors for constraining network inference and reverse engineering (15–18); (*ii*) specifies the appropriate null model relative to which one evaluates statistical significance (19–29); (*iii*) guides the development of improved network models; and (*iv*) reveals underlying design principles of evolved biological networks. It is therefore desirable to develop a method to determine which proposed mechanism models a given complex network without prior selection of features or null models.

Enumeration of subgraphs has been successfully used in the past few years to find network motifs (19, 20, 23–29) and is historically a well established method in the sociology community (30–32). Recently, the idea of clustering real networks based on their "significance profiles" has been proposed (33). The method assesses significance of given subgraphs relative to an assumed null model, generated by Monte Carlo sampling of networks with a degree distribution identical to that of the network of interest. The significance profiles are then shown to be similar for various groups of naturally occurring networks.

Both clustering and assessing statistically significant motifs can be characterized as schemes to identify reduced-complexity descriptions of the networks. We here present an approach that is instead *predictive,* using labeled graphs of known growth mechanisms as training data for a discriminative classifier. This classifier, then, presented with a new graph of interest, can reliably and robustly predict the growth mechanism that gave rise to that graph. Within the machine learning community, such predictive, *supervised learning*, techniques are differentiated from descriptive, *unsupervised learning*, techniques such as clustering.

We apply our method to the recently published *Drosophila melanogaster* protein–protein interaction network (34) and find that a duplication–mutation–complementation (DMC) mechanism (6) best reproduces *Drosophila*'s network. The prediction is robust against noise, even after random rewiring of up to 45% of the network edges. To validate, we also show that beyond 80% random rewiring the correct (Erdös–Rényi) classification is obtained.

## Methods

**The Data Set.** We use a protein–protein interaction map based on yeast two-hybrid screening (34). Because the data are subject to numerous false positives, Giot *et al.* (34) assign a confidence score $P \in [0, 1]$, measuring how likely the interaction occurs *in vivo*. To exclude unlikely interactions and focus on a core network that retains significant global features, we determine a confidence threshold $p^*$ based on percolation: measurements of the size of the components for all possible values of $p^*$ show that the two largest components are connected for $p^* = 0.65$ (see the supporting information, which is published on the PNAS web site). Edges in the graph correspond to interactions for which $p > p^*$. To reveal possible structural changes in *Drosophila* for less stringent thresholds, we also present results for $p^* = 0.5$ as suggested in ref. 34. We remove self-interactions from the network because none of the proposed mechanisms allow for them. After eliminating isolated vertices the resulting networks consist of 3,359 (4,625) vertices and 2,795 (4,683) edges for $p^* = 0.65$ (0.5).

**Network Mechanisms.** We generate 7,000 graphs, 1,000 for each of seven different models drawn from the literature, as training

Abbreviations: DMC, duplication–mutation–complementation; DMR, duplication–mutation using random mutations; LPA, linear preferential attachment; RDS, random static; RDG, random growing; AGV, aging vertex; SMW, small-world; ADT, alternating decision tree.

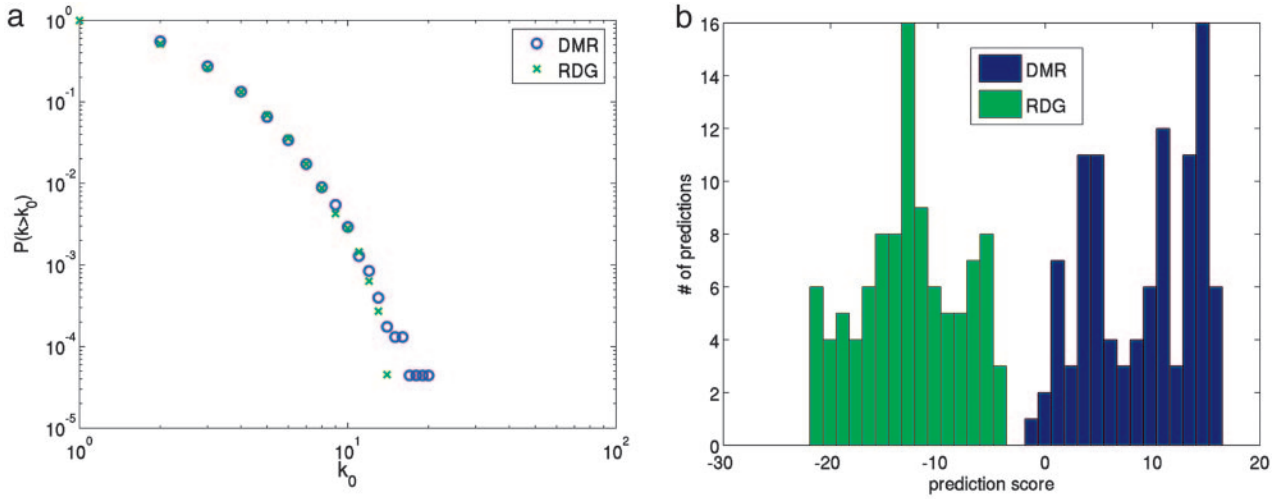
See Commentary on page 3173.

‖To whom correspondence should be addressed at: Department of Applied Physics and Applied Mathematics, Columbia University, 500 West 120th Street, New York, NY 10027. E-mail: chris.wiggins@columbia.edu.





www.pnas.org/cgi/doi/10.1073/pnas.0409515102




| | DMR | RDG |
|---|---|---|
| $\langle C \rangle$ | $2.6\ 10^{-4} \pm 1.3\ 10^{-4}$ | $5.4\ 10^{-4} \pm 3.7\ 10^{-4}$ |
| $\langle \ell \rangle$ | $10.4 \pm 0.1$ | $9.6 \pm 0.04$ |

**Fig. 1.** Discriminating similar networks. Ten graphs of two different mechanisms exhibit similar average geodesic lengths and almost identical degree distribution and clustering coefficients. (*a*) Cumulative degree distribution $p(k > k_0)$, average clustering coefficient $\langle C \rangle$ and average geodesic length $\langle \ell \rangle$, all quantities averaged over a set of 10 graphs. (*b*) Prediction scores for all 10 graphs and all five cross-validated (13) ADTs. The two sets of graphs can be perfectly separated by our classifier, even though none of these graphs is used in the classifier training.



data. Every graph is generated with the same number of edges and number of vertices as measured in *Drosophila*; all other existing parameters are sampled uniformly (see supporting information). The models, many of which were explicitly intended to model protein interaction networks, manifest various simple network growth mechanisms. As an example, the DMC algorithm (6) is inspired by an evolutionary model of the genome (35, 36) proposing that most of the duplicate genes observed today have been preserved by functional complementation. If either copy of the gene loses one of its functions (edges), the other becomes essential in ensuring the organism's survival. There is thus an increased preservation of duplicate genes induced by null mutations. The algorithm features a duplication step followed by mutations that preserve functional complementarity. At every iteration one chooses a vertex $v$ at random. A twin vertex $v_{\text{twin}}$ is then introduced, copying all of $v$'s edges. For each edge of $v$, one deletes with probability $q_{\text{del}}$ either the original edge or its corresponding edge of $v_{\text{twin}}$. The twins themselves are conjoined with an independent probability $q_{\text{con}}$, representing an interaction of a protein with its own copy. Note that no new edges are created by mutations. The DMC mechanism thus assumes that the probability of creating new advantageous functions by random mutations is negligible.

A slightly different implementation of duplication–mutation is realized in ref. 7 by using random mutations (DMR). Possible interactions between twins are neglected. Instead, edges between $v_{\text{twin}}$ and the neighbors of $v$ can be removed with a probability $q_{\text{del}}$ and new edges can be created at random between $v_{\text{twin}}$ and any other vertices with a probability $q_{\text{new}}/N$, where $N$ is the current total number of vertices. DMR thus emphasizes the creation of new advantageous functions by mutation.

In addition to (*i*) DMC and (*ii*) DMR, we generate training data for (*iii*) linear preferential attachment (LPA) networks (3, 5) (growing graphs with a probability of attaching new vertices to existing vertices proportional to $k + a$, $a$ being a constant parameter and $k$ being the degree of the existing vertex); (*iv*) random static (RDS) networks (37) (also known as Erdös–Rényi graphs; vertices are connected randomly); (*v*) random growing (RDG) networks (38) (growing graphs where new edges are created randomly between existing vertices); (*vi*) aging vertex (AGV) networks (39) (growing graphs modeling citation networks, where the probability for new edges decreases with the age of the vertex); and (*vii*) small-world (SMW) networks (4) (an interpolation between regular ring lattices and randomly connected graphs). For descriptions of the specific algorithms we refer the reader to the supporting information.

**Subgraph Census.** We quantify the topology of a network by exhaustive subgraph census (31) up to a given subgraph size; note that we do *not* assume a specific network randomization or test for statistical significance as in refs. 19, 20, 23–29, 31, and 32, but we instead *classify* network mechanisms by using the raw subgraph counts. Rather than choosing most important features *a priori*, we count all possible subgraphs up to a given cut-off, which can be made in the number of vertices, number of edges, or the length of a given walk. To show robustness to this choice, we present results for two different cut-offs. We first count all subgraphs that can be constructed by a walk of length eight (148 nonisomorphic[††] subgraphs); second, we consider all subgraphs up to a total number of seven edges (130 nonisomorphic subgraphs). Their counts are the input features for our classifier. It is worth noting that the mean geodesic length (average shortest path between two vertices) of the *Drosophila* network's giant component is 11.6 (9.4) for $p^* = 0.65$ (0.5). Walks of length eight are therefore able to traverse large parts of the network and can also reveal global structures.

**Learning Algorithm.** Our classifier is a generalized decision tree called an *alternating decision tree* (ADT) (40) by using the Adaboost (41) algorithm, which is related to additive logistic regression (42). Adaboost is a general discriminative learning algorithm proposed in 1997 by Freund and Schapire (41, 43) and has since been successfully used in numerous and varied applications [e.g., in text categorization (44, 45) and gene expression prediction (46)].

[††]Two graphs are isomorphic if there exists a relabeling of their vertices such that the two graphs are identical.

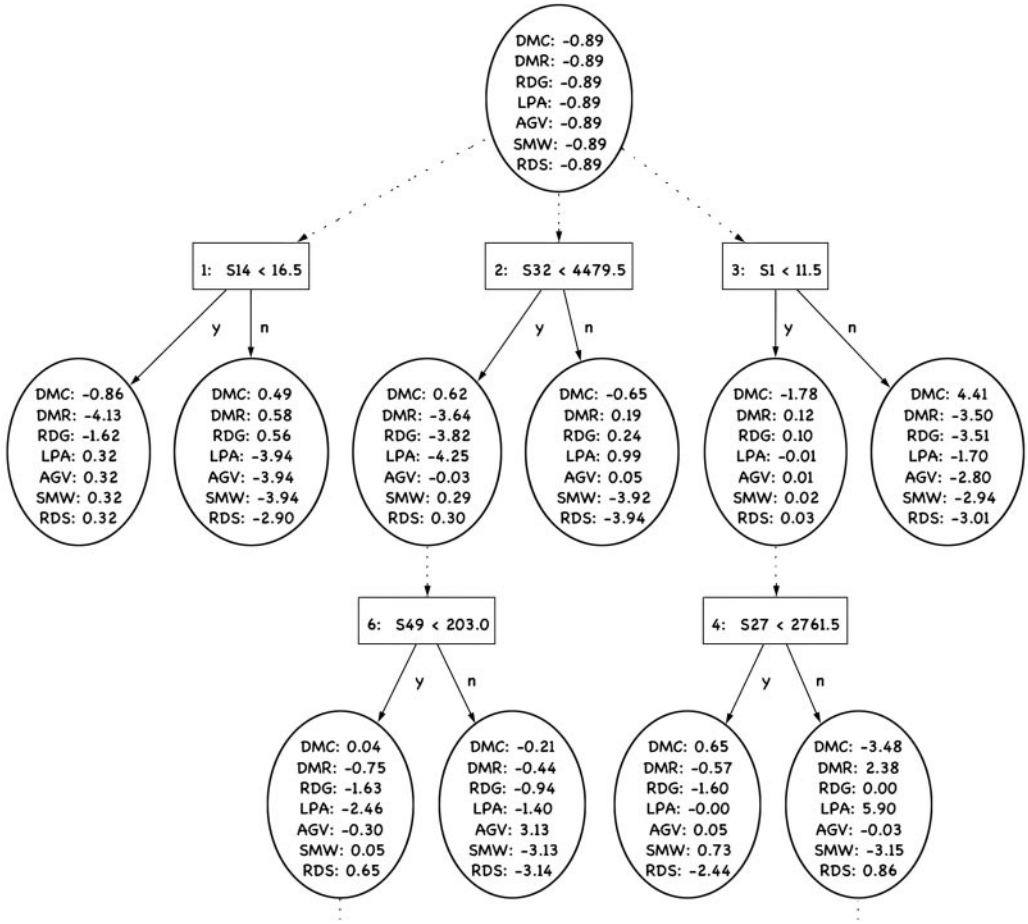



**Fig. 2.** ADT: The first few nodes of one of the trained ADTs are shown. At each boosting iteration one new decision node (rectangle) with its two prediction nodes (ovals) is introduced. Every test network follows multiple paths in the tree, dictated by the inequalities in the decision nodes (S# refers to a specific subgraph count; see Fig. 3). The final score is the sum of all prediction scores over all paths, and the class with the highest prediction score wins.

An example of an ADT is shown in Fig. 2. A given network's subgraph counts determine paths in the ADT dictated by inequalities specified by the *decision nodes* (rectangles) (subgraphs associated with Fig. 2 are shown in Fig. 3). For each class, the ADT outputs a real-valued *prediction score*, which is the sum of all weights over all paths. The class with the highest score wins. The prediction score $y(c)$ for class $c$ is related to the probability $p(c)$ for the tested network to be in class $c$ by $p(c) = e^{2y(c)}/(1 + e^{2y(c)})$ (42). (The supporting information gives additional details on the exact learning algorithm. Source code is available from C.H.W. on request.)

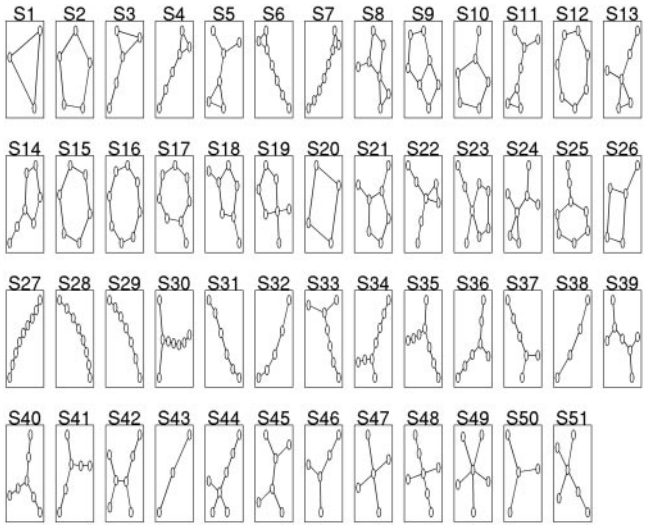



**Fig. 3.** Subgraphs associated with Figs. 2 and 4. Shown is the subset of 51 subgraphs (of 148) that appear in the learned ADT.

An advantage of ADTs is that they do not assume a specific geometry of the input space; that is, features are not coordinates in a metric space (as in support vector machines or $k$-nearest-neighbors classifiers), and the classification is thus independent of normalization. The algorithm assumes neither independence nor dependence among subgraph counts. The subgraphs reveal their importance themselves solely by their abilities to discriminate among different classes.

## Results

We perform cross-validation (ref. 13 and supporting information) with multiclass ADTs, thus determining an empirical estimate of the generalization error, i.e., the probability of mislabeling an unseen test datum. Table 1 relates truth and prediction for the test sets. Five of seven classes have nearly perfect prediction accuracy. Because AGV is constructed to be an interpolation between LPA and a ring lattice, the AGV, LPA, and SMW mechanisms are equivalent in specific parameter regimes and correspondingly show a nonnegligible overlap. Nevertheless, the overall prediction accuracy on the test sets still lies between 94.6% and 95.8% for different choices of $p^*$ and

**Table 1. Prediction accuracy (%) for tested networks using fivefold cross-validation (13)**

| | Prediction | | | | | | |
|---|---|---|---|---|---|---|---|
| Truth | DMR | DMC | AGV | LPA | SMW | RDS | RDG |
| DMR | 99.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.6 |
| DMC | 0.0 | 99.7 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 |
| AGV | 0.0 | 0.1 | 84.7 | 13.5 | 1.2 | 0.5 | 0.0 |
| LPA | 0.0 | 0.0 | 10.3 | 89.6 | 0.0 | 0.0 | 0.1 |
| SMW | 0.0 | 0.0 | 0.6 | 0.0 | 99.0 | 0.4 | 0.0 |
| RDS | 0.0 | 0.0 | 0.2 | 0.0 | 0.8 | 99.0 | 0.0 |
| RDG | 0.9 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 99.0 |

The (*i, j*) entry is the probability of predicting class *j* given that the true class is *i*. The training data are based on the size of the *Drosophila* protein network with a confidence threshold of $p^* = 0.5$, the input features of the classifier being counts of all possible walks of length eight. The overall prediction accuracy is 95.8%. Prediction errors among AGV, LPA, and SMW networks are due to equivalence of the models in specific parameter regimes.

subgraph size cut-off. Note that preferential attachment is completely distinguishable from duplication–mutation despite the fact that a duplication mechanism is sometimes described as an *effective* preferential attachment (ref. 47 and supporting information). Even models that are based on the same fundamental mechanism, such as duplication–mutation in DMC and DMR, are perfectly separable. Even small algorithmic changes in network mechanisms can thus give rise to easily detectable differences in substructures. Our results (see Fig. 1) confirm that although many of these models have similar degree distributions, clustering coefficients, or mean geodesic lengths, they have indeed distinguishable topologies.

Fig. 2 shows the first few decision nodes of a resulting ADT. The prediction scores reveal that a high count of 3-cycles suggests a DMC network (node 3). The DMC mechanism indeed facilitates the creation of many 3-cycles by allowing two copies to attach to each other, thus creating 3-cycles with their common neighbors. In particular a few combinations are good predictors for some classes. For example, a low count in 3-cycles combined with a high count in 8-edge linear chains is a good predictor for LPA and DMR networks (nodes 3 and 4). Because of the sparseness of the networks preferential attachment does not lead to a clustered structure. While LPA readily yields hubs, cycles are less probable. (Larger ADTs can be viewed in the supporting information.)

Having built a classifier enjoying good prediction accuracy, we can now determine the network mechanism that best reproduces the *Drosophila* protein network (or in principle any network of the same size) by using the trained ADTs for classification. Table 2 gives the prediction scores of the *Drosophila* network for each of the seven classes, averaged over folds.

The DMC mechanism is the only class having a positive prediction score in every case. In particular, for $p^* = 0.65$ the DMC classification has a high score of $8.2 \pm 1.0$ for eight-step subgraphs and $8.6 \pm 1.1$ for subgraphs with up to seven edges. Also, the comparatively small standard deviations over different folds indicate robustness of the classification against data subsampling. While the high rankings of both duplication–mutation classes confirm our biological understanding of protein network evolution, our findings strongly support an evolution restricted by functional complementarity over an evolution that creates and deletes functions at random.

Notably, for $p^* = 0.65$ the RDG mechanism of random growth (edges are connected randomly between existing vertices) has a higher prediction score than the LPA or AGV growing graph mechanisms. Growth without any underlying mechanism other than chance therefore generates networks closer in topology to the core network ($p^* = 0.65$) of *Drosophila* than growth governed by preferential attachment. We also emphasize that even though *Drosophila* exhibits the SMW *character* of high clustering and short mean geodesic length (34), the SMW *model* (4) (an interpolation between regular ring lattices and randomly connected graphs) does not accurately reproduce the *Drosophila* network. The classification for $p^* = 0.5$ is less confident, probably because of the additional noise present in the data when including low $p$ value (improbable) interactions, as we discuss below.

Although not necessary for the classification itself, visualizing the distribution for each model and each subgraph, compared with that subgraph's census in *Drosophila,* can give a qualitative and more intuitive way of interpreting the classification result and a better understanding of the topological differences between *Drosophila* and each of the seven mechanisms. To this end we determine *rank scores* for every subgraph and mechanism, defined as the percentages of sampled networks that have a subgraph count above *Drosophila*'s count. A rank score of 50% corresponds to a distribution whose median is equal to *Drosophila*'s subgraph count. Fig. 4 shows the color-coded rank scores for every mechanism and every subgraph (only the subset of 51 subgraphs, which appear in the learned ADT, is shown here; see the supporting information for the full set). The subgraphs are ordered by similarity in rank scores (see caption of Fig. 4). A few subgraphs (S36–S51) featuring hubs without cycles are best modeled by the LPA mechanism; i.e., these subgraphs have rank scores close to 50%. For almost all other subgraphs, both duplication–mutation mechanisms (DMC and DMR) consistently have better rank scores than the other models. Notably, the SMW and RDS mechanisms have rank

**Table 2. Prediction scores for the *Drosophila* protein network for different confidence thresholds $p^*$ and different cut-offs in subgraph size**

| | Eight-step subgraphs ($p^* = 0.65$) | | Subgraphs with up to seven edges ($p^* = 0.65$) | | Eight-step subgraphs ($p^* = 0.5$) | |
|---|---|---|---|---|---|---|
| Rank | Class | Score | Class | Score | Class | Score |
| 1 | DMC | $8.2 \pm 1.0$ | DMC | $8.6 \pm 1.1$ | DMC | $0.8 \pm 2.9$ |
| 2 | DMR | $-6.8 \pm 0.9$ | DMR | $-6.1 \pm 1.7$ | DMR | $-2.1 \pm 2.0$ |
| 3 | RDG | $-9.5 \pm 2.3$ | RDG | $-9.3 \pm 1.6$ | AGV | $-3.1 \pm 2.2$ |
| 4 | AGV | $-10.6 \pm 4.2$ | AGV | $-11.5 \pm 4.1$ | LPA | $-10.1 \pm 3.1$ |
| 5 | LPA | $-16.5 \pm 3.4$ | LPA | $-14.3 \pm 3.2$ | SMW | $-20.6 \pm 1.9$ |
| 6 | SMW | $-18.9 \pm 0.7$ | SMW | $-18.3 \pm 1.9$ | RDS | $-22.3 \pm 1.7$ |
| 7 | RDS | $-19.1 \pm 2.3$ | RDS | $-19.9 \pm 1.5$ | RDG | $-22.5 \pm 4.7$ |

*Drosophila* is consistently (independently of the cut-off in subgraph size) classified as a DMC network, with an especially strong prediction for a confidence threshold of $p^* = 0.65$.

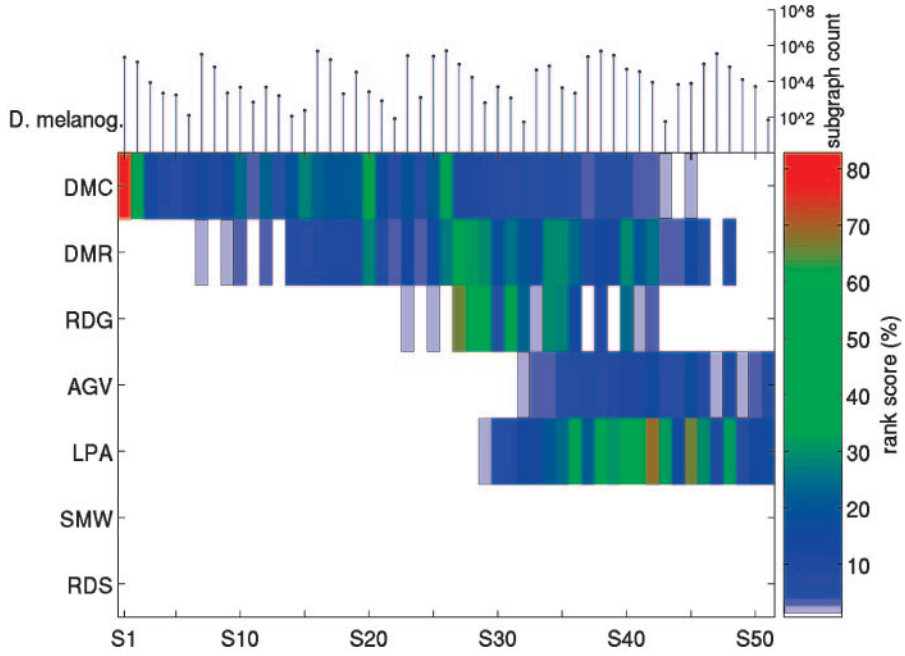



**Fig. 4.** Topological similarities and differences between *Drosophila* and each of the seven mechanisms. Color-coded rank scores are shown for a representative set of 51 subgraphs and every mechanism. The rank score $r_{i\alpha}$ for model $i$ and subgraph $\alpha$ is defined as the percentage of sampled networks having a subgraph count above *Drosophila*'s value. The matrix of correlation coefficients $\rho_{\alpha\beta}$ of rank scores is then a similarity matrix between subgraphs. The coordinates of the eigenvector corresponding to the smallest nonzero eigenvalue [or Fiedler vector (50)] of the Laplacian (51) $L_{\alpha\beta} = \rho_{\alpha\beta} - \delta_{\alpha\beta}\Sigma_\gamma \rho_{\gamma\beta}$ (where $\delta_{\alpha\beta}$ is the Kronecker symbol, equal to 1 iff $\alpha = \beta$ can then be used to sort the subgraphs according to similarity in rank scores (for details see the supporting information). A rank score of 50% (green) corresponds to a distribution whose median is equal to *Drosophila*'s subgraph count. The labels S1–S51 refer to Fig. 3. The histogram in the upper part of the figure shows *Drosophila*'s subgraph counts.

scores 0 for all subgraphs; i.e., all sampled networks have lower subgraph counts than *Drosophila*. For a few subgraphs that feature long linear chains (S27–S33), the DMR model has better rank scores than DMC, whereas for almost all other subgraphs DMC has the best rank scores. In particular, DMC is the only model that can reach *Drosphila*'s counts for subgraphs S1–S26, which show complex cyclic structure.

Because yeast two-hybrid data are known to be susceptible to numerous errors (34), network analyses are reliable only if they are robust against noise. To confirm that our method shows this robustness, we classify the *Drosophila* network for various levels of artificially introduced noise by replacing existing edges with edges chosen at random. Fig. 5 shows the prediction scores for all seven classes as functions of the fraction of edges replaced. As validation, the network is correctly and confidently ($p$ value $> 1 - 10^{-3}$) classified as an RDS graph when >80% of the edges are randomized. About 30% of *Drosophila*'s edges can be replaced without seeing any significant change in all seven prediction scores, and 40% can be replaced before *Drosophila* is no longer classified confidently as a DMC network. At this point the prediction scores of DMC, DMR, and AGV are very close, which is also observed for the prediction scores for $p^* = 0.5$ (see Table 2), where they rank top three in this order. The results therefore suggest that the less confident classification for $p^* = 0.5$ could be mainly due to the presence of more noise in the data after inclusion of low confidence-value edges.

We have presented a method to infer growth mechanisms for naturally occurring networks. Advantageous properties include robustness against both noise and data subsampling, and the absence of any prior assumptions about which network features are important. Moreover, because the learning algorithm does not assume any relationships among features, the input space can be generalized to include any additional statistics as potentially discriminative features. We find that the *Drosophila* protein interaction network is confidently classified as a DMC network, a result that strongly supports ideas presented by Vazquez *et al.* (6) and Force *et al.* (36) about the nature of genetic evolution,

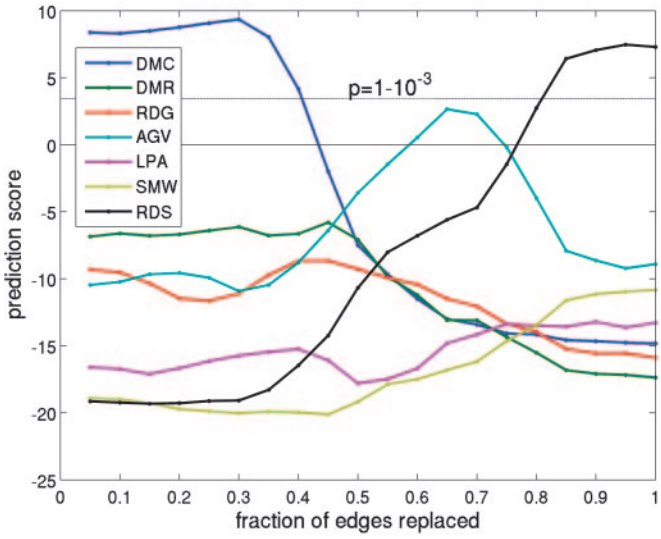



**Fig. 5.** Robustness against noise: Edges in *Drosophila* are randomly replaced, and the network is reclassified. Plotted are prediction scores for each of the seven classes as more and more edges are replaced. Each point is an average over 200 independent random replacements. For high noise levels (beyond 80%) the network is classified as an Erdös–Rényi (RDS) graph. Also note that the confidence in the classification as a DMC network for low noise (<30%) is even higher than in the classification as an RDS network for high noise. The prediction score $y(c)$ for class $c$ is related to the estimated probability $p(c)$ for the tested network to be in class $c$ by $p(c) = e^{2y(c)}/(1 + e^{2y(c)})$ (43). The dashed line indicates a $p$ value of $1 - 10^{-3}$.

as well as recent direct experimental evidence presented by Wang *et al.* (48) for a single DMC event in *Drosophila melanogaster*. We also showed that different mechanisms, such as DMR, LPA, and RDG, model *Drosophila* well for different sets of subgraphs, a result which suggests that a model that mixes several mechanisms might be able to reproduce *Drosophila* even more accurately. Preliminary studies on the yeast protein–protein interaction network, as produced by an analysis that integrates multiple data sources (49), also strongly favors the DMC mechanism. We anticipate that further use of machine learning techniques will answer a number of such questions of interest in systems biology.

We acknowledge insightful discussions with Christina Leslie and Yoav Freund and key suggestions on the manuscript by G. Stolovitzky. This work was supported by National Science Foundation Grants ECS-0332479, ECS-0425850, and DMS-9810750 and National Institutes of Health Grant GM036277 (to C.H.W.).

1. Strogatz, S. H. (2001) *Nature* **410,** 268–276.
2. Newman, M. (2003) *SIAM Rev.* **45,** 167–256.
3. Barabási, A. (1999) *Science* **286,** 509–512.
4. Watts, D. & Strogatz, S. (1998) *Nature* **363,** 202–204.
5. de Solla Price, D. J. (1965) *Science* **149,** 510–515.
6. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003) *ComPlexUs* **1,** 38–44.
7. Sole, R. V., Pastor-Satorras, R., Smith, E. & Kepler, T. B. (2002) *Adv. Complex Syst.* **5,** 43–54.
8. Berg, J., Lässig, M. & Wagner, A. (2003) arXiv:cond-mat/0207711.
9. Rzhetsky, A. & Gomez, S. M. (2001) *Bioinformatics* **17,** 988–996.
10. Qian, J., Luscombe, N. M. & Gerstein, M. (2001) *J. Mol. Biol.* **313,** 673–681.
11. Bhan, A., Galas, D. J. & Dewey, T. G. (2002) *Bioinformatics* **18,** 1486–1493.
12. Kumar, R., Raghavan, P., Rajagopalan, S. & Sivakumar, D. (2000) in *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*, ed. Blum, A. (Inst. Electrical Electronics Engineers, Piscataway, NJ), pp. 57–65.
13. Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning* (Springer, New York).
14. Devroye, L., Györfi, L. & Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition* (Springer, New York).
15. Saito, R., Suzuki, H. & Hayashizaki, Y. (2003) *Bioinformatics* **19,** 756–763.
16. Goldberg, D. S & Roth, F. P. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4372–4376.
17. Morris, Q. D., Frey, B. J. & Paige, C. J. (2004) in *Advances in Neural Information Processing Systems 16*, eds. Thrun, S., Saul, L. K. & Schölkopf, B. (MIT Press, Cambridge, MA) pp. 385–393.
18. Gomez, S. M. & Rzhetsky, A. (2002) *Pac. Symp. Biocomput.*, 413–424.
19. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002) *Nat. Genet.* **31,** 64–68.
20. Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N. & Alon, U. (2002) *Science* **298,** 824–827.
21. Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N. & Stone, L. (2004) *Science* **305,** 1107.
22. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R. & Alon, U. (2004) *Science* **305,** 1107.
23. Hasty, J., McMillen, D. & Collins, J. J. (2002) *Nature* **420,** 224–230.
24. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002) *Science* **298,** 799–804.
25. Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. (2003) *Nat. Genet.* **35,** 176–179.
26. Vespignani, A. (2003) *Nat. Genet.* **35,** 118–119.
27. Rosenfeld, N., Elowitz, M. & Alon, U. (2002) *J. Mol. Biol.* **323,** 785–793.
28. Mangan, S. & Alon, U. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 11980–11985.
29. Ziv, E., Koytcheff, R. & Wiggins, C. H. (2003) arXiv:cond-mat/0306610.
30. Holland, P & Leinhardt, S. (1976) *Sociological Methodology* **7,** 1–45.
31. Wasserman, S., Faust, K. & Iacobucci, D. (1994) *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, Cambridge, U.K.).
32. Connor, E. F. & Simberloff, D. (1979) *Ecology* **60,** 1132–1140.
33. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004) *Science* **303,** 1538–1542.
34. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al.* (2003) *Science* **302,** 1727–1736.
35. Hughes, A. L. (1994) *Proc. R. Soc. London B* **256,** 119–124.
36. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-L. & Postlethwait, J. (1999) *Genet. Soc. Am.* **151,** 1531–1545.
37. Erdös, P & Rényi, A. (1959) *Publicationes Mathematicae* **6,** 290–297.
38. Callaway, D., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. & Strogatz, S. H. (2001) *Phys. Rev. E.* **64,** 041902–041908.
39. Klemm, K. & Eguiluz, V. M. (2002) *Phys. Rev. E* **65,** 036123–036127.
40. Freund, Y. & Mason, L. (1999) in *Proceedings of the 16th International Conference on Machine Learning*, eds. Bratko, I. & Dzeroski, S. (Kaufmann, San Francisco, pp. 124–133.
41. Schapire, R. E. (2002) in *MSRI Workshop on Nonlinear Estimation and Classification*, eds. Denison, D. D., Hansen, M. H., Holmes, C. C., Mallick, B. & Yu, B. (Springer, New York), pp. 149–172.
42. Friedman, J., Hastie, T. & Tibshirani, R. (1998) *Ann. Stat.* **28,** 337–407.
43. Freund, Y. & Schapire, R. (1997) *J. Comput. Syst. Sci.* **55,** 119–139.
44. Schapire, R. E. & Singer, Y. (2000) *Machine Learning* **39,** 135–168.
45. Freund, Y & Schapire, R. (1999) *J. Jpn. Soc. Artif. Intell.* **14,** 711–780.
46. Middendorf, M., Kundaje, A., Wiggins, C., Freund, Y. & Leslie, C. (2004) *Bioinformatics* **20,** Suppl. 1, I232–I240.
47. Vazquez, A. (2003) *Phys. Rev. E* **67,** 056104–056118.
48. Wang, W., Yu, H. & Long, M. (2004) *Nat. Genet.* **36,** 523–527.
49. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8348–8353.
50. Fiedler, M. (1973) *Czech. Math. J.* **23,** 298–305.
51. Chung, F. R. K. (1997) *Spectral Graph Theory*, Regional Conference Series in Mathematics (Am. Math. Soc., Providence, RI).