High-Resolution Modeling of Cellular Signaling Networks

Michael Baym[1,2,⋆], Chris Bakal[3,4,⋆], Norbert Perrimon[3,4], and Bonnie Berger[1,2,⋆⋆]

[1] Department of Mathematics, MIT, Cambridge, MA 02139
[2] Computer Science and Artificial Intelligence Laboratory, MIT, 02139
[3] Department of Genetics, Harvard Medical School, Boston, MA 02115
[4] Howard Hughes Medical Institute, Boston MA 02215
bab@mit.edu

Abstract. A central challenge in systems biology is the reconstruction of biological networks from high-throughput data sets. A particularly difficult case of this is the inference of dynamic cellular signaling networks. Within signaling networks, a common motif is that of many activators and inhibitors acting upon a small set of substrates. Here we present a novel technique for high-resolution inference of signaling networks from perturbation data based on parameterized modeling of biochemical rates. We also introduce a powerful new signal-processing method for reduction of batch effects in microarray data. We demonstrate the efficacy of these techniques on data from experiments we performed on the *Drosophila* Rho-signaling network, correctly identifying many known features of the network. In comparison to existing techniques, we are able to provide significantly improved prediction of signaling networks on simulated data, and higher robustness to the noise inherent in all high-throughput experiments. While previous methods have been effective at inferring biological networks in broad statistical strokes, this work takes the further step of modeling both specific interactions and correlations in the background to increase the resolution. The generality of our techniques should allow them to be applied to a wide variety of networks.

1 Introduction

Biological signaling networks regulate a host of cellular processes in response to environmental cues. Due to the complexity of the networks and the lack of effective experimental and computational tools, there are still few biological signaling networks for which a systems-level, yet detailed, description is known [1]. Substantial evidence now exists that the architecture of these networks is highly complex, consisting in large part of enzymes that act as molecular switches to activate and inhibit downstream substrates via post-translational modification. These substrates are often themselves enzymes, acting in similar fashion.

⋆ These authors contributed equally to this work.
⋆⋆ To whom correspondence should be addressed.

M. Vingron and L. Wong (Eds.): RECOMB 2008, LNBI 4955, pp. 257–271, 2008.

In experiments, we are able to genetically inhibit or over-express the levels of activators, inhibitors and the substrates themselves, but rarely are able to directly observe the levels of active substrate in cells. Without the ability to directly observe the biochemical repercussions of inhibiting an enzyme in real-time, determining the true in vivo targets of these enzymes requires indirect observation of genetic perturbation and inference of enzyme-substrate relationships. For example, it is possible to observe downstream transcription levels which are affected in an unknown way by the level of active substrate [2].

The specific problem we address is the reconstruction of cellular signaling networks studied by perturbing components of the network, and reading the results via microarrays. We take a model-based approach to the problem of reconstructing network topology. For every pair of proteins in the network, we predict the most likely strength of interaction based on the data, and from this predict the topology of the network. This is computationally feasible as we are considering a subset of proteins for which we know the general network motif.

We demonstrate the efficacy of this approach by inferring from experiments the Rho-signaling network in *Drosophila*, in which some 40 enzymes activate and inhibit a set of approximately seven substrates. This network plays a critical role in cell adhesion and motility, and disruptions in the orthologous network in humans have been implicated in a number of different forms of cancer [3]. This structure, with many enzymes and few substrates (Fig. 1), is a common motif in signaling networks [4, 5].

To complicate the inference of the Rho-signaling network further, not every enzyme-substrate interaction predicted *in vitro* is reflected *in vivo* [6]. As such, we need more subtle information than is provided by current high-throughput protein-protein interaction techniques such as yeast two-hybrid screening [7, 8].

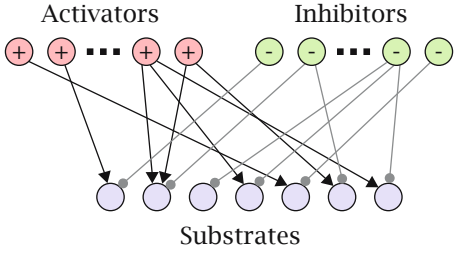

Fig. 1. The many enzyme-few substrate motif. A triangular arrowhead represents activation, a circular arrowhead inhibition.

To probe this network, we have carried out and analyzed a series of knockout and overexpression experiments in the *Drosophila* S2R+ cell line. We measure the regulatory effects of these changes using DNA microarrays. It is important to note that microarrays measure the relative abundance of the gene transcript, which can be used as a rough proxy for the total concentration of gene product. What they do not elucidate, however, is the relative fraction of an enzyme in an active or inactive state, which is crucial to the behavior of signaling networks.

To reconstruct the network from measurement, rather than directly use the microarray features corresponding to the proteins of interest, we instead use correlations in observations of the affected downstream gene products.

A number of related techniques for inferring global patterns based on high-throughput data exist. Many of these utilize the technique of probabilistic graphical models [9, 10, 11, 12, 13]. While these techniques are effective for inferring networks in broad statistical strokes, we increase the resolution and model the rate coefficients of individual reactions. The mathematics of our methodology is in fact isomorphic to a probabilistic graphical model approach; however as our parameters correspond directly to physical quantities or coefficients, we are able to dramatically narrow our model space when compared to a more general technique such as Bayesian or Markov networks [9]. In doing so we are able to gain both greater sensitivity, specificity, and robustness to noise. A related technique, based on modeling of rate kinetics in the framework of Dynamic Bayesian Networks has been effective in modeling genetic regulatory networks [14]. Techniques from information theory, such as ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [15, 16] and nonparameteric statistics, such as GSEA (Gene Set Enrichment Analysis) [17] have also been used to infer connections in high-throughput experiments. While not generally used for signaling network reconstruction, GSEA notably has been popular recently [18,19], in part for its efficacy in overcoming batch effect noise.

We take the novel approach of constructing and optimizing a detailed parameterized model, based on the biochemistry of the network we aim to reconstruct. For the first part of the network model, namely the connections of the enzymes to substrates, we know the specific rate equations for substrate activation and inhibition. By modeling the individual interactions in like manner to the well-established Michaelis-Mentin rate kinetics [20, 21, 14], we are able to construct a model of the effects of knockout experiments on the level of active substrate. Lacking prior information, we model the effect of the level of active substrate on the microarray data by a linear function. If the only source of error were uncorrelated Gaussian noise in the measurements, we could then simply fit the parameters of this model to the data to obtain a best guess at the model's topology.

However, noise and "batch effects" [18] in microarray data are a real-world complication for most inference methods, which we address in a novel way. Noise in microarrays is seemingly paradoxical. On one hand, identical samples plated onto two different microarrays will yield almost identical results [22, 23]. On the other hand, with many microarray data sets, when one simply clusters experiments by similarity of features, the strongest predictor of the results is to group by the day on which the experiment was performed. We hypothesize, in this analysis, that the batch effects in microarrays are in fact other cellular processes in the sample unrelated to the experimental state. Properly filtering the ever-present batch effects in microarray data requires more than simply considering them to be background noise. Specifically, instead of the standard approach of fitting the data to our signal and assuming noise cancels, we consider the data

to be a combination of the signal we are interested in and a second, structured signal of the batch effects.

Fitting this many-parameter model with physical constraints to the actual data optimizes our prediction for the signaling network, with remarkably good results.

To test this method we have constructed random networks with structure similar to the expected biology, and used these to generate data in simulated experiments. We find that when compared to reconstructions based on naïve correlation, GSEA, and ARACNE, we were able to obtain significantly more accurate network reconstructions. That is to say, at every specificity we obtained better sensitivity and vice-versa. The details of how GSEA and ARACNE were used in this manner can be found in Sec. 3.1.

We have also reconstructed the Rho-signaling network in *Drosophila* S2R+ cells from a series of RNAi and overexpression experiments we performed. While very little is experimentally known about this network, of the 40 pairs for which we have any biological evidence, we were able to predict 26 correctly, considerably better than chance (a p-value of 0.0079). It is important to remember that this standard is far from certain, and the known data represents a small fraction of the over 180 connections we aim to predict. Notably, many of the global features of the predicted network are in line with what is believed from biological experiments. While there is little doubt that with further experiments we will predict a more accurate network, this is the first detailed systems-level model of the *Drosophila* Rho-signaling network.

Contributions. We have introduced a novel parameterized model-based approach to signaling network inference from high-throughput data. We use this to provide testable predictions for connections in the *Drosophila* Rho-signaling network. Large-scale general statistical techniques have painted networks in broad strokes. Given the broad generality of such modeling, and the prevalence of similar motifs to the example studied here, the present approach is a crucial step in the program of systems biology.

Additionally we have developed a method for incorporating a noise model into this fit so as to greatly reduce the impact of batch effects in microarray data. This approach to noise in microarrays is widely applicable.

2 Models and Algorithms

In broad terms, we first aim to derive a model of the effects of our perturbations on the data whose parameters correspond to the edge weights of the cellular signaling network we wish to reconstruct. We first model how the level of active substrate changes in response to perturbations of the activators or inhibitors. To do this we derive an equilibrium condition based on well-known biochemical rate kinetics. We then make a linear model of how this affects the experimental data.

To fully understand the data, however, requires more than simply a model of the network. We need, as pointed out earlier, to model the noise, in order

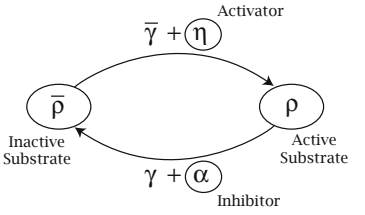

Fig. 2. The dynamics of an activator-inhibitor-substrate trio. The circled variables are proportional to protein concentrations.

to account for correlations in the background levels on unperturbed repeat experiments; we take a low-dimensional linear approximation of the batch effects present in microarray data. By fitting the parameters of the resultant model to the experimental data, we are able to predict both the topology and edge weights of the signaling network.

2.1 Biochemical Model

We first illustrate our approach for a single activator-inhibitor-substrate trio before extending to the many-node case. We start by deriving the time dependence of the concentration[1] ρ of active substrate in terms of the concentrations $\bar{\rho}$ of inactive substrate, η of activator, α of inhibitor, and the base rates $\bar{\gamma}$ of activation and γ of de-activation. Fig. 2 depicts these kinetics. As the rate at which inactive substrate becomes active is proportional to its concentration times the rate of activation and vice-versa,

$$\frac{d\rho}{dt} = -\frac{d\bar{\rho}}{dt} = \bar{\rho}\,(\bar{\gamma} + \eta) - \rho\,(\gamma + \alpha)\,. \tag{1}$$

We are primarily interested in ρ, the level of active substrate, as the downstream effects of the substrate are dependent on this concentration. As the measurements are taken several days after perturbation and are an average over the expression levels of many individual cells, by ergodicity we expect to find approximately the equilibrium ($d\rho/dt = 0$) concentration of substrate.

Solving for ρ at equilibrium yields:

$$\rho = \frac{\kappa\,(\bar{\gamma} + \eta)}{\bar{\gamma} + \eta + \gamma + \alpha}. \tag{2}$$

where $\kappa = \rho + \bar{\rho}$ is total concentration of the substrate, approximately available from the microarray data. By choice of time units we can let $\bar{\gamma} = 1$. This result, by no coincidence, is similar to the familiar Michaelis-Mentin rate kinetics.

[1] Choice of units of concentration is absorbed by scalar factors of the fit once the x_{jk} and y_{jk} coefficients are added; see Eq. 3.

We now generalize the model to multiple substrates κ_k, interchangeable activators η_j with relative strength x_{kj}, and inhibitors α_j with relative strength y_{kj}. The equilibrium concentration of the level of active substrate ρ_k then becomes:

$$\rho_k = \frac{\kappa_k \left(1 + \sum_j x_{kj}\eta_j\right)}{1 + \sum_j x_{kj}\eta_j + \gamma_k + \sum_j y_{kj}\alpha_j}. \tag{3}$$

Lacking more detailed biological information, and aiming to avoid the introduction of unnecessary parameters, we assume a linear response from features in the microarray. Specifically, for a vector of microarray feature data $\boldsymbol{\varphi}$, we model the effect as a general linear function of the levels of active substrate, of the form $\boldsymbol{a\rho} + \boldsymbol{r}$. Additionally we introduce a superscripted index z for those variables which vary by experiment. The level, φ_i^z, of the i^{th} feature in microarray z is in our model:

$$\varphi_i^z = \sum_k a_{ik} \left(\frac{\kappa_k^z \left(1 + \sum_j x_{kj}\eta_j^z\right)}{1 + \sum_j x_{kj}\eta_j^z + \gamma_k + \sum_j y_{kj}\alpha_j^z} \right) + r_i + \beta_i^z + \epsilon_i^z, \tag{4}$$

where the batch effects $\boldsymbol{\beta}$ and noise $\boldsymbol{\epsilon}$ are considered additively.

2.2 Noise Filtration

As batch effects in microarrays are highly correlated, our approach is to construct a linear model of their structure. Empirically, batch effects tend to have a small number, s, of significant singular values (from empirical data $s \simeq 4$). In the singular vector basis, we can model the batch effects as a (features $\times\, s$) matrix $\boldsymbol{c}$. To determine the background as a function of experiment batch, we rotate by an ($s \times$ batches) rotation matrix $\boldsymbol{u}$. Thus $\boldsymbol{cu} = \sum_j c_{ij}u_{jd}$ is a (features $\times$ batches) matrix whose columns are the background signal by batch. Finally to extract the batch effect for a given experiment z, we multiply by the characteristic function of experiments by batches, $\boldsymbol{\chi}$, where $\chi_d^z = 1$ if experiment z happened in batch d and is 0 otherwise. Our model of batch effects is then:

$$\beta_i = \sum_{l,d} c_{il}u_{ld}\chi_d^z. \tag{5}$$

All together, our detailed model for experimental data based on the network, experiments, and noise becomes:

$$\varphi_i^z = \sum_k a_{ik} \left(\frac{\kappa_k^z \left(1 + \sum_j x_{kj}\eta_j^z\right)}{1 + \sum_j x_{kj}\eta_j^z + \gamma_k + \sum_j y_{kj}\alpha_j^z} \right) + r_i + \sum_{l,d} c_{il}u_{ld}\chi_d^z + \epsilon_i. \tag{6}$$

2.3 Model Fitting

Having now constructed a model of our system, we minimize the least-squares difference between the model predictions and observed data (detailed in Sec. 3.2),

to obtain optimal model parameters. The resultant values of $\boldsymbol{x}$ and $\boldsymbol{y}$ predict the relative strengths of the activator-substrate interactions.

It is important to keep in mind which parameters are known and which we must fit. We know s and $\boldsymbol{\chi}$ from experiment. In lieu of detailed knowledge of the activity levels of the activator and inhibitor, we take κ_k^z, η_j^z and α_j^z to be 1 normally, 0 on those experiments for which the gene is silenced, and 2 for those in which it is overexpressed. The remaining fitting parameters of our model are $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{\gamma}, \boldsymbol{r}, \boldsymbol{c}$, and $\boldsymbol{u}$.

For a vector of experimental data $\boldsymbol{d}$, we construct, as above, a model for the predicted data $\boldsymbol{\varphi}$. Fitting the model to data is done by minimizing:

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{\gamma}, \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{u}) = \sum_{i,z} \left(d_i^z - \varphi_i^z\right)^2, \tag{7}$$

where φ_i^z is given in Eq. 6, subject to the constraints

$$x_{kj}, y_{kj}, \delta_k, \kappa_k \geq 0 \tag{8}$$

and the additional constraint that $\boldsymbol{u}$ is a rotation matrix. As the solution space is non-convex and likely has local minima, we use a general trust-regions [24] method for minimization starting at multiple starting points. The fit with lowest objective value is taken to be the best predictor of the network.

To verify that we have more data than parameters, we consider a microarray with Φ features and a network model with a total of θ activators and inhibitors and σ substrates. Additionally we consider a 4-dimensional noise model for λ batches. Then for ζ experiments, we have more data than parameters precisely when:

$$\zeta > \sigma + 4 + \frac{(\theta + 3)\sigma + 4\lambda - 10}{\Phi} \tag{9}$$

In a realistic setting, for 26 enzymes, six substrates, with on average six experiments per batch, and assuming each experiment has at least 50 features, then we need to perform at least 14 experiments in order to have more data than parameters. As the batch effect model has substantially lower rank than the number of batches, as long as there are at least five batches, over-fitting is unlikely.

In the above setting with 70 experiments, network optimization takes approximately 8 hours on a Powerbook G4 using an off-the-shelf constrained local nonlinear optimization routine in the MATLAB Optimization Toolbox [25] to a convergence tolerance of $1e-6$. While we aim to find the network which globally minimizes f, this trust-regions based local search technique occasionally reaches the convergence threshold at a demonstrably sub-optimal value. Continuing to optimize on a subset of the variables followed by repeated total optimization is often sufficient to pass these obstacles. Nevertheless, this still yields a good network prediction (see below). With more refined optimization tools, we will likely make even more accurate predictions. While we find that in very noisy cases the global minimum of f is smaller than that predicted by the actual connections, an overfit of the data, in practice this is a good guess.

3 Results

3.1 Simulations

We have generated simulated data on randomly created networks. The density of activator-substrate and inhibitor-substrate connections was chosen to reflect what is expected in the Rho-signaling network described in Sec. 3.2. From this, we have generated model experiment sets consisting of one knockout twice of each of the substrates and a single knockout of each activator and inhibitor in batches in random order. To further mimic our biological data set we included at least one baseline experiment in each batch. From this model we simulated experimental data with both noise and a batch-effect signal and attempted to fit the generated data.

To test against other techniques, we applied the statistics used by GSEA and ARACNE, modified for use on our model data sets. While GSEA is not typically used for signaling network reconstruct, its general usefulness in microarray analysis necessitates the comparison. ARACNE, on the other hand, while designed for a similar situation, does not directly apply, and so needs to be modified to make a direct comparison. As a baseline, we also computed the naïve (Pearson) correlation of experimental states.

GSEA starts by constructing, for each experimental condition, two subsets ("gene sets") of the features, one positive and one negative, which are used as indicators of the condition. To test whether a specific state is represented in a new experiment, the Kolmogorov-Smirnov enrichment score of those subsets in the new data is calculated (for details, see [17]). If the positive set is positively enriched and the negative set negatively enriched, the test state is said to be represented in the data. Likewise if the reverse occurs, the state is said to be negatively represented. If both are positively or negatively enriched, GSEA does not make a prediction. We are able to apply GSEA by computing positive and negative gene sets based on perturbation data for the substrates and then testing for enrichment in each of states in which we perturb an activator or inhibitor.

ARACNE, on the other hand, begins by computing the kernel-smoothed approximate mutual information (AMI) of every pair of features (for details, see [15]). In order to remove transitive effects, for every trio of features A, B, C, the pair with the smallest mutual information is marked to not be an edge. The remaining set of all unmarked edges is then a prediction of the network. As already discussed, we do not have features in our experiment that correspond directly to the levels we wish to measure. However, treating each experimental state as a feature, we are able to apply the AMI metric to obtain the relative efficacies of the activator and inhibitor perturbation experiments as predictors of the substrate perturbations. We know from the outset that the network we are trying to predict has no induced triangles, and so ARACNE would not remove any of the edges. However, the relative strengths of these predictions yield a predicted network topology.

On noiseless data, with only a minimal set of experiments and batch effects of comparable size to the perturbation signal, we are able to achieve a perfect network reconstruction which was not achieved by any of the other methods we consider. On highly noisy data, we cannot reconstruct the network perfectly; however we consistently outperform the other methods in both specificity and sensitivity (Fig. 3). Moreover, we find that while the model alone out-performs other techniques (comparably to AMI), the batch effect fit is of crucial importance. While this is clearly a biased result, as the simulated data is generated by the same model we assume in the fit, it does show that we are able to obtain a partial reconstruction even under high noise conditions. As this is a best-guess model from prior biological knowledge, the assumptions are far from unreasonable.

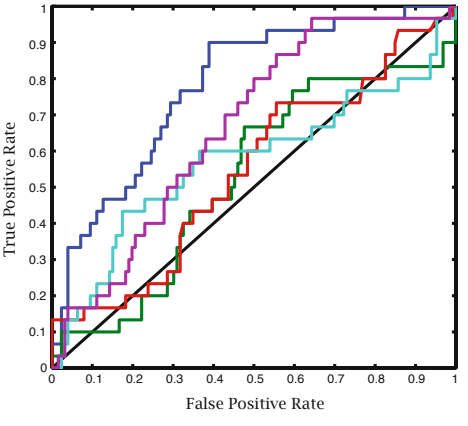

Fig. 3. Typical ROC curve for highly noisy simulated data. Our model (dark blue) is closest to the actual network, which would be a point at $[0, 1]$. Model fitting without batch effects (purple) is also considered. The other lines represent the predictions obtained by a GSEA-derived metric (red), an ARACNE-derived metric (light blue), and naïve correlation (green). The diagonal black line is the expected performance of random guessing. This particular set of simulated data has no repeat experiments for GAPs or GEFs, a batch signal of half the intensity of the perturbations, and an approximate total signal-to-noise ratio of 1.5.

3.2 Biological Data

We used our method, discussed above, on forthcoming microarray data collected from RNAi and overexpression experiments to predict the structure of the Rho-signaling network in *Drosophila* S2R+ cells. This network consists of approximately 47 proteins, divided roughly as 7 GTPases, 20 Guanine Nucleotide Exchange Factors (GEFs) and 20 GTPase Activating Proteins (GAPs). Importantly, we have the additional information that, despite their misleading names, the GEFs serve to activate certain GTPases and the GAPs serve to inhibit them. The exact connections, however, are for the vast majority, unknown.

Labeled aRNA, transcribed from cDNA, was prepared from S2R+ Drosophila cells following five days incubation with dsRNA or post-transfection of overexpression constructs. The aRNA was then hybridized to CombiMatrix 4x2k CustomArrays designed to include those genes most likely to yield a regulatory effect from a perturbation to the Rho-signaling network. After standard spatial and consensus Lowess [26] normalization, we k-means clustered [27] the data into 50 pseudo-features to capture only the large-scale variation in the data.[2]

After fitting, we have computed the significance of our fit using the Akaike and Bayesian Information Criteria (AIC and BIC) [28, 29]. These measure parameter fit quality as a function of the number of parameters, with smaller numbers being better. AIC tends to under-penalize free parameters while BIC tends to over-penalize, thus we computed both. As a baseline, we computed the AIC/BIC of the null model. While a direct fit of the pseudo-features yielded a lower AIC but not BIC, an iterative re-fit and solve technique, not unlike EM, produced a significant fit by both criteria (Table 1, prediction in Fig 5). This re-fitting was done by greedily resorting the groupings for meta-features based on the model fitness and refitting the model to the new meta-features. As each step strictly increases fit quality, and there are only finitely many sets of meta-features, this is naïvely guaranteed to converge in $O(n^k)$ iterations for n features and k meta-features. We find, however that the convergences generally to happens in around 5 iterations, leaving feature variance intact (an indication that this is not converging to a degenerate solution).

Table 1. AIC/BIC of the null model, best naïve fit, and best fit

Model	*Fit (f)*	*AIC*	*BIC*
Null Model ($\varphi_i^z = 0$)	0.9885	-8.389	-8.387
Best Fit	0.2342	-9.480	-8.366
Adapted Features	0.0328	-11.446	-10.332

To further test the accuracy of our model, we fit the model to four subsets of the 87 experiments and tested the prediction quality on the remaining experiments. The prediction error is calculated as the mean squared error of the predicted values divided by the mean standard deviation by feature. We tested on four sets: Sets 1 and 2 were chosen randomly to have nine (10.3% of experiments) and seventeen (19.5% of experiments) elements respectively, of which four of each are unduplicated experiments. Sets 3c and 4c were chosen randomly to have nine elements but were constrained not to have two elements from the same batch or experimental condition. We find that the model accurately predicts test set data (Table 2) for repeated experiments. Note that in Set 1, when

[2] The fact there are fewer than 50 significant singular values in the data and the linearity of $\boldsymbol{a}, \boldsymbol{r}$ and $\boldsymbol{\beta}$, indicates that we can not get more information from more clusters.

Table 2. Prediction error on test data.

Test Set	*Size*	*#Unduplicated*	*Total Fit (f)*	*Test Set Fit*	*Error*
1	9	4	0.0280	0.1307	14.6%
2	17	4	0.0288	0.0632	6.10%
3c	9	0	0.0302	0.0371	3.13%
4c	9	0	0.0301	0.0517	4.06%

44% of the experiments in the test set are non-duplicated, the prediction error is significantly higher. This indicates the necessity of both the batch and network components of the model.

While very little is known about the actual structure of the network, our reconstruction performed well when compared to previous biological data from in vivo experiments [30, 31, 32, 33, 34, 35, 36, 37] or mammalian homology, [38, 39, 40, 41, 42, 43, 44, 45, 46]. We predicted the existence of 57 of the 156 possible connections. Of the 23 known connections, both from in vivo experiments and inferred by orthology, we successfully predicted 11. Of the 17 pairs of proteins for which there is evidence they do not interact, we correctly predicted 15. This compares quite favorably to the predictions of other methods (Fig. 4). On this set of known interactions and non-interactions, the probability that our set of predicted connections overlapped correctly at least 26 times by chance is 0.0079. It is important to keep in mind that the known data represents less than a quarter of the testable connections predicted by our method.

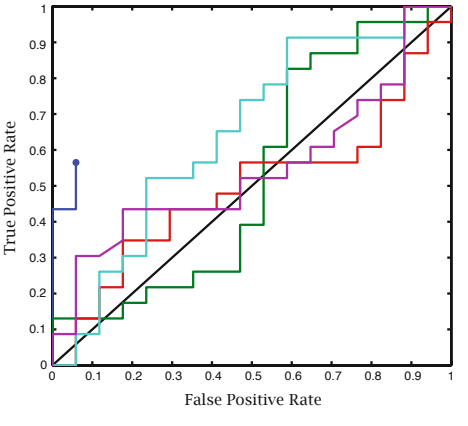

Fig. 4. ROC curve of network predictions vs. known data. Our model (dark blue) is closest, the curve discontinuity is on account of many of the predictions being zero. The other lines represent the predictions obtained by our model without a batch effect model (green), a GSEA-derived metric (purple), an ARACNE-derived metric (red), and naïve correlation (light blue).

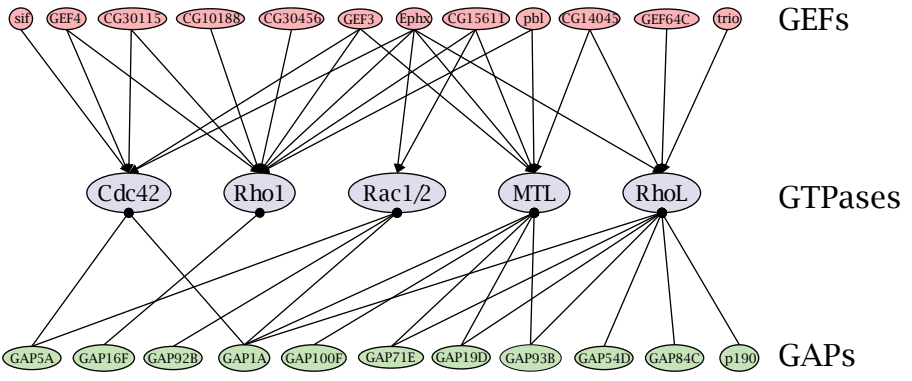

Fig. 5. The predicted Rho-signaling network in *Drosophila*

Two global network features of note, that the GTPase Rho1 is more highly connected than either Rac1/2 or Cdc42, and that the GEF Ephx has broad specificity, were reflected in our predictions as well. We also note that the prediction quality is not substantially different for GEFs (7 of 12 positives and 8 of 10 negatives) or GAPs (4 of 7 positives and 7 of 7 negatives).

4 Conclusion

In this paper we infer a signaling network from microarray data on perturbation experiments. We do so by constructing a detailed model of both the network and experimental background noise. We demonstrate the effectiveness of this technique on simulated data, and use it to make testable predictions of the connections in the *Drosophila* Rho-signaling network.

There are several natural extensions to our model. First, it is possible to backtrack errors in prediction in order to guide future experiments. We can also obtain a better fit on the unknown connections by incorporating further biological knowledge. For example, if it is known that a given enzyme-substrate pair does or does not interact, we can limit our model space to reflect this with an appropriate constraint on x_{jk} in Eq. 8. Recent advances in optimization promise greater efficiency and scalability than the method we used.

Our approaches have more general applicability. Since the many enzyme-few substrate motif is so common, we can use similar techniques to elucidate more networks as the data sets become available. Furthermore, microarray data is used in many contexts beyond network inference. The method of filtering batch effects proposed here will provide a potentially very useful tool for future exploration.

Acknowledgements. We are grateful to Jonathan Kelner, Kenneth Kamrin, and Nathan Palmer for helpful input. M.B. gratefully acknowledges support from the Fannie and John Hertz Foundation, and the National Defense Science and Engineering Program. C.B is a Fellow of the Leukemia and Lymphoma Society.

References

1. Friedman, A., Perrimon, N.: Genetic screening for signal transduction in the era of network biology. Cell 128, 225–231 (2007)
2. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H.: Functional discovery via a compendium of expression profiles. Cell 102, 109–126 (2000)
3. Sahai, E., Marshall, C.J.: Rho-gtpases and cancer. Nat. Rev. Cancer 2(2), 133–142 (2002)
4. Albert, R.: Scale-free networks in cell biology. J. Cell Sci. 118(21), 4947–4957 (2005)
5. Csete, M., Doyle, J.: Bow ties, metabolism and disease. Trends in Biotechnology 22(9), 446–450 (2004)
6. Michiels, F., Habets, G.G.M., Stam, J.C., van der Kammen, R.A., Collard, J.G.: A role for rac in tiaml-induced membrane ruffling and invasion. Nature 375, 338–340 (1995)
7. Fields, S., Song, O.-K.: A novel genetic system to detect protein-protein interactions. Nature 340, 245–246 (1989)
8. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L.: A protein interaction map of drosophila melanogaster. Science 302(5651), 1727–1736 (2003)
9. Friedman, N.: Inferring cellular networks using probabilistic graphical models. Science 303(5659), 799–805 (2004)
10. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721), 523–529 (2005)
11. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. J. of Computational Biology 7(3-4), 601–620 (2000)
12. Peõer, D., Regev, A., Elidan, G., Friedman, N.: Inferring subnetworks from perturbed expression profiles. Bioinformatics 17, S214–S224 (2001)
13. Li, C., Suzuki, S., Ge, Q.-W., Nakata, M., Matsuno, H., Miyano, S.: Structural modeling and analysis of signaling pathways based on petri nets. J. Bioinformatics and Computational Biology 4(5), 1119–1140 (2006)
14. Nachman, I., Regev, A., Friedman, N.: Inferring quantitative models of regulatory networks from expression data. Bioinformatics 20(suppl. 1), i248–i256 (2004)
15. Margolin, A.A., Wang, K., Lim, W.K., Kustagi, M., Nemenman, I., Califano, A.: Reverse engineering cellular networks. Nature Protocols 1, 662–671 (2006)
16. Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human B cells. Nature Genetics 37(4), 382–390 (2005)
17. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102(43), 15545–15550 (2005)

18. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R.: The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science 313(5795), 1929–1935 (2006)
19. Baur, J.A., Pearson, K.J., Price, N.L., Jamieson, H.A., Lerin, C., Kalra, A., Prabhu, V.V., Allard, J.S., Lopez-Lluch, G., Lewis, K., Pistell, P.J., Poosala, S., Becker, K.G., Boss, O., Gwinn, D., Wang, M., Ramaswamy, S., Fishbein, K.W., Spencer, R.G., Lakatta, E.G., Couteur, D.L., Shaw, R.J., Navas, P., Puigserver, P., Ingram, D.K., de Cabo, R., Sinclair, D.A.: Resveratrol improves health and survival of mice on a high-calorie diet. Nature 444(7117), 337–342 (2006)
20. Michaelis, L., Menten, M.: Die kinetik der invertinwirkung. Biochem. Z. 49, 333–369 (1913)
21. Briggs, G.E., Haldane, J.B.S.: A note on the kinetics of enzyme action. Biochem. J. 19, 339–339 (1925)
22. Borup, R., Zhao, P., Nagaraju, K., Bakay, E.P.H.M., Chen, Y.-W.: Sources of variability and effect of experimental approach on expression profiling data interpretation. BMC Bioinformatics 3(4) (2002)
23. Larkin, J.E., Frank, B.C., Gavras, H., Sultana, R., Quackenbush, J.: Independence and reproducibility across microarray platforms. Nature Methods 2(5), 337–344 (2005)
24. Coleman, T., Li, Y.: An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds. SIAM Journal on Optimization 6, 418–445 (1996)
25. The Mathworks: Optimization toolbox 3.1.2 (2007), `http://www.mathworks.com/products/optimization/`
26. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. J. Amer. Stat. Assoc. 74, 829–836 (1979)
27. Macqueen, J.B.: Some methods of classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathemtical Statistics and Probability, pp. 281–297 (1967)
28. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
29. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics 6(2), 461–464 (1978)
30. Sone, M., Hoshino, M., Suzuki, E., Kuroda, S., Kaibuchi, K., Nakagoshi, H., Saigo, K., Nabeshima, Y.-i., Hama, C.: Still life, a protein in synaptic terminals of drosophila homologous to gdp-gtp exchangers. Science 275(5299), 543–547 (1997)
31. Newsome, T.P., Schmidt, S., Dietzl, G., Keleman, K., Asling, B., Debant, A., Dickson, B.J.: Trio combines with dock to regulate pak activity during photoreceptor axon pathfinding in drosophila. Cell 101(3), 283–294 (2000)
32. Billuart, P., Winter, C.G., Maresh, A., Zhao, X., Luo, L.: Regulating axon branch stability: the role of p190 rhogap in repressing a retraction signaling pathway (2001)
33. Bashaw, G.J., Hu, H., Nobes, C.D., Goodman, C.S.: A novel dbl family rhogef promotes rho-dependent axon attraction to the central nervous system midline in drosophila and overcomes robo repulsion (2001)
34. Gonzalez, C.: Cell division: The place and time of cytokinesis. Current Biology 13(9), R363–R365 (2003)
35. Rossman, K.L., Der, C.J., Sondek, J.: Gef means go: turning on rho gtpases with guanine nucleotide-exchange factors. Nat. Rev. Mol. Cell Biol. 6(2), 167–180 (2005)

36. Hu, H., Li, M., Labrador, J.P., McEwen, J., Lai, E.C., Goodman, C.S., Bashaw, G.J.: Cross gtpase-activating protein (crossgap)/vilse links the roundabout receptor to rac to regulate midline repulsion (2005)
37. Nahm, M., Lee, M., Baek, S.-H., Yoon, J.-H., Kim, H.-H., Lee, Z.H., Lee, S.: Drosophila rhogef4 encodes a novel rhoa-specific guanine exchange factor that is highly expressed in the embryonic central nervous system. Gene 384, 139–144 (2006)
38. Reid, T., Bathoorn, A., Ahmadian, M.R., Collard, J.G.: Identification and characterization of hpem-2, a guanine nucleotide exchange factor specific for cdc42. J. Biol. Chem. 274(47), 33587–33593 (1999)
39. Shamah, S.M., Lin, M.Z., Goldberg, J.L., Estrach, S., Sahin, M., Hu, L., Bazalakova, M., Neve, R.L., Corfas, G., Debant, A., Greenberg, M.E.: Epha receptors regulate growth cone dynamics through the novel guanine nucleotide exchange factor ephexin. Cell 105(2), 233–244 (2001)
40. Hall, C., Michael, G.J., Cann, N., Ferrari, G., Teo, M., Jacobs, T., Monfries, C., Lim, L.: alpha2-chimaerin, a cdc42/rac1 regulator, is selectively expressed in the rat embryonic nervous system and is involved in neuritogenesis in n1e-115 neuroblastoma cells (2001)
41. Niu, J., Profirovic, J., Pan, H., Vaiskunaite, R., Voyno-Yasenetskaya, T.: G protein $\beta\gamma$ subunits stimulate p114rhogef, a guanine nucleotide exchange factor for rhoa and rac1: Regulation of cell shape and reactive oxygen species production. Circ. Res. 93(9), 848–856 (2003)
42. Nagata, K.-I., Inagaki, M.: Cytoskeletal modification of rho guanine nucleotide exchange factor activity: identification of a rho guanine nucleotide exchange factor as a binding partner for sept9b, a mammalian septin. Oncogene 24(1), 65–76 (2004)
43. Wells, C.D., Fawcett, J.P., Traweger, A., Yamanaka, Y., Goudreault, M., Elder, K., Kulkarni, S., Gish, G., Virag, C., Lim, C., Colwill, K., Starostine, A., Metalnikov, P., Pawson, T.: A rich1/amot complex regulates the cdc42 gtpase and apical-polarity proteins in epithelial cells. cell 125(3), 535–548 (2006)
44. Cho, Y.J., Cunnick, J.M., Yi, S.J., Kaartinen, V., Groffen, J., Heisterkamp, N.: Abr and bcr, two homologous rac gtpase-activating proteins, control multiple cellular functions of murine macrophages (2007)
45. Dalva, M.B.: There's more than one way to skin a chimaerin (2007)
46. Mitin, N., Betts, L., Yohe, M.E., Der, C.J., Sondek, J., Rossman, K.L.: Release of autoinhibition of asef by apc leads to cdc42 activation and tumor suppression. Nat. Struct. Mol. Biol. 14(9), 814–823 (2007)