A data integration methodology for systems biology

Daehee Hwang, Alistair G. Rust, Stephen Ramsey, Jennifer J. Smith, Deena M. Leslie, Andrea D. Weston, Pedro de Atauri, John D. Aitchison, Leroy Hood, Andrew F. Siegel, and Hamid Bolouri

PNAS 2005;102;17296-17301; originally published online Nov 21, 2005;
doi:10.1073/pnas.0508647102

A data integration methodology for systems biology

Daehee Hwang*, Alistair G. Rust*, Stephen Ramsey*, Jennifer J. Smith*, Deena M. Leslie*, Andrea D. Weston*†, Pedro de Atauri*, John D. Aitchison*, Leroy Hood*‡, Andrew F. Siegel§, and Hamid Bolouri*‡

*Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103; and §Departments of Management Science, Finance, and Statistics, University of Washington, Seattle WA 98195

Contributed by Leroy Hood, October 5, 2005

Different experimental technologies measure different aspects of a system and to differing depth and breadth. High-throughput assays have inherently high false-positive and false-negative rates. Moreover, each technology includes systematic biases of a different nature. These differences make network reconstruction from multiple data sets difficult and error-prone. Additionally, because of the rapid rate of progress in biotechnology, there is usually no curated exemplar data set from which one might estimate data integration parameters. To address these concerns, we have developed data integration methods that can handle multiple data sets differing in statistical power, type, size, and network coverage without requiring a curated training data set. Our methodology is general in purpose and may be applied to integrate data from any existing and future technologies. Here we outline our methods and then demonstrate their performance by applying them to simulated data sets. The results show that these methods select true-positive data elements much more accurately than classical approaches. In an accompanying companion paper, we demonstrate the applicability of our approach to biological data. We have integrated our methodology into a free open source software package named POINTILLIST.

Fisher's method | mixture distribution models

Systems biology (1, 2) aims to understand cellular behavior in terms of the spatiotemporal interactions among cellular components, such as genes, proteins, metabolites, and organelles. In systems biology, one typically perturbs a system and, with high-throughput measurements to identify all pertinent elements and their interactions, integrates them into a biological network to understand the system's behavior. As such, systems biology is predicated on the integration of experimental data from an ever increasing number of technologies, such as gene expression arrays, proteomics, and chromatin immunoprecipitation on chip assays (3). Integration achieves one of the most important imperatives of systems biology, namely it reduces the dimensionality of global data to deliver useful information about the system of interest.

A major challenge in systems biology is that technologies that globally interrogate biological systems have inherently high false-positive and false-negative rates (4); thus, each data type alone has a limited utility. The integration of data from different sources provides an effective means to deal with this issue by reinforcing bona fide observations and reducing false negatives. Moreover, because different experimental technologies provide different insights into a system, the integration of multiple data types offers the greatest information about a particular cellular process. For example, gene perturbation experiments (e.g., knockouts or RNA interference) reveal relationships between genes that may imply direct physical interactions or indirect logical interactions. In contrast, chromatin immunoprecipitation chip data can reveal direct protein–DNA interactions or cofactor associations with bound transcription factors (3). Combined together, these technologies can provide a much more detailed view of a transcriptional regulatory network than either alone (5).

There are a number of confounding problems that make data integration nontrivial. First, the types of data to be integrated range from discrete (e.g., a protein molecule may be localized to one or more organelles) to continuous (e.g., mRNA expression level). Second, each technology used has a different degree of reliability and different amounts of the various types of error. Even when considering multiple data sets generated by a common method, simply taking the intersection of these data sets does not remove random errors completely (6). Third, each data set includes its own systematic biases (4, 7). For example, labeling-based mass spectrometry approaches (e.g., isotope-coded affinity tag) tend to favor identification of highly abundant proteins. Small-scale experiments tend to provide strong evidence for a small portion of a network but say little about what may have been missed. Finally, in addition to data generated by high-throughput technologies, there are other attractive sources of data, such as small-scale experiments, curated databases, and computational predictions. To fully realize the potential of systems biology, it is imperative to draw from all of these sources of data. However, curated databases often favor widely studied proteins and genes. Curated databases may also merge data from different strains/cell types or from various experimental conditions, and they may contain considerable data on one part of a system while omitting other parts. Computational predictions that extrapolate from earlier experimental data [e.g., prediction of protein–protein interactions from known interactions of homologous proteins (8)] run the risk of perpetuating any systematic bias in the source data (in addition to false-positive and false-negative errors). *De novo* predictions are even more error prone.

There is therefore a pressing need for more effective methods of integrating data. These methods should accommodate various sources of binary, categorical, and continuous valued data acquired from high-throughput experiments, small-scale experiments, databases, and computational predictions; and should be suitable for dealing with missing data, high error rates, and systematic biases in each data set. In addition, there are few fully verified data sets available for training. Integration methods not requiring a training set have so far been limited to particular classes of data where specific assumptions hold true (9).

To address these concerns, we have developed a data integration methodology that can handle multiple data sets differing in type, size, and network coverage and does not require a training data set. This methodology uses an optimization algorithm to minimize the numbers of false positives and false negatives, and it makes no assumptions about the number of data sets integrated; rather, it is for general purposes and may be applied to integrate data from any existing and future technologies. We have integrated our methodology into a freely available software package named POINTILLIST.

In this paper, we describe our methodology and its statistical foundations using simple illustrative example data sets. Also in this

Conflict of interest statement: No conflicts declared.

Abbreviations: LS, Liptak–Stouffer's; MG, Mudholkar–George's; NP, nonparametric; MM, mixture model; cdf, cumulative density function; PDF, probability density function; ESA, enhanced simulated annealing.

†Present address: Pfizer Global Research and Development, Safety Sciences, Eastern Point Road, Groton, CT 06340.

‡To whom correspondence may be addressed: E-mail: hbolouri@systemsbiology.org or lhood@systemsbiology.org.

www.pnas.org/cgi/doi/10.1073/pnas.0508647102

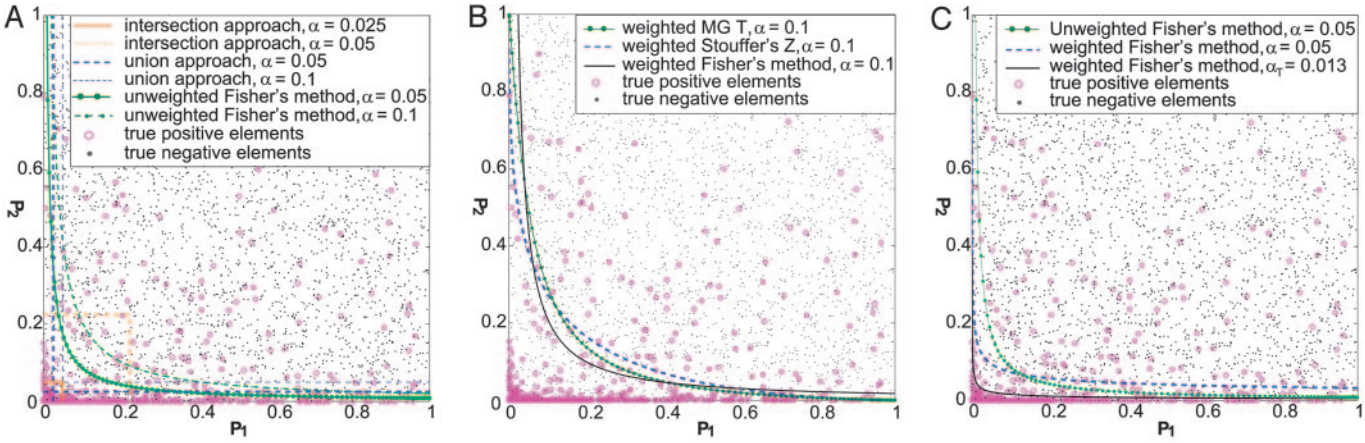

Fig. 1. Illustration of classical approaches and unweighted Fisher's method for integrating two simulated data sets. (*A*) Selection of significant elements using (*i*) intersection approaches with $\alpha = 0.05$ and $\alpha = 0.025$, (*ii*) union approaches with $\alpha = 0.05$ and $\alpha = 0.1$, and (*iii*) unweighted Fisher's method with $\alpha = 0.05$ and $\alpha = 0.1$. (*B*) Effects of integration statistics on the shape of the decision boundaries. (*C*) Selection of significant elements using weighted Fisher's method with (*i*) $\alpha = 0.05$ and (*ii*) $\alpha_T = 0.013$ to reduce false-positive errors based on the null hypothesis of the weighted Fisher's method. For these particular data sets, it can be seen that the selection using $\alpha_T = 0.013$ maximizes the precision (Table 1).

issue of PNAS (10), we demonstrate the utility and efficacy of POINTILLIST by applying it to the integration of 18 data sets to arrive at a network model of the galactose utilization network in yeast. The resulting network recapitulates the known biology of galactose utilization and provides new insights and predictions, some of which we verified experimentally.

Methods

Simulated Data Sets. Although we developed and tested our methodologies by using many sets of real experimental data, for clarity we base this paper on the simplest set of data that would be sufficient to illustrate the pertinent characteristics of the various methods presented. See the companion paper (10) for a demonstration of the applicability of our methods to a wide variety of experimental data. We generated simulated data mimicking real high-throughput data as follows. First, true differences (e.g., fold changes of gene expression levels) for the data elements affected by a perturbation (e.g., disease) were drawn with random signs from a noncentral distribution (gamma distribution with $a \geq 1.25$ and $b \geq 1.25$). We used two parameters (ϕ and η) to define the proportion of affected elements (ϕ) and the fraction of affected elements with negative differences (η), respectively. True differences for nonaffected elements ($1 - \phi$) were set to zero (Figs. 5–11, which are published as supporting information on the PNAS web site, illustrate characteristic features of the data and our methods). Second, normal random noise ($r_i e_i$) was added to the true differences (t_i) of each data set (T_i): $\mathbf{T}_i = \mathbf{t}_i + r_i \mathbf{e}_i$, where r_i is a noise level inversely related to the statistical power of the technology producing each data set, i, and e_i is standard normal. The affected data elements in a data set with a small noise level produce high absolute T_i values, resulting in low P values in the following test. Finally, a two-tailed test was performed on T_i to produce P values for each data set ($\mathbf{P}_i$): $\mathbf{P}_i = 2N_{\text{cdf}}(-|\mathbf{T}_i/r_i|)$, where the division using the scaling factor r_i is the scale of the noise. The same two data sets were used for Figs. 1–3 ($r_i = 0.666$ for data set 1 and 0.334 for data set 2, and $a = b = 2$). For the more noisy, three-dimensional data used in Fig. 4*B*, a and b values were randomly selected in the range 1.25–2.0 to generate $3 \times 10 = 30$ data sets (for the example shown in Fig. 4*A*, $a = 1.5$ and $b = 1.5$). The noise level (r_i) was selected from a uniform distribution and multiplied by a scaling factor of three for half the data and four for the rest.

Weighted Integration Methods. Several integration methods (11), such as Fisher's χ^2 (12) and Stouffer's Z (11), have been widely used in statistical metaanalysis to combine P values from k data sets. In this study, we used "weighted" versions of the following integration statistics to maximize the overall statistical power of the weighted sum of nonlinearly transformed P values in each method:

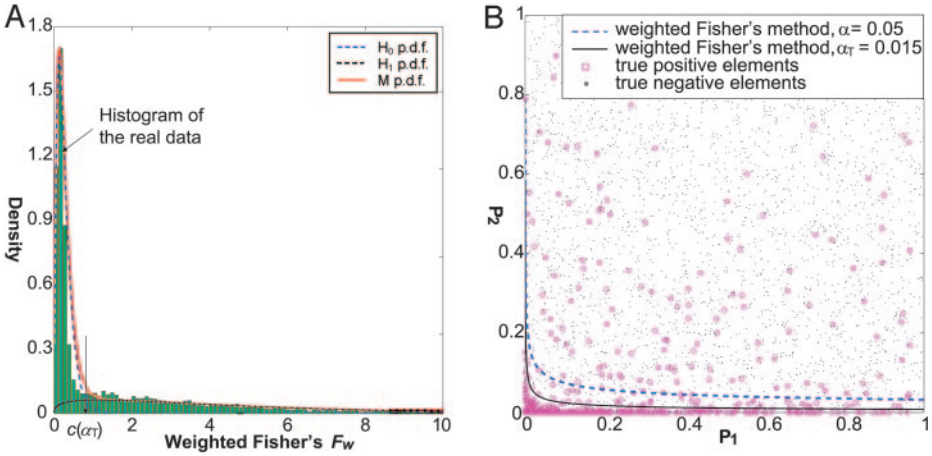

Fig. 2. Mixture distribution model. (*A*) The results of mixture distribution modeling, which indicates that the estimated H_0 and H_1 distributions represent the real data distribution (χ^2 of the residual was less than the cutoff with $\alpha = 0.05$). α_T was chosen as the area under the H_0 distribution from where the H_0 and H_1 distributions meet to the positive infinite. (*B*) Selection of significant elements using the α_T and its comparison with the weighted Fisher's method with $\alpha = 0.05$. The selection using α_T provides the best accuracy F measure (Table 1).

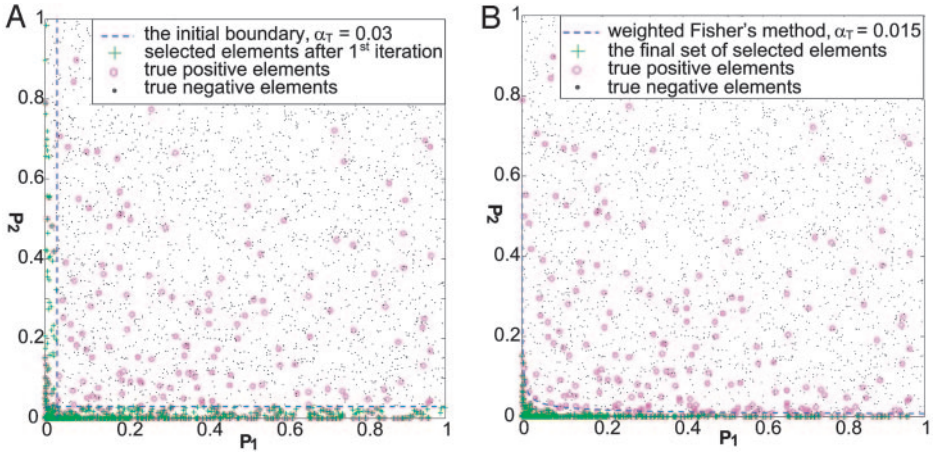

Fig. 3. NP weighted Fisher's method. (*A*) Selected elements after the first iteration. (*B*) Selected elements after the fifth iteration, which shows that the elements selected by P_1 (which has less statistical power) are being eliminated first. Thus, the iteration procedure correctly captures the relative importance of the data sets. (*C*) The final set of selected elements and its comparison with the selection using the decision boundary from $\alpha_T = 0.015$ determined by mixture distribution model (Fig. 3). Selections based on the mixture distribution model and the NP Fisher's method produce comparable accuracy (Table 1).

$$\text{Fisher's weighted } F{:}F_w = -2\sum_{i=1}^{k} w_i \ln(P_i).$$

$$\text{Mudholkar–George's (MG) weighted } T{:}T_w =$$

$$-\sqrt{\frac{15k+12}{(5k+2)k\pi^2}}\sum_{i=1}^{k} w_i \ln\left(\frac{P_i}{1-P_i}\right).$$

$$\text{Liptak–Stouffer's (LS) weighted } Z{:}Z_w =$$

$$\frac{1}{\sqrt{\sum w_i^2}}\sum_{i=1}^{k} w_i N^{-1}(1-P_i),$$

where $N^{-1}(\cdot)$ is the inverse of a standard normal cumulative density function (cdf). The weight (w_i) for the P values of each data set represents a relative measure of statistical power compared with the other data sets. Fig. 6 shows the definitions of false-positive and false-negative error rates and statistical power by using two distributions for nonaffected (background) and affected elements, when Fisher's weighted F was used. Also, Figs. 7 and 8 show the effect of random noise on the distribution of P values of the background and affected elements, respectively.

Determination of Weights. We determined the weights by maximizing the overall statistical power (observed in the data sets; Fig. 6) of the weighted integration statistics for a given significance level ($0 \le \alpha \le 1$). This maximization was implemented by using enhanced simulated annealing (ESA) (see ref. 13). (*i*) Guess an initial weight vector. (*ii*) For the weight vector, determine by using Monte Carlo simulations a cdf of background data elements (those satisfying the null hypothesis H_0) for a given integration statistic. The rejection method (14) was used to generate random numbers from a central distribution with $k = 1$ (i.e., χ^2 distribution with two degrees of freedom for Fisher's method, t-distribution with nine degrees of freedom for MG method, or standard normal distribution for LS method). (*iii*) Find an overall statistic value (c) corresponding to cdf(H_0) = $1 - \alpha$. (*iv*) Define a decision boundary for $F_w = c$ by generating a grid matrix $\boldsymbol{P}_{-j} = \{P_1, P_2, \ldots, P_{j-1}, P_{j+1}, \ldots, P_k\}$ using Gauss-Legendre quadrature (15) and then by computing $\boldsymbol{P}_j$ values for the grid points:

$$\boldsymbol{p}_j = \exp\left[\left(c + 2\sum_{i\in -j} w_i \ln(\boldsymbol{p}_i)\right) \Big/ (-2w_j)\right],$$

(*v*) Count selected data elements by using this boundary as the objective function to be maximized. (*vi*) Try another weight vector and reject or accept it by using the Metropolis algorithm until the ESA stopping criteria (13) are met. Finally, the overall P value for each data element was calculated using cdf(H_0). The selected elements should have overall P values less than the value of α.

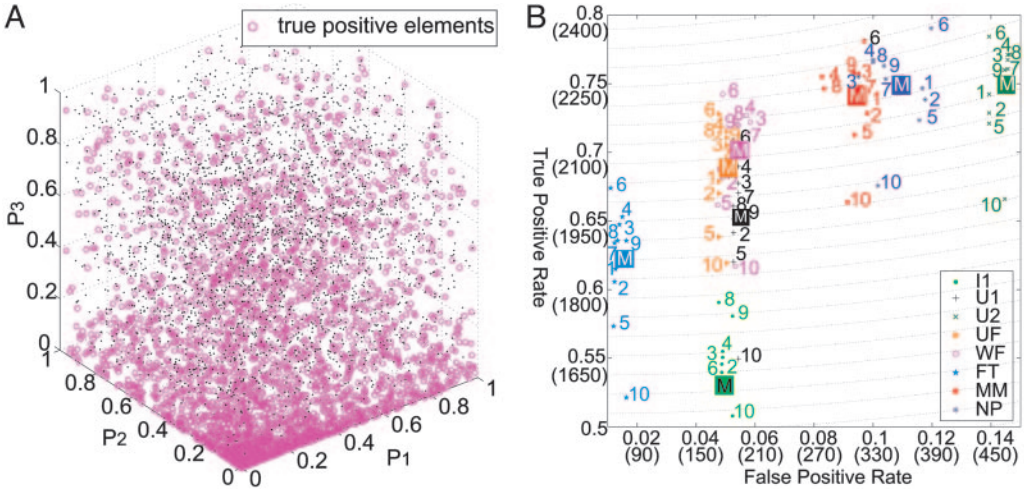

Fig. 4. Comparison of the performance of different integration methods. (*A*) P values for three complex data sets. (*B*) Receiver operating characteristic graph showing the relative performance of different integration methods in terms of false-positive and true-positive error rates. The MM and NP integration methods outperform the other methods (see text for details). UF, unweighted Fisher's method; WF, weighted Fisher's method; FT, threshold Fisher's method.

Determination of Significance Threshold (α_T) for Combined *P* Value. For a given weight vector, we compute the ratios of cumulative observed densities to cumulative expected densities by using the real data distributions in the range of α between 0 and a:

$$R(\alpha = a) = \frac{1 - D^{\mathrm{cdf}}(S_w = c(a))}{1 - H_0^{\mathrm{cdf}}(S_w = c(a))} = \frac{1 - D^{\mathrm{cdf}}(S_w = c(a))}{a},$$

where $D^{\mathrm{cdf}}(\cdot)$ is the cdf of the observed data distribution, and $c(a)$ corresponds to the integration statistic S_w (e.g., F_w) value when $\alpha = a$. Note $c(\alpha = 0)$ is infinite. As a decreases, the decision boundary moves toward the origin in *P*-value space. When there are true-positive elements, this results in a large cumulative density ratio because the expected density is small but the observed density is large. As a increases, however, the ratio decreases because the observed density slowly increases relative to the expected density. The plot of the ratio versus a shows that the curve can be split into two regions, one rich in true positives and the other in background data elements (Fig. 11). α_T was determined as the x axis value of the point nearest to the origin.

Mixture Distribution Model. For a given weight vector, we define the probability density function (PDF) of a mixture distribution model M as $\mathrm{PDF(M)} = (1 - \theta)\mathrm{PDF}(H_0) + \theta\mathrm{PDF}(H_1)$ (16, 17). We directly estimated an H_1 distribution by fitting to the distribution of the real data elements as follows: (*i*) guess a θ ($0 \leq \theta \leq 1$); (*ii*) define a $\mathrm{PDF}(H_0)$ by using Monte Carlo simulation; (*iii*) define a $\mathrm{PDF}(H_1)$ of integration statistic values (S_w) as a distribution of affected data elements by guessing two parameters (a and b) for a gamma distribution; (*iv*) compute the PDF(M) by using the equation above; (*v*) calculate the objective function as the squared sum of the difference between cdf(M) and the cdf of the real data $\|D^{\mathrm{cdf}}(S_w) - M^{\mathrm{cdf}}(S_w)\|_2^2$; (*vi*) repeat steps *ii*–*v* with a different set of gamma distribution parameters (a and b) and θ and then recompute the objective function; (*vii*) accept or reject these new trials by using the Metropolis algorithm until the ESA stopping criteria are met.

Nonparametric (NP) Decision Boundary. For a given α_T (see below), the initial set of elements is selected with the union method: $C = \cup\{\boldsymbol{P}_i < \alpha_T\}, i = \{1,2,\ldots,k\}$. A weight for each data set was then computed as the nonoverlapping area between the *P* value distributions of C and the set of random background elements (B):

$$w_i = \max_{S_i}\{\mathrm{CDF}(S_i,B) - \mathrm{CDF}(S_i,C)\},$$

where the instances of S_i are the transformed *P* values for data set i [e.g., when Fisher's method is used, $S_i = -2\ln(\boldsymbol{P}_i)$]. To quantify the overall statistical power for the two sets C and B, we compute as an objective function, f, the nonoverlapping area between distributions of weighted integration statistic values (S_w) of C and B:

$$f = \max_{S_i}\{\mathrm{CDF}(S_w, B) - \mathrm{CDF}(S_w, C)\}.$$

Note that f is related to statistical power of C, i.e., $1 - \mathrm{cdf}(S_w^{\max}, C)$, where $S_w^{\max}$ is the S_w value at which the cdf difference is at maximum. Also, the final set of selected elements depends on α_T. The initial α_T value was determined in a manner similar to the mixture distribution method above, except that we approximated the PDF of H_1 by using kernel density estimation (18).

Results

Example Data Sets. For the purposes of illustration, here we use two artificially generated example data sets. In this way, we know the data characteristics *a priori* and can illustrate and evaluate each step in our methodology unambiguously. The simulated data are designed to be as simple as possible while mimicking data from high-throughput technologies. In a companion paper (10), we demonstrate the application of the methodology to real biological data.

Fig. 1*A* presents two example data sets generated as described in *Methods*. Each of our data sets (assays for genes or proteins or their interactions) consists of two types of elements: true positives (the elements affected by an experimental perturbation, which we wish to detect) and true negatives (background elements not affected by the perturbation). As is common for high-throughput technologies, each data element is presented as a *P* value (the probability of observing a value when the corresponding data element is not affected by the experimental perturbation). Thus, data elements with lower *P* values are more likely to be true positives. In Fig. 1*A*, the true negatives are distributed uniformly (shown as gray dots). When plotting multiple data sets together, *P* values for true positives would ideally be expected to be distributed near the origin. In practice, because of the different noise characteristics of each measurement technology, *P* values for the same data element may be low in one measurement (along one axis) and not in another. This results in the distribution of true positives near the axes. Note that the true-positive elements in Fig. 1*A* are asymmetrically distributed. Data set two (measured along the *y* axis) was generated with a lower level of noise (thereby containing more statistical power) and therefore has more true-positive data elements with low *P* values. Finally, note that, although the true positives are highly correlated (as expected), the true negatives are not correlated. For real high-throughput data, true positives will comprise a small proportion of the total data. Thus data sets from different technologies will be largely uncorrelated (19). For the example presented in the companion paper, only 69 genes of 6,000 genes were selected as potential positive elements; 16,985 protein–protein interactions were selected out of $6{,}000^2$; and 8,555 protein–DNA interactions were selected out of $135 \times 6{,}000^2$. The correlation coefficients between the above data sets were all <0.3. To make it easier to understand and compare the methodologies presented, the same data are used for the analyses presented in Figs. 1–3.

Classical Approaches to Data Integration. Fig. 1*A* shows the decision boundaries for three commonly used approaches to data integration: intersection, union (20), and unweighted Fisher's method (12). As described above, true-positive data elements tend to be distributed near the axes and especially near the origin. Thus, data integration methods typically divide the total *P* value space into two regions by determining a decision boundary. Data elements in the region near the axes and/or origin, which is rich in true positives, are selected as significant. Data elements in the complementary region (away from the origin and axes) are considered true negatives. In the intersection and union methods, significant elements are identified independently in each data set using a cutoff *P* value (e.g., 0.05). For intersection, only the significant elements common to all data sets are selected. This method tends to miss many true-positive data elements near the P_1 axis and away from the origin (see Fig. 1*A*). The union method improves this situation by selecting elements significant in any one data set but includes many false positives near the P_2 axis and may still miss many elements with low geometric-mean *P* values. Note that the union and intersection methods treat each data set independently. In contrast, Fisher's method selects significant data elements by using a more sophisticated measure that results in a hyperbolic (curved) decision boundary, allowing selection of data elements near both axes and within a larger region near the origin.

Two additional integration methods widely used in metaanalysis to combine *P* values from multiple data sets are MG T and LS Z (see *Methods* for formulae). Fig. 1*B* illustrates the differences between these methods and Fisher's. The three methods involve different transformations of the *P* values (see *Methods*), resulting in different shapes of the decision boundary. As can be seen in Fig. 1*B*, the LS and MG methods favor data elements with low *P* values in all data sets (points near the origin). In contrast, Fisher's method

permits the selection of elements as long as there is at least one very small P value. Thus, Fisher's method is most effective when P values for true-positive data elements in different data sets tend not to agree (many data elements scattered near the axes), and the MG or LS methods when data sets tend to agree (many data elements scattered near the origin). Importantly, all of these methods treat both data sets equally, producing a symmetric decision boundary, even though our data are highly asymmetric. The methods we present below overcome this limitation by producing asymmetric decision boundaries (POINTILLIST provides all three methods. The user should select the appropriate method for the given data characteristics). For the remainder of this paper, for simplicity, we will use only Fisher's method without loss of generality.

Weighted Integration of *P* Values. To allow for various levels of statistical power (reliability) in different data sets, we use a weighted method for combining P values. In this approach, P values from different high-throughput assays are transformed, scaled [by a weight representing the reliability (statistical power) of the assay] and summed to generate a combined P value (see *Methods*). The combined P values are then tested by using Monte Carlo generated H_0 distributions as described in *Methods*. Data elements whose combined P values are below a threshold are accepted as components of the system of interest. All other data elements are rejected. The selection threshold corresponds to a significance level (α) reflecting the proportion of area below the decision boundary. Infinitely many combinations of weights can satisfy a given α (Fig. 9). For a given α, we select weights that maximize the number of elements chosen (Fig. 10, elements between the decision boundary and axes), which maximizes the number of true positives selected because the majority of data elements near the axes are true positives (for illustrative examples, see Fig. 1*C*). We used ESA to search for such optimal weights (see *Methods*). Fig. 1*C* compares the unweighted (green line) and weighted (blue line) Fisher's methods for our example data. In this example, P values from data set 2 are twice as reliable as those in data set 1, thus skewing the P_2 values correspondingly closer to 0 (see *Methods*). As shown, a weighted integration method produces an asymmetric decision boundary and can therefore capture more of the true positives.

Selecting a Joint-Significance Threshold (α). As shown in Figs. 7 and 8, the choice of the α-threshold is constrained by the opposing needs to minimize false positives and false negatives. A large α value can produce a high false-positive rate, whereas a small α value can lead to a high false-negative rate. To estimate a suitable threshold α value (α_T), we note that true-positive data elements cluster near the axes whereas the remaining data elements tend to be uniformly distributed. Therefore, one way to select α_T would be to find the value of α for which the resulting decision boundary best separates uniformly distributed data elements from the nonuniform distribution of data elements near the axes. For a given α, the cumulative density of the data elements measures the fraction of data elements selected. The cumulative expected density of the background (uniformly distributed) data elements is α. Thus to compute α_T, we calculate the ratios of cumulative observed densities to cumulative expected densities as α is increased from zero. α_T is the x-axis value of the point nearest to the origin (see *Methods* and Fig. 11). Because the data integration weights are calculated for a given α, it is necessary to recalculate the weights for α_T. For these recalculated weights, another α_T is determined. This procedure is repeated until α_T converges. The black line in Fig. 1*C* shows the decision boundary arrived at using this method for selecting α_T. Note how this boundary is much more stringent, resulting in fewer false positives.

Improved Threshold Significance Level Selection by Using Mixture Distribution Models. The above approach to estimating α_T uses only the distribution of the putative background data elements. As such, this approach focuses on minimizing the false positive rate only. By explicitly dividing the data elements into two populations (a putative true-positive set, H_1, and a putative background set, H_0), we can estimate false-positive and false-negative rates in each set. This approach allows optimization of the membership of H_0 and H_1 to minimize the false-positive and false-negative rates, as follows. (*i*) For an initial value of α (e.g., 0.05), determine the weight vector as described in *Weighed Integration of P Values.* (*ii*) Assume that the full set of observed data elements (green histogram in Fig. 2*A*) divides into two sets H_1 and H_0. (*iii*) The PDF of observed data elements can be approximated by a "mixture model" (MM), M: $\mathrm{PDF}(M) = (1 - \theta)\mathrm{PDF}(H_0) + \theta\mathrm{PDF}(H_1)$, where $0 \leq \theta \leq 1$ is the proportion of H_1 in M (red curve in Fig. 2*A*). (*iv*) For a given set of weights, the PDF of H_0 (dashed blue curve in Fig. 2*A*) can be evaluated numerically by Monte Carlo sampling of a uniform distribution (or the corresponding χ^2, T, or Z distributions, see *Methods* for details). (*v*) Guess an initial value for θ. (*vi*) Estimate the PDF of H_1 (black curve in Fig. 2*A*) as a gamma distribution that best satisfies the relationship in step 2 above using ESA (see *Methods*). (*vii*) Determine α_T as the right-hand tail of H_0 measured from the point where H_1 and H_0 PDFs meet [marked as $c(\alpha_T)$ in Fig. 2*A*]. (*viii*) Repeat steps 1–7 (repeat step 1 by using $\alpha = \alpha_T$) until α_T estimates converge. Finally, the significant elements are selected by using the decision boundary given by α_T.

Fig. 2*B* shows the decision boundaries arrived at through the above procedure (black curve, $\alpha_T = 0.015$) and for $\alpha = 0.05$ for comparison. Note that MM-based estimation of α_T reduces the number of false positives considerably while incurring relatively few additional false negatives (see Table 1, which is published as supporting information on the PNAS web site). The explicit MM also has fewer false negatives than the implicit method presented in the previous section (see Fig. 4 for a quantitative comparison using more demanding data that confirm that the MM method provides better overall performance).

Using a NP Decision Boundary. The above methods are all based on smooth parametric decision boundaries derived from assigning a reliability parameter (the weight) to each type of data. In practice, there are many sources of error for a given measurement error, in turn affecting the resulting P values in an irregular manner. In such cases, smooth parametric decision boundaries may be unable to accommodate these irregularities. To allow estimation of nonsmooth-decision boundaries, we have developed a heuristic NP method, as follows (see *Methods* for details).

First, we construct a first-pass set of candidate true-positive data elements C whose P values are less than a threshold α_T (see *Methods*) in at least one data set. In Fig. 3*A*, these are the data elements between the axes and the dashed blue lines. The remaining elements are grouped into the complement data set C^c. We also generate a set of random background elements (B, see *Methods*). Second, we compute a weight for C and B proportional to the nonoverlapping area between the two PDFs of transformed P values of the elements in C and B (see *Methods*). Third, we compute the combined P value statistic (S_w). Fourth, we quantify the statistical power of the current memberships of C and B as the nonoverlapping area between the corresponding S_w PDFs of C and B. Fifth, if the nonoverlapping area (f) is less than a user-specified threshold (e.g., $f_c = 0.99$), we remove putative false-positive elements with high S_w values from C and add them to C^c. The green crosses in Fig. 3*A* indicate data elements retained in C after this step. Sixth, steps 2 through 5 are repeated by using the new C and C^c sets until f reaches f_c. Finally, once the iteration process is terminated, potential false negatives are identified as elements in C^c whose S_w values exceed that of the point where the S_w PDFs of C and C^c meet. These elements are then added to the final set of candidate true-positive data elements C. Fig. 3*B* shows the final set of selected data elements (green crosses). For comparison, the dashed blue curve shows the decision boundary calculated using the MM (see previous section).

The above procedure is essentially the same as the parametric method described in the preceding section. The main difference is that, rather than moving a decision boundary (as in previous methods), here we move individual data elements between C and C^c. The final decision boundary is irregular because elements in a different part of the P value space are removed as the weights change in each iteration. To speed up the optimization process and avoid potential local minima in the search process, we have limited our optimization process to pruning. To be sure that the final set C includes as many of the true-positive data elements as possible, we initialize C by using a value for α_T that assures the inclusion of most true data elements at the expense of many false positives. These false positives are then removed iteratively. An adaptive rule is used to determine the fraction of C removed in step 5: $C_f = \min[(f_c - f)/f_c, 0.01]$. This rule removes 1% of C in the beginning but decreases the proportion of elements moved as f approaches f_c.

Comparison of the Performance of Different Integration Methods. The example data used in the previous figures were intentionally simplistic for ease of understanding. To highlight differences in the methods when applied to more complex data sets, here we compare the performance of the methods discussed using 10 examples of integration of three data sets. The individual data sets were generated by using the same scheme as described earlier, but with approximately three to four times more noise and less distinct populations of true positives (reduced a and b parameters for the gamma distributions). See *Methods* for details. Fig. 4*A* shows the distribution of the true-positive data elements for an example case.

Fig. 4*B* summarizes the performance of eight different data integration methods on the above 10 examples. For each run of each method, the false-positive rate is plotted along the x axis and the true-positive rate is plotted along the y axis. The 10 examples are identified by numbers 1–10. Each method is identified by a different marker shape and color (see legend). Instances of the letter M mark the mean performance of the correspondingly colored method. In a manner similar to receiver operating characteristic graphs (21), ideal performance is given by $X = 0, Y = 1$. The Euclidean distance from this point represents the degree of performance degradation from the ideal. The dashed contour lines in Fig. 4*B* represent loci of equal performance as defined by the above measure.

The NP and MM methods outperform all other methods on a case-by-case basis and based on average performance. The performance of the union method depends crucially on the choice of significance level [compare U1 ($\alpha = 0.05$) with U2 ($\alpha = 0.15$), which happens to be close to the α of the MM and NP methods]. The intersection method I1 (for which $\alpha = 0.05$) performs worse than all other methods. I2 performance ($\alpha = 0.05^3$) is not indicated because its true-positive rate was significantly <0.5. As expected, the thresholded Fisher's method, in which we minimize the number of false positives automatically, produced the smallest false-positive error rate at the expense of lowering the true-positive rate. However, as shown in Table 1, for less noisy data, the false-positive reduction method can be very effective. Finally, comparison of the weighted and unweighted Fisher's methods highlights the performance advantage of the weighted method for data sets with unequal amounts of noise, whereas the contour gradient between the weighted Fisher's method and the MM and NP methods highlights the importance of optimizing α_T. Overall, MM and NP perform best for minimization of false-positive and false-negative rates, followed by the thresholded Fisher's method, which mainly reduces the false-positive rate.

Handling Missing Values. Missing data points can arise from systematic biases in high-throughput technologies, e.g., favoring high abundance species. Missing values also occur when we integrate global data sets with curated database information. For methods that combine independent, uniformly distributed P values into a uniformly distributed P value, missing values may be handled by using the available P values to compute each combined P value. The simplest way of handling these missing values would be to assign a fixed P value in place of the missing data (e.g., $P = 0.5$ to indicate an equal chance of being a true positive or a true negative, or $P = 1.0$ to discourage elements with missing values from being selected). If one or more data sets include many missing values, the above approach can distort the calculated weights. To avoid such a scenario, we exclude data elements with missing values from the calculation of weights, thereby avoiding the above distortion problem. However, we include them in the data selection process so that data elements with missing values can still be selected.

Conclusions

We have presented a generalized framework for data integration in systems biology. The applicability of our methodology to different types and sizes of data and to different numbers of data sets is demonstrated by application to five different types of data integration in a companion paper (10). Although we focused here on presenting our methodology from the perspective of maximizing statistical power, it can also be applied to scenarios for which the different types of data being integrated have systematic differences between them, for example, combining mRNA and protein abundance measurements or *in vivo* and *in vitro* measurements. Examples of this type of integration are given in the companion paper (10). Data integration can never rule out inclusion of some false positives or loss of some true positives. Our methodology provides a framework for optimizing the tradeoff between these opposing demands. We have implemented all of the methods presented in a free open source software package (POINTILLIST) that allows users to select the integration method most appropriate to their needs.

Our methodology provides a simple and efficient means for combining multiple sets of noisy data to produce probabilistic models. The final outcome of our data integration procedure is a network model in which nodes represent biomolecular species (e.g., genes or proteins) and edges represent interactions (e.g., transcriptional regulation). Our methodology associates a P value with each node and edge in the network model. These P values indicate the degree of confidence in a node or edge being a true component of the system of interest (compared with background/control).

We thank Frederick Roth for insightful and critical analysis of an earlier version of our data integration methodology. A.F.S. holds the Grant I. Butterbaugh Professorship at the University of Washington.

1. Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. (2004) *Science* **306,** 640–643.
2. Ideker, T., Galitski, T. & Hood, L. (2001) *Annu. Rev. Genomics Hum. Genet.* **2,** 343–372.
3. 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.
4. von Mering, C. & Bork, P. (2002) *Nature* **417,** 797–798.
5. Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A. & Gifford, D. K. (2003) *Nat. Biotechnol.* **21,** 1337–1342.
6. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) *Nature* **417,** 399–403.
7. Mrowka, R., Patzak, A. & Herzel, H. (2001) *Genome Res.* **11,** 1971–1973.
8. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002) *Mol. Cell Proteomics* **1,** 349–356.
9. Gilchrist, M. A., Salter, L. A. & Wagner, A. (2004) *Bioinformatics* **20,** 689–700.
10. Hwang, D., Smith, J. J., Leslie, D. M., Weston, A. D., Rust, A. G., Ramsey, S., de Atauri, P., Siegel, A. F., Bolouri, H., Aitchison, J. D. & Hood, L. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 17302–17307.
11. Hedges, L. & Olkin, I. (1985) *Stat. Method Meta-Analysis* (Academic, San Diego).
12. Birnbaum, A. (1954) *J. Am. Stat. Assoc.* **49,** 559–574.
13. Siarry, P., Berthiau, G., Durdin, F. & Haussy, J. (1997) *ACM Trans. Math. Software* **23,** 209–228.
14. Devroye, L. (1997) *ACM Trans. Model. Comp. Simul.* **7,** 447–477.
15. Hildebrand, F. (1956) *Introduction to Numerical Analysis* (McGraw–Hill, New York).
16. Bailey, T. L. & Elkan, C. (1994) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36.
17. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. (2002) *Anal. Chem.* **74,** 5383–5392.
18. Bowman, A. W. & Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis* (Oxford Univ. Press, Oxford).
19. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. & Gerstein, M. (2003) *Science* **302,** 449–453.
20. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R. & Hood, L. (2001) *Science* **292,** 929–934.
21. Swets, J. A. (1988) *Science* **240,** 1285–1293.