that in a specific observation, the subnetwork n can be in any one of $R^{(n)}$ latent physiological states indexed by r. Each state is associated with a unique expression signature $\boldsymbol{s}_r^{(n)}$ over the subnetwork genes. Associations between the observations and the underlying physiological states are unknown, and treated as latent variables. This leads to a mixture model for gene expression in the subnetwork n:

$$\boldsymbol{x}^{(n)} \sim \sum_{r=1}^{R^{(n)}} w_r^{(n)} p(\boldsymbol{x}^{(n)} | \boldsymbol{s}_r^{(n)}, \Sigma_r^{(n)}), \qquad (1)$$

where each component distribution p is assumed to be Gaussian. In practice, we assume a diagonal covariance matrix $\Sigma_r^{(n)}$.

A particular transcriptional response is characterized by the triple $\{\boldsymbol{s}_r^{(n)}, \Sigma_r^{(n)}, w_r^{(n)}\}$. This defines the shape, fluctuations, and frequency of the associated gene expression signature in subnetwork n. The feasibility of the Gaussian modeling assumption is supported by the previous observations of (Kong *et al.*, 2006), where predefined gene sets were used to investigate differences in gene expression between two predefined sample groups. In our model, the subnetworks, transcriptional responses and the activating tissues are learned from data. In one-channel data such as Affymetrix arrays used in this study, the centroids $\boldsymbol{s}_r^{(n)}$ describe absolute expression signals of the preprocessed array data. Relative differences can be investigated by comparing the detected responses. The model is applicable also on two-channel expression data when a common reference sample is used for all arrays since the relative differences are not altered by the choice of comparison baseline when the same baseline is used for all samples.

Now the model has been specified assuming the subnetworks are given. In practice they are learned from the data. In order to do this we make two assumptions. First, we rely on the prior information in the global interaction network, and assume that co-regulated gene groups are connected components in this network. Second, we assume that the subnetworks are independent. This allows a well-defined algorithm, and the subnetworks are then interpretable as independent components of transcriptional regulation. In practice the algorithm, described below, is an agglomerative approximation for searching for locally independent subnetworks.

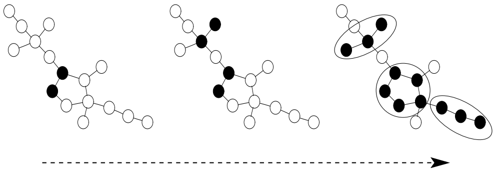

Fig. 2. The agglomerative subnetwork detection procedure. Initially, each gene is assigned in its own singleton subnetwork. Agglomeration proceeds by at each step merging the two neighboring subnetworks that benefit most from joint modeling of their transcriptional responses. This continues until no improvement is obtained by merging the subnetworks.

2.2 Implementation

Efficient implementation is crucial for scalability. For fast computation, we use an agglomerative procedure where interacting genes are gradually merged into larger subnetworks (Fig. 2). Joint modeling of dependent genes reveals coordinated responses and improves the likelihood of the data when compared to independent models, giving the first criterion for merging the subnetworks. However, increasing subnetwork size tends to increase model complexity and the possibility of overfitting since the number of samples remains constant while the dimensionality (subnetwork size) increases. To compensate for this effect, we use a Bayesian information criterion (Gelman *et al.*, 2003) to penalize increasing model complexity and to determine optimal subnetwork size.

The cost function for a subnetwork G is $C(G) = -2L + qLog(N)$, where L is the (marginal) log-likelihood of the data, given the mixture model in Eq. 1, q is the the number of parameters, and N denotes sample size. NetResponse searches for a joint model for the network genes that maximizes the likelihood of observed gene expression, but avoids increasing model complexity through penalizing an increasing number of model parameters. An optimal model is searched for by at each step merging the subnetwork pair that produces the maximal gain in the cost function. More formally, the algorithm merges at each step the subnetwork pair G_i, G_j that minimizes the cost $\Delta\mathcal{C} = -2(L_{i,j} - (L_i + L_j)) + (q_{i,j} - (q_i + q_j))Log(N)$. The agglomerative scheme is as follows:

Initialize: Learn univariate Gaussian mixture for the expression values of each gene, and bivariate joint models for all potential gene pairs with a direct link. Assign each gene into its own singleton subnetwork.

Merge: Merge the neighboring subnetworks G_i, G_j that have a direct link in the network and minimize the difference $\mathcal{C}$. Compute new joint models between the newly merged subnetwork and its neighbors.

Terminate: Continue merging until no improvement is obtained by merging the subnetworks ($\Delta\mathcal{C} \geq 0$).

The number $R^{(n)}$ of distinct transcriptional responses of the subnetwork is unknown, and is estimated with an infinite mixture model. Learning several multivariate Gaussian mixtures between the neighboring subnetworks at each step is a computationally demanding task, in particular when the number of mixture components is unknown. The Gaussian mixtures, including the number of mixture components, are learned with an efficient variational Dirichlet process implementation (Kurihara *et al.*, 2007). The likelihood L in the model is approximated by the lower bound of the variational approximation. The Gaussian mixture detects a particular type of dependency between the genes. In contrast to MATISSE (Ulitsky and Shamir, 2007) and other studies that use correlation or other methods to measure global co-variation, the mixture model detects coordinated responses that can be activated only in a few tissues. Tissue-specific joint regulation indicates functional dependency between the genes but it may have a minor contribution to the overall correlation between gene expression profiles. In principle, we could also model the dependencies in gene fluctuations within each individual response with covariances of the Gaussian components. However, this would heavily increase model complexity, and therefore we leave dependencies in gene-specific fluctuations within each response unmodeled, and focus on modeling differences between the responses. NetResponse provides a full generative model for gene expression, where each subnetwork is described with an independent joint mixture model. The maximum subnetwork size is limited to 20 genes to avoid numerical instabilities in computation. The infinite Gaussian mixture can automatically adapt model complexity to the sample size. We model subnetworks of 1-20 genes across 353 samples; similar dimensionality per sample size has previously been used with variational mixture models (Honkela *et al.*, 2008).

2.3 Data

Pathway interaction network. We investigate the pathway interaction network based on the KEGG database of metabolic pathways (Kanehisa *et al.*, 2008) provided by the SPIA package (Tarca *et al.*, 2009) of BioConductor (www.bioconductor.org). This implements the pathway impact analysis method originally proposed in (Draghici *et al.*, 2007), which is currently the only pathway analysis tool that considers pathway topology.