Supplement for cDREM: Inferring dynamic combinatorial gene regulation
Aaron Wise\textsuperscript{1} and Ziv Bar-Joseph\textsuperscript{1,2}
1 Lane Center for Computational Biology, Carnegie Mellon University
2 Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States
* E-mail: zivbj@cs.cmu.edu

**Supplementary Methods**

**Learning parameters for the group Lasso transition probability**

To solve the sparse group Lasso formulation discussed above we use a modified gradient descent algorithm. We follow the work of Vincent and Hansen (2012). This algorithm involves block gradient descent, where descent is performed group-at-a-time. For each group, we sequentially minimize

\[
\min_{w} Q(x) + \sum_{i=0}^{n_g} \lambda \|w_i\|_1 + \gamma \|w^{(g)}\|_2
\]

which is our objective function with respect to a given group \( g \). \( n_g \) is the number of elements in \( g \) and \( w^{(g)} \) is the block of the weight matrix involving members of group \( g \). \( Q(x) = (\nabla f(w) - Hw)^T x + \frac{1}{2} x^T H x \). Here \( f \) is our loss function and \( H \) is the Hessian matrix. \( Q \) is thus a quadratic approximation of the loss function. At each iteration, we minimize \( Q \) for a coordinate, and then take a step \( \delta \) in the direction of the minimum from our present weights. For more details, see Vincent and Hansen (2012). For each \( w_i \) in the group, this minimization problem can be written as

\[
\arg \min_{w_i} cw + \frac{1}{2} hw_i^2 + \gamma \sqrt{w_i^2 + r + \lambda |w_i|}
\]

where \( c = g^{(g)}_i + \sum_{j \neq i} (H_{gg})_{ij} w_j \) (\( g^{(g)}_i \) is the gradient of \( w_i \), and \( H_{gg} \) is the block Hessian), \( h \) is the \( i \)th diagonal element of \( H_{gg} \) and \( r = \sum_{j \neq i} x_j^2 \). In many cases, it is easy to tell when the gradient is 0 before the minimization has been solved; this results in a substantial speedup in practice. Still, this algorithm is more computationally intensive than the standard gradient descent algorithm in DREM, and so we use GPU computing (CUDA) to make the computations faster.
Supplementary Results

Comparison of cDREM combinatorial pairs to combinatorial pairs from DREM

The original version of DREM, like a number of other regulatory network reconstruction methods (Aderhold et al., 2013) relies on regression analysis that can only model additive TF activity. To test if cDREM, which can explicitly model logical AND and OR relationships, improves upon regression based methods, we compared the set of TF pairs identified by cDREM and DREM. To generate predicted pairs from DREM we performed the following steps. We first used DREM to learn models for each condition discussed above. We then extracted the sets of TFs identified by DREM for each split as important (based on p-value and model score) and considered all pairs of TFs in each set. We screened these sets following our pre-processing criteria to select pairs with a significant target overlap. This process resulted in 410 pairs identified by DREM over all four yeast data sets (compared to cDREM’s 93). We compared the resulting sets using the complementary data discussed above. We found that the BioGRID enrichment for the DREM pairs was much lower than cDREM’s pairs (20.5% vs. 30.1%). Also, the cDREM pairs had significantly higher knockout expression similarity ($p = 0.0005$, using the Mann–Whitney U test) and chIP-chip colocalization ($p = 0.047$), as well as borderline significantly higher knockout phenotype correlation ($p = 0.064$).

Comparison of AND and OR pairs

The above analysis was performed on the AND pairs. In contrast, for the OR pairs selected by cDREM we do not expect to see similar results when using the complementary datasets discussed above. For example, unlike AND pairs, OR pairs are not expected to have similar knockout expression patterns (since joint targets should not be affected by a single knockout) and the same holds for phenotypic screen outcomes. Further, since they work independently we do not expect them bind in close proximity. Our analysis confirms this hypothesis. In sharp contrast to the AND set, the OR pairs identified by cDREM are not significantly different from the overall set of TF pairs when using these complimentary datasets: knockout expression ($p = 0.428$), knockout phenotype correlation ($p = 0.994$) or chIP-chip colocalization ($p = 0.136$). Thus, cDREM was able to correctly identify not only pairs of jointly active TFs but also their mode of interaction.

GO analysis of the Human Flu Model

To analyze the model on a more global basis we looked at the GO enrichment of genes on paths in the model. Each path in the reconstructed network contains a set of genes that are predicted to be co-expressed and co-regulated. We list the 5 GO categories with the highest GO enrichment score among all paths in
Table 1. Top 5 enriched GO categories in the human flu reconstructed network.

<table>
<thead>
<tr>
<th>GO Term</th>
<th>P-value in cDREM model</th>
<th>P-value in baseline model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response to virus</td>
<td>4.6 * 10^{-32}</td>
<td>1.9 * 10^{-31}</td>
</tr>
<tr>
<td>Defense response to virus</td>
<td>4.1 * 10^{-30}</td>
<td>1.3 * 10^{-29}</td>
</tr>
<tr>
<td>Innate immune response</td>
<td>5.4 * 10^{-30}</td>
<td>3.1 * 10^{-29}</td>
</tr>
<tr>
<td>Defense response</td>
<td>3.7 * 10^{-29}</td>
<td>3.3 * 10^{-29}</td>
</tr>
<tr>
<td>Immune response</td>
<td>2.8 * 10^{-26}</td>
<td>2.2 * 10^{-26}</td>
</tr>
</tbody>
</table>

As can be seen, these GO terms are clearly the most relevant to the experiment, and include viral response, defense response and immune activity. We next compared the GO enrichments between a cDREM model that uses the combinatorial input set and a model that can only use individual factors. Interestingly, even though all experimental data used to generate the target sets came from single TF experiments, it turns out that by using our combinatorial analysis methods we can improve the assignment of genes to paths. For the top 3 GO categories the combinatorial model outperforms the individual TF regulation model indicating that the model captures part of the underlying mechanisms used to regulate genes.

References