Exploring the independence of gene regulatory modules

Janne Nikkilä, Antti Honkela, and Samuel Kaski

Adaptive Informatics Research Centre,
Helsinki University of Technology, FI-02015 TKK, Finland,
{janne.nikkila, antti.honkela, samuel.kaski}@tkk.fi,
WWW home page: http://www.cis.hut.fi/

Abstract. We study the discovery of gene regulatory modules based on transcription factor (TF) binding data and expression data from gene knockouts. We invoke the natural assumption that regulatory modules predominantly operate independently, which makes it possible to apply a new method for extracting them: the Independent Variable Group Analysis. We demonstrate that i) the independence assumption helps in discovering the regulatory modules from TF data, and ii) the independent gene modules discovered from TF-data can be found also in expression data from gene knockouts. This demonstrates that the regulatory effects by transcription factors are observable in knockout experiments. It additionally suggests that the difficult interpretation of the knockout experiments could be eased by taking into account the independent regulatory modules.

1 Background

Gene regulatory interactions are one of the main foci in systems biology at the moment. A regulatory interaction between a gene and a set of other genes is formed when the protein produced by the gene binds the other genes' DNA sequences and affects their expression. These proteins are called *transcription factors (TFs)*, and genes are known to often be regulated by them as groups or modules. The knowledge of the regulators and their targets is obviously of practical significance, for example, in various diseases in humans as well as in simple organisms like yeasts used in diverse tasks in bioprocess industry.

A direct laboratory technique to study gene regulatory interactions is to monitor which proteins bind which genes' DNA sequence with so-called ChIP on chip technique [1]. It produces data about the binding strength of the chosen TFs to all genes of the organism, and implicit information about which genes are regulated by the same TFs (i.e., potentially belong to the same regulatory module). However, the measurement techniques are imperfect and expensive, providing only scarce and noisy data about the *potential*, individual regulatory relationships. While a single TF-gene relation is not necessarily reliable, the TF binding data can be used to estimate the groupwise behaviour of the TFs. The

problem here is that the characteristics of the gene regulatory modules in terms of TF binding are not fully understood yet.

We will define and discover regulatory modules by invoking the assumption of the statistical independence of the regulatory modules. We apply a recently proposed method, *independent variable group analysis (IVGA)* [2, 3] to find independently regulated groups of genes from TF binding data. Although it would be possible to integrate various data sources to discover general regulatory modules, we focus here specifically on the regulatory modules represented by transcription factor binding. This has at least three justifications: i) ChIP on chip measurements are one of the most reliable data about the regulatory interactions in the cell, ii) they directly measure perhaps the most interesting type of regulatory interaction present in the cell, and iii) the relationship between the ChIP on chip measurements to other measurement techniques (for example, gene expression) is still unclear and the results can then be contrasted to these.

We validate the independence of the obtained gene groups by "simulating" wetlab experiments: we will measure whether the found regulatory modules are independent in new data from different measurement technique as well. As "new" data we will use previously published gene expression data from gene knockout experiments [4]. If the findings are favourable, the conclusion is that suppression of a gene from one module does not, on average, affect too much on other regulatory module activities i.e. expression. More importantly, this effect is then discernible in very commonly used knockout expression measurements.

In the literature, it has been observed previously that TF binding and knockout experiments from the same organism are dependent, see for example [5], and that the genes are regulated as modules, see for example [6, 7]. The novelty in our approach is that we focus on the gene modules discoverable from regulatory binding information by the independence assumption, and succesfully validate the modules with an independently measured knockout data. We thus provide evidence that i) independence assumption helps in discovering the gene regulatory modules from TF binding data, and ii) that the knockout experiments are capable of discovering independent, in terms of TF binding, regulatory modules.

2 Independent variable group analysis (IVGA)

Independent Variable Group Analysis (IVGA) [2, 3] is a principle for grouping variables that are mutually dependent together so that independent or only weakly dependent variables are placed to different groups. One natural criterion for solving the IVGA problem is to minimize the mutual information or in general multi-information, within the grouping evaluated by considering each group a separate random variable.

In more conventional terms, IVGA can be seen as a clustering method where samples are taken as random variables and the criterion is to minimize the mutual information between the groups. A similar criterion has been used for hierarchical clustering in [8]. Direct evaluation of mutual information within high dimensional distributions is difficult, but a connection to Bayesian meth-

ods demonstrated in [3] allows an efficient model-based approximation based on maximising the sum of marginal log-likelihoods of independent models for each group. This problem can be solved efficiently using variational Bayesian learning.

The IVGA algorithm is based on a heuristic combinatorial optimization method for finding an optimal grouping, combined with variational Bayesian learning to fit Gaussian mixture models for the groups and evaluate the objective function. The algorithm is initialized by placing each variable in a group of its own. It proceeds by moving variables, as well as splitting and joining groups, in such a way that the sum of the approximate marginal log-likelihoods always increases. More details of the algorithm can be found in [3]. A small change was made in order to help avoid local minima in the grouping phase by clearing the cache of mixture models for the groups at fixed intervals and all the models were refitted from scratch. This improved the training set marginal likelihood of the attained results significantly. The same IVGA method can also be used to evaluate the approximate mutual information to compare groupings by only fitting the Gaussian mixture models and evaluating the approximate marginal log-likelihood.

The actual objective function for IVGA can be derived by assuming that the data set $\boldsymbol{X}$ consists of vectors $\mathbf{x}(t), t = 1, \ldots, T$. The vectors are N-dimensional with the individual components denoted by $x_j, j = 1, \ldots, N$, and all observed x_j by $\boldsymbol{X}_j = (x_j(1), \ldots, x_j(T))$. The aim here is to find a partition of $\{1, \ldots, N\}$ to M disjoint sets $\mathcal{G} = \{\mathcal{G}_i | i = 1, \ldots, M\}$ such that the mutual information

$$I_{\mathcal{G}}(\mathbf{x}) = \sum_i H(\{x_j | j \in \mathcal{G}_i\}) - H(\mathbf{x}) \approx -\frac{1}{T} \sum_i \log p(\{\boldsymbol{X}_j | j \in \mathcal{G}_i\} | \mathcal{H}_i) - H(\mathbf{x}) \tag{1}$$

between the sets is minimised. Here $\mathcal{H}_i$ denotes the model for the ith group. In order to avoid evaluation of the constant $H(\mathbf{x})$ and to get more easily interpretable results, only the differences in estimated values of mutual information will be reported. Such differences are actually logarithms of Bayes factors between the models, divided by the number of samples.

3 Case study and results

3.1 Estimating the gene groups with TF-data

The TF binding data we used had 355 different combinations of TF and experimental conditions for 6229 yeast *Saccharomyces cerevisae* genes [1]. The intensity ratios in the data were $\log_2$ transformed, the detailed description of the generation of the data can be found in [1]. IVGA was estimated with TF-data once, starting from a random initialization. The densities of TFs were modeled with mixtures of Gaussians with diagonal covariance matrices, and the genes were grouped with a greedy search algorithm. The gene grouping from the best model achieved during training, in the sense of the cost function evaluated for the part of the TF-data consisting of the 215 knocked out genes in the validation data, was chosen for validation. The grouping consisted of 166 groups. The cost

function of only part of the data was used because optimization of the grouping is a very difficult problem and the result may occasionally improve other parts more at the cost of the part actually used for validation.

To get a comparison result, in an analogous fashion to IVGA estimation we grouped the genes with the standard clustering method K-means. Since the K-means is computationally less intensive, we fitted K-means 50 times with $K = 166$ from different initializations. Additionally, we made 50 random groupings for the genes to get a baseline result.

Additionally we estimated IVGA on the part of the TF-data consisting of the 215 knocked out genes in the validation data. The resulting grouping consisted of 5 groups. Similarly as before, we also fitted K-means to the smaller data set 50 times with $K = 5$.

3.2 Validating the groups with knockout data

The knockout expression data that was used to validate the obtained IVGA grouping consisted of genomewide (6308 genes) microarray expression measurements for 215 different gene knockouts [4]. We used only the knockout measurements made for the yeast strain grown in YPD medium, and in a form of logarithmic intensity ratios. Since groupings from the TF data were now for all the yeast genes, the 215 knocked out genes could be mapped to their respective groups. The experimental setup is illustrated in Fig. 1. The IVGA cost was then computed for the groupings from IVGA, K-means and random groupings of knockouts by fitting the mixture model into gene densities. In order to evaluate the accuracy of the cost function evaluation, the procedure was repeated for some test cases by fitting the models 5 times from different mixture model initializations. The resulting variation was found negligible in comparison to variation between different methods.

The knocked out genes mapped to 94 different groups in the IVGA grouping with 166 groups. The IVGA grouping was among the best K-means groupings, but not significantly better. Both of these were significantly better than random groupings with the difference between the means of K-means and the random groupings being 4.2 nats. The difference between the K-means and the random groupings is statistically significant (two-sample t-test, $p < 1.4 \cdot 10^{-48}$).

For the 5 groups the IVGA grouping yielded 0.56 nats better cost than the best K-means result. Overall, the difference between the methods was clearly significant (tail probability of the IVGA result for a normal distribution of K-means results, $p < 2.7 \cdot 10^{-6}$).

3.3 Interpreting the independence with Gene Ontology (GO)

The gene groups were also validated by studying the enrichments of the known Gene Ontology classes to the clusters, and the overlaps of the GO classes between different clusters. If statistically significantly many genes belong to the same GO class in a cluster, the cluster is then likely to have a meaningful biological interpretation. The other side of the coin is whether the same GO classes are

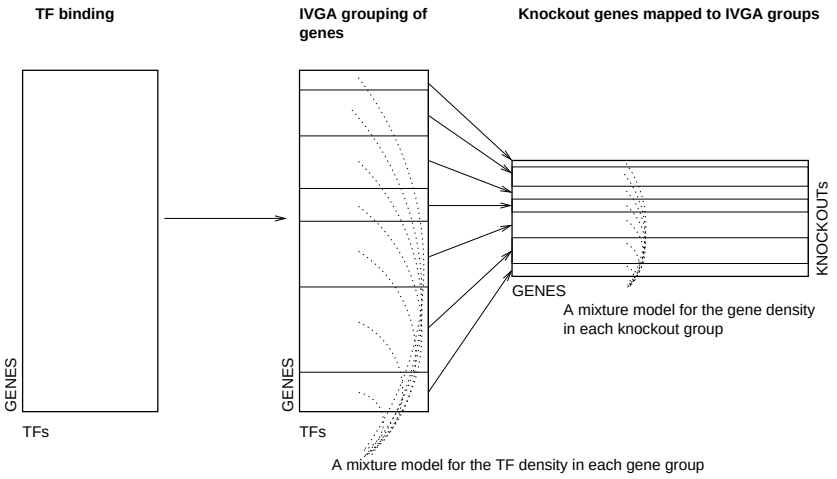

Fig. 1. A schematic illustration of the procedure of estimating the independence of gene regulatory modeles with IVGA applied on TF data, and the validation of the modules by applying IVGA with the fixed grouping to gene knockout expression data. The boxes with bold borders are data matrices, the thin lines represent the IVGA group borders.

represented in several clusters: if different clusters have different GO enrichments, the clusters are more likely to be independent in a biological sense.

The enrichments were tested for most independent groupings from both IVGA and K-means with the simple Fisher's exact test, revealing that in both the 159 out of 166 clusters in the bigger cluster set had statistically significant enrichments ($p < 0.01$). Moreover, out of 13695 possible cluster pairs, in both methods only around 400 pairs had an overlap in GO class enrichments.

In the smaller cluster set from both methods all 5 clusters had significant GO class enrichments ($p < 0.01$), and none of the GO classes were present in more than one cluster.

All cluster sets thus show clear evidence for independence also on the biological level.

4 Discussion

We presented a method for searching for the gene regulatory modules in yeast from TF-binding data based on their mutual indepedences. We showed that the found modules generalized to independently measured expression data from gene knockout experiments.

The results provide evidence that at least some regulatory modules can be assumed to function independently and, more importantly, this independence can be observed also in the expression data from knockout experiments. While some groups can be discovered using conventional clustering (here K-means) the

explicit minimization of their mutual information (IVGA) seems to improve the generalizability of the results.

The independence and meaningfulness of the groups in a biological sense was supported by strong enrichments of GO classes in clusters, and relatively small overlaps of the classes between the clusters. This result also suggests that the set of independent regulatory modules could be useful in interpreting regulatory effects in the knockout experiments. One of the usual problems in the interpretation of the knockout expression data is that the knockout often induces also secondary effects resulting from, for instance, the cell trying to compensate the missing gene, or being driven to another metabolic state.

IVGA shows promise in exploring the independences between gene groups. Future improvements on optimization and alternative choices for the probabilistic models should still improve the results. Note that reason for using IVGA cost function to evaluate groupings is that it measures precisely what we are interested in, i.e. mutual information between the groups.

References

1. Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., Yoo, J., Jennings, E., Zeitlinger, J., Pokholok, D., Kellis, M., Rolfe, P., Takusagawa, K., Lander, E., Gifford, D., Fraenkel, E., Young, R.: Transcriptional regulatory code of a eukaryotic genome. Nature **431**(7004) (2004) 99–104
2. Lagus, K., Alhoniemi, E., Valpola, H.: Independent variable group analysis. In G.Dorffner, H., K.Horenik, eds.: International Conference on Artificial Neural Networks - ICANN 2001. (2001) 203–210
3. Alhoniemi, E., Honkela, A., Lagus, K., Seppä, J., Wagner, P., Valpola, H.: Compact modeling of data using independent variable group analysis. Technical Report E3, Helsinki University of Technology, Publications in Computer and Information Science, Espoo, Finland (2006)
4. Mnaimneh, S., Davierwalak, A.P., Haynes, J., Moffat, J., Peng, W.T., Zhang, W., Yang, X., Pootoolaland, J., Chua, G., Lopez, A., Trochesset, M., Morse, D., Krogan, N.J., Hiley, S.L., Li, Z., Morris, Q., Grigull, J., Mitsakakis, N., Roberts, J., Greenblatt, J.F., Boone, C., Kaiser, C.A., Andrews, B.J., Hughes, T.R.: Exploration of essential gene functions via titratable allels. Cell **118** (2004) 31–44
5. Kaski, S., Nikkilä, J., Sinkkonen, J., Lahti, L., Knuuttila, J., Roos, C.: Associative clustering for exploring dependencies between functional genomics data sets. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2 **2**(3) (2005) 203–216
6. Slonim, N., Elemento, O., Tavazoie, S.: Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. Molecular Systems Biology **2** (2006)
7. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Friedman, N.: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics **34** (2003) 166–176
8. Kraskov, A., Stögbauer, H., Andrzejak, R.G., Grassberger, P.: Hierarchical clustering using mutual information. Europhysics Letters **70**(2) (2005) 278–284