Explaining gene responses by linear modeling

Yvonne Poeschl, Ivo Grosse, and Andreas Gogol-Döring

German Center of Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany
Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany

yvonne.poeschl@informatik.uni-halle.de

Abstract: Increasing our knowledge about molecular processes in response to a certain treatment or infection in plants, insects, or other organisms requires the identification of the genes involved in this response. In this paper, we propose the *Profile Interaction Finder* (PIF) to identify such genes from gene expression data which is based on a convex linear model, and we investigate its efficacy for two applications related to stimulus response. First, we seek to identify sets of putative regulatory genes that explain the expression levels of a gene under different stimuli best. Second, we aim at identifying genes that show a specific response to a stimulus or a combination of stimuli. For both applications, we study the expression response of two *Arabidopsis* species to treatment with the plant hormone auxin and of *Apis mellifera* to pathogen infection. The proposed approach may be of general utility for analyzing expression data with a focus on genes and gene sets that explain specific stimulus response.

1 Introduction

Genes in a living cell form a complex network in which the expression level of each gene, i.e., the concentration of messenger RNA molecules, depends on the expression level of other genes. For instance, the expression of a gene encoding a transcription factor (TF) could rise because of an external stimulus, which consequently influences – directly or indirectly – the transcription of tens or hundreds of other genes.

The investigation of the causal effects between one TF and its target genes is a difficult task requiring complex laboratory experiments. Fortunately, it is possible to get indications of potential regulatory relationships between genes by comparing their expression levels under different conditions, e.g., before and after stimulation. A variety of methods have been developed for this purpose, for a recent review see [WH14]. The higher the number of the involved data sets, and the more different conditions (treatments, time points, cell types, pathogens, etc.) are covered, the more detailed and accurate a prediction of the regulatory network could be. The underlying assumption is that genes that are closely connected in the regulatory network will also tend to have similar expression patterns under varying conditions. Mathematically speaking, gene expression levels obtained from M experiments can be represented by an M-dimensional vector, and if two genes are neighbors within this M-dimensional space, they presumably have a tight relation to each

other in terms of their regulation.

A conventional clustering method like HCLUST [MC11] relying only on the relation between pairs of genes sometimes fails to model cases in which one gene is jointly regulated by several other genes while it is only loosely correlated with each individual regulator. The Local Context Finder (LCF) introduced by Katagiri and Glazebrook [KG03] addresses this problem by reconstructing the M-dimensional expression profile of a gene as a linear combination of the expression profiles of other (neighboring) genes. One limitation of this approach is that it does not regard anti-correlated expression profiles. Although it is well known that, e.g., TFs could either increase or suppress the transcription of target genes, the latter case is not considered by the LCF.

In this paper we propose a new approach, the *Profile Interaction Finder* (PIF), that uses a distance metric that takes into account both positive and negative correlations. The approach selects for each gene a set of neighboring reference profiles that together explain the expression values of the gene best. Reference profiles are either expression profiles of other genes, which possibly have a regulatory influence on the current gene, or prototype profiles that reflect in which data set a certain experimental condition was present or absent. The proposed approach extends the LCF in two aspects, namely by considering both positive and negative interactions, and by using the flexible and generalizing notion of *reference profiles*. These extensions are instrumental in answering two central questions when analyzing expression data: (i) Which genes might have a positive or negative influence on the expression pattern of other genes?, And (ii) which genes respond positively or negatively to certain experimental conditions?

2 Methods

Supposed that we measure the expression of genes under varying conditions in M different experiments. To each gene we assign an *expression profile* $\underline{x} = (x_1, \ldots, x_M)$ containing the expression values of this gene. All expression profiles are normalized using a linear transformation such that the length $\|\underline{x}\| = 1$ and the mean $\bar{\underline{x}} = 0$. This normalization does not affect the Pearson correlation coefficient between two profiles $\underline{x}$ and $\underline{y}$, but it simplifies its calculation as the dot product $\underline{x} \cdot \underline{y}^T$ which can be interpreted as the cosine of the angle between the two vectors.

The goal of the proposed algorithm is to approximate a given expression profile by a linear model of *reference profiles* that could be either expression profiles of other genes or artificially created prototype profiles describing experimental conditions. Supposed for example that we set $n_m = 1$ if the m-th experiment is measured under a certain condition c, and $n_m = 0$ otherwise, then $\underline{n} = (n_1, \ldots, n_M)$ is after normalization a prototype profile for the condition 'measured on condition c'. More detailed examples for prototype profiles will be given in Section 3.2.

PIF returns for each gene $\underline{x}$ a set of *neighboring profiles* which are most informative for predicting $\underline{x}$. The proposed approach consists of three steps: (i) PIF first selects candidate reference profiles $\underline{n}_1, \ldots, \underline{n}_K$ related to $\underline{x}$ (Section 2.1), which (ii) are used to reconstruct $\underline{x}$

by a linear model (Section 2.2), and finally (iii) the results are filtered using bootstrapping (Section 2.3).

If gene expression profiles are used as references, the output could be interpreted as a gene regulatory network in which every gene is linked to all genes in its neighborhood. In case of prototype profiles, the genes could be sorted into clusters according to their neighborhoods. Examples for both applications are discussed in Section 3.

2.1 Selection of Reference Profiles

Fitting a linear model to a given input profile $\underline{x}$ could be computationally demanding, especially if the number of reference profiles is large. We therefore restrict the calculation to the subset of reference profiles that are most appropriate for reconstructing the input profile. This filtering process reduces computational costs and also improves the quality of the reconstruction by reducing noise.

For scoring the predictive power of a reference profile $\underline{n}$ relative to $\underline{x}$, we first compute the Pearson correlation coefficient between the two profiles. If this value is either close to 1 (positive correlation) or close to -1 (anti-correlation), then the two profiles are strongly connected, and in both cases the reference profile would be appropriate for reconstructing the input profile. A correlation coefficient of 0 on the other hand means that both vectors are orthogonal and no information about the input profile could be derived from the reference profile. The absolute value of the correlation coefficient $s = |\underline{x} \cdot \underline{n}^T|$ is a good indicator for the applicability of $\underline{n}$ for reconstructing $\underline{x}$. In contrast to the LCF given in [KG03], which only chooses reference profiles with maximum *positive* dot product, PIF also takes highly informative reference profiles with *negative* dot product into account.

We select at most K profiles with maximal score $s \geq t$, where t is a user-defined threshold. A high value of t ensures that only reference profiles in close proximity to the input profile are used, whereas with $t = 0$ the filtering step would be omitted completely. In this paper we use $K = 10$ and $t = 0.25$.

2.2 Linear Model Reconstruction

In the main step of our approach, we reconstruct the input profile $\underline{x}$ as a linear combination of the reference profiles $\underline{n}_1, \ldots, \underline{n}_K$ selected in step (i) (Section 2.1). We calculate non-negative weights $w_1, \ldots, w_K$ by a constrained linear fit such that the squared error function $f(\underline{w})$ is minimized,

$$f(\underline{w}) \quad = \quad \left\| \underline{x} - \sum_{k=1}^{K} w_k \mu_k \underline{n}_k \right\|^2 \quad \text{and} \quad 1 = \sum_{k=1}^{K} w_k, \tag{1}$$

where $\underline{\mu} = (\mu_1, \ldots, \mu_K) \in \{-1, 1\}^K$ denotes the signs of the dot products $\underline{x} \cdot \underline{n}_k^T$, i.e., $\mu_k = 1$ if $\underline{x} \cdot \underline{n}_k^T \geq 0$, and $\mu_k = -1$ if $\underline{x} \cdot \underline{n}_k^T < 0$. For reference profiles $\underline{n}_k$ that are anti-correlated to $\underline{x}$ the factor $\mu_k = -1$ reverts the direction of the reference profile such that the resulting profile $\underline{v}_k = \mu_k \underline{n}_k$ and $\underline{x}$ are *positively* correlated. This reduces the reconstruction to a convex linear combination, where all weights w_k are non-negative and sum to one.

We reformulate the optimization problem by including the constraint on the weights by introducing the Lagrangian multiplier λ:

$$L(\underline{w}, \lambda) \quad = \quad \left\| \underline{x} - \sum_{k=1}^{K} w_k \underline{v}_k \right\|^2 + \lambda \left(1 - \left(\sum_{k=1}^{K} w_k \right) \right) \tag{2}$$

We minimize $L(\underline{w}, \lambda)$ in eq. 2 by computing the derivatives for all w_k and then use the constraint in eq. 2 to compute λ, yielding

$$w_k = \sum_{j=1}^{K} s_{k,j}^{-1} \left(\frac{\lambda}{2} + \underline{v}_j \underline{x}^T \right) \qquad \lambda = 2 \cdot \frac{1 - \left(\sum_{j=1}^{K} \underline{v}_j \underline{x}^T \left(\sum_{k=1}^{K} s_{k,j}^{-1} \right) \right)}{\sum_{j=1}^{K} \sum_{k=1}^{K} s_{jk}^{-1}}, \tag{3}$$

where $\underline{v}_j = \mu_j \underline{n}_j$, and $\underline{\underline{s}}^{-1}$ is the inverse of $\underline{\underline{s}} = \underline{\underline{v}} \cdot \underline{\underline{v}}^T$ with $\underline{\underline{v}} = (\underline{v}_1, \ldots, \underline{v}_K)^T$. If $\underline{\underline{s}}$ becomes singular due to the linear dependence of some reference profiles, we compute the pseudo-inverse as suggested by [RS00].

The intended reconstruction of $\underline{x}$ is then given by the linear combination $\underline{r} = \sum_{k=1}^{K} w_k \mu_k \underline{n}_k$.

2.3 Determining Robust Neighborhoods

The weights $w_1, \ldots, w_K$ calculated in the previous section can be interpreted as degrees of relative importance of the reference profiles $\underline{n}_1, \ldots, \underline{n}_K$ for the explanation of an expression profile $\underline{x}$. Reference profiles $\underline{n}_k$ with a low weight w_k are likely expendable. Given a user-defined threshold r, we call the set $\{\underline{n}_k | w_k \geq r\}$ of all reference profiles with weights of at least r the *neighborhood* of $\underline{x}$. In this paper, we set $r = 0.1$.

The approach comprised of step (i) and (ii) (Section 2.1 and 2.2) described so far could be affected by noise in the gene expression data. Hence, we use bootstrapping in order to increase the reliability of the results. Given a data set with M samples, bootstrapping samples M out of these M samples with replacement, and we apply PIF to this sampled data set. We perform this bootstrapping step $L = 1000$ times and keep only reference profiles in the neighborhood of a gene which occurred in this neighborhood for at least p percent of the L repeats. In this paper, we use a thresholds of $p = 50\%$ for gene expression reference profiles (Section 3.1) and $p = 75\%$ for prototype profiles (Section 3.2).

3 Results

We will now investigate if PIF is capable of producing biologically relevant results when applied to reconstructing gene regulatory networks (Section 3.1) and to clustering genes according to experimental conditions (Section 3.2).

3.1 Reconstruction of Regulatory Networks

Auxin is one of the key phytohormones that controls plant development and growth. So far, only parts of auxin signaling are understood [DRQ08]. For the identification of novel candidate genes that might be involved in auxin signaling network, we applied PIF on a time-series of gene expression data of the two closely related plant species *Arabidopsis thaliana* and *Arabidopsis lyrata*, measured using expression microarrays at 0, 1, and 3 hours after auxin treatment. Each measurement was repeated three times, yielding $M = 2 \times 3 \times 3 = 18$ data sets.

We processed and normalized the raw data as described in [P$^+$13]. 9091 genes with a coefficient of variation above 0.05 were selected for further analysis. Each of these genes could be regulated either enhanced or repressed by any of the other genes, so we used the expression profiles of all 9091 genes as possible reference profiles.

Figure 1 shows a part of the reconstructed gene network connected to the well known auxin responsive gene *AT5G54510* that is up-regulated upon auxin stimulation. According to the PIF analysis, AT5G54510 is part of the neighborhoods of four other genes. The correlation coefficients between AT5G54510 and the two genes AT3G58190 and AT4G37295 are positive, so AT5G54510 might have an enhancing effect on their expression. In contrast to that, the correlation coefficients to the other two target genes AT4G10270 and AT3G10040 are negative, suggesting that AT5G54510 possibly suppresses their transcription.

None of the four genes related to AT5G54510 had been identified to be involved in the auxin signaling pathway. Nevertheless, especially AT3G58190 seems to be very likely involved in hormone signaling, since this gene is also connected to two more factors AT4G14560 and AT4G27260 both related to the hormone metabolism.

3.2 Prototype Analysis

In addition to the reconstruction of gene regulatory networks we can use the *Arabidopsis* data from the previous section to address various further questions. Examples are: 'Which genes respond quickly, or with a delay to auxin stimulation?' or 'Which genes are regulated differently in the two species?'. PIF is capable of answering these questions by using prototype profiles that reflect the different time points and species of the data sets (Figure 2A). Figure 2B-D shows an example of the results of this analysis, a cluster of 16 genes initially highly expressed in both species and later down-regulated, but more

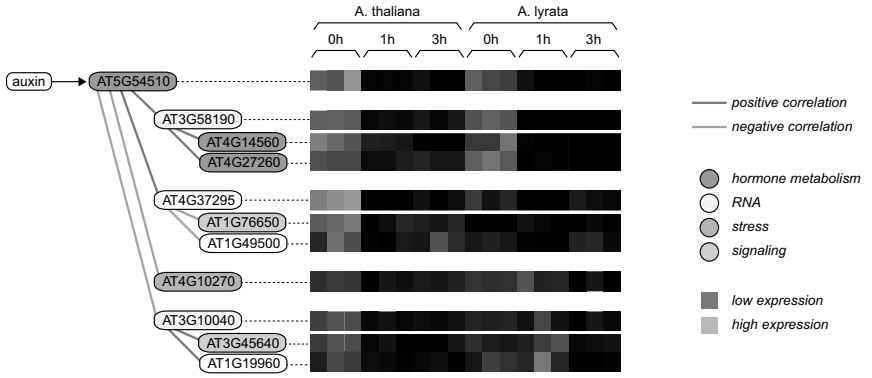

Figure 1: A part of the regulatory network for *Arabidopsis* reconstructed by PIF showing genes related to the auxin responsive gene *AT5G54510*. Genes connected with red edges are positively correlated; green edges mean negative correlation. The gene colors correspond to the GO-terms they are annotated with using MapMan [T$^+$04]. The heat map shows the expression levels of the 11 genes for each of the 18 data set. Red fields mean that the gene is highly expressed due to auxin treatment, while green fields mean low expression.

strongly in *Arabidopsis thaliana* than in *Arabidopsis lyrata*.

This expression pattern is described by a combination of three prototype profiles (Figure 2B). Each single prototype profile differs strongly from the expression profiles of the genes in this cluster (Figure 2C and D), so the cluster could only be found because PIF reconstructs expression profiles by combining several reference profiles (Section 2.2).

Statistical analysis reveals that for the GO-term [T$^+$04] 'RNA' the number of annotated genes in this cluster is significantly higher than expected (p-value > 0.05, Fisher's exact test). This indicates that PIF possibly sorted the genes into biological meaningful clusters.

To investigate if PIF could also handle more diverse input data, we applied it to multiple data sets collected for a metastudy [Tra14] concerning the impact of different pathogens on gene expression in honeybee (*Apis mellifera*), see Table 1. The expression data were collected from different sources, measured for different tissues and on different platforms, and preprocessed with different methods, so they have very different dynamic ranges. Hence, we decided to use relative ranks [BAAH04] instead of raw gene expressions as input for PIF.

We group 6242 genes present in all 9 data sets according to their response pattern to different experimental conditions, namely pathogens and tissues, see Figure 2A. Figure 2B-D show the example of a gene cluster containing 15 genes that respond positively to nosema infection in the fat body but negatively in the gut. Gut and fat body are distinct parts of the honeybee abdomen; genes in this group may be related to the immune response activated due to the infection. Although the individual genes within the clusters are more diverse than in the data set for *Arabidopsis*, their expression profiles broadly follow the pattern defined by the prototypes.

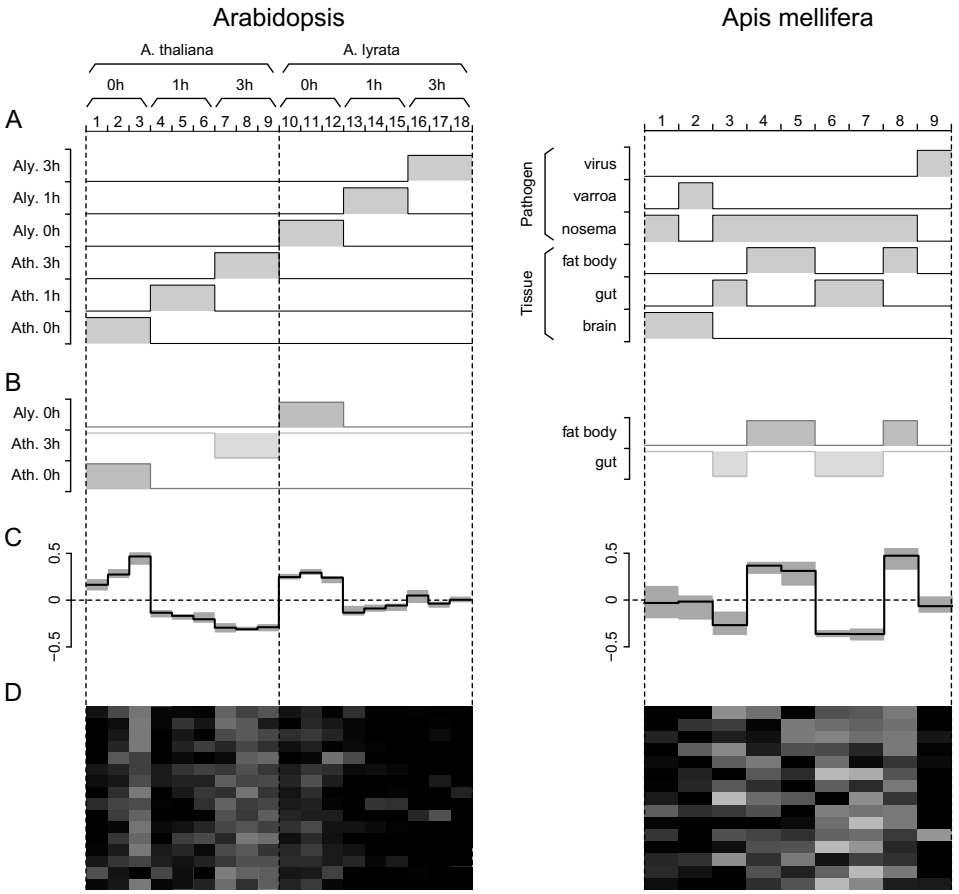

Figure 2: PIF analysis using prototype profiles as reference. The left panel shows results of the *Arabidopsis* data analysis, and the right panel shows results of the *Apis mellifera* metastudy. A: The complete set of prototype profiles (before normalization) used in the analysis. B: Neighboring prototype profiles for one selected gene cluster. Prototype profiles which correlate positively to the genes in the cluster are shown in red; anti-correlations are shown in green. C: Averaged expression profiles of the genes in the cluster. The orange boxes show the area between the first and the third quartile. D: Heat maps showing the expression profiles of genes in the cluster. Each line represents one gene. Red boxes show highly expressed/up-regulated genes, and green boxes show low expressed/down-regulated genes.

	Pathogen	Tissue	Platform	Source
1	Nosema	brain	RNA-seq	[M+13]
2	Varroa	brain	RNA-seq	[M+13]
3	Nosema	gut	tiling microarray	[D+12]
4	Nosema	fat body	expression microarray	[HAG13]
5	Nosema	fat body	expression microarray	[HAG13]
6	Nosema	gut	expression microarray	[HAG13]
7	Nosema	gut	expression microarray	[HAG13]
8	Nosema	fat body	expression microarray	[HAG13]
9	Virus	whole bee	expression microarray	[FA13]

Table 1: List of data sets used in the metastudy of *A. mellifera*.

4 Conclusions

The identification of genes acting as regulators of other genes or responding specifically to certain experimental conditions is an important aspect of gaining knowledge about gene regulatory processes in response to a treatment or infection. In this paper, we propose PIF, the Profile Interaction Finder, a novel approach that can be applied to expression data sets in order to tackle these questions.

Studying data sets of *A. thaliana* and *A. lyrata* after auxin treatment, and of *A. mellifera* after infection with different pathogens, PIF successfully identified genes related to the cell responses for the respective stimulus. In addition to that, PIF determined novel putative regulators that might affect several other genes in the downstream response. The detected targets of the *Arabidopsis* gene AT5G54510 for example had not yet been identified to be involved in the auxin signaling pathway. This shows that PIF is capable to discover previously unknown relationships between genes. The obtained results are highly relevant, as shown by linking them to already existing biological knowledge, represented for example in the gene ontology. Being capable to identify not only enhancing but also suppressing regulators is another advantageous feature of PIF. For example, with our method we were able to find two genes which are possibly down-regulated by AT5G54510.

Hence we conclude that PIF is a valuable tool for getting deeper insights into biological processes by analyzing gene expression data under varying experimental conditions.

5 Acknowledgements

We thank Carolin Delker, Jan Grau, Marcel Quint, Jana Trenner, and all participants of the Trans-Bee workshop for valuable discussions.

The honeybee transcriptome data used in Section 3.2 were collected and analyzed for the project Trans-Bee [Tra14], which was kindly supported by sDiv, the Synthesis Centre for Biodiversity Sciences – a unit of the German Centre for Integrative Biodiversity Research

(iDiv) Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118).

References

[BAAH04] Rainer Breitling, Patrick Armengaud, Anna Amtmann, and Pawel Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. {*FEBS*} *Letters*, 573(1–3):83 – 92, 2004.

[D+12] Claudia Dussaubat et al. Gut Pathology and Responses to the Microsporidium Nosema ceranae in the Honey Bee Apis mellifera. *PLoS ONE*, 7(5):e37017, 05 2012.

[DRQ08] Carolin Delker, Anja Raschke, and Marcel Quint. Auxin dynamics: the dazzling complexity of a small molecule's message. *Planta*, 227:929–941, 2008.

[FA13] Michelle L. Flenniken and Raul Andino. Non-Specific dsRNA-Mediated Antiviral Response in the Honey Bee. *PLoS ONE*, 8(10):e77263, 10 2013.

[HAG13] Holly Holt, Katherine Aronstein, and Christina Grozinger. Chronic parasitization by Nosema microsporidia causes global expression changes in core nutritional, metabolic and behavioral pathways in honey bee workers (Apis mellifera). *BMC Genomics*, 14(1):799, 2013.

[KG03] Fumiaki Katagiri and Jane Glazebrook. Local Context Finder (LCF) reveals multidimensional relationships among mRNA expression profiles of *Arabidopsis* responding to pathogen infection. *Proceedings of the National Academy of Sciences*, 100(19):10842–10847, 2003.

[M+13] Cynthia McDonnell et al. Ecto- and endoparasite induce similar chemical and brain neurogenomic responses in the honey bee (Apis mellifera). *BMC Ecology*, 13(1):25, 2013.

[MC11] Fionn Murtagh and Pedro Contreras. Methods of Hierarchical Clustering. *CoRR*, abs/1105.0121, 2011.

[P+13] Yvonne Poeschl et al. Optimized Probe Masking for Comparative Transcriptomics of Closely Related Species. *PLoS ONE*, 8(11):e78497, 11 2013.

[RS00] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[T+04] Oliver Thimm et al. MapMan: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, 37(6):914–939, 2004.

[Tra14] The Trans-Bee workshop, see http://www.idiv-biodiversity.de/sdiv/workshops/workshops-2013/stransbee (2014/07/23), 2014.

[WH14] Y.X. Rachel Wang and Haiyan Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of Theoretical Biology*, 2014.