Modeling gene regulation and spatial organization of sequence based motifs

Jochen Supper [1], Claas aufm Kampe [1], Dierk Wanke [2],
Kenneth W. Berendzen [2], Klaus Harter [2], Richard Bonneau [3], and Andreas Zell [1]

Abstract— **Reconstructing and modeling regulatory networks is an active area of research in bioinformatics and systems biology. Hence, various computational methods have been published, often successfully modeling one aspect of regulatory control. Gene regulation, however, is a process that depends on many different components such as transcription factors (TFs), *cis*-regulatory motifs and their temporal and spatial coordination. Accordingly, a promising new direction for computational analysis is the incorporation of multiple data types to discover, for instance, cluster membership, the spatial organization of *cis*-regulatory motifs and TFs that bind to these motifs.**

Here, we present such a data-driven framework, comprising four stages, to infer gene regulatory networks (GRNs) by modeling: 1. motif presence in the promoter, 2. spatial motif arrangement in co-regulated genes, 3. TFs that bind the respective motifs, and 4. dynamic properties of the GRN. A novel method is presented in stage 2, where we optimize for the spatial motif properties: orientation, occurrence of multiple motifs, relative distance between two motifs and distance to the Transcription Start Site (TSS). To find optimal distance based properties in efficient time we describe a dynamic programming approach. To combine multiple motif properties that are shared by genes with similar expression profiles a Hill-climber is employed. Subsequently, in stage 3 and 4, we infer GRNs by assigning TFs to the derived motifs and model time-dependent regulatory relationships between them with the Inferelator approach. None of the stages require the user to manually adjust any parameter, and thus derived properties can be analyzed without the bias introduced by parametrization. We applied this approach to *S. cerevisiae* data and obtained insight into individual and general properties of the spatial assembly of regulatory elements and inferred the corresponding GRN.

I. INTRODUCTION

Transcriptional and posttranscriptional regulation are important mechanisms for controlling protein abundance. During this process non-static protein complexes dynamically bind to regulatory DNA and RNA sequences, controlling the generation and degradation of mRNA. Currently the analysis of complex biochemical interactions is restricted to specific proteins of interest. For the analysis of regulatory control on a genome-wide scale it is, therefore, necessary to focus the analysis on DNA and RNA binding sequences which serve as proxy for more complex protein dynamics.

Accumulated mRNA levels and changes thereof after perturbations are strongly influenced by specific TFs. These changes can be monitored with microarrays on a genome-wide scale. One aim in analyzing such data is to determine relationships between TFs and genes, hence the inference of GRNs. This inference can be approached directly, by connecting TFs to genes based on expression data and additional *a priori* information ([12], [19]). This direct inference based solely on mRNA levels, however, is problematic as many methods assume causal relationships between the mRNA levels of the TFs and the regulated genes, that the system is at equilibrium, and that the rate of transcription (the direct target of TFs) is a function of accumulated mRNA level.

Other approaches were designed to circumvent this problem by a multi-step analysis integrating further data types, in addition to expression data. Typically, the first step thereby is (bi)clustering of all genes [7], with the hypothesis that clustered genes are under the same regulatory control. The second step is the detection of *cis*-regulatory motifs in the promoters of clustered genes by *de novo* methods [8], phylogeny based methods [10], or by searching for known binding sites [22]. Thereafter, TF-gene interactions can be established by associating binding sites of TFs to the respective motifs [4], without assuming expression correlation between TFs and genes. In a last step, a dynamical and predictive model can be inferred, in accordance to the GRN topology derived in the previous steps.

Bonneau *et al.* have described one path through this process, combining an integrative clustering and motif finding with a dynamical modeling approach (they model the change in transcript level over time-series, avoiding some of the problematic assumptions above) [5]. Another recent focus of attention, complementing these efforts, is the spatial organization of regulatory motifs. Several publications have pointed out spatial properties that are essential for the functionality of certain binding motifs [15]. In accordance to these findings computational methodologies have aimed at extracting regulatory modules with certain motif properties. In most cases, however, these studies were focused on one property like co-occurrence of other motifs or the positional influence was investigated by sliding window approaches [15]. Beer *et al.* [2] and Elemento *et al.* [9] presented valuable frameworks capable of modeling different motif properties in preclustered datasets. Preclustering, however, provides groups of co-expressed rather than co-regulated genes, thus this preprocessing step can lead to invalid assignments. Responding to this problem Nguyen *et al.* [14] presented a methodology called MED, which could overcome the potentially erroneous preclustering step, by

[1] Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Germany

[2] Center for Plant Molecular Biology (ZMBP), University of Tübingen, Germany

[3] Center for Genomics and Systems Biology, Biology Department, New York University, USA

Correspondence: jochen.supper@uni-tuebingen.de

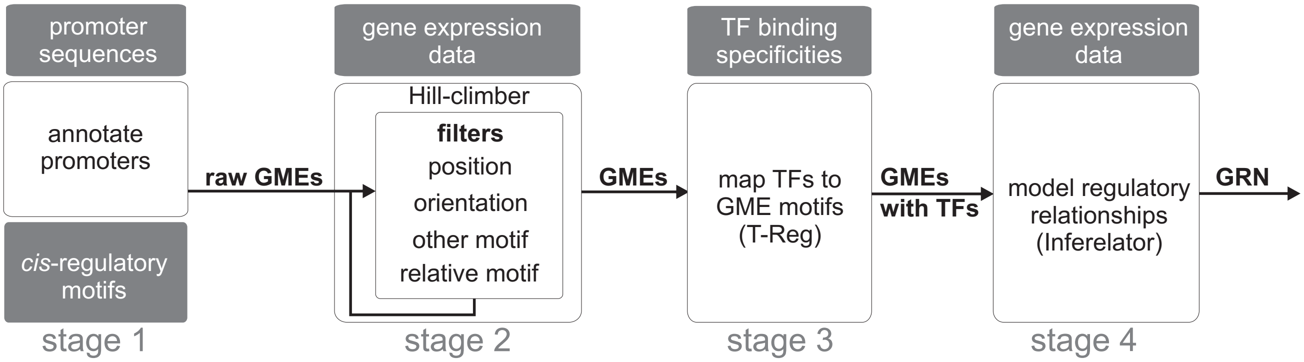

Fig. 1. Pipeline of the inference approach. The first stage is the promoter annotation, followed by the Hill-climber involved in learning the GMEs (stage 2). In the third stage TFs are assigned to the derived motifs. The fourth, and last, stage is the inference of the GRN through the Inferelator. The biological datasets used in each stage are plotted next to each stage.

calculating the activation of different motifs globally.

In this work we present a novel framework which is not dependent on preclustering of datasets, and does not require for parametrization by the user. We show the biological applicability on a *S. cerevisiae* dataset previously used by Beer *et al.* [2] and Nguyen *et al.* [14]. The resulting network contains rich information including motifs and their spatial arrangement, TFs binding to these sites, and a model that explains the network dynamics.

II. Approach

The aim of this work is twofold, first we present a method that reveals significant motif arrangements, and second we embed this method into a GRN inference pipeline (Fig. 1).

The method for revealing spatial motif properties combines motif and gene expression data, while optimizing an objective function (see Eq. 3), that is related to gene expression correlation. To describe biologically plausible motif arrangements we implemented four filter types – each optimizing one motif property –, that are iteratively applied and allow the accumulation of multiple properties.

The pipeline to infer GRNs is depicted in Fig. 1, where the motif analysis (stage 1 + 2) is supplemented with a down-stream pipeline. Thereby, TF binding specificities are mapped (T-Reg) to derived motifs (stage 3), and subsequently a dynamical modeling approach is applied (stage 4).

A. Biological data and promoter annotation - stage 1

To reveal biologically functional promoters (stage 1 + 2) we need gene expression data, motif data and promoter sequences. As gene expression dataset we choose the dataset of Spellman *et al.* [21] and Gasch *et al.* [11], that has been previously used to infer GRNs ([2], [14], [9]). This dataset covers response measurements to environmental stress stimuli and cell-cycle progressions – in total 255 conditions. We obtained the motif data from the publication of Nguyen *et al.* [14]. There they provide 62 motifs – 37 from literature and 25 from AlignACE and ScanACE. The promoter sequences are specified as -1 to -1 000 bp upstream of the TSS and annotated with motif data including the position and orientation of every motif (stage 1).

To map TFs to the derived motifs (stage 3) TF-binding specificities are needed. These are downloaded from TRANSFAC [13] and YEASTRACT [22]. For stage 4 no additional data is necessary, since the Inferelator only requests putative TFs and gene expression data.

B. Sampling gene-motif ensembles (GMEs) - Stage 2

First, we present the main methodology (Fig. 1, stage 2), which aims to grouping genes that are co-expressed and have a shared spacial motifs arrangement in their promoters. We start this method by seeding the optimization with genes sharing a certain motif. Such genes, that have one shared motif are referred to as *raw gene-motif ensembles* (raw RMEs), whereas genes that share a more complex motif pattern (i.e. multiple motifs, or specific spatial arrangement) are called *gene-motif ensembles* (GMEs).

To derive GMEs, raw GMEs are used as seeds and subsequently refined by adding spatial properties or further motifs. Thereby, for several iterations, the spatial property with the maximal score (Eq. 3) is added. The spatial motif properties we consider are: orientation, occurrence of multiple motifs, relative distance between two motifs and the distance of motifs to the transcription start site (TSS).

For each of these properties a filter is implemented that adds new constraints to the GMEs, and removes all genes that do not comply to it. Thereby, each filter is required to return the optimal GMEs with respect to its property. Thus, every filter performs an exhaustive search. To sample to overall search space a Hill-climber utilizes all implemented filters. Thereby, the Hill-climber proceeds as follows: for every GME apply the filter which returns the GME with the best score (Eq. 3). Then repeat this step iteratively, until the score cannot be further improved or every filter was applied. This procedure is repeated for every raw GME.

C. Scoring co-expression for gene-motif ensembles (GMEs)

If genes that have certain motif properties in common are actually co-regulated, one should expect to observe a significant correlation signal in the gene expression data. To quantify and score this correlation for a set of genes G, the average correlation of the gene expression profiles is determined by averaging over all pairwise Pearson correlations (see Eq. 1). Where $\rho_{i,j}$ is the correlation between gene i and j.

$$P_{av}(G) = \frac{\sum_{i=1}^{|G|} \sum_{j=i+1}^{|G|} \rho_{i,j}}{|G|^2 - |G|} \tag{1}$$

To determine if the average correlation of a gene set is significant, its size has to be taken into account. Thus, for all gene set sizes n from 2 to 1 000 a background probability density distribution f_n is calculated. This distribution is created for every n by randomly sampling 10 000 times from the complete expression dataset, and each time calculating the average correlation (Eq. 1). To obtain a density distribution, a log-normal function is fitted to these datasets, and verified by a χ^2 goodness-of-fit test.

To calculate the score for a gene set G of size n, we compare its average correlation $P_{av}(G)$ against the cumulative distribution of the background distribution:

$$S^*_{corr}(G) = \int_{P_{av}(G)}^{\infty} f_n(x)dx \tag{2}$$

Thus, we calculate the probability of obtaining an P_{av} as high or higher in the background model.

Adding filters that result in marginal increases in the score can result in over-fitted models. To maintain a trade-off between the number of filters applied and the score (Eq. 2), we employ a Bonferroni correction with respect to the number of filters applied. Hence, for each GME the corrected score (Eq. 3) is calculated by multiplying Eq. 2 with the number of filters k applied.

$$S_{corr}(G,k) = k \cdot S^*_{corr}(G) \tag{3}$$

D. Filters: spatial motif properties

We describe four filter types that, in combination, we expect to capture biologically plausible promoters. For every motif property one filter is implemented. These filters are designed in a modular way, so that filters can be removed or added, while accounting for dependencies that might exist (e.g. the *relative motif distance* filter depends on the *other motif* filter). In this design each filter obtains a GME as input and returns a GME as output, which has one additional rule assigned. Internally, each filter removes genes according to some criteria, and if multiple possibilities exist, each filter considers all of them. The following selection of filters was chosen because they were pointed out as important in several publications ([1], [14], [23], [3], [9], [18]).

1) Position: The *position* filter removes all genes that do not have a motif at a certain distance to the TSS, given a distance interval ($[min, max]$). In most previous works the optimal interval was determined by sliding window or binning approaches [14]. It is not clear, however, if a fixed window size is appropriate. Therefore, here a parameter free approach is chosen to determine the optimal distance interval.

One possible way to determine the optimal distance interval is to calculate all of them. Given that m is the number of motif positions annotated throughout a set of promoters, there exist $\frac{m^2-m}{2}$ distinguishable intervals. Thus, the average correlation (Eq. 1) has to be calculated for each interval. If this is done separately it results in a complexity of $O(m^4)$.

To avoid the quadratic runtime complexity a dynamic programming scheme is developed that exploits the structure of the average correlation formula (Eq. 1) and allows to perform this calculation in $O(m^2)$. In this approach the basic iteration step is to calculate the average correlation for each interval from smaller intervals. Thereby, each interval is referred to by the index of the starting motif i and the index of the ending motif j. Given such an interval defined by i and j, all genes therein are referred to by $G_{i,j}$. In this nomenclature motifs are sorted from $5'$ to $3'$.

To calculate P_{av} for a set of genes $G_{i,j}$ in one iteration, $P_{av}(G_{i,j-1})$, $P_{av}(G_{i+1,j})$ and $P_{av}(G_{i+1,j-1})$ have to be known. The corresponding iteration formula reads as follows:

$$\begin{aligned} P_{av}(G_{i,j}) = \frac{1}{|G_{i,j}|^2 - |G_{i,j}|} \cdot \big(P_{sum}(G_{i,j-1}) \\ + P_{sum}(G_{i+1,j}) - P_{sum}(G_{i+1,j-1}) + \rho_{i,j}\big) \end{aligned} \tag{4}$$

with

$$P_{sum}(G_{i,j}) = (|G_{i,j}|^2 - |G_{i,j}|) \cdot P_{av}(G_{i,j}) \tag{5}$$

In Eq. 4 the correlations of the intervals $[i, j-1]$ and $[i+1, j]$ are summed up and the correlations of the interval $[i+1, j-1]$ are subtracted, since its correlation values are covered twice. To this $\rho_{i,j}$ is added and the overall term is normalized, as in Eq. 1.

A special case can occur if a certain motif is assigned multiple times to one promoter. In this case the correlation of 1, is removed from Eq. 1. Whenever removing such a value the denominator is corrected for this effect.

The overall approach is implemented as an iterative procedure. In the first step all intervals containing two motifs are calculated. Thereupon, all intervals containing more than two motifs are calculated. Thus, in each step one $P_{av}(G_{i,j})$ value is calculated with a fixed number of simple algebraic operations (Eq. 4). Overall, this is repeated $m^2 - m$ times, leading to a runtime reduction from $O(m^4)$ to $O(m^2)$.

2) Orientation: Each motif is either oriented in 5' to 3' or 3' to 5' direction. Hence, the *orientation* filter considers both possibilities for all motifs and returns the orientation with the best score.

3) Other motif: This filter removes all genes which do not have a certain other motif in their promoter. Thus, the filter allows to select genes with multiple yet different motifs in their promoter, and thereby employs 'and' logic. All motifs are considered as additional motifs.

4) Relative motif distance: The *relative motif distance* filter uses the *position* filter, by calculating distances with respect to motif pairs, instead of the TSS. This filter considers the orientational arrangement of the motifs, and is capable of capturing relative motif orientations. This calculation is performed only if a second motif is already added to the GME.

E. Permutation analysis and functional enrichment

When sampling multiple times from large datasets, the problem of over-fitting must be addressed. Hence, we perform a permutation analysis to investigate if the signals found in biological data are distinguishable from signals found in a background dataset (obtained through permutation of the biological data). To construct the background dataset associations between the motifs and the expression data are randomly shuffled. To this shuffled dataset, the framework is applied and the score of the best GME for each raw GME is stored. This procedure is repeated 250 times. To evaluate results from the biological dataset their score S_{corr} is compared against the scores obtained from the permutation analysis $\vec{S}^p$ (Eq. 6). This score reflects the likelihood of obtaining a result as good as or better in a shuffled dataset, and thus is considered as a p-value approximation.

$$B\left(\vec{S}^p, S_{corr}\right) = \frac{\left|\left\{i|S_{corr} \geq \vec{S}^p_i, i = 1, \cdots, 250\right\}\right|}{\left|\vec{S}^p\right|} \qquad (6)$$

To analyze the functional enrichment of GMEs, a gene set enrichment analysis is performed with the GO::TermFinder [6].

F. Determining TF binding sites - stage 3

To infer GRNs from the GMEs learned by our procedure potential regulators (TFs) have to be determined. This is done by matching the PSSMs ('Position-Specific Scoring Matrix) of the TFs (from TRANSFAC and YEASTRACT) against the motifs of the GMEs with the T-Reg comparator [17]. Where a TF-motif relationship with a T-Reg score over 0.8 is considered as match.

G. Network inference by the Inferelator - stage 4

Ultimately, when inferring GRN the aim is to obtain a quantitative and predictive model of the network. Thus, we test the applicability of such an inference approach (Inferelator) to the derived GMEs and their TFs.

Given a set of potential regulators (TFs) for a GME, we consider two simple hypothesis:

1) The TF regulates this GME and the TF's activity is a simple function of its mRNA level.
2) The TF regulates this GME but the TF's activity is regulated post-transcriptionally in an un-observed manner.

If hypothesis 1 is true we expect a *detectable* regulatory relationship between the TF and the expression profile of the GME. To model this relationship for each GME we apply the Inferelator framework, and consider only binding TFs – from stage 3 – as potential regulators. The Inferelator proceeds by selecting the minimal number of regulatory influences that are predictive utilizing the LASSO methodology, selecting from complex dynamical relationships between TFs and target genes [20]. By integrating time-series and steady-state datasets and allowing for a combinatorial logic the Inferelator is able to learn models that can predict future time points (TF→target dynamics). For each GME we optimize the time constant τ (determining the response time of gene expression) over the interval 5 to 50 min and employ a sigmoidal activation function. The shrinkage parameter (used to constrain model size, enforce parsimonious models) is determined by choosing the model with the smallest 10-fold cross-validation (CV) error.

If hypothesis 2 is true we expect to obtain the null model from the Inferelator.

III. Results

A. Overview: (raw) gene-motif ensembles

To determine if the presence of one motif alone is sufficient to capture groups of co-expressed genes the correlation and score of the raw GMEs is analyzed. The average number of genes in a raw GME is 551 (see Table I), with a median correlation of 0.024 and a median score S_{corr} of $2.21 \cdot 10^{-2}$, indicating the probability of sampling such a score at random. Thus, only weak signals are detected in raw GMEs. For instance, only two out of 62 raw GMEs have an an average expression correlation over 0.1 and most GMEs do not have a statistically significant signal (39 out of 62), when considering a significance level of 0.01.

The optimized GMEs, on the other hand, contain 77 genes on average and have an average correlation of 0.22 and a median correlation of 0.33. Surprisingly, 46 of the 62 GMEs have better scores when derived from the biological dataset than in all 250 runs of the permutation analysis, leading to a median score S_{corr} of $2.06 \cdot 10^{-8}$.

B. Analysis of gene-motif ensemble properties

In each Hill-climber iteration every filter can potentially be applied to the GME. Nonetheless, a bias in filter utilization can be observed. The *position* and *other motif* filter, for instance, are applied frequently, and in most cases only one filter is applied. An overview of the filter utilization is given in Table II and several GMEs are depicted in Table I.

The results of the *position* filter are shown in Fig. 3. The position intervals accumulate near the TSS and no interval occurs more than 739 bp upstream of the TSS. The average start position is 127, the average end position 344, the average length 217, and the standard deviation 160.

The *orientation* filter is applied only two times, however, an orientational bias might be observable in more GMEs. To investigate this, a sign test is applied to all GMEs. In this analysis eight GMEs had a statistically significant orientational bias (with a significance level of 0.01). Furthermore, in six of these GMEs all motifs have the same orientation.

C. Building Gene Regulatory Networks (GRNs)

On average 2.51 TFs are linked to 45 of the 62 GMEs. For each GME with at least one assigned TF the Inferelator analysis is applied. For two GMEs the Inferelator returned the null model, indicating that no simple relationship between the TFs and the GMEs exists. For all other GMEs

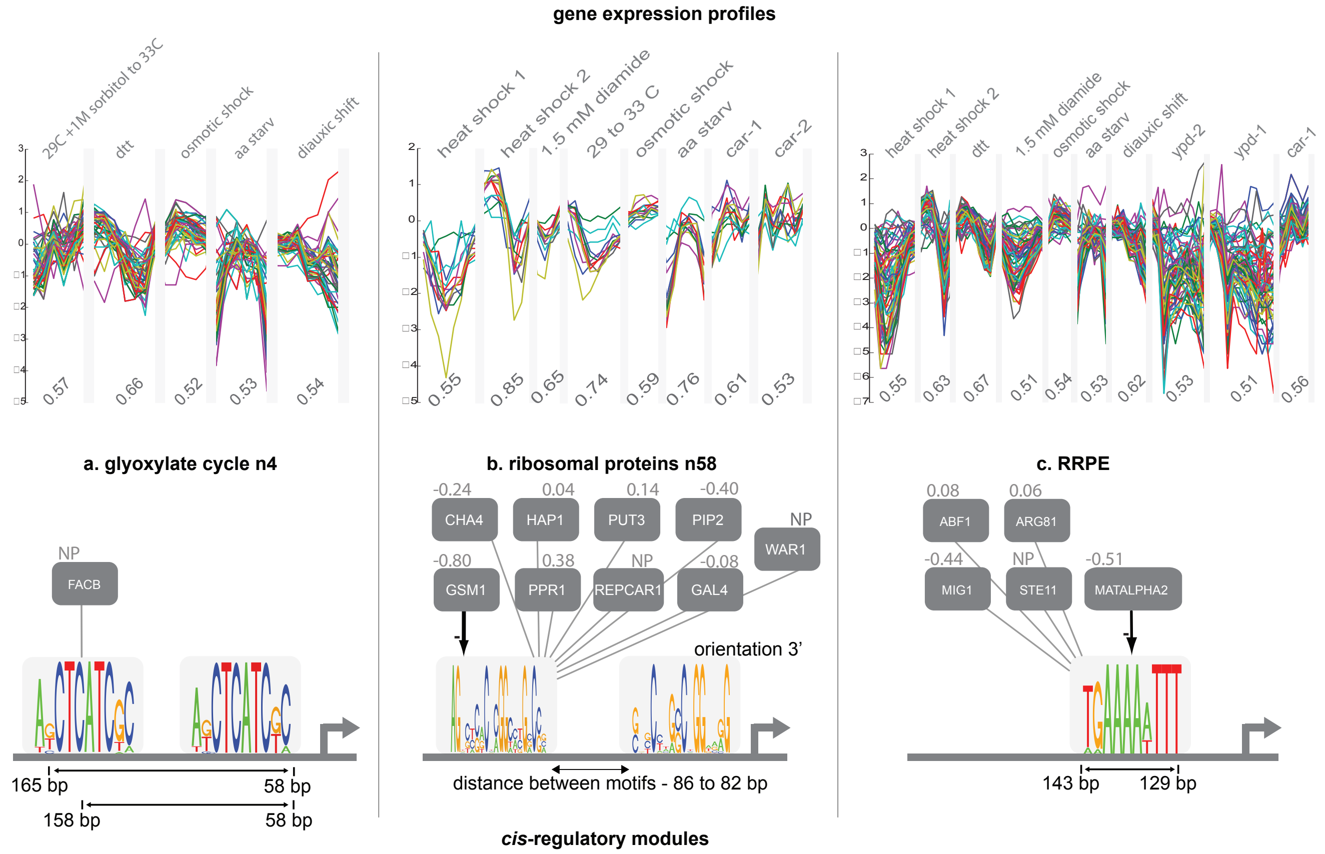

Fig. 2. Selected *gene-motif ensembles* (GMEs). The relative expression levels of the genes are displayed in the upper part of the figure, while only conditions with a P_{av} score over 0.5 are drawn. The lower part shows the derived motifs and the associated rules. All TFs which have a T-Reg score over 0.8 to one of the motifs are shown. For each TF the correlation to the GME is depicted above the box. The arrows indicate that these interactions were modeled by the Inferelator, where the (-) sign indicates an inhibitory interaction term. If a mapped TF was not contained in the expression data a *not present* (NP) is plotted. The statistical scores for each GME can be found in Table I.

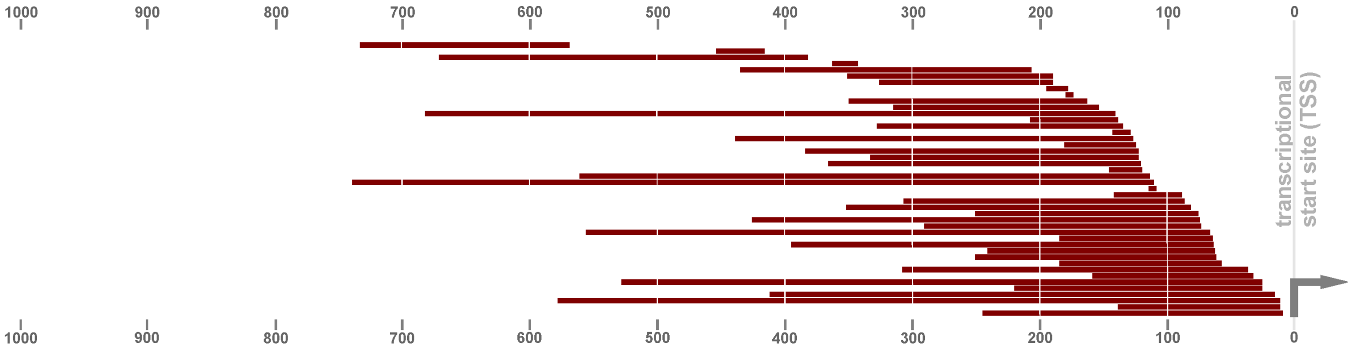

Fig. 3. Intervals in which binding motifs are allowed to occur for all GMEs. The considered upstream sequence is -1 to -1 000 bp and each interval is specified by a start and end position. Overall, 44 intervals from 42 GMEs are depicted, each on average containing 77 genes.

TABLE I

Statistical results for several raw GMEs. For each (raw) GME the average correlation (P_{av}) and the score (S_{corr}) is provided. For GMEs these statistical values are complemented by the score (S_{perm}) and the applied filters. The filters are specified as follows: *position* filter (PF), *other motif* filter (OMF), *relative motif distance* filter (RMDF) and *orientation* filter (OF). The index of the PF refers to the motif to which the interval filter was applied to. *Not present* (NP) entries occur for GMEs containing more than 1 000 genes (see Section II-C). The mean and median are calculated from the complete analysis on 62 GMEs.

	raw gene-motif ensembles			***gene-motif ensembles***				
Motif	#genes	P_{av}	score (S_{corr})	#genes	P_{av}	score (S_{corr})	p-value (S_{perm})	filters
glyoxylate cycle n4	1140	0.018	NA	53	0.43	$2.00 \cdot 10^{-11}$	0.016	OMF RMDF PF_2
ribosomal protein n58	84	0.029	0.06	13	0.63	$4.30 \cdot 10^{-7}$	0.00	OMF RMDF OF
RRPE	1061	0.105	NA	69	0.56	0.00	0.00	PF
MCB	213	0.969	0.00	99	0.21	0.00	0.00	PF
BAS1	541	0.021	0.21	20	0.49	$7.44 \cdot 10^{-8}$	0.00	OMF PF_2
MCM1	884	0.021	0.11	70	0.42	$5.21 \cdot 10^{-13}$	0.00	OMF
				⋮				
rrna processing n3	542	0.194	0.00	66	0.65	0.00	0.00	PF
mean	551	0.03	$2.18 \cdot 10^{-1}$	77	0.31	$2.05 \cdot 10^{-4}$	0.04	-
median	255	0.02	$2.21 \cdot 10^{-2}$	63	0.22	$2.06 \cdot 10^{-8}$	0.00	-

TABLE II

Overall filter statistics. The upper part shows how often each filter was applied. The lower part shows how many filters were applied to different GMEs.

filters			
44	31	14	2
position	other motif	relative motif distance	orientation
4	36	13	9
none	one	two	three

such models is established. Interestingly, 61% of the interaction terms had a negative effect on transcription rate, thus indicating an inhibitory relationship.

For instance, according to our learned Inferelator model, GSM1 inhibits the GME containing the ‘ribosomal proteins n58’ motif and the RRPE based GME is inhibited by MATALPHA2, which is known to act as inhibitor [16]. A model with OR logic was derived for the GME with the ‘metabolism of energy reserves n4’ motif, assuming a negative regulation by PPR1 and HAP1, both zinc fingers.

IV. Discussion

A. Gene-motif ensemble (GME) properties

We obtained numerous highly significant GMEs by applying our framework to *S. cerevisiae* data. Further analysis of the spatial organization of these modules provide insight into the characteristics and importance of different properties. For instance, the *position* filter was applied to the majority of the GMEs and is often displayed near the TSS, which could also be observed in previous studies [3]. Furthermore, the interval length and positioning display a high degree of flexibility. Other properties such as the conservation of orientation could be observed in 8 of 62 cases, implying that in most cases GMEs can be described irrespective of orientation.

B. Comparison to other methods

In 2006 Nguyen *et al.* [14] presented an approach called MED (motif expression decomposition), to calculate activity values for motifs from gene expression data. In a subsequent analysis they investigated motif positioning and orientation. To analyze the motif position they divided promoters into three partitions (short-, mid-, and long-range), and performed an analysis for each. One result they derived is based on the motif RRPE. After dividing the promoter into bins of length 150 bp they report the highest average correlation of 0.27 with an interval from -1 to -150 bp.

Our results refine this GME, as the *position* filter requires the motif to occur -129 to -143 bp upstream of the TSS. The corresponding GME contains 69 genes, and has an average correlation of 0.56 over all conditions. The most significant GO-term for this GME was ‘ribonucleoprotein complex biogenesis and assembly’ with a p-value of $2.90 \cdot 10^{-20}$.

Another GME described by Nguyen *et al.* [14] is based on the MCB motif, for which they report a positional and directional conservation. Accordingly, their motif was required to be within -150 to -300 bp and in 5-orientation. For this GME they report an average correlation of 0.2.

Our GME filters require the same motif to be within -87 to -302 bp of the TSS. 99 genes comply to this requirement displaying an average correlation of 0.21. To test for the orientational conservation reported by Nguyen *et al.* [14] we manually restricted the motifs to be oriented only in 5’ direction or in 3’ direction, leading to an average correlation of 0.27 and 0.15 respectively. The reason why the 5’ orientation filter is not applied in our analysis, is the lower number of genes. Given the trade-off between the number of genes, the average correlation (Eq. 3) and the additional Bonferroni correction (Eq. 6), the *orientation* filter is not considered significant in our framework. A GO-analysis of this module returned ‘DNA replication’ with a p-value of $2.8 \cdot 10^{-24}$.

Complementing the analysis of specific GMEs, a global view on the investigated properties should be discussed in light of modeling assumptions. For instance, the assumption of fixed interval sizes for motif positioning, is problematic when respecting the high degree of variance observed throughout the GMEs. This can also be observed for the *relative motif distance* filter.

V. CONCLUSION

We propose a parameter free and data-driven framework for inferring GRNs, with a novel method to reveal spatial motif arrangements (stage 2). This approach was realized with modular filters. The *position* filter was implemented as dynamic programming approach, this allowed for an exhaustive analysis of positional motif properties with respect to expression correlation. The statistical results obtained with this filter provide valuable information regarding the positional flexibility and biological utilization. Such results derived from system-level analysis can play an important role for understanding regulatory control mechanisms.

In comparison to previous publications we were capable of refining many GMEs, that upon more careful analysis showed high significance and have mechanistic descriptions of DNA binding sites and the dynamical effect of TF binding. Furthermore, the derived GMEs are accompanied by a wealth of information, such as statistical validation scores, condition specific average correlation, the occurrence and spatial constraints of motifs and TFs potentially binding these motifs (stage 3). In addition, for each GME a model of regulatory regulation was inferred with the Inferelator approach (stage 4). These dynamic response models allow to capture the type of regulation and the time scales these effects might be acting on. Many regulatory influences were inhibitory and all regulatory relationships were learned as effects on dynamical change in transcription from time series.

The GMEs support experimental validation, as the relevant experimental condition, the regulatory mode and the putative regulators are provided. Towards, the goal of inferring gene regulatory networks the derived GMEs provide a fairly comprehensive model.

FUNDING

National Genome Research Network (NGFN II) of the Federal Ministry of Education and Research in Germany under contract number 0313323 and the Deutsche Forschungs Gemeinschaft (SFB 446/AFGN HA2146/5-2).

ACKNOWLEDGEMENT

We thank Aviv Madar for his support on the Inferelator and Adrian Schröder for his advice on network inference.

REFERENCES

[1] Nilanjana Banerjee and Michael Q Zhang. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*, 31(23):7024–7031, Dec 2003.

[2] Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, Apr 2004.

[3] Kenneth W Berendzen, Kurt Stüber, Klaus Harter, and Dierk Wanke. Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics*, 7:522, 2006.

[4] K. Birnbaum, P. N. Benfey, and D. E. Shasha. cis element/transcription factor analysis (cis/tf): a method for discovering transcription factor/cis element relationships. *Genome Res*, 11(9):1567–1573, Sep 2001.

[5] Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, 7(5):R36, 2006.

[6] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, Dec 2004.

[7] Yizong Cheng and George M. Church. Biclustering of expression data. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103. AAAI Press, 2000.

[8] Eran Eden, Doron Lipson, Sivan Yogev, and Zohar Yakhini. Discovering motifs in ranked lists of dna sequences. *PLoS Comput Biol*, 3(3):e39, Mar 2007.

[9] Olivier Elemento, Noam Slonim, and Saeed Tavazoie. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell*, 28(2):337–350, Oct 2007.

[10] Fei Fang and Mathieu Blanchette. Footprinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res*, 34(Web Server issue):W617–W620, Jul 2006.

[11] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–4257, Dec 2000.

[12] Florian Geier, Jens Timmer, and Christian Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst Biol*, 1:11, 2007.

[13] V. Matys, E. Fricke, R. Geffers, E. Gssling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378, Jan 2003.

[14] Dat H Nguyen and Patrik D'haeseleer. Deciphering principles of transcription regulation in eukaryotic genomes. *Mol Syst Biol*, 2:2006.0012, 2006.

[15] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, 29(2):153–159, Oct 2001.

[16] S. D. Porter and M. Smith. Homoeo-domain homology in yeast mat alpha 2 is essential for repressor activity. *Nature*, 320(6064):766–768, 1986.

[17] Stefan Roepcke, Steffen Grossmann, Sven Rahmann, and Martin Vingron. T-reg comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res*, 33(Web Server issue):W438–W441, Jul 2005.

[18] Taewoo Ryu, Younghoon Kim, Dae-Won Kim, and Doheon Lee. Computational identification of combinatorial regulation and transcription factor binding sites. *Biotechnol Bioeng*, 97(6):1594–1602, Aug 2007.

[19] Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8 Suppl 6:S9, 2007.

[20] K. Sjöstrand. Matlab implementation of LASSO, LARS, the elastic net and SPCA, jun 2005. Version 2.0.

[21] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

[22] Miguel C Teixeira, Pedro Monteiro, Pooja Jain, Sandra Tenreiro, Alexandra R Fernandes, Nuno P Mira, Marta Alenquer, Ana T Freitas, Arlindo L Oliveira, and Isabel S-Correia. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res*, 34(Database issue):D446–D451, Jan 2006.

[23] Wei-Sheng Wu, Wen-Hsiung Li, and Bor-Sen Chen. Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics*, 7:421, 2006.