Inferring Transcriptional Regulators for Sets of Co-expressed Genes by Multi-Objective Evolutionary Optimization

Adrian Schröder*, Clemens Wrzodek*, Johannes Wollnik*, Andreas Dräger*, Dierk Wanke†, Kenneth W Berendzen† and Andreas Zell*

*Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, Sand 1, 72076 Tuebingen, Germany

†Center for Plant Physiology Tuebingen (ZMBP), University of Tuebingen, Auf der Morgenstelle 1, 72076 Tuebingen, Germany

***Abstract*—Higher organisms are able to respond to continuously changing external conditions by transducing cellular signals into specific regulatory programs, which control gene expression states of thousands of different genes. One of the central problems in understanding gene regulation is to decipher how combinations of transcription factors control sets of co-expressed genes under specific experimental conditions. Existing methods in this field mainly focus on sequence aspects and pattern recognition, e.g., by detecting *cis*-regulatory modules (CRMs) based on gene expression profiling data. We propose a novel approach by combining experimental data with a priori knowledge of respective experimental conditions. These various sources of evidence are likewise considered using multi-objective evolutionary optimization. In this work, we present three objective functions that are especially designed for stimulus-response experiments and can be used to integrate a priori knowledge into the detection of gene regulatory modules. This method was tested and evaluated on whole-genome microarray measurements of drug-response in human hepatocytes.**

I. Introduction

Higher organisms are able to respond to permanently changing environmental conditions by activating specific gene-regulatory programs. These condition-specific regulatory programs consist of positively and negatively regulating transcription factors (TFs) and coordinate the expression of sets of functionally related genes [1]. Although much is known about gene-regulatory relationships, knowledge about condition-specific regulatory programs is still limited. Genome-wide mRNA expression measurements, conducted under multiple experimental conditions, constitute powerful analytical tools to approach this challenge [1], [2]. Current computational methods, for deciphering gene-regulatory programs from high-throughput data, are mostly based on the detection of *cis*-regulatory modules (CRMs). CRMs are defined as patterns of transcription factor binding sites (TFBSs) that are present in promoter sequences of co-expressed genes. TFBSs are usually identified by mapping libraries of position frequency matrices (PFMs) to promoter sequences of genes. PFMs are widely used DNA-binding motif models that are mathematically represented as nucleotide frequency matrices [3]. The detection of common patterns of TFBSs makes it possible to draw conclusions about the transcription factors that cooperatively regulate the given set of genes [4]. Most CRM detection methods are restricted to mine for patterns of TFBSs at the sequence level and most often ignore the associated TFs and their relationships to the target genes on the mRNA expression level, which can partly be traced back to auto-regulatory transcriptional control mechanisms [5]. Recent approaches partly use these dependencies to discover regulatory relationships between TFs and their clustered target genes [6], [7]. CRM detection is mostly approached using heuristic optimization techniques, such as evolutionary algorithms [8], [9]. Alternative approaches mine for CRMs based on stochastic modeling using Bayesian Hidden Markov Models [10]. Most of these approaches, however, do not account for knowledge about the respective experimental conditions, like disease states, chemical stimuli, or other treatments. For this work, we developed a new algorithm that integrates multiple sources of biological knowledge, as gene expression profiles, position frequency matrices (PFMs), protein-protein interaction data, and protein-chemical interaction data into the search for CRMs. This is mathematically accomplished by formulating the search for CRMs as multi-objective optimization problem. Thus, the integration of additional sources of evidence drives the search for CRMs towards solutions that are specific for the underlying experimental condition. As depicted in Fig. 1, we propose three different sources of evidence, which are implemented as objective functions: (1) patterns of TFBSs, (2) multivariate relationships between TFs and target genes, (3) pathway connectivity scores between TFs and experimental treatment. The method was tested and evaluated on a microarray dataset of drug-treated samples of human hepatocytes.

II. Cis-regulatory module detection

The integration of the above mentioned sources of evidence is realized by formulating the CRM detection problem as multi-objective optimization problem. In order to account for the combinatorial complexity of this problem, we propose multi-objective genetic algorithms (MOGAs), which are heuristic search procedures that are inspired by natural evolution. In general, genetic algorithms (GAs), sample the search space by generating candidate solutions based on the three

978-1-4244-7833-0/11/$26.00 ©2011 IEEE

main principles of evolution: natural selection, mutation and recombination; MOEAs have been proven to be powerful for solving complex optimization problems [11], [12]. Multi-objective optimization problems are mathematically solved by investigating sets of so-called Pareto-optimal solutions, which represent appropriate solutions to the optimization problem that are not dominated by other solutions [11]. Since GAs are stochastic search procedures, different optimization runs may result in different solutions. In order to get robust results, the whole multi-objective optimization procedure is repeated multiple times.

As depicted in Fig. 1, the CRM-detection method expects as input sets of co-expressed genes, which can be derived from gene expression data by clustering [13]. The procedure is sub-divided into two main steps. In the first step, the promoter sequences of the input genes are scanned for TFBSs using libraries of PFMs. The resulting set of matching TFBSs constitutes the search space that is mined for CRMs by applying the MOGA in the second step.

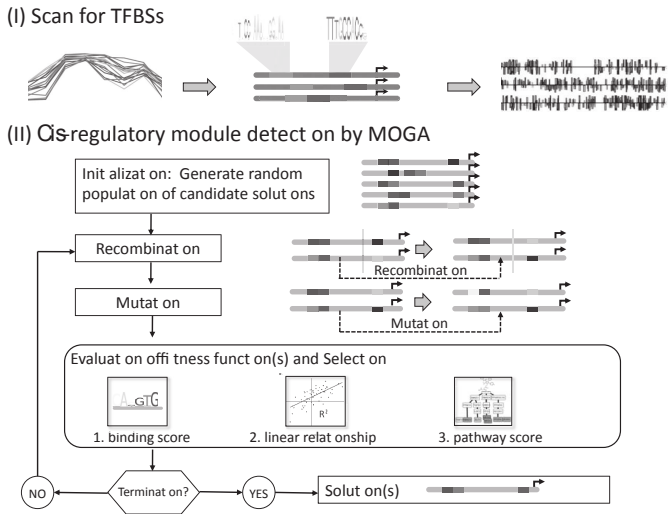

Fig. 1: **Multi-objective CRM-detection algorithm** (I) Scan for TFBSs: Given promoter sequences of clusters of co-expressed genes are scanned for TFBSs using PFMs, (II) *Cis*-regulatory module detection: First, the multi-objective genetic algorithm (MOGA) is initialized by generating an initial population of candidate CRMs. Within the main optimization routine the candidate solutions are recombined, mutated and evaluated according to the various fitness functions until the termination criterion is met. The recombination of two parent solutions is realized by cutting the parent CRMs randomly at one point and combining the first part of parent 1 with the second part of parent 2 and vice versa. Mutation of a candidate solution is performed by randomly exchanging one TFBS for another one.

A. Step I: Scan for TFBSs

In this step, as depicted in Fig. 1 (Step I), position frequency matrices (PFMs) are used to scan promoter sequences for TFBSs and calculate binding scores at each position. According to Aerts *et al.* (2003) [14], the binding score $W_{\mathbf{x}}(\Theta)$ of a single PFM Θ on subsequence $\mathbf{x}$ of sequence s can be calculated as

$$W_{\mathbf{x}}(\Theta) = \frac{\prod_{j=1}^{w} \Theta(b_j, j)}{\prod_{j=1}^{w} P(b_j|B_m)}, \quad \mathbf{x} = [b_1, ..., b_w], \qquad (1)$$

where b_j denotes the nucleotide found at position j in the subsequence $\mathbf{x}$, $\Theta(b_j, j)$ is the probability of finding b_j according to PFM Θ and $P(b_j|B_m)$ gives the probability of finding b_j according to the background model B_m. w is the length of the respective binding site. Eq. 1 calculates the likelihood that subsequence $\mathbf{x}$ is generated by the motif model with respect to the background model B_m. One TF can bind at multiple different positions to the DNA. The different binding sites that are matched by a PFM Θ on sequence s are referred to as instances of Θ on s. Binding scores are calculated according to Eq. 1 using a 4$^{\text{th}}$-order hidden Markov model as background. In order to decide, if a certain score $W_{\mathbf{x}}(\Theta)$ should be counted as match or not, cutoff levels need to be defined. Individual cutoff values are pre-calculated for each PFM individually rather than choosing a global cutoff score for all PFMs. To this end, a conservative cutoff strategy is implemented, which is based on scanning non-regulatory (i.e., exonic) sequences and calibrating the cutoff score such that no hits are found on non-regulatory sequences [15]. In this work, the databases JASPAR [16] and TRANSFAC [17] were used as PFM resources.

B. Step II: Multi-objective genetic algorithm

The core part of the proposed algorithm is to identify combinations of cooperatively regulating TFs for each cluster of genes. The crucial idea is to consider the problem of searching for CRMs in sets of co-expressed genes as a multi-objective combinatorial optimization problem, which allows us to combine various heterogeneous sources of evidence. Depending on the given microarray dataset, various sources of evidence come into consideration. In this work, we propose three different sources of evidence: (1) patterns of TFBSs, (2) relationships between regulators and clustered genes, and (3) pathway scores, representing the relationship between TFs and the respective experimental condition (see Fig. 1). Here, we propose a MOGA to search for optimal TF combinations for given clusters of genes. As described above, MOGAs search for so called Pareto-optimal solutions. To this end, as described in Section II-E, the final set of Pareto optimal solutions is evaluated by calculating genome-wide specificity scores.

1) First optimization objective: The first criterion that a CRM has to meet is the consistence of multiple good matching PFMs that all occur within a bounded region within the promoter sequences of the input genes. These properties are reflected by the cluster module score (CMS). Let $m^s_{i_k}$ be the k$^{\text{th}}$ instance of TFBS Θ_i on sequence s. The detection of a CRM can, according to Aerts *et al.* (2004) [8], be performed by sampling the search space for combinations $\mathbf{m} = \left(m^s_{1_k}, ..., m^s_{l_k}\right)$ of instances of TFBSs $\Theta_1, ..., \Theta_l$ that are found in sequence s. The fitness of one possible solution

$\mathbf{m}$ on a single sequence s can be evaluated by calculating the sequence module score (SMS) as follows

$$\mathrm{SMS}_{\mathbf{m}}(s) = \max_{1_k,\ldots,l_k} b(\mathbf{m}) \sum_{i=1}^{l} W_s\left(m_{i_k}^s\right), \tag{2}$$

Since TFBSs are assumed to lie in physical proximity in a promoter sequence, they are not allowed to overlap. Thus, the solution space of possible modules $\mathbf{m} = \left(m_{1_k}^s, ..., m_{l_k}^s\right)$ is restricted to those, which do not overlap and lie within a subsequence of window size ω. A module $\mathbf{m}$ does overlap if the PFMs of at least two instances of $\mathbf{m}$ match to overlapping regions of sequence s. In order to account for this restriction, a Boolean variable $b(\mathbf{m})$ is used. $b(\mathbf{m})$ indicates whether $\mathbf{m}$ is a valid module for sequence s_i or not. A module $\mathbf{m}$ is invalid, if two instances $m_{i_k}^s$ and $m_{j_k}^s$ overlap or if the configuration of $\mathbf{m}$ exceeds the specified window size ω.

Let $(s_1, ..., s_n)$ be a set of n promoter sequences that belong to a cluster $\mathbf{c}$ of n co-expressed genes. The module score for a cluster of promoter sequences (CMS) can be calculated by adding up the SMS of single sequences s_i:

$$\mathrm{CMS}_{\mathbf{m}}(\mathbf{c}) = \sum_{i=1}^{n} \mathrm{SMS}_{\mathbf{m}}(s_i), \tag{3}$$

Thus, the first objective f_1 of the proposed CRM-detection algorithm is to find a module $\mathbf{m}^*$ that maximizes the cluster module score $\mathrm{CMS}_{\mathbf{m}}(\mathbf{c})$, for a given cluster $\mathbf{c}$:

$$f_1 = \mathrm{CMS}_{\mathbf{m}}(\mathbf{c}) \tag{4}$$

2) Second optimization objective: Regulatory relationships between TFs and their target genes are often reflected at the mRNA level [5] and become visible in microarray experiments by strong correlations of their respective expression profiles. To this end, the multivariate relationship between TFs and their putative target genes (input genes) can serve as an additional source of evidence in the search for CRMs, which is realized by the second optimization objective. Let $\mathbf{m} = \left(m_{1_k}^s, ..., m_{l_k}^s\right)$ be a candidate solution consisting of one instance of each of the PFMs $\Theta_1, ..., \Theta_l$. Since every PFM Θ_i in our dataset is associated with a certain TF and we know the expression profiles of almost all TFs from the underlying whole-genome microarray dataset, we can include the multivariate relationship between the expression profiles of the transcription factors associated to a module $\mathbf{m}$ and the expression profile of the target genes in the cluster as second source of evidence.

A multivariate linear regression is used to model the combinatorial regulatory relationship between the expression profiles of TF combinations in module $\mathbf{m}$ and their putative target genes in the cluster. Let $\mathbf{\Theta}_{exp} = (\Theta_{exp_1}, ..., \Theta_{exp_n})^T$ be the expression profile of the associated TF to PFM Θ over all n data points.

$$\mathbf{y}_c = b_0 + b_1\mathbf{\Theta}_{exp_1} + ... + b_l\mathbf{\Theta}_{exp_n}, \tag{5}$$

where $\mathbf{y}_c = (y_{c_1}, ..., y_{c_n})^T$ is the mean expression profile over all genes in cluster $\mathbf{c}$, i.e., the cluster centroid. Since each cluster consists of highly correlated genes, taking the mean over all genes is reasonable at this point. The coefficient of determination R^2 is used to quantify the degree of linear relationship. R^2 ranges from 0 (no linear association) to 1 (perfect linear association). The sign of each regression coefficient b_i in the regression solution indicates if the corresponding TF is a positive or negative regulator. The method of least squares is applied [18] to solve the regression problem. Thus, the second objective f_2 is to find a module $\mathbf{m}^*$ with maximal coefficient of determination R^2 between the associated TFs and the cluster centroid $\mathbf{y}_c$:

$$f_2 = R^2 = 1 - \frac{\mathrm{E}_{\mathrm{err}}}{\mathrm{E}_{\mathrm{tot}}}, \tag{6}$$

where $\mathrm{E}_{\mathrm{err}} = \sum_{i=1}^{n}(y_{c_i} - \hat{y}_{c_i})^2$ gives the residual variation, i.e., the difference between data points y_{c_i} and regression values $\hat{y}_{c_i}$ and $\mathrm{E}_{\mathrm{tot}} = \sum_{i=1}^{n}(y_{c_i} - \overline{y}_c)^2$ with $\overline{y}_c = \frac{1}{n}\sum_{i=1}^{n} y_{c_i}$ gives the variation of the dependent variable $\mathbf{y}_c$.

3) Third optimization objective: The activation of TFs is generally the response of the cell to external stimuli, e.g. hormones, endogenous or exogenous chemicals, which trigger signaling cascades that finally promote the TF activation. To this end, when stimulus-response data are analyzed, signaling cascades between the external stimulus and the candidate set of TFs can serve as additional source of evidence. In this work, the BowTieBuilder algorithm [19] is used to evaluate the connection between a candidate set of TFs and external stimuli (e.g., chemicals or hormones). The BowTieBuilder is a pathway inference algorithm that identifies the most confident pathway between two given sets of proteins, i.e., the set of source chemicals/proteins S and set of target proteins T (TFs) based on a given network of known protein-protein and protein-chemical interactions. To this end, interaction networks from the protein-protein interaction database STRING [20] and the protein-chemical interaction database STITCH [21] have been integrated. These databases contain known and predicted interactions that are either physical or functional. Each interaction is associated with a confidence score between 0 and 1 depending on the reliability of the information sources. Formally, the third objective function f_3 is designed to assess the confidence of the interaction network between the respective stimulus d and the set of TFs $T_{\mathbf{m}}$ of a module $\mathbf{m}$. Therefore, the scores of the most confident interaction paths $cp(d, t)$ from the stimulus d to the TFs $t \in T_I$ are averaged. Accordingly, the more confident the stimulus is connected to the TFs, the better is the fitness of the candidate solution. Thus, the third objective f_3 is to find a module $\mathbf{m}^*$ that maximizes the confidence of the pathway between the stimulus and the candidate TFs. It is calculated as follows:

$$f_3 = \frac{1}{|T_{\mathbf{m}}|} \sum_{t \in T_{\mathbf{m}}} cp(d, t) \tag{7}$$

C. Multi-objective genetic algorithm with Pareto ranking

The objectives f_1, f_2 and f_3 are not correlated and in some cases, but not generally, conflicting. We do not recommend to optimize them individually, because this could result in biologically unacceptable results. The optimization of the second objective f_2, for example, without insurance that the respective TFBSs occur in the promoter sequences (f_1) will produce correlation networks, which often contain biologically implausible relationships. Therefore, multi-objective optimization can be applied to simultaneously optimize all objective functions in order to obtain a set of Pareto-optimal solutions. A solution is said to be Pareto-optimal if it cannot be improved with respect to one objective without worsening another objective [11]. In this work, we use a MOGA with Pareto ranking to detect a set of Pareto-optimal solutions with respect to f_1, f_2 and f_3. In Pareto ranking approaches, the population is ranked according to a dominance rule and the fitness value for each solution is assigned with respect to this ranking [22]. Here, we used a fast elitist non-dominated sorting genetic algorithm (NSGA-II) with crowding distance as proposed by Deb *et al.* [23], [24].

The algorithm starts with an initial population that consists of N candidate modules $\mathbf{m}^i$, which are randomly sampled from the pool of matching TFBSs. If the termination criterion is not satisfied, an offspring population $\mathcal{O}(t)$ is generated by applying crossover and mutation operators to the initial population $P(t)$ and the fitness of all individuals according to both fitness functions is evaluated. The mutation and crossover operators are called with a mutation rate of 0.1 and a crossover rate of 0.7, respectively (see Discussion). Mutation of a candidate solution is performed by randomly changing one TFBS with probability $\frac{1}{l}$. A 1-point crossover strategy is used to recombine the TFBS-configurations of two candidate solutions $\mathbf{m}^i$ and $\mathbf{m}^j$ to a new candidate solution $\mathbf{m}^k$. The fast non-dominated sorting algorithm NSGA-II is used to identify the non-dominated fronts $F_1, ..., F_k$ to rank the population in ascending order. Crowding distances $cd_{f_1}(\mathbf{m}^j)$, $cd_{f_2}(\mathbf{m}^j)$, and $cd_{f_3}(\mathbf{m}^j)$ are calculated for all candidate solutions $\mathbf{m}^j$ in F_i with respect to all objective functions f_1, f_2 and f_3. Candidate solutions are stored in archive A with respect to Pareto ranking and crowding distance [23]. Binary tournament selection, based on the crowding distances [24], is used to create the parent population $P(t+1)$ from A. If the termination criterion is satisfied, all solutions stored in archive A are returned to the user. The termination criterion is satisfied if $t = 30,000$ iterations have been evaluated or $5,000$ fitness evaluations did not yield any change to the Pareto front.
We used the NSGA-II implementation of EvA2 [25], which is a comprehensive heuristic optimization workbench implemented in Java™. EvA2 provides several optimization methods, such as evolution strategies, genetic algorithms, differential evolution, particle swarm optimization, as well as classical techniques such as multi-start hill climbing or simulated annealing [25].

D. Complexity

The size of the search space, and therefore the complexity of the combinatorial optimization problem, strongly depends on the following variables: (1) module size l, which has to be pre-defined (2) promoter sequence size, and (3) PFM cutoff strategy. Let N be number of PFM matches in all promoter sequences of one cluster and l be the maximum module size. The size of the search space Ω consists of all l-element subsets of N:

$$|\Omega| = \sum_{i=0}^{l} \binom{N}{i} \tag{8}$$

The maximal number l of different TFs within a CRM and the promoter size are strongly organism- and tissue-specific and therefore strongly depend on the underlying dataset. l directly influences the size of the search space according to Eq. 8. The promoter size influences the number of matching PFMs N after scan for TFBSs. Short promoter sequences of $1,000$ base pairs, as used in this work, lead to a relatively small N compared to promoter sequences of $10,000$ base pairs and more. All fitness functions proposed in this work can be calculated in reasonable time. The speed-determining optimization step is the evaluation of objective function f_3. This evaluation can be accelerated substantially by pre-calculating all shortest pathways between the disposed PFM library and the given stimulus and storing them in a look-up table. A module search in a gene cluster of about 20 genes with about $1,000$ potential transcription factors including $30,000$ generations can be performed in less than one day on an average dual core processor. In practice, convergence is most often reached after $10,000$ generations. Considering that it is impossible to test thousands of combinatorial TF relationships for a given gene cluster in a wet-lab approach, we feel it is well worth the wait, as the approach is providing new testable hypotheses.

E. Final specificity ranking

Multi-objective optimization algorithms generate, as mentioned above, not one single solution, but sets of Pareto optimal solutions. During optimization, NSGA-II with crowding distance was used to rank these Pareto optimal solutions, which is reasonable due to its rapid calculation. For the final set of solutions, however, we want to select a solution that is as close as possible to biology. According to Loo *et al.*, we define a CRM solution as biologically meaningful if it is specific for the gene cluster it was derived from [9]. To this end, we perform a genome-wide backwards search for each Pareto optimal solution and rank all genes in the genome according to presence and fitness of the respective CRM in each promoter of the genome. Then, the ranks of all genes of the given cluster are determined to derive a specificity score, which is defined as the rank of the worst gene of the respective cluster in the genome-wide ranking. The most specific CRM, i.e., the one with the best specificity score, is recognized as the final solution. This genome-wide ranking

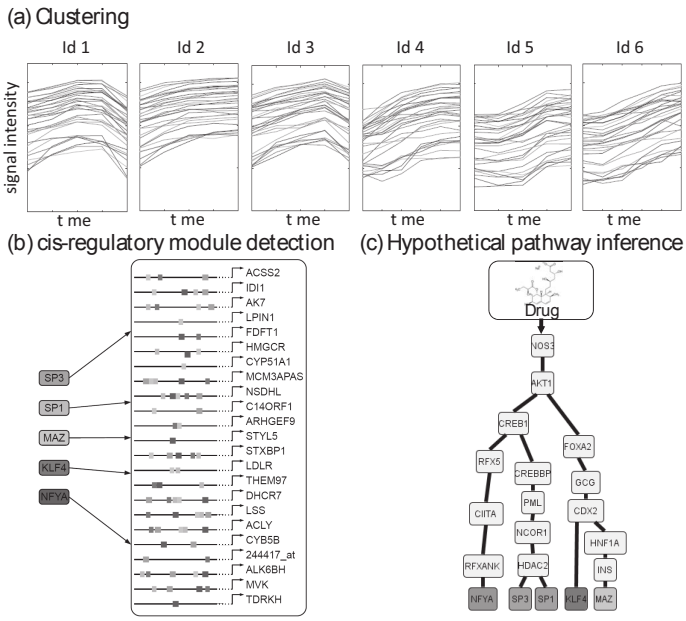

Fig. 2: **Multi-objective CRM detection results.** (a) Highly correlated expression time series of human hepatocytes from six individuals. Cluster 9 consists of 23 genes (see Table 1), which are for the most part involved in steroide metabolism (b) CRM-detection distills a combination of 5 different TFs, i.e., SP1, SP3, MAZ, KLF4, and NFYA. (c) Evaluation of the objective function f_3 reveals several co-factors and signaling molecules that are putatively involved in pravastatin-responsive regulatory mechanisms. The interaction of SP1, SP3 and HDAC2, for instance, has been reported previously [28]. KLF4 and NFYA are well known regulators of cholesterol and steroid metabolism [29], [30]. MAZ has not been reported before in the context of statin exposure.

procedure is computationally demanding and can therefore only be performed as a post-processing step.

III. Application to a *Homo sapiens* drug-response dataset

The multi-objective CRM-detection procedure was applied to a whole-genome microarray (Affymetrix U133 plus 2.0 chip) dataset of primary human hepatocytes from six individuals. Each sample was treated with pravastatin, a cholesterol lowering drug, and the control substance dimethylsulfoxide (DMSO). Microarray measurement were performed at up to six different time points (i.e., 0 h, 6 h, 12 h, 24 h, 48 h and 72 h) after the drug stimulus (measurements were conducted at the Microarray Facility Tuebingen, Germany). Microarray preprocessing, i.e., normalization, background correction and detection of differentially expressed genes, was conducted using standard procedures implemented in R Bioconductor [26]. Details about these experiments and the resulting dataset are described in [27].

A. Clustering

The Extended Dimension Iterative Signature Algorithm (EDISA) was used to find clusters of co-expressed genes. EDISA is a 3D clustering algorithm that is designed for gene-condition-time datasets, i.e., multiple experimental conditions performed at various time points [13], and mines the microarray data for subsets of genes, which are highly correlated over time. This algorithm is able to deal with subject-individual variability of gene expression profiles, which is often neglected in microarray analysis. EDISA is a heuristic search procedure that expects two main input parameters. The parameter τ_G defines the required similarity of each gene to the average trajectory of the cluster. The variable τ_C specifies required similarity of each condition with the average trajectory of the cluster. The number of clusters is not predefined but, as discussed in Supper *et al.* [13], strongly depends on the configuration of these input parameters as well as the strucure of the underlying dataset. As depicted in Table 1, EDISA identified 10 different clusters consisting of between 15 and 40 different genes.

In order to assess if sets of co-expressed genes are functionally related, gene set enrichment analysis was applied. To this end, a hypergeometric test was performed to each cluster of genes, testing for different Gene Ontology [31] categories and KEGG pathways [32]. In 8 of 10 clusters, genes of steroid-, lipid- and drug metabolism were found to be significantly enriched at significance level $\alpha = 0.05$ after Sidak multiple testing adjustment [33].

B. Application of multi-objective CRM-detection

For each cluster of co-expressed genes, up to five TFs were identified using the multi-objective algorithm presented above. The proximal promoter sequences ($1,000$ base pairs upstream from the transcription start site) were scanned using a window size ω of 200 base pairs. The CRM-detection method was run using all three fitness functions proposed above, whereas f_3 was configured to calculate pathway connectivity scores between the TFs of each candidate solution and the drug stimulus. In order to speed up the calculation of this function, shortest pathways between all individual TFs and drug stimulus are pre-calculated based on the databases STRING and STITCH [20], [21]. The optimization was started with a maximum number of $30,000$ evaluations of each objective function. Finally, the the Pareto optimal solutions were ranked according to the final ranking procedure presented above. All of the resulting CRMs contained several TFs that are well known in the context of statin exposure. Among the most frequent TFs are Kruppel-like factors (e.g., KLF4 and KLF11), as well as forkhead transcription factors (e.g., FOXD3, FOXO3). KLFs have previously been published in the context of statins [34], [35] and were recently found to be involved in the adipogenesis pathway [30], [36]. The hypoxia inducible factor 1 (HIF1A) was previously found to be involved in the regulation of the ATP-binding cassette (ABCA1) and plays a role in hypoxia mediated inhibition of cholesterol synthesis [37], [38]. As an example, the

TABLE I: **Clustering and CRM-detection results on pravastatin dataset.** Clustering by EDISA revealed 10 clusters of co-expressed genes that are mainly involved in steroid-, lipid- and drug-metabolism. Column two gives the number of genes within each cluster. Significantly enriched gene ontology categories and the corrected enrichment p-values are depicted in columns three and four, respectively. Column six shows the resulting TFs that correspond to the CRMs with best specificity values. Several forkhead transcription factors (e.g., FOXD3, FOXO3) and Kruppel-like factors were detected, which have previously been published in the context of statins [34], [35].

Id	Genes	Enrichment	p-value	CRMs
1	19	steroid metab.	$3.7 \cdot 10^{-5}$	NFYB, KLF4, SP1, ZNF343
2	36	drug metab.	$2.6 \cdot 10^{-2}$	GATA4, MYOD1, FOXO3, PITX1
3	24	lipid metab.	$6.3 \cdot 10^{-4}$	PITX2, FOXD3, ESR2, TCF3
4	21	steroid metab.	$8.6 \cdot 10^{-7}$	SP3, NFYA, BTD
5	32	steroid metab.	$5.8 \cdot 10^{-4}$	SP3, SP1, NFYA, KLF4
6	23	na	na	na
7	19	steroid metab.	$5.7 \cdot 10^{-3}$	KLF11, MAZ, ZBTB, SP1, NFYA
8	20	na	na	na
9	23	steroid metab.	$8.6 \cdot 10^{-7}$	SP3, KLF4, MAZ, SP1, NFYA
10	17	steroid metab.	$1.3 \cdot 10^{-9}$	BTD, NFYA, SP4, HIF1A, SP1

CRM-detection results of cluster 9 are depicted in Fig. 2. As depicted in Fig. 2 (c), several pravastatin specific TFs were found for steroid metabolism cluster 9, i.e., MAZ, KLF4, SP1, SP3 and NFYA. The detected TFs, i.e., NFYA, KLF4, SP1, and SP3 are known regulators of steroid metabolism in the context of statin treatment [28], [29], [35]. The MYC-associated zinc finger protein (MAZ) has not been reported before in this context. The complete multi-objective CRM-detection experiments were robustly replicated using 10 multi-runs.

IV. Implementation and availability

The new algorithm proposed in this work is entirely implemented in Java™, from which the following libraries and external programs are called: EvA2 [25], BioJava [39] and RSA-Tools [40]. It is integrated into a computational workbench, which supports nearly every step in the process from microarray gene expression data to the final CRMs [41]. The program is capable of retrieving sequences, performing TFBS scans on these sequences and finally, searching for CRMs. It can be launched via Java™ web start and possesses a graphical user interface based on SWT. For large datasets, the program is able to create jobs and processes them in parallel from the command line on various computer cluster systems. The results can be evaluated, visualized and validated in many ways. For more information on the program, please see the applications note [41] and the documentation which is also available from the URL stated below. The program is free of charge and can be accessed via the following link: http://www.ra.cs.uni-tuebingen.de/software/ModuleMaster/.

V. Discussion

The main idea behind this work is based on the observation that TFs and their target genes often correlate at the gene expression level. It provides a new mathematical concept to combine both sources of evidence in an appropriate way. The main outcome of this work is a new method for the detection of CRMs that can be used for the reconstruction of transcriptional regulatory networks from microarray data. The proposed method consists of a computational framework that covers all steps from promoter retrieval through CRM detection, CRM evaluation and transcriptional network visualization. The algorithm expects several crucial input parameters, which depend on the underlying dataset and need therefore to be specified by the user, i.e., promoter length and module size l. The length of promoter sequences strongly depends on the complexity of the underlying organism and the size of intergenic sequences contained in the genomes. For higher organisms (e.g. human) regulatory sequences may be several thousand base pairs long, whereas for lower organisms (e.g. yeast) promoter sequences may consist of only several hundred base pairs [42], [43]. Furthermore, promoter sequences may vary substantially even within organisms and depend on the orientation of the genes towards each other. Also the CRM sizes might strongly depend on the complexity of the model organism. Therefore, on the one hand side, promoter length and module size should be determined according to such biological issues. On the other hand, it should be kept in mind that both parameters strongly influence the size of the search space and, hence, the chance to find biologically meaningful results. For example, we took the human proximal promoters (1,000 bp upstream of the transcription start site) for two reasons: important regulation of human promoters by multiple TFs does occur close to the transcription start although some enhancers can be much farther away, and to reduce the overall complexity for the computational problem. Depending on how one generates their clusters of genes and therefore how many are obtained, the average molecular biologist can obtain new candidates for testing in the laboratory within short time compared to months and years sometimes needed in traditional molecular biology. Moreover, the success of this method depends on the amount and quality of PFMs that are characterized experimentally and used as a priori knowledge. If there are only few PFMs known for a certain organism, it does not make sense to screen for large sized *cis*-regulatory modules. As mentioned above, also the PFM cutoff strategy should be chosen according to these issues. The core algorithm is a MOGA, which optimizes three heterogeneous objectives in order to find biologically meaningful regulatory relationships. These objectives are neither correlated nor directly opposed. Additionally, no general preference of one objective over the other can be made in advance. Thus, Pareto-optimal solutions returned by the algorithm should be ranked according to their specificity. Since genetic algorithms are stochastic optimization procedures, the solutions of different runs may differ when different random seeds are used. Depending on the search space, the algorithm returns stable results, i.e., more or less the same solutions are found when the procedure is repeated multiple times. In genetic algorithms recombination is the main search operator. To this end, the crossover rate is

usually set to values between 0.7 and 1 [44], [45]. In order to avoid the loss of genetic variants, a certain amount of mutation is essential for any genetic algorithm but is usually very small. In this work, the crossover operator was applied with a probability 0.7 and the mutation operator was called with probability of 0.1, which are the default settings of the EvA2 workbench [25]. The actual probability of changing one bit after calling the mutation operator is set to $\frac{1}{l}$ according to [46]. In the end, all these parameters strongly depend on the specific optimization problem, i.e., on the complexity of the organism, the amount of available PFMs, and the underlying microarray dataset. The best parameter combinations can theoretically be determined by a grid search, which is computationally demanding. If a parameter grid search is computationally not feasible for a given problem, we recommend to use the default parameters of the EvA2 evolutionary optimization workbench [25]. As proposed in the applications note [41], the user may specify all parameters mentioned above through a graphical user interface and adapt them according to the particular optimization problem. This also holds for the termination criterion, which was set to $30,000$ overall evaluations or $5,000$ evaluations without Pareto front changes and is based on the observation that in most cases (i.e., all clusters within three different datasets) convergence has been achieved before the first $10,000$ fitness evaluations.

VI. Conclusion

In this work, we present a new algorithm for the detection of *cis*-regulatory modules and for the reconstruction of transcriptional regulatory networks from sets of co-expressed genes. The novelty of our approach is the integration of three sources of evidence: (1) patterns of transcription factor binding sites, (2) multivariate linear relationships between transcription factors and their target genes, and (3) pathway connectivity scores based on protein-protein- and protein-chemical interaction data. We show here, that multi-objective evolutionary optimization is well suited for the integration of heterogeneous sources of evidence to find new regulatory dependencies. A multitude of known specific regulatory relationships in *Homo sapiens* drug-response data was successfully confirmed.

Acknowledgment

We would like to thank all those, who have directly or indirectly contributed to this work. Special thanks go to Marcel Kronfeld for his EvA2 support. This work was funded by the Federal Ministry of Education and Research (BMBF, Germany). Virtual Liver network, grant number 0315756.

References

[1] E. Segal, M. Shapira, A. Regev, D. Pe'er, Dana Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat Genet*, vol. 34, pp. 166 – 176, 2003.

[2] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, and K. Harter, "The atgenexpress global stress expression data set: protocols, evaluation and model data analysis of uv-b light, drought and cold stress responses." *Plant J*, vol. 50, no. 2, pp. 347–363, Apr 2007.

[3] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements." *Nat Rev Genet*, vol. 5, no. 4, pp. 276–287, Apr 2004. [Online]. Available: http://dx.doi.org/10.1038/nrg1315

[4] X. Yu, J. Lin, D. Zack, and J. Qian, "Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors," *BMC Bioinformatics*, vol. 8, no. 1, p. 437, 2007.

[5] S. M. Kielbasa and M. Vingron, "Transcriptional autoregulatory loops are highly conserved in vertebrate evolution," *PLoS ONE*, vol. 3, no. 9, p. e3210, 2008.

[6] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biol*, vol. 7, no. 5, p. R36, 2006.

[7] M. Kazemian, C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. S. Hammonds, S. E. Celniker, S. Kumar, S. A. Wolfe, M. H. Brodsky, and S. Sinha, "Quantitative analysis of the drosophila segmentation regulatory network using pattern generating potentials." *PLoS Biol*, vol. 8, no. 8, 2010. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.1000456

[8] S. Aerts, P. Van Loo, Y. Moreau, and B. De Moor, "A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes," *Bioinformatics*, vol. 20, no. 12, pp. 1974–1976, 2004.

[9] P. V. Loo, S. Aerts, B. Thienpont, B. D. Moor, Y. Moreau, and P. Marynen, "ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues?" *Genome Biol*, vol. 9, no. 4, p. R66, Apr 2008.

[10] T.-h. Lin, P. Ray, G. Sandve, S. Uguroglu, and E. Xing, "Baycis: A bayesian hierarchical hmm for cis-regulatory module decoding in metazoan genomes," in *Research in Computational Molecular Biology*, ser. Lecture Notes in Computer Science, M. Vingron and L. Wong, Eds. Springer Berlin / Heidelberg, 2008, vol. 4955, pp. 66–81.

[11] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley, 2001.

[12] C. A. C. Coello, D. A. V. Veldhuizen, and G. B. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, 2002.

[13] J. Supper, M. Strauch, D. Wanke, K. Harter, and A. Zell, "EDISA: extracting biclusters from multiple time-series of gene expression profiles," *BMC Bioinformatics*, vol. 8, p. 334, 2007.

[14] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor, "Computational detection of cis-regulatory modules," *Bioinformatics*, vol. 19, pp. ii5–14, 2003.

[15] A. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. Kel-Margoulis, and E. Wingender, "MATCHTM: a tool for searching transcription factor binding sites in DNA sequences," *Nucl. Acids Res.*, vol. 31, no. 13, pp. 3576–3579, 2003.

[16] J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin, "Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update." *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D102–D106, Jan 2008. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm955

[17] E. Wingender, A. E. Kel, O. V. Kel, H. Karas, T. Heinemeyer, P. Dietze, R. Knueppel, A. G. Romaschenko, and N. A. Kolchanov, "TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation," *Nucleic Acids Res*, vol. 25, no. 1, pp. 265–268, Jan 1997.

[18] S. L. Crawford, "Correlation and regression," *Circulation*, vol. 114, no. 19, pp. 2083–2088, Nov 2006.

[19] J. Supper, L. Spangenberg, H. Planatscher, A. Dräger, A. Schröder, and A. Zell, "BowTieBuilder: modeling signal transduction pathways." *BMC Syst Biol*, vol. 3, p. 67, 2009. [Online]. Available: http://dx.doi.org/10.1186/1752-0509-3-67

[20] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Müller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, "STRING 8–a global view on proteins and their functional interactions in 630 organisms." *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D412–D416, Jan 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn760

[21] M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer, and P. Bork, "STITCH 2: an interaction network database for small molecules and proteins." *Nucleic Acids*

Res, vol. 38, no. Database issue, pp. D552–D556, Jan 2010. [Online]. Available: http://dx.doi.org/10.1093/nar/gkp937

[22] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, January 1989.

[23] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II," *Lecture Notes in Computer Science*, pp. 849–858, 2000.

[24] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[25] M. Kronfeld, *EvA2 Short Documentation*, Center for Bioinformatics Tübingen, University of Tübingen, 2008.

[26] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, pp. R80+, 2004. [Online]. Available: http://dx.doi.org/10.1186/gb-2004-5-10-r80

[27] A. Schröder, J. Wollnik, C. Wrzodek, A. Dräger, M. Bonin, O. Burk, M. Thomas, U. M. Zanger, and A. Zell, "Inferring statin-induced gene regulatory relationships in primary human hepatocytes," *Bioinformatics*, p. submitted, 2011.

[28] Y.-C. Lin, J.-H. Lin, C.-W. Chou, Y.-F. Chang, S.-H. Yeh, and C.-C. Chen, "Statins increase p21 through inhibition of histone deacetylase activity and release of promoter-associated hdac1/2." *Cancer Res*, vol. 68, no. 7, pp. 2375–2383, Apr 2008. [Online]. Available: http://dx.doi.org/10.1158/0008-5472.CAN-07-5807

[29] C. Qin, I. Samudio, S. Ngwenya, and S. Safe, "Estrogen-dependent regulation of ornithine decarboxylase in breast cancer cells through activation of nongenomic camp-dependent pathways." *Mol Carcinog*, vol. 40, no. 3, pp. 160–170, Jul 2004. [Online]. Available: http://dx.doi.org/10.1002/mc.20030

[30] K. Birsoy, Z. Chen, and J. Friedman, "Transcriptional regulation of adipogenesis by klf4." *Cell Metab*, vol. 7, no. 4, pp. 339–347, Apr 2008. [Online]. Available: http://dx.doi.org/10.1016/j.cmet.2008.02.001

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000.

[32] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res*, vol. 27, no. 1, pp. 29–34, Jan 1999.

[33] J. Ludbrook, "Multiple comparison procedures updated." *Clin Exp Pharmacol Physiol*, vol. 25, no. 12, pp. 1032–1037, Dec 1998.

[34] S. Sen-Banerjee, S. Mir, Z. Lin, A. Hamik, G. B. Atkins, H. Das, P. Banerjee, A. Kumar, and M. K. Jain, "Kruppel-like factor 2 as a novel mediator of statin effects in endothelial cells." *Circulation*, vol. 112, no. 5, pp. 720–726, Aug 2005. [Online]. Available: http://dx.doi.org/10.1161/CIRCULATIONAHA.104.525774

[35] T. T. Tuomisto, H. Lumivuori, E. Kansanen, S.-K. Hkkinen, M. P. Turunen, J. V. van Thienen, A. J. Horrevoets, A.-L. Levonen, and S. Yl-Herttuala, "Simvastatin has an anti-inflammatory effect on macrophages via upregulation of an atheroprotective transcription factor, kruppel-like factor 2." *Cardiovasc Res*, vol. 78, no. 1, pp. 175–184, Apr 2008. [Online]. Available: http://dx.doi.org/10.1093/cvr/cvn007

[36] C. W. Brey, M. P. Nelder, T. Hailemariam, R. Gaugler, and S. Hashmi, "Krppel-like family of transcription factors: an emerging new frontier in fat biology." *Int J Biol Sci*, vol. 5, no. 6, pp. 622–636, 2009.

[37] S.-H. Lee, K. H. Koo, J.-W. Park, H.-J. Kim, S.-K. Ye, J. B. Park, B.-K. Park, and Y.-N. Kim, "Hif-1 is induced via egfr activation and mediates resistance to anoikis-like cell death under lipid rafts/caveolae-disrupting stress." *Carcinogenesis*, vol. 30, no. 12, pp. 1997–2004, Dec 2009. [Online]. Available: http://dx.doi.org/10.1093/carcin/bgp233

[38] P. Ugocsai, A. Hohenstatt, G. Paragh, G. Liebisch, T. Langmann, Z. Wolf, T. Weiss, P. Groitl, T. Dobner, P. Kasprzak, L. Gbls, A. Falkert, B. Seelbach-Goebel, A. Gellhaus, E. Winterhager, M. Schmidt, G. L. Semenza, and G. Schmitz, "Hif-1beta determines abca1 expression under hypoxia in human macrophages." *Int J Biochem Cell Biol*, vol. 42, no. 2, pp. 241–252, Feb 2010. [Online]. Available: http://dx.doi.org/10.1016/j.biocel.2009.10.002

[39] R. C. G. Holland, T. A. Down, M. Pocock, A. Prlič, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer, and M. J. Schreiber, "BioJava: an open-source framework for bioinformatics," *Bioinformatics*, vol. 24, no. 18, pp. 2096–2097, Sep 2008.

[40] M. Thomas-Chollier, O. Sand, J.-V. Turatsinze, R. Janky, M. Defrance, E. Vervisch, S. Brohee, and J. van Helden, "RSAT: regulatory sequence analysis tools," *Nucleic Acids Res*, vol. 36, no. Web Server issue, pp. W119–W127, Jul 2008.

[41] C. Wrzodek, A. Schröder, A. Dräger, D. Wanke, K. W. Berendzen, M. Kronfeld, K. Harter, and A. Zell, "ModuleMaster: a new tool to decipher transcriptional regulatory networks," *Biosystems*, vol. 99, no. 1, pp. 79–81, October 2010.

[42] R. V. Davuluri, I. Grosse, and M. Q. Zhang, "Computational identification of promoters and first exons in the human genome." *Nat Genet*, vol. 29, no. 4, pp. 412–417, Dec 2001. [Online]. Available: http://dx.doi.org/10.1038/ng780

[43] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome." *Nature*, vol. 431, no. 7004, pp. 99–104, Sep 2004. [Online]. Available: http://dx.doi.org/10.1038/nature02800

[44] J. Grefenstette, "Optimization of control parameters for genetic algorithms," *IEEE Trans. Syst. Man Cybern.*, vol. 16, no. 1, pp. 122–128, 1986.

[45] J. D. Schaffer, R. A. Caruana, L. J. Eshelman, and R. Das, "A study of control parameters affecting online performance of genetic algorithms for function optimization," in *Proceedings of the third international conference on Genetic algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 51–60.

[46] T. Bäck, "Optimal mutation rates in genetic search," in *Proceedings of the fifth International Conference on Genetic Algorithms*. Morgan Kaufmann, 1993, pp. 2–8.