Gene Regulatory Network Inference as Relaxed Graph Matching

Deborah Weighill
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115
dweighill@hsph.harvard.edu

Marouen Ben Guebila
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115
benguebila@hsph.harvard.edu

Camila Lopes-Ramos
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115
camilaramos@hsph.harvard.edu

Kimberly Glass
Channing Division of Network Medicine,
Brigham and Women's Hospital,
Harvard Medical School,
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115
rekrg@channing.harvard.edu

John Quackenbush
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Department of Medicine,
Channing Division of Network Medicine,
Brigham and Women's Hospital,
Harvard Medical School,
Boston, MA 02115
johnq@hsph.harvard.edu

John Platig
Channing Division of Network Medicine,
Brigham and Women's Hospital,
Harvard Medical School,
Boston, MA 02115
rejhp@channing.harvard.edu

Rebekka Burkholz
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston, MA 02115
rburkholz@hsph.harvard.edu

Abstract

Gene regulatory network inference is instrumental to the discovery of genetic mechanisms driving diverse diseases, including cancer. Here, we present a theoretical framework for PANDA, an established method for gene regulatory network inference. PANDA is based on iterative message passing updates that resemble the gradient descent of an optimization problem, OTTER, which can be interpreted as relaxed inexact graph matching between a gene-gene co-expression and a protein-protein interaction matrix. The solutions of OTTER can be derived explicitly and inspire an alternative spectral algorithm, for which we can provide network recovery guarantees. We compare different solution approaches of OTTER to other inference methods using three biological data sets, which we make publicly available to offer a new application venue for relaxed graph matching in gene regulatory network inference. We find that using modern gradient descent methods

Preprint. Under review.

with superior convergence properties solving OTTER outperforms state-of-the-art gene regulatory network inference methods in predicting binding of transcription factors to regulatory regions.

1 Introduction

Next generation genome sequencing technology has revolutionized genetic research and provides data at an unprecedented scale. This progress facilitates large, genome-scale studies which provide new insights into gene regulation, including the control of protein production through the expression of genes. Proteins influence higher level cellular functions, which are often altered during the development and progression of different diseases, including cancer. To gain an understanding of the gene regulatory mechanisms perturbed by a disease, it is common practice to infer and compare associated gene regulatory networks (GRNs) [20, 22, 26, 34]. In many cases, these networks are weighted, bipartite, and have a representation as matrix W. W consist of two types of nodes – transcription factors (TFs) and genes. A TF is a protein that can bind to the DNA in the vicinity of a gene and regulate its expression, which constitutes a link in the gene regulatory network, see Fig. 1. Gene expression often leads to the production of associated proteins (including TFs) that can interact, form higher-order protein complexes, regulate genes, etc. Possible interactions of such proteins are studied in detail and are commonly summarized in a known matrix P. Also data on gene expression is widely accessible but is context specific, i.e. it depends on the tissue type, disease, etc. Based on this data, we can define a gene-gene co-expression matrix C. A more detailed explanation of gene regulation is given in the supplement. We also recommend the review article by Todeschini et al. [30]).

Our main goal is to infer a gene regulatory matrix W given the two matrices P and C. We pose this as a non-convex optimization problem, OTTER (**O**p**t**imize **t**o **E**stimate **R**egulation). It is related to relaxed inexact graph matching, which seeks agreement between two graphs (that are represented by matrices like P and C). The gene regulatory matrix W could be seen as relaxed permutation matrix that matches vertices in P and C. Similarly to graph matching, OTTER is theoretically tractable but the solutions are nonunique, which explains in part why GRN inference is challenging. We characterize the solution space and propose two distinct approaches for model selection, a spectral approach and gradient descent. For both, we provide theoretical network recovery guarantees. While the spectral method is robust to small noise, gradient descent is more reliable in higher noise settings and outperforms state-of-the-art gene regulatory network inference techniques.

Contributions 1) We pose a novel optimization problem for gene regulatory network inference, OTTER, which is analytically tractable. 2) We gain insights into a state-of-the-art GRN inference method, PANDA [12], as OTTER gradient descent resembles the related message passing equations. 3) We characterize OTTER's solution space and derive a spectral algorithm on this basis, for which we give network recovery guarantees. 4) We solve the gradient flow dynamics associated with gradient descent for OTTER. 5) We draw a connection from OTTER to relaxed graph matching and open a new application area for related algorithms. 6) OTTER gradient descent outperforms the current state of the art in GRN inference on three challenging biological data sets related to cancer. 7) We make the processed data publicly available to ease the use for researchers without a genetics background and to foster furthter innovation in relaxed graph matching and GRN inference.

Related work The OTTER objective is inspired by a state-of-the-art GRN inference method, PANDA (Passing Attributes between Networks for Data Assimilation) [12]. PANDA integrates multiple data sources through a message passing approach, which we find resembles the gradient descent of OTTER. A derivation is given in the supplement. PANDA has been used to investigate gene regulatory relationships in both tissue specific [28] as well as several disease contexts, including chronic obstructive pulmonary disease [20], asthma [26], beta cell differentiation [34], and colon cancer [22]. OTTER can be seen as a theoretically tractable simplification of PANDA, which is amenable to modern optimization techniques and draws connections to graph matching.

Graph matching [32] has strong theoretical foundations [16, 3] that can benefit a deeper understanding of OTTER and vice versa. In particular, the quadratic assignment problem (QAP) [1] and its variants [24] have a direct link to OTTER and can support similar biological theory. Graph matching has broad applications in computer science ranging from machine learning [7], pattern matching [36], vision

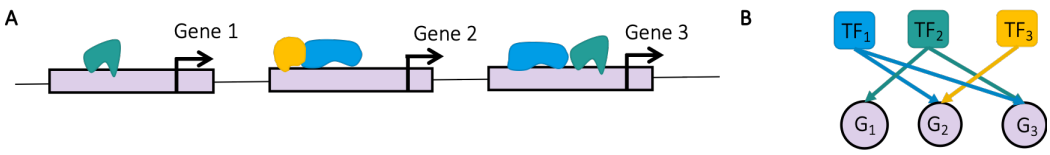

Figure 1: Gene regulation. A. Transcription factors (TFs) are represented by green, blue, and yellow objects that bind to the genome (gray band) in vicinity of the start site of a gene (black arrow) to regulate its expression. B. Representation of A as bipartite gene regulatory network.

[4, 35], and protein network alignment [27] to social network analysis [8], but has not been applied to gene regulatory network inference to the best of our knowledge. As we show, simple relaxed graph matching techniques outperform established GRN inference methods, which are detailed in the experiment section (Sec. 4).

2 Theoretical framework

Our goal is to learn a bipartite and weighted gene regulatory network which we represent as a matrix $W \in \mathbb{R}^{n_p \times n_c}$. An entry w_{ij} indicates whether TF protein i binds to a region of the DNA that is associated with gene j and has evidence that it regulates that gene. A larger weight w_{ij} is assumed to be associated with a higher probability of binding. We have n_p TFs and n_c genes, where the number of genes n_c is much larger ($n_p \ll n_c$). We assume that we observe only squares of W, which are given as symmetric matrices $P \in \mathbb{R}^{n_p \times n_p}$ (the protein-protein interaction matrix) and $C \in \mathbb{R}^{n_c \times n_c}$ (the gene-gene co-expression matrix). Thus, W explains protein interactions as $WW^T \approx P$ and gene co-expression as $W^T W \approx C$. We formulate this reasoning as the following optimization problem:

$$\min_{W \in \mathbb{R}^{n_p \times n_p}} f(W) \text{ with } \mathrm{f(W)} = \frac{(1-\lambda)}{4}\|\mathrm{WW^T} - \mathrm{P}\|^2 + \frac{\lambda}{4}\|\mathrm{W^T W} - \mathrm{C}\|^2 + \frac{\gamma}{2}\|\mathrm{W}\|^2, \quad (1)$$

which we call OTTER. $\lambda \in [0, 1]$ denotes a tuning parameter that moderates the influence of P versus C, and γ corresponds to a potential regularization. In principle, we could choose any matrix norm but limit our following discussion to the Frobenius norm $\|A\|^2 := \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 = \mathrm{tr}(A^T A)$ for a matrix $A = (a_{ij})$. For this choice, gradient descent resembles most closely the related message passing equations of PANDA and we can derive the corresponding solutions.

These solutions depend on the spectral decomposition of $P = U_p D_p U_p^T$ and $C = V_c D_c V_c^T$, which exist with respect to orthogonal U_p and V_c, as P and C are symmetric. Otherwise, the same results hold for the spectral decomposition of $(P + P^T)/2$ and $(C + C^T)/2$. D_p and D_c are diagonal matrices containing the eigenvalues of the respective matrix. In a slight abuse of notation, we denote with D_p a matrix $D_p \in \mathbb{R}^{n_p \times n_p}$ and, if convenient, a matrix $D_p \in \mathbb{R}^{n_c \times n_c}$, which is padded with zeros accordingly. Furthermore, let $M^{[n_p]} = (m_{ij})_{i \leq n_p, j \leq n_p}$ denote a submatrix of a larger matrix M with dimension $n_p \times n_p$. Without loss of generality, we assume that the eigenvalues $d_{p,ii}$ of P are indexed in descending order; $d_{p,ii} \geq d_{p,jj}$ for $i < j$. For C however, we require a good matching with P. We therefore assume implicitly that the distance of D_c to D_p is minimized with respect to permutations of the eigenvalues of C, that is $\|D_c - D_p\|^2 = \min_{\pi \in \mathcal{P}} \|D_{c,\pi} - D_p\|^2$, where $\mathcal{P}$ denotes the set of permutations of $\{1, \cdots, n_c\}$ and $D_{c,\pi}$ the corresponding ordering of eigenvalues on the diagonal. If D_p and D_c show little discrepancy, this will result in the eigenvalues of C being in descending order as well. Note that $\Delta := D_c^{[n_p]} - D_p$ measures the discrepancy in our biological hypothesis from the start. Usually, we can further assume that P and C are positive semi-definite so that $d_{ii} \geq 0$, meaning that our model is specified well enough such that C and P have roots. Based on this, we can characterize the solution space $\mathcal{S}$ in the general case.

Theorem 1. *For given $P \in \mathbb{R}^{n_p \times n_p}$ with $P = P^T$ and $C \in \mathbb{R}^{n_c \times n_c}$ with $C = C^T$, for any spectral decomposition $P = U_p D_p U_p^T$ and $C = V_c D_c V_c^T$, $\lambda \in [0, 1]$, the minimization problem (1) has solutions $W^* \in \mathcal{S}$ with singular value decomposition $W^* = U_p D_w V_c^T$, where*

$$d_{w,ii} = \sqrt{\max\left((1-\lambda)d_{p,ii} + \lambda d_{c,ii} - \gamma, 0\right)} \quad (2)$$

for $i \leq n_p$. For $d_{w,ii} = 0$, the corresponding columns of U_w and V_w are not restricted to the eigenvectors of P and C. The eigenvalues of C are ordered such that $D_c = D_{c,\pi}$, where the

permutation solves the minimization problem

$$\pi = \mathrm{argmin}_{\pi' \in \mathcal{P}} \frac{\lambda(1-\lambda)}{2} \|\mathrm{D_p} - \mathrm{D}_{\mathrm{c},\pi'}^{[\mathrm{n_p}]}\|^2 - \frac{\lambda}{2} \|\mathrm{D}_{\mathrm{c},\pi'}^{[\mathrm{n_p}]}\|^2 + (1-\lambda)\gamma \,\mathrm{tr}\left(\mathrm{D}_{\mathrm{c},\pi'}^{[\mathrm{n_p}]}\right), \tag{3}$$

For $\Delta := D_c^{[n_p]} - D_p$*, we further assume that*

$$\frac{1}{\lambda^2}\left(\|P\|^2 + \frac{1}{n_p}\,\mathrm{tr}\,(P)^2\right) + \frac{1}{n_p}\left(\mathrm{tr}(\Delta)\right)^2 > \|\Delta\|^2. \tag{4}$$

Condition (4) is usually met and it is a minor technicality to exclude alternative global minima of Objective (1) that defy our intuition. The nature of these alternatives is discussed in detail in the proof of the Theorem in the supplement.

According to Thm. 1, OTTER (1) has at least 2^{n_p} different solutions. Each column $u_{:i}$ of U_p has two optional signs that do not alter the spectral decomposition of P but can lead to a different W^*. The same applies to columns $v_{:i}$ of V_c. Only the product of corresponding columns ($u_{:i}$ and $v_{:,i}$) determines the respective solution W^*, as we have $w_{ij}^* = \sum_k d_{w,kk} u_{ik} v_{jk}$. This leaves us with 2^{n_p} alternatives. If the spectra are not simple, such that some eigenspaces have multiple choices of basis functions, we have additional degrees of freedom in constructing the solutions.

As a consequence, we face a model selection problem and require additional information to make an informed decision. Indeed, PANDA takes as an additional input an initial guess of a gene regulatory matrix $W_0 \in \mathbb{R}^{n_p \times n_c}$, which is typically based on data from a scan of the genome sequence using known transcription factor binding motifs. Assuming that W_0 provides good evidence, our first proposal selects the closest solution to W_0 using a spectral approach.

2.1 A spectral method for solving OTTER

For additional evidence W_0, we turn problem (1) into

$$\min_{W \in \mathcal{S}} \|W - W_0\|^2 = \min_{d_{s,ii} \in \{-1,1\}} \|D_w D_s - U_p^T W_0 V_c\|^2 \tag{5}$$

where we fix U_p and V_c and assume that W has simple singular values, meaning the singular values are different from each other. For simplicity, we write $M_0 = U_p^T W_0 V_c$. For simple spectra of P and C and matched eigenvalues, W^* is unique and given by $W^* = U_p D_w^* V_c^T$ with $d_{w,ii}^* = d_{w,ii}\mathrm{sign}(\mathrm{m_{0,ii}})$, where $d_{w,ii}$ is defined as in Thm. 1 and $\mathrm{sign(x)} = 1$ for $x \geq 0$ and $\mathrm{sign(x)} = -1$ for $x < 0$.

The question is how well this approach performs in the presence of noise. As the next proposition shows, even if only W_0 is noise corrupted so that we know D_W, perfect recovery of W is unlikely for large-scale problems. Let Φ denote the cumulative distribution function (cdf) of a standard normal and $X \sim \mathrm{Ber(p)}$ a Bernoulli random variable with success probability p.

Proposition 2. *Assume that we observe* $P = W^{*T}W^*$*,* $C = W^*W^{*T}$*, and* $W_0 = W^* + E$ *for a true underlying* $W^* \in \mathbb{R}^{n_p \times n_p}$ *and noise* $E \in \mathbb{R}^{n_p \times n_p}$ *with independent identically normally distributed components* $e_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$*. Further assume that* P *and* C *have a simple spectrum* $\{d_1, \ldots, d_{n_p}\}$*. Then, for the spectral approach* $\widehat{W} = \mathrm{argmin}_{\mathrm{W} \in \mathcal{S}} \|\mathrm{W} - \mathrm{W_0}\|^2$ *with* $\gamma = 0$*, the recovery loss is distributed as* $\|\widehat{W} - W^*\|^2 = 4\sum_{i=1}^{n_p} d_i^2 R_i$*, where* $R_i \sim \mathrm{Ber}\left(\Phi\left(-\mathrm{d_i}/\sigma\right)\right)$ *for* $d_i > 0$ *and* $R_i = 0$ *for* $d_i = 0$ *are independent. For any* $\epsilon > 0$*, the following holds with the usual Chernoff bound:*

$$\mathbb{P}\left(\|\widehat{W} - W^*\|^2 \leq \epsilon\right) \geq 1 - \exp\left(\epsilon - \mu - \frac{\epsilon}{4}\delta \log\left(\frac{\epsilon}{\mu}\right)\right),$$

where $\mu = \sum_i p_i$ *and* $\delta = \frac{1}{\max_i\left(d_{w,ii}^2\right)}$ *for* $\epsilon \leq \mu$ *and* $\delta = \frac{1}{\min_i\left(d_{w,ii}^2\right)}$ *otherwise.*

In particular, for the probability of perfect recovery ($\epsilon = 0$) we have $\mathbb{P}\left(\widehat{W} = W^*\right) = \prod_{i=1}^{n_p}\left(1 - \Phi\left(-d_i/\sigma\right)\right)$. In our examples, $n_p = 1636$ and $d_{\min} \approx 0.0001$. To achieve a probability of at least 0.5, we could allow for a noise variance of $\sigma \approx 3 \cdot 10^{-5}$. In many applications, this would be a reasonable range, considering that we have $n_p n_c \approx 4.4 \cdot 10^7$ matrix entries. Biological data is known to be noisy, so we expect levels of noise in P and C that will require regularization.

Regularization Accounting for noise in P and C complicates the analysis considerably, as the spectral decomposition of P and C (including their eigenvectors) are distorted. We focus on the population version (the average matrices) to motivate the need for additional regularization ($\gamma > 0$). Depending on the source of the noise, the spectrum can become biased. To see this, assume noise of the form $P = (W^* + E_p)(W^* + E_p)^T$ and $C = (W^* + E_c)^T(W^* + E_c)$ or $P = W^*W^{*^T} + E_pE_p^T$ and $C = W^{*T}W^* + E_c^TE_c$, respectively. If E_p and E_c have iid entries with zero mean, variance $\sigma^2_{p/c}$, and a symmetric distribution, we get $\mathbb{E}(P) = W^*W^{*T} + \sigma_p^2 I$ and $\mathbb{E}(C) = W^{*T}W^* + \sigma_c^2 I$, where I denotes the respective identity matrix. The choice $\gamma = \lambda\sigma_c^2 + (1-\lambda)\sigma_p^2$ in Eq. (1) can compensate for this spectral shift. It should be noted that, Thm. 1 states that such a l2 regularization alters the solutions to Problem (1) in two ways. Not only are the singular values of W^* shifted by $-\gamma$ to compensate for the biases introduced by the noise, also the matching of the eigenvalues of P and C is influenced by the additional penalty $\gamma(1-\lambda)\operatorname{tr}\left(D_{c,\pi}^{[n_p]}\right)$ in Eq. (3). Consequently, it may be optimal to pair the eigenvalues of P with smaller eigenvalues of C rather than larger ones if γ is large.

While the spectral method can be powerful in a setting in which noise is controlled such that our assumptions are met approximately, the gradient descent alternative gives us more tuning options, including the step size and early stopping, that will allow us to stay closer to the initial guess W_0.

2.2 Gradient descent for solving OTTER

Indeed, the message passing equations of PANDA resemble a gradient descent procedure, where the gradient of Objective (1) is given as $\nabla f(W) = WW^TW - (1-\lambda)PW - \lambda WC + \gamma W$. We explain this relationship in detail in the supplement. In our experiments, we used the ADAM method [18] for gradient descent but alternatives are equally applicable. From a theoretical perspective, we can understand the gradient dynamics for specific choices of W_0. For this purpose, we take the continuous time approximation (corresponding to infinitesimally small step size) and study the corresponding gradient flow:

$$\tau\frac{dW}{dt} = -\nabla f(W) = -WW^TW + (1-\lambda)PW + \lambda WC - \gamma W, \tag{6}$$

where we set the time unit $\tau = 1$ in the following for simplicity. If the initial W_0 has a similar singular value decomposition as a solution, the differential equation decouples and we can solve the resulting one-dimensional ordinary differential equations for the diagonal elements explicitly.

Proposition 3. *For initial $W_0 = U_pD_0V_c^T$ with $U_pD_pU_p^T$ and $V_cD_cV_c^T$, the solution of the gradient flow (6) is given by $W(t) = U_pD_tV_c^T$ with*

$$d_{t,ii} = \operatorname{sign}(\mathrm{d_{0,ii}})\mathrm{d_{w,ii}}\sqrt{\frac{1}{2}\mathrm{h}\left(\mathrm{d_{w,ii}^2 t} + \mathrm{h}^{-1}\left(2\frac{\mathrm{d_{0,ii}^2}}{\mathrm{d_{w,ii}^2}} - 1\right)\right) + 1},$$

where $h(x) = \tanh(\mathrm{x})$ if $d_{0,ii}^2 < d_{w,ii}^2$ and $h(x) = \coth(\mathrm{x})$ otherwise.

Note that the square root factor converges to 1 for $t \to \infty$ in both cases. Hence, the final solution inherits the signs $\operatorname{sign}(\mathrm{d_{0,kk}})$ of the initialization similar to our spectral approach. Thus, if we start from a reasonable guess W_0 that diagonalizes with respect to the same U and V as the global minima, gradient descent will converge to the closest global minimum for small enough learning rate. For general W_0, however, it is important to keep in mind that gradient descent can converge to different solutions. It does not necessarily stay close to our initialization and can even get stuck in a local minimum.

3 Relation to inexact graph matching

The biological justification for the OTTER objective function is also consistent with quadratic assignments [1], a more common choice in graph matching. From $P \approx WW^T$ and $C \approx W^TW$ we could deduce $PW \approx WW^TW \approx WC$ and thus minimize the Quadratic Assignment Problem (QAP) objective

$$g(W) = \frac{1}{2}\|PW - WC\|^2 + \frac{\gamma}{2}\|W\|^2 \tag{7}$$

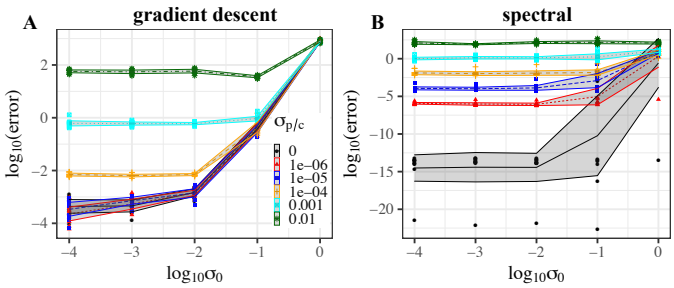

Figure 2: OTTER recovery error using (A) gradient descent and (B) spectral decomposition for artificial networks of size $n_p = 100$, $n_c = 200$ and Gaussian noise with variance $\sigma^2_{p/c}$ for P and C and σ^2_0 for W_0. Shaded regions correspond to the 0.95 confidence interval and lines to the average over 10 repetitions. The legend applies to both figures.

(with additional l2-regularization). In graph matching however, P and C are usually assumed to have the same dimension. They differ for inexact graph matching, but the smaller network is then supposed to be similar to a subgraph of the bigger one. Thus, the minimization is performed under the constraint that W is a permutation matrix. In contrast, we are interested not in a permutation matrix, but in a weighted network $W \in \mathbb{R}^{n_p \times n_c}$ that solves the relaxed QAP. The corresponding gradient is $\nabla g(W) = P^2 W + WC^2 - 2PWC + \gamma W$. While the related ADAM gradient descent approach is computationally more costly than the one solving OTTER, both approaches perform similarly well as we show in experiments (see Sec. 4).

This indicates that there is a great potential to perform gene regulatory network inference using alternative graph matching methods in future investigations. In particular, GRAMPA [9, 10] is a variant of QAP with strong recovery guarantees, also for the Wigner model where $P - C$ has iid noise entries. GRAMPA adds the term $-\delta \mathbf{1}^T W \mathbf{1}$ to the QAP objective (7), where $\mathbf{1}$ denotes a vector with all entries equal to one. As a consequence, the solution to the minimization problem becomes unique and explicitly solvable using a spectral approach. As the spectral version of OTTER, it does not perform well in estimating GRNs. Solving the GRAMPA objective function with gradient descent works better, though it is not competitive.

Graph matching can also be studied within the optimal transport framework [25, 29]. We could formulate the OTTER objective with respect to a nonstandard metric and regularization term. Since we are not searching for stochastic matrices W, this does not serve our purpose and we leave the transfer of related methods to gene regulatory network inference to future explorations.

4 Experiments

4.1 Experiments on synthetic data

To showcase the performance of OTTER for cases in which our assumptions are met, we create artificial data based on a ground truth W^* that we try to recover from noise corrupted inputs $W_0 = W^* + N_0$, $P = (W^* + N_p)(W^* + N_p)^T$, and $C = (W^* + N_c)^T (W^* + N_c)$. All noise entries are Gaussian and independently distributed with $n_{0,ii} \sim \mathcal{N}\left(0, \sigma_0^2\right)$, $n_{p,ii} \sim \mathcal{N}\left(0, \sigma_p^2\right)$, and $n_{c,ii} \sim \mathcal{N}\left(0, \sigma_c^2\right)$. To obtain a realistic ground truth for which we can repeat each experiment 10 times conveniently, we sub-sample (in each repetition) the ChIP-Seq network for the liver tissue to $n_p = 100$ and $n_c = 200$. (See the next section for more details.) As this is unweighted, we draw the weights iid from $\mathcal{N}(10^{-5}, 1)$. For each network, we use the spectral and the gradient descent version of OTTER and report the obtained recovery error $\||W - W^*\|^2$.

For simplicity, we do not align the eigenvalues of P and C optimally but arrange each of them in descending order in the spectral approach. We run ADAM gradient descent for the OTTER objective for 10^4 steps with the default ADAM parameters, as detailed in the supplement. For both the gradient decent and the spectral approach, we use parameters $\gamma = \sigma^2_{p/c} = \sigma^2_p = \sigma^2_c$ and $\lambda = 0.5$.

Table 1: TF binding prediction for different cancer tissues. The symbol † indicates that binding predictions were made only for TFs with ChIP-seq data due to high computational demands. The highest AUC-ROC/AUC-PR for each data set is shown in bold.

	AUC-ROC			AUC-PR		
METHOD	BREAST	CERVIX	LIVER	BREAST	CERVIX	LIVER
COR	0.5857	0.5692	0.5785	0.2774	0.2219	0.3094
PARTIAL COR†	0.5366	0.5209	0.5175	0.2361	0.1952	0.2525
ARACNE	0.6150	0.5234	0.5636	0.2858	0.2027	0.2986
GENIE3	0.4818†	0.4832	0.4846	0.2064†	0.1836	0.2437
TIGRESS†	0.4945	0.4808	0.5018	0.2088	0.1845	0.2523
OTTER SPECTRAL	0.5787	0.5420	0.5345	0.2555	0.2024	0.2614
GRAMPA GRAD	0.6000	0.5896	0.5670	0.2729	0.2214	0.2796
QAP GRAD	0.5722	0.6031	0.5335	0.2526	0.2346	0.2587
OTTER GRAD	0.6936	0.6826	0.6477	**0.3752**	0.3176	0.3717
PANDA	0.6739	0.6642	0.6211	0.3481	0.2960	0.3503
	TRANSFORMED P AND W_0					
OTTER* SPECTRAL	0.6123	0.6070	0.5973	0.2773	0.2367	0.3183
GRAMPA* GRAD	0.6312	0.6312	0.6021	0.3150	0.2551	0.3098
QAP* GRAD	0.7114	0.7096	0.6883	0.3568	0.3216	0.3764
OTTER* GRAD	**0.7161**	**0.7215**	**0.7166**	0.3574	**0.3280**	**0.4063**
PANDA*	0.7008	0.6642	0.6180	0.3497	0.2676	0.3340

The results are shown in Fig. 2. For small levels of noise in P and C, the spectral approach performs reliably and better than gradient descent. However, for high $\sigma^2_{p/c}$ gradient descent outperforms the spectral method. Since we biological data are inherently noisy, gradient descent seems to be the method of choice. Furthermore, it provides us with additional tuning options that we can leverage to outperform state of the art methods.

4.2 Experiments on cancer data

The most abundant data source for studying gene regulation is gene expression data. These data are often measured using bulk RNA-sequencing (RNA-seq) with samples corresponding to different individuals.

Datasets and experimental set-up We obtained bulk RNA-seq data from the Cancer Genome Atlas (TCGA) [31]. The data is downloaded from recount2 [6] for liver, cervical, and breast cancer tumors and normalized and filtered as described in the supplement. The corresponding Pearson correlation matrix defines the gene-gene co-expression matrix C consisting of $n_c = 31,247$ genes for breast cancer, $n_c = 30,181$ for cervix cancer and $27,081$ for liver cancer. The protein-protein interaction matrix P is derived using laboratory experiments and represents possible interactions; we use a padded version of that provided in [28]. It consists of $n_p = 1,636$ potential TFs. Our initial guess of a gene regulatory network, W_0, is similar across tissues but varies slightly depending on the number of genes (n_c) included after filtering and normalization. W_0 is a binary matrix where there is a "1" if there is a TF sequence motif in the promoter of the target gene, and a "0" otherwise. Sequence motif mapping was performed using the FIMO software [13] from the MEME suite [2] and the GenomicRanges R package [21]. In addition to the standard PANDA input P and W_0, we also consider transformations thereof, i.e., $\tilde{P} = P + 2.2$ with element-wise addition and $\tilde{W}_0 = \tilde{P}W_0$, that have been inspired by aspects of GRAMPA graph matching and increase the performance of most methods considerably. Note that neither W_0, $\tilde{P}W_0$ or P carry sign information about edge weights so that we cannot infer whether TFs inhibit or activate the expression of a gene. We therefore focus on the prediction of link existence with the understanding that the type of interaction can be estimated *post hoc*.

Validation of gene regulatory networks is a major challenge. Data from Chromatin immunoprecipitation (ChIP)-sequencing (ChIP-seq) experiments, which measures the binding of TFs to DNA in the genome, provides a validation standard against which to benchmark our results. It provides only partial validation as only a small fraction or the collections of TFs are typically profiled for any tissue.

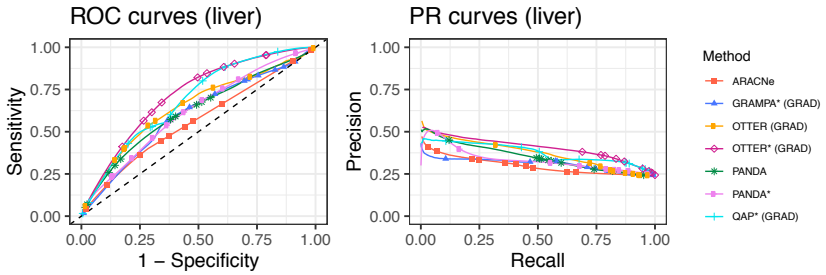

Figure 3: Performance curves for liver tissue.

However, because of the destructive nature of collecting molecular data from biological samples, ChIP-seq and RNA-seq data cannot be easily collected from the same biological samples, and so ChIP-seq data is generally collected from cell lines in the laboratory rather than from patient tumor samples. We used ChIP-seq data from the HeLa cell line (cervical cancer, 48 TFs), HepG2 cell line (liver cancer, 77 TFs) and MCF7 cell line (breast cancer, 62 TFs) available in the ReMap2018 database [5]. Because of the limitations in the available data we use AUC-ROC (area under the receiver operating characteristic curve) and AUPR (or AUC-PR) (area under the precision recall curve) to measure the performance of link classification on the subnetwork in each tissue that is constrained to the measured TFs.

Hyperparameter tuning of OTTER was assisted by MATLAB's bayesopt function utilizing a Gaussian process prior to maximize the joint AUC-PR for breast and cervix cancer, $\max \mathrm{AUPR}_{\mathrm{breast}} \cdot \mathrm{AUPR}_{\mathrm{cervix}}$. Breast and cervix data serve therefore as training data while the liver cancer data is an independent test set. The parameters of all compared methods are reported in the supplementary information.

Related literature and methods Many methods try to infer regulatory relationships solely based on gene expression with two possible (non-exclusive) objectives: structure learning and gene expression prediction. Note that the latter usually includes the former. TFs are proteins that are created from the mRNA expressed by their corresponding genes. Hence, predicting target gene expression from the expression of the genes coding for the TFs assumes a biologically reasonable structure.

The most common and basic approach is to analyse the Pearson correlation (COR) matrix or, if feasible, partial correlations (PARTIAL COR). Alternatives are based on mutual information, where ARACNe [19] is one of the most commonly used representatives. Among graphical models, mainly Gaussian graphical models are used because the learning algorithms have to scale to approximately $n_c = 30000$ genes (in the case of human tissue). The GLASSO [11] method is among the best performing candidates and uses LASSO regularization to enforce sparsity. However, it still does not scale to our setting, and thus we have omitted it from our analysis. Linear models [14] and random forests [15] have been used for a similar purpose, where TIGRESS [14] and GENIE3 [15] were top scorers at the DREAM5 challenge [23] (although the challenge was somewhat different from the GRN modeling we describe here). Both methods have high computational requirements and are less suitable for the human genome that consists of more than 25,000 genes. An alternative approach is to treat ChIP-seq binding predictions as a supervised learning problem [17, 33]. While such models can be quite accurate, they are limited to the small number of TFs for which ChIP-seq experiments are available and thus limited in their discovery of new gene regulatory relationships. Note that, in contrast, P and W_0 for PANDA-related objectives are always available and include a much larger set of known TFs.

Results Table 1 compares the feasible GRN inference and relaxed graph matching methods based on comparison to experimental ChiP-seq binding data. Overall, the OTTER* (or OTTER) gradient descent approach achieves the best performance on all tissues, in particular, on the liver test set (see also Fig. 3). An enrichment analysis of Gene Ontology terms between networks for healthy and cancerous liver tissue in the supplement provides additional evidence that OTTER* (GRAD) is biologically meaningful. The ADAM gradient descent solving the relaxed quadratic assignment problem (QAP* grad) is almost on par with OTTER* (GRAD).

In general, we observe better performance for the methods that incorporate additional biological evidence such as (transformed) protein-protein interactions and binding motifs, even though these are not tissue specific. A reason for this is that correlations in gene expression can be caused by many factors and that many TFs are expressed at very low levels but strongly activate their target genes, obscuring correlations between TFs and their targets Hence, graph matching approaches are a promising alternative to models that make predictions based on gene expression alone.

5 Discussion

We have gained theoretical insights into a state-of-the-art gene regulatory network inference method, PANDA. Our reformulation of the biological intuition behind PANDA as a non-convex optimization problem, OTTER, has multiple global minima, which we have characterized explicitly. Alternative solution approaches can therefore select different minima.

While a gradient descent approach resembles the original PANDA algorithm more closely, the spectral approach selects solutions closest to a biologically reasonable guess based on DNA motif information. The latter has network recovery guarantees in low noise settings and performs well on artificial data, while the former is more robust with respect to noise. The gradient descent outperforms state-of-the art gene regulatory network inference approaches on real world data sets corresponding to three human cancer tissues, which we make publicly available in their processed form for this task.

As we highlight, relaxed graph matching approaches like the relaxed quadratic assignment problem apply to this setting and achieve competitive performance. Hence, we see great potential in transferring alternative relaxed graph matching techniques to gene regulatory network inference in future investigations.

Data availability

OTTER is available in R, Python, and MATLAB through the netZoo packages: netZooR v0.7 (https://github.com/netZoo/netZooR), netZooPy v0.7 (https://github.com/netZoo/netZooPy), and netZooM v0.5 (https://github.com/netZoo/netZooM).

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`. DW, MG, CL, JQ, RB were supported by a grant from the US National Cancer Institute (1R35CA220523). JP acknowledges support from the US National Heart, Lung, and Blood Institute (NHLBI): K25HL140186 and KG from the K25 grant: K25HL133599. We thank Alkis Gotovos for helpful feedback on the manuscript.

References

[1] Yonathan Aflalo, Alexander Bronstein, and Ron Kimmel. On convex relaxation of graph isomorphism. *Proceedings of the National Academy of Sciences*, 112(10):2942–2947, 2015.

[2] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208, 2009.

[3] Boaz Barak, Chi-Ning Chou, Zhixian Lei, Tselil Schramm, and Yueqi Sheng. (nearly) efficient algorithms for the graph matching problem on correlated random graphs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9190–9198. Curran Associates, Inc., 2019.

[4] Alexander C. Berg, Tamara L. Berg, and Jitendra Malik. Shape matching and object recognition using low distortion correspondences. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 26–33 vol. 1, 2005.

[5] Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, and Benoit Ballester. Remap 2018: an updated atlas of regulatory regions from an integrative analysis of dna-binding chip-seq experiments. *Nucleic acids research*, 46(D1):D267–D275, 2018.

[6] Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. Reproducible RNA-seq analysis using recount2. *Nature biotechnology*, 35(4):319–321, 2017.

[7] Timothee Cour, Praveen Srinivasan, and Jianbo Shi. Balanced graph matching. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 313–320. MIT Press, 2007.

[8] Wenfei Fan. Graph pattern matching revised for social network analysis. In *Proceedings of the 15th International Conference on Database Theory*, ICDT '12, page 8–21, New York, NY, USA, 2012. Association for Computing Machinery.

[9] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations i: The gaussian model, 2019.

[10] Zhou Fan, Cheng Mao, Yihong Wu, and Jiaming Xu. Spectral graph matching and regularized quadratic relaxations ii: Erdős-rényi graphs and universality, 2019.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.

[12] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5), 2013.

[13] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[14] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6:145, 11 2012.

[15] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 5(9):1–10, 09 2010.

[16] Bo Jiang, Jin Tang, Chris Ding, Yihong Gong, and Bin Luo. Graph matching via multiplicative update algorithm. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3187–3195. Curran Associates, Inc., 2017.

[17] Mehran Karimzadeh and Michael M. Hoffman. Virtual chip-seq: predicting transcription factor binding by learning from the transcriptome. *bioRxiv*, 2019.

[18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[19] Alexander Lachmann, Federico M Giorgi, Gonzalo Lopez, and Andrea Califano. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 32(14):2233–2235, 2016.

[20] Taotao Lao, Kimberly Glass, Weiliang Qiu, Francesca Polverino, Kushagra Gupta, Jarrett Morrow, John Dominic Mancini, Linh Vuong, Mark A Perrella, Craig P Hersh, et al. Haploinsufficiency of hedgehog interacting protein causes increased emphysema induced by cigarette smoke through network rewiring. *Genome medicine*, 7(1):12, 2015.

[21] Michael Lawrence, Wolfgang Huber, Hervé Pages, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8), 2013.

[22] Camila M Lopes-Ramos, Marieke L Kuijjer, Shuji Ogino, Charles S Fuchs, Dawn L DeMeo, Kimberly Glass, and John Quackenbush. Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer research*, 78(19):5538–5547, 2018.

[23] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

[24] Haggai Maron and Yaron Lipman. (Probably) concave graph matching. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 408–418. Curran Associates, Inc., 2018.

[25] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML16, page 2664–2672. PMLR, 2016.

[26] Weiliang Qiu, Feng Guo, Kimberly Glass, Guo Cheng Yuan, John Quackenbush, Xiaobo Zhou, and Kelan G Tantisira. Differential connectivity of gene regulatory networks distinguishes corticosteroid response in asthma. *Journal of Allergy and Clinical Immunology*, 141(4):1250–1258, 2018.

[27] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

[28] Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. Understanding tissue-specific gene regulation. *Cell reports*, 21(4):1077–1088, 2017.

[29] Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal transport for structured data with application on graphs. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6275–6284, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[30] Anne-Laure Todeschini, Adrien Georges, and Reiner A Veitia. Transcription factors: specific dna binding and specific gene regulation. *Trends in genetics*, 30(6):211–219, 2014.

[31] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[32] Junchi Yan, Xu-Cheng Yin, Weiyao Lin, Cheng Deng, Hongyuan Zha, and Xiaokang Yang. A short survey of recent advances in graph matching. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, page 167–174, New York, NY, USA, 2016. Association for Computing Machinery.

[33] Ye Yuan and Ziv Bar-Joseph. Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, 116(52):27151–27158, 2019.

[34] Theodora Yung, Frankie Poon, Minggao Liang, Sabrina Coquenlorge, Emily C McGaugh, Chi-chung Hui, Michael D Wilson, M Cristina Nostro, and Tae-Hee Kim. Sufu-and spop-mediated downregulation of hedgehog signaling promotes beta cell differentiation through organ-specific niche signals. *Nature communications*, 10(1):1–17, 2019.

[35] Feng Zhou and Fernando De la Torre. Deformable graph matching. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, page 2922–2929, USA, 2013. IEEE Computer Society.

[36] Feng Zhou and Fernando De la Torre. Factorized graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1774–1789, 2016.