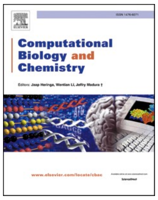
Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Inferring gene regulatory network from single-cell transcriptomic data by integrating multiple prior networks

Yanglan Gan [a], Yongchang Xin [a], Xin Hu [a], Guobing Zou [b], *

[a] *School of Computer Science and Technology, Donghua University, Shanghai, China*
[b] *School of Computer Engineering and Science, Shanghai University, Shanghai, China*

ARTICLE INFO

Keywords:
Gene regulatory network
Prior network
Gene expression
scRNA-seq transcriptome

ABSTRACT

Gene regulatory network models the interactions between transcription factors and target genes. Reconstructing gene regulation network is critically important to understand gene function in a particular cellular context, providing key insights into complex biological systems. We develop a new computational method, named iMPRN, which integrates multiple prior networks to infer regulatory network. Based on the network component analysis model, iMPRN adopts linear regression, graph embedding, and elastic networks to optimize each prior network in line with specific biological context. For each rewired prior networks, iMPRN evaluate the confidence of the regulatory edges in each network based on B scores and finally integrated these optimized networks. We validate the effectiveness of iMPRN by comparing it with four widely-used gene regulatory network reconstruction algorithms on a simulation data set. The results show that iMPRN can infer the gene regulatory network more accurately. Further, on a real scRNA-seq dataset, iMPRN is respectively applied to reconstruct gene regulatory networks for malignant and nonmalignant head and neck tumor cells, demonstrating distinctive differences in their corresponding regulatory networks.

1. Introduction

Cell types or stable states are defined by a particular combination of transcription factors (TFs) and their target genes (Lambert et al., 2018). Such regulatory interactions among TFs and their target genes are usually represented as a gene regulatory network (GRN), where nodes are TFs and their target genes, and edges represent the regulatory relationships (Fiers et al., 2018). In different cell types, some gene regulatory interactions may be very conserved and ubiquitous and many may only occur in certain tissues (Marbach et al., 2016; Iacono et al., 2019). Therefore, establishing accurate and comprehensive GRNs that express the regulatory interactions is essential to understand gene function in complex biological systems, providing key insights into gene-disease association (Fazilaty et al., 2019).

A plethora of methods have been developed to reconstruct gene regulatory networks over the past few years (Chen and Mar, 2018; Pratapa et al., 2020). While each method has its own unique characteristics in terms of underlying algorithm, one of the most distinctive differences between these GRNs methods are what types of data and information they are based on (Geurts et al., 2018). Existing network inference approaches are usually either in the form of steady-state expression data or time series expression data (Castro et al., 2019). Accordingly, these methods can be divided into two categories. Methods in the first category directly infer gene regulatory networks from input gene expression data. For example, GENIE3 uses a tree-based method to reconstruct GRNs between target genes and other genes (Irrthum et al., 2010). PIDC obtains the regulatory relationship between genes based on information theory (Chan et al., 2017). PPCOR calculates partial correlation coefficient and semi-partial correlation coefficient between all gene pairs and construct a GRN (Kim, 2015). SCENIC is a recent single-cell method for identifying stable cell states and network activity based on the estimated GRN model (Aibar et al., 2017). The second type of method is to use time series data of genes to construct gene regulatory networks in addition to expression data. LEAP defines a fixed-size time window on pseudo time-ordered data and calculates the Pearson correlation between genes to obtain the regulatory relationship between genes (Specht and Li, 2017). For given time-stamped single-cell transcription data, SINCERITIES uses the Kolmogorov-Smirnov(KS) statistic to calculate the temporal change in the expression of each gene through the distance of the marginal distribution between two consecutive time points (Papili Gao et al., 2018). SCODE is a method developed to reconstruct a GRN for single-cell transcription data. Specifically, the

* Corresponding author.

https://doi.org/10.1016/j.compbiolchem.2021.107512
Received 2 May 2021; Accepted 12 May 2021
Available online 19 May 2021

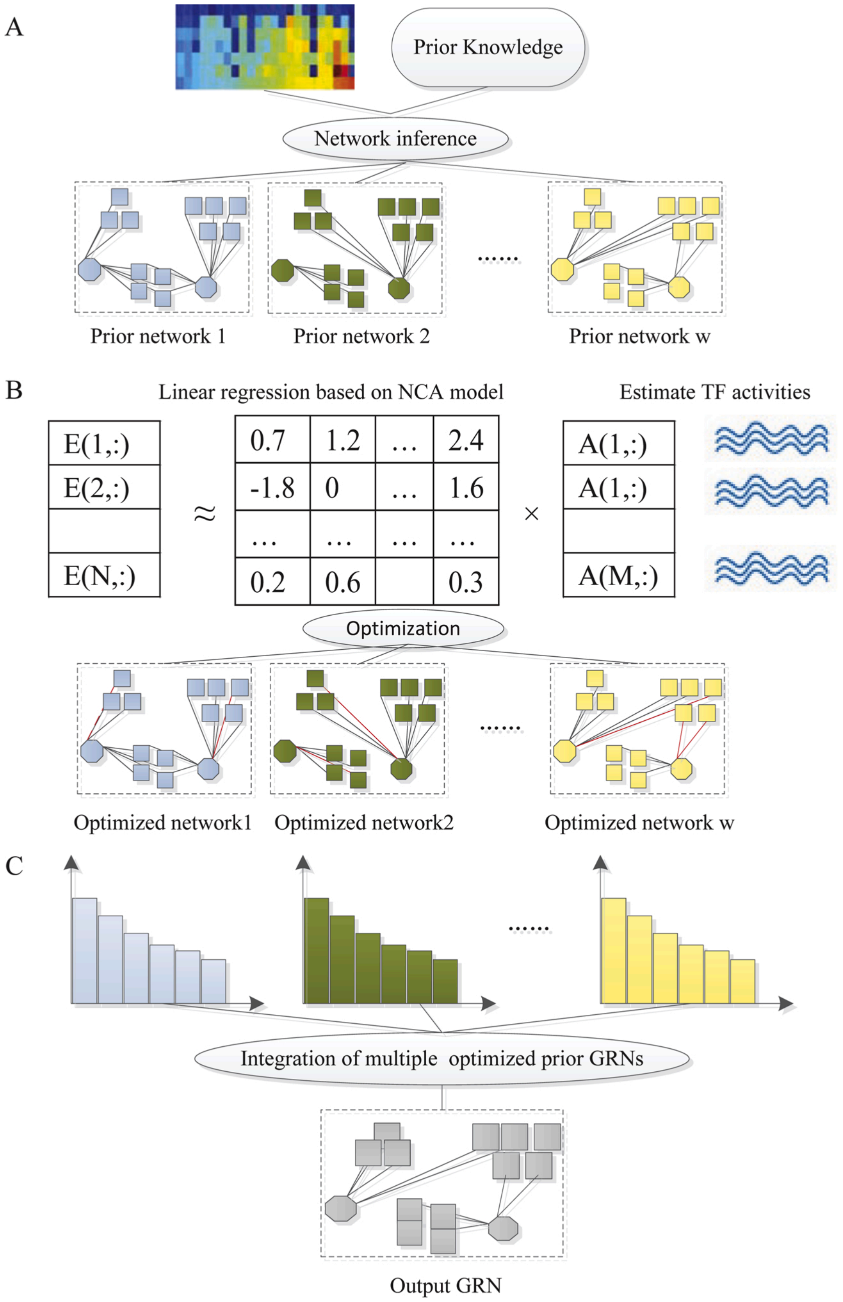


Fig. 1. The schematic flowchart of the proposed iMPRN. (A) Inferring an initial prior gene regulatory network with transcription factors and target genes represented as nodes and regulatory links as edges. (B) Estimating the activity of transcript factors and optimized the prior network. (C) Reevaluating regulatory links of each network and integrating the optimized prior GRNs to form the final high confidence GRN.

expression dynamics of transcription factors are described using linear ODEs (Matsumoto et al., 2017). SCNS takes single-cell gene expression data acquired over a period of time as in put and calculates logic rules (Boolean formulas) to simulate the progress and transformation of genes from the initial cell state to the later cell state (Woodhouse et al., 2018). BiXGBoost is a bidirectional-based method by considering both candidate regulatory genes and target genes for a specific gene (Zheng et al., 2019).

While these network inference methods have made significant progress, their performance on real datasets remains far from optimal. Due to the complex relationship between the expression of TF and its regulatory activity, it is still a major challenge to accurately inference gene regulatory network (Todorov et al., 2019). Previous analysis has shown that it would be helpful to further rely on the TF activity inferred from the data (Qiu et al., 2020). Network component analysis (NCA) has been proved to be a successful approach to infer such regulatory activities (Liao et al., 2003). As NCA requires prior knowledge of the GRN to infer TF activities, research effort has focus on integrating prior knowledge from different conditions to boost network inference (Arrieta-Ortiz et al., 2015; Miraldi et al., 2019). Methods such as Inferelator and Netrex both apply the NCA to generate context-specific GRNs improve the accuracy of network inference (Arrieta-Ortiz et al., 2015; Wang et al., 2018). However, reconstructing GRNs based on only a prior network may make the result dependent on the prior network.

Here we introduce a new computational method, named iMPRN, to infer regulatory network by integrating multiple prior networks. The proposed iMPRN first calculates the activities of transcript factors by NCA model based on the initial prior network, then adjust the edges of the prior network according to the activity of the regulatory genes to ensure that the generated network consistent with the biological context. For each the optimized network, the regulatory edges are sorted according to the calculated confidence scores. Then, iMPRN integrates these optimized GRNs to obtain the final gene regulatory network. In the model of iMPRN, we optimized and integrated multiple prior networks to decrease the impact of noise of scRNA-seq data and the dependence of the initial prior network. To evaluate the performance of iMPRN, we compare it with four widely-used methods on a simulation dataset of the hematopoietic stem cell differentiation process (HSC). The experimental results show that iMPRN outperforms state-of-the-art methods in terms of accuracy and scalability. Further, to investigate the differences in the regulatory mechanisms between malignant and non-malignant tumor cells, we apply iMPRN to single-cell RNA-seq data of head and neck tumor, identify the key genes of the two inferred GRNs, and observe distinctive differences in the gene functions of malignant and non-malignant tumor cells.

2. Materials and methods

2.1. The iMPRN method

To investigate the underlying mechanisms of gene regulation, we propose a new computational method iMPRN to infer gene regulatory network by integrating multiple prior networks. As reconstructing GRNs based on only a prior network may make the result dependent on the prior network, here we integrate multiple prior networks to decrease the dependence. The method consists of three main steps, as illustrated in Fig. 1. Firstly, we infer initial prior networks using different network inference methods. Secondly, using each prior network as the input to the NCA model to infer TF activities, and apply a Proximal Alternative Linearized Maximization algorithm to optimize the network based on relationships between the inferred TF activities and gene expression. Finally, we construct the final GRN by integrating the optimized prior GRN.

Step 1. Inferring initial prior regulatory networks

For the methods requiring prior knowledge of GRN, the prior network usually come from a related tissues or from the same organism without sufficient data. It may be far from the underlying true regulatory network and import certain bias into the downstream analysis. To reduce the bias and improve robustness, iMPRN infers regulatory network based on multiple prior networks. Given the expression data and prior information, we respectively adopt multiple network inference methods to reconstruct the initial networks. The prior information refers to any data that contains direct TF-target information. One source of such prior knowledge is an ever-growing collection of experimentally validated and manually curated databases of regulatory interactions. In addition, DNA-binding information also can be used to generate priors on mammalian regulatory network structure. According to previous evaluation, four state-of-the-art methods are respectively adopted, including GENIE3, PIDC, PPCOR and GRNBoost2. Specifically, these methods requires one or more parameters to be specified. As a previous method does (Siahpirani and Roy, 2017), we perform parameter estimation for each of these methods separately and provide them with the parameters that resulted in the best AUROC values. These inferred networks are then fed as prior information to the next step.

Step2. Optimizing the prior network based on NCA model and PALM algorithm

Given each prior network, we respectively approach network inference by modeling gene expression as a weighted sum of the activities of transcript factors. The goal is to estimate the TF activities and optimize the network from gene expression data as accurately as possible (Liao et al., 2003). As the previous method NetREX (Wang et al., 2018), the gene expression data X is represented by a linear relationship between regulatory activity A and regulatory relationship network R:

$$X(i,:) = \sum_j R(i,j) \times A(j,:) + \tau(i,:) \tag{1}$$

where $X(i, :)$ denotes the expression data of gene i, $R(i, j)$ denotes the potential regulatory relationship between gene i and TF j, $A(j, :)$ represent the regulatory activity of TF j and τ(i,:) represents noise.

Subsequently, estimation of the TF activities and optimization of the network structure of the prior GRN can be consider as finding the optimal linear model with several constraints. The constraints may delete or add the edges of the prior GRN to make the GRN topology related to biological context. The optimization problem after adding constraints is represented as:

$$\begin{gathered} min\frac{1}{2} \parallel E - \mathrm{SA}\|_F^2 + \lambda(\| R_0\|_0 - \parallel R_0 \odot R\|_0 + \parallel \overline{R_0} \odot R\|_0) \\ +\mathrm{ktr}(R^T\mathrm{LR}) + \eta \parallel R_0\|_0 + \xi \parallel R\|_F^2 + \mu \parallel A\|_F^2 \\ s.t. \parallel R\|_\infty \le a, \parallel A\|_\infty \le b. \end{gathered} \tag{2}$$

where λ, κ, η, and μ are parameters that control the importance of the relevant parts. The part controlled by λ constrains the number of edge added or deleted, where $\overline{R_0}$ is the complement adjacency matrix of the prior GRN φ_0, i.e $\overline{R_0} + R_0 = 1_{N\times M}$. $\|X\parallel_0$ is the l_0 norm, which counts the number of non-zero items in X. $\odot$ is the Hadamard product between matrices. Here,$\|R_0 \parallel_0 - \parallel R_0 \odot R \parallel_0$ represents the number of regulatory edges removed compared to φ_0. $\parallel \overline{R_0} \odot R\|_0$represents the number of regulatory edges added compared to φ_0. Therefor, λ can control the number of edges in the network structure. The larger the λ, the fewer edges can be added or deleted.

The part controlled by κ (the figure embedded part) encourages that if gene i and gene j are correlated with each other, they are more easily regulated by the same TF gene k. Where *tr*() is the rank of the matrix. η in Eq. (2) encourages sparsity in the final network structure. In addition, the part controlled by the parameter ξ uses the Frobenius norm to encourage all TFs to have non-zero values in R. In that case, $\eta \parallel R_0\|_0 + \xi \parallel R\|_F^2$ is similar to a l_1 elastic network (Zou and Hastie, 2005), here is called l_0 elastic network. According to the proof of NetREX, the l_0 elastic network may encourage TFs with similar activities to regulate the same set of genes. Finally, the part controlled by the variable μ controls the smoothness of the activity matrix A, which makes each element in A within the limits of {-b, b}.

To solve the convergence of the optimization, the proposed iMPRN relies on a Proximal Alternative Linearized Maximization (PALM) algorithm (Bolte et al., 2014). PALM can resolve the optimization problem of linear regression formulation, which is formulated as:

$$min : H(S,A) = F(R,A) + \phi(R) + \psi(A), \quad R \in \gamma, \quad A \in \Omega \tag{3}$$

where γ and Ω are the constraint sets of R and A and the PALM algorithm applies a technique called proximal forward-backward scheme to R and A.

Then, we converts the optimization problem Eq. (2) into the PALM algorithm framework introduced in Eq. (3) as below:

$$F(R,A) := \frac{1}{2} \parallel X - \mathrm{RA}\|_F^2 + \mathrm{ktr}(R^T\mathrm{LR}) \tag{4}$$

$$\phi(R) := \lambda(\| R_0\|_0 - \parallel R_0 \odot R\|_0 + \parallel \overline{R_0} \odot R \parallel + \xi \parallel R\|_F^2) \tag{5}$$

$$\psi(A) := \mu \parallel A\|_F^2 \tag{6}$$

The constraint sets γ and Ω are $\gamma = \{R| \parallel R \parallel_\infty \le a\}$ and $\Omega = \{A| \parallel A \parallel_\infty \le b\}$.

Based on the PALM algorithm, the optimization problem represent by Eq. (2) is transformed into the new form as Eqs. (4)–(6). Next, we iterate the GRN regulatory relationship matrix R and the TFs activity

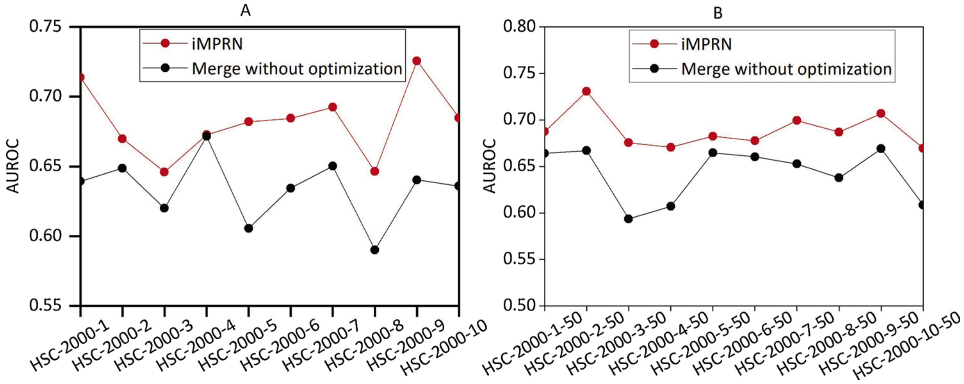


Fig. 2. Performance comparisons between iMPRN and iMPRN without optimization process. The performance was evaluated by AUROC on synthetic HSC datasets. (A) The HSC dataset, (B) The HSC dataset with 50% dropout rate.

matrix A according to Eqs. (4)–(6) until convergence, and the final R and A obtained are optimized GRN and TF activities output based on each prior network.

Step3. Integrating multiple optimized prior networks

After obtaining the individual optimized prior networks based on NCA model and PALM algorithm, they are further integrated to improve the inference performance (Fig. 1C). Aggregates are expected to include either the most significant edges from each individual networks and the most frequent ones. To select the most significant edges, we rank the edges in each optimized GRNs by confidence scores B, which measure their influence on the overall performance of the linear regression model. The confidence score B is computed by the following equation:

$$B(i,j) = 1 - \frac{\| X(i,:) - \sum_{k\neq j} R(i,k)A(k,:)\|_F^2}{\| E(i,:) - R(i,:)A\|_F^2} \tag{7}$$

In order to make the results more robust and avoid overfitting, the gene expression data is re-sampled in an alternative ways and the above process is repeated several times. We compute the average confidence score of each edge that have appeared in the networks and ranking the edges based on the average scores. Finally, for each optimized prior network, we can obtain a ranking of these regulatory edges based on the bootstrap strategy.

Next, we integrate these optimized prior regulatory networks into a final GRN. We seek to select the most significant and the most frequent edges across these multiple optimized networks. As the previous study (Marbach et al., 2010), we calculate the average score of the confidence scores of all the edges that appear in these multiple optimized regulatory networks. Finally, we sort the average score of the regulatory edges and select the top m edges as the output of the final gene regulatory network, where m is an given threshold. For the case without any pre-information, we apply different parameters and obtains the final ranking of the edges by reaching agreement.

2.2. Datasets and data preprocessing

For real biological scRNA-seq data, it is usually difficult to obtain the real edge tags of the gene regulatory network. In order to validate the effectiveness of iMPRN and compare it with existing methods, we first use two simulation datasets generated according to the Boolean model as the input expression data of regulatory network (Pratapa et al., 2020). The advantage of using the Boolean model is that it can be used as a real regulatory model to evaluate the performance of GRN (Giacomantonio and Goodhill, 2010). The data can simulate the interplay of gene regulation in various development processes. Here, we use two simulation datasets, including the hematopoietic stem cell differentiation process (HSC) and Gonadal Sex Determination (GSD). These two datasets respectively includes 10 data sets composed of 2000 cells, all of which are simulated by the Boolean model. In order to test the robustness of our method to zero-inflated data, common in transcripts with low abundance, we further simulate dropout datasets. Zero measurements appear to be a combination of technical errors and genuine lack of expression due to stochasticity or biological state (Kharchenko et al., 2014). We simulated dropout events at a certain rate. Here, expression values in the lowest 50% for each gene had a 50% probability of being recorded as 0 (Chan et al., 2017). For each dataset in HSC and GSD, we divide it into two parts, one is kept as the original dataset and the other part is to add a 50% dropout rate to the dataset to test the performance of iMPRN in the case of data noise.

To investigate the differences in the regulatory mechanisms between malignant and non-malignant tumor cells, we then use the single-cell gene expression data of human head and neck cancer cells (GSE103332) as the input data for reconstructing gene regulatory networks (Puram et al., 2017). The GSE103322 data was downloaded from the data repository NCBI Gene Expression Omnibus. Specifically, the dataset GSE103322 contains 5902 cells (2215 malignant tumor cells and 3363 non-malignant tumor cells) from human head and neck tumor. We divide the gene expression data of head and neck tumor cells into two datasets, malignant cells and non-malignant cells. In order to focus on highly variable genes, we sort these genes according to the variance of gene expression level from high to low. We take the data of genes with the top 5% variance as the input to iMPRN. Based on the gene expression of non-malignant and malignant tumor cells, iMPRN reconstructs two gene regulatory networks respectively. Subsequently, key genes in the inferred regulatory networks are extracted and analyzed.

2.3. Performance metrics

We evaluate the regulatory network inferred by each algorithm using a widely used evaluation criteria AUROC. By comparing the inferred regulatory networks with the true network used to simulate data, the numbers of correctly and incorrectly assigned edges can be computed as the threshold for edge inclusion is varied. Accordingly, AUROC is calculated from the area under the ROC curve, which is a plot of the false-positive rate (FPR) on the x axis versus the true-positive rate (TPR) on the y axis. These metrics are calculated as:

$$\begin{aligned} \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{FPR} &= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \end{aligned} \tag{8}$$

where TP and FP indicate the numbers of true and false positives, and TN and FN are true and false negatives.

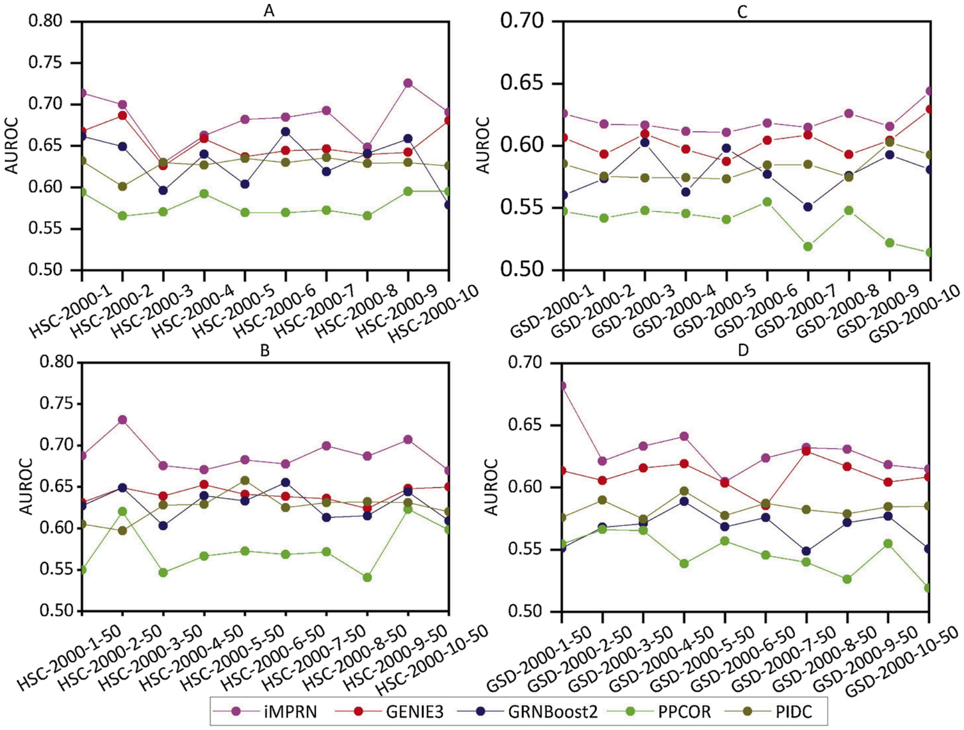


Fig. 3. Performance comparisons among iMPRN, GENE3, PIDC, PPCOR and GRNBoost2 on two synthetic datasets. (A) The HSC dataset, (B) The HSC dataset with 50% dropout rate, (C) The GSD dataset, (D) The GSD dataset with 50% dropout rate.

3. Results

3.1. Performance evaluation on Synthetic datasets

As iMPRN requires prior knowledge of multiple prior networks, here we respectively adopt four network inference method, including GENIE3 (Irrthum et al., 2010), PIDC (Chan et al., 2017), PPCOR (Kim, 2015) and GRNBoost2 (Moerman et al., 2019), to infer the prior networks. These algorithms are currently widely used gene regulatory network inference algorithms. Specifically, GENIE3 is an algorithm based on random forests to predict the interaction between target genes and other genes. PIDC obtains the regulatory relationship between genes based on information theory. PPCOR calculates the partial correlation coefficient and semi-partial correlation coefficient between all gene pairs, and constructs a regulatory network. GRNBoost2 is an algorithm for inferring the network by using stochastic gradient Boosting Machine regression and early-stopping regularization. Based on each initial prior networks, we utilize NCA model and PALM algorithm to simultaneously estimate the TF activities and optimize the structure of regulatory network, and integrate these optimized network into a final GRN. In the pipeline of iMPRN, the optimization of the prior network based on NCA model and PALM algorithm is an important step. To evaluate the importance of the optimization process, we compare the performance of iMPRN and that without optimization step. As shown in Fig. 2, compared with the results of iMPRN, the accuracy of the integrated networks without optimization decreases about 10% in most cases, which demonstrates that the optimizations step can indeed improve the performance. It is effective for the following integration step.

To evaluate the effectiveness of the iMPRN, we compare it with the four initially used algorithms. We apply the GRN inference algorithms to simulated datasets from curated models, which is created by previous study to evaluate different network inference methods (Pratapa et al.,

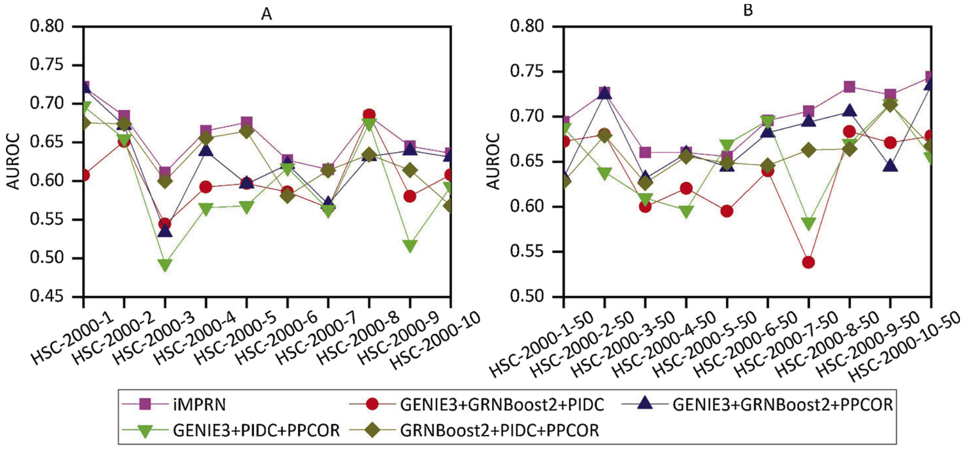


Fig. 4. Performance comparisons among different combinations of GENE3, PIDC, PPCOR, GRNBoost2 on the synthetic HSC datasets. (A) The HSC dataset, (B) The HSC dataset with 50% dropout rate.

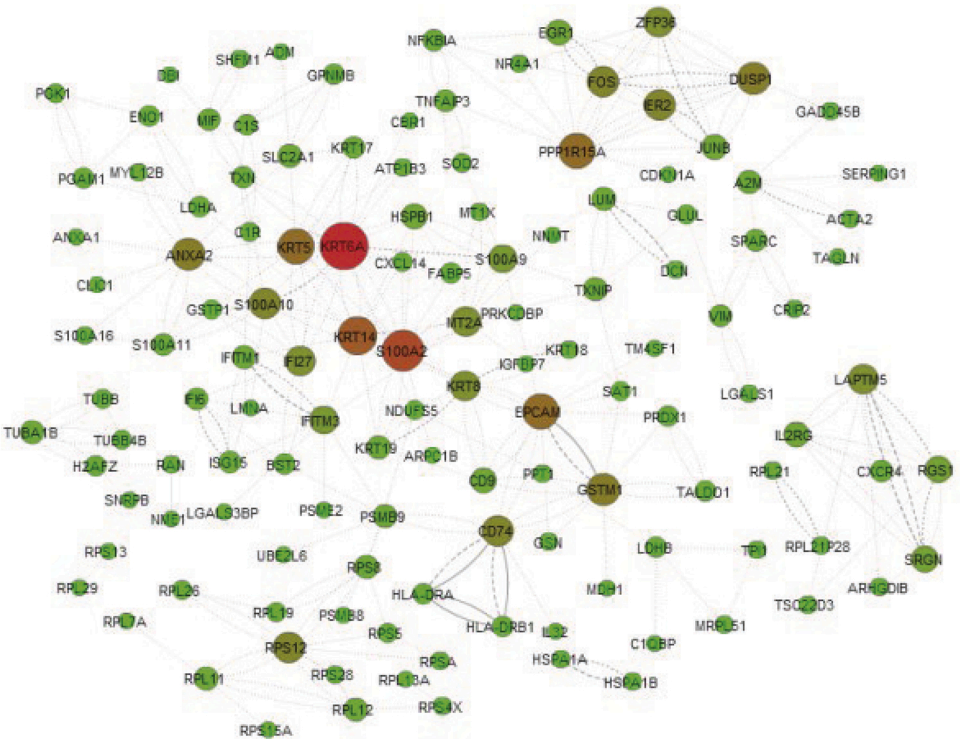


Fig. 5. The significant gene regulatory network reconstructed by iMPRN for malignant head and neck tumor cells.

2020). The simulated HSC dataset and GSD datasets respectively includes 10 data sets composed of 2000 cells, all of which are simulated by the Boolean model. For each dataset, we divide them into two parts. One part is the original dataset and the other is to add 50% dropout rate to the dataset, to test the performance of network inference methods in the case of data noise. We set each synthetic network as the ground truth and adopt the following strategy to evaluate the GRNs inferred by each algorithm. We respectively compare the GRNs for these 10 simulated datasets against the ground truth. We plotted ROC curves and measured the areas under these curves across the 10 different datasets of HSC and HSC with 50% dropout rate, GSD and GSD with 50% dropout rate. Fig. 3 plots AUROC values of these network inference algorithm. We observe that iMPRC has a higher AUROC across a majority of HSC, achieving an average AUROC about 0.70 on different dataset HSC, while PPCOR is less than 0.60. The performance of GENIE3 and PIDC are intermediate between iMPRN and PPCOR. We also evaluate the stability of the results

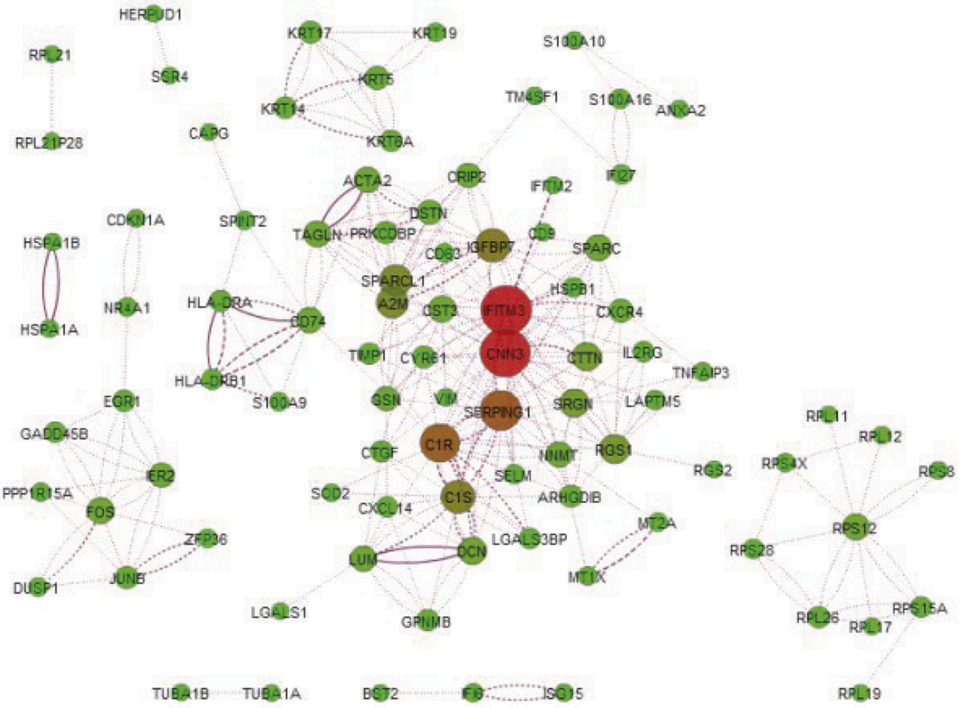


Fig. 6. The significant gene regulatory network reconstructed by iMPRN for non-malignant head and neck tumor cells.

Table 1
Functional enrichment of key genes in the GRNs inferred by iMPRN from malignant and non-malignant human head and neck tumors.

Category	Term name	Malignant	Non-malignant
	GO: Molecular function	*p*-Values (logP)	
GO:MF	Serine-type peptidase activity		5.24
GO:MF	Hydrolase activity, acting on acid phosphorus-nitrogen bonds		5.2103
GO:MF	Serine hydrolase activity		5.2103
GO:MF	Structural constituent of cytoskeleton	4.7807	
GO:MF	Protein-glutamine gamma-glutamyltransferase activity	4.6608	
GO:MF	Transferase activity, transferring amino-acyl groups	4.5704	
GO:MF	Cadherin binding involved in cell-cell adhesion	4.4344	
	GO: Biological process	*p* Values	
GO: BP	Membrane assembly	10	
GO: BP	Membrane biogenesis	10	
GO: BP	Regulation of complement activation, lectin pathway		10
GO: BP	Regulation of humoral immune response		10
GO: BP	Protein activation cascade		5.8827
GO: BP	Hemidesmosome assembly	4.9325	
GO: BP	Humoral immune response		4.8685
GO: BP	Cornification	4.8629	
GO: BP	Platelet degranulation		4.6891
GO: BP	Complement activation, classical pathway		4.5913
GO: BP	Blood coagulation, intrinsic pathway		4.568
GO: BP	Peptidyl-serine dephosphorylation	4.4741	
GO: BP	Humoral immune response mediated by circulating immunoglobulin		4.4659
GO: BP	Regulation of immune effector process		4.3587
GO: BP	Intermediate filament organization	4.2083	
GO: BP	Blood coagulation, fibrin clot formation		4.2083
	GO: Cellular component	*p* Values	
GO: CC	Blood microparticle		10
GO: CC	Platelet alpha granule lumen		5.5638
GO: CC	Keratin filament	5.0784	
GO: CC	Collagen-containing extracellular matrix		4.496
GO: CC	Extracellular matrix		4.1018
GO: CC	Intermediate filament	4.0215	

of each algorithm as we vary the samples. Among these methods, iMPRN and GENIE3 exhibit more stable performance. In addition, it is worthy noting that the AUROC of iMPRN on noisy HSC-Dropout50 data are as good as on the HSC dataset. For the simulated GSD datasets, we also observed the similar results. These results show that the proposed iMPRN is robust to noises and variations in sampling, which may attribute to integration of multiple prior networks.

Furthermore, to test the robustness of the proposed iMPRN, we respectively take one prior GRN out of the model and integrate the remaining prior networks. Then we compare the accuracy of these different integrated networks. The results is illustrated in Fig. 4. From the figure, we observe that each combination result in a partly inconsistent networks, which lead to different accuracy. In most cases, the integration of GENIE3+GRNBoost2+PPCOR achieves a higher AUROC than the other three combinations. Overall, iMPRN had the highest median AUROC scores across a majority of the networks, indicating that the integration of four prior networks can gain a higher robustness.

3.2. *Applying to real scRNA-seq tumor data sets*

Discovering cancer-related biological pathway is a critical task of cancer research. Reconstructing the gene regulatory network can promote the understanding of the regulatory mechanisms of cancer. To investigate the differences in the underlying regulatory relationship of malignant and non-malignant tumor cells, we respectively apply iMPRN to infer gene regulatory networks from the scRNA-seq data of malignant and non-malignant head and neck tumor (Argiris et al., 2008; Puram et al., 2017) (GSE103322). For visualization and subsequent analysis of the regulatory network, we take the output of the two regulatory networks of iMPRN and respectively rank the top 300 edges from each inferred gene regulatory networks.

The malignant tumor GRN and non-malignant tumor GRN are shown in Figs. 5 and 6 respectively. In these GRNs, the more edges that a gene is connected, the greater the degree is. The genes that have high degree will have larger size and darker color. According to the degree, we rank these genes and take the top 10 genes in a network as the key genes. The key genes in the malignant tumor GRN are: KRT6A, S100A2, KRT14, KRT5, PPP1R15A, EPCAM, ANXA2, RPS12 and DUSP1. The key genes in the non-malignant tumor GRN are: FITM3, CNN3, SERPING1, C1R, C1S, SPARCL1, A2M, IGFBP7, RPS12 and SRGN. These identified key genes are highly related to tumors. Most of these identified genes are closely related with a number of cancers, which may be a useful marker for cancers. Specifically, down-regulation of ANXA2 and SPARCL1 occurs in patients with head and neck squamous cell carcinoma (Choi et al., 2016). The gene S100A2 gene is significantly overexpressed in head and neck cancer (Imazawa et al., 2005). Only the gene RPS12 is shared in the two GRNs. From this perspective, there is a big difference between the GRNs of the malignant and non-malignant head and neck tumor.

Further, to investigate the functions of these key genes related to malignant and non-malignant tumor cells, we use TOPPCLUSTER (Kaimal et al., 2010) to perform gene function enrichment analysis respectively. The Go annotations of these key genes are divided into three aspects, including molecular function, cellular component and biological process, as listed in Table 1. In terms of molecular function, the key genes in the GRNs of non-malignant tumor are related to “serine-type peptidase activity” (A2M, C1R, R1S, SERPING1). Correspondingly, the key genes of the malignant cell are related to “transferase activity” (KRT14, KRT5, KRT6A), “cell adhesion mediator activity” (ANXA2, EPCAM) and “structural molecule activity” (KRT14, KRT5, KRT6A, RPS12). In terms of biological process, the key genes of the non-malignant tumor cells are related to “negative regulation of complement activation”, “negative regulation of humoral immune response and blood coagulation” (A2M SERPING1), “protein activation cascade” and “humoral immune response” (A2M, C1R, C1S, SERPING1). The key genes of the malignant tumor cells are enriched in “membrane assembly and membrane biogenesis” (ANXA2, KRT14, KRT5, KRT6A), “hemidesmosome assembly, cornification and cornified envelop assembly” (KRT14, KRT5, KRT6A).

4. Discussion

As gene regulatory relationship usually related with the biological context, the context-specific regulatory network inference is an important task to understand the regulatory mechanisms. Despite the advances in single-cell sequencing technology and genomics, it is still impractical to infer gene regulatory networks for each organism, tissue, cell and condition by accumulating a large number of measurements of specific conditions. We need a method that can make use of prior networks. We propose a gene regulatory network inference algorithm iMPRN based on integrated multiple prior networks. Starting with multiple prior networks, iMPRN uses gene expression data to interactively optimize new regulatory networks. This algorithm optimizes gene regulatory networks based on biological topology by adding and deleting edges. iMPRN adopt linear regression, graph embedding modules and elastic networks to make the network closer to the underlying true network. Meanwhile, we utilize the PALM algorithm to ensure the convergence of optimization process. To make the GRN robust against of the sampling error and the bias of individual prior network, we integrate the information of multiple prior networks. In that case, the robustness and accuracy of the final network are further improved. The comparison conducted on the simulated HSC dataset demonstrates that iMPRN performs well in the aspect of accuracy and robustness. Based on the real scRNA-seq data of human head and neck cancer, iMPRN reconstructs gene regulatory networks for malignant and non-malignant tumor cells. We further analyze the gene function enrichment of the key genes in the two inferred regulatory networks, further deepen the understanding the regulatory mechanisms of different tumor subtypes.

Funding

This work were sponsored in part by the National Natural Science Foundation of China (61772128,61772367), National Key Research and Development Program of China (2016YFC0901704) and Shanghai Natural Science Foundation (17ZR1400200,18ZR1414400).

Conflict of interest

The authors declare no conflict of interest.

References

Aibar, S., González-Blas, C.B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., et al., 2017. Scenic: single-cell regulatory network inference and clustering. Nat. Methods 14 (11), 1083–1086.

Argiris, A., Karamouzis, M.V., Raben, D., Ferris, R.L., 2008. Head and neck cancer. Lancet 371 (9625), 1695–1709.

Arrieta-Ortiz, M.L., Hafemeister, C., Bate, A.R., Chu, T., Greenfield, A., Shuster, B., Barry, S.N., Gallitto, M., Liu, B., Kacmarczyk, T., et al., 2015. An experimentally supported model of the bacillus subtilis global transcriptional regulatory network. Mol. Syst. Biol. 11 (11), 839.

Bolte, J., Sabach, S., Teboulle, M., 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. 146 (1–2), 459–494.

Castro, D.M., De Veaux, N.R., Miraldi, E.R., Bonneau, R., 2019. Multi-study inference of regulatory networks for more accurate models of gene regulation. PLOS Comput. Biol. 15 (1), e1006591.

Chan, T.E., Stumpf, M.P., Babtie, A.C., 2017. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5 (3), 251–267.

Chen, S., Mar, J.C., 2018. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinform. 19 (1), 1–21.

Choi, C.H., Chung, J.-Y., Chung, E.J., Sears, J.D., Lee, J.-W., Bae, D.-S., Hewitt, S.M., 2016. Prognostic significance of annexin a2 and annexin a4 expression in patients with cervical cancer. BMC Cancer 16 (1), 448.

Fazilaty, H., Rago, L., Youssef, K.K., Ocaña, O.H., Garcia-Asencio, F., Arcas, A., Galceran, J., Nieto, M.A., 2019. A gene regulatory network to control EMT programs in development and disease. Nat. Commun. 10 (1), 1–16.

Fiers, M.W., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., Aerts, S., 2018. Mapping gene regulatory networks from single-cell omics data. Brief. Funct. Genomics 17 (4), 246–254.

Geurts, P., et al., 2018. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. Sci. Rep. 8 (1), 1–12.

Giacomantonio, C.E., Goodhill, G.J., 2010. A boolean model of the gene regulatory network underlying mammalian cortical area development. PLoS Comput. Biol. 6 (9), e1000936.

Iacono, G., Massoni-Badosa, R., Heyn, H., 2019. Single-cell transcriptomics unveils gene regulatory network plasticity. Genome Biol. 20 (1), 110.

Imazawa, M., Hibi, K., Fujitake, S.-I., Kodera, Y., ITO, K., Akiyama, S., Nakao, A., 2005. S100a2 overexpression is frequently observed in esophageal squamous cell carcinoma. Anticancer Res. 25 (2B), 1247–1250.

Irrthum, A., Wehenkel, L., Geurts, P., et al., 2010. Inferring regulatory networks from expression data using tree-based methods. PLoS ONE 5 (9), e12776.

Kaimal, V., Bardes, E.E., Tabar, S.C., Jegga, A.G., Aronow, B.J., 2010. Toppcluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. Nucleic Acids Res. 38 (suppl_2), W96–W102.

Kharchenko, P.V., Silberstein, L., Scadden, D.T., 2014. Bayesian approach to single-cell differential expression analysis. Nat. Methods 11 (7), 740–742.

Kim, S., 2015. ppcor: an r package for a fast calculation to semi-partial correlation coefficients. Commun. Stat. Appl. Methods 22 (6), 665.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. The human transcription factors. Cell 172 (4), 650–665.

Liao, J.C., Boscolo, R., Yang, Y.-L., Tran, L.M., Sabatti, C., Roychowdhury, V.P., 2003. Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. USA 100 (26), 15522–15527.

Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G., 2010. Revealing strengths and weaknesses of methods for gene network inference. Proc. Natl. Acad. Sci. USA 107 (14), 6286–6291.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., Bergmann, S., 2016. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat. Methods 13 (4), 366.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S., Ko, S.B., Gouda, N., Hayashi, T., Nikaido, I., 2017. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. Bioinformatics 33 (15), 2314–2321.

Miraldi, E.R., Pokrovskii, M., Watters, A., Castro, D.M., De Veaux, N., Hall, J.A., Lee, J.-Y., Ciofani, M., Madar, A., Carriero, N., et al., 2019. Leveraging chromatin accessibility for transcriptional regulatory network inference in t helper 17 cells. Genome Res. 29 (3), 449–463.

Moerman, T., Aibar Santos, S., Bravo González-Blas, C., Simm, J., Moreau, Y., Aerts, J., Aerts, S., 2019. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics 35 (12), 2159–2161.

Papili Gao, N., Ud-Dean, S.M., Gandrillon, O., Gunawan, R., 2018. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics 34 (2), 258–266.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., Murali, T., 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods 17 (2), 147–154.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al., 2017. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 171 (7), 1611–1624.

Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., McFaline-Figueroa, J. L., Saunders, L., Trapnell, C., Kannan, S., 2020. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. Cell Syst.

Siahpirani, A.F., Roy, S., 2017. A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Res. 45 (4), e21.

Specht, A.T., Li, J., 2017. Leap: constructing gene co-expression networks for single-cell rna-sequencing data using pseudotime ordering. Bioinformatics 33 (5), 764–766.

Todorov, H., Cannoodt, R., Saelens, W., Saeys, Y., 2019. Network inference from single-cell transcriptomic data. Gene Regulatory Networks. Springer, pp. 235–249.

Wang, Y., Cho, D.-Y., Lee, H., Fear, J., Oliver, B., Przytycka, T.M., 2018. Reprogramming of regulatory network using expression uncovers sex-specific gene regulation in drosophila. Nat. Commun. 9 (1), 1–10.

Woodhouse, S., Piterman, N., Wintersteiger, C.M., Göttgens, B., Fisher, J., 2018. Scns: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. BMC Syst. Biol. 12 (1), 1–7.

Zheng, R., Li, M., Chen, X., Wu, F.-X., Pan, Y., Wang, J., 2019. Bixgboost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. Bioinformatics 35 (11), 1893–1900.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67 (2), 301–320.