Published in final edited form as:

ACM BCB. 2018 ; 2018: 1–10. doi:10.1145/3233547.3233551.

Target Gene Prediction of Transcription Factor Using a New Neighborhood-regularized Tri-factorization One-class Collaborative Filtering Algorithm

Hansaim Lim and

PhD program in Biochemistry, Graduate Center of the City University of New York NY 10016 United States

Lei Xie[§]

Department of Computer Science, Hunter College and Graduate Center, the City University of New York NY 10065 United States

Abstract

Identifying the target genes of transcription factors (TFs) is one of the key factors to understand transcriptional regulation. However, our understanding of genome-wide TF targeting profile is limited due to the cost of large scale experiments and intrinsic complexity. Thus, computational prediction methods are useful to predict the unobserved associations. Here, we developed a new one-class collaborative filtering algorithm tREMAP that is based on regularized, weighted nonnegative matrix tri-factorization. The algorithm predicts unobserved target genes for TFs using known gene-TF associations and protein-protein interaction network. Our benchmark study shows that tREMAP significantly outperforms its counterpart REMAP, a bi-factorization-based algorithm, for transcription factor target gene prediction in all four performance metrics AUC, MAP, MPR, and HLU. When evaluated by independent data sets, the prediction accuracy is 37.8% on the top 495 predicted associations, an enrichment factor of 4.19 compared with the random guess. Furthermore, many of the predicted novel associations by tREMAP are supported by evidence from literature. Although we only use canonical TF-target gene interaction data in this study, tREMAP can be directly applied to tissue-specific data sets. tREMAP provides a framework to integrate multiple omics data for the further improvement of TF target gene prediction. Thus, tREMAP is a potentially useful tool in studying gene regulatory networks. The benchmark data set and the source code of tREMAP are freely available at <https://github.com/hansaimlim/REMAP/tree/master/TriFacREMAP>.

Keywords

Gene regulatory network; Recommender system; Latent feature; TF-target interaction; tREMAP

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

[§]Corresponding author: lxie@isab.org.

1 INTRODUCTION

Transcription factors (TFs) regulate gene expression via complex interactions with the target genes, and the regulations are crucial for cellular organizations and development. TFs can activate or deactivate the target genes by binding to the recognition motifs in the promoter or enhancer regions of DNA, called cis-regulatory elements. TFs can interact with each other or recruit other protein components to form a protein complex, which is then recognized by RNA polymerase II (Pol II) to start transcription [1]. Such complex regulations explain the relative complexity of higher metazoans compared to lower organisms, such as unicellular eukaryotes or prokaryotes. In fact, the number of distinct genes itself cannot explain the complexity of organisms. For example, the unicellular *Caenorhabditis elegans* has approximately 20,000 genes, whereas *Drosophila* (fruit flies) has approximately 14,000 [2], [3]. It is known that the human genome contains only twice as many genes as *Drosophila*, and the difference is mainly from the duplication of the same gene rather than new ones [4]. Thus, the incredibly high complexity of humans cannot be understood without knowing the fact that human genome contains approximately one TF per every ten genes [5]. The complicated gene regulation by TFs seems to play an important role in development. In *Drosophila*, for example, deletion of one TF gene (*Antennapedia*) is known to cause a serious phenotypic defect – legs are on the head where antennae should be [6]. Therefore, understanding the associations between TFs and target genes is an important research topic in the biological and biomedical sciences.

Recent advancement of sequencing and molecular biology technology has led to laboratory techniques to identify gene-TF associations on a large scale, and the experimental data have been utilized for computational studies to integrate results from different experiments [7]. Chromatin immunoprecipitation (ChIP)-based methods include ChIP-chip [8], ChIP-seq [9], and ChIP-PET [10]. These techniques estimate the TF binding sites on DNA by crosslinking the TFs to DNA, immunoprecipitating the TF-DNA complex (with antibody), purifying and unlinking TFs from DNA and sequencing the DNA segments [7]. The sequences from ChIP methods are enriched around the binding sites for the TFs. Therefore, the target genes for TFs can be identified by mapping the sequence read peaks to the genome. However, the data from ChIP experiments are noisy and incomplete, containing sequence reads that are false positives and missing information about indirect TF-gene associations. Several studies have focused on identifying the true associations from the ChIP data by statistically comparing the sequence peaks to the background signal [11], [12], [13]. DamID is an alternative to ChIP techniques to identify TF-DNA interactions [14]. In DamID experiments, the adenosine in GATC motif near the TF binding sites are methylated, which is then amplified and detected to identify TF-DNA interactions. ChIP Enrichment Analysis (ChEA) [15] is a freely available tool that combines TF-DNA associations manually curated and automatically collected from 115 publications for the ChIP-X experiments, which are the three ChIP techniques and DamID. ChEA takes a set of genes (whose expression levels are significantly changed) and finds the potential TFs that are likely to interact with most of the genes [15]. ChEA represents a reliable but incomplete resource for known TF-target gene associations; thus, it can be used as a benchmark for algorithm development.

Although the laboratory techniques mentioned above are essential for studying TF-DNA associations, they are not complete. As mentioned above, sequencing data from experiments contain noisy reads that are not necessarily indicating the TF-DNA interactions. In addition, it is well known that the quality of the antibody used in ChIP protocols are crucial for successful experiments [16]. The antibody specificity may be insufficient, or it could block successive interactions between TFs, making it difficult to observe indirect interactions. DamID is largely limited by its resolution as the GATC motifs are required, although it does not require the use of antibodies and therefore has advantages over ChIP protocols [17]. Thus, the currently known TF-gene associations are incomplete due to the limitations of experimental techniques. Computational tools to infer missing TF-gene associations are needed to gain comprehensive understanding of the gene regulations.

Collaborative filtering methods are a group of computational algorithms that are widely used in many areas to infer unobserved associations based on the observed ones with or without additional information [18]. The early generations of collaborative filtering methods are based on probabilistic models and aimed for business concerns, such as recommending products for users in [Amazon.com](https://www.amazon.com) and Barnes and Noble [19], [20]. First proposed by Paatero and Tapper in 1994, nonnegative matrix factorization (MF) [21] has been a popular choice for recommendation problems, especially after the development of fast multiplicative update rules by Lee and Seung [22], [23]. One of the most successful collaborative filtering applications is the popular Netflix challenge, where the user-video preferences are predicted using the past activity of the users [24]. The early collaborative filtering methods heavily rely on the availability of the information about past activity, and it is difficult to make predictions for users without a history of their choices. To overcome the drawback, later generation collaborative filtering methods attempt to utilize additional information, including user-user or item-item similarities [18]. Recently, Yao et al. developed a collaborative filtering algorithm, wiZAN-dual, that utilizes both user-user and item-item similarity information as well as regularization and imputation parameters to improve prediction accuracy [25]. FASCINATE is an extension of wiZAN-dual on a multilayered network [26]. REMAP is an application of wiZAN-dual for biomedical problems [27]. REMAP predicts off-targets of drugs based on the drug-drug similarity and target-target similarity as well as the information about the known targets. In the comprehensive benchmark studies, REMAP outperforms other state-of-the-art methods. Thus, we will only use REMAP as a baseline for the performance evaluation.

As shown in REMAP, biomedical and biochemical association predictions can be modeled as collaborative filtering problems by replacing users with drugs and items with targets. Similarly, the unobserved TF-DNA associations can be predicted using REMAP. However, the drawback of MF-based collaborative filtering is that the factorized low-rank matrices for both users and items must have the same rank. That is, the user-side and item-side latent features must be in the same length, which is unrealistic, particularly if the number of users and items are very different. Moreover, the relationship between the user and the item is modeled by the inner product between two latent features. The inner product could be too simple to capture complex nonlinear relationships between two biological entities. In this study, we present tri-factorization REMAP (tREMAP), an extension of REMAP, that allows us to set different feature sizes for user and items as well as increase the power of modeling

complex relationships between them. We apply tREMAP to the target gene prediction of TF, in which the latent features of TFs and genes are set into different ranks. In the benchmark studies, tREMAP achieves better prediction accuracy for TF-gene association prediction, compared with REMAP. Many of our predicted associations are supported by evidence from the literature. Further development for tissue-specific gene-TF association prediction will significantly improve our understanding in transcriptional regulation.

2 RELATED WORKS

This section is a review of the existing methods for target gene identification tools and relevant databases, followed by methods mathematically similar to ours but applied to different biological problems. TF-related studies reviewed in this section attempted to prioritize the TF-DNA binding peaks to collect the putative TF-gene associations from ChIP-X experiments and the databases for the collected TF-gene associations. To the best of our knowledge, there are few TF-gene prediction tools that take known TF-gene associations as input. We found no such tool that takes the similarities between different genes or TFs. Our method utilizes matrix tri-factorization-based (MTF) collaborative filtering. Their relevant studies were discussed in the introduction of this paper and the last paragraph of this section.

Target identification from profile (TIP) is a probabilistic model that ranks target genes for TFs based on the relative binding signal strength from ChIP experiments, with an assumption that the binding signal is normally distributed [28]. Identifying target genes (iTAR) is an online server, which is designed to overcome the limitation from the normality assumption in TIP by applying Gaussian mixture model for p-value estimation [29]. Covariance based extraction of regulatory targets using multiple time series (CERMT) predicts TF target genes under an assumption that the true target genes for TFs will show similar response pattern to the TFs [30]. Targetfinder is a tool to predict target genes based on the assumption that the genes with similar expression profiles are likely to be regulated by the same TFs [31]. These methods either take ChIP experimental data as input or utilize gene expression data to compare the input genes. A recent study combining these ideas predicts functional TF-gene associations by correlating ChIP data and gene expression profiles [32].

TRANSFAC, a database initially developed in 1980s to list TF-gene interactions from experimental data, has been managed and updated to adopt new data across different organisms as well as tissue-specific regulations [33], [34]. In addition to the information about TFs, their binding sites, and target genes, TRANSFAC database now contains information about the control of gene expressions, the source cell line for TFs, and binding sites for different experimental conditions, if available [35]. JASPAR is another TF-gene database initially developed in 2004 to provide matrix-based TF binding sites from published experimental results [36]. JASPAR was recently updated to include multiple species in six taxonomic groups [37]. The encyclopedia of DNA elements (ENCODE) project, initiated in 2004, aimed to identify all functional elements in the human genome, which includes TF-gene associations [38]. ChEA provides a large collection of TF-gene association data manually curated and computationally extracted from over 100 publications

for ChIP-X experiments [15]. TRRUST is a more recently developed database for human TF-gene associations from text-mining a massive amount of literature abstracts [39]. TRRUST was updated to version 2 in 2017 to contain TF-gene associations in mice as well as more associations for humans[40]. TRANSFAC, JASPAR, ENCODE project and ChEA databases are listed in Harmonizome, an integrated knowledgebase about genes and proteins, developed to facilitate access to and learning from a large amount of biomedical data [41]. Human transcriptional regulation interactions database (HTRIdb) is claimed to be a freely available database containing experimentally verified human TF-gene associations [42].

As reviewed in the introduction, MF-based models have been applied to infer unknown associations such as unobserved drug-target binding. SymNMF is an MF-based method to integrate and infer missing similarity information between drugs and targets from multiple sources [43]. MTF differs from MF in that the input matrix is factorized into three smaller matrices (e.g. matrices F , G , and S in Table 1), instead of two (e.g. matrices F and G where $r = s$). Unlike aforementioned MF or MTF-based ranking methods, which heuristically optimizes the feature sizes (e.g. r and s in Table 1), MTF-based supervised clustering fixes the feature sizes and regularizes the network by prior knowledge. Hwang et al. developed an MTF-based clustering method (R-NMTF) for disease phenotypes and genes regularized by phenotype similarity and protein-protein interaction data [44]. Park et al. developed NTriPath to cluster cancer types and genes regularized by protein-protein interaction data [45]. While the output clusters may be used for certain ranking tasks, these methods require prior knowledge in the number of clusters and correct cluster labels in addition to the inputs for M(T)F- based ranking methods. Moreover, tREMAP incorporates sample weights and imputation into the optimization, making it better handle noisy and sparse data.

3 METHODS

3.1 Prediction method description

In this section, we first present a mathematical formulation of the one-class collaborative filtering problem. The optimization function for our prediction method, tri-factorization REMAP (tREMAP) in Equation 1 with the symbols described in Table 1. Then, we explain how tREMAP differs from REMAP [27], which is a single-ranked version of tREMAP. We also present the update rules for tREMAP, based on the multiplicative update rule by Lee and Seung [22].

$$J = \sum_{(u,i)} W_{(u,i)} \left(R_{(u,i)} + P_{(u,i)} - (FSG^T)_{(u,i)} \right)^2 + \lambda_r (\|F\|^2 + \|S\|^2 + \|G\|^2) + \lambda_F tr \left(F^T (D_M - M) F \right) + \lambda_G tr \left(G^T (D_N - N) G \right) \quad (1)$$

The problem tREMAP solves is to find the nonnegative low-rank matrices F , S , and G that minimizes the optimization function in Equation 1. The optimization function above consists of four terms. Although the formula is slightly different from that for REMAP, most ideas in the function are the same. The shared ideas are explained in the following paragraph.

Tri-factorization REMAP (tREMAP) is an extension of REMAP. REMAP [27] was applied to predict off-target drug-gene associations based on the wiZAN-dual algorithm [25]. REMAP and tREMAP share several ideas. They take the known user-item (drug-target in REMAP application) associations with user-user similarity scores and item-item similarity scores. The inputs are therefore three matrices: user-item association, user-user similarity, and item-item similarity matrices. The core MF algorithm tries to find the low-rank matrices containing the feature vector representations of users and items, such that the inner product of the matrices reconstructs the known association matrix. tREMAP and REMAP also commonly take a penalty weight, an imputation value, a regularization parameter, and importance weights for user-user and item-item similarity information as user-defined parameters. The penalty weight indicates the reliability of the known associations, and the imputation value indicates the probability of unknown associations to be positive. They can be either obtained from a priori knowledge, such as the false positive rate of high-throughput experiments or tuned as hyperparameters. The two importance weights control how much the corresponding similarity scores affect the optimization. In both tREMAP and REMAP is the homophily effect (i.e. similar users prefer similar items) is an important idea. The similarity scores, which can be measured by external methods (e.g. chemical structural similarity for different drugs), are used to reflect the homophily effect by updating the low-rank matrices so that the feature vectors for two similar users or items are close in their Euclidean distance (last two terms in Eq. 1).

The key difference between tREMAP and REMAP is that tREMAP finds three low-rank matrices to approximate the known association matrix, while REMAP and other traditional MF methods find only two. The optimization function for REMAP can be obtained by removing the matrix S in the Equation 1. Without the matrix S , however, one can easily see that the matrix inner product FG^T must be in the same dimension as the known association matrix R . Thus, the matrices F and G must have the same rank, meaning the feature size for both users and items are the same. The single-rank constraint is undesirable unless the actual feature sizes are coincidentally identical. The matrix S makes it possible to set the rank of F different from that of G . By introducing the matrix S into the traditional MF methods, better predictive performances are expected due to more flexible choices for feature sizes.

The optimization algorithm for tREMAP is based on the multiplicative update rule [22], similar to the algorithm for REMAP [27]. As in other MF problems, the optimization problem in Equation 1 is not convex due to the coupling of F , S , and G . Therefore, the multiplicative update rule finds a fixed-point solution for a local optimum of the problem with the nonnegativity constraint. The update rules for the three low-rank matrices are the following.

$$F_{(u,r)} \leftarrow F_{(u,r)} \sqrt{\frac{\left[(1-wp)RGS^T + wp\mathbf{1}_{m \times n}GS^T + \lambda_F MF \right]_{(u,r)}}{\left[(1-w)\hat{R}_\Theta GS^T + wF(SG^T GS^T) + \lambda_r F + \lambda_F D_M F \right]_{(u,r)}}} \quad (2)$$

$$G_{(i,s)} \leftarrow G_{(i,s)} \sqrt{\frac{[(1-wp)R^TFS + wp\mathbf{1}_{m \times n}FS + \lambda_G NG]_{(i,s)}}{[(1-w)\hat{R}_\Theta FS + wG(S^T F^T FS) + \lambda_r G + \lambda_G D_N G]_{(i,s)}}} \quad (3)$$

$$S_{(r,s)} \leftarrow S_{(r,s)} \sqrt{\frac{[(1-wp)F^T R G + wp(F^T \mathbf{1}_{m \times n} G)]_{(r,s)}}{[(1-w)F^T \hat{R}_\Theta G + wF^T(FSG^T)G + \lambda_r S]_{(r,s)}}} \quad (4)$$

The predicted score matrix for known associations, $\hat{R}_\Theta$, is defined as follows.

$$\hat{R}_{\Theta(u,i)} = \begin{cases} (FSG^T)_{(u,i)} & \text{if } (u,i) \in \Theta \\ 0 & \text{otherwise} \end{cases}$$

Note that we use a global scalar weight w and a global scalar imputation p , instead of position-specific weight matrix W and imputation matrix P . The update rules above are derived by considering the partial derivatives with regard to each low-rank matrix, while considering the other two low-rank matrices constant. Therefore, we update the low-rank matrices one at a time, while not changing the other two. The update process under the multiplicative update rules can be described as a gradient descent method with specially designed learning rates [46]. The derivation and proof of the update rules with the justification of using scalar weight and imputation values are in the Appendix.

Once the updates are complete, $\hat{R}$, the prediction score matrix for all TF-gene associations can be calculated by the inner product of the three low-rank matrices. The prediction score matrix for unknown associations, $\hat{R}_{\Theta^c}$, can be obtained by subtracting $\hat{R}_\Theta$ from $\hat{R}$, which contains prediction scores for both known and unknown associations.

$$\hat{R}_{\Theta^c} = \hat{R} - \hat{R}_\Theta, \text{ where } \hat{R} = FSG^T \quad (5)$$

3.2 Dataset description and performance evaluation

The TF-gene association data of our choice in this study is from ChEA, which contains manually curated as well as computationally extracted associations from more than 100 publications for ChIP-X experiments [15]. The ChEA dataset contains 386,776 TF-gene associations for 21,585 genes and 199 TFs for human. The gene-gene and TF-TF similarity matrices are calculated by assuming two interacting proteins are related to each other. The protein-protein interaction data is from the STRING database, which contains experimentally known and computationally predicted protein-protein interactions (PPIs)

with reliability scores [47]. The similarity score between the i^{th} and j^{th} genes (or TFs) is the protein-protein interaction reliability score divided by the maximum available score (1,000). If multiple reliability scores exist for a pair of proteins, they are averaged. This makes all similarity scores in between 0 and 1, standing for minimum and maximum similarity, respectively. Sequence-based protein-protein similarity scores can be used as it was done in the REMAP application [27]. Two proteins will have a high similarity score if their BLAST [48] alignment returns a high score. As a result, 9,207,162 and 12,775 PPI-based nonzero scores are obtained for gene-gene and TF-TF similarity, respectively.

To compare tREMAP with REMAP, we evaluated the performances of the two methods for the ChEA dataset described above. We performed 10-fold or 5-fold cross validation to measure four different performance metrics: area under receiver operating characteristic curve (AUC), mean average precision (MAP), half-life utility (HLU), and mean percentage ranking (MPR). AUC is one of the most widely used performance measurements that measures how quickly an algorithm achieves a high true positive rate while keeping low false positive rates. HLU measures the likelihood that a user accepts recommendation if the likelihood exponentially decreases with the ranking of recommended items [49]. MAP measures the average precision for all users at different true positive rates [50]. MPR is the average percentile rank of positive associations in the test samples [51]. The higher AUC, MAP, HLU, and the lower MPR, the better performance. We compared the performance with and without the similarity score matrices derived from protein-protein interactions.

3 RESULTS

Our benchmark tests under different conditions (e.g. different parameters and with/without similarity information) show tREMAP outperforms REMAP under all tested conditions (Table 2 and Table 3). Table 2 shows that regardless of the similarity matrices used, tREMAP performs significantly better than REMAP in all four metrics. Table 3 shows that the rank parameters affect the performances of both algorithms, and that tREMAP outperforms REMAP under any tested hyperparameters. Due to the number of parameters for both algorithms, it is impractical to compare the two algorithms with all possible combinations. Thus, we tested a limited number of combinations, evaluating the usefulness of different similarity measurements, and the effect of an additional low-rank matrix in tREMAP. We use the optimal regularization, reliability weight, imputation score, and maximum iteration parameters from our study with REMAP [27]. The optimal parameters are $w = p = \lambda_r = 0.1$, and $iter_{max} = 100$. In our previous study, alterations on these parameters did not significantly affect the performance, unless w , p , or λ_r is set to zero, or $iter_{max}$ is fewer than 50. Therefore, we set the similarity matrices and the hyperparameters for tREMAP based on Table 2, Table 3, and our previous study. Both of the gene-gene and TF-TF similarity matrices used for our tREMAP application are based on protein-protein interactions from STRING, as described in the method section [47]. The default hyperparameters are $w = p = \lambda_r = 0.1$, $iter_{max} = 100$, $rank_F = 1,000$, $rank_G = 100$, $\lambda_F = 0.01$, and $\lambda_G = 0.7$.

With the choice of parameters and similarity measurements described above, we performed tREMAP on the full ChEA gene-TF dataset. We first obtained the predicted score matrix $\hat{R}$

as described in the method section. To assess the statistical significance of the predicted scores, we randomly selected 100,000 scores in $\hat{R}$. Removing the scores for known associations, we plotted a histogram of the 909,924 predicted scores, which suggested that the predicted scores are not following a simple distribution, such as Gaussian or exponential distribution (Figure 1). Thus, we first removed the prediction scores for gene-TF associations that were already included in ChEA dataset, and we used Epanechnikov kernel to create a distribution that fits the sampled scores as shown in Figure 1. Then, we selected the predicted gene-TF pairs whose cumulative density is above 0.9808 under the kernel distribution. Our prediction and selection method returned 495 gene-TF associations that were not included in ChEA dataset (Additional file 1). We searched for TRANSFAC [35], ENCODE [38], and TRRUST2 [40] databases to evaluate the final prediction accuracy. As a result, 187 of the 495 (37.8%) associations were found in at least one of the three databases. Considering the fact that the chance of a correct prediction is 9.0% based on the known association matrix, we obtain an enrichment factor of 4.19 (37.8% divided by 9.0%) for our prediction accuracy.

Our method also predicted some associations that are strongly supported by published studies. The associations are listed in Table 4. While the association between NOTCH1 and MYC was previously known from studies regarding T-cell acute lymphoblastic leukemia and included in ChEA dataset, NOTCH2-MYC association was not included. Our prediction method suggests that NOTCH2 may also be association with MYC. In 2004, it was suggested that NOTCH2 and MYC are related in terms of cellular proliferation in mouse thymic lymphoma without strong evidence to conclude their association [52]. In 2016, a study concerned with hypoxia-induced signaling pathway showed that NOTCH2-knockdown murine mesenchymal stem cells cannot properly proliferate, which can be reverted by overexpression of MYC [53]. The collaboration of ZMIZ1 and activated NOTCH1 was found to cause T-cell acute lymphoblastic leukemia in mouse models, which was proposed to be a result of the interaction between ZMIZ1 and MYC at downstream [54]. ARID5B gene, whose role in T-cell acute lymphoblastic leukemia has been previously unknown, was found to directly bind MYC enhancer to promote the expression of MYC, which is a required step for the disease [55]. The concept of double protein lymphoma, characterized by the co-expression of MYC and BCL2 or BCL6, has been known to be aggressive [56], although the MYC/BCL6 biomarker is of less prognostic value [57]. Possibly due to rarity of studies involving MYC/BCL6, the association was not included in the ChEA dataset, while the MYC-BCL2 association was included. NDRG1, whose overexpression in tumor cells decreases the proliferation rate [58], is known to be suppressed by MYC in embryonic cells [59]. In a study concerned with genetic linkage in colon cancer cells, upregulation of glycosyltransferase genes, including ST3GAL1 by MYC was observed [60]. It was reported that EFNA5 was upregulated along with other genes in MYC-knockout mice neural stem and precursor cells [61]. The physical interaction between SPI1 and BCL6 was published in 2009. Interestingly, BCL6 acts as a repressor that binds to SPI1 in germinal center B cells [62]. Although the direct association is unknown and thus excluded from ChEA dataset, SOX2 and HES1 (with other genes) have been studied as markers of neural stem cells [63]. A more recent study added evidence that SOX2 and HES1 are at least members of the same regulatory pathway in rat anterior pituitary cells [64]. In the study, it

was also found that SOX2-expressing cells have significantly lower levels of NOTCH2 expression, suggesting a potential repression of NOTCH2 by SOX2 [64]. The direct association between CREM and MEIS1 was not known although they are involved in the myogenesis, the growth of skeletal muscle. A recent study suggests that although CREM and MEIS1 may not directly interact, they seem to regulate the growth process through another transcription factor, NF-Y [65]. A recent ChIP-seq experiment showed that RUNX1 is a target of AR, which is important for AR-dependent transcription and cell growth in androgen-dependent prostate cancer [66]. These studies support our claim that tREMAP can predict unobserved, but positive associations based on the known associations. NOTCH2-MYC, ZMIZ1-MYC, BCL6-MYC, and EFNA5-MYC associations are in the ENCODE database [38], but not in the TRANSFAC [35] or TRRUST2 [40] database. MYC-NDRG1, MYC-ST3GAL1, and MYC-BCL6 associations are found in both ENCODE and TRRUST2 databases. ARID5B-MYC, SOX2-HES1, and CREM-MEIS1 associations are not found in any of the three databases, suggesting that our method can predict novel gene-TF associations from known ones with proper similarity measurements.

3 DISCUSSION

Comparing the condition C and D in Table 2, the improvement using protein-protein interaction information to measure similarities between TFs is significant. However, similar information for gene-gene similarity did not seem to improve the performance (Conditions A and B compared with C in Table 2). We used a larger importance weight for the TF-TF similarity, implying that our assumption that multi-TF complexes regulating one gene is more likely than TF regulating two different target genes interacting with each other. This does not mean that we ignore one TF regulating multiple genes. In fact, it is well known that more than one TF can regulate a gene through multi-TF complexes [6]. For better performance as well as interpretability, other types of gene-gene similarity scores may be used. The similarity may be based on the sequence alignment scores of the regulatory elements of the genes, which assumes that the DNA sequences of the regulatory elements have evolved to efficiently recruit the TFs. Differential gene expression data can also be used to measure similarities between genes. The hypothesis in such a case is that two genes showing similar patterns of expression under the same conditions are likely to be regulated by the same TFs. A combination of the two types of similarity scores may improve the predictions.

Our benchmarks in Table 3 suggest that the improved performance of tREMAP is from the existence of the matrix S . While the main purpose of introducing matrix S is to set different ranks for genes and TFs, it is not clear whether the ranks must be very different. The condition B in Table 3 shows that tREMAP performs better than REMAP even if all rank parameters are set to 100. In practice, the rank parameters are heuristically optimized. On the other hand, the matrix S can be viewed as a hidden layer introduced to REMAP. Thus, the matrix S may have worked similarly to the hidden layers for the popular deep learning methods, characterized by multiple layers of neural networks with activation functions and regularization steps. Increasing the number of low-rank matrices in tREMAP to mimic deep learning may be an interesting future study. A more interesting combination is to integrate neural network techniques with matrix factorization, as shown in a recent study where the

matrix inner product is considered an additional layer to a multilayer neural network [67]. The time complexity due to the introduction of an additional low-rank matrix as well as large number of parameters from multilayer neural network can be overcome by factorizing smaller submatrices and projecting to the original feature space [68]. In addition, the algorithm to optimize the cost function of matrix factorization may be improved. Simultaneous perturbation stochastic approximation is a potential algorithm to improve the performance as well as the speed of optimization since it requires a dramatically low number of evaluations per iteration and randomness to potentially find the global minimum solution [69], [70]. Such work will enable larger scale applications of the association prediction method with improved accuracy and interpretability.

4 CONCLUSIONS

In this study, we develop a tri-factorization-based collaborative filtering algorithm, tREMAP, that allows users to set different low-ranks for users and items. Compared with its single-rank analog, tREMAP showed better performances measured by four different metrics. We apply tREMAP to predict unobserved TF-gene associations using a collection of known associations. Many of the predicted associations by tREMAP are supported by evidence in the literature or listed in existing databases. Therefore, tREMAP is a powerful tool for TF-gene association prediction, and it can be directly applied to tissue-specific tasks to yield further refined predictions.

ACKNOWLEDGMENTS

This work was supported by Grant Number R01LM011986 from the National Library of Medicine (NLM) of the National Institute of Health (NIH), and Grant Number R01GM122845 from the National Institute of General Medical Sciences (NIGMS) of the National Institute of Health (NIH).

A MATHEMATICAL JUSTIFICATION

The mathematical justifications for our algorithm are presented in this Appendix. Equations that appeared in the main manuscript are numbered the same as they first appeared. New equations are numbered with a prefix ‘A.’ In the main method section, we proposed updated rules for the three low-rank matrices that finds a local minimum of the cost function in Eq. 1.

$$J = \sum_{(u,i)} W_{(u,i)} \left(R_{(u,i)} + P_{(u,i)} - (FSG^T)_{(u,i)} \right)^2 + \lambda_r (\|F\|^2 + \|S\|^2 + \|G\|^2) + \lambda_F tr \left(F^T (D_M - M) F \right) + \lambda_G tr \left(G^T (D_N - N) G \right) \quad (1)$$

The cost function above is nonconvex. Thus, we updated one low-rank matrix at a time, while considering the others as constant. When S and G are fixed, the cost function above becomes a simpler form.

$$J_F = \|W \odot ((R + P) - FSG^T)\|^2 + \lambda_r \|F\|^2 + \lambda_F \text{tr}(F^T(D_M - M)F), \quad \text{s.t. } F \geq 0 \quad (\text{A1})$$

The partial derivative of the J_F with regards to F is the following.

$$\frac{1}{2} \nabla_F J = \frac{1}{2} \frac{\partial J_F}{\partial F} = -W \odot W \odot (R + P)GS^T + W \odot W \odot FSG^T GS^T + \lambda_r F + \lambda_F D_M F - \lambda_F MF \quad (\text{A2})$$

Based on the multiplicative update rule proposed by Lee and Seung [22], we obtain the update rule for F as follows.

$$F_{(u,r)} \leftarrow F_{(u,r)} \sqrt{\frac{[W \odot W \odot (R + P)GS^T + \lambda_F MF]_{(u,r)}}{[W \odot W \odot (FSG^T)GS^T + \lambda_r F + \lambda_F D_M F]_{(u,r)}}} \quad (\text{A3})$$

In the method section, we proposed simplified update rules that reduce the computational complexity from the large dimension of the weight and imputation matrices. The simplified update rule for A3 is the Eq. 2.

$$F_{(u,r)} \leftarrow F_{(u,r)} \sqrt{\frac{[(1 - wp)RGS^T + wp\mathbf{1}_{m \times n}GS^T + \lambda_F MF]_{(u,r)}}{[(1 - w)\hat{R}_\Theta GS^T + wF(SG^T GS^T) + \lambda_r F + \lambda_F D_M F]_{(u,r)}}} \quad (2)$$

In the remainder of this section, we first show that the fixed-point solution of A3 satisfies the KKT condition, and that Eq. 2 and A3 are mathematically equivalent. Then, we show that the cost function in A1 decreases monotonically under the update rule in A3.

THEOREM 1. The fixed-point solution of A3 satisfies the KKT condition.

PROOF. The Lagrangian of A1 is the following (Λ is the Lagrange multiplier).

$$L_{J_F} = \|W \odot ((R + P) - FSG^T)\|^2 + \lambda_r \|F\|^2 + \lambda_F \text{tr}((F^T D_M F) - (F^T MF)) - \text{tr}(\Lambda F) \quad (\text{A4})$$

Let $\frac{\partial}{\partial x} L_{J_F} = 0$, we obtain the following.

$$2(-W \odot W \odot (R + P)GS^T + W \odot W \odot FSG^T GS^T + \lambda_r F + \lambda_F D_M F - \lambda_F MF) = \Lambda \quad (A5)$$

From the KKT complementary slackness condition, we obtain the following.

$$\begin{aligned} & [-W \odot W \odot (R + P)GS^T + W \odot W \odot FSG^T GS^T + \lambda_r F + \lambda_F D_M F - \lambda_F MF]_{(u,r)} \cdot F_{(u,r)} = 0 \\ & (A6) \end{aligned}$$

A7 is the fixed-point solution of A3, which satisfies A6.

$$[W \odot W \odot (R + P)GS^T + \lambda_F MF]_{(u,r)} = [W \odot W \odot (FSG^T)GS^T + \lambda_r F + \lambda_F D_M F]_{(u,r)} \quad (A7)$$

Next, we show that Eq. 2 is equivalent to A3. We use $\mathbf{1}^A$, $\mathbf{1}^\Theta$, and $\mathbf{1}^{\Theta^c}$ as the indicator matrices for full, observed, and unobserved data, respectively, so that $\mathbf{1}_{m \times n} = \mathbf{1}^A = \mathbf{1}^\Theta + \mathbf{1}^{\Theta^c}$, and $\mathbf{1}^\Theta = R$. Based on that, the weight and imputation values are for unobserved associations only, the equations below turn the weight matrix W and imputation matrix P into scalar weight w and scalar imputation value p , respectively. Note that the weight matrix W contains the square root of the global weight w on unobserved positions and zero on observed ones.

$$\begin{aligned} W \odot W \odot (R + P)GS^T &= (\mathbf{1}^\Theta \odot R + wp \cdot \mathbf{1}^{\Theta^c})GS^T = (R + wp \cdot \mathbf{1}^A - wp \cdot \mathbf{1}^\Theta)GS^T = (1 - wp) \cdot RGS^T + wp \\ &\cdot \mathbf{1}_{m \times n} GS^T \end{aligned}$$

and

$$\begin{aligned} (W \odot W \odot FSG^T)GS^T &= W \odot W \odot (\hat{R}_\Theta + \hat{R}_{\Theta^c})GS^T = (\hat{R}_\Theta + w \cdot \mathbf{1}^{\Theta^c} \odot \hat{R}_{\Theta^c})GS^T \\ &= (\hat{R}_\Theta + w \cdot \mathbf{1}^A \odot FSG^T - w \cdot \mathbf{1}^\Theta \odot \hat{R}_\Theta)GS^T = (1 - w) \cdot \hat{R}_\Theta GS^T + w \cdot FSG^T GS^T \end{aligned}$$

Substituting the two equations above into A3 proves that Eq. 2 is equivalent to A3.

THEOREM 2. The cost function in A1 decreases monotonically under the update rule in A3.

PROOF. To prove theorem 2, we start from the cost function A1. According to the auxiliary function strategy [71], $H(F, \tilde{F})$ is an auxiliary function of $J(F)$ if it satisfies the following conditions.

$$H(F, F) = J(F), \quad \text{and} \quad H(F, \tilde{F}) \geq J(F) \quad (\text{A8})$$

Defining $F^{(t+1)} = \arg \min_F H(F, F^{(t)})$ proves that $J(F^{(t)})$ monotonically decreases since the following condition is met by the design of the auxiliary function.

$$J(F^{(t)}) = H(F^{(t)}, F^{(t)}) \geq H(F^{(t+1)}, F^{(t)}) \geq J(F^{(t+1)}) \quad (\text{A9})$$

We first find an auxiliary function satisfying the conditions in A8, and then solve for the auxiliary function, which is the global minimum of the auxiliary function.

$$\begin{aligned} H(F, \tilde{F}) = & -2 \sum_{u=1}^m \sum_{k=1}^r [(W \odot W \odot (R+P))GS^T]_{(u,k)} \tilde{F}_{(u,k)} \left(1 + \log \left(\frac{F_{(u,k)}}{\tilde{F}_{(u,k)}} \right) \right) \quad (\text{A10}) \\ & - \sum_{u=1}^m \sum_{v=1}^m \sum_{k=1}^r \lambda_F M_{(u,v)} \tilde{F}_{(v,k)} \tilde{F}_{(u,k)} \left(1 + \log \left(\frac{F_{(v,k)} F_{(u,k)}}{\tilde{F}_{(v,k)} \tilde{F}_{(u,k)}} \right) \right) \\ & + \sum_{u=1}^m \sum_{k=1}^r \lambda_r F_{(u,k)}^2 + \sum_{u=1}^m \sum_{k=1}^r \frac{[(W \odot W \odot \tilde{F}SG^T)GS^T]_{(u,k)} F_{(u,k)}^2}{\tilde{F}_{(u,k)}} \\ & + \sum_{u=1}^m \sum_{k=1}^r \frac{[\lambda_F D_M \tilde{F}]_{(u,k)} F_{(u,k)}^2}{\tilde{F}_{(u,k)}} \end{aligned}$$

It is trivial to show $H(F, F) = J(F)$. To show $H(F, \tilde{F}) \geq J(F)$, we name the five terms in A10 as H1, H2, H3, H4, and H5, respectively. Then, using the inequality $x \geq 1 + \log(x)$, the H1 becomes the following.

$$\begin{aligned} \text{H1} \geq & -2 \sum_{u=1}^m \sum_{k=1}^r [(W \odot W \odot (R+P))GS^T]_{(u,k)} F_{(u,k)} = -2 \text{tr} \quad (\text{A11}) \\ & [(W \odot W \odot (R+P))GS^T F] \end{aligned}$$

$$\text{H2} \geq - \sum_{u=1}^m \sum_{v=1}^m \sum_{k=1}^r \lambda_F M_{(u,v)} F_{(v,k)} F_{(u,k)} = -\lambda_F \text{tr}(F^T M F) \quad (\text{A12})$$

Then, for H3 we get $H3 = \lambda_r \text{tr}(FF^T)$ (A13)

For H4, let $F_{(u,k)} = \tilde{F}_{(u,k)} Q_{(u,k)}$ and we have the following.

$$\begin{aligned}
 H4 &= \sum_{u=1}^m \sum_{i=1}^n \sum_{k=1}^r \sum_{l=1}^r \frac{\tilde{F}_{(u,l)} (SG^T)_{(i,l)} W_{(u,i)}^2 (GS^T)_{(i,k)} F_{(u,k)}^2}{\tilde{F}_{(u,k)}} \quad (A14) \\
 &= \sum_{u=1}^m \sum_{i=1}^n \sum_{k=1}^r \sum_{l=1}^r \tilde{F}_{(u,l)} (SG^T)_{(i,l)} W_{(u,i)}^2 (GS^T)_{(i,k)} \tilde{F}_{(u,k)} Q_{(u,k)}^2 \\
 &= \sum_{u=1}^m \sum_{i=1}^n \sum_{k=1}^r \sum_{l=1}^r \tilde{F}_{(u,l)} (SG^T)_{(i,l)} W_{(u,i)}^2 (GS^T)_{(i,k)} \tilde{F}_{(u,k)} \left(\frac{Q_{(u,k)}^2 + Q_{(u,l)}^2}{2} \right) \\
 &\geq \sum_{u=1}^m \sum_{i=1}^n \sum_{k=1}^r \sum_{l=1}^r \tilde{F}_{(u,l)} (SG^T)_{(i,l)} W_{(u,i)}^2 (GS^T)_{(i,k)} \tilde{F}_{(u,k)} (Q_{(u,k)} + Q_{(u,l)}) \\
 &= \sum_{u=1}^m \sum_{i=1}^n \sum_{k=1}^r \sum_{l=1}^r F_{(u,l)} (SG^T)_{(i,l)} W_{(u,i)}^2 (GS^T)_{(i,k)} F_{(u,k)} \\
 &= \text{tr}[(W \odot W \odot (FSG^T))GS^T F^T]
 \end{aligned}$$

We use the inequality below, where $A_{n \times n}$, $B_{k \times k}$, $S_{n \times k}$, and $S_{n \times k}^*$ are nonnegative, and A and B are symmetric [72].

$$\sum_{i=1}^n \sum_{p=1}^k \frac{(AS^*B)S_{(i,p)}^2}{S_{(i,p)}^*} \geq \text{tr}(S^T ASB) \quad (A15)$$

Thus, $H5 = \sum_{u=1}^m \sum_{k=1}^r \frac{[\lambda_F D_M \tilde{F}]_{(u,k)} F_{(u,k)}^2}{\tilde{F}_{(u,k)}} \geq \lambda_F \text{tr}(F^T D_M F)$

Substituting A11 – A15 into A10 shows that the auxiliary function satisfies the second condition in A8.

The gradient of the auxiliary function is the following.

$$\begin{aligned}
\frac{1}{2} \cdot \frac{\partial H(\mathbf{F}, \tilde{\mathbf{F}})}{\partial \mathbf{F}_{(u,k)}} = & - \frac{[W \odot W \odot (R + P)GS^T]_{(u,k)} \cdot \tilde{\mathbf{F}}_{(u,k)}}{\mathbf{F}_{(u,k)}} - \frac{[\lambda_F M \tilde{\mathbf{F}}]_{(u,k)} \cdot \tilde{\mathbf{F}}_{(u,k)}}{\mathbf{F}_{(u,k)}} \\
& + \frac{\lambda_r \mathbf{F}_{(u,k)} \cdot \tilde{\mathbf{F}}_{(u,k)}}{\tilde{\mathbf{F}}_{(u,k)}} + \frac{[W \odot W \odot \tilde{\mathbf{F}}SG^TGS^T]_{(u,k)} \cdot \mathbf{F}_{(u,k)}}{\tilde{\mathbf{F}}_{(u,k)}} + \frac{[\lambda_F D_M \tilde{\mathbf{F}}]_{(u,k)} \cdot \mathbf{F}_{(u,k)}}{\tilde{\mathbf{F}}_{(u,k)}} = \\
& - \frac{[W \odot W \odot (R + P)GS^T + \lambda_F M \tilde{\mathbf{F}}]_{(u,k)} \cdot \tilde{\mathbf{F}}_{(u,k)}}{\mathbf{F}_{(u,k)}} \\
& + \frac{[W \odot W \odot \tilde{\mathbf{F}}SG^TGS^T + \lambda_r \tilde{\mathbf{F}} + \lambda_F D_M \tilde{\mathbf{F}}]_{(u,k)} \cdot \mathbf{F}_{(u,k)}}{\tilde{\mathbf{F}}_{(u,k)}}
\end{aligned} \tag{A16}$$

The Hessian of $H(\mathbf{F}, \tilde{\mathbf{F}})$ is a diagonal matrix with positive diagonal elements. Thus, we can obtain the global minimum by setting A16 to be zero, which results in the following solution.

$$\mathbf{F}_{(u,k)}^2 = \tilde{\mathbf{F}}_{(u,k)}^2 \cdot \frac{[W \odot W \odot (R + P)GS^T + \lambda_F M \tilde{\mathbf{F}}]_{(u,k)}}{[W \odot W \odot (\tilde{\mathbf{F}}SG^TGS^T + \lambda_r \tilde{\mathbf{F}} + \lambda_F D_M \tilde{\mathbf{F}})]_{(u,k)}} \tag{A17}$$

Setting $\mathbf{F}^{(t+1)} = \mathbf{F}$ and $\mathbf{F}^{(t)} = \tilde{\mathbf{F}}$ proves that the update rule A3 monotonically decreases the cost function. With the equivalence between Eq. 2 and A3, A1 monotonically decreases under the update rule Eq. 2.

The update rule for G can be proved analogously to the proof above. The matrix S -equivalent of the cost function A1 is the following.

$$J(S) = \text{tr}(-2W \odot W \odot (R + P)GS^T F^T) + \text{tr}(W \odot W \odot (FSG^T)GS^T F^T) + \lambda_r \text{tr}(S^T S)$$

Therefore, we choose an auxiliary function for matrix S , which is missing two terms corresponding to H2 and H5 in A10. The auxiliary function and its gradient are the following.

$$\begin{aligned}
H(S, \tilde{S}) = & -2 \cdot \sum_{i=1}^r \sum_{j=1}^s [F^T(W \odot W \odot (R + P))]G]_{(i,j)} \cdot \tilde{S}_{(i,j)} \cdot \left(1 + \log \frac{S_{(i,j)}}{\tilde{S}_{(i,j)}}\right) \\
& + \sum_{i=1}^r \sum_{j=1}^s \lambda_r S_{(i,j)}^2 + \sum_{i=1}^r \sum_{j=1}^s \frac{[F^T(W \odot W \odot (F\tilde{S}G^T))]G]_{(i,j)} \cdot S_{(i,j)}^2}{\tilde{S}_{(i,j)}} \cdot \frac{1}{2} \cdot \frac{\partial H(S, \tilde{S})}{\partial S_{(i,j)}} = \\
& - \frac{[F^T(W \odot W \odot (R + P))]G]_{(i,j)} \cdot \tilde{S}_{(i,j)}}{S_{(i,j)}} + \frac{[\lambda_r \tilde{S}]_{(i,j)} \cdot S_{(i,j)}}{\tilde{S}_{(i,j)}} + \frac{[F^T(W \odot W \odot (F\tilde{S}G^T))]G]_{(i,j)} \cdot S_{(i,j)}}{\tilde{S}_{(i,j)}}
\end{aligned}$$

Setting the gradient to zero, we obtain the global minimum solution.

$$S_{(i,j)}^2 = \tilde{S}_{(i,j)}^2 \cdot \frac{\left[F^T (W \odot W \odot (R + P)) G \right]_{(i,j)}}{\left[F^T (W \odot W \odot (F \tilde{S} G^T)) G + \lambda_r \tilde{S} \right]_{(i,j)}}$$

Combining the theorems in the Appendix, we proved that the proposed update rules satisfy the KKT condition and converge to the solution.

REFERENCES

- [1]. Levine M, and Tjian R, "Transcription regulation and animal diversity," *Nature*, vol. 424, no. 6945, pp. 147–51, 7 10, 2003. [PubMed: 12853946]
- [2]. Ruvkun G, and Hobert O, "The taxonomy of developmental control in *Caenorhabditis elegans*," *Science*, vol. 282, no. 5396, pp. 2033–41, 12 11, 1998. [PubMed: 9851920]
- [3]. Adams MD, Celniker SE, Holt RA et al., "The genome sequence of *Drosophila melanogaster*," *Science*, vol. 287, no. 5461, pp. 2185–95, 3 24, 2000. [PubMed: 10731132]
- [4]. Baltimore D, "Our genome unveiled," *Nature*, vol. 409, no. 6822, pp. 814–816, 2 15, 2001. [PubMed: 11236992]
- [5]. Lander ES, Consortium IHGS, Linton LM et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2 15, 2001. [PubMed: 11237011]
- [6]. Phillips T, and Hoopes L, "Transcription factors and transcriptional control in eukaryotic cells," *Nature Education*, vol. 1, no. 1, pp. 119, 2008.
- [7]. Collas P, "The current state of chromatin immunoprecipitation," *Mol Biotechnol*, vol. 45, no. 1, pp. 87–100, 5, 2010. [PubMed: 20077036]
- [8]. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, and Brown PO, "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, vol. 409, no. 6819, pp. 533–538, 2001. [PubMed: 11206552]
- [9]. Johnson DS, Mortazavi A, Myers RM, and Wold B, "Genome-wide mapping of in vivo protein-DNA interactions," *Science*, vol. 316, no. 5830, pp. 1497–502, 6 8, 2007. [PubMed: 17540862]
- [10]. Wei CL, Wu Q, Vega VB et al., "A global map of p53 transcription-factor binding sites in the human genome," *Cell*, vol. 124, no. 1, pp. 207–19, 1 13, 2006. [PubMed: 16413492]
- [11]. Kharchenko PV, Tolstorukov MY, and Park PJ, "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nat Biotechnol*, vol. 26, no. 12, pp. 1351–9, 12, 2008. [PubMed: 19029915]
- [12]. Nix DA, Courdy SJ, and Boucher KM, "Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks," *BMC Bioinformatics*, vol. 9, pp. 523, 12 5, 2008. [PubMed: 19061503]
- [13]. Tuteja G, White P, Schug J, and Kaestner KH, "Extracting transcription factor targets from ChIP-Seq data," *Nucleic Acids Res*, vol. 37, no. 17, pp. e113, 9, 2009. [PubMed: 19553195]
- [14]. Vogel MJ, Peric-Hupkes D, and van Steensel B, "Detection of in vivo protein-DNA interactions using DamID in mammalian cells," *Nat Protoc*, vol. 2, no. 6, pp. 1467–78, 2007. [PubMed: 17545983]
- [15]. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, and Ma'ayan A, "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments," *Bioinformatics*, vol. 26, no. 19, pp. 2438–44, 10 1, 2010. [PubMed: 20709693]
- [16]. Furey TS, "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions," *Nat Rev Genet*, vol. 13, no. 12, pp. 840–52, 12, 2012. [PubMed: 23090257]

- [17]. McClure CD, and Southall TD, "Getting Down to Specifics: Profiling Gene Expression and Protein-DNA Interactions in a Cell Type-Specific Manner," *Adv Genet*, vol. 91, pp. 103–51, 2015. [PubMed: 26410031]
- [18]. Su X, and Khoshgoftaar TM, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, pp. 4, 2009.
- [19]. Linden G, Smith B, and York J, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.
- [20]. Hofmann T, "Latent semantic models for collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 89–115, 2004.
- [21]. Paatero P, and Tapper U, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.
- [22]. Lee DD, and Seung HS, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, 10 21, 1999. [PubMed: 10548103]
- [23]. Lee J, Sun M, and Lebanon G, "A comparative study of collaborative filtering algorithms," *arXiv preprint arXiv:12053193*, 2012.
- [24]. Bennett J, and Lanning S, "The netflix prize." p. 35.
- [25]. Yao Y, Tong H, Yan G et al., "Dual-regularized one-class collaborative filtering," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. pp. 759–768.
- [26]. Chen C, Tong H, Xie L, Ying L, and He Q, "FASCINATE: Fast Cross-Layer Dependency Inference on Multi-layered Networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016, pp. 765–774.
- [27]. Lim H, Poleksic A, Yao Y et al., "Large-scale off-target identification using fast and accurate dual regularized one-class collaborative filtering and its application to drug repurposing," *PLoS Comput Biol*, vol. 12, no. 10, pp. e1005135, 10, 2016. [PubMed: 27716836]
- [28]. Cheng C, Min R, and Gerstein M, "TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles," *Bioinformatics*, vol. 27, no. 23, pp. 3221–3227, 2011. [PubMed: 22039215]
- [29]. Yang C-C, Andrews EH, Chen M-H et al., "iTAR: a web server for identifying target genes of transcription factors using ChIP-seq or ChIP-chip data," *BMC Genomics*, vol. 17, pp. 632, 2016. [PubMed: 27519564]
- [30]. Redestig H, Weicht D, Selbig J, and Hannah MA, "Transcription factor target prediction using multiple short expression time series from *Arabidopsis thaliana*," *BMC Bioinformatics*, vol. 8, pp. 454, 11 18, 2007. [PubMed: 18021423]
- [31]. Kielbasa SM, Bluthgen N, Fahling M, and Mrowka R, "Targetfinder.org: a resource for systematic discovery of transcription factor target genes," *Nucleic Acids Res*, vol. 38, no. Web Server issue, pp. W233–8, 7, 2010. [PubMed: 20460454]
- [32]. Banks CJ, Joshi A, and Michael T, "Functional transcription factor target discovery via compendia of binding and expression profiles," *Sci Rep*, vol. 6, pp. 20649, 2 9, 2016. [PubMed: 26857150]
- [33]. Wingender E, "Compilation of transcription regulating proteins," *Nucleic acids research*, vol. 16, no. 5 Pt B, pp. 1879, 1988. [PubMed: 3282223]
- [34]. Matys V, Fricke E, Geffers R et al., "TRANSFAC ® : transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003. [PubMed: 12520026]
- [35]. Matys V, Kel-Margoulis OV, Fricke E et al., "TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes," *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D108–D110, 2006. [PubMed: 16381825]
- [36]. Sandelin A, Alkema W, Engström P, Wasserman WW, and Lenhard B, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D91–D94, 2004. [PubMed: 14681366]
- [37]. Mathelier A, Fornes O, Arenillas DJ et al., "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 44, no. D1, pp. D110–D115, 2016. [PubMed: 26531826]

- [38]. Consortium EP, "The ENCODE (ENCyclopedia Of DNA Elements) Project," *Science*, vol. 306, no. 5696, pp. 636–40, 10 22, 2004. [PubMed: 15499007]
- [39]. Han H, Shim H, Shin D et al., "TRRUST: a reference database of human transcriptional regulatory interactions," *Sci Rep*, vol. 5, pp. 11432, 6 12, 2015. [PubMed: 26066708]
- [40]. Han H, Cho JW, Lee S et al., "TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions," *Nucleic Acids Res*, 10 26, 2017.
- [41]. Rouillard AD, Gunderson GW, Fernandez NF et al., "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins," *Database*, vol. 2016, no. 2016, pp. baw100-baw100, 2016. [PubMed: 27374120]
- [42]. Bovolenta LA, Acencio ML, and Lemke N, "HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions," *BMC Genomics*, vol. 13, pp. 405, 8 17, 2012. [PubMed: 22900683]
- [43]. Chen H, and Li J, "A Flexible and Robust Multi-Source Learning Algorithm for Drug Repositioning," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Boston, Massachusetts, USA, 2017, pp. 510–515.
- [44]. Hwang T, Atluri G, Xie M et al., "Co-clustering phenome-genome for phenotype classification and disease gene discovery," *Nucleic Acids Res*, vol. 40, no. 19, pp. e146, 10, 2012. [PubMed: 22735708]
- [45]. Park S, Kim SJ, Yu D et al., "An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types," *Bioinformatics*, vol. 32, no. 11, pp. 1643–51, 6 1, 2016. [PubMed: 26635139]
- [46]. Li L-X, Wu L, Zhang H-S, and Wu F-X, "A fast algorithm for nonnegative matrix factorization and its convergence," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 10, pp. 1855–1863, 2014. [PubMed: 25291738]
- [47]. Szklarczyk D, Franceschini A, Wyder S et al., "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D447–52, 1, 2015. [PubMed: 25352553]
- [48]. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ, "Basic local alignment search tool," *J Mol Biol*, vol. 215, no. 3, pp. 403–10, 10 5, 1990. [PubMed: 2231712]
- [49]. Breese JS, Heckerman D, and Kadie C, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, Madison, Wisconsin, 1998, pp. 43–52.
- [50]. Li Y, Hu J, Zhai C, and Chen Y, "Improving one-class collaborative filtering by incorporating rich user information." pp. 959–968.
- [51]. Hu Y, Koren Y, and Volinsky C, "Collaborative filtering for implicit feedback datasets." pp. 263–272.
- [52]. Lopez-Nieva P, Santos J, and Fernandez-Piqueras J, "Defective expression of Notch1 and Notch2 in connection to alterations of c-Myc and Ikaros in gamma-radiation-induced mouse thymic lymphomas," *Carcinogenesis*, vol. 25, no. 7, pp. 1299–304, 7, 2004. [PubMed: 14976135]
- [53]. Sato Y, Mabuchi Y, Miyamoto K et al., "Notch2 Signaling Regulates the Proliferation of Murine Bone Marrow-Derived Mesenchymal Stem/Stromal Cells via c-Myc Expression," *PLoS One*, vol. 11, no. 11, pp. e0165946, 2016. [PubMed: 27855169]
- [54]. Rakowski LA, Garagiola DD, Li CM et al., "Convergence of the ZMIZ1 and NOTCH1 pathways at C-MYC in acute T lymphoblastic leukemias," *Cancer Res*, vol. 73, no. 2, pp. 930–41, 1 15, 2013. [PubMed: 23161489]
- [55]. Leong WZ, Tan SH, Ngoc PCT et al., "ARID5B Activates the TAL1-Induced Core Regulatory Circuit and the MYC Oncogene in T-Cell Acute Lymphoblastic Leukemia," *Am Soc Hematology*, 2017.
- [56]. Pillai RK, Sathanoori M, Van Oss SB, and Swerdlow SH, "Double-hit B-cell lymphomas with BCL6 and MYC translocations are aggressive, frequently extranodal lymphomas distinct from BCL2 double-hit B-cell lymphomas," *Am J Surg Pathol*, vol. 37, no. 3, pp. 323–32, 3, 2013. [PubMed: 23348205]

- [57]. Ye Q, Xu-Monette ZY, Tzankov A et al., "Prognostic impact of concurrent MYC and BCL6 rearrangements and expression in de novo diffuse large B-cell lymphoma," *Oncotarget*, vol. 7, no. 3, pp. 2401–16, 1 19, 2016. [PubMed: 26573234]
- [58]. Guan RJ, Ford HL, Fu Y, Li Y, Shaw LM, and Pardee AB, "Drg-1 as a differentiation-related, putative metastatic suppressor gene in human colon cancer," *Cancer research*, vol. 60, no. 3, pp. 749–755, 2000. [PubMed: 10676663]
- [59]. Qu X, Zhai Y, Wei H et al., "Characterization and expression of three novel differentiation-related genes belong to the human NDRG gene family," *Mol Cell Biochem*, vol. 229, no. 1–2, pp. 35–44, 1, 2002. [PubMed: 11936845]
- [60]. Sakuma K, Aoki M, and Kannagi R, "Transcription factors c-Myc and CDX2 mediate E-selectin ligand expression in colon cancer cells undergoing EGF/bFGF-induced epithelial-mesenchymal transition," *Proc Natl Acad Sci U S A*, vol. 109, no. 20, pp. 7776–81, 5 15, 2012. [PubMed: 22547830]
- [61]. Martinez-Cerdeno V, Lemen JM, Chan V et al., "N-Myc and GCN5 regulate significantly overlapping transcriptional programs in neural stem cells," *PLoS One*, vol. 7, no. 6, pp. e39456, 2012. [PubMed: 22745758]
- [62]. Wei F, Zaprazna K, Wang J, and Atchison ML, "PU.1 can recruit BCL6 to DNA to repress gene expression in germinal center B cells," *Mol Cell Biol*, vol. 29, no. 17, pp. 4612–22, 9, 2009. [PubMed: 19564417]
- [63]. Takanaga H, Tsuchida-Straeten N, Nishide K, Watanabe A, Aburatani H, and Kondo T, "Gli2 is a novel regulator of sox2 expression in telencephalic neuroepithelial cells," *Stem Cells*, vol. 27, no. 1, pp. 165–74, 1, 2009. [PubMed: 18927476]
- [64]. Batchuluun K, Azuma M, Fujiwara K, Yashiro T, and Kikuchi M, "Notch Signaling and Maintenance of SOX2 Expression in Rat Anterior Pituitary Cells," *Acta Histochem Cytochem*, vol. 50, no. 2, pp. 63–69, 4 27, 2017. [PubMed: 28522881]
- [65]. Grade CVC, Mantovani CS, Fontoura MA, Yusuf F, Brand-Saberi B, and Alvares LE, "CREB, NF-Y and MEIS1 conserved binding sites are essential to balance Myostatin promoter/enhancer activity during early myogenesis," *Mol Biol Rep*, vol. 44, no. 5, pp. 419–427, 10, 2017. [PubMed: 28956216]
- [66]. Takayama K, Suzuki T, Tsutsumi S et al., "RUNX1, an androgen- and EZH2-regulated gene, has differential roles in AR-dependent and -independent prostate cancer," *Oncotarget*, vol. 6, no. 4, pp. 2263–76, 2 10, 2015. [PubMed: 25537508]
- [67]. He X, Liao L, Zhang H, Nie L, Hu X, and Chua T-S, "Neural collaborative filtering," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 173–182.
- [68]. Mackey LW, Talwalkar A, and Jordan MI, "Distributed matrix completion and robust factorization," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 913–960, 2015.
- [69]. Spall JC, "An overview of the simultaneous perturbation method for efficient optimization," *Airport Modeling and Simulation*, pp. 141–154, 1999.
- [70]. Maryak JL, and Chin DC, *Global optimization via SPSA*, Johns Hopkins Univ. Laurel MD Applied Physics Lab, 2002.
- [71]. Lee DD, and Seung HS, "Algorithms for non-negative matrix factorization." *Adv Neural Inf Process Syst*. pp. 556–562.
- [72]. Ding C, Li T, Peng W, and Park H, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 126–135.

Figure 1: Probability density of 909,924 randomly sampled tREMAP prediction scores for gene-TF associations that are excluded from training data. Epanechnikov kernel fits to the sampled scores.

Table 1:

Symbol definitions and descriptions. Matrices are capitalized and italicized, and scalars are in lower-cases.

Symbols	Definition
m, n	Number of unique genes and TFs.
r, s	Feature sizes for genes and transcription factors, respectively. $r < m$, and $s < n$.
w	Scalar reliability weight. $w \in [0, 1]$
p	Scalar imputation score. $p \in [0, 1]$
Θ, Θ^c	Set of observed and unobserved associations.
$R_{(i,j)}$	Element at the i^{th} row and j^{th} column of matrix R .
R	Known association matrix. $R_{(i,j)} = 1$ if $(i,j) \in \Theta$, 0 otherwise. $R \in \mathbb{R}^{m \times n}$
F	Low-rank feature matrix for genes. $F \in \mathbb{R}^{m \times r}$
G	Low-rank feature matrix for TFs. $G \in \mathbb{R}^{n \times s}$
S	Low-rank compressed feature matrix. $S \in \mathbb{R}^{r \times s}$
M	Gene-gene similarity score matrix. It is a symmetric, positive matrix. $M \in \mathbb{R}^{m \times m}$
N	TF similarity score matrix, defined similarly to M . $N \in \mathbb{R}^{n \times n}$
D_M, D_N	Degree matrices for M and N , respectively. D_M and D_N are diagonal, positive matrices.
W	Weight matrix. $W_{(i,j)} = 1$ if $(i,j) \in \Theta$, w otherwise. $W \in \mathbb{R}^{m \times n}$
P	Imputation matrix. $P_{(i,j)} = 0$ if $(i,j) \in \Theta$, p otherwise. $P \in \mathbb{R}^{m \times n}$
$\mathbf{1}_{m \times n}$	Indicator matrix containing 1 at every position. $\mathbf{1}_{m \times n} \in \mathbb{R}^{m \times n}$
λ_r	Regularization parameter. $\lambda_r \in [0, 1]$
λ_F, λ_G	Importance weights for genes and TFs.
$tr(M)$	Trace of matrix M .
$\ M\ $	Frobenius norm of matrix M .

Table 2:

Performance comparison for tREMAP and REMAP with different similarity information (mean and standard deviation for 10-fold cross validation).

$I_{\text{Cond.}}$	Algo.	AUC	HLU	MAP	MPR
A	tREMAP	.740(1.2e-6)	16.6(.002)	.125(3.6e-5)	.259(6.2e-6)
	REMAP	.704(.001)	11.5(.030)	.092(.001)	.310(.001)
B	tREMAP	.740(1.4e-6)	16.6(.004)	.125(4.7e-5)	.259(8.7e-6)
	REMAP	.714(0.001)	12.5(.050)	.095(.001)	.297(.001)
C	tREMAP	.740(1.5e-6)	16.6(.004)	.125(4.1e-5)	.259(8.8e-6)
	REMAP	.715(0.001)	12.5(.058)	.099(.001)	.296(.001)
D	tREMAP	.725(2.2e-6)	13.9(.003)	.102(4.0e-5)	.259(5.2e-6)
	REMAP	.687(.001)	9.4(.087)	.074(.001)	.312(.001)

$I_{\text{Condition A}}$: TF similarity scores are based on sequence similarity only, and gene similarity scores are not used. **Condition B**: TF similarity scores are the average of sequence-based and protein-protein interaction-based scores, and gene similarity scores are based on protein-protein interactions only. **Condition C**: TF similarity scores are based on protein-protein interactions only, and gene similarity scores are not used. **Condition D**: No similarity information used.

Table 3:

Performance comparison for tREMAP and REMAP with different hyperparameters (mean and standard deviation for 10-fold cross validation).

² Cond.	Algo.	AUC	HLU	MAP	MPR
A	tREMAP	.740(1.5e-6)	16.6(.004)	.125(4.1e-5)	.259(8.8e-6)
	REMAP	.715(.001)	12.5(.058)	.099(1e-4)	.296(.001)
B	tREMAP	.740(5.1e-6)	16.6(.010)	.125(.001)	.259(1.1e-6)
	REMAP	.704(0.001)	11.5(.144)	.092(.001)	.310(.001)
C	tREMAP	.725(1.8e-6)	13.9(.014)	.102(2e-4)	.259(3.2e-5)
	REMAP	.688(0.001)	9.8(.093)	.076(7e-4)	.310(.001)
D	tREMAP	.725(2.2e-6)	13.9(.003)	.102(4.0e-5)	.259(5.2e-6)
	REMAP	.687(.001)	9.4(.087)	.074(.001)	.312(.001)

²**Condition A:** Default parameters. **Condition B:** tREMAP ranks=(100,100), REMAP rank=100. **Condition C:** tREMAP ranks=(100,50), REMAP rank=50. **Condition D:** $\lambda_F = \lambda_G = 0$.

Table 4:

Predicted gene-TF associations with cumulative distribution function (CDF) of the Epanechnikov kernel fitted to the tREMAP prediction scores.

TF	Gene	CDF	Database	Reference
MYC	N0TCH2	0.99213	ENCODE	[52], [53]
MYC	ZMIZ1	0.99906	ENCODE	[54]
MYC	ARID5B	0.99928	ENCODE	[55]
MYC	BCL6	0.99909	ENCODE, TRANSFAC	[56], [57]
MYC	NDRG1	0.98982	ENCODE, TRRUST2	[59]
MYC	ST3GAL1	0.9992	ENCODE, TRRUST2	[60]
MYC	EFNA5	0.99607	ENCODE	[61]
SPI1	BCL6	0.99864	ENCODE, TRRUST2	[62]
SOX2	HES1	0.99993	None	[63], [64]
SOX2	N0TCH2	0.99335	None	[64]
CREM	MEIS1	0.99955	None	[65]
AR	RUNX1	0.99837	None	[66]