Active module identification from multilayer weighted gene co-expression networks

Li, Dong; Pan, Zhisong; Hu, Guyu; Anderson, Graham; He, Shan

DOI:
10.1109/TCBB.2020.2970400

Active module identification from multilayer weighted gene co-expression networks: a continuous optimization approach

Dong Li
School of Computer Science
University of Birmingham
Birmingham, UK B15 2TT
Email: donggeat@gmail.com

Guyu Hu
and Zhisong Pan
Army Engineering University of PLA
Nanjing, China, 210007

Graham Anderson
Institute of Immunology
and Immunotherapy
University of Birmingham
Birmingham, UK B15 2TT

Shan He
School of Computer Science
University of Birmingham
Birmingham, UK B15 2TT
Email: s.he@cs.bham.ac.uk

***Abstract*—Searching for active modules, i.e., regions showing striking changes in molecular activity in biological networks is important to reveal regulatory and signaling mechanisms of biological systems. Most existing active modules identification methods are based on protein-protein interaction networks or metabolic networks, which require comprehensive and accurate prior knowledge. On the other hand, weighted gene co-expression networks (WGCNs) are purely constructed from gene expression profiles. However, existing WGCN analysis methods are designed for identifying functional modules but not capable of identifying active modules. There is an urgent need to develop an active module identification algorithm for WGCNs to discover regulatory and signaling mechanism associating with a given cellular response.**

To address this urgent need, we propose a novel algorithm called active modules on the multi-layer weighted (co-expression gene) network, based on a continuous optimization approach (AMOUNTAIN). The algorithm is capable of identifying active modules not only from single-layer WGCNs but also from multilayer WGCNs such as cross-species and dynamic WGCNs. We first validate AMOUNTAIN on a synthetic benchmark dataset. We then apply AMOUNTAIN to WGCNs constructed from Th17 differentiation gene expression datasets of human and mouse, which include a single layer, a cross-species two-layer and a multilayer dynamic WGCNs.

The identified active modules from WGCNs are enriched by known protein-protein interactions, and more importantly, they reveal some interesting and important regulatory and signaling mechanisms of Th17 cell differentiation.

1. Introduction

One of the most important problems in network biology is searching for **active modules**, i.e. connected regions of the molecular interaction network showing striking changes in molecular activity or phenotypic signatures that are associated with a given cellular response [1]. The activities of a network are usually measured by high-throughput omics data, e.g., microarray or NGS gene expression data. By identifying the activated parts of multiple functional modules and their interrelationships, active modules are able to reveal regulatory and signaling mechanisms [2]. In recent years, many active module identification algorithms have been developed [2], [3], [4], [5], [6], [7], [8].

However, most of the existing active module identification algorithms, including our own work [7], [8] can only work with protein-protein interaction (PPI) or metabolic networks. These networks are constructed from prior knowledge databases, which might not be comprehensive and accurate. Moreover, for some non-model species or some new model species such as Daphnia, their PPI or metabolic networks are not available, which limited the application of active module identification algorithms.

In contrast, gene co-expression network is a pure data-driven gene network, which only relies on gene expression profile. Given a gene expression profile, a similarity matrix is calculated, in which each element measures the correlation of a pair of genes, i.e., how similar their expression levels change together. In this paper, we focus on the weighted gene co-expression networks (WGCNs), fully connected graphs where the weights correspond to the correlations between pairs of genes.

However, there are no active module identification algorithms for WGCNs. Most of the existing WGCNs module detection algorithms are based on clustering, i.e., grouping similar genes based on their correlations or edge weights into modules [9], [10]. These identified modules are considered to participate in some biological process [9], and those with significant biological meaning are regarded as functional modules. Since the genes are clustered without considering their activity, unlike active modules, these functional modules cannot reveal the dynamic mechanisms associating with a given cellular response.

We develop the first active module identification algorithm `AMOUNTAIN` for WGCNs. The aim of this algorithm is to identify active modules to reveal not only the dynamic biological processes but also the regulatory and signaling mechanisms underlying a given cellular response. To this end, we propose a new definition of the active module in a WGCN. Based on the definition, we formally formulate the active modules identification problem in single-layer WGCNs and generalize the problem to multilayer WGCNs.

Apart from formulating the active module identification

problem in single and multilayer WGCNs, another main contribution of this paper is the new continuous optimization formulation based on the elastic net penalty. This continuous formulation can be solved more efficiently than the original combinatorial optimization formulation.

We evaluate the proposed framework on both simulated data and real-world data, including multiple species and time-course gene expression datasets. These results indicate that the identified active modules can reveal not only the dynamic biological processes but also the regulatory and signaling mechanisms that underlie a given cellular response.

2. Methods

2.1. Defining active module in WGCNs

To identify active modules from WGCNs, which are essentially weighted and fully connected graphs, we need to define what is an active module. Our core criterion is that it should be significantly different from random subnetworks in two perspectives: 1) From the topological point of view, the nodes in the active module should be densely connected with each other, i.e., significantly co-expressed, which is quantified by the module score based on edge weights. 2) From the regulatory and signaling mechanism point of view, an active module should show a significant change in molecular activity which can be measured by the module score based on the activities, i.e., expression levels of the genes (node scores).

2.2. Single-layer network

2.2.1. Problem definition. Based on the above criteria, for a single-layer network, our active module definition considers 1) the node scores of the genes as the measures their activities under certain conditions; 2) the edge weights which represent the topology or co-expression relationship among those genes.

More specifically, we aim to find an active module or subgraph of size k (otherwise it corresponds to a trivial case containing all top-scored nodes) that has both maximal aggregated node score and maximum aggregated edge weight, which can be formally defined as:

Problem 1. *Given a complete graph $G = (V, E)$, with vertex weight $\mathbf{z}_v \in \mathbb{R}$ for each $v \in V$ and non-negative edge weights $W = [w_{ij}]$ for each edge (i, j), find a subgraph T of size k with large vertex weights $\sum_{i \in T} z_i$ and also edge weights $\sum_{i,j \in T} w_{ij}$.*

A module is represented by a membership vector $\mathbf{x} \in \{0, 1\}^n$, where n is the total number of nodes and $x_i = 1$ means the i-gene belongs to the module. Thus the optimization is naturally expressed as:

$$\begin{aligned} &\max_{\mathbf{x}} \ F(\mathbf{x}) = \mathbf{x}^\top W \mathbf{x} + \lambda \mathbf{z}^\top \mathbf{x} \\ &\text{Subject to} \\ &\quad \sum_{i=1}^{n} x_i = k \\ &\quad x_i \in \{0, 1\}, \qquad\qquad i = 1, ..., n, \end{aligned} \tag{1}$$

where parameter λ controls the trade-off between edges score and nodes score.

2.2.2. Continuous optimization formulation with elastic net penalty. The NP-hardness of equation (1) can be proved by reducing it to the well-known k-clique problem (See supplementary text S1 section 1), which is NP-complete. To solve this NP-hard problem, similar to [6], one might apply linear relaxation and then use integer programming methods. However, the running time is not guaranteed, especially for large-scale networks.

In this paper, we relax the integer constraints of $\mathbf{x}$ to continuous constraints [11], [12] and control the module size by introducing a vector norms of $\mathbf{x}$. Specifically, in solution $\mathbf{x} \in \mathbb{R}^n_+$ when $x_i > 0$ means the i-th node is in the module, it becomes a nonnegative and equality constrained quadratic programming (QP) problem (2), which can be solved by various existing continuous optimization techniques in polynomial time.

$$\begin{aligned} &\max_{\mathbf{x} \in \mathbb{R}^n_+} \ F(\mathbf{x}) = \mathbf{x}^\top W \mathbf{x} + \lambda \mathbf{z}^\top \mathbf{x} \\ &\text{Subject to,} \\ &\quad f(\mathbf{x}) = 1, \end{aligned} \tag{2}$$

where $f(\mathbf{x})$ is the vector norm. The ℓ_p-norm ($p > 0$) of $\mathbf{x}$ is defined as $(\sum_i |x_i|^p)^{1/p}$.

The choice of vector norm affects the structure of the solution (2). For example, the ℓ_0-norm and ℓ_1-norm can produce a sparse solution which corresponds to modules with small size. This is desirable since we aim to identify smaller modules which are easy to verify in the follow-up experiments. Since the optimization of ℓ_0-norm is also an NP-hard combinatorial problem, the ℓ_1-norm has been widely used as an alternative [13]. However, ℓ_1-norm tends to produce too sparse solutions which are again not desired.

Recently, elastic net penalty [14], which is a linear combination of ℓ_1-norm and ℓ_2-norm, i.e., $\alpha\|\mathbf{x}\|_1 + (1-\alpha)\|\mathbf{x}\|_2^2$, has been introduced. In the context of the least square problem with elastic net penalty, $\alpha = 1$ corresponds to lasso [15] and $\alpha = 0$ corresponds to ridge regression. Therefore, the elastic net is considered to enjoy the advantages of both lasso and ridge regression, i.e., the sparsity and accuracy, by tuning the parameter α.

In `AMOUNTAIN`, we use the elastic net penalty [14] to control the sparsity and improve the efficiency of our module

identification algorithm. Therefore, the problem (2) becomes

$$\begin{aligned} &\max_{\mathbf{x}\in\mathbb{R}^n_+} F(\mathbf{x}) = \mathbf{x}^\top W\mathbf{x} + \lambda \mathbf{z}^\top \mathbf{x} \\ &\text{Subject to,} \\ &\qquad f(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\|\mathbf{x}\|_2^2 = 1. \end{aligned} \tag{3}$$

2.2.3. Optimization method. We use a projected gradient method to solve (3) since the objective function is smooth and differentiable and the constraint, i.e., elastic net, is strictly convex. In addition to gradient ascend to find the local maximum, the projected gradient method employs projection operation to project the current candidate solution to the nearest point in the convex feasible region [16], [17]. The projected gradient method is guaranteed to converge to the stationary points of the problem (3) [18]. Specifically, We use the following sequence to approximate the final solution:

$$\mathbf{x}^{(k+1)} = \Pi_C\big(\mathbf{x}^{(k)} + \beta^{(k)}\nabla F(\mathbf{x}^{(k)})\big), \tag{4}$$

where $\beta^{(k)}$ is the step size which can be fixed or tuned to improve the convergence rate [16]. Π_C is the Euclidean projection of a vector $\mathbf{g} = \mathbf{x}^{(k)} + \beta^{(k)}\nabla F(\mathbf{x}^{(k)})$ on the convex set C, and the subproblem is thus defined as:

$$\begin{aligned} &\Pi_C(\mathbf{g}) = \arg\min_{\mathbf{x}\in\mathbb{R}^n_+} \frac{1}{2}\|\mathbf{x}-\mathbf{g}\|_2^2 \\ &\text{Subject to,} \\ &\qquad \alpha\|\mathbf{x}\|_1 + (1-\alpha)\|\mathbf{x}\|_2^2 = 1. \end{aligned} \tag{5}$$

Solving subproblem (5) involves a root finding procedure [17] which can be done in linear time, and the details can be found in section 2 of Supplementary text S1.

The Euclidean projection based optimization for problem (3) is summarized as algorithm 1.

Algorithm 1 Euclidean projections optimization

Input: Network edge weight $W \in \mathbb{R}^{n\times n}$, node score $\mathbf{z} \in \mathbb{R}^n$ and initial solution $\mathbf{x}^{(0)} \in \mathbb{R}^n_+$ which is randomly sampled from the uniform distribution [0,1] and then projected to the feasible region.

Output: Module indicator vector $\mathbf{x}$

1: Update $\mathbf{g}$ in equation (5) by the gradient of $F(\mathbf{x})$ in equation (3).
2: Solve optimal $\mathbf{x}$ in equation (5) by Algorithm 1 in supplementary text S1. Convergence or reach maximal iterations

In order to identify N modules from one network, similar to [19], [20], we run the algorithm 1 N times. Each time the algorithm extracts a module and subsequently deletes the module from background network. The general procedure for identifying N modules from given gene expression profile is summarized in Algorithm 2.

Algorithm 2 Active modules identification of GCN

Input: Gene expression profile $X \in \mathbb{R}^{n\times p}$, number of modules M

Output: M modules

1: **Construction**: Construct a weighted gene expression network G.
2: **Nodes scores**: Perform gene differential analysis to calculate fold-changes or p-values and assign to the genes as node scores. iterations less than M
3: **Optimization**: Find solution $\mathbf{x}$ for (1) using algorithm (1)
4: **Extraction**: Extract nodes in $\mathbf{x}$ and corresponding edges from G and delete them from G afterwards.

2.3. Multilayer network

We start from a simple case where inter-layer interactions only exist between neighborhood layers, then derive a compact form for multilayer networks without interlayer links.

2.3.1. Multilayer network with inter-layer interactions. We first generalize the single layer active module identification problem to two-layer WGCNs. We define a two-layer active module as two modules in two different networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ connected by inter-layer edges. The inter-layer edges were defined by $A = [a]_{ij} \in \mathbb{R}^{n_1\times n_2}$ where n_1 and n_2 are the numbers of nodes in G_1 and G_2. The two-layer WGCN active module identification problem is formally defined as

Problem 2. *Given two complete graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, with vertex weights $\mathbf{z}_{1v} \in R$ for each $v \in V_1$ and $\mathbf{z}_{2v} \in R$ for each $v \in V_2$, and non-negative edge weights $W_1 \in \mathbb{R}^{n_1\times n_1}$ for edges in G_1 and $W_2 \in \mathbb{R}^{n_2\times n_2}$ for edges in G_2. The intensity of inter-layer interactions were measured by $A = [a]_{ij} \in \mathbb{R}^{n_1\times n_2}$. The goal is to find two subgraphs $T_1 \in G_1$ and $T_2 \in G_2$ which both have large vertices weights and edges weights as well as intensive inter-layer interactions with each other.*

We use two variables $\mathbf{x} \in \mathbb{R}^{n_1}$ and $\mathbf{y} \in \mathbb{R}^{n_2}$ to represent the memberships of active modules in two different networks, and $x_i > 0$ means the i-th node in the first network is in the module. Thus the optimization problem can be expressed as an extension to (2),

$$\begin{aligned} &\max_{\mathbf{x}\in\mathbb{R}^{n1}_+,\mathbf{y}\in\mathbb{R}^{n_2}_+} F = \mathbf{x}^\top W_1\mathbf{x} + \lambda_1\mathbf{z}_1^\top\mathbf{x} + \mathbf{y}^\top W_2\mathbf{y} \\ &+\lambda_2\mathbf{z}_2^\top\mathbf{y} + \lambda_3\mathbf{x}^\top A\mathbf{y} \\ &\qquad \text{Subject to} \\ &\qquad\qquad f_1(\mathbf{x}) = 1 \\ &\qquad\qquad f_2(\mathbf{y}) = 1, \end{aligned} \tag{6}$$

where $f_1(\mathbf{x})$ and $f_2(\mathbf{y})$ are the vector norms on two vectors respectively. For simplicity we use the same elastic net penalty $f(\mathbf{x}) = \alpha\|\mathbf{x}\|_1 + (1-\alpha)\|\mathbf{x}\|_2^2$ for both $\mathbf{x}$ and $\mathbf{y}$.

There is an additional parameter λ_3 in (6) controlling how much the inter-layer links in the resulting modules. Large λ_3 leads to conserved modules across two layers.

Optimization method. In order to solve (6) we use alternating optimization, i.e., iteratively optimizing one variable while fixing another each time [16]. Dealing with one variable has the same form as in (2), so at each iteration in Algorithm 2, the optimization (Line 4) is replaced with:

- Find $\mathbf{x}^{(k+1)}$ such that $F(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k)}) \leq F(\mathbf{x}^{(k)}, \mathbf{y}^{(k)})$
- Find $\mathbf{y}^{(k+1)}$ such that $F(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)}) \leq F(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k)})$,

while other parts of the algorithm remain the same.

Similar to the two-layer case, we can formulate the identification problem for multilayer WGCNs with interlayer links. Alternating optimization can be used as the same way.

2.3.2. Multilayer network without inter-layer interactions. For multilayer WGCNs without inter-layer links, we formulated the module identification problem based on the tensor computational paradigm in [12] . However, different from [12], we use node activity to search for modules and our method is based on elastic net regularization. Formally, the multilayer WGCN module identification problem is defined as

Problem 3. *Given an L-layers network with each layer a complete graph $G = (V, E)$ where $|V| = n$. The vertices weight and non-negative edges weight in the i-th layer are $\mathbf{z}^{(i)} \in \mathbb{R}^n$, $W^{(i)} \in \mathbb{R}^{n \times n}$ respectively. The goal is to find a conserved subgraph T with large vertices weight $\sum_{k=1}^{L} \sum_{i \in T} z_i^{(k)}$ and also edges weights $\sum_{k=1}^{L} \sum_{i,j \in T} w_{ij}^{(k)}$.*

The corresponding objective function becomes

$$\max_{\mathbf{x} \in \mathbb{R}_+^n} F = \sum_{k=1}^{L} (\mathbf{x}^\top W^{(k)} \mathbf{x} + \lambda_k \mathbf{z}^{(k)\top} \mathbf{x}) \quad (7)$$

Subject to

$$f(\mathbf{x}) = 1$$

where λ_k controls the trade-off between edges weights and nodes activity in the k-th layer and $f(\mathbf{x})$ is the vector norm such as elastic net penalty $f(\mathbf{x}) = \alpha \|\mathbf{x}\|_1 + (1 - \alpha)\|\mathbf{x}\|_2^2$. We can solve the optimization problem using Algorithm 1.

2.4. Parameters selection

The parameter λ in equation (3) represents the trade-off between the aggregated edges and nodes scores, the larger λ would include more high scored genes in the module. If there is no prior knowledge or preference about edge scores and nodes scores, we suggest using the default value $\lambda = 1$. We use a binary search method to select α for the elastic net penalty which controls the sparsity of the results, which determines the size of the modules. A suitable size of modules will facilitate their biological interpretation and validation. We empirically set the module size to be around 50-100 and 50-500 for single layer WGCN. See supplementary text S1 section 6 for usage of parameter selection to search the desired size module.

2.5. Implementation

`AMOUNTAIN` is implemented in R as a Bioconductor package. The principle is to provide all functions mentioned above with minimal dependencies. It turns out that pure R runs slowly for large-scale networks. Therefore we provide C implementation of core modules identification functions, in which the matrix operations are based on the GNU Scientific Library (GSL). Overall, C libraries called by R run 10x faster than pure R functions when identifying modules on a single network with 10,000 nodes.

2.6. Data collecting

2.6.1. Synthetic data. Several related works have used artificially generated data [10], [12], [21] to test their algorithms in single network module identification. We follow [12] to construct gene co-expression networks for simulation study (See supplementary text S1 section 3.2). Our simulated networks have a clear topological structure as well as node scores.

2.6.2. Real-world data: Th17 cell differentiation gene expression. Datasets and experimental procedure. We downloaded the gene expression profiles of human Th17 cell differentiation (GSE35103) from Gene Expression Omnibus (GEO) [22]. The dataset was collected to identify transcriptional changes induced by vitro polarization of human cord blood CD4+ cells towards Th17 subtype with the combination of IL6, IL1b, and TGFb [23]. There are 57 samples, consisting of 3 biological replicates of time series data (0, 0.5, 1, 2, 4, 6, 12, 24, 48 and 72 hours) of Th17 polarized cells and control Th0 cells [24].

Inspired by `xHeinz` [25], in addition to *Homo sapiens* dataset GSE35103, we use *Mus musculus* Th17 cell differentiation dataset (GSE43955) for the multilayer cross-species co-expression study. The original papers [26] and [23] reported the expression profiles identification controlled by the differentiation of Th17 cell.

Miroarray preprocessing. To deal with missing or invalid values, we discarded probes with more than 20% missing values or NAs, and replaced them with positions in a valid probe with k-nearest neighbors (KNN, $k = 10$) [27] output of the rest samples of that probe. We did not filter out genes by only selecting significantly expressed genes using a linear model, as `xHeinz` does [25], because the algorithm 2 requires as more information about genes correlation relationship to construct co-expression networks. Furthermore, the whole objective in (2) consists of two parts, and the gene activities only contribute half of it.

Single layer co-expression network construction. Algorithm 1 requires a weighted gene co-expression network as input. Here we just use the co-expression matrix as the

edge score, where each entry W_{ij} in the symmetric means the correlation value of gene i and j, using all samples. In particular, we use the Pearson correlation coefficient estimate, which is widely used in co-expression network construction [28]. And we remove the negative correlation values since positive correlation are preferred to the objective of equation (3) as well as biological relevance [29]. The node score vector $\mathbf{z}$ is computed using limma [30] by comparing the specific group with the control group. In each time point, the expression level measurement p-values represent gene activities for Algorithm 1. As we want to maximize the objective, p-values are replaced by z-scores in practice. Correlation-based similarity requires as many samples while gene activities are closely related to certain conditions, including the exposed time period.

Multilayer cross-species co-expression network construction. To evaluate the performance of our algorithm on multilayer WGCNs with interlayer links, we test it on a cross-species co-expression network, with the aim to find evolutionarily conserved modules. Following the case study in `xHeinz` [25], we use GSE35103 and GSE43955 to construct a two-layer cross-species network, of which each layer is the WGCN of a species. For both layers, we select the gene expression profiles at time point 1h because of two reasons: 1) there are more activities in the early phase of Th17 differentiation in both in mouse [26], [31] and human [23]; and 2) there are enough replicates for both species at this time point.

After gene expression data pre-processing, we obtained 28870 genes in the human layer and 22192 genes in the mouse layer. We then performed the orthologous mapping between human and mouse genes to obtain interlayer connections using from Ensembl [32]. The orthologous mapping resulting 11039 interlayer links and their weights represent the confidence of the mapping.

Multilayer dynamic co-expression network construction. To evaluate the performance of our algorithm on multilayer networks without interlayer links, we applied `AMOUNTAIN` to a dynamic co-expression network to identify the conserved and time-point specific active modules. The dynamic co-expression network was constructed from the human Th17 differentiation gene expression time course dataset GSE35103. The network consists of multiple layers and each layer represent the co-expression network of a time point. Ideally, layer x should be constructed from the samples belong to time point x. But there are only three replicates of each time point which makes the correlation values suspicious. Therefore we use all samples of all the time points to construct the co-expression network for each layer and calculate gene activities from corresponding time points. In other words, each layer shares the same edges scores but with different nodes scores.

The node scores of our dynamic co-expression network are calculated from the gene expression profiles at three later time points, i.e., 12h, 24h, and 48h. This is because Th17 differentiation showed that the effective secretion of Th17 hallmark cytokines only happens after several days of polarization [23], [26]. In essence, we constructed a three-layer dynamic co-expression network of three later time points during human Th17 differentiation.

3. Results and discussion

3.1. Results from Synthetic data

We first evaluated the accuracy of our algorithm on the single-layer WGCN constructed from the synthetic data (See Materials). We compared the accuracy of our algorithm with that of the Multi-Stage Convex Relaxation (MSCR) algorithm [12]. The details of the MSCR algorithm [12] and parameter setting can be found in Section 3 of supplementary text S1.

Our results (See Table S2 in supplementary text S1) show that the proposed method outperformed the MSCR algorithm [12], especially when network size is larger. We also found that `AMOUNTAIN` is less sensitive to the parameters, i.e., it can accurately identify the ground true modules with a range of different parameters (See Table S3-S4 and Figure S1). We also test the robustness of the proposed method by introducing small perturbations to the edge scores and node scores. Our results (See Table S5) show that even with small perturbations to the network, the proposed method can find the ground module with the same parameters.

3.2. Modules of the single-layer human Th17 co-expression network

We first applied `AMOUNTAIN` with default $\lambda = 1$ to the Th17 single layer WGCN, using the time point 2h to compute the node score (see Materials). We provided the identified modules as gene lists in Supplementary files "**Table SS1.xlsx**" in S4.

Since cellular signaling mechanisms involve protein-protein interactions (PPIs) that transmit information, we first investigate whether the co-expression active module modules identified by `AMOUNTAIN` are enriched by PPIs. To this end, we use PPI enrichment analysis provided by database STRING [33]. These PPIs include curated databases and experimentally determined and predicted interactions such as gene neighborhood and gene co-occurrence. Our results show that 33 of 100 modules have significant PPI enrichment ($p<0.05$). If we relax the maximal module size to 500, the number of `AMOUNTAIN` modules with significant PPI enrichment is 50 (Supplementary file **Table SS2.xlsx** in S4). These results indicate that `AMOUNTAIN` was able to find co-expression modules that are enriched by known PPI, which might reveal some signaling mechanisms of a given cellular response.

In addition to PPI enrichment analysis, we also conduct KEGG pathways enrichment analysis to check whether the identified modules are enriched by known signaling pathways. The numbers of modules significantly enriched by KEGG pathways are 88 and 90 out of 100 identified modules, for maximal size 50 and 500 constraints respectively. The top enriched KEGG pathways include 1) Influenza A,

which inhibits Th17 pathway activation by secondary bacterial challenge [34]; 2) Hepatitis C, a common virus infection that could introduce Th17 cells [35]; 3) Prolactin signaling, which may induce the production of Th17 [36] and 4) JAK-STAT signaling, which plays a central role in orchestrating of immune system, especially for cytokines involved in T helper cell differentiation [37], [38]. (See supplementary file "**Table SS4**.xlsx" (size 50-100) and "**Table SS4**.xlsx" (size 50-500) in S4 for all modules).

We investigate the biological function of the identified modules using functional enrichment analysis with STRING [33]. Our results show that 55 out of 100 identified modules (with module size 50-100) are enriched by at least one GO term (biological processes) at a given FDR ($\leq$0.05) cutoff. If we relax the maximal module size to 500, 62 modules that are significantly enriched (FDR $\leq$0.05) by GO terms are found.

We list the first 10 modules with the PPI and GO enrichment information in Table 1. We can see except for the 10th module, all the top 9 modules were enriched by PPIs and biological processes (See supplementary file "**Table SS3**.xlsx" (size 50-100) and "**Table SS4**.xlsx" (size 50-500) in S4 for all modules). We also find that there is a strong correlation between PPI and biological processes, i.e., modules enriched by more protein interactions tend to have more significant GO terms.

Among these 10 modules, the first identified module is enriched by biological processes and pathways related to Th17 differentiation in the early stage [23]. For example, we find that this module consisted of several important transcription factors such as STAT1/STAT2/STAT3, which are known regulators of the Th17 differentiation process [39]. These regulators were surrounded by other genes in the same cytokine signaling pathway (See Table SS1).

It is worth mentioning that the first identified module overlaps with the 29 differentially expressed genes (DEGs) identified by `limma` [30], (See in Table SS1). As shown in Figure 1, there are 26 nodes (annotated with green color) shared between the first module identified by `AMOUNTAIN` and the DEGs. There are 3 DEGs were not included in the first identified module (annotated with red color). However, they were included in the second module identified by `AMOUNTAIN`. We speculate the reason for the significant overlap is partly because Th17 differentiation exhibits a high level of activity in the early stage [23]. However, from a two-layer cross-species network below or another study on ankylosing spondylitis disease samples (See Supplementary S2), the first identified module was different from DEGs.

3.3. Active module identification in cross-species networks

We applied our algorithm to the cross-species network of human and mouse Th17 differentiation at 1 hour (See Materials), with default value $\lambda_1 = \lambda_2 = 1$. We set $\lambda_3 = 1000$ (Details of tuning λ_3 are listed in supplementary text S1, section 4). The identified 100 modules with their size and shared inter-layer links information are stored in supplementary "**Table SS5**.xlsx". Due to the space limit, we select the first identified conserved module for further analysis.

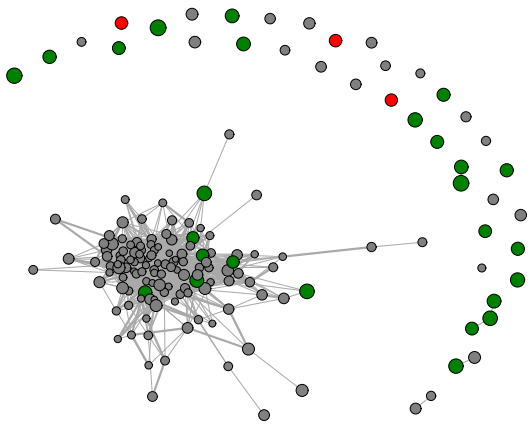

Figure 1: The first identified active module of single layer co-expression network, and the DEGs. We applied the cuff-off threshold (correlation coefficient $\leq$ 0.8) to delete those edges with low correlations. In the module, node sizes are proportional to the intensities of gene activities and edge widths to the correlation coefficients. The green nodes are shared by identified module and DEGs, and the red nodes are only in DEGs.

In the first identified conserved module, there are 52 and 57 genes in human and mouse layers, respectively. There are 52 conserved genes which include several key genes such as STAT2/SOCS3/IRF1 in Th17 cell differentiation [40], [41], [42]. The complete gene list of the 2-layer modules is provided in supplementary file "**Table SS1**.xlsx".

In order to illustrate the potential conserved signaling mechanisms, we overlaid known interactions from STRING to the (nodes) genes extracted from the two layers as shown in Figure 2 and 3. (See the corresponding co-expression modules in Figure S6 and Figure S7 in supplementary text S1). We can see that genes in both layers centered around STAT2/Stat2 and IRF1/Irf1, which are the key transcriptional regulators of Th17 cell differentiation [40], [41], [42].

Our functional and KEGG pathway enrichment results show that both modules share some common pathways such as the JAK-STAT signaling pathway [37], [38] and those pathways are relevant to Th17 differentiation. The detailed top enriched biological functions and KEGG pathways enriched by modules from both species are listed in Table S6 in supplementary text S1.

3.4. Results from the dynamic multilayer networks

In order to unveil the dynamic regulatory and signaling mechanisms of the Th17 differentiation, we applied our `AMOUNTAIN` algorithm to the Th17 dynamic multilayer network (See Materials). `AMOUNTAIN` can readily identify conserved modules. Also, to depict the dynamic changes

TABLE 1: Overview of top 10 modules identified from single layer WGCN of human, at 2 hour time point. "PPI P-value" indicates if there are any significant known protein interactions in the module. "P-value" is for the corresponding GO term.

ID	Size	PPI P-value	Representative GO term (BP) and description	P-value
1	161	0	GO:0019221, cytokine-mediated signaling pathway	2.03E-18
2	190	0	GO:0044711, single-organism biosynthetic process	1.43E-9
3	294	4.02E-3	GO:0032479, regulation of type I interferon production	9.51E-6
4	150	1.54E-6	GO:0050860, negative regulation of T cell receptor signaling pathway	1.66E-6
5	234	0	GO:0000278, mitotic cell cycle	3.31E-21
6	301	0	GO:0000122, negative regulation of transcription from RNA polymerase II promoter	3.12E-8
7	248	4.88E-5	GO:0051726, regulation of cell cycle	1.77E-10
8	182	9.13E-3	GO:0006955, immune response	4.04E-8
9	73	3.31E-13	GO:0002764, immune response-regulating signaling pathway	1.01E-5
10	54	0.37	None	None

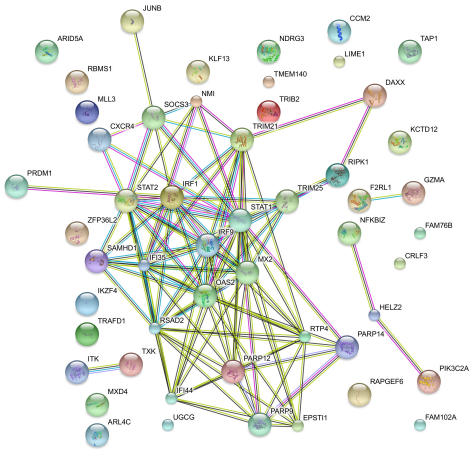

Figure 2: The first identified module in the human layer at 1 hour time point, plotted by STRING, where edges represent the known interactions. Colored nodes standard for query proteins and first shell of interactors, and white nodes for the second shell of interactors.

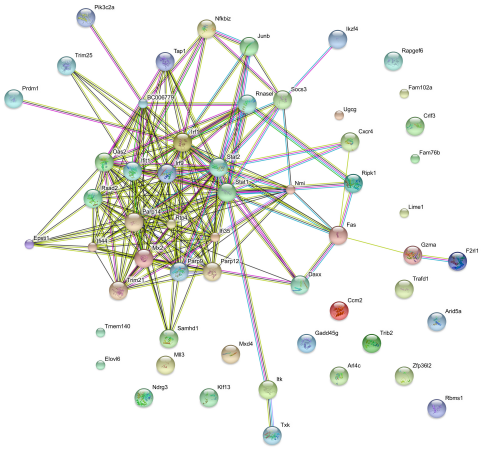

Figure 3: The first identified module in the mouse layer at 1 hour time point, plotted by STRING, interactions are denser compared with human layer. Key transcriptional factors Stat2/Irf1 are densely surrounded by interactions.

of co-expression networks, we identified time point specific modules by applying `AMOUNTAIN` to each layer. Figure **??** shows the first identified conserved module and Figure S8-S10 in the supplementary text S1 show the three-time point specific modules respectively.

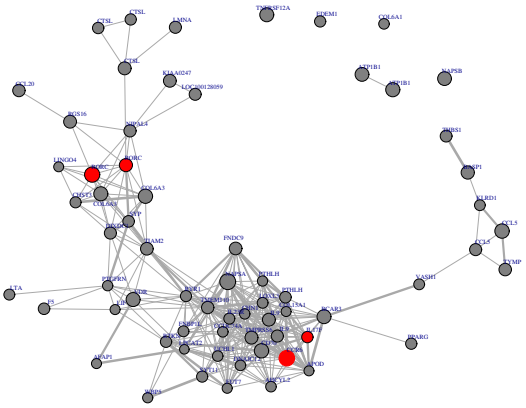

Figure 4: The first identified conserved module for a three layer network where each layer represents nodes from 12h, 24h and 48h respectively. The red nodes are two probes of gene RORC, a signature gene w.r.t. Th17 lineage commitment. Plotted by `igraph` [43].

The conserved module identified from the three-layers dynamic network across the later time points (12h, 24h, and 48h) includes several signature genes of Th17 lineage commitment, e.g., RORC and RUNX1 [44], [45]. These genes showed significantly different expression profile compared with the Th0 group (See Sheet 1 of "**Table SS6**.xlxs" in supplementary files). We found that in the conserved module, RORC gene always interacted with VDR (Vitamin D Receptor), which is very relevant to T cell development and differentiation [46], [47]. Another interesting finding is that, in the conserved module, RORC gene also interacted with BHLHE40, a transcription factor that controls cytokine production by T cells [48].

By comparing the conserved modules with the three-time point specific modules, we found that some genes only connected to RORC at specific time points. For example, RBPJ was only identified in the time point specific module from the network at 24h, which is a known regulator of the Notch signaling pathway [49] and plays an important role

in lineage fate decisions in cells. The above result indicates that our algorithm can identify co-expression active modules from the multilayer dynamic network to provide insights into the dynamics of Th17 lineage commitment.

3.5. Discussion

In addition to the Th17 datasets, we have applied AMOUNTAIN to gene expression profiles of ankylosing spondylitis samples [50] as another case study. Our results indicate the identified modules consists of enriched biological processes and pathways which are consistent with the results of the original study [50]. We provide the detail results in Supplementary text S2.

We have demonstrated the performance of AMOUNTAIN using WGCNs constructed using the Pearson correlation coefficient, which is a standard way for constructing WGCNs. However, the users can apply it on WGCNs constructed from Spearman correlation or other association methods [51].

Related works. Although algorithms in [25], [52] can identify evolutionarily conserved modules from two-layer cross-species PPI networks, to our best knowledge, general active module identification algorithms for multilayer gene co-expression networks do not exist. The most relevant algorithm is the algorithm in [12], which was proposed to identify heavy recurrent modules from multiple gene co-expression networks. However, this algorithm differs from AMOUNTAIN in the following perspectives: 1) The algorithm in [12] is specifically designed for multi-slice (multiplex) networks, which share exactly the same set of nodes, while the AMOUNTAIN algorithm is designed for more general multilayer networks. For example, different layers could have different sets of nodes and the inter-layer interactions can be considered. 2) The algorithm [12] only considers edge weights while AMOUNTAIN considers both edge weights and node scores (hence the modules are called active modules). 3) the algorithm in [12] is based on a non-convex regularization while AMOUNTAIN algorithm adopts a convex regularization which is more efficient to achieve sparsity of solutions.

The optimization problem. Maximizing the constrained quadratic function (3) with indefinite matrix is NP-hard [53]. In a different context, i.e., shape matching in computer vision, Rodola et al. [54] solved the same problem as the objective function (3) using a projected gradient method. The only difference is the procedure to solve the subproblem (5) since their target solution was not sparse as here.

In [12], Li et al. solved a similar problem using power method, followed by a normalization step. We found our projected gradient method performed better than the power method in terms of convergence rate and accuracy (See Figure S11 in supplementary text S1).

Difference with WGCNA. Although both WGCNA [10] and AMOUNTAIN can analyze WGCNs, there is a crucial difference between them: WGCNA partitions the whole network and groups the genes with similar biological functions into co-expression modules (hence functional modules), while AMOUNTAIN extracts active modules with significant node activities. These active modules could be used to generate hypotheses of the signaling and regulatory mechanisms of a given cellular response [1]. By controlling the size, AMOUNTAIN can identify small modules which facilitate follow-up experiments to test the hypotheses.

Another technical difference between WGCNA and AMOUNTAIN is the network construction. WGCNA uses absolute value of the correlation while AMOUNTAIN only uses the positive correlations. To see the effect of this difference, we executed the WGCNA algorithm on the same single layer network constructed from GSE35103. As discussed in supplementary text S3 and presented in supplementary file "**Table SS7**.xlsx"), WGCNA identified 38 modules with average size 760, of which 12 had significant PPI enrichment (p-values$<$0.05. These modules also consists of a large proportion of isolated genes. In comparison, AMOUNTAIN identified more modules with significant PPI enrichment and less genes (See Table 1).

Limitations of AMOUNTAIN. Although the AMOUNTAIN is robust to the noise in the WGCNs (See Table S5) and not sensitive to the parameters (See Table S4), deciding the size of the active modules will affect the identification performance. In our current implementation of AMOUNTAIN, the users need to determine the module size according to the preference or prior knowledge. It is our future work to find a way to determine the module size rigorously and automatically.

Another limitation of AMOUNTAIN is that it can only identify non-overlapping modules. This is due to the simplicity of our algorithm which keeps the optimization procedure unchanged by deleting each module from the network once it is identified. Since in real networks modules do overlap, our future work will extent AMOUNTAIN to overlapping modules identification.

4. Conclusion

This paper presents AMOUNTAIN, a general and efficient active modules identification algorithm for single layer and multilayer WGCNs. Our algorithm is based on a new definition of active modules in WGCNs. This definition enables us to formulate the module identification problem that not only considers the correlation between genes but also their activity. We also generalized the active module identification problem in single layer WGCNs to multilayer WGCNs. Another main contribution of our paper is the continuous optimization formulation of the problem, which achieves better efficiency when dealing with large-scale networks.

We provide AMOUNTAIN as an R package which is freely available at Bioconductor. We expect our AMOUNTAIN algorithm can be applied to a wide range of problems that involve identifying dynamic and evolutionary mechanisms associating with a cellular response.

Acknowledgments

The authors would like to thank Dr. James B. Brown and Dr. Luisa Orsini for discussions.

References

[1] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[2] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. suppl 1, pp. S233–S240, 2002.

[3] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, no. 1, p. 140, 2007.

[4] Y.-Q. Qiu, S. Zhang, X.-S. Zhang, and L. Chen, "Detecting disease associated modules and prioritizing active genes based on high throughput data," *BMC bioinformatics*, vol. 11, no. 1, p. 26, 2010.

[5] H. Ma, E. E. Schadt, L. M. Kaplan, and H. Zhao, "Cosine: Condition-specific sub-network identification using a global optimization method," *Bioinformatics*, vol. 27, no. 9, pp. 1290–1298, 2011.

[6] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Müller, "Identifying functional modules in protein–protein interaction networks: an integrated exact approach," *Bioinformatics*, vol. 24, no. 13, pp. i223–i231, 2008.

[7] W. Chen, J. Liu, and S. He, "Prior knowledge guided active modules identification: an integrated multi-objective approach," *BMC systems biology*, vol. 11, no. 2, p. 8, 2017.

[8] D. Li, Z. Pan, G. Hu, Z. Zhu, and S. He, "Active module identification in intracellular networks using a memetic algorithm with a new binary decoding scheme," *BMC Genomics*, vol. 18, no. 2, p. 209, 2017.

[9] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[10] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[11] Y. Wang and Y. Xia, "Condition specific subnetwork identification using an optimization model," *Proc Optim Syst Biol*, vol. 9, pp. 333–340, 2008.

[12] W. Li, C.-C. Liu, T. Zhang, H. Li, M. S. Waterman, and X. J. Zhou, "Integrative analysis of many weighted co-expression networks using tensor computation," *PLoS Comput Biol*, vol. 7, no. 6, p. e1001106, 2011.

[13] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[16] C.-b. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[17] P. Gong, K. Gai, and C. Zhang, "Efficient euclidean projections via piecewise root finding and its application in gradient projection," *Neurocomputing*, vol. 74, no. 17, pp. 2754–2766, 2011.

[18] P. H. Calamai and J. J. Moré, "Projected gradient methods for linearly constrained problems," *Mathematical programming*, vol. 39, no. 1, pp. 93–116, 1987.

[19] Y. Zhao, E. Levina, and J. Zhu, "Community extraction for social networks," *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7321–7326, 2011.

[20] Y. Liu, D. A. Tennant, Z. Zhu, J. K. Heath, X. Yao, and S. He, "Dime: a scalable disease module identification algorithm with application to glioma progression," *PloS one*, vol. 9, no. 2, 2014.

[21] D. Rajagopalan and P. Agarwal, "Inferring pathways from gene lists using a literature-derived network of biological relationships," *Bioinformatics*, vol. 21, no. 6, pp. 788–793, 2005.

[22] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[23] S. Tuomela, V. Salo, S. K. Tripathi, Z. Chen, K. Laurila, B. Gupta, T. Äijö, L. Oikari, B. Stockinger, H. Lähdesmäki *et al.*, "Identification of early gene expression changes during human th17 cell differentiation," *Blood*, vol. 119, no. 23, pp. e151–e160, 2012.

[24] T. Pramila, S. Miles, D. GuhaThakurta, D. Jemiolo, and L. L. Breeden, "Conserved homeodomain proteins interact with mads box protein mcm1 to restrict ecb-dependent transcription to the m/g1 phase of the cell cycle," *Genes & development*, vol. 16, no. 23, pp. 3034–3045, 2002.

[25] M. El-Kebir, H. Soueidan, T. Hume, D. Beisser, M. Dittrich, T. Müller, G. Blin, J. Heringa, M. Nikolski, L. F. Wessels *et al.*, "xheinz: An algorithm for mining cross-species network modules under a flexible conservation model," *Bioinformatics*, p. btv316, 2015.

[26] N. Yosef, A. K. Shalek, J. T. Gaublomme, H. Jin, Y. Lee, A. Awasthi, C. Wu, K. Karwacz, S. Xiao, M. Jorgolli *et al.*, "Dynamic regulatory network controlling th17 cell differentiation," *Nature*, vol. 496, no. 7446, pp. 461–468, 2013.

[27] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[28] L. L. Elo, H. Järvenpää, M. Orešič, R. Lahesmaa, and T. Aittokallio, "Systematic construction of gene coexpression networks with applications to human t helper cell differentiation process," *Bioinformatics*, vol. 23, no. 16, pp. 2096–2103, 2007.

[29] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets," *Genome research*, vol. 14, no. 6, pp. 1085–1094, 2004.

[30] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

[31] M. Ciofani, A. Madar, C. Galan, M. Sellars, K. Mace, F. Pauli, A. Agarwal, W. Huang, C. N. Parkurst, M. Muratet *et al.*, "A validated regulatory network for th17 cell specification," *Cell*, vol. 151, no. 2, pp. 289–303, 2012.

[32] A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil *et al.*, "Ensembl 2016," *Nucleic acids research*, vol. 44, no. D1, pp. D710–D716, 2015.

[33] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou *et al.*, "String v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic acids research*, p. gku1003, 2014.

[34] A. Kudva, E. V. Scheller, K. M. Robinson, C. R. Crowe, S. M. Choi, S. R. Slight, S. A. Khader, P. J. Dubin, R. I. Enelow, J. K. Kolls *et al.*, "Influenza a inhibits th17-mediated host defense against bacterial pneumonia in mice," *The Journal of Immunology*, vol. 186, no. 3, pp. 1666–1674, 2011.

[35] A. G. Rowan, J. M. Fletcher, E. J. Ryan, B. Moran, J. E. Hegarty, C. O'Farrelly, and K. H. Mills, "Hepatitis c virus-specific th17 cells are suppressed by virus-induced tgf-β," *The Journal of Immunology*, vol. 181, no. 7, pp. 4485–4494, 2008.

[36] C. Hau, N. Kanda, Y. Tada, S. Shibata, S. Sato, and S. Watanabe, "Prolactin induces the production of th17 and th1 cytokines/chemokines in murine imiquimod-induced psoriasiform skin," *Journal of the European Academy of Dermatology and Venereology*, vol. 28, no. 10, pp. 1370–1379, 2014.

[37] J. J. O'Shea, S. M. Steward-Tharp, A. Laurence, W. T. Watford, L. Wei, A. S. Adamson, and S. Fan, "Signal transduction and th17 cell differentiation," *Microbes and Infection*, vol. 11, no. 5, pp. 599–611, 2009.

[38] F. Seif, M. Khoshmirsafa, H. Aazami, M. Mohsenzadegan, G. Sedighi, and M. Bahar, "The role of jak-stat signaling pathway and its regulators in the fate of t helper cells," *Cell Communication and Signaling*, vol. 15, no. 1, p. 23, 2017.

[39] L. Durant, W. T. Watford, H. L. Ramos, A. Laurence, G. Vahedi, L. Wei, H. Takahashi, H.-W. Sun, Y. Kanno, F. Powrie *et al.*, "Diverse targets of the transcription factor stat3 contribute to t cell pathogenicity and homeostasis," *Immunity*, vol. 32, no. 5, pp. 605–615, 2010.

[40] J. Zhu, H. Yamane, and W. E. Paul, "Differentiation of effector cd4 t cell populations," *Annual review of immunology*, vol. 28, p. 445, 2010.

[41] H. Yang, S.-M. Lee, B. Gao, J. Zhang, and D. Fang, "Histone deacetylase sirtuin 1 deacetylates irf1 protein and programs dendritic cells to control th17 protein differentiation during autoimmune inflammation," *Journal of Biological Chemistry*, vol. 288, no. 52, pp. 37 256–37 266, 2013.

[42] K. Karwacz, N. Yosef, and V. Kuchroo, "Irf-1 is a key transcriptional regulator of tr1 differentiation (p1135)," *The Journal of Immunology*, vol. 190, no. 1 Supplement, pp. 50–10, 2013.

[43] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[44] V. Lazarevic, X. Chen, J.-H. Shim, E.-S. Hwang, E. Jang, A. N. Bolm, M. Oukka, V. K. Kuchroo, and L. H. Glimcher, "T-bet represses th17 differentiation by preventing runx1-mediated activation of the gene encoding ror [gamma] t," *Nature immunology*, vol. 12, no. 1, pp. 96–104, 2011.

[45] A. V. Villarino, E. Gallo, and A. K. Abbas, "Stat1-activating cytokines limit th17 responses through both t-bet–dependent and–independent mechanisms," *The Journal of Immunology*, vol. 185, no. 11, pp. 6461–6471, 2010.

[46] S. H. Chang, Y. Chung, and C. Dong, "Vitamin d suppresses th17 cytokine production by inducing c/ebp homologous protein (chop) expression," *Journal of Biological Chemistry*, vol. 285, no. 50, pp. 38 751–38 755, 2010.

[47] M. Kongsbak, T. B. Levring, C. Geisler, and M. R. Von Essen, "The vitamin d receptor andt cell function," *Lipid Signaling in T Cell Development and Function*, p. 119, 2015.

[48] C.-C. Lin, T. R. Bradstreet, E. A. Schwarzkopf, J. Sim, J. A. Carrero, C. Chou, L. E. Cook, T. Egawa, R. Taneja, T. L. Murphy *et al.*, "Bhlhe40 controls cytokine production by t cells and is essential for pathogenicity in autoimmune neuroinflammation," *Nature communications*, vol. 5, 2014.

[49] K. Tanigaki and T. Honjo, "Chapter seven-two opposing roles of rbp-j in notch signaling," *Current topics in developmental biology*, vol. 92, pp. 231–252, 2010.

[50] F. M. Pimentel-Santos, D. Ligeiro, M. Matos, A. F. Mourão, J. Costa, H. Santos, A. Barcelos, F. Godinho, P. Pinto, M. Cruz *et al.*, "Whole blood transcriptional profiling in ankylosing spondylitis identifies novel candidate genes that might contribute to the inflammatory and tissue-destructive disease aspects," *Arthritis research & therapy*, vol. 13, no. 2, p. R57, 2011.

[51] S. Kumari, J. Nie, H.-S. Chen, H. Ma, R. Stewart, X. Li, M.-Z. Lu, W. M. Taylor, and H. Wei, "Evaluation of gene association methods for coexpression network construction and biological knowledge discovery," *PloS one*, vol. 7, no. 11, 2012.

[52] G. E. Zinman, S. Naiman, D. M. O'Dee, N. Kumar, G. J. Nau, H. Y. Cohen, and Z. Bar-Joseph, "Moduleblast: identifying activated subnetworks within and across species," *Nucleic acids research*, vol. 43, no. 3, pp. e20–e20, 2015.

[53] S. Burer and A. N. Letchford, "On nonconvex quadratic programming with box constraints," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 1073–1089, 2009.

[54] E. Rodola, A. Torsello, T. Harada, Y. Kuniyoshi, and D. Cremers, "Elastic net constraints for shape matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1169–1176.