Subject Section

BEM: Mining Coregulation Patterns in Transcriptomics via Boolean Matrix Factorization

Lifan Liang[1], Kunju Zhu[1,2] and Songjian Lu[1,*]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, 15206-3701, United States., [2]Clinical Medicine Research Institute, Jinan University, Guangzhou, 51063, Guangdong, China.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The matrix factorization is an important way to analyze co-regulation patterns in transcriptomic data, which can reveal the tumor signal perturbation status and subtype classification. However, current matrix factorization methods do not provide clear bicluster structure. Furthermore, these algorithms are based on the assumption of linear combination, which may not be sufficient to capture the coregulation patterns.

Results: We presented a new algorithm for Boolean matrix factorization via expectation maximization (BEM). BEM is more aligned with the molecular mechanism of transcriptomic coregulation and can scale to matrix with over 100 million data points. Synthetic experiments showed that BEM outperformed other Boolean matrix factorization methods in terms of reconstruction error. Real world application demonstrated that BEM is applicable to all kinds of transcriptomic data, including bulk RNAseq, single cell RNAseq, and spatial transcriptomic datasets. Given appropriate binarization, BEM was able to extract coregulation patterns consistent with disease subtypes, cell types, or spatial anatomy.

Availability: Python source code of BEM is available on https://github.com/LifanLiang/EM_BMF

Contact: songjian@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Grouping genes or samples according to their shared expression patterns was an important task. On the genes' side, similar expression profiles across conditions indicated coregulation of gene expression, which can be used to infer upstream pathway activities (Tai *et al.*, 2018) and the regulatory relationship between transcription regulators and target genes (Paul *et al.*, 2015). On the samples' side, clusters of samples helps reveal the heterogeneity in disease population. For example, (Sørlie *et al.*, 2001) has identified clinically relevant breast cancer subtypes from expression profiles alone. This task has become ever more prevalent since the emergence of new technologies such as single cell RNAseq (Patel *et al.*, 2014) and spatial transcriptomics (Berglund *et al.*, 2018), which enable us to interrogate tumor heterogeneity with finer granularity.

However, clustering directly on only one side (either on the sample side or the gene side) yields limited performance. That is because computational distance between objects are contaminated by the noise in irrelevant features. As illustrated in Fig. 1a, given that coregulation mechanism is prevalent in expression profile, the similarity between samples/conditions should only depend on a small group of genes with common upstream factors. Since at most several hundred genes can be coregulated, the distinctive expression profiles for a cluster of samples is no more than several hundred genes. This means all the other genes acted as random noise for the identification of this one cluster. Most studies handled such issues with feature selection. This approach required prior knowledge or external information, which potentially hinders the identification of novel and interesting features. Moreover, the issue of contamination cannot be resolved even with perfect feature selection. A selected subset of genes can be informative features for one cluster while being random noise to another.

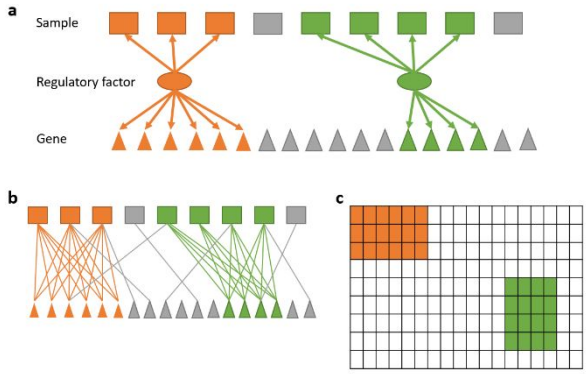
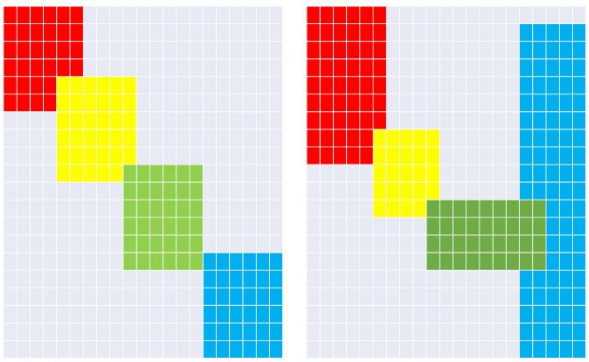


Fig. 1. Figure (a) shows the assumed generative process of transcriptomic data, where a regulatory factor present in a subset of samples causes a similar expression status in a subset of genes. Hence identifying the group of samples with the same regulatory factor require only the corresponding subset of regulated genes. In Fig (b) the identification of regulatory factors / biclusters are transformed into dense bipartite graph problem, where an edge between sample and gene encodes gene activation or differential expression in a sample. Fig (c) showed that by transforming the bipartite graph as an unweighted adjacency matrix, Boolean matrix factorization can be viewed as a problem of permuting rows and columns to uncover dense block patterns corresponding to dense bipartite subgraphs.

Fig. 2. Figure on the left showed a scenario where dense blocks (most values are 1) comply



with uniform Bernoulli prior and thus have roughly the same shape (6×6); figure on the right showed a more realistic scenario where blocks are allowed to have various shapes.

Thus, biclustering should be a natural choice when it comes to high dimensional gene expressions analysis. By finding clusters and their corresponding features simultaneously, biclustering directly resolved the contamination issues above. Since first proposed by (Cheng and Church, 2000), various biclustering algorithms have been developed and applied to gene expression data (Xie *et al.*, 2019). However, most biclustering algorithms (Tanay *et al.*, 2002; Bergmann *et al.*, 2003; Li *et al.*, 2009) are heuristic-based with local iterative search. Thus these algorithms are mostly used to identify subtle gene-sample substructure, rather than tasks requiring systematic analysis of sample heterogeneity such as subtype classification or cell type deconvolution.

Another popular approach is to factorize the gene-sample matrix (Stein-O'Brien *et al.*, 2018). As an example, nonnegative matrix factorization (NMF) has been widely used in gene expression analysis in the past decade (Brunet *et al.*, 2004). By performing dimension reduction on both columns and rows, the matrix factorization approach provides more information about the global heterogeneity. Recently, a comprehensive evaluation (Saelens *et al.*, 2018) showed that matrix factorization outperforms clustering and biclustering in terms of identifying coexpression modules. However, this approach does not explicitly provide clustering structures. Even for methods that provides bicluster structure (Hochreiter *et al.*, 2010), the assumption of linear combination may not be sufficient to capture the coregulation patterns in transcriptomic data.

A less well-known alternative, Boolean matrix factorization, performs systematic dimension reduction like the conventional matrix factorization methods while providing a clear clustering structures through the inferred Boolean factors. As illustrated in Fig. 1B&C, Boolean matrix factorization can be viewed as solutions to dense subgraph extraction, which directly connects to the graph theoretic biclustering model (Tanay *et al.*, 2002). Although dichotomization of gene expression matrix incurs loss of information, it greatly simplifies possible bicluster patterns (e.g. shift, scale, plaid) and noise patterns in the original data. Its performance on gene expression data was favorably compared with popular biclustering algorithms (Zhang *et al.*, 2010). Furthermore, recent advances on Boolean matrix factorization (Rukat *et al.*, 2017; Ravanbakhsh *et al.*, 2016) are able to find solutions robust against noise and distortion with linear complexity, which is of great practical values for high-throughput data analysis. By applying their methods to single cell RNAseq data, LoM (Rukat *et al.*, 2017) has accurately recovered certain cell types and corresponding markers.

However, to achieve such robustness and efficiency, recent Boolean matrix factorization algorithms (Rukat *et al.*, 2017; Ravanbakhsh *et al.*, 2016; Neumann, 2018) have made strong assumptions about the shape of the latent factors. As illustrated in Fig. 2, the fully Bayesian approach (Rukat *et al.*, 2017) assumed that Boolean factors were generated by a uniform Bernoulli distribution. Another algorithm (Neumann, 2018) assumed that one side of the matrix only contains clusters with relatively large size. Unfortunately, prior knowledge on factors' sizes are usually unavailable in transcriptomic data. Previous studies have shown that the number of genes within a coexpression module can be less than 100 (Padilha and Campello, 2017) or close to 1000 (van Dam *et al.*, 2018). The number of samples within a cluster are even more unpredictable. Thus, assumptions about bicluster sizes may impose strong bias in gene expression analysis.

In this study, we presented a new algorithm free of assumptions about Boolean factors while retaining the advantages of previous algorithms. As illustrated in Fig. 1, our algorithm aimed to identify Boolean factors accurately in a more realistic scenario where latent Boolean factors can take any shape.

As described in Section 3, our approach consists of three novel ideas: (1) allow the latent factors to vary by relaxing the parameters to be continuous values within [0, 1] sampled from Beta distribution; (2) reparameterize the parameters from [0, 1] to $(-\infty, +\infty)$ thus making simple gradient ascent feasible; (3) uniform noise is directly modeled and jointly estimated with latent factors in an EM algorithm.

In Section 4, our algorithm was compared with LoM and message passing. Synthetic experiments showed that our algorithm outperformed both of them when latent factors' sizes varied considerably with each other and observation is not overwhelmed by noise. Real data experiment also indicated that our algorithm has extracted more information with the same number of latent factors

2 Methods

2.1 Problem formulation

The problem of Boolean matrix factorization (BMF) is to identify two

binary matrices, U and Z, with rank L such that every element in the N by M binary matrix, X, is an OR mixture of AND product:

$$X_{nm} = \vee_{l \le L}(U_{nl} \wedge Z_{jl})$$

here $\vee$ is the OR operator and $\wedge$ is the AND operator. In this study, however, a different formulation was adopted. We assume that each element of X, X_{nm}, is sampled from a different Bernoulli distribution. Similarly, every element in the latent factors is sampled from different Bernoulli distributions. The generative process of X can be described as follows:

$$U_{nl} \sim Bernoulli(\mu_{nl})$$

$$Z_{ml} \sim Bernoulli(\zeta_{ml})$$

$$P_{nm} = 1 - P(X_{nm} = 0) = 1 - \prod_{l \le L}(1 - \mu_{nl} * \zeta_{ml})$$

$$X_{nm} \sim Bernoulli(P_{nm})$$

where μ is a N×L matrix with values in [0, 1], ζ is a M×L matrix with values in [0, 1]. Clearly, by forcing μ and ζ to be binary, our formulation will be identical to previous Bayesian approaches. Thus, our formulation is a generalized version of previous ones. With this approach, our goal for Boolean matrix factorization is to estimate the parameter μ and ζ instead of their samples U and Z.

2.2 Maximum likelihood estimation

μ and ζ can be estimated by maximizing the log likelihood of X, which is:

$$LL(\mu,\zeta;X) = \sum_{n \le N,\, m \le M}[X_{nm} log P_{nm} + (1 - X_{nm}) \log (1 - P_{nm})]$$

Conventional gradient descent is not application because μ and ζ need to be within the interval [0, 1]. Thus, μ and ζ are reparameterized as $\sigma(A)$ and $\sigma(B)$ elementwise:

$$\mu_{nl} = 1/(1 + e^{-A_{nl}})$$

$$\zeta_{nl} = 1/(1 + e^{-B_{nl}})$$

With reparameterization, it becomes a problem of unconstrained nonlinear programming. A simple gradient ascent algorithm is sufficient to jointly optimize the estimators of A and B. The partial likelihood gradients regarding A is:

$$\partial LL/\partial A_{il} = \sum_{m \le M}[\mu_{nl}\zeta_{ml}/(1 - \mu_{nl}\zeta_{ml})(1 - \mu_{nl})(1 - X_{nm}/P_{nm})]$$

Note that A and B are symmetric, thus the partial gradient of B can be computed similarly as A. In subsequent description, equations related to B and Z were also neglected due to this symmetry.

2.3 Noise estimation

We further introduced a parameter, ϵ, to explicitly model the probability that elements in X is contaminated by noise. In this scenario, the observed data, X*, is generated as:

$$C_{nm} \sim Bernoulli(\epsilon)$$

$$X^*_{nm} = ABS(X_{nm} - C_{nm})$$

where C_{nm} is a N×M binary matrix with every element as a i.i.d sample from a Bernoulli distribution parameterized by a scalar ϵ lying between [0, 1]. ABS is the function of taking absolute values. To reflect the addition of noise in the model, we need to add one step in the generative process:

$$P^* = (1 - \epsilon)P + \epsilon(1 - P)$$

The noisy observation, X^*, is sampled from P* instead of P:

$$X^*_{nm} \sim Bernoulli(P^*)$$

Thus, the model likelihood becomes:

$$LL(\mu,\zeta;X^*) = \sum_{n \le N,\, m \le M}[X^*_{nm} log P^*_{nm} + (1 - X^*_{nm}) \log (1 - P^*_{nm})]$$

To optimize the likelihood function regarding μ, ζ and ϵ, we applied the expectation maximization algorithm. In M step, μ and ζ are estimated with the same approach as described in Section 2.2. The difference is the presence of a fixed ϵ, leading to a different equation for likelihood gradients:

$$\partial LL/\partial A_{il} = \sum_{m \le M}[\mu_{nl}\zeta_{ml}(1 - \mu_{nl})(1 - P_{nm})(1 - 2\epsilon)(P^*_{nm} - X'_{ij})/(1 - P^*_{nm})/P^*_{nm}/(1 - \mu_{nl}\zeta_{ml})]$$

In E step, based on the modified generative process described in the beginning of this section, the expected value of ϵ is equivalent to the difference between the noisy observation, X*, and the reconstructed data without noise, $\hat{X}$:

$$\epsilon = |C|/(NM) = |\hat{X} - X^*|/(NM)$$

The estimate above is only approximate. The exact estimate should be the average difference between X* and P*. However, the exact estimate requires a much more stringent convergence criterion in the M step. During synthetic experiments, the performance of approximate estimate is not significantly different from the exact one. Thus, the approximate estimate of ϵ was adopted.

2.4 MAP Estimation as Regularization

We further impose prior distribution on μ and ζ:

$$\mu_{ml} \sim Beta(\alpha, \beta)$$

$$\zeta_{ml} \sim Beta(\alpha, \beta)$$

In practice, μ and ζ can comply with different Beta distributions. For the convenience of notation, we simply assume they have a common prior distribution. Thus μ and ζ are estimated based on Maximum a Posteriori (MAP) estimator. The posterior probability function of μ and ζ is:

$$\log [\Pr (X \mid \mu,\zeta,\epsilon)] = LL + (\alpha - 1) \left[\sum_{m \le M,\, l \le L} log \mu_{ml} + \sum_{n \le N, l \le L} log \zeta_{nl}\right] + (\beta - 1)[\sum_{m \le M,\, l \le L} \log (1 - \mu_{ml}) + \sum_{n \le N, l \le L} \log (1 - \zeta_{nl})]$$

where LL is described in Section 2.3. We applied gradient ascent to the objective function. The partial gradient for $\Pr(X \mid \mu,\zeta,\epsilon)$ is:

$$\partial \Pr(X \mid \mu,\zeta,\epsilon)/\partial A_{nl} = \partial LL/\partial A_{nl} + (\alpha - 1)(1 - \mu_{nl}) - (\beta - 1)\mu_{nl}$$

Clearly, when α and β are set to 1, the MAP estimator will be identical to the maximum likelihood estimator. When α and β $ are larger than 1, latent factors will be skewed towards 0.5; when α and β are less than 1, latent factors are pushed towards 0 or 1. Alternatively, the entropy of μ and ζ can be used as penalty and the objective becomes minimizing KL divergence. However, users can push the sparsity of latent factors by making α and β asymmetric, which is not available with entropy.

2.5 Handling missing values

As briefly mentioned in Section 2.1, our approach to matrix completion is simple. During training, parameters are only updated based on the gradients from the observed data points. When convergence is reached, missing data are imputed by the reconstructed data without noise.

2.6 Model Selection

During synthetic experiments, we found that minimum Akaike information criterion (AIC) always corresponded to the correct number of factors. Thus, the number of Boolean factors in real data experiments was chosen based on AIC:

$$AIC = 2(N + M)L - 2\log[\Pr(X \mid \mu,\zeta,\epsilon)]$$

where $\log[\Pr(X \mid \mu,\zeta,\epsilon)]$ is the objective function of the model. Thus, it can be easily calculated.

2.7 Implementation

The pseudo code for the model proposed in this study is shown in Algorithm 1. In addition to the theoretical aspects illustrated in previous sections, here we illustrated several practical decisions based on the algorithm's performance during synthetic experiments: (1) resilient propagation on full batch was adopted to optimize the estimator of latent factors. This is because of its superior performance in terms of convergence rates and optimum loss when compared to vanilla gradient ascent, SGD, and ADAM; (2) the convergence criteria for M-step is whether the reconstructed data is the same as the previous iteration; (3) the priors, α and β, are set to 0.95 across all the experiments, slightly pushing parameters towards 0 and 1; (4) parameters A and B were clipped after each update, meaning that all of them are bounded within [-5, 5]. This is necessary given the setting of priors and the convergence criteria.

As for the computational complexity, the most time-consuming step is to compute the partial gradient for each element in the factor. The computational complexity in one iteration is O(NML). The size of latent factors, L, is fixed and usually small. Thus, the complexity of our algorithm is still linear to the size of the matrix.

Algorithm 1: Expectation Maximization

Input : X an $N \times M$ binary matrix; L number of latent factors; α, β, Beta priors

Output: μ, ζ, latent factors; ϵ, flip probability

$\epsilon \leftarrow 0$;
$A^{M\times L} \leftarrow Gaussian(mean = 0, std = 0.01)$;
$B^{N\times L} \leftarrow Gaussian(mean = 0, std = 0.01)$;
while $|\epsilon - \epsilon^*| > 1e-3$ **do**
  $\epsilon \leftarrow \epsilon^*$
  while *True* **do**
    $X' \leftarrow Reconstruct(A, B)$;
    $G_A^{M\times L}, G_B^{N\times L} \leftarrow ComputeGradient(X, A, B, \epsilon)$;
    $A^*, B^* \leftarrow RPROP(A, B, G_A, G_B)$;
    if $X' == Reconstruct(A^*, B^*)$ **then** break
    $A \leftarrow clip(A^*)$;
    $B \leftarrow clip(B^*)$;
  end
  $\epsilon^* \leftarrow Diff(X, Reconstruct(A, B))$;
end
return $\sigma(A), \sigma(B), \epsilon$

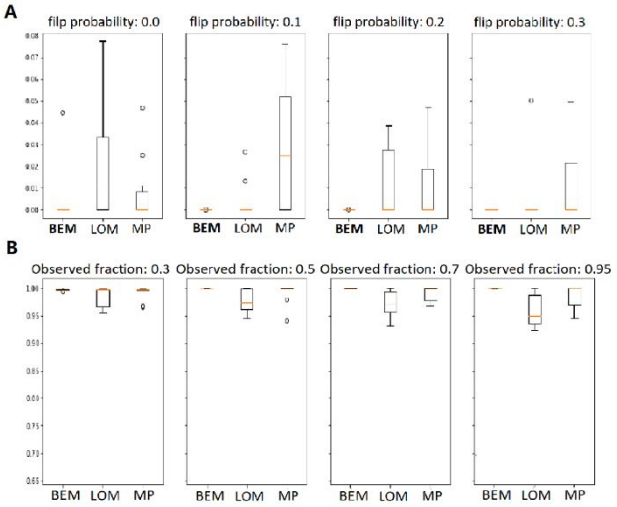


Fig. 3. Results of synthetic experiments. (A) is the reconstruction error (8% max) of synthetic data in 3.1.1 when Bernoulli priors varied. Synthetic matrices were 1000×1000 with rank 5. BEM (Left) is the algorithm proposed in this study; MP (right) is short for message passing; LOM (middle) is the Logical factorization machine. (B) is the correctly inferred fraction on synthetic data, Y axis ranged from 1.00 to 0.65.

3 Results

Our algorithm was compared with the message passing approach (Ravanbakhsh *et al.*, 2016) and the full Bayesian approach (Rukat *et al.*, 2017), referred to as LoM/OrM below. The Bernoulli prior for the two algorithms were estimated using empirical Bayes approach described in (Rukat *et al.*, 2017). During synthetic experiments, we evaluated the three algorithms on two tasks: noisy matrix factorization and noisy matrix completion. In real data experiment, the three algorithms were compared by the subtype classification accuracy on RNAseq datasets from TCGA (Tomczak *et al.*, 2015). Finally, we demonstrated our algorithm's real-world application to two datasets generated by scRNAseq and in situ hybridization (ISH) respectively.

3.1 Synthetic Experiment

The observed matrices with noise, X*, were synthesized based on the same sampling scheme as our probabilistic problem formulation, except that each scalar value in latent factors were samples from a uniform distribution on interval [P-0.3, P+0.3]. P was determined by the preset matrix density Pr(X=1):

$$\Pr(X=1) = 1 - (1-P^2)^L$$

3.1.1 Matrix factorization

We evaluated the three algorithms on five different noise levels (flip probability): 0.0, 0.1, 0.2, 0.3. The sampling scheme was repeated 10 times for each noise level. The performance was measured by the reconstruction error rates, which is comparing the reconstructed matrix with the synthesized matrix without noise:

$$err = |\hat{X} - X|/(NM)$$

As shown in Fig. 3A, although EM algorithm is likely to reach a local optimum, the performance of BEM is more stable across different noise levels compared with the other probabilistic approaches. BEM has achieved zero error in 9 out of 10 synthetic datasets with lower noise levels (flip probability $\leq$ 0.3), while the other two can only perfectly reconstruct the noiseless matrix in 6 to 9 synthetic samples. However, when the flip probability is above 0.3, LoM performed slightly better than message passing and BEM. Such comparison results remained the same when matrix density was 0.3 and 0.7 (see supplement Fig 1 & 2).

When tested against various matrix size and Boolean ranks, the degree of freedom versus sample size, $(N+M)L/(NM)$, is important for the relative performance of BEM. As shown in supplement Fig 3, when Boolean rank was increased from 5 to 10, LoM achieved the best performance across different noise levels. However, when matrix sizes increased from 1000 to 2500 (Supp Fig 4), LoM's performance has a much greater variance than message passing and BEM.

3.1.2 Matrix completion

We evaluated the three methods with various observed fraction (i.e. 30%, 50%, 70%, 95%). Synthetic data was generated with the same sampling scheme as above. The noise was set at 20%. All the algorithms were evaluated by the fraction of correctly inferred values. As shown in Fig. 3B, LoM accurately inferred 5% more of the missing data when the fraction of observed data is less than 30%.

In summary, BEM outperformed other Boolean matrix factorization method when noise level is less than or equal to 30% and the size of the matrix is sufficiently large relative to the number of latent factors Since most gene expression datasets satisfy these conditions, BEM is more applicable to transcriptomic data than other Boolean matrix factorization methods.

3.2 Real data experiments

3.2.1 Classification of breast cancer subtypes

We downloaded transcriptomic profiles of breast cancer patients from TCGA (Tomczak *et al.*, 2015). The dataset contained 1095 tumor samples and 114 normal samples. Each sample is a transcriptomic profile of 19665 genes Expression values in the matrix was log(count+1), where the count was TPM values. Assuming that expression of normal genes complied with Normal distribution, we can calculate the Z score for each gene in the tumor sample by:

$$Z_i = (E_i - \mu_i) / \sigma_i$$

where μ_i is the mean of gene i in normal samples and σ_i is the standard deviation of gene i in normal samples. Each individual expression, x, was determined as differential expression (1) or not (0) with two criteria: (1) absolute fold change between x and the normal mean was larger than 2; (2) the Z score of x was larger than 1.645, corresponding to the one-sided normal p value of 0.05.

From this binary matrix, 15 factors were extracted with our algorithm and others for comparison. The number of factors was determined by AIC. To examine the effectiveness of BEM, we investigated the proportion of patient subtypes in the factors on the samples' side.

To compare the performance of these factorization methods, we used factors about the samples (or meta-samples) as features for tumor subtype classification. Non-negative matrix factorization (NMF) was also included for comparison. Classification was conducted with Multinomial logistic regression. Both NMF and logistic regression come from a python package named "scikit-learn". Performance was evaluated with leave-one-out cross validation. As shown in Table. 1, our algorithm has achieved the highest classification performance among algorithms for Boolean matrix factorization.

Table 1. Breast cancer subtype classification accuracy

Matrix Factorization	Accuracy (%)
BEM	**81.3**
MP	77.7
LoM	77.8
NMF	50

Table 2. Accuracy for each subtype with 15 factors

Subtype	Normal	LumA	LumB	Her2	Basal
# of Samples	23	47	190	64	140
LoM	0.0	81.1	**74.7**	48.4	**98.5**
MP	0.0	**91.1**	53.7	53.1	94.3
BEM	**34.8**	88.7	64.7	**65.6**	96.4

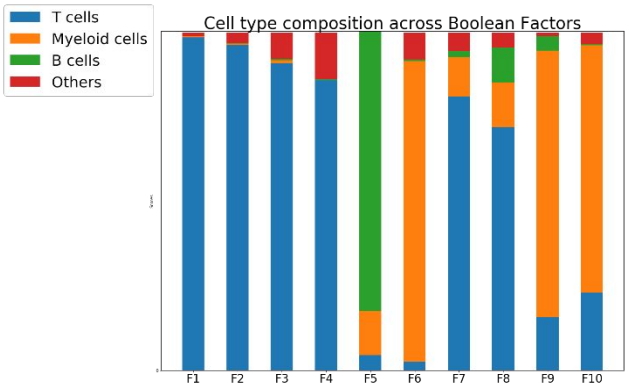


Fig. 4. Each column shows the proportions of each cell type in one factor.

We further investigated classification accuracy with other Boolean matrix factorization methods in each tumor subtype. As shown in Table. 2, all the Boolean matrix factorization methods achieved high accuracy in

the subtype of LumA and Basal. It indicates the genes expression data and the subsequent differential expression analysis has provided abundant discriminative information about the two subtypes. However, LoM and Message Passing are not able to discriminate Her2 and Normal-like tumors effectively while BEM is somewhat capable of. This result showed that by getting rid of assumptions about factors' sizes, BEM is more likely to capture subtle patterns that have greater variance on factor sizes.

3.2.2 Cell type deconvolution from single cell RNAseq

The single cell RNAseq data about melanoma patients was collected from Gene Expression Omnibus (GSE120575). This dataset contained 55737 genes on 16291 cells across 48 patient samples. On average 339 cells were measured in each sample. 19 samples were measured before therapy and were used to predict therapy responsiveness. The rest were measured after therapy.

The expression values were encoded as 1 if the gene has nonzero expression values, otherwise 0. Genes that expressed in less than 1% of the cells or over 99% of the cells were removed. Only 10474 genes remained for matrix factorization analysis. We chose 10 factors based on the Akaike information criteria (AIC), which is close to the choice of 11 clusters in the original study. The R package, "ccfindR"(Woo *et al.*, 2019), was also used to perform non-negative matrix factorization (NMF) and its variation (vbNMF). Model likelihood for vbNMF did not change significantly as the number of factors increased from 9 to 12. Thus, we still used 10 factors for both NMF approaches. We constructed the gold standard for major cell types from the marker gene sets provided in the original study (Sade-Feldman *et al.*, 2018). In addition, due to high overlap in the gold standard, CD4 T cells and CD8 T cells were merged into T cells (87.8% of CD8 T cells were also CD4 T cells); cDCs dendritic cells, pDCs, macrophage, neutrophils, and myeloid were merged as myeloid cells. (Over 50% of each of these cell types were also classified as myeloid cells). 3764 cells that cannot be classified by the gold standard marker genes were excluded in this step. Note that the excluded cells were still used in matrix factorization analysis and subsequent responsiveness prediction. Cell types with small cell counts and little overlap with each other were simply denoted as "Others" in Fig. 4.

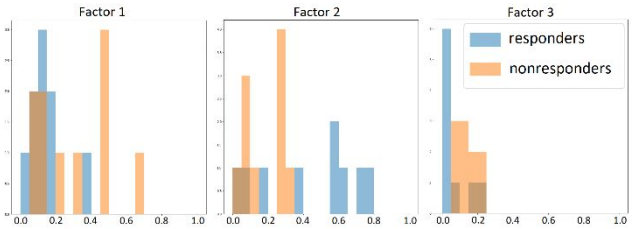


Fig. 5. The distribution of responders (blue) and nonresponders (brown) over aggregated Boolean factor values. Bins in deep brown is the overlapping proportions. Nonresponders tend to have higher aggregated values in factor 1 and factor 3, while responders have higher values in factor 2. However, this is not statistically significant due to limited sample size (19 samples.)}

The cell-side factors were dichotomized with 0.5 as cutoff. After dichotomization, if the factor value of the ith cell in the jth factor was 1, then the ith cluster contained the ith cell. Clearly, the clusters were not mutually exclusive. As shown in Fig. 4, the first four factors corresponded to T cells exclusively (from 98.7% to 85.9%). 94.2% of T cells belonged to at least one of these four factors. The fifth factor corresponded to myeloid cells mostly (97.1%). Also 77.6% of myeloid cells belonged to this factor. The sixth factor corresponded to B cells mostly (82.9%). And 97.7% of B cells belonged to this factor. The other four factors were mixtures of various cell types, which might capture cellular functions across cell types. It showed that our algorithm was able to identify high levels of expression patterns accurately.

We also aggregated the factors on the cell side into sample level features by taking average of all the factor values in cells belonging to the same sample. (Note that even cells that cannot be classified by markers were included). Aggregation was performed with both our Booelan factors and the NMF factors. These aggregated values for 19 samples before therapy were used as features in logistic regression to predict responsiveness of patients. The target variable was binary, either responsive, or nonresponsive. The accuracy was evaluated with leave-one-out cross validation. As shown in Table. 3, using the 10 features from our algorithm was significantly better than the original cell type information. our algorithm had probably extracted information related to therapy responsiveness beyond merely cell types.

We further performed transcription factor (TF) enrichment analysis and gene ontology (GO) enrichment analysis on factors on the gene side to investigate therapy related gene regulation mechanism. More specifically, we analyzed the four factors that consists of T cells. Listed in Table. 4&5, were the top TFs and GOs significantly enriched (P<0.001). Factor 1 and factor 3 were similar as they shared one transcription factor (EZH1) and one Biological process (cytokine-related). Research (Abdalkader *et al.*, 2016) suggested that the absence of EZH1 was important in controlling proliferation / resting of lymphoid cells. The disruption of EZH1 / EZH2 ratio signified abnormal immune cell state. Factor 1 might represent T cells in response to INF-gamma, the viral response. We further analyzed genes uniquely activate in factor 3 to distinguish the two. It seemed factor 3 may capture expression patterns of T cells in hypoxia. Actually, patients with high aggregated values in these two factors tended to be nonresponders. We suspected that the presence of imbalanced EZH1/EZH2 ratio and hypoxia have negative effects on patient responsiveness. Factor 2 was characterized by IL-21 response, which activates T cell. Patient with high Factor 2 values tended to be responders. This was consistent with current knowledge ((Santegoets *et al.*, 2013). Two enriched TFs in factor 2, IKZF4 and FOXP3, often collaborated in immunosuppressive activities (Jia *et al.*, 2019). Since the two TFs were enriched for knockout experiment, the expression pattern of factor 2 indicated the immunostimulatory state of immune cells. The enrichment analysis, combined with Fig. 5, showed that the three T cell dominant factors described above had captured distinct expression patterns that could discriminate immunotherapy responsiveness.

Table 3. Prediction accuracy of immunotherapy responsiveness

Features	Accuracy (%)
Gold standard cell type	42.1
Aggregated NMF factors	47.4
Aggregated vbNMF factors	47.4
Aggregated Boolean factors (continuous)	63.2
Aggregated Boolean factors (0.5 cutoff)	78.9

Table 4. TF Enrichment analysis of single cell gene factors

Genes' factors	Size of enriched genes	Enriched TFs
1	**407**	EZH1; FOXP3
2	260	IKZF4; XBP1; FOXP3

3	834	EZH1; NKX25; MEIS2
4	316	ZBTB7B; STAT1; XBP1
3(unique)	492	NKX25

Table 5. GO Enrichment analysis of single cell gene factors

Factors	Enriched GO
1	cellular response to interferon-gamma (GO:0071346) cytokine-mediated signaling pathway (GO:0019221)
2	T cell activation (GO:0042110) interleukin-21-mediated signaling pathway (GO:0038114)
3	T cell receptor signaling pathway (GO:0050852) cytokine-mediated signaling pathway (GO:0019221)
4	transcription regulation in response to hypoxia (GO:0061418) neutrophil degranulation (GO:0043312)
3(unique)	T cell activation (GO:0042110) cytokine-mediated signaling pathway (GO:0019221)

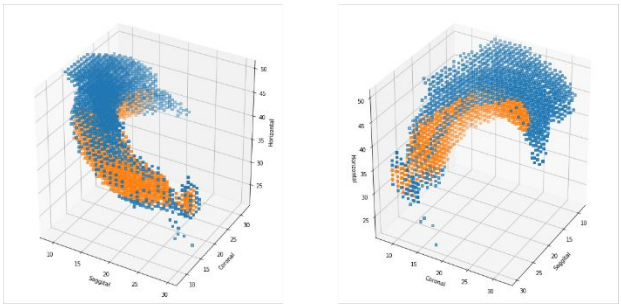


Fig. 6. 2-factorization of spatial transcriptomics in mouse hippocampal formation.

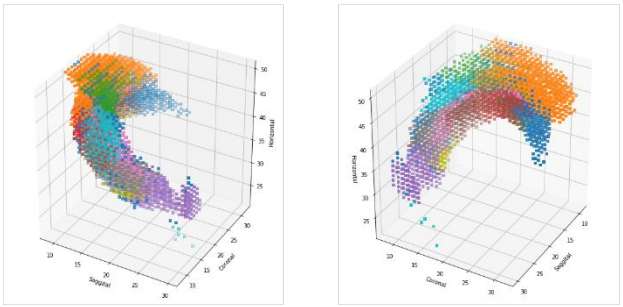


Fig. 7. 10-factorization of spatial transcriptomics in mouse hippocampal formation.

3.2.3 Segmentation of Spatial Transcriptomics

Spatial transcriptomic data about hippocampal formation in adult mouse brain was downloaded from Allen Brain Atlas (Lein *et al.*, 2007). Our selected region had 2510 voxels. Each voxel contained an expression profile of 19908 genes. Gene expression values were measured with in situ hybridization (ISH) technology. As shown in supplement Fig. 5, the number of non-expressed genes was consistent within the same Sagittal section. Thus, we believed that most of non-expressed genes are actually missing values and masked them as is. Sagittal sections with less than 3000 expressed genes were excluded from the analysis. Nonzero expressions were dichotomized based on individual average of each gene. Clearly, this dataset contained both missing values and noisy measurements, which was suitable to test our algorithm's performance.

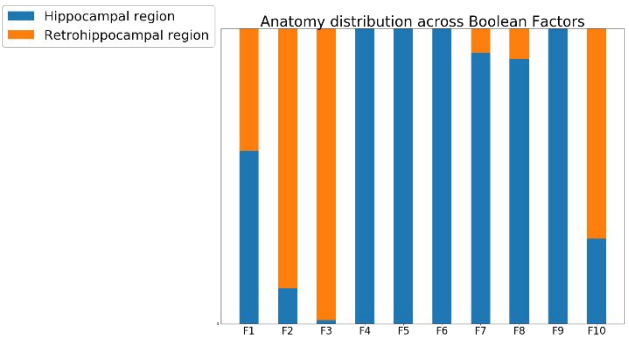


Fig. 8. 10-factorization of spatial transcriptomics mapped to two high level anatomical labels

Several different sizes of latent factors were attempted, including 2, 5, 10, and 15 factors. As shown in Fig. 6&7, our algorithm produced spatially tight clusters without the aids of spatial information. We also tried 5 factors, 15 factors and 30 factors, the voxels assigned to each factor were still close together (see Supp Fig 6-10).

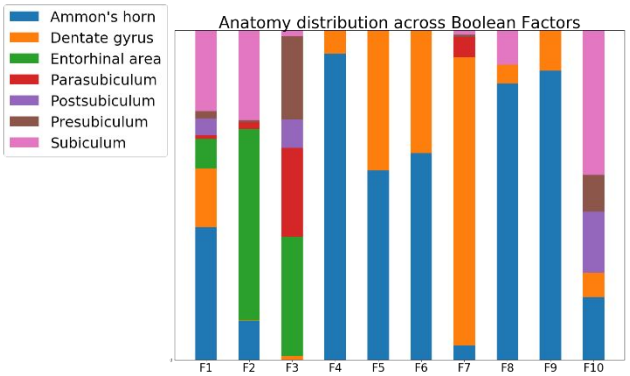


Fig. 9. 10-factorization of spatial transcriptomics mapped to seven low level anatomical labels

We also investigated the alignment of our voxel factors with the anatomical labels. On a high-level anatomical category (Fig. 8), most factors were dominant in either hippocampal region or retro hippocampal region, except factor 1 and factor 10. These two factors might represent the area bridging the two parts or expression patterns unrelated to anatomy labels. With finer granularity (Fig. 9) factor 1 and factor 10 were indeed a mixture of different areas. Other factors seemed to be somewhat aligned with anatomical structure. Only two out of seven labels were dominant in most of those factors.

4 Discussion

In this study, we presented BEM, a new Boolean matrix factorization (BMF) algorithm free of assumptions about Boolean factors' sizes. Synthetic results showed that our algorithm outperformed recent BMF algorithms (Rukat *et al.*, 2017; Ravanbakhsh *et al.*, 2016). When looking into classification accuracy in each breast cancer subtype (Table. 2), BEM is able to detect some signals from minor groups including normal-like subtype and HER2, where LoM and Message passing ignored. This showed that by removing the assumption of factor shape, BEM can capture coregulation patterns even with the issue of class imbalance.

When comparing with NMF, BEM performed much better in terms of accuracy in both cancer subtype classification and immunotherapy

responsiveness prediction. Although NMF has been widely used in transcriptomic data analysis and yielded fruitful insights (Zhu *et al.*, 2017; Woo *et al.*, 2019; Noto *et al.*, 2017), our results supported that the Boolean mechanism illustrated in Fig. 1 is more aligned with the biological mechanisms in transcriptomics than the linear relationship assumed in NMF. However, given the vast research efforts in NMF, there are many techniques that Boolean factorization can borrow and adapt. For example, ccfindR(Woo *et al.*, 2019) has adopted the Bayes factor approach to select the number of latent factors. Similarly, BEM also incorporated the AIC metric to assist users in model selection. However, we have not yet established an effective way to rank genes like the D-score proposed in NMFEM (Zhu *et al.*, 2017). Since a gene-side factor often has thousands of genes activated. Future research need to investigate how to select discriminative genes from latent Boolean factors.

When analyzing gene expression data with BEM, binarization is the most important preprocessing step. In the real data experiment, we demonstrated three different binarization strategies for different types of transcriptomic data. We recommend future users to consider three aspects when performing binarization: (1) the semantics of binary values. For example, in TCGA breast cancer analysis, binary values indicated the status of differential expression. But in scRNA-seq, binary values corresponded to expression and non-expression; (2) the presence of missing values. For example, we determined that the sparsity in spatial transcriptomics were mostly missing values and masked all as missing instead of actually being 0; (3) if users are certain that the data is multimodal by nature, then one-hot encoding should be adopted instead. For example, a gene G has states A, B, and C, then it should be encoded by 3 Boolean factors. If the first factor has 1, then G in the corresponding sample is in state A. However, the order among the modes would be lost. Therefore, when the data is ordinal, users need to decide to sacrifice the order information or multimodal information, otherwise BEM is not applicable.

Another potential issue in preprocessing is the integration of different datasets. Although BEM is not tested in this scenario, integrated datasets may result in mixed semantics of binary values. One way to deal with this is to regress out sample covariates and batch effects. The other is to perform binarization on each dataset separately.

In addition, users should be aware of two practical details. One is hyperparameter tuning. The hyperparameters in this algorithm, the Beta prior, should be set to 0.95 when binary factor values are preferred. If users need to estimate the uncertainty of the output, the Beta prior should be set to 1, hence the probabilistic estimates returned by the algorithm is not biased towards 0 or 1. The other is the step of noise estimation, since we assume that noise was symmetric. That is, the probability of 1 flipped to 0 is the same as that of 0 flipped to 1. When the error mechanism is known to be asymmetric, users have two ways to deal with such issues: (1) mask certain values to be missing values. Take scRNA-seq for instance, if we know that certain genes are prone to drop-out, then we may mask all its nonexpressed values as zeros; (2) estimate the probability with known binarization errors or prior knowledge. For example, if both single cell transcriptomic and proteomic profiles are available for the same condition, users can examine the dropout ratio of scRNA-seq with proteomic profiles about the same gene. Future research could further look into ways to alleviate such assumptions.

Our algorithm is applicable to various high-throughput data as long as the tasks of latent variable inference can be represented as dense bipartite subgraph problem or the tiling problem. We applied BEM to three transcriptomic datasets generated with bulk RNAseq, single cell RNAseq, and ISH respectively. Given appropriate dichotomization, results in Bulk RNAseq and single cell RNAseq showed that BEM is able to extract information related to expression patterns such as disease subtypes and cell types. Results in ISH expression data also indicated that our algorithm can extract neuron expression patterns without the aids of spatial information.

Funding

This work has been partially supported by awards R00LM011673 and U54HG008540 from NIH. The project used the Hillman Cancer Bioinformatics Services and the UPMC Hillman Cancer Center Developmental Funding that are supported in part by award P30CA047904 from the NCI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Abdalkader,L. *et al.* (2016) Aberrant differential expression of EZH1 and EZH2 in Polycomb repressive complex 2 among B- and T/NK-cell neoplasms. *Pathology*, **48**, 467–482.

Berglund,E. *et al.* (2018) Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.*, **9**, 2419.

Bergmann,S. *et al.* (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E, Stat. Nonlin. Soft. Matter. Phys.*, **67**, 031902.

Brunet,J.-P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA*, **101**, 4164–4169.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Hochreiter,S. *et al.* (2010) FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, **26**, 1520–1527.

Jia,H. *et al.* (2019) The expression of FOXP3 and its role in human cancers. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Li,G. *et al.* (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.*, **37**, e101.

Neumann,S. (2018) Bipartite Stochastic Block Models with Tiny Clusters. In, *Advances in Neural Information Processing Systems*., pp. 3867–3877.

Noto,T. *et al.* (2017) Genome-scale investigation of olfactory system spatial heterogeneity. *PLoS One*, **12**, e0178087.

Padilha,V.A. and Campello,R.J.G.B. (2017) A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, **18**, 55.

Patel,A.P. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.

Paul,F. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.

Ravanbakhsh,S. *et al.* (2016) Boolean Matrix Factorization and Noisy Completion via Message Passing. In, *ICML*., pp. 945–954.

Rukat,T. *et al.* (2017) Bayesian Boolean matrix factorisation. In, *Proceedings of the 34th International Conference on Machine Learning-Volume 70*., pp. 2969–2978.

Sade-Feldman,M. *et al.* (2018) Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell*, **175**, 998–1013.e20.

Saelens,W. *et al.* (2018) A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.*, **9**, 1090.

Santegoets,S.J. *et al.* (2013) IL-21 in cancer immunotherapy: At the right place at the right time. *Oncoimmunology*, **2**, e24522.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, **98**, 10869–10874.

Stein-O'Brien,G.L. *et al.* (2018) Enter the Matrix: Factorization Uncovers Knowledge from Omics. *Trends Genet.*, **34**, 790–805.

Tai,Y. *et al.* (2018) Gene co-expression network analysis reveals coordinated regulation of three characteristic secondary biosynthetic pathways in tea plant (Camellia sinensis). *BMC Genomics*, **19**, 616.

Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18 Suppl 1**, S136–44.

Tomczak,K. *et al.* (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, **19**, A68–77.

van Dam,S. *et al.* (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformatics*, **19**, 575–592.

Woo,J. *et al.* (2019) De novo prediction of cell-type complexity in single-cell RNA-seq and tumor microenvironments. *Life Sci. Alliance*, **2**.

Xie,J. *et al.* (2019) It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Brief. Bioinformatics*, **20**, 1449–1464.

Zhang,Z.-Y. *et al.* (2010) Binary matrix factorization for analyzing gene expression data. *Data Min Knowl Discov*, **20**, 28–52.

Zhu,X. *et al.* (2017) Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ*, **5**, e2888.