BAYESIAN BICLUSTERING FOR PATIENT STRATIFICATION

SAHAND KHAKABIMAMAGHANI and MARTIN ESTER

School of Computing Science, Simon Fraser University, 8888 University Drive
Burnaby, BC, V5A 1S6, Canada
E-mail: {*sahandk, ester*}*@sfu.ca*

The move from Empirical Medicine towards Personalized Medicine has attracted attention to Stratified Medicine (SM). Some methods are provided in the literature for patient stratification, which is the central task of SM, however, there are still significant open issues. First, it is still unclear if integrating different datatypes will help in detecting disease subtypes more accurately, and, if not, which datatype(s) are most useful for this task. Second, it is not clear how we can compare different methods of patient stratification. Third, as most of the proposed stratification methods are deterministic, there is a need for investigating the potential benefits of applying probabilistic methods. To address these issues, we introduce a novel integrative Bayesian biclustering method, called B2PS, for patient stratification and propose methods for evaluating the results. Our experimental results demonstrate the superiority of B2PS over a popular state-of-the-art method and the benefits of Bayesian approaches. Our results agree with the intuition that transcriptomic data forms a better basis for patient stratification than genomic data.

1. Introduction

In Empirical Medicine every patient of a particular disease receives the same treatment. However, although working for simpler diseases to a degree, this approach has not been successful for more complex diseases like cancer. Therefore, the paradigm in medicine is shifting from Empirical to so called Personalized Medicine, which is a patient derived approach with the goal of providing individual treatments for each patient according to his/her particular conditions and features. As an intermediate step currently being investigated, "Stratified Medicine is an approach by which groups of patients with the same disease are subdivided into different categories depending on the underlying mechanism of disease and their probable response to a therapeutic intervention [1]."

According to the definition of stratified medicine, a cohort of patients is divided into subgroups, called subtypes, and the specific features of each subtype that constitute the disease mechanism for that subtype are identified. These features will then be used to design subtype-specific treatments. One possible approach to patient stratification is Biclustering, which is proven useful for this task [2] and is commonly in use for it. A comprehensive discussion of bi-clustering methods can be found in [3]. Most of the biclustering algorithms proposed in the literature utilize an optimization method to find the solution. They can be categorized into two main classes:

1. *Deterministic:* Examples are Singular Value Decomposition (SVD) [4] and Non-negative Matrix Factorization (NMF) [5], which try to optimize the value of latent variables indicating the clustering structure. Although these methods initialize the latent factors randomly, given the same initial random parameters, they will always produce the same final result.
2. *Probabilistic:* this family of methods models the data as a Bayesian network of variables with cluster ids being a latent variable. Examples are Plaid [6] and SAMBA [7]. These methods

also use random initialization; however, since they use stochastic optimization, they might produce different solutions in different executions given the same initial values.

Methods in the second group usually return a probabilistic assignment of objects to clusters. This is more desirable for patient stratification, because first, it provides a model-based (rather than ad-hoc) approach to predict subtypes for new patients with unknown subtypes, and second, patients in one subtype often share features with patients in other subtypes and probabilistic assignments to subtypes capture these similarities and are more informative than strict assignments [8]. Furthermore, stochastic optimization methods are less prone to get stuck in local optimums. In addition, probabilistic models allow for introduction of prior knowledge into model.

In terms of the diversity of data types used as stratification input, methods can be categorized into single-input and integrative. Hofree et al. [5] and Hochreiter et al. [9] are examples of single-input approaches. They, respectively, use somatic point mutation and gene expression data. While some (but not all) of these publications provide comparisons between their methods and existing methods, these comparisons were conducted using either synthetic data or real databases with clinically known subtypes and, as also discussed in [2] and to the best of our knowledge, no suitable metric is provided for benchmarking when the data are real and unlabeled.

Some single-input stratification methods use a different approach by finding the subtypes based on only a single data type, fixing the detected subtypes, and then integrating other data types to investigate subtype-specific features in those datasets. Examples are two prominent references Verhaak et al. [10] and Cho and Przytycka [8], both of which used gene expression data as the main datatype for patient stratification, but they did not discuss the logical reasons for this choice.

As an example of integrative methods, Shen et al. [11] proposed a Bayesian method, namely iCluster, for integrative clustering of genomic data and applied it to breast and lung cancer data. In another study, Sun et al. [4] proposed a multi-view SVD method and applied it for integrating genomic and clinical data to find disease subtypes and their associated genetic variations. We note that these publications do not compare with competitors and do not demonstrate the merits of the integrative approach compared to single-input patient stratification through benchmarking experiments. Although Sun et al. [4] used AUC scores for discussing this point, we believe that their results are not an indicator of superiority of the integrative method, but are the natural result of their experimental setup. Furthermore, they only examine combining clinical and point mutation data and do not consider other genomic, transcriptomic, or proteomic data types.

Table 1 summarizes the mentioned approaches to patient stratification and compares them according to the discussed aspects. According to our discussion, the merit of integrating different datasets for patient stratification is still an open issue. Furthermore, no systematic methods and metrics have been presented in the literature for evaluating patient clustering results and efforts have been focused rather on gene clustering (as in Prelic et. al [2]). Moreover, as also seen in Table 1, the utility of probabilistic methods in patient stratification is overlooked, although they are frequently applied for gene clustering. As discussed earlier, these methods provide potential solutions for the problems in patient stratification.

In this paper, we address these open issues by proposing a novel Probabilistic Graphical Model (PGM), which we call B2PS (Bayesian Biclustering for Patient Stratification), and appropriate

evaluation metrics. To the best of our knowledge, the model provided here is the first Integrative Bayesian Biclustering model. While there are solutions for Integrative Biclustering [12] as well as Bayesian Biclustering [13] in the literature, no work so far combines integrative, Bayesian, and Biclustering concepts in one model.

Table 1. Existing and proposed methods

Method	Probabilistic or Deterministic	Clustering/ Biclustering	Stratification Input Datatypes
Verhaak et al. (2010) [10]	Deterministic (HC)	Clustering	Expression
Hochreiter et al. (2010) [9]	Deterministic (FA)	Biclustering	Expression
Hofree et al. (2013) [5]	Deterministic (NMF)	Biclustering	Mutation
Shen et al. (2009, 2012) [11, 14]	Deterministic (FA)	Clustering	Multiple
Sun et al. (2014) [4]	Deterministic (SVD)	Clustering	Multiple
Cho & Przytycka (2013) [8]	Probabilistic (PGM)	Clustering	Multiple
B2PS	Probabilistic (PGM)	Biclustering	Multiple

Abbreviations used in this table: HC (Hierarchical Clustering) – FA (Factor Analysis)

The main contributions of this paper are as follows:

- The proposed model allows for **incorporation of prior knowledge**, which is useful for dealing with noisy data. Our experimental results show that this ability is useful for processing noisy biological data and improves the stratification performance.
- The proposed method is able to **detect the natural number of clusters** for each dimension (i.e., row and column), identification of which requires an iterative trial process in deterministic methods. Measured evaluation metrics indicates that the natural sample clusters detected by our method form a better partitioning than the one detected by conventional NMF.
- Unlike conventional bi-clustering methods, **the number of row and column clusters is not assumed to be the same** in our model. This is a useful assumption that is more consistent with typical biological datasets and, according to our experimental results, provides a more informative clustering across both dimensions.
- The integrative method proposed here allows for examination of patient stratification results when using **different combinations of diverse datatypes** with no theoretical limitation on the number of data types. This makes it possible to identify the datatypes that are more useful for patient stratification. Experimental results with two TCGA datasets suggest that gene expression data is more informative than genomic data for patient stratification.

We compare the performance of B2PS against NMF, a state-of-the-art deterministic method. Experimental results demonstrate the superiority of B2PS over NMF regarding both patient stratification and feature clustering in different experimental settings. We believe that the outputs

of the proposed method can be a useful basis for detecting the subtype-specific driver aberrations, which is one of the goals of stratified and personalized medicine.

2. Methods

2.1. Model

To perform patient stratification using different datatypes, an integrative probabilistic graphical model for biclustering is proposed. The model is shown in Figure 1. Observed variables are shaded and hyperparameters are in dotted circles. Table 2 includes a detailed description of the variables of the model.

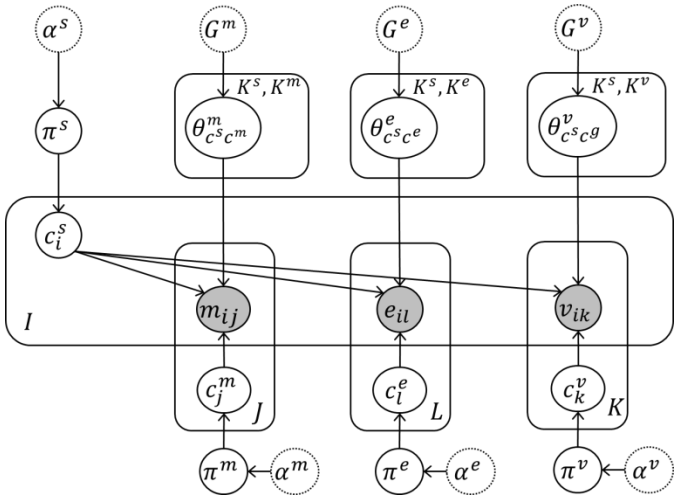

Fig. 1. The probabilistic graphical model of B2PS

Because the goal is to integrate different datatypes about the same set of patients/samples, in our model, datasets of different datatypes are assumed to have the same rows/samples but can have different columns/features. Accordingly, the row clustering is shared across different datatypes, but each dataset has its particular column clustering. However, column clusterings of different datatypes are indirectly related to each other through the shared row clustering. While, no direct dependency is assumed between sample clusters c_i^s and gene clusters c_l^e,c_j^m, and c_k^v in this model, they are indirectly dependent given the observed data variables. In terms of clustering structures discussed in [13], B2PS produces a single non-overlapping clustering, meaning that each row/column belongs to a single cluster that has no overlap with other clusters.

2.2. Parameter Learning and Inference

The Gibbs sampling method [15] is used for parameter learning and latent variable inference. After random initialization, the latent variables (see Table 2) are iteratively sampled one by one based on computed marginal conditional probabilities. Eq. 1 shows the conditional probability for sample/row clusters. Parameters π^m, π^e, and π^v and hyperparameters α^m, α^e, and α^v are not

included in this equation for they are conditionally independent from c_i^s given c^m, c^e, and c^v (refer to the model in Figure 1). Other absent parameters are integrated out.

Table 2: Parameters and variables included in B2PS

Type	Name	Description	Distribution
Observed Variables	e_{il}	Expression status of gene l of sample i	$e_{il} \sim \text{Multinomial}_3\left(\theta^e_{c_i^s c_j^g}\right)$
	m_{ij}	Mutation status of gene j of sample i	$m_{ij} \sim \text{Bernoulli}\left(\theta^m_{c_i^s c_j^g}\right)$
	v_{ik}	Copy number variation of gene k of sample i	$v_{ik} \sim \text{Multinomial}_5\left(\theta^v_{c_i^s c_j^g}\right)$
Hyperparameters	α^s	The parameter of prior Dirichlet distribution for samples. K^s is the number of sample clusters. K^s and p^s are provided as input.	$\alpha^s = [\frac{p^s}{K^s} \dots \frac{p^s}{K^s}]_{1\times K^s}$
	α^x	The parameter of prior Dirichlet distribution for fetures of data type x. K^x is the number of feature clusters. K^x and p^x are provided as input.	$\alpha^x = [\frac{p^x}{K^x} \dots \frac{p^x}{K^x}]_{1\times K^x}$
	G^x	The parameters for prior distributions of $\theta^x_{c^s c^x}$ for data type x. β values are provided as input.	$G^m = \{\beta_0^m, \beta_1^m\}$ $G^v = \{\beta_{-2}^v, \beta_{-1}^v, \beta_0^v, \beta_1^v, \beta_2^v\}$ $G^e = \{\beta_{-1}^e, \beta_0^e, \beta_1^e\}$
Model Parameters	π^s	Distribution of the probability of belonging to different sample clusters	$\pi^s \sim \text{Dirichlet}_{K^s}(\alpha^s)$
	π^x	Distribution of the probability of belonging to different feature clusters for data type x	$\pi^x \sim \text{Dirichlet}_{K^x}(\alpha^x)$
	$\theta^x_{c^s c^x}$	Parameters for distribution of the values of the entities belonging to bicluster (c^s, c^x) datatype x	$\theta^m_{c^s c^\mu} \sim \text{Beta}(G^m)$ $\theta^v_{c^s c^v} \sim \text{Dirichlet}_5(G^v)$ $\theta^e_{c^s c^e} \sim \text{Dirichlet}_3(G^e)$
Latent Variables	c_i^s	Cluster id for ith sample (sampled variable)	$c_i^s \sim \text{Multinomial}_{K^s}(\pi^s)$
	c_l^e,c_j^m, c_k^v	Cluster id for lth, jth, and kth gene in corresponding datasets (sampled variable)	$c_r^x \sim \text{Multinomial}_{K^x}(\pi^x)$

In the above table, x can be m (point mutation), e (gene expression), or v (copy number variation).

All variables used in Eq. 1 and Eq. 2 (below) are described in Table 3. The right side of Eq. 1 has generally two terms; the first term accounts for the size of clusters (i.e., larger clusters are assigned greater probability) and the second term incorporates the similarity of row i to the members of each cluster (i.e., giving higher probability for assigning row i to clusters with more similar members). Values of the hyperparameters control the balance between these two terms. Feature clusters for different data types are sampled similarly. As an example, the Eq. 2 is the conditional probability of feature clusters according to gene expression data.

$$P(c_i^s = q | c_{-i}^s, c^m, c^e, c^v, m, e, v; \alpha^s, G^m, G^v, G^e)$$
$$\propto P(c_i^s = q, c_{-i}^s, c^m, c^e, c^v, m, e, v; \alpha^s, G^m, G^v, G^e)$$
$$\propto \frac{ns_q^{-i} + \alpha_q^s}{ns^{-i} + p^s} \times \prod_{x \in \{m,e,v\}} \left[\prod_{t=1}^{K^x} \prod_{\{r | c_r^x = t\}} \left(\frac{nx_{qt}^{x_{ir},-i} + \beta_{x_{ir}}^x}{nx_{qt}^{-i} + \beta_x} \right) \right]^{D_x} \quad (1)$$

$$P\left(c_j^e = q \middle| c_{-j}^e, c^s, e; \alpha^e, G^e\right) \propto \frac{ne_q^{-j} + \alpha_q^e}{ne^{-j} + p^e} \times \prod_{t=1}^{K^s} \prod_{\{i | c_i^s = p\}} \left(\frac{ne_{tq}^{e_{ij},-j} + \beta_{e_{ij}}^e}{ne_{tq}^{-j} + \beta_e} \right) \quad (2)$$

Table 3: The variables included in sampling conditional probabilities

Variable	Description
c_{-i}^s	Cluster id variables for all samples except ith sample
c_{-j}^e	Cluster id variables for all features of expression datatype except jth feature
ns^{-i}	The total number of samples minus one (the ith sample)
ns_a^{-i}	The number of samples in sample cluster a excluding the ith sample
nx^{-r}	The total number of features in database x minus one (the rth feature)
nx_b^{-r}	The number of features in feature cluster b of dataset x excluding the rth feature
nx_{ab}^{-i}, nx_{ab}^{-r}	The number of elements in bicluster (a,b) in dataset x except those elements related to the ith sample or rth feature, respectively
$nx_{ab}^{x_{ir},-i}$, $nx_{ab}^{x_{ir},-r}$	The number of elements in bicluster (a,b) in dataset x whose value equals x_{ir} except those elements related to the ith sample or rth feature, respectively
β_x	$\beta_x = \sum_d \beta_d^x$, where d is the values that a data point of type x can take (e.g., for point mutation $d \in \{0,1\}$)
D_x	A binary variable indicating inclusion ($D_x = 1$) or exclusion ($D_x = 0$) of data type x in or from the conditional probability, when examining different combinations of datatypes.

In the above table, x can be m (point mutation), e (gene expression), or v (copy number variation).

The number of clusters for samples and genes are denoted respectively by K^s and K^x, where x can be m, e, or v (see Table 2). The random initialization of cluster id variables produces a uniform distribution of entities to these clusters. However, according to the terms included in above conditional probabilities, sampling tends to minimize the number of clusters such that the members of a cluster are highly similar. So, as the biclustering converges throughout the iterations, some clusters become empty with no entities assigned to them, if the values for K^s and K^x are set large enough. Accordingly, after each execution of learning algorithm (until convergence) the natural number of clusters can be determined as the number of occupied clusters.

2.3. Computing Final Clusters and Model Parameters

Due to the stochastic nature of Gibbs sampling, the results of two distinct executions can be different. Therefore, as in [5] and [16], a consensus method based on repeated execution of the learning algorithm is used to yield a more robust clustering. This method is based on a similarity matrix, where the similarity is measured as the number of times (out of several executions) that two entities (samples or genes) belong to the same cluster at the end of an execution. Then, the consensus matrices (one for each dimension) are used to perform UPGMA hierarchical clustering to identify the final sample and gene clusters. The number of clusters used for hierarchical clustering is the average of the number of clusters occupied at the end of different executions. After finding the final clustering structures, the model parameters can be estimated as maximum a posteriori probabilities.

2.4. Comparison Partner

To compare the performance of the proposed probabilistic model with deterministic methods, we use a popular method for patient stratification based on Non-negative Matrix Factorization (NMF). We used the multiplicative NMF algorithm of Lee and Seung [17]. We downloaded the MATLAB implementation by Zhang et al. [12], who modified and used the algorithm for biclustering genomic and transcriptomic data. We amended the code to produce consensus matrices for further post-processing described in section 2.6.

2.5. Evaluation

Between two main categories of internal and external measures used to evaluate clustering results, we used external measures, which are more suitable for assessing the performance of patient or gene clustering algorithms [2]. According to the goal of patient stratification, different patient groups are expected to exhibit distinctive responses to treatments. Therefore, for evaluating the patient clustering results, we use clinical data and perform survival analysis. We use the log-rank test [18] implemented in R 'survival' package. The smaller the log-rank *p*-value, the more distinctive the survival behavior of different patient clusters. This measure is a popular measure for validating stratification results, but, to the best of our knowledge, it has not been used for comparing different clustering algorithms.

Since the main goal of this study is sample stratification, we also measure the stability and robustness of sample clustering outputs regarding the Cophenetic Correlation Coefficient using the method described by Brunet et al. [16]. This is a measure between 0 and 1 and approaches 1 as results of an experiment are more repeatable and robust. Since almost all of the features of the datasets used in our experiments are genes, the Gene Ontology Term Overlap (GOTO) [19] criterion is used for evaluating the feature clustering. Larger values of this metric imply more meaningful clustering in terms of biological relationship between cluster members.

2.6. Parameter Tuning

To determine the best number of clusters for NMF, the method proposed by Brunet et al. [16] is used, which is based on the Cophenetic Correlation Coefficient briefly described in section 2.5. Similar to method described in section 2.3 for B2PS, a consensus matrix is computed throughout

execution of NMF for the same number of times as for B2PS. This experiment is repeated with different numbers of clusters and the Cophenetic Correlation Coefficient is recorded for each experiment. Finally, a chart showing the trend of the Cophenetic Correlation Coefficient versus the increasing number of clusters is drawn and the number after which the coefficient value decreases considerably is chosen as the optimal number of clusters.

The parameters of B2PS are the hyperparameters of prior distributions of values for data points and cluster assignment probabilities. Sample clustering hyperparameter α^s is common among all datatypes, however, feature clustering and data value priors are distinct for different datatypes. Clustering hyperparameters are set uniformly as shown in Table 2 and depend on the values of p^s (for samples) and p^x (for features). For weak or non-informative priors, these values are set to 1 and for strong or informative priors they are set according to the number of samples and features of the dataset being analyzed. Data value prior hyperparameters are set according to their real distribution in the dataset under investigation. When weak, they are scaled such that β values (see Table 2) of the data types being analyzed sum to one. Strong priors are adjusted according to the size of the dataset under analysis.

The optimal values of hyperparameters for each datatype are selected through a trial process that optimizes for log-rank *p*-value. For integrated analysis of several datatypes, the prior settings of individual data types are used. For common hyperparameter α^s, the value used for the datatype producing the best sample clustering in its independent analysis is used.

3. Experiments

3.1. Data

Data for this research are obtained from The Cancer Genome Atlas (TCGA) online dataset [20]. Data include genomic data, namely somatic point mutation and genome-wide copy number variation, and transcriptomic gene expression data. Data are about Glioblastoma Multiform (GBM) and Breast Invasive Carcinoma (BRCA) patients. For each disease, data of a subset of patients/samples having records for all three datatypes mentioned above is downloaded.

To be analyzable with our method, data are preprocessed into three matrices where rows refer to samples and columns refer to features (i.e., genes or miRNAs). According to different properties of the three datatypes, different preprocessing methods are used. Final values are 0 (for genes not containing any non-silent mutation) and 1 (otherwise) for point mutation data, {-2, -1, 0, 1, 2} (the change in the normal number of copies of a gene or miRNA computed by GISTIC2.0 [21]) for CNV, and -1 (under-expression), 0, and +1 (over-expression) for gene expression data (capturing changes more than two fold). Number of features of preprocessed final datasets for somatic point mutation, CNV, and expression data were respectively 4117, 23082, 11874 for 102 GBM samples and 13776, 23082, and 17814 for 501 BRCA samples. Because NMF only accepts non-negative values, for experiments with NMF these data are further preprocessed using the method described in [12]. Clinical data were also available for the patients and contained information required for survival analysis. We retrieved gene ontology data for GOTO analysis using the 'biomaRt' R package [22].

3.2. Results

The experiments are designed with three goals in mind: 1) to show the benefit of the ability to incorporate prior knowledge enabled by the Bayesian approach, 2) to identify the best combination of datatypes for patient stratification, and 3) to compare the proposed method with a state-of-the-art method. In all experiments, the learning algorithm is executed 50 times for both B2PS and NMF. To set the number of iterations for each execution, the learning algorithm is first applied with a large number of iterations, the point of (relative) convergence of the objective function is detected manually, and then the algorithm is run with that number of iterations.

3.2.1. Effects of Priors

To investigate the effects of priors on performance of B2PS, different combinations of strong and weak values for hyperparameters are examined. As an example, the results of a subset of different possible settings for GBM expression dataset are shown in Table 5. Since the main goal of this research was sample stratification, final selected priors (bolded in table) favor better sample clustering over better gene clustering.

According to these and similar results for the BRCA dataset (not reported due to page limit), strong data priors increase the performance regarding the sample clustering with a slight decrease in gene clustering score. This can be explained by the fact that strong priors cancel the noise of gene expression data to a degree, which generally, is expected to increases the sizes of sample and gene clusters. For sample clusters, this effect is somewhat attenuated according to strong patterns in expression profiles of each cluster and the number of clusters remain almost the same. However for gene clusters, this effect merges more similar gene clusters resulting in fewer clusters.

Strong priors for clustering have a reverse effect on clustering structure. As the clustering priors increase, tendency to create clusters with higher similarity among their members increases. So, we should expect smaller and more precise clusters and, consequently, larger number of clusters. Once more, for the same reasons mentioned for data prior, this is more observable for gene clustering rather than sample clustering. Generally the results endorse the usefulness of ability to include prior knowledge in patient stratification.

3.2.2. Informative Datatypes for Patient Stratification

To identify the most informative datatypes for patient stratification we examined different combinations of three datatypes: somatic point mutation, copy number variation and gene expression. Results are summarized in Table 6 for GBM and BRCA datasets. Here, no results are reported for point mutation data, because, due to high heterogeneity of these data, independent experiments with point mutation dataset did not converge to any stable results and, moreover, point mutation data did not have any effects on the output of integrative experiments.

According to the results, gene expression data, when used alone, produces the best result according to both sample clustering (log-rank p-value) and gene clustering (GOTO score). For sample clustering, this can be related to the fact that gene expression profiles are closer to final phenotypes and reflect the cumulative effects of molecular aberrations occurred in earlier steps of central dogma of biology better than other mentioned data types. For gene clustering, higher

GOTO score for expression data compared to others is interpretable according to the fact that genes with similar expression patterns across different samples are more likely to share the same functions in cell than genes with similar CNV.

Table 5: Different prior settings for experiments with GBM gene expression dataset

Priors			Num. of Sample Clusters	Num. of Feature Clusters	Log-rank p-value	GOTO
Data	Sample Clustering	Gene Clustering				
weak	weak	weak	8	66	0.018	3.444
strong	**weak**	**weak**	**8**	**25**	**0.004**	**3.408**
strong	strong	weak	9	21	0.017	3.404
strong	weak	strong	8	73	0.019	3.415
strong	strong	strong	8	70	0.008	3.418

Table 6: Results of integrative and single input experiments for GBM and BRCA

Dataset	Data Types	Sample Clusters	Feature Clusters		Log-rank p-value	Cophenetic Corr. Coef.	GOTO	
			Exp.	CNV			Exp.	CNV
GBM	Exp.	8	25	NA	**0.004**	0.958	**3.408**	NA
	CNV	19	NA	86	0.411	0.976	NA	1.820
	Exp. and CNV	7	22	68	0.292	0.799	3.403	1.802
BRCA	Exp.	8	69	NA	**0.140**	0.935	**2.598**	NA
	CNV	20	NA	63	0.353	0.913	NA	1.854
	Exp. and CNV	11	69	68	0.535	0.897	2.580	1.857

Moreover, according to the results, combination of expression and CNV data types introduces noise and decreases the robustness (the Cophenetic Correlation Coefficient) of the results and, deteriorates performance of sample and gene clustering compared to when gene expression is used alone. This is related to the inconsistency between different data types and the fact that different genotypes can be transcribed and translated into similar phenotypes.

3.2.3. B2PS vs. NMF

Comparison between the proposed method and NMF is conducted using gene expression data, which is here detected as the most informative datatype for patient stratification. To identify the number of clusters of NMF, the method described in section 2.6 is used. The results of NMF with the selected number of clusters and B2PS with the detected number of clusters are included in Table 7 for GBM and BRCA datasets. According to the results, although NMF produces slightly

more robust results (which can be related to the higher number of clusters for B2PS), B2PS produces remarkably more meaningful stratification and feature clusters.

Table 7. Comparison between B2PS and NMF

Dataset	Method	Sample Clusters	Feature Clusters	Log-rank p-value	Cophenetic Corr. Coef.	GOTO
GBM	B2PS	8	25	**0.004**	0.958	**3.408**
	NMF	3	3	0.458	0.965	2.535
	B2PS	3	29	0.047	0.967	**3.405**
	B2PS	3	6	0.217	0.999	3.392
BRCA	B2PS	8	69	**0.140**	0.935	**2.598**
	NMF	3	3	0.226	0.991	2.541
	B2PS	3	101	0.120	0.998	**2.603**
	B2PS	3	6	0.489	0.983	2.548

To see whether B2PS can also perform as well when the numbers of sample clusters are the same for both methods, in another experiment, B2PS is forced to find the clustering structure with the number of subtypes detected by NMF. Results shown in Table 7 approves that B2PS performs better stratification and, interestingly, when the number of sample clusters of B2PS is restricted, the number of detected feature clusters increases and the quality of feature clusters remain almost the same as (slightly better than) the unrestricted case. To examine if this flexibility in the number of clusters across two different dimensions is an advantage that is effective in superior performance of B2PS, the results are compared with the case when this flexibility is discarded by simulating the inflexibility of NMF. For this, the numbers of sample and feature clusters are set "logically" equal for B2PS. Since, unlike NMF, B2PS inputs consists of both negative and positive values, then "logically" equivalent setting for B2PS is when the number of feature clusters is twice the number of sample clusters. The results of these double-restricted experiments are also included in Table 7. As it can be seen, this additional restriction distorts the performance in both aspects of sample and feature clustering considerably. Accordingly, results support the hypothesis that flexibility in the number of clusters improves the performance.

4. Conclusions

We proposed a novel probabilistic graphical model, called B2PS, for Bayesian integrative biclustering of biological data for patient stratification. Our experimental results demonstrate the effectiveness of the Bayesian approach for inclusion of prior knowledge and detection of a natural number of clusters. Our experiments also show that B2PS is more effective in patient stratification than NMF, due to the probabilistic nature of B2PS and its flexibility in the number of clusters across two dimensions. In cases where gene expression data is collectible (e.g., cancer), this type of data turns out to be more informative than other genomic data for patient stratification at least

for the datasets used in this study. For diseases where gene expression data cannot be gathered from the relevant tissue, methods like the one proposed in [5], which preprocess the genomic data to reduce their heterogeneity, can be useful. B2PS helps achieving the ultimate goal of stratified medicine by providing more robust subtypes and gene clusters, which can serve as a starting point to find subtype-specific gene expression profiles and consequently subtype specific pathways or subnetworks. This information together with the mutation profiles can then be employed to find the driver genetic variations for each subtype (the hallmark of stratified medicine). Future research may explore the integration of other data types (e.g., methylation, miRNA expression, and other structural variations like gene fusion) as well as increasing the resolution of the current datatypes (e.g., modeling gene expression as continuous distribution).

References

[1] J. C. D. Willis and G. M. Lord, *Nature Reviews (Immunology)* 15, 323 (2015).

[2] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, et al., *Bioinformatics* 22, 1122 (2006).

[3] A. Oghabian, S. Kilpinen, S. Hautaniemi and E. Czeizler, *PLoS ONE* 9 (2014).

[4] J. Sun, J. Bi and H. R. Kranzler, *BMC Genetics* 15 (2014).

[5] M. Hofree, J. P. Shen, H. Carter, A. Gross and T. Ideker, *Nature Methods,* 1108 (2013).

[6] L. Lazzeroni and A. Owen, *Statistica Sinica* 12, 61 (2002).

[7] A. Tanay, R. Sharan and R. Shamir, *Bioinformatics* 18, 136 (2002).

[8] D. Y. Cho and T. M. Przytycka, *Nucleic Acids Res.* 41, 8011 (2013).

[9] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr et al., *Bioinformatics* 26, 1520 (2010).

[10] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, et al., *Cancer Cell* 17, 98 (2010).

[11] R. Shen, A. B. Olshen and M. Ladanyi, *Bioinformatics* 25, 2906 (2009).

[12] S. Zhang, C. Liu, W. Li, H. Shen et al., *Nucleic Acids Res* 40, 9379 (2012).

[13] E. Meeds and S. Roweis, UTML TR 2007–001, University of Toronto, Toronto (2007).

[14] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, et al., *PLoS One* 7, (2012).

[15] G. Casella and E. I. George, *The American Statistician* 46, 167 (1992).

[16] D. D. Lee and H. S. Seung, *Adv. Neural Inform. Process. Syst* 13, 556 (2001.

[17] N. Mantel, *Cancer Chemotherapy Reports* 50, 163 (1966).

[18] J. P. Brunet, P. Tamayo, T. R. Golub and P. M. Jill, *PNAS* 12, 4164 (2004.

[19] M. Mistry and P. Pavlidis, *Bioinformatics* 9 (2008).

[20] "The Cancer Genome Atlas," [Online]. Available: http://cancergenome.nih.gov/.

[21] C. Mermel, S. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim and G. Getz, *Genome Biology* 12 (2011).

[22] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis et al., *Bioinformatics* 21, 3439 (2005).