Bioinformatics, 37(16), 2021, 2398–2404
doi: 10.1093/bioinformatics/btaa1076
Advance Access Publication Date: 26 December 2020
Original Paper

Systems biology

BiCoN: network-constrained biclustering of patients and omics data

Olga Lazareva [1,*], Stefan Canzar[2], Kevin Yuan [1], Jan Baumbach[1], David B. Blumenthal [1], Paolo Tieri [3,4], Tim Kacprowski [1,5,‡] and Markus List[1,‡]

[1]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Weihenstephan, 80333 Munich, Germany, [2]Gene Center, Ludwig-Maximilians-University of Munich, 81377 Munich, Germany, [3]CNR National Research Council, IAC Institute for Applied Computing, Rome 00185, Italy, [4]La Sapienza University of Rome, Rome 00185, Italy and [5]Division of Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, Brunswick 38106, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Authors.
Associate Editor: Alfonso Valencia

Received on July 20, 2020; revised on November 25, 2020; editorial decision on December 12, 2020; accepted on December 15, 2020

Abstract

Motivation: Unsupervised learning approaches are frequently used to stratify patients into clinically relevant subgroups and to identify biomarkers such as disease-associated genes. However, clustering and biclustering techniques are oblivious to the functional relationship of genes and are thus not ideally suited to pinpoint molecular mechanisms along with patient subgroups.

Results: We developed the network-constrained biclustering approach Biclustering Constrained by Networks (BiCoN) which (i) restricts biclusters to functionally related genes connected in molecular interaction networks and (ii) maximizes the difference in gene expression between two subgroups of patients. This allows BiCoN to simultaneously pinpoint molecular mechanisms responsible for the patient grouping. Network-constrained clustering of genes makes BiCoN more robust to noise and batch effects than typical clustering and biclustering methods. BiCoN can faithfully reproduce known disease subtypes as well as novel, clinically relevant patient subgroups, as we could demonstrate using breast and lung cancer datasets. In summary, BiCoN is a novel systems medicine tool that combines several heuristic optimization strategies for robust disease mechanism extraction. BiCoN is well-documented and freely available as a python package or a web interface.

Availability and implementation: PyPI package: https://pypi.org/project/bicon.

Web interface: https://exbio.wzw.tum.de/bicon.

Contact: olga.lazareva@tum.de

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

1 Introduction

Biomarkers are essential for stratifying patients for diagnosis, prognosis, or treatment selection. Currently, individual or composite molecular biomarkers based on, e.g. expression, methylation, mutation status or copy number variation are used. Biomarker discovery has greatly benefited from supervised methods that identify molecular features that have a strong association with disease-relevant variables such as drug response, relapse, survival time or disease subtype. However, supervised methods are strongly biased by our current understanding of diseases, in particular by disease definitions that were established before rich molecular data became available. While classical unsupervised methods such as clustering have been successfully applied in the past, e.g. to reveal gene signatures predicting breast cancer subtypes (Nielsen *et al.*, 2010; Parker *et al.*, 2009), they group patients based on the entire molecular profile and overlook meaningful differences limited to a subset of genes.

Biclustering aims to discover rows in a matrix which exhibit similar behavior across a subset of columns and *vice versa* (Hartigan, 1972). It is suited for identifying disease-associated genes from gene expression data while stratifying patients at the same time (Prelić et al., 2006). As an NP-hard problem (Tanay *et al.*, 2002), biclustering is typically solved via heuristics. A gene expression matrix describes the expression of genes (rows) across samples

(columns), which can reflect individual patients, time points or conditions. In patient stratification (i.e. splitting patients into clinically relevant subgroups), samples typically stem from different patients with a disease phenotype. In gene expression data, a bicluster defines a set of genes and a set of patients for which these genes are co-expressed (Cheng and Church, 2000). Gene co-expression does not imply a direct functional connection and, hence, genes identified by biclustering are often difficult to interpret. In contrast, molecular interaction networks such as protein–protein interaction (PPI) networks capture direct and functional interactions.

Many diseases are caused by aberrations in molecular pathways or modules of functionally related genes (Berg *et al.*, 2002). This suggests to focus on gene modules for delivering more interpretable and robust mechanistic explanations of disease phenotypes. Network enrichment methods leverage prior information of molecular interactions for identifying gene modules as subnetworks (Batra *et al.*, 2017). Gene modules are robust features for classification and disease subtyping (Alcaraz *et al.*, 2017). Few methods exist that can utilize molecular interaction networks along with gene expression for patient stratification. Two integer linear programming methods were suggested (Liu *et al.*, 2014; Yu *et al.*, 2017) both of which rely on the GeneRank (Morrison *et al.*, 2005) algorithm to incorporate network information. GeneRank depends on a parameter θ describing the influence of the network whose choice is not straightforward and was shown to have a notable impact on the results (Yu *et al.*, 2017). While these methods propagate the gene expression signal among the connected genes in a network, they generally do not produce connected subnetworks. Thus, they are not suited for discovering disease modules with mechanistic interpretation. To overcome this issue, we present biclustering constrained by networks (BiCoN), a tool that accepts gene expression data as input and stratifies patients into two subgroups while identifying, for each group, a subnetwork of genes that can be interpreted as a shared molecular mechanism. In contrast to the classical definition of biclustering, BiCoN extracts a fixed number of non-overlapping biclusters, which are connected in a molecular interaction network. BiCoN delivers meaningful results on real-world datasets on par with other state-of-the art methods. We have validated our results on breast cancer (TCGA Pan-Cancer) and non-small cell lung carcinoma (NSCLC) datasets (Rousseaux *et al.*, 2013) and found that BiCoN is robust to batch effects and delivers biologically interpretable mechanistic insights into disease subtypes.

2 Materials and methods

2.1 Problem statement

BiCoN aims at stratifying patients into two subgroups while extracting two sets of genes which are connected in a molecular interaction network and show opposite behavior (i.e. similar to conventional differential expression analysis). The resulting subnetwork can thus be interpreted as a biological function jointly carried out by these genes which is active in one patient group and inactive in the other one. This assumption is reflected in our objective function and formally described below.

Consider a matrix of expression values $X^{n\times m}$ with n genes and m patients as well as $G=(V,E)$, a molecular interaction network of gene set V and protein–protein or gene–gene interactions E. We further consider P as the set of m patients (samples) and construct a complete bipartite graph $B=((V,P),E_w)$ with genes V and patients P as node types connected by weighted edges E_w. Edge weights reflect the expression strength for a given patient from expression values $X^{n\times m}$. We construct a joint graph $J=((V,P),(E,E_w))$ by mapping G onto B via the shared genes in V. Our goal is to partition P into clusters P_1, P_2 and to find two connected subnetworks $G_1(V_1,E_{,1}),G_2(V_2,E_2)$ each of minimal size $L_{\min}$ and of maximal size $L_{\max}$. Size constraints can be adapted by users to the expected size of the molecular pathways, i.e. small subnetworks will represent more specific and large subnetworks more general molecular functions or biological processes. Thus, we aim to derive patient groups (clusters P_1 and P_2) which are characterized by maximally differential expression in the extracted subnetworks:

$$f(X,V,P,c)=\sum_{(i,j)\in(1,2),(2,1)} w_i(\overline{X}[V_i,P_i]-\overline{X}[V_i,P_j]) \quad (1)$$

Where $\overline{X}[V_i,P_j]$ is the average expression of genes of module i for patients in cluster j, w_i is a weight for $G_i(V_i,E_i)$ which penalizes too small or too large, disconnected solutions:

$$w_i=\begin{cases}\dfrac{|\mathrm{LCC}_{G_i(V_i,E_i)}|}{L_{\min}} & \text{if } |\mathrm{LCC}_{G_i(V_i,E_i)}|\le L_{\min}\\ \dfrac{L_{\max}}{|\mathrm{LCC}_{G_i(V_i,E_i)}|} & \text{if } |\mathrm{LCC}_{G_i(V_i,E_i)}|\ge L_{\max}\\ 1 & \text{otherwise}\end{cases} \quad (2)$$

Where $|\mathrm{LCC}_{G_i(V_i,E_i)}|$ is the size of the largest connected component (LCC) in a subnetwork $G_i(V_i,E_i)$. Thus, w_i is always equal to 1 if the size of LCC corresponds to the user defined $L_{\min}$ and $L_{\max}$ and $w_i<1$ means that the obtained solution does not fit into the desired range. Smaller w_i means larger deviation from a user's preferences. The motivation for implementing a fuzzy threshold is that the user-selected $L_{\min}$ and $L_{\max}$ parameters may not always lead to a viable solution, i.e. if $w_i<1$ BiCoN could not find any differentially expressed subnetworks of the selected size in the given data.

To obtain more than two clusters, BiCoN can in principle be applied recursively to further split clusters as also shown in the application in Section 4.2.

2.2 Bicon algorithm

BiCoN is a heuristic algorithm that finds differentially expressed subnetworks that can mechanistically explain patient stratification. This combinatorial problem can be addressed by various metaheuristic frameworks such as e.g. Genetic Algorithm (Banzhaf *et al.*, 1998) or Swarm Intelligence (Eberhart and Kennedy, 1995). We have chosen Ant Colony Optimization (ACO) (Stützle, 2009) as the main framework that performs exploration of the search space and Local Search (Aarts *et al.*, 2003) to ensure local optimality of the final solution. The combination of ACO and Local Search was shown to be very efficient in finding near-optimal solutions to hard combinatorial optimization problems (Stützle and Hoos, 1999) and leads to significant improvements compared to ACO or local search alone (Stutzle and Hoos, 1997). As we already had good prior experiences with ACO on similar problems (Alcaraz *et al.*, 2012), we expected that combination with local search will lead to high quality results.

ACO is a nature-inspired probabilistic technique for solving computational problems which can be reduced to finding optimal paths through graphs. We use ACO to identify a set of relevant genes for each patient which we subsequently aggregate into a global solution. A full description of the algorithm and the pseudo-code can be found in the Supplementary Material, section 'Algorithm description'. We also describe the full workflow in Figure 1. Briefly, ants travel the joint graph J in three phases which are repeated until convergence:

i. An ant performs a random walk within nodes that are highly connected to a patient-node and makes greedy choices according to the objective function [Equation (1)] by choosing genes which are most relevant to a patient [orange edges in Fig. 1 (Step 2)]. The probability of selecting a gene for a certain patient depends on the combined information from gene expression values (which are encoded in the heuristic information matrix) and the ant's 'memories' on whether the choice of this gene has led to a quality solution in the previous rounds (pheromone matrix). More details about the implementation can be found in Supplementary Material, section 'Algorithm description'.

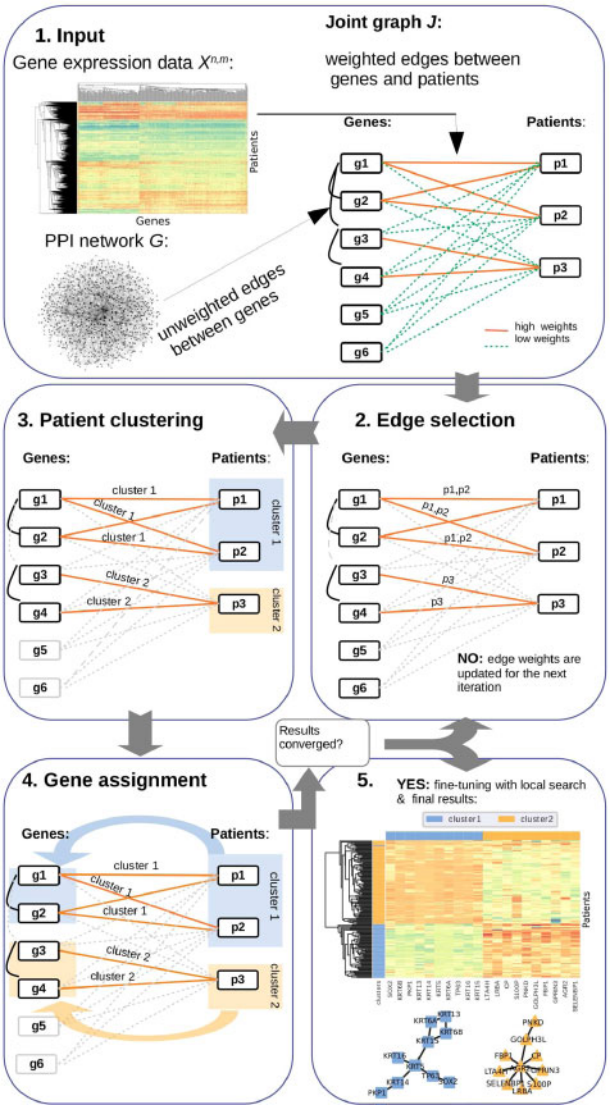

Fig. 1. The algorithmic framework of BiCoN. (Step 1) Gene expression data X is converted to a bipartite graph B and PPI interactions (black) are added as additional edges between genes to form a joint graph J. (Step 2) ACO determines the most relevant features for each patient where edges are annotated with patient IDs. The selected genes are used for patient clustering in Step 3. Next, (Step 4) genes are reassigned from individual patients to their corresponding clusters. Multiple possible solutions are computed in parallel and then evaluated and reinforced. As a result (Step 5), patients are stratified-based only on subnetworks that can be interpreted as disease mechanisms

ii. The selected genes are then used for clustering patients with the k-means algorithm where $k = c = 2$ (Step 3, Fig. 1). Relevant genes for each patient cluster are extracted at Step 4 (Fig. 1). A candidate solution is evaluated by the objective function score.
iii. The best solution is used for updating the pheromone and probability matrices for the next iteration.

When the best solution is obtained we perform local search for possible local improvements, i.e. iteratively apply changes to subnetworks (such as node insertion, deletion or substitution) and keep changes that lead to objective function maximization. This allows us to retrieve robust and stable solutions as well as to ensure local optimality.

Even though BiCoN uses several hyperparameters, our experiments have shown that those do not have a large impact on the results and the optimal combination is determined automatically based on the dimension and distribution of the expression matrix. Therefore, the user only has to specify the desired size of the solution subnetworks ($L_{\min}$ and $L_{\max}$).

2.3 Data collection and processing

2.2.1 Gene expression data

TCGA breast cancer data were obtained through the UCSC Xena browser (https://xenabrowser.net/). The NSCLC dataset [accession number GSE30219, (Rousseaux *et al.*, 2013)] was obtained using GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/). Both datasets were retrieved together with the corresponding metadata which contained annotated cancer subtypes.

For the NSCLC dataset, gene probes were mapped to Entrez gene IDs. If multiple probes corresponded to a single gene, the median value was used. We applied a $\log_2$ transformation to account for skewness of the data. Data were z-score transformed to indicate the magnitude of changes in gene expression in individual samples and conditions compared to the background. In most gene expression datasets, a majority of genes is lowly expressed and does not vary to a larger extent. To account for this and to improve run-time, BiCoN filters out genes with a small variance preserving only the n most variant genes (here, $n = 3000$).

2.2.2 Molecular interaction network

We used physical and genetic PPIs in *Homo sapiens* from BioGRID (version 3.5.176). The network consisted of 343 563 unique interactions between 16 830 genes.

2.4 Simulation of batch effects

Batch effects are technical variations that have been introduced by external factors during handling of the samples (e.g. personnel effects, environmental conditions and different experiment times) (Goh *et al.*, 2017; Luo *et al.*, 2010). While some of those effects can be minimized, batch effects are still almost inevitable in practice (Chen *et al.*, 2011). Many methods have been proposed for removing batch effects from data (Lazar *et al.*, 2012). However, removing batch effects may also remove biologically relevant group differences from the data. Batch effect correction methods that are designed to retain group differences can lead to exaggerated confidence in downstream analyses (Nygaard *et al.*, 2016). In unsupervised analysis, this issue is critical, since we, by definition, do not know the relevant sample or patient groups *a priori*.

To demonstrate that BiCoN is robust to batch effects, we simulate data using a linear mixed effect model. We consider two variables: *cluster* and *batch*. The variable *cluster* indicates whether a gene is part of the foreground (*cluster* = 1 or *cluster* = 2) or the background (*cluster* = 0), i.e. it is not differentially expressed. The variable *batch* indicates the study or batch of expression values (*batch* = 1 or *batch* = 2). The expression values are simulated as follows:

$$g_i = 1 + 2 \times batch + 2 \times cluster + \gamma_1 \times cluster + \gamma_2 \times batch + \varepsilon_i \quad (3)$$

where the first part of the equation ($1 + 2 \times batch + 2 \times cluster$) is fixed and shared by all genes. Errors ε_i are independent and identically distributed (with zero mean). The random effects parameters γ_1 and γ_2 follow a bivariate normal distribution with zero mean, and variance 1 and 2, respectively, i.e. the technical variance is twice the biological variance.

The network was simulated as three disjoint Barabasi–Albert graphs (one for each of genes biclusters and one for background genes) (Barabási and Albert, 1999) which were connected by random edges until they have reached the same density as the BioGRID network (0.0013). Barabasi–Albert graphs have similar node degree distribution as the BioGRID network and were thus considered suitable for the simulation study.

2.5 Benchmarking

To show how BiCoN results compare to commonly used clustering and biclustering algorithms, we selected several popular biclustering and clustering methods (listed in Supplementary Table S3) and performed multiple assessments:

i. To show how BiCoN can recover PAM50 annotated breast cancer subtypes (using TCGA data as a source), we computed Jaccard index (an intersection of two sets over the union) between the known subtypes and the resulting patients clusters/biclusters.
ii. To show how BiCoN can handle batch effect in comparison to other methods, we simulated data as described in Section 3.2 and computed the overlap between known classes of patients and the resulting clusters/biclusters. To avoid favoring the assumption of genes connectivity used by BiCoN, we also repeated the simulation such that the signal-carrying foreground genes are randomly distributed over the network.

As a metric for comparison, we used Jaccard index rather than Matthews Correlation Coefficient (MCC) as it allows to measure relationship between resulting biclusters and the actual classes even when the patients biclusters overlap and do not include all patients. All data were normalized and processed as described in Section 3.1 for all methods (including BiCoN).

Even though we use classical clustering methods for benchmarking, we emphasise key differences between the suggested approach and classical clustering. BiCoN extracts biological mechanisms that explain patient stratification. Even though subnetworks extraction after clustering of patients is feasible, to our knowledge there is no gold standard for this procedure. While it is possible to extract subnetworks and disease mechanisms subsequent to clustering or by relying on known disease subtypes (Alcaraz *et al.*, 2017), we argue that such clusters are driven by global differences and not by the activity of a single disease mechanism. Hence, extracting disease mechanisms along with patient stratification is better suited to identify patient subgroups affected by key disease mechanisms. Moreover, clustering performed on the whole genome is also not advisable as the use of multi-dimensional data can lead to multiple negative effects, which are often referred to as 'curse of dimensionality' (Thangavelu *et al.*, 2019).

For all selected algorithms, we chose parameters that maximize performance for each of the methods.

3 Results and discussion

We evaluated BiCoN on simulated and real data with respect to the robustness of patient clustering and gene selection as well as robustness to batch effects. Furthermore, two application cases illustrate the practical use of BiCoN.

3.1 Noise robustness

To introduce varying levels of noise to a dataset, we randomly select between 0 and 90% of the genes and randomly permute their expression values. A noise level of 0.1 means that the expression vectors of 10% of genes were permuted. For each noise level, we average results over 10 independent runs.

We use the NSCLC dataset with two annotated subtypes as gold standard: adenocarcinoma and squamous cell carcinoma. As evaluation metrics, we consider the value of BiCoN objective function as well as MCC (Matthews, 1975) between the proposed clusters and cancer subtype labels. The latter is meant to demonstrate that BiCoN is able to recover cancer subtypes while inferring a mechanistic explanation for the subtype differences. For all described results, we retain the 3000 most variant genes and set parameters $L_{\min} = 10$ and $L_{\max} = 25$ to control the size of the solution.

Figure 2a shows a consistent decline in the objective function with increasing noise, indicating that the algorithm is reacting reasonably to the decline in data quality. Figure 2b shows that the algorithm is able to recapture the cancer subtypes almost perfectly (average MCC higher than 0.9) up to a noise level of 0.5 where 50% of the data have been permuted. Figure 2c shows a strong positive correlation between the objective function value and MCC, which confirms that the objective function is high when cancer subtypes are well separated.

3.2 Batch effect robustness

BiCoN is a graph-based method and, hence, it is not as strongly affected by the global distribution of expression values as classical clustering methods. Pre-processing methods that scale data to a certain range enforce it to have certain mean and variance (e.g. z-scores) or make the distribution more symmetrical (e.g. $\log_2$ transformation) are not ideal for batch effect correction as they do not differentiate between signal and noise. In this scenario, a graph-based method benefits from the assumption that the joint signal of the genes in a subnetwork is stronger than that of individual genes.

To study if BiCoN can indeed tolerate batch effects, we simulate gene expression data (see Section 3 for details) where we introduce a batch effect with a larger variance than for the group difference. Our aim is to show that BiCoN can leverage the network to recover the signal even if it is overshadowed by batch effects.

We have simulated expression data for 2×20 foreground genes (two biclusters) and for 1000 background genes. We also tested the performance with 2×30, 2×40 and 2×60 foreground genes.

Figure 3a shows that the batches differ in their distribution, causing hierarchical clustering to group samples by batch rather than by disease phenotype. Figure 3b shows that differences due to batch effects are eliminated after z-score normalization. We can also see that the difference between the sample groups is now lost and cannot be recovered by hierarchical clustering. Figure 3c shows that in spite of this noise, BiCoN can recover the disease phenotype together with the foreground genes. Thus, when two datasets can be normalized separately (e.g. z-scores are applied to each dataset), BiCoN is uniquely suited to cluster patients where individual gene modules are disturbed. Even when the signal is obscured by batch effects, the functional connection of solution genes in the network (Fig. 3d) helps to robustly recover the signal.

To show how BiCoN results align with other clustering and biclustering methods, we have simulated 10 datasets with batch effect and evaluated the performance. To make sure that we do not put BiCoN in favor by enforcing connectivity of genes, we also performed simulations with a single Barabasi–Albert graph, where foreground genes were randomly distributed (Fig. 4).

Among the considered biclustering algorithms (Supplementary Table S3), only Bimax was capable of finding any clusters, while Plaid and QUBIC could not find any structure in the given data regardless of chosen parameters and therefore was excluded from further assessment. The experiments showed that even though the quality of the results drops when the foreground genes are not directly connected, BiCoN still performs significantly better than other methods. The simulated network had power-law node degree distribution which means that the network diameter is rather small and therefore many foreground genes are still reachable through hub-nodes even when they are not directly connected. Thus, BiCoN performance dropped when using random networks (due to the noise of the hub-nodes) but still outperformed other methods that are not network-restricted.

3.3 Application to TCGA breast cancer data

We applied BiCoN to the TCGA breast cancer dataset. We expected BiCoN to be able to recover known subtypes assigned via the PAM50 gene panel (Nielsen *et al.*, 2010; Parker *et al.*, 2009) which is a gold standard in breast cancer subtype prediction. For the analysis, we focused on patients with the most common molecular subtypes, luminal (estrogen-receptor and/or progesterone-receptor positive) and basal (hormone-receptor-negative and HER2 negative).

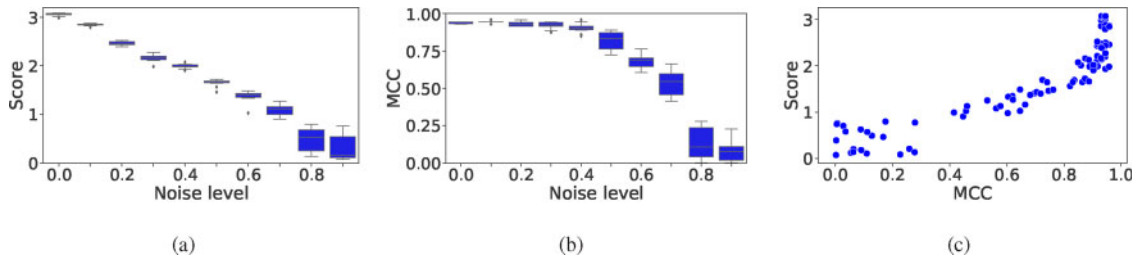

Fig. 2. Robustness analysis. (**a**) Objective function score versus the percentage of noisy data. (**b**) MCC with respect to the known classes versus the percentage of noisy data. (**c**) Correlation of objective function scores and MCC

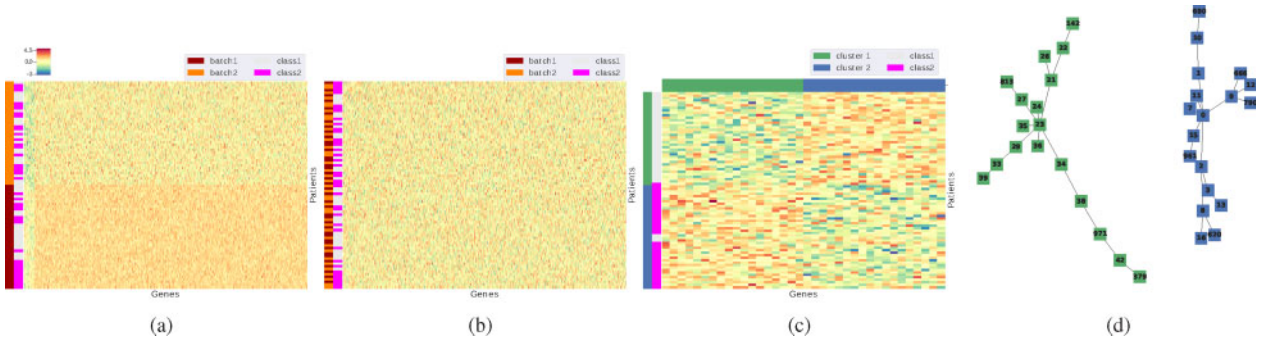

Fig. 3. (**a**) Hierarchical clustering of two datasets with different distributions due to batch effects. (**b**) The merged datasets after z-scores normalization. The batch effect vanishes, but the disease phenotype is still not distinguishable. (**c**) BiCoN is able to recover the initial disease phenotypes with Jaccard index of 0.92 (in average after 10 runs) while extracting the 40 foreground genes out of 1000 background genes. (**d**) The resulting subnetworks for two corresponding patient clusters

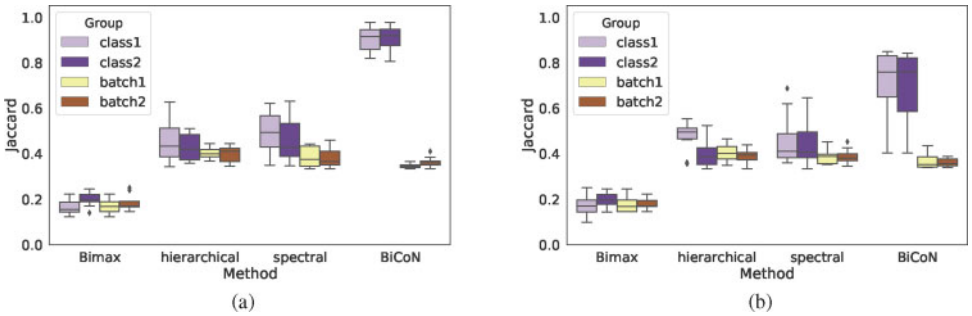

Fig. 4. Jaccard indices between the patients' clusters and actual subgroups (class 1 or class 2) as well as with batches of patients (batch 1 and batch 2) for 10 simulated datasets with a strong batch effect. (**a**) When foreground genes are connected in a network, BiCoN clusters patients almost perfectly based on the actual signal. (**b**) When the foreground genes are randomly distributed in the network, BiCoN still achieves higher performance than other methods that were capable to find any clusters. Plaid and QUBIC were not able to find any clusters and were excluded from further assessment

As a proof of concept, we first showed that BiCoN can separate patients into the two clinically well distinguishable subtypes luminal and basal. Next, we applied BiCoN separately for patients with luminal and basal subtype to investigate how patients are stratified in a more challenging scenario. For each subgroup, we ran the algorithm 10 times and selected a solution with the highest score based on the previous observation that the highest objective function score corresponds to the highest correlation between the resulting biclusters and the expected patient groups. We conducted gene set enrichment using genes from both subnetworks together using the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) as a background. We used the same hyperparameters as for our previous analysis: 3000 the most variant genes and $L_{min} = 10$ and $L_{max} = 25$ to control the size of the solution.

3.3.1 Luminal versus basal separation

As expected, the separation between patients with luminal and basal breast cancer subtypes is straight-forward. The clusters correspond

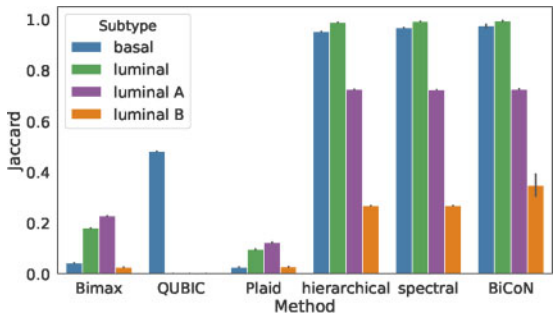

Fig. 5. TCGA breast cancer subtypes identification by various algorithms (for 10 runs). Jaccard index was computed as a best match between produced patients clusters and the known breast cancer subtypes for BiCoN and other well-known clustering and biclustering algorithms. BiCoN shows performance which is comparable with other clustering algorithms while also revealing functionally connected subnetworks which explain the phenotype

to the subtype labels and the separation between patients groups matches the PAM50 classification (average Jaccard index is equal to 0.99, Supplementary Fig. S1a). BiCoN not only performs as well as methods like hierarchical clustering (Fig. 5, where the Jaccard index is 0.96 for the luminal and basal subtypes) but also yields two differentially expressed subnetworks (Supplementary Fig. S1b). The extracted subnetworks explain subtype differences with a vastly lower number of genes than a classical clustering method while offering a mechanistic explanation of subtype differences. Note that while BiCoN restricts genes inside a bicluster to be connected, it does not impose any relationships between two biclusters. As a consequence, it is possible that the resulting subnetworks overlap.

In contrast to methods yielding gene signatures such as PAM50, BiCoN focuses on revealing specific pathways. Enrichment analysis of cancer-related pathways (Supplementary Fig. S6) confirms strong association of the resulting genes with breast cancer subtype-specific signaling, in particular estrogen signaling pathway (adjusted P-value $= 0.018$) and ErbB signaling pathway (adjusted P-value $= 0.025$).

Random-walks on scale-free networks are biased toward hub-nodes since these have a high degree (Gillis *et al.*, 2014). BiCoN avoids this hub bias as it performs random walks on the joint graph of a PPI and expression data which is not scale-free. Consequently, the selected nodes have approximately the same degree distribution as the input network (Supplementary Fig. S4).

3.3.2 Luminal patient stratification

Next, we consider only patients that were originally classified as luminal subtype to see if we can further stratify them into subtypes luminal A and luminal B which are known to be difficult to separate on the level of gene expression. Here, our solution does not agree with the PAM50 classes [Fig. 5, mean Jaccard index 0.49 for the luminal A (lumA) and luminal B (lumB) subtypes], although we observe two clearly separable groups and that most of the luminal B patients were placed in cluster 1 (Supplementary Fig. S2). We hypothesize that contributions of the tumor-microenvironment may explain the observed clusters. To test this hypothesis, we used the signature-based deconvolution method xCell (Aran *et al.*, 2017) to estimates contributions of 64 immune and stromal cell types in the two clusters. xCell summarizes the contribution of tumor-infiltrating leukocytes to the microenvironment via aggregated scores such as an immune score, a stromal score and a microenvironment score. Clusters reported by BiCoN show significant differences between cell type scores. The strongest difference between patients is found in the stromal score ($-\log_{10}$ P-value is over > 55), hematopoietic stem cells ($-\log_{10}$ P-value > 50) and CLP cells ($-\log_{10}$ P-value > 50). See Supplementary Figures S7a and S8a for details. These results indicate that some of the luminal A and luminal B patients share similar tumor microenvironments and, consequently, the further stratification of luminal subtypes is not straight-forward. These results are corroborated by other studies which investigate immune-related subtypes of luminal breast cancer (Jiang *et al.*, 2020; Zhu *et al.*, 2019).

3.3.3 Basal patients' stratification

Bertucci *et al.* (2012) characterized basal, also known as triple negative, breast cancer as the most challenging breast cancer subtype with poor prognosis despite relatively high chemosensitivity. Currently, there is no targeted therapy and no routine diagnostic procedure specifically for this subtype. Although no clinically relevant subgroups of the basal subtype are known, BiCoN achieved a clear separation into two subgroups (Supplementary Fig. S3).

Derived subnetworks show robust correlation with immune system response functions which is reasonable given that tumor samples are infiltrated with leukocytes. All three enriched pathways (primary immunodeficiency, hematopoietic cell lineage and B cell receptor signaling pathway) have a direct connection to the immune response (Supplementary Fig. S5a). Molecular function enrichment also confirms the relation between the selected genes and immune response (Supplementary Fig. S5b). Cell type deconvolution analysis with xCell shows a high correlation of the clusters with aDC, CD4+ memory T-cells, B-cells, CD8+ T-cells and other immune response related cells (Supplementary Figs S7b and S8b). Similar to the results in luminal patients, our results indicate that basal breast cancer patients can be clustered by the contribution of tumor-infiltrating leukocytes, which is a clinical key factor for prognosis and treatment via immunotherapy.

4 Conclusion and outlook

Classical biclustering methods were shown to perform sub-optimally when non-intersecting, large patient subgroups are of interest as is often the case in patient stratification. Clustering methods, on the other hand, are more suited for this task, but they use the whole gene set and do not provide a mechanistic explanation of patient stratification (Fig. 5). Therefore, BiCoN is uniquely suited to cluster patients along with extracting fixed-size subnetworks capable of mechanistically explaining the patient stratification. Moreover, simultaneous clustering of gene expression and networks make BiCoN robust to noise and more robust to batch effect than typical clustering and biclustering methods.

BiCoN leverages molecular interaction networks in the analysis of gene expression data to faithfully produce known subtypes as well as novel, clinically relevant patient subgroups, as we could demonstrate using data from TCGA. We stress that BiCoN and the concept of network-constrained biclustering are not limited to gene expression data or PPI networks. We plan to apply BiCoN to other types of omics data such as DNA methylation, copy number variation or single nucleotide polymorphisms. We envision BiCoN to be useful for single-cell RNA-seq data for uncovering differences in signaling between clusters of cells and for the discovery of novel cell types. BiCoN, which is available as a web-interface and a PyPI package, has great potential to enhance our understanding of diseases, cellular heterogeneity and putative drug targets.

Acknowledgements

The results shown here are in whole- or part-based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors thank Quirin Heiß for his contributions to the source code of the web-interface and Hoan Van Do for a fruitful discussion about the algorithm.

Funding

This work was supported by the Bavarian State Ministry of Science and the Arts as part of the Bavarian Research Institute for Digital Transformation (BIDT) [to O.L.]; H2020 project RepoTrial [777111 to J.B. and T.K.]; VILLUM Young Investigator Grant [to J.B.]; and COST CA15120 OpenMultiMed [to O.L.].

Conflict of Interest: none declared.

References

Aarts,E. *et al.* (2003) *Local Search in Combinatorial Optimization*. Princeton University Press, Chichester, UK.

Alcaraz,N. *et al.* (2012) Efficient key pathway mining: combining networks and omics data. *Integr. Biol.*, **4**, 756–764.

Alcaraz,N. *et al.* (2017) *De novo* pathway-based biomarker identification. *Nucleic Acids Res.*, **45**, e151.

Aran,D. *et al.* (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.*, **18**, 220.

Banzhaf,W. *et al.* (1998) *Genetic Programming: An Introduction*. Vol. **1**, Morgan Kaufmann, San Francisco.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Batra,R. *et al.* (2017) On the performance of *de novo* pathway enrichment. *NPJ Syst. Biol. Appl.*, **3**, 6.

Berg,J. *et al.* (2002) Defects in signaling pathways can lead to cancer and other diseases. In: *Biochemistry*, 5th edn, Section, 15. W.H. Freeman, New York.

Bertucci,F. *et al.* (2012) Basal breast cancer: a complex and deadly molecular subtype. *Curr. Mol. Med.*, **12**, 96–110.

Chen,C. *et al.* (2011) Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One*, **6**, e17238.

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

Eberhart,R. and Kennedy,J. (1995) Particle swarm optimization. In: *Proceedings of the IEEE International Conference on Neural Networks*. Vol. **4**, Citeseer, pp. 1942–1948, Perth, Western Australia.

Gillis,J. *et al.* (2014) Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *J. Proteomics*, **100**, 44–54.

Goh,W.W.B. *et al.* (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.*, **35**, 498–507.

Hartigan,J.A. (1972) Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, **67**, 123–129.

Jiang,J. *et al.* (2020) Tumour-infiltrating immune cell-based subtyping and signature gene analysis in breast cancer based on gene expression profiles. *J. Cancer*, **11**, 1568–1583.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Lazar,C. *et al.* (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, **14**, 469–490.

Liu,Y. *et al.* (2014) A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, **15**, 37.

Luo,J. *et al.* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. *Pharmacogenomics J.*, **10**, 278–291.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Morrison,J.L. *et al.* (2005) GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.

Nielsen,T.O. *et al.* (2010) A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.*, **16**, 5222–5232.

Nygaard,V. *et al.* (2016) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, **17**, 29–39.

Parker,J.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Prelić,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Rousseaux,S. *et al.* (2013) Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Trans. Med.*, **5**, 186ra66.

Stützle,T. (2009) Ant colony optimization. In: Ehrgott, M., Fonseca, C.M., Gandibleux, X., Hao, J.-K. and Sevaux, M. (eds.) *Evolutionary Multi-Criterion Optimization*. Springer, Berlin Heidelberg.

Stutzle,T. and Hoos,H. (1997) Max-min ant system and local search for the traveling salesman problem. In: *Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC'97)*, IEEE, Indianapolis, IN, USA, pp. 309–314.

Stützle,T. and Hoos,H. (1999) The max-min ant system and local search for combinatorial optimization problems. In: *Meta-Heuristics*, Springer, pp. 313–329, Boston, MA.

Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.

Thangavelu,S. *et al.* (2019) Feature selection in cancer genetics using hybrid soft computing., IEEE, Palladam, India, India, pp. 734–739.

Yu,G. *et al.* (2017) Network-aided bi-clustering for discovering cancer subtypes. *Sci. Rep.*, **7**, 1046.

Zhu,B. *et al.* (2019) Immune gene expression profiling reveals heterogeneity in luminal breast tumors. *Breast Cancer Res.*, **21**, 147.