Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models [version 1; peer review: awaiting peer review]

Katharina Baum¹, Jagath C. Rajapakse³, Francisco Azuaje¹

¹Bioinformatics and Modelling, Luxembourg Institute of Health, Strassen, Luxembourg
²Mathematical Modelling of Cellular Processes, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
³School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

Abstract

Background: Biological entities such as genes, promoters, mRNA, metabolites or proteins do not act alone, but in concert in their network context. Modules, i.e., groups of nodes with similar topological properties in these networks characterize important biological functions of the underlying biomolecular system. Edges in such molecular networks represent regulatory and physical interactions, and comparing them between conditions provides valuable information on differential molecular mechanisms. However, biological data is inherently noisy and network reduction techniques can propagate errors particularly to the level of edges. We aim to improve the analysis of networks of biological molecules by deriving modules together with edge relevance estimations that are based on global network characteristics.

Methods: We propose to fit the networks to stochastic block models (SBM), a method that has not yet been investigated for the analysis of biomolecular networks. This procedure both delivers modules of the networks and enables the derivation of edge confidence scores. We apply it to correlation-based networks of breast cancer data originating from high-throughput measurements of diverse molecular layers such as transcriptomics, proteomics, and metabolomics. The networks were reduced by thresholding for correlation significance or by requirements on scale-freeness.

Results and discussion: We find that the networks are best represented by the hierarchical version of the SBM, and many of the predicted blocks have a biological meaning according to functional annotation. The edge confidence scores are overall in concordance with the biological evidence given by the measurements. As they are based on global network connectivity characteristics and potential hierarchies within the biomolecular networks are taken into account, they could be used as additional, integrated features in network-based data comparisons. Their tight relationship to edge existence probabilities can be exploited to predict missing or spurious edges in order to improve the network representation of the underlying biological system.
Keywords
biomolecular networks, co-expression networks, edge relevance, hierarchical stochastic block model, missing and spurious edges, module detection, network-based omics analysis

Corresponding authors: Jagath C. Rajapakse (asjagath@ntu.edu.sg), Francisco Azuaje (francisco.azuaje@lih.lu)

Author roles: Baum K: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Rajapakse JC: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Review & Editing; Azuaje F: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was funded by Luxembourg National Research Fund, FNR, SINGALUN project. KB acknowledges funding by an Add-on Fellowship for Interdisciplinary Life Sciences of the Joachim Herz Stiftung.

Copyright: © 2019 Baum K et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Baum K, Rajapakse JC and Azuaje F. Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models [version 1; peer review: awaiting peer review] F1000Research 2019, 8:465 (https://doi.org/10.12688/f1000research.18705.1)

First published: 14 Apr 2019, 8:465 (https://doi.org/10.12688/f1000research.18705.1)
Introduction
High-throughput measurement techniques are advancing and become ever less expensive, enabling the screening of multiple biological data layers in single patients as almost standard clinical diagnostic tools. The wealth of the biological data can only be understood if treating the measured entities – gene promoters, mRNA, metabolites, proteins and their activity – not separately, but in their network context. Thereby, one method to capture the interdependencies of the intracellular machinery relies on the hypothesis that strongly connected molecular entities are either co-activated or co-repressed, i.e. their measured abundances should be correlated. Fully connected, weighted biomolecular networks can be established, in which each node corresponds to a molecular entity and is connected to each other node by an edge. The weight of the edge is the correlation between the measurements and is considered to represent how strongly the nodes are connected, interact or regulate each other.

Approaching a network-level analysis of a biological system by correlation-based interactions has the advantage that it does not require a priori knowledge, and thus it is focused on the interaction profile for which evidence can be found in the measurements of the considered condition. However, this is a blessing and a curse: correlation-based networks suffer from the fact that the biological measurements are inherently noisy, even more so for small sample sizes as in rare diseases or in personalized medicine. This affects the values of the edge weights and the results of network reduction, and can be detrimental for subsequent analyses of the networks. In these cases, considering in addition alternative sources of edge relevance that are based on more global characteristics of the system might be beneficial and could make the network representations of the system and their analysis more robust.

One of the commonly performed analyses of correlation-based networks relies on the observation that biological networks exhibit a modular structure, in which tightly regulated modules are loosely connected to other modules. An important readout from correlation-based networks therefore is the biological and functional characterization of condition-specific modules, i.e. communities of co-regulated entities. A plethora of methods for module detection in networks has been proposed. Also for the inference of edge relevance, or edge prediction, as being tightly linked to the problem of the detection of missing or spurious edges, numerous methods have been suggested. In this work, we showcase a method that is able to derive, within a single framework, modules as well as scores of edge relevance: representing the biomolecular correlation-based networks as stochastic block models (SBMs). SBMs are the simplest form of generative network models for community structures and can also accommodate hierarchies, which are key to convey robustness to biological function. Other generative model approaches relying on scale-freeness of the network architecture have been used for edge prediction in protein-protein interaction networks, and the SBM has been used for representing other types of networks, but not yet for biomolecular networks. In generative network model approaches, the network is described by a stringent mathematical framework based on statistical assumptions on network characteristics. For the case of the stochastic block model, this step delivers already the modular structure of the network as the nodes are assigned to blocks according to their connectivity properties to all other nodes. In contrast to other module detection methods, blocks in the SBM are not necessarily formed by tightly intra-connected entities but by entities which interact similarly with the nodes from all other blocks. Therefore, comparing SBM-derived modules between networks representing different (e.g. biomedical) conditions could be especially informative to shed light on regulatory changes. In a second step, the mathematical representation of the networks by SBMs is exploited to estimate edge probabilities. Specifically, it is assessed whether the existence of an edge in the network improves or reduces the fit of the network to the SBM. The resulting edge confidence scores are based on global network structure and can be used as alternative measure of edge relevance.

Here we showcase the SBM-based analysis (overview in Figure 1A) for networks of three different molecular types, derived from transcriptomic, proteomic and metabolomics data of breast cancer tumours. We assessed which of the different versions of SBM fits each data layer best. Then, we investigated whether the SBM representation is able to capture functionally relevant structures in our biomolecular networks. In detail, we determined the agreement between the predicted blocks and independent biomolecular functional annotations from databases. Finally, we took advantage of the description of the networks as SBMs for the computation of an edge confidence score for each edge as measure of edge relevance. The edge confidence scores can be exploited to re-establish erroneously removed edges or to remove spurious edges, or they could serve by themselves for deriving disease-relevant differences when comparing groups of patients.

All code is freely available on https://gitlab.com/biomodlih/ sbm-for-correlation-based-networks.

Methods
mRNA, protein, metabolite data for ER-breast cancer tumors
Breast cancer mRNA expression from RNAseq was obtained from the TCGA BRCA cohort via RTCGA downloading TCGA level 3 preprocessed BRCA files (search term: mRNAseq_Prep. Preprocess) on Nov 2, 2017. We used the normalized RSEM values. Protein data was obtained via the CPTAC homepage. We used the first replicate of samples measured in duplicates. We employed the unshared log ratio value for each sample to be used as alternative measure of edge relevance.

We used the following files for mRNA, protein and clinical data, respectively:
- *gdac.broadinstitute.org_BRCA.mRNAseq_Preprocess.
Level_3.2016012800.0.0/ BRCA.uncv2.mRNAseq_RSEM_normalized_log2.txt*
Figure 1. Analysis of correlation-based networks from different omics data. (A) Pipeline for the approach. Given the matrix of measurements of mRNA expression, protein expression or metabolite abundance for a group of samples (I), we compute a fully-connected weighted network (II) of the molecular species for each data layer separately using a correlation-based approach. We reduce the networks by setting a threshold to the edge weights and binarize the networks (III). Each network is fitted to different types of stochastic block models in which the network nodes are partitioned into blocks (IV), the best fitting model is employed for deriving SBM-based edge confidence scores (V). (B) We established correlation-based networks using Spearman’s correlation within mRNA expression, protein expression and metabolite biomolecular data from a subgroup of cancer patients. Shown are histograms of the correlations obtained for all edges of the networks. (C) We only kept edges in the network that had a correlation which differed significantly from zero (sign. corr.). Based on this criterion, for different multiple testing correction methods (BH: Benjamini-Hochberg, BF: Bonferroni) and significance levels (0.05, 0.01), different degrees of reduction can be achieved. We chose for each data layer the most stringent threshold (highlighted in blue) that reduced the edge count to less than $10^7$ edges (left) while keeping the percentage of nodes in the largest connected component high (right). (D) Histogram of the node degrees of the networks reduced by criterion on significance of correlation (red) or on scale-freeness (blue) on double log-scale. For scale-free networks, a linear relationship between log-frequency and log-node-degree is expected.
We converted the adjacency matrices of the reduced networks to obtain for the posterior odds ratio \( \Lambda \):

\[
\Lambda = \frac{P(b_2, SBM_2)P(G | b_2, SBM_2)}{P(b_1, SBM_1)P(G | b_1, SBM_1)},
\]

We can consider the SBM type with the lowest minimal description length most likely; the distance of the posterior odds ratio to 1 determines how much more likely it is than the other SBM type.

We performed 500 runs with random initial partitions for each of the four SBM types and each of the six networks. The SBM type with lowest minimal description length was used for further analyses.

### Overrepresentation analysis

Within the mRNA and protein networks, biological annotation of the blocks and overrepresentation was performed using Reactome pathways and the package ReactomePA\(^8\). We restricted our analysis to Reactome pathways containing at least 10 and at most 500 annotated genes, all entities of the network were used as background. We employed Benjamini-Hochberg multiple-testing correction. Pathways were considered enriched for the parameter \( \lambda \) for a given partition \( b \) and thus

\[
P(G | b) = P(G | \lambda, b) \cdot P(\lambda).
\]

Because the probability of the network itself, \( P(G) \), is constant for a fixed observed network, the description length is monotonically inversely related to the probability that partition \( b \) is responsible for the observed network \( G \), \( P(b|G) \). Therefore, finding the partition which minimizes the description length is equivalent to finding the partition with maximal posterior probability, \( P(b|G) \). Furthermore, as can be seen in its second term in the first line, the description length \( DL \) contains a “penalty” term for the complexity of the SBM description. Thus, it can be used to distinguish which of the four examined SBM types (classical, hierarchical, degree-corrected, hierarchical+ degree-corrected) is most suitable to describe the network \( G \).

Due to the large sizes of the networks, sampling over the posterior distribution is costly. Therefore, we decided to compare different SBM versions using only the estimated maximum of the posterior, i.e. at the partition \( b \) with the lowest minimal description length, in the expression of the posterior odds ratio\(^8\). Assuming that both SBM types that are compared, \( SBM_1 \) and \( SBM_2 \), are equally probable \( a \) priori, \( P(SBM_1) = P(SBM_2) \), we obtain for the posterior odds ratio \( \Lambda \):

\[
\Lambda = \frac{P(b_1, SBM_1)P(G | b_1, SBM_1)}{P(b_2, SBM_2)P(G | b_2, SBM_2)} = \frac{P(G | b_1, SBM_1)P(b_1)}{P(G | b_2, SBM_2)P(b_2)} = 2^{-(DL_1-DL_2)}.
\]

We performed 500 runs with random initial partitions for each of the four SBM types and each of the six networks. The SBM type with lowest minimal description length was used for further analyses.

---

**References**

1. TCGA_Breast_BI_Proteome_CDAP_Protein_Report.r3/Protein_data/CDAP/TCGA_Breast_BI_Proteome.itraq.tsv
2. gdac.broadinstitute.org_BRCA.Merge_Clinical_Level_1.2016012800.0.0/BRCA.clin.merged.txt
3. Network reduction
   - The Hmisc package\(^7\) was used to determine the p-value associated to the correlation (significance of the correlation being different from zero). We corrected for multiple testing using Bonferroni (mRNA expression) or Benjamini-Hochberg (protein, metabolite) correction. Correlations were considered significantly different from zero for corrected p-values lower than 0.05 (mRNA, metabolite) or 0.01 (protein). For the reduction by imposing a scale-free architecture of the reduced network, we employed the `pick_hard_threshold` function of the WGCNA package\(^1\) with the default requirement (0.85) on goodness of fit to a power-law degree distribution of the nodes.

4. Fit to SBM
   - We converted the adjacency matrices of the reduced networks to edge lists in csv format, added the disconnected nodes and fed the resulting networks as graphs into the Python graph-tool\(^3\) framework. Using an agglomerative multi-level Markov Chain Monte Carlo algorithm\(^5\), that module allows for determining a (potentially hierarchical) partition, \( b \), of the network, \( G \), which minimizes the description length, \( DL \), of the SBM\(^4\):

\[
DL = -\log_2(P(G | \lambda, b) - \log_2(P(\lambda, b)) = -\log_2(P(G | \lambda, b) \cdot P(\lambda, b)) = -\log_2(P(G | b) \cdot P(b)) = -\log_2(P(b | G) \cdot P(G)).
\]

Thereby, \( P(A | C, D) \) denotes the probability of \( A \) given \( C \) and \( D \), and \( \lambda \) captures all parameters of the model apart from the partition, such as the number of edges between blocks and degree distribution parameters for the degree-corrected SBM, in their planar or hierarchical version. Note that the above relationship holds only under a microcanonical formulation of the priors, i.e. hard constraints imposed on the values of the model parameters \( \lambda \) by the structure of the network \( G \) the SBM is fitted to (see 34), which enables considering only one value for the parameter \( \lambda \) for a given partition \( b \) and thus

---

**F1000Research 2019, 8:465 Last updated: 14 APR 2019**
For the metabolite dataset, we mapped the pubchem IDs and metabolite names to KEGG compound IDs using the MBRole webservice version 2.36 and merged them semi-manually, thereby preferring metabolite names (in case of mismatch with pubchem record) and KEGG IDs with pathway annotation. We downloaded the human KEGG pathway annotation for the mapped metabolites from MBRole and used it as user-defined annotation for overrepresentation analysis with the enricher function from the package clusterProfiler.37 We only considered pathways with a minimal size of 2 and otherwise the same settings as for mRNA and protein overrepresentation analysis.

Distance measures of Reactome terms and within hierarchical SBMs

We employed two distance measures of Reactome pathways considering the graph given by the Reactome familial hierarchy tree: (i) the distance in terms of number of steps necessary in the Reactome hierarchy graph to reach from one pathway to the other (distance 1), and (ii) the hierarchy level of the lowest common ancestor, or least common subsumer, for ontology graphs (distance 2). We incorporated an artificial top-Reactome pathway into the hierarchy to connect all pathways with each other and to have our distance measures well-defined. It lies at the 10th hierarchy level, so the largest distance between two pathways is 10 for distance 1 and 18 for distance 2 (nine steps to the highest level and nine back). For all comparisons, only blocks with at least one overrepresented pathway were considered. For comparing the distances of Reactome pathway annotations between blocks, i.e. to associate a distance to a pair of blocks, we median-averaged the distances over all combinations of Reactome pathways associated to the two blocks. For distances of pathway annotations within blocks, we median-averaged the distances between all possible combination of different Reactome pathways associated to the block. The trivial distances of zero for the distance of a Reactome pathway to itself were omitted, as well as blocks with only one enriched Reactome pathway for intra-block distances. Distances between blocks in the SBM hierarchy, as employed in Figure 3D, were defined analogously to the distance measure (i) based on the graph of the SBM hierarchy.

WGCNA clustering

Alternative clustering using the WGCNA package3 was performed following the WGCNA tutorial on clustering. We employed the full correlation matrices including self-edges. First, the correlation values were scaled to values between zero and 1 (by (1+corr)/2), and the soft thresholding delivering scale free network topology was determined using the pickSoftThreshold fromSimilarity function with default settings. Power estimates were 8, 16 and 12 for the mRNA, protein and metabolite network, respectively. We calculated the dissimilarity using TomSimilarity on the soft thresholded correlation matrix, used hclust with method “average” and cutreeDynamic with “deep-Split” parameter of 4 (mRNA) or 2 (else), “panRespectsDendro” set to FALSE and “minClusterSize” of 3 (for mRNA, metabolite) or 4 (for proteins) to make the clustering most similar to the one obtained from the SBM.

Edge prediction

We aim to derive edge confidence scores for each single edge in the network exploiting the representation of the network as SBM. Let us consider a fixed network given as graph $G$. If $\Delta G$ is a set of edges which do or do not occur in the network, the probability that these edges belong to the observed network (for edges missing in $G$) or do not belong to the observed network (for edges from $G$), $P(\Delta G, G)$, can be written as:

$$P(\Delta G|G) \propto \sum_b P(G+\Delta G|b) P(b|G)$$

for $b$ being partitions of the network $G$ (please refer to Fit to SBM for further information on notation). The derivation assumes that the original network, $G$ and the altered network with edges in $\Delta G$ added or removed, $G + \Delta G$, has been generated by some SBM type (and all probabilities are conditional on that SBM type), and that the set $\Delta G$ has been chosen by some uniform distribution among all possible edges. The proportionality factor between the expressions depends on the network $G$ and the number of edges in $\Delta G$, and thus can be in particular neglected if only comparing edge confidences between single edges of a network. Because we aim to score all edges, and due to the sizes of the networks making computations slow, we refrained from sampling over the posterior distribution of the partitions and instead employed the single-point approximation for edge prediction proposed in 27:

$$P(\Delta G|G) \propto \sum_b P(G+\Delta G|b) P(b|G) \approx P(G+\Delta G|b^*) = P(G|b^*) P(\Delta G|b^*)$$

It resorts to neglecting the summands for all partitions except for the one, $b^*$, which contributes most to the posterior distribution, $h^* = \max_b P(b|G)$. In addition, the approximation relies on the assumption that the estimated optimal partition for the representation of $G$ by the SBM is the same for $G$ and its altered version with added or removed edge, $G + \Delta G$, which is reasonable for our application case of single edge predictions, i.e. for $\Delta G$ being composed of a single putatively missing or spurious edge. The term $P(b^*|G)$ can be considered constant for a fixed G and SBM type, so we can shift it to the proportionality factor. Considering the microcanonical formulation of the SBM (see Fit to SBM), it becomes clear that the edge predictions for $\Delta G$ directly depend on the difference between the description length of the original network $G$ with partition $b^*$, $DL_{\Delta G,b^*}$, and the description length of the altered network $G + \Delta G$ with partition $b^*$, $DL_{G + \Delta G,b^*}$:

$$P(\Delta G|G) \approx \frac{P(G+\Delta G|b^*)}{P(G|b^*)} = \frac{P(G+\Delta G|b^*) P(b^*)}{P(G|b^*) P(b^*)} = \frac{2^{2DL_{\Delta G,b^*}}}{2^{DL_{\Delta G,b^*}}} = 2^{DL_{\Delta G,b^*} - DL_{\Delta G,b^*}}.$$ 

The difference between the description lengths, $DL_{\Delta G,b^*} - DL_{\Delta G,b^*}$, was computed via the function get_edges_prob from graph-tool.32 Note that as of time of writing, get_edges_prob does not work for weighted SBMs with real-normal edge covariate (see file issue #452 at graph-tool.skewed.de). In addition, this function employs the natural (instead of the dual)
logarithm and consequently, the obtained value has to be scaled by \(\log_e\) to obtain the plain difference of description lengths.

In order to make clear that neither these edge predictions nor their dual (or natural) logarithm correspond to actual probabilities, we used the term edge confidence scores or simply edge scores throughout the manuscript.

**Results**

**Data preparation, network generation, and reduction**

We showcase the methods for data from a subgroup of breast cancer patients: those with estrogen receptor negative (ER-) tumours. ER status is predictive of patient outcome, with ER-leading to unfavourable prognoses, and its assessment is part of standard breast cancer patient screening\(^{36}\). The Cancer Genome Atlas (TCGA) initiative has provided data on the mRNA level (from RNAseq) for several hundred patient tissue samples; mass-spectrometry-based proteomic data (4plex iTRAQ) are available for a subgroup of 36 patients\(^{41}\). Metabolomics data have been measured by GC-TOF-MS in a different breast cancer cohort study for 67 samples\(^{29}\).

We treated each measured entity as node and established a correlation-based, weighted, biomolecular network for each single measurement layer. Therein, each pair of nodes is connected by an edge for which the weight is determined by Spearman’s correlation of the measurements of the nodes (over the samples), delivering values between -1 and 1. Thus, an edge that connects nodes with a correlation close to 1 or close to -1 represents strong positive regulation and strong negative regulation, respectively; edges connecting nodes with a correlation close to zero represent weak or absent regulation. We employed Spearman’s correlation because it captures also non-linear relationships between measurements, and it is robust to outliers and any monotonous transformation (e.g. logarithmization). We dealt with missing values in the data by replacing them by small values (NAs due to log transformation of zero counts in the mRNA data) or removing entities with >20% missing values for all samples (for the protein data). Subsequently, we computed the correlation only considering pairwise complete observations. The distributions of the computed correlation values for the three data types are shown in Figure 1B.

The resulting biomolecular networks capture the relationships of the intracellular machinery, and their analyses deliver important insights on altered regulations in disease states\(^{37}\). However, because these networks are fully connected, i.e., every entity is connected to each other entity within a layer, the networks become very large and their analyses difficult. A common approach is to reduce the networks, either by selecting a subset of entities as nodes prior to network establishment, mainly by using criteria on abundance, or by using the assumption that weak regulations are less important for the biological network and can be omitted without impairing the represented function of the network. We decided on the latter approach in order not to bias our choice of considered molecular entities.

We used two different techniques of network reduction by thresholding: In the first, we only kept edges for which the correlation was significantly different from zero (“sign. corr.”), i.e., the regulation being sufficiently strong. We adapted the significance thresholds for the different molecular data layers to their respective sizes. Larger networks were subjected to more stringent cut-offs in order to achieve an appropriate degree of reduction (Figure 1C, employed thresholds marked in blue). In the second reduction method, we removed weak links until the reduced network met a criterion of scale-freeness (3, “scale-free”). Scale-freeness is considered a key property of self-organized networks and thus also of biological molecular networks\(^1\). In scale-free network formation, highly connected nodes tend to attract more connections than lowly connected nodes leading to a degree distribution following a power-law; the degree distributions of the reduced networks are shown in Figure 1D. Both reduction techniques are hard-thresholding techniques meaning that edges are removed from the networks. The resulting six reduced networks, two for each data layer, were used in a binary form, i.e., weight information was discarded after reduction. Some characteristics of the original and reduced networks are shown in Table 1; the node count of the reduced networks coincides with the number of entities considered after missing value removal. In the following sections, we describe how to analyze these different homogeneous cancer networks by fitting them to SBMs.

**Fitting the reduced networks to stochastic block models**

Biological networks are known to be modular and hierarchical. Different molecular entities, such as genes, mRNAs, and proteins, are interconnected and form different modules to fulfill a specific function. Modularity can convey more robustness to the overall system, e.g. by preventing perturbations in single modules to spread fast and to cause erroneous behavior in other modules and thus functions. Hierarchies capture two characteristics of biological systems: (i) the ordered combination of functions, i.e., multiple simple functions resulting in more complex behavior or responses; and (ii) the inherent levels of complex organization of life, from single molecules to cell organelles, cells, tissue, organs and whole organisms.

The stochastic block model (SBM) is the simplest form of a generative network model based on communities, i.e., group structures of the nodes. Thereby, nodes are assigned to blocks according to their connectivity properties (Figure 2A left); the block associations of two nodes fully determine the probability of an edge between them. A shortcoming of the classical SBM for representations of real networks is that nodes within one block need to have similar degrees. The degree-corrected version of the SBM\(^{42}\) accounts for that and enables different degrees for nodes within a block. Another extension of the SBM is its hierarchical version, in which the blocks are further partitioned into blocks of higher levels\(^{27}\) (Figure 2A right). This model is especially suitable to represent large networks with many nodes. Fitting biological networks to hierarchical (also called nested) SBMs is therefore most appropriate.

We assessed which of the four following types of stochastic block models could best represent the biological networks: the classical SBM, the degree-corrected SBM, the hierarchical SBM or the degree-corrected and hierarchical SBM (Figure 2B). We used the Python module graph-tool\(^{12}\) to fit SBMs to the
networks, i.e., to find which partition (basal and/or hierarchically ordered) describes the network best as SBM.

Note that we also examined the performance of weighted stochastic block models for our purpose of edge prediction as they have been successfully employed before for non-biomolecular networks\cite{18,27,43}. However, the characteristics of the optimal weighted SBM (number of blocks) were severely impacted by prior assumptions on the edge weight distributions, and the derived edge confidence scores did not coincide well with evidence on edge relevance given by the correlations of the edges for fully-connected weighted networks (see Figure S1\cite{44}). Taking in addition the increased computational effort for fitting the weighted SBM compared to the binary version into account, we restricted our further analyses to non-weighted SBMs.

The model fit is performed following the rationale of Occam’s razor: The simplest model describing the data is the best. Thus, we searched for the partition that minimized the description length, i.e., the amount of information necessary to describe the network as an SBM. Additional information required to capture degree-correction and/or hierarchies compared to the classical SBM is thereby taken into account. Consequently, it can be directly concluded which of the four SBM model types is most appropriate for a certain network: The one with the lowest minimal description length. The optimization of the description length runs via an agglomerative Markov Chain Monte Carlo algorithm\cite{33}. It is non-deterministic and multiple initiations of the underlying partition of the SBM are required to obtain globally instead of locally optimal partitions. We performed optimizations for 500 initial partitions for each network and SBM type.

For all four mRNA and protein networks, the classical SBM delivered the worst fit and the degree-corrected, hierarchical SBM fitted best (Figure 2B left and middle). Degree correction did not prove necessary to describe the metabolite networks, for which the hierarchical SBMs fitted slightly better than the classical SBM (Figure 2B right). A graphical representation of the best fitting SBMs in circular layout, showing the blocks from the lowest layer (for mRNA, protein) or the metabolites (metabolite network) and their connections in color, and the hierarchical structure on top in black, is given in Figure 2C.

SBM-derived communities of the biomolecular networks capture biological function

We wanted to assess how well the biological content of the biomolecular networks is captured in the stochastic block models, i.e., if the clustering of the nodes into the blocks is biologically meaningful. To that purpose, we estimated whether the blocks show common biological function based on Reactome or KEGG pathways. In particular, we performed overrepresentation analyses of Reactome pathways for the blocks at each level of the SBMs of the four protein and mRNA networks, and overrepresentation of KEGG pathways for the blocks predicted for the two metabolite networks. We find that a high percentage of blocks in each level has at least one associated Reactome or KEGG pathway, i.e., the genes of the pathway are

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>mRNA</th>
<th>protein</th>
<th>metabolite</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples</td>
<td>237</td>
<td>36</td>
<td>67</td>
</tr>
<tr>
<td>entities (nodes, before NA removal)</td>
<td>18321</td>
<td>10625</td>
<td>162</td>
</tr>
<tr>
<td>edges (before reduction)</td>
<td>167820360</td>
<td>56440000</td>
<td>13041</td>
</tr>
<tr>
<td>reduced network: scale-free</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimal correlation</td>
<td>0.603</td>
<td>0.788</td>
<td>0.375</td>
</tr>
<tr>
<td>edges</td>
<td>287790</td>
<td>85980</td>
<td>1825</td>
</tr>
<tr>
<td>entities of degree zero</td>
<td>8183</td>
<td>4967</td>
<td>4</td>
</tr>
<tr>
<td>entities in largest connected component</td>
<td>9111</td>
<td>3187</td>
<td>158</td>
</tr>
<tr>
<td>reduced network: sign. corr.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minimal correlation</td>
<td>0.395</td>
<td>0.534</td>
<td>0.310</td>
</tr>
<tr>
<td>edges</td>
<td>4260267</td>
<td>2434159</td>
<td>2811</td>
</tr>
<tr>
<td>entities of degree zero</td>
<td>1061</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>entities in largest connected component</td>
<td>17190</td>
<td>8427</td>
<td>161</td>
</tr>
</tbody>
</table>
Figure 2. Stochastic block models for representing correlation-based biomolecular networks. (A) In a stochastic block model (SBM) representation, the nodes of a network are partitioned into blocks according to their similarity in connectivity. The hierarchical version of the SBM (right) imposes in addition hierarchical partitions onto the blocks. (B) Six biomolecular networks derived from transcriptomic, proteomic or metabolomics data of breast tumours were fitted to four different types of SBM: the classical SBM (SBM, light blue), the degree-corrected SBM (dcSBM, light green), the hierarchical SBM (hSBM, dark blue) and the degree-corrected hierarchical SBM (dchSBM, dark green). We performed fits for 500 initial partitions for each network. A fit consists of altering the partitions underlying the SBM such as to minimize the description length. The smaller the description length the better the fit. Hierarchical SBMs outperform non-hierarchical SBMs, degree correction is required for the mRNA and protein networks. (C) Graphical representations of the best fitting hierarchical stochastic block models with degree correction (dchSBM) or without (hSBM) for the networks reduced by significance of correlation (mRNAs, metabolites) or by scale-freeness (protein). The lowest layer (genes, proteins) is truncated in the mRNA and protein networks, colored lines denote edges between blocks of the first level (for mRNA, protein network) or between metabolites (level 0, metabolite network). Edges of blocks or metabolites which belong to the same block in the level above have the same colour. The higher-order hierarchical structure is shown in black.

overrepresented within the block, they occur more frequently than expected by chance (Figure 3A bars). Except for the highest hierarchy levels that consist only of a few blocks, this percentage is decisively higher than for a random clustering of exactly the same structure (results for 3 random clusterings shown as black crosses in Figure 3A).

Many blocks, however, have not only one but multiple Reactome pathways assigned. If blocks are biologically meaningful, we would expect to observe similar Reactome pathways within blocks but less similar pathways when comparing the pathways of different blocks.

The dissimilarity of Reactome pathways is naturally represented by their distance in the Reactome familial hierarchy structure; the more distant pathways are, the less connected are their biological functions. We used two closely related measures to assess it: (i) the distance of two pathways is the length of the
Figure 3. Modules derived from hierarchical SBM represent biological function. (A) Percentage of blocks with at least one overrepresented Reactome (KEGG) pathway for the best fitting SBM (see Figure 2B) for the mRNA and protein (metabolite) networks, reduced by condition on scale-freeness (scale-free) or on significance of correlation (sign. corr.), for each hierarchical level (bars). Black crosses denote the percentages of blocks with at least one overrepresented Reactome (KEGG) pathway for three SBMs each with exactly the same structures but randomly shuffled mRNAs or proteins (metabolites). (B) For the lowest hierarchy level clustering of each of the four mRNA and protein SBMs, we calculated the average distances between every pair of Reactome pathways between blocks and those within blocks, for two different distance measures based on the Reactome hierarchy. The lower the distances the more similar are the pathways. For both distance measures, the pathways associated to one single block (within) are significantly more similar than those associated to different blocks (Welch’s t-test p-value < 0.01) suggesting that biological functions are consistent within blocks and distinct between blocks. (C) Distances of Reactome pathways (as in B) between blocks (or within blocks, for distance of blocks in SBM being zero) versus the distance of the blocks in the SBM. We do not find evidence that the Reactome hierarchy is reflected in the SBM hierarchy. (D) Distribution of predicted module (block) sizes for the six networks using SBM (lowest hierarchical level only) or the three networks using network generation and module detection from WGCNA. (E) Percentage of modules detected by the WGCNA approach with at least one enriched Reactome or KEGG pathway. Black crosses denote the results for three similar clusterings with randomly shuffled mRNAs, proteins, or metabolites.
shortest path in the hierarchy tree from one pathway to the other, and (ii) the higher the hierarchy level of the lowest common ancestor (least common subsumer) of two pathways, the more distant they are. Please note that we restricted this analysis to the mRNA and protein networks because the KEGG pathways employed for metabolite data are only assorted into a very shallow hierarchy.

We compared the distances of Reactome pathways using the average of the two distance measures over pairs of Reactome pathways associated with one block (within blocks) and over pairs of Reactome pathways associated with two different blocks (between blocks) for the lowest hierarchy level blocks, as shown in Figure 3B. Thereby, for all four networks and both measures, we observe significantly smaller distances of Reactome pathways within blocks than between blocks (Welch’s test \( p < 0.01 \)). This suggests that biological function is coherent within and distinct between blocks, thus further enhancing the notion that SBMs represent well the biological function of the networks.

We also examined the relationships of the hierarchies within the SBMs to that of the Reactome pathways. In analogy to the distance of Reactome pathways, we defined the distance of blocks within a hierarchical SBM by counting the number of steps necessary to reach a block from the other (via their lowest common higher-level block). We compared the distance of each pair of blocks within the SBM on the lowest level to the average of the pairwise distances of their associated Reactome pathways (Figure 3C). Therein, we found only a weak positive correlation for Reactome pathway distance measure (i), that is even further reduced if neglecting intra-block coherence, i.e., neglecting blocks with distance zero. We concluded that the hierarchy within the SBM does not strongly coincide with the hierarchy within the Reactome pathways.

In order to assess how the clustering by SBM relates to established clustering techniques in correlation-based networks, we also performed module detection by using the WGCNA package\(^1\). Note that for this approach, no model reduction is necessary, which means that we obtain only one result for each data layer. After soft thresholding to enforce scale-free network architecture, we employed WGCNA module detection to obtain comparable numbers of modules as for the SBM-based approach: 333 (mRNA), 109 (protein) and 12 (metabolite) for WGCNA, which are similar to the 438, 113 and 15 blocks detected by SBM for the corresponding scale free networks. Overall, the WGCNA clusterings show a larger diversity of module sizes than those obtained for the SBM approach (Figure 3D). For all three data layers, a higher percentage of WGCNA-derived modules have a biological annotation compared to the blocks from the SBM (compare Figure 3E to 3A). However, for the mRNA layer, many of the blocks also have an enriched pathway annotation for randomly shuffled gene names which hints on reduced significance of the WGCNA results and better performance for the modules detected by fitting to SBM. For protein and metabolites, in terms of detection of biologically annotated modules, the WGCNA approach seems to provide slightly better results than the SBM approach.

The observed differences could stem from conceptual differences in the approaches: Instead of detecting clusters of entities with highly positively correlated abundances only as in WGCNA, nodes from the same block are characterized by common connectivity characteristics in the SBM. Clearly, as proteins interact and bind directly for complex formation or regulation, and metabolites are interconverted into one another, entities which act together tend to have similar abundances and thus the modularization by WGCNA shows good results. For the detection of modules for mRNAs whose interaction can be considered less direct, assorting entities with similar connectivity patterns as in the SBM-derived modules is beneficial. In addition, the WGCNA framework, as most other module detection methods, cannot aid in assessing edge relevance - which is enabled by the fit of the networks to SBMs.

Assessing edge relevance by SBM-based edge confidence scores

The description of the biomolecular networks as SBMs were exploited to determine a confidence score for each existing or non-existing edge of the network. This score would capture how well the existence (or non-existence) of the edge fits to the network’s description by the SBM: Erroneously kept or removed edges lead to a worse fit of the SBM to the network, and removing or reinstalling these edges lead to fit improvement. Indeed, under certain assumptions (see Methods), the derived edge confidence score is proportional to the actual absolute probability that the edge belongs to the network, and thus signifies relative edge relevance. Therefore, the confidence score could be used to predict whether an edge is spurious (or missing)\(^{13,18,27}\). A high missing edge confidence score suggests that the edge in question is missing and should be restored, it has a high relevance; a high spurious edge confidence score suggests that the edge in question is spurious and should be removed from the network, it has a low relevance. The SBM-based edge confidence scores rely on global network connectivity characteristics and complement the correlation-based weights of the edges which stem from the local, measured characteristics of their nodes.

For each putatively missing and spurious edge, we used the Python module graph-tool\(^1\) to compute its edge confidence score (Figure 4). Thereby, we took advantage of the following fact: Entities of degree zero, i.e., nodes that have no connection to any other node in the network after reduction, are indistinguishable to the SBM. Consequently, also all putatively missing edges connecting any of these nodes to a specific second node are only distinguishable by this second node, and thus carry the same missing edge confidence score. This reduces the number of scores we need to compute considerably, depending on the degree of reduction (see counts of entities of degree zero in Table 1). We display the scores for the different types of missing edges in different histograms (Figure 4A, B middle and bottom). For very big networks with a low degree of reduction (the mRNA and protein networks reduced by significance of correlation), it
Figure 4. SBM-derived edge prediction: Missing and spurious edge confidence scores. (A) Histograms of edge confidence scores for the best fitting SBM of the three networks reduced by criterion on scale-freeness for all existing, i.e., putatively spurious edges (spurious edge score, top), for the putatively missing edges between nodes with degree > 0 (missing edge score, middle), for the putatively missing edges connecting a node of degree 0 to a node with a larger degree (missing edge (disc) score, bottom – only one edge for each node with degree > 0 is shown). (B) Edge confidence scores as described in (A) for the networks reduced by criterion on significance of correlation. For the mRNA and protein network, missing edge scores (middle) were only computed for the edges with largest absolute correlations of the node measurements.
is still computationally not feasible to compute the scores for all
missing edges. We therefore resorted to computing it for as
many missing edges as we have existing edges in the network
(i.e. approx. 2.4×10⁶ for the protein network, 4.6×10⁶ for the
mRNA network, see Table 1), and chose those with highest
absolute edge weights (Figure 4B left and middle).

Recall that the edge confidence scores are relative, i.e., they
serve for comparing relevance between edges only. In addition,
computation of the scores relies on the assumption that the
partition of the originally fitted SBM is correct for the net-
work. Different edge confidence scores might be obtained for
SBMs with different partitions but with similarly good fit to the
network. We neglect this complication for the sake of computa-
tional efficiency. Still, we have to keep both facts in mind for the
interpretation and usage of the edge confidence scores. For
example, an evident threshold for declaring an edge as relevant
(“missing”) or not relevant (“spurious”) would be to have the
respective edge confidence score larger than zero, as this indi-
cates an improvement in fit quality if adding or removing the
dge, respectively. However, we observe an imbalance of our
computed scores which is inherent to the approach: Because
having less edges reduces the complexity of the underlying
network, removing edges preferentially reduces also the amount
of information required to describe the SBM, i.e., its description
length turns smaller, its goodness of fit improves. Therefore, the
eedge confidence score distributions of missing edges are shifted
to the left - the great majority of edges would be predicted as not
missing for the threshold of zero, they are not relevant and should
be left out (Figure 4A, B, 2nd line). Conversely, the spurious
edge confidence score distributions are shifted to the right - the
great majority of existing edges would be predicted as spurious,
they are not relevant and should be removed for an edge score
threshold of zero (Figure 4A, B, top). Both measures point to
making the networks smaller.

**Evaluation of edge confidence scores based on correlation**

In order to determine whether the edge confidence scores are
overall a reasonable assessment of edge relevance, we compared
the predicted scores to the edge weights of the edges as derived
directly from Spearman’s correlation of the measurements of
the nodes. It is important to note that these edge weights
(correlations) were used exclusively for the reduction of the
correlation-based networks. The edge correlations were at no
point provided to the SBM, neither during the fitting to the
SBM (except for the weighted version and only for Figure S1,
Figure S2) nor during the computation of the edge scores.
Thus, they are close to an independent validation of the edge
relevance, edges with large correlation being more relevant to
the system encoded by the network than edges with a correlation
close to zero.

Indeed, for all six networks and SBMs, we find an overall posi-
tive correlation between the (absolute) edge correlations and the
missing edge confidence scores: Edges with high correlation are preferentially predicted as missing, i.e., relevant to the
network, also in terms of edge confidence score (Figure 5).
Similarly, we find an overall negative correlation between

**Discussion**

Using example cases of correlation-based transcriptomic,
proteomic and metabolomics networks from breast cancer
tumour samples, here we show how stochastic block models
can be employed for the analysis of biomolecular networks. The
networks can be best represented by the hierarchical version
of the stochastic block models. This gives rise to biologically
meaningful separation of the biomolecules into many function-
ally relevant blocks. In addition, the SBM framework enables
the computation of edge confidence scores that can be used to
predict missing or spurious links.

The representation of the networks by SBMs poses a challenge:
The model fit and the derivation of edge confidence scores can
be computationally very demanding, especially for networks
with many edges. This was the case here for the mRNA networks
and the protein network reduced by significance of correlation.
Therefore, the approach of network analysis by SBM seems most
suitable for smaller and/or sparser networks. On the other hand,
it delivers two opportunities. First, modules derived from blocks
in SBMs are not only defined as clusters of tight relationships,
i.e. co-expression clusters, as obtained with other clustering
approaches⁴⁵-⁶µ. In an SBM, nodes from the same block are
characterized by common connectivity characteristics, i.e.,
common interaction profiles with nodes from the same and from
other blocks. Consequently, comparing SBM-derived modules
between conditions naturally points towards detecting altered
regulations. This has not been investigated here, the biological
role and disease-relevance of the blocks in our example case of
ER- breast cancer still need to be assessed.

Second, the derived edge predictions can reproduce the interac-
tion strengths as estimated from measurements. SBM-based edge
confidences tend to score missing edges higher that have high
correlation, i.e., that would be considered a strong regulation,
an important edge. Analogously, existing edges in the network
with low correlation tend to be predicted as spurious with high
correlation, i.e., as dispensable. Keep in mind that the correla-
tions corresponding to the edges were not provided at any point
to the SBM construction. Thus, the SBM approach proves very
strong in delivering relevant edge predictions. A further biological
validation of the predicted edge confidence scores, for example
by comparison to interaction databases, would be an interesting
next step.

Still, a question remains: How should the edge confidence scores
be translated into edge predictions to alter the network, i.e. to
actually remove or re-install edges? There is a natural threshold
for declaring an edge as missing or spurious, namely if the
Figure 5. SBM-derived edge prediction: Validation by correlation. Relationship between edge confidence scores and edge correlations (Spearman’s correlation of the measurements at their nodes) for putatively spurious (top) or putatively missing edges between nodes of degree > 0 (bottom) for the networks reduced by criterion on scale-freeness (A) or significance of correlation (B). Edge confidence scores correlate well to edge correlations, with edges with high SBM-predicted confidence in spuriousness – edges predicted as irrelevant – having lower correlations, edges with high SBM-predicted confidence in missingness – edges predicted as relevant – having higher correlations. Please note that for the protein network reduced by significance of correlation, due to missing values in the dataset, a fixed significance threshold leads to different correlation thresholds and thus the correlation value boundary is blurred (B, middle).

respective confidence score would be larger than zero. However, due to the minimal description length approach for SBM-based edge relevance assessment, the confidence scores are shifted towards reducing the networks as much as possible (Figure 4), such that this natural threshold for deriving the prediction of actual missingness or spuriousness from the confidence scores is not valid. Further examinations on other possible thresholds are required. However, for relative comparisons of relevance between edges in a network the SBM-based scores are suitable.

Additional directions of SBM-based analysis of biomolecular networks remain to be explored.

(i) For edge relevance assessment in other network types, it has been proposed that edge predictions with SBMs are more reliable if resorting to an ensemble of good fits instead of using the best fit only\textsuperscript{27}. Due to the sizes of the employed correlation-based biomolecular networks, this approach is computationally not of practical relevance here, but it would be an interesting point to assess in the future.

(ii) We considered the weighted version of the SBM that could be an interesting option for the analysis of biomolecular networks because it enables representing fully-connected weighted networks as SBM without prior reduction. It could prove exciting especially for smaller networks, e.g. primary
metabolites, or in more targeted data analysis approaches. However, in our case, we found hints that the weighted version of the SBM might not be appropriate for the task of edge prediction from SBM-based confidence scores for fully-connected networks (see Figure S1). For reduced networks, the results for weighted SBMs seem more promising (see Figure S2). A final assessment on the usefulness of the weighted version of the SBMs is still pending, as long as nothing is known about interaction strength distributions between and within modules: The fitted SBM strongly depends on the prior assumptions chosen for these distributions. An extensive analysis of different choices would be required, which is beyond the scope of this study.

We propose to employ SBM-based modules and edge confidence scores as additional pieces of information to make the results of follow-up analyses more robust that rely on relationships between nodes, i.e., on edges and their strength. Edges and their strength play a key role in interpreting biomolecular effects, e.g., of mutations or drugs. If a drug “activates a protein”, this typically means alteration of the protein’s interaction strengths with other proteins or the DNA. The “mutation of a gene” may severely alter the binding properties of the corresponding translated protein to interaction partners. In sum, interactions and their strengths are at the heart of biologically relevant alterations in biomolecular networks, and characterizing the former is required in order to understand the effects of the latter. The SBM framework enables assessing the edge relevance or interaction strength on the basis of consistent, global network characteristics, instead of on the basis of correlation of measurements. Especially for personalized analyses which generally rely on a characterization by only few error-prone measurements of each molecule, this will be crucial to derive more reliable predictions.

Data availability

Underlying data

In this study, data from TCGA and CPTAC was used. Proteomics data stem from 41, metabolomics data stem from 29. The metabolomics raw data can only be obtained upon accessing the cited article; processed data (Spearman’s correlations and associated p-values) can be found in the gitlab repository below.

Code used to perform the analyses together with a detailed work-flow documentation: https://gitlab.com/biomedliib/sbm-for-correlation-based-networks

Archived code as at time of publication: http://doi.org/10.5281/zenodo.262041

License: GNU GPLv3 license.

Extended data

Zenodo: Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models. http://doi.org/10.5281/zenodo.262041. This project contains the following extended data files:

- Figure S1: The weighted SBM seems not appropriate for edge prediction from edge confidence scores for fully connected networks
- Figure S2: Edge predictions for a reduced weighted network with planar or hierarchical weighted SBMs

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Grant information

This research was funded by Luxembourg National Research Fund, FNR, SINGALUN project. KB acknowledges funding by an Add-on Fellowship for Interdisciplinary Life Sciences of the Joachim Herz Stiftung.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References


Harrell FE Jr and with contributions from Charles Dupont and many others: Hmisc: Harrell miscellaneous. 2018. Reference Source


Peixoto TP: The graph-tool python library. figshare. 2014. Publisher Full Text


The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com