RESEARCH ARTICLE

REVISED Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models [version 2; peer review: 1 approved, 2 approved with reservations]

Katharina Baum[1,2], Jagath C. Rajapakse[3], Francisco Azuaje[1]

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

v2 **First published:** 14 Apr 2019, **8**:465
https://doi.org/10.12688/f1000research.18705.1
Latest published: 27 Aug 2019, **8**:465
https://doi.org/10.12688/f1000research.18705.2

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: The key challenge we address here is investigating the capability of stochastic block models (SBMs) for representing and analyzing different types of biomolecular networks. Fitting them to SBMs both delivers modules of the networks and enables the derivation of edge confidence scores, and it has not yet been investigated for analyzing biomolecular networks. We apply SBM-based analysis independently to three correlation-based networks of breast cancer data originating from high-throughput measurements of different molecular layers: either transcriptomics, proteomics, or 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 biologically and phenotypically relevant

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 27 Aug 2019	✓ view		? view
version 1 14 Apr 2019	? view	? view	✗ view

1. **Lei Xie**, The City University of New York, New York, USA
Yue Qiu, City University of New York, New York, USA

2. **Silvia Pineda San Juan**, University of California, San Francisco (UCSF), San Francisco, USA
Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

3. **Ana Conesa**, University of Florida, Gainesville, USA
Manuel Ugidos, Institute of Biomedicine of

functional annotation. The edge confidence scores are overall in concordance with the biological evidence given by the measurements. We conclude that biomolecular networks can be appropriately represented and analyzed by fitting SBMs. As the SBM-derived edge confidence scores are based on global network connectivity characteristics and potential hierarchies within the biomolecular networks are considered, they could be used as additional, integrated features in network-based data comparisons.

Keywords
biomolecular networks, co-expression networks, edge relevance, hierarchical stochastic block model, missing and spurious edges, module detection, network-based omics analysis

This article is included in the Bioinformatics gateway.

Valencia (IBV), CSIC, Valencia, Spain

Any reports and responses or comments on the article can be found at the end of the article.

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

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 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2019, **8**:465 https://doi.org/10.12688/f1000research.18705.2

First published: 14 Apr 2019, **8**:465 https://doi.org/10.12688/f1000research.18705.1

REVISED **Amendments from Version 1**

In the revised version of our manuscript, we reformulated parts of the Abstract and Introduction in order to clarify the goal of our analysis: Assessing whether stochastic block models (SBMs) can be used to represent and analyze single-layer biomolecular networks derived from different, clinically relevant datatypes that provide complementary perspectives of cellular or tumour tissue properties.

We included a more detailed description of the network reduction procedures and the applied reasoning in the Methods and Results sections, along with new Figure 1D, as well as the relationship between the two employed reduction methods in new Figure S1. We incorporated a description of the SBM types and their associated formalisms that connect them to the properties of the modelled network into the Methods section.

We clarified the contents of Figure 3 by adapting its description and shifting contents to new Figure S3, and we included pathway distances of the higher-level blocks (new Figure 3D).

The most comprehensive extensions consist of new analyses with a focus on biological predictions. First, we compare biological insights derived from SBM-based clustering to known characteristics of breast cancer, that is the phenotype we employ for our case study, as derived from MSigDB (new Figure 4). Second, we summarize predictions of the SBM-based clustering (new Figure 5, new Table 2) and relate them to (breast) cancer literature. Together with the strikingly tight relationship between SBM-based edge scores and data-derived edge interaction strength (correlation), these analyses support our conclusion that SBMs are suitable to represent and analyze biomolecular networks and to derive relevant predictions.

Any further responses from the reviewers can be found at the end of the article.

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 promotors, mRNA, metabolites, proteins and their activity – not separately, but in their network context[1]. 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[2–4]. 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 deleterious 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[5–14]. 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[15–23]. 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)[24].

SBMs are the simplest form of generative network models for community structures and can also accommodate hierarchies[25], 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[15], and the SBM has been used for representing other types of networks[17,18,26,27], 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.

The key challenge we address here was to investigate the capability of SBMs for representing and analyzing different types of biomolecular networks. We aimed to assess to which extent the SBM is applicable to derive useful information in terms of (i) relevant clustering as well as (ii) network-based, alternative edge scores. Therefore, we showcased the SBM-based analysis (overview in Figure 1A) for three biomolecular networks of different molecular types, derived from either transcriptomic,

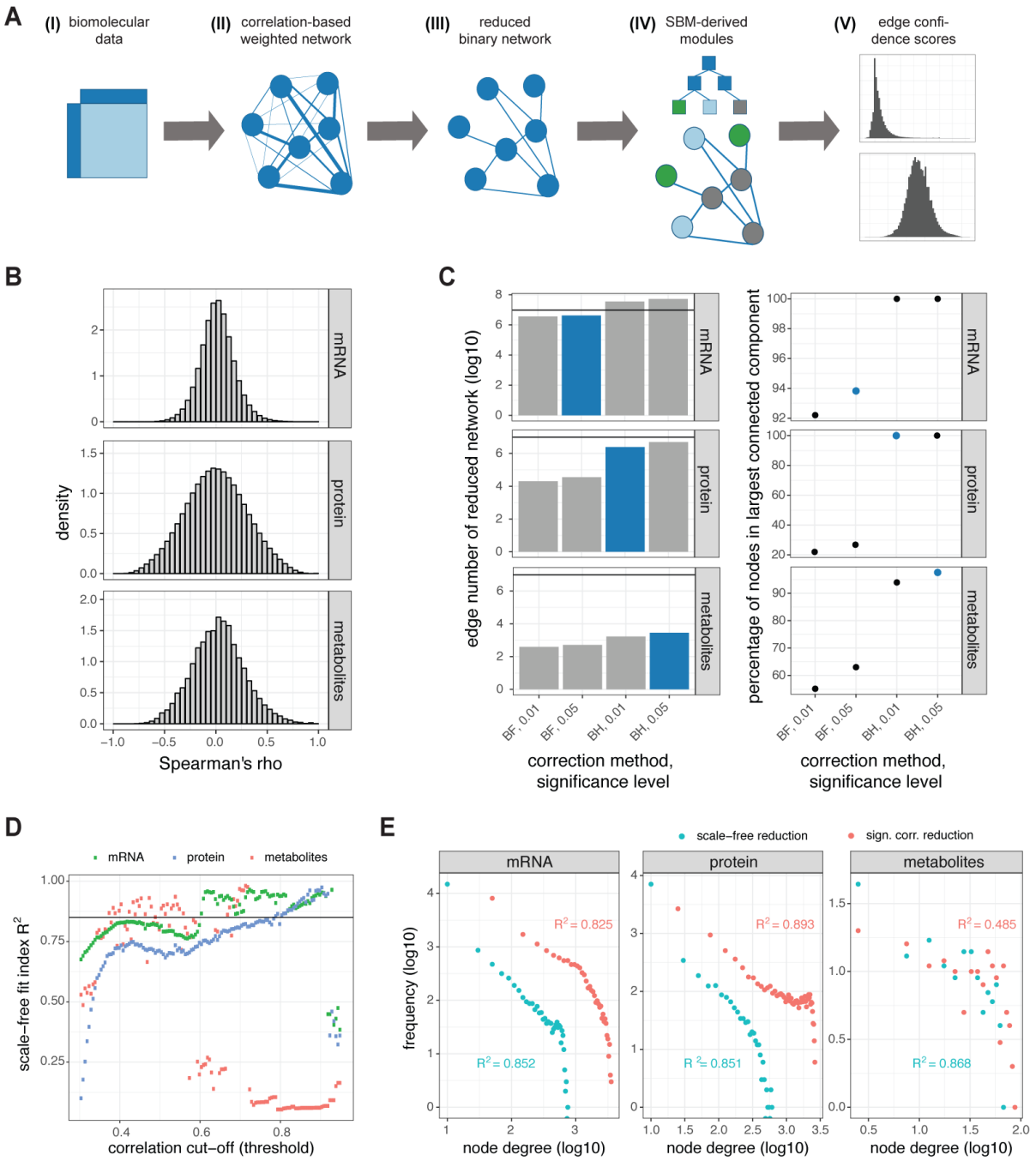

Figure 1. Analysis of correlation-based networks from different omics data: Pipeline and network preparation. (**A**) Pipeline for the approach. Given the matrix of measurements of mRNA expression, protein expression or metabolite abundance for a group of samples (I), we compute a fully-connected weighted 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.). 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**) Scale-free fit indices R^2 according to 3 for the mRNA (green), protein (blue), and metabolite (red) network reduced by different correlation thresholds between 0.3 and 0.95. The employed threshold of 0.85 for the scale-free fit index is indicated by a black line. (**E**) 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 together with the scale-free fit indices R^2. Networks are the more scale-free the more linear the relationship between log-frequency and log-node-degree is.

proteomic or 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, compared the SBM-predicted function to breast cancer signatures and showed how to derive additional predictions from SBM-based clusters. 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[28] downloading TCGA level 3 preprocessed BRCA files (search term: *mRNAseq_Preprocess*) on Nov 2, 2017. We used the normalized RSEM values. Protein data was obtained via the CPTAC homepage from the data generated in 29. We used the first replicate of samples measured in duplicates. We employed the unshared log ratio value for each sample to maximize reliability of protein identification. Clinical data for both TCGA and CPTAC data was retrieved and evaluated using the RTCGA package. Specifically, 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*
- *TCGA_Breast_BI_Proteome_CDAP_Protein_Report.r3/Protein_data/CDAP/TCGA_Breast_BI_Proteome.itraq.tsv*
- *gdac.broadinstitute.org_BRCA.Merge_Clinical.Level_1.2016012800.0.0/BRCA.clin.merged.txt.*

For both mRNA and protein data, we only used samples from patients whose entry "patient.breast_carcinoma_estrogen_receptor_status" in their clinical data was "negative". In addition, we restricted our analysis to solid tumour samples, i.e. TCGA sample identifiers ending with 01. This gave rise to 237 samples with 18321 measured genes for the mRNA data and 36 samples with 10625 measured proteins for the protein data.

Metabolite data was used from the Excel file provided in 30 using the measurements of 162 metabolites in the 67 samples containing ERn in their label.

Missing values

We removed missing values in the mRNA data by replacing them by -10 to account for the fact that they arose from logarithmizing read counts of zero. In the protein data, we removed the 2195 proteins which had more than 20% of missing values over the considered samples (i.e. more than 7 NAs among the 36 samples), resulting in 8430 measured proteins that we analyzed further. The metabolite data did not contain missing values as imputation had been performed in the original publication[30].

Network generation

We used the correlation computation from the Hmisc package[31] (function *rcorr*) to determine Spearman's correlation of the measurements of each pair of entities (mRNA, protein or metabolite) over all samples. Only pairwise complete observations were employed. Unless stated otherwise, we neglected self-edges.

Network reduction

The Hmisc package[31] was used to determine the p-value associated to the correlation (significance of the correlation being different from zero). For each data layer, we assessed four different combinations of multiple testing correction and significance thresholds: Bonferroni or Benjamini-Hochberg[32] multiple testing correction combined with significance thresholds of 0.01 or 0.05. The resulting reduced networks were characterized in terms of edge count and largest connected component. For each data layer, the network was chosen that yielded a sufficient degree of reduction ($< 10^7$ edges) to enable a sufficiently short computation time and memory consumption for the fit to SBM while maintaining at the same time a high percentage of nodes being connected to each other as would be expected in biological networks. Finally, Bonferroni correction was chosen for the mRNA network, Benjamini-Hochberg correction for the smaller protein and metabolite networks. 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 *pickHardThreshold* function of the WGCNA package[3] with the default requirement (0.85) on goodness of fit to a power-law degree distribution of the nodes. Given the symmetric absolute correlation matrix of the network edges, this function reduces the network by one of a given set of edge thresholds at a time and determines the scale-free fit index R^2 which lies between 0 (bad fit) and 1 (perfect fit) by comparing the resulting degree distribution of the reduced network to a power-law degree distribution. The lowest of the tested edge thresholds that gives a scale-free fit index > 0.85 is reported as estimated threshold. For the edge thresholds, we started with a grid with stepsize 0.05 between 0.3 and 0.95, refining according to the resulting estimates to vectors with stepsize 0.001 between 0.5 and 0.625 for the mRNA network, between 0.7 and 0.82 for the protein network, and between 0.3 and 0.4 for the metabolite network. Finally estimated edge correlation thresholds were 0.603 (mRNA), 0.788 (protein), and 0.375 (metabolite).

Fit to SBM

We employed four versions of the stochastic block model (SBM) derived from three SBM types (classical, degree-corrected, hierarchical SBM) and their Bayesian description[33,34]. In the

classical SBM[24], the model is fully defined by a partition b of the nodes into blocks and the matrix $e = \{e_{rs}\}$ of the numbers of edges between blocks (with e_{rr} being double the edge count within a block for convenience). We employ microcanonical formulations which imposes hard constraints on the values of the model parameters according to the observed graph G[34]. Given a graph G and a partition b, the probability that the graph was generated with the observed edge count matrix e, $P(G|b)$, can be described as

$$P(G|b) = P(G \mid e,b) \cdot P(e).$$

Thereby, $P(A|C, D)$ denotes the probability of A given C and D. With $A = \{A_{ij}\}$ being the adjacency matrix of the (multi)graph G, with $e_r = \sum_s e_{rs}$ the number of edges adjacent to block r, and n_r the number of nodes in block r according to the partition b, we employ

$$P(G \mid e,b) = \frac{\prod_{r<s} e_{rs} \prod_r e_{rr}!!}{\prod_r n_r^{e_r} \prod_{i<j} A_{ij}! \prod_i A_{ii}!!}$$

using the definition $(2m)!! = 2^m \cdot m!$. In addition, for $n_e = 2E/(B(B+1))$ the expected total number of edges with E the number of edges in graph G, and B the number of non-empty blocks in partition b the prior distribution on the edges is given by:

$$P(e) = \frac{n_e^E}{(n_e+1)^{E+B(B+1)/2}}.$$

Please refer to Peixoto, 2017[34] for a detailed derivation and explanation. Note that for our networks, as we only use simple graphs and do not consider self-edges, i.e. $A_{ij} \in \{0, 1\}$ and $A_{ii} = 0$, the product $\prod_{i<j} A_{ij}! \prod_i A_{ii}!!$ simplifies to 1. The prior for the partition b is derived from sampling the number of non-empty blocks B as $P(B) = 1/N$ with N the number of nodes in graph G, then sample the distribution of block sizes n_r conditioned on B and finally the partition b conditioned on the former two which yields in total[34]:

$$P(b) = \frac{\prod_r n_r!}{N!} \binom{N-1}{B-1}^{-1} \frac{1}{N}.$$

A drawback of the classical SBM is that all nodes within a block are forced to have similar degrees which is not appropriate for most encountered real networks. Karrer and Newman proposed the degree-corrected SBM[35] that is, in addition to the partition b and the edge count matrix e, defined by a degree sequence $k = \{k_i\}$ setting a degree for each node. Using the microcanonical formulation[34], we have

$$P(G \mid b) = P(G \mid k,e,b) \cdot P(k \mid e,b) \cdot P(e).$$

with $P(e)$ as above. $P(G|k, e, b)$ is the probability of generating a graph G where the edge counts as well as the degree sequence is fixed to a specific value given a certain partition b and is given by

$$P(G \mid k,e,b) = \frac{\prod_{r<s} e_{rs}! \prod_r e_r r!! \prod_i k_i!}{\prod_r e_r!!}.$$

The prior we employ for the degree sequence is conditioned on the degree frequencies which are in turn sampled from a uniform hyperprior[34]:

$$P(k \mid e,b) = \prod_r \frac{\prod_k \eta_k^r!}{n_r!} \prod_r q(e_r, n_r)^{-1}$$

with η_k^r the number of nodes of degree k in block r and $q(n, m)$ the number of partitions of n elements into at most m groups.

The Bayesian approach as detailed by Peixoto[33,34] prevents overfitting but is prone to underfitting, meaning that statistically relevant structures may not be detected. In particular, the number of groups which can be resolved is limited such that small groups in very large networks can hardly be detected[34]. A solution has been proposed via the hierarchical (or nested) SBM[25]. Therein, the edge count matrix $e = \{e_{rs}\}$ is described with another SBM, i.e. the blocks of the first SBM are considered as nodes a second-level SBM which are partitioned into a second level of blocks, with the according second-level edge count matrix. The second-level SBM can again be described as another third-level SBM and so on, L number of times, forming a nested hierarchy of SBMs.

The joint distribution for the hierarchical microcanonical degree-corrected SBM is given by[34]

$$P(A,k,\{e_l\},\{b_l\}) = P(A \mid k,e_1,b_1) \cdot P(k \mid e_1,b_1) \cdot P(\{e_l\}) \cdot P(\{b_l\})$$

for $l \in \{0,\ldots,L\}$, with e_l the edge count matrix at level l, b_l the partition at level l. The according prior distributions are given as described above and in addition for the edge count by[34]

$$P(\{e_l\} \mid \{b_l\}) = \prod_{l=1}^{L} P(e_l \mid e_{l+1}, b_l)$$

and imposing the boundary conditions $B_L = 1$ for the number of nonempty blocks at the highest level L, and $P(b_L) = 1$. Thereby,

$$P(e_l \mid e_{l+1}, b_l) = \prod_{r<s} \binom{n_r^l n_s^l}{e_{rs}^{l+1}}^{-1} \prod_r \binom{n_r^l (n_r^l+1)/2}{e_{rr}^{l+1}/2}^{-1}$$

with $\binom{n}{m}$ the number of combinations of m elements with repetitions from a set of n elements. For the prior distribution of the hierarchical partitions, we employ

$$P(\{b_l\}) = \prod_{l=1}^{L} P(b_l)$$

using the equation for $P(b)$ from above with the block and node counts for the respective level l and the boundary condition $B_0 = N$. The hierarchical model without degree correction is obtained by replacing $P(A|k, e_1, b_1) \cdot P(k|e_1, b_1)$ by $P(A|e_1, b_1)$.

Here, we examine four SBM versions: the classical SBM, the degree-corrected SBM, the hierarchical SBM and the degree-corrected & hierarchical SBM. For fitting the SBM to one of these four SBMs (i.e. in order to determine the most probable (hierarchical) partition given the data), 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[36] framework. Initializing a partition according to the prior distribution of the partitions from above and using an agglomerative multi-level Markov Chain Monte Carlo algorithm[37], that module allows for determining a (potentially hierarchical) partition, b, of the network, G, which minimizes the description length, DL, of the SBM[33]:

$$\begin{aligned} DL &= -\log_2 P(G\,|\,\lambda, b) - \log_2 P(\lambda, b) \\ &= -\log_2 (P(G\,|\,\lambda, b) \cdot P(\lambda, b)) \\ &= -\log_2 (P(G|b) \cdot P(b)) = -\log_2 (P(b\,|\,G) \cdot P(G)). \end{aligned}$$

Thereby, λ 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 λ by the structure of the network G the SBM is fitted to (see 33, 34), which enables considering only one value for the parameter λ 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 monotonously 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[33]. 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 Λ[33]:

$$\begin{aligned} \Lambda &= \frac{P(b_1, SBM_1|G)}{P(b_2, SBM_2|G)} = \frac{P(G|b_1, SBM_1)P(b_1)P(SBM_1)}{P(G|b_2, SBM_2)P(b_2)P(SBM_2)} \\ &= \frac{P(G|b_1, SBM_1)P(b_1)}{P(G|b_2, SBM_2)P(b_2)} = 2^{-(DL_1 - DL_2)}. \end{aligned}$$

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[38]. 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 overrepresented for default settings (p-value < 0.05, q-value < 0.2), and only human pathways occurring in the file from the Reactome database, https://reactome.org/download/current/ReactomePathwaysRelation.txt (downloaded June 6, 2018, 39), were used. This file was also employed as representation of the Reactome hierarchy to determine distances between Reactome pathways.

For the metabolite dataset, we mapped the pubchem IDs and metabolite names to KEGG compound IDs using the MBRole webserver version 2[40] 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[41]. We only considered pathways with a minimal size of 2 and otherwise the same settings as for mRNA and protein overrepresentation analysis.

Comparison to breast cancer signatures and summarizing biological function

For comparison to known biological functions relevant to breast cancer, we downloaded those gene sets from the MSigDB database v6.2[42], gene set collection C6, oncogenic signatures, that arose for the search terms "breast AND (cancer OR carcinoma)" on the MSigDB web interface (17 gene sets in total, search and download on July 31, 2019). We computed the Reactome pathways that are overrepresented in any of these gene sets (as described in section "Overrepresentation analysis"). The resulting 43 Reactome pathways were assorted according to their Reactome parent pathways (top level Reactome pathways) and, for each single pathway, its occurrence among pathways overrepresented in SBM-predicted blocks was counted.

For summarizing biological function, we mapped each overrepresented Reactome pathway to its parent Reactome pathway (top-level pathway). For each network and each hierarchy level, we counted the number of occurrence of each Reactome parent pathway among the Reactome pathways overrepresented in any of the SBM-derived blocks. We normalized these counts to the total number of overrepresented pathways in the blocks of the network and hierarchy level in order to obtain the percentage of

each Reactome parent pathway. Note that we counted the same pathway multiple times if it was overrepresented in multiple SBM-derived blocks. In order to account for the different sizes of the Reactome parent pathways, we further divided the percentage of a Reactome parent pathway for a network and hierarchy level as described above by the percentage of all Reactome pathways that are associated to the Reactome parent pathway. This latter percentage is for example high for "metabolism" and "signal transduction" and low for "chromatin organization" and "circadian clock".

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 overrepresented 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 package[3] 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 threshold 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 "deepSplit" parameter of 4 (mRNA) or 2 (else), "pamRespectsDendro" 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 δG is a set of edges which do or do not occur in the network G, 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[27]:

$$P(\delta G|G) \propto \sum_b \frac{P(G+\delta G|b)}{P(G|b)} P(b|G)$$

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 δG 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 δG 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 δG, 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:

$$P(\delta G|G) \propto \sum_b \frac{P(G+\delta G|b)}{P(G|b)} P(b|G) \approx \frac{P(G+\delta G|b^*)}{P(G|b^*)} P(b^*|G).$$

It resorts to neglecting the summands for all partitions except for the one, b^*, which contributes most to the posterior distribution, $b^* = max_b P(b|G)$. 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 δG 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 δG directly depend on the difference between the description length of the original network G with partition b^*, DL_{G,b^*}, and the description length of the altered network $G + \delta G$ with partition b^*, $DL_{G+\delta G,b^*}$[27]:

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

The difference between the description lengths, $DL_{G,b^*} - DL_{G+\delta G,b^*}$, was computed via the function *get_edges_prob* from graph-tool[36]. Note that as of time of writing, *get_edges_prob* does not work for weighted SBMs with real-normal edge covariate (see filed 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 $\log_2(e)$ 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 SBM-based analysis 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[43–45]. We used three types of molecules reflecting different characteristics of tumorous tissue or cells: mRNAs, proteins or metabolites. These are key cellular molecules that are widely applied, in isolation or in combinations, in different biomedical research domains. Their abundance and interconnectedness in networks are therefore of high interest if aiming to characterize cells or tumorous tissue.

Other potential data types could be e.g. mutations, copy number variation, DNA methylation or miRNAs, which are interesting avenues to further explore. While they could be useful, there are some caveats associated with them, e.g.: the derived interactions within layers are even less directly interpretable than for mRNA, protein or metabolite; networks generated from mutations and copy number variation are extremely sparse; the functional interpretation of DNA methylation data relies on mRNA expression and the resulting networks are extremely big; the roles of miRNAs are less well known. Therefore, we decided to restrict our analyses to the three biomolecular entities mRNA, proteins and metabolites.

The Cancer Genome Atlas (TCGA) initiative has provided data on the mRNA level (from RNAseq) for 237 ER- breast cancer patient tissue samples; mass-spectrometry-based proteomic data (4plex iTRAQ) are available for a subgroup of 36 patients[29]. Metabolomics data have been measured by GC-TOF-MS in a different breast cancer cohort study for 67 samples[30]. 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 (only for 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[2,3]. 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 and because we wanted to focus on the connections between species, i.e. the covariation of expression. Thereby, also lowly abundant species can exhibit strong connections, and indeed they are found to play a role, as indicated by a non-zero degree, in our reduced networks (see Figure S1D[46]).

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. Therefore, we computed the p-value associated to each correlation value in the networks. Then, for each of the three networks, we applied both Bonferroni and Benjamini-Hochberg multiple testing correction methods along with the classical significance thresholds 0.01 and 0.05 (see Figure 1C). We finally chose the correction method and significance threshold for each data layer considering a trade-off between minimal network size (i.e. minimal computation time for the subsequent fit to SBM) and maximal connectedness of the reduced network: We used the combination of multiple testing correction and significance threshold that provided a high degree of reduction while maintaining a high percentage of nodes within the largest connected component of the network. While the stringent Bonferroni correction was necessary to achieve a sufficient degree of reduction for the mRNA network, it severely disrupted the connectedness for the protein and metabolite data layer leading to less than 30% or 65% of the nodes being in the largest connected component for protein or metabolite, respectively (see Figure 1C, employed thresholds marked in blue).

In the second reduction method, we systematically removed links weaker than increasingly stringent correlation thresholds until the reduced network met a criterion of scale-freeness (using a function from the WGCNA R package[3]), see Figure 1D. 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 with a negative

exponent. The least stringent correlation threshold for which a sufficiently good fit (scale-free fit index $R^2 > 0.85$, 3) between the degree distribution of the reduced network and a power-law degree distribution with negative exponent was obtained was used for network reduction (networks named "scale-free"). The degree distributions of all six reduced networks are shown in Figure 1E. The relationship between the two reduction by thresholding approaches are further illustrated in Figure S1A-C[46]. 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. 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[35] 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[25] (Figure 2A right). This model is especially suitable to represent large networks with many nodes as it counteracts underfitting. 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[36] 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[18,27,47]. However, the characteristics of the optimal weighted SBM (number of blocks) were severely impacted

Table 1. Characteristics of the networks derived from the three data layers. Note that genes/proteins/metabolites were removed if having >20% NA values (which was the case only for proteins), all other entities were kept as nodes in the reduced network even if their degree was zero (i.e. having no edge) after network reduction.

Characteristic	mRNA	protein	metabolite
samples	237	36	67
entities (nodes, before NA removal)	18321	10625	162
edges (before reduction)	167820360	56440000	13041
reduced network: scale-free			
minimal correlation	0.603	0.788	0.375
entities (nodes)	18321	8430	162
edges	287790	85980	1825
entities of degree zero	8183	4967	4
entities in largest connected component	9111	3187	158
reduced network: sign. corr.			
minimal correlation	0.395	0.534	0.310
entities (nodes)	18321	8430	162
edges	4260267	2434159	2811
entities of degree zero	1061	3	1
entities in largest connected component	17190	8427	161

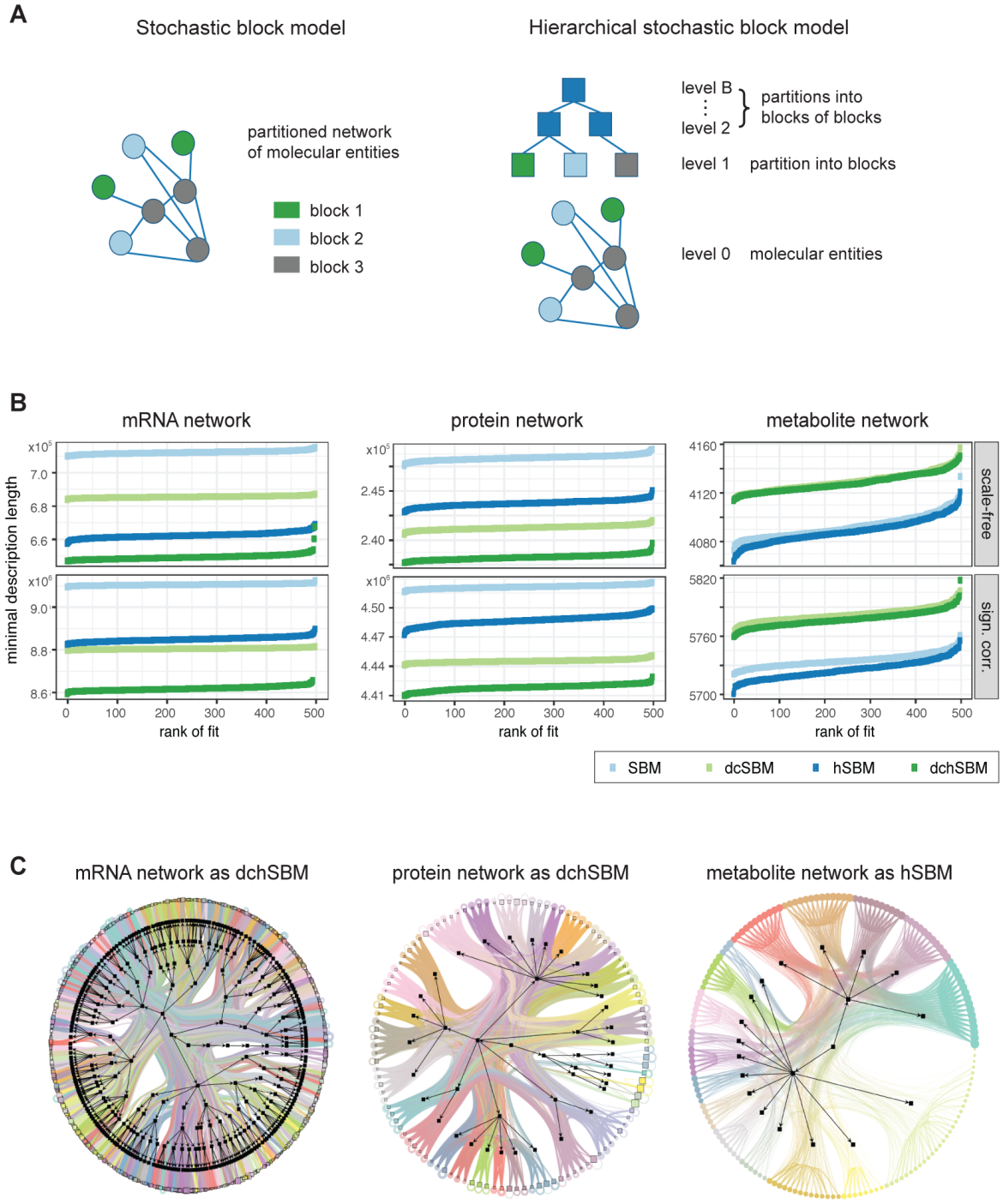

Figure 2. Stochastic block models for representing correlation-based biomolecular networks. (**A**) In a stochastic block model (SBM) representation, the nodes of a network are partitioned into blocks according to their similarity in connectivity. The hierarchical version of the SBM (right) imposes in 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.

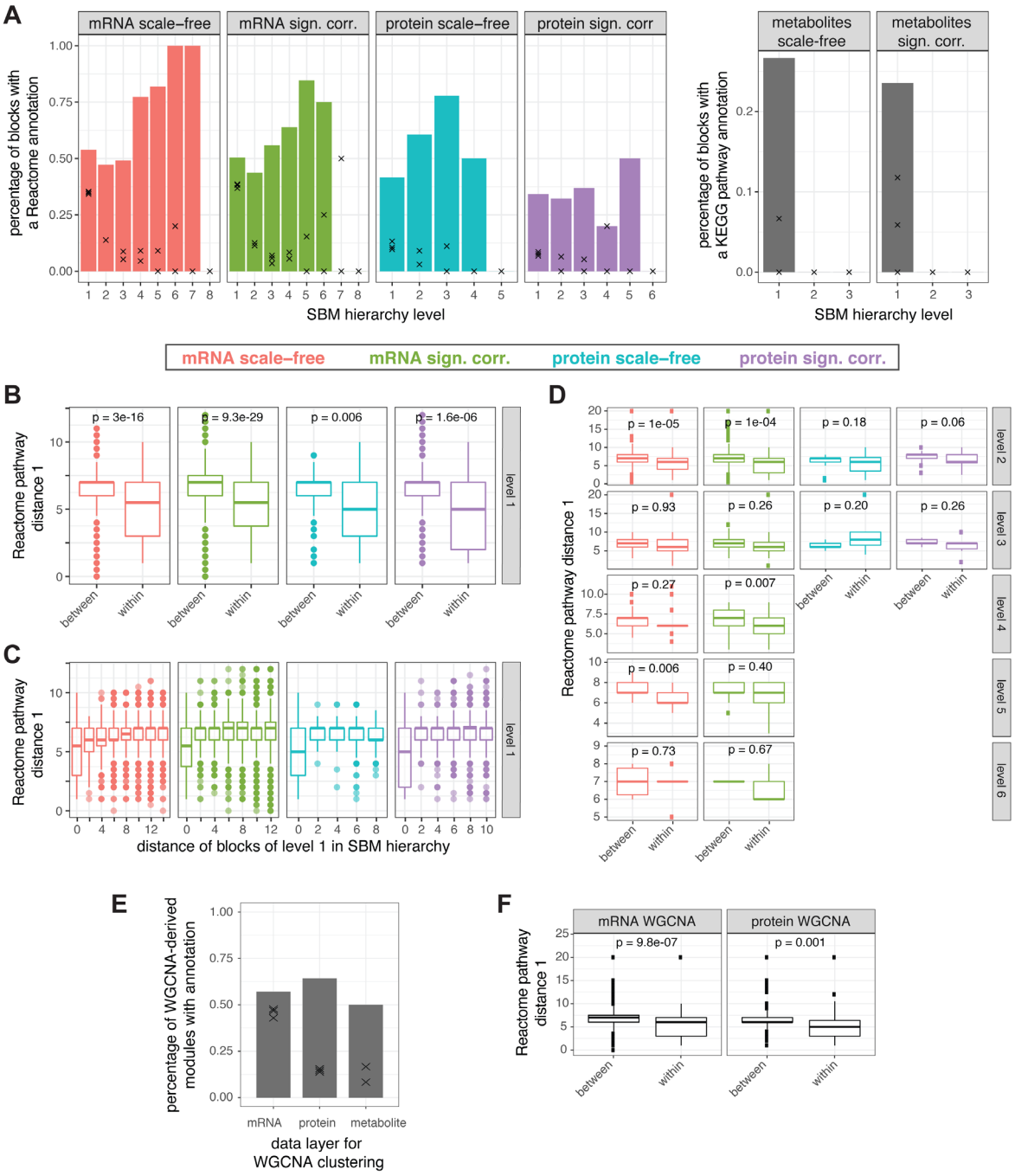

Figure 3. Modules derived from hierarchical SBM represent biological function. (**A**) Percentage of blocks with at least one overrepresented Reactome (KEGG) pathway for the best fitting SBM (see Figure 2B) for the mRNA and protein (metabolite) networks, reduced by condition on scale-freeness (scale-free) 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 (level 1) 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 a distance measures (i) based on the Reactome hierarchy (see Figure S3[46] for the results using the alternative distance measure (ii)). The lower the distances the more similar are the pathways. 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**) Distance 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 hierarchy for blocks on level 1. We do not find evidence that the Reactome hierarchy is reflected in the SBM hierarchy. (**D**) Between-module vs. within-module distances of Reactome terms as in (**B**) for blocks of SBM hierarchy levels 2-6. (**E**) Percentage of modules detected by the WGCNA approach with at least one overrepresented Reactome or KEGG pathway. Black crosses denote the results for three similar clusterings with randomly shuffled mRNAs, proteins, or metabolites. (**F**) Between-module vs. within module distances of Reactome terms as in (**B**) for the clusterings predicted from WGCNA-based module detection.

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 S2[46]). 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[37]. 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 and predict 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. Reactome provides a hierarchical annotation which enables a definition of a distance between terms and a comparison to the hierarchical structure given by the SBM, KEGG is one of the annotation databases used most for metabolites. 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 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 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 distance measure 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 (level 1, see Figure 3B for distance measure (i), Figure S3[46] for distance measure (ii)). 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.

In addition, we compared Reactome pathway distances between blocks and within blocks for the blocks of the higher levels (levels 2-6, Figure 3D). Please note that this analysis could only be performed for the subset of SBM hierarchy levels with at least three blocks with more than one overrepresented Reactome pathway (otherwise, a Welch's test cannot be performed due to low sample count). For higher levels, within-block distances are only in some cases significantly lower than between-block distances, no general effect can be observed.

Furthermore, we examined the relationship between the hierarchies within the SBMs and the hierarchy of the Reactome pathways. First, 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). Second, we compared the distance within the SBM hierarchy of each pair of blocks (on level 1) to the distance of their associated Reactome pathways within the Reactome hierarchy (Figure 3C, Figure S3B). Thereby, 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 in the SBM hierarchy. From both analyses (Figure 3C–D), 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[3]. 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; these numbers 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 S3D[46]). 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 overrepresented 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. Comparison of within-module distances and between-module distances of the annotated Reactome terms for the mRNA and protein networks delivers significantly lower within-module distances as for the SBM clusterings (Figure 3F).

The observed differences derived from WGCNA vs. SBM 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.

We furthermore compared the results from the overrepresentation analysis of the SBM-derived blocks to known biological insights from breast cancer. To that purpose, for the mRNA and protein networks, we compared the SBM-derived biological Reactome terms to those found for oncogenic signatures obtained from the database MSigDB[42]. The 43 Reactome pathways known to be related to breast cancer according to MSigDB (see *Methods*) stem mainly from the categories extracellular matrix (ECM) organization, signal transduction, cell cycle, hemostasis, DNA repair, and few others. All but one of these pathways were found as overrepresented, partially with high frequency, in SBM-derived blocks of one of the mRNA or protein networks (Figure 4). The exception is “Defensins” that are relevant for antimicrobial immune response and therefore the immune system. Its occurrence in oncogenic signatures might originate from immune cells measured together with tumour tissue. Overall, the two protein networks exhibit less overrepresented pathways, but especially the categories ECM organization and DNA repair are well represented in both networks, and cell cycle in the network reduced by significance of correlation (lower 2 panels in Figure 4). Thus, biological functions related to breast cancer are well captured in the SBM-derived clustering.

For the metabolite networks, we investigated the KEGG pathways found as overrepresented in the SBM-derived blocks (Table 2). Therein, especially “Biosynthesis of unsaturated fatty acids” is very prominent, it is indeed overrepresented in two blocks for each network (not shown) and occurs in different variations in the scale-free network (e.g. further overrepresented terms “Fatty acid biosynthesis”, “Linoleic acid metabolism”, Table 2). Indeed, fatty acids synthesis has been related to metastasis, therapeutic resistance and relapse in cancer[48]. The SBM-derived predicted importance of valine, leucine and isoleucine metabolism for breast cancer (Table 2) has been suggested before, in particular with respect to leucine[49]. In addition, ABC transporters occur as overrepresented in both SBM-derived clusterings of the metabolite networks: They have been suggested to play a role in chemoresistance[50] and therefore point to one possible underlying reason of the bad prognosis of ER- breast cancer patients.

In order to identify further biological functions that could play a role according to SBM-derived network structures in mRNA and protein networks, we summarized the overrepresented Reactome pathways on the level of the parent Reactome pathways (on the top level of the Reactome hierarchy, Figure 5). Thereby, we could retrieve the categories known to be of relevance to breast cancer, in particular ECM organization and cell cycle, for all four networks. We found additional categories that were observed more frequently than they are represented in the Reactome hierarchy: “Metabolism of RNA”, “Metabolism of proteins”, and “Chromatin organization” (columns with darker color in Figure 5A). Please note that tRNA synthesis has occurred as overrepresented KEGG pathway of the SBM-derived structures of the metabolite networks (Table 2) - since tRNA is subject to RNA metabolism and is required for the translation of proteins, it lies at the interface those two relevant metabolism categories predicted from mRNA and protein networks and thus complements their predictions. Metabolism of proteins relies on amino acids and could be also related to the leucine addiction reported for breast cancer[49], and it is supported by the SBM-derived predictions in the metabolite network on the relevance of the metabolism of further amino acids (Table 2). Of note, metabolism in general did not occur with a high frequency (relative to the size of the category “Metabolism”, see Figure 5B) revealing a certain specificity of the predictions obtained by fitting to SBM. The highlight on chromatin organization is an interesting

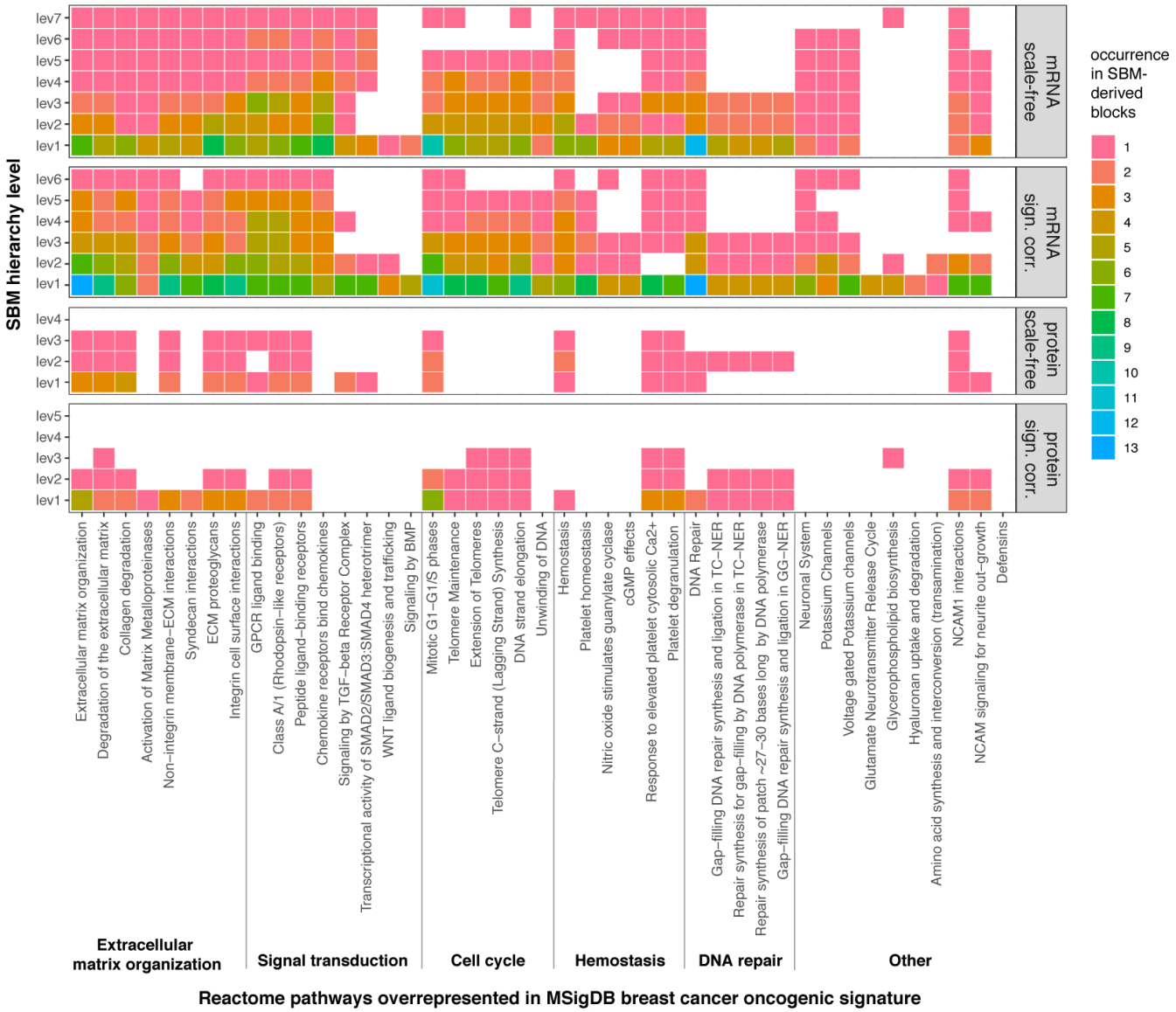

Figure 4. SBM-derived blocks of mRNA and protein networks exhibit biological functions related to breast cancer signature. Oncogenic signature gene sets related to breast cancer were retrieved from MSigDB[42] and mapped to Reactome pathways. The color code shows for how many SBM-derived blocks each of these pathways was overrepresented, for the four mRNA and protein networks and each hierarchy level. Hierarchy levels without any blocks with overrepresented pathways are omitted.

SBM-derived prediction complementary to the metabolism motif and invites further exploration, e.g. by examining related blocks and their connectivity structure in relation to other blocks in the SBM.

To summarize, biological functions related to breast cancer are well captured in the SBM-derived clustering. Networks derived from different datalayers enable different perspectives of the phenotype that can support each other or provide complementary predictions. After the assessment of the biological content in the SBM-derived clustering, we moved to the second deliverable of the fit of the networks to SBMs: finding alternative edge relevance based on global network properties.

Assessing edge relevance by SBM-based edge confidence scores

The descriptions 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

Table 2. Overrepresented KEGG pathways for the SBM-derived blocks of the metabolite networks (reduced by significance of correlation, sign. corr., or by a criterion on scale-freeness, scale-free).

metabolite network: sign. corr.	metabolite network: scale-free
Biosynthesis of unsaturated fatty acids	Biosynthesis of unsaturated fatty acids
Aminoacyl-tRNA biosynthesis	Aminoacyl-tRNA biosynthesis
ABC transporters	ABC transporters
Valine, leucine and isoleucine biosynthesis	Valine, leucine and isoleucine biosynthesis
Valine, leucine and isoleucine degradation	Valine, leucine and isoleucine degradation
Amino sugar and nucleotide sugar metabolism	Starch and sucrose metabolism
	Fatty acid biosynthesis
	Linoleic acid metabolism
	Ubiquinone and other terpenoid-quinone biosynthesis
	Biosynthesis of secondary metabolites
	Glycine, serine and threonine metabolism

could be used to predict whether an edge is spurious (or missing)[17,18,27]. 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 the six reduced networks, all edges that existed in the network were considered as "putatively spurious". Similarly, all edges that were not in the network because they had been removed from the (fully connected) correlation-based network during the reduction procedure were considered as "putatively missing". For each putatively missing and spurious edge, we used the Python module graph-tool[36] to compute its edge confidence score (Figure 6). 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 6A, 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 is still computationally not feasible to compute the scores for all putatively missing edges. We therefore resorted to computing it for as many putatively missing edges as we have existing edges in the network (i.e. approx. 2.4×10^6 for the protein network, 4.6×10^6 for the mRNA network, see Table 1), and chose those with highest absolute edge weights (Figure 6B 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 network. 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 computational 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 indicates an improvement in fit quality if adding or removing the edge, 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 edge 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 6A, 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 6A, 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

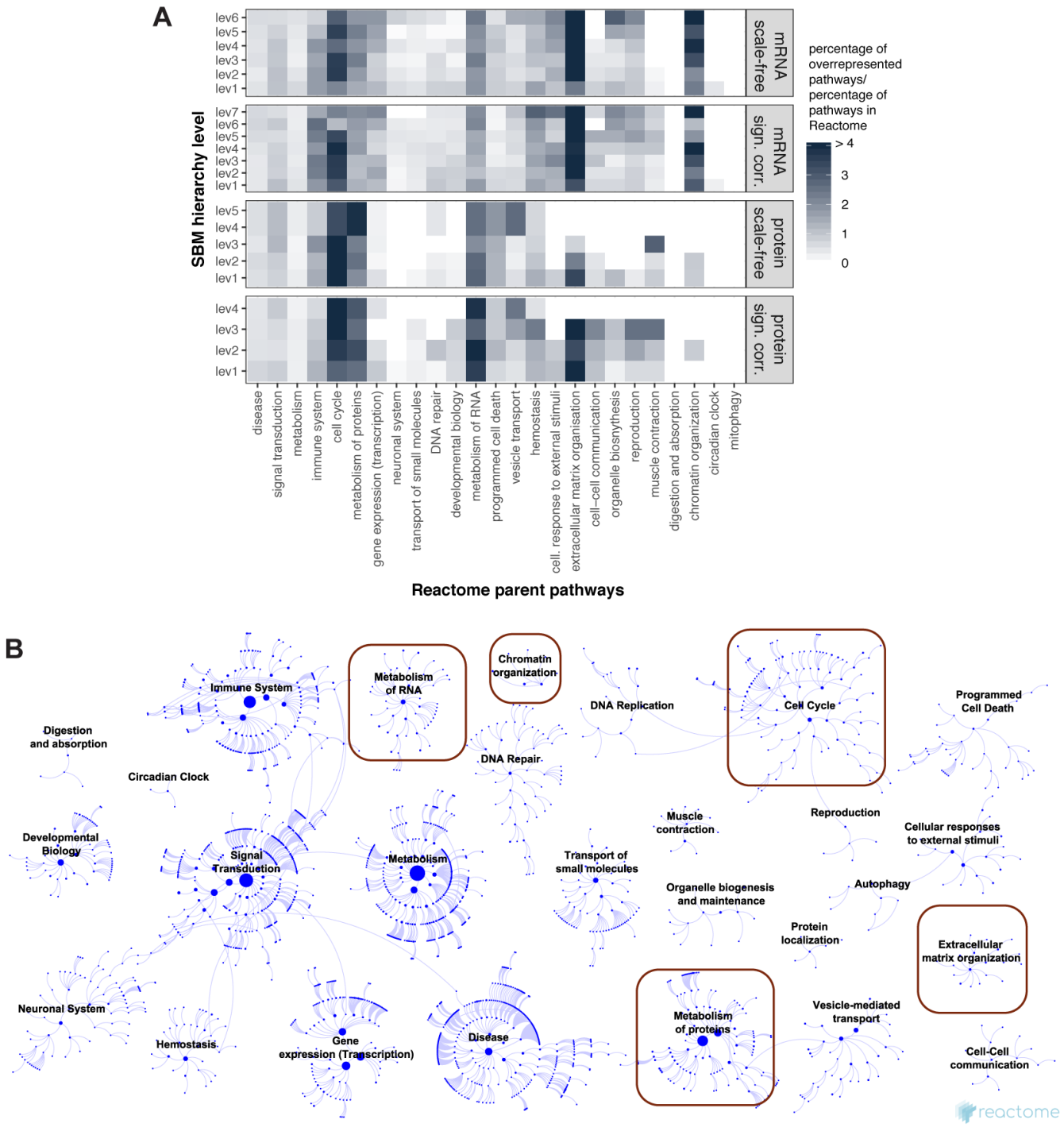

Figure 5. SBM-derived blocks of mRNA and protein networks predict further pathways relevant for breast cancer. (**A**) We mapped the Reactome pathways to their Reactome parent pathways and counted their occurrence as overrepresented pathway in SBM-derived blocks of the four mRNA and protein networks and each hierarchy level. Reactome parent pathways are sorted by the number of pathways they summarize (see (**B**)). The color code gives the percentage of occurrence as overrepresented of each parent pathway (counted occurrence divided by total number of overrepresented pathways obtained for the SBM-clustering of the hierarchy level and network) relative to the percentage of each parent pathway in the Reactome annotation (number of pathways associated to the parent pathway relative to the number of total Reactome pathways). These values indicate whether certain parent pathways occur more frequently than suggested by the size of the parent pathways cluster in Reactome. The higher the value (i.e. the darker the color), the more relevant the parent pathway is predicted by the SBM-derived network structure. (**B**) Reactome pathway organization (downloaded from the Reactome website Pathway Viewer, 39). The parent pathways that are predicted as especially relevant for breast cancer according to (**A**) are highlighted.

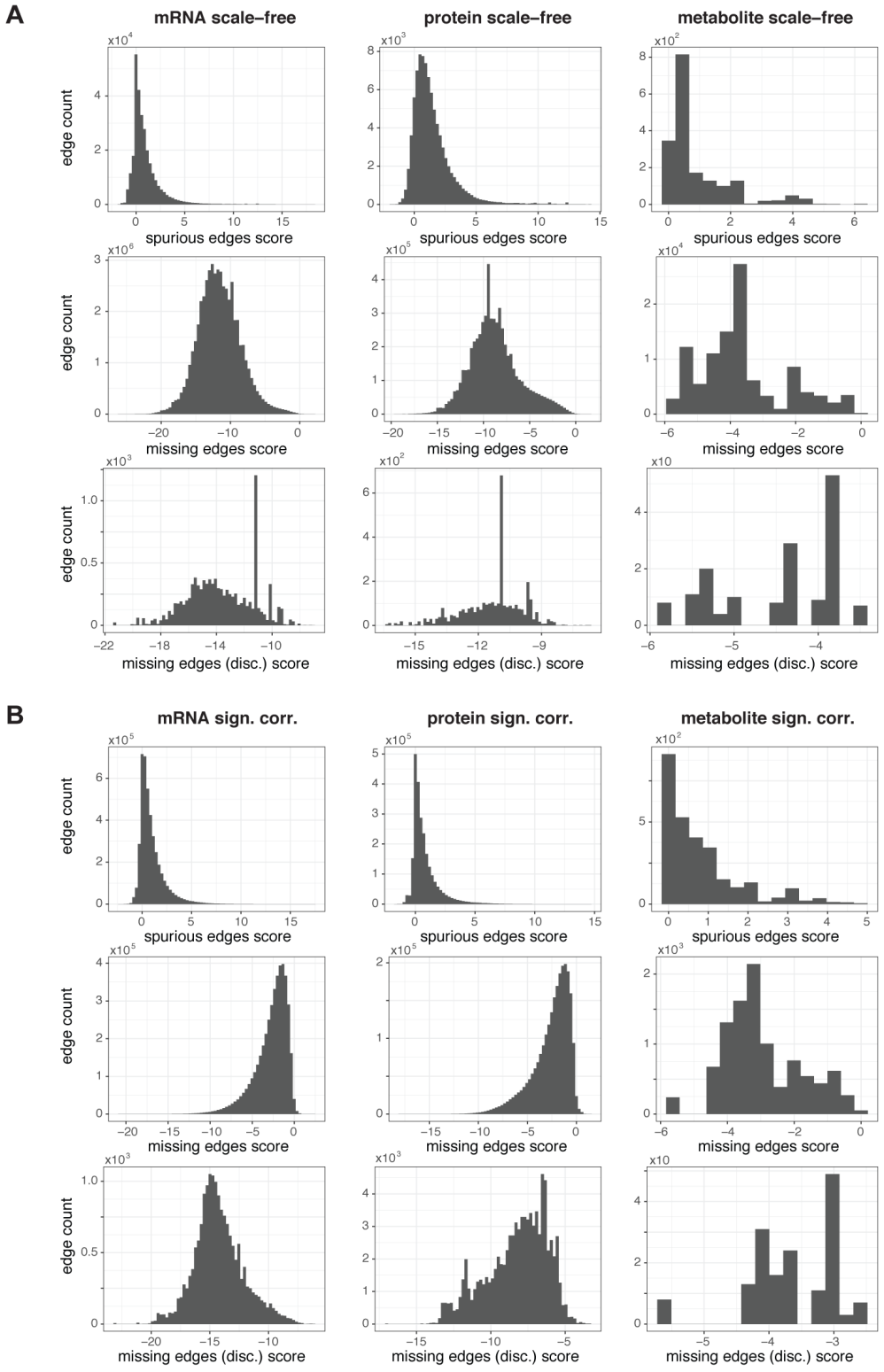

Figure 6. SBM-derived edge prediction: Missing and spurious edge confidence scores. (**A**) Histograms of edge confidence scores for the best fitting SBM of the three networks reduced by criterion on scale-freeness for all existing, i.e., putatively spurious edges (spurious edge score, top), for the 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.

at no point provided to the SBM, neither during the fitting to the SBM (except for the weighted version and only for Figures S2, S4[46]) 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 positive 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 7). Similarly, we find an overall negative correlation between (absolute) edge correlations and spurious edge confidence scores: Edges with high correlation are preferentially *not* predicted as spurious, i.e., they are predicted as relevant to the network, also in terms of edge confidence score. Consequently, the comparison to the edge correlation suggests that SBM-derived edge confidence scores could be used as additional information for assessing the relevance of edges for multiple omics correlation-based networks.

Discussion

Using example cases of correlation-based transcriptomic, proteomic and metabolomics networks from breast cancer tumour samples, here we show that and 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 functionally relevant blocks. Biological functions related to breast cancer are well captured in the SBM-derived clusters and the networks derived from the different data layers shed light on different perspectives of the phenotype that can support each other and result in complementary predictions. 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 approaches[3,9,14]. 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. Indeed, we found our SBM-derived blocks to be biologically relevant to our examined phenotype, and comparing SBM-derived structures between different phenotypes is an intriguing next step.

Second, the derived edge predictions can reproduce the interaction 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 confidence, i.e., as dispensable. Keep in mind that the correlations 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 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 6), 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[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 S2[46]). For reduced networks, the results for weighted SBMs seem more promising (see Figure S4[46]). 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

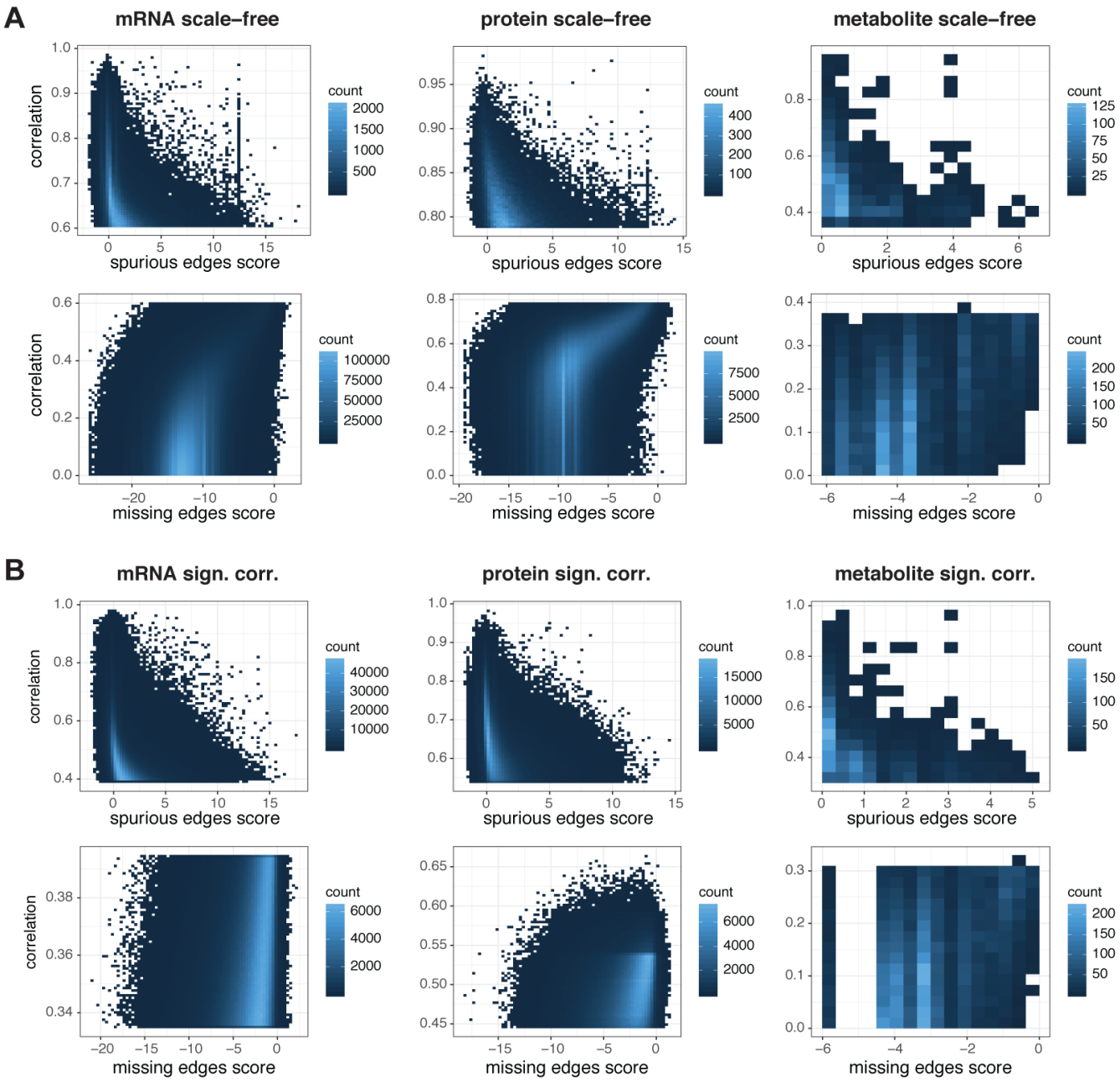

Figure 7. SBM-derived edge prediction: Validation by correlation. Relationship between edge confidence scores and edge correlations (Spearman's correlation of the measurements at their nodes) for putatively spurious (top) or putatively missing edges between nodes of degree > 0 (bottom) for the networks 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).

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 were used. Proteomics data stem from 29, metabolomics data stem from 30. The metabolomics raw data can only be obtained upon accessing the cited article[30]; 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/biomodlih/sbm-for-correlation-based-networks

Archived code as at time of publication: https://doi.org/10.5281/zenodo.3363060[46]

License: GNU GPLv3 license.

Extended data

Zenodo: Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models. https://doi.org/10.5281/zenodo.3363060[46]. This project contains the following extended data files:

- Figure S1: The relationship between network reduction by significance of correlation or by scale-freeness
- Figure S2: The weighted SBM seems not appropriate for edge prediction from edge confidence scores for fully connected networks
- Figure S3: Pathway characteristics for alternative distance measure, and block size distributions
- Figure S4: Edge predictions for a reduced weighted network with planar or hierarchical weighted SBMs

Grant information

This research was funded by Luxembourg National Research Fund, FNR, SINGALUN project. This research was partially supported by Tier-2 MOE2016-T2-1-029 grant by the Ministry of Education, Singapore. 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

1. Barabási AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet.* 2011; **12**(1): 56–68.
PubMed Abstract | Publisher Full Text | Free Full Text

2. van Dam S, Võsa U, van der Graaf A, *et al.*: **Gene co-expression analysis for functional classification and gene-disease predictions.** *Brief Bioinform.* 2018; **19**(4): 575–592.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics.* 2008; **9**: 559.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Toubiana D, Xue W, Zhang N, *et al.*: **Correlation-Based Network Analysis of Metabolite and Enzyme Profiles Reveals a Role of Citrate Biosynthesis in Modulating N and C Metabolism in *Zea mays*.** *Front Plant Sci.* 2016; **7**: 1022.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Pellegrini M: **Community Detection in Biological Networks.** *Encyclopedia of Bioinformatics and Computational Biology.* Elsevier, 2019; **1**: 978–987.
Publisher Full Text

6. Langfelder P, Horvath S: **Fast R Functions for Robust Correlations and Hierarchical Clustering.** *J Stat Softw.* 2012; **46**(11): pii: i11.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Jeub LGS, Sporns O, Fortunato S: **Multiresolution Consensus Clustering in Networks.** *Sci Rep.* 2018; **8**(1): 3259.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Lancichinetti A, Fortunato S: **Consensus clustering in complex networks.** *Sci Rep.* 2012; **2**: 336.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Su Y, Wang B, Zhang X: **A seed-expanding method based on random walks for community detection in networks with ambiguous community structures.** *Sci Rep.* 2017; **7**: 41830.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Rosvall M, Bergstrom CT: **An information-theoretic framework for resolving community structure in complex networks.** *Proc Natl Acad Sci U S A.* 2007; **104**(18): 7327–31.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Reichardt J, Bornholdt S: **Statistical mechanics of community detection.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2006; **74**(1 Pt 2): 016110.
PubMed Abstract | Publisher Full Text

12. Newman ME, Girvan M: **Finding and evaluating community structure in networks.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2004; **69**(2 Pt 2): 026113.
PubMed Abstract | Publisher Full Text

13. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res.* 2002; **30**(7): 1575–84.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Blondel VD, Guillaume J, Lambiotte R, *et al.*: **Fast unfolding of communities in large networks.** *J Stat Mech Theory Exp.* 2008; **2008**(10): P10008.
Publisher Full Text

15. Zhu Y, Zhang XF, Dai DQ, *et al.*: **Identifying spurious interactions and predicting missing interactions in the protein-protein interaction networks via a generative network model.** *IEEE/ACM Trans Comput Biol Bioinform.* 2013; **10**(1): 219–25.
PubMed Abstract | Publisher Full Text

16. Wang H, Zhang F, Min H, *et al.*: **SHINE: Signed heterogeneous information network embedding for sentiment link prediction.** *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.* 2018; 592–600.
Publisher Full Text

17. Guimera R, Sales-Pardo M: **Missing and spurious interactions and the reconstruction of complex networks.** *Proc Natl Acad Sci U S A.* 2009; **106**(52): 22073–8.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Aicher C, Jacobs AZ, Clauset A: **Learning latent block structure in weighted networks.** *J Complex Netw.* 2015; **3**(2): 221–248.
Publisher Full Text

19. Williamson SA: **Nonparametric network models for link prediction.** *J Mach Learn Res.* 2016; **17**.
Reference Source

20. Zhu B, Xia Y, Zhang XJ: **Weight prediction in complex networks based on neighbor set.** *Sci Rep.* 2016; **6**: 38080.
PubMed Abstract | Publisher Full Text

21. Navlakha S, Gitter A, Bar-Joseph Z: **A network-based approach for predicting missing pathway interactions.** *PLoS Comput Biol.* 2012; **8**(8): e1002640.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Shakibian H, Moghadam Charkari N: **Mutual information model for link prediction in heterogeneous complex networks.** *Sci Rep.* 2017; **7**: 44981.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Pan L, Zhou T, Lü L, *et al.*: **Predicting missing links and identifying spurious links via likelihood analysis.** *Sci Rep.* 2016; **6**: 22955.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Holland PW, Laskey KB, Leinhardt S: **Stochastic blockmodels: First steps.** *Soc Networks.* 1983; **5**(2): 109–137.
Publisher Full Text

25. Peixoto TP: **Hierarchical block structures and high-resolution model selection in large networks.** *Phys Rev X.* 2014; **4**(1): 011047.
Publisher Full Text

26. Zhang X, Wang XJ, Zhao CL, *et al.*: **Degree-corrected stochastic block models and reliability in networks.** *Physica A-Statistical Mechanics and Its Applications.* 2014; **393**: 553–559.
Publisher Full Text

27. Vallès-Català T, Peixoto TP, Sales-Pardo M, *et al.*: **Consistencies and inconsistencies between model selection and link prediction in networks.** *Phys Rev E.* 2018; **97**(6–1): 062316.
PubMed Abstract | Publisher Full Text

28. Kosinski M, Biecek P: **RTCGA: The cancer genome atlas data integration.** 2016.
Reference Source

29. Mertins P, Mani DR, Ruggles KV, *et al.*: **Proteogenomics connects somatic mutations to signalling in breast cancer.** *Nature.* 2016; **534**(7605): 55–62.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Budczies J, Brockmöller SF, Müller BM, *et al.*: **Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism.** *J Proteomics.* 2013; **94**: 279–88.
PubMed Abstract | Publisher Full Text

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

32. Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *J R Statist Soc B.* 1995; **57**(1): 289–300.
Publisher Full Text

33. Peixoto TP: **Bayesian stochastic blockmodeling.** *eprint arXiv:1705.10225.* 2017; arXiv:1705.10225.
Reference Source

34. Peixoto TP: **Nonparametric Bayesian inference of the microcanonical stochastic block model.** *Phys Rev E.* 2017; **95**(1–1): 012317.
PubMed Abstract | Publisher Full Text

35. Karrer B, Newman MEJ: **Stochastic blockmodels and community structure in networks.** *Phys Rev E.* 2011; **83**(1): 016107.
Publisher Full Text

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

37. Peixoto TP: **Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models.** *Phys Rev E Stat Nonlin Soft Matter Phys.* 2014; **89**(1): 012804.
PubMed Abstract | Publisher Full Text

38. Yu G, He QY: **ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.** *Mol Biosyst.* 2016; **12**(2): 477–9.
PubMed Abstract | Publisher Full Text

39. Fabregat A, Jupe S, Matthews L, *et al.*: **The Reactome Pathway Knowledgebase.** *Nucleic Acids Res.* 2018; **46**(D1): D649–D655.
PubMed Abstract | Publisher Full Text | Free Full Text

40. López-Ibáñez J, Pazos F, Chagoyen M: **MBROLE 2.0-functional enrichment of chemical compounds.** *Nucleic Acids Res.* 2016; **44**(W1): W201–W204.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Yu GC, Wang LG, Han YY, *et al.*: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *OMICS.* 2012; **16**(5): 284–287.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Liberzon A, Birger C, Thorvaldsdóttir H, *et al.*: **The Molecular Signatures Database (MSigDB) hallmark gene set collection.** *Cell Syst.* 2015; **1**(6): 417–425.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Harvey JM, Clark GM, Osborne CK, *et al.*: **Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer.** *J Clin Oncol.* 1999; **17**(5): 1474–81.
PubMed Abstract | Publisher Full Text

44. Samaan NA, Buzdar AU, Aldinger KA, *et al.*: **Estrogen receptor: a prognostic factor in breast cancer.** *Cancer.* 1981; **47**(3): 554–60.
PubMed Abstract

45. **Clinical practice guidelines for the use of tumor markers in breast and colorectal cancer. Adopted on May 17, 1996 by the American Society of Clinical Oncology.** *J Clin Oncol.* 1996; **14**(10): 2843–2877.
PubMed Abstract | Publisher Full Text

46. Baum K, Rajapakse JC, Azuaje F: **Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models (version v3) [Data set].** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3363060

47. Peixoto TP: **Nonparametric weighted stochastic block models.** *Phys Rev E.* 2018; **97**(1–1): 012306.
PubMed Abstract | Publisher Full Text

48. Kuo CY, Ann DK: **When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance.** *Cancer Commun (Lond).* 2018; **38**(1): 47.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Xiao F, Wang C, Yin H, *et al.*: **Leucine deprivation inhibits proliferation and induces apoptosis of human breast cancer cells via fatty acid synthase.** *Oncotarget.* 2016; **7**(39): 63679–63689.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Begicevic RR, Falasca M: **ABC Transporters in Cancer Stem Cells: Beyond Chemoresistance.** *Int J Mol Sci.* 2017; **18**(11): pii: E2362.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 2

Reviewer Report 12 September 2019

https://doi.org/10.5256/f1000research.22342.r52997

?

Ana Conesa
Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences (IFAS), University of Florida, Gainesville, FL, USA
Manuel Ugidos
Institute of Biomedicine of Valencia (IBV), CSIC, Valencia, Spain

Revised manuscript satisfied most of the previous comments.

1. **Data preprocessing:**
   Datasets are well characterize by the number of samples and features, the missing data imputation of the three datasets is clarified.

2. **Network generation and reduction:**
   An explanation is included about the use of 10^7 as a threshold for the number of edges and its relation with the test-multiple correction rates performed. Furthermore, the scale-free reduction is better explained.

3. **Stochastic block models:**
   On the one hand, the SBM model definition is more extensive and the underfitting problem is addressed. On the other hand, although the comment about the number of partitions is correctly answered, this explanation is not included in the body text.

4. **Functional enrichment:**
   The use of Reactome and KEGG is now justified. The description of Figure 3 has improved and a new panel was added. Moreover, two new Figures were included dealing with the biological insights derived from the SBM blocks. Finally, the edge confident scores section is explained in more detail.

New comments about revised manuscript:

Figure 3C and 3D both suggest that SBM block hierarchy does not relates with Reactome's hierarchy. As mentioned in text, SBM blocks are characterized by common connectivity

characteristics which may not be completely related with traditional classification based on functions. This is the key point of SBM derived blocks and it is not fully evaluated in terms of biological meaning, i.e, which pathways are more related to the SBM blocks? Also, a figure with the SBM networks (network topology) as in Figure 2C but including tags with enriched pathways should be included in order to evaluate the biological information that could be extracted from the SBM network. In this sense, Figures 4 and 5 are more suitable than Figure 3 as a "validation" of the capability of SBM-derived communities to capture biological function related to breast cancer. The validation step previously proposed: "A good approach could be to select a gene X and see if the rest of the genes of the pathways, where gene X is annotated, are in the same block of gene X (or in near blocks in the hierarchy). And repeat this for a high number of genes." has not been neither performed or considered in the response. If not suitable or possible, please explain why.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, functional genomics, transcriptomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 09 September 2019

https://doi.org/10.5256/f1000research.22342.r52998

Yue Qiu
City University of New York, New York, NY, USA
Lei Xie
Department of Computer Science, Hunter College, The City University of New York, New York, NY, USA

Authors have addressed the issues raised. No further comments to make

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, systems biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 June 2019

https://doi.org/10.5256/f1000research.20484.r47834

Ana Conesa
Department of Microbiology and Cell Science, Institute of Food and Agricultural Sciences (IFAS), University of Florida, Gainesville, FL, USA
Manuel Ugidos
Institute of Biomedicine of Valencia (IBV), CSIC, Valencia, Spain

This manuscript describes the application of stochastic block models (SBMs) as a method to construct biological networks from different omics data. Starting from correlation-based networks of breast cancer data, SBM delivers different modules (or blocks) of the network. In order to assess the usefulness of this approach, the authors evaluate the biological meaning of the modules obtained performing a functional enrichment analysis for each module of the network. Finally, the authors include the edge confidence score computation for each edge in the resulting network.

This novel application of SBMs with omics data is interesting but its utility is not clearly explained in the manuscript. First, the title is confusing, since I understood a single network is obtained from different omics data, i.e. a multiomic integration analysis, however it is not performed. Furthermore, why are these three datasets used? It seems the authors want to evaluate the SBM-based networks starting from datasets with different complexity (different number of features/samples), but the reason is not properly explained.
As for the SBM model definition, because of the microcanonical formulation because of its hard constraints, network underfitting could be an important issue. Although underfitting is solved using nested (hierarchical) or degree-corrected SBMs, an explanation of this phenomena and how to deal with it would be suitable in “Fit to SBM” methodological section.
Regarding the functional enrichment, the choice of the pathway database is not clear. Why is Reactome chosen for genes annotation and KEGG for metabolites? Taking into account that breast cancer related biological findings is not the scope of this manuscript, a deeper analysis of biological information revealed by network blocks is necessary: which pathways are block-specific? Which pathway or pathways are present in the bigger block? Do the pathway distances between blocks have any relation with previous knowledge about breast cancer? Furthermore, the pathway distances within and between blocks are computed just for the lower level of blocks. Since the network is hierarchical, pathway distances can be obtained for each level. This would indicate if the hierarchy is biological meaningful too. Moreover, the same genes are known to be annotated to related pathways. Therefore, this expected result does not represent a validation to determine the biological meaning of the blocks in the network. A good approach could be to select a gene X and see if the rest of the genes of the pathways, where gene X is annotated, are in the same block of gene X (or in near blocks in the hierarchy). And repeat this for a high number of genes.
Finally, the edge confidence score looks like important to improve the network characterization, however they do not use this score to optimize the final network. I was expecting the comparison

between the original SBM-based network and the corrected SBM-based network via edge confidence score optimization. In order to use the edge confidence scores, a way to compute thresholds for missing and spurious edges should be proposed.

Resuming, SBM-based biological networks are a new way to represent omics data and it represents a novel approach for SBMs. However, a novel application should be coupled with a suitable interpretation of its results. Moreover, the main conclusion or this manuscript is not clear and the main question for a new methodological development is unanswered, why should I use this method?

Specific points:

1) Data preprocessing:
- Why are the three datasets used? Is it just a matter of different complexity? Datasets are not completely characterized as the number of features is not indicated.
- Authors say missing values of metabolite data were imputed previously. It is not clear whether they performed the imputation or not. In case they did, the algorithm used is not indicated. If missing values of metabolite data were imputed, why weren't the missing values of protein data?

2) Network generation and reduction:
- They use different test-multiple correction rates for mRNA, protein and metabolites in order to obtain a similar degree of reduction. The authors should demonstrate this does not lead to any biases. The reduction objective is set at 10^7 edges, is there any reason?
- In Figure 1.D, a linear relationship is expected for the blue dots but it not perfectly linear, why? Could the scale-free reduction be modified to improve it?

3) Stochastic block models:
- How are the number of initial partitions (500) defined? Is there any relation between the suitable number of initial partitions and the network complexity (size)?

4) Functional enrichment:
- They use Reactome database for genes annotation and KEGG for metabolites, but there is no an explanation of why that choice.
- Figure 3.C is confusing. I do not understand the objective of this analysis. In order to demonstrate the biological meaning of the hierarchical SBM, the pathway distances comparison within and between blocks at every hierarchical level (as in Figure 3.B) would be better than the provided analysis.
- As mentioned before, a deeper biological interpretation of the network is necessary.
- I am not sure what the clustering analysis is contributing to. Modules obtained by WGCNA shows good results looking at the number significant of pathways, but are they biologically relevant? Are they related to breast cancer?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, functional genomics, transcriptomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 12 Aug 2019

Katharina Baum, Luxembourg Institute of Health, Strassen, Luxembourg

We thank Dr. Ana Conesa and Manuel Ugidos for their time and effort they invested into commenting our work. They raised important points that we carefully addressed in our revised version 2 of the manuscript. Please find our specific replies and the description of introduced changes below.

>This manuscript describes the application of stochastic block models (SBMs) as a method to construct biological networks from different omics data. Starting from correlation-based networks of breast cancer data, SBM delivers different modules (or blocks) of the network. In order to assess the usefulness of this approach, the authors evaluate the biological meaning of the modules obtained performing a functional enrichment analysis for each module of the network. Finally, the authors include the edge confidence score computation for each edge in the resulting network.
>This novel application of SBMs with omics data is interesting but its utility is not clearly explained in the manuscript. First, the title is confusing, since I understood a single network is obtained from different omics data, i.e. a multiomic integration analysis, however it is not performed.
>Furthermore, why are these three datasets used? It seems the authors want to evaluate the SBM-based networks starting from datasets with different complexity (different number of features/samples), but the reason is not properly explained.
>As for the SBM model definition, because of the microcanonical formulation because of its hard constraints, network underfitting could be an important issue. Although underfitting is solved using nested (hierarchical) or degree-corrected SBMs, an explanation of this

phenomena and how to deal with it would be suitable in 'Fit to SBM' methodological section.
>Regarding the functional enrichment, the choice of the pathway database is not clear. Why is Reactome chosen for genes annotation and KEGG for metabolites?
>Taking into account that breast cancer related biological findings is not the scope of this manuscript, a deeper analysis of biological information revealed by network blocks is necessary: which pathways are block-specific? Which pathway or pathways are present in the bigger block? Do the pathway distances between blocks have any relation with previous knowledge about breast cancer?
> Furthermore, the pathway distances within and between blocks are computed just for the lower level of blocks. Since the network is hierarchical, pathway distances can be obtained for each level. This would indicate if the hierarchy is biological meaningful too. Moreover, the same genes are known to be annotated to related pathways. Therefore, this expected result does not represent a validation to determine the biological meaning of the blocks in the network. A good approach could be to select a gene X and see if the rest of the genes of the pathways, where gene X is annotated, are in the same block of gene X (or in near blocks in the hierarchy). And repeat this for a high number of genes.
>Finally, the edge confidence score looks like important to improve the network characterization, however they do not use this score to optimize the final network. I was expecting the comparison between the original SBM-based network and the corrected SBM-based network via edge confidence score optimization. In order to use the edge confidence scores, a way to compute thresholds for missing and spurious edges should be proposed.
> Resuming, SBM-based biological networks are a new way to represent omics data and it represents a novel approach for SBMs. However, a novel application should be coupled with a suitable interpretation of its results. Moreover, the main conclusion or this manuscript is not clear and the main question for a new methodological development is unanswered, why should I use this method?

Response:
We thank the reviewers for the nice summary of our work. In order to sharpen the focus, we clarified in the abstract and at the end of the introduction that our goal is to pave the ground for the usage of the SBM model for different types of biomolecular networks. Investigating the capability of SBMs for representing and analysing different types of biological networks was the key challenge addressed in our article. We did not intend to compare the networks between layers, but rather assess to which extent the SBM is applicable to derive useful information in terms of (i) relevant clustering as well as (ii) network-based, alternative edge scores.
We have shown that a lot more SBM-predicted blocks have biological counterparts (i.e. more genes or metabolites associated with certain Reactome or KEGG terms are clustered together) than expected by chance, and we show in our revised version that biological processes known to be relevant to the examined phenotype can be derived (new Figs. 4, 5, Table 2). In addition, we showed that the SBM-based edge relevance scores coincide with the correlation values (which have not been given to fit the network to the SBM). These results support our hypothesis that the SBM is suitable to represent and analyze biomolecular networks in which interactions are derived from correlations. This opens the avenue to new types of analyses using the SBM and its output for which this work lays the foundation.
We hope that the reviewers will also find that the revised version supports this reasoning and the derived conclusion.

In response to other points that were raised only here but not in the specific points include below:

To improve the clarity of the description of the data usage from the beginning, we reformulated the according sentences in the abstract and introduction which now read:
'We apply SBM-based analysis independently to three correlation-based networks of breast cancer data originating from high-throughput measurements of different molecular layers: transcriptomics, proteomics, or metabolomics.'
'Here we showcase the SBM-based analysis (overview in Fig. 1A) for three networks of different molecular types, derived from either transcriptomic, proteomic or metabolomics data of breast cancer tumours.'

In order to facilitate the understanding of the different SBM versions, we added an explanation on the hierarchical and degree-corrected version of the SBM into the methods part and comment on the relationship between underfitting and the hierarchical SBM in the revised methods section.

>Specific points:

>1) Data preprocessing:
Why are the three datasets used? Is it just a matter of different complexity? Datasets are not completely characterized as the number of features is not indicated.

Response:
Protein, mRNA and metabolites are key molecules in cells, which are widely applied, in isolation or in combinations, in different biomedical research domains. Their abundance and interconnectedness in networks are therefore of high interest if aiming to characterize cells or tumorous tissue. We used them to illustrate three examples for cellular interaction networks, whose interactions have different biological meanings. Moreover, in the revised article now we explain that we use these networks to reflect different characteristics of tumorous tissue/cells. The number of features (i.e. number of mRNAs, proteins and metabolites) are given in the network characteristics in Table 1 ('entities' before NA removal and afterwards). In order to improve accessibility to the reader, we also state these numbers in the methods part of the revised manuscript and clearly identify them as feature count ('mRNA, protein, metabolite data for ER-breast cancer tumors').

>Authors say missing values of metabolite data were imputed previously. It is not clear whether they performed the imputation or not. In case they did, the algorithm used is not indicated. If missing values of metabolite data were imputed, why weren't the missing values of protein data?

Response:
We used the metabolite data as provided in Budczies et al. 2009 in which imputation had been performed. In order to clarify this, we reformulated the sentence in the methods part to 'The metabolite data did not contain missing values as imputation had been performed in the original publication[Budczies et al. 2009].'
We used the processed protein data as provided in the original publication (Mertins et al, 2014). In general, in contrast to missing values in metabolite data which are usually considered to occur

due to abundance below detection limit, missing values in protein data can have multiple different reasons making imputation less straightforward. Because we do not focus on data pre-processing here, we kept the data as close as possible to those originally published and rather took the opportunity to provide an idea of how to incorporate datasets with missing data into our proposed analysis framework.

>2) Network generation and reduction:
They use different test-multiple correction rates for mRNA, protein and metabolites in order to obtain a similar degree of reduction. The authors should demonstrate this does not lead to any biases. The reduction objective is set at 10^7edges, is there any reason?

Response:
We reformulated and describe this part in more detail in the revised manuscript. In fact, we intended to provide examples for different corrections methods and thresholds which are frequently used, and to achieve a high degree of reduction (to reduce runtime for SBM fit) while still ensuring that the resulting biomolecular network is as connected as possible (see Fig. 1C). Similarity of the degree of reduction between the three data layers was not a goal to achieve. The border of 10^7 edges arose out of computation time considerations (runtime scales with edge count for SBM fit in graph-tool, they correspond to several hours and Gbs of memory consumption for a single initialization), and thus could be adapted to each user's case.

>In Figure 1.D, a linear relationship is expected for the blue dots but it not perfectly linear, why? Could the scale-free reduction be modified to improve it?

Response:
Usually, a perfect fit to scale-freeness cannot be achieved because altering the cut-off threshold leads to removal of multiple edges at the same time; the achievable combinations of which edges belong to the network are fully determined by the correlation values associated to the edges. We show possible scale-free fit indices (that index is a measure of how closely the node degree distribution resembles a power-law, i.e. how close the network is to scale-freeness) of the networks reduced by edge thresholding in new Fig. 1D. Therein, it also becomes clear that other thresholds could be used and can lead to 'better' scale-freeness in the resulting networks – while at the same time making the resulting network less connected. Other specific requirements on the degree of scale-freeness can be adapted on a case-by-case basis.

>3) Stochastic block models:
How are the number of initial partitions (500) defined? Is there any relation between the suitable number of initial partitions and the network complexity (size)?

Response:
The number of initializations was chosen as first assessment whether this is sufficient to be able to distinguish between the different SBM types – which was the case. As is clear, for some networks different SBM types are more closely related whereas for others, they are well more separated. This will depend on the actual network/data that is fitted and serves as orientation. Of course, more initializations are always better, but it is subject to a trade-off between computation time and finding a good partition.

>4) Functional enrichment:
They use Reactome database for genes annotation and KEGG for metabolites, but there is no an explanation of why that choice.

Response:
We added the explanation to the revised version. Reactome provides a hierarchical annotation which qualifies it for comparison to the hierarchical structure given by the SBM, KEGG is one of the annotation databases used most for metabolites.

>Figure 3.C is confusing. I do not understand the objective of this analysis. In order to demonstrate the biological meaning of the hierarchical SBM, the pathway distances comparison within and between blocks at every hierarchical level (as in Figure 3.B) would be better than the provided analysis.

Response:
We altered the accompanying description of Fig. 3C. What we intended is to relate the distance of two SBM blocks in the SBM hierarchy (x-axis) to the distances between the Reactome terms associated to the blocks (y-axis). For a matching hierarchy structure of Reactome and SBM, we would expect a positive correlation, i.e. more distant SBM blocks in the SBM hierarchy having more distantly related Reactome terms. We do not find convincing evidence for this.
To clarify the approach, we restructured Fig. 3 (by moving the results for the second distance measure into the supplement, new Fig. S3).
In addition, we now also include the analysis of within-block distance vs. between-block distances for the other hierarchy levels as suggested by the reviewer (new Fig. 3D). As in our previous analysis in Fig. 3C, we cannot find strong evidence for the hierarchy of the SBM coinciding with the Reactome hierarchy. Please note that the numbers of blocks with annotation get low in higher hierarchy levels, thereby reducing the number of hierarchy levels that can be considered for this analysis.

>As mentioned before, a deeper biological interpretation of the network is necessary.
I am not sure what the clustering analysis is contributing to. Modules obtained by WGCNA shows good results looking at the number significant of pathways, but are they biologically relevant? Are they related to breast cancer?

Response:
We provide additional biological interpretations of the clustering results in the revised version of the manuscript (see new Figs. 4, 5, Table 2). They show that the SBM-derived clustering can detect biological pathways known to be implicated in breast cancer according to oncogenic signatures from MSigDB, such as extracellular matrix organization or the cell cycle (see new Fig. 4), or fatty acid biosynthesis (Table 2). In addition, the relevance of further processes can be predicted, e.g. chromatin organization (new Fig. 5).

Competing Interests: No competing interests were disclosed.

Reviewer Report 29 April 2019

https://doi.org/10.5256/f1000research.20484.r47223

?

Silvia Pineda San Juan
[1] Division of Transplant Surgery, Department of Surgery, University of California, San Francisco (UCSF), San Francisco, CA, USA
[2] Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

The manuscript provides an innovative way of using stochastic block models to perform networks for different omics data.
The idea is interesting but the application should be better characterized. I was expecting a network integrating the different datasets, but the network is performed separately per each dataset. This fact should be addressed at the very beginning to avoid confusion to the readers.
Also, the use of the three datasets is not justified, why these three and no others?
It looks like the authors want to compare the three networks applied to the three datasets (mRNA, proteins, and metabolites) but the different sample size among other factors makes them not comparable.
Finally, one of my main concern is that there is not a clear conclusion of the study, are they proposing a network that is better than the ones that already exist? (this is not assessed). Are they finding new biological insights for breast cancer? (this is not shown). The final conclusion is not clear since the authors did not give a biological example of the method application. It is not clear if the method outperforms others or if the method is able to find new biological interpretations, etc. Why people should use this method? What type of information will they obtain?

More specific points:

Data preparation and network generation:

- It is not clear to me why the authors used mRNA, protein and metabolites for this network study. Is it just because they were available, or is there a hypothesis under this selection? Why don't they use other omics data available in the TCGA? The sample size is pretty reduced for the proteomics data with only 36 samples. The other thing I do not understand is why they used metabolomics data measured in other individuals. And for the mRNA they do not give an exact sample size.

- Why do they use Bonferroni for mRNA expression and Benjamini-Hochberg for protein, metabolite and just 0.01 for protein? This should be better justified.

- They replace 0 values by NA, why? There is a big difference between a lack of expression and a missing value.

- Do they filter for those genes that have a very low expression among samples? They only

specify this for protein data, but what about mRNA expression? Are they considered only 0 for low expression or a very small cut-off normally used in mRNA analysis?

- Regarding the scale-free reduction. They explain that the technique removes weak links until met a criterion based on WGCNA package, but it does not well explain the process for this and how they applied this to the data. Please explain.

Fitting SBM:

- They built the network based on a stochastic block model (SBM) representation, the nodes of a network are partitioned into blocks according to their similarity in connectivity. It is not clear how the SBM is applied to the data and how the similarities are obtained.

For the SBM representing biological function:

- The whole module is unclear to me since they do not provide any biological or functional interpretation. I don't understand the goal of this. Figure 3 is also quite confusing.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistics, Computational biology, Data analysis, Genomics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Aug 2019

Katharina Baum, Luxembourg Institute of Health, Strassen, Luxembourg

We thank Dr. Silvia Pineda San Juan for the careful revision of our manuscript. Her

comments helped us enhancing the clarity and stringency of our presentation and the reader's accessibility to our work. We respond to her specific comments and describe the introduced changes to version 2 of our manuscript below.

>The manuscript provides an innovative way of using stochastic block models to perform networks for different omics data.
The idea is interesting but the application should be better characterized. I was expecting a network integrating the different datasets, but the network is performed separately per each dataset. This fact should be addressed at the very beginning to avoid confusion to the readers.

Response:
We thank the reviewer for her appreciation of our idea. In order to improve the clarity of the description of the data usage from the beginning, we reformulated the corresponding sentences in the abstract and introduction which now read:
'We apply SBM-based analysis independently to three correlation-based networks of breast cancer data originating from high-throughput measurements of different molecular layers: either transcriptomics, proteomics, or metabolomics.'
'Here we showcase the SBM-based analysis (overview in Fig. 1A) for three networks of different molecular types, derived from either transcriptomic, proteomic or metabolomics data of breast cancer tumours.'

>Also, the use of the three datasets is not justified, why these three and no others?
It looks like the authors want to compare the three networks applied to the three datasets (mRNA, proteins, and metabolites) but the different sample size among other factors makes them not comparable.
Finally, one of my main concern is that there is not a clear conclusion of the study, are they proposing a network that is better than the ones that already exist? (this is not assessed). Are they finding new biological insights for breast cancer? (this is not shown). The final conclusion is not clear since the authors did not give a biological example of the method application. It is not clear if the method outperforms others or if the method is able to find new biological interpretations, etc. Why people should use this method? What type of information will they obtain?

Response:
In order to sharpen the focus, we clarified in the abstract and at the end of the introduction that our goal is to pave the ground for the usage of the SBM model for different types of biomolecular networks. Investigating the capability of SBMs for representing and analysing different types of biological networks was the key challenge addressed in our article. We did not intend to compare the networks between layers, but rather assess to which extent the SBM is applicable to derive useful information in terms of (i) relevant clustering as well as (ii) network-based, alternative edge scores.
We have shown that a lot more SBM-predicted blocks have biological counterparts (i.e. more genes or metabolites associated with certain Reactome or KEGG terms are clustered together) than expected by chance, and biological processes known to be relevant to the examined phenotype are can be derived (new Fig. 4, 5, Table 2). In addition, we showed that the SBM-based

edge relevance scores coincide with the correlation values (which have not been given to fit the network to the SBM). These results support our hypothesis that the SBM is suitable to represent and analyze biomolecular networks in which interactions are derived from correlations.
This opens the avenue to new types of analyses using the SBM for which this work lays the foundation.
To strengthen our findings, as suggested by the reviewer, we now also include more biological interpretations of the clustering results (see Figs. 4, 5, Table 2). They show that the SBM-derived clustering can detect biological pathways known to be implicated in breast cancer according to oncogenic signatures from MSigDB, such as extracellular matrix organization and the cell cycle (see new Fig. 4), or fatty acid biosynthesis (Table 2). In addition, the relevance of further processes can be predicted, e.g. the chromatin organization (new Fig. 5).

>More specific points:

>Data preparation and network generation:
It is not clear to me why the authors used mRNA, protein and metabolites for this network study. Is it just because they were available, or is there a hypothesis under this selection? Why don't they use other omics data available in the TCGA? The sample size is pretty reduced for the proteomics data with only 36 samples. The other thing I do not understand is why they used metabolomics data measured in other individuals. And for the mRNA they do not give an exact sample size.

Response:
Protein, mRNA and metabolites are key molecules in cells, which are widely applied, in isolation or in combinations, in different biomedical research domains. Their abundance and interconnectedness in networks are therefore of high interest if aiming to characterize cells or tumorous tissue. We used them to illustrate three examples for cellular interaction networks, whose interactions have different biological meanings. Moreover, in the revised article now we explain that we use these networks to reflect different characteristics of tumorous tissue/cells.
Other potential data types could be e.g. mutations, copy number variation, DNA methylation or miRNAs, which are interesting avenues to further explore. While they could be useful, there are some caveats associated with them, e.g.: the derived interactions within layers are even less directly interpretable than for mRNA, protein or metabolite; networks generated from mutations and copy number variation are extremely sparse; the functional interpretation of DNA methylation data relies on transcription and the resulting networks are extremely big; the roles of miRNAs are less well known. Therefore, we decided to restrict our analyses to the three biomolecular entities mRNA, proteins and metabolites. We included these considerations when introducing the employed data layers in the results part of the revised manuscript.
Concerning the metabolite data: We are convinced that metabolites characterize a highly interesting layer of intra-tumor processes which is complementary to the gene expression associated layers mRNA and proteins. Unfortunately, metabolomics have not been measured for TCGA samples which is why we resorted to an alternative cohort for this data layer.
Please note that the number of tumour samples for each data layer were stated in the methods section (mRNA: 237, protein: 36, metabolite: 68). We now included that value for the mRNA data layer also into the main text.

>Why do they use Bonferroni for mRNA expression and Benjamini-Hochberg for protein, metabolite and just 0.01 for protein? This should be better justified.

Response:
We explain our approach for network reduction more in detail in the methods and results part of the revised version of the manuscript. In fact, at first, we applied both Bonferroni and Benjamini-Hochberg correction methods, both of which are widely used and accepted, along with the classical significance thresholds 0.01 and 0.05 to the networks of all three data layers (see Fig. 1C). We finally chose the correction method and threshold for each data layer considering a trade-off between minimal network size (i.e. minimal computation time for the subsequent fit to SBM) and maximal connectedness of the reduced network: We used the combination which provided a high degree of reduction (less than 10 million edges in the network) while maximizing the size of the largest connected component in the network. While the stringent Bonferroni correction is necessary to achieve a sufficient degree of reduction for the mRNA network, it severely disrupted the connectedness for the protein and metabolite data layer leading to less than 30% or 65% of the nodes being in the largest connected component for protein or metabolite, respectively (see Fig 1C).
Using different correction methods and significance thresholds serve as examples of typical scenarios which could be envisioned during network reduction. Finally, every user could use their own thresholds reasonable for network reduction.

>They replace 0 values by NA, why? There is a big difference between a lack of expression and a missing value.

Response:
We replaced NAs by -10 in the log-counts of the RNAseq data. As stated in the methods part, the reason for this is the following: The RNAseq data are logarithmized (relative) counts. In this case the NAs are therefore artefacts from logarithmizing a zero count. Our replacement served to reverse this artefact.

>Do they filter for those genes that have a very low expression among samples? They only specify this for protein data, but what about mRNA expression? Are they considered only 0 for low expression or a very small cut-off normally used in mRNA analysis?

Response:
Apart from the replacement of NAs by very small values (-10 in log-counts) in the mRNA data, we use the data as provided by TCGA. The rationale for this is to reduce bias with respect to which nodes we consider. Since we do not focus on expression strength, but on the connection between molecular species, i.e. co-variation of expression, also lowly abundant species could, and indeed do, play a role, i.e. they have non-zero degree in the reduced networks. We dedicate new Figure S1D to an illustration of this fact.
One can imagine multiple other criteria of entity removal prior to analysis (e.g. tissue-specific GTEx, using pathways of interest) but this is not the focus of our work.

>Regarding the scale-free reduction. They explain that the technique removes weak links until met a criterion based on WGCNA package, but it does not well explain the process for this and how they applied this to the data. Please explain.

Response:
We incorporated a more detailed explanation of this reduction technique in the methods part and added an additional panel to Fig. 1 (Fig. 1D) which illustrates intermediate results of the process. This is the revised methods part: 'For the reduction by imposing a scale-free architecture of the reduced network, we employed the pickHardThreshold *function of the WGCNA package (Langfelder et al, 2008) with the default requirement (0.85) on goodness of fit to a power-law degree distribution of the nodes. Given the symmetric absolute correlation matrix of the network edges, this function reduces the network by one of a given set of edge thresholds at a time and determines the scale-free fit index* R^2 *which lies between 0 (bad fit) and 1 (perfect fit) by comparing the resulting degree distribution of the reduced network to a power-law degree distribution. The lowest of the tested edge thresholds that gives a scale-free fit index > 0.85 is reported as estimated threshold. For the edge thresholds, we started with a grid with stepsize 0.05 between 0.3 and 0.95, refining according to the resulting estimates to vectors with stepsize 0.001 between 0.5 and 0.625 for the mRNA network, between 0.7 and 0.82 for the protein network and between 0.3 and 0.4 for the metabolite network. Finally estimated edge correlation thresholds were 0.603 (mRNA), 0.788 (protein), and 0.375 (metabolite).'*

>Fitting SBM:
They built the network based on a stochastic block model (SBM) representation, the nodes of a network are partitioned into blocks according to their similarity in connectivity. It is not clear how the SBM is applied to the data and how the similarities are obtained.

Response:
We extended the corresponding section describing the SBM in the Methods, in which we now also summarize the equations underlying the building of the SBM and the relationship between model likelihood and properties of the network graph derived from the data.

>For the SBM representing biological function:
The whole module is unclear to me since they do not provide any biological or functional interpretation. I don't understand the goal of this. Figure 3 is also quite confusing.
Response:
The assumption behind the analysis is that a predicted block structure is meaningful if many of the predicted blocks have a biological counterpart, e.g., biological pathway. Such an analysis was performed as a starting point for assessing potential biological relevance. Nevertheless, because we agree that additional biological interpretations would benefit the reader, we included other examples of biological interpretations (new Figs. 4, 5, Table 2 and accompanying description in the results section). In addition, in order to improve accessibility, we reduced the contents of Fig. 3 (by only showing results for one distance measure) and clearly indicated which SBM hierarchy level the examined blocks stem from.

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 April 2019

https://doi.org/10.5256/f1000research.20484.r47226

?

Lei Xie
Department of Computer Science, Hunter College, The City University of New York, New York, NY, USA

Yue Qiu
City University of New York, New York, NY, USA

The paper by Baum et al describes a method to construct biological networks from omics data using stochastic block models (SBMs). The application of SBM on mRNA, protein and metabolic data gives a new way of deriving information from correlation based biological networks. The proposed method could be a useful addition to network biology. To strengthen the manuscript, I would suggest the following points to be addressed:

1. In the over representation analysis, while the statistics on distance of Reactome/KEGG terms shows the block generated here is significantly better than random. Is it possible to provide a such comparison between the clustering result by SBM and WGCNA?
2. In the network reduction step, a linear relationship is expected between log-frequency and log-node-degree. While the scale-free fit index (R^2) will be > 0.85 with WGCNA default requirement, how different threshold affect that linearity is not clear from Fig 1D. Also, how well can the network be reduced by significance of correlation fit to a scale free network?
3. To demonstrate that SBM can provide biological insight, more detailed analysis on the benchmark data sets could be useful. For example, what are unique and common pathways for different breast cancer subtypes, how well are these findings consistent with existing knowledge?
4. Figure 4 shows SBM based confidence score can be used to predict the existence of an edge. However, it is not clearly stated how the edges are determined as putatively missing or spurious. Also, the description for figure 4A is hard to understand.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, systems biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 12 Aug 2019

Katharina Baum, Luxembourg Institute of Health, Strassen, Luxembourg

We thank Prof. Lei Xie and Yue Qiu for their comments that helped improving the accessibility of the presented contents and strengthened the manuscript. In the following, we answer to the specific points they raised and how we addressed them in version 2.

>The paper by Baum et al describes a method to construct biological networks from omics data using stochastic block models (SBMs). The application of SBM on mRNA, protein and metabolic data gives a new way of deriving information from correlation based biological networks. The proposed method could be a useful addition to network biology. To strengthen the manuscript, I would suggest the following points to be addressed:
>In the over representation analysis, while the statistics on distance of Reactome/KEGG terms shows the block generated here is significantly better than random. Is it possible to provide a such comparison between the clustering result by SBM and WGCNA?

Response:
We thank the reviewers for their appreciation of our work. We now also provide the computation of the SBM distances for the mRNA and protein WGCNA clustering results (see new Fig. 3F and new Fig. S3C). They show also a bigger distance between clusters than within clusters. We added the resulting p-values to the comparison of the distances for both SBM (Fig. 3B, D) and WGCNA (Fig. 3F, S3C) approach.

>In the network reduction step, a linear relationship is expected between log-frequency and log-node-degree. While the scale-free fit index (R2) will be > 0.85 with WGCNA default requirement, how different threshold affect that linearity is not clear from Fig 1D. Also, how well can the network be reduced by significance of correlation fit to a scale free network?

Response:
Fig. 1D (now 1E) has been included for illustrative purposes only to show some further characteristics (in this case: degree distribution) apart from those in Table 1 of the six networks employed for the SBM-based analysis. It can be used to appreciate the fact that the networks reduced by requirement on significance of correlation are not optimized for it but still may exhibit a close to linear relationship between log-frequency and log-node-degree. For clarification, we added the scale-free fit indices obtained for each network into old Fig. 1D - now Fig. 1E.
We also extended the description of the scale-freeness fitting procedure and approach by using

the WGCNA package (see revised Methods section). Please note that less strict correlation thresholds than the employed here lead to networks being further away from a network with scale-free characteristic (this is inherent to the WCGNA fitting approach as the least stringent threshold is determined leading to a scale-free fit index of at least 0.85) and more stringent correlation thresholds can lead to networks being closer to a scale-free network (i.e. to higher values of R^2). We give now these dependencies for the three networks in new Fig. 1D.
Please note that reducing by a threshold on significance of correlation is a different reduction method and it is neither expected nor intended that the resulting networks are scale-free. In order to illustrate the relationship between the two reduction methods, we introduced new illustrations in supplemental figure S1. Therein,
- we show in new panel S1A that for the metabolite and mRNA layer, due to the datasets not having NA values and thus the sample size being the same for each and every pair of metabolite or mRNA species, the p-values of the correlation being different from zero depend monotonously on the absolute correlation values. This is not the case for the protein dataset for which different sample sizes may occur. Therefore, the correlation value for some interactions will be backed by fewer data points only which leads to increased p-value despite the same correlation value.
- we show in new S1B that, for the mRNA and metabolite data layer, the scale-free fit index for networks reduced by different significance thresholds can be directly compared to those from Fig. 1D.
- we show in new S1C how the significance threshold of correlation being different from zero relates to the scale-free score for the protein dataset and where the four networks reduced by the four considered significance thresholds locate therein.

>To demonstrate that SBM can provide biological insight, more detailed analysis on the benchmark data sets could be useful. For example, what are unique and common pathways for different breast cancer subtypes, how well are these findings consistent with existing knowledge?

Response:
We agree that a comparison of results between breast cancer subgroups to provide biological insights into the utility of the clustering would be useful. However, our work focusses on determining whether SBMs are suitable to represent biomolecular networks as they have not been explored for molecular network analysis.
We indeed expect additional and complementary biological findings to what is already known with the SBM approach. To strengthen our work, we also added some showcase examples of biological interpretation (new Figs. 4, 5, Table 2 and Description in the text). They show that the SBM-derived clustering can detect biological pathways known to be implicated in breast cancer according to oncogenic signatures from MSigDB, such as extracellular matrix organization or the cell cycle (see new Fig. 4), or fatty acid biosynthesis (Table 2). In addition, the relevance of further processes can be predicted, e.g. the chromatin organization (new Fig. 5).

>Figure 4 shows SBM based confidence score can be used to predict the existence of an edge. However, it is not clearly stated how the edges are determined as putatively missing or spurious. Also, the description for figure 4A is hard to understand.

Response:
We extended the description of the edge detection by adding the following sentences in the

section 'Assessing edge relevance by SBM-based edge confidence scores':
'For the six reduced networks, all edges that exist in the network were considered as 'putatively spurious'. Similarly, all edges that were not in the network because they had been removed from the (fully connected) correlation-based network during the reduction procedure were considered as 'putatively missing'.'
We clarified the legend to new Fig. 6A (old Fig. 4A) that now reads:
'(A) Histograms of edge confidence scores for the best fitting SBM of the three networks reduced by criterion on scale-freeness. Top: spurious edge confidence scores, computed for all edges existing in the network; middle and bottom: missing edge confidence scores, computed for all edges that were removed during the reduction procedure (middle: missing edges between nodes with degree>0 in the reduced network, bottom: missing edges adjacent to a node of degree zero in the reduced network).'

Competing Interests: No competing interests were disclosed.