Generative network model of transcriptome patterns in disease cohorts with tunable signal strength

Piotr Nyczka and Marc-Thorsten Hütt*
Department of Life Sciences and Chemistry, Jacobs University, D-28759 Bremen, Germany

(Received 3 September 2019; accepted 13 May 2020; published 23 July 2020)

Algorithmic methods for interpreting the collective transcriptome (gene expression) patterns of disease cohorts in the context of biological networks are a cornerstone of systems medicine. The calibration of these algorithms using synthetic data with predefined statistical properties can be a relevant benchmarking procedure, facilitating the choice of the appropriate algorithm and the detailed mechanistic interpretation of the results. Here we present a generative model producing patterns of significantly up- and down-regulated genes for synthetic disease cohorts, in which the statistical agreement between the given biological network and the transcriptome patterns can be tuned. Parameters of this generative model are, among others, the size of the cohort, the number of disease-associated genes, the clustering of differentially expressed genes in the network and the network size. Several properties of the model can be analyzed analytically. In a first application of this generative model to produce test instances, we show that considering the subset of significant expression changes occurring in more than one patient of the cohort as an additional filtering step serves as an efficient noise suppression mechanism to enhance the recall of the signal contained in the data by the network connectivity.

DOI: 10.1103/PhysRevResearch.2.033130

I. INTRODUCTION

Over the last two decades a range of algorithmic methods has been developed for interpreting the collective gene expression (transcriptome) patterns of disease cohorts in the context of biological networks (e.g., Refs. [1–7]). These methods are in the broader context of interpreting clinical data using biological networks [5,8–12] as a path towards a systems-level understanding of medical data, i.e., the emerging discipline of systems medicine. In this way, medical research undergoes a transformation similar to the one we have seen in Biology with the advent of systems biology [13–15].

One of the components, which research in systems medicine is currently lacking, are generative models capable of producing data sets with similar statistical properties as real clinical (cohort) data, where the signal type and signal strength can be tuned via specific model parameters, thus allowing medical researchers to test, compare—and hence better select—and calibrate their data analysis methods.

Creating such generative models for medical data sets can be a novel and highly transformative contribution of statistical physics to medical research. Over decades already, physics has contributed to systems biology in a multitude of ways (see Refs. [16–19] for overviews; see Refs. [20–23] for specific examples). The design of generative models has for a long time been an important component of statistical physics. Examples include suitable random graphs models to benchmark the analysis of real networks [24–26], as well as the multitude of interdisciplinary applications of coupled phase oscillators [27] and the Ising model, as well as its derivations [28–30].

The goal of "network medicine" [31], the statistical analyses of high-throughput ("omics") data from a network perspective, is to identify systematic differences between healthy and disease states or between different diseases [9]. The algorithms differ in the processing of the given biological network and the transcriptome data characterizing the disease cohort and the controls, as well as in the quantification of the network patterns obtained from the individual transcriptome profiles and the various types of filtering and binning steps applied to the experimental data (see, e.g., Refs. [4,32–34]). Due to the often small cohort sizes, the unknown numbers of disease-associated genes, the noise in the transcriptome data and uncertainties in the biological networks it is in general unclear, what "signal strength" can be expected in such an analysis.

Here we introduce a simple model for generating dichotomized gene expression (transcriptome) profiles for disease cohorts. By "dichotomized" we mean that for each gene and patient we generate only the presence or absence of this gene in the set of differentially up- or down-regulated (see Ref. [4]) genes. Furthermore, we study the properties of this stochastic model and the statistical features of the generated disease-specific transcriptome profiles. For some key features of the model we have been able to derive semianalytical expressions, as well. Such disease cohort transcriptome data will differ substantially in signal strength: often it will not be possible to extract any information on the set of disease genes (no discernible signal).

Sometimes, statistical methods applied to the individual patient level will reveal features of the set of disease genes without the need to include any network information (strong signal). Often only the combination of network information

*m.huett@jacobs-university.de

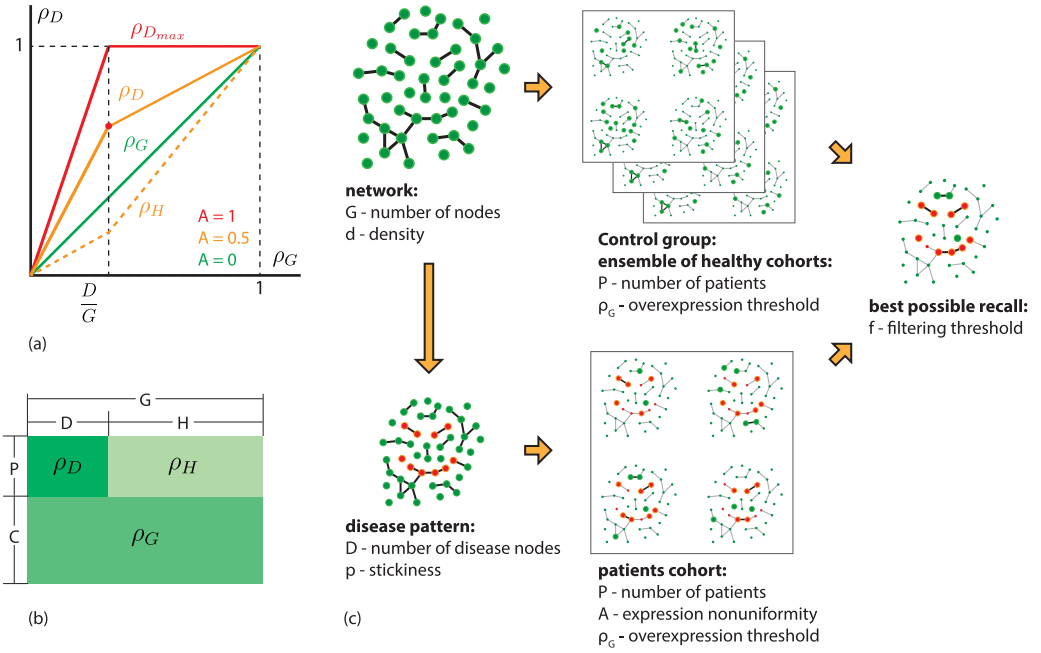

FIG. 1. [(a) and (b)] Depiction of expression densities ρ_G, ρ_D, and ρ_H and their dependence on the parameters G, D, H, and P. Here, G is the number of genes in the network, D is the number of disease related genes, H is the number of disease unrelated genes, $H = G - D$, while P is the number of patients and C is the number of controls. For the densities, ρ_G is the expression density in the control group, ρ_D is the expression density of disease-related genes in the patient group and ρ_H is the expression density of other genes in the patient group. The quantity A is the parameter scaling the statistical over-representation of disease genes among patients. (c) Schematic representation of our data generation and analysis pipeline, including disease pattern creation, generation of cohorts and recalling the disease pattern via filtering.

and cohort information will allow access to some properties of the set of disease genes (weak signal). This weak signal scenario is the one addressed here. Our generative model of transcriptome patterns in disease cohorts has been designed to create data sets with very small differences between patients and controls. For this case of weak signals, we show how the interplay of statistical properties and network information allows extracting information on the set of disease-associated genes.

If the biological network is meaningful for the situation at hand (i.e., for characterizing the gene expression data) we can expect that the differentially expressed genes display a certain amount of clustering in the network. By now there is ample evidence linking neighborhood in biological networks with disease-gene associations [10,35] and gene coexpression [1,3].

In our generative model, we mimic this expected clustering by creating a set of V_D disease-associated genes via a version of a random walk algorithm on the given network. With this minimal model we thus provide a conceptual tool for exploring key ideas in systems medicine.

In the following, we will describe the model (Sec. II), next we will illustrate the capabilities of the generative model using a simple case study (Sec. III). Then we provide broad description of the results discussing disease cohorts and filtering, together with some analytical insights (Sec. IV). Lastly, we put our results in the context of systems medicine (Sec. V).

II. THE GENERATIVE MODEL

The foundation of our generative model of stylized transcriptome profiles for disease cohorts is the existence of a small systematic difference between patients and controls. The control group can be expected to be phenotypically highly diverse. Patients, however, even though similarly diverse, at least agree in one phenotype (the disease). Thus if the systematic gene expression differences between patients and controls can be functionally linked to the disease, we should see a clustering of the corresponding genes in a biological network relevant to this functional interpretation.

The generative part of the model consists of generating a disease, then a cohort and control group. These steps are described in the present section. The resulting sets then enter the analysis pipeline, which is described and further studied in Sec. IV. All these steps are illustrated in the following schematic depictions: disease and cohort generation: Fig. 1 (right), details are depicted in Appendix B.

A. Disease generation

Let $\Gamma(V_G, E)$ be a graph representing a biological network, where vertices V_G denote genes linked (via some biological relationship) by edges E.

Examples of such networks are gene-centric representations of metabolic networks [4,36], transcriptional regulatory networks [37] and protein-interaction networks translated into genes via gene-to-protein mappings [38,39]. Such relationships between gene expression patterns and network properties have for a long time been a topic of interest in physics. In [40] random matrix theory has been applied to gene expression patterns. Reconstructing gene expression levels pertaining to individual cell types in samples containing

mixtures of cell types has been studied in Ref. [41]. In an application of the theory of stochastic processes to gene expression, the impact of noise correlation time on regime shifts (sudden changes in systemic behavior) has been analyzed [42]. The broad question, how connectivity can be inferred from the response dynamics of a system, has been addressed in Ref. [43]. In Ref. [44], constraints on gene expression have been derived from fluctuation theorems. In an application of Boolean networks to real gene regulatory systems, the link of network states to an epigenetic landscape has been outlined [20].

As discussed above, we assume that the signal discriminating between patients and controls in a disease cohort is associated with a set of disease-associated genes. If the network Γ is relevant for the disease mechanisms, we can expect that the disease genes display some type of (occasionally weak) clustering on the network.

We generate suitable gene sets via a random walk with occasional jumps (sometimes also called a "teleporting random walk" [45]) with a teleportation probability $1-p$. In this way, the amount of clustering displayed by the set of genes (nodes) generated by the random walk is regulated by the parameter p. For the purpose of our investigation, this set of genes represents the *disease*. The full procedure can be summarized as follows. (1) Randomly draw one gene from the network and mark (color) it as disease-related. (2) With probability p color one randomly chosen neighbor of the last colored gene and repeat step 2 or with probability $1-p$ go to the step 1.

Vertices colored in the course of this procedure indicate disease-related genes and form the set V_D. The parameter p regulates statistical properties of the set V_D such as the average cluster size. The greater p, the larger is the average cluster, which sizes follow a Poissonian distribution. For $p=0$, there is no clustering at all and disease genes are distributed randomly, while for $p=1$ only one large cluster forming a connected subgraph of $\Gamma(V_G, E)$ is created.

B. Cohort generation

We now turn to the generation of the *disease cohort*. A typical step in the analysis of real (experimental) transcriptome data is to identify significantly up- and down-regulated genes via some suitable statistical test, leading to a binarized version (a gene is part of the differentially expressed set or not) of the data. In our model, the transcriptome data are already generated in this binarized form.

We deliberately decided not to distinguish between up- and down-regulated genes. There is a multitude of methods for extracting gene sets from expression patterns with differential gene expression being only one possibility [see the diverse discussions in 46–49]. An alternative is to consider the highest percentile of expression levels of a gene, as it was done in [4] and [7].

In the following, we describe how the corresponding sets of differentially expressed genes for an individual patient or member of the control group (sets $V^*_{G_i}$) are generated. Here and in the following the asterisk indicates that these gene sets are the *marked* or selected sets (as opposed to the "whole" static sets like V_G).

We assume that for some diseases, the disease-related genes $V_D \subseteq V_G$ form clusters in the given biological network Γ. Furthermore, we assume that patients have a slightly higher probability for genes from the set V_D to be differentially expressed than the controls. Across the whole cohort this results in a set $V^*_D \subseteq V_D$, which is the data representation of the (static) set V_D. Additionally, patients can also differentially express genes unrelated to the disease (the rest of the set V_G), $V_H = V_G \setminus V_D$.

The whole set of differentially expressed genes of the patient j is $V^*_{G_j} = V^*_{H_j} \cup V^*_{D_j}$. Usually for clinical cohorts focusing on a specific disease a group of patients $S_P = \{P_j\}$ and the control group $S_C = \{C_i\}$ exist, where $|S_P| = P$ and $|S_C| = C$.

An advantage of large control sets, $C \geqslant P$, is that subsets of S_C of the same size as the patient set can be created and, hence, the comparison of S_C and S_P can be done on a broader statistical level (for several subsets of S_C) and hence leads to a more reliable estimate of V_D. As long as $C \geqslant P$, the size of the control group will not show up as a separate parameter of our analysis.

The process of generation of artificial expression profiles adheres to the following algorithm.

(1) Control group. For each member of the control group take a random subset of the set of all genes $V^*_{C_i} \equiv V^*_{G,i} \subseteq V_G$ with the probability for each gene to be chosen being ρ_G.

(2) Patients. For each patient randomly select two subsets of genes with different probabilities: $V^*_{D_j} \subseteq V_D$ with probability ρ_D and $V^*_{H_j} \subseteq V_H$ with probability ρ_H. Then take the union of these sets, $V^*_{P_j} \equiv V^*_{G_j} = V^*_{D_j} \cup V^*_{H_j}$, as an individual expression profile.

When filtering (see below) is applied to these sets, we obtain the patient and control gene sets V^*_P and V^*_C discussed in the illustrative example above. The following shorthand notations for the sizes of sets:

$$\begin{array}{lllllll} G & \equiv & |V_G|, & G^*_i & \equiv & \left|V^*_{G_i}\right|, \\ H & \equiv & |V_H|, & H^*_j & \equiv & \left|V^*_{H_j}\right|, \\ D & \equiv & |V_D|, & D^*_j & \equiv & \left|V^*_{D_j}\right|, \\ G & = & H+D, & G^*_j & = & H^*_j + D^*_j. \end{array} \tag{1}$$

There are two important assumptions with regard to probabilities ρ_x. (1) There is no difference in the average number of expressed genes between individual patients from S_P and controls from S_C. (2) Patients S_P tend to express disease related genes V_D slightly more often than the rest V_H and than the healthy group S_C. Thus $\rho_H \leqslant \rho_G \leqslant \rho_D$.

Within our generative model and under these two assumptions the densities can be parameterized in the following way:

$$\begin{aligned} \rho_{D_{\max}} &= \min\left[1, \rho_G \frac{G}{D}\right], \\ \rho_D &= \rho_G + A(\rho_{D_{\max}} - \rho_G), \\ \rho_H &= \rho_G - \frac{D}{H}(\rho_D - \rho_G), \end{aligned} \tag{2}$$

where A is the parameter scaling over-representation of disease related genes in the group of patients, hence

$\rho_D(A=0)=\rho_G$ and $\rho_D(A=1)=\rho_{D_{\max}}$. For reference, see Fig. 1 (bottom left), where the relation between densities is depicted. Artificial expression profiles created in this way are ready for further analysis.

Note that we distinguish two steps: generating the disease (where V_D appears) and generating the cohort (where affected genes are randomly sampled from V_D and, to a lesser probability, from $V_G \setminus V_D$). The set V_P is then the set of genes extracted (e.g., via filtering) from the patient cohort. The corresponding set for the controls is V_C.

III. A FIRST ILLUSTRATION OF THE GENERATIVE MODEL

Before describing results generated by the model and the network-based data analysis method, it is worth to analyze a specific example, illustrating how a signal embedded in the data generated with our model can be extracted via the tuning of the data analysis ("filtering") parameters.

Let us assume a disease with a certain set V_D of disease-related genes, $D=|V_D|$. In the "disease generation" part of the model, this set is generated by a random walk on a given biological network $\Gamma=\Gamma(V_G,E)$. The nodes of the network are genes (a set V_G, with $V_D \subset V_G$) and a link describes some biological interaction (forming the edge set E).

In the "cohort generation" part of the model, two subsets of data are generated. (1) Patients. These are represented by sets of genes with a slight over-representation of disease-associated genes (for a patient, the probability of selecting a gene from the set V_D, rather than from the remaining set $V_H = V_G \setminus V_D$, is regulated by the parameter A). (2) Controls. The gene sets for the healthy controls are randomly drawn from the whole set V_G.

Cohorts generated in such a way vary in cohort sizes (numbers of patients $|P|$, number of controls $|C|$), clustering of disease genes in the network (clustering parameter p), numbers of genes $G=|V_G|$, and disease-associated genes D.

Another parameter is the number of differentially expressed genes (regulated via the density parameter ρ_G). This parameter can be varied during the statistical analysis via a (fractional) threshold for determining differentially expressed genes (see, e.g., Ref. [4]).

The set of differentially expressed genes, together with their multiplicities, in the patient group and the control group are then subjected to further statistical analysis. Here, any data analysis technique from the literature can in principle be applied to this stylized representation of transcriptome profiles of a disease cohort. Here we illustrate this procedure using an extension of the "network coherence" method discussed in Refs. [3,4], where the connectivity of subgraphs spanned by gene sets is evaluated. The two ingredients of our statistical analysis are *multiplicity filtering* and the computation of *subgraph connectivity*.

Multiplicity filtering of strength f means that a set of candidate disease genes is constructed from the patient gene sets, which consists only of genes occurring more than f times in the patient gene sets. The same is done for the controls.

Subgraph connectivity essentially assesses the agreement of a set of genes with the network. Here we use the following definition: Given a set of genes V_G^*, we consider the induced subgraph $\Gamma|_{V_G^*}$ consisting only of nodes from V_G^* and the subset of edges from E among nodes from V_G^*. The connectivity $c_{V_G^*}$ is the number of nodes in V_G^* with nonzero degree in this induced subgraph divided by the total number of nodes in the subgraph (i.e., by $|V_G^*|$).

Variation of f (selecting genes from the individual gene sets into a common set according to their multiplicities) and ρ_G (varying the number of differentially expressed genes and, hence, the average size of the individual gene sets) now allows us to estimate the set of disease-associated genes V_D. At a fixed choice of f and ρ_G we obtain one common set V_P^* for the disease group and one common set V_C^* for the controls.

Figure 2 shows the comparison of the two sets V_P^* and V_C^* with the disease gene set V_D for different values of ρ_G and f. Subsets of nodes, together with the links among them, are highlighted in the following way: true positives (genes in the intersection of V_P^* and V_D): red; false negatives (genes in V_D, but not in V_P^*): yellow; false positives (genes in V_P^*, but not in V_D): green. Furthermore, the connectivities c_P^* and c_C^*, asf well as their difference Δc are indicated. One can see the variation of the reconstruction quality (many red, few green and yellow subgraphs) with the data analysis parameters f and ρ_G. Also, the figure provides a first indication that the Δc can serve as a measure for this reconstruction quality.

The generative model described so far is also made available via a web application, see Ref. [50]. Note that default settings of this app allow to reproduce Fig. 2

In the following, we will compute this difference of connectivities between the gene sets derived from patients and controls as a quality measure. The analysis strategy will then be to vary the data analysis parameters such that Δc is maximized. We expect that the resulting set V_P^* is the best predictor of V_D.

IV. RESULTS

A. Individual connectivity

The first and simplest analysis strategy is to study, how the distributions of connectivity c (defined as a fraction of nonisolated nodes in the subnetwork spanned by differentially expressed genes) vary between the patients and controls. Within the framework of our model of disease cohorts, the averages of these connectivities can be obtained analytically (see Appendix A for details) and are given by the following expressions:

$$\begin{aligned} c_{C_i^*} &= 1-(1-d)^{G_i^*-1}, \\ c_{D_j^*} &= 1-(1-\rho_D\ p)^2\,(1-d)^{H_j^*+D_j^*-3\rho_D}, \\ c_{H_j^*} &= 1-(1-d)^{H_j^*+D_j^*-1} \\ c_{P_j^*} &= \frac{D_j^*\ c_{D_j^*}+H_j^*\ c_{H_j^*}}{D_j^*+H_j^*}, \end{aligned} \tag{3}$$

where $c_{C_i^*}$ is connectivity for the healthy individual i, and $c_{P_j^*}$ is connectivity of the patient j.

This attempt does not always proof successful, as the difference in c between these two groups might rarely be detectable, see Fig. 3 (where three of four panels present no discernible difference between patients and controls). The

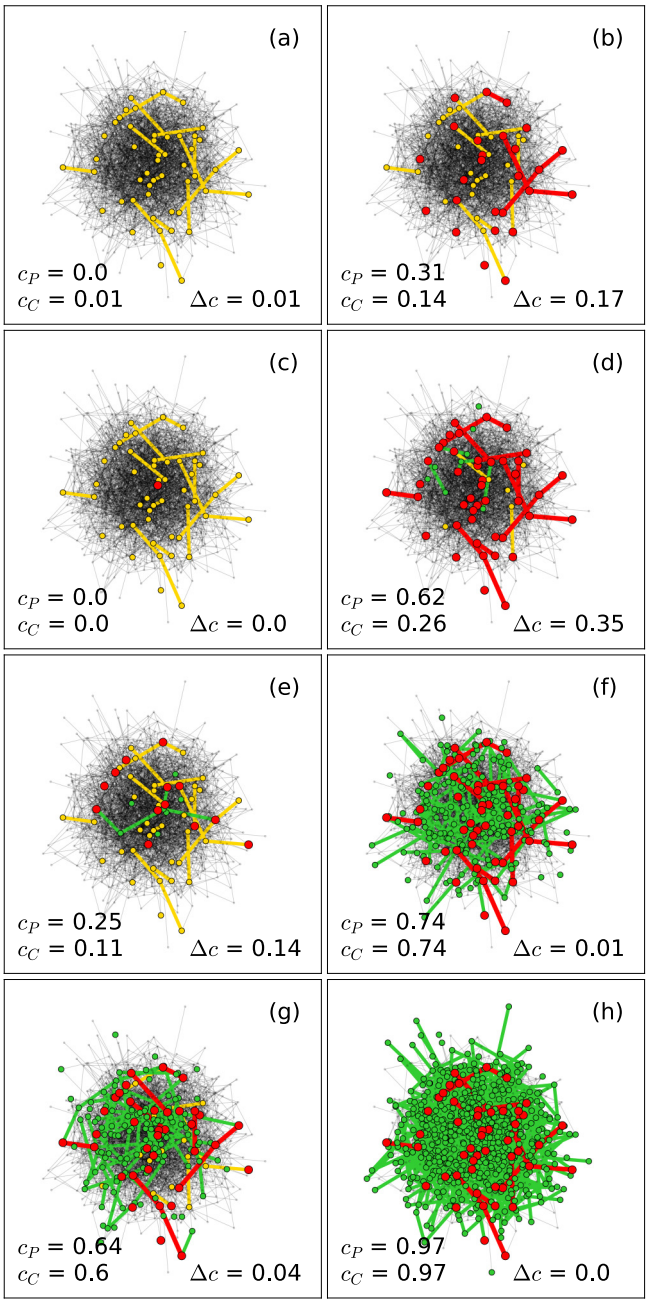

FIG. 2. Eight different combinations of the parameters ρ_G and f resulting in eight different sets, i.e., groups of selected ("colored") nodes. Color coding is as follows: red: true positives $V_D \cap V_P^*$, yellow: false negatives $V_D \setminus V_P^*$, and green: false positives $V_P^* \setminus V_D$. A perfect set would have only red nodes and none of green or yellow. The closest to it is (d). It also has the highest Δc (where Δc indicates how strongly the ratio of nonisolated nodes (connectivity) of the patient group exceeds the one derived from the control group). Other parameter values: $G = 1000$, $d = 0.006$, $D = 50$, $P = 20$, $p = 0.3$, and $A = 0.25$. Different columns stand for subsequent values of ρ_G: [(a), (c), (e), and (g)] 0.01 and [(b), (d), (f), and (h)] 0.05. Rows denote different values of f: [(a) and (b)] 5, [(c) and (d)] 3, [(e) and (f)] 1, and [(g) and (h)] 0.

reason is that usually D, p, and A are rather low (small set of disease genes, low clustering in the network, small enhancement of differentially expression for disease genes).

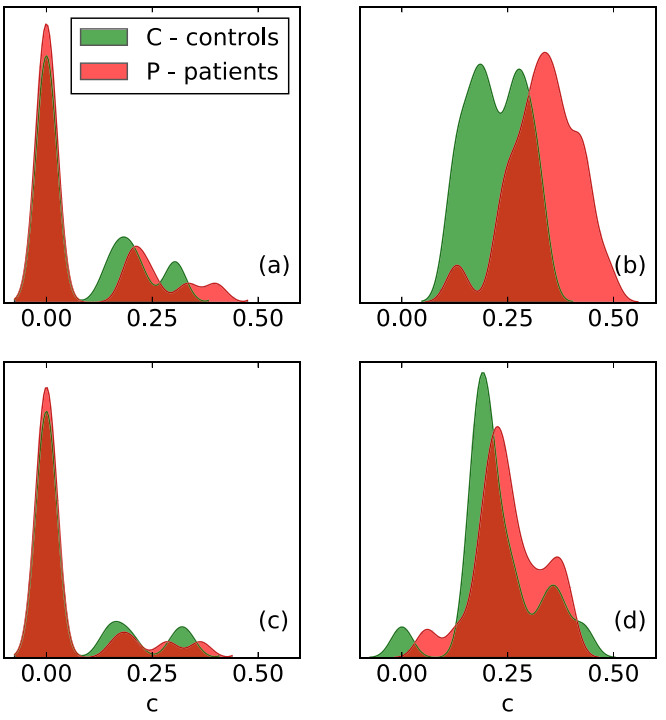

FIG. 3. Comparison of individual connectivity distributions for $P = 20$ patients (red) and $C = 20$ controls (green) for otherwise the same parameters as in Fig. 2: $G = 1000$, $d = 0.006$, $D = 50$, and $p = 0.3$. The dark red color indicates the overlap of the two histograms (red+green). From left to right: more differentially expressed genes; from bottom to top: stronger disease-control differences. Different columns stand for subsequent values of ρ_G: [(a) and (c)] 0.01 and [(b) and (d)] 0.05. Rows denote different values of A: [(a) and (b)] 0.5 and [(c) and (d)] 0.25.

Comparison of the bottom row of Fig. 3 with Fig. 2 shows that even in the cases of weak signal and no noticeable difference on the individual level, the filtering procedure can allow extracting information on the disease gene set.

B. Collective connectivity

In the following approach, there is no attempt of interpreting the data on an individual level. Instead, the group of the patients is analyzed as a whole. While individual information is lost in this analysis, the set of disease-associated genes, as well as functional clusters in the given biological network can be extracted with higher quality. Extraction of this information consists roughly of three steps.

(1) Filtering. This step combines information on differential gene expression from single patients into one group, and then filters genes by the number of occurrences. This is a noise reduction step in order to unveil general underlying biological mechanisms driving the disease for more details, see Fig. 8 in Appendix B.

(2) Projection. This step is the process of putting filtered data onto a given network.

(3) Maximising. The first two steps are meant to be performed as a function of ρ_G and f, such that a set of these parameters is searched maximising the difference between connectivities of the patients group c_S and the controls c_C.

This maximization allows to get the best possible recall of the set of disease-related genes V_D for more details, see Fig. 9 in Appendix B.

The exact implementation of these steps is rather straightforward and intuitive. Filtering is summing up sets of expressed genes to create multiset of all differentially expressed genes among control C and patients group P:

$$M_C^* = \sum_{i=1}^{C} V_{C_i^*} = \left\{v_k^{m(v_k)}\right\}_C,$$
$$M_P^* = \sum_{j=1}^{P} V_{P_j^*} = \left\{v_k^{m(v_k)}\right\}_P. \tag{4}$$

Notice, that while $V_{C_i^*}$ or $V_{P_j^*}$ are the sets of differentially expressed genes, M_C^* and M_P^* are multisets, which means that the same elements can be present multiple times in the same multiset, where $v_{P_j^*}$ is a single gene and $m(v_{P_j^*})$ is the multiplicity of this gene in combined multiset M_P. The noise suppression step now consists of accepting only genes repeating itself more than f times, where f is the filtering threshold:

$$V^{*f} = \{v_k | m(v_k) > f\}, \tag{5}$$

where V^{*f} is not a multiset anymore but becomes just an ordinary set again. In the following we will show that this filtering procedure performs well in the task of identifying disease-related genes. We expect this to be true not only for the synthetic data studied here, but also for real disease cohorts, particularly when the signal is weak.

On this collective level, the calculation of connectivity is slightly more complicated than in the individual case, but still remains feasible. At first, the probability $R_X \equiv R_X(P, \rho_X, f)$ of given gene to remain in the set V^{*f} after filtering is calculated:

$$R_X = \begin{cases} \sum_{i=f+1}^{P} \binom{P}{i} \rho_X^i (1-\rho_X)^{P-i} & \text{for } f \geqslant 0 \\ \sum_{i=1}^{|f|} \binom{P}{i} \rho_X^i (1-\rho_X)^{P-i} & \text{for } f < 0 \end{cases}, \tag{6}$$

where $X \in \{G, D, H\}$ and f is the filtering level. While the typical filtering procedures will use positive integer values for f, the procedure can also be extended to nonpositive integer values.

(1) For $f > 0$, all the genes appearing more than f times in the multiset M^* are taken: $V^{*f} = \{v_k | m(v_k) > f\}$.

(2) For $f = 0$ all the genes from the multiset M^* are taken: $V^{*f} = \{v_k | m(v_k) \neq 0\}$.

(3) For $f < 0$ all the genes appearing less or equal than $-f$ times in the multiset M^* are taken: $V^{*f} = \{v_k | m(v_k) \leqslant -f\}$.

Hence it is obvious that $V_P^{*f} \cup V_P^{*-f} = V_P^{*0}$. We expect that, while positive values of f will allow an identification of the *disease mechanisms*, negative values of f will address the diverse *disease coping mechanisms* of the individuals, as well as other individual traits, in the cohort.

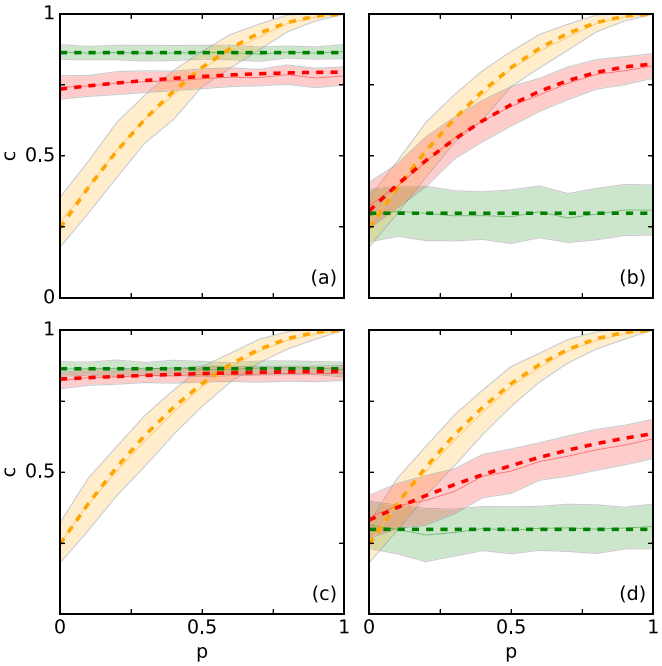

FIG. 4. Connectivity c for collective expression for different A, p, and f (while parameters D, P, and ρ_G are fixed). Simulations are made on an Erdős-Rényi graph with an edge density of $d = 0.006$. Red: c_P^{*f}, green: c_C^{*f} and orange: connectivity of the complete pure disease related gene set $c^f(V_D)$. Solid lines are from the numerical experiment with filled areas denoting the span of $\pm\sigma$, while dashed lines are the average values obtained analytically. Different columns stand for subsequent values of f: [(a) and (c)] 0 and [(b) and (d)] 1. Rows denote different values of A: [(a) and (b)] 0.5 and [(c) and (d)] 0.25.

By analogy with the previous expressions [Eqs. (3)], we obtain

$$\begin{aligned} c_C^{*f} &= 1 - (1-d)^{G^*-1}, \\ c_{D^*}^{f} &= 1 - (1 - R_D\, p)^2\, (1-d)^{H^*+D^*-3R_D}, \\ c_{H^*}^{f} &= 1 - (1-d)^{H^*+D^*-1}, \\ c_P^{*f} &= \frac{D^*\, c_{D^*}^{f} + H^*\, c_{H^*}^{f}}{D^* + H^*}, \end{aligned} \tag{7}$$

where we use $X^* = R_X\, pX$, which is true on average in the realizations of cohorts, with $X \in \{G, D, H\}$.

Here, c_P^{*f} is the collective connectivity calculated from the group of patients, while c_C^{*f} is the collective connectivity obtained from the control group in the way that P individuals from the control group are taken randomly. If the size of the control group allows, this step is repeated many times, such that c_C^{*f} becomes the average connectivity of the control group, providing us with a null model to properly assess the patient-derived collective connectivity, c_P^{*f}. In particular, it can be assessed, how the difference between c_P^{*f} and c_C^{*f} changes with the filtering threshold f, densities ρ_x and the number of patients P. In case of a statistically significant difference between the two connectivities one can claim that the disease related cluster of genes has been detected. Figure 4

depicts how signal (red) and noise (green) are changing as a function of the disease signal parameter A and the filtering strength f. Also, the connectivity of the full set V_D (orange) is provided as a reference. Thus, the orange curve indicates the maximal signal strength, which can be obtained by data analysis. Note that at low clustering (low p) due to differences in the sizes of gene sets, the connectivity of the full set can be systematically *smaller* than the connectivities derived from the other two sets. Furthermore, the numerical results (solid lines, together with their standard deviations) are compared with the analytical predictions (dashed lines). It is clear, from this Fig. 4, that agreement between numerical and analytical results is clearly seen. At $A = 0.5$, we can see that an increase of f reduces the background signal c_C^{*f} and enhances the disease signal c_P^{*f}. This is not always the case, because for higher f signal can degrade again, for more details see Figs. S1–S4 in Ref. [51]. In Fig. 4 with increasing p and a suitable filtering, the red curve, representing signal, should be more and more different from the green curve, representing noise. The orange curve is the connectivity of the subgraph spanned by V_D and hence the maximally achievable signal strength.

Figures S1–S4 [51] show these results for a wider range of A and f, confirming the general observations from Fig. 4. Furthermore, they show the interplay, in the generated data, between multiplicities, patient cohort size, and the signal strengths A and p, resulting in a nontrivial dependence of the accuracy on filtering: Moving along a row (i.e., at fixed A, varying f), we see in many cases the red curve (signal) moving away from the green curve (noise) towards the orange curve (maximally achievable signal strength) and then back again. A more detailed view on this nontrivial dependence is offered by Fig. S9–S14.

At lower values of A the disease signal c_P^{*f} is only visible at high clustering p of disease genes. This points to the "multiplicative" nature of the two disease-related parameters (disease signal parameter A and network signal parameter p); if A is not very high, a strong clustering of disease genes in the network is needed to detect the set of disease genes.

The difference between the red and the green curves is an indicator, how well the network-based analysis is capable of discriminating between the patient group and the control group. Hence, it is helpful to study this difference Δc as a function of the parameters involved. This will be addressed in the following sections.

C. Optimal filtering

The aim of the method described above is to extract systematic differences between gene expression patterns of patients and controls with respect to a given network and, furthermore, obtain information on the set of disease-associated genes, V_D. As a result of the data analysis one obtains filtered sets of genes V^{*f}. These sets then allow to calculate c values. For the group of patients S_P, $V_P^{*f} \to c_P^{*f}$ and for the control group S_C, $V_C^{*f} \to c_C^{*f}$. It has been stated before, that in the case of the control group, the average over the ensemble of many group compositions having P individuals can be taken.

When confronted with real data only two parameters which can be set: the threshold for differential gene expression leading to the gene expression density ρ_G and the filtering threshold f. The rest of the parameters of the model is coming from the experimental setup or can be obtained (or estimated) from biological knowledge, and cannot be changed in the process of the data analysis. We therefore need to understand, how the reconstruction of the set V_D depends on the values of ρ_G and f, in a balance of maximizing information extracted from the data while minimizing noise.

It is helpful to summarize the parameters and observables of our generative model again, as the optimal values of ρ_G and f will depend on these parameters, as well as on our capability to extract the connectivity differences from the observables.

In general, the parameters of the model can be divided into three categories: *unknown* (D, A, and p) parameters are coming from the biological reality; they can be estimated from the data after detailed analysis (Sec. IV D), but are unknown in the beginning; *known* (d, P, and G) parameters are the properties of the experimental settings and are fixed in the analysis process; and *variable* (ρ_G and f) parameters can be set during the process of data analysis. The observables are classified in two categories: *direct observables* (G^*, c_P^{*f}, and c_C^{*f}), which are obtained directly from the cohort data, and *hidden observables* (D^*, H^*, c_D^{*f}, and c_H^{*f}), which need to be estimated indirectly from the data and, hence, are procedure-dependent (see Sec. IV D).

1. Calculation of optimal filtering from model parameters

If all parameters are available, the optimal or near-optimal values of f and ρ_G can be calculated directly. Optimal values of f and ρ_G will be denoted as $\bar{f}$ and $\bar{\rho_G}$ and the values are obtained as follows. From Fig. 1, we see that the fraction ρ_D/ρ_H has its maximum for $\rho_G \in (0, \frac{D}{G}\rangle$, while the difference $\rho_D - \rho_H$ is greatest for $\rho_G = \frac{D}{G}$ and this is the searched value of this parameter, because in this specific point both the fraction and the difference of ρ_D and ρ_H have their maximal values. Hence,

$$\bar{\rho_G} = \frac{D}{G}. \tag{8}$$

Having $\bar{\rho_G}$, we can now compute $\bar{f}$. We start from the number of occurrences of a given gene in the patient multiset M_P^*. In the patient cohort there are two types of genes, V_D and V_H. Optimization strategy is about extracting the greatest possible number of V_D genes, while keeping the extracted V_H genes on the lowest possible level. The probability of a given gene to be present in the patient cohort more than f times is

$$R_X = \sum_{i=f+1}^{P} \binom{P}{i} \rho_X^i (1-\rho_X)^{P-i} \tag{9}$$

and the probability of exactly n appearances is

$$r_X(n) = \binom{P}{n} \rho_X^n (1-\rho_X)^{P-n}. \tag{10}$$

Furthermore, the average number of appearances is just

$$\langle n_X \rangle = P\,\rho_X. \tag{11}$$

Without a detailed knowledge about the shape of the distributions involved, the most reasonable choice is to put the filtering threshold f in between of $\langle n_H \rangle$ and $\langle n_D \rangle$ in equal distance from both:

$$\bar{f} \approx \tfrac{1}{2}(\rho_D + \rho_H)P. \tag{12}$$

This choice provides a near-optimal recall.

2. Extraction of optimal filtering via data analysis

The situation of identifying the optimal choices of parameters ρ_G and f is substantially different, when the key parameters characterizing the disease, D, A, and p are unknown. In particular, the network now becomes a crucial part of the data analysis procedure. As discussed above, the main observable of our analysis is the connectivity signal strength defined as the difference between the connectivity obtained from the group of patients, c_P^{*f}, and from the control group, c_C^{*f}:

$$\Delta c = c_P^{*f} - c_C^{*f}. \tag{13}$$

As shown in Appendix C, the quality Q (also defined in Appendix C) of the retrieved set of genes is positively correlated with Δc, which is easily obtained from the data analysis. Moreover, this correlation persists at relatively high levels for a broad range of (known and unknown) parameters.

On these grounds, the full method can now be established. In the model, Δc depends on all the parameters, but in the data analysis part only two are changed, then $\Delta c \to \Delta c(\rho_G, f)$. and the same goes for $Q \to Q(\rho_G, f)$. A sweep of the parameter space $\rho_G \times f$ is required, in order to find $\Delta c_{\max}$ which is very likely to give also $Q_{\max}$ (or a quality value close to it) and, hence, the set of the genes of the best possible quality. $Q_{\max} = 1$ means that whole set of disease related genes is extracted and none of the unrelated ones.

D. Estimation of disease parameters

From the data analysis and the analytical equations discussed so far, we can estimate the disease parameters. After the scan of the $\rho_G \times f$ space, the point where $\Delta c(\rho_G, f)$ is maximal, has been obtained. For the sake of simplicity, we assume that this point is unique. In practice (in particular in small empirical data sets) noise can lead to multiple, coexisting optimal or near-optimal points, which would then need to be evaluated separately. Furthermore, we assume that this point coincides with highest quality Q. In this case, the maximization of Δc leads to the optimal parameters $\bar{\rho_G}$ and $\bar{f}$: $\Delta c_{\max}(\rho_G, f) = \Delta c(\bar{\rho_G}, \bar{f})$. This enables us to subsequently estimate values of the disease-related parameters: D, A, and p as $\tilde{D}$, $\tilde{A}$, and $\tilde{p}$.

This estimation procedure is performed in a hierarchical order. First $\bar{\rho_G}$ yields $\tilde{D}$, then $\bar{f}$ and $\bar{\rho_G}$ yield $\tilde{A}$ and lastly $\tilde{D}$, $\tilde{A}$, and $\Delta c_{\max}$ yielding $\tilde{p}$. This procedure acknowledges the underlying dependence scheme of the quantities involved: $\bar{\rho_G} \to \bar{\rho_G}(D)$, $\bar{f} \to \bar{f}(D, A)$ and $\Delta c_{\max} \to \Delta c_{\max}(D, A, p)$. Or in reverse fashion: $\tilde{D} \to \tilde{D}(\bar{\rho_G})$, $\tilde{A} \to \tilde{A}(\bar{\rho_G}, \bar{f})$, $\tilde{p} \to \tilde{p}(\bar{\rho_G}, \bar{f}, \Delta c_{\max})$.

1. Estimation of D

At first, the estimation of $\tilde{D}$ from $\bar{\rho_G}$, is based on Eq. (8) and yields

$$\tilde{D} = \bar{\rho_G}\, G, \tag{14}$$

where $\bar{\rho_G}$ is the value of ρ_G for which Δc is maximal. This is easily done with the data analysis routine and gives an almost perfect estimation of D, where precision depends on the resolution of the ρ_G parameter.

2. Estimation of A

The next step is the estimation of $\tilde{A}$, which is obtained from the $\bar{f}$:

$$\bar{f} \approx \tfrac{1}{2}(\bar{\rho_D} + \bar{\rho_G})P. \tag{15}$$

It is similar to Eq. (12), but now with ρ_G instead of ρ_D. This approximation gives a better estimation and simpler analytical form. The reason is because in Eq. (12) the quality Q of the signal was maximized, while in Eq. (15) the actual network signal Δc is maximized. Even though they are close to each other, they are not always perfectly aligned. Putting $\bar{\rho_G}$ to the Eqs. (2) gives

$$\bar{\rho_D} = \bar{\rho_G} + A(1 - \bar{\rho_G}). \tag{16}$$

Now plugging Eq. (16) into Eq. (15) and combining with the restriction that A cannot exceed 1 yields

$$\tilde{A} = \min\left\{2\frac{\frac{\bar{f}}{P} - \bar{\rho_G}}{1 - \bar{\rho_G}}, 1\right\}. \tag{17}$$

It should be noted that this is an approximation. However, it works reasonably well, as can be seen in Figs. 5 and 6, where also the estimate for p is depicted.

3. Estimation of p

Having obtained $\tilde{D}$ and $\tilde{A}$, the last step is the estimation of p as a $\tilde{p}$ from $\Delta c_{\max}$. Equations (7) yield

$$\tilde{p} = \max\left\{\frac{1 - \sqrt{(1 - \tilde{c}_{D^*}^f)(1 - d)^{-\tilde{H}^* - \tilde{D}^* + 3\tilde{R}_D^*}}}{\tilde{R}_D^*}, 0\right\}, \tag{18}$$

where also from Eq. (7) one has

$$\tilde{c}_{D^*}^f = \frac{1}{\tilde{D}^*}\left[c_P^{*f}(\tilde{H}^* + \tilde{D}^*) - \tilde{H}^*\, \tilde{c}_{H^*}^f\right] \tag{19}$$

$$\tag{20}$$

and again

$$\tilde{c}_{H^*}^f = 1 - (1 - d)^{\tilde{H}^* + \tilde{D}^* - 1}, \tag{21}$$

where

$$\tilde{R}_D = \sum_{i=\bar{f}+1}^{P} \binom{P}{i} \bar{\rho_D}^i (1 - \bar{\rho_D})^{P-i}, \tag{22}$$

$$\tilde{R}_H = \sum_{i=\bar{f}+1}^{P} \binom{P}{i} \bar{\rho_H}^i (1 - \bar{\rho_H})^{P-i}, \tag{23}$$

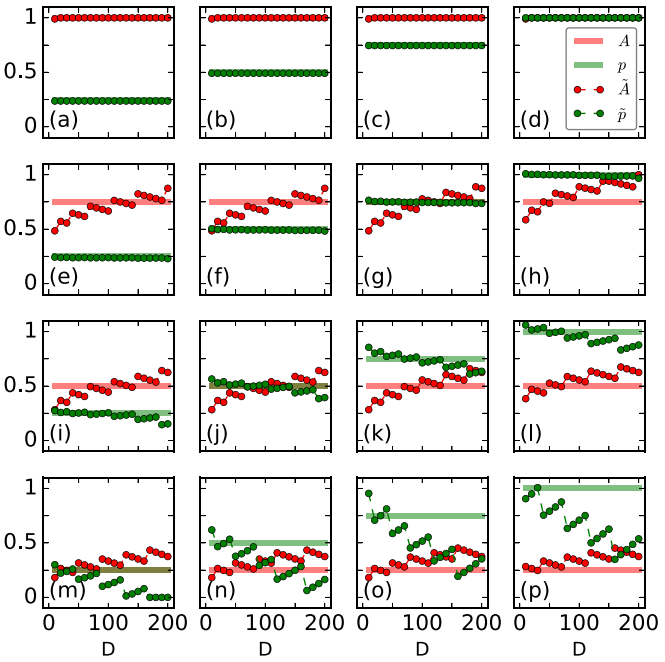

FIG. 5. Estimation of parameters A and p as a function of D, where $G = 1000$, $P = 20$, and $d = 0.006$. Solid lines indicate original parameters and dashed lines with dots are estimates. Different columns stand for subsequent values of p: [(a), (e), (i), and (m)] 0.25, [(b), (f), (j) and (n)] 0.5, [(c), (g), (k), and (o)] 0.75, and [(d), (h), (l), and (p)] 1. Rows denote different values of A: [(a)–(d)] 1, [(e)–(h)] 0.75, [(i)–(l)] 0.5, and [(m)–(p)] 0.25.

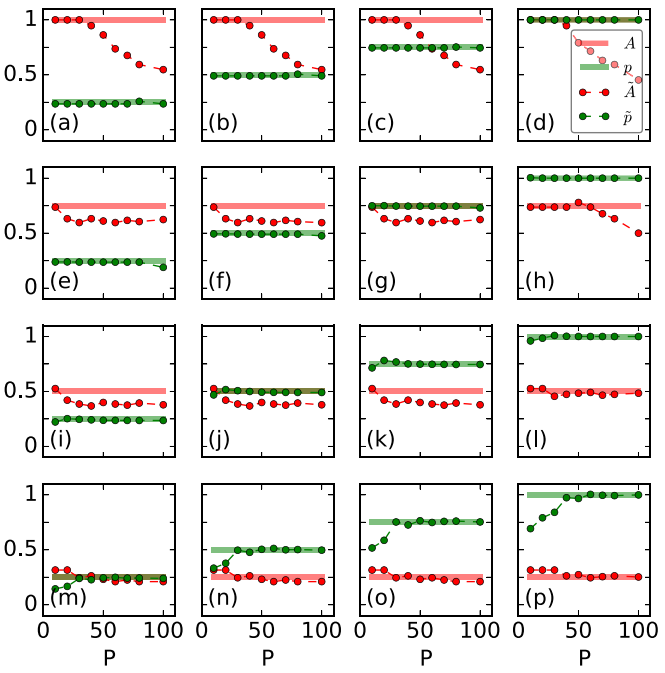

FIG. 6. Estimation of parameters A and p as a function of P, where $G = 1000$, $D = 50$, and $d = 0.006$. Solid lines indicate original parameters and dashed lines with dots are estimates. Different columns stand for subsequent values of p: [(a), (e), (i), and (m)] 0.25, [(b), (f), (j), and (n)] 0.5, [(c), (g), (k), and (o)] 0.75, and [(d), (h), (l), and (p)] 1. Rows denote different values of A: [(a)–(d)] 1, [(e)–(h)] 0.75, [(i)–(l)] 0.5, and [(m)–(p)] 0.25.

and

$$\tilde{D}^* = \tilde{R}_D\,\tilde{D}, \tag{24}$$

$$\tilde{H}^* = \tilde{R}_H\,(G - \tilde{D}). \tag{25}$$

In this way, all the hidden parameters can finally be estimated from the data. Having these values also the quality Q can be evaluated. See Figs. S5–S8 in Ref. [51] for more details regarding above estimation.

V. DISCUSSION

The challenge of systems medicine is to employ modeling and data analysis tools from systems biology for the interpretation of medical data. Statistical physics has routinely contributed data analysis techniques for gene expression data (see, e.g., Refs. [40,42,44])

In contrast to the many successful applications of mathematical and computational methods in Systems Biology, data sets in systems medicine are often very small and highly heterogeneous with few well documented cases of evidence converging towards universal principles. We can think of the "signal strength" in such data as the maximal amount of information an ideal analysis method would be capable of extracting from a given data set (e.g., discriminating the disease state from healthy controls). In typical medical data sets, we can expect this signal strength to be rather small. At the same time, we see a tremendous diversity in computational methods, often to the point where data analysis tools are tailor-made for a specific data set and only applied to this data set alone [52]. In particular, there are few comparative studies, which could allow a medical researcher an informed choice in this diversity of computational methods. Notable exceptions are [53,54].

We employ well established tools, like a random walks and networks, to formulate a simple yet flexible model of gene expression profiles for clinical cohorts typically encountered in medical research. This abstract, generative model can be used to create test instances of data to try out and calibrate existing analysis tools. This model can also facilitate the design of new data analysis methods. With the filtering-based connectivity assessment described here, we give an example of such a new method. Here we ask, what the cohort-level, collective analysis (in contrast to the sequential analysis of individual patients within a disease cohort) can offer. We exploit here that the disease subgroup can be expected to be more uniform than the controls, because they share a phenotypic feature. Qualitatively, in our generative model, this leads to an enrichment of disease-associated genes in the filtered set.

Here we show that, surprisingly, key parameters of the data, like the number of disease-associated genes, can be reconstructed from the data, by using a comparatively small number of quantitative parameters: (1) the "recall" of disease-associated genes in the expression data (i.e., the offset in likelihood of being differentially expressed in the disease state, (2) the amount of clustering of differentially expressed genes in the given biological network. We can furthermore clearly delineate the region of signal strength, where the

network facilitates the analysis, e.g., the identification of the disease-associated genes.

The filtering-based data analysis method presented here can be seen in competition with a multitude of related methods from the corresponding literature in bioinformatics, systems biology, and systems medicine [55]. An example is the "key pathway miner" method [56–58], which shares similarities with the filtering approach described above.

The data generation part of our study can be used to test and compare the performance of these different analysis methods in a quantitative fashion under variation of the main parameters (like cohort size, network clustering, size of the disease gene set, etc.).

The generative model is flexible enough to simulate many different types of diseases and to draw statistical conclusions and, as a consequence, some insight into the disease. Furthermore, analytical solutions for the main quantities of the model are also provided, which facilitates our understanding of the parameter interdependencies. The most surprising fact is that results are mostly independent of the network structure, but rather depends solely on the network density. This is true especially for relatively sparse networks, which is the case of most biological networks.

In practice, the model can be used to create cohort data and explore the various parameter dependencies via a web-based PYTHON application, see Ref. [59]. The method of analysis of the simulated data presented here works remarkably well and reveals estimates of the disease gene set with high accuracy even for small cohorts and weak signals. We believe that the method is capable of contributing to new discoveries in biology and medicine.

An important resource of disease-associated genes currently are genome-wide association studies (GWAS), as discussed in [60,61]. Ultimately, disease associations of genes derived from such approaches based on population genetics need to converge with the more functional associations obtained from transcriptome profiles of patients (which is the type of data set discussed here). Biological networks can be expected to play an important role in uniting these two categories of disease associations of genes [11]. Our generative model is intended as a computational resource along this way.

ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (Grant No. 01ZX1306D).

APPENDIX A: AVERAGE CONNECTIVITY IN ERDÖS-RÉNYI GRAPHS

1. Connectivity *c*

Connectivity c plays central role in our data analysis procedure. The definition of connectivity we employ here is simply the fraction of nonisolated nodes in the given graph $\Gamma(V, E)$:

$$c = \frac{1}{N}\sum_{v\in V}\{1|k(v) > 0\}, \tag{A1}$$

where N is the number of vertices and $k(v)$ is the degree of the vertex v. In our analysis, the connectivity of random graphs is used to quantify a deviation of the network signal from randomness.

2. 1D chain of events

The "teleportation random walk" (TRW) algorithm is well conceptualised by a linear chain of events, with only two possibilities for each step in this chain; step A: with $P(A) = p$ move to the neighboring vertex and color it, and step B: with $P(B) = 1 - p$ jump (teleport) to the randomly selected vertex and color it. Hence a chain of events may look like:

$$AAABABBAABB$$

having graphical representation:

$$\bullet - \bullet - \bullet - \bullet \cdots \bullet - \bullet \cdots \bullet \cdots \bullet - \bullet - \bullet \cdots \bullet \cdots \bullet,$$

where the link to the neighbor $-$ is A, and random jump (lack of the link) $\cdots$ is B and a $\bullet$ is a vertex. Probability of such a chain of events is

$$p\,p\,p\,(1-p)\,p\,(1-p)\,(1-p)\,p\,p\,(1-p)\,(1-p) = p^6(1-p)^5.$$

Calculation of c in a 1D connected chain is straightforward. It is dependent only on the probability of jump p, hence,

$$c_1(p) = 1 - (1-p)^2, \tag{A2}$$

where $(1-p)^2$ is the probability for a given colored vertex to have no colored neighbors in the 1D chain, therefore $1-(1-p)^2$ is a probability of having at least one neighbor, which is sufficient to regard this vertex as connected.

3. Assumptions

The above chain of events is precise on a 1D chain of vertices and is approximately valid for relatively sparse networks with low clustering coefficients. In such a case, the colored cluster is very well approximated by a 1D chain of colored vertices, forming pathlike structure, embedded in the structure of the network.

The connectivity c created by the TRW depends on the density of the links d in the network, but is rather independent of other aspects of the network structure. This is due to the fact that the TRW algorithm follows the structure of the network itself, effectively canceling structural influence on connectivity c, which quantifies the number of breaks in the path.

A second assumption is related to network size. For $p \neq 1$, teleportation B causes the formation of multiple clusters. In small networks it is very likely that some of them may overlap. This overlap is difficult to incorporate in analytical calculations. In order to reduce effects of clusters overlapping, network size should be much larger than the number of colored nodes $N \gg n$.

4. Background contribution

In case of the real graph, c_1 is a contribution to the c from the 1D chainlike arrangement of colored vertices. In

the graph different from the 1D chain, some colored vertices could be connected via additional edges lying outside of 1D arrangement. It creates possibility for vertex isolated in 1D chain to be connected with some other vertex via the "network background. This affects the connectivity outcome c, and is described by

$$c_2(d,n) = 1-(1-d)^{n-3}, \tag{A3}$$

where n is a number of colored vertices, d is density of edges, and 3 is there because two potential edges were already used by 1D cluster-chain. The final expression, obtained from the combination of factors c_1 and c_2, has the form of

$$\begin{aligned} c \equiv c_c(p,d,n) &= 1-(1-c_1(p))(1-c_2(p)) \\ &= 1-(1-p)^2(1-d)^{n-3}. \end{aligned} \tag{A4}$$

It describes average c_c which is also depicted on Fig. 7, where three different colorings are plotted. The network and the number of colored nodes are the same, but the values of p are different. For $p=0$, colored nodes are spread completely randomly, $p=0.5$ displays moderate clustering of the colored nodes, while for $p=1.0$, there is just one connected cluster of colored nodes. Furthermore, the connectivity c as a function of p is shown.

5. Subgraphs

Previously, c_c was defined as the ratio of colored vertices having at least one colored neighbor and the number of all colored vertices in the graph. Via the concept of subgraphs, we can look at the connectivity from another perspective. In the subgraph consisting of all colored vertices extracted from the bigger original graph $\Gamma_c(V_c,E_c)\subseteq\Gamma(V,E)$, the connectivity c_c is the percentage of nonisolated vertices. Connectivity of the random sample taken from the original graph $\Gamma^*(V^*,E^*)\subseteq\Gamma(V,E)$ is given by

$$c \equiv c^*(d,n) = 1-(1-d)^{n-1}, \tag{A5}$$

where n is a number of vertices in the subgraph, d is density of edges, and * denotes random sampling over a given set. We can now look at the connectivity obtained from randomly selected vertices of the colored subgraph $\Gamma_c^*(V_c^*,E_c^*)\subseteq\Gamma_c(V_c,E_c)$ and another subgraph consisting of nodes randomly selected from the noncolored ones $\Gamma_c'^*(V_c'^*,E_c'^*)\subseteq\Gamma_c'(V_c',E_c')$. Where $\Gamma_c'(V_c',E_c')=\Gamma(V,E)\setminus\Gamma_c(V_c,E_c)$. This unified subgraph is

$$\Gamma_u(V_u,E_u)=\Gamma(V_c'^*\cup V_c^*, E(V_c'^*\cup V_c^*)). \tag{A6}$$

Regarding the subset of randomly selected vertices $V_c'^*$:

$$c_c'^*(d,n) = 1-(1-d)^{n_c'^*+n_c^*-1}, \tag{A7}$$

where n_c^* is a number of randomly selected colored vertices and $n_c'^*$ is a number of vertices randomly selected from the uncolored rest of the graph. Hence $n_c^*+n_c'^*$ is a total number of vertices in the graph Γ_u. For randomly selected subset of colored vertices V_c^*, one has

$$\begin{aligned} c_c^*(p,d,n) &= 1-(1-pR)^2(1-d)^{n_c^*+n_c'^*-3R}, \\ c_c^*(p,d,n) &= 1-\left(1-p\frac{n_c^*}{n_c}\right)^2(1-d)^{n_c^*+n_c'^*-3\frac{n_c^*}{n_c}}, \end{aligned} \tag{A8}$$

where n_c is a number of all colored vertices and n_c^* is a number of randomly selected colored vertices, $n_c'^*$ is the number of vertices randomly selected from the uncolored subset and $R=\frac{n_c^*}{n_c}$ is probability of a colored node to be randomly selected. Final expression for c_u comes from the weighted average of c_c^* and $c_c'^*$:

$$c \equiv c_u = \frac{n_c^*c_c^*+n_c'^*c_c'^*}{n_c^*+n_c'^*}. \tag{A9}$$

The above expression is an approximation for the average c over the ensemble of graphs and subgraphs. This approximation works very well for the most of the graphs used in the systems biology, as they tend to be sparse and to have low clustering coefficients. The value of c is an observable coming directly from the data analysis. It has been shown here that it can be also easily calculated for the TRW model presented in the main part of our investigation.

APPENDIX B: FILTERING AND MAXIMIZATION

This section depicts details of the data analysis procedure. It refers to two crucial parts of it. Expression filtering and connectivity computation is illustrated in Fig. 8. Tuning of analysis parameters to maximize Δc in order to extract the set

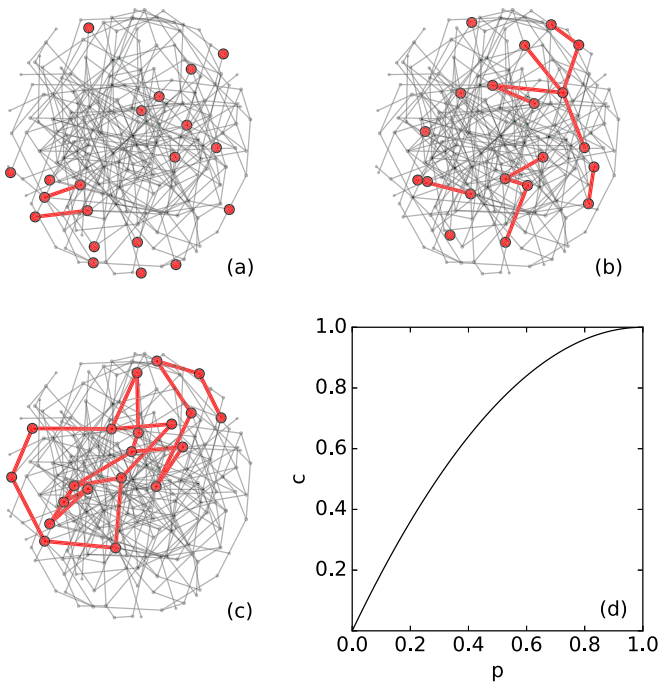

FIG. 7. Example of the clustering (and connectivity) of selected ("colored") nodes for different values of the TRW parameter p: (a) 0, (b) 0.5, and (c) 1.0. (d) depicts the connectivity $c(p)$ as a function of p.

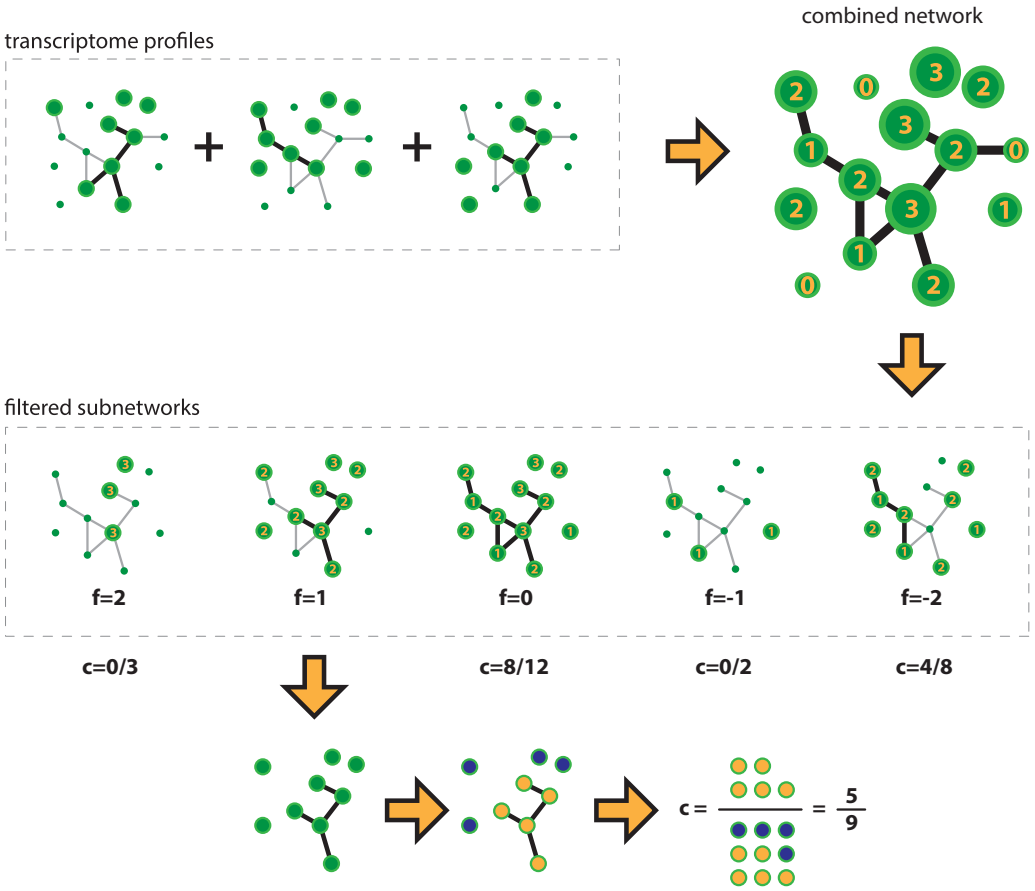

FIG. 8. Illustration of the filtering procedure. First, dichotomized individual transcriptome profiles (obtained by evaluating threshold ρ_G) are summed up to form combined multiset on the network (combined network). Then several filtered subnetworks are created based on the number of occurrences of nodes in the multiset (employing the filtering threshold f). Next, for each subnetwork, the fraction of nonisolated nodes (connectivity c^f) is calculated. This ensemble of connectivity values is subject of further analysis depicted in Fig. 9. For positive values of f, all vertices appearing more than f times are taken, while for negative values all vertices appearing equal or less times than f are selected. $f = 0$ is treated as a positive number as all the vertices appearing more than 0 times survive filtering, which is equivalent to no filtering at all.

of genes with highest match to the disease-gene set is depicted in Fig. 9.

APPENDIX C: QUALITY ASSESSMENT

1. Quality measures

As discussed above, maximizing Δc is a good strategy for finding the optimal values of the parameters ρ_G and f. These optimal values, $\bar{\rho_G}$ and $\bar{f}$, allow us to extract the set of genes possibly closest to the disease gene set V_D. In this Appendix, the assessment of the reconstruction quality of the disease gene set is discussed. The aim of our data analysis method is the extraction of a candidate set of disease genes, V_P^{*f}, possibly closest to the true disease gene set V_D. Ideally, $V_G^{*f} = V_D$. In reality it is rarely achieved. What can be achieved is a maximization of $\frac{D^*}{G^*}$ ratio, which is by definition $\frac{D^*}{G^*} = \frac{D^*}{H^*+D^*}$. Hence, for $\frac{D^*}{H^*+D^*} = 1$, only true disease genes have been extracted such that $V_G^{*f} \subset V_D$.

Filtering has the goal of retaining only highly reliable candidate genes. It can happen that most of the set will be filtered out and as a result only a small part of V_D will remain. This contradicts the aim to retrieve the largest possible number of disease genes. Therefore also $\frac{D^*}{D}$ should be as high as possible.

Summarizing the task is to find a reasonable compromise between two contradictory factors: maximization of the *precision* PPV $= \frac{D^*}{G^*}$ (positive predictive value) and maximization of *sensitivity* TPR $= \frac{D^*}{D}$ (true positive rate).

The method described in the main text provides a heuristic, how to obtain a good solution. Using the quantities introduced above, we can now evaluate the quality of the candidate set of disease genes and its proximity to the optimal set. In order to deal with this task, it is convenient to define single parameter measuring quality. One of the most natural choices is the product of sensitivity and precision:

$$Q = \text{PPV} \times \text{TPR} = \frac{D^*}{G^*}\frac{D^*}{D} = \frac{D^{*2}}{D\,G^*}. \tag{C1}$$

A property of PPV is that PPV $\in \langle 0, 1\rangle$. For PPV $= 0$, there are no disease genes retrieved, $V_P^{*f} \cap V_D = \emptyset$, while for PPV $= 1$ only disease genes are retrieved, $V_P^* \subseteq V_D$ and $V_P^* \cap V_H = \emptyset$.

A property of TPR is that TPR $\in \langle 0, 1\rangle$, where for TPR $=$ 0 also no disease genes are retrieved $V_P^{*f} \cap V_D = \emptyset$, but for

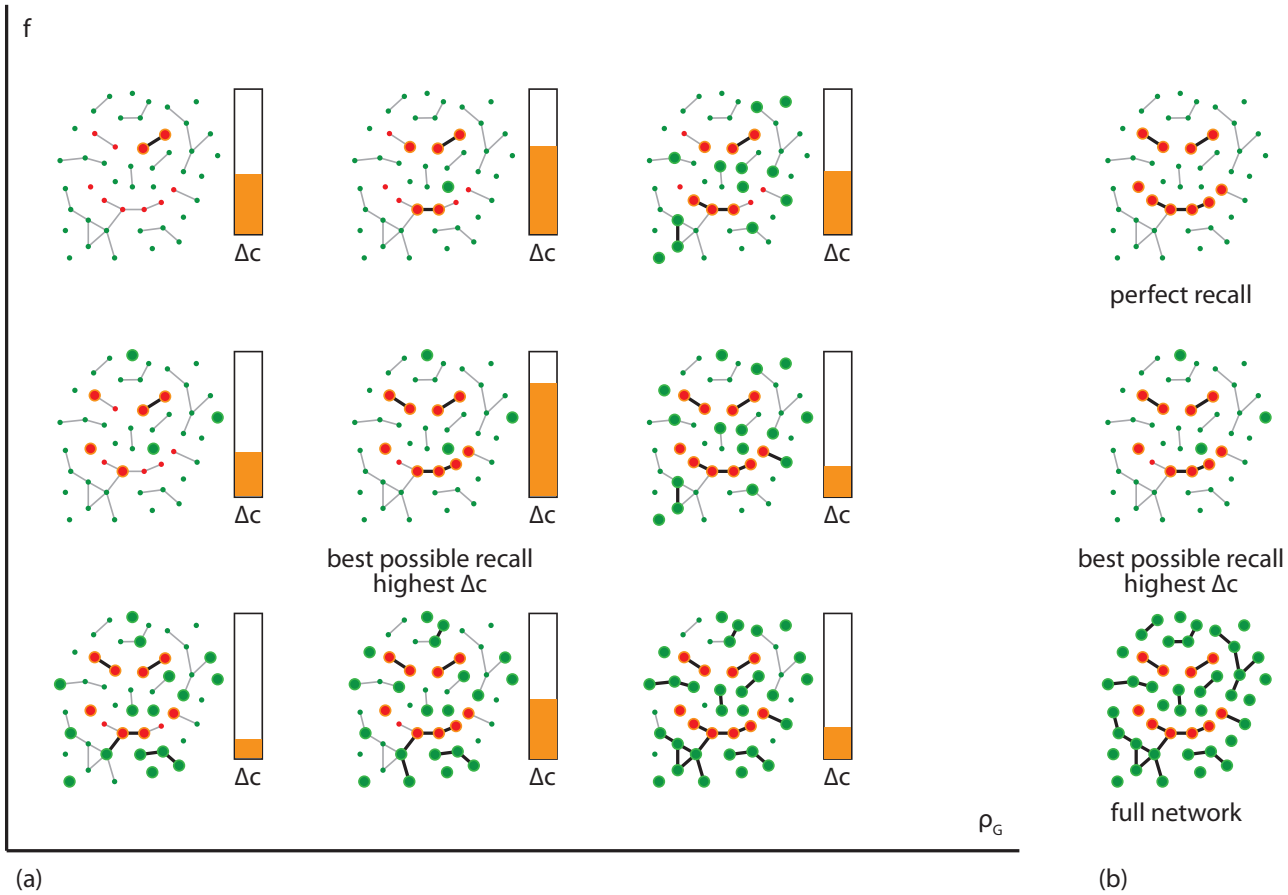

FIG. 9. Illustration of the selection of optimal data analysis parameters. The figure shows the ensemble of $V_P^*(\rho_G, f)$ sets, together with the respective Δc values. Maximal value of Δc most likely gives the best possible recall of the set of disease genes V_D. The procedure is as follows. First, screen over $\rho_G \times f$ space (where $f \geqslant 0$) and generate ensemble of filtered sets $V_P^*(\rho_G, f)$ and $V_C^*(\rho_G, f)$ along with associated connectivity values $c_P^*(\rho_G, f)$ and $c_C^*(\rho_G, f)$. The screen should be relatively broad but within reasonable range, according to a rule of thumb rather than to precise prescription. For each pair of parameters, $\Delta c(\rho_G, f) = c_P^*(\rho_G, f) - c_C^*(\rho_G, f)$ is calculated. Then from all these values $\Delta c_{\max} = \max(\Delta c(\rho_G, f)) = \Delta c(\rho_G', f')$ is selected along with the respective set $V_P^*(\rho_G', f')$, which is likely to yield the best possible recall of the set of disease related genes V_D. (a) depicts ensemble of retrieved sets spanning $\rho_G \times f$ space. (b) is a cheat sheet of the important sets.

TPR $= 1$ all of them are found, $V_P^{*f} \cap V_D = V_D$ (although other genes may have been found in addition, thus $V_P^{*f} \supseteq V_D$).

The product of these two: $Q =$ PPV TPR has this property that $Q \in \langle 0, 1\rangle$. For $Q = 0$, there are no disease genes extracted, $V_P^{*f} \cap V_D = \emptyset$, while for $Q = 1$, the whole set of disease genes and only these genes are retrieved, $V_P^{*f} = V_D$. Therefore $Q = 1$ means the best possible quality has been achieved.

The quantity Q has another crucial property. It correlates reasonably well with Δc (see Figs. 10–12). This is an important observation, as it shows that maximization of Δc is the right strategy for obtaining highest possible quality. It is worth to mention that Q or PPV and TPR cannot be measured from the data directly, but can only be evaluated via approximation of the model parameters described in the main text. See Ref. [51] for more figures depicting relationships between Q and Δc (Figs. S9–S14) and also calculated and estimated parameters (Figs. S5–S8).

2. Quality of the test

The binary classifier employed here, together with the quantities introduced above, allows us to visualize the quality of the disease gene set prediction in terms of a ROC (receiver operating characteristic) curve (see Fig. 13). In general the ROC curve shows the true positive ratio against the false positive ratio. The larger the area under the curve (AUC), the better is the prediction. Figure 13 proves the high performance of the presented prediction scheme for low to moderate numbers of the disease related genes. This curve is independent on p as it does not have any kind of relation to the network, but is a test of classifying power of the statistical part of the presented method.

For low A and high D, the performance of the test is much worse (as it should be), but in some cases can still be assumed to be acceptable. It should be stressed, however, that the situation where $D = 200$ out of 1000 is very unlikely in a real situation, because it would mean that 1/5 of all the genes present in the network are related to the disease.

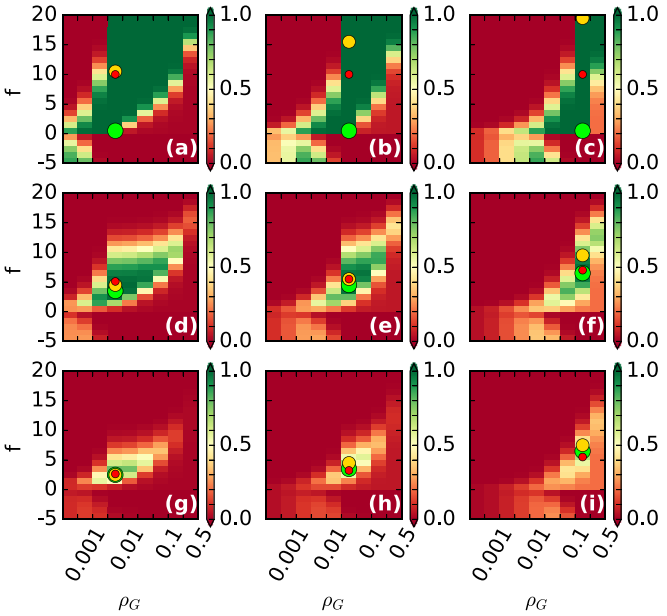

FIG. 10. Quality Q as a function of ρ_G and f. Colored circles indicate position of maximums of Q (green), Δc (yellow), and analytically calculated from the model parameters (12) (red). Different columns stand for subsequent values of D: [(a), (d), and (g)] 10, [(b), (e), and (h)] 50, and [(c), (f), and (i)] 200. Rows denote different values of A: [(a)–(c)] 1, [(d)–(f)] 0.5, and [(g)–(i)] 0.25.

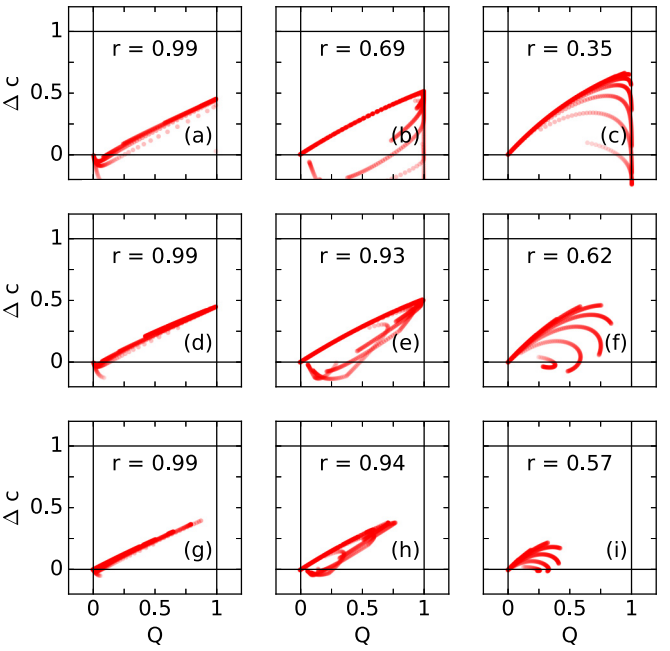

FIG. 12. Figure $\Delta c(Q)$ shows high correlation of both variables, see values of r. Different branches visible in the figure originate from different filtering levels f. On this figure $p = 0.25$. Different columns stand for subsequent values of D: [(a), (d), and (g)] 10, [(b), (e), and (h)] 50 and [(c), (f), and (i)] 200. Rows denote different values of A: [(a)–(c)] 1, [(d)–(f)] 0.5, and [(g)–(i)] 0.25.

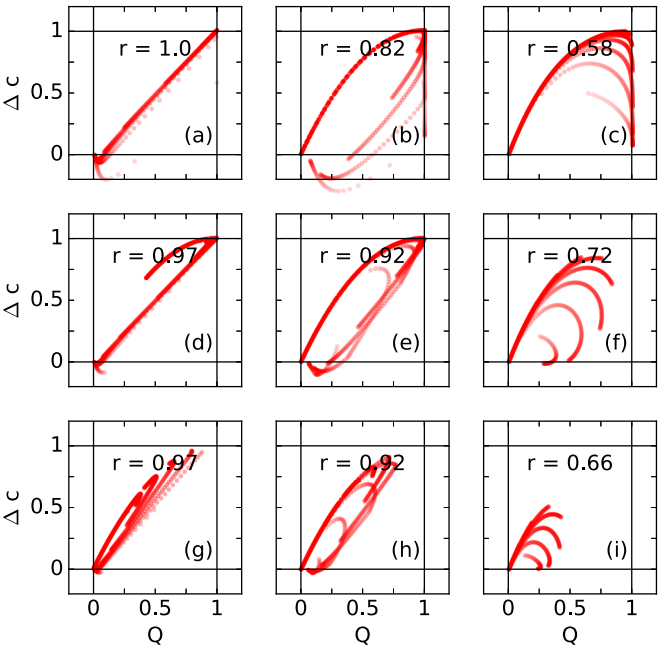

FIG. 11. Figure $\Delta c(Q)$ shows high correlation of both variables, see values of r. Different branches visible in the figure originate from different filtering levels f. On this figure $p = 1.0$. Different columns stand for subsequent values of D: [(a), (d), and (g)] 10, [(b), (e), and (h)] 50 and [(c), (f), (i)] 200. Rows denote different values of A: [(a)–(c)] 1, [(d)–(f)] 0.5, and [(g)–(i)] 0.25.

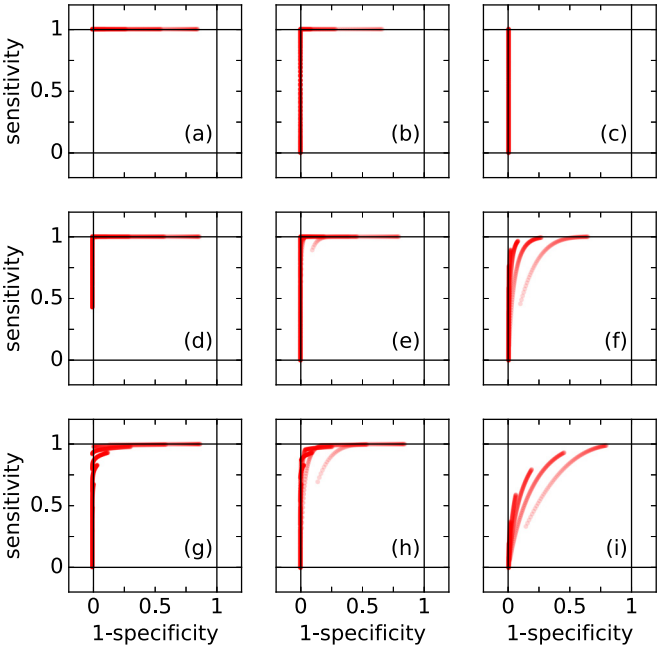

FIG. 13. Receiving operator characteristic curve for presented statistical test. Each branch of the plot is a function obtained by sweeping along ρ_G, while different branches are for different values of f. Different columns stand for subsequent values of D: [(a), (d), and (g)] 10, [(b), (e), and (h)] 50, and [(c), (f), and (i)] 200. Rows denote different values of A: [(a)–(c)] 1, [(d)–(f)] 0.5, and [(g)–(i)] 0.25.

One may assume that the best value of parameters ρ_G and f can be extracted solely based on the ROC curve (Fig. 13). This is false, because the optimal values of the parameters change with other parameters of the model about which prior knowledge is inaccessible. Therefore the network is essential in identifying the optimal prediction.

[1] S. A. Becker and B. O. Palsson, PLoS Comput. Biol. **4**, e1000082 (2008).
[2] N. Sonnenschein, M. Geertz, G. Muskhelishvili, and M.-T. Hütt, BMC Syst. Biol. **5**, 40 (2011).
[3] N. Sonnenschein, J. F. G. Dzib, A. Lesne, S. Eilebrecht, S. Boulkroun, M.-C. Zennaro, A. Benecke, and M.-T. Hütt, BMC Syst. Biol. **6**, 41 (2012).
[4] C. Knecht, C. Fretter, P. Rosenstiel, M. Krawczak, and M.-T. Hütt, Sci. Rep. **6**, 32584 (2016).
[5] A. Oulas, G. Minadakis, M. Zachariou, K. Sokratous, M. M. Bourdakou, and G. M. Spyrou, Briefings Bioinf. **20**, 806 (2017).
[6] R. Häsler, R. Sheibani-Tezerji, A. Sinha, M. Barann, A. Rehman, D. Esser, K. Aden, C. Knecht, B. Brandt, S. Nikolaus *et al.*, Gut **66**, 2087 (2017).
[7] K. Schlicht, P. Nyczka, A. Caliebe, S. Freitag-Wolf, A. Claringbould, L. Franke, U. Võsa, S. L. Kardia, J. A. Smith, W. Zhao *et al.*, Human Genetics **138**, 375 (2019).
[8] T. Ideker and N. J. Krogan, Mol. Syst. Biol. **8**, 565 (2012).
[9] M.-T. Hütt, Br. J. Clin. Pharmacol. **77**, 597 (2014).
[10] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, Nat. Rev. Genet. **18**, 551 (2017).
[11] J. K. Huang, D. E. Carlin, M. K. Yu, W. Zhang, J. F. Kreisberg, P. Tamayo, and T. Ideker, Cell Systems **6**, 484 (2018).
[12] B. Wang, A. Pourshafeie, M. Zitnik, J. Zhu, C. D. Bustamante, S. Batzoglou, and J. Leskovec, Nat. Commun. **9**, 3108 (2018).
[13] H. Kitano, Science **295**, 1662 (2002).
[14] H. Kitano, Nature (London) **420**, 206 (2002).
[15] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman and Hall/CRC, 2006).
[16] K. Sneppen, *Models of Life* (Cambridge University Press, 2014).
[17] N. E. Radde and M.-T. Hütt, EPJ Nonlinear Biomed. Phys. **4**, 7 (2016).
[18] R. Cortini, M. Barbi, B. R. Caré, C. Lavelle, A. Lesne, J. Mozziconacci, and J.-M. Victor, Rev. Mod. Phys. **88**, 025002 (2016).
[19] Y.-Y. Liu and A.-L. Barabási, Rev. Mod. Phys. **88**, 035006 (2016).
[20] C. Villarreal, P. Padilla-Longoria, and E. R. Alvarez-Buylla, Phys. Rev. Lett. **109**, 118102 (2012).
[21] K. Kaneko, C. Furusawa, and T. Yomo, Phys. Rev. X **5**, 011014 (2015).
[22] T. E. Ouldridge, C. C. Govern, and P. R. ten Wolde, Phys. Rev. X **7**, 021004 (2017).
[23] T. P. Peixoto, Phys. Rev. X **8**, 041011 (2018).
[24] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[25] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[26] U. Alon, Nat. Rev. Genet. **8**, 450 (2007).
[27] F. A. Rodrigues, T. K. D. Peron, P. Ji, and J. Kurths, Phys. Rep. **610**, 1 (2016).
[28] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[29] D. Stauffer, J. Stat. Phys. **151**, 9 (2013).
[30] P. Nyczka and K. Sznajd-Weron, J. Stat. Phys. **151**, 174 (2013).
[31] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, Nat. Rev. Genet. **12**, 56 (2011).
[32] A. Suratanee and K. Plaimas, Bioinf. Biol. Insights **11**, 1 (2017).
[33] S. Patkar, A. Magen, R. Sharan, and S. Hannenhalli, PLoS Comput. Biol. **13**, e1005793 (2017).
[34] S. Lee, C. Zhang, Z. Liu, M. Klevstig, B. Mukhopadhyay, M. Bergentall, R. Cinar, M. Ståhlman, N. Sikanic, J. K. Park *et al.*, Mol. Syst. Biol. **13**, 938 (2017).
[35] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, Science **347**, 1257601 (2015).
[36] I. Thiele, N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe *et al.*, Nat. Biotechnol. **31**, 419 (2013).
[37] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander *et al.*, Nature (London) **489**, 91 (2012).
[38] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork *et al.*, Nucleic Acids Res. **47**, D607 (2018).
[39] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam *et al.*, Nucleic Acids Res. **47**, D529 (2018).
[40] S. Jalan, N. Solymosi, G. Vattay, and B. Li, Phys. Rev. E **81**, 046118 (2010).
[41] N. Riedel and J. Berg, Phys. Rev. E **87**, 042715 (2013).
[42] Y. Sharma and P. S. Dutta, Phys. Rev. E **96**, 022409 (2017).
[43] M. Timme, Phys. Rev. Lett. **98**, 224101 (2007).
[44] J. Berg, Phys. Rev. Lett. **100**, 188101 (2008).
[45] M. Rosvall and C. T. Bergstrom, Proc. Natl. Acad. Sci. USA **105**, 1118 (2008).
[46] D. K. Slonim, Nat. Genet. **32**, 502 (2002).
[47] J. Quackenbush, Nat. Genet. **32**, 496 (2002).
[48] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, Nat. Rev. Genet. **7**, 55 (2006).
[49] J. J. Goeman and P. Bühlmann, Bioinf. **23**, 980 (2007).
[50] http://sysbio.jacobs-university.de/website/cohortizer
[51] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.2.033130 for further examples and statistical tests.
[52] M. Cassman, Nature (London) **438**, 1079 (2005).
[53] D.-Y. Cho, Y.-A. Kim, and T. M. Przytycka, PLoS Comput. Biol. **8**, e1002820 (2012).
[54] Y. Lichtblau, K. Zimmermann, B. Haldemann, D. Lenze, M. Hummel, and U. Leser, Briefings Bioinf. **18**, 837 (2016).
[55] C. Soneson and M. Delorenzi, BMC Bioinf. **14**, 91 (2013).

[56] N. Alcaraz, H. Kücük, J. Weile, A. Wipat, and J. Baumbach, Internet Math. **7**, 299 (2011).

[57] N. Alcaraz, J. Pauling, R. Batra, E. Barbosa, A. Junge, A. G. Christensen, V. Azevedo, H. J. Ditzel, and J. Baumbach, BMC Sys. Biol. **8**, 99 (2014).

[58] N. Alcaraz, T. Friedrich, T. Kötzing, A. Krohmer, J. Müller, J. Pauling, and J. Baumbach, Integrative Biology **4**, 756 (2012).

[59] The web-based PYTHON application can be accessed at http://sysbio.jacobs-university.de/website/cohortizer.

[60] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, Am. J. Hum. Genet. **101**, 5 (2017).

[61] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales *et al.*, Nucleic Acids Res. **45**, D896 (2016).