Wieczorek *et al. BMC Systems Biology* (2015) 9:24
DOI 10.1186/s12918-015-0170-2

BMC Systems Biology

METHODOLOGY ARTICLE Open Access

Uncovering distinct protein-network topologies in heterogeneous cell populations

Jakob Wieczorek[1†], Rahuman S Malik-Sheriff[2,3,4†], Yessica Fermin[1], Hernán E Grecco[2,5*], Eli Zamir[2*] and Katja Ickstadt[1*]

Abstract

Background: Cell biology research is fundamentally limited by the number of intracellular components, particularly proteins, that can be co-measured in the same cell. Therefore, cell-to-cell heterogeneity in unmeasured proteins can lead to completely different observed relations between the same measured proteins. Attempts to infer such relations in a heterogeneous cell population can yield uninformative average relations if only one underlying biochemical network is assumed. To address this, we developed a method that recursively couples an iterative unmixing process with a Bayesian analysis of each unmixed subpopulation.

Results: Our approach enables to identify the number of distinct cell subpopulations, unmix their corresponding observations and resolve the network structure of each subpopulation. Using simulations of the MAPK pathway upon EGF and NGF stimulations we assess the performance of the method. We demonstrate that the presented method can identify better than clustering approaches the number of subpopulations within a mixture of observations, thus resolving correctly the statistical relations between the proteins.

Conclusions: Coupling the unmixing of multiplexed observations with the inference of statistical relations between the measured parameters is essential for the success of both of these processes. Here we present a conceptual and algorithmic solution to achieve such coupling and hence to analyze data obtained from a natural mixture of cell populations. As the technologies and necessity for multiplexed measurements are rising in the systems biology era, this work addresses an important current challenge in the analysis of the derived data.

Keywords: Bayesian analysis, Cluster analysis, Intercellular variability, Network analysis, Protein networks, Reverse engineering, Unmixing

Background

In order to understand how a protein network gives rise to a cellular function it is essential to quantify the states of the involved proteins and their causal relations. However, it is actually not possible to strictly define out of the proteome the subset of all proteins which are involved in a certain cellular process since these will always have interactions with proteins not included in this subset. In spite of major advances in proteomic [1, 2] and cytometric [3–7] methods, quantification of the levels and post-translational modifications of all proteins of the proteome in the same cell is still beyond reach. Therefore, we fundamentally cannot observe the whole system at once (i.e in the same cell) but only a small part of it (Fig. 1a) [8]. This limit, by itself, could have been overcome by looking at different parts of the system in different cells and building a model of the whole system step by step. However, such a strategy is fundamentally hampered by natural cell-to-cell variability which makes the integration of information highly challenging. Several studies have addressed the challenge of network reconstruction in the presence of intrinsic and extrinsic noise [9] around one prototypic network structure [10–12]. However, in many physiological cases the cell-to-cell variance is not only due to noise around one cellular state but also due to subpopulations which are in qualitatively distinct types of

*Correspondence: grecco@mpi-dortmund.mpg.de; eli.zamir@mpi-dortmund.mpg.de; ickstadt@statistik.uni-dortmund.de
† Equal contributors
[2] Department of Systemic Cell Biology, Max-Planck Institute of Molecular Physiology, Dortmund, Germany
[1] Faculty of Statistics, TU Dortmund University, Dortmund, Germany
Full list of author information is available at the end of the article

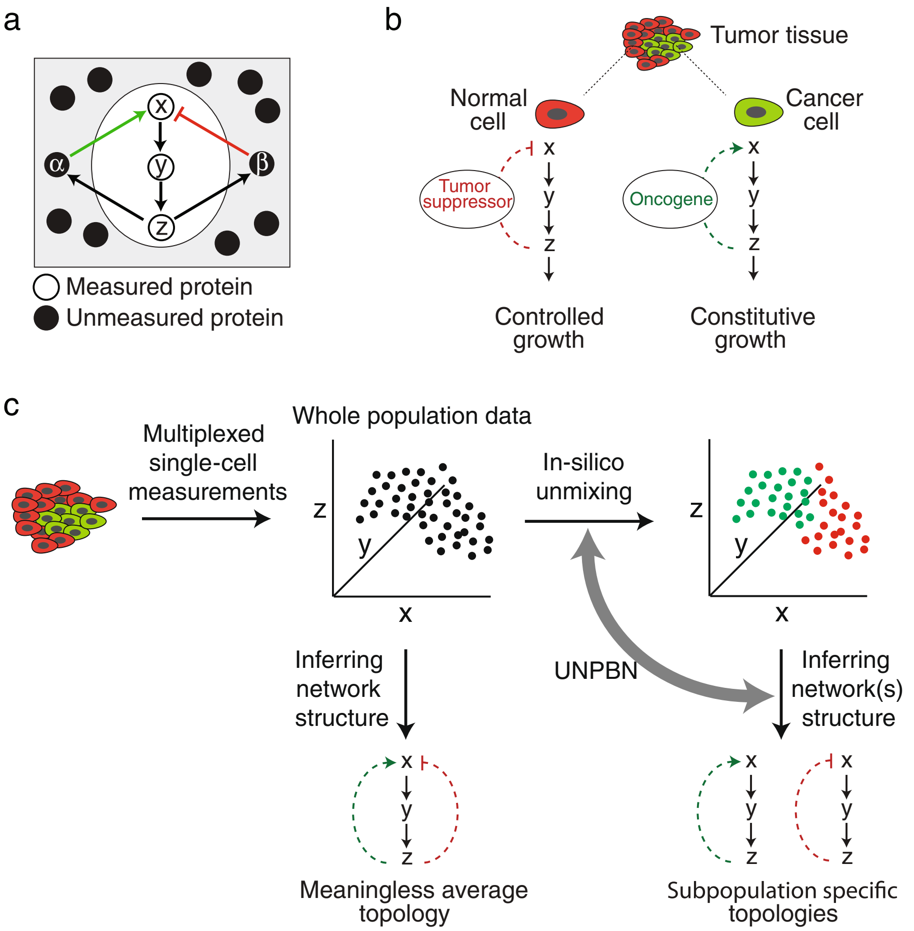

Fig. 1 UNPBN addresses the challenge of studying intracellular protein networks caused by unmeasured proteins and inter-cellular heterogeneity. **a** A biochemical system for which three proteins (*x*, *y*, *z*) are being measured in the same cell while the other proteins are unmeasured. Note that the effects of *z* on *x* are mediated by unmeasured proteins (α and β). **b** Depending on the level and state of these unmeasured proteins, the measured causality between *x* and *z* can differ qualitatively between cells. For example, normal and cancer cells have different activity levels of oncogenes and tumor suppressors which here lead to a negative or a positive causal effect of *z* on *x*, respectively, thereby to a controlled growth or a constitutive growth. **c** Left, multiparametric high-throughput single-cell measurements (e.g., flow-cytometry) of a heterogenous sample of cells (e.g., cancer and normal cells). Middle, attempts to statistically infer a single set of relations (here, causal topology) between the measured proteins fail because there are two distinct subpopulations having two distinct sets of relations. At the same time, it is also impossible to identify the two distinct subpopulations as two distinct proximity-based clusters. Right, UNPBN performs unmixing and inference of statistical relations as one process, thus finds the set of sets-of-relations (network topologies) that explains best the observations

states. Such qualitative variabilities within the same cell population are generated by epigenetic commitment of cells to different fates (e.g., proliferation versus differentiation) as well as by genetic alterations (Fig. 1b) as in cancer [13, 14]. In many cases the distinct cell subpopulations are spatially intermixed and therefore are harvested together and co-measured within the same sample (e.g., by flow-cytometry). In such cases, causal relations and correlations between measured proteins can be qualitatively different in different cells if they are mediated by non-measured proteins which have different states at each subpopulation. Therefore, integration of observations over the cell population toward one model would be invalid and will yield uninformative average relations (Fig. 1c, middle). Ultimately, in order to solve this fundamental problem one should identify the number of qualitatively different subpopulations in the data, thus unmix the cells in-silico and resolve separately for each subpopulation the relations between the measured proteins. A recent work suggested to use a mixture model to unravel subpopulations in biochemical systems based on ordinary differential equations and prior knowledge about the number of subpopulations as well as about kinetic constants underlying the differences between them [15]. In this work we developed a Bayesian method for achieving this goal without such prior knowledge.

To unmix observations of cells from different subpopulations, we are taking advantage of the high-dimensionality of the observations, as typically obtained from cell-based high-content measurements such as flow-cytometry [3, 13, 16, 17], mass-cytometry [4, 5] and toponome imaging [6, 7]. Within each subpopulation, stochastic cell-to-cell variability in protein expression levels gives rise to high-dimensional probability distributions with the same dimensionality as the number of biochemical species (e.g., proteins) measured in each cell. To this extent, network inference approaches, like Gaussian Bayesian networks (GBN) [18–20], to resolve a single statistical model that fits best the data, have been already developed [21–23]. In this work we use as a basis our previously described nonparametric Bayesian network analysis (NPBN, [21]) and expand it to allow for different network structures in a mixture of different cell subpopulations (Fig. 1c, right). In this method, termed hereafter unmixing-via-NPBN (UNPBN), a flexible number of Gaussian Bayesian networks is being fitted to the data and thereby iteratively identifying the number of distinct subpopulations, unmixing the observations and resolving the statistical model for each subpopulation. As a model system to assess and demonstrate our method we simulated the canonical MAPK Raf-Mek-Erk kinases cascade in the context of PC12 cells stimulated by either epidermal growth factor (EGF) or nerve growth factor (NGF) [24]. We show that our method identifies better than common clustering approaches the presence of two subpopulations within a mixture of EGF and NGF stimulated PC12 cells based on the levels of active Raf, Mek and Erk in each cell. This enabled to resolve correctly the statistical relations between Raf, Mek and Erk in each subpopulation.

Methods

Simulation

The EGF and NGF signaling network was simulated based on a previously described model [24]. The SBML format of this model (BIOMD0000000049, www.ebi.ac.uk, retrieval date Oct. 24, 2011) was imported into the Matlab Symbiology platform to simulate the dynamics of the signaling network using *ode15s* (stiff/NDF) solver. To introduce intra-subpopulation cell-to-cell variability (termed herein noise), for each run of the simulation we sampled the values for the total Raf, Mek and Erk levels from a Normal distribution $N(\mu, \sigma)$ around the respective initial concentration for a given set of fractional deviation from the mean ($\sigma = \mu \cdot fd$, where $fd \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$). The values of fd represent here the degree of stochastic variance in the expression levels of Raf, Mek and Erk. Simulations were repeated 175 times with random sampling of total Raf, Mek and Erk levels to generate the data for each cell subpopulation. In each individual simulation repeat, the response of the network to EGF or NGF was simulated for 600 seconds after stimulation and the levels of c-Raf-Ras-GTP (hereafter referred as pRaf, reflecting the consequently activated Raf), ppMek and ppErk (the active, double phosphorylated, forms of Mek and Erk, respectively) were sampled every 1 minute as the observed parameters for the unmixing analysis. Mixtures containing two distinct cell subpopoulations were generated by mixing an equal number, unless indicated otherwise, of simulated observations obtained upon EGF and NGF stimulations. Mixtures containing four distinct cell subpopoulations were generated by altering the parameter in the SBML model corresponding to the catalytic activity (k_{cat}) of Mek (J136) from its wild-type (Mek^{wt}) value (k_{cat} = 0.15 s^{-1}) to a value depicting a mutant Mek (Mek^{mut}) with a lower activity (k_{cat} = 0.015 s^{-1}). Thus, by having two different stimulations and two different levels of Mek activity, observations of four distinct cell subpopulations were generated: EGF-Mek^{wt}, EGF-Mek^{mut}, NGF-Mek^{wt} and NGF-Mek^{mut} (Additional file 1a-d).

UNPBN

Methodologically, UNPBN is based on the nonparametric Bayesian networks (NPBN) approach [21]. It allows to avoid the assumption of underlying Gaussian distributions for the data and to find networks with nonlinear relations between the nodes. The UNPBN method combines a nonparametric mixture model incorporating the Dirichlet process [21, 25] and an allocation sampler [26, 27]. Prior to the description of the UNPBN approach a short introduction of GBNs [28] is provided here, as they are a basis of the presented method. We define the data X, consisting of n observations of a system/network with d species/nodes ($X \in \mathbb{R}^{n \times d}$), such that x_j represents an n-dimensional vector containing the observed concentrations of the j species ($j = 1, \ldots, d$). In the Bayesian networks approach the relations between the nodes in a graph $\mathcal{G}$ are modeled as conditional probability distributions (CPDs) p. If the CPDs for all nodes in $\mathcal{G}$ are given by Normal distributions of the form $x_j | pa_{\mathcal{G}}(x_j) \sim N(\mu_j + \sum_{\mathcal{K}_j} \beta_{j,k}(x_k - \mu_k), \sigma_j^2)$, where $pa_{\mathcal{G}}(x_j)$ denotes the parents of node x_j, $\mathcal{K}_j = \{k | x_k \in pa_{\mathcal{G}}(x_j)\}$, the μ_j and σ_j^2 are the unconditional means and variances of x_j, and $\beta_{j,k}$ are real-valued coefficients determining the influence of x_k on x_j, and, if in addition, $\mathcal{G}$ is a directed acyclic graph (DAG) then the pair $(p, \mathcal{G})$ is called a GBN. The network structure is inferred using Gaussian distributions with a Normal-Wishart prior [20]. The estimation of $\mathcal{G}$ is embedded in a Markov Chain Monte Carlo (MCMC) framework, conducted by maximizing the sampling distribution of the sampled graph

$$L(\mathcal{G}|X) = \prod_{j=1}^{d} \int L\left(\sigma_j^2, \beta_j | X^{(\{j\} \cup \mathcal{K}_j)}\right) p\left(\sigma_j^2, \beta_j\right) d\sigma_j d\beta_j,$$

with $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,j-1})$, $(\beta_2, \ldots, \beta_d) = B$ and $X^{(\mathcal{J})}$ denotes the columns of X with indices in $\mathcal{J}$. The MCMC algorithm uses so called single edge operations [29].

UNPBN generalizes the GBN approach as it is based on flexible nonparametric Bayesian mixture models for networks [21] which in turn combine different GBNs for different subsets of the data. The mixture is taken with respect to all parameters $(\mu, \sigma, B, \mathcal{G})$. The model for the data can be written as $p(x) = \int p(x|\mu, \sigma, B, \mathcal{G}) dP(\mu, \sigma, B, \mathcal{G})$ with μ and σ vectors of the unconditional means μ_j and variances σ_j^2, respectively. The discrete mixing measure P is distributed according to $\mathbb{P}$, a random probability measure, and $p(x|\mu, \sigma, B, \mathcal{G})$ is a multivariate Normal distribution with a conditional independence structure compatible with $\mathcal{G}$. According to the discrete nature of P, support points $\mu_h, \sigma_h, B_h, \mathcal{G}_h$ and probabilities w_h, the mixture can be written as

$$p(x) = \sum_{h=1}^{N} w_h \, p(x|\mu_h, \sigma_h, B_h, \mathcal{G}_h).$$

The prior distribution of the mixing weights w_h is assigned by P and the prior for $\mu_h, \sigma_h, B_h, \mathcal{G}_h$ is given by the base measure P_0 of $\mathbb{P}$ for all h. The N different mixture components h can be interpreted as subpopulations in the data set. Accordingly, here such subpopulations are referred to as components. The assignment of each data point to its corresponding component is described by the allocation vector $l = (l_1, \ldots, l_n)'$ [26].

The network structure $\mathcal{G}$ and the allocation vector l are the main focus of our UNPBN procedure. The remaining parameters μ_h, σ_h and B_h are integrated out and the MCMC algorithm iterates by updating the DAG $\mathcal{G}$, the number of components N and the latent allocation vector l, leading to the posterior distribution

$$p(l, \mathcal{G}, N|X) = \prod_{h=1}^{N} L(\mathcal{G}|X_{(\mathcal{I}_h)}) p_N(m) p(N) p(\mathcal{G}),$$

where $L(\mathcal{G}|X) = \int L(\sigma, B|X) p(\sigma, B) d\sigma dB$ is the marginal sampling distribution for $\mathcal{G}$, $p_N(m)$ is a probability distribution on the space of allocation vectors, $p(N)$ is the distribution of the number of components and $\mathcal{I}_h = \{i \in \{1, \ldots, n\} | l_i = h\}$ and $X_{(\mathcal{I})}$ denotes the rows of X with indices in $\mathcal{I}$.

In our UNPBN analysis a prior is needed for $\theta_h = (\mu_h, \sigma_h, B_h, \mathcal{G}_h)$ and for $w_1, \ldots, w_N$. For $\mathcal{G}_h$ the prior which was used is uniform over the cardinality of the parent sets [30], for σ_h and B_h we employed the Normal-Wishart prior distribution with the identity matrix as the precision matrix and $d + 2$ degrees of freedom. The mean vector of the multivariate Normal distribution (μ_h) was chosen as a vector of zeros. For N we used a Poisson distribution with parameter $\lambda = 1$ and the w_h were obtained from a Dirichlet distribution with parameter vector $(\alpha, \ldots, \alpha)$ with $\alpha = 1$. Further details for the sampling distribution, posterior distribution and the MCMC sampling scheme were discussed in previous publications [21, 26, 31]. The approach is implemented in Matlab (R2009b, The MathWorks Inc., Natick, Massachusetts). The presented results are obtained from MCMC runs with $2.8 \cdot 10^6$ iterations with a thinning of 350 and a burn in of $1.4 \cdot 10^6$ iterations for networks with two subpopulations and from runs with $5 \cdot 10^6$ iterations with a thinning of 500 and a burn in of $2 \cdot 10^6$ iterations for networks with four subpopulations.

Postprocessing of graphs

Although it is possible to use the output from the UNPBN analysis directly, for example to choose the most frequent DAG or allocation vector as a representative, it is preferable to perform an additional postprocessing step that takes into account all MCMC samples and improves the results considerably. The inferred graphs in the iterations of the MCMC simulation are stored in the form of adjacency matrices. Such a matrix A consists of the elements a_{ij} $(i, j = 1, \ldots, d)$, $a_{ij} = 1$ if nodes i and j are conditionally dependent (an arrow leading from node i to node j) and $a_{ij} = 0$ if nodes i and j are conditionally independent (no arrow between them) or if $i = j$. For each pair of nodes the MCMC output of the UNPBN analysis can be summarized by the posterior edge probability $pep_{ij} = \sum_{s=1}^{r} a_{ij}^{s}/r$ where s is the index of the r iteration steps in the MCMC simulation. The resulting *pep* number ranges from 0 (i.e., strong evidence for the absence of a connection) to 1 (i.e., strong evidence for a connection between the corresponding nodes). These *pep* values are used in Fig. 5 for the presentation of the obtained results.

Postprocessing of allocations

For the allocation vector, however, it is not possible to summarize the sampled vectors in the same way as for the edges, because of the so called label switching problem. During the sampling procedure the labels of the components change randomly, so that if two allocation vectors are compared it is not clear if a particular observation has been allocated to a different component or if the label of the component has changed. We employed a method based on maximizing the adjusted *Rand index* that bypasses this obstacle and that combines the allocation vectors of each MCMC iteration into one single vector [32]. This method is implemented in R [33] in the package 'mcclust' and was used in cases where it was necessary to fix the number of components to a particular value (Fig. 4). In cases where the analysis is focused on the unmixing performance of UNPBN (Fig. 3), the sampled allocation vectors are evaluated regarding the homogeneity of the resulting components. For each entry $l_i^{s,h}$ in the allocation vector sampled in iteration s, in each component h, the true component is determined by comparison with the simulation setting. Based on this, componentwise, observations originating from the same true component are considered as allocated correctly, (the indicator function $I\left(l_i^{s,h}\right)$ is set to 1) while the remaining observations in that component are considered as wrongly allocated ($I\left(l_i^{s,h}\right) = 0$). The percentage of correctly allocated observations (*pco*) for a particular UNPBN outcome is derived by

$$pco = \frac{1}{r}\sum_{s=1}^{r}\frac{1}{N^s}\sum_{h=1}^{N^s}\frac{1}{n_h^s}\sum_{i=1}^{n_h^s} I\left(l_i^{s,h}\right)\cdot 100$$

with r considered MCMC iterations, size n_h^s of component h in iteration s and N^s number of components in the allocation vector in iteration s.

Cluster analyses

In order to compare UNPBN with clustering methods, k-means and hierarchical clustering were used in this work. The k-means clustering method finds the partition that divides the data to n clusters (where n is given by the user) such that the sum distances of all observations to the corresponding cluster mean is minimized [34]. The k-means cluster analysis was performed in Matlab, using the function "kmeans" with the distance parameter being set to squared Euclidean distance. To obtain stable results, the clustering was repeated 500 times with randomly chosen different starting points. Hierarchical clustering is an agglomerative procedure which merges in each step the two closest objects, repeatedly till the whole data set is in one single cluster. The hierarchical clustering was performed in Matlab using the functions "pdist", with the distance parameter being set to Euclidean, followed by "linkage", with the method parameter being set to inner squared distance ("ward", [35]).

Silhouette analysis

We used the average silhouette width (ASW) [36], to assess the quality of a given clustering and to compare the results of clusterings with different parameter settings. For a given clustering result, the silhouette value is calculated as

$$sil(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}}\,.$$

For each observation, x_i, $a(x_i)$ is the average dissimilarity between x_i and all other data points within the same cluster, and $b(x_i)$ is the smallest average dissimilarity between x_i and the data points in the remaining clusters, calculated for each cluster separately. Any measure of dissimilarity can be used, but distance measures are the most common. In this work the Euclidean distance was employed. The silhouette value is ranging between -1 and 1. Negative values indicate that a particular observation will fit better in another cluster, so it has been matched wrongly and the quality of the clustering result can be improved. High positive values indicate a good clustering result. The ASW is computed by averaging all $sil(x_i)$ values, thus it provides an overall evaluation of the regarded clustering. While ASW enables to compare clustering performed with the same method with different parameters, ASW values are not comparable between different clustering methods.

Results

Simulation of inter and intra cell-population variabilities

In order to evaluate the performance of the method we simulated the MAPK module in PC12 cells using a previously described model [24]. This model captures the different temporal profiles of Erk activation upon EGF and NGF stimulation, attributing it to the differential activation and dynamics of Ras and Rap (Fig. 2a) [24, 37]. Both stimulations activate via Sos and Ras the upstream kinase, Raf and thereby the whole MAPK cascade. However, each stimulation has a different effect on other proteins which affect the MAPK module and its dynamics. In EGF stimulation, Erk inhibits Sos and thereby forms a negative feedback loop leading to a transient Erk activation which encodes a proliferation signal. In NGF stimulation this negative feedback is overcome by a nested positive feedback loop [38] formed due to the activation of PKCδ which phosphorylates RKIP and thus leads to its release from Raf and thereby enabling Raf activation by Erk. The model used here considers another difference attributed to a sustained activation of another activator of the MAPK cascade, Rap1, by NGF but not by EGF [24, 37]. Thus, NGF leads to a sustained Erk activation, encoding a signal for differentiation. As a source for inter-population variability, we simulated the dynamics of the complete network upon either EGF or NGF stimulation. For the aim of this work, we based our analyses on snapshots of the simulation, and, in turn, analyzed each time point independently. As a source of intra-population variability (hereafter referred to as noise), we added stochastic noise in total protein levels mimicking natural cell-to-cell variance in protein expression (see Methods). However, unlike instrumental noise that only affects the readout, noise in expression levels affects the system itself. Thus, although the introduced noise was generated as Gaussian, its propagation through the system generates asymmetric high-order patterns shaped by the topology of the network (Fig. 2b, c).

In the absence of noise, the levels of phosphorylated (thus activated) Raf, Mek and Erk follow the expected profiles, exhibiting a clear difference between EGF or NGF stimulations (Fig. 2d, red and green solid lines). With intra-population variance, the profiles get broader and

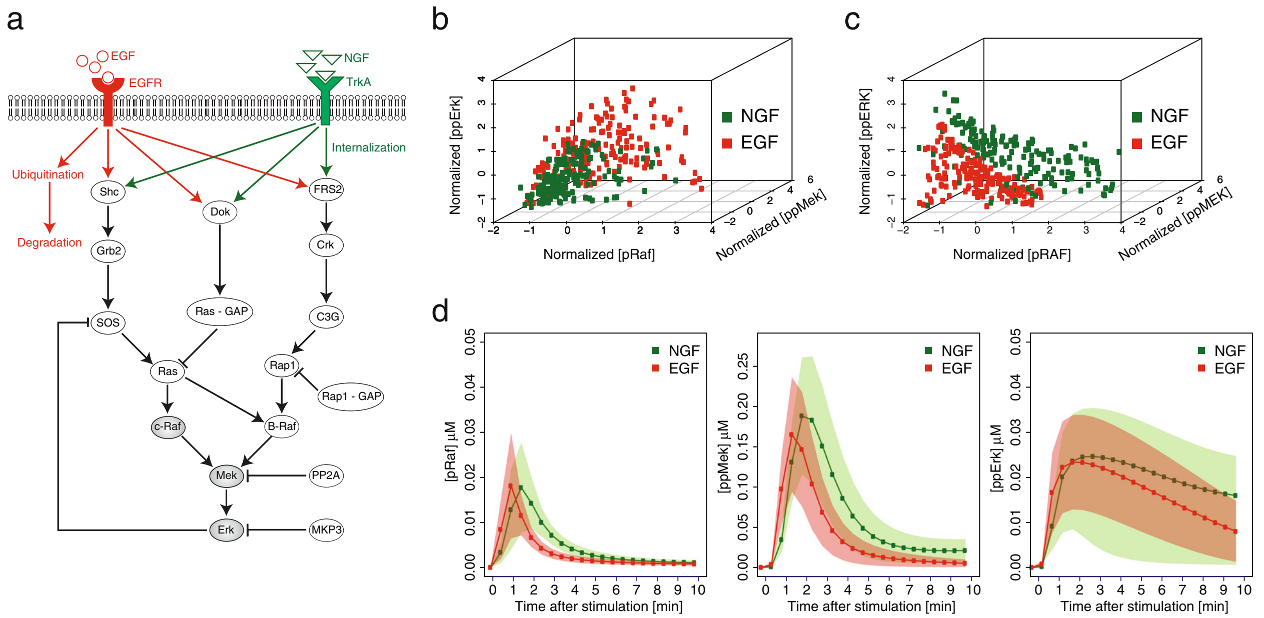

Fig. 2 Simulations of the EGF and NGF signaling network in PC12 cells. **a** The simulated protein network, as previously described [24]. The elements (proteins, interactions and processes) that are unique to either EGF or NGF are colored red and green, respectively. The components that we later consider as the measured components of that system (i.e. Raf, Mek and Erk) are colored grey. **b** A 3-dimensional scatter plot showing the normalized levels of active (i.e. phosphorylated) Raf (pRaf), Mek (ppMek) and Erk (ppErk) versus each other at 2 minutes after NGF (green) or EGF (red) stimulation with noise level of 0.7. **c** The same as **b** but for 9 minutes after stimulation. **d** Time profiles of pRaf, ppMek and ppErk levels as a function of time after NGF (green) or EGF (red) stimulations, without noise (solid line) or with 0.7 noise level (green and red shadows)

overlap between the two stimulations (Fig. 2d, red and green shadows), making it difficult to allocate individual observations to the corresponding stimulation (as for example at 2 minutes after stimulation, Fig. 2b). To impose the fundamental experimental limit of observing only part of the system, for the subsequent analysis we considered an observation to be the triplet formed by the concentrations of phosphorylated species of Raf, Mek and Erk per cell, ignoring all other information. Finally, to generate heterogeneous cell-populations, we mixed observations randomly selected in equal amounts from the EGF and NGF datasets.

UNPBN unmixes observations of distinct subpopulations

We first wanted to test whether UNPBN can classify correctly observations coming from distinct cell subpopulations. We applied UNPBN on mixed cell populations having different levels of noise and counted the observations correctly allocated to the EGF and NGF stimulated subpopulations (Fig. 3a). For noise levels of 0.1, 0.5 and 0.7, around 100 %, 93 % and 85 % of the observations are correctly allocated, respectively (Fig. 3a). To assess the accuracies of UNPBN, we compared them to those achieved by two widely used clustering approaches - hierarchical clustering and k-means clustering. When the noise is low (0.1 and 0.3), the two subpopulations are well separated by all methods (Fig. 3a). As expected, the performance of all three methods is negatively affected when the noise level is increased. However, the UNPBN considerably outperforms the other reference methods for all noise levels above 0.2 (Fig. 3a). If the relative abundance of the two subpopulations is 1:9, all methods classify about equally well for a low noise level (0.2), while for a high noise level (0.7) UNPBN classifies as good as k-means but better than hierarchical clustering (Additional file 2).

We next focused on the high noise level of 0.7 and compared the performance of the methods as a function of time after stimulation (Fig. 3b). Along the different time points the performance of all methods varies, reflecting a changing difficulty to identify the two subpopulations based on the levels of pRaf, ppMek and ppErk. UNPBN constantly outperforms the clustering methods in all the time points and is more robust with its performance level (Fig. 3b). Furthermore, the performances of the two clustering methods along the time points have a similar profile, which differs from the profile of UNPBN (Fig. 3b, time points 3-10 minutes). These results are consistent with the fact that, unlike the clustering approaches, UNPBN uses high-order patterns, rather than merely distances between observations. This additional information is shown here to be indeed valuable for the ability to unmix subpopulations based on high-dimensional observations.

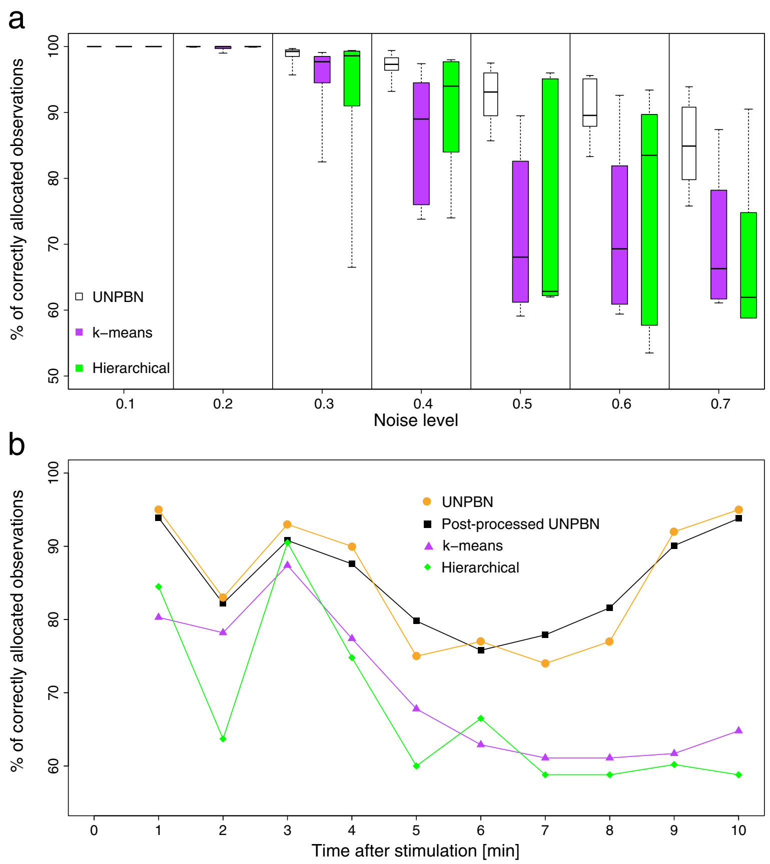

Fig. 3 Unmixing observations of a mixed cell population by UNPBN in comparison to clustering approaches. **a** Mixtures of observations of EGF and NGF stimulated cells with different noise levels were generated as described. Observations were sampled at one-minute intervals for 10 minutes after stimulation. For each noise level and sampled time point, observations were unmixed using UNPBN, k-means clustering (with $k = 2$) and hierarchical clustering (taking the final two clusters). The percentages of correctly allocated observations, averaged over all time points, are indicated by boxplots for the different methods as a function of the noise level (line within the box, the median; box, the 0.25 and 0.75 quartiles; whiskers, the largest and smallest data points which are still within the interval of 1.5 times the interquartile range from the box). **b** Comparison of the unmixing accuracy with noise level 0.7 along the different sampled time points, as achieved by UNPBN, post-processed UNPBN limited to two components, k-means (with $k = 2$) and hierarchical clustering (taking the final two clusters)

UNPBN identifies the number of subpopulations in a mixture

In many cases, when a sample of cells is derived it is unknown a priori how many distinct subpopulations it contains. Therefore, a comprehensive unmixing approach should also be able to identify the number of subpopulations without such a priori knowledge. Indeed, while the clustering approaches were guided to search for two subpopulations, UNPBN was not given this information but found it independently (Fig. 3b). Moreover, the performance of UNPBN does not change significantly if it is forced to identify exactly two subpopulations, indicating the ability of UNPBN to correctly determine by itself the number of distinct subpopulations in a mixture (Fig. 3b).

In order to compare the capability of the different methods to identify the number of subpopulations we used the ASW to determine the quality of the clusters and thereby the number of clusters (i.e. subpopulations) in the data as could be inferred by each method [36]. The ASW of a cluster is a measure of how tightly grouped are the data points in the cluster, such that larger ASW values denote tighter clusters. For a cell population containing two subpopulations (EGF and NGF stimulated cells, Additional file 1a,b) we calculated the ASW as a function of the number of clusters derived by UNPBN (here constraints by postprocessing yield an imposed number of components), k-means and hierarchical clustering (Fig. 4a-c). For UNPBN and k-means clustering, the maximal ASW is found when the number of clusters is 2, the actual number of subpopulations in the data (Fig. 4b,c, red bars). However only in UNPBN there is a significant and robust difference with the other cluster sizes, while with k-means

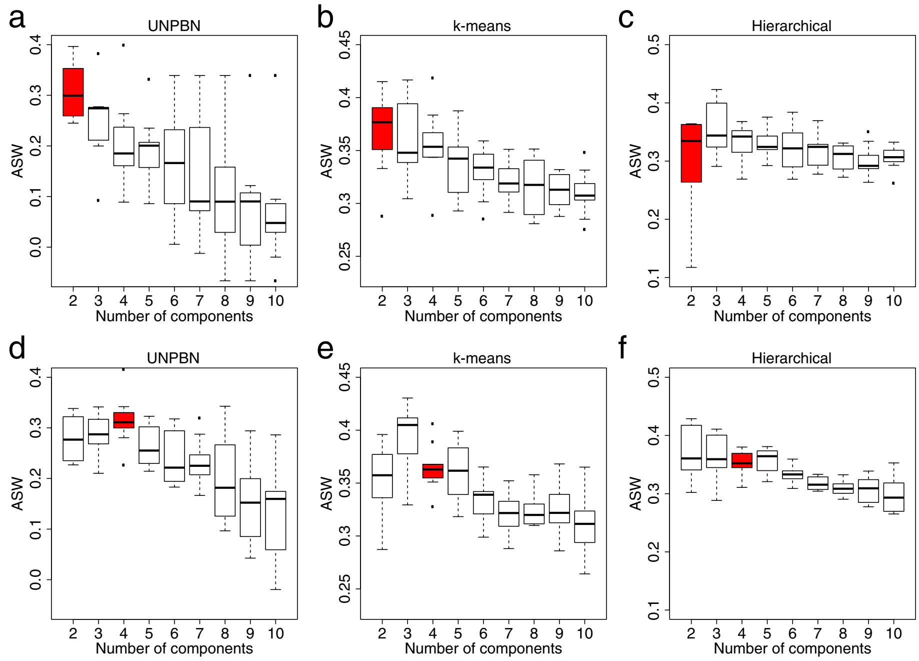

Fig. 4 The success in identifying the correct number of distinct cell subpopoulations (i.e. components) in a mixture by UNBPN in comparison to clustering approaches. **a** A boxplot showing the ASW versus the tested number of components obtained by UNBPN analysis (here constrained in the postprocessing step to the imposed number of components) of a mixture of two subpopulations (EGF and NGF stimulated cells). The boxplot indicates the median (line within the box), the 0.25 and 0.75 quartiles (box), margined by the largest and smallest data points which are still within the interval of 1.5 times the interquartile range from the box (whiskers), and the outliers (dots) obtained from pooled values over all time points with noise level of 0.5. **b** and **c**, the same as in **a** but for ASW obtained following k-means clustering and hierarchical clustering, respectively. **d**, **e** and **f**, the same as the corresponding **a**, **b** and **c**, but for a mixture of 4 subpopulations: EGF-Mek^{wt}, EGF-Mek^{mut}, NGF-Mek^{wt} and NGF-Mek^{mut} (Additional file 1a-d). It should be noted that silhouette widths are incomparable between different clustering approaches. However, silhouette widths are comparable between different parameters of the same clustering approach and thereby indicate the identified number of distinct subpopoulations as the one providing the largest ASW

clustering the ASWs obtained for 2 and 3 clusters are not robustly separable. With hierarchical clustering the performance is further worse since ASWs obtained for 3 and 4 clusters are comparable, or even higher than those obtained for 2 clusters (Fig. 4c). UNPBN successfully identified the number of subpopulations also if their relative abundance was significantly different (1:9, see Additional file 3).

We next tested the performance of the method with a more complex mixture of cells containing four distinct subpopulations. To simulate these subpopulations, the catalytic rate constant, k_{cat}, of Mek in the model was changed, mimicking a wild-type Mek (Mek^{wt}) and a mutant Mek (Mek^{mut}) that phosphorylate Erk at different rates. Thus, together with the two different stimulations, EGF and NGF, four distinct subpopulations were generated, denoted by EGF-Mek^{wt}, EGF-Mek^{mut}, NGF-Mek^{wt} and NGF-Mek^{mut} (Additional file 1a-d). UNPBN correctly identified that the data contains four distinct subpopulations, in contrast to k-means and hierarchical clustering (Fig. 4d-f).

UNPBN uncovers distinct topologies for distinct subpopulations

The causal relations between the components of a system are constant, since the set of biochemical reactions and constants that describe the whole system remains constant. However, in practice, only part of the components of a system can be co-measured and therefore the reaction constants become apparent constants that depend on the unmeasured components. Here we intentionally simulated the fundamental limit of looking on only a small part of a system. Therefore we expected the apparent strength of the causal connection between pRaf, ppMek and ppErk, as reflected by the undirected posterior edge probabilities among them, to change as a function of the stimulation and time. Indeed, when we analyzed separately EGF and NGF stimulated cells we observed different posterior edge probabilities between the two treatments, as well as within each treatment at the different time points (Fig. 5). When analyzing the mixed population with a standard GBN approach (i.e. without the possibility of unmixing) [39], we obtained posterior edge probabilities exhibiting,

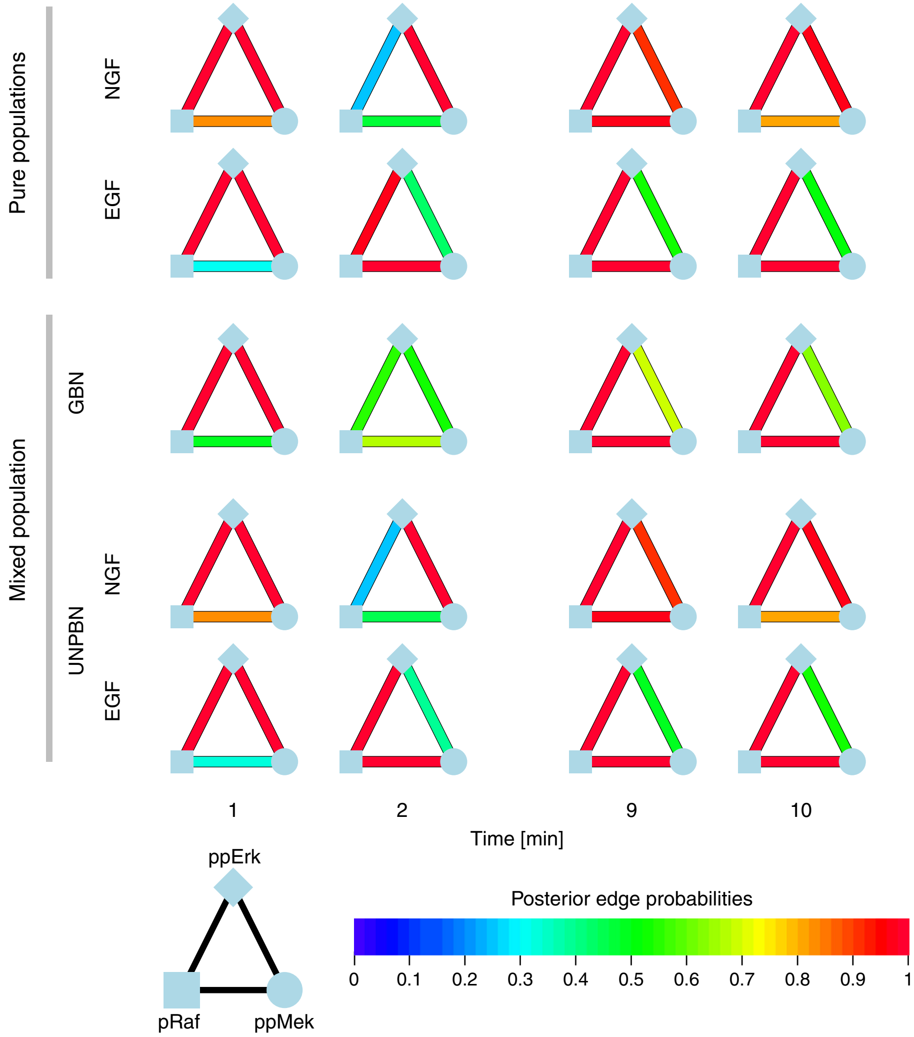

Fig. 5 Recovering correctly the edge probabilities network between pRaf, ppMek and ppErk for each cell subpopulation in a mixture by UNPBN. The triangles show, color encoded, the posterior edge probabilities between pRaf, ppMek and ppErk at different time points after stimulation as derived by UNPBN analysis of pure NGF-stimulated or EGF-stimulated cell populations (top two rows) with noise level of 0.2 and of a mixture of these populations by GBN versus UNPBN (bottom three rows). Note, for the pure subpopulations, the changes in the edge probabilities between the same components at different time points and upon the different stimulations. Analyzing a mixture of NGF-stimulated or EGF-stimulated cells without unmixing yields uninformative average edge probabilities, not representing any of the distinct two subpopulations. Employing UNPBN recovers precisely the pure edge probabilities for each of the two subpopulations. For clarity, we show here the edge probabilities for 4 time points after stimulation that illustrate the need for unmixing

in general, an average behavior of the two subpopulations. Naturally, these average values become meaningless when the two subpopulations exhibit very different posterior edge probabilities (e.g., at 2 minutes, Fig. 5). In contrast, when analyzing the mixed population with UNPBN (i.e., with unmixing), the unmixing step enabled to uncover the true network of posterior edge probabilities for each stimulation and at each time point (Fig. 5). This also demonstrates that the performance of the unmixing process (Fig. 3) was sufficiently good to enable correct inference of protein-protein relations in each subpopulation. Since more than one DAG may represent exactly the same set of conditional independence relationships [40], given static data without perturbations it is more reliable to infer the causal strengths between proteins, regardless of the direction of these causalities. Extending the UNPBN approach to dynamic data, or using perturbation data or adding prior information, will further facilitate the inference of directionality in the causal relations for each cell subpopulation.

Discussion

In the era of systems biology, single-cell measurement techniques are rapidly expanding with respect to the number of cells that can be analyzed and the number of biochemical species that can be co-measured per cell. The approaches to explore these data have focused so far either on identifying different subpopulations of cells based on multiparametric proximities or on inferring the topology of statistical relations between the parameters for the population as a whole. However, the aim to reach each of these two goals in separate has fundamental problems. In one direction, ignoring the heterogeneity between cell subpopulations will lead to inferring a meaningless average topology of statistical relations of the population as a whole. In the other direction, since statistical relations are inferred from the correlation between the measured parameters, the identification of cell subpopulations based on multiparametric proximities inherently conflicts with the capability to resolve the topology of relations within each subpopulation. Furthermore, protein networks with distinct topologies can be at the same state (i.e., to have high multiparametric proximity) and protein networks with the same topology can be at different states (e.g., at different phases along an oscillatory response). Therefore, attempts to identify cell subpopulations based on multiparametric proximities may actually identify different cellular states but not different types of cells. The method presented here pioneers a comprehensive solution to these fundamental problems by performing the identification of cell subpopulations (i.e. unmixing) and the inference of statistical relations between the measured parameters in one joint analysis.

Intentionally, we used snapshot data of a dynamic process (the response of cells to EGF or NGF stimulations) and, respectively, the method we developed does not rely on temporal information nor intends to give a model description of the dynamic itself. Due to that, this method can be applied on the type of single-cell multiparametric measurements currently available such as multicolor flow-cytometry [3, 16], multiplexed mass cytometry [4, 5] and toponome imaging [6, 7]. The classification of the distinct subpopulations in cell populations sampled at different time points along an experiment can hint toward the dynamic behavior of each subpopulation. However, such traceability of subpopulations along the time points depends on how different is their relative abundance within the whole population and on the sampling rate in comparison to the timescale of the biological process. Advances in multicolor live cell imaging in combination with high-throughput automated microscopy gradually facilitate monitoring increasing numbers of parameters in individual live cells over many cells. The data obtained from such measurements will enable not only tracking the dynamics of the measured parameters in each cell subpopulation but also tracking them in individual cells. This kind of temporal information will help to further improve the identification of the distinct cell subpopulations and the inference of statistical relations between measured parameters in each subpoulation. As indicated by this work, it would be important also for the analysis of such live cell measurement data that unmixing and inference of protein-protein relations will be performed as one process.

The importance to recover single-cell phenotypes out of an uninformative average cell population behavior has been established and exemplified in many systems. Notably, in these examples there was only one measured parameter per cell, often the output of the system, and, therefore, the unmixing was straightforward. However, in order to obtain mechanistic insight into how a biochemical system works it is required to examine the protein network itself, and not only its output. For this, multiple parameters should be co-measured per cell to overcome uncorrelated cell-to-cell variability between these parameters (e.g., due to stochastic noise in expression levels as simulated in this work). We demonstrated here that in such a case unmixing cannot be achieved anymore using the proximity between the values of these parameters, while it can be successfully achieved using the high-order relations between them as captured by UNPBN. Importantly, UNPBN can be straightforwardly extended to incorporate prior knowledge about parts of the network in the individual subpopulations.

Conclusions

Our results show that the coupling between unmixing of observations and inference of statistical relations is essential and effective. With respect to the unmixing, our method was capable to identify the number of qualitatively distinct subpopulations considerably better than multiparametric proximity based approaches (hierarchical clustering and k-means clustering). Consequently, the statistical relations in each unmixed subpopulation were also correctly recovered, while, without unmixing, uninformative average relations were inferred. As systems biology and personalized medicine are aiming toward reverse-engineering and re-engineering signaling networks, they are increasingly challenged by the inter-cellular variability and the large size of the relevant biochemical system. The work presented here offers a conceptual solution as well as an applicable statistical method to address this challenge.

Additional files

Additional file 1: The four distinct simulated topologies of the EGF and NGF signaling network used in Fig. 4. (a) NGF-Mek^{wt}: the wild-type network (see Methods) with NGF stimulation. (b) EGF-Mek^{wt}: as in (a) but with EGF stimulation. (c) NGF-Mek^{mut}: the wild-type network with NGF stimulation, beside that here the SBML model parameter corresponding to the k_{cat} of Mek (J136) is altered from its wild-type value ($k_{cat} = 0.15\ s^{-1}$) to a value depicting a mutant Mek with a lower activity ($k_{cat} = 0.015\ s^{-1}$), as indicates the thinner arrow from Mek to Erk. (d) EGF-Mek^{mut}: as in (c) but with EGF stimulation.

Additional file 2: Unmixing observations of cell subpopulations, mixed in a 1:9 ratio, by UNPBN in comparison to clustering approaches. Mixtures of observations of EGF stimulated cells (90 %) and NGF stimulated cells (10 %) were generated with noise levels of 0.2 and 0.7. Observations were sampled at one-minute intervals for 10 minutes after stimulation. For each noise level and sampled time point, observations were unmixed using UNPBN, k-means clustering (with $k = 2$) and hierarchical clustering (taking the final two clusters). The percentages of correctly allocated observations, averaged over all time points, are indicated by boxplots for the different methods for both noise levels (line within the box, the median; box, the 0.25 and 0.75 quartiles; whiskers, the largest and smallest data points which are still within the interval of 1.5 times the interquartile range from the box). (a) The percentages of correctly allocated observations of the EGF-stimulated subpopulation. (b) The percentages of correctly allocated observations of the NGF-stimulated subpopulation.

Additional file 3: The success of UNPBN in identifying the correct number of cell subpopulations (i.e. components) mixed in a 1:9 ratio. Mixtures of observations of EGF stimulated cells (90 %) and NGF stimulated cells (10 %) were generated with noise levels of 0.2 and 0.7. Observations were sampled at one-minute intervals for 10 minutes after stimulation. (a) A boxplot showing the ASW versus the tested number of components obtained by UNBPN analysis (here constrained in the postprocessing step to the imposed number of components). Each boxplot indicates the median (line within the box), the 0.25 and 0.75 quartiles (box), margined by the largest and smallest data points which are still within the interval of 1.5 times the interquartile range from the box (whiskers), and the outliers (dots) obtained from pooled values over all time points with a noise level of 0.2. (b) The same as (a) but with a noise level of 0.7.

Abbreviations

ASW: Average silhouette width; DAG: Directed acyclic graph; EGF: Epidermal growth factor; Erk: Extracellular signal regulated kinase; GBN: Gaussian Bayesian network; MAPK: Mitogen-activated protein kinase; MCMC: Markov Chain Monte Carlo; Mek: MAPK/Erk kinase; NGF: Nerve growth factor; NPBN: Nonparametric Bayesian network analysis; PKC: Protein kinase C; UNPBN: Unmixing via NPBN.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JW, RSM-S, HEG, EZ and KI conceived the project, JW and KI developed the UNPBN algorithm, RSM-S, HEG and EZ generated the simulated data, JW, RSM-S, YF, HEG, EZ and KI analyzed the data, JW, RSM-S, HEG, EZ and KI wrote the paper. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by BMBF grant no. 0315507. YF was supported by a German academic exchange service (DAAD) fellowship.

Author details

[1]Faculty of Statistics, TU Dortmund University, Dortmund, Germany. [2]Department of Systemic Cell Biology, Max-Planck Institute of Molecular Physiology, Dortmund, Germany. [3]Present address: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK. [4]Present address: MRC Clinical Sciences Centre, Imperial College London, London, UK. [5]Present address: Department of Physics, FCEN, University of Buenos Aires and IFIBA, CONICET, Buenos Aires, Argentina.

Received: 15 November 2014 Accepted: 18 May 2015
Published online: 04 June 2015

References

1. Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010;11(6):427–39. doi:10.1038/nrm2900.
2. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol. 2011;7:548. doi:10.1038/msb.2011.81.
3. Schulz KR, Danna EA, Krutzik PO, Nolan GP. Single-cell phospho-protein analysis by flow cytometry. Curr Protoc Immunol. 2012;Chapter 8: 8–17120. doi:10.1002/0471142735.im0817s96.
4. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. Trends Immunol. 2012;33(7):323–2. doi:10.1016/j.it.2012.02.010.
5. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol. 2012;30(9):858–67. doi:10.1038/nbt.2317.
6. Friedenberger M, Bode M, Krusche A, Schubert W. Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome imaging system: sample preparation and measuring procedures. Nat Protoc. 2007;2(9):2285–94. doi:10.1038/nprot.2007.320.
7. Schubert W, Bonnekoh B, Pommer AJ, Philipsen L, Böckelmann R, Malykh Y, et al. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. Nat Biotechnol. 2006;24(10): 1270–8. doi:10.1038/nbt1250.
8. Sachs K, Itani S, Carlisle J, Nolan GP, Pe'er D, Lauffenburger DA. Learning signaling network structures with sparsely distributed data. J Comput Biol. 2009;16(2):201–12. doi:10.1089/cmb.2008.07TT.
9. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002;297(5584):1183–6. doi:10.1126/science.1070919.
10. Hasenauer J, Waldherr S, Doszczak M, Radde N, Scheurich P, Allgöwer F. Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinformatics. 2011;12:125. doi:10.1186/1471-2105-12-125.
11. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J, et al. Moment-based inference predicts bimodality in transient gene expression. Proc Natl Acad Sci U S A. 2012;109(21):8340–5. doi:10.1073/pnas.1200161109.
12. Zechner C, Unger M, Pelet S, Peter M, Koeppl H. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. Nat Methods. 2014;11(2):197–202. doi:10.1038/nmeth.2794.
13. Irish JM, Hovland R, Krutzik PO, Perez OD, Bruserud Ø, Gjertsen BT, et al. Single cell profiling of potentiated phospho-protein networks in cancer cells. Cell. 2004;118(2):217–8. doi:10.1016/j.cell.2004.06.028.
14. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer. 2012;12(5):323–4. doi:10.1038/nrc3261.
15. Hasenauer J, Hasenauer C, Hucho T, Theis FJ. ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics. PLoS Comput Biol. 2014;10(7):1003686. doi:10.1371/journal.pcbi.1003686.
16. Krutzik PO, Crane JM, Clutter MR, Nolan GP. High-content single-cell drug screening with phosphospecific flow cytometry. Nat Chem Biol. 2008;4(2):132–42. doi:10.1038/nchembio.2007.59.
17. Zamir E, Geiger B, Cohen N, Kam Z, Katz BZ. Resolving and classifying haematopoietic bone-marrow cell populations by multi-dimensional analysis of flow-cytometry data. Br J Haematol. 2005;129(3):420–31. doi:10.1111/j.1365-2141.2005.05471.x.
18. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. Inference in Bayesian networks. Nat Biotechnol. 2006;24(1):51–3. doi:10.1038/nbt0106-51.
19. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7(3-4):601-20.
20. Geiger D, Heckerman D. Learning Gaussian networks In: de Mántaras RL, Poole D, editors. Uncertainty in Artificial Intelligence Proceedings of the Tenth Conference. San Francisco, CA: Morgan Kaufmann; 1994. p. 235–43.
21. Ickstadt K, Bornkamp B, Grzegorczyk M, Wieczorek J, Sheriff MR, Grecco HE, et al. Nonparametric Bayesian Networks (with discussion) In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M, editors. Bayesian Statistics 9. Oxford, UK: Oxford University Press; 2011.
22. Pe'er D. Bayesian network analysis of signaling networks: a primer. Sci STKE. 2005;2005(281):4. doi:10.1126/stke.2812005pl4.
23. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523–9. doi:10.1126/science.1105809.
24. Sasagawa S, Ozaki Y.-i, Fujita K, Kuroda S. Prediction and validation of the distinct dynamics of transient and sustained Erk activation. Nat Cell Biol. 2005;7(4):365–73. doi:10.1038/ncb1233.
25. Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Stat. 1973;1:209–30.

26. Nobile A, Fearnside A. Bayesian finite mixtures with an unknown number of components. Stat Comput. 2007;17:147–62.
27. Grzegorczyk M, Husmeier D, Edwards KD, Ghazal P, Millar AJ. Modelling non-stationary gene regulatory processes with a non-homogeneous Bayesian network and the allocation sampler. Bioinformatics. 2008;24(18): 2071–078. doi:10.1093/bioinformatics/btn367.
28. Pearl J. A model of self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society. Irvine, CA: University of California; 1985. p. 329–34.
29. Madigan D, York J. Bayesian graphical models for discrete data. Int Stat Rev. 1995;63:215–32.
30. Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Mach Learn. 2003;50:95–125. doi:10.1023/A:1020249912095.
31. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. J Comput Graphical Stat. 2000;9:249–65.
32. Fritsch A, Ickstadt K. Improved criteria for clustering based on the posterior similarity matrix. Bayesian Anal. 2009;4:367–92. doi:10.1214/09-BA414.
33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
34. MacQueen JB. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Oakland, California: University of California Press; 1967. p. 281–97.
35. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58:236–44.
36. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.
37. Vaudry D, Stork PJS, Lazarovici P, Eiden LE. Signaling pathways for PC-12 cell differentiation: making the right connections. Science. 2002;296(5573):1648–9. doi:10.1126/science.1071552.
38. Santos SDM, Verveer PJ, Bastiaens PIH. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. Nat Cell Biol. 2007;9(3):324–0. doi:10.1038/ncb1543.
39. Grzegorczyk M. An introduction to Gaussian Bayesian networks. Methods Mol Biol. 2010;662:121–47.
40. Verma T, Pearl J. Equivalence and synthesis of causal models. In: Bonissone P, Henrion M, Kanal LN, Lemmer JF, editors. Uncertainty in Artificial Intelligence 6. Cambridge, MA: Elsevier Science Publishers; 1991. p. 225–68.