Publication III

Juuso A. Parkkinen and Samuel Kaski. Probabilistic drug connectivity mapping. *BMC Bioinformatics*, 15:113, 2014.

Parkkinen and Kaski *BMC Bioinformatics* 2014, **15**:113
http://www.biomedcentral.com/1471-2105/15/113

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Probabilistic drug connectivity mapping

Juuso A Parkkinen[1] and Samuel Kaski[1,2*]

Abstract

Background: The aim of connectivity mapping is to match drugs using drug-treatment gene expression profiles from multiple cell lines. This can be viewed as an information retrieval task, with the goal of finding the most relevant profiles for a given query drug. We infer the relevance for retrieval by data-driven probabilistic modeling of the drug responses, resulting in *probabilistic connectivity mapping*, and further consider the available cell lines as different data sources. We use a special type of probabilistic model to separate what is shared and specific between the sources, in contrast to earlier connectivity mapping methods that have intentionally aggregated all available data, neglecting information about the differences between the cell lines.

Results: We show that the probabilistic multi-source connectivity mapping method is superior to alternatives in finding functionally and chemically similar drugs from the Connectivity Map data set. We also demonstrate that an extension of the method is capable of retrieving combinations of drugs that match different relevant parts of the query drug response profile.

Conclusions: The probabilistic modeling-based connectivity mapping method provides a promising alternative to earlier methods. Principled integration of data from different cell lines helps to identify relevant responses for specific drug repositioning applications.

Keywords: Connectivity mapping, Data integration, Gene expression, Latent variable models, Probabilistic modeling

Background

Current widespread application of high-throughput transcriptional profiling has made large collections of drug-treatment gene expression data both possible and feasible. One of the most important such databases is the Connectivity Map (CMap) [1] that allows users to match transcriptional profiles elicited by drug treatments and diseases. The idea is that any perturbation to the genome-wise gene expression can be summarized by a proper gene signature. Such signatures can be obtained using microarray data and used as proxies of disease phenotypes and drug effects. Matching drugs and diseases based on these signatures is known as *connectivity mapping*, and it has shown promise in drug discovery and repositioning [2-5]. CMap's successor, the Library of Integrated Network-based Cellular Signatures (LINCS, http://www.lincsproject.org/), will offer data for thousands of compounds on tens of cell lines in the near future, providing a unique resource for connectivity mapping -based drug discovery.

Connectivity mapping can be seen as an information retrieval problem, where the task is to find the most relevant gene expression profile for a given query drug profile. The key to successful retrieval is a good definition of the relevance measure. Current connectivity mapping methods define relevance based on similarity in the sets of top up- and down-regulated genes between the two measurement profiles [1] or the consensus profiles constructed by combining all measurement samples for a given drug [2]. Using non-parametric rank-based statistics to define the similarity [6], these methods can integrate data from multiple measurement platforms while reducing batch effects. Alternatively, one could use the Pearson correlation to compute the similarity, but it is more sensitive to platform differences [1].

Transcriptional drug-treatment databases, such as CMap and LINCS, provide measurement data for various experimental factors, including multiple cell types,

*Correspondence: samuel.kaski@aalto.fi
[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland
[2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

doses, and time points. So far, data over multiple experimental factors has been aggregated into a consensus view [2], but this method intentionally ignores possible cell-line-specific effects of the drugs [4]. With the number of experimental factors growing notably in the future, data integration methods capable of distinguishing cell-line-specific effects and various types of consensus or common effects would be needed to bring out the full benefits from connectivity mapping.

In this paper, we propose an alternative, probabilistic model-based approach for defining relevance, with the assumption that a suitably chosen probabilistic model can detect relevant effects from the noisy data. If the representation that the model provides is more informative and less noisy than the input data, retrieval is then more precise based on the model instead of based on the noisy original data. For tractability, we assume that the transcriptional effects caused by drug treatments consists of a set of processes that generate partly overlapping patterns in the observations, and model each process as a probabilistic latent *factor* of data.

Assume then that some of the factors are shared by subsets of the cell lines, and some are specific to individual cell lines. When searching for drugs for a specific type of cancer, for instance, effects in those cell lines are then relevant, and it would be natural to define relevance as activity in those factors.

Relevance stems from the goal of the analyst, and can alternatively be to find effects specific to one cell line. If there are several relevant cell lines, however, a nice side benefit follows: The data contains noise from various sources in addition to the signal, such as measurement batch effects, and the noise is, by definition, specific to individual cell lines. If relevance is defined in terms of the shared activity, it is more tolerant to noise.

What remains now is to find a method to integrate data sources to identify shared patterns. A classical method is the Canonical Correlation Analysis (CCA, [7]), which seeks statistical dependencies between two data sets with paired samples. CCA has been applied for multiple biological problems [8-10]. However, for the general connectivity mapping problem, CCA is not sufficient as it only searches for the shared factors and needs to be generalized to multiple data sources.

A recent data integration method, called Group Factor Analysis (GFA, [11]), is a generalization of CCA directly suitable for the task. GFA decomposes the transcriptional response data into factors specific to individual cell lines and factors shared by two or more cell lines. The name comes from the analysis of groups of variables, here one group for one cell line. Besides being a generalization of CCA, the method generalizes standard factor analysis from finding relationships between scalar variables to finding relationships between groups of variables, or data sources.

Data integration with GFA is one key novel aspect in our method, as the earlier connectivity mapping methods intentionally did not study which responses generalize across the cell lines and which do not. The consensus-based method [2] assumes that only the general effects of drugs are relevant, effectively discarding any specific effects as noise. This is optimal only in the case of drugs with similar effects across cell lines, but this is not always true and hence the consensus-based method is overly restrictive. GFA scales to an arbitrary number of data sources, and the Bayesian probabilistic modeling makes it possible to cope with the biggest problem of gene expression data, the "large p small n" problem of having a relatively large number of variables (genes) compared to the number of samples.

Given the probabilistic model, retrieval of the relevant drug response profiles is then performed based on an activity profile over the factors, or alternatively the latent factor representation, the model has learned from data. We call the approach *probabilistic connectivity mapping* (Figure 1). A suitable relevance measure is the Pearson correlation, as it focuses on the active (non-zero) factors of the query and ignores the inactive ones. Depending on the goal, the analyst can choose to focus on factors shared by cell lines, specific factors, or both.

We apply the method to the CMap data and show that it outperforms earlier connectivity mapping approaches in finding functionally and chemically similar drugs. Additionally, the careful data integration helps: Shared factors are the most relevant for the retrieval, but some specific factors are relevant as well. This indicates that while most drugs exhibit similar responses across cell lines, there are also some important differences that are captured by our model.

Alternatively to GFA, a more straightforward probabilistic factor analysis can also be used by simply concatenating data from all cell lines and not taking into account the grouping of the variables according to cell lines. We will consider this alternative as well; GFA is expected to have the advantage that interpretation of the factors should be easier as they explicitly specialize to a subset of cell lines, but the retrieval performances are expected to be similar.

In addition to retrieval of single drugs, we demonstrate how the model-based approach can be extended to retrieve combinations of drugs. The idea is to retrieve a set of drugs, where each drug matches a different part of the relevant query response. This is beneficial for polypharmacology, where drugs have multiple target effects [12,13]. We demonstrate that combinatorial retrieval can provide complementary information to single-drug retrieval for polypharmacologic drugs.

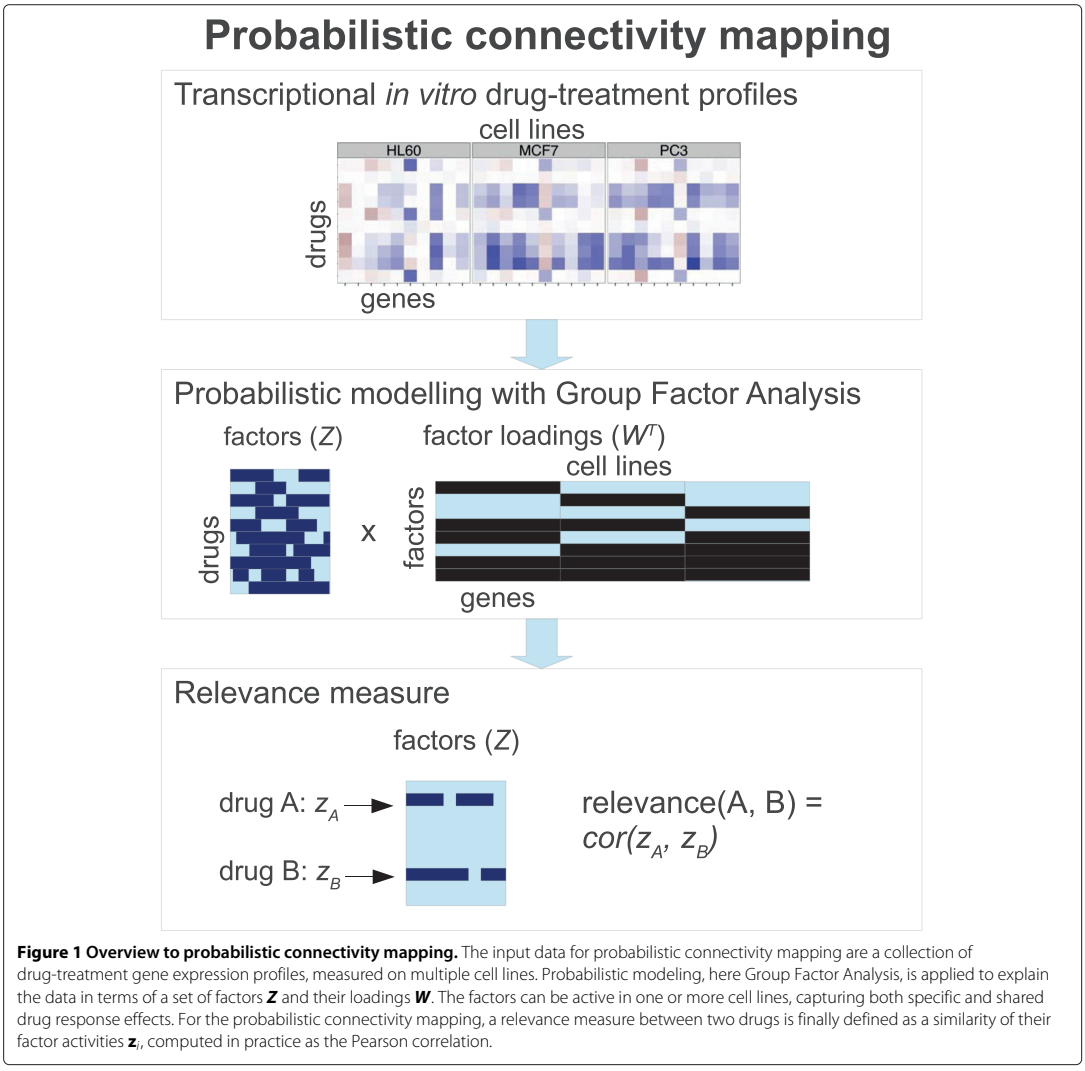

Figure 1 Overview to probabilistic connectivity mapping. The input data for probabilistic connectivity mapping are a collection of drug-treatment gene expression profiles, measured on multiple cell lines. Probabilistic modeling, here Group Factor Analysis, is applied to explain the data in terms of a set of factors $\boldsymbol{Z}$ and their loadings $\boldsymbol{W}$. The factors can be active in one or more cell lines, capturing both specific and shared drug response effects. For the probabilistic connectivity mapping, a relevance measure between two drugs is finally defined as a similarity of their factor activities $\mathbf{z}_i$, computed in practice as the Pearson correlation.

Data integration via probabilistic modeling is expected to bring a couple of further benefits. As the strengths of the responses vary widely, and the data is expected to be heteroskedastic, fixed signature sizes used in current connectivity mapping approaches may lose important information. The probabilistic modeling approach copes with varying sample norm in a natural fashion. A final benefit is the ability to cope with batch effects that plague microarray experiments. They are view-specific by nature, so retrieval that focuses on the shared effects can help to further reduce the batch effects, complementing preprocessing procedures such as mean-centering [14].

Results and discussion

Connectivity mapping results

We evaluated the proposed probabilistic connectivity mapping approach by applying it to a collection of 718 compounds and three cell lines from the CMap database, normalized with mean-centering [14]. The gene expression profiles were modeled across the set of 930 Landmark genes identified in the LINCS project. Three probabilistic models were used: Group Factor Analysis (GFA), sparse factor analysis (sFA), and Bayesian principal component analysis (BPCA). As comparison, we used two earlier connectivity mapping methods: rank-based average

enrichment-score distance (AESD, [2]) and correlation (COR) on the differential expression data averaged over the cell lines. We evaluated the retrieval performance based on two external "ground truths" on relevance: how many of the retrieved samples have the same fourth level ATC codes as the query drug and chemical similarity. We measured retrieval performance with two complementary goodness measures: partial area under the ROC curve and top-10 mean average precision (MAP).

Probabilistic connectivity mapping with GFA and sFA clearly outperform the other methods (Figure 2) on both ground truths and goodness measures. The sFA was slightly better with the partial AUC measure and GFA for the top-10 MAP measure. Bayesian PCA clearly performed worse, indicating that the sparsity assumptions made in GFA and sFA are important for capturing the relevant responses from the data.

In the experiments of Figure 2, we used all factors, as that turned out to produce the best absolute retrieval performance for this data. We next investigated the possible benefits of focusing on the factors shared by the cell lines. The retrieval was based on the most active shared factors (from GFA), and compared to the performance with an equal number of the most active factors that are specific to one cell line. Additionally, we compared this to the most active factors from sFA. Figure 3 shows that the shared factors produce better retrieval almost everywhere. These results suggest that the explicit group-wise sparsity assumption in GFA, resulting in the decomposition to shared and specific effects, is beneficial in modeling data from multiple cell lines.

Combinatorial retrieval results

We next studied how well the method extends to combinatorial retrieval, that is, retrieval of multiple drugs that together are relevant to the query. We queried with drugs having multiple ATC codes, and the ground truth (unknown to the model) was the set of ATC codes. Our hypothesis was that if some of the ATC codes represent minor response effects, drugs with those codes would not get a high relevance score when retrieving single drugs, as the drugs with the other code(s) would dominate. However, the minority codes could show up in combinatorial retrieval. We also expect the combinatorial retrieval to work better when the multiple effects of the query are more varied, as the effects would then get less mixed up. Figure 4 shows an example of combinatorial retrieval results and compares them to single-drug retrieval results. Comparisons of the retrieval performance are summarized in Figure 5. We see that combinatorial retrieval improves the results for a good proportion of the polypharmacologic drugs, and that performance is better with lower ATC levels, that is, more distinguished effects.

As single-drug retrieval is expected to work, even for polypharmacologic drugs, when searching for drugs with

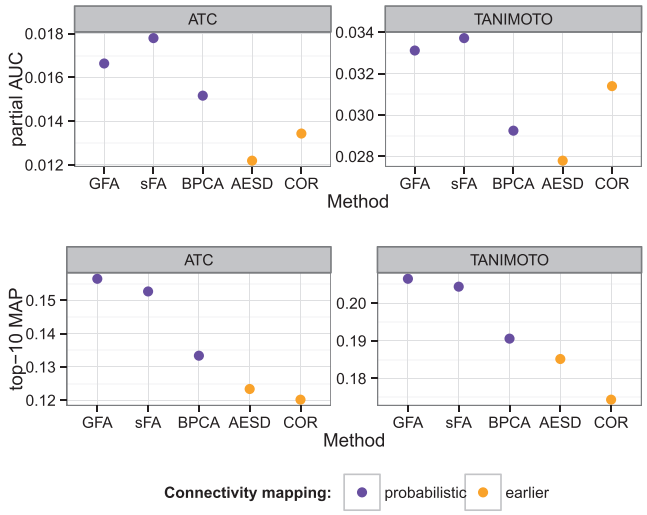

Figure 2 Probabilistic connectivity mapping outperforms earlier alternatives in retrieving similar drugs. The retrieval performance is indicated with two goodness measures (top row: partial AUC, bottom row: top-10 MAP) and two ground truths (left: ATC codes, right: Tanimoto similarity of the 2D fingerprints of the drugs). Probabilistic connectivity mapping (purple color) is performed with three models: Group Factor Analysis (GFA), sparse factor analysis (sFA), and Bayesian PCA (BPCA). These are compared to two earlier connectivity mapping methods (orange color): rank-based average enrichment-score distance (AESD) and the Pearson correlation over the differential expression profiles (COR).

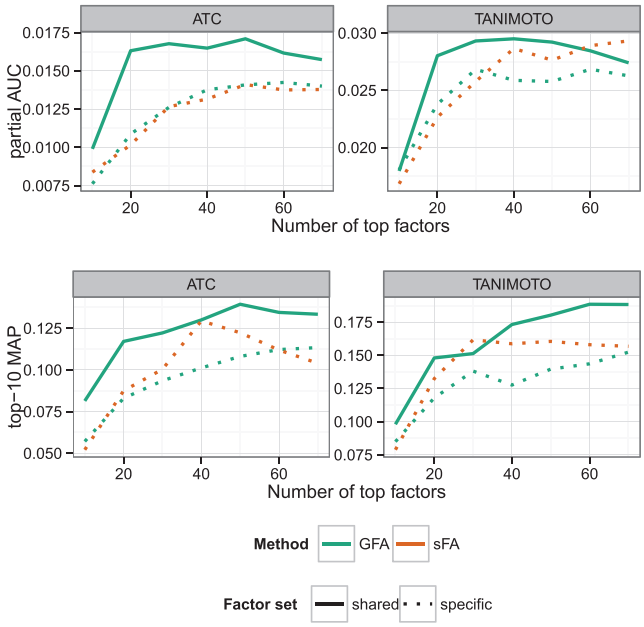

Figure 3 Factors shared across multiple cell lines are more informative for retrieval performance than cell-line-specific factors. Retrieval performance is shown for the top shared (solid line) and specific (dotted line) factors from GFA (green color) and sFA (brown color), as a function of the number of top factors. Factors were selected based on the highest α parameter values.

precisely the same combination of effects, we removed drugs having multiple ATC matches with the query drug from the retrieved set. After that, performance compared to single-drug retrieval clearly improved (Figure 5), indicating that combinatorial retrieval was able to find additional drug combinations and provide complementary information to single-drug retrieval.

Conclusions

We introduced *probabilistic connectivity mapping*, a model-based alternative to earlier drug connectivity mapping methods. Our first contribution was to define the relevance for the information retrieval task based on a probabilistic model that captures the relevant gene expression effects for the query drug in the form of

Query: scopolamine, ATC codes: a N05 a S01

Rank	Single	CombDrug1	CombDrug2	CombRank
1	zuclopenthixol	fusidic acid	zuclopenthixol	fusidic acid
2	zalcitabine	calcium pantothenate	zuclopenthixol	zuclopenthixol
3	buflomedil	succinylsulfathiazole	zalcitabine	calcium pantothenate
4	cyclopenthiazide	fusidic acid	zalcitabine	succinylsulfathiazole
5	succinylsulfathiazole	amphotericin B	zuclopenthixol	zalcitabine
6	benzocaine	netilmicin	zalcitabine	amphotericin B
7	nicotinic acid	megestrol	zuclopenthixol	netilmicin
8	nystatin	pramocaine	zuclopenthixol	megestrol
9	netilmicin	clebopride	zalcitabine	pramocaine
10	felbinac	flunisolide	zuclopenthixol	clebopride

Figure 4 Combinatorial retrieval example. Using scopolamine as the query drug, the top-10 retrieval results are shown for single-drug and combinatorial retrieval, with ATC codes shared with the query indicated by colors. For combinatorial retrieval, the drugs are ranked (CombRank) based on their first appearance in the retrieved pairs (either CombDrug1 or CombDrug2). In the example, using both single-drug and combinatorial retrieval, a match for ATC code N05 is found at the first rank. However, combinatorial retrieval also provides a match for the other ATC code S01 already at the first rank, whereas single-drug retrieval finds a match only at rank 9. The result demonstrates that the combinatorial retrieval approach can be beneficial for polypharmacologic queries.

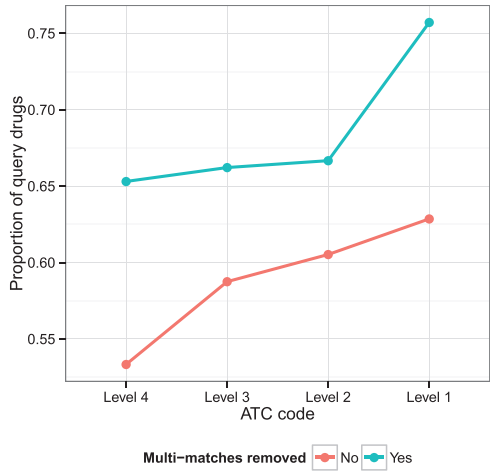

Figure 5 Combinatorial retrieval provides additional information to complement single-drug retrieval. The y-axis indicates the proportion of the query drugs for which combinatorial retrieval improves the rank for the first hit for at least one ATC code (random performance: 0.5). The results are shown for four different ATC code levels (x-axis). Red: retrieval from the full set; blue: retrieval after removing drugs having multiple ATC matches with the query.

probabilistic latent factors inferred from data. The chosen model integrates data over available experimental factors, here cell lines, which has not been considered in earlier connectivity mapping approaches. We showed that probabilistic connectivity mapping outperforms earlier alternatives in finding functionally and chemically similar drugs, based on transcriptional response profiles. We additionally showed that gene expression response factors shared across cell lines, identified by a multi-source probabilistic model, were the most relevant for retrieval. We also confirmed the utility of the Landmark genes identified in the LINCS project.

In addition to single-drug retrieval, we showed how probabilistic connectivity mapping naturally allows retrieval of sets of drugs, and showed how such combinatorial retrieval provides complementary information to single-drug retrieval for drugs with multiple mechanisms of action.

Connectivity mapping has also been proposed for predicting synergistic drug combinations given a disease query [3]. A straightforward assumption is that drugs with similar gene expression signatures could be synergistic, and a successful *in vivo* proof-of-concept of this approach has been reported by Hassane *et al.* [15]. An alternative assumption is to search for drug combinations with either completely independent actions or actions on different but related targets or pathways [16,17], and our proposed combinatorial retrieval method could provide hypotheses for such combinations.

Based on the drug similarity validation with the CMap data, probabilistic connectivity mapping provides a promising alternative for earlier methods. Next, the method could be applied to matching known drugs and drug combinations to disease samples, providing hypotheses of novel therapies.

For the current CMap data, the absolute retrieval performance was at its best when all factors were used for defining the relevance, even though for smaller numbers of factors the shared ones were more informative. We expect this to change when the datasets become larger and more heterogeneous, requiring more expertise from the user to choose a set of informative cell lines, or even more advanced tools to model the users' interests.

As the LINCS-project will generate data over tens of cell lines, we also expect other benefits of the Group Factor Analysis -based probabilistic connectivity mapping to become even more apparent. Being able to identify both shared responses across a large number of cell types, and on the other hand responses specific only to few cell lines, will be highly valuable to drug development and discovery. It would be even possible to impose more structure on the Group Factor Analysis model, inferring which cell lines response similarly to the drugs, providing potentially highly relevant information for personalized medicine approaches.

The recent work by Iskar *et al.* [5] used a biclustering approach to identify important response modules from the CMap data, and identify shared modules based on

overlapping genes as a post-processing step. They proposed using the modules to match drugs, even though they did not proceed to recommending particular metrics. They did, however, demonstrate drug repositioning by validating some examples from both shared and cell-line-specific modules, suggesting that a suitable probabilistic biclustering method (such as [18]) could be usable for probabilistic connectivity mapping as well.

Methods

Data

We used the Connectivity Map (CMap) build 2 drug-treatment transcriptional data [1]. The data was RMA-normalized [19], and we included measurements only from the HT-HG_U133A microarray platform, for drugs that were measured on all three of the most prominent cell lines (MCF7, PC3, HL60). To follow the state-of-the-art preprocessing procedure by Iskar *et al.* [14] we included treatments only from the large CMap batches with around 40 measurements, ignoring the small batches with at most 6 measurements.

For each drug and cell line pair, we included only the highest concentration. Differential expression was computed against the mean of the treatment measurements for each batch, instead of the biological controls, as suggested by Iskar *et al.* [14]. Remaining replicates of drug and cell line pairs were merged by averaging. This resulted in drug-treatment gene expression profiles for 718 drugs for the three cell lines. We additionally re-computed the preprocessing by including treatments from all batches. This resulted in the addition of only 1.5 % more treatments and no new chemicals, and hence the results for all methods, and conclusions, were expectedly practically identical to those using only the large batches.

Instead of the full genome, we used the set of Landmark genes provided by the LINCS project (http://lincscloud.org/the-landmark-genes/). This set of about 1000 genes has been curated based on large gene expression compendium to be minimally redundant, widely expressed in various cellular contexts, and largely representative of the full genome. Using this particular set of genes is thus expected to result in a higher signal-to-noise ratio in the data, as compared to the full genome. The retrieval performance using the Landmark genes was indeed better for all methods as compared to using the full genome (results not shown), confirming that using them is a sensible choice for connectivity mapping. Of the 968 Landmark genes provided by LINCS, 930 were present in the CMap data.

Rank-based connectivity mapping

Existing connectivity mapping methods use a gene set enrichment-based [6] measure for matching drugs [1]. In this paper, we use the method described by Iorio *et al.* [2]: The genes were first ranked based on differential expression. For each drug, the ranked gene lists from the different cell lines were then merged by the KruBor rank aggregation method. A consensus gene signature was then produced by taking the top up- and down-regulated genes from the merged list. The query drug was matched to other drugs in the database by computing the Kolmogorov-Smirnov statistics based enrichment score between the query signature and the ranked lists of the other drugs. We tried both average and maximum enrichment-score distances (AESD, MESD), AESD giving better retrieval performance. Using the full genome, Iorio *et al.* [2] identified 250 genes as an optimal signature size. However, as we are using only the 930 Landmark genes, we re-validated the signature size, resulting in the best retrieval performance with a signature size of 50 genes (results not shown).

Probabilistic connectivity mapping with Group Factor Analysis

Factor analysis (FA) is a standard data analysis tool for capturing and understanding linear relationships between variables [20]. It uses a set of K factors to explain dependencies between the features in a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$:

$$\mathbf{X} = \mathbf{Z}\mathbf{W}^T + \mathbf{E}\,, \tag{1}$$

where the columns of $\mathbf{Z}$ are the K unobserved factors, $\mathbf{W} \in \mathbb{R}^{D \times K}$ contains their loadings, and $\mathbf{E}$ is Gaussian residual noise. Different factor analysis variants can be defined by choosing specific priors for the loadings $\mathbf{W}$ and structure for the residual noise $\mathbf{E}$.

Group Factor Analysis (GFA) was recently introduced [11] for generalizing from modeling of dependencies between scalar variables, which FA does, to modeling dependencies between data sets. In the machine learning community, learning from multiple sources of data has been called *multi-view* learning, *views* referring to data sets with shared (or co-occurring) samples. Given a collection $\mathbf{X}_1, \ldots, \mathbf{X}_M$ of M views, here cell lines, with shared samples and dimensionalities $D_1, \ldots, D_M$, the task is to find K factors that describe the collection and in particular the dependencies between the data views $\mathbf{X}_m$. For simplicity, we assume normally distributed data. This choice can of course be tailored if there's more prior knowledge. In this paper, the assumption is validated based on external retrieval validation. The likelihood for observed data $\mathbf{X}$ is

$$p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \boldsymbol{\tau}) = \prod_{m=1}^{M} \mathcal{N}(\mathbf{X}_m|\mathbf{Z}\mathbf{W}_m^T, \boldsymbol{\tau}_m^{-1}\mathbf{I})\,. \tag{2}$$

Now the noise $\mathbf{E}$ in Equation 1 is diagonal $[\boldsymbol{\tau}_1^{-1}, \ldots, \boldsymbol{\tau}_M^{-1}]$ with each $\boldsymbol{\tau}_m^{-1}$ repeated D_m times. Hence, every dimension within view m has the same noise variance, whereas

the views may have different variances. A Gamma prior is used for the inverse variances $\boldsymbol{\tau}_m$:

$$p(\boldsymbol{\tau}|a^\tau, b^\tau) = \prod_{m=1}^{M} \mathcal{G}(\boldsymbol{\tau}_m|a^\tau, b^\tau) \,. \quad (3)$$

The factors $\mathbf{Z}$ are assumed to be normally distributed with zero mean and unit covariance:

$$p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \,. \quad (4)$$

The weight matrix $\mathbf{W}$ is made group-sparse by a group-wise automatic relevance determination (ARD) prior,

$$p(\boldsymbol{\alpha}|a^\alpha, b^\alpha) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}(\boldsymbol{\alpha}_{m,k}|a^\alpha, b^\alpha) \quad (5)$$

$$p(\mathbf{W}) = p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{N}(\mathbf{w}_{m,k}(d)|0, \alpha_{m,k}^{-1}) \,, \quad (6)$$

where $\mathbf{w}_{m,k}(d)$ denotes the dth element in the projection vector $\mathbf{w}_{m,k}$. The inverse variance of each vector is controlled by the parameter $\alpha_{m,k}$ with a Gamma prior. The hyperparameters a^τ, b^τ, a^α and b^α are set to very small values, here 10^{-14}.

The ARD makes groups of variables inactive for specific factors by driving their $\alpha_{m,k}^{-1}$ to zero, providing factors that are active for only a specific subset of the views. The ability of GFA to separate shared and specific effects is the core of the model, distinguishing it from earlier factor analysis models. The ARD prior is simultaneously used to control the model complexity, that is, the number of factors, by shutting down unused factors during the inference. There are other alternatives for the ARD prior that could be explored in the future. Model inference is carried out with a variational approximation, using the R package CCAGFA available in CRAN [11]. Details of the inference are given in the Appendix.

To evaluate the benefits from the multi-view Group Factor Analysis for probabilistic connectivity mapping, we compare it to two alternative formulations of the factor analysis problem that do not use the multi-view information. For this, we concatenate all data into a single data matrix $\boldsymbol{X}$. First, we assume that the noise variance is equal over the variables, reducing the factor analysis to the Bayesian principal component analysis (BPCA) [21]. Second, we assign each feature an independent ARD-prior, resulting in a sparse factor analysis model (sFA, [22]).

Given the set of factors $\mathbf{Z}$, identified by the model applied on a collection of drug-treatment measurements from multiple cell lines, the probabilistic connectivity mapping procedure is completed by computing the relevance measures between pairs of drugs. We define the relevance between drugs i and j as the Pearson correlation between the latent variables $\mathbf{z}_i$ and $\mathbf{z}_j$. The correlation-based relevance measure has the favorable property of focusing on the active (non-zero) factor values, representing relevant activity for the query. The measure is additionally normalized by definition, removing the effects of varying norms of the samples. Depending on the task of the analyst, the relevance can be computed over all or a subset of the factors, for example only the factors shared by two or more views. In this paper, we use data from all three cell lines in the CMap data, preprocessed as in [14], to allow fair comparison with the alternative methods. However, the model could be learned from only a subset of the cell lines as well.

Combinatorial retrieval

There are many situations where single-drug connectivity mapping does not provide fully satisfactory results. For example, many drugs activate multiple targets and biological processes, which is called polypharmacology [12,13]. If we assume that a query drug q activates two distinct biological processes, single-drug retrieval would tend to provide relevant matches to only the most dominant one of them, whereas an optimal retrieval result would cover them both. This can be achieved with combinatorial retrieval, where pairs (or more) of drugs are searched for instead of single drugs, such that each drug in the pair matches to one of the active processes of the query. This can be formulated as an extension of the probabilistic connectivity mapping to combinatorial retrieval. The goal is then to search for the pair p of drugs i and j that jointly explain the query activity better than any single drug. This is achieved by combining the factor profiles of the pair of drugs into a single factor profile $\mathbf{z}_p$ such that it maximizes the relevance, i.e. $cor(\mathbf{z}_p, \mathbf{z}_q)$. Formally, $\mathbf{z}_p = \{z_{p,k}\}, k \in \{1, \ldots, K\}, z_{p,k} \in \{z_{i,k}, z_{j,k}\}$. In other words, each factor value $z_{p,k}$ is chosen from either $\mathbf{z}_i$ or $\mathbf{z}_j$, and the choices are made to maximize $cor(\mathbf{z}_p, \mathbf{z}_q)$.

Validation

To validate the probabilistic connectivity mapping approach, we use two external ground truth data sets of known drug similarity as in [14]: Shared ATC codes and chemical similarity. According to the first set, drugs are considered functionally similar if they share the level four Anatomic Therapeutic Chemical (ATC) classification codes [23]. The ATC is a hierarchical grouping of drugs based on the organ or systems on which they act, and their therapeutic, pharmacological, and chemical properties. The alternative is to consider two drugs (chemically) similar if the Tanimoto similarity between their 2D fingerprints is higher than 0.8. Tanimoto similarities are computed using the rcdk R package [24].

Two different goodness measures are computed for the retrieval, given a ranked list of other drugs for the query drug, and external ground truth stemming from either Tanimoto or ATC. The first is partial area under the ROC curve ($FPR < 0.1$) over the pooled set of all drug pair similarities, as in [14]. The second is top-10 mean average precision (MAP), a standard goodness measure in information retrieval. The two goodness measures focus on different, complementary aspects of retrieval performance: Partial AUC focuses on the overall shortest distances, which the user might want to explore, emphasizing the cases where relevant matches for the drugs are easily found. The top-10 MAP, in contrast, is a mean over all query drugs, giving equal weight also to those drugs for which a match is harder to find.

To validate the combinatorial retrieval approach, we constructed a setup for testing the ability of the model to retrieve relevant drugs for a given polypharmacologic query drug. In particular, we used the subset of drugs with multiple ATC code assignments as queries. The results were ranked based on both single-drug retrieval and combinatorial retrieval, and the top rank positions in which each ATC code shared with the query first appeared in the lists were found. We then computed the proportion of query drugs with at least one ATC code for which the combinatorial retrieval gives an improved ranking compared to single retrieval, using ATC code levels from one to four. The rationale is that if one ATC label dominates the effects, it is likely to appear high in the standard (single drug) retrieval, whereas other minor effects related to other ATC(s) may be further down in the results list. Combinatorial retrieval, however, also allows minor results to appear in the top ranks. By jointly evaluating all the ATC codes for the query drug, we can see whether combinatorial retrieval finds drugs that match the ATC codes but do not show up high on standard retrieval.

As there are some drugs that share the same multiple ATC codes, those are likely to be found by single-drug retrieval more easily. We thus additionally evaluated the setup where such drugs are removed from the set of drugs retrieved; this should highlight how many additional drugs the combinatorial retrieval can find.

Availability and requirements

Project name: Probabilistic connectivity mapping
Project home page: http://research.ics.aalto.fi/mi/software/ProbCMap/
Operating systems: Platform independent
Programming language: R
Other requirements: None
License: FreeBSD
Any restrictions to use by non-academics: No

Appendix

The full posterior distribution of the GFA model is

$$p(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}|\mathbf{Z},\mathbf{W},\boldsymbol{\alpha},\boldsymbol{\tau})p(\mathbf{Z})p(\mathbf{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|a^{\alpha},b^{\alpha}) \\ p(\boldsymbol{\tau}|a^{\tau},b^{\tau})/p(\mathbf{X}) \; . \tag{7}$$

For the variational inference, the posterior is approximated as

$$p(\boldsymbol{\theta}|\mathbf{X}) \approx q(\boldsymbol{\theta}) = q(\mathbf{Z})q(\mathbf{W})q(\boldsymbol{\alpha})q(\boldsymbol{\tau}) \; . \tag{8}$$

The latent factors are updated as

$$q(\mathbf{Z}) = \prod_{i=1}^{N} q(\mathbf{z}_i) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i|\mathbf{m}_i^{(z)}, \boldsymbol{\Sigma}^{(z)}) \; , \tag{9}$$

where the parameters are:

$$\boldsymbol{\Sigma}^{(z)} = \left(\mathbf{I}_k + \sum_{m=1}^{M} \langle \boldsymbol{\tau}_m \rangle \langle \mathbf{W}^{(m)} \mathbf{W}^{(m)\top} \rangle \right)^{-1}$$

$$\mathbf{m}_i^{(z)} = \sum_{m=1}^{M} \boldsymbol{\Sigma}^{(z)} \langle \mathbf{W}^{(m)} \rangle \langle \boldsymbol{\tau}_m \rangle \mathbf{x}_i^{(m)} \; .$$

The projection matrices are updated as

$$q(\mathbf{W}) = \prod_{m=1}^{M} \prod_{j=1}^{D_m} \mathcal{N}(\mathbf{w}_{:,j}^{(m)}|\mathbf{m}_{m,j}^{(w)}, \boldsymbol{\Sigma}_m^{(w)}) \; , \tag{10}$$

where $\mathbf{w}_{:,j}^{(m)}$ denotes the jth column of matrix $\mathbf{W}^{(m)}$,

$$\boldsymbol{\Sigma}_m^{(w)} = \left(\langle \boldsymbol{\tau}_m \rangle \sum_{i=1}^{N} \langle \mathbf{z}_i \mathbf{z}_i^{\top} \rangle + \langle \overline{\overline{\boldsymbol{\alpha}}}_m \rangle \right)^{-1}$$

$$\mathbf{m}_{m,j}^{(w)} = \boldsymbol{\Sigma}_m^{(w)} \langle \boldsymbol{\tau}_m \rangle \left(\sum_{i=1}^{N} \mathbf{x}_{ij}^{(m)} \langle \mathbf{z}_i \rangle \right) \; ,$$

and $\overline{\overline{\boldsymbol{\alpha}}}_m$ is the mth row of $\boldsymbol{\alpha}$ transferred into a diagonal $K \times K$ matrix.
The noise precision $q(\boldsymbol{\tau}) = \prod_{m=1}^{M} \mathcal{G}(\boldsymbol{\tau}_m|a_m^{\tau}, b_m^{\tau})$ parameters are updated as

$$a_m^{\tau} = a^{\tau} + \frac{D_m N}{2}$$

$$b_m^{\tau} = b^{\tau} + \frac{1}{2} \sum_{i=1}^{N} \left\langle (\mathbf{x}_i^{(m)} - \mathbf{W}^{(m)\top} \mathbf{z}_i)^2 \right\rangle .$$

The ARD precision $q(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}(\alpha_{mk}|a_m^{\alpha}, b_{m,k}^{\alpha})$ parameters are updated as

$$a_m^{\alpha} = a^{\alpha} + \frac{D_m}{2}$$

$$b_{m,k}^{\alpha} = b^{\alpha} + \frac{\langle \mathbf{w}_k^{(m)} \mathbf{w}_k^{(m)\top} \rangle}{2} \; .$$

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
The authors developed the method and designed the experiments together. JP implemented the method and carried out the experiments. Both authors read and approved the final manuscript.

Acknowledgements
We'd like to thank Suleiman A. Khan for his useful discussions and insights. *Funding:* The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170; Computational Modeling of the Biological Effects of Chemicals, 140057), Helsinki Doctoral Programme in Computer Science.

Received: 18 December 2013 Accepted: 14 April 2014
Published: 17 April 2014

References

1. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR: **The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease.** *Science* 2006, **313**(5795):1929–1935.
2. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D: **Discovery of drug mode of action and drug repositioning from transcriptional responses.** *Proc Nat Acad Sci* 2010, **107**(33):14621–14626.
3. Qu XA, Rajpal DK: **Applications of connectivity map in drug discovery and development.** *Drug Discov Today* 2012, **17**(23-24):1289–1298.
4. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J: **Transcriptional data: a new gateway to drug repositioning?** *Drug Discov Today* 2013, **18**(7-8):350–357.
5. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, Kaminska K. H, Runz H, Gavin A-C, Pepperkok R, van Noort V, Bork P: **Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding.** *Mol Syst Biol* 2013, **9**(1). [http://msb.embopress.org/content/9/1/662.long]
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Nat Acad Sci USA* 2005, **102**(43):15545–15550.
7. Hotelling H: **Relations between two sets of variates.** *Biometrika* 1936, **28**(3/4):321–377.
8. Khan S, Faisal A, Mpindi J, Parkkinen J, Kalliokoski T, Poso A, Kallioniemi O, Wennerberg K, Kaski S: **Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs.** *BMC Bioinformatics* 2012, **13**(1):112.
9. Lin D, Zhang J, Li J, Calhoun V, Deng HW, Wang YP: **Group sparse canonical correlation analysis for genomic data integration.** *BMC Bioinformatics* 2013, **14**(1):245.
10. Huopaniemi I, Suvitaival T, Nikkilä J, Orešič M, Kaski S: **Multivariate multi-way analysis of multi-source data.** *Bioinformatics* 2010, **26**(12):391–398.
11. Virtanen S, Klami A, Khan SA, Kaski S: **Bayesian Group Factor Analysis.** In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. Volume 22.* JMLR W&CP; 2012:1269–1277. Implementation in R available at [http://research.ics.aalto.fi/mi/software/CCAGFA/]
12. Hopkins A, Mason J, Overington J: **Can we rationally design promiscuous drugs?** *Curr Opin Struct Biol* 2006, **16**(1):127–136.
13. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB, Whaley R, Glennon RA, Hert J, Thomas KL, Edwards DD, Shoichet BK, Roth BL: **Predicting new molecular targets for known drugs.** *Nature* 2009, **462**(7270):175–181.
14. Iskar M, Campillos M, Kuhn M, Jensen LJ, van Noort V, Bork P: **Drug-induced regulation of target expression.** *PLoS Comput Biol* 2010, **6**(9):1000925.
15. Hassane DC, Sen S, Minhajuddin M, Rossi RM, Corbett CA, Balys M, Wei L, Crooks PA, Guzman ML, Jordan CT: **Chemical genomic screening reveals synergism between parthenolide and inhibitors of the PI-3 kinase and mTOR pathways.** *Blood* 2010, **116**(26):5983–5990.
16. Jia J, Zhu F, Ma X, Cao Z, Cao ZW, Li Y, Li YX, Chen YZZ: **Mechanisms of drug combinations: interaction and network perspectives.** *Nat Reviews Drug Discov* 2009, **8**(2):111–128.
17. Al-Lazikani B, Banerji U, Workman P: **Combinatorial drug therapy for cancer in the post-genomic era.** *Nat Biotech* 2012, **30**(7):679–692.
18. Caldas J, Kaski S: **Hierarchical generative biclustering for microRNA expression analysis.** *J Comput Biol: J Comput Mol Cell Biol* 2011, **18**(3):251–261.
19. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**(4):e15.
20. Thurstone L: **Multiple factor analysis.** *Psychol Rev* 1931, **38**(5):406–427.
21. Bishop CM: **Variational principal components.** In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470), Volume 1.* IEE; 1999. [http://research.microsoft.com/apps/pubs/default.aspx?id=67241]
22. Archambeau C, Bach F: **Sparse probabilistic projections.** In *Advances in Neural Information Processing Systems, Volume 21*. Cambridge, MA: MIT Press; 2009:73–80. [http://dblp.uni-trier.de/rec/bibtex/conf/nips/ArchambeauB08]
23. WHO Collaborating Centre for Drug Statistics Methodology: **ATC classification index with DDDs, 2013 (Oslo 2012).** [http://www.whocc.no/atc_ddd_index/]
24. Guha R: **Chemical informatics functionality in R.** *J Stat Softw* 2007, **18**(5):1–16.

doi:10.1186/1471-2105-15-113
Cite this article as: Parkkinen and Kaski: Probabilistic drug connectivity mapping. *BMC Bioinformatics* 2014 **15**:113.