Searching for Synergies: Matrix Algebraic Approaches for Efficient Pair Screening

OPEN ACCESS Freely available online

PLOS ONE

Searching for Synergies: Matrix Algebraic Approaches for Efficient Pair Screening

Philip Gerlee[1,2], Linnéa Schmidt[1,3], Naser Monsefi[1], Teresia Kling[1,3], Rebecka Jörnsten[2], Sven Nelander[1,3]*

1 Sahlgrenska Cancer Center, University of Gothenburg, Gothenburg, Sweden, **2** Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, Gothenburg, Sweden, **3** Department of Immunology, Genetics and Pathology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Abstract

Functionally interacting perturbations, such as synergistic drugs pairs or synthetic lethal gene pairs, are of key interest in both pharmacology and functional genomics. However, to find such pairs by traditional screening methods is both time consuming and costly. We present a novel computational-experimental framework for efficient identification of synergistic target pairs, applicable for screening of systems with sizes on the order of current drug, small RNA or SGA (Synthetic Genetic Array) libraries (>1000 targets). This framework exploits the fact that the response of a drug pair in a given system, or a pair of genes' propensity to interact functionally, can be partly predicted by computational means from (i) a *small* set of experimentally determined target pairs, and (ii) pre-existing data (e.g. gene ontology, PPI) on the similarities between targets. Predictions are obtained by a novel matrix algebraic technique, based on cyclical projections onto convex sets. We demonstrate the efficiency of the proposed method using drug-drug interaction data from seven cancer cell lines and gene-gene interaction data from yeast SGA screens. Our protocol increases the rate of synergism discovery significantly over traditional screening, by up to 7-fold. Our method is easy to implement and could be applied to accelerate pair screening for both animal and microbial systems.

Citation: Gerlee P, Schmidt L, Monsefi N, Kling T, Jörnsten R, et al. (2013) Searching for Synergies: Matrix Algebraic Approaches for Efficient Pair Screening. PLoS ONE 8(7): e68598. doi:10.1371/journal.pone.0068598

Editor: Shu-Dong Zhang, Queen's University Belfast, United Kingdom

Received December 15, 2012; **Accepted** May 31, 2013; **Published** July 25, 2013

Funding: The research received support from the Swedish Research Council (vr.se), the Swedish Cancer Society (cancerfonden.se), and Science for Life Laboratory (scilifelab.uu.se). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sven.nelander@igp.uu.se

Introduction

System-scale chemical and genetic screens have progressed from testing single targets to testing combinations of targets. Pairwise tests can reveal functional couplings, such as drug-drug synergism and pathway modules, that cannot be captured by single target screens. In a typical setting, the functional interaction between two targets i and j (drugs or genes) is calculated as an interaction score X_{ij}, commonly defined as:

$$X_{ij} = W_{ij} - W_i W_j, \quad (1)$$

where W_i and W_j are the relative phenotypes after perturbations of single targets i, j and W_{ij} is the response to perturbation of the i and j combination.

System-scale mapping of all interaction scores X_{ij} can serve several important purposes. First, positive and negative values of X_{ij} can be interpreted within the framework of epistasis analysis to deduce pathway relationships between the targets i and j, or to define functional modules in the system [1–7]. Second, both negative and positive interactions are of considerable therapeutic interest. Negative interactions reveal synergistic target pairs that can increase efficiency and widen the therapeutic window of a treatment. Positive interactions can reveal redundant target pairs that may slow down the acquisition of drug resistance [8,9]. Screens in several cellular systems, e.g. cancer cells, have revealed that combination effects are prevalent [10]; thus, mapping interaction scores in cellular systems presents an important challenge for systems biology [11–14].

In a traditional pair screening process, an interaction score, X_{ij}, is experimentally obtained for *every* pair (i,j), and pairs are considered interacting if the interaction score (or some relevant statistic that captures functional coupling) exceeds a threshold. Exhaustive screening is a very costly strategy, since the number of experiments needed grows quadratically with the number of targets, n. The largest pair screening reported [4] is of a magnitude of $n \approx 4500$. However, to screen drug libraries ($n > 100{,}000$) or human shRNA libraries ($n > 5{,}000$), the experimental burden would be prohibitive for standard labs.

Here, we therefore recast the screening problem in terms of a different goal: can we find a *reasonably high fraction* of all synergistic pairs (e.g. 75%), by testing a *relatively low fraction* of all pairs (e.g. 20%)? The acceleration of pairwise interaction mapping was previously proposed in the context of pulldown experiments for PPI mapping [15,16], but also methods specific to genetic interactions have been proposed [17,18]. Our method differs from these in that it exploits properties of interaction networks common to both PPIs and genetic networks, and hence has wider applicability. In addition, the method does not assume a particular experimental design as in pulldown experiments.

We introduce a mathematical notion of screening efficiency and methods to maximize this efficiency, based on alternation between

gradual experimental testing and a matrix algebraic technique to predict synergism. The functioning of this novel algorithm does not rely on the degree of target specificity, or a particular choice of interactions measure, and using several data sets from yeast and cancer cell lines, we demonstrate that our method greatly improves screening efficiency and is both computationally efficient and easy to implement. Further, the performance of the algorithm can be improved by including similarity between drugs/genes, such as target of action or functional interactions.

Results

Quantifying screening efficiency by the fractional discovery rate

To characterize screening efficiency, we propose to use the *fractional discovery rate*. Since the algorithms we propose are stochastic in nature we suggest to use a metric which quantifies the average behaviour of an algorithm when applied to a certain data set.

Consider the following hypothetical scenario: an idealized screen is carried out in which the experimenter tests a growing fraction z of all possible pairs of drugs/genes. At a given z a fraction f of all synergistic pairs have been discovered. Now imagine that we repeat the screening process many times and calculate the average fraction of discovered pairs at a given z, described by the curve $f(z)$ (Figure 1A,B). We define the fractional discovery rate as the derivative $f'(z)$. If an experimenter screens in a systematic, "brute force" fashion, the expected value of the fractional discovery rate will be given by $(1-f)/(1-z)$, i.e. the ratio between the remaining fraction of synergies $(1-f)$ and the remaining screenable fraction $(1-z)$. This implies that the relation

$$f'(z)=\frac{1-f(z)}{1-z} \tag{2}$$

holds for all $z\in[0,1]$.

Let us now consider a screening principle, such as our proposed method, that enriches for synergism. We summarize this enrichment by a factor of τ, where now instead

$$f'(z)=\tau\frac{1-f(z)}{1-z}, \tag{3}$$

i.e. the fractional discovery rate is τ times higher as compared to "brute force" experimentation. We refer to τ as the *screening efficiency*. We solve this differential equation, with boundary conditions $f(0)=0, f(1)=1$, and obtain the explicit relationship:

$$f(z)=1-(1-z)^{\tau}. \tag{4}$$

For $\tau=1$ this function simply describes a line with slope 1, going from (0,0) to (1,1). Efficient screening procedures should identify a large fraction of synergies from a relative low fraction of all pair experiments, thus resulting in higher values of τ (Figure 1B). An oracle screen (knowledge of which pairs are synergistic) achieves the maximum possible τ, given by 1 divided by the prevalence of synergistic pairs.

We will now discuss how to construct screening procedures that improve synergism discovery. We thus proceed to formulate an experimental protocol that incorporates the following three ideas; (i) concurrent estimation of the synergism propensity; (ii) a novel interaction score imputation framework which performs well in cases where the screened fraction is low and can take biological database information into account, and; (iii) an adaptive strategy that toggles between principles (i) and (ii) to optimize screening efficiency. These three components of the experimental protocol are described in the sections below.

We assess the performance on nine data sets, comprising seven cancer cell lines and two yeast data sets (Methods and Table 1) We reason that achieving a high value of τ across a range of screens of different size and type of data should extrapolate to future screens.

Propensity-based sampling improves screening efficiency

Based on the assumption that some targets are more likely to interact than others (so-called "hubs" in a system), one should be able to increase the screening efficiency, τ, by prioritizing targets that have been identified as synergistic in the early phases of the screen. In previous work, Myers and co-workers have used such methods to predict the number of interactors of yeast genes [4].

Here, we formulate an concurrent estimation scheme to prioritize targets likely to be involved in synergies. We denote a target i's propensity to interact by P_i, $i=1,...,n$. Given current estimates of the P_i, we select pairs (i,j) for testing with probability P_{ij}, where

$$P_{ij}\propto P_iP_j. \tag{5}$$

This simple screening protocol is random, but biased towards the likely hubs of interactions (Algorithm 1, Methods). To make the screening protocol adaptive, we use a Bayesian estimate of P_i, as follows. We first assume that the likelihood to observe X_i synergies for target i, is binomial distributed with parameters P_i, N_i. We subsequently assume that the parameter of this distribution, P_i, was drawn from a conjugate prior beta distribution with parameters α and β. The estimate of P_i is thus given by:

$$\hat{P}_i=\frac{X_i+\alpha}{N_i+\alpha+\beta}, \tag{6}$$

where X_i is the number of synergies found, and N_i the total number of interactions tested for gene (or drug) i. This estimate is the mean of the posterior distribution of P_i, given a beta-distributed prior for P_i with parameters α and β. Maximum marginal likelihood estimates of α and β from data suggest using $\alpha=0.5$ and $\beta=n$ in our protocol (see Methods). This corresponds to a prior belief that relatively few targets constitute hubs of interaction. Applying this simple protocol to our set of 9 different interaction score matrices, and averaging over a large number of realisations, we achieve a τ in the range 2.0 to 4.6 (Figure 1D). At experimental fraction 20%, we are able to discover 37% to 67% of all synergies (compared with 20% for brute-force screening). Using of a flat prior ($\alpha=\beta=1$, prior belief that roughly half the targets interact) reduces τ substantially (Figure 1D). The difference in performance on the different data sets is mainly attributed to the size of the data sets, where the method performs better on larger sets of data (see Discussion).

Matrix imputation for highly incomplete interaction score data

To improve screening efficiency further, we use matrix completion to predict likely synergies from limited amounts of screening data. Matrix completion methods to impute missing values have been tested on interaction score matrices [19,20] when a small percentage of data is missing. Recent results, however, have shown that surprisingly few entries are needed for imputation

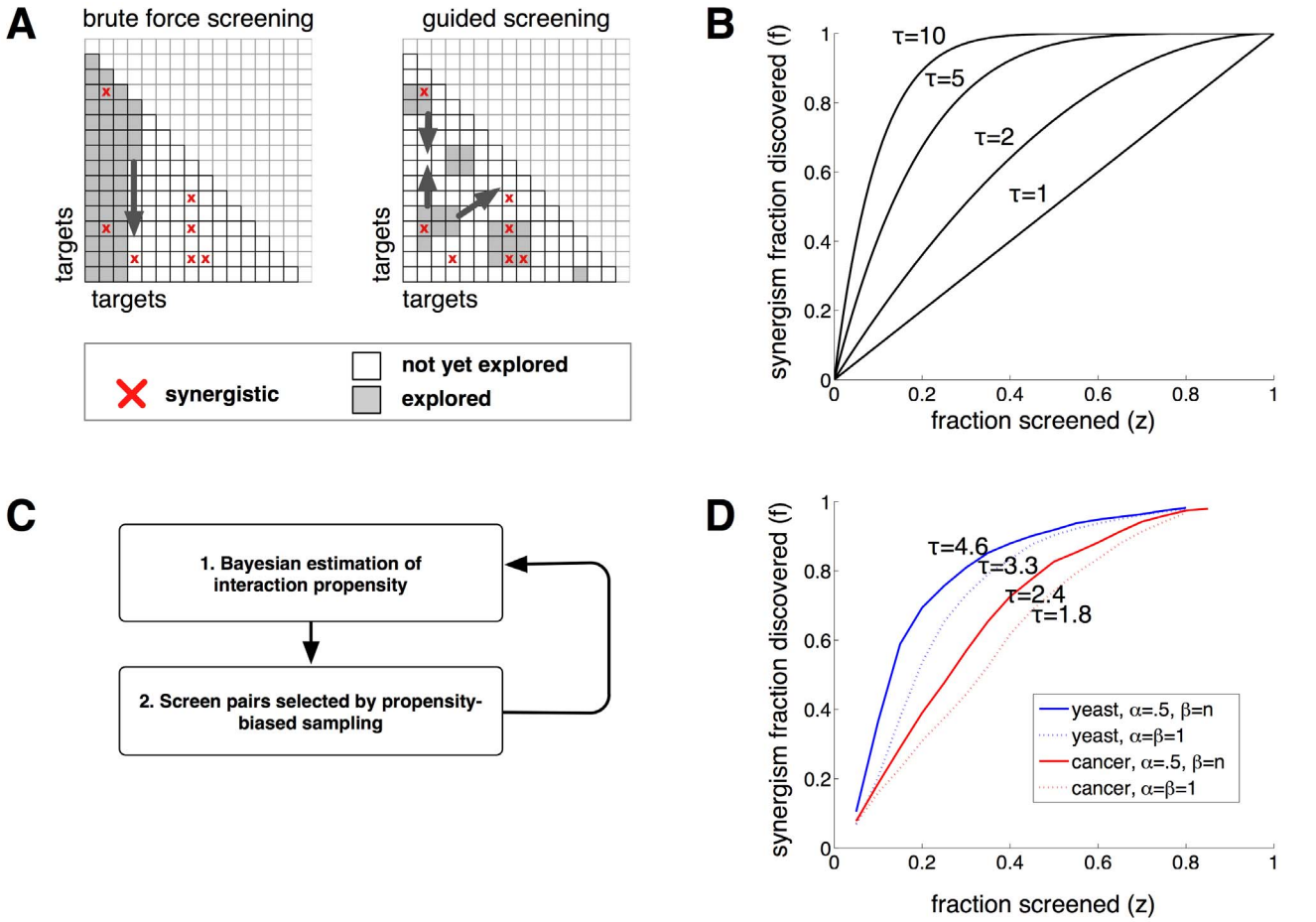

Figure 1. Efficient experimental screening. A: Principal difference between systematic screening (testing all pairs sequentially) and guided screening (letting discovered synergistic pairs, marked as red X's guide the subsequent steps of the screening process). B: We characterize the screening process by the fractional discovery rate $f=1-(1-z)^{\tau}$ attained at experimental fraction z, where τ denotes the screening efficiency. In a screening process with high τ value, a large fraction of all synergies is found by testing a small fraction of all possible target pairs. $\tau=1$ corresponds to systematic screening. C: A simple protocol to increase τ is to direct the screen towards targets that seem prone to synergism. For this, we propose a sampling protocol based on Bayesian estimates of a target's propensity to interact. D: We use a beta-prior with parameters $\alpha=0.5,\beta=n$, where n is the number of targets. A flat prior ($\alpha=\beta=1$) reduces performance.
doi:10.1371/journal.pone.0068598.g001

in cases where a matrix possesses some underlying 'block-like' structure [21,22]. This type of structure is know to be prevalent among interaction score matrices [2,23].

Prediction of interaction scores using set projections. We proceed to define a customized matrix completion method for interaction score data which, unlike standard matrix completion algorithms, encompasses an option to include prior information on functional similarity between targets. For instance, our method can be used to include information on shared mechanism of action between drugs, or shared pathway membership between genes, both likely to give rise to similar interaction behavior. We give a brief description of the method here (details in Methods). The interaction score matrix

Table 1. Benchmarking data sets.

Data set	Assay	Number of targets
Costanzo et al.	Synthetic Genetic Array	4457 (944 with prior information used here)
Schuldiner et al.	Synthetic Genetic Array - like method	427
Colon cancer cell line (HCT116)	Drug-drug interaction	190
Lung cancer cell line (A549)	Drug-drug interaction	190
Glioblastoma cell lines (T98G, U343MG, U87MG, U373MG, A172)	Drug-drug interaction	31 for each cell line

doi:10.1371/journal.pone.0068598.t001

X represents a point in the space of all symmetric $n \times n$ matrices. We assume that this point lies in the intersection of three convex sets, termed R_{data},$R_{modular}$ and R_{sim}, each encoding a different type of evidence:

1. The set R_{data} contains all symmetric matrices that agrees with the currently available data, within an error tolerance level (Methods, eq. 8). For example, if half the entries of the matrix are known, then R_{data} consists of all matrices where the known entries (within error) are equal to the data, and where all other entries are allowed to vary freely.
2. The set $R_{modular}$ contains all matrices with a sufficiently block-like structure. Both gene-gene and drug-drug interaction scores form clusters (blocks), thought to reflect some degree of intrinsic modular organization of cellular pathways [24]. Following [21], we define modularity as a constraint on the nuclear norm of the matrix X (Methods, eq. 9).
3. The set R_{sim} contains all matrices X that conform with externally defined information on functional similarities between the targets. We expect some degree of correlation between the rows/columns i and j in X, when targets (i,j) have a similar biological function [4]. We represent functional similarity by a matrix K, with the property $X \approx KX$, derived from data sources such as PPI networks, GO terms and drug mechanism of action (Methods, eq. 10).

We find a feasible point, i.e. a solution X located in the intersection $R_{data} \cap R_{modular} \cap R_{sim}$, by a cyclical sequence of projections onto these convex sets from a given starting point (Figure 2A). This iterative algorithm is highly efficient and can be applied to data where the number of targets n is quite large, ranging up to a couple of thousand targets (see Implementation in Methods).

Comparison to existing matrix completion methods. To assess the performance of the novel imputation technique, we compare our method to two other techniques: (i) The recently proposed Local Least Squares (LLS) [19] which was advantageously compared to other approaches such as Bayesian Principal Components Analysis [25]; (ii) A meta-predictor, EMDI, which combines LLS and several other methods by a weighted average [20]. For each method, we separately assess the prediction accuracy for each sample fraction z of observed matrix entries, $z \in [0.1,0.9]$ (randomly sampled from the data). Our model gives more accurate predictions from sparse data (10–20% observation) across all data sets (Figure 2B), and is highly competitive even for larger sample fractions z. We compared the prediction accuracy measured as Pearson correlation over 20 independent runs. This gave a strong significance ($p < 0.001$) in 7 datasets (both yeast SGA screens, colon cancer, lung cancer, and A172, T98G, U343 glioma cells), and a moderate/weak significance in 2 datasets (U373 and U87; $p = 0.02$ and 0.06 respectively) For our comparisons, we also considered the APN method by Battle et al. [5]. However, this method is specifically aimed at predicting buffering/antagonistic relationships, and is computationally heavy.

Combining propensity-based sampling with prediction-driven screening

Our final screening procedure (Algorithm 2, Methods) combines and alternates between the two different "search modes" described above: (i) propensity-based random sampling, biased toward untested pairs comprising targets that have hitherto exhibited a high propensity for interacting with other targets (in steps 1 and 2); and (ii) greedy collection of target pairs for which matrix completion has predicted a high degree of synergism (steps 3 and 4). The balance between the two search modes is determined by their performance (step 5) (Figure 3A).

We find that combining interaction prediction and propensity-based sampling to guide the screening process results in a screening efficiency τ much better than the propensity-based sampling alone (Figure 3C). For instance, our procedure detects 62% and 72% of the synergistic pairs in the yeast data sets by testing only 20% of the interactions (Figure 3B). This drops to 46% and 67% when we use propensity-based sampling only (compared with expected 20% for brute-force screening). For the largest data set [4], we see a 4.5-fold increase in efficiency τ for the full protocol over brute-force screening, which drops to about four-fold when the prediction method uses no prior information, and to only three-fold with propensity-based sampling alone (Figure 3D).

Discussion

Performance and data requirements: the impact of modularity and screen size

We note that screening efficiency is most improved for the larger screens. Intuitively, for small systems, there is not much room for targeted screening to improve over brute-force methods. However, as the number of possible interaction pairs grows, it becomes more important to choose experiments carefully to speed up detection. More formally, imputation methods based on nuclear norm constraints (as in our $R_{modular}$) work best when the matrix rank k is much lower than the matrix size n. A theoretical result by Candes et al. [26] states that sample size requirements to make accurate predictions is proportional to $kn \log_2 n$. Assuming the number of modules (k) remains relatively constant for the different biological systems, n is thus the major determining factor for screening efficiency. We repeat the analysis on subsets of the largest data set, showing that screening efficiency indeed is proportional to the number of targets (Figure S1 in File S1). Thus, we expect the screening efficiency obtained with our protocol will grow with the number of targets.

One reason why our prediction approach performs well plausibly lies in the fact that all data sets we analyze exhibit a strong degree of modularity. Previously, both the two yeast data sets and the HCT116 and A549 cancer cell lines have been shown to contain functional clusters [4,13]. To assess whether our own measurements in five glioma cell lines also show some functional modularity, we calculated the correlation (in the X matrix) between drugs that belonged to the same vs. different categories. This analysis confirmed that drugs in the same category (e.g. RTK inhibitors) have higher correlations (Figure S2 in File S1).

In terms of other approaches proposed for accelerated pair screening our method bears some resemblance to the algorithm suggested by Lappe et al. [15]. That method was designed for exploring PPI networks, and gathers information during the screen to accelerate the process. The method exploits the fact that the network contains hubs with high connectivity, and uses as bait for the next experiment the protein that has been seen as prey most often so far. That is similar to our propensity based approach, but only works in the case where a target is tested against all other targets, as is the case of mass spectrometry pull-down experiments. Our method is different, and in a sense more refined, since it not only selects an entire row of the matrix, but a specific matrix element for testing.

The other component of our algorithm, which allows for the incorporation of prior knowledge about the targets, is somewhat similar to the method proposed by Wong et al. [17]. That method uses a wide variety of prior data such as subcellular localisation of the proteins, chromosomal distance etc. With this type of

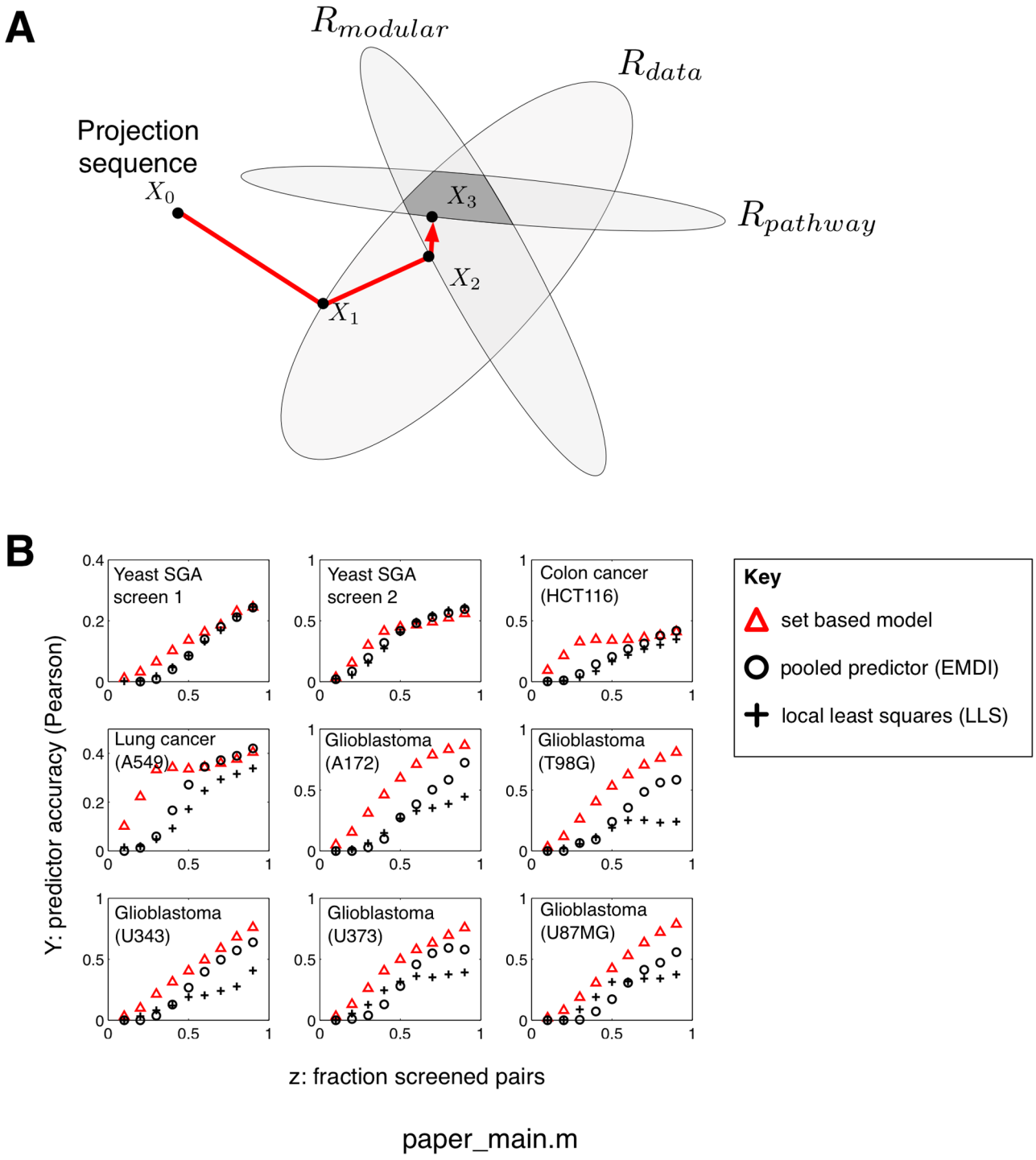

Figure 2. Predicting synergism scores from highly incomplete data via cyclical set projection. A: To improve screening efficiency further, we introduce a projection-based predictor of synergism scores. An initial guess of a synergism score matrix X_0 is projected first onto the set R_{data}, which corresponds to known interaction scores, then onto the set $R_{modular}$, which contains matrices of approximately low rank, and finally onto R_{sim}, holding the matrices consistent with known functional similarity. The projections are applied cyclically until convergence to a final prediction of X is reached, which is guaranteed due to convexity of the three sets (here illustrating convergence in one iteration). B: Prediction accuracy in five glioblastoma cell lines and reference data sets. Comparison between our projection-based method and two state-of-the-art methods for interaction score imputation methods, LLS and EMDI. Generally, set based projections outperform the other methods (predictions correlate more with true values), especially when the screened fraction z is small.
doi:10.1371/journal.pone.0068598.g002

information they can predict the probability that a gene pair exhibits synthetic sick or lethal interactions. However their method is not suited for a screening process and is not optimised for handling incremental data. The main conclusions when comparing our algorithm with other methods is hence that it combines both a subroutine for defining single experiments based on

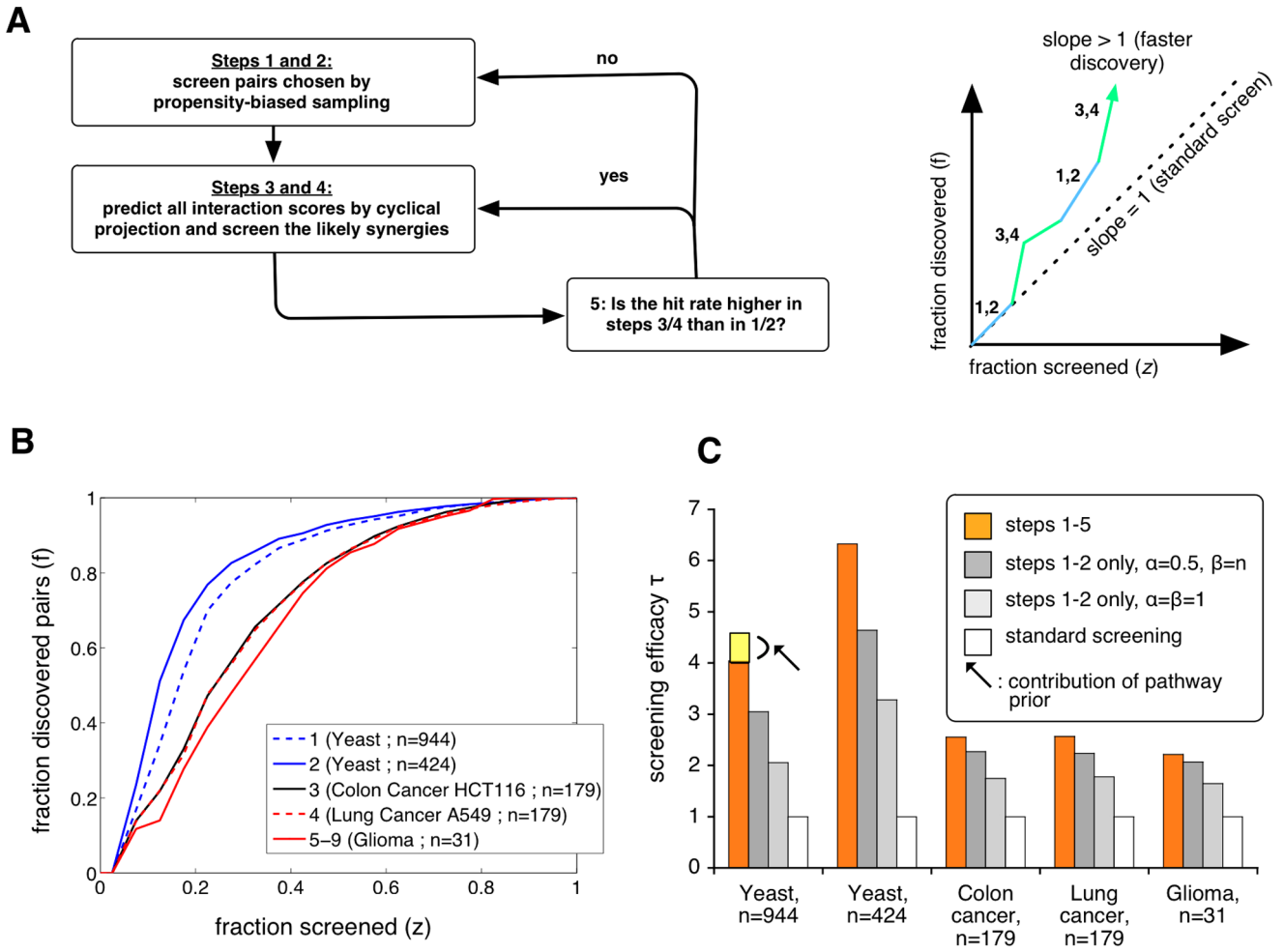

Figure 3. Improving screening efficiency by combining propensity-based sampling with interaction score prediction via matrix completion. A: We extend the simpler protocol (propensity-based sampling only, Figure 1C), adding a projection-based predictor to choose likely synergistic pairs (steps 3 and 4). If the prediction-driven screening discovery rate is higher than the preceding propensity-based screening, a new prediction-driven screening cycle is started (step 5). We switch between propensity-based sampling and prediction to increase the fractional discovery rate. B: Fractional discovery rate across 9 data sets show marked improvement over brute-force screening. C: Estimates of the screening efficiency τ demonstrate that the full protocol (steps 1–5) gives better performance than propensity-based sampling only (steps 1–2). Yellow block: additional contribution by projecting onto R_{sim} in the largest yeast SGA screen.
doi:10.1371/journal.pone.0068598.g003

previous findings in the screen, and allows for the incorporation of prior knowledge about the targets.

In a natural sense every discovery algorithm is limited by the type of assumptions that are made during the construction of the method, or in other words what type of patterns the data is expected to contain. In our case we have exploited the fact that interactions matrices tend to exhibit block-like structure, which mathematically corresponds to a low matrix rank, and also made use of the observation that certain hubs exists in the data. This implies that interactions which deviate from these patterns will be less likely to be detected, and this means that the algorithm is not geared towards "true" discovery, but limited by the assumptions made.

Another difficulty that arises when screening a novel system is to decide on an appropriate synergy threshold value, which effectively determines the number of targets in the screen. Data sets from many diverse system do however suggest that interaction scores have a characersitic long-tailed, non-normal distribution, which lends some hope to transferring knowledge from one system to another. In most cases some preliminary data is also available, e.g. from the SGA data in yeast [4] we could derive a reasonable threshold value (see Methods), and this information can be carried over to other genetic systems.

Conclusions

We have presented a novel method for screening of gene or drug-pairs with the aim of finding synergistic interactions as quickly and cost-efficiently as possible. We expect that advanced matrix imputation methods and prediction based screening procedures, as outlined here, may find several applications. For the five cancer cell lines here analyzed, the proposed methods can serve to rapidly map the interaction landscape for multiple drugs, helping guide discovery screens, and defining combination therapies that overcome some of the shortcomings of current monotherapies for cancer.

We conclude that our approach exhibits good performance on real experimental data. The proposed approach is distinct from previous matrix completion methods, since it also incorporates prior molecular information, and also distinct from methods that

rely completely on molecular data [27–30]. In principle, the approach can be generalized to incorporate additional constraining sets to further improve the solution; this is reserved for future work.

The developed method is geared toward rapid discovery of synergistic pairs, and, in order to achieve this, modular and structural similarities between targets are exploited. The method is thus more likely to discover synergistic interactions that follow this modular pattern, whereas "unexpected" interactions will be harder to find.

Future directions include the exploration of higher order combinations [14], and to introduce improved, target specific estimation of the propensities P_i, by for example taking into account the observed negative correlation between single mutant fitness and number of interactions [4]. It may also be interesting to investigate formal techniques for experimental planning [31,32] and refine strategies to define our functional similarity matrix, K by including e.g. drug side-effect similarity [33]. These measures might improve screening efficiency even further.

Methods

Preparation and generation of benchmarking data sets

Gold standard/public data. Our data consists of 4 publicly available data sets (Table 1 and main text). The first two sets of measurements that we study are standard two-gene synthetic gene array data [4,19], which both contain interaction scores (eq. 1) defined in terms of yeast viability under gene single/double gene knockouts. We used interaction scores as provided, without further normalization, obtained from the supplements of Costanzo et al. [4] (SGA experiments, using the rigorous cutoff preparation of the data; and SGA/ESP data as provided in the supplement of Colm et al. [19]). We used QQ plots against a normal distribution to choose a point where negative interaction scores deviated significantly from a normal distribution. This gave us a threshold value roughly 3 standard deviations from the mean, giving a prevalence of synergies of 1%. The next two data sets were obtained from Zalicus (previously CombinatoRx, a company that pursues drug pair screening) and represent drug pair responses in HCT116 and A549 colon cancer and lung cancer cells, respectively. Here, the interaction scores quantify drug-drug interaction across multiple doses, using a customized metric defined as in Lehar et al. [13]. For the CombinatoRx data, we used the synergism thresholds defined in the original publication (an S-index less than -0.29) which corresponds to a prevalence of synergies of 6%.

Experiment in five glioblastoma cell lines. In addition, we generated data for five glioblastoma cell lines, as follows. Glioma cell line T98G was obtained from ATCC and A172, U-343MG, U-373MG and U-87MG were obtained from Cell lines Services, Germany. All cell lines were grown in monolayer and maintained in high-glucose (4.5 g/l) DMEM supplemented with 10% FBS (Fetal Bovine Serum), 1% PEST (Penicillin/Streptomycin) and 2 mM L-Glutamine and incubated at 37°C with 5% CO_2 in a New Brunswick Galaxy R Incubator. For the experiments we selected a set of 31 compounds, some of which were selected uniformly at random from a library, and some of which had a similar or related mechanism of action. The drugs used are listed in Supplementary Table 1 in File S1. Tumor cells were plated at 1.5×10^3 cells/well in a TPP 96-well plate 24 hours prior to treatment. Cells were treated with drugs diluted in media, single or in combination, and incubated for 48 hours. For combinations, 4 replicates were performed with 3–6 replicate negative controls of equal amount of DMSO (0.1–0.2%). Viability studies were performed using the alamar blue assay (Invitrogen Corp.). At end of experiment cells were incubated for 4 hours with alamar blue reagent (Invitrogen Corp.) for cell viabilty measurements. Fluorescence was read at Exc544/Em590 on a microplate reader (SPECTRAmax GEMINI XS, Molecular Devices). From viability assays, we quantified the drug response as the ratio $W=\bar{Y}_{treated}/\bar{Y}_{control}$, where Y represents the fluorescence signal and bar represents average across replicates. We measured interaction scores for 465 drug pairs (corresponding to all pairs chosen from 31 compounds) using eq. (1), which is usually referred to as the Bliss interaction score. To identify synergies we proceed as with the public SGA screens, lowered the threshold to 1.5 standard deviations, and obtain $6-14\%$ synergies in the different cell lines.

Estimating a target's marginal propensity to interact

A target's propensity to interact is estimated concurrently during the screen using the formula

$$\hat{P}_i=\frac{X_i+\alpha}{N_i+\alpha+\beta} \tag{7}$$

where X_i is the number of synergies found so far, N_i the total number of interactions tested for gene (or drug) i, and we assume a beta-prior, with parameters α and β, for a target's propensity to interact. The prior mean, $\alpha/(\alpha+\beta)$, signifies the *a priori* expected interaction frequency of each target, here assumed to be the same for all targets.

We estimated the parameters α and β from the data sets using a maximum marginal likelihood estimate of the probability distribution P_i^A, i.e. the probability to find a gene (or drug) with i synergistic interactions in data set A (Empirical Bayes) [34]. The obtained values were for α in the range 0.26–1.05 and β close to the number of targets (genes or drugs) n. As a rule of thumb, we therefore suggest to use $\alpha=0.5$ (the median observed value for all data sets) and $\beta=n$ as a prior in our protocol. These values of α and β correspond to a prior skewed toward few interaction hubs.

We also compare the screening procedure obtained with a so-called flat prior ($\alpha=\beta=1$). This prior corresponds to a prior belief that half of the targets are involved in a synergistic interaction and is slow to adapt to findings in the early phases of the screen. Concurrent estimates $\hat{P}_i=X_i/N_i$ (no prior), on the other hand, are too sensitive to early findings in the screen. In principle, α and β could be chosen in a gene-specific manner, but this is reserved for future work.

Matrix completion for interaction scores

The interaction score matrix should be in the intersection of three sets. We view the interaction matrix X as an unknown point in the space of all real-valued symmetric matrices of size $n\times n$, denoted Sym_n. Our goal is to use different kinds of available evidence to define constrained subsets of Sym_n, which contain the feasible values of X. Given these subsets, we will predict X by finding a single feasible point that is located in the intersection of all three constrained subsets.

The first subset, $R_{data}\subset\mathrm{Sym}_n$, contains all symmetric matrices that are consistent with our experimental observations. This set is defined by the sum of square distance from the experimental points, i.e.

$$R_{data}=\{X;\ \|(X_\Omega-M_\Omega)\|_F^2<\varepsilon_d\} \tag{8}$$

where M is the experimental data, and the notation M_Ω denotes that M is only determined (observed) for a subset Ω of matrix

elements, which correspond to the known interaction scores. The norm $\|\cdot\|_F^2$ denotes the Frobenius norm (sum of the squared elements) for matrices, and ε_d is an upper bound on the acceptable disagreement (tolerance) with the experimental data.

The second subset, $R_{modular} \subset \mathrm{Sym}_n$, is the subset of matrices that fulfill the criterion of having the characteristic 'modular' structure typically seen in interaction score data. Clusters of interaction scores are frequently attributed to shared biological functions. Previously, this principle has been used to interpret interaction scores as functional modules, or make predictions of the function of particular targets or drugs [2,4,24,35]. Here, we instead aim to define a matrix algebraic constraint on X that ensures modularity. To define a set of symmetric matrices that have a modular structure, we apply the nuclear norm constraint

$$R_{modular} = \{X; \|X\|_\star < \varepsilon_m\}, \tag{9}$$

where $\|X\|_\star$ denotes the nuclear norm of X, defined as the sum of the singular values of X. In practice, constraining the nuclear norm of a matrix is used as a technique to constrain the rank of X [22]. A small value of ε_m thus implies few modules in the data. Obvious alternatives to this constraint would be e.g. the monochromaticity score by [24] and likelihood-based scores [36] or constraining the rank of X. However, the nuclear norm, which is a convex function of X, is very well suited for rapid optimization techniques [21,22].

The third subset, R_{sim}, contains all matrices consistent with prior pathway information. In contrast to the previous subset, $R_{modular}$, which contains *any* matrix with *any* modular structure, R_{sim} contains more specific information, i.e. it defines *a particular* modular structure defined by external data. We define this set by:

$$R_{sim} = \{\|X - KX\|_F^2 < \varepsilon_s\} \tag{10}$$

Here, K is a matrix that reflects the expected degree of similarity between rows and columns of X. To motivate this definition, consider an unknown interaction score x_{ij} and a linear interpolation function that predicts x_{ij} from any available "neighboring", scores x_{rj}, $(r,j) \in \Omega$. In other words,

$$\hat{x}_{ij} = \frac{\sum_r k_{ir} x_{rj}}{\sum_r k_{ir}} \tag{11}$$

here, k_{ir} is a non-negative weight that quantifies the functional similarity between target i and r. We organize these weights into a matrix format $K = \{k_{ir}\}$ and scale the rows/columns to sum to 1, i.e. K is a bistochastic matrix. We note that a fully observed X is consistent with the above kernel estimate if $X \approx KX$, i.e. when $\|X - KX\|_F^2$ is small, which motivates the definition of the set R_{sim} (eq. 10).

Functional similarity data. We explored which available data sources can be used to construct a matrix K with the property that $X \approx KX$. Here aiming for a heuristically defined K, we first defined K from different data sources, as the properly scaled matrix formed from (i) protein-protein interaction networks, (ii) co-expression networks, (iii) naïve GO term correlations; and, (iv) GO-term derived semantic scores [37–39] using 18 alternative tables from Yang et al. [40]. To gain insight about the usefulness of each data type as a prior to predict gene-gene interactions in yeast, we evaluated a total of 22 different K-matrices (listed in Supplementary Table 2 in File S1) using the metric

$$d_K = 1 - \|X - KX\| / \|X\|,$$

which will assume the value 0 if K fails to capture the contents of X and 1 if X is perfectly explained by K. (We remind the reader that K is a bistochastic matrix with zero diagonal, which excluded the trivial solution $K = I$, the identity matrix). Overall, the results show that PPI, GO term correlations and GO semantic scores were relatively equal in explanatory power (explaining up to 19% of X, Supplementary Table 2 in File S1) and while mRNA from one particular compendium were slightly less efficient. In the simulations below, we thus computed the average K for PPI (MINT) [41], mRNA, GO correlation and GO semantic data. Averaging was done using identical weights for each of the data types. The possibility of readjusting such weights during an ongoing screen, is reserved for future work.

For the glioma experiments, we defined five functional groups among the 31 drugs (Supplementary Table 1 in File S1). We thus defined $K_{ij} = 0$ when drug were in two different groups, and $K_{ij} = 1$ when they were in the same group. This matrix was subsequently scaled by bistochastic scaling.

Predicting interaction scores by cyclical projection onto convex sets. Our next task is to find an interaction matrix, which is located in the intersection of all three subsets in Sym_n, i.e. it fits the data ($X \in R_{data}$), it is modular ($X \in R_{modular}$); and, it is consistent with database information ($X \in R_{sim}$). In other words,

$$X \in R_{data} \cap R_{sim} \cap R_{modular} \tag{12}$$

There are highly efficient numerical methods to find the intersection of sets. Here, we find the solution by a cyclical sequence of projections, a method which has previously been applied to signal recovery and feasibility problems with multiple constraints [42–44].

Our algorithm starts with $X_0 = 0$ (a matrix with all entries equal to zero) and subsequently alternates between these three steps:

$$\text{For } t = 0, \ldots, T_{max} \begin{cases} X_{3t+1} \leftarrow \mathrm{proj}_{R_{data}}(X_{3t}, \varepsilon_d) \\ X_{3t+2} \leftarrow \mathrm{proj}_{R_{modular}}(X_{3t+1}, \varepsilon_m) \\ X_{3t+3} \leftarrow \mathrm{proj}_{R_{sim}}(X_{3t+2}, \varepsilon_s) \end{cases} \tag{13}$$

where $proj(\cdot,\cdot)$ denotes projection onto (or towards) the respective set (in the Frobenius norm). Each function in equation 13 thus maps a point X_q in the space of matrices to a new point X_{q+1}, which lies within the current set of interest R_{q+1} and also is at a minimal distance to the previous point X_q.

We cycle over projections until a converge criterion is met (see below). If the sets have a nonempty intersection, convergence is guaranteed by that fact that the three operations are cyclically applied projections onto convex sets [42].

By default, our method starts from an initial guess of a matrix of zeroes. To assess the robustness of the algorithm to differences in the initial guess in matrix space (i.e. the starting point in figure 2A), we performed a simulation in which *Saccharomyces cerevisiae* data with 80% missing values were imputed, using randomized matrices as X_0 (each matrix containing iid normally distributed random values with $\mu = 0$ and $\sigma = 100$). For each of the 100 simulations, the maximum deviation of 2 matrix elements between any two simulations was always less than 10^{-7}, with a mean of 3×10^{-9}, i.e. within the numerical precision (Figure S3 in File S1). This suggest that the performance of the algorithm is insensitive to the choice of initial condition.

Our method is a heuristic extension of previously described matrix completion algorithms Softimpute and SVT [21,22,26]. As a special case (relaxing constraints on modularity and functional similarity) our algorithm corresponds to the Softimpute algorithm [22]. The inclusion of functionality similarity is not a feature of this previous method, nor of the other methods considered in our comparison study (LLS [19] and EMDI [20]). Moreover, the general framework of cyclical projections we employ may have other extensions (e.g. by including additional convex set constraints for other data types), but this exploration is reserved for future work. All comparisons presented are based on Pearson correlation. However, using sum of squares prediction error did not alter the ranking of the tested methods.

Implementation

Our cyclical projection algorithm, explained above, starts with an empty interaction score matrix $X_0=0$ and subsequently applies three projection operations (onto the sets R_{data}, $R_{modular}$ and R_{sim}) to obtain a sequence of iterates $X_1,X_2,..$ until convergence, here defined as a small fractional change of X in terms of the Frobenius norm, i.e. $\|X_t-X_{t-1}\|_F^2/\|X_{t-1}\|_F^2<\delta$, with δ set to 0.0001.

For practical purposes, the projection functions (eq. 13) are not parameterized with the tolerance constants ε ($\varepsilon_d,\varepsilon_m,\varepsilon_s$) used to define the sets, but with penalties $\lambda>0$. This has no consequence for the solution of the problem, since for a given ε there is a λ which produces the same solution and vice versa [45]. We provide the derivation of the explicit projection formulae and parametrization using λ in the Supplement (Lemma 1–3 in File S1). The projection operations have the following computational forms:

$$\begin{aligned}\mathrm{proj}_{R_{data}}(X)&=X+\frac{\lambda_d}{1+\lambda_d}(M_\Omega-X_\Omega)\\ \mathrm{proj}_{R_{modular}}(X)&=st(X,\lambda_m)\\ \mathrm{proj}_{R_{sim}}(X)&=(I-((I-K)^2+I/\lambda_s)^{-1}(I-K)^2)X\end{aligned}\qquad(14)$$

Here, M is the experimental data, and the notation M_Ω denotes that M is only determined (observed) for a subset Ω of matrix elements, which correspond to the known interaction scores. I denotes the identity matrix and st($\cdot$,$\cdot$) is a soft-thresholding operation on the singular values of X, defined as $st(X,\lambda_m)=U\tilde{S}V^T$, where $X=USV^T$ is the singular value decomposition (SVD) of X, and $\tilde{S}$ is defined as $\tilde{s}_{ij}=0$ for $i\neq j$, and $\tilde{s}_{ii}=\max(0,s_{ii}-\lambda_m)$ [21,22].

We implemented this algorithm in MATLAB, using the PROPACK package to calculate the Singular Value Decompositions necessary for projection onto the set $R_{modular}$. We choose λ constants as follows. λ_d is kept constant at a default value of 10 (changing this value did not affect the results in a significant manner, although values close to zero should be avoided as $\lambda_d=0$ would imply that the experimental data are non-informative). λ_m,λ_s are chosen by five-fold cross-validation, in which 20% of data points in Ω are left out and predicted for a series of (λ_m,λ_s) pairs. We chose the pair that maximizes predictive power, measured by the Pearson correlation between observed and predicted values.

One should also note that for some choices of λ, the three R sets will become too small, and not overlap; in such cases the algorithm will instead converge onto a limit cycle, alternating between a limited number of solutions. In these cases, we recommend decreasing the λ values, alternatively using the average X over the three cycling steps as a solution that lies close to all sets [42].

In terms of algorithmic speed, the most time-demanding step is the calculation of the first few components of a singular value decomposition (SVD) of X (to project onto $R_{modular}$). The projections onto R_{data} and R_{sim} merely require matrix multiplications. The MATLAB implementation uses the PROPACK package to compute the SVD. As an example of a running time, the method requires 2 seconds to converge for a 500 matrix and about 10 minutes for a 4000 matrix. However, in cases where $n\gg10{,}000$ improved SVD methods are needed and we consider adding this to future versions of the implementation. The code is available from the authors upon request.

Screening via propensity-based sampling and interaction score prediction, Algorithms 1 and 2

Algorithm 1 outlines the screening strategy that incorporates prior knowledge or observed marginal interaction propensity for each gene or drug, i.e. how frequently an individual target is involved in a synergistic interaction. Algorithm 1 consists of iteration of steps 1 and 2 below. Algorithm 2 is the screening principle that incorporates both marginal propensity and interaction score prediction via matrix completion. Algorithm 2 is defined via steps 1 through 5. Two tuning parameters, n_1 and n_2, determine the number of experiments to perform in each step of the screen. While it is theoretically possible to step through the screen one experiment at a time, it is probably not the most practical strategy. As a default we used $n_1=n_2=1\%$ of all pairs.

Propensity-based sampling: steps 1 and 2.

1. Estimate probabilities $P_1,P_2,\ldots,P_n$ for the targets to have synergism with any other target (see main text for definition of Bayes estimates of P_i).

2. Perform n_1 experiments sampled from the untested fraction experiments, where the sampling probability for each pair is proportional to P_iP_j.

Prediction-driven screening: steps 3 and 4.

3. Use the matrix completion method defined by equations (14) to predict interaction scores X.

4. Pick the n_2 most extreme predicted interactions and test them by experiment.

Switching between the propensity-based and prediction-driven paradigms.

5. Estimate the hit rates H_2 for the n_2 most recent prediction-based experiments and H_1 for the n_1 previous random experiments (sampled as in step 2). If $H_2>H_1$ go to step 3. Otherwise go back to step 1. The hit rates are defined as the ratio between the number of identified synergies and the total number of experiments.

Supporting Information

File S1 Supporting figures and tables.
(PDF)

Author Contributions

Conceived and designed the experiments: SN. Performed the experiments: LS NM. Analyzed the data: SN PG RJ TK. Contributed reagents/materials/analysis tools: SN. Wrote the paper: PG LS NM TK RJ SN. Computational methodology: SN PG RJ. Glioblastoma experiments and data analysis: LS NM TK.

References

1. Zupan B, Demsar J, Bratko I, Juvan P, Halter JA, et al. (2003) Genepath: a system for automated construction of genetic networks from mutant data. Bioinformatics 19: 383–9.
2. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303: 808–13.
3. Boutros M, Brás LP, Huber W (2006) Analysis of cell-based rnai screens. Genome Biol 7: R66.
4. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, et al. (2010) The genetic landscape of a cell. Science 327: 425–31.
5. Battle A, Jonikas MC, Walter P, Weissman JS, Koller D (2010) Automated identification of pathways from quantitative genetic interaction data. Mol Syst Biol 6: 379.
6. Horn T, Sandmann T, Fischer B, Axelsson E, Huber W, et al. (2011) Mapping of signaling networks through synthetic genetic interaction analysis by rnai. Nat Methods 8: 341–6.
7. Axelsson E, Sandmann T, Horn T, Boutros M, Huber W, et al. (2011) Extracting quantitative genetic interaction phenotypes from matrix combinatorial rnai. BMC Bioinformatics 12: 342.
8. Chait R, Craney A, Kishony R (2007) Antibiotic interactions that select against resistance. Nature 446: 668–71.
9. Komarova NL, Wodarz D (2009) Combination therapies against chronic myeloid leukemia: shortterm versus long-term strategies. Cancer Res 69: 4904–10.
10. Cokol M, Chua HN, Tasan M, Mutlu B, Weinstein ZB, et al. (2011) Systematic exploration of synergistic drug pairs. Mol Syst Biol 7: 544.
11. Nelander S, Wang W, Nilsson B, She QB, Pratilas C, et al. (2008) Models from experiments: combinatorial drug perturbations of cancer cells. EMBO/Nature Molecular Systems Biology 4: 216.
12. Lehár J, Krueger A, Zimmermann G, Borisy A (2008) High-order combination effects and biological robustness. Mol Syst Biol 4: 215.
13. Lehár J, Krueger AS, Avery W, Heilbut AM, Johansen LM, et al. (2009) Synergistic drug combinations tend to improve therapeutically relevant selectivity. Nat Biotechnol 27: 659–66.
14. Zinner RG, Barrett BL, Popova E, Damien P, Volgin AY, et al. (2009) Algorithmic guided screening of drug combinations of arbitrary size for activity against cancer cells. Mol Cancer Ther 8: 521–32.
15. Lappe M, Holm L (2004) Unraveling protein interaction networks with near-optimal efficiency. Nat Biotechnol 22: 98–103.
16. Schwartz AS, Yu J, Gardenour KR, Finley RL Jr, Ideker T (2009) Cost-effective strategies for completing the interactome. Nat Meth 6: 55–61.
17. Wong S, Zhang L, Tong A, Li Z, Goldberg D, et al. (2004) Combining biological networks to predict genetic interactions. Proceedings of the National Academy of Sciences of the United States of America 101: 15682–15687.
18. Qi Y, Suhail Y, Lin YY, Boeke JD, Bader JS (2008) Finding friends and enemies in an enemiesonly network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. Genome Research 18: 1991–2004.
19. Colm R, Derek G (2010) Missing value imputation for epistatic maps. BMC Bioinformatics 11: 197.
20. Pan XY, Tian Y, Huang Y, Shen HB (2011) Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. Genomics 97: 257–64.
21. Cai JF, Candès EJ, Shen Z (2010) A singular value thresholding algorithm for matrix completion. SIAM J Optim 20: 1956–1982.
22. Mazumder R, Hastie T, Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res 11: 2287–2322.
23. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23: 561–566.
24. Segrè D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. Nat Genet 37: 77–83.
25. Oba S, Sato M, Takemasa I, Monden M, Matsubara K, et al. (2003) A bayesian missing value estimation method for gene expression profile data. Bioinformatics 19: 2088–96.
26. Candes EJ, Recht B (2008) Exact matrix completion via convex optimization. arXiv cs.IT.
27. Zhong W, Sternberg PW (2006) Genome-wide prediction of c. elegans genetic interactions. Science 311: 1481–4.
28. Spitzer M, Griffiths E, Blakely KM, Wildenhain J, Ejim L, et al. (2011) Cross-species discovery of syncretic drug combinations that potentiate the antifungal uconazole. Mol Syst Biol 7: 499.
29. Hu Z, Hu B, Collins JF (2007) Prediction of synergistic transcription factors by function conservation. Genome Biol 8: R257.
30. Jansen G, Lee AY, Epp E, Fredette A, Surprenant J, et al. (2009) Chemogenomic profiling predicts antifungal synergies. Mol Syst Biol 5: 338.
31. Vatcheva I, Jong HD, Mars NJI (2000) Selection of perturbation experiments for model discrimination.
32. King R, Whelan K, Jones F, Reiser P, Bryant C, et al. (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. Nature.
33. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 6: 343.
34. Hahn GJ, Shapiro S (1994) Statistical Models in Engineering. Hoboken, NJ: John Wiley & Sons, 95 pp.
35. Yeh P, Tschumi AI, Kishony R (2006) Functional classification of drugs by properties of their pairwise interactions. Nat Genet 38: 489–94.
36. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23: 561–6.
37. Schlicker A, Domingues FS, Rahnenfhrer J, Lengauer T (2006) A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics 7: 302.
38. Frhlich H, Speer N, Poustka A, Beissbarth T (2007) Gosim-an r-package for computation of information theoretic go similarities between terms and gene products. BMC Bioinformatics 8: 166.
39. Couto FM, Silva MJ, Coutinho PM (2007) Measuring semantic similarity between gene ontology terms. Data & Knowledge Engineering 61: 137–152.
40. Yang H, Nepusz T, Paccanaro A (2012) Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics 28: 1383–1389.
41. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) Mint: the molecular interaction database. Nucleic Acids Res 35: D572–D574.
42. Combettes P (1993) The foundations of set theoretic estimation. Proceedings of the IEEE 81: 182–208.
43. Serbes A, Durak L (2010) Optimum signal and image recovery by the method of alternating projections in fractional fourier domains. Communications in Nonlinear Science and Numerical Simulation 15: 675–689.
44. Bauschke H, Borwein J (1996) On projection algorithms for solving convex feasibility problems. Siam Review 38: 367–426.
45. Osborne MR, Presnell B, Turlach BA (2000) On the lasso and its dual. J Comput Graph Statist 9: 319–337.