CHALMERS

Efficient experimental screening: epistatic interaction identification using information theory

Master's Thesis in Bioinformatics and Systems Biology

Mariana Buongermino Pereira

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2012

Efficient experimental screening:
epistatic interaction identification using information theory

Mariana Buongermino Pereira

Department of Mathematical Sciences
Chalmers University of Technology
SE - 412 96 Gothenburg, Sweden
Telephone: +46 (0)31 - 772 10 00

Thesis for the Degree of Master of Science

Efficient experimental screening:
epistatic interaction identification using information theory

MARIANA BUONGERMINO PEREIRA

Department of Mathematical Sciences
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden 2012

Mathematical Sciences
Gothenburg, 2012

Preface

This report is a Master's Thesis in Bioinformatics and Systems Biology and it is a part of Master's Programme at Chalmers University of Technology and University of Gothenburg.

The research project has been carried out at Sahlgrenska Cancer Center, part of the Sahlgrenska Academy at University of Gothenburg.

Supervisor:
Sven Nelander
Sahlgrenska Academy
University of Gothenburg
Gothenburg, Sweden

Examiner:
Olle Nerman
Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden

Abstract

Screening for epistatic interactions is relevant in several contexts such as evolution, genetics and cancer treatment. The present work uses information theory heuristics to increase the efficiency of such screens. The problem is mathematically described as a $n \times n$ sparse matrix, where n is the number of genes or drugs of interest, $1's$ correspond to the epistatic interactions and $0's$ to the non-epistatic ones. Based on the assumption that certain nodes (genes or drugs) are hubs in a network, the approximate conditional probabilities of a node to interact with any other are approximate by a function of the amount of synergies found for that node among few initial observations. These probabilities are updated according to the observed epistatic interactions for each gene or drug. Considering the change in the approximate updated probabilities if one new pair is tested, the expected information gain for each possible pair is calculated. Finally, the pair with maximum average expected information gain is experimentally tested and probabilities are updated. The algorithm proposed here improves screening efficiency over random screening procedures up to sixfold.

Keywords: Information Theory, Cancer Treatment, Epistasis, Synergistic Interactions, Bayesian Experimental Planning.

Acknowledgements

First of all, I thank my parents for supporting my decisions and encouraging me to study what I was really passionate about, since, in their words: “who does what loves is always going to be successful”.

I thank Mikael for being a good friend!

I thank Sven Nelander for his supervision and the interesting topic for my thesis. Rebecka Jörnsten for her inputs. Linnea, Teresia and Caroline at the Sahlgrenska Cancer Center for nice company during the spring.

A special thank to Philip Gerlee for his careful proof reading, ideas and suggestions. To Olle Nerman for his wonderful guidance during the whole programme. And, to Erik Kristiansson for his great advices.

Mariana Buongermino Pereira
Gothenburg, 2012.06.04

Contents

List of Figures

List of Tables

Outline

This thesis is divided into 5 chapters.

In Chapter 1, Introduction, the biological background of the problem and motivation for writing this thesis are described along with concepts from information theory that are used to solve the biological question. Also a performance measure is introduced.

In Chapter 2, Methods, the mathematical foundation introduced in the Chapter 1 is applied the problem and two algorithms are developed.

Chapter 3 presents the Results and Discussion comparing the two algorithms performance, followed by Chapters 4 and 5 with Conclusion and suggestion for Future Work, respectively.

1 Introduction

1.1 Biological Background

In 1909, the term *epistasis* was first used by Bateson to describe how one allele at a certain locus could interfere with the expression of another allele at another locus [1]. This qualitative definition has also been used to describe proteins whose action is affected by variation in the amount of another protein [2]. In 1918, Fisher further developed the *epistasis* concept, formally describing it as the deviation from the additive combination of each locus [3]. Such a definition is commonly found in quantitative genetics [2, 4], to describe the fitness of mutants. In a similar way, *epistasis* can describe the gain or loss in fitness caused by the action of a drug pair in an individual, compared to the action of the single agents.

Even though the model proposed by Fisher was an additive model, the multiplicative model of interaction is commonly used since it is a relative measure that is adequat to rule out experimental variation and is more sensitive to detect small values of interaction. Thus, epistasis ε_{ij} between two genes (or drugs) i and j is given by [4, 5]:

$$\varepsilon_{ij} = W_{ij} - W_i W_j \tag{1}$$

where W_i and W_j are phenotype changes after a single perturbation and W_{ij} is phenotype after the two genes i and j have been mutated or the drugs i and j have perturbed the system. Note that if the phenotypes are nonnegative one may transform the multiplicative definition to an additive one by logarithmic transformation.

When $\varepsilon_{ij} = 0$ there is no epistatic effect between two mutations or drugs. When $\varepsilon_{ij} \neq 0$, the interaction effect can be positive ($\varepsilon_{ij} > 0$) or negative ($\varepsilon_{ij} < 0$). Positive interactions, also called buffering or antagonistic interactions, appear redundant in the system, but they can slow down resistance acquisition in the case of drug interaction. Negative or synergistic interaction, on the other hand, signifies a *synergy* between the two drugs in such a way that they intensify the effects of the mutations or drugs [4, 5]. Note that the very definition of the phenotype can vary so that the interpretations are opposite. As an example consider optimal growth rate of a cell culture and the related culture size doubling time under optimal conditions. Small relative growth rates correspond to long doubling times.

1.2 Motivation

Epistatic interactions are important in both evolution and genetics. In evolution, they can explain speciation and the emergence of different mechanism in living beings. In genetics, in turn, they reveal the mechanism and function of gene associations [5]. Similarly, in complex diseases, such as cancer, synergistic drug pair can increase the treatment effectiveness. Therefore, identifying epistasis has motivated several screening processes that aim to characterize epistasis values for a large number of gene mutations or drug pairs. Due to the large size of gene or drug library ($n > 100,000$), exhaustive screening of all possible pair is often unfeasible. Consequently, algorithms able to efficiently recognize epistatic pairs are needed.

In this context, the problem can be described as a matrix $\mathbf{S} \in \{0,1\}^{n \times n}$, where the entries S_{ij} are defined as epistatic, 1, or non-epistatic, 0. Thus, it is relevant to define the efficiency with which we can find epistatic pairs in the matrix $\mathbf{S}$. This can be achieved by relating the fraction of found epistatic pairs f to the fraction of screened pairs z, as a function $f(z)$. For a brute-force system, where the fraction of discovered pairs is identical to the screened fraction, the discovery rate f' can be defined as $f' = (1-f)/(1-z)$, i.e. the remaining epistatic pairs fraction $(1-f)$ divided by the remaining screenable fraction $(1-z)$. Similarly, a more efficient method would have a discovery rate given by $f' = \tau(1-f)/(1-z)$, where $\tau > 1$ and the fraction of found pairs is higher than the screened fraction. Solving the differential equation with boundary conditions $f(0) = 0$ and $f(1) = 1$, i.e. respectively no interactions are found when no pairs are tested and all pairs are found when all pairs have been screened, we have:

$$f(z) = 1 - (1-z)^{\tau} \tag{2}$$

where τ is defined as *screening efficiency*.

Equation 2 tell us that in a brute-force screening procedure, for any percentage of screened pair, the same percentage of epistases is found and $\tau = 1$. When epistases are efficiently found, the percentage of found epistatic pairs is higher than the screened pair percentage, resulting in $\tau > 1$ (Figure 1).

Observe that we did not argue in any strict way that a specific method necessarily must satisfy a differential equation of the kind above.

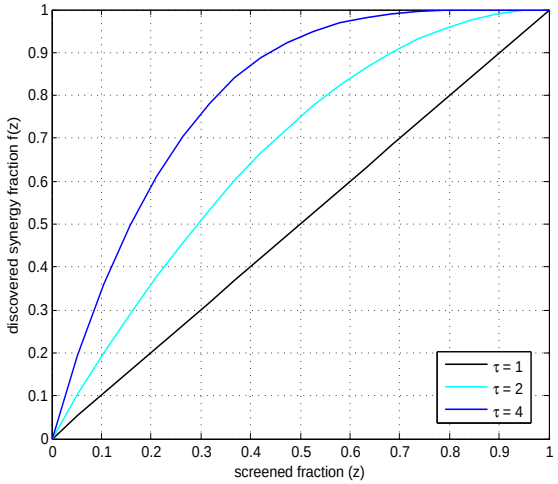

Figure 1: Screening efficiency: The plot of discovered epistatic pairs fraction f as function of screened pairs fraction z shows the faster epistases are found the higher the screening efficiency τ value is.

1.3 Aims

In 2012, Gerlee *et al.* [6] proposed an algorithm to efficiently screen matrix $\mathbf{S}$ in order to maximize screening efficiency τ. Their algorithm alternates between two strategies: i) propensity-based sampling, which exploits the fact that some genes (or drugs), also known as hubs, have a higher probability of interacting than others and ii) prediction-based screening, which combines experimental data with knowledge from bioinformatics databases and exploits the modular nature of the data to predict which drugs are more likely to present a epistatic interaction and therefore be selected to be experimentally tested.

The present work aims to improve the first part of Gerlee's algorithm and find a method for efficiently deciding on which pair to test. In other words, part i) in Gerlee's algorithm test the drug pair with maximum probability of being synergistic (or antagonistic). However, the problem of searching for epistatic interactions alternates between trying pairs that are very likely to present epistatic interactions and exploring interactions with less available evidence of synergy (or antagonism). Thus, the algorithm developed during the present Master's project searches for interactions whose values provide the maximum information necessary to generate a more accurate prediction in the next round. In short, we suggest choosing the experiment, i.e. pair (i, j), which maximizes the *average expected information gain*, as measured by the sum of the entropy of heuristically updated probabilities in the experimental cell and the sum of entropy changes in all cells with any one of the two experimental conditions i or j involved. The ideas are first that the entropy of the probabilities in a cell measures how well one can predict that particular cell specific interaction

status, and second that the main information gain from an experiment mainly concerns the interactions included in this sum. Alternatively, one could consider minimizing the entropy of the whole system, but we are not tracking the multivariate distributions in our heuristic algorithm, and moreover for the prediction of a single but random remaining experiment the sum of the marginal entropies should be more relevant.

1.4 Information Theory

The concept of information theory was first introduced by Fisher in 1925 [7], however it was Shannon in 1948 [8] who generalized the concept and established it the way it is used today. Information theory is commonly used in communication engineering as a way to predict the messages being conveyed based on the probabilities of each character in the alphabet in a certain language. However, information theory is a branch of mathematical probability and statistics and therefore has applications beyond communication theory [9] as will be evident in the present work.

1.4.1 Definitions

The information I provided by an event X is the inverse of the logarithm of the event's probability $P(X)$:

$$I(X) = \log_2 \frac{1}{P(X)} = -\log_2 P(X) \tag{3}$$

where $\log_2$ is used in order to give the information in bits. From now on, the index 2 will be omitted to simplify the notation. Intuitively, Equation 3 indicates that an event that is less likely to occur provides us with more information.

Now, if X is a random variable that can assume k possible states, so that $X = \{x_1, x_2, ..., x_k\}$ with probabilities $P = \{p_1, p_2, ..., p_k\}$, where P is the probability distribution over all possible states and $p_1 = p(x_1)$, the *entropy* of X can be defined as the expected information gained if we learn the value of X:

$$H(X) = -\sum_{i=1}^{k} p_i \log p_i \tag{4}$$

$H(X)$ is usually called the *Shannon entropy*, and can also be interpreted to quantify our lack of knowledge about the variable X.

In order to better understand the entropy concept, let us consider the case in which X

follows a Bernoulli distribution, i.e. $X = \{0, 1\}$ and with probabilities $P = \{1 - p, p\}$ respectively. Clearly, $k = 2$ and Equation 4 can be written as:

$$H(X) = -(p \log p + (1 - p) \log(1 - p))$$

Figure 2 shows the entropy H as function of probabilities p for the Bernoulli case. There, it can be seen that when the two possible outcomes, 0 and 1, have the same probability, i.e. $p = 0.5$, the entropy is maximum. In other words, when $p = 0.5$ both outcomes are equally likely and, thus, the outcome uncertainty is maximal (this situation is analogue to a fair coin for which heads and tails provide maximum information once observed). However, the other extreme situations happen when $p = 1$ or $p = 0$ (analogue to the case of a completely unfair coin) zero information is gained when the system is observed, given that there is no uncertainty.

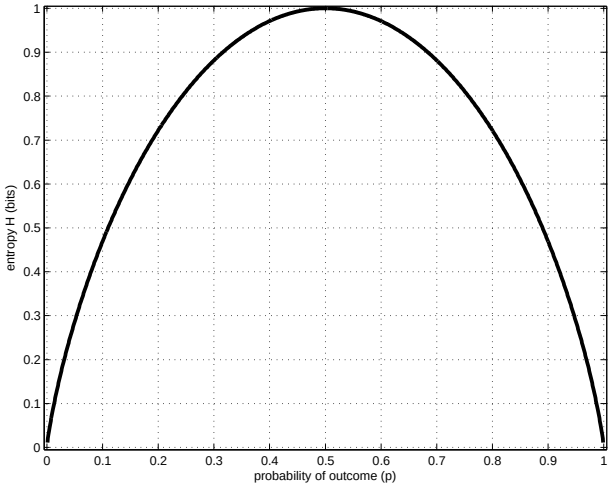

Figure 2: Entropy for a Bernoulli random variable: for a system with only two possible outcomes, entropy is maximum when uncertainty is maximum, i.e. both outcomes are equally likely, i.e. $p_i = 0.5$.

When an experiment E is performed on a system, the information gained decreases our *lack of knowledge* of the system. We can interpret this change in the knowledge about the system as a modification of the probability distribution $P(X)$ that is assigned to the system. The observation E will change the prior distribution $P(X)$ into a posterior distribution $P(X|E)$. The information gain G for such observation E, also called Kullback-Leibler divergence [10], is defined as:

$$G[P(X), P(X|E)] = \sum_{i=1}^{k} p(x_i|E) \log_2 \frac{p(x_i|E)}{p(x_i)} \tag{5}$$

In other words, the information gain G is a measure of what is learnt about the probability distribution X when an experiment E is performed. G represents, therefore, the *divergence* between the two probability distributions $P(X)$ and $P(X|E)$ in terms of the available information on the variable X. Please note that $G \geq 0$, with equality when $P(X) = P(X|E)$.

2 Methods

2.1 A Mathematical Description of the Problem

The level of interaction within a set of genes or drugs can be represented by a matrix $\mathbf{I} \in \mathbb{R}^{n \times n}$, where n is the number of genes or drugs considered. Each matrix entry is the epistasis ε_{ij}, given by Equation 1. Often, it is interesting to identify strong synergistic pairs whose epistasis is found below a threshold λ, i.e. $\varepsilon_{ij} \leq \lambda < 0$. Similarly, we could think about the problem when we are interested in identify relevant antagonistic effects, defined by $0 < \lambda < \varepsilon_{ij}$. Here we are interested in synergistic drug pair's identification and hence we define a binary synergy matrix $\mathbf{S} \in \{0, 1\}^{n \times n}$ whose entries S_{ij} are defined as synergistic, 1, or non-synergistic, 0, according to:

$$S_{ij} = \begin{cases} 1 & \text{for} \quad \varepsilon_{ij} \leq \lambda \\ 0 & \text{for} \quad \varepsilon_{ij} > \lambda \end{cases} \tag{6}$$

Since synergistic pairs are rare, our problem can be mathematically described as finding $1's$ in a sparse matrix. Figure 3 displays the situation: a 31×31 matrix, which is symmetric allowing the bottom part to be ignored (shown in grey). The black cells indicate synergistic pairs, i.e. cells whose entries are $1's$, whereas white cells represent the non-synergistic pairs, i.e. entries that are 0's. Also, the diagonal is not considered, since a drug, by definition, cannot interact with itself to have a synergistic effect [11].

2.1.1 Synergy probabilities

Here, we are interested in finding $1's$ in sparse matrices as the one presented in Figure 3. For that, we are going to use a heuristic Bayesian experimental design approach [10, 12, 13, 14], and therefore it is necessary to define the probabilities with which each matrix entry contains a synergistic value.

2.1.1.1 Preliminary considerations Consider a statistical model where to each drug or gene, i, there belongs a certain probability parameter $P(i)$. Given these ***P(i)-parameters*** the probability of having an interaction between drug i and j is assumed to be $P(i) * P(j)$. Now, we may also imagine a single fix distribution from which the $P(i)$´s are initially drawn independently, and assume that we have so far observed whether we have synergies for a certain subset of all the possible drug pair combinations.

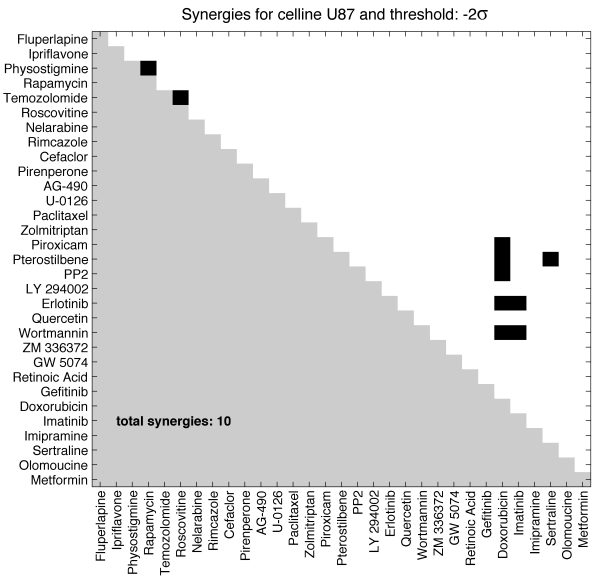

Figure 3: Drug-drug synergy matrix $\mathbf{S}$: Epistasis values converted to synergistic or non-synergistic entries, i.e. 1 or 0, respectively. Data for 31 drugs, indicated on the axes, tested on the glioblastoma cell line U-87MG. The matrix is symmetric, so that the lower part is not taken into account, and neither is the diagonal (grey zone in the plot), since a drug does not have a synergistic effect with itself. Black cells are the synergistic pairs, i.e. entries whose values are 1, and the white cells are non-synergistic pairs, i.e. cells with $0's$. In total, out of the 465 ($n(n-1)/2$) possible pairs there are only 10 that are synergistic.

In principle one may then use Bayes formula to update the full distribution of all the $P(i)$, and the probabilities of all the so far untested synergy indicators. Next we can, in principle, calculate the expected sum of marginal cell-wise information gain from observing the real synergy status in any particular single cell, by conditioning on the outcome and summing over all conditional entropy changes (differences) in the so far untested cells. These potential expected gains can then be compared for the different possible single experiments and the best combination i, j, i.e. the combination with highest expected gain could be chosen as the next experiment.

However the sketched approach is quite computationally hard because dependencies are quickly establishing between all the marginal distributions of the $P(i)$´s and all the indicators. Thus, we have instead considered an heuristic updating procedure for the cell probabilities that is not a real Bayesian one, but inspired by Bayes fomula for a simple binomial experiment with Beta-distributed a priori distribution for the parameters $P(i)$. If the $P(i)$'s does not vary much the resulting estimate we have used.

The following equation:

$$P(i) \approx \frac{X(i) + \alpha}{N(i) + \alpha + \beta}$$

where $X(i)$ is the number of 1's found for each i when $N(i)$ interactions have been tested for this i, should be a reasonable estimate of $P(i) * \overline{P}$ rather than of $P(i)$, which thus should be higher. (See Section 2.1.1.2 for more information). Unfortunately we missed the second factor in all our implementations, and moreover the fact that the experiments are selected to be especially informative make the pairing at least initially quite atypical so that the approach is anyhow quite questionable.

The result is under all circumstances that all our estimated $P(i)$'s used in the optimization are too small! A quick, but non-rigorous, way out would be to simply modify the estimates of $P(i)$ with a common scale factor C so that the relative frequency of synergies so far matches the expected in the natural but in fact also approximate formula using the sum over cells tested so far of $P(i) * P(j) * C^2$.

The *C-factor* will thus vary, but slowly, with the relative frequency at least after some initial steps. But observe that the *C-factor* has an impact on the entropy in all the untested cells. It remains to investigate whether this arguing will have any positive impact on the performance of the experimental selection procedure.

Using computer-intensive techniques it should of course be possible to study full Bayesian rigorous information theory based approaches, along the lines in this paper. At least this should work in sparse synergy systems, a topic that may be studied in the future.

2.1.1.2 Our approximate method In a network, certain nodes (in our case certain drugs) are often more likely to interact. Drugs with this behaviour are called hubs of the system. Based on this assumption, we can define $P(i)$ as the probability of drug i interact with any other drug. These probabilities can be defined as the approximate conditional probability as explained in Section 2.1.1.1:

$$P(i) = \frac{X(i) + \alpha}{N(i) + \alpha + \beta} \tag{7}$$

where $X(i)$ is the number of synergies found for drug i when $N(i)$ interactions have been tested for this drug. Equation 7 correspond to the mean of a posterior distribution of $P(i)$,

where the prior is the mean of beta-distribution with parameters α and β. In other words, we assume that *a priori* all drugs have the same probability $\frac{\alpha}{\alpha+\beta}$ to synergistically interact. Here we use $\alpha = 0.5$ and $\beta = n$ as suggested by Gerlee *et al.* [6].

Consequently, the probability $P((i,j) = 1)$ of a pair (i,j), being a synergistic pair is given by:

$$P(ij = 1) = P(i) * P(j) \tag{8}$$

where $P(i)$ and $P(j)$ are the probabilities of each drug interact synergistically (Equation 7). Note that, in order to simplify the notation, $P((i,j) = 1)$ and (i,j) are represented by $P(ij = 1)$ and ij, respectively.

2.1.2 Maximum Average Expected Gain of Information Criteria

Our task is to find $1's$ in an efficient way, as defined by Equation 2. If we imagine that initially we have few observations of epistasis values, $P(i)$ can be calculated using Equation 7. In turn, the synergy probabilities $P(ij = 1)$ can be calculated according to Equation 8. However, we still need to decide which drug pairs to test next. Clearly, we could consider the pair whose synergy probability is highest. However, this approach will quickly trap the search procedure around the same area, even when all the $1's$ there have already been found. In other words, this approach can overlook synergistic pairs whose drugs have little or no information available (few interactions tested). Therefore, our decision criteria has to explore areas of the matrix for which little is known. In this context, the information gain (Equation 5) seems an appropriate measure.

The information provided by an experiment decreases our lack of knowledge of the system. In our case, the system is the matrix $\mathbf{S} \in \{0,1\}^{n \times n}$, where n is the number of drugs (or genes) whose interactions are the entries defined as 1, when there is a synergistic effect, or 0 for non-synergistic interaction (Equation 6). Our knowledge about the system is given by probabilities $P(ij = 1)$ (Equation 8) which can be described as a matrix $\mathbf{P} \in [0,1]^{n \times n}$. However, $P(ij = 1)$ is given by the product of the approximate conditional probabilities $P(i) * P(j)$, which are the ones to be in fact updated when an experiment is performed. Therefore, when a drug pair is tested we can define the information gain G (Equation 5) to express the average information given by the marginal distributions instead of the exact value for each possible matrix synergy matrix $\mathbf{S}$. Thus G can be written as:

$$G[\mathcal{P}, \mathcal{P}^{ab}] = \sum_{k=0}^{1} \sum_{ij} P(ij = k|ab) \log \frac{P(ij = k|ab)}{P(ij = k)} \tag{9}$$

where the first sum runs over all the possible states each pair can assume, i.e. 0 for a non-synergistic pair and 1 for a synergistic pair, the second sum runs over all pairs in the matrix, and ab is the pair that is tested on the experiment. The information gain G is the relative information between the prior and posterior distributions $\mathcal{P}$ and $\mathcal{P}^{ab}$, respectively, for the probabilities before and after one pair ab is experimentally tested. In other words, G is the information provided by the change in the probabilities when a pair ab is tested and yields an average information across all the entries in the system.

When we simulate experimental procedure, each experiment tests one drug pair ab and returns its synergistic interaction value, $S_{ab} = 0$ or $S_{ab} = 1$, with associated probabilities $P(ab = 0)$ and $P(ab = 1)$, respectively. Each possible outcome, $S_{ab} = 0$ or $S_{ab} = 1$, has an associated information gain, $G(ab = 0)$ and $G(ab = 1)$, both given by Equation 9. Therefore, the average expected information gain $E[G_{ab}]$ for each possible experiment ab is approximated by:

$$\begin{aligned} E[G_{ab}] &= P(ab = 1) * G(ab = 1) + P(ab = 0) * G(ab = 0) = \\ &= P(ab = 1) * \sum_{k=0}^{1} \sum_{ij} P(ij = k|ab = 1) \log_2 \frac{P(ij = k|ab = 1)}{P(ij = k)} \\ &+ P(ab = 0) * \sum_{k=0}^{1} \sum_{ij} P(ij = k|ab = 0) \log_2 \frac{P(ij = k|ab = 0)}{P(ij = k)} \end{aligned} \tag{10}$$

The algorithm picks the pair ab with maximum $E[G_{ab}]$ to be experimentally tested.

2.1.3 Summary

Figure 4 shows an overview of the data used in the mathematical description of the problem. It starts with the experimental data, which is the epistasis ε_{ij} between two drugs i and j, organized in a matrix $\mathbf{I} \in \mathbb{R}^{n \times n}$. From this data the epistasis standard deviation σ is calculated and used to establish the synergy threshold λ, below which synergy is taken as existent and above as non-existent. Thus, λ is used to convert matrix $\mathbf{I}$ into a binary matrix $\mathbf{S} \in \{0, 1\}^{n \times n}$. The number of tested pairs N_i and synergies found X_i for each drug i is computed and used to calculate the approximate conditional probability $P(i)$ of each drug

interact in a synergistic way (Equation 7). These data is then used to calculate the probabilities $P(ij = 1)$ of each pair ij yielding a synergistic interaction (Equation 8). Finally, experiments are simulated and the one with maximum average expected information gain (Equation 10) is select in the experimental design as the next one to be tested.

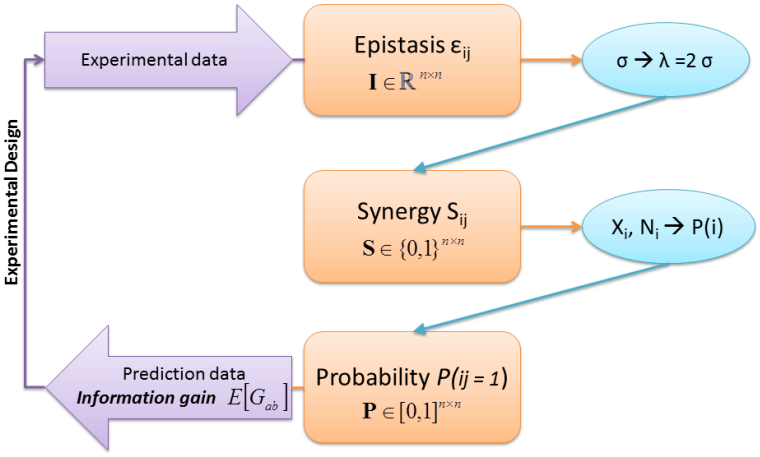

Figure 4: Overview of the mathematical description of the problem: a sketch on how experimental data is used to make predictions (See text on Section 2.1.3).

2.2 Implementation

In order to make the calculations faster, it was necessary to simplify Equation 10. For that, we considered the fact that our experiment outcome follows a Bernoulli distribution, where:

$$P(ij = 0) = 1 - P(ij = 1), \forall ij.$$

Using the above relationship and a Taylor approximation, Equation 10 can be simplified to (Appendix 5):

$$\begin{aligned} E[G_{ab}] =& P(ab = 1) * \sum_{ij} \left[\frac{(P(ij = 1|ab = 1) - P(ij = 1))^2}{P(ij = 1).(1 - P(ij = 1))} \right] \\ & P(ab = 0) * \sum_{ij} \left[\frac{(P(ij = 1|ab = 0) - P(ij = 1))^2}{P(ij = 1).(1 - P(ij = 1))} \right] \end{aligned} \tag{11}$$

2.2.1 Basic Algorithm

In order to optimize the fraction of synergies found versus the fraction of screened pairs, a basic algorithm was implement to predict the next *single* drug pair to be tested (Figure

5 and Algorithm 1). In this algorithm, for every non-tested pair, the average expected information gain is calculated according to Equation 11 and the pair with maximum average expected information gain is tested. Values of synergies found $X(i)$ and interactions tested $N(i)$ for each drug are updated as well as the approximate conditional probability $P(i)$ and the synergy probabilities $P(ij = 1)$ and again the non-tested pairs have their expected information evaluated until the matrix has been completed.

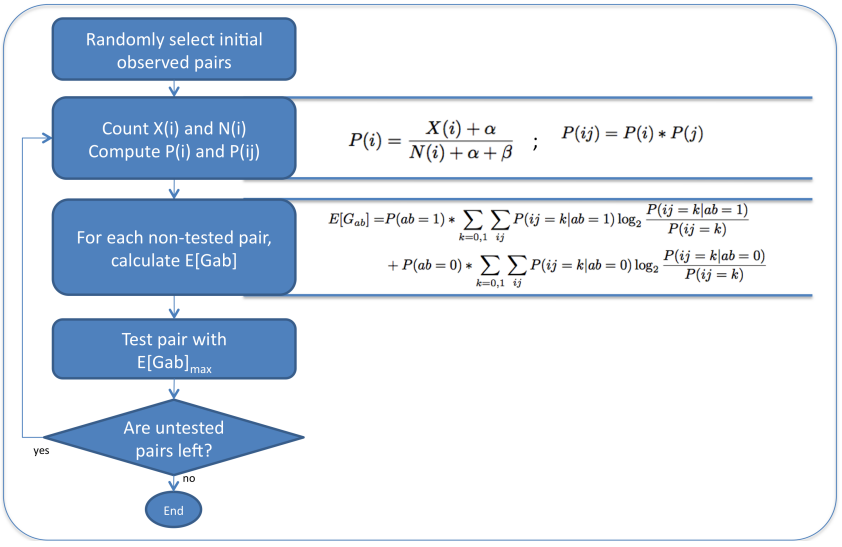

Figure 5: Basic algorithm: for the selection of a *single* drug (or gene) pair to be tested next.

Algorithm 1 Basic algorithm

1: Start with few known interactions.
2: **for all** drugs i **do**
3: Compute the number of known synergies $X(i)$ and tested interactions $N(i)$.
4: Calculate the approximate conditional probability of synergistic interaction $P(i)$ (**Equation 7**).
5: **end for**
6: **for all** drug pairs ij **do**
7: Calculate probability $P(ij = 1)$ of resulting in a synergistic interaction (**Equation 8**).
8: **end for**
9: **for all** non-tested drug pairs ab **do**
10: Simulate the impact a 1 or a 0 found in cell ab have in all $P(i)$ by calculating the average expected information gain (**Equation 11**).
11: **end for**.
12: Test the pair ab with $E[G(ab)]_{max}$ and add their synergistic values (0 or 1) to the matrix.
13: Repeat steps 2 to 12 until the entire matrix has been screened.
14: Calculate screening efficiency τ fitting data to **Equation 2**.

2.2.2 Combinatorial Algorithm

In real life, many times several drug pairs are tested at a time in the lab. Therefore, the basic algorithm was modified in order to output a *set* of drug pairs to be simultaneously tested instead of one at a time. Our *combinatorial algorithm* simply listed the top 25% of pairs with highest average expected information gain (Equation 11). The highest of this list was chosen as the first pair to be tested. Next, the average expected information gain was calculated assuming the first pair was simultaneously tested with each of the other pairs in the list. This gives us a list with a combination of two pairs to be tested with respective information gain. Again the two pairs that together provide the maximum average expected information gain are selected. Then, each of the other pairs in the top list is simultaneously tried with the two first pairs, and again the set of three pairs with maximum average expected information gain is selected. The process is repeated until m pairs have been selected. Those pairs are then experimentally tested, the approximate conditional probability $P(i)$ are updated and a new set of m pairs is picked based on the maximum average expected information gain. The process is repeated until the entire matrix is screened (Figure 6 and Algorithm 2).

2.2.3 Evaluation on Real Data

Algorithm 1 and 2 were tested on data from drug-drug interaction on 5 different glioblastoma cell lines: A172, T98G, U-87MG, U-343MG, U-373MG[1]. For each cell line, 31 drugs have their epistasis tested with each one of the other drugs (Equation 1). Epistasis was measured using Alamar blue assay and calculated from relative viability after 48 hours of exposure to single drug and drug pairs. For further details on the experimental procedure please refer to [15].

In order to define if the interaction score was synergistic or not, i.e. converting the continuous epistasis values ε_{ij} from matrix $\mathbf{I} \in \mathbb{R}^{n \times n}$ into a discrete matrix $\mathbf{S} \in \{0,1\}^{n \times n}$, three criteria had to be accomplished. First, epistasis value should be lower than a threshold $\lambda = -2\sigma$, where σ is standard deviation of epistasis for all drug pair in all cell lines. Thus, first condition can be stated as:

$$\varepsilon_{ij} = W_{ij} - W_i W_j < -2\sigma \tag{12}$$

Second, the ratio between the drug-drug effect and the control should be smaller than 0.8. The control is the measure of cell growth after the same amount of time but without the use of any drug. Since in cancer treatment drugs aim to stop cell growth, the smaller the

[1]T98G was obtained from ATCC and A172, U-343MG, U-373MG and U-87MG were obtained from Cell Lines Services, Germany.

Algorithm 2 Combinatorial algorithm

Start with few known interactions.
for all drugs i **do**
    Compute the number of known synergies $X(i)$ and tested interactions $N(i)$.
    Calculate the probabilities of synergistic interaction $P(i)$ (**Equation 7**).
end for
for all drug pairs ij **do**
    Calculate probability $P(ij = 1)$ of resulting in a synergistic interaction (**Equation 8**).
end for
for all non-tested drug pairs ab **do**
    Simulate the impact a 1 or a 0 found in cell ab have **in all** $P(i)$ **by calculating the average expected information gain for the pair** ab (**Equation 11**).
end for
Include the pair ab with $E[G(ab)]_{max}$ in the **set Y** of selected pairs.
Set **bound B** to be maximum expected information gain.
Set **candidates C** as x% top list of pairs with highest expected information gain.
for m best pairs **do**
    for all candidate C_i **do**
        Calculate $E[G(Y, C_i)]$ for the case when $\{\mathbf{Y}, C_i\}$ are tested simultaneously.
        if $E[G(Y, C_i)]$ > **B then**
            B = $E[G(Y, C_i)]$
            solution = C_i
        end if
    end for
    Include **solution** in the set **Y** of best pairs.
    Exclude **solution** from candidates **C**.
end for
Test the set **Y** of best pairs and add their synergistic values (0 or 1) the matrix.
Repeat steps 2 to 26 until the entire matrix has been screened.
Calculate screening efficiency τ fitting data to **Equation 2**.

ratio the better. In Equation 12 W_{ij} is given by the ration between the cell growth under the effect of drugs i and j V_{ij}, and the cell growth without any drug $V_{control}$. The second condition is then:

$$W_{ij} = \frac{V_{ij}}{V_{control}} < 0.8 \tag{13}$$

Third, the growth V_{ij} and $V_{control}$ were replicated 3 times and its means were tested to be statistically different using a two-sample t-test and synergy was accepted if *p-value* was smaller than 0.01. Accordingly, a drug pair that would fulfil all these three criteria was assumed to be synergistic, i.e. value set to 1, otherwise it was set to 0 (non-synergistic).

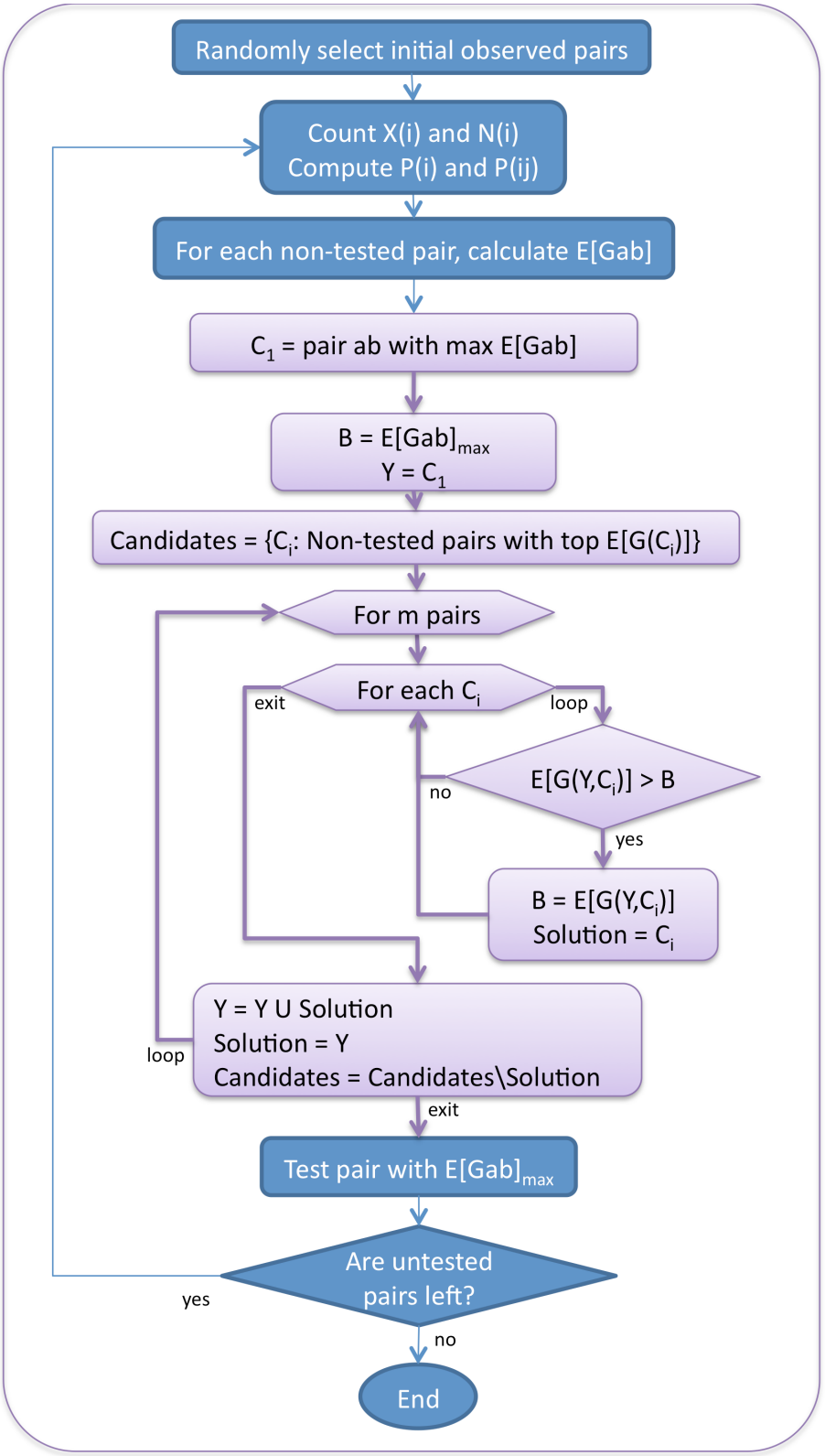

Figure 6: Combinatorial algorithm: for the selection of a *set* of m drugs (or genes) pair to be simultaneously tested. The blue parts are the same as in the basic algorithm (Figure 5).

Since this synergy definition is not standardized, we varied the threshold λ in a range of $[-2.4\sigma, -1.5\sigma]$. This analysis provided, for each cell line, one matrix for each λ with different amount of $1's$ to be found. Also, the percentage of initial points in the matrix was varied for 9 points in the interval $[0.01, 0.20]$.

Screening efficiency was calculated pooling together data from 1000 or 500 simulations, for basic and combinatorial algorithm respectively, for each cell line and each λ and fitting the results using Equation 2.

3 Results and Discussion

As explained in Section 2.1.1.1, our results were obtained using an heuristic method that approximated the conditional probabilities to Equation 7 without taking in consideration the *C-factor*. This imply that $P(i)$'s used were underestimated. Since we just observed this in the end of the work and the results were satisfactory, as it will be shown, we present the work without the *C-factor*.

The standard deviation for all epistasis values ε_{ij} across all cell lines was $\sigma = 0.0988$.

For the experimental procedures, synergy threshold λ was considered -2σ. However, due to the lack of a standard on the synergy definition, the synergy threshold λ varied in a range from $[-2.4\sigma; , -1.5\sigma]$ including 16 points for the basic algorithm and 3 for the combinatorial. The λ choice impacts the number of synergies in the matrix. For the fifth cell line U-373MG, there were no synergies in this range, and therefore it will not be consider in the further discussions. For the cell line U-343MG, there are only 2 to 4 synergies in the matrices. For the other cell lines, A172, U-87MG and T98G, the number of synergies varied from 5 to 45. When there were few synergies in the matrix, it might happen that all of then were randomly chosen as initial observations and in this case, there were no synergies left to be found and the algorithm efficiency could not be calculated. Consequently, these simulations were discarded.

3.1 Basic Algorithm

For the basic algorithm, 1000 simulations were run for each of the 16 different $\lambda's$ and each of the 4 cell lines. The screening efficiency for all these simulations (excluding the ones when all $1's$ were chosen as initial observations) was $\boldsymbol{\tau = 3.32}$. Figure 7 shows the overall result for all cell lines and all thresholds pooled together and all threshold together for each cell line individually. In Table 1, there is information about amount of synergies (first row) and τ (second row) for each cell line and each λ. In the bottommost row, there is overall results for each λ value including all cell lines and in rightmost column, there are overall results for each cell line including all thresholds λ. Highlighted in pink, there are the value for -2.4σ, -2.0σ and -1.5σ. Overall for all the cell lines (bottommost row), it can be seen that the screening efficiency τ is in general better for lower synergies thresholds λ. This happens due to the fact that higher $\lambda's$ may propitiate the appearance of *isolated synergies*, i.e. synergies that are the unique synergy for both drugs i and j involved in the synergistic pair (Figure 8). Isolated synergies contradict our assumption that drugs tend to be hubs in the network, and that are thus expected to have more than one synergy. However, due to the

characteristic of the algorithm to pick the cell that yields the maximum information gain, it can still do a good job and find isolated synergies with worst efficiency $\tau = 1.90$ for cell line U-343MG and $\lambda = -1.6\sigma$.

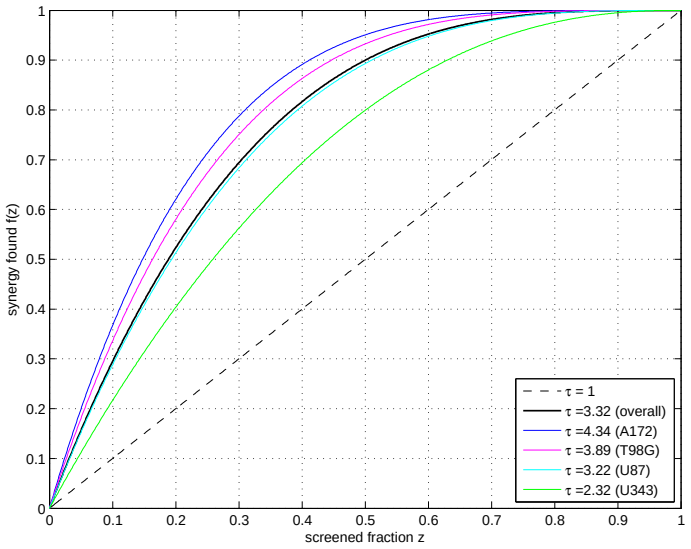

Figure 7: Screening efficiency for the basic algorithm: τ for all threshold and all cell lines pooled together (full black line), and for each cell line individually (overall $\lambda's$), considering 1000 repetitions with different initial points and data fitted to Equation 2. The difference in results is due to number of synergies in each cell line and also synergies configurations.

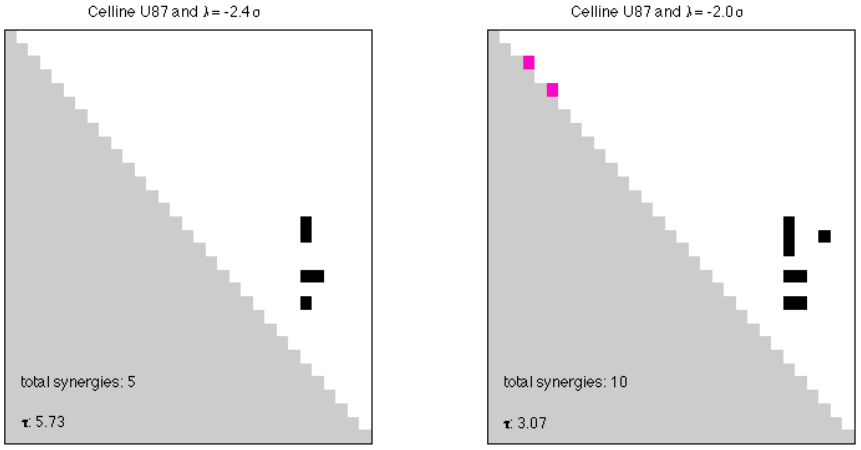

Figure 8: Isolated synergy: Example for cell line U-87MG and thresholds -2.4σ (left) and 2.0σ (right). In the left panel, performance is better $\tau = 5.73$, since at least one of the drugs i or j in each synergy has more than one synergy across all pairs, following the assumption that certain drugs are hubs in the network. In right panel, on the other hand, the pink synergies indicate an isolated synergy, i.e. a synergy which is unique for both drugs i and j involved in it. This configuration decreases τ to 3.07, since the isolated synergies are harder to discover. However, due to the algorithm characteristic to pick the pair that yields the maximum information gain, all synergistic pairs are still relatively rapidly found and screening efficiency is good.

Table 1: Screening efficiency τ and amount of synergies for each cell line and each synergy threshold λ: for each cell line, the first row shows the synergies in the matrix and the second shows τ. The rightmost column shows overall τ pooled for each cell line across all thresholds, while the bottommost row shows overall τ pooled for each threshold across all cell lines.

	-2.40σ	-2.34σ	-2.28σ	-2.22σ	-2.16σ	-2.10σ	-2.04σ	-2.00σ	-1.92σ	-1.86σ	-1.80σ	-1.74σ	-1.68σ	-1.62σ	-1.56σ	-1.50σ	overall
A172	6	8	11	11	12	12	15	15	15	15	16	16	16	16	16	17	
	3.81	2.83	4.51	4.51	5.40	5.43	4.41	4.41	4.47	4.48	4.17	4.19	4.30	4.18	4.19	6.15	4.34
T98G	19	21	21	22	23	24	26	28	33	34	37	38	41	41	44	45	
	3.80	3.98	4.01	4.10	3.86	4.31	4.56	4.39	3.75	3.60	3.86	3.75	3.61	3.62	3.79	3.56	3.89
U-87	5	5	6	7	9	9	10	10	11	13	13	15	16	18	20	20	
	5.73	5.28	6.62	3.94	3.82	3.75	2.74	2.70	2.25	2.61	2.58	2.28	2.04	2.75	2.85	2.82	3.22
U-343	2	2	2	2	2	2	2	2	2	2	2	3	4	4	4	4	
	2.38	2.30	2.37	2.36	2.30	2.47	2.54	2.35	2.50	2.49	2.58	3.28	1.92	1.84	1.90	1.95	2.32
overall	3.73	3.38	4.00	3.56	3.60	3.75	3.42	3.30	3.09	3.19	3.23	3.33	2.77	2.91	3.00	3.15	**3.32**

Figure 9 shows, for each cell line, screening efficiency values versus synergy threshold λ (left) and versus number of synergies in the matrix (right), which is just a conversion from threshold into amount of synergies. It can be noticed that τ is quiet stable for λ variation and more sensitive to the actual amount of synergies in the matrix, implying that small changes in λ don't change the number of synergies for the same cell line. In the right plot, it can be noticed that cell line T98G is the most stable while U-87MG and A172 present opposite behaviours, i.e. for U-87MG τ decreases with number of synergies while for A172 τ increases. This difference in the behaviour can be explained by synergy configuration in the different cell lines and existence of isolated synergies at certain λ values (Figure 8). Cell line U-343MG has a small variation in number of synergies, making it hard to access its stability.

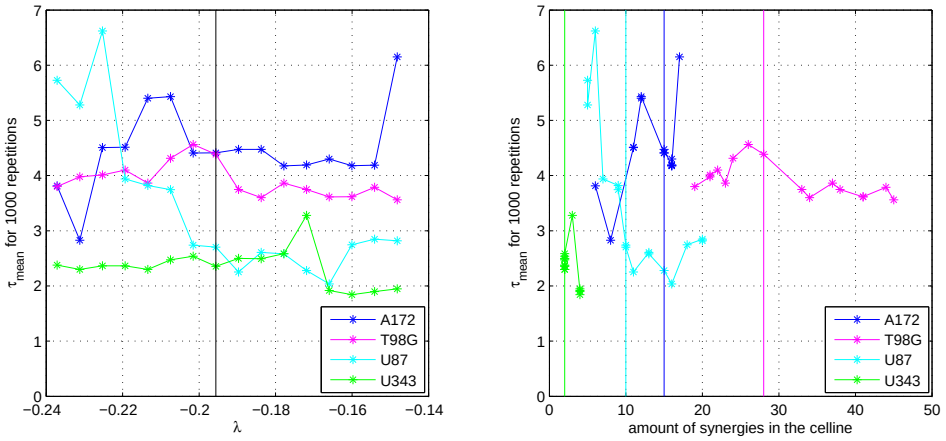

Figure 9: Screening efficiency versus synergy threshold (left) and amount of synergies (right): When synergy thresholds λ are converted into amount of synergies, it can be seen that τ is more sensitive to the number of synergies than to λ itself. Each cell line presents, however, a different behaviour when the number of synergies varies. The losses in efficiency are mainly explained by different synergys configuration and presence of isolated synergies (Figure 8). The vertical lines indicate $\lambda = -2\sigma$.

For the experimental data, λ was set as -2σ. Also, it has been shown (Figure 9) that when the number of synergies varies τ has different behaviours for each cell line. Thus, we investigate the changes in the synergy matrix $\mathbf{S}$ configuration for each cell line at $\lambda = -2\sigma$ and -1.5σ in Figure 10. There, isolated synergies can be observed (in pink) for cell lines U-87MG at $\lambda = -2.0\sigma$ and U-343MG at $\lambda = -1.5\sigma$, configurations whose τ is also smaller compared to the same cell lines without isolated synergies. Therefore, we can conclude that isolated synergies explain part of the reduction in τ. However, for the other cell lines, A172 and T98G, the difference in τ cannot be explained by a direct feature in the synergies configuration.

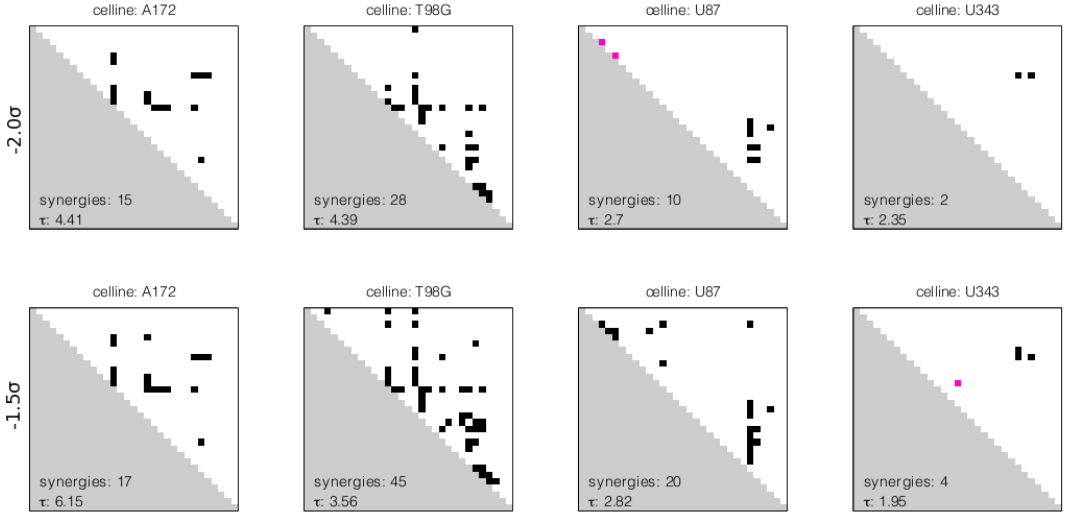

Figure 10: Synergy matrices for each cell line at $\lambda = -2\sigma$ and -1.5σ: isolated synergies (Figure 8) are highlighted in pink. It can be seen that the isolated synergies happen in the lower λ, explaining at least part of the loss in screening efficiency.

When the percentage of initial known epistasis values varies (data not shown), τ increases linearly with the increase of known initial observations. This is expected, since more information allows us to make better decisions. However, the slope of the curve is small, which indicates that it is not a crucial parameter. Another parameter evaluated was the amount of initial $1's$ observed for the same percentage of initial observation. In this case, a tendency for τ to increase when more $1's$ were observed among the initial pairs was noticed. This can be explained by the fact that since $1's$ are rare, i.e. each cell in the matrix has low probability of being 1, therefore they are also the most informative events, yielding better predictions. In other words, the presence of $0's$ in the matrix is not very informative, since it is expected to be mainly $0's$ in the synergistic matrix $\mathbf{S}$.

3.2 Combinatorial Algorithm

For the combinatorial algorithm, 500 simulations were run for 3 values of threshold, $\lambda = \{-2.4\sigma, -2.0\sigma, -1.5\sigma\}$, and for each of the 4 cell lines. Sets of 2 and 4 pairs were simultaneously tested. The screening efficiency for all these simulations (excluding the ones when all $1's$ were chosen as initial observations) were $\boldsymbol{\tau = 3.22}$ and $\boldsymbol{\tau = 3.41}$, respectively for 2 and 4 combinations of pairs simultaneously tested. Table 2 shows overall efficiency to each cell line for $m = 1, 2, 4$ simultaneous pairs. Please note that $m = 1$ is equivalent to the basic algorithm. It can be seen that the performance is similar regardless of m. In fact, a two-sample t-test for equality in $\hat{\tau}$ for $m = 2$ *or* 4 compared to $m = 1$ did not showed a significat difference in the algorithms' performance. The tests' hypotheses were:

$$H_0 : \hat{\tau}_{m=1} = \hat{\tau}_{m=2\ or\ 4}$$
$$H_a : \hat{\tau}_{m=1} \neq \hat{\tau}_{m=2\ or\ 4} \tag{14}$$

The resulting *p-value's* were much higher than 0.01, confirming that there is no statistically significant difference between the two algorithms, and it is not relevant to consider simultaneous tests.

Table 2: Screening efficiency τ for different amount m of drug pairs tested simultaneously: it can be seen that m does not have a relevant impact in the results.

	$m = 1$	$m = 2$	$m = 4$
A172	4.34	4.41	4.35
T98G	3.89	4.04	3.99
U-87MG	3.22	3.53	3.53
U-343MG	2.32	1.67	2.23
overall	**3.32**	**3.22**	**3.41**

Since the combinatorial algorithm is computationally more time consuming and the improvement in performance is not significant, predictions for the next pair(s) to test can be done by simply selecting m pairs with top $E[G_{ab}]$ from the basic algorithm.

4 Conclusion

Epistases are relevant in several fields such as evolution, genetics and cancer treatment. Due to the large genes and drugs libraries, algorithms that provide an experimental design that increases the epistatic interactions discovery rate are necessary to reduce experimental time and cost.

The present work proposes two versions of an algorithm based on information theory to decide on which pairs to experimentally test. The basic algorithm is based on the assumption that certain drugs are hubs in the network, i.e. they are more likely to present epistatic interaction. Thus, an interaction probability is calculated based on the synergies observed for each drug. The epistatic probability for each pair is given by the product of the interaction probability for each drug involved in the pair. Next, each non-tested pair is considered to have an epistatic or non-epistatic interaction. This modifies the interaction probability, whose change is measured using the average expected information gain for each pair. The pair with maximum average expected information gain is chosen for an experiment. The combinatorial algorithm does the same but selects a set of pair to be tested at a time instead of only one.

The results of the two algorithms presented no statitiscally significant difference. Therefore, the basic algorithm is going to be further discussed since it is faster than the combinatorial one. The algorithm was tested for 4 cell lines and 16 different synergies thresholds, resulting in approximately 64 different synergy configurations (some thresholds yielded the same number of synergies to the same cell line and consequently output the same configuration, therefore there were less than 64 configurations).

The overall screening efficiency was threefold better than traditional methods. For the different configurations efficiency was in a range from two- to sixfold improvement. These differences could be partially explained by isolated synergies, i.e. synergies that were the only synergy for both drugs involved in the pair. Such a feature clearly does not follow the hub assumption compromising the result. However, even for the worst case, we still obtained a satisfactory improvement of twofold improvement. This can be explained by the characteristic of the algorithm to pick pairs that yield maximum information, instead of maximum probability, resulting in a better screening of the matrix.

5 Future Work

Even though we did not scale the approximate probabilities properly (Section 2.1.1.1), we got very encouraging results of improved detection rates compared to random search. It is definitely worth to study the effects both of the proposed ad hoc scaling, and try to implement a more rigorous Bayesian probability update.

Next, we are interested in applying the algorithm to larger datasets. For that, it is necessary to increase computational speed. We can recommend that either parallel computing or a non-exhaustive simulation of tested pairs is used. For a non-exhaustive simulation of tested pairs an optimization method, such as simulated annealing, can be used. Also, it is interesting to incorporate the present basic algorithm into Gerlee *et al.*'s algorithm [6].

References

[1] Bateson W. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge, 1909.

[2] Cordell HJ. Epistasis: what it means, what it does not mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002. doi:10.1093/hmg/11.20.2463.

[3] Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edin.*, 52:399–433, 1918.

[4] Phillips PC; Otto SP; Whitlock MC. *Epistasis and the Evolutionary Process*, chapter Beyond the Average: The Evolutionary Importance of Gene Interactions and Variablity of Epistatic Effects, pages 20–38. Oxford University Press, New York, 2000.

[5] Segre D; DeLuna A; Church GM; Kishony R. Modular epistasis in yeast metabolism. *Nature Genetics*, 37(1):77–83, 2005. doi:10.1038/ng1489.

[6] Gerlee P; Schmidt L; Monsefi N; Kling T; Jörnsten R; Nelander S. Searching for synergies: matrix algebraic approaches for efficient pair screening. *Submitted*, 2012.

[7] Fisher RA. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725, 1925. doi:10.1017/S0305004100009580.

[8] Shannon CE. A mathematical theory of communication. *Bell System Techical Journal*, 27:379–423; 623–656, 1948. Reprinted at: `http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf`.

[9] Kullback S. *Information theory and statistics*. John Wiley and Sons, New York, 1959.

[10] Kullback S; Leibler RA. On information and sufficiency. *Ann. Math. Statist.*, 22(1): 79–86, 1951. doi:10.1214/aoms/1177729694.

[11] Loewe S; Muischnek H. Effect of combinations: mathematical basis of the problem. *Arch. Exp. Pathol. Pharmakol*, 114:313–326, 1926.

[12] Lindley DV. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. URL `http://www.jstor.org/stable/2237191`.

[13] Chaloner K; Verdinelli I. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. URL `http://www.jstor.org/stable/2246015`.

[14] Clyde MA. Experimental design: A bayesian perspective. *Int. Encyc. Social and Behavioral Sciences*, 2001.

[15] Monsefi N. Mapping of anticancer drug combinations. Master's thesis, University of Gothenburg, 2010.

Appendix

Approximation of expected information gain

Our system is matrix $\mathbf{S} \in \{0,1\}^{n\times n}$, where n is the number of drugs and 0 indicates a non-synergistic interaction while 1 indicates a synergistic one. Each experiment correspond to try the epistasis value for a drug pair ab.

The average information gain G when a drug pair ab is tested is given by [10]:

$$G[\mathcal{P}, \mathcal{P}^{ab}] = \sum_{k=0}^{1} \sum_{ij} P(ij = k|ab) \log \frac{P(ij = k|ab)}{P(ij = k)} \tag{A-1}$$

where the first sum runs over all the possible states each pair can assume, i.e. 0 for a non-synergistic pair and 1 for a synergistic pair, the second sum runs over all pairs in the matrix.

The Maclaurin series expansion for the logarithm function, when $x << 1$, can be approximate to:

$$\begin{aligned} \log(1-x) &= -\sum_{k=1}^{\infty} \frac{x^k}{k} \\ &= -x - \frac{x^2}{2} - \frac{x^3}{3} - ... \\ &\approx -x \end{aligned}$$

If we now define $y = 1 - x$, the Maclaurin expansion can be re-written as:

$$\log y \approx y - 1 \tag{A-2}$$

Let us define:

$$\delta = P(ij = k|ab) - P(ij = k) \tag{A-3}$$

Using Equations A-2 and A-3, the right-hand side of Equation A-1 can be re-written as:

$$
\begin{aligned}
P(ij=k|ab)\log_2\frac{P(ij=k|ab)}{P(ij=k)} \approx \\
&\approx P(ij=k|ab)\left[\frac{P(ij=k|ab)}{P(ij=k)}-1\right]= \\
&= P(ij=k|ab)\left[\frac{P(ij=k)+\delta}{P(ij=k)}-1\right]= \\
&= P(ij=k|ab)\left[\frac{P(ij=k)}{P(ij=k)}+\frac{\delta}{P(ij=k)}-1\right]= \\
&= \frac{P(ij=k|ab)}{P(ij=k)}\cdot\delta \\
&= \frac{P(ij=k)+\delta}{P(ij=k)}\cdot\delta= \\
&= \delta+\frac{\delta^2}{P(ij=k)}
\end{aligned}
\tag{A-4}
$$

If we use this approximation in Equation A-1 the information gain G becomes :

$$
G[\mathcal{P},\mathcal{P}^{ab}]=\sum_{ij}\left[P(ij=1|ab)\log\frac{P(ij=1|ab)}{P(ij=1)}+P(ij=0|ab)\log\frac{P(ij=0|ab)}{P(ij=0)}\right]
\tag{A-5}
$$

Since in our case ij follows a Bernoulli distribution, we know that $P(ij=1)=1-P(ij=0)$ and we can define:

$$
\begin{aligned}
A &= P(ij=1|ab) \\
(1-A) &= P(ij=0|ab) \\
B &= P(ij=1) \\
(1-B) &= P(ij=0)
\end{aligned}
$$

Replacing the above relationships and approximation from Equation A-4 in Equation A-5, we have:

$$
\begin{aligned}
G[\mathcal{P}, \mathcal{P}^{ab}] &= \sum_{ij} \left[A \log \frac{A}{B} + (1-A) \log \frac{(1-A)}{(1-B)} \right] \\
&\approx \sum_{ij} \left[A - B + \frac{(A-B)^2}{B} + (1-A) - (1-B) + \frac{[(1-A)-(1-B)]^2}{(1-B)} \right] \\
&= \sum_{ij} \frac{(A-B)^2}{B(1-B)} \\
&= \sum_{ij} \frac{(P(ij=1|ab) - P(ij=1))^2}{P(ij=1)(1-P(ij=1))}
\end{aligned}
\tag{A-6}
$$

and consequently the average expected gain of information $E[G_{ab}]$ (Equation 10) can be expressed as:

$$
\begin{aligned}
E[G_{ab}] &= P(ab=1) * G(ab=1) + P(ab=0) * G(ab=0) \approx \\
&\approx P(ab=1) * \sum_{ij} \frac{(P(ij=1|ab=1) - P(ij=1))^2}{P(ij=1)(1-P(ij=1))} \\
&\quad + P(ab=0) * \sum_{ij} \frac{(P(ij=1|ab=0) - P(ij=1))^2}{P(ij=1)(1-P(ij=1))}
\end{aligned}
\tag{A-7}
$$