Computer analysis of perturbations of gene expression patterns in response to small molecules

M.Sc. Thesis

Molecular, immuno- and microbiology program

Author:

Gábor Szegvári

Supervisor:

András Málnási-Csizmadia, associate professor
Department of Biochemistry

EÖTVÖS LORÁND UNIVERSITY OF SCIENCES
FACULTY OF NATURAL SCIENCES
INSTITUTE OF BIOLOGY

Budapest 2012

Table of contents

Appendix available as separate files. Read appendix_guide.rtf for description of the files.

„The only real voyage of discovery
consists not in seeking new landscapes,
but in having new eyes."

Marcel Proust

Scientific research has always moved in a spiral of Observation – Hypothesis – Experiment – Data – Observation…, which, as indicated, classically starts with an observation. In recent years, though, with the dawn of more efficient and high-throughput techniques a new option was created, namely to start with an experiment instead. The former is usually called hypothesis-driven and the latter data-driven research. The distinction between the two can be considered both justified and not, depending on the point of view. The separation is valid because the methodology of the two can be greatly different, as the use of the data-driven approach requires the analysis of massive amounts of data as its foundation and thus introduces new tools from mathematics and informatics. On the other hand the two can be considered theoretically the same, differing only in where they obtain an observation, data-driven research trading fields of flowers for matrices of data and tails of reptiles for tails of distribution functions. Thus the high-throughput experiment can be viewed as a new eye with which we can look at nature and make observations. With the rise of techniques providing incredible amounts of data researchers start to divert their attention to more complex problems. One of such tasks is the complete characterization of the effect drugs have when administered to a whole organism.

Because of the limitation in resources the classical approach of mapping a drug's (or drug candidate's) complete range of effects (its "effect profile") consists of a low number of highly specialized experiments testing the molecule for effects based on the scientist's choice. Even assuming a highly knowledgeable experimenter it is more than likely that effects will remain undiscovered. In the best case it will lead to the potential of the compound remaining untapped. In the worst case it will lead to further (unnecessary) testing of a chemical unsuitable for treatment due to unforeseen side effects consuming time, energy and money. It is not unheard of that serious side effects get recognized only after the release of the drug onto the market, jeopardizing the health and lives of patients. Contergan (thalidomide) is a well-known example (Newman, 1986). This does not change the fact that the only dependable method of determining whether a drug has a certain effect remains a specific in vitro and/or in vivo test. What can we gain from the modern techniques then?

While it is true that until we decode the workings of life on the molecular level in its entirety (towards which all biologists work, but it would not be wise to count on it in a reasonable amount of time), indirect and computational methods will only provide approximations and predictions and will require validation by the conventional techniques. However they have much greater data processing power and as such may base their results on data previously unavailable to scientists. Instead of having to rely completely on the decisions of the researcher, a preliminary screening can be performed to suggest which specific test should be applied reducing the subjective element of the experiment. Even with the possibility of some results being faulty, the efficiency of experiments based on them is still significantly higher than what the classical approach can offer (i.e. stopping when the researchers run out of ideas). Moreover, their uncertainty can be quantified and accounted for and thus they can actually increase the reliability of the results.

It is obvious that one cannot handle this new type of information the same way as the usual data. There are two main causes for this: 1. as I have mentioned, the first batch of data to analyze here forms the observation instead of the result 2. the sheer amount of data makes it technically impossible to interpret or use it without a computer (“manually”). Both of these lead to a need for new types of analysis. There are many ways to overcome this and certainly many more to come, but here we will restrict our discussion to the two approaches that form a crucial part of the work presented.

The first one is a natural choice, utilized for ages when dealing with a multitude of objects: classification. At the same time classification both reduces the number of components and assigns more direct meaning to them.

The second one bypasses the problem entirely by treating all the data for an observation as a single entity, usually referred to as a fingerprint, profile or pattern. We will use the term ‘pattern analysis’ for this approach. By their nature, patterns cannot be interpreted individually, but serve as a tool to identify and compare observations. As this does not rely on a detailed understanding of the elements of the pattern or their relation to the attributes of the observations is question, many different sources can be used to acquire patterns without the need for further extensive studies. With the right experimental design we might even be able to deduce some of this information.

Nevertheless, while less limited by our knowledge, naturally the analysis will yield better results if there is a strong connection between the information in the pattern and the attributes of interest. Let's take the case of drug design for example. One can create a pattern from the physicochemical properties of a compound to derive information on its medical effects. However when our research group tried this, it was not quite efficient (unpublished data). Patterns that describe the interaction of the molecule with a model of the target system proved to be more useful, even if the model seems somewhat artificial.

DRUG PROFILE MATCHING (DPM)

While proteins are not the only possible interaction partners of a compound introduced to the organism as a drug, as most enzymatic activities belong to them, restricting our investigation to small molecule-protein relations is an acceptable approximation. In their work Simon et al. (Simon et al., 2012) explored two concepts regarding the use of in silico (computer-generated) protein-ligand patterns: first, whether it is actually an appropriate basis for prediction of drug effect profiles; second, how much does the quality of the proteins used to create the patterns affect the results, especially if non-target proteins are suitable.

They generate patterns with a technique called docking. In general this utilizes the 3D structure of a small molecule and a large one and probes the surface of the large with the small, approximating the parameters of a possible bond between the two. In this case they used only the naturally occurring ligand-binding pocket of selected non-drug-target proteins (from the PDB, chosen based on suitability to their methods) as the "large molecule" (149 different proteins, one pocket each). Drugs and drug candidates (theoretically any compound) take on the role of the "small molecule". The proteins' structure is kept static, while the drug is allowed to change conformation and rotate freely in a water-free box restricting it to the pocket. Binding free energies are calculated for each drug-protein pair 25 times and the minima form the matrix of Interaction Patterns (IP-s).

They used data on 1177 FDA approved drugs, extracted from the DrugBank database. Over all of them 559 medical effects are defined, of which 177 have at least 10 drugs registered to them. The others are excluded for not having sufficient information for classification and the 177 forms the binary Effect Profile (EP) matrix. This is an earlier version of the database we use and describe in *Materials and Methods* (see pg. 22).

The matrices are then analyzed with linear discriminant analysis (LDA, a.k.a. Fisher's approach). In LDA one creates a linear discriminating function on the explanatory variables using which two groups (here: drugs which have a particular effect vs. those which not) can be separated. In essence it computes a new axis in the variable space so that if the points are projected onto it the two groups have the least variance and the distance between their averages is as big as possible (see figure 1). Then the probability of belonging to a group can be calculated for each observation based on their and the group's position on this axis. If a cutoff value is defined it can be thought of as the hyperplane perpendicular to this axis best separating the groups.

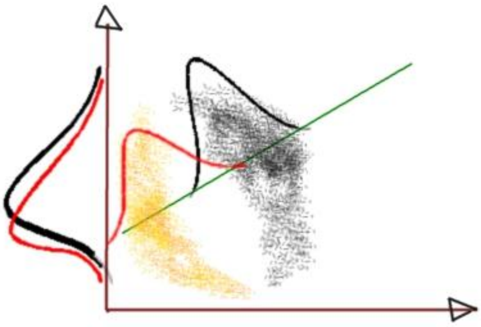

Figure 1: **Demonstrating the principle of LDA on a 2D example. Distribution along the initial axes is not suitable to separate the groups, but introducing a theorethical new axis largely improves discrimination.**

As each category is tested separately and independent of others, there are no constrictions on the number of categories a drug can be sorted into, neither on the relation of these. The output is the solution of an exact mathematical problem and thus the time necessary to perform it is short too. It is usually recommended for multivariate normal data (values in every variable are distributed normally and the same is true for any linear combinations of them too). It is not a strict requirement, but if the data has an unusual distribution it might not perform well. Since it tries to separate groups with a plane, it cannot fit to "weird" forms, which could possibly be easily accounted for if a more complex shaped surface could be used.

The results of this work are validated using both statistical methods (Simon et al., 2012) and in vitro testing of a number of predictions made (manuscript under preparation).

The whole concept of DPM is summarized well on figure 2. (Note: here canonical correlation analysis is listed separately from LDA, but it can be considered a part of it too.)

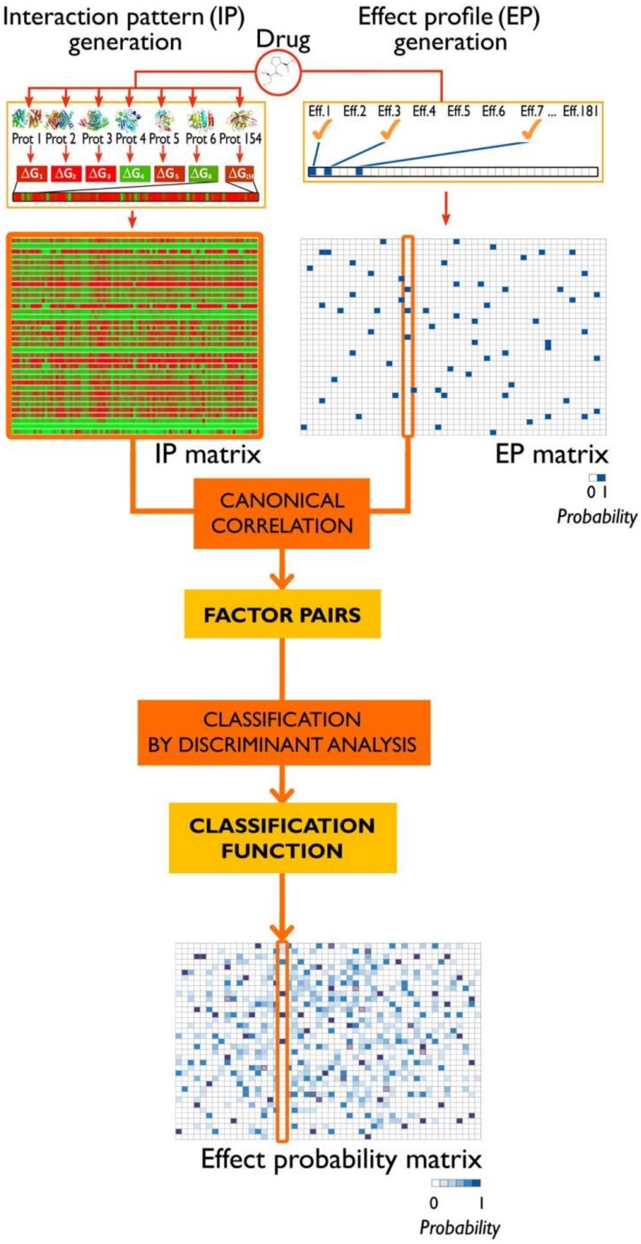

Figure 2: Overview of the Drug Profile Matching method. Reprinted with permission from Simon et al.: Drug effect prediction by polypharmacology-based interaction profiling. Journal of chemical information and modeling 2011. Copyright 2011 American Chemical Society.

While DPM is a potent technique on its own, there are advantages to creating data based on multiple model systems. Different sources might work better for other effect categories. If the method is “closer to life”, i.e. in vitro or in vivo, the information it contains about the model system is valuable too and we may be able to extract it. We can expect a technique suitable for the systematic investigation of drug effect profiles to satisfy the following criteria:

I. Measurement design must be independent of what is known about the drug’s effect profile.

II. The method must be resource-efficient, i.e. produce sufficient amount of data with reasonable time and energy investment.

III. The measured data must show a connection with the drugs’ effects.

In the following chapter we describe the use of DNA chips to show that it matches the criteria.

DNA MICROARRAY

DNA microarray technology (also called DNA chip or gene chip) is based on the fact that given appropriate conditions complementary strands of DNA will hybridize with each other regardless of the source we gained them from. So if we synthesize a short strand of DNA (an oligonucleotide) with a known sequence, we can identify complementary DNA in a sample by detecting the interaction of the two. The detection can be achieved with a method similar to immunological blots: First the synthesized interaction partner (the probe) is immobilized on a solid surface in a well-defined spot. The partners in the sample (the targets) are labeled and added to the surface in a liquid phase. After allowing the interactions to form, the sample is washed away and only the targets which have found a suitable partner remain as they are indirectly bound to the solid phase through the probe. After this the labels are detected on the surface and the signals are identified by their position (see figure 3). Since signals can only come from probe-target hybrids and we know which probes were attached at which positions this also defines the quality of the targets involved. The first experiment utilizing these principles was conducted in 1982 (Augenlicht and Kobrin, 1982) and the first use of microarray technology was reported in 1995 (Schena et al., 1995).

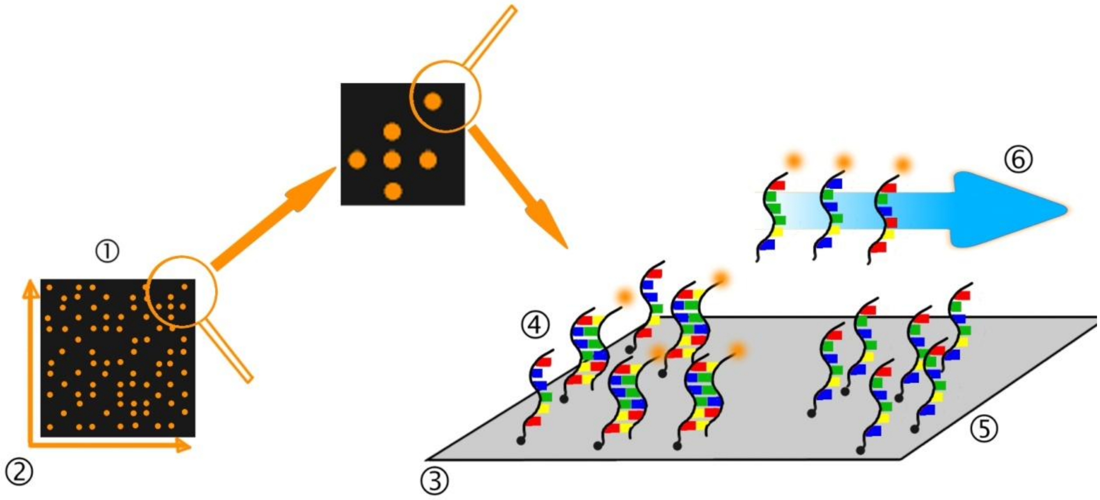

Figure 3: **The scheme of DNA microarrays. 1-the chip, with positive spots showing fluorescence 2-position alog these axes is used to identify spots 3-solid surface 4-spot showing a positive result 5-spot showing a negative result 6-not hybridizing strands are washed off the chip before detection**

There are several different designs which all follow the above principles. The solid surface can be manufactured from different materials, like glass, silicone or plastic. The form of the solid phase can be a flat plate or a set of microscopic beads. In the first case positions are identified with two coordinates on the plane. In the latter, each bead carries one type of probe and is labeled with two different fluorescent dyes (that do not interfere with the target label) in different amounts and the level of these serve as “coordinates” to identify the bead. Targets are usually labeled with fluorescent dyes (fluorophores), but alternative methods can be used such as chemiluminescent molecules. With advances in technology the size of the instrument is not a limiting factor and the number of probes can vary greatly depending on the problem the array was made to answer. Probe spots on plate-style arrays can be affixed as close as 11 μm-s, the detection resolution can be 1.56 μm and quantities of DNA as small as 0.75 pM can be detected [i3]. In addition to this high sensitivity, quantitative measurements can be performed too.

A certain item of interest is not represented with a single sequence (type of oligonucleotide), but rather a group of them, called a probe set. The signals of the different probes in a set can be analyzed individually, but genome-wide studies aiming to query the whole (or most of) the genetic material of a cell often use the average (and possibly the variance) of them instead to ease interpretation. To reduce errors stemming from the inhomogenity of the liquid phase, on a plate-style array the spots of a set are not physically close to each other. The quantity of targets hybridized to the probes is measured relative to control probe sets. These numbers constitute the output of the microarray experiment.

Microarrays can be used to evaluate the results of many types of tests. Using parts of a reference genome as probes, mutations can be detected in the genome used as target, as differences in sequence lead to a lower efficiency of hybridization. The sets of sequences gained from DamID (DNA adenine methyltransferase identification) or chromatin immunoprecipitation (ChIP) (Ren et al., 2000), that show association with certain proteins of interest can be identified with a genome array of the species in question. Messenger RNA extracted from a sample can be converted into DNA using a reverse transcriptase (for higher stability and better hybridization to DNA probes, the result is called a copy DNA or cDNA) and be used as target to measure expression levels on a genomic scale (Schena et al., 1995).

The microarray design has also been used with other biologically relevant molecules besides DNA. Protein and peptide arrays are used to measure protein-protein type interactions. The scheme of the experiment is the same with proteins (or parts of them) taking the place of DNA and their various types of connections replacing hybridization (MacBeath and Schreiber, 2000). Chemical compound microarrays can also be constructed (Freiberg et al., 2004). In this version small molecules are immobilized onto a surface and a protein is used as the target to reveal possible modulators of it from the compound library.

We have established that we are capable of generating multiple types of broad scale patterns. While these patterns still do not describe every possible aspect of a particular state of a cell, they provide us with a method that can take a considerable amount of samples from the underlying pattern and with great diversity. Because of the interconnectedness of the components of the cell, it is highly likely that even if the parameter directly creating or most highly influencing an attribute of the system is not measured, we could still detect its indirect effects. Through this it may be possible to discover the “source” of the attribute, but even if not, the pattern can be used as a fingerprint to identify the different states. The same applies to measuring a difference in states, i.e. comparing a state of interest to an adequately chosen control state.

THE CONNECTIVITY MAP

BUILD 01

Researchers at the Broad Institute were the first to highlight the problem of finding connections between diseases, the molecular mechanisms of cells and the effect of small molecule drugs on them and in proposing a way to tackle this problem (Lamb et al., 2006). While these aspects of the same system are already studied separately from each other, communication between the fields is hindered by the difference in the methods used. To explore the relations of these three, we need some extent of uniformity of approach. Therefore their aim was to find a suitable platform on which data obtained from these sources can be compared.

Since direct comparison is naturally out of the question, the creation of an “adaptor”, a common point of reference is needed. This should be created using a type of data that is attainable from all three aspects and with great diversity of it (many variables) to accommodate as many of external data sources (experiments) for comparison as possible, even if they were not created specifically for this purpose (and thus use their own set of variables). In essence we will need a pattern that is present in these different experimental systems and shows a relation to their state.

Every living cell in every moment has a gene expression pattern, and it plays a pivotal role in determining the attributes of the cell as it is one of the major regulators of its composition. Thus we can assume that similar changes in it indicate a similar response of the cell to perturbation. Microarray technology provides us a practical solution to measuring its first step, transcription, as mentioned above (Schena et al., 1995). This method does not have any requirements regarding the treatment of the cells serving as a source of RNA, except that it should not introduce unidentified mRNA into the system. Thus a measurement can be taken practically irrespective of the type of condition we want to examine. By comparing the sample under the condition in question with a suitable control, a pattern related to the condition can be obtained.

Thus we arrive at the design of the Connectivity Map Project (CMAP) (Lamb et al., 2006). Change in the level of transcription of genes was chosen as a common language for the different disciplines researching the aspects of disease and medicine and a database was created to serve as a reference.

The database is built from the microarray fingerprints of the changes cells undergo in response to bioactive small molecules (referred to as “perturbagenes”). To obtain a pattern representing the change the values from the expression patterns of the treated cells are divided by the values of a control treated only with the vehicle (see figure 4). A particular treatment, defined by the cell line, perturbagene, its concentration and the duration of treatment, paired with the appropriate controls is termed an instance. In the first set of experiments 164 different perturbagenes were used in 453 instances. Four different cell lines were used (MCF7, PC3, HL60 and SKMEL5), all laboratory strains derived from cancerous human cells. Such lines were chosen for ease of handling. Measurements were made with human genome chips created by Affymetrix (code HG-U133A). Due to the large number of arrays used, a new group of controls were made for every batch of experiments. This set is now referred to as build 01 or CMAP01.

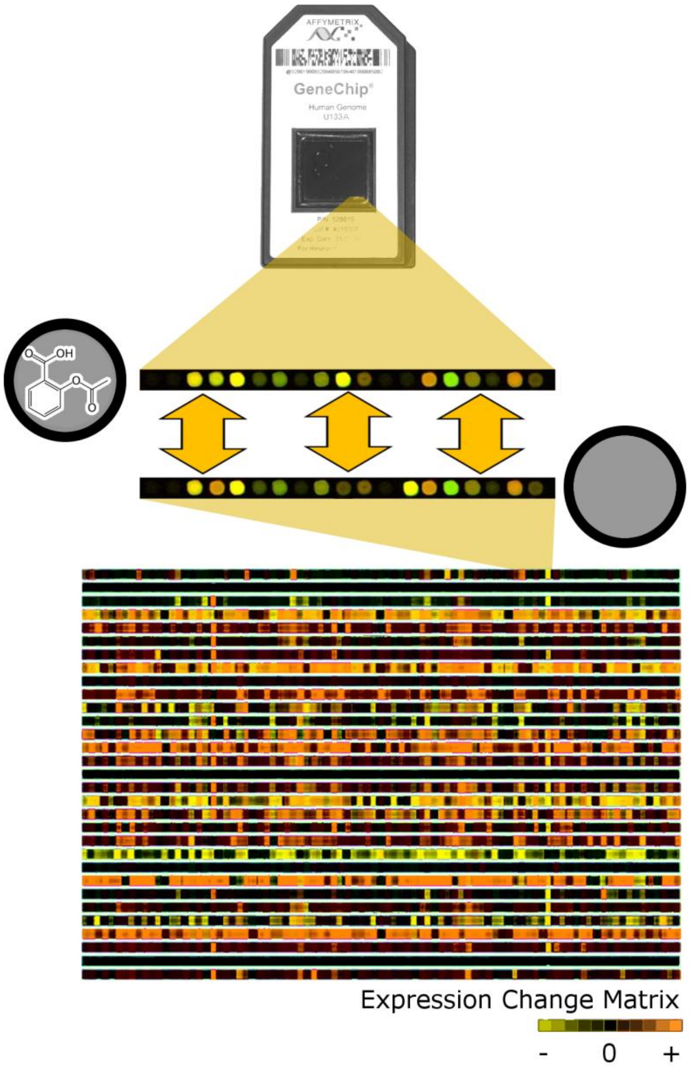

Figure 4: **Generating the Connectivity Map reference database.**

Next, a way to compare single expression profiles with the reference dataset has to be decided on. To be in concordance with the original aim of the project it has to be compatible with different types of experiments, not just profiling done with the same microarray. To this end they utilize a nonparametric, rank-based method based on the Kolmogorov-Smirnov statistic called Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005). The profile of interest ("query") is compared with all the instances in the database and a similarity statistic is derived from each comparison ("connectivity score"). To create a query only the fact that a particular sequence got up- or downregulated is needed to be specified and it is not necessary for it to contain information on all the probe sets used in the measurements that yielded the database. Thus the query consists only of two sets of variables. This way the amount of data that can be accommodated is significantly increased. Then the position of these variables in the list of all probe sets ranked by the values in the particular instance is inspected. The connectivity score of the query and the instance is calculated from this distribution. The score can range from -1 to +1. Zero connectivity indicates that the two patterns are completely unrelated. +1 means the two are highly connected: genes upregulated in the query are usually ranked high in the instance and vice-versa. -1 also implies a strong connection, but of the opposite nature. In the end, instances are ranked based on their connectivity score and this list is considered the output of the method (see figure 5).

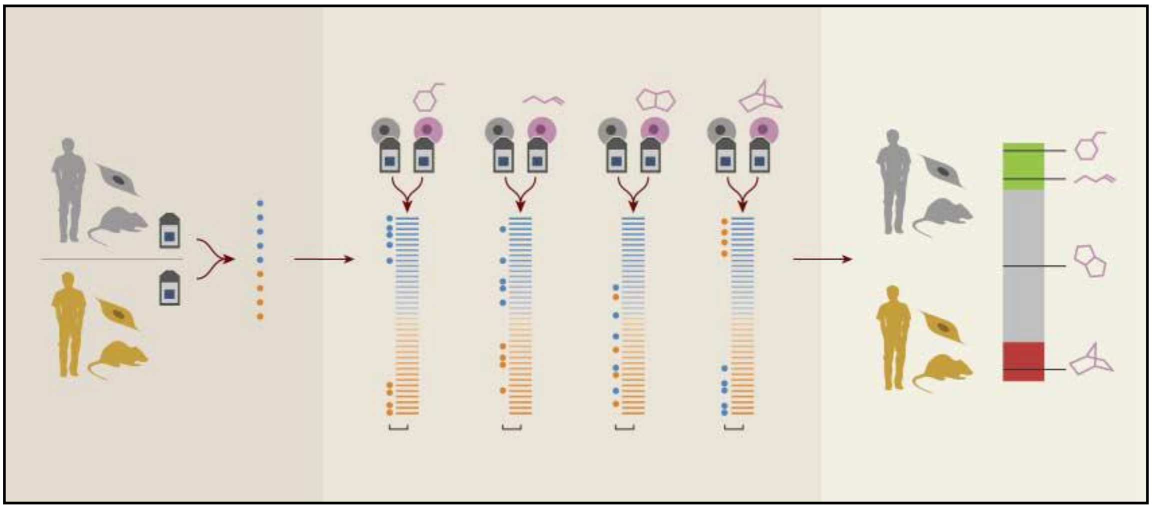

Figure 5: The Connectivity Map concept. Left: Expression change data is translated into a query. Center: Data points of the query are compared to the ranked reference lists. Right: A ranked list of connectivity scores is generated. From (Lamb et al., 2006)**. Reprinted with permission from AAAS.**

Data from several sources was used to prove and demonstrate the usefulness of this system (Lamb et al., 2006). Amongst others, by comparison with the database, researchers were able to better understand the mechanism of action of gedunin, a natural product derived from the Meliacae family of plants (Hieronymus et al., 2006). Its suppressive effect on androgen receptor was described, but the exact relation was unknown. In the top ranked results of their query, several instances of geldanamycin and its derivatives were found, all inhibitors of HSP90, a known interaction partner of the receptor. This prompted the scientists to test gedunin for HSP90 inhibitory activity, which was indeed confirmed. With CMAP the connection between gedunin and geldanamycins could be revealed, even though there is no structural similarity.

In another study, by measuring the difference in expression patterns between two types of cells and using this as input, they were able to identify a drug that proved to be able to induce a change in one of them so as it became similar to the other (Wei et al., 2006). The object of interest was acute lymphoblastic leukemia (ALL), more specifically its resistance or lack thereof to dexamethasone treatment. The expression patterns of resistant and sensitive lines were measured and the differences calculated (like treating one as “control” and the other “perturbed”). The signature acquired was used to query the CMAP reference database. One of the high scoring signatures was that of rapamycin (also known as sirolimus, inhibitor of the kinase mTOR, which is a central molecule in many signaling pathways and is known to regulate apoptosis among others). The existence of this match suggests that the changes induced by rapamycin are highly similar to the changes we would expect to see if we managed to transform a resistant cell into a sensitive one. Further (in vitro) testing confirmed their hypothesis and now rapamycin is registered for use in the treatment of dexamethasone resistant ALL.

Despite its usefulness and success, there are several limitations of the technique (Lamb, 2007). There are only a limited number of cell types used to create the reference set. Since no cell expresses its every gene it is possible that a perturbagene will have no effect on our cells because they lack the necessary apparatus (e.g. receptor) to react. While certain compounds would require specific cell lines, others with more ubiquitous effect show similar results on different cells (Lamb et al., 2006). Thus expanding the set of cells would yield a relatively low amount of information. Besides the lack of diversity in cells, the fact that they are measured outside of their natural environment is of serious concern. Some drugs act on cells indirectly, e.g. by modifying the level of a hormone. Molecules with an antagonistic activity can show no effects in absence of an agonist. Correcting for this (like adding the agonist to both the treated and control samples) would require a priori knowledge of the mechanism of action. This is not always available about perturbagenes used to create the reference set and goes against the basic purposes of the technique to expect it about new molecules tested. Another important issue is the interpretation of the results. What can be considered a high enough or low enough connectivity score varies query by query and thus the identification of “hits” is not trivial. CMAP does not assign any index of statistical validity to help with that and to avoid false hits (Lamb et al., 2006; Zhang and Gant, 2008). While there are certain cases where the enrichment of perturbagenes having a common effect in the output list is striking and easily noticeable, we cannot expect this every time. This reintroduces a decent amount of subjectivity to the process too.

BUILDING ON

Another shortcoming of build 01 is its lack of complexity, i.e. low number of instances (Lamb, 2007). This, in addition to that the validity of the approach was confirmed, led to the extension of the reference set with data from further experiments and thus the creation of build 02. (It was not published in a separate article, nor could I find an exact date of its creation. From the publication date of papers using cmap01 and 02 I would approximate 2009.) As stated in the previous paragraph the introduction of new cell lines was expected to be not efficient and thus they concentrated on raising the number of perturbagenes. Version 2 contains information on 1309 compounds acquired in 6100 instances. Most of the experiments were conducted using three cell lines, MCF7, PC3 and HL60.

As the original authors worked on expanding the data set, other researchers who saw promise in the project contributed too. Zhang and Gant created a new way of scoring connections to address the lack of a measure of statistical validity that can also make use of more information from the query (Zhang and Gant, 2008). They aimed to assign an index value to the probe sets in the signatures that reflects their level of perturbation and thus importance in describing the change compared to the control. To this end, they are ordered based not on the ratio of treated versus control, but the absolute value of the logarithm of that. This transformation allows for the equal treatment of up- and downregulation. Before, upregulation was represented by a value from (1;∞), while downregulation was “crammed into” [0;1), making a direct comparison difficult. Now both range from 0 to infinity, in the same direction. The highest ordered probe set gets rank ±N, where N is the number of probe sets in the reference signatures and the lowest gets ±1. The sign of the rank depends on whether the sequence in question was up- (+) or downregulated (-). If sufficient information is available, ranks can be assigned to the probe sets or their equivalents in the query the same way. If no such data is present, all ranks for the query should be considered ±1. After these calculations a new connectivity score can be acquired with the following equation:

$$c(R,s) = \sum_{i=1}^{m} R(g_i)s(g_i)$$

where m is the number of probe sets or genes in the query, g_i is the i-th probe set or gene in the query signature, $s(g_i)$ is its rank in the query and $R(g_i)$ is its rank in the reference signature. There are some important attributes of this value that should be noted. Every probe set or gene that was perturbed in the same direction in the two signatures will yield a positive value to the sum, while those affected differently will yield a negative. This also means that they can cancel each other out. Probe sets or genes with higher absolute rank in either signature will sway the result more than their lower ranked counterparts. This value can be normalized by dividing it with the highest theoretically possible score for given N and m. This way the new index is from the interval [-1;1] too. With comparing the gained connectivity scores to ones created using random queries, a p value can be assigned to each, making it possible to detect false hits. A computer program implementing this method was also created (Zhang and Gant, 2009).

There is a large amount of gene expression data produced worldwide. In the Gene Expression Omnibus (GEO), the largest deposit of data of this nature, the collection of the results of more than 20,000 experiments was reported in 2011 (Barrett et al., 2011). In spite of this, the creators of CMAP had very good reasons not to rely on external resources and create their own data instead. While GEO collects the information, the results of independent experiments are not necessarily compatible or comparable. They are created on numerous different platforms, differing in what they test and how they are identified. In some cases controls are not properly listed or there are alternative ways comparisons can be made. Despite this, other groups have tried to find a way to utilize these data and create larger databases to serve as basis for similarity searches inspired by CMAP.

The first such approach is the GEM-TREND (Gene Expression data Mining Toward RElevant Network Discovery) search engine (Feng et al., 2009). They filtered GEO for experiments where control and treatment statuses were clearly stated to gain a subset consisting of 1540 entries (“batches”) reporting on 41516 samples (“instances”). Probe set ID-s were translated to UniGene ID-s using the relevant annotations of the different platforms. Leaving out samples lacking the necessary annotation 995 entries and 25974 samples remained. This was used as a reference dataset and the searching method was adopted from CMAP. The results are tested for statistical significance based on comparisons to random queries similar to (Zhang and Gant, 2008).

The creators of SPIED (Searchable Platform Independent Expression Database) chose a different approach (Williams, 2012). Instead of leaving out insufficiently described data, they introduce the concept of “effective fold” to interpret it. Values are compared to the average of the experimental series in place of the missing (or at least undefined) control values. Similarity scoring is based on Pearson regression analysis. While the computation methods may show a large difference between CMAP and SPIED, the underlying principles and hypotheses are the same. Both the input (query) and output are identical in nature too, indicating (what is also stated in the article,) that the aim of the two is the same and SPIED is an expansion of the idea behind CMAP.

We wish to further the design of methods capable of predicting the full effect profiles of drugs and drug candidates or more generally any bioactive substance. We chose to focus our attention on the database generated in build 02 of the Connectivity Map project because:

1. The data contained in it is related to drug effects.
2. The handling of the amount of data contained in other, extended databases could easily surpass our technical limitations and we should confirm that this type of data is useful for our purposes before committing resources to solve such a problem.

3. It was already demonstrated that it is capable of revealing connections between molecules with similar effects, even when their mechanism of action is not necessarily the same.
4. Because of the nature of the data (expression profiles) there is a possibility that it has the potential to link drug effects with the behavior of the components of the cellular machinery.

To investigate this last aspect we need to link the probe sets of the microarray to genes and through them to molecular properties. For this we need the annotation of the chip [i2], and a database that connects genes with their, or rather their products' properties.

GENE ONTOLOGY

Luckily we are not the first ones to show a need for a database annotating genes and their products with their various attributes. Indeed, since the invention of high-throughput sequencing techniques making us able to read the genetic code on a genomic scale, the focus of studies has shifted to put together the "dictionary" needed to fully comprehend the information in the letters we read.

We use the publicly available Gene Ontology database to annotate our list of genes (Ashburner et al., 2000). GO provides a unified tripartite hierarchical classification for gene product attributes. The three types of attributes are Molecular Function, Biological Process and Cellular Component. For each gene the codes most precisely describing it are listed in the database, according to our knowledge to date. The resources available online are updated weekly (we use the 2012.03.19. version of the files). For our purposes we consider each gene to also belong to all categories that are above the listed ones in the hierarchy (can be reached through backwards steps on the directed graph of relations).

PREDICTION METHODS

Our main hypothesis is that a similar expression-change profile (XCP) is indicative of a similar effect profile (EP). While it has been established that CMAP is capable of recognizing molecules with similar effects, this cannot be considered proof in itself. If we accept the above hypothesis, pattern based prediction methods will be able to reveal yet unseen parts of EPs. Instead of investigating this statement before performing the predictions, we will examine the predictive power and validity of our calculations as a means to assess the legitimacy of our assumption.

There are several prediction techniques we can choose from. Hierarchical clustering might seem a valid option because it creates another grouping or categorization of our observations, which could subsequently be compared to EPs. Drugs not registered for an effect, but belonging to the same cluster as several others who are could possibly be marked as targets of further investigations. But clustering techniques to date produce disjoint clusters, thus we would get only one prediction for every drug and most of them would have to be the restatement of known drug-effect associations so that we could base the interpretation of the other results on them.

Another option is the already mentioned Linear Discriminant Analysis (see page 6). While our variables fail normality (e.g. Anderson-Darling) tests even alone, investigation of the distributions in a randomly selected sample showed that values tend to cluster around a single value and form a single peak, even if their slopes differ from that of the normal distribution (data not shown). Thus we do not have to reject LDA right away. Also, Fisher's approach assumes that the two groups have the same covariance matrix. We expect drugs categorized with the same effect to have XCP-s more similar to each other's than to the other drugs'. While these abate the expected effectiveness of the method, it is important to note that they do not *invalidate* its results if it is able to find a hyperplane suitable to distinguish between the groups. In accordance with this, LDA was performed, but another prediction method was considered too.

Logistic regression is a prediction algorithm that has been shown to perform better under nonnormality (Efron, 1975). It is based on the assumption that what we observe as our category variable (telling about all drugs whether they are registered for an effect or not in our case) is based on another, underlying (and unobserved) variable. This variable sets a probability that the outcome of a Bernoulli trial will be positive (based on the logistic function – hence the name) and it is this outcome we detect as the category variable. The unobserved variable is estimated based on the linear combination of the explanatory variables (here the XCP) and this estimate is corrected over several iterations to maximize the likelihood of our observed outcome. The output is a probability for each observation that it belongs to the positive category (see figure 6).

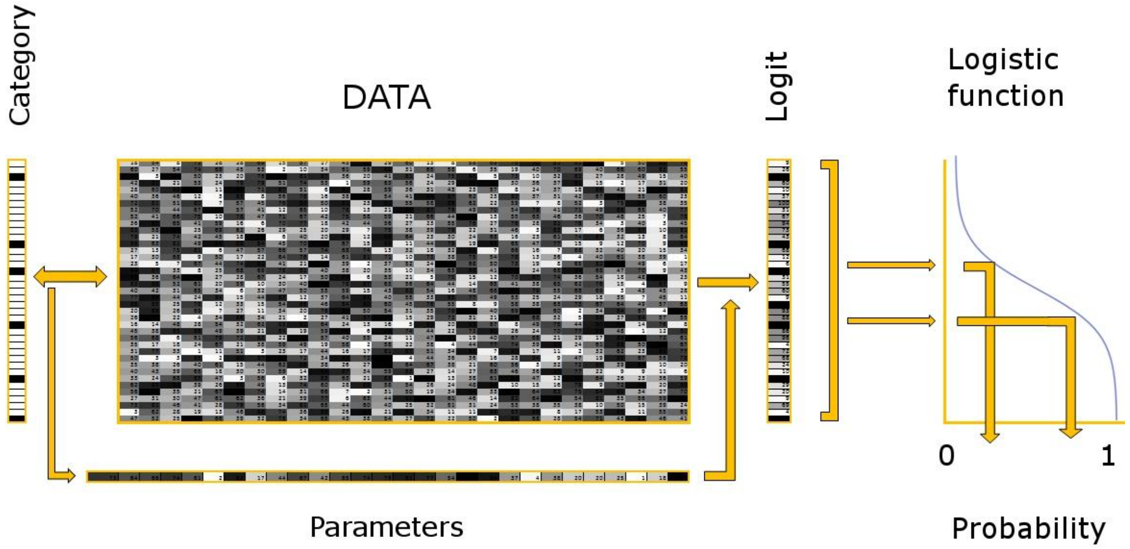

Figure 6: **Logistic regression. The relation of the category and explanatory variables is used to estimate the parameters of the logit function so as to maximize the likelihood of the generation of the observed pattern in the category variable. „Explanatory variable", „underlying variable" and „logit" are synonymous in this case.**

While having a large number of variables means we have more information, having many variables compared to the number of observations can lead to overfitting (see figure 7). It means that our model "learns" to recognize the particular positive samples provided instead of the category they belong to.

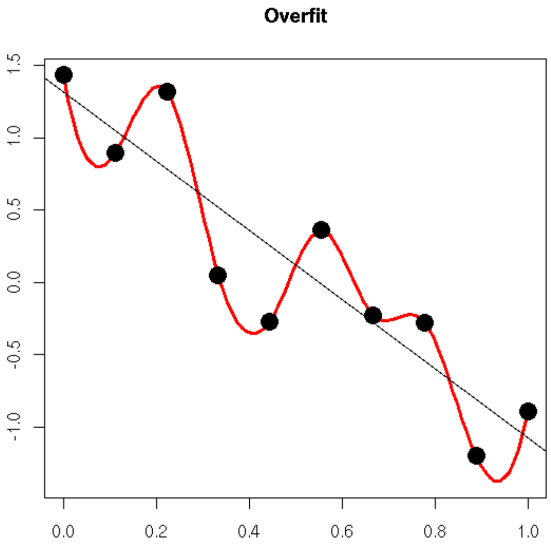

Figure 7: **Overfitting. While the red function fits the data better, it is not able to grasp the linear relation between the varibles disturbed by error. (Image slightly modified after Vincent Zoonekynd and used under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License)**

Because of this and the fact that handling over 10,000 variables with these procedures would pose technical difficulties, the use of dimensionality reduction is necessary. We used stepwise discrimination. It selects a subset of the variables based on their (possible) contribution to the discriminating power of the model. It is thus dependent on the category variable investigated.

One of the main advantages of pattern based methods is that there is no need to grasp the connection between the predicted attributes and each and every variable used for creating the prediction. Surely, such a condition would render methods based on data generated by high-throughput experiments powerless. On the other hand, its reverse, i.e. stipulating these connections based on the pattern-similarities might prove quite interesting. We attempt to use stepwise discrimination to find genes connected to certain drug effects.

OUR AIMS

To summarize, in this work we:

1. Test whether gene expression data is suitable as input for pattern based prediction of drug effects.
2. Attempt to design a method to investigate the relation between drugs' effects and genes differentially expressed in response to them.

MATERIALS AND METHODS

MATERIALS

GENE EXPRESSION DATA

In the focus of our investigations is a database containing information on the change in the level of transcription of 22283 loci in response to 1309 different small molecules. The measurements were conducted by Justin Lamb and his colleagues for their project, the Connectivity Map (http://www.broadinstitute.org/cmap, (Lamb et al., 2006; Lamb, 2007)), using the HG-U133A gene chip from Affymetrix [i1]. Four different cell lines (MCF7, PC3, HL60, and SKMEL5) were used in 6100 experiments (also called instances).

The list of instances [i4], raw microarray data (.cel files) and the matrix of ranks used for the calculation of connectivity are available for free on their website. At the same place CMAP can be queried using only a standard web browser. The table of treated/control ratios we use was kindly provided by Justin Lamb upon our request.

DRUG EFFECT DATA

Since this is not the first project in this particular field of research in our laboratory, some resources (both tangible and intangible) were readily available to us. The drug effect profile database is one of them. It is a tool our colleagues are continuously improving. The usage of an earlier version can be seen in (Simon et al., 2012) and there is another article, recently accepted for publication, that employs a more recent dataset (Peragovics et al., 2012).

The information on the drugs' effect profiles forms a database originally based on the one from DrugBank. In addition to containing the data in its "predecessor", our database has been further expanded and refined. Manual addition and checking of data has been performed multiple times. In its current form it has a two-level hierarchy and contains information on 1879 drugs and has 65 level 1 ("main") and 319 level 2 ("sub-") effect categories. Since we have data only on a subset of these drugs and have to drop categories with less than 10 drugs in them because they do not provide sufficient data for prediction, we can only use 40 main and 31 subcategories in these analyses.

In data tables each category was referred to with a unique code of two letters and five numbers. The first letter is E for effect in all cases. The second is M for main categories and S for subcategories. The numbers have no particular meaning. Outside of datasets a shortened version is also used, for example M35 instead of EM00035.

GENE (PRODUCT) FUNCTION DATA

Gene function data was acquired from the Gene Ontology database (Ashburner et al., 2000)[i5, i6]. We use the 2012.03.19. version of the files. Table 1 presents the size of the portion of the database we were able to use: the number of genes in it and the number of GO terms (different attributes) they are annotated with.

Table 1: Dimensions of the Gene Ontology database under different filtering criteria. (*): if we consider each gene to also belong to all categories that are above the listed ones

	Whole DB	Human genes	Genes in CMAP
Number of genes	225,891	17,979	11,494
Number of terms (1≤ gene(s) listed)	19,070	12,464	11,750 (11,842*)

METHODS

Calculations have been performed with the Statistical Analysis System for Windows, version 9.2 (SAS) unless noted otherwise. For details on the methods used, the book "Multivariate Data Reduction and Discrimination with SAS Software" (Khattree and Naik, 2000) was used as a reference.

STEPWISE DISCRIMINATION

Stepwise discrimination is a dimensionality reduction technique. It selects a subset of variables so that their ability to discriminate between two categories of observations is the highest. The calculation is executed in steps (hence the name). In every step it measures every variable's (possible) contribution to the discrimination using the index variable F, which is a function of the ratio of Wilks' lambda for one-way MANOVA-s with and without the variable in question:

$$F = \frac{n-g-j}{g-1} \cdot \left(\frac{\Lambda_F(j)}{\Lambda_F(j+1)} - 1 \right) \text{ for including and}$$

$$F = \frac{n-g-j+1}{g-1} \cdot \left(\frac{\Lambda_B(j-1)}{\Lambda_B(j)} - 1 \right) \text{ for excluding,}$$

where j is the number of variables currently in the model, g is the number of populations (=2 here), n is the number of observations and the argument of Wilks' lambda denotes the number of variables in the model considered. The statistical significance of each variable is calculated and they are added or removed based on that. We use $p \geq 0.15$ as the condition. To calculate p the procedure assumes multivariate normality, which is not true for our data. However, we expect this does not affect the order of the variables, only how long a list will we get. When no more changes are possible (or necessary), the procedure ends. There are three versions of this method termed forward, backward and stepwise selection. The first starts with zero variables and builds the model by adding those satisfying the criteria. The second does the opposite, starting with all the variables and discarding those below the line. The third starts with zero variables too, but test for both adding and removing in each step. The only reason to use the former two alternatives would be if that part of the procedure ran too slow, but we did not encounter such problem. Another part of the calculations (correlation tables of the ~12000 variables), however was done with a separate program specifically created for this purpose because of technical difficulties in SAS.

The F values gained in the last step were also used to rank the selected variables (genes) where necessary. A high F value means we lose much from the discriminatory power of our model if we exclude the variable and thus implies greater importance.

We use the output of this method as input for multiple techniques. First, we base our predictions with LDA and logistic regression on the variables selected with stepwise discrimination. Furthermore, we hypothesize that the list of genes selected this way can serve as a basis to reveal connections between drug effect categories and genes or groups of genes.

Linear discriminant analysis

LDA is used to create a linear discriminating function on the explanatory variables using which two groups can be separated. It computes a new axis on which the groups differ the most (have the least variance and the most distance between the averages). The function is defined with the equation of the axis and the probability of an observation belonging to a group can be calculated from the distances of the point representing it and the centers of the groups. If we choose a cutoff value, that will define a hyperplane separating the groups.

Predictions were based on the variables selected with stepwise discrimination, and were created including the first 25 and 15 by F value for the given effect category.

Logistic regression

Logistic regression is used to estimate the probability that the object of interest, which we describe with a series of variables, has a certain attribute. The 600 recognized drugs are used as the training set (in which all observations are already categorized) to create a linear function on the describing variables (parameters are calculated with maximum likelihood estimation). The result is then transformed into the [0;1] interval using the logistic function. This is both normalization and allows for the result to be interpreted as a probability. The choice to use this particular function is, while conventional, purely intuitive. Others are used for these purposes too, particularly the probit function.

We apply the function gained to all the observations (1294 perturbagenes). For the training set we assume that most of the drugs are properly categorized and the few miscategorized ones will not alter the function significantly. Thus we expect to see “false” positives, marking probable new effects.

Predictions were based on the variables selected with stepwise discrimination, and were created including the first 50, 25 and 15 by F value for the given effect category.

ERROR MESSAGES

Logistic regression is not based on an exact calculation, the regression parameters are approximated through a series of iterations. Because of this, problems can arise when running the procedure that could not have been predicted beforehand. SAS reports such problems in the log file. If there was no hold-up, we get the message that:

```
‶Convergence criterion (GCONV=1E-8) satisfied.″
```

It is possible that the parameters do not converge based on the data, meaning a maximum likelihood estimate cannot be reached. To limit runtime, a maximum number of iterations is set. The log reports that:

```
‶Convergence was not attained in 25 iterations.″
```

In some situations the parameters reach a point where the likelihood cannot be improved with the method used and thus get “stuck” before convergence could be reached. In the log the following appears:

```
‶Ridging has failed to improve the loglikelihood.″
```

Another way for the iteration to be forced to stop prematurely is to find a set of parameters which provide complete separation of the sample. This means that the two groups can be divided based on the predicted values without overlaps. While this is the aim of the procedure, in this case it is reached before convergence. Because of the separation the iteration stops, meaning the parameters do not represent the approximation attainable from the data, but the first case of separation encountered by the program. Thus the model will perform well on this sample, but may have trouble with observations outside of it. In the log we find that:

```
"There is a complete separation of data points. The maximum
likelihood estimate does not exist."
```

After the error messages, SAS always warns the user, that:

```
"Results shown are based on the last maximum likelihood
iteration. Validity of the model fit is questionable."
```

Ten-fold cross-validation

In cross-validation we separate our originally classified observations (here: the 600 drugs) into a training and a test set randomly, perform the prediction many times and summarize the results. We use the so-called ten-fold cross validation method (TFXV). We divide the observations into ten approximately equal parts randomly and use each as the test set once with the other nine as the training set. The division and prediction is repeated 100 times and the mean (MPV – mean probability value) and standard deviation of the predicted probability values are calculated for each observation. MPV is informative of one drug-effect pair. Since the values are calculated from less data and many repeats get averaged, this is indicative of the robustness of our model. It is also quite sensitive to overfitting: as the observation being evaluated does not contribute to the model, MPV will drop (data not shown).

To describe the quality of the method examined on a given (effect) category, we calculate the mean of the MPV-s (MMPV), separately for those drugs that are registered for an effect and those that are not (these types of observations are also referred to as events and nonevents, respectively).

Receiver Operating Characteristic Analysis

ROC analysis is used to measure a technique's ability to predict a binary attribute of observations. The observations are sorted by the predicted probability that they have the attribute, then for every possible cutoff value the true positive and false positive rates ($TPR = \frac{\#(events\ predicted\ positive)}{\#(events)}$, $FPR = \frac{\#(nonevents\ predicted\ positive)}{\#(nonevents)}$) are calculated and these pairs are plotted. One of the most telling parameters of this curve is the area under it (AUC). If the method tested predicts completely randomly then with a change of the cutoff value there is an equal chance that TPR or FPR will change (if the observation “skipped” is positive then TPR, otherwise FPR; with random prediction there is a 0.5 chance the next value is positive). This will approximately yield a line with a slope of 1 and the area under it will be 0.5. If the prediction is perfect, first the TPR will rise from 0 to 1 while we get all the positives over the cutoff with no change in FPR. After that we start to include negatives over the cutoff and the FPR will rise. The AUC in this case will be 1 (see figure 8).

It should be noted that the AUC in itself is not sufficient to measure the performance of the prediction. Particularly AUC is not sensitive to overfitting.

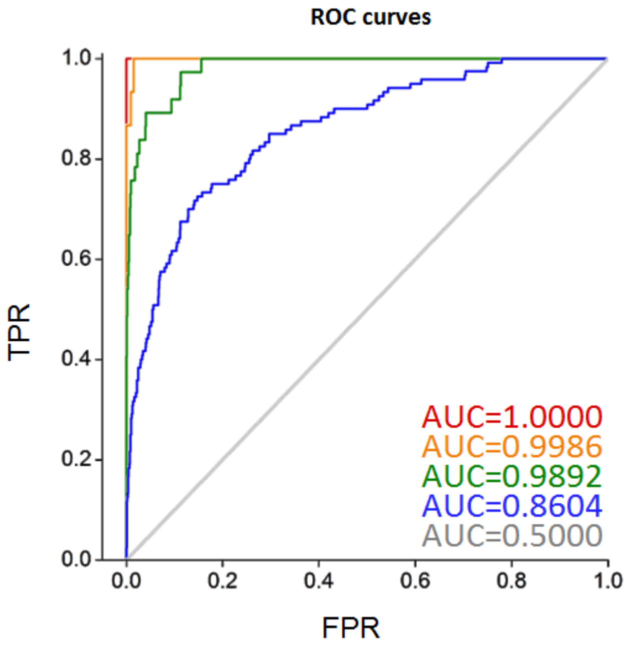

Figure 8: **Receiver Operating Characteristic curves, with the area under them (AUC) also displayed. TPR=True Positive Rate, FPR=False Positive Rate.**

RESULTS AND DISCUSSION

PROCESSING CMAP DATA

VALUES

The values in the tables originally shared with us were average perturbed level/control level ratios, but before analysis we took the natural logarithm of these values. We did this so that the suppression and enhancement of expression are treated equally. For example if we compare two substances, one increasing the expression of a gene 2-fold and another decreasing it 2-fold, we want them to have the same weight in our analyses. The values associated with these effects are not at equal distances from the unit in the original data as $\left|\frac{1}{2}-1\right| \neq |2-1|$. But with this transformation: $\left|ln\left(\frac{1}{2}\right)-ln(1)\right| = |ln(2)-ln(1)|$, and the same holds true for any similar relation. This step was also taken by others working with CMAP datasets (Zhang and Gant, 2008).

SELECTING INSTANCES

From the 3 cell lines prominently used in CMAP02 we chose to work with only the one they did the highest number of tests on, MCF7. While this lowers the amount of data accessible we have two reasons for doing so: first, handling even one 12000-variable data set proved to be challenging because of technological limitations (see *Technical issues*, pg. 46); second, we have no information on how results gained from different cells are related. Thus, merging data gained from different cells would be practically unfavorable, while keeping them separate would create interpretation problems we did not wish to address in these early stages of the work on the subject.

Since in the table listing the experimental conditions used in CMAP02 [i4] there is no uniform identifier listed for the substances used, filtering for drugs was based on the common name of the perturbagenes (variable 'cmap_name'). They were compared to the names listed in our effect database, both formatted lowercase. It is highly unlikely that different molecules are listed with exactly the same name, so we consider all the 600 matches correct (list available in the appendix). It is possible that there are more to be found (e.g. listed under alternate names). However, this knowledge is not necessary in the current stage of the project and so the time consuming work of manual checking was not performed yet. Where multiple instances were listed for a substance one was chose from them, based on concentration used. The aim was to choose a medium (not too high, not too low) concentration. To this end, the highest concentration under 45µM was selected.

SELECTING PROBE SETS

To assign biological meaning to the probe sets, their connection to genes was investigated. The annotation of the microarray lists both UniGeneID and EntrezGeneID [i2]. In previous studies, both were used for similar purposes (Uni: Feng et al., 2009 Entrez: Williams, 2012). We chose EntrezGeneID because that is the one the Gene Ontology database uses [i5]. Two factors were considered to assure the specificity of the probe sets. First, some sets have multiple GeneID-s associated with them. If results are interpreted manually, these could be left in and, if they have relevance, examined later. As we expect to get larger list of genes though, we do not interpret results on that level and thus need to be able to unequivocally annotate our list. Multiple ID-s would lead to ambiguity and complicate implementation. Second, Affymetrix uses three different suffixes of the probe set ID to indicate non-unique sets [i3]. "_a" indicates probe sets that recognize alternate transcripts from the same gene, but is not used in the array CMAP02 is generated with. "_s" sets hybridize with products from different genes. "_x" means even lower specificity. We excluded all "_x" sets and included "_s" only if no unique set was available for the given gene. (Sets with no EntrezGeneID linked and control sequences were also, obviously, excluded.) If, after applying the above filters, there still were multiple sets associated with the same gene, the one with the most variance on the observations recognized as drugs was selected.

From the 22283 probe sets 12261 were selected as representatives of genes (of that 3416 have an "_s" suffix).

PROCESSING GO DATA

To be able to analyze categories higher in the hierarchy, we had to “reformat” the gene annotation data downloaded from the Gene Ontology website [i5]. As mentioned previously, in that genes are associated only with the most precise GO term available for each of their attributes. For example, if a protein coded by a gene is proved to be an estrogen receptor (GO:0030284), we would not find it in GO:0003707 “Steroid hormone receptor activity”. To solve this, another file, listing every edge in the directed graph of the relations between terms was acquired [i6]. Using that the notation was extended to fit our purposes (see figure 9).

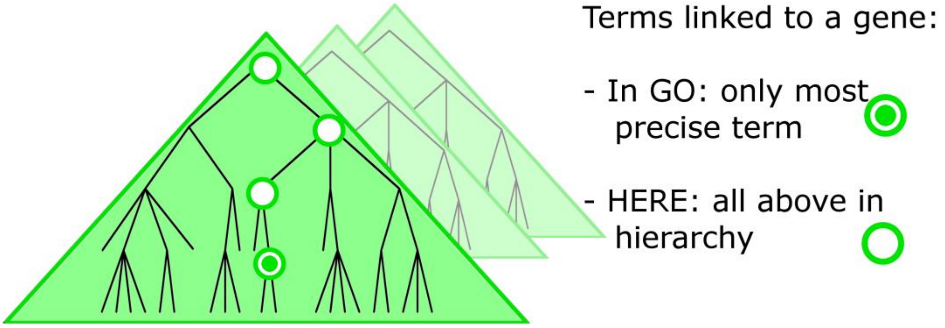
Terms linked to a gene:
- In GO: only most precise term
- HERE: all above in hierarchy

Figure 9: **Sketch of the structure of the GO database, with the changes applied marked.**

While GO is hierarchical, levels are not defined in the database. Since the categories on different “branches” differ in how broad they are, connecting them this way would yield groups that have no biological meaning. Creating levels so that they have such meaning would need a human decision regarding every category and is thus not a viable option. Still, we need a way to roughly group categories based on their size and so define levels based purely on the graph of relations. For each type the level 1 category is the one containing all genes (number 3674 for Molecular Function, 8150 for Biological Process and 5575 for Cellular Component). For every category the level is defined as the length of the shortest path leading to the level 1 category plus 1.

PREDICTIONS

COMPARING METHODS

To determine which prediction technique should we trust the most and use on the non-drug (or not recognized) perturbagenes we tested many variations.

First, the number of variables used. In all cases variables were selected with stepwise discrimination, ranked by F value the first n was used. We have tested how many variables logistic regression can handle. In tests with more variables many runs of the procedure ended with error messages. To get a clearer picture, we ran ten-fold cross-validation with 5 repeats (71 effects * 10 parts * 5 repeats = 3550 runs of the procedure across all effect categories) and tallied the different outcomes of proc logistic (see table 2). As in other cases we use this cross-validation method to ensure the robustness of our results.

Table 2: Outcomes of logistic regression. Data from a 5 repeat test run of ten-fold cross-validation. No case of separation was observed. Please refer to Methods (pg. 26) for the explanation of the outcomes.

#(variables)	Convergence		No conv. in 25 iter.		Ridging failed	
50	348	9.8%	2923	82.3%	279	7.9%
25	1533	43.2%	1819	51.2%	198	5.6%
15	3404	95.9%	57	1.6%	89	2.5%

The parameters reaching convergence is the “proper” way for the procedure to end and means the resulting model is reliable. The ratio of such successful runs is too low with 50 or 25 variables. From this we conclude that performing logistic regression with considerably more variables than 15 is unadvised in our case. We will get back to this problem in “*Concerns about EPV*” (pg. 37).

As the next step we calculated the AUC of ROC for three different methods: logistic regression with 15 variables and LDA with 15 and 25 variables (see figure 10). We chose this few variables for LDA so that it can be compared to the results from logistic regression, but also, as a rule of thumb, a number more than 10% of the number of observations is to be avoided as it may lead to overfitting more easily. AUC is the attribute of a category (here: drug effect) and is a measure of the predictive power of the model in question.

A random prediction’s AUC would average at 0.5, and one providing complete separation would score 1.0. With no category under 0.74, all three show decent power in every effect category. The models based on 15 variables have a similar distribution while LDA(25) stands out with >60% over 0.97.

After this, we performed a “complete” (100 repeat) TFXV for the three methods. We use MPV-s to evaluate the results’ robustness and to test for overfitting, mainly in the case of LDA(25) as that has a high number of categories that get an AUC near 1. MPV-s’ means over events and nonevents (MMPV-s) describe the effect in question (see figure 11). As a rule of thumb, an MMPV over 0.5 can be considered good for events. For logistic regression, the distribution of event-MMPV-s peaks at 0.25, which is rather low. On the other hand, with LDA(15) a decent 59% of effects are ≥0.5 and with LDA(25) a considerable 76% scores over half.

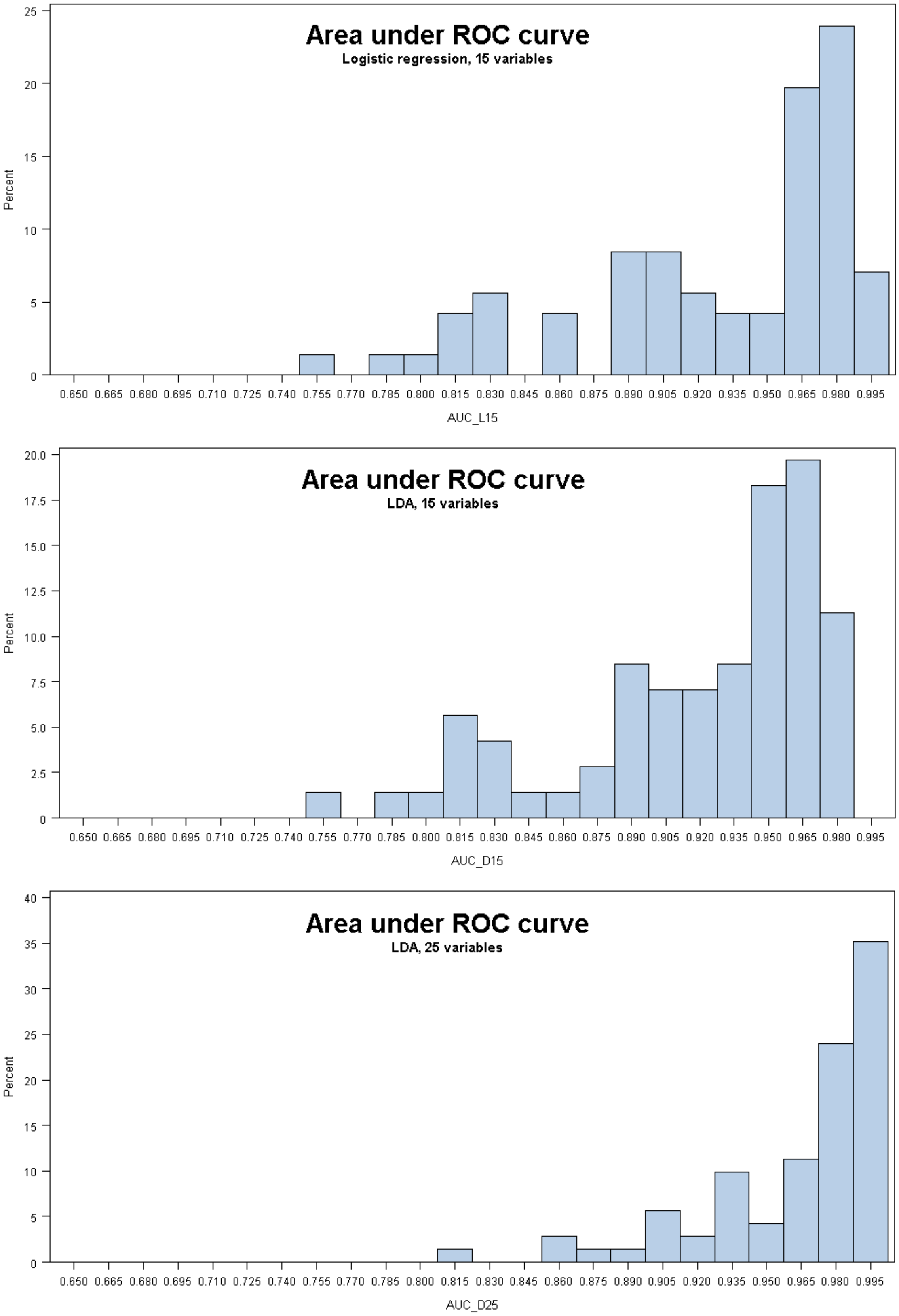

Figure 10: **Distribution of ROC-AUC gained from three different methods: logistic regression with 15 variables (L15), LDA with 15 and 25 variables (D15, D25). Random selection would yield 0.5, complete separation gives 1.0.**

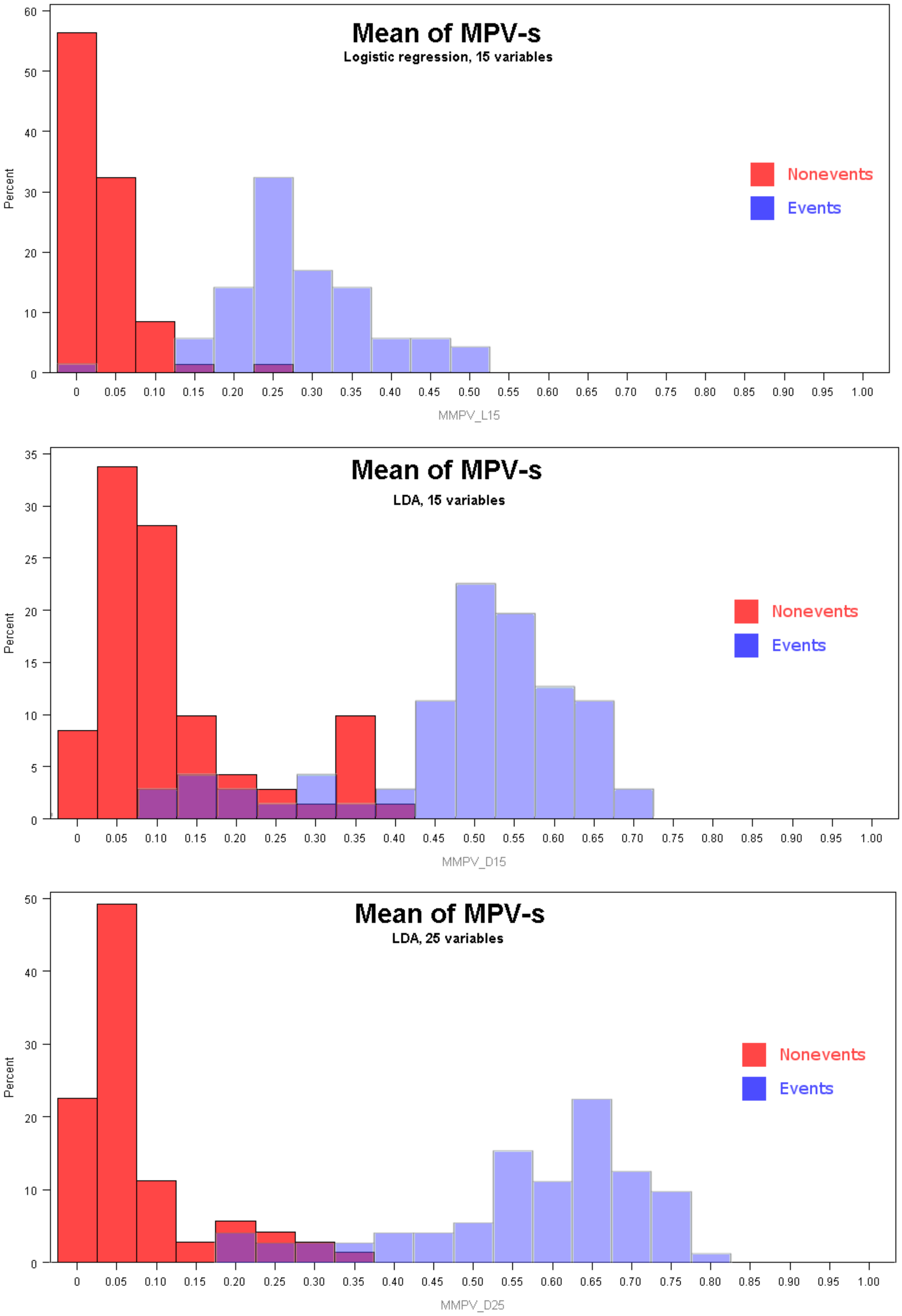

Figure 11: Distribution of MMPV gained from three different methods: logistic regression with 15 variables (L15), LDA with 15 and 25 variables (D15, D25). Event: observation originally classified positive, here: drug registered for the effect category. Nonevents: all other observations.

Based on the above results we chose to use linear discriminant analysis with 25 variables to acquire predictions for the rest of the perturbagenes (the ones not recognized as drugs, 694 molecules) (see figure 12). Results are available in the appendix. We choose to refer to this method as ECPEC, for Expression Change based Prediction of Effect Categories. Proper evaluation of them, with in vitro experiments, is outside of the scope of this work, but is part of our future plans.

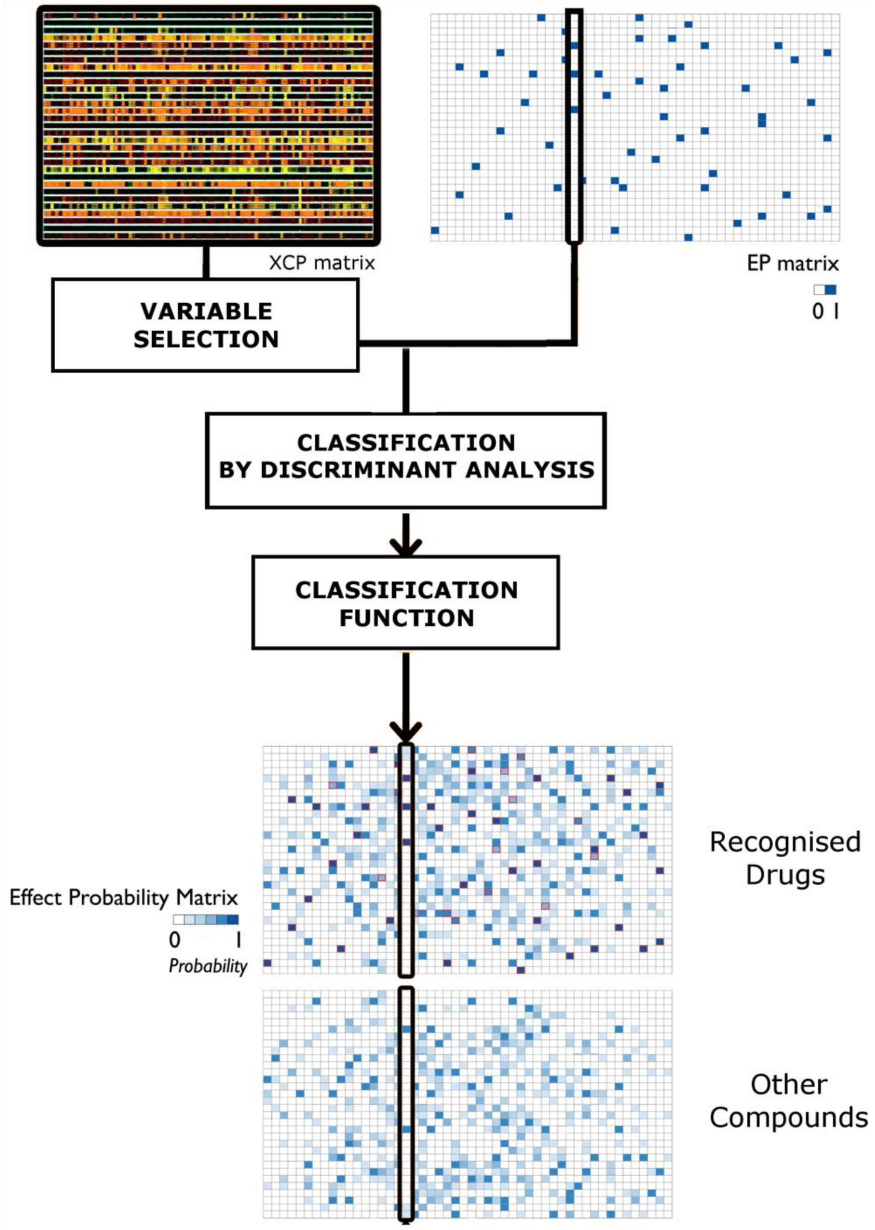

Figure 12: **Scheme of ECPEC – Expression Change based Prediction of Effect Categories**

CONCERNS ABOUT EPV

It has been reported that a low (<10) ratio of originally positive observations (events) to the number of variables (usually referred to as EPV – event per variable) can have adverse effects on the quality of the prediction with logistic regression (Peduzzi et al., 1996). This is usually counteracted by lowering the number of variables used. However, a large portion of our categories have less than 30 events (see figure 13), so in our case this would usually mean that we can select up to 2 variables to use for prediction. The power and usefulness of such a setup is at the least questionable.

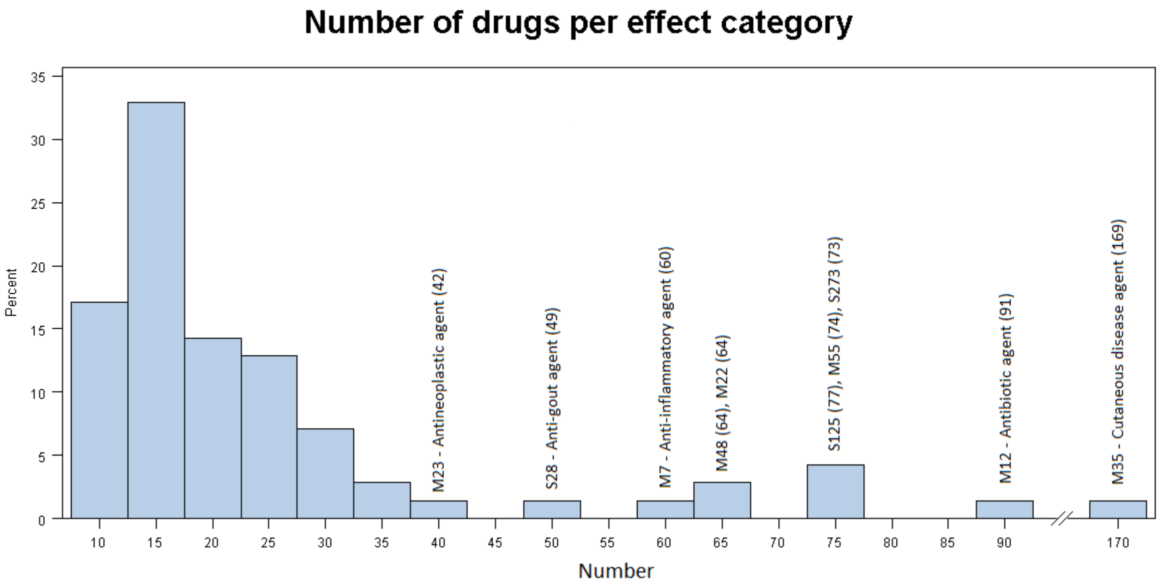

Figure 13: Distribution of effect categories by how many drugs of the 600 are registered under them. Large categories are identified too. S125 is „Cutaneous disease agent. Antibiotic agent", M55 is „Sedative and/or hypnotic agent", S273 is „Sedative and/or hypnotic agent. Not otherwise specified", M48 is „Neurodegenerative disease agent", M22 is „Antihypertensive agent"

For now, logistic regression is performed with 15 variables, but we will have to account for the issue. There are three possible solutions to this:

1. To prove that the conclusions we wish to derive from our results are not or just minimally dependent on the problematic factors.
2. To find out my above assumption is wrong and even with few variables the procedure is efficient.
3. To include more instances in the analysis, not just one per drug. Maybe even use multiple cell lines.

Theoretically, there is one more option, namely to use a bigger database with a higher number of observations (like SPIED (Williams, 2012)). The problem with that is that event count is not much higher in the whole effect database either, especially for subcategories. More than 95% of subcategories have ≤40 drugs listed, and more than 90% has ≤28. The main categories fare a bit better with 35 (53.8%) of them including ≥40 drugs. Combining this with solution #3 above could prove interesting though.

COMPARING WITH DPM

Since the methods used and thus the measures for the evaluation of the quality of the outcome are similar, DPM presents us with an opportunity to properly compare our results with another study (Peragovics et al., 2012). They have not calculated the mean of MPV-s for nonevents, so we are left with two indices, AUC and the MMPV-s of events. In the following tables we show the performance of DPM and ECPEC side by side and sorted according to the difference between them and highlight the effect categories that seem to favor one over the other (see tables 3 to 5, data available in the appendix). There is no theoretical basis to expect a correlation between the two. The Spearman rank correlation coefficient is 0.5423 for AUC-s and 0.3208 for MMPV-s, both positive but not outstandingly high, which can prompt us to expect some decent differences with a background of overall similar tendencies (see figures 14 and 15).

Table 3: **Comparision of AUC-s from the DPM and ECPEC (LDA with 25 variables). Coloring is based on the values (minimum: red, maximum: green). Sorted by the difference, starting with the categories the CMAP data based prediction performed better on.**

Effect category	ID	AUC		difference
		DPM	ECPEC	
antineoplastic agent	EM00023	0.88	0.981	0.101
antiasthmatic agent	EM00011	0.94	0.996	0.056
antiarrhythmic agent	EM00010	0.94	0.987	0.047
alzheimer disease agent	ES00013	0.94	0.986	0.046
serotonin agent	EM00056	0.95	0.995	0.045
obstipant	EM00049	0.95	0.988	0.038
sexual hormone and sexual activity agent	EM00057	0.93	0.968	0.038
antihyperlipidemic agent	EM00021	0.95	0.983	0.033
sclerosis multiplex agent	ES00272	0.95	0.982	0.032
anti-glaucoma agent	EM00006	0.96	0.992	0.032
muscarinic antagonist	ES00223	0.96	0.991	0.031
cutaneous disease agent. antifungal agent	ES00126	0.97	0.994	0.024
antifungal agent	EM00018	0.97	0.994	0.024
antiparasitic agent	EM00024	0.97	0.991	0.021
cholinolytic	EM00032	0.97	0.990	0.020
sedative and/or hypnotic agent. nos	ES00273	0.9	0.920	0.020
gastrointestinal ulcer agent	EM00041	0.98	0.997	0.017
central striated muscle relaxant	ES00119	0.98	0.996	0.016
calcium and bone metabolism agent	EM00030	0.98	0.994	0.014
tardive dyskinesia agent	ES00301	0.98	0.994	0.014
striated muscle agent	EM00059	0.97	0.983	0.013
cutaneous disease agent. antihistamine	ES00127	0.98	0.992	0.012
parkinson disease agent	ES00250	0.96	0.972	0.012
immunosuppressive agent	EM00045	0.97	0.982	0.012
antitussive and expectorant	EM00027	0.98	0.990	0.010
neurodegenerative disease agent	EM00048	0.88	0.889	0.009
primer headache treatment	EM00053	0.9	0.907	0.007
anti-inflammatory agent	EM00007	0.93	0.936	0.006
antiepileptic agent	EM00017	0.97	0.976	0.006
antianginal agent	EM00009	0.93	0.936	0.006
antihistamine	EM00020	0.96	0.966	0.006
carbohydrate metabolism agent	EM00031	0.98	0.985	0.005
heart failure agent	EM00042	0.96	0.963	0.003
norepinephrine liberation blocker or stimulant	ES00232	0.99	0.992	0.002
cutaneous disease agent. immunosuppressive agent	ES00130	0.96	0.961	0.001
antiprotozoal agent. nos	ES00079	0.98	0.981	0.001
smooth muscle agent	EM00058	0.96	0.960	0.000

Effect category	ID	AUC		difference
		DPM	ECPEC	
cell proliferation agent. hormone system associating agent	ES00116	0.99	0.989	-0.001
adrenal gland hormone	EM00001	0.98	0.979	-0.001
non-selective beta receptor antagonist	ES00231	1	0.998	-0.002
anesthetic agent. local	EM00003	0.99	0.987	-0.003
histamine h1 receptor antagonist	ES00187	0.98	0.975	-0.005
glucocorticoid	ES00176	1	0.994	-0.006
anti-inflammatory agent. glucocorticoid	ES00030	1	0.994	-0.006
anti-cytokine agent	ES00024	1	0.994	-0.006
antimigraine agent. prevention	ES00075	0.95	0.944	-0.006
antidepressant and antimanic agent	EM00014	0.97	0.963	-0.007
dopamine d2 receptor antagonist	ES00154	1	0.992	-0.008
beta-lactam antibiotic. cephalosporin	ES00092	1	0.992	-0.008
prokinetic agent	EM00054	0.98	0.972	-0.008
antihypertensive agent. diuretic	ES00069	0.99	0.980	-0.010
sympatholytic	EM00060	0.95	0.940	-0.010
standard antipsychotic. phenothiazine	ES00291	1	0.987	-0.013
beta-lactam antibiotic. penicillin	ES00094	1	0.986	-0.014
sympathetic blocker	ES00296	0.97	0.955	-0.015
antimalarial agent	ES00071	0.99	0.975	-0.015
antihypertensive agent	EM00022	0.93	0.912	-0.018
antiprotozoal agent	EM00025	0.97	0.946	-0.024
antiemetic agent	EM00016	0.97	0.937	-0.033
sedative and/or hypnotic agent	EM00055	0.9	0.861	-0.039
nsaid. non-selective cox inhibitor	ES00226	0.98	0.935	-0.045
anti-inflammatory agent. nsaid	ES00031	0.98	0.935	-0.045
antipsychotic	EM00026	0.98	0.934	-0.046
cutaneous disease agent	EM00035	0.86	0.814	-0.046
sympathomimetic	EM00061	0.95	0.903	-0.047
anti-gout agent	ES00028	0.96	0.911	-0.049
diuretic	EM00039	0.97	0.918	-0.052
cutaneous disease agent. antibiotic agent	ES00125	0.95	0.871	-0.079
antibiotic agent	EM00012	0.94	0.860	-0.080

Table 4: **Comparision of MMPV-s from the DPM and our study (LDA with 25 variables). Coloring is based on the values (minimum: red, maximum: green). Sorted by the difference, starting with the categories the CMAP data based prediction performed better on.**

Effect category	ID	MMPV (events)		difference
		DPM	ECPEC	
antiprotozoal agent. nos	ES00079	0.04	0.648	0.608
antiparasitic agent	EM00024	0.02	0.563	0.543
alzheimer disease agent	ES00013	0.15	0.675	0.525
antiarrhythmic agent	EM00010	0.31	0.760	0.450
cutaneous disease agent. antifungal agent	ES00126	0.33	0.728	0.398
antifungal agent	EM00018	0.33	0.724	0.394
parkinson disease agent	ES00250	0.3	0.680	0.380
gastrointestinal ulcer agent	EM00041	0.44	0.735	0.295
serotonin agent	EM00056	0.45	0.744	0.294
calcium and bone metabolism agent	EM00030	0.53	0.800	0.270
antidepressant and antimanic agent	EM00014	0.42	0.641	0.221
antihyperlipidemic agent	EM00021	0.45	0.671	0.221
norepinephrine liberation blocker or stimulant	ES00232	0.08	0.300	0.220
antiemetic agent	EM00016	0.38	0.593	0.213
antiprotozoal agent	EM00025	0.17	0.374	0.204
prokinetic agent	EM00054	0.35	0.550	0.200
antihypertensive agent. diuretic	ES00069	0.38	0.564	0.184
primer headache treatment	EM00053	0.38	0.559	0.179
antineoplastic agent	EM00023	0.49	0.664	0.174
cutaneous disease agent. antihistamine	ES00127	0.58	0.752	0.172
striated muscle agent	EM00059	0.35	0.510	0.160
central striated muscle relaxant	ES00119	0.38	0.508	0.128
neurodegenerative disease agent	EM00048	0.49	0.617	0.127
sympatholytic	EM00060	0.5	0.627	0.127
obstipant	EM00049	0.45	0.567	0.117
antimigraine agent. prevention	ES00075	0.43	0.546	0.116
sympathetic blocker	ES00296	0.47	0.579	0.109
sedative and/or hypnotic agent. nos	ES00273	0.61	0.714	0.104
antihypertensive agent	EM00022	0.55	0.645	0.095
carbohydrate metabolism agent	EM00031	0.48	0.568	0.088
antiasthmatic agent	EM00011	0.39	0.476	0.086
heart failure agent	EM00042	0.46	0.545	0.085
diuretic	EM00039	0.4	0.473	0.073
antihistamine	EM00020	0.59	0.661	0.071
sclerosis multiplex agent	ES00272	0.19	0.259	0.069
anti-inflammatory agent	EM00007	0.63	0.691	0.061
sexual hormone and sexual activity agent	EM00057	0.6	0.656	0.056
antimalarial agent	ES00071	0.15	0.200	0.050
dopamine d2 receptor antagonist	ES00154	0.26	0.307	0.047

Effect category	ID	MMPV (events)		difference
		DPM	ECPEC	
nsaid. non-selective cox inhibitor	ES00226	0.6	0.632	0.032
cholinolytic	EM00032	0.6	0.631	0.031
anti-inflammatory agent. nsaid	ES00031	0.6	0.628	0.028
cutaneous disease agent	EM00035	0.6	0.608	0.008
anti-gout agent	ES00028	0.68	0.674	-0.006
anesthetic agent. local	EM00003	0.45	0.442	-0.008
antitussive and expectorant	EM00027	0.27	0.242	-0.028
sympathomimetic	EM00061	0.58	0.552	-0.028
sedative and/or hypnotic agent	EM00055	0.64	0.608	-0.032
tardive dyskinesia agent	ES00301	0.66	0.626	-0.034
immunosuppressive agent	EM00045	0.76	0.710	-0.050
cutaneous disease agent. immunosuppressive agent	ES00130	0.74	0.667	-0.073
antibiotic agent	EM00012	0.72	0.643	-0.077
cutaneous disease agent. antibiotic agent	ES00125	0.71	0.620	-0.090
histamine h1 receptor antagonist	ES00187	0.64	0.521	-0.119
antianginal agent	EM00009	0.5	0.378	-0.122
muscarinic antagonist	ES00223	0.68	0.543	-0.137
cell proliferation agent. hormone system associating agent	ES00116	0.52	0.381	-0.139
smooth muscle agent	EM00058	0.48	0.337	-0.143
anti-inflammatory agent. glucocorticoid	ES00030	0.91	0.726	-0.184
glucocorticoid	ES00176	0.91	0.726	-0.184
anti-cytokine agent	ES00024	0.91	0.724	-0.186
beta-lactam antibiotic. penicillin	ES00094	0.79	0.602	-0.188
non-selective beta receptor antagonist	ES00231	0.8	0.595	-0.205
antipsychotic	EM00026	0.69	0.467	-0.223
beta-lactam antibiotic. cephalosporin	ES00092	0.87	0.644	-0.226
standard antipsychotic. phenothiazine	ES00291	0.94	0.714	-0.226
anti-glaucoma agent	EM00006	0.43	0.182	-0.248
adrenal gland hormone	EM00001	0.7	0.406	-0.294
antiepileptic agent	EM00017	0.54	0.186	-0.354

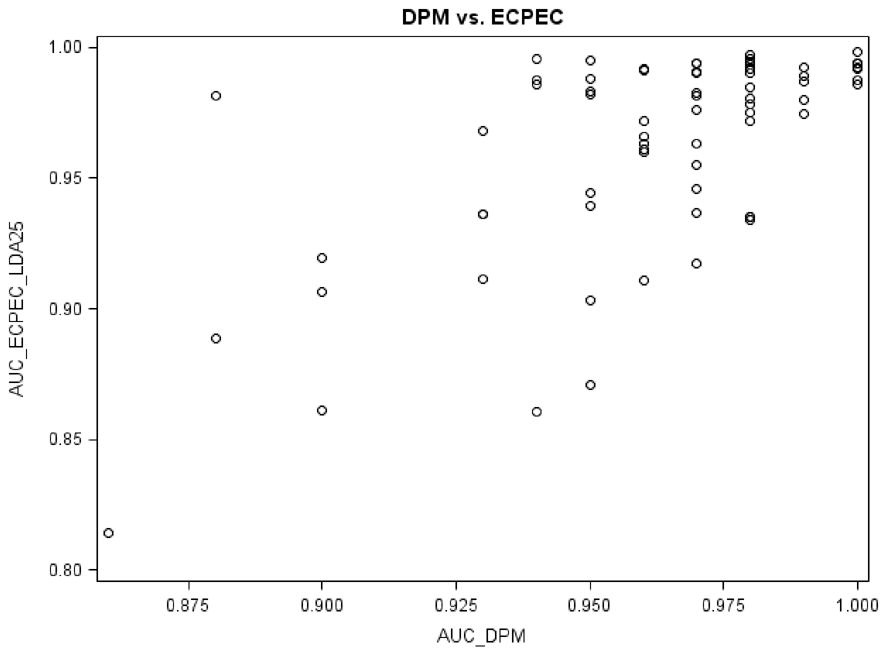

Figure 14: **Comparing AUC-s from DPM and ECPEC (LDA with 25 variables). Spearman rank correlation coefficient r=0.5423.**

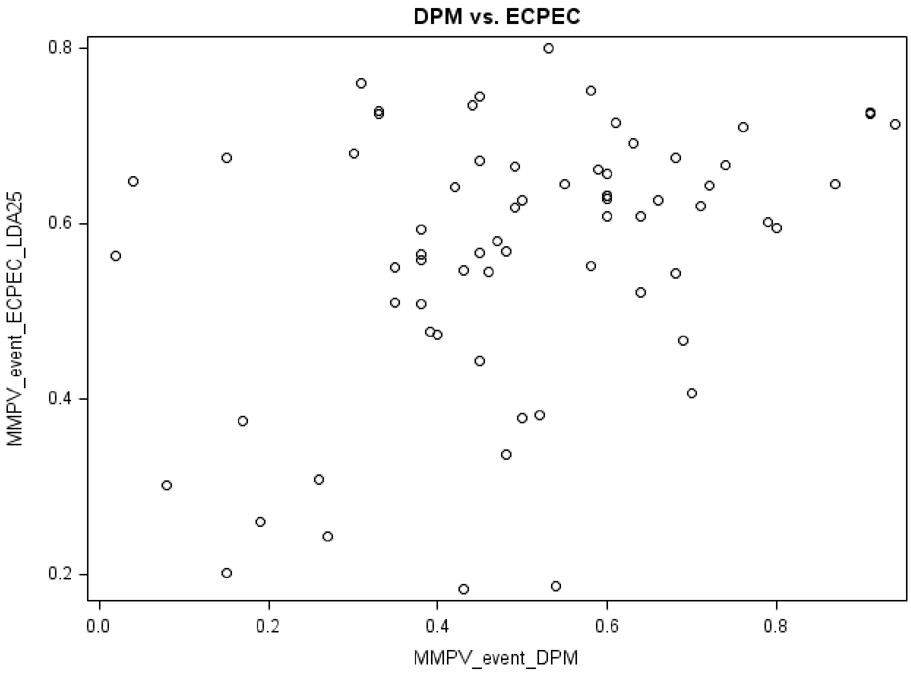

Figure 15: Comparing MMPV-s (of events) from DPM and ECPEC (LDA with 25 variables). Spearman rank correlation coefficient r=0.3208.

Table 5: **Accentuating the 10 effects that favor either method the most. The number is the maximum (worse) of the two ranks of the effect category (for AUC and MMPV), when sorted by difference in performance, like in the above tables.**

In the top X for both AUC and MMPV,			
favoring ECPEC	**X**	**favoring DPM**	**X**
alzheimer disease agent	4	antipsychotic	6
antiarrhythmic agent	4	beta-lactam antibiotic. penicillin	15
serotonin agent	9	standard antipsychotic. phenothiazine	16
antihyperlipidemic agent	12	cutaneous disease agent. antibiotic agent	16
cutaneous disease agent. antifungal agent	12	antibiotic agent	17
antifungal agent	13	beta-lactam antibiotic. cephalosporin	20
antiparasitic agent	14	sedative and/or hypnotic agent	21
gastrointestinal ulcer agent	17	sympathomimetic	22
antineoplastic agent	19	anti-cytokine agent	24
calcium and bone metabolism agent	19	anti-inflammatory agent. glucocorticoid	25

We can see a distinct enrichment of more structure-based effect categories on the DPM-favoring side of the spectrum. This does not come as a surprise as DPM works with structural information as primary input. Most of these effects are also results of drug actions not expected to manifest on breast cells like MCF7 (antibiotics, neural agents). It should be noted, however, that these are not absent from the other side (e.g. antifungal agent) and that many of these have decent AUC and MPV values from the CMAP-based calculations too. Aside from the above mentioned lack of structure-based categories there is no clear tendency emerging from the list of effects favoring the expression data. Another point of interest is that the two calculations were carried out on a different set of drugs and thus the number of elements in each category differs. As such it is better to consider this a preliminary test.

RELATING GENES AND EFFECTS

We attempt to use stepwise discrimination (see pg. 24) to create a list of genes associating with a particular effect category. The genes selected have a response profile similar to the binary variable describing the effect category membership of drugs for the given category. In other words the change in their transcription levels correlates well with whether the drug used had that effect or not. But the list has to be expanded, since lowering redundancy is also an aim of dimensionality reduction techniques. We have to counteract this if we want to get a comprehensive list of associated genes. So we use the selected genes as starting points and "expand the selection" around them by including genes highly correlating with them. "High" was defined as Spearman correlation coefficient r≥0.6 for our first calculations. Based on the results we may consider changing it later. We plan to use the GO database to evaluate our results (see figure 16). How to interpret the results for each GO term-effect category pair is still an open question. The raw output of this method can be found in the appendix.

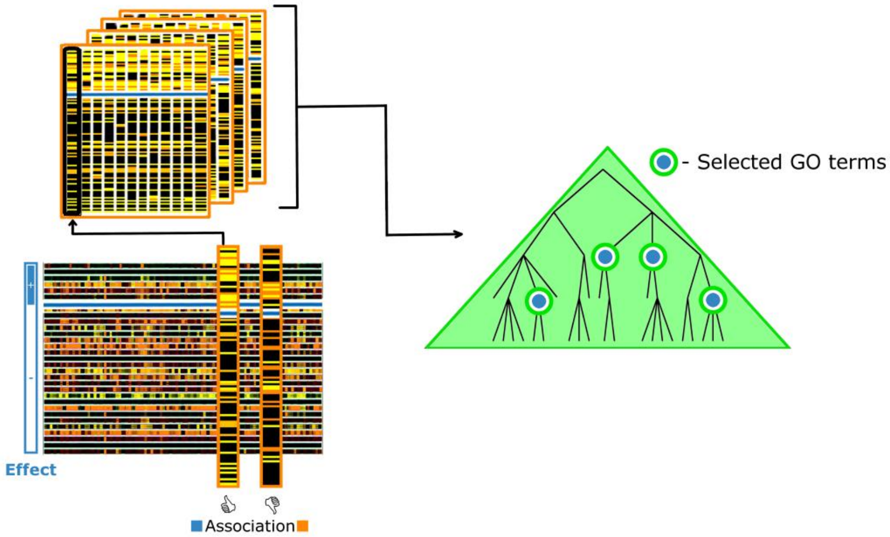

Figure 16: **Connecting GO terms and effect categories. 1. Stepwise discrimination selects gens associating with the effect. 2. List is expanded with highly correlating genes. 3. Result analyzed on the level of GO terms.**

TECHNICAL ISSUES

After detailing our theoretical tools, we feel it is appropriate to mention that dealing with big databases like these is not a trivial matter from a practical point of view either. While the quality of the CPU used affects the speed with which the calculations are completed, the amount of RAM available was a more pressing concern in the course of our work. Memory limits the program’s ability to handle large data sets. The available amount is dependent on two factors: RAM physically present (hardware) and the amount the program, in our case SAS, can handle (software). We started on a regular laptop with SAS 9.1 32bit (4GB RAM), but realized early on that this will be insufficient. The first procedure returning an error message about memory was proc princomp (Principal Component Analysis, a form of dimensionality reduction) when we asked it to calculate the principal components from a roughly 22k variables-x-514 observations matrix (probe sets on drugs (less here because the early screens were faulty)). The new platform includes a more advanced computer (available to us courtesy of Delta Informatika Zrt.), with considerably faster CPU and 18GB RAM, plus SAS 9.2 64bit. Unfortunately even this version of SAS cannot use more than 8GB-s of memory. This setup proved sufficient for most of our needs to date. The only exception so far was the calculation of correlation tables from the filtered CMAP02 matrix (12261x600) needed for stepwise discrimination (inside proc stepdisc or separately with proc corr). With our possibilities in improvement of instruments exhausted, we had to ask a computer programmer to perform this part with a separate program.

FUTURE PLANS

First and foremost we want to further analyze the results gained here with linear discriminant analysis. We want to compare our predictions to DPM calculations specifically created for this purpose, i.e. with the same set of drugs and effect categories. Testing promising predictions in vitro is a logical next step and is a part of the proper validation of the method.

As for gene-effect relations, since utilizing stepwise discrimination this way is a new idea, interpreting the results is not a trivial matter. We will have to test our ideas using well-established connections from the literature as references.

Besides directly reinforcing our initial hypothesis, we can also explore the potential in this platform and ask new types of questions. If we "turn the tables" and regard genes as observations instead of drugs, we could create a tool for the prediction of gene product function. Microarray technology poses no restraint on the type of cell used and treatment administered. If we can derive meaningful observations utilizing only one cell type, it can be possible to predict the effectiveness of drugs on a given patient from primary cell cultures. Instead of examining drugs separately, treating cell lines with combinations of them is not a difficult task. However, the knowledge we can gain about drug interactions is of great clinical importance.

I am certain this list of possibilities is not comprehensive and that adjusting to these new eyes will be a long and greatly rewarding venture.

SUMMARY

Determining the full spectrum of effects a drug will induce when administered to a complete organism is a crucial task in drug design. To deduce something of such complex nature, an equally complex set of data is necessary. After applying high throughput techniques to efficiently acquire this information, the opposite problem arises: understanding the meaning or significance of every single data point is impossible. Pattern based methods offer a solution by handling all the data at once and deriving information from similarities.

We investigated the possibilities of using gene expression data from microarray experiments for this purpose. The CMAP database we use records the change in cells' expression patterns in response to perturbation with small bioactive compounds. From the 22283 variables we identified 12261 specifically representing certain genes and annotated 11494 of them with gene product function using the Gene Ontology database. From the 1294 small molecules used on the cell line we chose to work with, we recognized 600 as active substances of drugs and used a database based on DrugBank to annotate them with their known effects.

We examined linear discriminant analysis' (LDA) and logistic regression's performance on our data set. Logistic regression has shown serious limitations, mainly in the number of variables it can handle. It is most likely that this stems from the low number of positive observations in the data (events, here: drugs registered for an effect). LDA outperformed logistic regression and scored well on both tested indices of prediction quality (ROC-AUC and MPV of cross-validation). We also proposed a way to use a dimensionality reduction technique, stepwise discrimination to reveal connections between drug effects and the expression of (groups of) genes.

The work presented here is considered the first step of a larger project. While it does not answer every question, it furthers development with exploring the prospects of using gene expression data as input for pattern based methods.

ÖSSZEFOGLALÁS

(Summary in Hungarian)

Az utóbbi évtizedek rohamos technikai fejlődése és különösen a nagy áteresztőképességű módszerek megjelenése a tudományos kutatás egy új ágát hívta létre, amelyet “adat alapú” kutatásnak hívunk. A nagy adattömegekből, mint megfigyelésből kiinduló irányzat természetesen ezen a területen új, matematikai és informatikai analitikai módszerek használatát teszi szükségessé. A számítógépek segítségével azonban lehetőségünk nyílik egészen új módon tekinteni a vizsgált rendszerekre.

A gyógyszertervezés egyik nagy kihívása, hogy a molekulák összes hatását (teljes hatásprofilját) felderítse még azok forgalomba hozatala előtt. A hatások szisztematikus térképezése, azok sokfélesége miatt, a klasszikus hatás-specifikus módszerekkel gyakorlatban nem kivitelezhető. Bár megbízhatóságukban a számítógépes módszerek elmaradnak, előzetes szűrésként alkalmazva nagyban javíthatják a kísérletezés hatékonyságát.

A nagy adattömegek értelmezéséhez két módszert mutattunk be: a csoportosítást (vagy kategorizálást) és a mintázatelemzést. Előbbi egyszerre csökkenti az elemek számát és rendel hozzájuk (a feltett kérdés szintjéhez viszonyítva) közvetlenebb információt. Utóbbi pedig teljesen megkerüli a problémát és az összes változót egy egységként kezelve a köztük lévő hasonlóság mértéke alapján von le következtetéseket. Nagy előnye, hogy nem szükséges minden adatpont jelentőségét részletesen ismernünk.

Felmerül a kérdés, hogy milyen mintázat alkalmas a gyógyszerek hatásainak jóslására. Hiszen, bár a kutatónak nem kell mélységeiben tisztában lennie azzal, hogy a változók és a jósolt tulajdonság között milyen kapcsolat van, a kapcsolatnak jelen kell lennie és minél erősebb, annál jobb minőségű eredményt kaphatunk. Azt tapasztaltuk, hogy a gyógyszer és a célsejt egy megfelelően komplex modelljének interakciói megfelelő alapot biztosítanak az ilyen jellegű vizsgálódásokhoz.

Az általunk felhasznált adatbázis DNS-chip technológiával készült (eredetileg a CMAP projekt keretein belül), és azt tartalmazza, hogy az alkalmazott molekulák hatására hogyan változott meg az egyes szekvenciák kifejeződésének szintje, a kontroll és kezelt állapotokban mért értékek hányadosaként. Az összesen 22283 változóból 12261-et sikerült egyértelműen megfeleltetnünk egy génnek, ebből 11494-hez pedig funkciót tudunk rendelni a Gene Ontology adatbázis segítségével. A legtöbb kísérletükben használt sejtvonalhoz 1294 expressziós mintázat-változás profilt rögzítettek, ebből 600 esetben a használt molekulához gyógyszerhatás-profilt tudtunk rendelni.

A két profil közti kapcsolatot lineáris diszkriminancia-analízissel (LDA) és logisztikus regresszióval vizsgáltuk. Utóbbi alkalmazhatóságában jelentős korlátokat ismertünk fel, különösen a felhasználható változók számában. Az tűnik a legvalószínűbbnek, hogy ez a pozitív megfigyelések (események, itt: az adott hatásra regisztrált gyógyszerek) alacsony számából fakad. A diszkriminancia-analízis jól teljesített és felülmúlta az LDA-t a jóslás minőségének mindkét vizsgált mérőszámában (ROC-AUC és a kereszt-validációból nyert átlagos valószínűségi érték (MPV)). A jóslások mellett javaslatot tettünk egy dimenzionalitás-csökkentő módszer (az ún. stepwise discrimination, "lépésenkénti diszkrimináció") alkalmazására a gyógyszerhatások és gének (csoportjainak) kifejeződésének változása közötti kapcsolatok feltárására.

Jelen munkát egy nagyobb projekt első lépésének tekintjük. Bár nem válaszol meg minden kérdést, előremozdítja a kutatást a génexpressziós adatok mintázat-elemzéssel való feldolgozásában rejlő lehetőségek vizsgálatával.

ACKNOWLEDGEMENTS

I want to thank my coworkers Ágnes Peragovics and Zoltán Simon for introducing me to this field and for their help throughout my work.

I want to thank András Málnási-Csizmadia for offering me a position in his group and for providing me with the necessary technical background. I also want to thank him for the opportunity to attend a conference on the subject while still an undergraduate student.

I also want to express my gratitude to the above mentioned people for their invaluable help in creating present thesis, along with Ildikó Radnai who helped in creating the figures.

I am grateful to Delta Informatika Zrt. for being able to borrow their equipment and employees.

I want to thank László Nyitray as Head of the Department for allowing me to work there.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25–29. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract [Accessed March 8, 2012].

Augenlicht, L. H., and Kobrin, D. (1982). Cloning and screening of sequences expressed in a mouse colon tumor. Cancer research *42*, 1088–1093. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7059971 [Accessed April 29, 2012].

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., et al. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic acids research *39*, D1005–10.

Efron, B. (1975). The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. Journal of the American Statistical Association *70*, 892–898.

Feng, C., Araki, M., Kunimoto, R., Tamon, A., Makiguchi, H., Niijima, S., Tsujimoto, G., and Okuno, Y. (2009). GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. BMC genomics *10*, 411. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2748096&tool=pmcentrez&rendertype=abstract [Accessed April 12, 2012].

Freiberg, G., Wilkins, J., David, C., Kofron, J., Jia, Y., Hirst, G. C., Burns, D. J., and Warrior, U. (2004). Utilization of microarrayed compound screening (microARCS) to identify inhibitors of p56lck tyrosine kinase. Journal of biomolecular screening *9*, 12–21. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15006144 [Accessed May 9, 2012].

Hieronymus, H., Lamb, J., Ross, K. N., Peng, X. P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S. M., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. Cancer cell *10*, 321–330. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17010675 [Accessed May 9, 2012].

Khattree, R., and Naik, D. N. (2000). Multivariate Data Reduction and Discrimination with SAS Software (Cary, NC: SAS Institute Inc.).

Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. Nature reviews. Cancer *7*, 54–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17186018.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science (New York, N.Y.) *313*, 1929–1935. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17008526 [Accessed July 6, 2011].

MacBeath, G., and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. Science (New York, N.Y.) *289*, 1760–1763. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10976071 [Accessed March 30, 2012].

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature genetics *34*, 267–273. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12808457 [Accessed March 1, 2012].

Newman, C. G. (1986). The thalidomide syndrome: risks of exposure and spectrum of malformations. Clinics in perinatology *13*, 555–573. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3533365 [Accessed May 9, 2012].

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology *49*, 1373–1379. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8970487 [Accessed March 14, 2012].

Peragovics, A., Simon, Z., Brandhuber, I., Jelinek, B., Hári, P., Hetényi, C., Czobor, P., and Málnási-Csizmadia, A. (2012). Contribution of 2D, 3D structural features of drug molecules in the prediction of Drug Profile Matching. Journal of chemical information and modeling.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000). Genome-wide location and function of DNA binding proteins. Science (New York, N.Y.) *290*, 2306–2309. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11125145 [Accessed March 7, 2012].

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.) *270*, 467–470. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7569999 [Accessed March 25, 2012].

Simon, Z., Peragovics, A., Vigh-Smeller, M., Csukly, G., Tombor, L., Yang, Z., Zahoránszky-Kohalmi, G., Végner, L., Jelinek, B., Hári, P., et al. (2012). Drug effect prediction by polypharmacology-based interaction profiling. Journal of chemical information and modeling *52*, 134–145. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22098080 [Accessed June 10, 2012].

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545–15550. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1239896&tool=pmcentrez&rendertype=abstract [Accessed March 1, 2012].

Wei, G., Twomey, D., Lamb, J., Schlis, K., Agarwal, J., Stam, R. W., Opferman, J. T., Sallan, S. E., den Boer, M. L., Pieters, R., et al. (2006). Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. Cancer cell *10*, 331–342. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17010674 [Accessed May 10, 2012].

Williams, G. (2012). A searchable cross-platform gene expression database reveals connections between drug treatments and disease. BMC genomics *13*, 12. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3305579&tool=pmcentrez&rendertype=abstract [Accessed April 26, 2012].

Zhang, S.-D., and Gant, T. W. (2008). A simple and robust method for connecting small-molecule drugs using gene-expression signatures. BMC bioinformatics *9*, 258. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2464610&tool=pmcentrez&rendertype=abstract [Accessed May 9, 2012].

Zhang, S.-D., and Gant, T. W. (2009). sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures. BMC bioinformatics *10*, 236. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2732627&tool=pmcentrez&rendertype=abstract [Accessed May 9, 2012].

RESOURCES ON THE INTERNET

HG-U133A gene chip

[i1] Manufacturer’s page:

http://www.affymetrix.com/estore/browse/products.jsp?navMode=34000&productId=131537&navAction=jump&aId=productsNav

[i2] Annotation:

http://www.affymetrix.com/Auth/analysis/downloads/na31/ivt/HG-U133A_2.na31.annot.csv.zip

[i3] Technical Note:

http://media.affymetrix.com/support/technical/technotes/hgu133_p2_technote.pdf

Connectivity Map

[i4] List of instances (experiments, with repeats) in the CMAP project:

http://www.broadinstitute.org/cmap/cmap_instances_02.xls

Gene Ontology

The GO database is constantly being improved. We used the 2012.03.19. version of these files. The links below always point to the newest version.

[i5] Annotation of genes with GO categories:

ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz

[i6] Relations of GO categories:

http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo

ABBREVIATIONS

GO – Gene Ontology, refers to the project or the database created by it, see pg. 18

MCF7 – Michigan Cancer Foundation-7, refers to a breast cancer cell line. The gene expression data we use in our studies was created on this cell line, see pg. 12 and 15

LDA – Linear discriminant analysis, one of the techniques with which predictions are made, see pg. 6

ROC – Receiver Operating Characteristic, see pg. 27

AUC – Area Under the Curve, see pg. 27

TPR – True Positive Rate, the ratio of observations classified as positive in all the positive samples.

FPR – False Positive Rate, the ratio of observations classified as positive in all the negative samples.

XCP – eXpression Change Profile, see figure 4 on pg. 12

EPV – Event Per Variable, events are observations listed as positive in the category variable (here: drugs registered for a certain effect) , see pg. 37

TFXV – ten-fold cross-validation, see Methods, pg. 27

MPV – Mean Probability Value, the arithmetic average of the results for one observation (here: molecule) in cross-validation, see pg. 27

MMPV – Mean of MPV-s, usually calculated for events and nonevents separately, used to describe the predictions for a category (here: drug effect), see pg. 27, 33 and figure 11 on pg.35