BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 15 2005, pages 3286–3293
doi:10.1093/bioinformatics/bti515

Data and text mining

A latent variable model for chemogenomic profiling

Patrick Flaherty[1,*], Guri Giaever[3], Jochen Kumm[3], Michael I. Jordan[2] and Adam P. Arkin[4]

[1]Department of Electrical Engineering and Computer Science and [2]Division of Computer Science, Department of Statistics, University of California, Berkeley, CA 94720, USA, [3]Stanford Genome Technology Center, Stanford University School of Medicine, Palo Alto, CA 94304, USA and [4]Department of Bioengineering, University of California and Physical Biosciences Division, Lawrence Berkeley National Laboratory, Howard Hughes Medical Institute, Berkeley, CA 94720, USA

Received on July 28, 2004; revised on May 11, 2005; accepted on May 23, 2005
Advance Access publication May 26, 2005

ABSTRACT

Motivation: In haploinsufficiency profiling data, pleiotropic genes are often misclassified by clustering algorithms that impose the constraint that a gene or experiment belong to only one cluster. We have developed a general probabilistic model that clusters genes and experiments without requiring that a given gene or drug only appear in one cluster. The model also incorporates the functional annotation of known genes to guide the clustering procedure.

Results: We applied our model to the clustering of 79 chemogenomic experiments in yeast. Known pleiotropic genes *PDR5* and *MAL11* are more accurately represented by the model than by a clustering procedure that requires genes to belong to a single cluster. Drugs such as miconazole and fenpropimorph that have different targets but similar off-target genes are clustered more accurately by the model-based framework. We show that this model is useful for summarizing the relationship among treatments and genes affected by those treatments in a compendium of microarray profiles.

Availability: Supplementary information and computer code at http://genomics.lbl.gov/llda

Contact: flaherty@berkeley.edu

1 INTRODUCTION

Haploinsufficiency profiling (HIP) in *Saccharomyces cerevisiae* is used to identify genes that, when deleted, confer sensitivity on small molecules *in vivo*. Specifically, this genome-wide screen employs the heterozygous deletion collection to identify strains that exhibit a significant inhibition of growth in the presence of compound. It has been demonstrated that drug-induced haploinsufficiency is effective in identifying the gene product that is the target of a compound.

The results from studies of HIP have demonstrated its efficacy in uncovering the mechanisms of action of individual compounds (Giaever *et al*., 1999, 2004; Lum *et al*., 2004). In these studies, the *S.cerevisiae* heterozygous deletion library is used to measure the sensitivity of each strain as quantified by the fitness defect score (defined in the Methods section) in response to drug treatment (Giaever *et al*., 2004). The current study focuses on these chemogenomic experiments as a corpus to discover common targets and associated cellular functions for multiple classes of drugs.

*To whom correspondence should be addressed.

The two main contributions of this paper are as follows:

- A statistical model for chemogenomic experiments that incorporates gene function information and properly handles pleiotropic genes.
- An analysis of HIP chemogenomic experiments using the statistical model to summarize the relationship among treatments and genes affected by those treatments.

The dataset used in this study comprises 79 experiments, each measuring the fitness defects of all heterozygous yeast deletion strains in the presence of 1 of 13 different drugs. The assay was performed at a variety of drug concentrations and most concentrations are replicated at least once. In each experiment, all 5918 strains of yeast, each deleted for one copy of a specific gene, was assigned a fitness defect score (Giaever *et al*., 2004).

The drugs used in this study span a wide range of clinical applications. Table 1 shows a brief summary of the compounds and the genes encoding their putative target proteins, if known. The applications of these drugs include cholesterol lowering medications (lovastatin and atorvastatin), anticancer agents [5-fluorouracil (5-FU), methotrexate], an agricultural fungicidal agent (fenpropimorph), systemic and topical antifungal medications [caspofungin, itraconazole, miconazole, fluconazole, amphotericin B and 5-fluorocytosine (5-FC)] and two drugs that have other applications (dyclonine and alverine-citrate).

Numerous computational and experimental studies have used the analysis of compendia to discover consensus patterns in experiments and to uncover functional information (Bergmann *et al*., 2003; Eisen *et al*., 1998; Hughes *et al*., 2000). The underlying statistical problem that is addressed in many of these analyses is that of finding clusters in data. A variety of classical clustering algorithms, including hierarchical clustering (HC) (Eisen *et al*., 1998) and self-organizing maps have been deployed for this purpose.

It is important to note, however, that many clustering algorithms are based on an underlying assumption that each data point belongs to only a single cluster—a mutual exclusivity assumption that may not be a good match to biological reality. In particular, the implicit assumption that a gene can belong to only one cluster clashes with the fact that a gene often has multiple functions in the cell. Similarly, an assumption that a drug can belong to only one cluster neglects

Table 1. Drug targets and applications

Drug	Application	Target	Topic	Replicates
5-Fluorocytosine	Antifungal	Cdc21, RNA, DNA	2	2
5-Fluorouracil	Anticancer	Cdc21, RNA, DNA	2	11
Alverine-citrate	Anticholinergic	Unknown, (Erg24)	6	3
Amphotericin B	Antifungal	Ergosterol	7	4
Atorvastatin	Anticholesterol	Hmg1, Hmg2	3	5
Caspofungin	Antifungal	Cell wall	1	5
Dyclonine	Anesthetic	Erg2	6	6
Fenpropimorph	Plant antifungal	Erg2, Erg24	6	5
Fluconazole	Antifungal	Erg11	5	2
Itraconazole	Antifungal	Erg11	5	8
Lovastatin	Anticholesterol	Hmg1, Hmg2	3	4
Methotrexate	Anticancer	Dfr1	8, 9	16
Miconazole	Antifungal	Erg11	5	8

Thirteen drugs with a wide variety of clinical applications were tested in this compendium (Bennett, 2001; Chabner *et al*., 2001; Coelho *et al*., 2001; Giaever *et al*., 2004; Katzung, 1998). The putative targets in yeast, if known, are shown in the third column. The LLDA model uses an allocation of topics to cluster experiments and these clusters tend to reflect the known class structure of these drugs. The topic(s) most associated to each drug are shown with the number of replicates in the compendium for that drug.

the fact that different drugs can have similar off-target effects. In this paper, we present a model-based approach to clustering that avoids such mutual exclusivity assumptions.

The singular value decomposition (SVD) method has also been used for clustering biological data (Peterson, 2003). While the SVD does not make a 'mutual exclusivity' assumption, it also is not a model-based approach. We return to a comparison with SVD in the discussion.

We refer to the model underlying our clustering methodology as Labeled Latent Dirichlet Allocation (LLDA). The LLDA model is an instance of a general family of probabilistic models, known as probabilistic graphical models. Probabilistic graphical models provide a general Bayesian framework for representing joint probability distributions over collections of variables, and for computing posterior distributions on subsets of those variables (Jordan, 1999). Graphical models have been successfully employed in other applications (Alexandersson *et al*., 2003; Jaakkola and Jordan, 1999; Jansen *et al*., 2003; Segal *et al*., 2003b); the model that we present here is designed specifically to identify consensus targets and mechanisms of drug action in HIP experiments. A component of the model is a set of 'allocation variables' that link experiments and genes, such that the computation of the posterior probability distributions over these variables defines model-based clusters. The framework also facilitates the incorporation of external information into the clustering procedure; in particular, we show how to incorporate Munich Information Center for Protein Sequences (MIPS) functional annotations (Mewes *et al*., 2002) as part of the model.

2 LABELED LATENT DIRICHLET ALLOCATION MODEL

The key entities in the LLDA model are a set of discrete random variables that we refer to as 'topics'. Associated with each topic is a pair of probability distributions: a probability distribution over genes and a probability distribution over MIPS function categories. These distributions, in essence, define the biological meaning of each topic. The notion of 'topic' is similar to the notion of 'cluster' in traditional treatments; we use a different terminology to emphasize that there is no assumption that the topics partition the set of genes. Thus the same gene can have high probability under each of several topics. An experiment is represented by a set of choices, or 'allocations', among the available topics. Here again, there is no mutual exclusivity—the same topic can have a high probability of being allocated to each of several experiments.

After fitting the LLDA model to the experimental data, we find that genes showing a significant growth inhibition or fitness defect in a group of experiments are assigned a high probability to the topics that are allocated to those experiments. Experiments with similar fitness defect profiles have similar distributions over topics in the LLDA model. The genes were ranked according to their probability under each topic in the model and the highly ranked genes comprise the consensus sensitive genes for the drugs allocated that topic.

3 METHODS

3.1 Microarray dataset

As a control set, cells were grown for 20 generations in standard optimal (YPD) medium and genomic DNA was isolated. Using universal PCR primers we amplified barcode tags from the genomic DNA. The PCR product was hybridized to custom Affymetrix TAG3 microarray chips. These 36 standardized control chips provided the average control intensity for each position on the microarray corresponding to a tag in the pool. The specifics of the deletion cassette used and the 79 experiments that were performed by growing the pool of heterozygous deletion strains for 20 generations with the drug being tested are described elsewhere (Giaever *et al*., 2004).

3.2 Data analysis

A fitness defect score for each strain in each experiment was computed by the method described in other work (Giaever *et al*., 2004). The fitness defect score quantifies the square difference between the amount of strain in the population and that expected under standard optimal conditions. The profiles for each experiment were standardized to lie between 0 and 100. Then the fitness defect scores for each gene were rounded to the nearest integer.

The LLDA model can be represented as a graphical model (Jordan, 2003). In this formalism, each node represents a random variable. The connectivity of the graph implies a factorization of the joint probability distribution among the random variables. In Figure 1, we have used black nodes to represent fully observed random variables, gray nodes for partially observed variables and unshaded nodes for unobserved random variables. The box or 'plate' indicates that the structure contained within is replicated the number of times indicated in the corner of the plate. This replication handles the repeated measurements of the data. The number of the fitness defect scores is denoted, N_{d}, and the number of experiments is denoted, D, in the LLDA model.

Each node in the graphical model is endowed with a conditional probability distribution. The central object of the model, the topic, is represented by Z and is an unshaded node indicating it is a latent or unobserved random variable. The fully observed variable S represents a gene and is distributed as a multinomial with parameter conditional on the choice of topic. We denote this conditional probability relationship as $S|Z=i \sim \mathrm{Multi}(\beta_i)$, where β_i is the multinomial parameter associated with topic i. The conditional probability of a partially observed label variable is denoted similarly as $L|Z=i \sim \mathrm{Multi}(\pi_i)$. This variable encodes the MIPS category assigned to the gene represented by the S variable. For some genes the MIPS category is known and for others it is not, rendering the node partially observed. The topic node Z is also distributed as a multinomial with parameter θ, $Z|\theta = \mathrm{Multi}(\theta)$. The remaining unshaded node, θ, is a latent Dirichlet random variable that depends on the parameter α. This variable allows the distribution over topics

Table 2. Top 15 genes in each topic ranked by probability

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	*YBR293W*	*YPR143W*	*HMG1*	*SAM50*	*PDR5*	*ERG24*	*DIA4*	*DFR1*	*DFR1*
2	*NGL1*	*NOP4*	*ERG13*	*MGM1*	*SET6*	*SET6*	*POP1*	*FOL2*	*FOL2*
3	*GLC7*	*YPL044C*	*PDR5*	*YKL023W*	*PDR16*	*TVP18*	*FUN12*	*YBT1*	*FOL1*
4	*FEN2*	*DIS3*	*FEN2*	*DRE2*	*ERG11*	*MAL11*	*GCD2*	*LST4*	*YBT1*
5	*LCB1*	*RRP6*	*UPC2*	*APE3*	*MAL11*	*YPR090W*	*FKS1*	*YKT6*	*YAP6*
6	*CUP5*	*MTR4*	*YGR205W*	*MPS2*	*TVP18*	*SBE22*	*RIT1*	*YOR227W*	*YOR072W*
7	*YBR113W*	*MAK21*	*YLF2*	*YNL184C*	*YOR345C*	*UMP1*	*YAR047C*	*YOR225W*	*WHI2*
8	*YCR025C*	*FUI1*	*NRP1*	*BRE4*	*YDR467C*	*YDR467C*	*ARC18*	*SPI1*	*YLR281C*
9	*YCR024C*	*RRP42*	*FKS1*	*VHT1*	*YPR090W*	*YLR173W*	*SEC28*	*YGR054W*	*YLR312C*
10	*CTP1*	*YPR142C*	*TOA1*	*YDR514C*	*HFI1*	*SCL1*	*GAL1*	*YKL207W*	*SIP3*
11	*IMG2*	*MAK5*	*YCR025C*	*CTP1*	*YOR331C*	*YPR089W*	*MEF2*	*YOR072W*	*CDA2*
12	*RET2*	*MAS2*	*YBR281C*	*YML122C*	*VID22*	*KRE33*	*GCD1*	*GDI1*	*EXG1*
13	*ATG15*	*GLC7*	*EXG1*	*XDJ1*	*NCP1*	*YGR205W*	*SAM4*	*ELC1*	*RPS30A*
14	*YBR284W*	*RRP45*	*YML122C*	*IRA2*	*PDR1*	*NEO1*	*YLR312C*	*RFT1*	*YLR280C*
15	*FUN12*	*ITR1*	*YLR296W*	*FMP13*	*PAC2*	*MRPL35*	*RPC25*	*WHI3*	*YIL064W*

Each topic is composed of a group of genes in the heterozygous deletion library that show a sensitivity or fitness defect to the drug. Each gene is permitted to be a member of more than one topic and the genes are ranked according to the probability assigned to that gene according to a parameter of the LLDA model. The table shows the top 15 genes ranked by probability for each topic in the model.

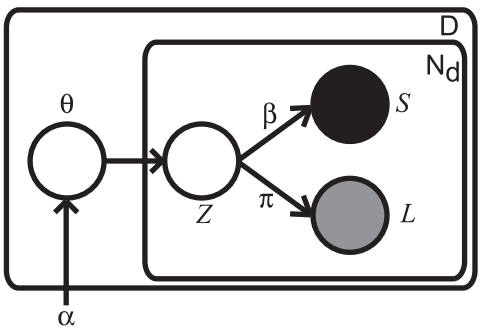

Fig. 1. As a graphical model, the labeled latent Dirichlet allocation model has four main component random variables. Each mode (L, S, Z, θ) in the model represents a random variable. Nodes shaded in black represent fully observed data. Gray indicates partially observed data and unshaded nodes are latent or unobserved random variables. The edges connecting the nodes correspond to a particular conditional independence structure assured for the random variables. The structure chosen here is designed for the chemogenomic experiments in order to cluster the fitness defect profiles using function category assignments for each gene. The variable S represents the gene, L represents the function category label, Z represents the topic and θ is a latent Dirichlet random variable.

to vary with each experiment. Detailed comparative analyses of the effects of this modeling choice are presented elsewhere (Blei *et al*., 2003).

The parameters of the model were estimated by an iterative variational expectation–maximization (EM) procedure (Dempster *et al*., 1977; Jordan, 1999). This iterative solution procedure requires a random initialization. Local optima are unavoidable due to the non-convexity of the log-likelihood function. The variational EM algorithm was repeated 10 times using random initial conditions and the best performing parameters were selected. To select the number of topics we used a minimum description length score (Hansen *et al*., 2001). Further details regarding the statistical inference procedures are provided in the Supplementary Information.

The distribution of MIPS functions for each experiment was computed by marginalizing the joint probability distribution (PDF) in the model over genes and topics. The function category distribution for each gene was obtained by marginalizing the joint PDF over the topics. The drug function distributions were inferred by marginalizing over topics for each posterior distribution given the experiment.

4 RESULTS

Only a small number (72) of genes have probability >0.01 in any topic. Nine genes are assigned with probability >0.01 in more than one topic: *FEN2*, *GLC7*, *FOL2*, *MAL11*, *YBT1*, *TVP18*, *PDR5*, *DFR1* and *SET6*. More than 10 genes appear in the list of top 15 genes, ranked by probability, in more than one topic. For different drug treatments there is little overlap in topics used to describe those experiments, except for some interesting cases that we describe in detail below:

4.1 Consensus sensitive genes for classes of drug

Using a minimum-description length score we found that nine topics yielded the most accurate model. Given this choice of the number of topics, the parameters of the model were estimated from the 79 individual drug experiments. Table 2 shows the top 15 genes sorted by decreasing probability for each topic and Figure 2 shows the posterior probability distribution over topics, for each experiment in the compendium. Only 3734 of a total of 5918 strains scored as significant in at least one experiment by the fitness defect analysis (Giaever *et al*., 2004), and are included in this model.

Since our analysis is a model-based approach we can use function annotations for each gene from the MIPS catalogue to assist in the formation of topics. A conditional dependence among the topic, function category and gene is incorporated in the model reflecting the expectation that drugs often impact only a few cellular functions. This conditional dependence constrains the topics to contain genes that have a large fitness defect for the experiments allocated that topic. The topics also tend to be cohesive in the impacted cellular functions. Genes ranked highly in each topic tend to be involved in the same cellular functions and are descriptive of the drugs allocated that topic.

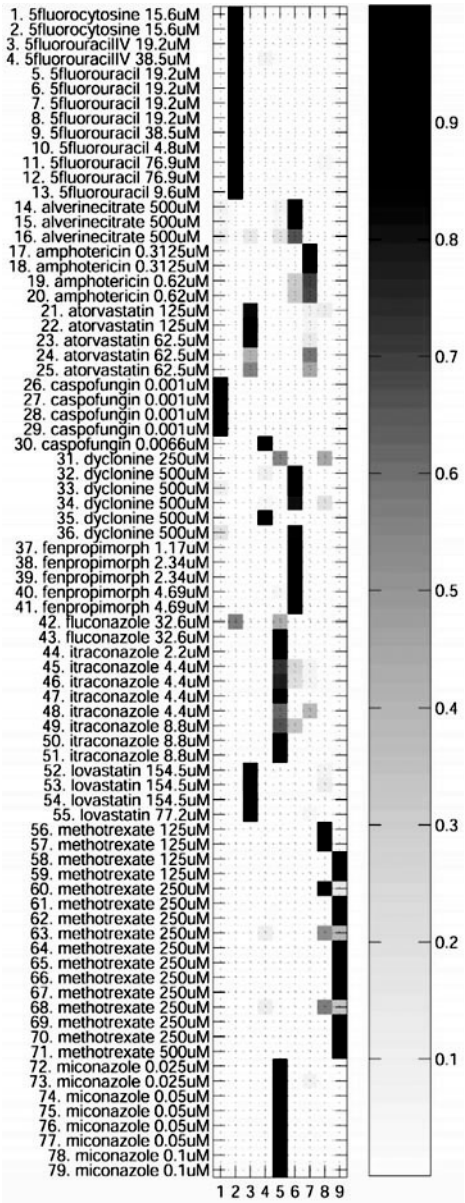

Fig. 2. Experiments cluster based on the topics that are allocated to them. Each row represents the normalized fraction of topic allocated to the experiment. Topic 5 is allocated to the azole class drugs. Dyclonine, fenpropimorph and alverine-citrate target the ergosterol pathway at a different point than the azole drugs and are represented by topic 6. Topic 2 is allocated to 5-FU and 5-FC experiments. 5-FU is an anticancer and 5-FC is an antifungal because humans lack a cytosine deaminase that converts 5FC into 5FU. Amphotericin B experiments are allocated topic 7. Topic 3 is mainly allocated to the anti-cholesterol drugs, lovastatin and atorvastatin. Topics 8 and 9 are allocated to the highly replicated methotrexate experiments and the antifungal caspofungin is allocated topic 1. Clustering using the LLDA model does not require that genes or experiments be allocated to mutually exclusive groups.

4.1.1 The azole drugs affect drug transporters, Erg11 and genes represented in topic 5 The azoles (itraconazole, miconazole and fluconazole) generally have a large allocation of topic 5. This topic contains the pleiotropic drug resistant genes *PDR5* and *PDR16* known to confer sensitivity on the azoles when deleted, a gene of uncharacterized function, *SET6*, and the established target *ERG11* (van den Hazel *et al*., 1999; Wills *et al*., 2000). Interestingly, several other genes, not previously known to confer sensitivity on the azoles, were also uncovered (Table 2).

4.1.2 Fenpropimorph, alverine-citrate and dyclonine affect ERG24 and genes represented in topic 6 These therapeutically distinct compounds are found to cluster by being allocated the same topic in the LLDA model. Fenpropimorph, a morpholine antifungal, is thought to target *ERG2* and *ERG24* (Baloch *et al*., 1984; Lai *et al*., 1994; Marcireau *et al*., 1990), and the target of alverine-citrate, an anticholinergic is not known. Dyclonine is a human anesthetic thought to affect the ergosterol pathway through *ERG2* (Hughes *et al*., 2000). Interestingly, these three drugs all have a similar chemical backbone which may explain their similar fitness defect profiles (Giaever *et al*., 2004). *ERG24*, the top gene in topic 6, is a putative target of fenpropimorph but the topic is also allocated to alverine-citrate and dyclonine. The *ERG2* strain is expected to be sensitive to fenpropimorph but is not observed to be in these *in vivo* experiments (Giaever *et al*., 2004). Most of the experiments on these three drugs are allocated topic 6, but one dyclonine experiment is allocated topic 4 almost exclusively. A closer examination of that experiment reveals a very different fitness defect profile than the other dyclonine experiments. This experiment is a possible outlier, perhaps, due to poor hybridization or a microarray protocol error. Both topics 5 and 6 contain *MAL11*, *TVP18* and *SET6* in the top 15 genes. The topics also differ significantly because topic 5 contains the target of the azole antifungals, *ERG11*, but not *ERG24* and vice versa.

4.1.3 Amphotericin B affects genes involved in cell membrane and wall integrity, represented in topic 7 Amphotericin B binds ergosterol-like lipids in the membrane and increases membrane permeability (Katzung, 1998). Topic 7 is allocated to experiments on this drug and includes genes involved in cell membrane integrity and membrane-associated proteins. Interestingly, *ARC18*, in topic 7, is a member of the Arp2/3 actin polymerization complex. The fitness defect of these strains in amphotericin B is consistent with the observation that cortical patch mobility, mediated by the Arp2/3 complex, is required for cell wall remodeling (Machesky and Gould, 1999). Highlighting the interaction between cell wall and cell membrane where amphotericin B is thought to bind, *FKS1*, which is involved in cell wall organization and biogenesis, is in the topic list and has been found to co-localize with the Arp2/3 complex (Utsugi *et al*., 2002). The observation of a fitness defect of these four genes demonstrates the necessity of the stability of the coupling between the cell wall and the cell membrane for survival in amphotericin B. Furthermore, the sensitivity of *POP1*, categorized as a protein synthesis gene, indicates that there is more to the mechanism of action to be understood.

4.1.4 Statin compounds affect HMG1, ERG13, UPC2 and genes represented in topic 3 Atorvastatin and lovastatin are cholesterol reduction medications. Of the known targets, Hmg1 and Hmg2, the *HMG1* heterozygous deletion strain is the only one that shows a significant fitness defect to the drugs. This is consistent with the observation that the Hmg1 protein contributes the majority of the HMG-CoA reductase activity in the cell (Basson *et al*., 1986). Furthermore, HMG-CoA reductase is the rate-limiting step in the sterol biosynthesis pathway in yeast (Basson *et al*., 1986). *ERG13* is also

significantly sensitive to both drugs, according to this analysis, and is a novel discovery further discussed elsewhere (Giaever *et al*., 2004). Consistent with an effect on the ergosterol pathway, the fifth highest ranking gene, *UPC2*, is a transcription factor that regulates ergosterol biosynthesis (Vik and Rine, 2001). *PDR5* also appears in the top 15 genes in this topic but would not have, had it been constrained to be assigned to only one cluster of genes.

4.1.5 Flurouracil and 5-fluorocytosine affect FUI1, genes involved in RNA production and genes represented in topic 2 While humans lack the cytosine deaminase required to convert 5-FC into 5-FU, yeast contain this enzyme (Zhang *et al*., 2003). Therefore, the effect of both drugs in yeast is similar and the experiments are, indeed, allocated to the same topic. Both drugs are thought to act via three mechanisms: (1) inhibition of thymidylate synthase, (2) direct incorporation into DNA and (3) direct incorporation into RNA (Longley *et al*., 2003). The first annotated gene in this topic, *NOP4*, codes for a protein involved in RNA binding and rRNA processing. The next gene in order, *YPL044C*, is antisense to *NOP4* and the deletion of this gene may disrupt the function of *NOP4*. Indeed many of the genes in this topic (*RRP6*, *RRP42* and *RRP45*) are involved in ribosomal RNA processing. Ribosomal RNA processing genes are generally classified by MIPS to be involved in the transcription process. Surprisingly, we do not observe the putative target (Lum *et al*., 2004) *CDC21* as being sensitive. Although this does not preclude *CDC21* as a target, these results suggest that direct incorporation into the RNA may be the primary mechanism of action. Other studies support this finding (Giaever *et al*., 2004; Lum *et al*., 2004; Scherf *et al*., 2000).

4.1.6 Methotrexate affects YBT1, YAP6, DFR1, FOL1/2 and genes in topic 8 and 9 The anticancer drug methotrexate, a folic acid antagonist (Katzung, 1998), is primarily allocated topics 8 and 9. An investigation of the genes that group into topics 8 and 9 reveals the ABC transporter *YBT1*, which is the yeast homolog of a known methotrexate transporter (Zeng *et al*., 2001). The pleiotropic drug resistant genes *PDR5/PDR16* are sensitive to the azole class of drugs but not to methotrexate, suggesting that the mechanism by which yeast exports methotrexate and azole drugs are distinct. *YAP6*, the fifth gene in topic 9 which codes for a transcriptional regulator, is hypothesized to be involved in pleiotropic drug resistance (Furuchi *et al*., 2001). The target of methotrexate, Dfr1, is the first item in both topics 8 and 9. The *FOL1* and *FOL2* deletion strains are also sensitive to the drug. *FOL1* and *FOL2* are involved in folic acid biosynthesis and act upstream of *DFR1* in the targeted pathway.

4.1.7 Caspofungin affects LCB1 and genes represented in topic 1 Caspofungin is a semisynthetic compound that is a derivative of a natural product and is a recently approved echinocandin antifungal (Letscher-Bru and Herbrecht, 2003). Experiments on caspofungin in this study have a large allocation of topic 1. Caspofungin is thought to target proteins involved in the β(1,3)-D-glucan synthase activity, which is involved in the synthesis of cell wall glucan (Letscher-Bru and Herbrecht, 2003). Specifically, mutation of the *FKS1* gene confers resistance to caspofungin, and *FKS1* is thought to be a large subunit of the target of the drug (Douglas *et al*., 1994, 1997). Currently, the mechanism of caspofungin is not completely understood (Letscher-Bru and Herbrecht, 2003). The HIP profiles and model implicate *LCB1* as sensitive to the compound. Further experimentation may reveal whether or not this gene is involved in the mechanism of action. This protein is involved in the first step of the biosynthesis of sphingolipids, an essential component of the cell membrane (Wills *et al*., 2000).

4.2 Model sensitivity

In any clustering model, we must control the tradeoff between more detailed resolution for each experiment and better high-level pattern discovery in the whole corpus. By choosing to focus more on pattern discovery over all experiments, we may miss a gene that is somewhat sensitive to an individual treatment. A closer examination of the dyclonine experiments reveals that the *NEO1* heterozygous deletion strain has a fitness defect for the drug, but is only ranked 14th in the distribution for topic 6. *NEO1* is a neomycin-resistance gene involved in intracellular protein transport. Since the objective of the LLDA model is to uncover genes that show sensitivity to groups of drugs, it is robust to the individualities of each experiment that are evident with a detailed examination. This robustness is advantageous in filtering experimental false positives, as well in identifying higher order patterns in the compendium.

In any microarray experimental study, the fitness defect profiles of some experiments show variability; several such experiments are included in the methotrexate and itraconazole replicates. They are evident by the relative spread of their topic allocations. Also, one each of the dyclonine and caspofungin experiments does not clearly group with the other replicates. They are allocated topic 4 and are the only experiments allocated that topic. We have observed that selection of proper concentration of the drug is critical to yield meaningful results. The cells may not be inhibited enough at these lower concentrations to observe significantly sensitive strains. Similarly, non-specific effects may appear with too much drug.

4.3 Drug function prediction

We use the estimate of the posterior topic distribution, given the fitness defect scores for an experiment in the LLDA model, to compute the distribution over function categories for each experiment in the dataset. Figure 3 shows that the strains that are sensitive to 5-FU are likely to be involved in transcription. There is evidence the mechanism of action is specific to the disruption of rRNA processing (Longley *et al*., 2003). The following highly ranked genes in topic 2 are indeed classified by the MIPS function ontology to be involved in rRNA processing: *NOP4*, *DIS3*, *MTR4* and *RRP42*. The model predicts that amphotericin B affects protein synthesis and that methotrexate is an anti-metabolite. The miconazole, itraconazole and fluconazole experiments show an enhancement for transport facilitations—most probably due to the pleiotropic drug resistance genes *PDR5* and *PDR16*. Amphotericin B is allocated topic 7 which is enriched for genes involved in protein synthesis including *DIA4* and *GCD2*. While there is no convenient base distribution to compare each experiment beyond the naïve uniform distribution, this plot allows for a qualitative comparison of the functions a drug impacts in the cell.

4.4 Comparison with hierarchical clustering

In this section, we present a comparison of the results from the LLDA model with the results from HC. While any such comparison is necessarily qualitative, the pattern of differences revealed by the empirical results has proved useful in attempting to understand the practical consequences of the different perspectives on clustering taken by LLDA and HC.

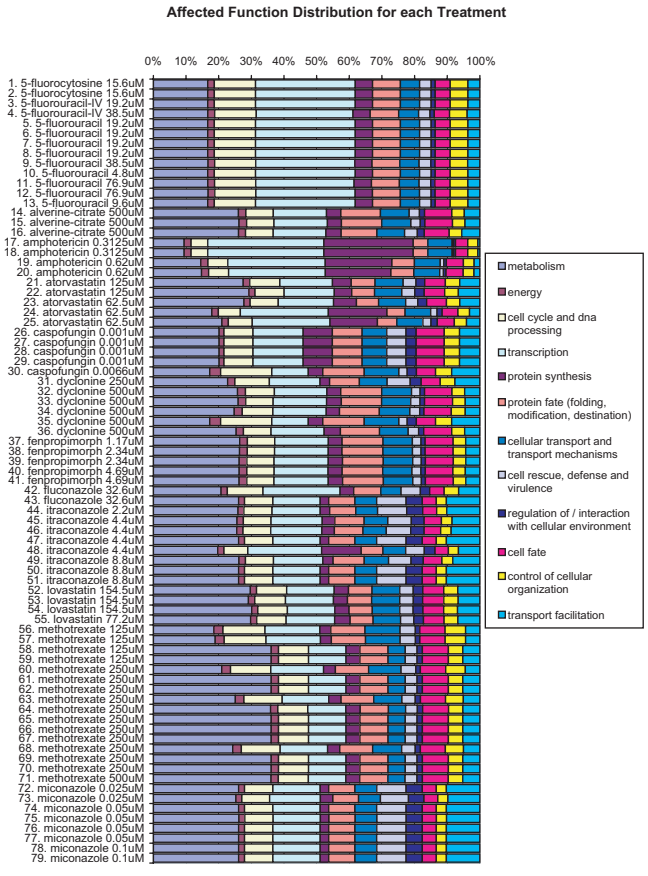

Fig. 3. The LLDA model is used to infer the function categories affected by each experiment. Methotrexate is a known anti-metabolite and most replicates show a large fraction of the metabolism function. The azole drug experiments (itraconazole, fluconazole and miconazole) have a large allocation of metabolism genes and a larger fraction of drug transporters than the other drugs in the compendium. 5-FU experiments are enriched for the transcription function. The fitness defect profiles for these experiments show that ribosomal RNA processing genes (classified as transcription genes by MIPS) exhibit a growth defect with the drug.

Figure 4a shows a HC of all 79 experiments based on the same data used for the LLDA model. Figure 4b shows a dendrogram over 27 relevant genes also clustered using HC. All HC dendrograms were generated by average linkage of the correlation distance using the R statistical language (R Development Core Team, 2004). HC was also computed using Euclidean and Manhattan distance metrics as well as complete linkage clustering. The linkage method and distance metric were chosen here to yield the best clustering of all combinations tested. We have found three important differences between HC and LLDA. First, HC often separates experiments involving the same treatment, when the concentration of the compound is varied. Second, genes that show a fitness defect for compounds with different targets are clustered with only one of those targets by HC. Third, the greedy agglomeration of genes into clusters by HC leads genes that are affected by the same treatment to be separated from the target, when they are more similar to each other than to the target.

With regards to the first point, consider the dendrogram shown in Figure 4a for the clustering of experiments. Four methotrexate treatments (58, 61, 67, 71 shown in orange) are separated from a much larger cluster (57, 59, 62, 60, 63, 68, 66, 69, 65, 70, 64). Similarly, four azole treatments (42, 48, 49, 50 shown in green) are distant from the other replicates (43, 45, 46, 51, 78, 79, 44, 77, 75, 74, 76, 47). LLDA and HC both capture the similarity among alverine-citrate, dyclonine, fenpropimorph and the azoles. LLDA discriminates between the two groups of treatments because *ERG11* has a large fitness defect in the azole treatments, but not in the other treatments. Two fenpropimorph experiments (40, 41 shown in red) are distant from the other replicates in the dendrogram, possibly because they are at a higher concentration than the other fenpropimorph experiments. Non-specific genes are affected when the concentration of drug is increased too much. Many genes with a small fitness defect score can increase the Euclidean, Manhattan or correlation distance between otherwise similar experiments.

The second point—that genes that show a fitness defect for compounds with different targets are clustered with only one of those targets by HC—is demonstrated in Figure 4b. *SET6* and *PDR5* are distantly separated from *ERG11* but are close to *ERG24*. *MAL11* and *TVP18* are close to *ERG11*, but far from *PDR5* and *SET6*. These pleiotropic genes (*PDR5*, *SET6*, *TVP18* and *MAL11* shown in green) exhibit a fitness defect in both the azole treatments and the alverine-citrate, dyclonine and fenpropimorph treatments. Genes affected by the same drugs should be clustered together, but this is not the case for pleiotropic genes using HC. The HC algorithm associates the pleiotropic genes with the closest target, but not with both targets.

The final observation is that the profiles of some genes are more similar to each other than to the target gene, even though they are all affected by the same drug. Figure 4b shows the target of methotrexate, *DFR1*, is distantly separated from other genes that are sensitive to methotrexate: *YBT1* and *FOL1/2* (shown in orange). These three genes are closer to each other than to the target of methotrexate, when the distance is measured using all of the treatments.

5 DISCUSSION

HIP allows us to detect putative molecular mechanisms underlying the action of particular drugs. Our objective is to screen many novel and well-known compounds using this technique to learn more about the mechanisms by which they act. For example, some compounds such as dyclonine, alverine-citrate and fenpropimorph have similar chemical structures and do cluster in the model based on their common fitness defect profiles. Other compounds such as methotrexate and 5-FU operate by very different mechanisms as evidenced by their dissimilar fitness defect profiles, and are appropriately allocated different topics.

The graphical model used here allows us to model the HIP profiles of each experiment as an allocation of topics. The genes and experiments are not partitioned into mutually exclusive clusters. This allows the model to represent the similarities among compounds and also allows the sets of genes affected by different drugs to overlap in a model-based analysis. Incorporating MIPS annotation of genes into the LLDA model enables us to incorporate known functional classifications in a probabilistically well-founded manner, providing additional guidance for the clustering. The model is not only limited to MIPS as an annotation source; it can also similarly employ Gene Ontology annotations (Ashburner *et al*., 2000) or any other classification scheme for genes.

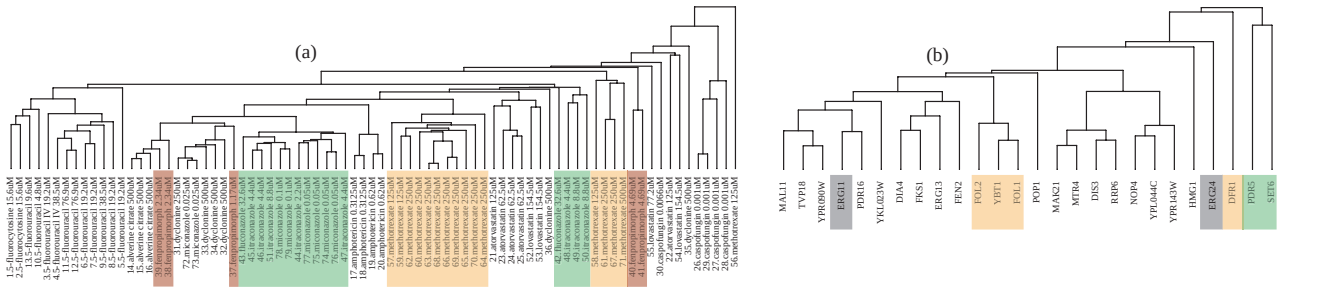

Fig. 4. The dendrogram of average linkage HC of experiments with a correlation distance is shown in (***a***). A subset of genes are clustered in (***b***) using the same method. (a) Hierarchical clustering separates fenpropimorph experiments (40, 41 shown in red) from the other experiments with the same drug. Several azole experiments (42, 48, 49, 50 shown in green) and methotrexate experiments (shown in orange) are also distantly separated from replicate treatments. (b) In the dendrogram for the HC of genes, pleiotropic genes are shown in green. Genes that are haploinsufficient in methotrexate are shown in orange. The targets of fenpropimorph and the azoles, *ERG11* and *ERG24*, are shown in gray. *ERG11* is separated from *PDR5* and *SET6* even though all three strains have a fitness defect in azole treatments. The pleiotropic gene, *SET6*, is also haploinsufficient in fenpropimorph and is closer to that drug's target, *ERG24*, than it is to the azole target, *ERG11*. Since HC is a greedy agglomerative procedure, pleiotropic genes such as *SET6* only associate with one of the two targets. Other pleiotropic genes *MAL11* and *TVP18* are separated from *SET6* and *PDR5* even though they all have a fitness defect in the azoles, dyclonine, fenpropimorph and alverine-citrate. The constraints imposed by HC prevent *SET6*, *PDR5*, *MAL11* and *TVP18* from being close to both *ERG11* and *ERG24* in the tree.

This model successfully clusters different groups of drugs that target various components of the ergosterol pathway (topics 3, 5 and 6 in Fig. 2). The LLDA model also groups together drugs with similar chemical backbones, doing so because of the commonality in their fitness defect profiles. In most cases, the known targets for various drug classes are found within the top 15 genes associated with each topic. In addition, some previously uncharacterized genes, such as *SET6*, which is ranked highly in topics 5 and 6, may be related to the mechanisms of actions of the associated drugs. This statistical modeling tool is most useful as new drugs are screened against the deletion library to classify small molecules that operate by a novel mechanism, or are similar to a mechanism that has already been screened.

This compendium study includes 79 experiments, and the methods and statistical models developed are scalable to many hundreds of chemogenomic experiments. Though the focus of this study is on induced haploinsufficiency experiments, the LLDA model and the graphical model framework are also readily applicable to more traditional expression datasets and other high-throughput genomic screens. We expect that the large-scale genomic studies that explore changes in gene expression when cells are exposed to drug are complementary to those of haploinsufficiency screens.

The LLDA model can be understood as a member of a general family of methods known as bi-clustering methods—a family which includes the SVD when the SVD is viewed as a clustering method (Alter *et al*., 2000; Cheng and Church, 2000; Hofmann and Puzicha, 1998; Lazzeroni and Owen, 2002; Tanay *et al*., 2002). Viewing a dataset as a matrix (e.g. genes by treatments), these methods differ from classical clustering techniques in that they treat the rows and columns symmetrically. This symmetry provides a dual notion of 'cluster'; in particular, for each column cluster one can group those row variables that are highly associated with that cluster (large 'weights'); these groupings will overlap in general. We view LLDA as a particular natural expression of this idea. In particular, in LLDA the symmetry is a direct consequence of Bayes' theorem and the 'weights' are posterior probabilities. Moreover, the probabilistic framework underlying LLDA has advantages in terms of extensibility; as we have seen, LLDA naturally incorporates functional labels within the clustering procedure.

Focusing on specifically probabilistic approaches, LLDA is also related to the notion of a 'module network' due to Segal *et al*. (2003b). Module networks and LLDA are both instances of the general family of probabilistic graphical models; both involve probabilistic clustering of genes. The specific assumptions underlying these models are different, however, reflecting different data-analytic goals. In particular, module networks make a mutual exclusivity assumption—a gene can only appear in a single module. [This assumption can be removed via a further extension of the module network framework; see Segal *et al*. (2003a)].

In summary, we have presented a model-based approach to the analysis of data from HIP experiments. While simple, the model properly handles gene pleiotropy, a biological phenomenon that is often overlooked when off-the-shelf clustering methods are applied to biological data. Our results demonstrate the utility of the compendium-based analysis of HIP experiments to reveal treatments that have related mechanisms of action.

ACKNOWLEDGEMENTS

We would like to acknowledge Daniel Jaramillo, David Blei and Denise Wolf for their careful reading of the manuscript. A.P.A. and P.F. would also like to acknowledge support from the National Cancer Institute of the National Institutes of Health and the Howard Hughes Medical Institute for support during the period of this work. GG would like to acknowledge support from the National Human Genome Research Institute.

Conflict of Interest: none declared.

REFERENCES

Alexandersson,M. *et al*. (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.*, **13**, 496–502.

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Baloch,R.I. *et al*. (1984) Inhibition of ergosterol biosynthesis in *Saccharomyces cerevisiae* and *Ustilago maydis* by tridemorph, fenpropimorph and fenpropidin. *Phytochemistry*, **23**, 2219–2226.

Basson,M.E. *et al*. (1986) *Saccharomyces cerevisiae* contains two functional genes encoding 3-hydroxy-3-methylglutaryl-coenzyme A reductase. *Proc. Natl Acad. Sci. USA*, **83**, 5563–5567.

Bennett,J. (2001) *Antimicrobial Agents: Antifungal Agents*. McGraw-Hill Professional, New York, Chapter 49, pp. 1295–1312.

Bergmann,S. *et al*. (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **67**, 031902.

Blei,D.M. *et al*. (2003) Latent Dirichlet Allocation. *J. Mach. Learning Res.*, **3**, 993–1022.

Chabner,B.A. *et al*. (2001) *Chemotherapy of Neoplastic Diseases*. McGraw-Hill Professional, New York, Chapter 52, pp. 1389–1146.

Cheng,Y. and Church,G.M. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 93–103.

Coelho,A.M. *et al*. (2001) Rectal antinociceptive properties of alverine citrate are linked to antagonism at the 5-HT1A receptor subtype. *J. Pharm. Pharmacol.*, **53**, 1419–1426.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–39.

Douglas,C.M. *et al*. (1994) The *Saccharomyces cerevisiae* FKS1 (*ETG1*) gene encodes an integral membrane protein which is a subunit of 1,3-beta-D-glucan synthase. *Proc. Natl Acad. Sci. USA*, **91**, 12907–12911.

Douglas,C.M. *et al*. (1997) Identification of the FKS1 gene of *Candida albicans* as the essential target of 1,3-beta-D-glucan synthase inhibitors. *Antimicrob. Agents Chemother.*, **41**, 2471–2479.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Furuchi,T. *et al*. (2001) Two nuclear proteins, Cin5 and Ydr259c, confer resistance to cisplatin in *Saccharomyces cerevisiae*. *Mol. Pharmacol.*, **59**, 470–474.

Giaever,G. *et al*. (1999) Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.*, **21**, 278–283.

Giaever,G. *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

Giaever,G. *et al*. (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc. Natl Acad. Sci. USA*, **101**, 793–798.

Hansen,M. and Yu,B. (2001) Model selection and the principle of minimum description length. *J. Am. Stat. Assoc.*, **96**, 746–774.

Hardman,J.G. *et al*. (2001) The pharmacological basis of therapeutics. 10e. McGraw-Hill Professional, New York.

Hofmann,T. and Puzicha,Y. (1998) Statistical models for co-occurrence data. *Technical Report* AIM-1625, Artifical Intelligence Laboratory, Massachusetts Institute of Technology.

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Jaakkola,T.S. and Jordan,M.I. (1999) Variational probabilistic inference and the QMR-DT network. *J. Artif. Intell. Res.*, **10**, 291–322.

Jansen,R. *et al*. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.

Jordan,M.I. (1999) Learning in Graphical Models. The MIT Press, Cambridge, MA.

Jordan,M.I. (2004) Graphical Models. *Stat. Sci.*, **19**, 140–155.

Katzung,B.G. (1998) *Basic and Clinical Pharmacology*. Appleton & Lange, Stanford, CT.

Lai,M.H. *et al*. (1994) The identification of a gene family in the *Saccharomyces cerevisiae* ergosterol biosynthesis pathway. *Gene*, **140**, 41–49.

Lazzeroni,L. and Owen,A.B. (2002) Plaid models for gene expression data. *Stat. Sin.*, **12**, 61–86.

Letscher-Bru,V. and Herbrecht,R. (2003) Caspofungin: the first representative of a new antifungal class. *J. Antimicrob. Chemother.*, **51**, 513–521.

Longley,D.B. *et al*. (2003) 5–Fluorouracil: mechanisms of action and clinical strategies. *Nat. Rev. Cancer*, **3**, 330–338.

Lum,P.Y. *et al*. (2004) Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell*, **116**, 121–137.

Machesky,L.M. and Gould,K.L. (1999) The Arp2/3 complex: a multifunctional actin organizer. *Curr. Opin. Cell Biol.*, **11**, 117–121.

Marcireau,C. *et al*. (1990) *In vivo* effects of fenpropimorph on the yeast *Saccharomyces cerevisiae* and determination of the molecular basis of the antifungal property. *Antimicrob. Agents Chemother.*, **34**, 989–993.

Mewes,H.W. *et al*. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Peterson,L.E. (2003) Partitioning large-sample microarray-based gene expression profiles using principal components analysis. *Comput. Methods Programs Biomed.*, **70**, 107–119.

R Development Core Team (2004) R: a language and environment for statistical computing. Manual, R Foundation for Statistical Computing.

Scherf,U. *et al*. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.

Segal,E. *et al*. (2003a) Decomposing gene expression into cellular processes. *Pac. Symp. Biocomput.*, 89–100.

Segal,E. *et al*. (2003b) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Tanay,A. *et al*. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18** (Suppl. 1), S136–S144.

Utsugi,T. *et al*. (2002) Movement of yeast 1,3-beta-glucan synthase is essential for uniform cell wall synthesis. *Genes Cells*, **7**, 1–9.

van den Hazel,H.B. *et al*. (1999) PDR16 and PDR17, two homologous genes of *Saccharomyces cerevisiae*, affect lipid biosynthesis and resistance to multiple drugs. *J. Biol. Chem.*, **274**, 1934–1941.

Vik,A. and Rine,J. (2001) Upc2p and Ecm22p, dual regulators of sterol biosynthesis in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, 21, 6395–6405.

Wills,E.A. *et al*. (2000) New potential targets for antifungal development. *Emerg. Ther. Targets*, **4**, 1–32.

Zeng,H. *et al*. (2001) Transport of methotrexate (MTX) and folates by multidrug resistance protein (MRP) 3 and MRP1: effect of polyglutamylation on MTX transport. *Cancer Res.*, 61, 7225–7232.

Zhang,M. *et al*. (2003) Regional delivery and selective expression of a high-activity yeast cytosine deaminase in an intrahepatic colon cancer model. *Cancer Res.*, **63**, 658–663.