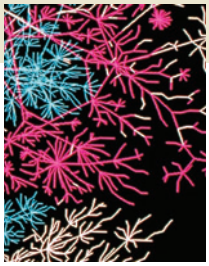

Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks

Diego di Bernardo[1,5], Michael J Thompson[2,5], Timothy S Gardner[2,5], Sarah E Chobot[3], Erin L Eastwood[3,4], Andrew P Wojtovich[3], Sean J Elliott[3], Scott E Schaus[3,4] & James J Collins[2]

A major challenge in drug discovery is to distinguish the molecular targets of a bioactive compound from the hundreds to thousands of additional gene products that respond indirectly to changes in the activity of the targets[1–8]. Here, we present an integrated computational-experimental approach for computing the likelihood that gene products and associated pathways are targets of a compound. This is achieved by filtering the mRNA expression profile of compound-exposed cells using a reverse-engineered model of the cell's gene regulatory network. We apply the method to a set of 515 whole-genome yeast expression profiles resulting from a variety of treatments (compounds, knockouts and induced expression), and correctly enrich for the known targets and associated pathways in the majority of compounds examined. We demonstrate our approach with PTSB, a growth inhibitory compound with a previously unknown mode of action, by predicting and validating thioredoxin and thioredoxin reductase as its target.

A critical step in drug development is the optimization of therapeutic efficacy and the minimization of undesirable side effects of a candidate drug. Ideally, optimization is carried out using knowledge of the drug's mode of action, that is, the molecular targets that mediate its therapeutic effects and side effects. For many drug candidates, however, the targets are unknown and difficult to identify among the thousands of gene products in a typical genome.

DNA microarray technology enables the observation of all genes with a transcriptional response to a compound treatment, and thus provides an opportunity to efficiently identify a compound's targets. However, whole-genome expression profiles do not distinguish the genes targeted by a compound from the indirectly regulated genes. To overcome this problem, we have developed a model-based approach that is able to accurately distinguish a compound's targets from the indirect responders, and, in contrast to association analysis techniques[1,9,10], haploinsufficiency profiling[5–7] and chemical-genetic interaction mapping[8], does not require libraries of genetic mutants or fitness-based assays of drug response. With this approach, called mode-of-action by network identification (MNI), we first reverse-engineer a network model[11–27] of regulatory interactions in the organism of interest using a training data set of whole-genome expression profiles (**Fig. 1**). We then use the model to analyze the expression profile of compound-treated cells to determine the pathways and genes targeted by the compound. The reverse-engineered model is a directed graph relating the concentrations of transcripts to each other. An edge in the graph means that the activity of one gene product influences the transcription of another gene (**Fig. 2**). Multiple genes may influence the activity of a particular gene; these influences are integrated in the model as a weighted sum of the transcript concentrations (**Fig. 2**). Because the model is trained using transcription data only, regulatory influences between genes may be mediated through protein or metabolite species that are not explicitly represented.

The algorithm assumes that training profiles are obtained in steady state following a variety of treatments, including compounds, RNAi, and gene-specific mutations (**Fig. 1**). The ability to use varied treatment types in the training data is an important advance over earlier model estimation techniques[15,22,23], which required knowledge of the gene targets of each training perturbation. This improved flexibility may enable application of the MNI approach to higher model organisms, where gene-specific perturbations are more difficult to implement. To infer a network model without requiring gene-specific perturbations, the algorithm employs an iterative procedure: it first predicts the targets of the treatment using an assumed network model, and then uses those predicted targets to estimate a better model. The procedure repeats until convergence criteria are met (see **Supplementary Notes** online). This approach is analogous to the Expectation Maximization (EM) algorithm[28] commonly used to train Bayesian networks.

Once the regulatory model is trained, we apply it to the expression profile of a test compound to predict its targets. The model acts as a filter, in essence, checking the expression level of each gene in the cell (relative to the level of all other genes in the cell) for consistency with

[1]Telethon Institute for Genetics and Medicine, Naples, Italy. [2]Center for BioDynamics and Department of Biomedical Engineering, [3]Department of Chemistry, [4]Center for Chemical Methodology and Library Development, Boston University, Boston, Massachusetts, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to J.J.C. (jcollins@bu.edu).

Published online 4 March 2005; doi:10.1038/nbt1075

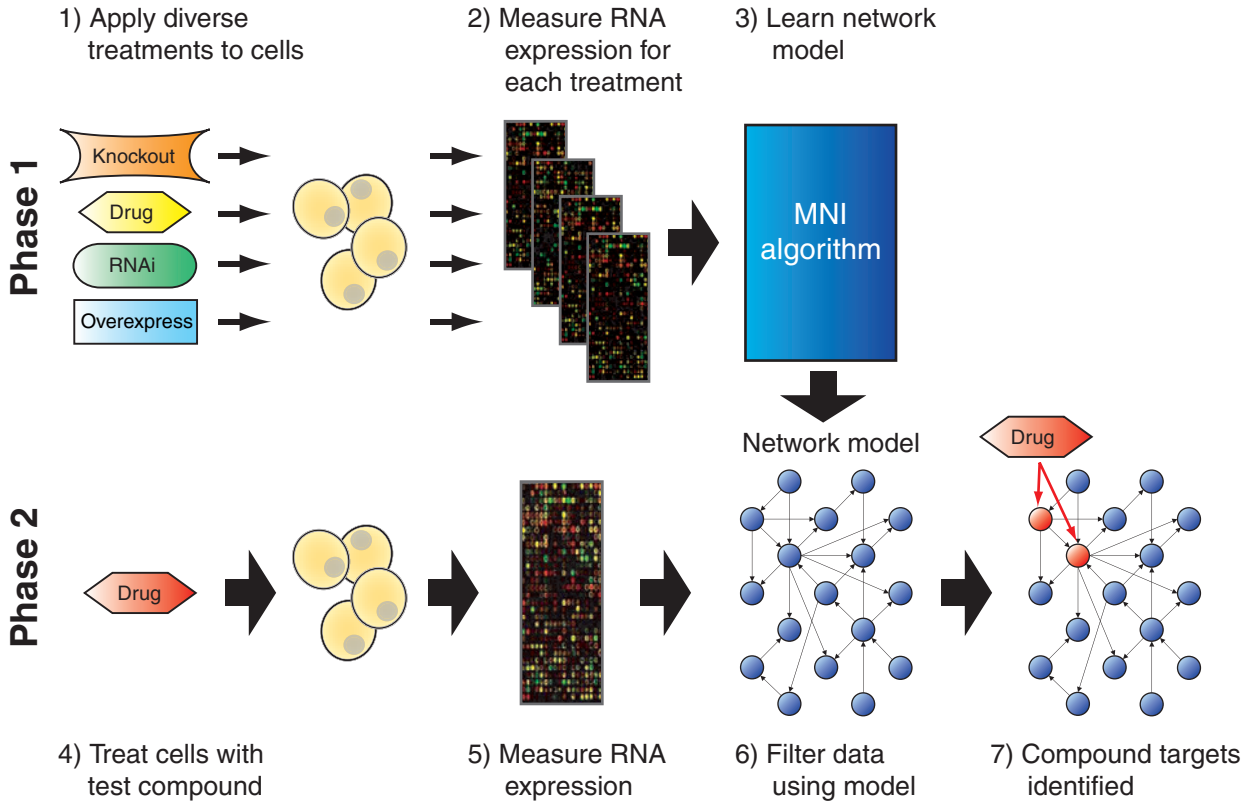

Figure 1 Overview of the MNI method. In phase 1, a set of treatments, including knockouts, compounds, overexpressions and/or RNAi, is applied to an organism. Cells or tissues are sampled, and mRNA is collected. The abundance changes of all mRNA species in the organism are measured. The data are used by the MNI algorithm to infer a model of the regulatory influences between genes in the organism (blue-filled circles indicate genes; arrows indicate regulatory influences). In phase 2, a test treatment, such as a drug, is applied to the cells and expression changes of all mRNA species are measured. The expression data are then filtered using the network model to distinguish the targets of the test treatment (red-filled circles) from secondary responders.

regulatory influences embodied in the trained regulatory model. The genes are then ranked by a *z*-statistic that measures their level of consistency (see Methods and **Supplementary Notes**). The highest-ranked genes are those whose expression is most inconsistent with the model, and this inconsistency is attributed to the external influence of the compound on those genes (see **Supplementary Notes** for a detailed outline of the algorithm).

We tested our approach by combining two publicly available, whole-genome yeast expression data sets: a compendium of 300 profiles of gene deletions, titratable promoter insertions and drug compound treatments from Hughes *et al.*[1] and a recent set of 215 titratable promoter insertions in essential genes from Mnaimneh *et al.*[29]. For each treatment/perturbation, a single profile was obtained from yeast cells grown to steady state after the perturbation. A log-transformed expression ratio was computed for each gene in each profile relative to untreated, wild-type yeast strains. The algorithm was blinded to any information regarding the gene targets of the treatments and mutations.

To evaluate the performance of the MNI algorithm, we tested its ability to predict the gene targets of the 11 promoter insertions from the Hughes compendium (**Table 1**). For the 11 mutant profiles tested, the algorithm ranked the targeted gene as the most likely affected gene in 8 out of 11 cases. Two of the remaining perturbed genes were correctly ranked in the top 10 of most likely affected genes (*RHO1* and *PMA1*). The final perturbed gene, *ERG11*, was ranked 42nd, which is a substantial enrichment over its ranking based on the significance of its expression change alone (it was ranked 2,820 by *z*-score of expression change; **Table 1**). In contrast, ranking by expression change identified the affected gene with high significance (ranked among the top 10 out of 6,000 genes) for only three of the 11 mutations, which is significantly worse than the MNI algorithm.

We compared the performance of the MNI algorithm to two association analysis approaches: a correlation method[1,9] and a linear combination method[10] (**Table 1** and **Supplementary Table 1** online). The correlation method computes the correlation coefficient between the expression profile of a test compound and each mutant profile in the training data set. The mutant profiles with the greatest similarity to the compound profile are considered the most likely targets. The linear combination approach finds a weighted sum of mutant profiles that best match the profile of the test compound. The most heavily weighted mutants are considered the most likely targets. The primary limitation of these methods is that they can only identify the target of a compound if a mutant strain for that target has been included in the training data set. For nine of the 11 titratable promoter profiles, no corresponding profile exists.

We next applied the MNI algorithm to identify probable targets of drug compounds. Unlike promoter insertions, which directly influence transcription, compounds predominantly affect protein activity and only indirectly influence transcription. As a result, the algorithm is more likely to identify genes in the same pathway as the affected protein rather than the target itself, such as transcriptionally regulated genes downstream of the target protein. On the other hand, when transcriptional feedback regulation is present in the pathway containing the

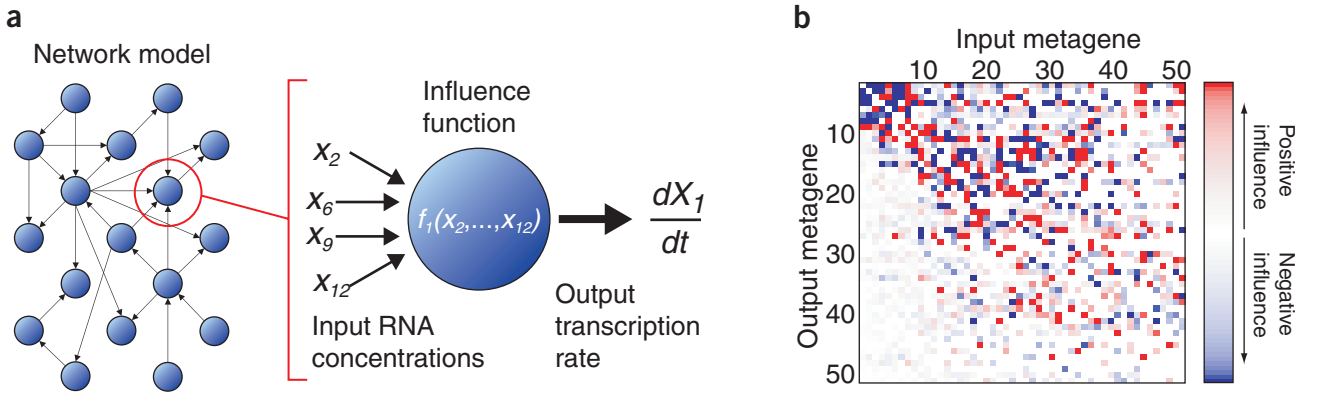

Figure 2 Structure of the network model. (**a**) The network model represents regulatory influences (arrows) between transcripts as influence functions for each gene (blue nodes). During phase 1, the MNI algorithm identifies the subset of transcripts (the input RNA concentrations) that influence the rate of transcription (the output transcription rate) of every other transcript. The algorithm also learns the coefficients of the interaction function that relates the inputs to outputs. (**b**) The colored matrix represents a portion of the yeast gene-network model identified by the MNI algorithm. Gene expression profiles are first reduced to a lower-dimensional set of *metagenes* (as described in the **Supplementary Notes**; the first 50 metagenes are shown) and a network model is trained for the metagenes. The metagenes represent characteristic expression profiles, which can be combined to approximate the expression profile of each transcript in the cell. Each pixel in the matrix represents a positive influence (red), negative influence (blue) or no influence (white) of the metagenes on each other. The metagene model, which can be transformed to describe regulatory influences between true genes, is used in phase 2 of the algorithm to distinguish compound targets from secondary responders.

targeted gene, it is likely that the algorithm will also assign a high rank to the targeted gene product. Thus, in analyzing the MNI predictions for compound treatments, we consider as targets both the pathways that are significantly overrepresented among the highly ranked genes and the highly ranked genes within those pathways. Pathways are identified as significantly overrepresented Gene Ontology (GO) (http://www.geneontology.org/) processes among the highly ranked genes.

We used the MNI algorithm to identify probable targets of 15 compounds, 13 of which were drawn from the Hughes compendium[1] and two from other studies[1,30]. Of the 15 compounds examined, nine have previously determined targets, while the targets of the other six compounds are unknown. The pathways and protein targets of the nine compounds of known mode of action are shown in **Table 2**. For each of these compounds, we used the MNI algorithm to rank more than 6,000 yeast genes by the likelihood that they were the targets of each drug treatment. We then subjected the 50 highest ranked genes to pathway analysis, using the GO Term Finder tool (http://www.yeastgenome.org), to identify overrepresented GO biological process annotations. The most significant annotation for each case is reported in **Table 2**, along with the highly ranked genes in that pathway.

The most overrepresented pathways identified among the genes ranked by the MNI algorithm matched the known targeted pathway for seven of the nine compounds (**Table 2**). The four compounds that target ergosterol biosynthesis (terbinafine, lovastatin, itraconazole and dyclonine) affect genes that are enriched for steroid and lipid metabolism, of which ergosterol biosynthesis is a more specific sub-category that also shows significant enrichment. The top pathways identified for each of the four compounds contain a high preponderance of ergosterol biosynthetic enzymes, and the gene encoding the known target protein for each respective compound is ranked near the top for each pathway (**Table 2** and **Fig. 3**). In determining the targets of hydroxyurea, a ribonucleotide reductase inhibitor, the algorithm identifies 'DNA replication,' the primary pathway of hydroxyurea's targets (Rnr2 and Rnr4), as the second most significant unique annotation. The algorithm identified *RNR4* and *RNR2* as the top ranked genes in that pathway (second and sixth overall, respectively), as well as two other genes encoding proteins in the ribonucleotide reductase complex (*RNR1* and *RNR3*). The highest ranked annotated processes were related to DNA repair, in which the RNR complex plays an important role[31]; the 'heteroduplex formation' genes *RAD51* and *RAD54* act in double-strand break repair through homologous recombination, and are highly ranked by the MNI algorithm. In the case of cycloheximide, the most significant annotation did not match the known pathway, but the MNI algorithm ranked two genes (*RPL26b* and *RPS29a*) in the top 50 that are members of the ribosome complex, which is targeted by the drug.

Table 1 Results of the MNI approach in identifying targets of genetic perturbations

Promoter mutant	Target	rank MNI	rank LC	rank C	rank R
tet-*IDI1*	*IDI1*	1	–	–	1
tet-*RHO1*	*RHO1*	4	–	–	1
tet-*YEF3*	*YEF3*	1	–	–	116
tet-*AUR1*	*AUR1*	1	–	–	14
tet-*FKS1*	*FKS1*	1	89	2	41
tet-*KAR2*	*KAR2*	1	–	–	64
tet-*CDC42*	*CDC42*	1	278	22	141
tet-*HMG2*	*HMG2*	1	–	–	19
tet-*PMA1*	*PMA1*	6	–	–	22
tet-*ERG11*	*ERG11*	42	–	–	2,820
tet-*CMD1*	*CMD1*	1	–	–	1

Results of association methods are provided for comparison. Promoter mutants are obtained by replacing the endogenous promoter with a tet-regulatable promoter. LC, linear combination; C, correlation; R, RNA change (*z*-score); –, association analysis methods do not identify target genes that are not themselves perturbed.

For three of the nine compounds with known modes of action (tunicamycin, nikkomycin, and 3-aminotriazole), the MNI algorithm did not identify the known target. However, for tunicamycin and 3-aminotriazole, the MNI algorithm did identify the targeted biosynthetic pathways and gene products acting adjacent to the known targeted proteins (Alg7 and His3, respectively) in those pathways (**Table 2**). The target of tunicamycin, Alg7, is an integral membrane protein of the endoplasmic reticulum (ER) that catalyzes the transfer of N-acetylglucosamine-1-P from UDP-N-acetylglucosamine to dolichol phosphate in the first step of lipid-linked oligosaccharide synthesis[32]. The MNI algorithm identified several protein-ER targeting proteins (Sec62, Sil1 and Sec59) among the top 50 most likely targets for tunicamycin. The final step in the synthesis of dolichol phosphate, the substrate of Alg7, is catalyzed by Sec59, which is ranked third in the top-ranked pathway by MNI (**Table 2**). Similarly, a target of 3-aminotriazole, His3, catalyzes the sixth step in the synthesis of histidine from 5-phosphoribosyl 1-pyrophosphate[32]. The following (seventh) step in that biosynthetic pathway is catalyzed by His5, which is ranked tenth in the top-ranked pathway by MNI (**Table 2**).

The MNI algorithm requires that the training perturbations influence a diversity of cell functions. If a particular cellular pathway does not show a response in any experiment, then a regulatory model for that pathway cannot be learned and thus no predictions can be made about that pathway. For instance, although in principle it is possible to use expression response profiles from environmental stimuli and stresses with this algorithm, we have found that even large data sets[16] sampling many unique environmental stresses can yield training data with low information content. Thus, the failure to identify the target of nikkomycin may be due to insufficient stimulation of the pathway related to its function.

We also examined the predicted target pathways and genes for the six compounds with currently unknown targets (**Supplementary Tables 3** and **4** online). For example, methyl methanesulfonate (MMS) is an alkylating agent that damages DNA; it is not thought to have a direct protein target. However, prior studies have shown that *rnr3* deletion strains are most sensitive to MMS treatment[1], and thus Rnr3 is a likely mediator of the effects of MMS. The MNI algorithm ranks *RNR3* as the sixth most likely target of MMS. Interestingly, the most significant pathway among the top ranked genes was 'sterol biosynthesis' ($P < 5.0 \times 10^{-5}$), containing the highly ranked genes *ERG5*, *CYB5*, *HMG1*, and *MVD1*. Previous studies have shown that disruption of ergosterol biosynthesis leads to MMS sensitivity[33], possibly due to defective mitochondrial mitogenesis, as discussed in the examination of the membrane-associated progesterone receptor family protein and probable sterol synthesis regulator, Dap1[34].

Table 2 Pathways and associated genes targeted by drug compounds

Drug	Known pathway	Known target	Significant GO ontology (rank, *P*-value)	Ranked pathway genes (rank)
Terbinafine	Ergosterol biosynthesis[41]	Erg1	**Steroid metabolism** (1, 10^{-14})	*ERG7* (4), ***ERG1*** **(5)**, *ERG8* (11), *ERG26* (13), *UPC2* (17), *ERG28* (18), *ERG11* (20), *DAP1* (33), *HES1* (34), *ATF2* (36), *ERG5* (49)
Lovastatin	Ergosterol biosynthesis[42]	Hmg2, Hmg1	**Lipid metabolism** (1, 10^{-4})	*BST1* (1), *ERG1* (18), *YSR3* (23), ***HMG2*** **(30)**, *LCB5* (31), *ERG13* (36), *VRG4* (48)
Itraconazole	Ergosterol biosynthesis[43]	Erg11	**Steroid metabolism** (1, 10^{-8})	***ERG11*** **(2)**, *ERG24* (4), *ERG1* (6), *ERG25* (13), *CYB5* (16), *ERG27* (19), *ATF2* (23)
Hydroxyurea	DNA replication[44]	Rnr2, Rnr4	Heteroduplex formation (1, 10^{-4})	*RAD51* (15), *RAD54* (47)
			DNA replication (2, 10^{-2})	***RNR4*** **(2)**, ***RNR2*** **(6)**, *RNR1* (14), *RNR3* (23)
Cycloheximide	Protein biosynthesis[45]	Ribosome	Nuclear mRNA splicing, via spliceosome (1, 10^{-4})	*SYF1* (3), *SMD3* (19), *HSH49* (42)
			–	***RPL26B*** **(32)**, ***RPS29A*** **(34)**
Tunicamycin	N-linked glycosylation[46]	Alg7	**Protein-ER targeting** (1, 10^{-3})	*SEC62* (1), *SIL1* (31), *SEC59*[a] (43)
Nikkomycin	Cell wall chitin biosynthesis[47]	Chs3	Protein amino acid alkylation(1, 10^{-3})	*SWD2* (3), *RMT2* (6)
Drugs not in the original compendium data set				
3-aminotriazole	Histidine biosynthesis[48]	His3	**Organic acid metabolism** (1, 10^{-7})	*FRM2* (8), *BIO5* (9), *YAT2* (10), *ARO10* (18), *ARO9* (20), *CHA1* (21), *BIO3* (31), *ARG1* (33), *ARG4* (37), *HIS5*[b] (42), *LYS1* (47), *SAM2* (50)
	Oxygen and reactive oxygen species metabolism[30]	Cta1		
Dyclonine	Ergosterol biosynthesis[1]	Erg2	**Sterol biosynthesis** (1, 10^{-18})	*ERG3* (1), *ERG6* (2), *CYB5* (3), ***ERG2*** **(4)**, *ERG11* (6), *ERG28* (10), *ERG1* (12), *ERG5* (13), *ERG27* (18), *MVD1* (23), *ERG24* (30), *ERG26* (37)
Novel drug with unknown mode of action				
PTSB	–	–	**Cell redox homeostasis** (1, 10^{-3})	***TRR1*** **(32)**, ***TRX2*** **(36)**

Bold text indicates matches with previously reported targets and pathways for each compound. –, known target pathway is not significantly overrepresented among ranked genes, or the target pathway or gene is unknown.

[a]Sec59 catalyzes the reaction immediately preceding Alg7 (tunicamycin's target) in the dolichol pathway of N-linked glycosylation. [b]His5 catalyzes the reaction immediately following His3 (3-aminotriazole's target) in the histidine biosynthesis pathway.

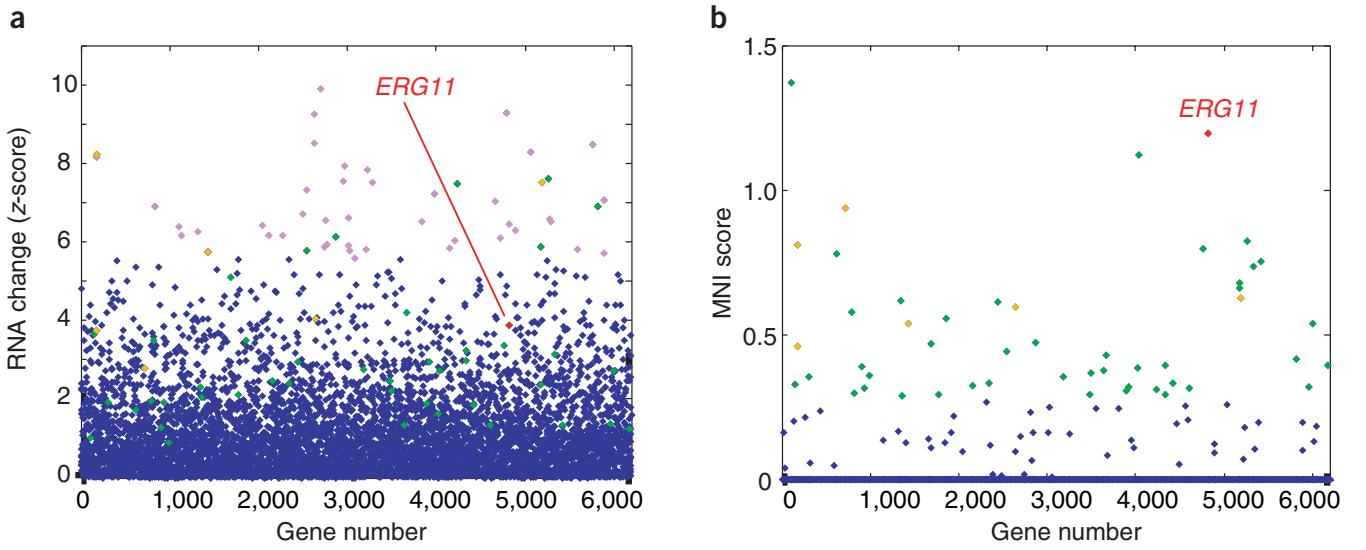

Figure 3 Predicted targets of itraconazole. (**a**) mRNA expression changes of 6194 yeast genes following treatment with itraconazole[1]. Changes are plotted as the *z*-score, x/σ_x, where x is the log(expression ratio) and σ_x is the standard error on the log expression ratio. (**b**) Targets of itraconazole predicted by the MNI algorithm using the expression changes in panel **a**. Higher MNI scores indicate higher likelihood that the gene is a target. *ERG11* (red), a known target of itraconazole, is the second most likely target identified with the MNI algorithm. Those genes ranked in the top 50 by MNI and annotated with the top ranked GO process, 'steroid metabolism' (in addition to *ERG11*), are shown in orange. The remaining genes ranked in the top 50 by MNI are shown in green. Genes ranked in the top 50 by *z*-score of mRNA expression change are shown in purple.

Gene and pathway rankings for MMS and all other drugs are provided in **Supplementary Tables 3** and **5** online. In the **Supplementary Notes**, we also examine the performance of the association analysis approaches and the raw expression change ranking in identifying target genes and pathways for all of the compounds considered (**Supplementary Tables 1–4**); we note that the MNI algorithm outperforms these approaches.

Overall, our results show that for most compounds the MNI algorithm is successful in correctly identifying the target pathway with the highest significance. Moreover, within a significant pathway, the algorithm typically ranks the target gene product higher than other genes in the pathway. This performance is likely due to the 'tournament' strategy used to rank genes (see **Supplementary Notes**). For a particular test compound profile, the algorithm is applied repeatedly to rank the genes. In each application of the algorithm, gene profiles are collapsed into a small number of principal components ('metagenes'). The metagenes represent the behavior of a group of similarly expressed genes. Such genes are likely to be involved in the same pathway. Thus, in initial rounds, genes within a pathway may be treated and ranked similarly. In each subsequent application of the algorithm, the one-third most highly ranked genes are selected and reanalyzed. Thus, a fewer number of gene profiles are collapsed into the representative metagenes, and the resolution of the predictions is improved. Therefore, in later iterations of the algorithm, genes within a pathway can be differentiated.

The MNI algorithm's ability to rank both genes and pathways suggests that the most probable targets of novel compounds can be identified as those that act within the most significantly overrepresented GO processes (pathways) and are highly ranked within those processes. The resulting small list of probable targets can then be validated for interaction with the compound by direct biochemical assays.

Here we demonstrate the use of this strategy on a tetrazole-containing compound, 1-phenyl-1H-tetrazol-5-ylsulfonyl-butanenitrile (PTSB), found to inhibit growth in both wild-type *Saccharomyces cerevisiae* (BY4743, IC_{50} of 25 μM) and human small lung carcinoma cells (A549, IC_{50} of 5 μM). We first determined the changes in steady-state gene expression in *S. cerevisiae* upon treatment with PTSB using oligonucleotide arrays. We used the MNI algorithm and the reverse-engineered network model described above to obtain a ranking of the most likely targets of PTSB. The most highly overrepresented GO process among the top 50 most likely perturbed genes was the 'cell redox homeostasis' annotation ($P < 2.2 \times 10^{-3}$). Two genes with that annotation are ranked in the top 50: thioredoxin reductase (*TRR1*, rank = 32) and thioredoxin (*TRX2*, rank = 36). To validate the predictions made by the MNI algorithm, we performed a biochemical assay to monitor the NADPH-dependent reduction of dithio(bis)nitrobenzoic acid (DTNB) by thioredoxin and thioredoxin reductase[35] (**Fig. 4**; Methods). The accumulation of the reduced DTNB product, a thiolate anion, was observed spectroscopically (λ_{max} = 412 nm) in the presence of 0, 1, 5 and 50 μM PTSB. The results

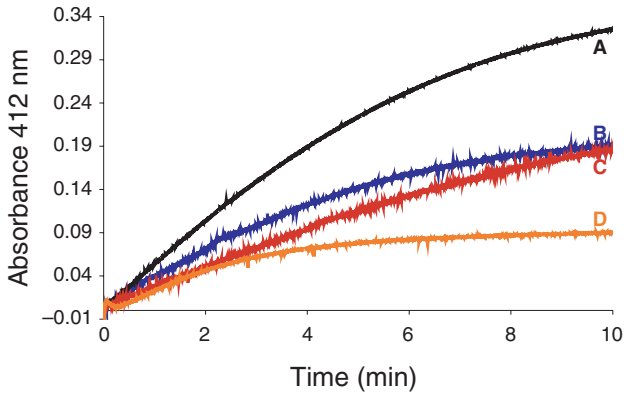

Figure 4 Thioredoxin/thioredoxin reductase activity assay. The reduction of 5,5′-dithio-bis-2-nitrobenzoic acid (DTNB)[35] was monitored in the presence of (A) 0 μM (B) 1 μM (C) 5 μM and (D) 50 μM PTSB. Concentrations of DTNB, thioredoxin reductase and thioredoxin are given in the Methods. Temperature = 20 °C; pH = 8.0.

demonstrate that PTSB efficiently inhibits the thioredoxin/thioredoxin reductase system.

We have presented an approach for the identification of drug targets using a computational model of genetic network interactions determined with gene expression array data. Data sets appropriate for analysis with the MNI algorithm are becoming increasingly available in several model organisms[36–38]. The reverse-engineered gene network models that are at the core of the MNI algorithm will also become more informative with an increase in data over a wider range of cellular behavior. We anticipate that such refined models will contain even greater predictive power for drug target identification.

METHODS

Public expression data. Two publicly available sets of gene expression profiles[1,29] served as the training data set for the MNI algorithm, with two primary modifications. First, information regarding the identity of compounds used to treat the cells and the identity of the mutated genes in each profile was not provided to the MNI algorithm. Thus, the data set was representative of an experimental situation where only treatments with unknown modes of action were applied to the model organism. Second, if the test expression profile (that is, the profile for which targets were to be identified) was part of the 515 profiles in the compendium, it was removed from the training data set before analysis. Note also that an additional public data set[30] was used as the source of expression data for one compound, 3-aminotriazole. All expression profiles were preprocessed before analysis: missing expression ratios were set to zero, and missing standard errors were estimated as described in the **Supplementary Notes**.

DNA microarray construction for PTSB experiments. A set of 6,307 synthesized oligonucleotide 70-mer probes including ten controls was obtained from Operon Technologies. The plates of DNA were suspended in 3× SSC (0.45 M NaCl, 45 mM sodium citrate, pH 7.0) to make printable aliquots. The DNA solutions were spotted on CMT-GAPS II slides (Corning) using OmniGrid Accent (GeneMachines) microarraying robot equipped with a Stealth Printhead (SPH32, Telechem International) containing 16-Stealth Micro Spotting Pins (SMP4, Telechem International). Postprocessing of the slides was accomplished according to published procedures[39].

Drug treatment and preparation of microarray sample. An overnight culture of a drug-sensitive strain of *S. cerevisiae* was diluted to an OD_{600} of 0.1, treated with 5 μM PTSB and then grown to an OD_{600} of 0.8. Total RNA was isolated from the flash-frozen cultured yeast cells using the acidic phenol method. Poly(A) RNA was isolated using an oligo(dT) resin (Oligotex, Qiagen). cDNA was synthesized followed by double-strand synthesis. *In vitro* transcription was then used for amplification of antisense RNA (aRNA) (Amino Allyl MessageAMP aRNA kit, Ambion). The *in vitro* transcription employed 5-(3′-amino-allyl)-dUTP for dye conjugation. The control and experimental probes were coupled with Cy3- and Cy5- N-hydroxysuccinamide esters (Amersham Biosciences), respectively, and purified using a MEGAclear kit (Ambion). The samples were concentrated and fragmented before hybridization. Each experiment was conducted in duplicate.

Data acquisition and analysis. The microarrays were scanned with a GenePix 4000B array scanner (Axon Instruments) using GenePix 3.0 software to quantify the Cy3- and Cy5-fluorescence intensities at each spot and determine the background signal intensities. Signal intensities greater than three standard deviations above the average background were considered for analysis. A scaling factor was calculated using the ratio of the Cy3 average mean signal intensity to the Cy5 average mean signal intensity. The scaling factor was applied to normalize the two channels. The Yeast Protein Database (YPD) and the GeneSpring software package (Silicon Genetics) were used for data analysis.

Validation of PTSB target: thioredoxin/thioredoxin reductase assay. The solution assay of coupled thioredoxin-thioredoxin reductase activity using DTNB was carried out by the method of Holmgren and Reichard[35], with the following modifications. To an assay mixture of 10 mM Tris, we added 3.12 mM EDTA (pH 8.0), NADPH and DTNB to final concentrations of 0.05 mM and 0.33 mM, respectively. DTNB was prepared before the experiment, in ethanol, as a 100 mM stock solution. To this mixture, *Escherichia coli* thioredoxin reductase was added to a concentration of 1 μM, as was a variable amount of PTSB (see text for details). The reaction was initiated by the addition of *E. coli* thioredoxin (250 μM, final concentration), and monitored by absorption change due to the thiolate anion at 412 nm (at pH = 8.0, ε_{412} = 13.6 $mM^{-1}cm^{-1}$). The *E. coli* thioredoxin I protein (*trxA* gene product) is 34% identical to *S. cerevisiae* thioredoxin (Trx2; identified by MNI) and 30–40% identical to human thioredoxins. The *E. coli* thioredoxin reductase protein (*trxB* gene product) is 47% identical to *S. cerevisiae* thioredoxin reductase (Trr1; identified by MNI) and approximately 25% identical to human thioredoxin reductases.

MNI algorithm. The algorithm and underlying assumptions are described in detail in the **Supplementary Notes**. Here we provide a brief summary. The algorithm operates in two phases. In the first phase (the training phase), a model of regulatory influences in the cell is learned from an $N \times M$ data matrix, X, consisting of measurements of steady-state expression ratios of N genes in M experiments. In prior work[15], we showed that such a regulatory model can be constructed provided that specific genes are perturbed in each of the M experiments. The gene-specific perturbations enable the construction of an $N \times M$ matrix, P, of external influences on the genes. Regulatory influences are obtained as coefficients in the matrix A that provide a sparse solution to a linearized steady-state model of the regulatory network: $A(X - 1) = P$. In the MNI algorithm, a similar strategy is used. However, gene-specific perturbations are assumed to be unavailable. Thus the matrix P is unknown and our prior approach is inapplicable. To estimate the network model A, with no data on P, the MNI algorithm uses a recursive strategy. The algorithm begins by using a naive model of the regulatory structure (i.e., no genes regulate any other genes) to estimate P from the expression data X. The estimate of P is then used, along with X, to determine A by principal components regression[40]. The estimates of A and P are then used to recursively reestimate one another until the estimates converge. The recursive approach is much like the EM algorithm[28] commonly used to train Bayesian networks. The estimation of P corresponds to the 'E-step,' and the estimation of A corresponds to the 'M-ste'.

In past work[15], expression-ratio data were used to compose the data matrix X, thereby allowing the inference of a linearized model of regulatory influences. The MNI algorithm, however, uses log-transformed expression-ratio data in the data matrix X. This transformation improves the statistical properties of the data by stabilizing the variances of the expression ratios, and it enables the identification of a log-linear model[19] of gene regulation. The log-linear model enables the capture of some nonlinear properties of the regulatory network, providing better predictive power.

In the second phase of the algorithm, the A matrix, representing a model of regulatory influences in the cell, is used to estimate the targets of a test compound. The test compound is incorporated in the model as an $N \times 1$ vector, p, of gene-specific influences that result in the log-transformed expression-ratios, x, measured for the compound. The p vector is then calculated directly from the log-linear regulatory model as: $P = Ax$. The significance of each element of the p vector is then calculated as a z-score. Genes are ranked according to the z-score of their corresponding element in the p vector, and the top-ranked genes and pathways are selected as probable targets of the test compound.

Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

Support for this work was provided by the Department of Energy, the National Institutes of Health, the National Heart, Lung and Blood Institute Proteomics Initiative, the Whitaker Foundation, the National Science Foundation, the Fondazione Telethon, Boston University and the Pharmaceutical Research and Manufacturers of America Foundation.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details)

Published online at http://www.nature.com/naturebiotechnology/

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

2. Courcelle, J., Khodursky, A., Peter, B., Brown, P.O. & Hanawalt, P.C. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics* **158**, 41–64 (2001).
3. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5**, 262–275 (2004).
4. Miklos, G.L.G. & Maleszka, R. Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.* **22**, 615–621 (2004).
5. Giaever, G. *et al.* Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* **21**, 278–283 (1999).
6. Giaever, G. *et al.* Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc. Natl. Acad. Sci. USA* **101**, 793–798 (2004).
7. Lum, P.Y. *et al.* Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell* **116**, 121–137 (2004).
8. Parsons, A.B. *et al.* Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat. Biotechnol.* **22**, 62–69 (2004).
9. Marton, M.J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* **4**, 1293–1301 (1998).
10. Stoughton, R. & Friend, S.H. Methods for identifying pathways of drug action. *US* Patent No. 5,965,352 (2003).
11. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21**, 1337–1342 (2003).
12. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
13. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
14. de la Fuente, A., Brazhnik, P. & Mendes, P. Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.* **18**, 395–398 (2002).
15. Gardner, T.S., di Bernardo, D., Lorenz, D. & Collins, J.J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105 (2003).
16. Gasch, A.P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
17. Herrgard, M.J., Covert, M.W. & Palsson, B.O. Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol.* **15**, 70–77 (2004).
18. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
19. Liao, J.C. *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA* **100**, 15522–15527 (2003).
20. Rice, J. & Stolovitzky, G. Making the most of it: pathway reconstruction and integrative simulation using the data at hand. *Drug Discov. Today: BioSilico* **2**, 70–77 (2004).
21. Ronen, M., Rosenberg, R., Shraiman, B.I. & Alon, U. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA* **99**, 10555–10560 (2002).
22. Tegner, J., Yeung, M.K., Hasty, J. & Collins, J.J. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. USA* **100**, 5944–5949 (2003).
23. Yeung, M.K.S., Tegner, J. & Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**, 6163–6168 (2002).
24. Kholodenko, B.N. *et al.* Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. USA* **99**, 12841–12846 (2002).
25. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Combining location and expression data for principled discovery of genetic regulatory network. *Proc. Pacific. Symp. Biocomp.* **7**, 437–449 (2002).
26. Kalir, S. & Alon, U. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell* **117**, 713–720 (2004).
27. Schmitt, W.A.J., Raab, R.M. & Stephanopoulos, G. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Res.* **14**, 1654–1663 (2004).
28. Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. [Ser. B]* **39**, 1–38 (1977).
29. Mnaimneh, S. *et al.* Exploration of essential gene functions via titratable promoter alleles. *Cell* **118**, 31–44 (2004).
30. Ueda, M. *et al.* Effect of catalase-specific inhibitor 3-amino-1,2,4-triazole on yeast peroxisomal catalase *in vivo*. *FEMS Microbiol. Lett.* **219**, 93–98 (2003).
31. Chabes, A. *et al.* Survival of DNA damage in yeast directly depends on increased dNTP levels allowed by relaxes feedback inhibition of ribonucleotide reductase. *Cell* **112**, 391–401 (2003).
32. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
33. Bennett, C. *et al.* Genes required for ionizing radiation resistance in yeast. *Nat. Genet.* **29**, 426–434 (2001).
34. Hand, R.A., Jia, N., Bard, M. & Craven, R.J. *Saccharomyces cerevisiae* Dap1p, a novel DNA damage response protein related to the mammalian membrane-associated progesterone receptor. *Eukaryot. Cell* **2**, 306–317 (2003).
35. Holmgren, A. & Reichard, P. Thioredoxin 2: cleavage with cyanogen bromide. *Eur. J. Biochem.* **2**, 187–196 (1967).
36. NCBI Gene Expression Omnibus: http://www.ncbi.nlm.nih.gov/geo.
37. Stanford Microarray Database: http://genome-www5.stanford.edu.
38. The Alliance for Cellular Signaling: http://www.signaling-gateway.org/.
39. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Meth. Enzymol.* **303**, 179–205 (1999).
40. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer, New York, 2001).
41. Klobucnikova, V. *et al.* Terbinafine resistance in a pleiotropic yeast mutant is caused by a single point mutation in the *ERG1* gene. *Biochem. Biophys. Res. Commun.* **309**, 666–671 (2003).
42. Rine, J., Hansen, W., Hardeman, E. & Davis, R.W. Targeted selection of recombinant clones through gene dosage effects. *Proc. Natl. Acad. Sci. USA* **80**, 6750–6754 (1983).
43. Daum, G., Lees, N.D., Bard, M. & Dickson, R. Biochemistry, cell biology and molecular biology of lipids of *Saccharomyces cerevisiae*. *Yeast* **14**, 1471–1510 (1998).
44. Rittberg, D.A. & Wright, J.A. Relationships between sensitivity to hydroxyurea and 4-methyl-5-amino-1-formylisoquinoline thiosemicarbazone (MAIO) and ribonucleotide reductase *RNR2* mRNA levels in strains of *Saccharomyces cerevisiae*. *Biochem. Cell Biol.* **67**, 352–357 (1989).
45. Stocklein, W. & Piepersberg, W. Binding of cycloheximide to ribosomes from wild-type and mutant strains of *Saccharomyces cerevisiae*. *Antimicrob. Agents Chemother.* **18**, 863–867 (1980).
46. Barnes, G., Hansen, W.J., Holcomb, C.L. & Rine, J. Asparagine-linked glycosylation in *Saccharomyces cerevisiae*: genetic analysis of an early step. *Mol. Cell Biol.* **4**, 2381–2388 (1984).
47. Gaughran, J.P., Lai, M.H., Kirsch, D.R. & Silverman, S.J. Nikkomycin Z is a specific inhibitor of *Saccharomyces cerevisiae* chitin synthase isozyme Chs3 *in vitro* and *in vivo*. *J. Bacteriol.* **176**, 5857–5860 (1994).
48. Anderson, R.M. *et al.* Yeast life-span extension by calorie restriction is independent of NAD fluctuation. *Science* **302**, 2124–2126 (2003).