Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*

Sriram Chandrasekaran[a,b] **and Nathan D. Price**[a,b,c,1]

[a]Center for Biophysics and Computational Biology, [b]Institute for Genomic Biology, and [c]Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana, IL 61801

Edited by Peter Schuster, University of Vienna, Vienna, Austria, and approved September 2, 2010 (received for review April 15, 2010)

Prediction of metabolic changes that result from genetic or environmental perturbations has several important applications, including diagnosing metabolic disorders and discovering novel drug targets. A cardinal challenge in obtaining accurate predictions is the integration of transcriptional regulatory networks with the corresponding metabolic network. We propose a method called probabilistic regulation of metabolism (PROM) that achieves this synthesis and enables straightforward, automated, and quantitative integration of high-throughput data into constraint-based modeling, making it an ideal tool for constructing genome-scale regulatory-metabolic network models for less-studied organisms. PROM introduces probabilities to represent gene states and gene–transcription factor interactions. By using PROM, we constructed an integrated regulatory-metabolic network for the model organism, *Escherichia coli*, and demonstrated that our method based on automated inference is more accurate and comprehensive than the current state of the art, which is based on manual curation of literature. After validating the approach, we used PROM to build a genome-scale integrated metabolic-regulatory model for *Mycobacterium tuberculosis*, a critically important human pathogen. This study incorporated data from more than 1,300 microarrays, 2,000 transcription factor–target interactions regulating 3,300 metabolic reactions, and 1,905 KO phenotypes for *E. coli* and *M. tuberculosis*. PROM identified KO phenotypes with accuracies as high as 95%, and predicted growth rates quantitatively with correlation of 0.95. Importantly, PROM represents the successful integration of a top-down reconstructed, statistically inferred regulatory network with a bottom-up reconstructed, biochemically detailed metabolic network, bridging two important classes of systems biology models that are rarely combined quantitatively.

constraint-based modeling | flux balance analysis | metabolic networks | transcriptional regulation | probabilistic regulation of metabolism

Systems biology seeks to quantitatively simulate cell behavior by studying interactions between various biomolecules and the properties that emerge from these interacting networks. Reconstruction, simulation, and validation of biological networks thus represent key activities in systems biology. Large research communities have formed around different network classes, largely using quite different tools for their analysis. Metabolic networks are typically reconstructed and analyzed by using detailed chemistry-based methods (1, 2), whereas transcriptional regulatory networks have typically been reconstructed based on high-throughput protein–DNA interaction data and statistical inference of functional relationships from genomics and transcriptomics data (3–6). Although both these fields have garnered considerable attention, significant work remains to be done in integrating these two types of biological networks and the different styles of systems models that are used to describe them (7). A seamlessly integrated metabolic-regulatory network would enable us to better predict how genetic mutations and transcriptional perturbations are translated into flux responses at the metabolic level. There have been significant successes in this endeavor (8, 9), yet substantial challenges remain.

Currently the most commonly used genome-scale analysis method that links the transcriptome of an organism with metabolism is regulatory flux balance analysis (RFBA) (8, 10). RFBA incorporates regulatory constraints into traditional constraint based modeling of metabolic networks [i.e., flux balance analysis (FBA)]. In constraint-based analysis, the functional states a network can choose are limited by various physicochemical constraints (11). By using matrix formalism to depict the metabolic network, the tools of linear programming, linear algebra, and convex analysis can be used to study the network properties. In the case of RFBA, the metabolic network is not only restricted by mass, thermodynamic, and energy constraints, but also by the gene regulatory network that controls it. Steady-state RFBA (SRFBA) (12) and integrated FBA (iFBA) (13) are similar methods based on Boolean logic. SR-FBA uses the same genome-scale integrated metabolic regulatory network as RFBA but characterizes its steady-state behavior, whereas iFBA uses differential equations to model a subset of the regulatory network. Methods based on stochastic models or differential equations (14, 15) are usually restricted to modeling small systems and have not been extended thus far to the genome scale.

There are several shortcomings in performing integrated modeling using RFBA and other Boolean logic-based methods. The RFBA model simplifies the relationship between the transcriptome and the metabolome to a binary process, wherein genes and reaction fluxes can only have two states in the population: on or off. However, perhaps the biggest impediment to performing integrative modeling using RFBA across a large number of species is the absence of an automated algorithm for determining the Boolean rules for relating the regulator with its target. Although the manual process can be accurate in modeling metabolic regulation, manual reconstruction greatly limits the number of interactions that can be modeled, and thus very few genome-scale metabolic-regulatory models existed before our study (8, 9, 16). Partly because of the manual nature of this process, the interaction rules are also qualitative in nature, with genes being turned completely on or off, and cannot take intermediate values. Given the large number of interactions, it is extremely difficult to write Boolean rules and identify significant interactions at the genome scale. Finally, this process also requires extensive literature search, which is why the first two metabolic-regulatory models were made for the extremely well studied model organisms *Escherichia coli* and *Saccharomyces cerevisiae*. Hence there is a critical need for an algorithm that can build and model integrated metabolic-regulatory networks in an automated fashion. It is also imperative that the algorithm make full use of high-

Author contributions: S.C. and N.D.P. designed research; S.C. performed research; S.C. and N.D.P. analyzed data; and S.C. and N.D.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ndprice@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1005139107/-/DCSupplemental.

throughput data, as much of the data generated currently for most new organisms are high-throughput in nature.

Given these needs, we present a method called probabilistic regulation of metabolism (PROM), which overcomes some of the drawbacks of RFBA and other similar methods that use Boolean rules. PROM enables direct integration of the transcriptional and metabolic networks for modeling and overcomes the need for manually writing the Boolean rules by automatically quantifying the interactions from high-throughput data, thereby greatly increasing the capacity to generate genome-scale integrated models. The model framework, based on constraint-based analysis, is designed to circumvent the need for kinetic parameters for metabolic modeling, and importantly does not assume direct correlation between enzyme activity and mRNA expression. PROM is robust to noise in high-throughput data and can be easily integrated with automated algorithms for network inference. The PROM algorithm uses conditional probabilities for modeling transcriptional regulation, similar to the probabilistic Boolean networks of Shmulevich et al. (17) and uses FBA (18) for modeling metabolic networks. More broadly, PROM represents an important step that unifies two key systems biology approaches: biochemical reaction networks and statistical inference networks (7).

Model Framework: Genome-Scale Integrative Modeling Using PROM

The construction of an integrated metabolic-regulatory network using PROM requires the following: (*i*) reconstructed genome-scale metabolic network (19); (*ii*) regulatory network structure, consisting of transcription factors (TFs) and their targets (20); (*iii*) abundant gene expression data, in which the transcriptome has been measured under various environmental and genetic perturbations; and (*iv*) additional interactions involving enzyme regulation by metabolites and proteins. Table 1 summarizes the data sources and features for the models presented.

PROM introduces probabilities to represent gene states and gene–TF interactions. For example, the probability of gene A being on when the regulating TF B is off is given by $P(A = 1|B = 0)$; similarly, $P(A = 1|B = 1)$ gives the probability of A being on when B is on. The relationship between TF and target gene is then quantified by using microarray data. By using this interaction data, we can model the effect of perturbations to the regulatory network using PROM. To predict the effect of a TF KO on a gene A, which is the probability $P(A = 1|B = 0)$, we estimate the number of microarray samples wherein the target gene A is on when the TF B is off. If the data set is large enough, we can get a robust estimate of the probability for the interaction. So, if the probability associated with a gene being on is 0.8, then we estimate that in 80% of the samples we find the gene to be on, and 20% of the samples it is off or not expressed. To model the effect of the KO at the genome scale, the states of all its target genes are determined. These probabilities are then used to constrain the fluxes through the reactions controlled by the target genes. For the example discussed earlier, the flux through the reaction regulated by gene A would not exceed the maximum flux possible through the reaction [i.e., maximum velocity (V_{max})] if it is on, and would be zero when it is off. Hence, on average, the maximum flux through the reaction in the population would be $0.8 \times V_{max}$ or, in general, the upper bound for the flux is $p \times V_{max}$, where p is the probability of the gene being on. The systemic reaction V_{max} is estimated by flux variability analysis (FVA) (21) (*SI Methods*) on the unregulated metabolic model. To account for other factors that may affect enzyme activity such as translational, posttranslational, and metabolite interactions, we use an approach analogous to that of Shlomi et al. (22) and use these constraints as cues to determine the most likely flux through a particular enzyme. Unlike thermodynamic or environmental constraints that cannot be violated, the regulatory constraints are "soft" constraints, so the system can exceed these constraints to maximize growth, but with a penalty. Following this procedure, we arrive at an optimal model, which satisfies most or all of the regulatory constraints (Figs. 1 and 2).

Table 1. Model features and accuracy in predicting KO phenotypes in each organism

Feature	*E. coli*	*M. tuberculosis*
Metabolic model	IAF1260	iNJ661
Metabolic reactions	2,382	1,028
Regulatory data	RegulonDB	Balazsi et al. (27)
Regulatory interactions	1,773	218
Microarrays	907	437
Total genes in the model	1,400	691
Validation data set	1,875 growth phenotypes	30 TF KO
Accuracy, %	85	95
Sensitivity, %	70	83
Specificity, %	91	100

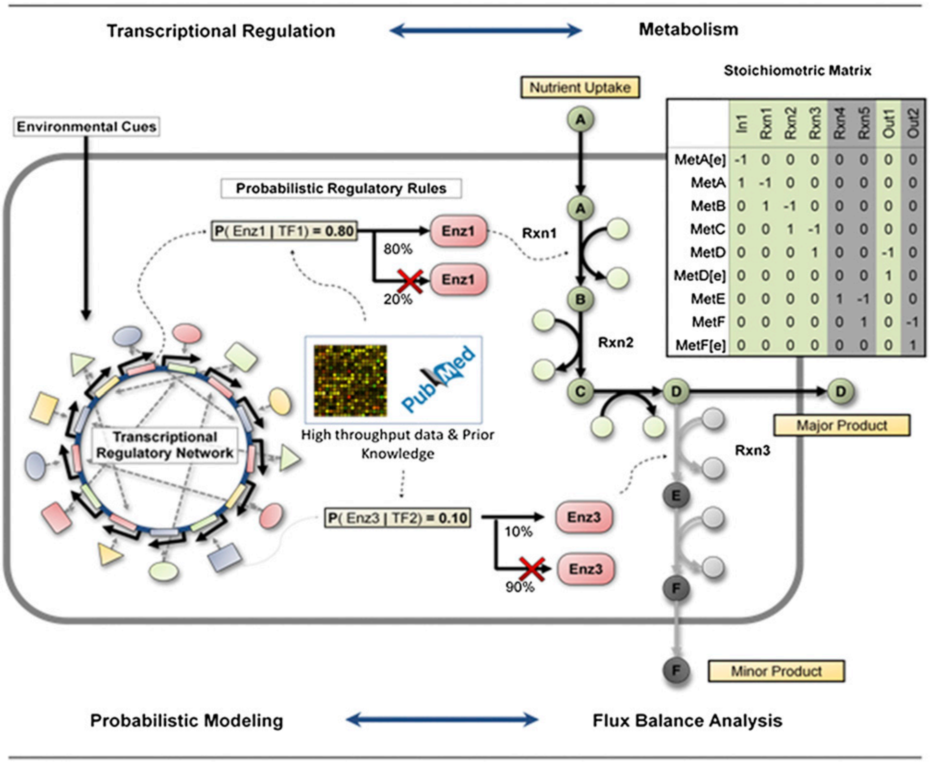
When the constraints have been set, the optimal growth of the regulated network is determined by solving a linear optimization problem as in FBA (18). Briefly, constraint-based analysis involves solving the linear optimization problem $max\ w^{T}v$, subject to constraints $S.v = 0$ and $lb \leq v \leq ub$, where S is the stoichiometric matrix, v is a flux vector representing a particular flux configuration, $w^{T}v$ is the linear objective function, and lb and ub are vectors containing the minimum and maximum fluxes through each reaction. PROM finds a flux distribution that satisfies the same constraints as FBA plus additional constraints resulting from the transcriptional regulation: $\min(\kappa.\alpha + \kappa.\beta)$, subject to constraints $lb' - \alpha \leq v \leq ub' + \beta$ and $\alpha,\beta \geq 0$, where lb' and ub' are constraints based on transcriptional regulation, α and β are positive constants that represent deviation from those constraints, and κ represents the penalty for such deviations. The higher the value of κ, the greater is the constraint on the system based on transcriptional regulation. For values of κ significantly greater than 1, the regulatory constraints become "hard," and for values less than 0.1 they become less pronounced (Fig. S1). We have chosen a κ value of 1 for all our simulations as it represents a tradeoff between the two extremes. The optimization problem was solved using the GNU Linear Programming Kit solver.

The probabilistic framework used in PROM has many other advantages to model regulation, apart from the fact that it can be readily learned from high-throughput data. The first is the ability to cope with uncertainty inherently present in biological systems (23)—both intrinsic, at the level of gene expression; and extrinsic, caused by other components in the cell—resulting in cell-cell variation even in isogenic populations (24). Hence no two sets of biological measurements are identical, and by considering genes to be completely on or off, we ignore these variations. By using a metric to quantify the interactions, we indirectly account for the presence of noise in the data and in the cell.

Unlike in RFBA, in which the gene can only be completely on or off, in PROM the gene state can take intermediate values. So for a gene A with multiple TFs B, C, and D, we could have $P(A = 1|B = 1) = 0.3$, $P(A = 1|C = 1) = 0.5$ and $P(A = 1|D = 1) = 0.8$. The use of probabilities would thus enable the model to differentiate between a strong regulator and a weak one, and the relative influence of each TF on the target can be modeled.

PROM's implementation of probabilities, constraint-based analysis, and on/off formalism also precludes the need for knowing individual enzyme parameters like V_{max} and kinetic constants. In both RFBA and PROM, the maximum flux through a reaction is determined by the topology of the network and no additional parameters are needed for metabolic modeling. Nonetheless, additional constraints can be incorporated into the model when available.

An added advantage of the use of probabilistic on/off formalism is that it does not assume that mRNA levels and enzyme



Stoichiometric Matrix

	In1	Rxn1	Rxn2	Rxn3	Rxn4	Rxn5	Out1	Out2
MetA[e]	-1	0	0	0	0	0	0	0
MetA	1	-1	0	0	0	0	0	0
MetB	0	1	-1	0	0	0	0	0
MetC	0	0	1	-1	0	0	0	0
MetD	0	0	0	1	0	0	-1	0
MetD[e]	0	0	0	0	0	0	1	0
MetE	0	0	0	0	1	-1	0	0
MetF	0	0	0	0	0	1	0	-1
MetF[e]	0	0	0	0	0	0	0	1

Fig. 1. Overview of the process used to integrate the metabolic and regulatory network using PROM. The metabolic network is represented using a stoichiometric matrix and regulatory interactions are represented as probabilities. The TF states are determined based on environmental conditions; the state of TF is then used to determine the on/off state of the target genes based on probabilities estimated from microarray data. The probabilities are then used to constrain the fluxes through the metabolic network.

levels are directly correlated. That is, a change in expression does not result in a proportional change in flux or the flux bounds. Instead, PROM considers only changes in gene expression that turn the activity of the enzyme on or off. If the mRNA coding for a particular protein is absent, it is reasonable to assume that the protein is also not present in the cell. Also, the model does not restrict the flux state to be perfectly correlated with the on/off probabilities as well. They are used only used as cues to determine the most likely upper bound on the system. As they are just bounds, the optimal flux level could be well below the bounds and in our case, as the bounds are soft, they could to some extent be higher as well. Given the limited knowledge we have on the state of various other factors that affect enzyme activity, the use of gene expression would be a powerful constraint on the system. We demonstrate by using PROM that we can predict phenotypes qualitatively and quantitatively by applying regulatory constraints on the metabolic network derived from microarrays.

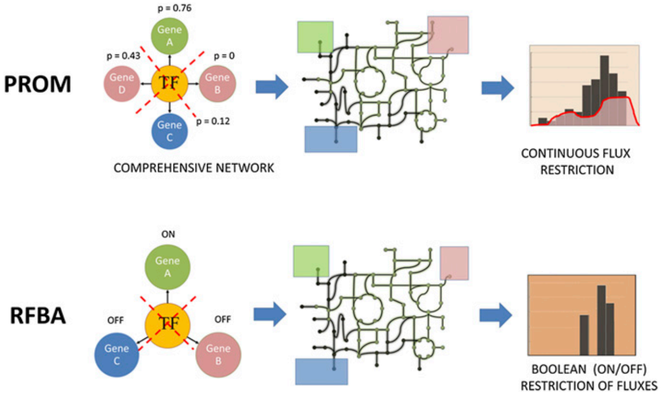


Fig. 2. Comparison between PROM (*Top*) and RFBA (*Bottom*): a perturbation to a TF results in alteration in expression of its target genes. These are then mapped onto the metabolic network. Depending on the gene state, the fluxes through the reactions are constrained and the optimal growth rate is determined by using FBA. In PROM, the constraints based on gene expression (red) are used as cues to obtain the optimal flux state, whereas in RFBA, genes and fluxes can have only two states (on/off). Further, PROM's automated inference of interactions and probabilistic formalism enables it to create comprehensive models.

Results and Discussion

Comparison with RFBA: PROM's Automated Quantification of Interactions Is More Accurate than Manual Curation in Predicting Phenotypes. We compared PROM's ability to predict the growth phenotypes of TF KO against RFBA using data from Covert et al. (8), who predicted growth phenotypes from A Systematic Annotation Package (ASAP) for community analysis of genomes database (25). As both SRFBA and RFBA models use the same Boolean network, we expect them to give the same phenotype results. The ASAP database has growth phenotypes of several *E. coli* gene KOs under various conditions. From the database, we identified 15 TFs whose phenotypes were measured under 125 different growth conditions. PROM was more accurate than RFBA in predicting these growth phenotypes. The predictions made by both models were nearly identical except in the phenotypes involving the TF KO, ilvY. RFBA predicted the phenotype to be lethal in all 125 conditions in which the gene ilvY was knocked out, PROM predicted it to be lethal in 33 cases, whereas in fact it was lethal in 56 cases. PROM's prediction was closer to the actual value than RFBA's. Overall, RFBA had an accuracy of 82.5% whereas PROM had an accuracy of 85% in predicting phenotypes (Table 1). The difference in accuracy stems from the "strict" regulatory rules in RFBA whereby genes can only be considered completely on or off within the population. Because of this rigid way of determining the gene state, RFBA wrongly predicts some KOs to be lethal or vice versa. PROM, on the contrary, is "softer" than RFBA, yet sensitive enough to identify suboptimal and lethal KOs. This is exemplified in the TF KO discussed earlier in which RFBA predicted the phenotype to be lethal in all conditions whereas PROM more accurately predicted it to be lethal only in a subset of the conditions. Fig. S2 contains the phenotype predictions by both RFBA and PROM on all KOs and discusses other minor differences between the two models.

PROM's accuracy in comparison with RFBA is highly significant, given that PROM computationally quantified the interactions using high throughput data whereas the Boolean rules for RFBA were constructed through detailed manual curation of literature. We also compared our results using the same metabolic model (ijr904) used previously (8) and found that PROM outperformed RFBA with the older model as well (Fig. S2). The same parameters were used for prediction of KOs for both RFBA and PROM to ensure fair comparisons. The only additional interactions added to the *E. coli* PROM model were those with metabolites in the environment. These were identical to those reported in the RFBA model of Covert et al. (8), and were included because these interactions would play a major role when simulating the effect of various growth conditions.

Quantitative Phenotype Prediction in *E. coli*. As PROM enables continuous restriction of flux, it can also be used to predict growth phenotypes quantitatively. We used quantitative growth rate data from Covert et al. (8), in which six strains with KOs of key transcriptional regulators in the oxygen response (Δ*arcA*, Δ*appY*, Δ*fnr*, Δ*oxyR*, Δ*soxS*, and the double KO Δ*arcA*Δ*fnr*) were constructed and growth rates were measured in aerobic and anaerobic glucose minimal medium conditions. PROM predicted the growth rates of the 14 KO phenotypes accurately with a correlation coefficient of 0.95 ($P < 1e\text{-}3$; Table 2). This correlation is the same as was achieved by a second iteration of building the RFBA model (8) that was updated specifically to improve predictive power for these sets of TF KOs. Our results are hence very significant given the scale of the problem—we are predicting perturbations in the TRN and its effect on a different system—the metabolic network based on automated quantification of interactions using high-throughput data.

PROM Helps Create Comprehensive Genome-Scale Cellular Network Models. The critical advance of PROM is that not only does it show equivalent or improved accuracy, but it utilizes high-throughput data and thus can be used to construct much more comprehensive models. In this case, the PROM model for *E. coli* had 136 TFs regulating 708 metabolic genes through 1,773 interactions, whereas the RFBA model had 104 TFs regulating 479 metabolic genes through 724 interactions. Although it is certainly true that a larger RFBA model could be manually reconstructed now from additionally generated data, the PROM approach is appealing because of its derivation from high-throughput data, allowing for the rapid reconstruction of integrated metabolic-regulatory networks for a broader range of organisms. Although the regulatory data used from RegulonDB were high-confidence interactions, we demonstrate in the subsequent modeling of *Mycobacterium tuberculosis* that the algorithm performs well even with noisy or statistically inferred interactions.

Table 2. Growth rate predicted by PROM in 16 phenotype conditions

Culture	Actual	PROM
WT + O2	0.71	0.7382
WT − O2	0.49	0.385
ΔarcA + O2	0.69	0.7651
ΔarcA − O2	0.38	0.3224
Δfnr + O2	0.63	0.5635
Δfnr − O2	0.41	0.2181
Δfnr/ΔarcA + O2	0.65	0.6596
Δfnr/ΔarcA − O2	0.3	0.204
ΔappY + O2	0.64	0.7152
ΔappY − O2	0.48	0.3287
ΔoxyR + O2	0.64	0.7876
ΔoxyR − O2	0.48	0.3287
ΔsoxS + O2	0.72	0.7687
ΔsoxS − O2	0.46	0.379

PROM predicted the growth rates with a correlation coefficient of 0.95.

Integrative Modeling in *M. tuberculosis* H37Rv Using PROM. Tuberculosis is a major cause of death in the third world (26), yet much remains to be learned about this pathogen, especially with respect to its regulatory and metabolic networks. PROM helps to expand our knowledge about this organism by integrating various high-throughput data and can potentially determine the functions of less-studied genes and gene networks and the impact of perturbing those on global metabolism. We used PROM to build a genome-scale integrated metabolic-regulatory network for *M. tuberculosis*. The regulatory data for *M. tuberculosis* was compiled by Balazsi et al. (27) by using gene regulatory interactions in databases such as MTbreglist (28), interactions documented in the literature, and enlarged this core network to include gene pairs that have orthologues with confirmed relationship in *E. coli*. To our knowledge, this is the largest transcriptional regulatory network of M. *tuberculosis* that has been assembled to date, comprising approximately 20% of its genome. For estimating probabilities between regulators and metabolic genes, we use the gene expression data consisting of 437 whole-genome microarrays of *M. tuberculosis* H37Rv measuring the effects of 75 different drugs (29).

By using this model, we demonstrate the various capabilities of the PROM algorithm. Specifically, we show that PROM can make accurate predictions of cellular phenotype, identify functions of less studied regulatory genes, and discover alternate drug targets. We systematically knocked out all the TFs in *M. tuberculosis* that regulate metabolic genes and their phenotypes were predicted using PROM. We compared the prediction phenotypes with the results obtained by Sassetti et al. (30), Gao et al. (31), and Lamichhane et al. (32). The Sassetti et al. data set (30) identifies genes needed for optimal growth of *M. tuberculosis* strain H37Rv in vitro using transposon site hybridization mutagenesis. The data of Gao et al. (31) contain a set of genes that are consistently expressed under different growth conditions and the Lamichhane et al. data set (32) has genes that were found to be inessential or dispensable for growth using random mutagenesis. There was less than 30% overlap between the Gao et al. (31) and Sassetti et al. (30) data, suggesting that both studies were not comprehensive enough to identify all the essential genes. We use the term "essential" to denote the genes that are needed for optimal growth, as determined by these experimental studies. Also, we did not include the genes that were needed for in vivo growth as the growth conditions inside the human body (in vivo) would be very different from the ones simulated by our model.

Genome-Scale in Silico Phenotype Prediction in *M. tuberculosis*. PROM correctly predicted the phenotypes of 23 of the 24 TF KOs for which data were available and identified five of the six TFs that are required for optimal growth in the model (Table S1). The gene KdpE, which was mispredicted by PROM to be inessential, regulates the cell's turgor pressure, and its KO phenotype could not be predicted correctly by the model. As PROM predicts phenotypes based on a gene's effect on metabolism, it cannot determine correctly the phenotypes of genes with major nonmetabolic functions. As these genes were also found to regulate some metabolic genes by Balazsi et al. (27), they were included in the integrated model.

PROM also predicted six other TF KOs to be suboptimal (Table S1). These genes were not found to be essential for optimal growth by Gao et al. (30) and Sassetti et al. (31), yet, they were not found to be inessential by Lamichhane et al. (32) either. As all three data sets were not exhaustive and had little overlap, we considered these six genes to be candidate essential genes. The orthologues of some of the genes—ArgR, lrp/AsnC, SigH—were also found to be essential for growth in well studied organisms like *E. coli* and *Bacillus subtilis* using the Database of Essential Genes (33), which further supports our results. The overall accuracy of the model was 95%. Table 1 summarizes the model results in *M. tuberculosis*.

Despite the lack of complete biological knowledge about *M. tuberculosis*, PROM was still able to predict the phenotypes with relatively high accuracy. For comparison, a similar study (34) using the metabolic model alone predicted phenotypes of metabolic gene KOs with an accuracy of only 55%. So, the predictive power of PROM is highly significant, although clearly much more knowledge about the integrated metabolic-regulatory networks in *M. tuberculosis* is needed to expand the model to achieve the comprehensiveness observed in *E. coli*. A reason for the higher accuracy of the integrated model despite using the same metabolic network used in the previous study is the fact that most regulators in our analysis control pathways that are well studied and the metabolic model had high accuracy predicting those pathways.

Modeling Prior Knowledge. Microarrays are a powerful tool for identifying regulatory interactions, yet not all interactions can be inferred or quantified using them. Other techniques such as chip-chip or KO assays can detect additional interactions. As regulatory interactions used in our study were identified using a large set of expression arrays, chromatin immunoprecipitation, or a combination of these and other high-throughput methods (27, 35), some of the incorrect predictions by PROM in the genome-scale phenotype prediction analysis might be attributed to those interactions that could not be properly quantified using the available microarray data. For this reason, PROM allows the user to add interactions based on prior knowledge in to the model with defined probabilities. For example, the *M. tuberculosis* model could be simulated by assigning a user-defined probability of 0 or 1 for all interactions in the model that were based on strong literature evidence. This type of addition forces the model to set the genes to be completely on or off for those interactions with strong literature evidence and uses microarrays to determine the probability for the rest. The modified model was able to identify all essential genes except kdpE, and was almost as accurate as the original model in which all interactions were estimated using microarrays (Table S2). PROM thus allows users to add high confidence interactions into the model based on prior knowledge. Further, this approach can also be used for model refinement. We can add new interactions and check their validity with experiments. The discrepancies with actual data will help to further refine the model and generate new hypotheses. PROM thus provides the opportunity to not only integrate data from various sources but also the ability to test hypotheses.

Predicting Drug Targets Using PROM. TFs interact with several genes and are major hubs in cellular networks. By targeting these highly connected hubs, it is possible to perturb the entire cellular network and cause cell death. A drug that targets the gene Rv0485 would also affect its regulatory targets—glmU, glmS, and nagA, which were all found to be essential (30). Additional targets could be those TFs whose KOs were predicted to be suboptimal, like Rv3575c, Rv3291c, Rv1395, and argR. These genes, along with those that were found to be essential, could be targeted for treatment. Among the drug targets in the dataset identified based on essentiality in *M. tuberculosis*, PROM identified five of the six candidate TF drug targets (hypergeometric test, $P = 0.0004$). Similarly, it detected seven of the 10 candidate drug targets based on essentiality in *E. coli* or *M. tuberculosis* (hypergeometric $P = 0.01$). Overall, of the 11 predicted essential genes by PROM, seven were drug targets, which is significant ($P = 0.01$) in comparison with the background, in which only 10 of the 30 genes in the model were drug targets. Although we already know the essentiality of some of these genes through experimental evidence, PROM could theoretically be used to determine candidate essential genes missed by these methods and also for other pathogens with comparatively little experimental data.

PROM for Functional Annotation. PROM can also help determine the functions of less studied genes. Two genes—Rv0485 and Rv3575c—do not have significant functional annotation in the Tuberculist database (36). Their functions have been inferred based on sequence alignment as transcriptional regulators. Using PROM, we find that the protein Rv3575c likely regulates enzymes involved in nucleotide biosynthesis and glycine/serine metabolism. In silico KO of this regulator led to a 50% reduction in growth rate and to altered flux changes in amino acid and nucleotide metabolism. The protein Rv0485, based on the regulatory data, controls enzymes involved in membrane biosynthesis. It could also serve as a potential drug target, as most drugs target membrane biosynthesis in microbes (37). Perturbation to this protein caused a reduction in growth rate of 20%. Fig. S3 displays the metabolic networks regulated by Rv0485 and Rv3575c. Even though we can infer the predicted function of these genes through in vitro binding studies, we cannot be sure if these interactions really occur in vivo. Further, even if we do have conclusive evidence that a TF regulates an enzyme, alteration in the activity or expression of that enzyme might not have a significant effect on the cell's metabolism. The metabolic network is known to be robust to perturbations, and changes at the transcriptional level may not be translated to flux responses at the metabolic network level. For example, the regulator regX3 was predicted to regulate the phosphate transporter phoT. The probability quantified using microarray data for the phoT gene to be on when regX3 was off was found to be 0.53. Yet, at the metabolic level, this did not have any impact on the growth rate, suggesting that there is an alternate route for phosphate to enter the cell and perturbing the TF does not have any significant effect on metabolism. This could not have been inferred if we had used only the binding data for regX3, which could have suggested that the TF has a significant impact on phosphate metabolism. Thus, PROM adds value by being able to discriminate between expression changes that are likely to matter for metabolic phenotype and those that are not. This also highlights the fact that the flux response predicted by PROM is not correlated with gene expression or the probabilities directly.

Conclusion

In summary, we have created an approach for integrating transcriptional regulatory and metabolic networks called PROM. In this study, we use the PROM algorithm first to successfully integrate the metabolic and transcriptional regulatory networks of *E. coli*, demonstrating that our method shows improved accuracy over the current state of the art, and, most importantly, it greatly increases the comprehensiveness of the integrated regulatory-metabolic model through the automated integration of high-throughput data. The (mostly) automated reconstruction of integrated metabolic-regulatory networks is critically important to harnessing the exponentially increasing amount of genomics and transcriptomics data to form quantitative, testable models that elucidate the genotype-to-phenotype relationship. The PROM method can be extended easily to any organism in which a large number of microarray experiments have been run. As a proof of concept, we used it to construct a genome scale integrated regulatory-metabolic model for the critically important human pathogen *M. tuberculosis*. We demonstrated PROM's ability to model the effect of perturbations to transcriptional regulators and subsequently used it to predict microbial growth phenotypes under various environmental conditions. PROM may also identify putative functions of less annotated genes and help discover alternate targets for therapy.

Methods

For each TF–target pair, the probability of activation was calculated from RMA normalized microarray data. Regulatory interactions not involving metabolic genes were not included in our models, resulting in a final regulatory model for *M. tuberculosis* with 218 TF–gene interactions and an *E. coli* model with 1,773 TF–gene interactions. These interactions were then screened for false positives and condition-specific interactions using the Kolmogorov–Smirnov statistic (38) (*SI Methods*) and only significant interactions (i.e., $P < 0.05$) were retained. The interactions were quantified using several hundred gene expression arrays (904 for *E. coli* and 437 for *M. tuberculosis*) obtained from the Many Microbes Database (39) and Gene Expression Omnibus (40). The microarrays were chosen in such a way as to maximize the number of conditions under which gene expression is measured while reducing array platform-induced variations, resulting in accurate quantification of TF–target interactions (*SI Methods* and Fig. S8). The metabolic data are represented in silico in the form of a stoichiometric matrix, wherein every column corresponds to a reaction and every row corresponds to a metabolite. The stoichiometric matrix thus contains both chemical and topological information.

For determining on/off states of a gene, the data were binarized with respect to a fixed low value threshold for all genes similar to the method of Becker and Palsson (41). Gene expression values less than a threshold (Fig. S4) were considered to be off and the remaining values were set to on. We then estimate the number of times the target state is on or off for each TF state. For example, $P(target = ON|TF = OFF)$ would be as follows:

$$\frac{N(Target = 1TF = 0)}{N(TF = 0)} \qquad [1]$$

where N is the number of times the event is observed. Furthermore, for cases in which the probability of interaction cannot be estimated by using microarray data because of unavailability of expression, or if the gene was on or off in all of the conditions, we set the probability to a default value of 1. A value of 1 implies that the bound for the reaction would not be adjusted and the flux bound would be V_{max}, the same as an unregulated model, whereas if the probability is set to 0, the reaction is considered to be off. Fig. S1 shows the distribution of the probabilities in *E. coli* for target genes when their corresponding regulator was turned off. For turning a reaction off, the bounds for the reaction were set to a low value (10^{-6}). TF KOs were predicted to be lethal by using the method of Shlomi et al. (42). If the respective prediction of the mutated organism's maximal growth rate was less than 5% of the WT growth rate, it was considered as lethal. Changing the threshold did not alter the accuracy of the model significantly, suggesting that there is a clear demarcation between lethal and nonlethal KOs (Figs. S5 and S6). Any KO that resulted in a growth rate lower than the WT was considered to be suboptimal (Fig. S7). All simulations were performed in MATLAB. The COBRA toolbox (43) was used to load and optimize the metabolic model. The PROM algorithm is available for download at the authors' Web site (http://www.igb.uiuc.edu/labs/price/downloads).

ACKNOWLEDGMENTS. We thank Zan Schulten and Caroline Milne for critical readings of the manuscript, Aswin Kannan for sharing ideas on optimization, Matthew Gonnerman for assistance with computation, and James Eddy for help with making the figures. This work was supported by a National Institutes of Health/National Cancer Institute Howard Temin Pathway to Independence Award in Cancer Research, a Young Investigator Grant from the Roy J. Carver Charitable Trust, a National Science Foundation CAREER grant, and US Department of Defense Grant W81XWH-09-1-0107.

1. Terzer M, Maynard ND, Covert MW, Stelling J (2009) Genome-scale metabolic networks. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 1:285–297.
2. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat Rev Microbiol* 2:886–897.
3. Schlitt T, Brazma A (2007) Current approaches to gene regulatory network modelling. *BMC Bioinformatics* 8(suppl 6):S9.
4. Davidson EH, et al. (2002) A genomic regulatory network for development. *Science* 295:1669–1678.
5. Lee TI, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298:799–804.
6. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 31:64–68.
7. Price ND, Shmulevich I (2007) Biochemical and statistical network models for systems biology. *Curr Opin Biotechnol* 18:365–370.
8. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:92–96.
9. Herrgård MJ, Lee BS, Portnoy V, Palsson BO (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. *Genome Res* 16:627–635.
10. Covert MW, Schilling CH, Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* 213:73–88.
11. Price ND, Papin JA, Schilling CH, Palsson BO (2003) Genome-scale microbial in silico models: The constraints-based approach. *Trends Biotechnol* 21:162–169.
12. Shlomi T, Eisenberg Y, Sharan R, Ruppin E (2007) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* 3:101.
13. Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics* 24:2044–2050.
14. Kaleta C, Centler F, di Fenizio PS, Dittrich P (2008) Phenotype prediction in regulated metabolic networks. *BMC Syst Biol* 2:37.
15. Lee JM, Min Lee J, Gianchandani EP, Eddy JA, Papin JA (2008) Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLOS Comput Biol* 4:e1000086.
16. Goelzer A, et al. (2008) Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis. *BMC Syst Biol* 2:20.
17. Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic Boolean Networks: A rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 18:261–274.
18. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. *Curr Opin Biotechnol* 14:491–496.
19. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* 7:129–143.
20. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D (2007) How to infer gene networks from expression profiles. *Mol Syst Biol* 3:78.
21. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5:264–276.
22. Shlomi T, Cabili MN, Herrgård MJ, Palsson BO, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol* 26:1003–1010.
23. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186.
24. Spudich JL, Koshland DE, Jr. (1976) Non-genetic individuality: Chance in the single cell. *Nature* 262:467–471.
25. Glasner JD, et al. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res* 31:147–151.
26. Dye C (2006) Global epidemiology of tuberculosis. *Lancet* 367:938–940.
27. Balázsi G, Heath AP, Shi L, Gennaro ML (2008) The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. *Mol Syst Biol* 4:225.
28. Jacques PE, et al. (2005) MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. *Bioinformatics* 21:2563–2565.
29. Boshoff HI, et al. (2004) The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: Novel insights into drug mechanisms of action. *J Biol Chem* 279:40174–40184.
30. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 48:77–84.
31. Gao Q, et al. (2005) Gene expression diversity among Mycobacterium tuberculosis clinical isolates. *Microbiology* 151:5–14.
32. Lamichhane G, et al. (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. *Proc Natl Acad Sci USA* 100:7213–7218.
33. Zhang R, Lin Y (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res* 37(database issue):D455–D458.
34. Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol* 1:26.
35. Salgado H, et al. (2004) RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res* 32(Database issue):D303–D306.
36. Camus JC, Pryor MJ, Médigue C, Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology* 148:2967–2973.
37. Walsh C (2000) Molecular mechanisms that confer antibacterial drug resistance. *Nature* 406:775–781.
38. Young IT (1977) Proof without prejudice: Use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. *J Histochem Cytochem* 25:935–941.
39. Faith JJ, et al. (2008) Many Microbe Microarrays Database: Uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res* 36(Database issue):D866–D870.
40. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30:207–210.
41. Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. *PLOS Comput Biol* 4:e1000082.
42. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 102:7695–7700.
43. Becker SA, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* 2:727–738.