Genome Biology

METHOD Open Access

MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks

Edward Vitkin and Tomer Shlomi*

Abstract

Genome-scale metabolic network reconstructions are considered a key step in quantifying the genotype-phenotype relationship. We present a novel gap-filling approach, MetabolIc Reconstruction via functionAl GEnomics (MIRAGE), which identifies missing network reactions by integrating metabolic flux analysis and functional genomics data. MIRAGE's performance is demonstrated on the reconstruction of metabolic network models of *E. coli* and *Synechocystis* sp. and validated via existing networks for these species. Then, it is applied to reconstruct genome-scale metabolic network models for 36 sequenced cyanobacteria amenable for constraint-based modeling analysis and specifically for metabolic engineering. The reconstructed network models are supplied via standard SBML files.

Background

Genome-scale metabolic network reconstructions are considered a key step in quantifying the genotype-phenotype relationship [1]. While the process of manually reconstructing genome-scale metabolic network models is complex [2], such networks have already been manually reconstructed for more than 50 organisms [3], including common model microorganisms [4,5], industrially relevant microbes [6-9], various pathogens [10-13], and recently for human cellular metabolism [14]. A modeling approach called constraint-based modeling serves to analyze the function of such networks by solely relying on simple physical-chemical constraints [15,16] and is frequently used to predict various phenotypes of microorganisms (reviewed in [3,17-20]). Several applications of constraint-based modeling for metabolic engineering of microbial species has been presented (reviewed in [17,21]).

The two major computational challenges in metabolic network reconstruction are (i) the identification of missing reactions in a metabolic network, and (ii) the association of genes with network reactions. The identification of missing reactions, referred to as 'gap-filling', is commonly performed based on a pre-defined metabolic capability that the network is expected to be able to fulfill. For example, the capability to generate essential biomass products under various genetic and environmental conditions [22-25], the synthesis of specific compounds identified via metabolomics [26], the flow of specifically measured flux rates [27], and the ability to activate a core set of reactions under a steady-state assumption [28]. Missing reactions are obtained either from other species [22,23,26-28], or via computational chemistry methodologies that enumerate possible metabolic routes [29], aiming to identify a minimal number of missing reactions to fulfill the required objective. A specific approach for network reconstruction that is based on the concept of elementary flux modes [30] was previously applied to successfully recover missing network reactions [16,31,32]. Another gap-filling approach that integrates some of these principles has been recently used to reconstruct 130 genome-scale metabolic network models of various bacteria [22]. While most of the above gap-filling methods rely strictly on metabolic flux analysis and do not utilize functional genomics data to guide the search for missing reactions, computational methods that aim to address the second challenge of gene-reaction assignment do rely intensively on functional genomics data. Specifically, several methods predict gene assignment based on genomic data, utilizing principles such as conserved chromosomal proximity [31,33,34] and similarity in phylogenetic profiles with neighboring genes in the same

* Correspondence: tomersh@cs.technion.ac.il
Department of Computer Science, Technion - Israel Institute of Technology, Haifa 32000, Israel

pathway [32,35,36]. Others rely on an additional array of functional genomics data, including gene co-expression and protein-protein interactions [37-43].

Here, we present a novel approach, MetabolIc Reconstruction via functionAl GEnomics (MIRAGE), for reconstructing metabolic network models and specifically addressing the problem of gap-filling, by searching for missing reactions whose presence is supported by various functional genomic data. Specifically, to reconstruct a metabolic network model for an organism of interest, MIRAGE starts from a core set of reactions, whose presence is established via strong genomic evidence, and identifies missing reactions that are required to activate the latter core reactions (in addition to biomass requirement) by identifying additional reactions, whose presence is further supported by phylogenetic profiles and gene expression data. The performance of MIRAGE, in comparison to previous methods, is demonstrated on the reconstruction of network models for *Escherichia coli* and the cyanobacteria *Synechocystis* sp. PCC 6803, validated via existing networks for these species. Then, it is applied to reconstruct genome-scale metabolic network models for 36 sequenced cyanobacteria (supplied via standard Systems Biology Markup Language (SBML) files [44]), amenable for constraint-based modeling analysis and specifically for metabolic engineering. To demonstrate the utility of the reconstructed cyanobacteria networks, a strain design method was applied to predict gene knockouts whose implementation is expected to significantly elevate the production rate of an important nutritional product, astaxanthin.

Results and discussion

MIRAGE

MIRAGE is a functional genomics-based model reconstruction approach that aims to generate a genome-scale metabolic network model for an organism of interest, given a core set of reactions that are known to exist in its network, and optionally, a definition of a biomass reaction. The core set of reactions can be automatically derived strictly from genomic data, based on strong sequence similarity with known enzyme-coding genes in other species. The method then aims to find missing reactions (from a universal database of candidate gap-filling reactions such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)), supported by functional genomics data, whose addition to the network would lead to a functional model. The method follows a two-step procedure, starting with the utilization of functional genomics data to estimate the probability of including each reaction from the universal database in the reconstructed network, and then, metabolic flux analysis that selects the most likely set of reactions whose addition to the network would satisfy the above described objectives.

For the first step, we utilize two functional-genomics data sources to estimate the likelihood that a reaction from a universal reactions database should be included in the target metabolic network: (i) enzymes' phylogenetic profiles, and (ii) gene expression. Specifically, we define a weight for each reaction in the universal database (that is not already included in the reconstruction's core reactions set), based on the functional similarity between neighboring enzymes, in terms of resemblance of phylogenetic profiles, and correlation in gene expression of the enzyme-coding genes (Materials and methods).

Enzyme phylogenetic profiles were extracted from KEGG, representing a pattern of enzyme presence or absence across an available collection of species. For each reaction in KEGG, we computed a phylogenetic weight, representing the likelihood for its inclusion in the network reconstruction. Specifically, the phylogenetic weight of a certain reaction is calculated based on the maximal Jaccard coefficient between its phylogenetic profile and the corresponding profiles of its neighboring core reactions in the network (Materials and methods). Similarly, an expression weight for each reaction was calculated by evaluating gene expression profiles (measured in the target organism) of potential enzyme-coding genes (considering all non-annotated genes in the genome), compared with the expression profiles of known genes associated with neighboring core reactions. The sum of the phylogenetic and expression weights after proper normalization was used as input for the second reconstruction step (Materials and methods).

The second reconstruction step aims to find a set of high weight gap-filling reactions that satisfy the objectives described above. Towards this goal, we employed the following reaction pruning procedure. Starting from a metabolic network model consisting of all reactions in the universal reaction database, we iteratively remove potential gap-filling reactions, as long as the removal does not affect the consistency of the model. In each iteration, the probability of choosing a certain reaction for removal is inversely proportional to its weight - that is, low weight reactions have a higher probability to be chosen first for removal. The model consistency check procedure involves the usage of constraint-based modeling to verify that the remaining network (i) enables each core reaction to carry non-zero metabolic flux within a stoichiometrically balanced flux distribution, accounting for reaction directionality constraints, (ii) enables the production of all essential biomass constituents, and (iii) accounts for the growth-associated dilution of all network metabolites (that is, guaranteeing that the network consists of complete pathways for either the transport or *de novo* synthesis of all metabolites that exist in the network) [45]. Since the reactions' scanning order may affect the resulting model, the algorithm is executed several times with different, random pruning orders (Materials and methods). The fraction of obtained models that contains a certain reaction reflects the confidence that it should be included in the final model. Hence, to construct the final metabolic network model, we run the reactions removal procedure again, based on an ordering defined by the received confidence values (Materials and methods).

Notably, the presented method extends upon the Model Building Algorithm (MBA) of Jerby *et al.* [28] that was recently used to reconstruct a model of human liver metabolism. The MBA method addresses only the first objective from the above list, while not accounting for biomass production and growth-associated metabolite dilution, which are of less importance for the modeling of human tissue metabolism. Furthermore, it accounts for functional genomics data in a more limited manner, by using them only to define two core sets of reactions with either a moderate or high probability to be retained in a specific tissue model. In contrast, MIRAGE assigns a continuous score per each reaction that reflects its probability to be retained in a specific species model, allowing us to make better use of these data.

The described method is computationally demanding since each trial of the random reaction pruning procedure (out of the 500 trials performed to gather sufficient confidence statistics), requires eliminating each reaction from the universal reactions set in turn, and checking the consistency of the resulting model. Implementing the speedup heuristic suggested by Jerby *et al.* [28], which aims to minimize the number of linear optimizations required in each model consistency check, provided some improvement in running time. However, each random pruning trial still took around 35 hours, which made the entire method computationally intractable. The significant increase in running time in comparison to the method of Jerby *et al.* resulted from the markedly large size and complexity of the universal reaction database in comparison to the human network model used by Jerby *et al.*, and the additional reconstruction objectives previously not accounted for.

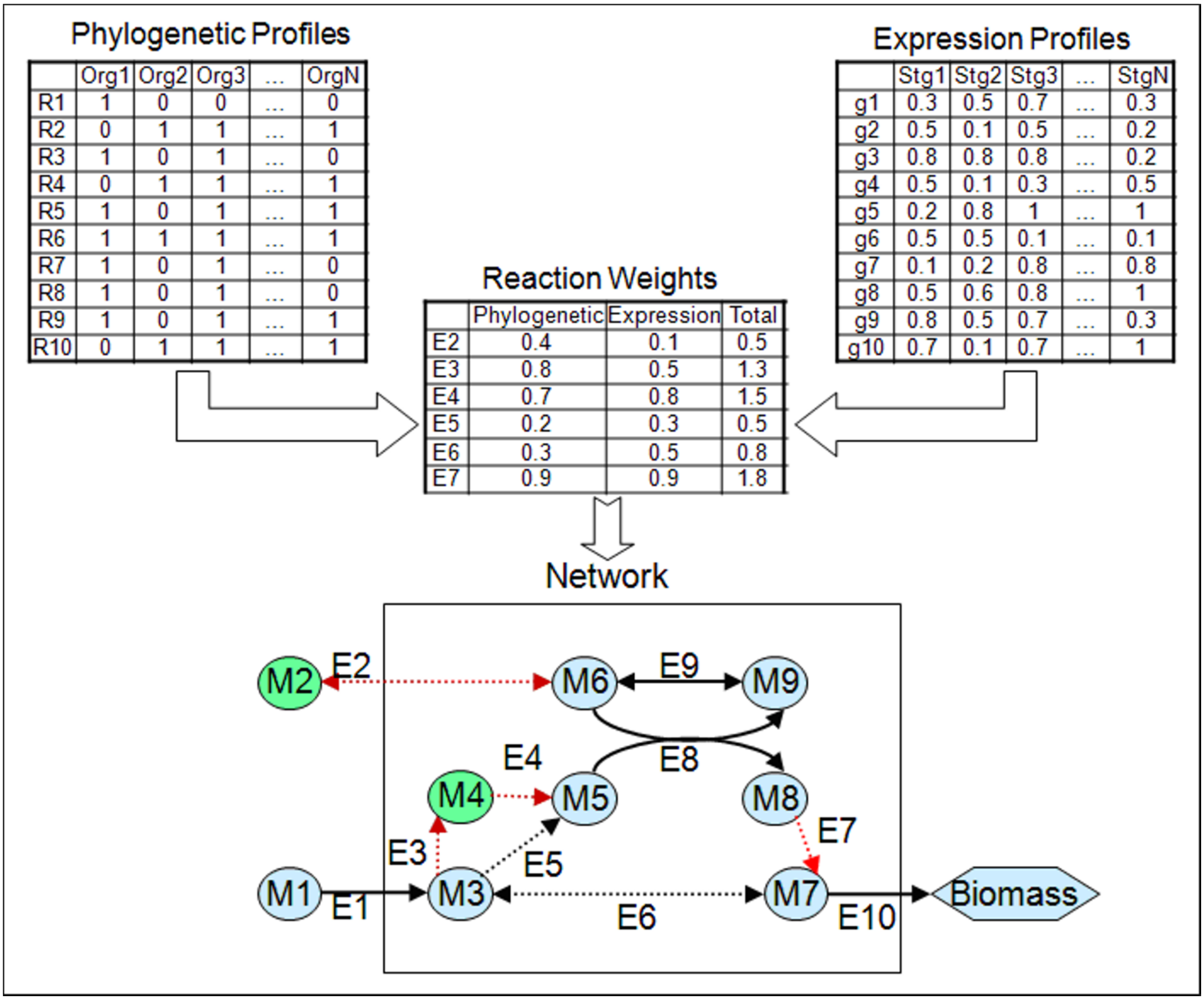
To overcome this, we implemented the following additional speed-up techniques (Materials and methods). First, the model consistency check procedure is based on identifying a set of flux distributions in which all core reactions are activated (that is, have non-zero flux), and is applied following the removal of each reaction in the reaction pruning procedure. The first speed-up involved the utilization of flux distributions computed in one call to the model consistency check procedure in subsequent calls to this procedure (testing the potential removal of subsequent reactions in the pruning order) to avoid time-consuming linear programming optimizations. Second, to further minimize the number of performed linear optimizations, the latter are now formulated with the objective of minimizing flux through subsequent gap-filling reactions in the pruning order. These two speed-up techniques, significantly elaborated upon in Additional file 1, provide a 100-fold improvement in running time.

Figure 1 illustrates the working of MIRAGE on a toy model. Reactions E1, E8, E9 and E10 are core reactions, while all the other reactions are candidates for gap-filling. MIRAGE predicts the addition of reactions E2, E3, E4 and E7 to enable flux activation of all core reactions, biomass production, and accounting for growth dilution of all metabolites in the core. The inclusion of reactions leading from M3 to M5 is required to enable flux activation of core reactions E8 and E9. In this case, the choice of including both reactions E3 and E4 for gap-filling, instead of the single reaction E5, is based on higher support for the former reactions in the functional-genomic data. Reaction E2 is predicted for gap-filling to compensate for growth-associated dilution of metabolites M6 and M9 [45].

Validation of MIRAGE in the reconstruction of a metabolic network for *E. coli*

To evaluate the performance of MIRAGE, we applied it to reconstruct a metabolic network model for *E. coli*, for which a comprehensively curated model (iAF1260) is already available for validation [46]. Towards this end, we extracted a cross-species reactions dataset from KEGG having 7,211 reactions (referred to as the universal reactions set). To define a core set of known *E. coli* reactions to be used by MIRAGE, we considered KEGG reactions annotated as existing in *E. coli* and also belonging to iAF1260, plus the known biomass and all exchange reactions from iAF1260. Then we removed dead-end reactions that cannot be activated within a feasible flux distribution when considering the entire universal reactions set, yielding a core set of 812 reactions. Performing standard flux variability analysis [47] when focusing only on this set of 812 core reactions revealed that 45% (365/812) of these reactions are on dead-ends. MIRAGE's task is hence to identify gap-filling reactions that would resolve these dead-ends, aiming to identify a remaining set of 109 reactions from iAF1260. Notably, our analysis did not account for subcellular localization of metabolic processes, and hence duplicated reactions in iAF1260 that correspond to multiple compartments were removed.

Comparison of MIRAGE's reconstructed network model for *E. coli* with iAF1260 shows a predictive precision of 41.9% and recall of 24.3%, which is significantly better than random sampling of gap-filling reactions (hyper-geometric *P*-value < 10^{-16}; Figure 2; Additional file 1, part 6, and Supp. Table 1 in Additional file 1). As controls, we assessed the predictive performance of using



Phylogenetic Profiles

	Org1	Org2	Org3	...	OrgN
R1	1	0	0	...	0
R2	0	1	1	...	1
R3	1	0	1	...	0
R4	0	1	1	...	1
R5	1	0	1	...	1
R6	1	1	1	...	1
R7	1	0	1	...	0
R8	1	0	1	...	0
R9	1	0	1	...	1
R10	0	1	1	...	1

Expression Profiles

	Stg1	Stg2	Stg3	...	StgN
g1	0.3	0.5	0.7	...	0.3
g2	0.5	0.1	0.5	...	0.2
g3	0.8	0.8	0.8	...	0.2
g4	0.5	0.1	0.3	...	0.5
g5	0.2	0.8	1	...	1
g6	0.5	0.5	0.1	...	0.1
g7	0.1	0.2	0.8	...	0.8
g8	0.5	0.6	0.8	...	1
g9	0.8	0.5	0.7	...	0.3
g10	0.7	0.1	0.7	...	1

Reaction Weights

	Phylogenetic	Expression	Total
E2	0.4	0.1	0.5
E3	0.8	0.5	1.3
E4	0.7	0.8	1.5
E5	0.2	0.3	0.5
E6	0.3	0.5	0.8
E7	0.9	0.9	1.8

Figure 1 The application of MIRAGE on a toy model. Core reactions (E1, E8, E9 and E10) are marked with straight lines, while gap-filling reactions are marked with dashed lines. A weight for each reaction is computed based on the correlation of its phylogenetic and expression profiles with those of neighboring core reactions in the network. Reactions predicted for gap-filling by MIRAGE are in red. Specifically, E2 is chosen to enable the growth-associated dilution of metabolites M6 and M9. E3 and E4 are chosen (instead of E5, which has a significantly lower weight) to enable the flux activation of E8 and E9. E7 is chosen to enable flux activation of E8 and E9 under steady-state. Reaction E6 is not chosen for gap-filling as it is redundant given the above-mentioned chosen essential reactions.

only the functional genomics data based on the computed reaction weights (by ordering potential gap-filling reactions based on their computed weights), and the predictive performance of MIRAGE without utilizing functional genomics data (by assigning reactions with random weights; as done in the MBA algorithm). Using only the functional genomics data, the resulting predictive performance was significantly lower than that of MIRAGE (Figure 2), reaching a precision of 6.1%, under a recall level of 19.6% (P-value = 2×10^{-9}). Without utilizing functional genomics data, the predictive performance was also markedly lower, with a precision of 27.5% and recall of 20.6% (P-value < 10^{-16}). Using only gene expression [48] or phylogenetic weights (based on all species in KEGG) provided lower precision of 31.8% and 36.9%, respectively, with slightly lower recall levels (19.6% and 22.4%, respectively) to those achieved when utilizing both (Figure 2), demonstrating the importance of integrating multiple functional-genomics data sources. As a further control, we applied MIRAGE to reconstruct a metabolic network model for *E. coli*, without prior knowledge of exchange reactions (which in the above analysis were taken from the model of iAF1260), finding an overall similar predictive performance, showing an improvement of MIRAGE compared to other approaches (Supp. Table 2 in Additional file 1).

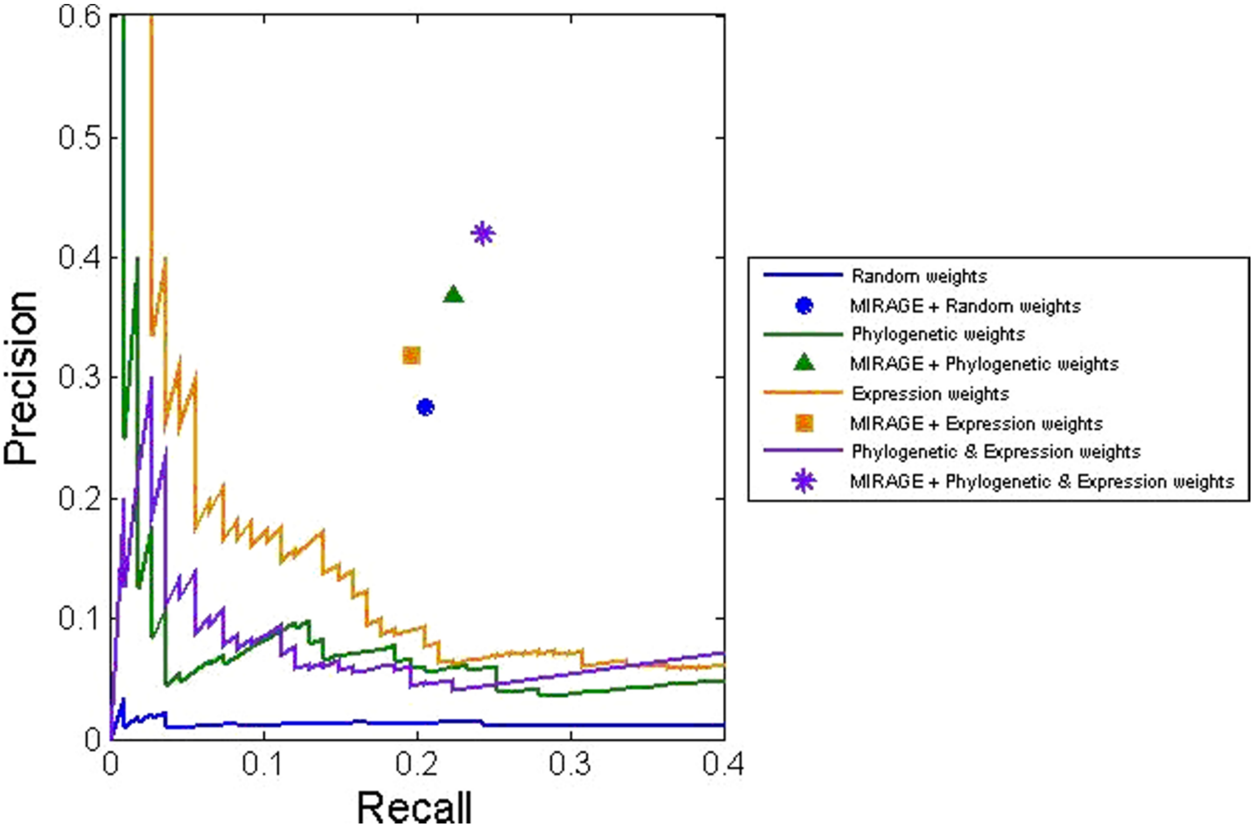


Figure 2 MIRAGE's predictive performance on reconstructing a known metabolic network of *E. coli*. The precision and recall of MIRAGE is marked with a star symbol. The precision and recall of several controls, including variants of MIRAGE that utilize only phylogenetic data, only expression data, or no functional-genomics data, are marked with a triangle, bar, and circle, respectively. The predictive performance of the functional genomic data (that is, by ordering potential gap-filling reactions based on their computed functional genomic weights, without utilizing metabolic flux analysis) is shown by the straight lines: the performance of the phylogenetic data, gene expression, and both data sources are colored green, yellow, and purple, respectively. The performance of random predictions of gap-filling reactions is colored blue.

Comparing the predictive performance of MIRAGE on reconstructing the metabolic network of *E. coli* with that of Model SEED [22] has shown a marked advantage to the former. While the number of core reactions considered by MIRAGE and the SEED algorithm in the reconstruction of a metabolic network model of *E. coli* is close (812 and 826 reactions for MIRAGE and SEED, respectively), the number of predicted gap-filling reactions by MIRAGE was 62, in comparison to only 10 by SEED. This results from MIRAGE's aim to resolve all gap-filling problems instead of just enabling biomass production as performed by SEED. The precision of MIRAGE's predictions was significantly higher than that of SEED, reaching 41.9% for MIRAGE versus 10% for SEED. Re-running MIRAGE given the very same definition of a biomass reaction used in the SEED reconstruction of *E. coli*'s model (rather than the biomass definition taken from iAF1260) still resulted in a higher number of 76 predicted gap-filling reactions, with a significantly higher precision of 34.2% than that achieved by SEED.

Applying MIRAGE to reconstruct metabolic network models for cyanobacteria

To demonstrate the utility of MIRAGE, we applied it to reconstruct genome-scale metabolic network models for 36 cyanobacteria for which genomic data are available to define core reactions sets. Our analysis spans all cyanobacteria for which enzyme annotations are available in KEGG, including *Synechocystis*, *Synechococcus*, *Cyanobacteria*, *Prochlorococcus*, *Anabaena*, and so on [49]. For all species, we considered the same biomass function, obtained from a previously reconstructed model of *Synechocystis* sp. PCC 6803 [50], assuming that CO_2 is the sole carbon source. Due to lack of comprehensive gene expression for most cyanobacteria species, we utilized here only phylogenetic data (considering all species in KEGG) to define reaction weights.

The average size of a core reactions set for a cyanobacteria network is 570 reactions (Figure 3a), out of which, 331 reactions belong to all of the 36 network cores (Figure 3b). The high degree of similarity between the reaction cores of the various cyanobacteria species reflects the current knowledge on common metabolic processes across these species, obtained mostly from sequence comparisons. These shared core reactions belong to highly conserved metabolic pathways, such as glycolysis, gluconeogenesis, and the TCA cycle among others. MIRAGE's predictions extend these networks in a species-specific manner, with many reactions predicted to belong to a small number of species (Figure 3b). These species-specific reactions belong to more peripheral pathways, for example, diterpenoid biosynthesis, fluorene degradation and others.

To evaluate the performance of MIRAGE in reconstructing cyanobacteria models, we compared a reconstructed network model for *Synechocystis* sp. PCC 6803 with the manually curated models of Knoop *et al.* [50] and iSyn811 [51,52]. In this case, MIRAGE was applied to reconstruct a *Synechocystis* model by further utilizing gene expression data obtained from Tu *et al.* [53] as part of the reconstruction process (Materials and methods). The comparison shows a predictive precision of 70% and recall of 24.6% for the Knoop *et al.* model [50] and precision of 37.5% and recall of 45% for iSyn811 [51,52]. These results are significantly better than random sampling (hypergeometric *P*-values are 2.99×10^{-27} and 3.59×10^{-31} for Knoop *et al.*'s model and iSyn811, respectively). Again, we find that the predictive performance of either the functional genomics data or the flux analysis alone is far worse (Figure 4). A comparison with the predictive performance of Model SEED was not possible in this case, as the SEED algorithm was not applied to reconstruct cyanobacteria models (focusing only on well-studied and annotated genomes). As a further evaluation criterion, we performed a BLAST [54] search of the known enzyme sequences catalyzing the predicted gap-filling reactions in other species against the genomes of the corresponding cyanobacteria. Reassuringly, we found that the resulting BLAST *E*-scores show significantly higher sequence similarity for the set of predicted reactions in comparison to a random set of reactions (*t*-test of 1.04×10^{-74}). Moreover, 20.3% of predicted reactions showed E-values below 10^{-100}, compared to 9.7% of randomly sampled reactions, testifying the overall correctness of the predicted set of reactions.

As a further evaluation of our reconstructed *Synechocystis* model, we applied it to predict gene knockout lethality data provided by [50]. We find that the prediction performance of our model is comparable with that of Knoop *et al.* (Supp. Table 3 in Additional file 1): out of 39 genes known to be non-essential, Knoop *et al.* correctly predicted 35, while our model correctly predicts 38. Out of 11 known essential genes, Knoop *et al.* correctly predicted 7, while our model correctly predicts 6. The fact that our automatically generated model reaches a similar level of prediction performance to that of a manually curated model demonstrates the applicability and importance of our model reconstruction approach.

Utilizing the reconstructed cyanobacteria networks for metabolic engineering

To demonstrate the applicability of the reconstructed cyanobacteria networks, we applied a computational metabolic engineering approach called Optknock [55] on these networks to rationally design genetic modifications that would increase the production of astaxanthin, which is a powerful antioxidant belonging to the carotenoid family. These metabolites are known to be produced by various cyanobacteria [56,57]. Optknock works by searching for gene knockouts that would couple the maximal production and secretion of a molecule of interest with a naturally selected trait of maximizing growth rate. Notably, 24 of the original core networks extracted from KEGG include an astaxanthin production reaction, though only 8 of these are not dead-end. In contrast, 25 of the network models reconstructed by MIRAGE have a functional astaxanthin production pathway, amenable for Optknock analysis.

The application of Optknock for astaxanthin production identified double gene knockouts in 15 species that are expected to lead to astaxanthin secretion (Table 1). For 12 out of these, Optknock predicts the knockout of dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase (EC: 2.5.1.1), which consumes an essential precursor for astaxanthin biosynthesis (1-hydroxy-2-methyl-2-butenyl4-diphosphate). The maximal achievable astaxanthin production rate reaches 2.18 μmol gDW^{-1} h^{-1} in Prochlorococcus marinus 9601, representing a carbon utilization of 40% for astaxanthin production (considering a CO_2 uptake rate of 0.22 mmol gDW^{-1} h^{-1} [50]). This utilization of CO_2 to produce astaxanthin is predicted to reduce growth rate by 57% relative to the wild-type Prochlorococcus strain (Table 1).

Conclusions

Our paper presents a novel method, MIRAGE, for reconstructing metabolic network models by integrating metabolic flux analysis and functional genomics data to resolve network gaps. MIRAGE was validated based on a comparison of its predictions with manually curated metabolic networks for *E. coli* [46] and *Synechocystis* sp. PCC 6803 [50-52]. Then it was applied to reconstruct metabolic network models for an ensemble of cyanobacteria, with the resulting networks shown to be amenable for metabolic engineering applications of astaxanthin secretion.

Our results show that functional genomics data enable the marked improvement of gap-filling in metabolic networks. Furthermore, we show that the integration of more than one type of functional genomics data can further improve the performance of MIRAGE. Naturally,

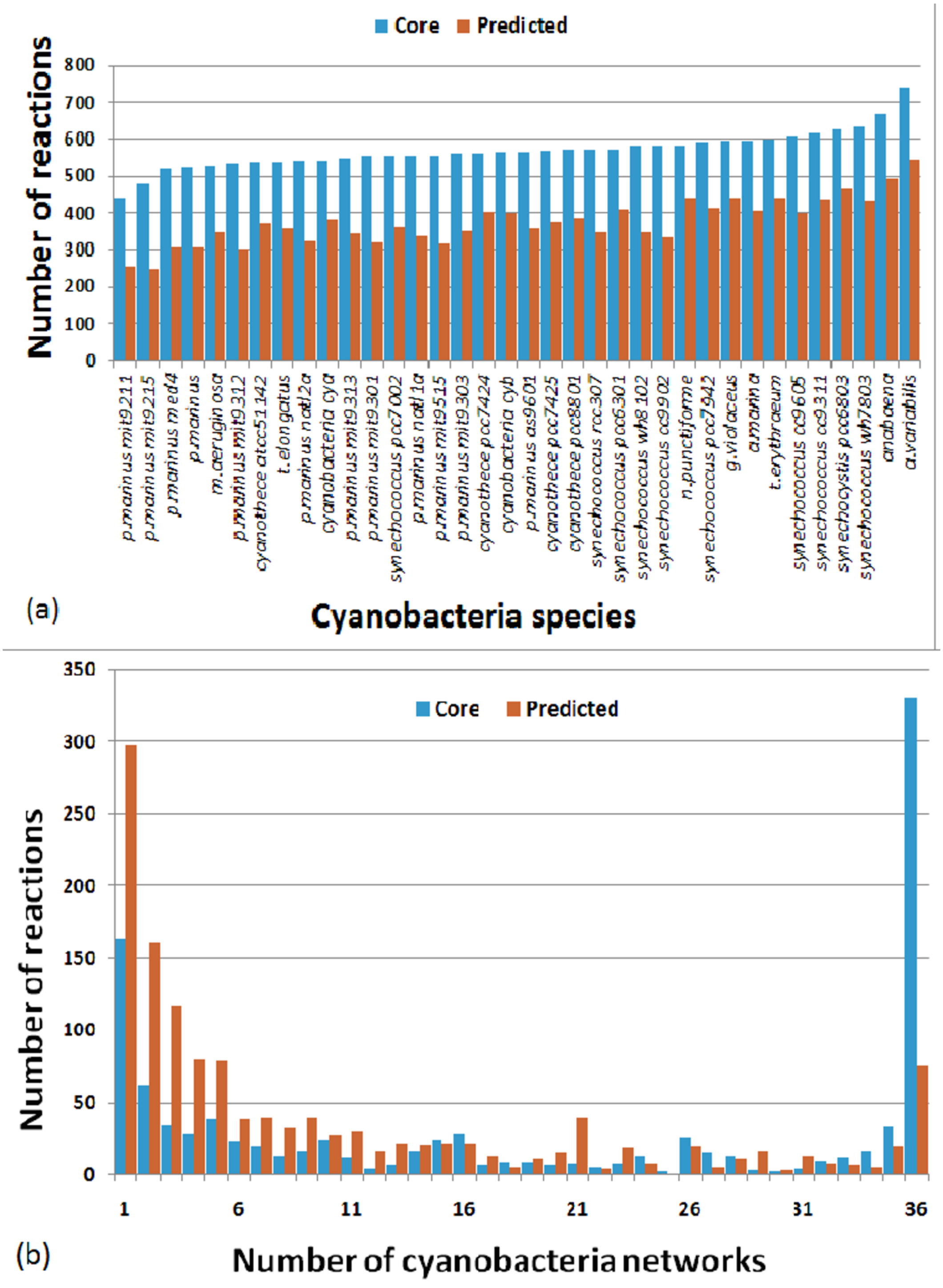


Figure 3 Statistics on MIRAGE's reconstructed cyanobacteria metabolic networks. **(a)** The number of core reactions (blue) and predicted gap-filling reactions (red) in the various reconstructed cyanobacteria models. **(b)** A histogram of core reactions (blue) and predicted gap-filling reactions (red) that participate in different numbers of reconstructed cyanobacteria models. As shown, cyanobacteria network cores consist of many reactions that are known to exist in all 36 species, while many of the predicted gap-filling reactions are species-specific.

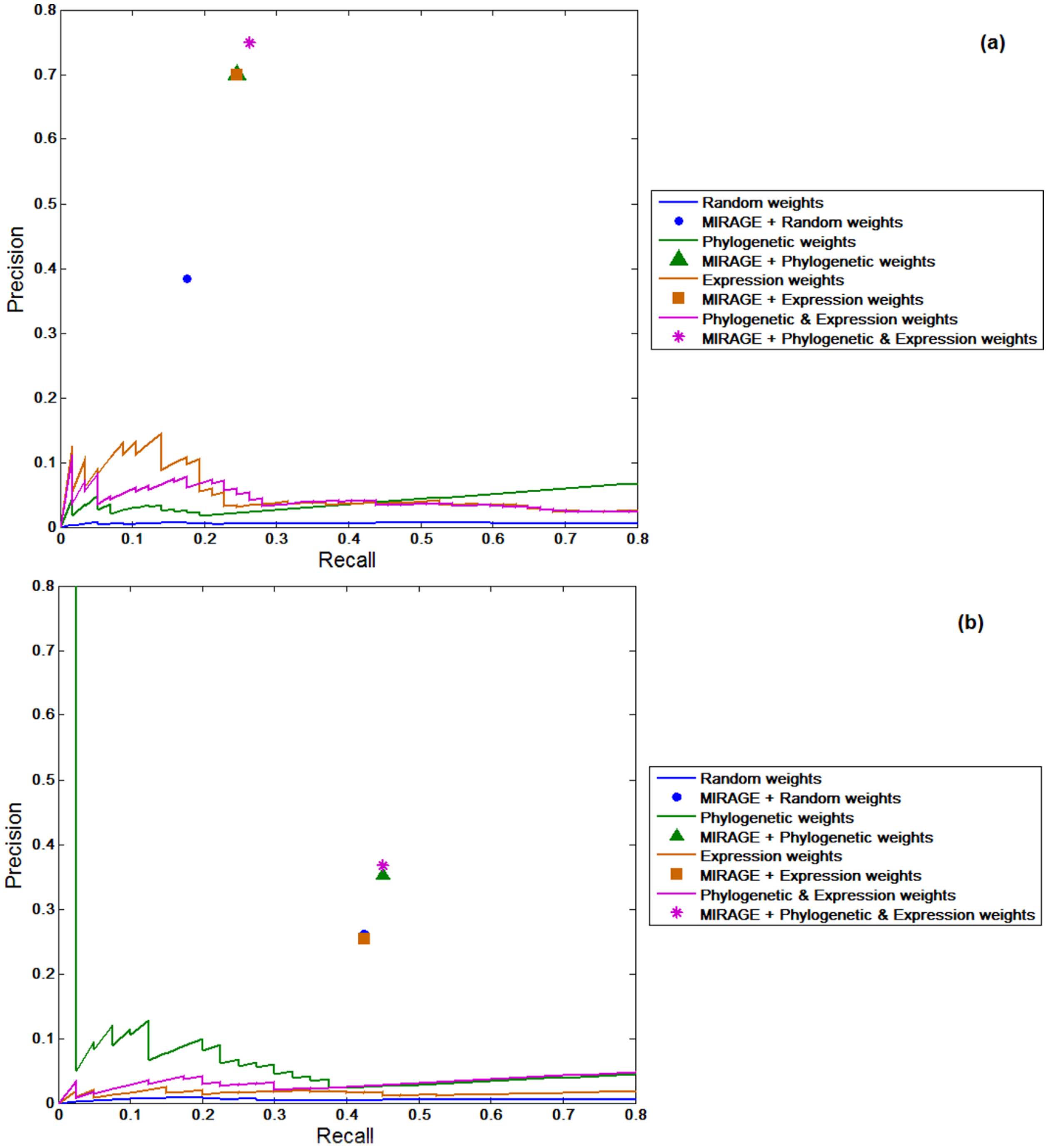


Figure 4 MIRAGE's predictive performance on reconstructing a known metabolic network of the cyanobacteria *Synechocystis* sp. PCC6803. **(a)** Metabolic network after Knoop *et al.* [50]; **(b)** metabolic network after Montagud *et al.* [51,52]. The precision and recall of MIRAGE is marked with a star symbol. The precision and recall of a variant of MIRAGE that does not utilize functional-genomics data is marked with a circle. The predictive performance of the functional genomic data (without metabolic flux analysis) is shown by the purple line. The performance of random predictions of gap-filling reactions is colored blue.

MIRAGE can be extended to account for additional functional genomics data, including protein-protein interactions and genomic context data, which were previously used for the identification of missing gene annotations in metabolic networks [38]. Metabolomics data can also be integrated within MIRAGE, to enable the definition of a metabolite core, consisting of metabolites that are known to be synthesized, and hence corresponding pathways

Table 1 Predicted knockout strategies for Astaxanthin over-production in cyanobacteria

Organism	Growth rate (mutant/wt) [h^{-1}]	Astaxanthin production [μmol gDW^{-1} h^{-1}] (min/max)	Knockout reaction names	KEGG reaction ID/EC number
P. marinus as9601	0.026/0.061	2.18/2.18	Dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase	R01658/EC 2.5.1.1
			5-O-(1-Carboxyvinyl)-3-phosphoshikimate phosphate-lyase	R01714/EC 4.2.3.5
T. elongatus	0.016/0.048	2.08/2.08	Dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase	R01658/EC 2.5.1.1
			Phosphoenolpyruvate: D-erythrose-4-phosphate C-(1-carboxyvinyl)transferase	R01826/EC 2.5.1.54
G. violaceus	0.033/0.061	1.64/1.64	Succinate:(acceptor) oxidoreductase	R00408/EC 1.3.99.1
			Dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase	R01658/EC 2.5.1.1
Synechococcus pcc7942	0.033/0.061	1.64/1.64	Succinate:(acceptor) oxidoreductase	R00408/EC 1.3.99.1
			Dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase	R01658/EC 2.5.1.1
Cyanobacteria cyb	0.032/0.058	1.57/1.57	Succinate:(acceptor) oxidoreductase	R00408/EC 1.3.99.1
			Dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase	R01658/EC 2.5.1.1
T. erythraeum	0.051/0.061	0.24/0.24	Hydrogen-carbonate: L-glutamineamido-ligase	R00575/EC 6.3.5.5
			Dimethylallyl-diphosphate: isopentenyl-diphosphate dimethylallyl transtransferase	R01658/EC 2.5.1.1
A. variabilis	0.051/0.061	0/0.96	4-Methyl-2-oxopentanoate: NAD+ oxidoreductase	R01651/EC 1.2.1.25
			1-Deoxy-D-xylulose-5-phosphate pyruvate-lyase	R05636/EC 2.2.1.7
Anabaena	0.051/0.061	0/0.96	3-(4-Methylpent-3-en-1-yl)-pent-2-enedioyl-CoA hydrolyase	R03493/EC 4.2.1.57
			1-Deoxy-D-xylulose-5-phosphate pyruvate-lyase	R05636/EC 2.2.1.7

For each species the table shows the predicted reactions whose knockout is expected to provide maximal astaxanthin production rate, the expected astaxanthin production rate, and the expected decline in growth rate.

that connect them to the rest of the network must be identified [26].

Several existing gap-filling methods work by searching for a minimal set of missing reactions that would enable the network to perform a certain task [22]. MIRAGE extends upon these methods by enabling the identification of pathways that are not necessarily minimal in size, if supported by functional genomics data. However, MIRAGE is still limited in being unable to predict the presence of alternative pathways, in case either one is sufficient to fulfill its defined objectives. This may explain the relatively low recall levels achieved by MIRAGE and the other tested approaches. For example, this was demonstrated in Figure 1, where reaction R6 will not be predicted for gap-filling, as an alternative pathway that fulfills the required metabolic objectives was chosen. The identification of alternative pathways based on more complex integration of functional genomics data with metabolic flux analysis is currently an open challenge for all known gap-filling algorithms. An additional limitation of MIRAGE is that it does not explicitly account for thermodynamic considerations as part of the network reconstruction process. Future implementation of this approach may formulate additional thermodynamic constraints as part of the model consistency check, as suggested in Thermodynamic Metabolic Flux Analysis (TMFA) [58] (which would require further speedups to obtain reasonable running times).

Metabolic models generated by automated methods such as MIRAGE should be regarded as first draft models, requiring further manual curation to bring them up to comparable level with standard manually curated models. The growing interest in reconstructing metabolic network models for hundreds of species raises the challenge of developing improved such gap-filling approaches that could speed up the reconstruction process, while the approach presented here shows a marked improvement in this direction over the state-of-the-art, supporting the advantage of integrating functional genomic data as part of model reconstruction. We expect MIRAGE to be used for automatic reconstructions of many other species, leading to a significant boost in the understanding of their metabolism.

Materials and methods

Step I: calculation of functional genomics weights

Binary vectors describing reaction phylogenetic profiles were acquired from KEGG. A phylogenetic weight for each non-core reaction is defined as the maximal Jaccard similarity with a phylogenetic profile of a core reaction that shares a metabolite substrate with the reaction at hand. The Jaccard values are normalized based on the frequency of appearance of the shared metabolite in the universal reaction database (see Additional file 1 for details).

An expression weight for a given non-core reaction is computed by evaluating the correlation between profiles of genes that may potentially code for an enzyme catalyzing the reaction at hand and expression profiles of genes associated with neighboring reactions. Specifically, for each gene in the genome of the target species, we compute the average Pearson correlation between its expression profile and profiles of genes associated with neighboring reactions, with the expression weight defined as the maximal such correlation obtained. All Pearson correlations are normalized by the frequency of appearance of the connecting metabolites (as done above).

The distribution of phylogenetic and expression weights are normalized to having the same mean and standard deviation. Final edge weights are defined based on the sum of normalized phylogenetic and expression weights. Reactions for which either weight is missing are assigned the median normalized value.

Step II: finding gap-filling reactions supported by functional genomics weights

First, we create a random reaction pruning list by iteratively sampling the next reaction with probability proportional to its weight normalized by the sum of weights of the remaining non-sampled reactions. Next, we scan through the obtained reaction list and try to remove each reaction in turn from the model, as long as the resulting model remains consistent. The consistency check involves verifying that: (i) each core reaction can carry non-zero flux under steady-state and reaction directionality constraints; (ii) there can be non-zero flux through the biomass reaction; and (iii) the growth-dilution of each metabolite in the network is accounted for. Once we finish scanning through the pruning list, we are left with a minimal functional model. We repeat the random pruning procedure 500 times and count the number of times that each non-core reaction appeared in the final model. Finally, we order the non-core reactions based on their frequencies (from low to high) and repeat the pruning step to obtain the final model.

The details of the above and the implementation of the various speed-up techniques that makes this algorithm computationally tractable are described in Additional file 1. The implementation of MIRAGE is available at [44].

Abbreviations

KEGG, Kyoto Encyclopedia of Genes and Genomes; MBA, Model Building Algorithm; MIRAGE, MetabolIc Reconstruction via functionAl Genomics; SBML, Systems Biology Markup Language.

Authors' contributions

EV and TS conceived the research and wrote the paper. EV performed the computational analysis. Both authors read and approved the final manuscript.

Acknowledgements
TS was supported by grants from the Israel Science Foundation and Israel Ministry of Science.

Received: 9 March 2012 Revised: 25 September 2012
Accepted: 29 November 2012 Published: 29 November 2012

References

1. Frazier ME, Johnson GM, Thomassen DG, Oliver CE, Patrinos A: **Realizing the potential of the genome revolution: the genomes to life program.** *Science* 2003, **300**:290-293.
2. Thiele I, Palsson BO: **A protocol for generating a high-quality genome-scale metabolic reconstruction.** *Nat Protoc* 2010, **5**:93-121.
3. Oberhardt MA, Palsson BO, Papin JA: **Applications of genome-scale metabolic reconstructions.** *Mol Syst Biol* 2009, **5**:320.
4. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: **A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Mol Syst Biol* 2007, **3**:121.
5. Mo M, Palsson B, Herrgard M: **Connecting extracellular metabolomic measurements to intracellular flux states in yeast.** *BMC Systems Biol* 2009, **3**:37.
6. Durot M, Le Fevre F, de Berardinis V, Kreimeyer A, Vallenet D, Combe C, Smidtas S, Salanoubat M, Weissenbach J, Schachter V: **Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data.** *BMC Systems Biol* 2008, **2**:85.
7. Ryan SS, Eleftherios TP: **Genome-scale model for *Clostridium acetobutylicum*: Part I. Metabolic network resolution and analysis.** *Biotechnol Bioeng* 2008, **101**:1036-1052.
8. Izallalen M, Mahadevan R, Burgard A, Postier B, Didonato R, Sun J, Schilling CH, Lovley DR: ***Geobacter sulfurreducens* strain engineered for increased rates of respiration.** *Metab Eng* 2008, **10**:267-275.
9. Mahadevan R, Bond DR, Butler JE, Esteve-Nunez A, Coppi MV, Palsson BO, Schilling CH, Lovley DR: **Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling.** *Appl Environ Microbiol* 2006, **72**:1558-1568.
10. Kjeld Raunkj⊠r K, Jens N: **In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network.** *Biotechnol Bioeng* 2009, **102**:583-597.
11. Jamshidi N, Palsson B: **Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets.** *BMC Systems Biol* 2007, **1**:26.
12. Schilling C, Covert M, Famili I, Church G, Edwards J, Palsson B: **Genome-scale metabolic model of Helicobacter pylori 26695.** *J Bacteriol* 2002, **184**:4582-4593.
13. Becker S, Palsson B: **Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation.** *BMC Microbiol* 2005, **5**:8.
14. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *Proc Natl Acad Sci USA* 2007, **104**:1777-1782.
15. Price ND, Papin JA, Schilling CH, Palsson BO: **Genome-scale microbial in silico models: the constraints-based approach.** *Trends Biotechnol* 2003, **21**:162-169.
16. Ruppin E, Papin JA, de Figueiredo LF, Schuster S: **Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks.** *Curr Opin Biotechnol* 2010, **21**:502-510.
17. Price ND, Reed JL, Palsson BO: **Genome-scale models of microbial cells: evaluating the consequences of constraints.** *Nat Rev Microbiol* 2004, **2**:886-897.
18. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO: **Reconstruction of biochemical networks in microorganisms.** *Nat Rev Micro* 2009, **7**:129-143.
19. Schuster S, Pfeiffer T, Fell DA: **Is maximization of molar yield in metabolic networks favoured by evolution?.** *J Theor Biol* 2008, **252**:497-504.

20. Papp B, Teusink B, Notebaart RA: A critical view of metabolic network adaptations. *HFSP J* 2009, **3**:24-35.
21. Feist AM, Palsson BO: The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol* 2008, **26**:659-667.
22. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL: High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 2010, **28**:977-982.
23. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO: Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* 2006, **103**:17480-17484.
24. Kumar VS, Maranas CD: GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput Biol* 2009, **5**: e1000308.
25. Satish Kumar V, Dasika MS, Maranas CD: Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 2007, **8**:212.
26. Christian N, May P, Kempa S, Handorf T, Ebenhoh O: An integrative approach towards completing genome-scale metabolic networks. *Mol Biosyst* 2009, **5**:1889-1903.
27. Herrgard MJ, Fong SS, Palsson BO: Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* 2006, **2**:e72.
28. Jerby L, Shlomi T, Ruppin E: Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol Syst Biol* 2010, **6**:401.
29. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ: Exploring the diversity of complex metabolic networks. *Bioinformatics* 2005, **21**:1603-1609.
30. Schuster S, Fell DA, Dandekar T: A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* 2000, **18**:326-332.
31. Lee JM, Sonnhammer EL: Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* 2003, **13**:875-882.
32. Green ML, Karp PD: Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics* 2007, **23**:i205-211.
33. Dandekar T, Snel B, Huynen M, Bork P: Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998, **23**:324-328.
34. Pachkov M, Dandekar T, Korbel J, Bork P, Schuster S: Use of pathway analysis and genome context methods for functional genomics of Mycoplasma pneumoniae nucleotide metabolism. *Gene* 2007, **396**:215-225.
35. Green ML, Karp PD: A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 2004, **5**:76.
36. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.
37. Kharchenko P, Vitkup D, Church GM: Filling gaps in a metabolic network using expression information. *Bioinformatics* 2004, , **20 Suppl 1**: i178-185.
38. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM: Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* 2006, **7**:177.
39. Chen L, Vitkup D: Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol* 2006, **7**:R17.
40. Overbeek R, Disz T, Stevens R: The SEED: a peer-to-peer environment for genome annotation. *Commun ACM* 2004, **47**:46-51.
41. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Krause L, Kubal M, Larsen N, Linke B, Mchardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, *et al*: The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 2005, **33**:5691-5702.
42. Hsiao TL, Revelles O, Chen L, Sauer U, Vitkup D: Automatic policing of biochemical annotations using genomic correlations. *Nat Chem Biol* 2010, **6**:34-40.
43. Breitling R, Vitkup D, Barrett MP: New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol* 2008, **6**:156-161.
44. Tomer Shlomi's Research Group. [http://www.cs.technion.ac.il/~tomersh/tools].
45. Benyamini T, Folger O, Ruppin E, Shlomi T: Flux balance analysis accounting for metabolite dilution. *Genome Biol* 2010, **11**:R43.
46. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO: A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007, **3**:121.
47. Gudmundsson S, Thiele I: Computationally efficient flux variability analysis. *BMC Bioinformatics* 2010, **11**:489.
48. Traxler MF, Chang DE, Conway T: Guanosine 3',5'-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in Escherichia coli. *Proc Natl Acad Sci USA* 2006, **103**:2374-2379.
49. Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000, **28**:27-30.
50. Knoop H, Zilliges Y, Lockau W, Steuer R: The metabolic network of Synechocystis sp. PCC 6803: systemic properties of autotrophic growth. *Plant Physiol* 2010, **154**:410-422.
51. Montagud A, Navarro E, Fernandez de Cordoba P, Urchueguia JF, Patil KR: Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst Biol* 2010, **4**:156.
52. Montagud A, Zelezniak A, Navarro E, de Cordoba PF, Urchueguia JF, Patil KR: Flux coupling and transcriptional regulation within the metabolic network of the photosynthetic bacterium Synechocystis sp. PCC6803. *Biotechnol J* 2011, **6**:330-342.
53. Tu CJ, Shrager J, Burnap RL, Postier BL, Grossman AR: Consequences of a deletion in dspA on transcript accumulation in Synechocystis sp. strain PCC6803. *J Bacteriol* 2004, **186**:3889-3902.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, **215**:403-410.
55. Burgard AP, Pharkya P, Maranas CD: Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003, **84**:647-657.
56. Liang C, Zhao F, Wei W, Wen Z, Qin S: Carotenoid biosynthesis in cyanobacteria: structural and evolutionary scenarios based on comparative genomics. *Int J Biol Sci* 2006, **2**:197-207.
57. Müller CP, Jacobs BL: *Handbook of the Behavioral Neurobiology of Serotonin.* 1st edition edition. London: Academic Press; 2010.
58. Henry CS, Broadbelt LJ, Hatzimanikatis V: Thermodynamics-based metabolic flux analysis. *Biophys J* 2007, **92**:1792-1805.

doi:10.1186/gb-2012-13-11-r111
Cite this article as: Vitkin and Shlomi: MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biology* 2012 **13**:R111.