BIOINFORMATICS

Vol. 25 ISMB 2009, pages i247–i252
doi:10.1093/bioinformatics/btp209

Network-based prediction of metabolic enzymes' subcellular localization

Shira Mintz-Oron[1,*], Asaph Aharoni[1], Eytan Ruppin[2,3] and Tomer Shlomi[4,*]

[1]Department of Plant Sciences, Weizmann Institute of Science, Rehovot 76100, [2]School of Computer Science, [3]School of Medicine, Tel-Aviv University, Tel-Aviv 69978 and [4]Faculty of Computer Science, Technion, Haifa 32000, Israel

ABSTRACT

Motivation: Revealing the subcellular localization of proteins within membrane-bound compartments is of a major importance for inferring protein function. Though current high-throughput localization experiments provide valuable data, they are costly and time-consuming, and due to technical difficulties not readily applicable for many Eukaryotes. Physical characteristics of proteins, such as sequence targeting signals and amino acid composition are commonly used to predict subcellular localizations using computational approaches. Recently it was shown that protein–protein interaction (PPI) networks can be used to significantly improve the prediction accuracy of protein subcellular localization. However, as high-throughput PPI data depend on costly high-throughput experiments and are currently available for only a few organisms, the scope of such methods is yet limited.

Results: This study presents a novel constraint-based method for predicting subcellular localization of enzymes based on their embedding metabolic network, relying on a parsimony principle of a minimal number of cross-membrane metabolite transporters. In a cross-validation test of predicting known subcellular localization of yeast enzymes, the method is shown to be markedly robust, providing accurate localization predictions even when only 20% of the known enzyme localizations are given as input. It is shown to outperform pathway enrichment-based methods both in terms of prediction accuracy and in its ability to predict the subcellular localization of entire metabolic pathways when no a-priori pathway-specific localization data is available (and hence enrichment methods are bound to fail). With the number of available metabolic networks already reaching more than 600 and growing fast, the new method may significantly contribute to the identification of enzyme localizations in many different organisms.

Contact: shira.mintz@weizmann.ac.il; tomersh@cs.technion.ac.il

1 INTRODUCTION

Eukaryotic cells contain several membrane-bound compartments called organelles that perform specialized biological functions. Subcellular compartmentalization allows the cell to maintain different environments that bring enzymes and substrates into physical proximity, participating in compartment-specific processes. Revealing the subcellular localization of proteins is of a major importance for inferring protein functions (Huh *et al.*, 2003) and for the discovery of drug targets (as some compartments are more easily accessible than others for drug molecules). Systematic protein localization experiments based on green fluorescent protein (GFP) tagging have been performed for several microbial species (Kumar *et al.*, 2002; Matsuyama *et al.*, 2006). Such large-scale experiments are both costly and time-consuming, and due to technical difficulties are commonly not applicable to higher Eukaryotes.

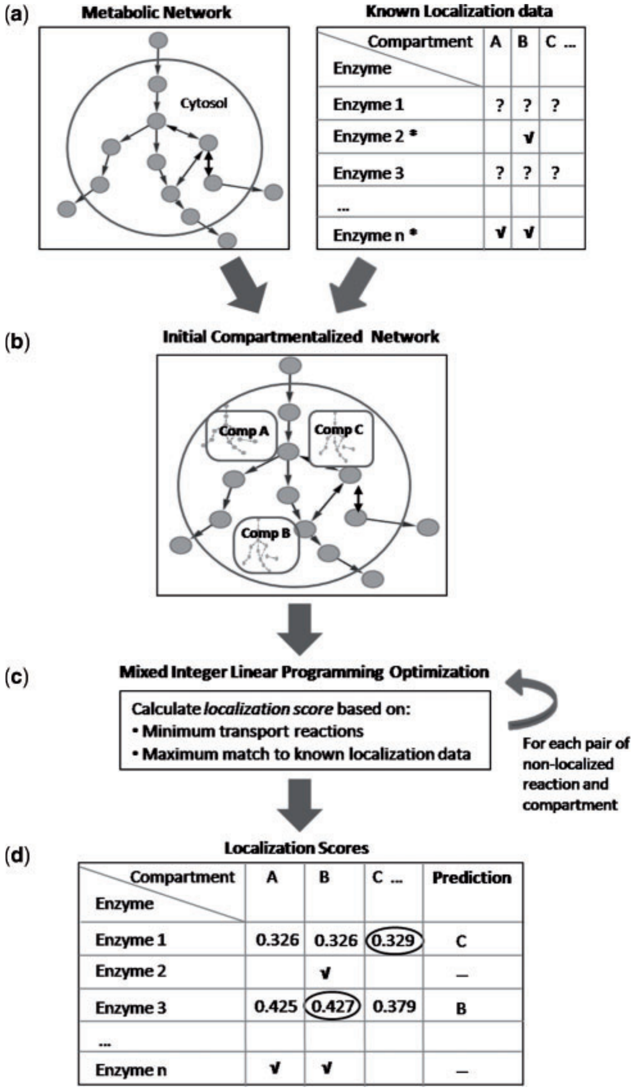
Current limitations of experimental procedures for identifying protein subcellular localization have given rise to ongoing development of computational methods for predicting localization data (Bhasin and Raghava, 2004; Emanuelsson *et al.*, 2000; Nakai and Horton, 1999; Scott *et al.*, 2004; Shatkay *et al.*, 2007). Such methods rely on lists of features that characterize a protein, such as its amino acid composition, their physio-chemical and structural properties, codon-bias, protein motifs and targeting signals (short stretches of amino-acid residues predominantly located at the N-terminus). The various localization methods apply different supervised classification approaches (e.g. artificial neural networks, nearest neighbor, SVM, etc.) to predict protein subcellular localization based on training data of experimentally determined protein localization. The performance of these methods significantly varies between different organisms and compartments, requiring specific calibration in each context. Recent studies have investigated a complementary approach for predicting subcellular localization, utilizing large-scale protein–protein interaction (PPI) networks (Lee *et al.*, 2008; Scott *et al.*, 2005). These methods are based on the assumption that two proteins should be localized within the same or adjacent compartments in order to interact. The work of Lee *et al.* has shown that in some cases, utilizing a PPI network may provide accurate localization predictions even without relying on common protein characteristics such as those described above.

This study presents the first method that predicts the subcellular localization of enzymes based on a metabolic network. Relying on a metabolic network rather than a PPI network is highly advantageous as metabolic networks are readily available for hundreds of species based on cross-species enzyme sequence homology (Kanehisa and Goto, 2000), while large-scale PPI networks depend on costly high-throughput experiments that are currently available for only a few organisms. Metabolic enzymes are also less likely to yield PPI interactions (Uetz *et al.*, 2000); e.g. the probability of a PPI between metabolic enzymes in yeast is less than half of the probability of an interaction between other non metabolic proteins (based on data extracted from the DIP database (Salwinski *et al.*, 2004)). Thus, PPI-based localization methods have far less data to bootstrap upon the localization of metabolic enzymes. Yet, constraint-based modeling of metabolic networks was previously shown to predict strong functional associations between enzymes (Notebaart *et al.*, 2008; Rokhlenko *et al.*, 2007), and to successfully predict various

[*]To whom correspondence should be addressed.

metabolic phenotypes in microorganisms [see Price *et al.* (2004) for a review] and recently in human (Duarte *et al.*, 2007; Shlomi *et al.*, 2008).

Considering a metabolic network view of metabolic processes, the potential activity of an enzyme in a certain compartment depends on the activity of many other enzymes in the compartment synthesizing and degrading its substrate metabolites and on the activity of membrane transporters that move metabolites between compartments. Indeed, metabolic processes tend to be clustered within various compartments without extensive usage of cross-membrane metabolite transporters. For example, 36% and 47% of the metabolic pathways in yeast are confined within only one or two compartments, respectively (based on pathway annotation data in the model of (Duarte *et al.*, 2004). This may be explained by the fact that the cross-membrane exchange of many of the metabolites depends on transporter proteins, imposing an energetic cost (e.g. demanding an ATP or GTP molecule per each metabolite translocation (Palmieri *et al.*, 2000) or requiring the maintenance of a membrane potential (Wada *et al.*, 1987). *Accordingly, our method predicts the subcellular localization of enzymes relying on a parsimony principle of a minimal number of cross-membrane metabolite transports.* The utility of the method is demonstrated via a cross-validation test of predicting known subcellular localization of yeast enzymes.


(a)
Metabolic Network
Cytosol
Known Localization data
(b)
Initial Compartmentalized Network
Comp A
Comp C
Comp B
(c)
Mixed Integer Linear Programming Optimization
Calculate *localization score* based on:
• Minimum transport reactions
• Maximum match to known localization data
For each pair of non-localized reaction and compartment
(d)
Localization Scores

Compartment / Enzyme	A	B	C ...
Enzyme 1	?	?	?
Enzyme 2 *		✓	
Enzyme 3	?	?	?
...			
Enzyme n *	✓	✓	

Compartment / Enzyme	A	B	C ...	Prediction
Enzyme 1	0.326	0.326	0.329	C
Enzyme 2		✓		–
Enzyme 3	0.425	0.427	0.379	B
...				
Enzyme n	✓	✓		–

Fig. 1. A schematic representation of the enzyme subcellular localization prediction method. (**a**) The input data is a metabolic network, representing a set of enzyme-catalyzed reactions, and the known localization data for a subset of enzymes. (**b**) Integrating the given network and localization data yields an initial compartmentalized network, consisting of several compartments. Localized reactions appear in the corresponding compartments while the non-localized reactions are duplicated to all compartments. (**c**) Mixed Integer Linear Programming (MILP) is applied for each pair of non-localized reaction and compartment to calculate a *localization score*, reflecting the likelihood of this reaction to be present in that compartment. (**d**) Enzymes are predicted to be localized in compartments achieving the highest localization scores.

2 METHODS

We present a new constraint-based modeling (CBM) method for systematically predicting subcellular localization of enzymes in a metabolic network, based on a-priori localization data for a subset of the enzymes, relying on a parsimony principle of minimal number of cross-membrane metabolite exchange. A schematic representation of the method is presented in Figure 1. The input data for this method is a metabolic network, representing a set of enzyme-catalyzed reactions, and the known localization of a subset of the enzymes (Fig. 1a). Enzymatic reactions whose subcellular localization is given as input are referred to as *localized reactions*. We refer to the remaining reactions as *non-localized reactions*. The first step of our method involves the integration of the given metabolic network and the known localization data to construct an *initial compartmentalized network* (Fig. 1b). This network consists of several compartments, in which localized reactions may be activated in the corresponding compartments, and non-localized reactions are duplicated to be present in all compartments. Next, we apply a Mixed Integer Linear Programming (MILP) method to compute a *localization score* for each pair of non-localized reaction and compartment, reflecting the likelihood of this reaction to be present in that compartment (Fig. 1c). These scores are used to determine the localization of each non-localized reaction, which is the output of the method (Fig. 1d). Next, we provide a brief overview on constraint-based modeling, followed by a detailed description of the various steps of our compartment prediction method.

2.1 Constraint-based modeling of metabolic networks

A metabolic network consisting of *n* metabolites and *m* reactions can be represented by a *stoichiometric matrix*, denoted *S*, in which the $S_{i,j}$ represents the stoichiometric coefficient of metabolite *i* in reaction *j* (Price *et al.*, 2004). A steady-state flux distribution (i.e. an assignment of flux rates to all reactions in the network), denoted *v*, should satisfy the following mass-balance constraint:

$$S\cdot v=0$$

The exchange of metabolites with the environment is represented as a set of *exchange reactions,* enabling for a pre-defined set of metabolites to be either taken-up or secreted from the growth media. Available metabolite localization data can be incorporated within *S*, by considering each row as an instance of a metabolite in a specific compartment, and each column as a reaction that involves metabolites in specific compartments (Duarte *et al.*, 2004). In this case, additional *transport reactions* (incorporated as additional columns in *S*) are used to represent metabolite translocation across compartments. In addition to mass-balance, a-priori data on reaction directionality can be used to enforce flux rates to have positive values for non-reversible reactions. In some cases, additional flux capacity constraints are available and can be further applied. The stoichiometric mass-balance, directionality and capacity constraints give rise to a space of feasible flux distributions that has been analyzed by various optimization methods (Price *et al.*, 2004).

2.2 Constructing an initial compartmentalized network

We construct an initial compartmentalized network model with k compartments, in which (i) all metabolites are present in all k compartments; (ii) transport reactions enable metabolite exchange between cytoplasm and all k compartments. We denote by $\bar{t}^i$ the transport reactions between the cytoplasm and compartment i; (iii) all reactions are present in all k compartments. (iv) localized reactions can be activated (i.e. carry non-zero flux) only in their associated compartments, while non-localized reactions can be activated in all compartments. We denote the set of compartments in which a localized reaction i can be activated by Ci. In this model, a steady-state flux distribution, $v=(\bar{v}^1,\bar{v}^2,\ldots,\bar{v}^k,)$, should enforce the following stoichiometric mass-balance, flux directionality and capacity constraints:

$$S\cdot\bar{v}^1=\sum_{i=2}^{k}\bar{t}^i+\bar{e}\cdot\bar{t}^1 \tag{1}$$

$$S\cdot\bar{v}^i=-\bar{t}^i,\quad i=2\ldots k \tag{2}$$

$$v_{\min}\leq\bar{v}^i\leq v_{\max},\quad i=1\ldots k \tag{3}$$

$$v_j^i,\quad \forall_{i,j}|C_i|>0 \text{ and } \quad j\notin C_i \tag{4}$$

where, $\bar{v}^1$ denotes the flux distribution in the cytoplasm, and $\bar{e}$ denotes the set of enabled exchange reactions (i.e. all zero, except for ones for exchange reactions) that move cytosolic metabolites across the model boundaries. Equation (1) enforces mass-balance in the cytosol, where all metabolites may be transported to all other compartments, and a subset of them exchanged with the growth environment. To enable the activation of a significant fraction of the reactions in the network, we considered a rich media in which all exchange reactions in the network model are enabled. Equation (2) enforces mass-balance for each compartment i. Equation (3) enforces flux directionality and capacity constraints, once available in the model. Equation (4) restricts flux activity of localized reactions only to their known compartments (for which the set of associated compartments Ci consists of at least a single compartment, i.e. $|Ci|>0$).

2.3 A MILP optimization for predicting enzyme localization scores

To predict a subcellular localization of non-localized reactions, we employ MILP to predict a feasible flux distribution v in the initial compartmentalized network, in which: (i) a maximal number of localized reactions are activated (i.e. carry non-zero flux) in their known compartment, and (ii) a minimal number of transport reactions are activated. The predicted flux distribution is used to infer the localization score for each reaction and compartment pair, as described below. To enable counting of the number of localized reactions that are correctly activated in the right compartments, for each localized reaction i and its known compartment j, we define Boolean auxiliary variables, $y_j^{i,+},y_j^{i,-}$, representing whether reaction i indeed carries non-zero flux in either the forward or backward directions in compartment j, respectively [following (Shlomi *et al.*, 2008)]:

$$v_j^i+y_j^{i,+}(v_{\min,j}-\varepsilon)\geq v_{\min,j},\forall|C_i|>0 \tag{5}$$

$$v_j^i+y_j^{i,-}(v_{\max,j}-\varepsilon)\geq v_{\max,j},\forall|C_i|>0 \tag{6}$$

where ε denotes a minimal flux rate in which a reaction is considered as active. We used ε=0.1, following previous studies (Shlomi *et al.*, 2008), though other choices provided qualitatively similar results. Using a similar method to count the number of activated transport reactions is problematic as it requires $k\cdot n$ additional Boolean auxiliary variables, which is computationally intractable. Instead, we employ a relaxation based on an $L1$ metric (Shlomi *et al.*, 2005), minimizing the absolute value of fluxes rate through transport reactions (which requires no additional Boolean variables). The resulting objective function of our MILP formulation is hence:

$$\max_{\bar{v}^i},\bar{y}^{i,+},\bar{y}^{i,-},\bar{t}^i\sum_{j=1}^{m}(y_j^{i,+}+y_j^{i,-})-\sum_{j=1}^{n}p|t_j^i| \tag{7}$$

where p denotes penalty per unit of flux through a transport reaction. We choose a penalty value satisfying $p<<1/(k\cdot n)$ (where $k\cdot n$ equals the total number of transport reactions), such that the optimization criteria of maximizing the activity of localized reactions is given a higher priority over the optimization criteria of minimizing the activity of transport reactions. Using this parameter, the optimization problem is equivalent to first, maximizing the number of localized reactions that are activated and only then, minimizing the flux through transport reactions. The transport of *currency metabolites* that participate in a high number of reactions in the network (above 20; e.g. ATP, NADH, H2O) was not associated with a penalty score, as these metabolites are assumed to be present in all compartments. The commercial CPLEX solver was used for solving MILP problems, on a Pentium-4 machine running Linux in dozens of seconds per problem.

Once an optimal flux distribution is computed via our MILP formulation, the localization of enzymes can be predicted based on their flux activity in various compartments. However, due to the existence of alternative pathways in the network, alternative optimal flux distributions may exist, resulting in different localization predictions in each case. To handle such alternative MILP solutions we employ a method similar to flux variability analysis (FVA) (Mahadevan and Schilling, 2003). Specifically, for each non-localized reaction i and compartment j, we compute an optimal flux distribution using the above MILP method, while forcing the reaction to have non-zero flux specifically in that compartment. The resulting optimization value is referred to as the localization score. In the cross-validation tests described below, we predict for each non-localized enzyme either one or two compartments (based on the actual number of compartments the reaction is known to be localized to), by considering the compartments achieving the highest localization scores. In case of an enzyme having the same optimal localization score for both cytoplasm and another compartment, the enzyme is predicted to be localized in the other compartment (to avoid bias due to the large size of the cytoplasm in which reaction activation is a-priori more likely to require a lower number of transporters). When the same localization score is obtained for more than two compartments, we provide no localization prediction for that enzyme (lowering the prediction coverage).

2.4 An illustrative example of enzyme subcellular localization predictions

An illustrative example of the prediction method is depicted in Figure 2. In this example, the metabolic network contains 11 metabolites, 8 enzymatic reactions and 4 exchange reactions. The initial compartmentalized network is shown in the figure, depicting three compartments (cytoplasm, compartment-A, compartment-B), and transport reactions for metabolites between cytoplasm and the two other compartments. Localized reactions, given as input, are marked with thick edges ($R2$, $R4$, $R6$ and $R8$), while non-localized reactions ($R1$, $R3$, $R5$ and $R7$), whose localizations we wish to predict, appear in all three compartments.

We applied our MILP method to predict an optimal flux distribution for this example, and show the derived flux distribution within the network (i.e. marking activated, non-zero flux, reactions). As shown, the flux distribution spans all three compartments, starting from the uptake of metabolite $M1$ and $M10$ from the growth environment, and ending with the secretion of metabolites $M5$ and $M9$. Reactions $R1$ and $R5$ are predicted to be localized to compartment-A, as the transport of a single metabolite $M1$ into the compartment enables the production of metabolites $M2$ and $M7$, required for activating reactions $R2$ and $R6$ in this compartment. Reaction $R7$ is predicted

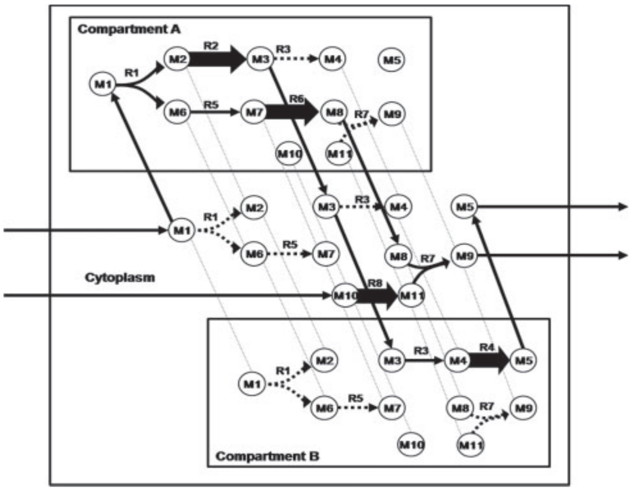


Fig. 2. An illustrative example of our method for enzyme subcellular compartment prediction. The initial compartmentalized network consists of three compartments, with instances of 11 metabolites and 8 reactions in each compartment. Thin edges connecting different instances of a metabolite in various compartments represent transport reactions that move metabolites across membrane boundaries. Wide arrows represent localized reactions whose known localization is given as input to the prediction method. Solid arrows represent reactions that are predicted to have non-zero flux by our method reflecting their predicted localization. Dashed arrows represent reactions predicted to have zero flux.

to be localized in the cytoplasm, as its substrate $M11$ is produced by $R8$ solely in the cytoplasm, and hence activating $R7$ in a different compartment would require an unnecessary transport of $M11$ out of the cytoplasm. An example in which the method cannot uniquely predict the localization of a certain reaction is in the case of reaction $R3$ that produces metabolite $M4$ from $M3$. Activation of the localized reaction $R4$ in compartment-B requires that its substrate metabolite $M4$ would be present in this compartment. Metabolite $M3$ is produced solely in compartment-A, and hence reaction $R3$ can be activated in compartment-A (with $M4$ being transported via two transporters to compartment-B), or $M3$ can be transported to the cytoplasm or to compartment-B and $R3$ activated in the cytoplasm or compartment-B, respectively. In all three cases, the activation of $R4$ in compartment-B would have the same total cost of activating two transport reactions.

3 RESULTS

3.1 Validating the localization prediction via a metabolic network of Saccharomyces cerevisiae

To evaluate the performance of our method, we applied it to predict enzyme localization for metabolic enzymes in the yeast *S.cerevisiae*. Both the metabolic network and subcellular localization data for *S.cerevisiae* are available within the genome-scale, fully compartmentalized metabolic network model of (Duarte *et al.*, 2004). This network model accounts for 750 genes, 1062 metabolites and 1149 reactions, acting in seven compartments: cytosol, mitochondrion, peroxisome, nucleus, endoplasmic reticulum, golgi apparatus, and vacuole. The cytosol is the largest compartment, consisting of 65% of the total metabolic reactions, followed by the mitochondrion, consisting of 16% of the reactions.

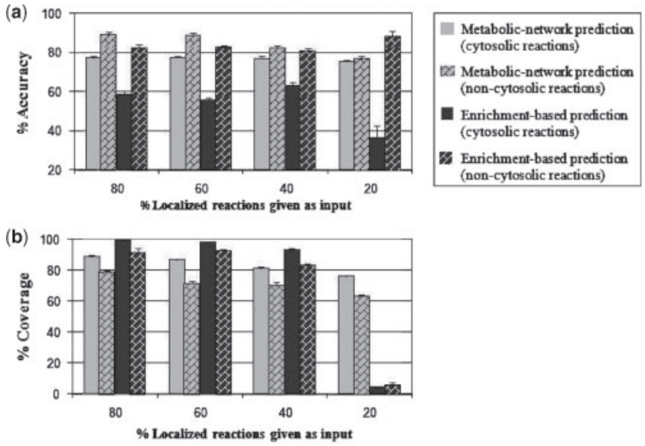


Fig. 3. Accuracy (**a**) and coverage (**b**) of enzyme subcellular localization predictions in a cross-validation test in the yeast *S.cerevisiae*. The average and standard error of the accuracy and coverage measures were calculated based on 10 applications of the prediction methods over randomly sampled sets of localized enzymes of similar size that are used as input.

To evaluate our method, we first removed all existing localization data from the network model of Duarte *et al.*, forming a new stochiometric matrix with a single merged compartment that can be given as input to our prediction method. Then, we applied a cross-validation test, in which we randomly partitioned the enzymes to localized and non-localized sets and applied our method to predict the localization of the non-localized enzymes, given the localized enzymes and the metabolic network. For the non-localized enzymes, we assumed that prior knowledge (obtained for example, via sequence-based prediction methods such as those described above) narrows down the list of potential localizations of enzymes to one out of four compartments, and hence restricts the activity of non-localized reactions in the model to three randomly chosen compartments in addition to the correct compartment. The specific choice of restricting non-localized reactions to four compartments was made based on a comprehensive analysis of ten prediction methods applied to *Arabidopsis thaliana*, which showed that over 90% of its enzymes are predicted to be localized to no more than four compartments (Heazlewood *et al.*, 2007). To further evaluate the performance of our method, we compare it with an enrichment-based method that predicts subcellular localization based on an assumption that pathways are coherently localized in various compartments. Specifically, for each non-localized reaction participating in a certain pathway we compute a hyper-geometric p-value reflecting the pathway's enrichment with enzymes localized within each compartment. The localization of the reaction is predicted based on the compartment that yields the lowest p-value. The metabolic pathway data was also obtained from the metabolic network model of Duarte *et al.*

The accuracy and coverage of our method in comparison with the enrichment-based method, for various fractions of localized enzymes input sets, are shown in Figure 3. Since the known distribution of enzyme compartment localization is significantly skewed towards the cytoplasm, we present the accuracy and coverage statistics separately for cytosolic reactions and non-cytosolic reactions (showing that our method correctly predicts localization in both cases). The accuracy of our method is markedly robust, remaining above 78% for both cytosolic and non-cytosolic reactions when the percentage of given localized reactions is as low as 20%. The coverage shows a moderate decline from 88% and 79% for cytosolic and non-cytosolic reactions, respectively, given the 80% localized reactions as input, towards 78% and 63% in the case of 20% localized reactions. This decline in coverage is quite expected, as when the set of localized reactions used as input is small, many reactions have the same likelihood of being active in various compartments. The pathway enrichment-based method constantly achieves a markedly lower accuracy (except for the case in which only 20% localized reactions are used as input in which its coverage is minute), especially for the prediction of cytosolic enzymes. The coverage of this method shows a slight advantage over our network-based approach for high fractions of localized reactions used as input. However, the pathway enrichment-based method is shown to be highly non-robust when the fraction of given localized reactions decreases—when it reaches 20% its coverage significantly drops to below 10%. The failure of the pathway enrichment-based method to match our network-based approach is somewhat expected, considering that many metabolic pathways do cross compartmental boundaries (as discussed above) and as it does not account for pathways intersection as shown below.

3.2 An example of predicting the subcellular localization of complete pathways

We further tested our method in predicting subcellular localization for a set of enzymes with unknown localization that covers two complete pathways. In such cases, a-priori localization data is not available for any enzyme in a certain pathway, and hence pathway enrichment-based methods are bound to fail, requiring a network-based view of pathways connectivity. Specifically, we aim to predict the localization of a group of enzymes composing the TCA cycle and the glyoxylate cycle, given the known localization all of the remaining enzymes in the network (Fig. 4). The set of enzymes to predict consist of 12 reactions: three ethanol fermentation reactions (with isozymes localized to both cytoplasm and mitochondria), four TCA cycle reactions (only in mitochondria; referred to as mitochondria-specific TCA cycle reactions), two glyoxylate cycle reactions [one in both peroxisome and cytosol and another only in cytosol) and three reactions involving both gloxylate and TCA (with isozymes localized in all three compartments; (Regev-Rudzki *et al.*, 2005)]. The three reactions involved in the ethanol oxidation pathway are correctly predicted to be localized both in the cytosol and mitochondria. The four mitochondria-specific TCA cycle reactions are correctly predicted to be localized in mitochondria and the three reactions involving both gloxylate and TCA pathways are correctly predicted to be localized in all three compartments. The glyoxylate reaction localized to both peroxisome and cytosol is correctly predicted to be localized in the cytosol, though its second most likely localization is falsely predicted to be mitochondria, while only its third predicted localization is peroxisome. A similar problem arises with the localization prediction of the glyoxylate reaction that is known to be localized only in the cytosol, where the method predicts a most likely mitochondrial localization. The two false predictions for these reactions result from the utilization of their substrate metabolites also in mitochondria. However, in both cases, the next most likely localization prediction given by our method is the correct one.

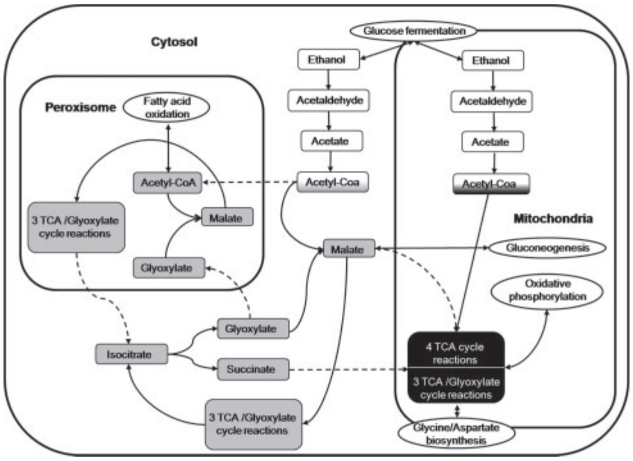


Fig. 4. Enzyme subcellular localization prediction of two complete metabolic pathways, including the TCA cycle (black rectangles) and glyoxylate cycle (grey rectangles), and a subset of the ethanol oxidation pathway (white rectangles), given localization data for enzymes in other connected pathways (white ellipses) as input. Transport reactions are marked by dotted arrows.

3.3 Validating emergent subcellular localization predictions via GO annotation

Following the cross-validation test, we turned to predict novel subcellular localizations of enzymes in the metabolic network. Towards this goal, we re-ran our method on the same network in a leave-one-out cross validation setup, in which localization data for all reactions but one was used to predict the localization of that single reaction. We found that our method predicts a non-cytosolic localization for 22 reactions in the model although they are localized in the model to the cytosol. Inspecting the GO cellular localization annotation for these reactions revealed that the localization of 10 out of the 22 was correctly predicted. This prediction accuracy is statistically significant ($p < 0.05$) compared with a random assignment of genes to compartments (with an assignment probability for each compartment relative to its size in GO).

An example of such emergent localization predictions that are not accounted for in the model is the case of enzymes SUR2 (dihydrosphingosine C-4 hydroxylase), and TSC10 (3-ketosphinganine reductase), which catalyze consecutive reactions in the ceramide biosynthesis pathway. Ceramides are formed as the key intermediates in the biosynthesis of sphingolipids, essential components of the plasma membrane. This pathway is known to be accomplished by ER enzymes, some of which can also be localized to the cytosol (Natter *et al.*, 2005). SUR2 is known to be localized exclusively to the ER, although mistakenly it is localized strictly to the cytosol in the model. TSC10 is experimentally localized to both the ER and the cytosol, though again, mistakenly it is localized in the model only to the cytosol. Our method predicts the correct localization of both enzymes in the ER.

4 DISCUSSION

This study presents a novel constraint-based modeling method for predicting subcellular localization of enzymes embedded in a metabolic network. While constraint-based modeling was previously employed to predict many different metabolic phenotypes, this is the first application of this approach to predict subcellular localization. The method is based on a parsimonious principle of minimal number of metabolite transports across compartments—a novel concept in the context of metabolic network analysis, which enables the prediction of plausible flux distributions that are clustered within various compartments, but does not strictly enforce such a clustering in a 'hard-wired' manner.

While previous methods have used PPI networks to improve localization predictions, relying on metabolic networks is advantageous as metabolic networks are readily available for hundreds of species. Another advantage is in regard to the prediction of metabolic enzyme localization, as metabolic enzymes are less likely to yield PPIs. However, an evaluation of the performance of PPI-based localization prediction methods when applied strictly to metabolic enzymes has yet to be performed. An inherent limitation of metabolic network-based localization prediction is that it is strictly limited to metabolic enzymes. Furthermore, we note that our method is computationally more demanding than common methods that rely on standard supervised classification approaches, requiring multiple mixed integer optimizations.

We intend to apply this method to predict subcellular localizations in other organisms in which the true localization is unknown. For example, in the plant model organism *A.thaliana* for which experimentally determined localization data is available for only 50% of the metabolic enzymes (Heazlewood *et al.*, 2007). Different aspects of protein's localization can be explored using this method, such as the dependency of localization on various growth media (John *et al.*, 2006). While here we rely on this principle of minimal metabolite exchange to partition a network to its subcellular compartments, future research may employ the same principle to partition metabolic networks into different subsystems. For example, with the availability of a genome-scale human metabolic network model, a similar approach may be used to partition it to various tissue-specific subsystems. Or in the realm of meta-genomics, similar approaches may be used to partition a metabolic network of population of species to species-specific subsystems.

Funding: Israeli Science Foundation grant (to E.R.); A.A. is the incumbent of the Adolpho and Evelyn Blum Career Development Chair of Cancer Research; the work in the Aharoni laboratory was supported by Mrs Louise Gartner, Dallas, TX and Mr and Mrs Mordechai Segal, Israel.

Conflict of Interest: none declared.

REFERENCES

Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414–W419.

Duarte,N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA*, **104**, 1777–1782.

Duarte,N.C. *et al.* (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.*, **14**, 1298–1309.

Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.

Heazlewood,J.L. *et al.* (2007) SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res.*, **35**, D213–D218.

Huh,W.K. *et al.* (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.

John,M. *et al.* (2006) Growth substrate dependent localization of tetrachloroethene reductive dehalogenase in Sulfurospirillum multivorans. *Arch. Microbiol.*, **186**, 99–106.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Kumar,A. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.*, **16**, 707–719.

Lee,K. *et al.* (2008) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic Acids Res*, **36**, e136.

Mahadevan,R. and Schilling,C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, **5**, 264–276.

Matsuyama,A. *et al.* (2006) ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. *Nat. Biotechnol.*, **24**, 841–847.

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.

Natter,K. *et al.* (2005) The spatial organization of lipid synthesis in the yeast Saccharomyces cerevisiae derived from large scale green fluorescent protein tagging and high resolution microscopy. *Mol. Cell Proteomics*, **4**, 662–672.

Notebaart,R.A. *et al.* (2008) Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput. Biol.*, **4**, e26.

Palmieri,L. *et al.* (2000) Yeast mitochondrial carriers: bacterial expression, biochemical identification and metabolic significance. *J. Bioenerg. Biomembr.*, **32**, 67–77.

Price,N.D. *et al.* (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, **2**, 886–897.

Regev-Rudzki,N. *et al.* (2005) Yeast aconitase in two locations and two metabolic pathways: seeing small amounts is believing. *Mol. Biol. Cell*, **16**, 4163–4171.

Rokhlenko,O. *et al.* (2007) Constraint-based functional similarity of metabolic genes: going beyond network topology. *Bioinformatics*, **23**, 2139–2146.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Scott,M.S. *et al.* (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res.*, **14**, 1957–1966.

Scott,M.S. *et al.* (2005) Refining protein subcellular localization. *PLoS Comput. Biol.*, **1**, e66.

Shatkay,H. *et al.* (2007) SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, **23**, 1410–1417.

Shlomi,T. *et al.* (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl Acad. Sci. USA*, **102**, 7695–7700.

Shlomi,T. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.

Uetz,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Wada,Y. *et al.* (1987) Vacuolar ion channel of the yeast, Saccharomyces cerevisiae. *J. Biol. Chem.*, **262**, 17260–17263.