

An automated approach to assigning subcellular localization in genome-scale metabolic network reconstructions

Master of Science Thesis in the Master Degree Program in Innovative and Sustainable Chemical Engineering

SAEED SHEYKHSHOAIEEKHTIARABADI

Department of Chemical and Biological Engineering
Systems and Synthetic Biology
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden, 2010

An automated approach to assigning subcellular localization in genome-scale metabolic network reconstructions

SAEED SHEYKHSHOAIEEKHTIARABADI

Department of Chemical and Biological Engineering
Chalmers University of Technology
Göteborg, Sweden, 2010

An automated approach to assigning subcellular localization in genome-scale metabolic network reconstructions

Saeed Sheykhshoaieekhtiarabadi
Supervisor(s): Jens Nielsen, Rasmus Ågren
Examiner: Jens Nielsen
Department of Chemical and Biological Engineering
Systems and synthetic biology

Chalmers University of Technology

Abstract:

Genome-scale models take a simplified view on metabolism by only considering the stoichiometry of the metabolic reactions. Despite this simplification, these models have proved useful in many different areas, such as in identification of metabolic engineering targets, for analyzing metabolite connectivity and pathway redundancy or for studying metabolic interactions between species. Due to the increasing popularity of this type of models, and the necessity to reconstruct models for less well-studied organisms, there is a need for tools to automate the reconstruction process. Recently there have been attempts to automate the network reconstruction based on protein sequence homology. However, these strategies are mainly aimed at prokaryotic systems where there is no subcellular compartmentalization of enzymes. Here we present a method for assigning subcellular localization to enzymatic reactions in an automated fashion. The algorithm aims at assigning localization in a manner that is consistent with signal peptide composition and physiochemical protein properties, while at the same time maintaining a well-connected and functional network. Non-enzymatic reactions, such as diffusion across membranes, are inferred based on connectivity. We believe that this technique can significantly speed up the otherwise very time consuming and laborious task of model reconstruction for eukaryotic organisms.

Contents

1.1	Reconstruction of metabolic networks.....	5
1.2	Automated approaches in genome-scale metabolic network reconstruction	9
1.3	Localization in genome-scale metabolic network reconstruction	10
1.3.1	Subcellular localization predictors.....	10
1.4	Project summary	13
2	Method.....	14
2.1	Problems	14
2.2	Heuristic optimization.....	14
2.3	Fully-connected Localization Assignment (F-LocA)	14
2.1.3	Preliminary steps.....	14
2.1.4	Automated assigning based on full-connectivity and simulated annealing	15
3	Result	17
3.1	Evaluation	17
3.1.1	<i>Saccharomyces cerevisiae</i>	17
3.1.2	<i>Penicillium chrysogenum</i>	18
3.2	Case study for <i>Pichia stipitis</i>	19
4	Discussion.....	22
5	References.....	24

1. Introduction

Network reconstruction plays a central role in the field of systems biology. Depending on the application, several different types of networks are used, including; metabolic networks, protein-DNA networks, regulatory networks, and signaling networks (Palsson 2006).

Metabolic networks are debatably the most extensive type of networks, due to good availability of high-quality data, and the type that most closely relates the molecular function of the cell to the observed phenotype.

Metabolic networks consist of a series of metabolic process such as metabolism of complex carbohydrates, nucleotides, amino acids, cofactors and vitamins, lipids and many other part of metabolism (Francke, Siezen et al. 2005). A metabolic network can be an important facilitator in many metabolic engineering applications. They have been used to find sets of gene deletions, amplification targets, or heterologous genes that can be introduced to result in the production of some desired chemical (Palsson 2006). These models can be applied for the interpretation of experimental data and also metabolic network analysis (Patil, Akesson et al. 2004)s.

The first step in constructing a mathematical model for (some part of) metabolism is defining a metabolic network. From there, there are two distinctly different types of models: Stoichiometric models and kinetic models (Patil, Akesson et al. 2004). The simpler types of models are stoichiometric models, where metabolic reactions are represented only by the stoichiometry of the metabolites involved. Kinetic models try to capture the cellular behavior over time, which requires that kinetic parameters for each enzyme can be determined (Borodina and Nielsen 2005). This type of data is not readily available and therefore it is often necessary to use the simpler stoichiometric model, which essentially models only steady state situations. This is especially true for large metabolic models, the extreme case being genome-scale metabolic models (GSMMs) which aim at incorporating all metabolic reactions for a given organism (Reed, Vo et al. 2003).

1.1 Reconstruction of metabolic networks

Broadly speaking, there are two different approaches to network reconstruction: bottom-up and top-down. The end goal of both of these approaches is a mathematical model, however with different origin and evolution. A bottom-up reconstruction is based on knowledge about single reactions, which are stitched together to form a functional model. In contrary, the top-down approach takes a more holistic view of the system and considers the whole network during the reconstruction (Palsson 2006). Figure 1 illustrates this aspect schematically. The bottom-up approach is carried out via direct methods, and in contrary top-down approach is normally carried out via inference methods.

Since complete genome sequences of many organisms are accessible, possible opportunities arises to generate a metabolic models for less-studied organisms (Patil, Akesson et al. 2004). Table 1 shows reconstructed GSMMs for some important model-species: *E. coli*, *S. cerevisiae*,

A. nidulans, *A. thaliana* and *M. musculus* as a first bacteria, eukaryotes, filamentous fungi, plant and mammalian respectively for which GSMMs were reconstructed. As Table 1 shows, by duration of time due to availability of more data and literature, the number of identified genes, reactions and metabolites has increased. Consequently, the number of included subcellular compartments increased in eukaryote models.

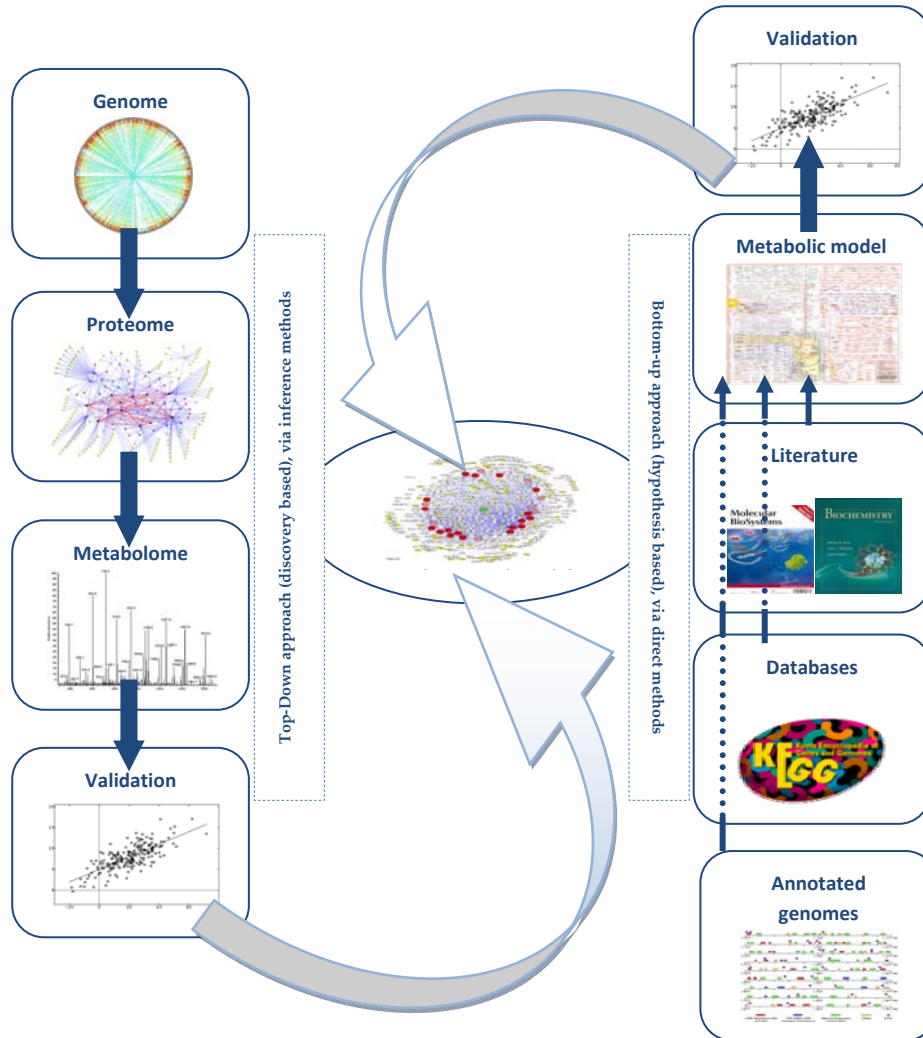


Figure 1: Top-down and bottom-up approaches for network reconstruction. Adapted from (Palsson 2006).

The genome-scale metabolic network reconstruction can be considered to consist of two parts; the extraction of metabolic biochemistry and gene annotation of an organism for network reconstruction; and supplementing this reconstruction by cell physiology of the organism to build the microbial metabolic model (Covert, Schilling et al. 2001).

Table 1: Reconstructed genome-scale metabolic model properties for different model organisms.

Model	Organism	References	Reactions	Metabolites	Genes	Compartments
iJR904	E. coli	(Reed and Palsson 2004)	931	761	904	1
iFF708	S. cerevisiae	(Forster, Famili et al. 2003)	1175	584	708	3
iHD666	A. nidulans	(David, Hofmann et al. 2006)	1213	732	666	4
-	M. musculus	(Quek and Nielsen 2008)	2037	2104	1399	2
AraGEM	A. thaliana	(Baerenfaller, Grossmann et al. 2008)	1320	1438	1427	5

The process of genome-scale metabolic network reconstruction can be compromised in to four essential steps and one additional step as a passage to design and discovery. These steps are summarized in Figure 2.

1) Draft reconstruction based on annotated genome

For a specific organism, a draft genome annotation is the preliminary step for reconstruction and can be achieved from different databases depending on target organism such as SGD (Saccharomyces Genome databases) (Christie, Weng et al. 2004), Ecocyc (Karp, Keseler et al. 2007) for E. coli or comprehensive databases like IMG (Integrated Microbial Genomes) (Markowitz, Korzeniewski et al. 2006) or EntrezGene (Maglott, Ostell et al. 2007). For less well studied organisms, it can be necessary to infer gene function from homology. In either case, all genes that encode for metabolic enzymes are collected and included in the model. Setting up biochemical reactions that are carried out by those enzymes can be done manually or by some automated approaches. In this stage many metabolic databases such as BRENDA (Schomburg, Chang et al. 2004), MetaCyc (Krieger, Zhang et al. 2004) or KEGG (Kyoto Encyclopedia of Genes and Genome) (Kanehisa and Goto 2000) function to relate observed reactions in target organism to EC codes (International Union of Biochemistry and Molecular Biology. Nomenclature Committee. and Webb 1992). Although this transferring of information can also be done through some automated tools, the information relating to compartmentalization and reaction reversibility might be difficult or impossible to obtain. (Feist, Herrgard et al. 2009)

2) Curation of the reconstruction

Automatically reconstructed metabolic model include misinterpreted reactions which in reality are not carried out *in vivo*. In addition, the first step results in a number of candidate biochemical reactions and the network will include lots of gaps. To deal with these deficiencies specific knowledge of the target organism is crucial (Thiele and Palsson 2010). Important sources to find this data are textbooks, literature, reviews and organism specific databases. This curation process involves lots of manual work and is a very time consuming process (Thiele and Palsson 2010). Therefore the overall function of this fundamental step is filling gaps in the network, based on evidence from organism specific databases or other available literature. By assimilation of

automated reconstruction (genome based) and manual curation (literature based), a high quality network reconstruction can be achieved. Based on biochemical reactions, the stoichiometric matrix is constructed in this step, which comprises the interconnection of reactions and metabolites in the network (Feist, Herrgard et al. 2009). This mathematical representation of reconstructed networks gives the relation between high-throughput data and *in silico* modeling which is applied to determine the network connectivity and pathway redundancy (Oberhardt, Palsson et al. 2009; Thiele and Palsson 2010).

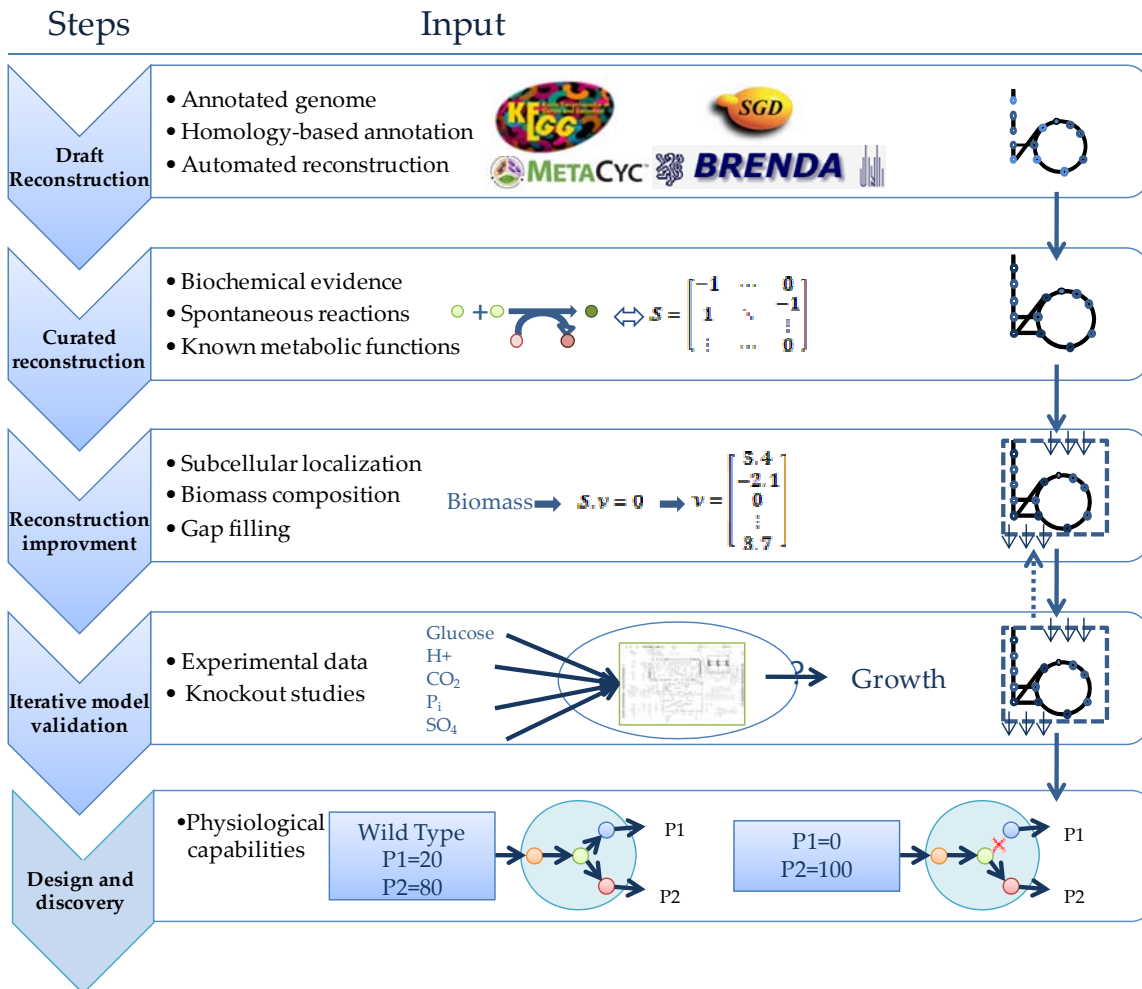


Figure 2: Fundamental steps in reconstruction of genome-scale metabolic network. In each step specific input is required. The whole process of reconstruction is done in an iterative manner, and the refining part is carried out by comparison to experimental data. The functional model generated in the fourth step can be used step as a platform for design and discovery. Adapted from (Feist, Scholten et al. 2006; Feist, Herrgard et al. 2009; Thiele and Palsson 2010)

3) Reconstruction improvement

The curated metabolic network is converted to a computational model and biomass composition data is introduced by accounting for all known biomass precursors (Feist, Herrgard et al. 2009). Growth on defined media and production of biomass constituents is simulated to check the functionality of model. Generating biomass composition often leads to the identification of additional gaps, which indicates that some of the biomass precursors are not produced. Gaps can be filled to account for the missing pathways or disregarded if they are mistakes from the first draft reconstruction. (Feist, Herrgard et al. 2009). The conversion of the metabolic network into a computational model is done via the stoichiometric matrix. A crucial stage in the formulation of the mathematical models is adding constraints to model, which are categorized in different types: mass balance, steady-state, thermodynamic (directionality), regulatory and environmental constraints (Becker, Feist et al. 2007).

4) Network evaluation and iterative validation.

The general errors in the model are identified in the step. Typical errors consist of lacking transport reactions, lacking exchange reactions, mistake in reaction constraints or inability of the model to consume to produce or consume cofactors (Thiele and Palsson 2010). In addition the ability of the model to correctly simulate the growth rate is evaluated, which may lead to identification of lacking metabolic function in the model (Thiele and Palsson 2010). There are a numbers of issues resolved during this step, such as identifying dead end metabolites, an extensive literature reviewing, and re-annotation of the genome to identify candidate reactions for filling the gaps and knock out studies that can lead discovery of new biological information (Thiele and Palsson 2010). This step is done in an iterative manner by returning the model to previous step for additional gap filling. The iteration stops after the model achieves the goals of the reconstruction.

5) A platform for discovery and design.

The evaluated GSMM is applied for different carbon sources and medium to expand the metabolic content. In this step different type of studies are interested such as analyzing network topology properties, phenotypic prediction by constraint-based model, define candidate genes for gap-filling and alternative pathway discovery (Thiele and Palsson 2010).

1.2 Automated approaches in genome-scale metabolic network reconstruction

The fundamental steps for reconstruction that are described above can be broken down to 96 steps (Thiele and Palsson 2010) which shows the complexity and slow pace of generating GSMM. To speed up the creation of models, lots of efforts have recently gone into establishing automated approaches to accelerate the creation of models, such as; Model SEED (Henry, DeJongh et al. 2010) and RAVEN Toolbox (Unpublished data).

The Model SEED is a web-based resource that is based on the SEED scaffold for accurate genome annotation (DeJongh, Formsma et al. 2007). It integrates gene-protein-reaction associations, generates the biomass constituents and reactions, convening reactions in a network, study reactions directionality through thermodynamic knowledge and finally optimizes the model, and attempts to create the draft GSMM (Henry, DeJongh et al. 2010). According to reports the model accuracy before optimization is 66% and after optimization it is 87% (Henry, DeJongh et al. 2010). The Model SEED is carried out for bacteria.

RAVEN Toolbox is a comprehensive toolbox for reconstruction, simulation and visualization of metabolic networks. The RAVEN Toolbox can generate draft models based on template models or KEGG by means of protein homology.

1.3 Localization in genome-scale metabolic network reconstruction

Gene and reaction localization is the one of the step for reconstruction of genome-scale model that mainly is considered for reconstruction of eukaryotes. The compartments that are most relevant for metabolism are normally considered to be cytoplasm, mitochondria, peroxisome and extracellular space. This step is a crucial step in reconstruction, since assigning a reaction to the wrong compartment lead to decreased connectivity of the model and further on cause additional gaps. In addition it might require the addition of intra-cellular transport reactions without associated evidence (Thiele and Palsson 2010).

1.3.1 Subcellular localization predictors

Different types of algorithm for predicting the cellular localization have been introduced such as CELLO (Yu, Chen et al. 2006), PSORT (Nakai and Horton 1999) or PASUB (Lu, Szafron et al. 2004). The algorithms focus on different aspects, but they have in common that localization of proteins is based on nucleotide or amino acid sequences. There are a number of different issues associated with subcellular prediction based on only gene/protein sequences, such as that some proteins have multiple locations. Due to high accuracy of predictions in CELLO and WoLFPSORT (Horton, Park et al. 2006), these two algorithms are briefly introduced.

The CELLO approach is done in two classes: in the first class coding regions are scored. The scores are classified through SVM (support vector machine) classifiers (Yu, Chen et al. 2006). Each of classifiers derived individual vectors from sequence. In the second class, the responses from each classifiers of first class are translated as a probability distribution (Yu, Chen et al. 2006). These scores are entered to last classifier, which works as a jury vote, to assign the protein's final localization. The most effective part of CELLO is that it takes advantage of four different types of sequence coding; the amino acid constituent, the di-peptide constituent, the segmented amino acid constituent and sequence built on physic-chemical properties of amino acids (Yu, Chen et al. 2006). The general approach of this method is summarized in Figure 3.

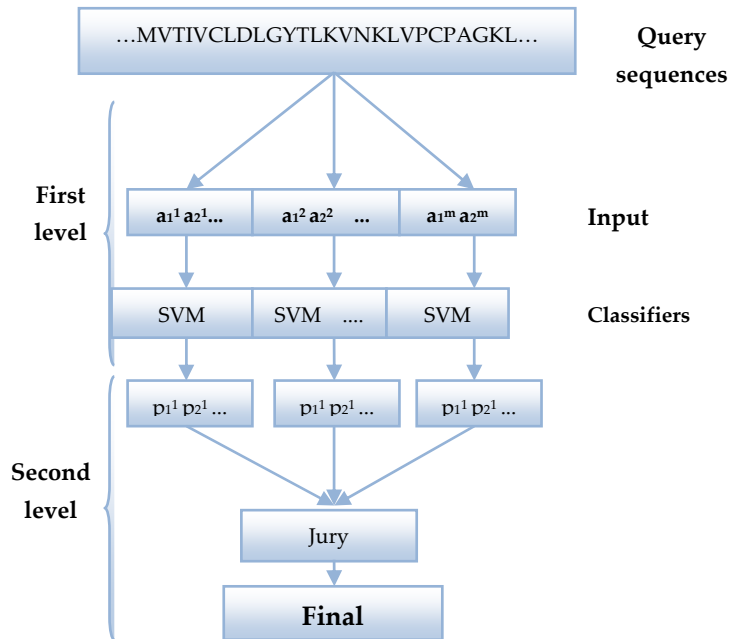


Figure 3: The general overview of CELLO procedure. The software starts with encoding the query sequence to computable codes as an input for the SVM classifiers in first level. Then it takes the votes and combine them in jury votes to get the final location. Classification in first level is composed of m number of SVMs which inputs to them are variety of vectors from different coding region ($a_1^1 a_2^1 \dots$) and other columns. Output from each classifier is probability distribution of L localization. The final classifier is the jury to create final probability distribution. Adapted from (Yu, Chen et al. 2006).

WoLFPSORT is the combination of WoLF and PSORT II (Gardy, Laird et al. 2005). In fact WoLFPSORT algorithm is the extension of PSORT II which works based on known protein sorting signals (Gardy, Laird et al. 2005). The main method behind the WoLFPSORT is training the sequence by a list of proteins with known localization and scoring them based on sorting signals, amino acid composition and functional motifs. At the end WoLFPSORT predicts the final localization based on simple k NN (k -nearest neighbor) for different classifiers. WoLFPSORT is a complementary algorithm for PSORT II to support fungi, animal and plant sequences. WoLFPSORT is also included correlative sequence features (Horton, Park et al. 2006). The pipeline of WoLFPSORT has been summarized in the Figure 4.

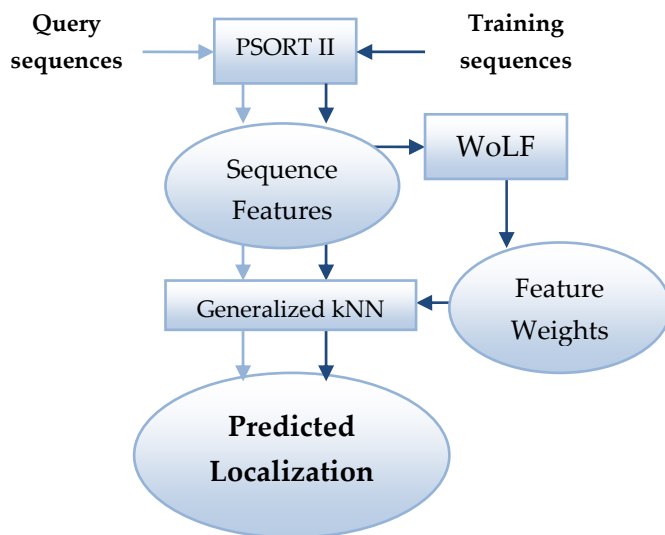


Figure 4: Pipeline for WoLF PSORT system. It is an extended version of PSORT II that has been adopted k -Nearest Neighbors algorithm for classification. In the figure ovals indicate quantities and rectangles represent procedures (Horton, Park et al. 2006).

According to Table 2 & Table 3, which compares different prediction performances for subcellular localization of proteins, CELLO has the high accuracy of 87% to assign the proteins in the right subcellular location. WoLF PSORT also has the highest sensitivity, specificity and Mathew's correlation coefficient among the other tested subcellular localization predictor.

Table 2: The comparison of subcellular localization predictor with different approaches for eukaryotes. It indicates CELLO is the significant among other predictors, all values are in percent.(Yu, Lin et al. 2004)

	CELLO	Reinhardt & Hubbard	Yuan	SubLoc
Cytoplasm	85.1	55	78.1	76.9
Extracellular	84.3	75	62.2	80
Mitochondrial	63.2	61	69.2	56.7
Nuclear	96	72	74.1	87.4
Accuracy	87	66	73	79.4

Table 3: Sensitivity, specificity and Mathew's correlation coefficient (MCC) for different subcellular localization predictors. it indicates the WoLFPSORT is the significant one among the predictors (**Klee and Sosa 2007**).

Parameter	BaCelLoa	SignalP3.0	HSLpred	MultiLoc	WoLFPSORT	PredSL
Sensitivity	0.60	0.60	0.67	0.78	0.85	0.69
Specificity	0.98	0.96	0.94	0.87	0.94	0.94
MCC	0.65	0.63	0.65	0.66	0.79	0.67

1.4 Project summary

As mentioned above, recently there have been attempts to automate the reconstruction of genome scale models based on functional gene annotation from gene sequence homology, mapping these genes for metabolic enzymes and generating of the network. However these techniques are mainly aimed at prokaryotic systems where there is no subcellular compartmentalization of reactions, such as model for SEED which is applied on different set of bacteria.

In this project, we used the scores from subcellular localization predictors, which were introduced above. Based on these algorithms two extreme solutions of the localization problem exist. One extreme solution would be to assign all reactions in one compartment, which lead to losing some specific features of metabolism of the organism (Figure5.a), so the model has low match to localization but high connectivity. In other extreme solution reaction are assigned to compartments based solely on the scores from the localization predictors (Figure5.b). This solution has a high match to localization but low connectivity.

The optimal solution is to assign reactions in a manner that is in good agreement with localization still maintains full-connectivity. Here we present a method for assigning subcellular localization to enzymatic reactions based on connectivity in an automated fashion. The algorithm aims at assigning localization in a manner that is consistent with signal peptide composition and physiochemical protein properties, while at the same time maintaining a well-connected and functional network (Figure5.c).

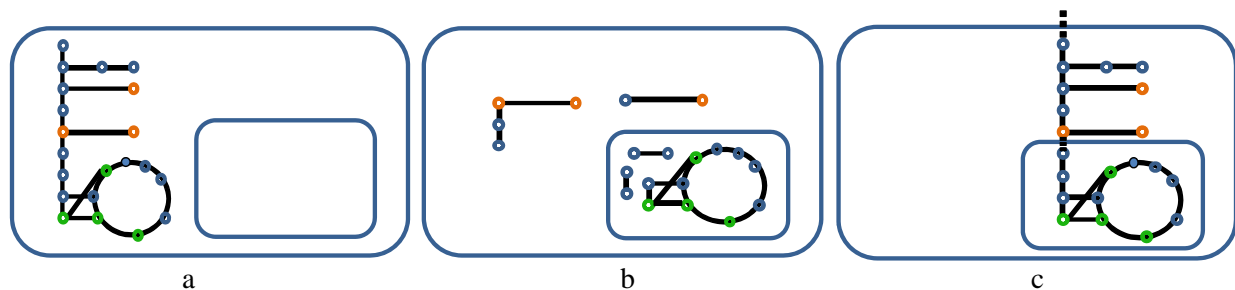


Figure 5: Possible solutions for assigning localization while at the same time maintaining a well-connected and functional model; a) One extreme solution would be that the reactions are assigned to compartments based only on predicted localization. This will correspond to a badly connected network. **b)** The other extreme is that all reactions are assigned to one compartment. This will correspond to a high connectivity but with a bad fitting to the predicted protein localizations. **c)** The purpose of the presented algorithm is to assign reactions to compartments in a way that is in good agreement with the predictions, while at the same time maintaining full-connectivity.

In following section, the technique and the algorithm is introduced in details. In addition integration of this algorithm in RAVEN toolbox is studied. Two eukaryotes, *Saccharomyces cerevisiae* and *Penicillium chrysogenum* have been chosen for evaluation of the algorithm and in next step a fully-connected compartmentalized genome-scale metabolic model for *Pichia stipitis* has been reconstructed as a case study while written algorithm for localization has been used to assign the location of reactions.

2 Method

2.1 Problems

To accomplish the aim of algorithm, it is crucial to determine the attributes of problem and select an optimization algorithm well-suited to the particular problem. One important issue is that the fitness landscape non-continuous. The variables in the problem are discrete, meaning that genes cannot be 50% present in one compartment and 50% in another compartment. The search space is very large; therefore there are many local optima for the solution that make it hard to find the optimal solution. In addition, the input from the localization predictors is noisy and of low quality.

2.2 Heuristic optimization

Heuristic methods are well-suited to deal with these kinds of problems. Two stochastic methods, genetic algorithm (GA) and simulated annealing (SA), which are mostly applied on laborious optimization problems were considered in this case (Kirkpatrick, Gelatt et al. 1983). SA was selected for this problem since SA is easy to implement with difficult formulation problems (Kohonen 1999). In addition SA can find good solutions by local moving in neighbors but not necessarily the global optimum. Since SA involves stochastics the solution can be changed every time. So in each iteration SA substitute the current solution by the previous one based on a global parameter which is called temperature (Fang 2000). The temperature is reduced based on introduced cooling function until no further changes happen, which is called freeze point (Fang 2000). In each step in cooling stage, there should be criteria to whether to keep or neglect the perturbation. It is crucial to set the initial temperature properly to ensure in each stage simulation takes enough time to reach thermal equilibrium and consequently reach the lowest energy state (Fang 2000).

2.3 Fully-connected Localization Assignment (F-LocA)

An algorithm for fully-connected localization assignment F-LocA was implemented in MATLAB. It is capable of importing metabolic models from output model of RAVEN Toolbox. F-LocA has two foci: 1) model compartmentalization, 2) model well-connectivity. The implementation maintains a fully connected model at all time. Overall the algorithm requires two inputs; a model structure and a scoring structure which can be achieved from CELLO and WoLFPSORT.

2.1.3 Preliminary steps

Two main inputs for the algorithm are generated in preliminary step, first is draft model which is generated by RAVEN Toolbox. Second is a scoring structure which is achieved by submitting the protein sequence of target organism in subcellular localization predictors (WoLF PSORT, CELLO). The predictors score the possibility of a given gene to exist in each of the compartments. In order to get a more reliable scoring, and to account for biases in each of the localization predictors, the average score from the two predictors was used.

The draft model for the target organism is merged to one compartment, which means all reactions are assigned to default compartment at the initiation of the optimization (Figure 6.a), which is derived by RAVEN Toolbox from KEGG database. The number of iterations in the optimization loop and transport cost are also the inputs for main function of algorithm. Transport costs were introduced in algorithm due to; lacks of transport information in KEGG, transport genes are not well-known and identification of them by homology is difficult.

2.1.4 Automated assigning based on full-connectivity and simulated annealing

The draft model is first expanded so that reactions with iso-enzymes are split in to several reactions where each of them is under the control of only one gene. The expanded model is then converted to an irreversible format, so all reversible reactions are split into two irreversible reactions. For all reactions without associated genes, dummy genes are introduced with equal scores for all compartments.

The algorithm always maintains the full-connectivity of network, which means the reactants of all reactions, can be synthesized. So if the template model has reactions that are not connected, and therefore nonfunctional, they are removed prior to optimization. According to number of compartments the model is enlarged, which means the number of metabolites and reactions are increased, although all reactions are still assigned to default compartment. One of the crucial issues in this concept is adjustment of transport reactions; here the transport cost function has been introduced. This function is a tradeoff between inferring several transport reactions without evidence and having a high localization score. With a very low transport cost all reactions would assigned according to their predictors scores.

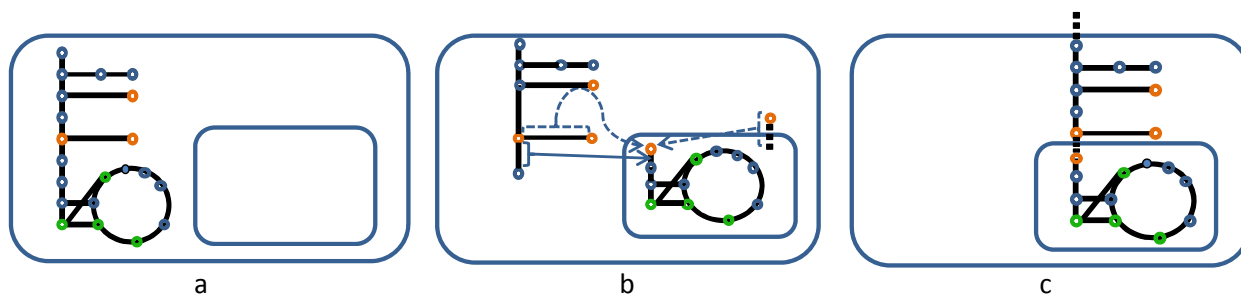


Figure 6: Overview of automated assigning based on full-connectivity and simulated annealing. a) All reactions are assigned to a default compartment at the initiation of the optimization. The algorithm depends on having a network where the reactants of all reactions can be synthesized, which means that the model should be fully connected. Any non-connected reactions are removed prior to the optimization. b) Based on current localization score; choose a gene that should be moved to another compartment. If the network is not fully connected, move reactions in recursive way until connected. If there is no way to connect a metabolite without significantly lowering the overall fitness, a transport reaction is included for that metabolite. c) The overall score is calculated based on agreement with predicted localizations and the number of transport reactions used. The simulated annealing algorithm initially allows for keeping bad solutions, but as the optimization proceeds only improvements in fitness are allowed.

After initiation, the optimization loop is started by randomly choosing a gene that should be moved to another compartment, based on current localization score. The target compartment is

then chosen randomly but weighted towards compartments with good scores for that gene (Figure 6.b). The gene and all corresponding reactions which the gene encodes are then moved to the new compartment. After assigning those reactions to the new compartment, it is possible that the model is unconnected. By monitoring the stoichiometric matrix the consumed and produced metabolites can be identified and from there the unconnected metabolites. All reactions that could reconnect the network are then identified. The cost of reconnecting the network by moving reactions from other compartments is calculated. If this cost is higher than the cost of introducing a transport reaction, the corresponding genes are moved. Otherwise a transport reaction is introduced. It is important to identify the current unconnected metabolites had been connected by transport reactions before, and also determine if metabolites are become unconnected in the original compartment. So iteratively other possible reactions are moved according to their gene and associated scores. For moving each gene, the transport reactions that are no longer necessary are tracked. Reactions are moved recursively until the network is fully connected (Figure 6.c). Based on predicted localization scores and number of transport reactions that has been added to model, the overall score is calculated. In this part simulated annealing algorithm is implemented, which at first allow the algorithm to keep all solutions and by preceding the optimization allows, it forces algorithm to keep good solutions. As initializing the simulated annealing the initial temperature is set to 5, and cooling function is done by linear decreasing, here the slope is 0.995. The overall approach of F-LocA has been summarized in Figure 7, which indicated the input, output and settings of algorithm.

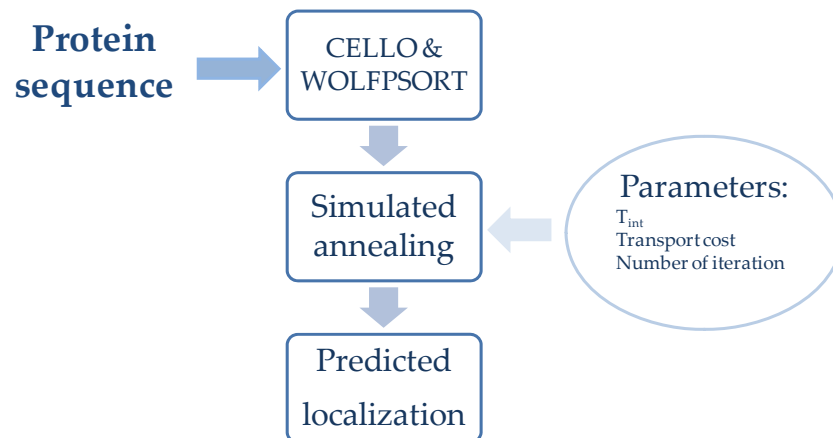


Figure 7: Overview of F-LocA process. In the first step the protein sequence is submitted to subcellular predictors, the output scores and simulated annealing assist the algorithm to find final predicted localization based on full-connectivity. The parameters are used to initialize the simulated annealing.

3 Result

F-LocA was evaluated by applying the algorithm to two models for eukaryotes. In addition F-LocA has been integrated to the RAVEN Toolbox as a complementary step to expand the functionality of RAVEN and enable automated reconstruction of a fully-connected compartmentalized model.

3.1 Evaluation

In order to verify the functionality of F-LocA, the technique has been carried out for two well-studied eukaryotes: *Saccharomyces cerevisiae* and *Penicillium chrysogenum*. The GSMM of these two eukaryotes have been reconstructed in our group and projects are still ongoing. By applying F-LocA new models were reconstructed and a comparison was carried out between existing models, the models generated by F-LocA and the output of the subcellular localization.

3.1.1 *Saccharomyces cerevisiae*

Based on the diversity of genome-scale models that have been reconstructed to date, *Saccharomyces cerevisiae* is favorite organism among researchers, where different versions have been published such as; iFF708 (Forster, Famili et al. 2003), iND750 (Duarte, Herrgard et al. 2004), iLL672 (Kuepfer, Sauer et al. 2005) and iIN800 (Nookaew, Jewett et al. 2008) . Hence, the required knowledge for evaluation of F-LocA is available for yeast, and the accuracy of compartmentalization in current models is believed to be high.

The protein sequence of yeast was submitted to subcellular localization predictors (CELLO & WoLFPSORT), to get the scores for existence of ORF in different compartments. These two scores were normalized.

The updated yeast model which is an ongoing project in our group (Table 4) has been used as draft model to F-LocA and existing model for further comparison.

Table 4: Structure of updated genome-scale model for *Saccharomyces cerevisiae* (unpublished data, ongoing project)

Model	Genes	Reactions	Metabolites	Compartments
Updated model (ongoing Project)	877 (16.66%)*	1873	1353	4

* Percentage of associated ORFs in the model to ORFs in the genome

All reactions in the draft model were merged in to the default compartment, here cytoplasm. F-LocA was applied by using the merged model and the localization scores from CELLO and WoLFPSORT. Here two compartments, cytoplasm and mitochondria have been considered for assigning the reactions. The highest score for each gene in other compartments is assigned to cytoplasm. By the end of the pipeline the full-connected compartmentalized model can be achieved. This model was then compared to the original model before merging it in default compartment and also the output of the subcellular localization predictors (Figure 8).

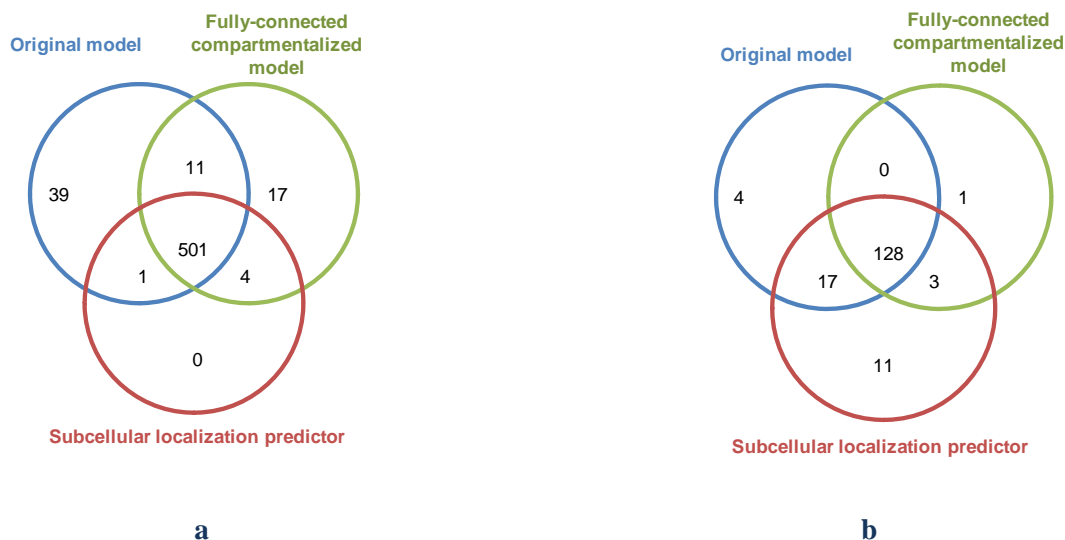


Figure 8: Overlapping of genes in yeast GSMM with respect to gene localization between original model, full-connected compartmentalized model and what was predicted using subcellular localization predictors. a) Number of genes in cytoplasm, b) In mitochondria.

As Figure 8.a indicates, 501 genes were overlapping between the original model, fully-connected compartmentalized model and predictors for cytoplasm. 11 genes were still agreed between full-connected and original model and 4 genes between full-connected model and predictor. 128 genes assigned to mitochondria are overlapping between the three models (Figure 8.b).

3.1.2 *Penicillium chrysogenum*

F-LocA was used to generate fully-connected compartmentalized model for *Penicillium chrysogenum*, by using the same procedure as for yeast. The arisen model from F-LocA was then compared to the GSMM that has been reconstructed in our group (Table 5) and also to the output of the subcellular localization predictors.

Table 5: Structure of updated genome-scale model for *Penicillium chrysogenum* (unpublished data; ongoing project)

Model	Genes	Reactions	Metabolites	Compartments
iAL1008	1008 (7.89%)*	1544	1320	4

Figure 9, indicates that F-LocA is capable of assigning almost all genes to the correct compartment while maintaining a fully connected model.

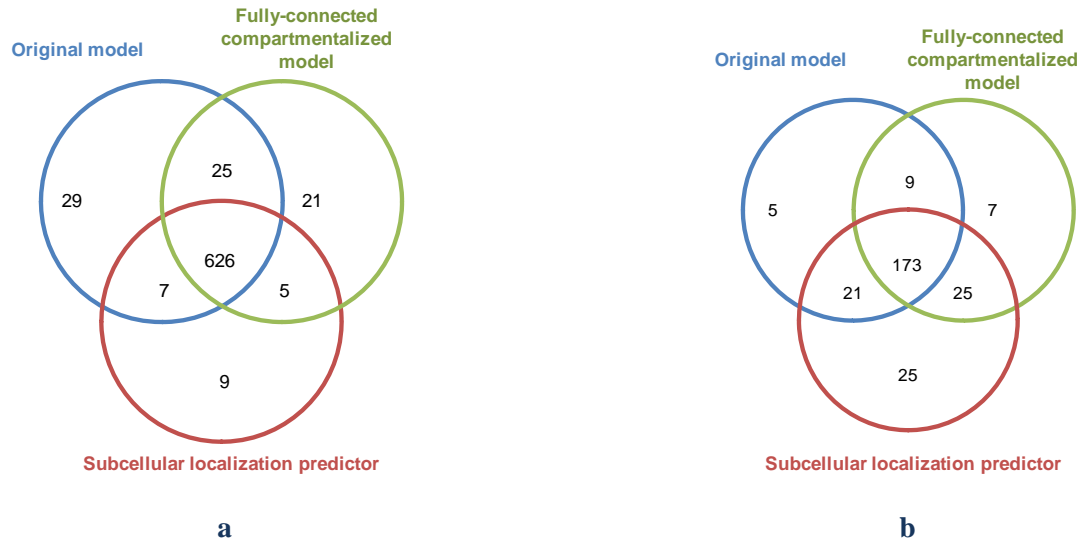


Figure 9: Overlapping of genes in *Penicillium chrysogenum* with respect to gene localization between original model, full-connected compartmentalized model and what was predicted using subcellular localization predictors. a) Number of genes in cytoplasm, b) In mitochondria.

3.2 Case study for *Pichia stipitis*

According to evaluations that have been carried out on the F-LocA, the method is capable of assigning the reactions in correct location and in the mean time return a well-connected network by introducing appropriate transport reactions. Hence, in the next step it has been tried to take advantage of F-LocA in RAVEN Toolbox. The scaffold of this procedure has been indicated in Figure 10. In summary, RAVEN Toolbox trains hidden Markov models to represent metabolic enzymes based on KEGG and HMMER. The protein sequences of the target organism are then matched to this model to create an initial reaction list to generate a metabolic model for target organism. This draft model can be used as input to F-LocA, which is carried out by importing the subcellular localization prediction scores for protein sequence of target organism. At the end of the pipeline the fully connected and compartmentalized model is generated (Figure 10).

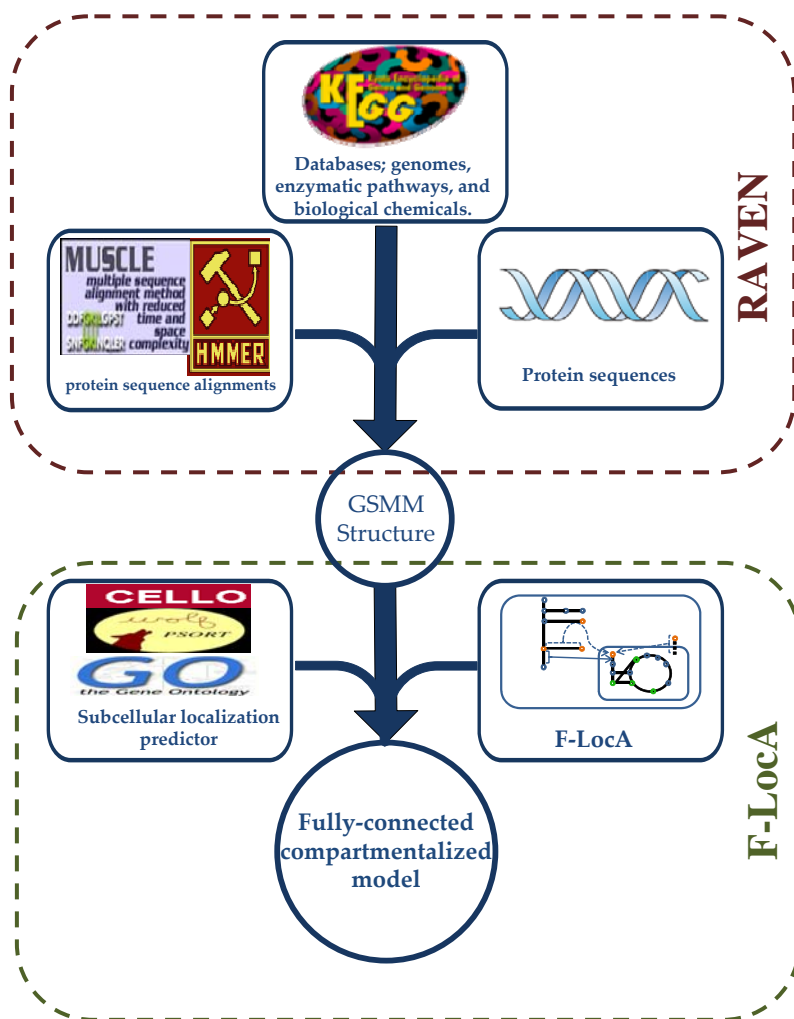


Figure 10: Integration of F-LocA in RAVEN Toolbox. In the beginning based on protein sequence, the HMMER is implemented which use “profile hidden Markov models” to investigate homolog of sequence and aligning the protein sequence. Then based on this genome annotation and databases for genomes, enzymatic pathways and biochemical reactions a draft reconstruction is built. In this step the information about compartmentalization is added to model from subcellular localization predictors. Further on F-LocA is tried to assign localization in a manner that is consistent with signal peptide composition and physiochemical protein properties, while at the same time maintain a well-connected and functional network. At the end of the pipeline, full-connected compartmentalized model is a final output.

Here we applied the described pipeline for *Pichia stipitis* which is capable to ferment in aerobic and anaerobic condition and also converting xylose to ethanol in direct fermentation (Jeffries, Grigoriev et al. 2007). The draft model for *Pichia stipitis* was generated automatically based on the KEGG database and protein sequences of *Pichia stipitis* using RAVEN (Table 6). An E-values of 10^{-50} was used as a cut off value to match the genes and associated enzyme function.

Table 6: Structure of draft models for *Pichia stipitis*

Model	Genes	Reactions	Metabolites
Draft model based on KEGG	902 (15.51%)*	1269	1423

* Percentage of associated ORFs in the model to ORFs in the genome

The properties of exported model are summarized in Table 7. The number of genes, reactions and metabolites has been decreased due to the full connectivity requirement in F-LocA since unconnected reactions are removed prior to optimization.

Table 7: Structure of fully-connected compartmentalized model for *Pichia stipitis*

Model	Genes	Reactions	Metabolites
Model based on F-LocA	688 (11.83%)*	981	695

* Percentage of associated ORFs in the model to ORFs in the genome

Figure 11 indicates the agreement of gene localization between fully-connected compartmentalized model and predictors in cytoplasm and mitochondria. The case study shows that the F-LocA can be easily integrated in RAVEN Toolbox for full-connectivity and compartmentalization of model, therefore accelerate the whole process of reconstruction. The fully-connected compartmentalized model for *Pichia stipitis* was entered to manual curation step in reconstruction. After improving the model during different step of reconstruction, and achieving the finalized model, the physiological capabilities of *P.stipitis* and *P. pastoris* (unpublished data) are currently being studied.

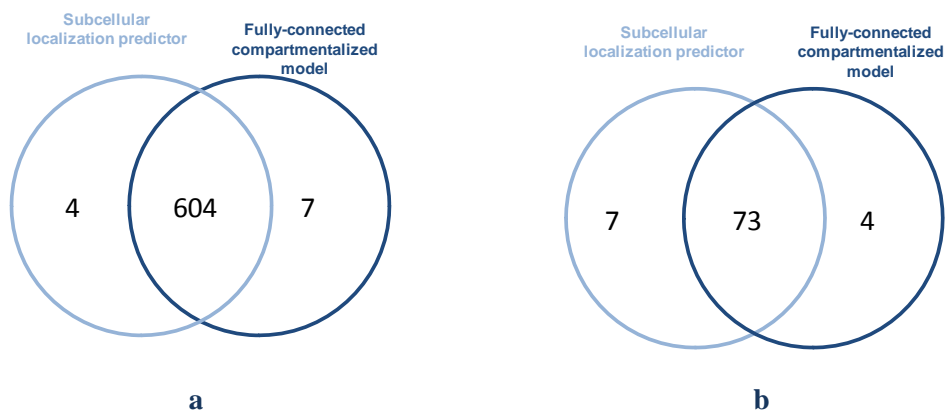


Figure 11: Overlapping of genes in *Pichia stipitis* with respect to gene localization between full-connected compartmentalized model and what was predicted using subcellular localization predictors a) Number of genes in cytoplasm b) In mitochondria.

4 Discussion

The subcellular localization is a one of the most time consuming task in the genome-scale metabolic network reconstructions. After evaluating the F-LocA and carrying out the case study, we believe that F-LocA can significantly speed up the task of model reconstruction for eukaryotic organisms. Also F-LocA has been incorporated to the RAVEN Toolbox, which leads to expansion of RAVEN and consequently the functionality of RAVEN.

We considered some assumptions for simplify the problem before generating the algorithm. One simplification is the gene products can be presented only in one compartment at the same time. This assumption constrains the problem to a smaller dimension and when it comes to metabolism such enzymes are normally present as iso-enzymes. The second simplification is that, all transport reactions are facilitated by diffusion, and sym/anti porters are not considered. The third simplification concerns reversible reactions, which are split in to two separate reactions. This might lead to false connectivity in some cases, because the metabolites can be determined as connected whereas the metabolite in reality is connected via the reverse reaction.

One issue in F-LocA is low accuracy of prediction in subcellular localization for some compartments, which leads to some difficulties in assigning proteins to. As an example the accuracy of assigning peroxisomal genes is very low in CELLO and WoLFPSORT, which translates to poor prediction ability in F-LocA.

Overall, F-LocA is unique technique to solve the issues of compartmentalization and connectivity in genome-scale metabolic reconstruction. The introduced algorithm was generated based on full-connectivity approach, so it is possible for F-LocA to return a functional model.

Acknowledgment

This project was done in the Systems and Synthetic Biology group at the Department of Chemical and Biological Engineering at Chalmers University of Technology under the supervisions of Prof. Jens Nielsen and PhD student Rasmus Agren.

I am very grateful to Professor Jens Nielsen to give me the opportunity to work in his lab and for all his support, and also Rasmus Agren for all his help, comments, discussion and support during my project.

I would like to thank, Intawat Nookaew, Sergio Bordel Velasco, Luis Caspeta, Siavash Partow, Payam Ghiaci, Fredrik Karlsson, Tobias Österlund, Juan Octavio Valle Rodriguez, António Roldão, Wanwipa Vongsangnak, Dina Petranovic Nielsen, Martina Butorac, Erica Dahlin, Pegah Khorramzadeh, Amir Feizi and Antje Berger for their discussions and support.

5 References

- Baerenfaller, K., J. Grossmann, et al. (2008). "Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics." *Science* **320**(5878): 938-941.
- Becker, S. A., A. M. Feist, et al. (2007). "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." *Nat Protoc* **2**(3): 727-738.
- Borodina, I. and J. Nielsen (2005). "From genomes to in silico cells via metabolic networks." *Current Opinion in Biotechnology* **16**(3): 350-355.
- Christie, K. R., S. Weng, et al. (2004). "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms." *Nucleic Acids Research* **32**(Database issue): D311-314.
- Covert, M. W., C. H. Schilling, et al. (2001). "Metabolic modeling of microbial strains in silico." *Trends Biochem Sci* **26**(3): 179-186.
- David, H., G. Hofmann, et al. (2006). "Metabolic network driven analysis of genome-wide transcription data from Aspergillus nidulans." *Genome Biology* **7**(11): -.
- DeJongh, M., K. Formsma, et al. (2007). "Toward the automated generation of genome-scale metabolic networks in the SEED." *BMC Bioinformatics* **8**: 139.
- Duarte, N. C., M. J. Herrgard, et al. (2004). "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model." *Genome Research* **14**(7): 1298-1309.
- Fang, M. (2000). Layout Optimization for Point-to-Multi-point Wireless Optical Networks via Simulated Annealing & Genetic Algorithm
- Feist, A. M., M. J. Herrgard, et al. (2009). "Reconstruction of biochemical networks in microorganisms." *Nature Reviews Microbiology* **7**(2): 129-143.
- Feist, A. M., J. C. M. Scholten, et al. (2006). "Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri." *Molecular Systems Biology*: -.
- Forster, J., I. Famili, et al. (2003). "Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network." *Genome Research* **13**(2): 244-253.
- Francke, C., R. J. Siezen, et al. (2005). "Reconstructing the metabolic network of a bacterium from its genome." *Trends in Microbiology* **13**(11): 550-558.
- Gardy, J. L., M. R. Laird, et al. (2005). "PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis." *Bioinformatics* **21**(5): 617-623.
- Henry, C. S., M. DeJongh, et al. (2010). "High-throughput generation, optimization and analysis of genome-scale metabolic models." *Nature Biotechnology* **28**(9): 977-U922.
- Horton, P., K. J. Park, et al. (2006). "Protein subcellular localization prediction with WOLF PSORT." *Proceedings of the 4th Asia-Pacific Bioinformatics Conference* **3**: 39-48
- 363.
- International Union of Biochemistry and Molecular Biology. Nomenclature Committee. and E. C. Webb (1992). *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. San Diego, Published for the International Union of Biochemistry and Molecular Biology by Academic Press.
- Jeffries, T. W., I. V. Grigoriev, et al. (2007). "Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis." *Nature Biotechnology* **25**(3): 319-326.
- Kanehisa, M. and S. Goto (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* **28**(1): 27-30.
- Karp, P. D., I. M. Keseler, et al. (2007). "Multidimensional annotation of the Escherichia coli K-12 genome." *Nucleic Acids Research* **35**(22): 7577-7590.
- Kirkpatrick, S., C. D. Gelatt, et al. (1983). "Optimization by Simulated Annealing." *Science* **220**(4598): 671-680.

- Klee, E. W. and C. P. Sosa (2007). "Computational classification of classically secreted proteins." Drug Discov Today **12**(5-6): 234-240.
- Kohonen, J. (1999). A brief comparison of simulated annealing and genetic algorithm approaches. Helsinki, University of Helsinki-Department of Computer Science: 4.
- Krieger, C. J., P. F. Zhang, et al. (2004). "MetaCyc: a multiorganism database of metabolic pathways and enzymes." Nucleic Acids Research **32**: D438-D442.
- Kuepfer, L., U. Sauer, et al. (2005). "Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*." Genome Research **15**(10): 1421-1430.
- Lu, Z., D. Szafron, et al. (2004). "Predicting subcellular localization of proteins using machine-learned classifiers." Bioinformatics **20**(4): 547-556.
- Maglott, D., J. Ostell, et al. (2007). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Research **35**: D26-D31.
- Markowitz, V. M., F. Korzeniewski, et al. (2006). "The integrated microbial genomes (IMG) system." Nucleic Acids Research **34**: D344-D348.
- Nakai, K. and P. Horton (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization." Trends Biochem Sci **24**(1): 34-36.
- Nookaew, I., M. C. Jewett, et al. (2008). "The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism." Bmc Systems Biology **2**: -.
- Oberhardt, M. A., B. O. Palsson, et al. (2009). "Applications of genome-scale metabolic reconstructions." Molecular Systems Biology **5**: -.
- Palsson, B. (2006). Systems biology : properties of reconstructed networks. Cambridge ; New York, Cambridge University Press.
- Patil, K. R., M. Akesson, et al. (2004). "Use of genome-scale microbial models for metabolic engineering." Current Opinion in Biotechnology **15**(1): 64-69.
- Quek, L. E. and L. K. Nielsen (2008). "On the Reconstruction of the *Mus Musculus* Genome-Scale Metabolic Network Model." Genome Informatics 2008, Vol 21 **21**: 89-100
- 236.
- Reed, J., T. Vo, et al. (2003). "An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)." Genome Biology **4**(9): R54.
- Reed, J. L. and B. O. Palsson (2004). "Genome-scale in silico models of *E-coli* have multiple equivalent phenotypic states: Assessment of correlated reaction subsets that comprise network states." Genome Research **14**(9): 1797-1805.
- Schomburg, I., A. Chang, et al. (2004). "BRENDA, the enzyme database: updates and major new developments." Nucleic Acids Research **32**: D431-D433.
- Thiele, I. and B. O. Palsson (2010). "A protocol for generating a high-quality genome-scale metabolic reconstruction." Nature Protocols **5**(1): 93-121.
- Yu, C.-S., Y.-C. Chen, et al. (2006). "Prediction of protein subcellular localization." Proteins: Structure, Function, and Bioinformatics **64**(3): 643-651.
- Yu, C. S., C. J. Lin, et al. (2004). "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions." Protein Sci **13**(5): 1402-1406.