Report on structural comparison and validation of the genome scale metabolic models for Saccharomyces cerevisiae

Lu, C.; Whelan, KE.; King, RD.

Publication date:
2010

Citation for published version (APA):
Lu, C., Whelan, KE., & King, RD. (2010). *Report on structural comparison and validation of the genome scale metabolic models for Saccharomyces cerevisiae*. European Commission. http://hdl.handle.net/2160/4635

Report on structural comparison and validation of the genome scale metabolic models for Saccharomyces cerevisiae

C. Lu, KE. Whelan, RD. King

Department of Computer Science, Aberystwyth University, Wales, UK

{cul, knw, rdk}@aber.ac.be
April 2009

1 Introduction

One of the objectives in the Unicellsys project is to develop computational tools to automatically modify the metabolic pathway models based on the experimental data and the bioinformatics information available. The machine learning tool could be developed within the framework of constraint-based optimization and/or inductive logic programming (ILP). Meanwhile a good starting point for the yeast metabolic pathway model is important for the quality of the modified model.

There are a couple of existing genome scale reconstructed metabolic models for Saccharomyces cerevisiae from different research groups and of different revisions. Here we focus on the models from the latest reconstruction efforts, which include the Aber model [2, 3], the iIN800 model [7] and the consensus model [1].

The Aber model, which is a logical model for the yeast metabolism, has been constructed by RobotScientist's group at Aberystwyth University, UK. The iIN800 model came out of an international joint work and was constructed as a scaffold to query the lipid metabolism. The consensus model came from a latest curation effort by YSBN consortium in 'Jamboree' style, which was mainly conducted at the University of Manchester, UK.

In this work, we initially compared the structural information between the Aber model and the consensus model, and tried to unify the two if possible. Logical model simulation for single and double gene deletion were performed based on the two network models and results were validated by the existing experimental data. Furthermore, flux balance analyses have been performed utilizing the consensus model. Flux range analysis has been used to identify the network gaps including blocked reactions and dead-end metabolites.

Relevant work

Other relevant work out of the collaboration of this project is as follows.

1. A detailed manual checking of the reactions in the Aber model have been done with emphasis on comparison to the iIN800 model in lipid metabolism.

2. A detailed manual curation effort has been made to identify a list of dead-end metabolites in iIN800 model together with model revision proposals.

3. A manual cross-check between the dead-end metabolites identified in the consensus model and iIN800 model has also been done. It was observed that the two models actually share many dead-end metabolites, whereas one model can be complementary to another in some parts. The above work has been done by Dr Pinar Pir at University of Cambridge.

4. There is an ongoing work in the University of Manchester, where the lipid metabolism in the consensus model has been gradually filled in with added information from the iIN800 model and other sources, and the whole consensus model has been under continuous curation.

1.1 Model origins

The three models share similar roots in the model construction. The Aber logical model was constructed from iFF708, an existing flux balance analysis model [5], with additional information from KEGG database. The consensus models use two separatedly developed metabolic networks as the starting point, namely iMM904 and iLL672. Moreover, iLL672 was derived fromm iFF708 with extensively curation in order to improve the ability of the corresponding FBA model in gene deletion phenotype prediction. The iMM904 model was constructed from iND750 [6] with further curation and extension.

The iIN800 model was originated from iFF708 and has been expanded with a much more detail description of lipid metabolism using databases (such as KEGG and SGD) and literatures. It also provides an improved biomass equation under different growth conditions for FBA simulation.

1.2 Difficulties in structural comparison

Both models have been formulated in to exchangeable format: the aber model in prolog files and the consensus model in SBML. Thus it is not difficult to extract the chemical entities and reactions and their annotations in the models. In the Aber model the majority of the compounds are KEGG compounds, but some have only an abbreviation and a name. The consensus model has already been unified to a certain level in terms of annotation of chemical entities and reactions, though a few compounds have only names and formulas. There are still potential difficulties in matching between the two models.

- No standard nomenclature is followed for various biochemical entities, especially for metabolites.
- No standard way of representing the chemical reaction equations. The inclusion of ubiquitous metabolites such as H2O, H+ are not uniform within the model or across models.
- Errors and inconsistencies are to be expected.

2 Methods

To compare the structure information in the models, we first tried to match the chemical entities. For metabolites in consensus model, a KEGG compound ID has been given if possible, either by taking directly the ChEBI DB source information or through synonyms and chemical formula matching. Special characters in names or synonyms have been handled as well before matching. Then the common unique chemical reactions without compartment information have been searched by matching both the substrate reactants and the product reactants.

We also compared the reactions by checking the enzymes and the genes that encode the enzymes. As the aber model has no information of multi-enzyme complexes, the comparison with the consensus model in iso-enzyme will be problematic. The annotated EC code for the reactions have also been compared.

The assignment of pathway names in the consensus model is somehow arbitrary, it turns out difficult to compare the pathways involved directly for the two models. The list of possible iso-enzymes for each reaction in both models have been extracted and compared.

We also try to map all the chemical species to a common pathway database with ontology and visualization support, i.e. YeastCyc database (or the yeast pathway database in MetaCyc). A quick and dirty overview as well as the detail pathway information of the differences between the two models will be available by mapping the species from two models to MetaCyc. However due to similar problems of naming convention in species and the incompleteness of the model, there might be quite a lot of species and reactions not being able to found in YeastCyc database.

In a logical model simulation, growth/lethality prediction is basically equivalent to finding a path in the metabolic graph from a set of initial compounds (mimicking the growth medium) to a set of essential compounds for cell growth/division. Logical models were built for both models, simulations for logical models were run for prediction of yeast growth (viable or inviable) for all single gene deletants and double gene deletants. Prediction results were validated with experimental data from literature.

Flux balance analysis (FBA) is a constraint-based approach using linear programming to identify a flux distribution that optimize the given objective function (such as maximization of the flux for biomass formation, that is the growth rate) [6]. FBA applies mass and energy balance constraints to model

steady state behavior of reconstructed networks. Further utilized constraints involve irreversibility of reactions and the maximum flux through any reaction or transporter. The network of reactions is defined by a stoichiometric matrix S, whose element represents the stoichiometry of metabolite in the reactions. Exchange fluxes for boundary metabolites that are allowed to be in and out of the cell boundary are also incorporated in the S matrix. The steady state constraints for FBA is thus expressed as: $Sv = 0$ and $v\mathrm{min_i} \leq \mathrm{v_i} \leq \mathrm{vmax_i}$, $i = 1, ..., N$ where v is the vector of N reaction fluxes. $v\mathrm{min_i}$ and $v\mathrm{max_i}$ correspond to the lower and upper bound for individual flux v_i.

In order to systematically validate the consensus model structure, we constructed an FBA model based on the consensus stoichiometric information while using similar biomass composition borrowed from iIN800 model.

3 Results

3.1 Topological comparison

Matching of compound list

All reactants and modifiers (catalyzers) in the consensus model have been given an ID specific to the model, each was annotated by specific information, including name/synoyms, compound ID from standard database such as ChEBI, KEGG, PubChem or HMDB, and chemical structures in form of InChI or SMILES if possible. Some metabolites have only names and empirical formulas linked with.

Reactants of the aber model have been given a KEGG compound ID or (when a KEGG ID is not possible) a unique ID of the abbreviation of the metabolite name (originated from the iFF708. All compounds were provided a name and/or synonyms.

Out of total 1168 metabolites, there are 931 unique ones in the consensus model with compartment information taken into account. Whereas, after removing compartment information, out of total 1092 metabolites, 663 ones are unique and, 616 (92.9%) out of the 663 unique metabolites could be referenced by a KEGG compound ID. This has been done by either using original KEGG ID provided by the model, checking links to ChEBI compound or simply matching between name/synonyms and formulas).

Aber model contains 820 (or 810 after excluding transport reactions) unique metabolites among which 761 (92.8%) have a KEGG compound ID. Both models share 482 KEGG compounds. With further matching with names only, 3 additional matched non-KEGG compounds were found. The compound matching ended up with 485 shared compounds for the two models (73.2% for the consensus model and 63% for the aber model).

Note that, one reason for the poor shareness in the two models even if they could both highly annotated to KEGG ID might be the different choice of the metabolite states and focus on different subpathways. For some individual

molecules with multiple states (e.g. because of acid-base reactions), the consensus model attempts to use the chemical entities believed to be most common at the pH of the relevant compartment. However, in this version of consensus construction, all species are assumed to be in the form that corresponds to the most common protonation state at pH 7.2. The metabolites were annotated with a database entry with the correct protonation state, but in a number of cases, the databases only contained the metabolite in a neutral form or otherwise in an incorrect protonation state. However, it is not clear about the chemical entities used in aber model (essentially KEGG and iFF708).

Compartment information

Both models contains compartment information, however, 15 compartments have been used in consensus model while only 4 in aber model. It is not directly comparable if we compare the reactions combined with compartment information. Therefore, in our comparison, the compartment information has been ignored.

Matching of reactions

In the consensus model, formation of of protein complex for catalyzing has been represented in as reactions without modifiers. Enzymes or genes were annotated by references to SGD and UniProt. Each reaction was given a name and was assigned to a pathway names (with a certain arbitrariness). 738 and 478 unique transformation in the networks were annotated with EC number and PubMed references. All reactions are set to be reversible.

This is still difficult to have any perfect match without proper adjustment of the reaction equations. Without change of reactants, 150 reactions have perfect match, which means reactions in both models share the same sets of substrates and products; 70 of these reactions are transport reactions.

For each reaction in the consensus model, all of its cofactors, i.e. currency metabolites such as ATP, NADH and CoA, have been included. Reactions with Markush structures or ambiguities have been removed in consensus model. Thus the lipids are under-represented in the consensus model. However in aber model, some currency metabolites have been ignored from the reactions, such as water and H+ which has much lower connectivity than the ones in the consensus model.

Matching of catalyzers and reaction annotation

Each reactions for both aber model or consensus model was annotated with a unique EC code if possible. The 96 reactions for protein complex formation in consensus model were not considered here.

In consensus model 960 out of 1761 reactions have been annotated with an EC code. By borrowing information from the enzyme, 1054 reactions could be annotated with an EC code (many with top level classes). And each reaction has

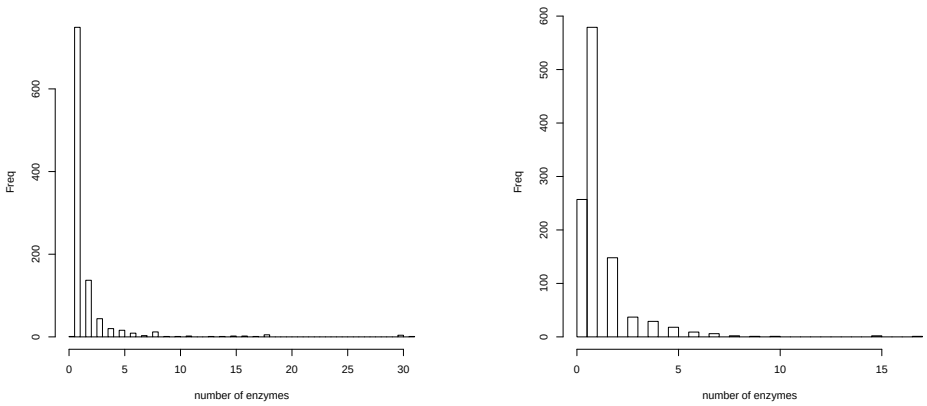

Figure 1: Histogram for the number of iso-enzymes in aber model (left) and consensus model (right) respectively.

been provided either a single gene enzyme, an enzyme complex, or no catalyzer at all. Ignoring the compartment information, there are in total 1090 unique chemical reactions for the consensus model, and 833 (76.4%) of which were catalyzed by at least one catalyzer (either an single gene enzyme or a protein complex). This consensus model doesn't provide annotation for the unknown enzymes.

In aber model, 1515 out of 1894 reactions have been annotated with an EC code according to the catalyzers in the reaction. There are in total 1012 unique chemical reactions. Almost all unique reactions have been annotated with at least one enzyme (except for one). In total 1146 unique enzymes have been involved as catalyzers and 920 of which are known yeast ORF and 226 are unknown. The unknown enzymes were given a name starting with 'U' and end up with '_'. Each enzyme has been given a gene name, EC code or description if possible.

However, as the aber model contains no information about enzyme complexes, we would compare only the individual ORFs. There in total 833 individual yeast genes involved as catalyzers, if protein complexes are considered as unique enzymes the number of unique enzymes goes down to 771. The two models share 659 yeast genes without taking into account enzyme complexes information. 261 of enzymes with known genes in aber model are not included in the consensus model; while 174 yeast genes involved in the consensus model are not included in aber model.

See Fig. 1 for comparison of the number of iso-Enzymes in the consensus model and in the aber model. Due to lack of muti-enzyme complexes information, the number of iso-enzymes won't be correct for the aber model.

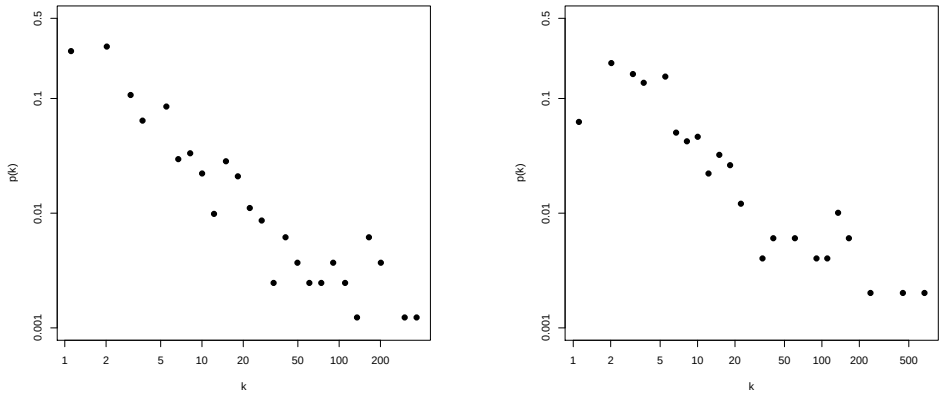

Figure 2: Connection degree of metabolites in aber model (left) and consensus model (right) respectively.

3.1.1 Comparison of metabolite connectivity

We have compared the connection degree (the number of reactions that a metabolite participates in) between the two models (Table 1). Fig. 2 shows the log-log plot of the degree distribution for both models. It is not difficult to observe that ubiquitous metabolites such as water and H+ have much lower connectivity in aber model than in the consensus model. Therefore, it is important to include the ubiquitous metabolites rather than ignoring them.

3.2 Logical model simulation for gene deletion study

Setting of starting compounds and essential products

The logical model simulation follows the same mechanism as described in [**?**]. Similar settings have been used for the logical Aber model as in [**?**] with some adaptions in case a certain compounds are not covered in the consensus model.

The growth status of the simulated experiment is determined as follows:

- Continued growth (viable) is predicted iff all essential compounds are present in the Cytosol compartment of the model.
- Retarded growth (inviable) is predicted if any essential compound missing from the Cytosol compartment.

The starting compound set mimics the components in a minimal medium (MMD+ura+hist+leu), which contains the minimum sets of compounds required for wild type growth as well as additional nutritional requirements for uracil, histidine and leucine. Table 2 lists the starting compounds used for the Aber and consensus model.

The essential compound set consists of the following five classes: amino acid, nucleic acid, polysacharide, membrane and intermediate as listed in Table 3.

Gene essentiality prediction

A gene is predicted to be essential if a retarded growth is predicted by the logical model simulation under a defined mimimum medium condition. Giaever et al provided the wet lab gene essentiality data for YPD medium [4]. The sensitivity of the model to detect essential genes are our main evaluation measure here. See Table 4 for the results of the model prediction for single gene deletion study.

Yeast growth prediction under mimimum medium

The same simulation results from the two models have also been verified with the experimental data obtained under the defined medium (MMD+ura+hist+leu)[4].

Two different cases of gene deletion sensitivity were considered, 1) genes found to be significantly sensitive after 5 generations and 2) genes found to be significantly sensitive after 5 generations that remain significantly sensitive after 15 generations. The 1106 essential genes (in YPD) were also added to the experimental data to increase the coverage of verifiable ORFs in the models. The two different viability criteria resulted into two different experimental data sets, namely set A (74 inviable vs 4669 viable outcome) and set B (19 inviable vs 4724 viable). The third experimental data set C (122 inviable and 4602 viable) was derived basically from set B, except that in case of viable outcome in set B, a more relaxed criteria is used to decide whether the gene deletion strain is 'invialble' or not [6][2].

The performance of the models were evaluated by checking the number of True Positive (TP), False Negative (FP), True Negative (TN), False Negative (FN), the proportion of majority class, the accuracy, sensitivity, specificity, Positvie Predictive Rate (PPV) and Negative Prediction Rate (NPV). The case with 'inviable' out is referred to as positive event and 'viable' as negative event. The false negative cases are of great interest which might deserve further study to in order to improve the model. For automatic model refinement, the false predictions are also of the focus for generating model refinement hypotheses. For each model, six sets of results were presented for use of different experimental sets (A, B or C) and for different sets of validated ORFs (depended on whether verifiable by the experimental data and shared within the two models).

Simulation for double gene deletion

Simulations have also been done on gene double deletion in order to find higher order essential genes. All possible pairs of genes in the model (excluding essential single genes predicted by the model) have been checked and a list of inviable double deletions were obtained. There are 60 gene pairs predicted to be essential by the Aber model and 29 essential gene pairs by the consensus model. These predicted lethal gene pairs needs further proper validation either based on literature or further experimental tests.

3.3 Flux balance analysis on the consensus model

To construct a flux balance model using the consensus model, a biomass formation reaction was added to the model. Additionally all extracellular metabolites were set to be boundary metabolites and their relevant exchange reactions were added in as well. The biomass formation reaction was taken directly from iIN800 model and only the one under carbon limit condition has been considered here. Some of the lipid components that are not present in the consensus model have been removed from this biomass formation reaction. Table 6 has listed the biomass composition used for the consensus model and the iIN800 model.

The maximum uptake rates for the exchange metabolites for the minimal media (MM) was adapted from Snitkin et al. A quick FBA validation of the the consensus model under minimal media has be conducted using the same sets of experimental data (gene essentiality data and yeast growth data). Similar performance has been obtained for the FBA model as for the logical model.

Next, the structural gaps within networks of the consensus model have been identified using flux range analysis (or flux variability analysis). The network gaps are mainly caused by the two connected problems: 1) blocked reactions that are unable to carry any fluxes; and 2) problem metabolites that are either non-producible or non-consumable. For this flux range analysis, we used the unique reactions only, i.e., different reactions of the same chemical transformation but catalyzed by different iso-enzymes have been combined to a unique reaction. And only the structural constraints have been used, i.e., the constraints for the media condition (maximum uptake rates) have been ignored and the minimum growth rate was set to zero. The list of blocked reactions were first identified by flux range analysis, from which the un-producible or un-consumable metabolites were then derived.

From this analysis, 277 (19%) reactions are found blocked and 328 (24%) metabolites are identified as dead-end. In cytoplasm alone, there are 156 (26%) problem metabolites and most of them (123 out of 156) are isolated metabolites, which are involved in only one unique reaction.

Based on flux range analysis, we have identified a list of network gaps within the reconstructed consensus network model, filling in these gaps should improve the model at least in terms of more accurate description of the cell growth metabolism and should potentially increase the predictive power of the model. The main mechanisms for restoring the connectivity within the network include: 1) reaction reversing; 2) adding reactions from other sources such as reference models; 3) adding external/intracellular transport reactions. And this could be done in a systematic and automatic way within the constrain-based optimization framework by minimizing the metabolic model modification [8, 9]. Automatic gap filling will generate a list of hypotheses which can be checked manually through literature or further tested experimentally. Meanwhile, manual modification can also been done to restore the network connectivity, which might be more accurate and with higher confidence. In summary, an iterative semi-automatic approaches would be suitable for the network connectivity restoration problem.

A computational tool for basic flux balance analysis and structural validity checking including network gap finding and filling have been implemented in Python, utilizing the open source third-party solvers such as OpenOpt and lpSolve. Further development and testing is still needed for the computational tool for automatic network gap filling and model refinement based on the experimental data within the constraint-based optimization framework.

4 Concluding remarks

The consensus model is more accurate and less ambiguous in terms of annotation of the species and reactions and inclusion of reactions and enzymes. And the uncertain enzymes and reactions in the ancestor model have been excluded from the consensus model. It might be a good starting point for our future model refinement. Model refinement could be realized by incrementally adding more reactions or enzymes or encoded genes into the model based on the background information and experimental data. Apart from the base model, a set of additional background information need to be compiled from various biochemical data resources. The Aber model has provided an example of model and background information formalization and a mechanism of a more stable and faster simulation, which might suit for abductive / inductive reasoning in the logic programming framework.

A computational tool should be developed in the future for integration of constraint-based optimization and probabilistic logic programming methods to refine the model using various sources of data and information. In general, the model refinement procedure should contain iterative semi-automatic cycles from hypotheses generation, experimental testing (e.g. by employment of Robot Scientist "Adam" [3]), automatic model modification to manual curation and verification.

The model simulation results show that by removing some uncertain parts in the pathway model (as for consensus model) did improve the overall accuracy of the prediction however decrease the detection rate for the gene essentiality of the model . Further detailed comparison between the models can be done by focusing on the sub pathways that the two models differ and checking the functional distribution of the predicted essential genes/gene pairs.

Moreover, ontology support for the reactions and pathways are still missing for the consensus model. One of the future work will be to collect reaction/pathway ontology information as well as evidence information for the background information needed for hypothesis generation in automatic model improvement procedure. Also it would be of interest to have a comparison of the consensus model with the yeast pathway model in MetaCyc, in which the ontology information is available for all metabolites/enzymes/proteins and reaction/pathways.

References

[1] Herrgard Markus J., Swainston Neil, et al., Consensus model manuscript. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nature Biotechnol. 2008, 26, 1155-1160.

[2] Whelan, K. E. and King, R. D., Using a logical model to predict the growth of yeast. BMC Bioinformatics 2008, 9:97.

[3] King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E., Clare, A. The Automation of Science. Science, 2009, 324(5923):85-89.

[4] Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature, 2002, 418: 387-391.

[5] Forster J, Famili I, Fu P, Palsson BO:, Nielsen J. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic, Network. Genome Research, 2003, 13:244-253.

[6] Duarte NC, Herrgard MJ, Pallson BO, Reconstruction an Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. Genome Res, 2004, 14:1298-1309.

[7] Nookaew I. , Jewett M., Meechai A., Thammarongtham, C., Laoteng, K., Cheevadhanarak S., Nielsen, J. and Bhumiratana S. The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism. BMC Systems Biology, 2008, 2:71, 1752-0509.

[8] Reed J. L., Patel T.R., Chen K.H., Joyce A.R., Applebee M.K., Herring C.D., Bui O.T., Knight E.M., Fong S.S., and Palsson B.O. Systems approach to refining genome annotation. PNAS 2006, 103:46, 17480-17484.

[9] Kumar V.S., Dasika M.S., and Maranas C.D. Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics. 2007; 8: 212.

Table 1: Comparison of the top connected metabolites in yeast metabolic networks for aber modle and consensus model.

	aber model			consensus model		
Rank	CompoundID	Degree	Name.Aber	CompoundID	Degree	Name.Consensus
1	C00013	387	Pyrophosphate	C00080	675	H+
2	C00002	294	ATP	C00001	464	H2O
3	C00080	213	H+	C00002	252	ATP
4	C00009	197	Orthophosphate	C00010	179	CoA
5	C00001	183	H2O	C00008	162	ADP
6	C00006	171	NADP+	C00003	157	NAD+
7	C00011	164	CO2	C00006	147	NADP+
8	C00005	163	NADPH	C00005	146	NADPH
9	C00008	160	ADP	C00004	144	NADH
10	C00046	152	RNA	C00009	141	Orthophosphate
11	C00003	122	NAD+	C00013	133	Pyrophosphate
12	C00020	117	AMP	C00011	107	CO2
13	C00004	108	NADH	C00020	105	AMP
14	C00229	94	Acyl-carrier protein	C00007	90	Oxygen
15	C00039	92	DNA	C00024	84	Acetyl-CoA
16	C01209	83	Malonyl-[acyl-carrier protein]	C00025	66	L-Glutamate
17	C00010	78	CoA	C01342	60	NH4+
18	C00173	76	Acyl-[acyl-carrier protein]	C00027	58	H2O2
19	C00025	61	L-Glutamate	C00026	44	2-Oxoglutarate
20	C00044	56	GTP	C00342	42	Thioredoxin
21	C00014	54	NH3	C00343	42	Oxidized thioredoxin
22	C00024	52	Acetyl-CoA	C00022	35	Pyruvate
23	C00007	50	Oxygen	C00031	24	D-Glucose
24	C00022	44	Pyruvate	C00033	23	Acetate
25	C00063	44	CTP	C00229	22	Acyl-carrier protein
26	C00075	42	UTP	C00015	22	UDP
27	C00035	38	GDP	C00084	21	Acetaldehyde
28	C00026	38	2-Oxoglutarate	C00083	21	Malonyl-CoA
29	C00496	30	Ubiquitin	C00042	20	Succinate
30	C00201	30	Nucleoside triphosphate	C00035	19	GDP

Table 2: Comparison of the model starting compound settings for the aber and consensus model.

KEGGid	Name	inStartAber	inStartCons	ubiqAber	ubiqCons
C00123	L-Leucine	Yes	Yes		
C00135	L-Histidine	Yes	Yes		
C00137	myo-Inositol	Yes	Yes		
C00073	L-Methionine	Yes	Yes		
C00106	Uracil	Yes	Yes		
C00342	Thioredoxin	Yes	Yes	Yes	Yes
C00864	Pantothenate	Yes	Yes		
C00238	Potassium	Yes	Yes		
C00115	Chloride	Yes	Yes		
C08219	Potassium iodide	Yes	Yes		
C00504	Folate	Yes	Yes		
C00059	Sulfate	Yes	Yes		
C00305	Magnesium	Yes	Yes		
C00568	4-Aminobenzoate	Yes	Yes		
C08130	Calcium chloride anhydrous	Yes	Yes		
C01330	Sodium	Yes	Yes		
C00120	Biotin	Yes	Yes		
C00009	Orthophosphate	Yes	Yes		
C00008	ADP	Yes	Yes	Yes	Yes
C00007	Oxygen	Yes	Yes	Yes	Yes
C00034	Manganese	Yes	Yes		
C00001	H2O	Yes	Yes		
C06232	Molybdate	Yes	Yes		
C00253	Nicotinate	Yes	Yes		
C00255	Riboflavin	Yes	Yes		
C00314	Pyridoxine	Yes	Yes		
C00070	Copper	Yes	Yes		
C00076	Calcium	Yes	Yes		
C00378	Thiamin	Yes	Yes		
C00038	Zinc	Yes	Yes		
C00080	H+	Yes	Yes		
C06266	Boron	Yes	Yes		
C00023	Iron	Yes			
C00014	NH3	Yes			
C00267	alpha-D-Glucose	Yes			
C14818	Fe2+		Yes		
C01342	NH4+		Yes		
C00670	sn-glycero-3-Phosphocholine		Yes*		
C00006	NADP+		Yes		Yes
C00031	D-Glucose		Yes		

Note: Some ubiquitous compounds have been added to the model at the initial stage to make sure the wild type model is viable and they must be one of the producible products in the model as well. C00670 is added to the starting compound set such that the model for the wild type is viable as the consensus model is incomplete in lipid metabolism and choline relevant productsare not producible from the current consensus model.

Table 3: Comparison of the essential compound settings for aber model and consensus model.

KEGGid	Name	inGoalAber	inGoalCons
C00123	L-Leucine	Yes	Yes
C00135	L-Histidine	Yes	Yes
C00137	myo-Inositol	Yes	Yes
C00073	L-Methionine	Yes	Yes
C00025	L-Glutamate	Yes	Yes
C00103	D-Glucose 1-phosphate	Yes	Yes
C00024	Acetyl-CoA	Yes	Yes
C00063	CTP	Yes	Yes
C00062	L-Arginine	Yes	Yes
C00065	L-Serine	Yes	Yes
C00148	L-Proline	Yes	Yes
C00286	dGTP	Yes	Yes
C01694	Ergosterol	Yes	Yes
C00022	Pyruvate	Yes	Yes
C00668	alpha-D-Glucose 6-phosphate	Yes	
C00114	Choline	Yes	Yes
C00116	Glycerol	Yes	Yes
C00356	(S)-3-Hydroxy-3-methylglutaryl-CoA	Yes	Yes
C00152	L-Asparagine	Yes	Yes
C00157	Phosphatidylcholine	Yes	
C00422	Triacylglycerol	Yes	
C00183	L-Valine	Yes	Yes
C00064	L-Glutamine	Yes	Yes
C01120	Sphinganine 1-phosphate	Yes	Yes
C00416	Phosphatidate	Yes	
C00458	dCTP	Yes	Yes
C00459	dTTP	Yes	Yes
C00096	GDP-mannose	Yes	Yes
C00097	L-Cysteine	Yes	Yes
C00037	Glycine	Yes	Yes
C00078	L-Tryptophan	Yes	Yes
C00079	L-Phenylalanine	Yes	Yes
C00188	L-Threonine	Yes	Yes
C00189	Ethanolamine	Yes	Yes
C00075	UTP	Yes	Yes
C00002	ATP	Yes	Yes
C00131	dATP	Yes	Yes
C00082	L-Tyrosine	Yes	Yes
C00407	L-Isoleucine	Yes	Yes
C00043	UDP-N-acetyl-D-glucosamine	Yes	
C00041	L-Alanine	Yes	Yes
C00047	L-Lysine	Yes	Yes
C00044	GTP	Yes	Yes
C00049	L-Aspartate	Yes	Yes
C00092	D-Glucose 6-phosphate		Yes
C00588	Choline phosphate		Yes
C00093	sn-Glycerol 3-phosphate		Yes
C00203	UDP-N-acetyl-D-galactosamine		Yes

Table 4: Comparison of gene essentiality prediction for the Aber and consensus model. Here only the 1106 essential genes in YPD medium were used to validate the logical model prediction.

Model	Aber	Consensus	Shared	Union
ORFs in model	919	833	659	1093
ORFs both in model and in in essential gene set	168	113	104	177
Predicted essential genes	59	53	30	87
Corr essential prediction	30	29	26	40
Sensitivty for essentiality detection	17.9%	25.7%	25.0%	22.6%
Corr rate for essentiality prediction	50.8%	49.2%	86.7%	46.0%

Table 5: Performance measures of logical model simulation for yeast growth prediction for the Aber and consensus model, validated by three experimental data sets using defined medium

Model (#ORF)	Exp	TP	FN	TN	FP	rMaj %	accuracy %	sens %	spec %	PPV %	NPV %
Aber	A	40	167	626	17	75.65	78.35	19.32	97.36	70.18	78.94
	B	35	148	645	22	78.47	80.00	19.13	96.70	61.40	81.34
(850)	C	30	156	632	27	77.99	78.34	16.13	95.90	52.63	80.20
Cons	A	34	117	607	6	80.24	83.90	22.52	99.02	85.00	83.84
	B	30	98	626	10	83.25	85.86	23.44	98.43	75.00	86.46
(764)	C	29	110	613	11	81.78	84.14	20.86	98.24	72.50	84.79
Aber-	A	40	101	448	14	76.62	80.93	28.37	96.97	74.07	81.60
Shared	B	35	84	465	19	80.27	82.92	29.41	96.07	64.81	84.70
(603)	C	30	91	457	24	79.90	80.90	24.79	95.01	55.56	83.39
Cons-	A	33	108	457	5	76.62	81.26	23.40	98.92	86.84	80.88
Shared	B	29	90	475	9	80.27	83.58	24.37	98.14	76.32	84.07
(603)	C	28	93	471	10	79.90	82.89	23.14	97.92	73.68	83.51

Table 6: Biomass composition used for flux balance analysis on the consensus model in comparison to the reference model iIN800.

MetaboliteID in consensus	MetaboliteID in iIN800	Metabolite Name	Coefficients under carbon limited condition
M_143	M_629	L-Alanine	0.35734
M_163	M_632	L-Arginine	0.13579
M_166	M_634	L-Asparagine	0.17152
M_168	M_637	L-Aspartate	0.17152
M_209	M_644	L-Cysteine	0.04288
M_308	M_650	L-Glutamate	0.268
M_305	M_653	L-Glutamine	0.268
M_315	M_547	Glycine	0.32518
M_344	M_658	L-Histidine	0.075041
M_363	M_663	L-Isoleucine	0.17152
M_380	M_668	L-Leucine	0.25014
M_383	M_672	L-Lysine	0.23942
M_395	M_676	L-Methionine	0.050027
M_448	M_681	L-Phenylalanine	0.11435
M_470	M_684	L-Proline	0.12864
M_499	M_688	L-Serine	0.25371
M_531	M_691	L-Threonine	0.19653
M_559	M_695	L-Tryptophan	0.028
M_565	M_699	L-Tyrosine	0.096481
M_578	M_702	L-Valine	0.25728
M_319	M_549	Glycogen	0.51852
M_537	M_249	alpha,alpha-Trehalose	0.023371
M_392	M_728	Mannan	0.82099
M_2	M_29	1,3-beta-D-Glucan	1.1358
M_172	M_277	ATP	59.276
M_151	M_262	AMP	0.051
M_321	M_555	GMP	0.051
M_201	M_347	CMP	0.05
M_571	M_1090	UMP	0.067
M_215	M_400	dAMP	0.003587
M_223	M_407	dCMP	0.002432
M_252	M_454	dTMP	0.003587
M_230	M_429	dGMP	0.002432
M_505	M_982	Sulfate	0.02
Lipids			
N/A	M_861	Phosphatidylcholine	0.002884
N/A	M_38	1-Phosphatidyl-D-myo-inositol	0.001531
N/A	M_866	Phosphatidylserine	0.000373
N/A	M_862	Phosphatidylethanolamine	0.000697
N/A	M_220	Acyl_acids	0.000206
N/A	M_1050	Triacylglycerol	0.000781
N/A	M_477	Ergosterol-ester	0.000812
M_265	M_473	Ergosta-5,7,22,24(28)-tetraenol	0.000125
M_263	M_476	Ergosterol	0.005603
M_589	M_1116	Zymosterol	0.000015
M_261	M_466	Episterol	0.000096
M_278	M_491	Fecosterol	0.000114
M_379	M_706	Lanosterol	0.000032
M_61	M_153	4,4-Dimethylzymosterol	0.000056
N/A	M_606	Ceramide-I	0.000351
N/A	M_605	Ceramide-II	0.000066