


RESEARCH

Open Access



# Whole-genome metabolic model of *Trichoderma reesei* built by comparative reconstruction

Sandra Castillo<sup>1\*</sup> , Dorothee Barth<sup>1</sup>, Mikko Arvas<sup>1</sup>, Tiina M. Pakula<sup>1</sup>, Esa Pitkänen<sup>2</sup>, Peter Blomberg<sup>1</sup>, Tuulikki Seppänen-Laakso<sup>1</sup>, Heli Nygren<sup>1</sup>, Dhinakaran Sivasiddharthan<sup>1</sup>, Merja Penttilä<sup>1</sup> and Merja Oja<sup>1</sup>

## Abstract

**Background:** *Trichoderma reesei* is one of the main sources of biomass-hydrolyzing enzymes for the biotechnology industry. There is a need for improving its enzyme production efficiency. The use of metabolic modeling for the simulation and prediction of this organism's metabolism is potentially a valuable tool for improving its capabilities. An accurate metabolic model is needed to perform metabolic modeling analysis.

**Results:** A whole-genome metabolic model of *T. reesei* has been reconstructed together with metabolic models of 55 related species using the metabolic model reconstruction algorithm CoReCo. The previously published CoReCo method has been improved to obtain better quality models. The main improvements are the creation of a unified database of reactions and compounds and the use of reaction directions as constraints in the gap-filling step of the algorithm. In addition, the biomass composition of *T. reesei* has been measured experimentally to build and include a specific biomass equation in the model.

**Conclusions:** The improvements presented in this work on the CoReCo pipeline for metabolic model reconstruction resulted in higher-quality metabolic models compared with previous versions. A metabolic model of *T. reesei* has been created and is publicly available in the BIOMODELS database. The model contains a biomass equation, reaction boundaries and uptake/export reactions which make it ready for simulation. To validate the model, we demonstrate that the model is able to predict biomass production accurately and no stoichiometrically infeasible yields are detected. The new *T. reesei* model is ready to be used for simulations of protein production processes.

## Background

*Trichoderma reesei* is a filamentous fungus widely used for commercial scale production of biomass-degrading enzymes. Cellulose is the most abundant organic compound in the biosphere and is used as a raw material by many industries such as paper, food, textile and for biofuel production. *Trichoderma reesei* is the main industrial source for cellulases and hemicellulases, enzymes that hydrolyze the cellulose component of lignocellulosic materials. There is a need for reducing the cost and maximizing the yield of these cellulose-degrading enzymes.

Whole-genome stoichiometric metabolic models aim to fully explain the metabolism of an organism. The model is a collection of interconnected metabolic reactions that represent the biochemical possibilities of the organism. Whole-genome metabolic reaction networks have been reconstructed for many species, such as human (Recon 2) [1], yeast *Saccharomyces cerevisiae* [2], *Pichia stipitis* [3], *Pichia pastoris* [3], model plant *Arabidopsis thaliana* [4], bacteria *Escherichia coli* [5], *Bacillus subtilis* [6], fungus *Aspergillus niger* [7], *Aspergillus oryzae* [8], *Aspergillus nidulans* [9] and cyanobacteria *Synechocystis* [10]. To date, the BIOMODELS database [11] contains 1483 published models, some of which are whole-genome stoichiometric metabolic models. Additionally, the Path2models branch of the BIOMODELS

\*Correspondence: sandra.castillo@vtt.fi

<sup>1</sup> VTT Technical Research Centre of Finland, Tietotie 2, P.O. Box FI-1000, 02044 Espoo, Finland

Full list of author information is available at the end of the article

database hosts an automatic reconstruction of 2641 whole-genome models based on the pathway information already included in KEGG [12] or MetaCyc [13] for the particular organism in question.

Metabolic modeling can aid as a tool in the development of microbial strains capable of high efficiency production of chemicals or proteins. Metabolic models are used to simulate the performance of the production strain in different scenarios (genetic modifications, cultivation set-ups) [14, 15]. High-quality metabolic models are required for successful metabolic simulations and predictions. So far, metabolic modeling for protein production scenarios has been rare, probably because the lack of good quality metabolic models for typical protein production hosts. Metabolic modeling of super oxide dismutase production in *Komagatella phaffii* (*P. pastoris*) has been done [16]. Metabolic modeling of protein production in *T. reesei* has been reported the first time in our accompanying paper [17], where the metabolic model from this paper has been used.

The reconstruction of metabolic models can be a tedious process including as many as 96 steps [18]. The main steps include the annotation of enzymes encoded by the organism's genes, and the subsequent assembly of the metabolic reactions that are supported by these enzymes into a network. Recently, several automatic reconstruction tools have been proposed. Model SEED [19] is a web based pipeline for creating and analyzing metabolic models using techniques that automate the process. RAVEN [20] is a program suite for semi-automated reconstruction of a metabolic model based on a reference whole-genome metabolic model (or KEGG). RAVEN includes tools for gap-filling, quality control, compartmentalization and visualization of the models to speed up the manual curation work required on top of the automatic reconstruction step.

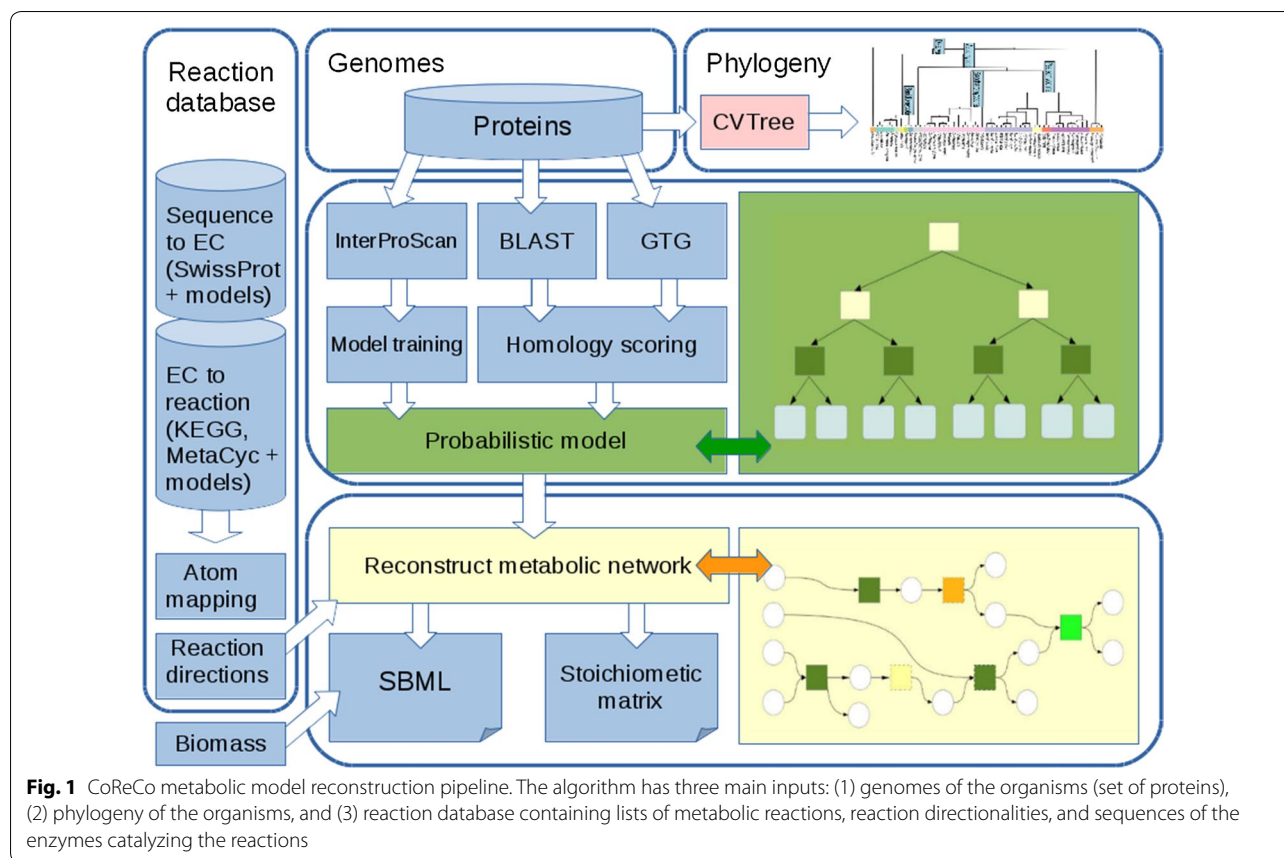
The CoReCo algorithm [21] automatically reconstructs gapless metabolic models for several related species at once. The CoReCo metabolic model reconstruction pipeline has two parts (see Fig. 1). In the first part of the pipeline, the enzyme content of the input organisms is scored. The score for each enzyme is computed based on a probabilistic model that combines homology-based scoring from BLAST [22] and GTG [23] to InterPro annotations [24]. Including the phylogeny of the input organisms improves the prediction of enzymes in these organisms. In the interface between the two parts, a score for each reaction is computed as the maximum of the scores of the enzymes catalyzing the reaction. In the second part of the algorithm, a gapless metabolic model is created by adding high scoring reactions to the model, ensuring at each step the connectivity of the network. Low-scoring reactions may be added to avoid gaps. Enforcing gaplessness in the

network reconstruction process and taking into account atom mappings [25] results in good quality models. However, the performance of previous CoReCo models had two main caveats: a number of cofactors or their precursors like biotin, pantothenate and choline could not be synthesized by the models and when testing for production of individual molecules slightly higher than stoichiometrically possible yields were often detected. Such problems were also common in early manually reconstructed metabolic models.

In this work, we present several improvements to the CoReCo metabolic model reconstruction method. CoReCo creates metabolic models by combining reactions given to it as inputs. So far, the KEGG database has been used as the source of reactions. Unfortunately, it has proven to be inadequate in several ways (1) several reactions are missing, (2) reactions are not balanced or contain undefined number of atoms (generic 'R') or (3) are electron imbalanced. Electron imbalance negatively affects the calculation of reaction directions and flux balances. Missing reactions cause unnecessary gaps or long detour pathways. To resolve these issues, a new comprehensive database of metabolic reactions was created by combining information from public databases and whole-genome metabolic models. Furthermore, reaction directions were included as additional constraints for the gapless network reconstruction step of the algorithm. The scoring of enzymes was updated from previously having been the mean of the two evidence sources (GTG and BLAST) to being the maximum of the two evidence sources.

The whole-genome metabolic models can be used to simulate bioprocesses, for example, in a protein production set up. Central to metabolic modeling of bioprocesses, is the simulation of cell growth. To do this successfully, the cellular biomass composition needs to be described accurately in the metabolic model [26]. Cell growth is coupled to the metabolic network via a biomass equation that lists the components required for growth: amino acids, cell wall components, RNA, DNA, etc. Biomass composition has previously been measured and used to create biomass equations for genome-scale metabolic models of *S. cerevisiae* [27], *Scheffersomyces stipitis* [28] and *A. thaliana* [4]. The biomass equations in whole-genome metabolic models are similar to each other, but there are species specific differences.

The updated CoReCo metabolic model reconstruction pipeline was used to create a whole-genome metabolic model for the industrially important fungus *T. reesei*. CoReCo draws power from a phylogeny of input organisms, thus the *T. reesei* model has been reconstructed as part of a set of 56 fungi, allowing comparison of the fungal metabolic models. To complement the *T. reesei*



metabolic model, we measured the biomass composition of *T. reesei*.

## Results

### Database of balanced reactions

Metabolic reactions from several public reaction databases and numerous relevant genome-scale metabolic models were gathered to create a high-quality database of balanced metabolic reactions. KEGG [12] and MetaCyc [13] list metabolic reactions in pathways, possibly leading to gaps between disconnected pathways. The whole-genome metabolic models of biotechnologically relevant fungi were included to provide a gapless set of metabolic reactions.

To combine the reactions from several different sources, first a unified list of all metabolites included in these reactions was created. Compounds were collected from YMDB [29], HMDB [30], ChEBI [31], KEGG Compound [12], and Rhea databases [32]. Metabolites from whole-genome metabolic models were mapped via database crosslinks, or via metabolite names when name-based matching allowed unique mapping. However, from each source, various metabolites remained unmapped, thus resulting in reactions not having all reactants

mapped; see column “% of reactions fully resolved” in Table 1.

For the purpose of reconstructing models for fungi, genome-scale metabolic models of *S. cerevisiae* (community model v. 6.06) [2], *A. niger* (iMA871) [7], *A. oryzae* (iWV1314) [8], *A. nidulans* (iHD666) [9], *K. (Pichia) pastoris* (iLC915), *P. stipitis* (iSS884) [3], and *Penicillium chrysogenum* (iAL1006) [20] were included. In addition to these, reactions from KEGG and MetaCyc databases were included. The corresponding reactions from the different sources were combined, leaving only one copy of each unique reaction. In an effort to keep track of the source of the reactions, a representative for each unique reaction was selected. The representative was selected preferably from the whole-genome models, then from KEGG, and finally from MetaCyc. Column “Selected as representative” in Table 1, lists the number of reactions selected from each source. We selected 1020 reactions from the *S. cerevisiae* model, and then additional reactions, not present in *S. cerevisiae* model, from the other whole-genome models. 6618 KEGG reactions, not present in the whole-genome models, were selected. And finally, 3145 MetaCyc reactions, not present in the other sources, were included. Note that even though only 1020

**Table 1 Shows the origin of the reactions added to the reaction database**

Species	Source	Reactions	% of matched reactions	% of balanced reactions	Selected as representative	Representatives that balance
<i>S. cerevisiae</i>	ymn6.06	1888	96	90	1020	923
<i>A. niger</i>	iMA871	1399	57	95	266	262
<i>A. nidulans</i>	iHD666	1303	85	69	88	72
<i>A. oryzae</i>	iWV1314	2360	93	74	265	235
<i>K. (Pichia) pastoris</i>	iLC915	1423	88	91	237	207
<i>P. stipitis</i>	iSS884	1332	88	92	6	5
<i>P. chrysogenum</i>	iAL1006	1636	83	95	75	67
KEGG reaction	KEGG	9236	93	90	6618	6376
MetaCyc	MetaCyc	11,181	62	90	3145	3054
Total		31,758	77	88	11,720	11,201

A fully resolved reaction is a reaction having all reactants identified by metabolites in the updated database. All representative reactions balance elements other than hydrogen. Reactions that also balance electrons are denoted "reactions that balance"

of the 1888 reactions from the yeast *S. cerevisiae* model were selected, a good coverage of the core metabolism is achieved. The *S. cerevisiae* model includes identical reactions in several compartments and which have been combined in our reaction database. Lipid metabolism is not well-represented in neither the public databases (KEGG and MetaCyc) nor the whole-genome models. For the CoReCo reaction database, reactions for lipid metabolism were collected from v7.0 of the yeast community model [33]. Lipid reactions from other sources were removed.

Surprisingly, a large fraction of the metabolic reactions from the various sources showed problems in atom balances or charges (see column "% of balanced reactions" in Table 1). To tackle the issue of charge imbalance, the number of hydrogen atoms was replaced with the total amount of electrons in the compound. For some reactions, atom and electron balances were achieved using a balancing procedure that changed the stoichiometric coefficients while allowing the addition of water to the reaction equations. Still, some of the selected reactions had to be rejected due to balancing problems (compare columns "Selected as representative" and "Representatives that balance" columns in Table 1). Correct atom balances are required for the atom-mapping procedure, where each carbon atom in the substrates is mapped to the corresponding atom in the products [34]. Atom-mapping constraints aid in the gapless network reconstruction step of the CoReCo algorithm. Electron balance is important for the computation of reaction directions via thermodynamics. For this work, reaction directions were extracted from the selected, manually curated, full-genome metabolic models. However, our set of atom- and electron-balanced reactions would be ready for thermodynamic calculations using good group-contribution methods, such as [35–38].

In the CoReCo metabolic model reconstruction pipeline, the sequence information is coupled to the metabolic reactions via E.C. numbers. Previously, the sequence to E.C. information was extracted from SwissProt under the assumption that the enzyme annotations in this manually curated section of the UniProt database are reliable. The sequence to E.C. information for whole-genome metabolic models was extracted from the reaction gene rules included in these models and added to the Swissprot-derived information.

#### Improvements to the CoReCo algorithm

Two main updates have been made to the CoReCo algorithm. First, the scoring of enzymes has been refined. In the first step of the CoReCo algorithm, the proteins of the input organisms are compared to sequence information using BLAST and GTG. Using the homology search results, the probability of observing each enzyme is scored for each species. This score is based on the probabilistic model that takes into account both BLAST and GTG and it has been shown that the inclusion of GTG improves the reconstructions [21]. Previously, the score was computed as an average of these two sources of information. In some rare cases, the BLAST evidence was very good, but the GTG evidence low. Thus, resulting in a score so low that the reaction was rejected from the reconstructed model even though BLAST clearly found the enzyme to be present in the organism. Low GTG evidence typically arises when GTG found the right enzyme but not the phylogenetically closest one. For example, for "2.7.8.1 Choline/ethanolaminephosphotransferase 1" the best match by BLAST for fungal sequences in UniProt is the *S. cerevisiae* protein P22140 EPT1, while GTG found the *Homo sapiens* protein Q9Y6K0 CEPT1 as the best match. Both EPT1 and CEPT1 carry out the same reaction, but CEPT1 (due to phylogenetic distance) has

far lower sequence similarity, hence the GTG score is close zero. Consequently, 2.7.8.1 was excluded from the models. In the updated version of CoReCo, the score is computed as the maximum of the BLAST and the GTG scores.

The second update is in the network reconstruction step where reaction directions are now taken into account. In the second phase of the CoReCo algorithm, the metabolic network is reconstructed using an algorithm that creates gapless metabolic networks [21]. The reactions are added to the network in an iterative manner, starting from the highest-scoring reactions, and subsequently adding reactions until the remaining reactions have a score lower than a user-defined threshold  $\alpha$ . At each stage, the connectivity of the network is guaranteed by requiring the new reactions to be connected to the already established network as well as to a predefined list of source metabolites. To ensure connectedness, a limited number of low-scoring reactions may be added to fill potential gaps. In the updated version of the CoReCo algorithm, reaction directions are taken into account in the gap-filling process, thus ensuring that the network reconstruction is proceeding in a more sensible manner.

Zymosterol production exemplifies how the improvements made in the CoReCo pipeline and the reaction database result in a more accurate metabolic model. In KEGG's zymosterol pathway as shown in Fig. 2, the reactions containing a *T. reesei* gene annotation are colored blue. Notice that the E.C.: 1.3.1.70, a  $\delta(14)$ -sterol reductase, is not found in *T. reesei* according to KEGG, however, CoReCo predicts that *T. reesei* contains E.C.: 1.3.1.70. The best evidence for CoReCo comes from the *T. reesei* gene tre81049 (JGI v2.0 genome annotation identifier), manually curated as C-14 sterol reductase that matches *N. crassa* erg-3 (P38670) in the bidirectional BLAST scoring scheme of CoReCo. *N. crassa* erg-3 has been manually curated to be a E.C.: 1.3.1.70 enzyme.

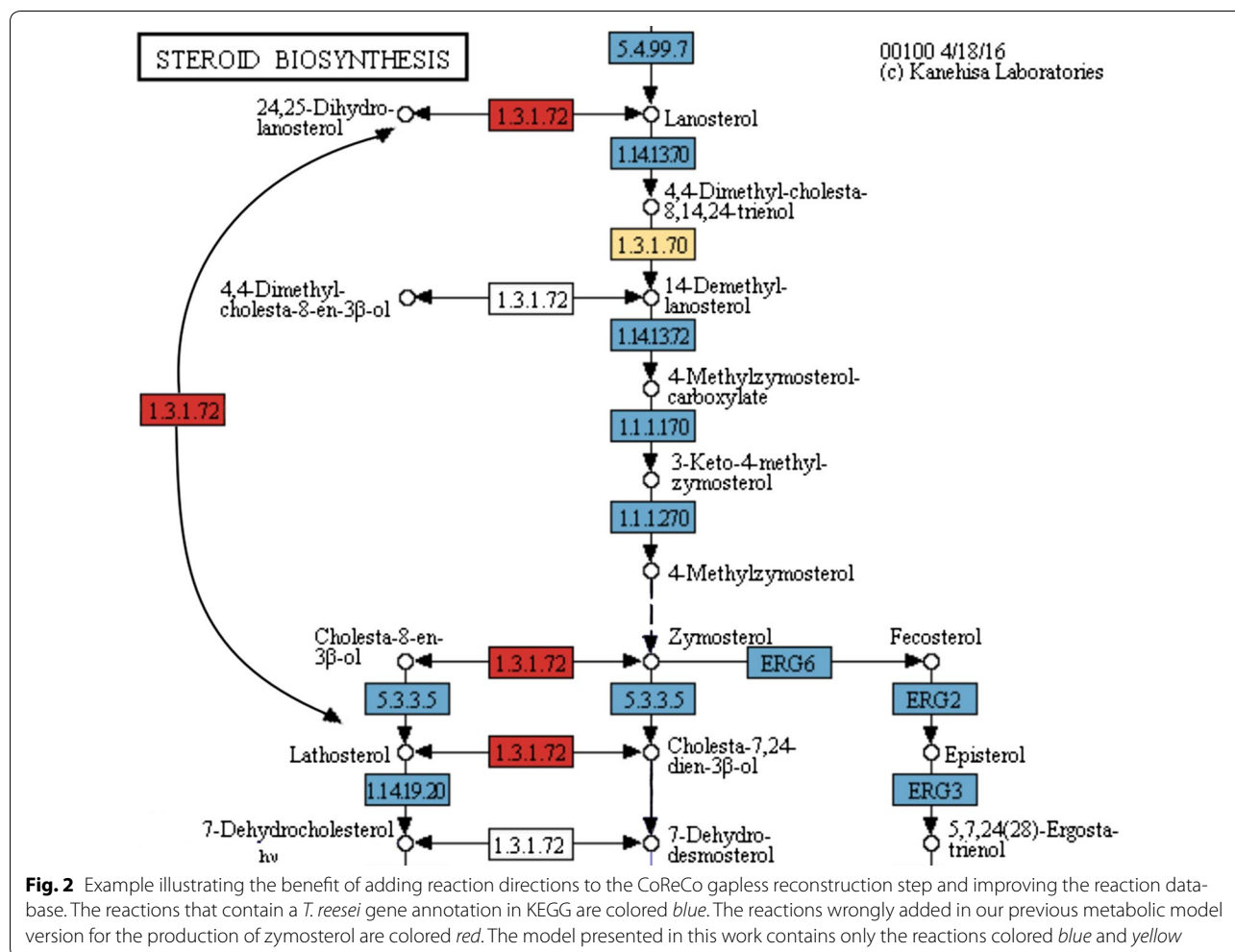
In the previous version of CoReCo, due to the fact that KEGG is missing reactions in the path going from 4-methylzymosterol to zymosterol, the CoReCo gap-filling step added low score reactions (marked red) to complete the path. The model connected lanosterol to lathosterol, and subsequently to zymosterol through the reactions with E.C. numbers 1.3.1.72 and 5.3.3.5. After reaction bounds were added to the completed model, it was unable to produce zymosterol since the reactions assigned to E.C. 5.3.3.5 are irreversible. However, after the reaction boundaries were included already in the gapless network assembly step and the additional use of the newly created reaction database, the reconstructed models in this specific case did not contain low-scoring reactions anymore and overall performed much better in similar situations.

The last step of the CoReCo pipeline is the writing of the models in SBML format. Exchange reactions, reaction directions and two biomass equations are included in each reconstructed metabolic model.

#### Reconstruction of *T. reesei* metabolic model

A whole-genome metabolic model for *T. reesei* was built using the updated CoReCo algorithm and reaction database described above. CoReCo reconstructs models for several species at once by taking into account the phylogeny of the input species in the enzyme scoring step. Here, genomes of 56 fungi were used as input to CoReCo. This set includes the 49 fungi used in the first CoReCo publication [21]. For the purpose of reconstructing the *T. reesei* model, we added several *Trichoderma* genomes into the set of input organisms. A phylogeny of the input organisms was created using the CVtree algorithm [39]. For *T. reesei* metabolic model reconstruction, the CoReCo  $\alpha$  parameter, i.e., the threshold for reaction scores was selected based on the biomass compound production in the simulations (see below). The CoReCo  $\beta$  parameter, controlling how many reactions with a score lower than  $\alpha$  may be added in the gap-filling step, was set to 2.

The reconstructed *T. reesei* model has 3926 reactions (including 148 exchange reactions and 261 orphan reactions) and 3348 metabolites. The orphan reactions are added to the network because their score exceeds the  $\alpha$  parameter. However, the subsequent gap-filling process fails to connect them to the rest of the network. The automatically reconstructed model was able to create most of the biomass components. In the manual curation process, 16 reactions were added, rendering the model fully functional. Of the added reactions, 13 were related to lipid metabolism. These lipid metabolism reactions were not included automatically because they were not coupled to sequence information in the CoReCo input, and thus the algorithm was unable to include them. An additional reaction had to be included to allow the synthesis of L-threonine. This reaction's score was just below the algorithm threshold. Finally, the P/O ratio was set to 1 by coupling ATP production to the reduction of oxygen in the electron transport chain adding two reactions. Correct reaction directions proved to be critical for a successful reconstruction. The reaction directions were originally harvested from the whole-genome models, but were iteratively updated during repeated rounds of reconstruction of models, simulating growth, and refining bounds. During this process, the boundaries of 112 reactions included in the *T. reesei* model were closed, and the direction of 24 reactions changed, respectively. The final models have been automatically reconstructed using the manually curated reaction directions.



### Construction of the biomass equation for *T. reesei*

A biomass equation is essential part of a metabolic model. It is necessary to have an accurate biomass equation, before the metabolic model can be used to simulate cellular metabolism and growth. To create the equation, the major components of biomass (total cellular protein, carbohydrates, DNA, RNA, esterified and free fatty acids and the major lipid classes) were measured in *T. reesei* grown on minimal medium containing cellobiose as a carbon source. The measurement data in this study was supplemented with data on the codon frequency in the transcriptome [17] to estimate the amino acid composition of proteins, and the amount of ash measured from chemostat cultures of *T. reesei* [40]. An overview of the biomass composition measurements is given in Table 2. The compound coefficients used in the biomass equation of the models can be seen in Table 3. The biomass equation was then coupled to the reconstructed metabolic model of *T. reesei*.

### Biomass production by the reconstructed fungal metabolic models

The quality of the reconstructed metabolic models was assessed by simulations. To be able to use a metabolic model to simulate protein production, the model should be able to accurately predict cell growth and the load of protein production. The concept of growth is modeled via the biomass equation, whereas protein production load can be estimated as a combination of amino acids, ATP, ribonucleotides, and nucleotides required for the product protein, see for example [41]. Thus, to validate the quality of the models, we estimated their capability to produce the biomass components (including amino acids, ATP, RNA, and DNA). The goal of this analysis is to verify that the models contain all the necessary metabolic pathways.

First, the models' ability to grow (Fig. 3), i.e., to produce all necessary biomass components in correct ratios, was tested with a simple biomass equation for *S. cerevisiae*

**Table 2 The macromolecular composition of *T. reesei***

Biomass component	% (w/w)
Proteins	45.100 <sup>a</sup>
Ala	2.715 <sup>b</sup>
Arg	3.944 <sup>b</sup>
Asn	1.613 <sup>b</sup>
Asp	2.677 <sup>b</sup>
Cys	0.549 <sup>b</sup>
Gln	2.063 <sup>b</sup>
Glu	3.200 <sup>b</sup>
Gly	1.649 <sup>b</sup>
His	1.346 <sup>b</sup>
Ile	2.186 <sup>b</sup>
Leu	4.101 <sup>b</sup>
Lys	2.595 <sup>b</sup>
Met	1.202 <sup>b</sup>
Phe	2.220 <sup>b</sup>
Pro	2.363 <sup>b</sup>
Ser	2.808 <sup>b</sup>
Thr	2.354 <sup>b</sup>
Trp	1.144 <sup>b</sup>
Tyr	1.846 <sup>b</sup>
Val	2.523 <sup>b</sup>
Carbohydrates	23.190 <sup>a</sup>
Chitin	7.820 <sup>c</sup>
Other carbohydrates	15.370 <sup>d</sup>
RNA	6.122 <sup>ab</sup>
DNA	0.912 <sup>a</sup>
Lipids	4.176 <sup>a</sup>
Fatty acids-esters	
Myristic acid (C14:0) est	0.004 <sup>a</sup>
Palmitic acid (C16:0) est	0.613 <sup>a</sup>
Palmitoleic acid (C16:1n-7) est	0.011 <sup>a</sup>
Stearic acid (C18:0) est	0.070 <sup>a</sup>
Oleic acid (C18:1n-9) est	0.126 <sup>a</sup>
Linoleic acid (C18:2n-6) est	1.425 <sup>a</sup>
$\alpha$ -linolenic acid (C18:3n-3) est	0.292 <sup>a</sup>
Arachidic acid (C20:0) est	0.004 <sup>a</sup>
Lignoceric acid (C24:0) est	0.005 <sup>a</sup>
Fatty acids-free	
Palmitic acid (C16:0) FFA	0.060 <sup>a</sup>
Stearic acid (C18:0) FFA	0.012 <sup>a</sup>
Oleic acid (C18:1n-9) FFA	0.107 <sup>a</sup>
Linoleic acid (C18:2n-6) FFA	0.274 <sup>a</sup>

from the model iMM904 [42]. The FBA was run such that the models were provided an input of a minimal media containing one unit of glucose and an unlimited amount of nitrogen, phosphate, water, oxygen, iron and sulfate. Models reconstructed with different reaction score inclusion thresholds ( $\alpha$  parameter) were also compared. *T. reesei*

**Table 2 continued**

Biomass component	% (w/w)
Ergosterol	0.278 <sup>a</sup>
Triacylglycerol	1.792 <sup>e</sup>
Phosphatidylethanolamine	0.551 <sup>e</sup>
Phosphatidylcholine	1.102 <sup>e</sup>
Ash	5.100 <sup>f</sup>

<sup>a</sup> Measured as described in "Methods" section

<sup>b</sup> Amino acid ratios calculated based on codon ratios in RNAseq transcriptome data, with transcripts encoding secreted proteins removed

<sup>c</sup> Estimated based on the chitin content (% w/w) of *A. oryzae* biomass [8], and corrected for different ash contents in *A. oryzae* and *T. reesei* prepares

<sup>d</sup> Carbohydrates other than chitin

<sup>e</sup> Calculated based on the measured ratio of TAG:PE:PC 52:16:32 and the measured amount of fatty acid esters

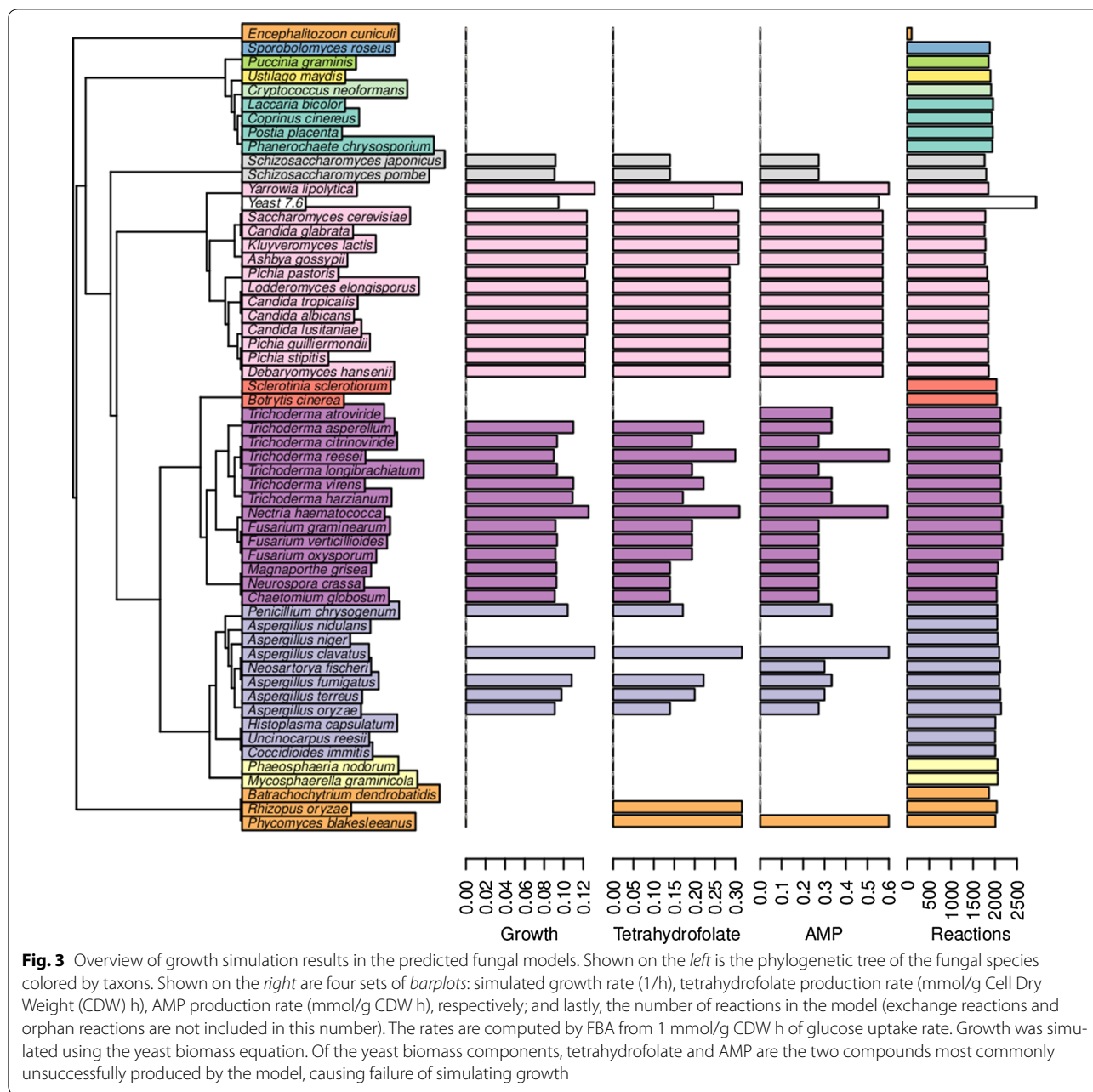
<sup>f</sup> Measured from lactose-limited chemostat cultures of *T. reesei* Rut-C30 ( $D = 0.051/h$ ) described in [40]

<sup>g</sup> Sum of measured free fatty acids, ergosterol, triacylglycerol, phosphatidylethanolamine and phosphatidylcholine

model reconstructed with  $\alpha = 0.5$  was able to produce the highest number of biomass components. With this  $\alpha$  value, 33 species were able to grow in simulations. In contrast, with an  $\alpha$  of 0.4 38 species were able to grow. The simulated growth rates, with  $\alpha = 0.5$ , are shown in Fig. 3. In the simulations, *Saccharomycotina* species typically grow at a rate of 0.12, while *T. reesei* grows at a rate of 0.09.

FBA was also carried out for simulating the production of each individual biomass compound included in the *T. reesei* biomass equation. For these simulations, the objective was set to maximize the production of a single biomass component, for example L-valine. As input, the models were given ten units of glucose and the same minimal media as described above. The simulation results are shown in Fig. 4 (see also Additional file 1) for a carbon normalized version of the same figure). In FBA simulations, the *T. reesei* model is able to produce all biomass components with stoichiometrically plausible yield, for example from ten units of glucose ( $C_6H_{12}O_6$ ), the model creates 12 units of L-valines ( $C_5H_{11}NO_2$ ).

Most of the other fungal metabolic models, created fully automatically using the same settings (reaction directions, addition of 13 lipid reactions,  $\alpha = 0.5$ ) as for *T. reesei*, are to a large extent able to create most of the biomass components as well (Fig. 4). Most importantly, the production rates for all tested compounds and in all species are stoichiometrically plausible, i.e., no extra carbon is created by physically unrealistic reactions. The biomass component production rates correlates with the phylogeny of the species. As a comparison, the identical FBA simulation setup was also applied to the Yeast 7.6 *S. cerevisiae* consensus model [43, 44] from <http://yeast.sourceforge.net/>. The compound yields are very similar in



**Fig. 3** Overview of growth simulation results in the predicted fungal models. Shown on the *left* is the phylogenetic tree of the fungal species colored by taxons. Shown on the *right* are four sets of *barplots*: simulated growth rate (1/h), tetrahydrofolate production rate (mmol/g Cell Dry Weight (CDW) h), AMP production rate (mmol/g CDW h), respectively; and lastly, the number of reactions in the model (exchange reactions and orphan reactions are not included in this number). The rates are computed by FBA from 1 mmol/g CDW h of glucose uptake rate. Growth was simulated using the yeast biomass equation. Of the yeast biomass components, tetrahydrofolate and AMP are the two compounds most commonly unsuccessfully produced by the model, causing failure of simulating growth

*T. reesei* and Yeast 7.6 with the notable exception of phosphoethanolamine and phosphocholine that are produced at far higher levels in the current CoReCo models. One compound, linoleic acid, currently included in the experimentally determined *T. reesei* biomass is not produced by Yeast 7.6.

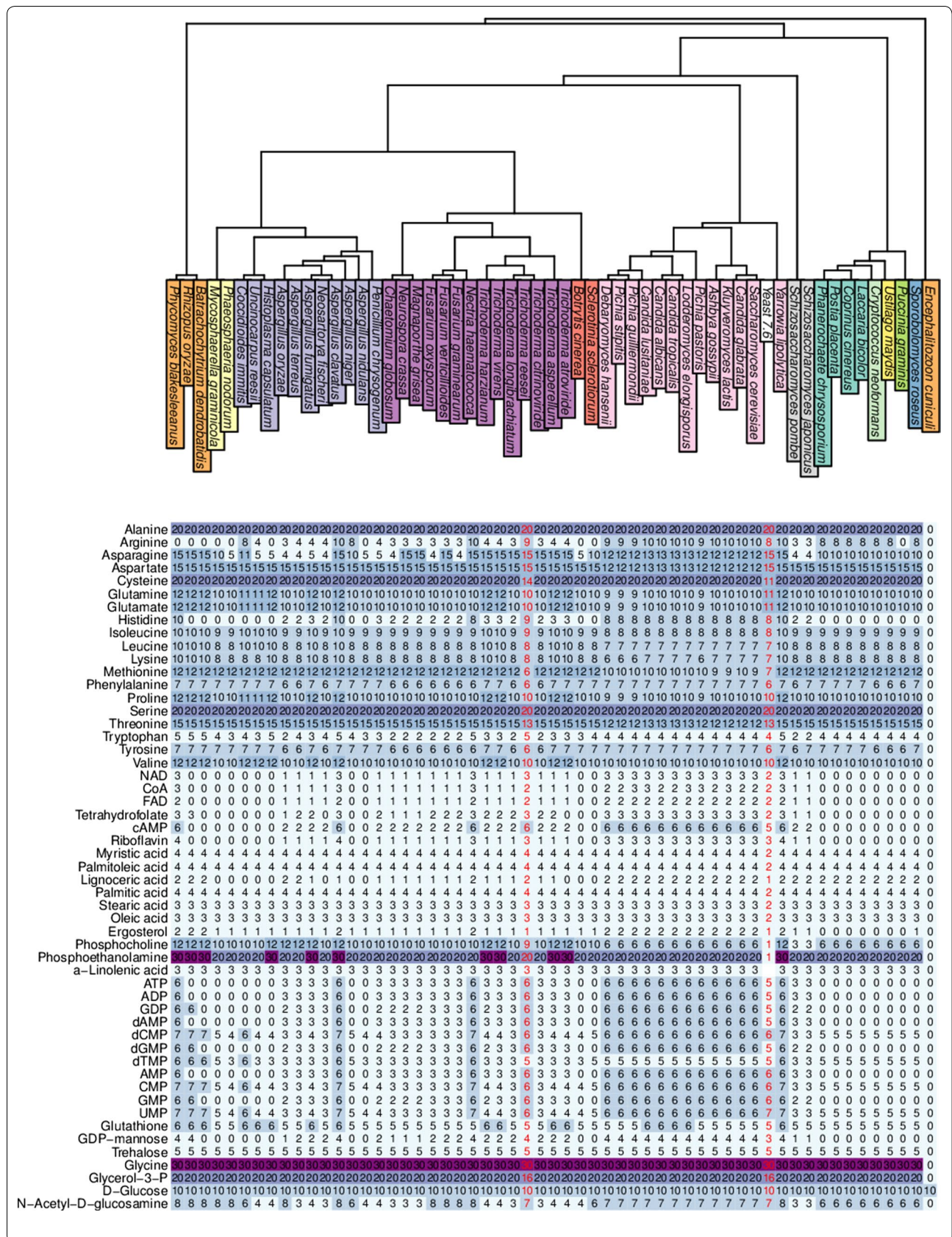
**Curation of the automatically reconstructed metabolic model based on the literature**

An extensive literature search about *T. reesei* was made to analyze the reliability of the reaction content in the metabolic model. A list of carbon sources was extracted from [45] (See Table 4). For each carbon source, an FBA was

(See figure on next page.)

**Fig. 4** Production rates for components of the *T. reesei* biomass equation, simulated by FBA. As carbon sources, the models were given tenunits of glucose. Shown at the *top* is the phylogenetic tree of the fungal species colored by taxons. Shown *below* is a heatmap of production rates for each compound and species. For reference, the compound production rates have also been computed for the Yeast 7.6 model [43, 44]. Numbers have been rounded to integers





performed on the models maximizing for growth. The model achieved biomass production with each carbon source tested, but five out of the 18 sources found in the literature were not included in the model.

Two cases could be identified, in which a specific enzymatic function was wrongly included in the model. *T. reesei* is known to lack the enzyme invertase needed to hydrolyze sucrose to form glucose and fructose [46, 47], as well as the enzyme glucose oxidase [45]. However, in the automatically created models, the reactions catalyzed by these enzymes were present. These wrongly assigned functions may arise from the influence of the phylogenetic tree on the final reaction scores, since many of the other *Trichoderma* species contain the enzymes that are missing in *T. reesei*. Both reactions were fixed manually in the model closing their bounds.

### Simulation of *T. reesei* metabolism

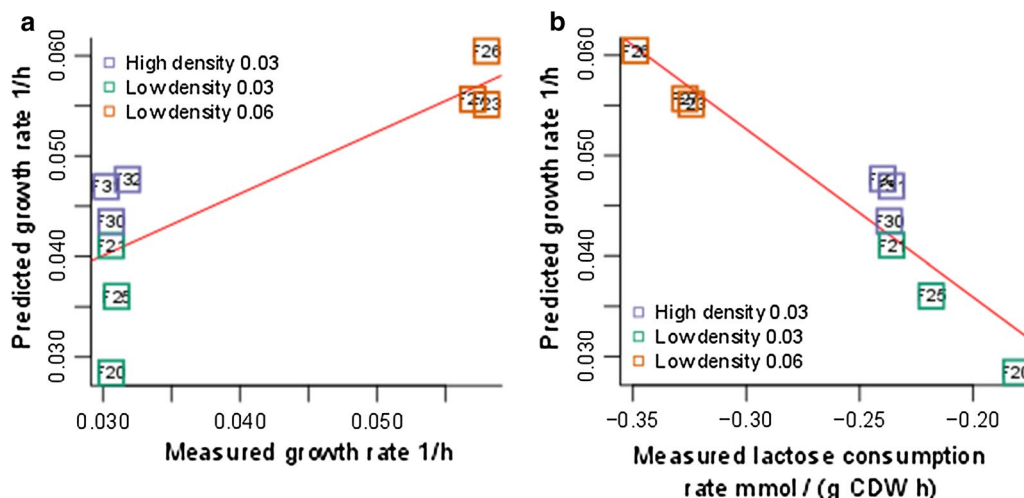
To assess the capability of the *T. reesei* model to predict the growth rate, chemostat cultivation data from a protein production experiment [48] was used for comparison. The model was constrained with the measured carbon source (lactose) consumption, CO<sub>2</sub> evolution and O<sub>2</sub> uptake rates. The cultivation data includes a set of nine chemostat cultures carried out under three different conditions: (1) high cellular density, growth rate of 0.03 h<sup>-1</sup>, (2) low cellular density, growth rate of 0.03 h<sup>-1</sup>, and (3) low cellular density, growth rate of 0.06 h<sup>-1</sup>. The high cellular density corresponds to a cellular dry weight (CDW) of 13 g/l and low cellular density to a CDW of 4 g/l. These conditions were selected to study the low growth rate protein phenotype of *T. reesei*, where the highest specific protein production rate is achieved at a relatively low growth rate

of 0.03 h<sup>-1</sup> and low cell density [49]. A Pearson correlation of 0.96 ( $p < 0.0075$ ) was found between measured and predicted growth rates (Fig. 5a). A slightly more significant Pearson correlation of  $-0.94$  ( $p < 0.0001$ ) was found between predicted growth rate and measured lactose consumption rate (Fig. 5b). In case 1 (high cellular density, growth rate 0.03 h<sup>-1</sup>), the growth rate prediction is on average 0.015 h<sup>-1</sup> higher than the measured rate. In case 2 (low cellular density, growth rate 0.03 h<sup>-1</sup>), the growth rate prediction is on average 0.004 h<sup>-1</sup> smaller than the measured rate. In case 3 (low cellular density, growth rate 0.06 h<sup>-1</sup>), the growth rate prediction is on average 0.0005 h<sup>-1</sup> larger than the measured rate. Furthermore, highest variation in the growth rate predictions is detected for case 2.

The assessment of this *T. reesei* model's capability to predict protein production was carried out separately. The model was successfully used to predict production of native cellulases and of heterologous proteins in [17], where the experimental data used for the simulations is also thoroughly discussed.

### Discussion

The improvements to the CoReCo algorithm include the introduction of a better reaction database, update of the reaction scoring scheme, and inclusion of reaction directionalities into the reconstruction step. A new post-processing step adds exchange reactions, two optional biomass reactions, reaction bounds and the additional manually curated reactions (P/O ratio and required lipid reactions) to the models. With these additions the models are usable in simulations directly after being reconstructed.



**Fig. 5** Growth rate simulations. On the y-axes, the predicted growth rate (h<sup>-1</sup>) and on x-axes **a** measured growth rate (h<sup>-1</sup>) and **b** measured carbon source, i.e., lactose consumption rate (mmol/(g CDW h)) are shown. The data points are marked with the fermentation identifiers (i.e., F32 is fermentation number 32) and surrounded by a box, colored to indicate the case corresponding to the fermentation. See further details from text

### The quality of the reconstructed models

To successfully model cultivations, including protein production experiments, the model needs to have all the relevant biochemical pathways included. Amino acids and energy metabolism being especially relevant for protein production. It is also necessary to be able to accurately model cell growth during the bioprocess. To evaluate the quality of the models, and to assess the overall improvement of the database and the algorithmic updates to the simulation capabilities of CoReCo models, growth rate and individual biomass compound yields were simulated. Previously published CoReCo fungal models required trace amounts of eight different carbon containing compounds in the growth media of computational experiments to successfully simulate growth on the main carbon source, i.e., glucose. Even with these supplements only 17 out of 49 species could predict growth. With improvements presented in this work, 38 out of 56 species can grow and no other carbon containing compound, except glucose, is required in the medium. Furthermore, the biomass function of *T. reesei*, based on the measurements presented in this paper, contains ten more compounds than the previously used yeast biomass equation. The predicted growth rate of the new *T. reesei* model differs only by  $0.004 \text{ h}^{-1}$  from the simulated Yeast 7.6 growth rate. As measured in experiments, the maximal growth rate of *T. reesei* is lower than that of *S. cerevisiae*, but this is likely to result from other factors than stoichiometry.

The improvement presented in this work on the reaction database and the CoReCo algorithm has affected drastically the quality of the models (see a quantification of the effect of these improvements in Table 5). In previously published CoReCo fungal models [21], stoichiometrically unfeasible yields of biomass components were common. In average, 85% of the biomass compounds produced by the old models exhibited some level of stoichiometrically unrealistic yields, i.e., slightly more carbon was produced than was available from the carbon source. With the improvements presented in this work, for 3080 cases (55 compounds in 56 species), no stoichiometrically unfeasible yields were detected (see last row in Table 5). The number of reactions needed to fill the gaps in the networks (i.e., the number of low-scoring reactions) after adding the high-scored reactions was higher in the old models than in the new ones. In addition, the number of dead end metabolites and dead end reactions were higher in the old models compared with the new ones. We believe that the lack of reactions connecting some metabolites in the reaction database was forcing the algorithm to

find alternative paths containing a major number of low scored reactions.

Furthermore, in comparison to Yeast 7.6 (Fig. 4) the biomass component yields of current CoReCo models appear very similar. The differences to Yeast 7.6 yields are likely to stem from problems in the reaction database. Hence, problematic cases like ethanolamine phosphate and choline phosphate production will give information for the starting point for the next round of reaction database improvement. However, real metabolic differences between the species could also be present.

When comparing the reconstructed fungal models with the Yeast 7.6 model as a benchmark, the *T. reesei* model appears to perform slightly better than our other reconstructed models. For example, the Yeast 7.6 model yields 11 units of cysteine from ten units of glucose, while *T. reesei* yields 14 units of cysteine and all other models 20 units, respectively. Similar issues can be seen for methionine, threonine and glycerol-3-phosphate. This is likely due to the fact that the reaction bounds assigned to the reaction database were evaluated based on the performance of the *T. reesei* model. Hence, although all the models are produced with the same automatic process, reaction database curation is slightly biased towards producing a functional *T. reesei* model.

To assess the predictive capabilities of the CoReCo built *T. reesei* model, FBA was performed to simulate growth with experimentally determined rate data from protein production chemostat cultivations from [48] as input. The correlation between measured and predicted growth rates was good, but overall the model predicted a higher growth rate than what was measured (Fig. 5a). This is likely due to the fact that the energetically demanding protein production was not considered in this simulation. Protein production prediction has been made using the *T. reesei* model produced in this manuscript. The work made in [17] highlights with numerous examples how useful stoichiometric modeling can be for enzyme production. In particular, it was found both experimental and modeling based evidence for issues in sulfur assimilation.

Notably, in the condition where highest protein production occurs, the predictions show the highest variation. In this condition, the experimentally determined carbon source, i.e., lactose, consumption rate has lowest correlation to the measured growth rate. This discrepancy could stem from cellular regulatory choices of balancing the metabolism either towards growth or protein production. Further dissection of such discrepancies could lead to completely new insight on how to improve protein production.

**Table 3 The coefficients (mmol/g CDW) used in the biomass equation of the *T. reesei* model**

Kegg Id	Compound name	Coefficient
C00133	Alanine	-0.382 <sup>a</sup>
C00062	Arginine	-0.253 <sup>a</sup>
C00152	Asparagine	-0.141 <sup>a</sup>
C00049	Aspartate	-0.233 <sup>a</sup>
C00097	Cysteine	-0.053 <sup>a</sup>
C00064	Glutamine	-0.161 <sup>a</sup>
C00025	Glutamate	-0.248 <sup>a</sup>
C00037	Glycine	-0.289 <sup>a</sup>
C00135	Histidine	-0.098 <sup>a</sup>
C00407	Isoleucine	-0.193 <sup>a</sup>
C00123	Leucine	-0.362 <sup>a</sup>
C00047	Lysine	-0.202 <sup>a</sup>
C00073	Methionine	-0.092 <sup>a</sup>
C00079	Phenylalanine	-0.151 <sup>a</sup>
C00148	Proline	-0.243 <sup>a</sup>
C00065	Serine	-0.322 <sup>a</sup>
C00188	Threonine	-0.233 <sup>a</sup>
C00078	Tryptophan	-0.061 <sup>a</sup>
C00082	Tyrosine	-0.113 <sup>a</sup>
C00183	Valine	-0.254 <sup>a</sup>
C06424	Myristic acid	-0.00015 <sup>b</sup>
C08362	Palmitoleic acid	-0.00043 <sup>b</sup>
C06427	$\alpha$ -Linolenic acid	-0.01 <sup>b</sup>
C00219	Arachidic acid	-0.00013 <sup>b</sup>
C08320	Lignoceric acid	-0.00014 <sup>b</sup>
C00249	Palmitic acid	-0.026 <sup>c</sup>
C01530	Stearic acid	-0.003 <sup>c</sup>
C00712	Oleic acid	-0.008 <sup>c</sup>
C01595	Linoleic acid	-0.061 <sup>c</sup>
C01694	Ergosterol	-0.007 <sup>d</sup>
C00093	Glycerol-3-P	-0.04 <sup>e</sup>
C00588	Phosphocholine	-0.014 <sup>e</sup>
C00346	Phosphoethanolamine	-0.007 <sup>e</sup>
C00031	D-Glucose	-0.385 <sup>f</sup>
C00140	N-acetyl-D-glucosamine	-0.948 <sup>f</sup>
C00360	Deoxyadenosine monophosphate	-0.007 <sup>g</sup>
C00239	Deoxycytidine monophosphate	-0.008 <sup>g</sup>
C00362	Deoxyguanosine monophosphate	-0.008 <sup>g</sup>
C00364	Deoxythymidine monophosphate	-0.007 <sup>g</sup>
C00020	Adenosine-monophosphate	-0.047 <sup>h</sup>
C00055	Cytidine monophosphate	-0.045 <sup>h</sup>
C00144	Guanosine monophosphate	-0.053 <sup>h</sup>
C00105	Uridine monophosphate	-0.045 <sup>h</sup>
C00001	Water	-59.276 <sup>i</sup>
C00002	Adenosine triphosphate	-59.276 <sup>i</sup>
C00008	Adenosine diphosphate	59.276 <sup>i</sup>
C00009	Organic phosphorous	59.305 <sup>i</sup>
C00059	Sulfate	-0.02 <sup>i</sup>
C00255	Riboflavin	-0.001 <sup>i</sup>

**Table 3 continued**

Kegg Id	Compound name	Coefficient
C00010	Coenzyme A	-0.000001 <sup>j</sup>
C00003	Nicotinamide adenine dinucleotide (NAD)	-0.000001 <sup>j</sup>
C00016	Flavin adenine dinucleotide (FAD)	-0.000001 <sup>j</sup>
C00051	Glutathione	-0.000001 <sup>j</sup>
C00101	tetrahydrofolate	-0.000001 <sup>j</sup>
C00575	3',5'-cyclic AMP	-0.000001 <sup>j</sup>
C00096	GDP-mannose	-0.000001 <sup>j</sup>
C01083	$\alpha,\alpha$ -Trehalose	-0.000001 <sup>j</sup>

The coefficients correspond to the measured or estimated molar amounts of the compound in the cells, as described in "Methods" section

<sup>a</sup> The amount of amino acids calculated based on the measured cellular protein and the ratio of amino acids in cellular proteins calculated based on the codon abundance in the RNAseq data of transcriptome

<sup>b</sup> Esterified fatty acid measured using GC-MS

<sup>c</sup> Sum of measured esterified and free fatty acid measured using GC-MS

<sup>d</sup> Measured using GC-MS

<sup>e</sup> The amount estimated to be needed for synthesis of triacylglycerols, phosphatidylethanolamines or phosphatidylcholines (1 mol of glycerol-3-P, phosphoethanolamine or phosphocholine per 1 mol of triacylglycerol, phosphatidylethanolamine or phosphatidylcholine, respectively)

<sup>f</sup> The measured total carbohydrate was assumed to consist of polymers of D-glucose subunits (glucan) and polymers of N-acetyl-D-glucosamine (chitin). Chitin content of the cells was estimated based on the amount of chitin in *Aspergillus oryzae* [8], and the rest of the measured carbohydrate as glucan

<sup>g</sup> The amount of deoxyribonucleotides in DNA calculated based on the cellular DNA amount and the GC content of the genome

<sup>h</sup> The amount of ribonucleotides in RNA was estimated based on the measured total RNA amount the nucleotide ratio in genome region encoding ribosomal 18S-28S pre-rRNA

<sup>i</sup> As in *S. cerevisiae* model iMM904 [42]

<sup>j</sup> Trace amount of the compound was added

## Conclusions

In this paper, we have presented algorithmic improvements to CoReCo, a tool for reconstructing whole-genome metabolic networks. The new models are more streamlined than previously: they have less reactions, and include reaction directionality constraints that render the solutions space for constrained based metabolic model applications smaller. The CoReCo software was updated such that the SBML-models received as output of the reconstruction pipeline are truly simulation ready containing a biomass equation, an objective function, and uptake and export reactions, respectively. We have demonstrated that the models can be used to simulate growth or metabolite production.

The CoReCo methodology was used to create a full-genome metabolic model for *T. reesei*. This model enables simulations of the metabolic aspects of protein production for this commonly used host organism in protein production applications. Simulations with the *T. reesei* model correspond well with experimental data. The applicability of the model for protein production has

**Table 4 Carbon sources used by *T. reesei* found in the literature**

Carbon source	Growth rate (per unit of compound)
$\alpha$ -methyl-D-mannoside	Not found in the model
$\beta$ -methyl-D-glucoside	Not found in the model
Arbutin	0.171
Cellobiose	0.190
D-arabitol	0.080
D-Fructose	0.090
D-Galactose	0.095
D-Glucose	0.090
D-Mannitol	0.095
D-Mannose	0.090
D-Sorbitol	0.100
D-xylose	0.075
Esculin	Not found in the model
Glycerol	0.049
Glycerol-1-monoacetate	Not found in the model
L-arabinose	0.075
L-sorbose	0.095
N-acetyl- $\beta$ -D-glucosamine	0.115
Salicin	0.185
Trehalose	0.190

The exchange reaction for each of these compounds has been opened to allow the model to have one unit of uptake. Growth has been maximized using FBA

already been demonstrated in [17]. The model contains the reconstructed metabolic network, and a biomass equation that has been created based on the measurements of *T. reesei* biomass.

Current results indicate problem areas to which future improvement of the reaction database needs to be directed. With minor database improvement, CoReCo model reconstruction for all available genome sequences

emerges as a tool to discover new stoichiometrically superior production organisms and pathways.

## Methods

### Updated database of metabolic reactions

Compounds were collected from YMDB, HMDB, ChEBI, KEGG Compound, and Rhea databases. Compounds from the different sources were mapped to each other based on database identifiers (KEGG\_CID, ChEBI\_CID, etc) and structure (e.g., InChI\_string). Only metabolites for which the composition could be linked to a molecular data file (.mol file) were included in the updated database since a .mol file for each metabolite was required by the atom-mapping algorithm. Some sources contained only the names of the metabolites, i.e., the metabolites were not linked to any database identifier. If these metabolites could not be reliably mapped to compounds having a structure, the corresponding reactions from that particular source were left out of the updated database. For the purpose of the reconstruction algorithm, a unique identifier was selected to represent each metabolite. Information about the corresponding metabolite identifiers in the different sources was also kept.

A list of unique metabolic reactions was created by replacing the source-specific metabolite identifiers in reaction expressions with the unique metabolite identifiers mentioned above, and subsequently compared to one another. Reactions from different sources were deemed equal if they had the exact same set of participating metabolites. Protons were ignored during the comparison.

Out of the 11,720 unique reactions selected from the sources, 11,201 could be balanced to the level of electrons. The remaining reactions could only be element-balanced, but not electron-balanced. Despite the improvement made in the balancing algorithm, i.e.,

**Table 5 Comparison of the previously reconstructed models [21] with the new models produced by CoReCo in this article**

CoReCo model	<i>T. reesei</i>		<i>S. cerevisiae</i>		Average of all models <sup>a</sup>	
	Pitkänen [21] (%)	This article (%)	Pitkänen [21] (%)	This article (%)	Pitkänen [21] (%)	This article (%)
Dead end reactions	58	44	61	44	59	44
Dead end metabolites	68	59	70	58	69	58
Reactions added during the gap-filling step	37	20	35	22	39	21
Biomass components produced	7	100	7	100	7	89
Biomass components produced*	98*	100	100*	100	91*	89
Biomass components produced with stoichiometrically unrealistic yield	98*	0	64*	0	85*	0

\* When eight extra compounds were added as a source metabolites into the model simulations

<sup>a</sup> The obligate intracellular parasite *Encephalitozoon cuniculi* has been excluded from the averages

balancing the total number of electrons rather than balancing hydrogen atoms and charge, 519 metabolic reactions could not be fully balanced.

Reaction directions were harvested from the source whole-genome metabolic models. When conflicting reaction directions were encountered, the reactions were set to be bidirectional. KEGG- and MetaCyc-derived reactions were also set to be bidirectional, since these databases do not provide reliable information on the directionality of reactions. The reaction directions were iteratively updated until insensible flux distributions were no longer encountered in model simulations. After each update of reaction directions, the models were reconstructed anew using the updated reaction directions.

The reactions describing lipid metabolism in the updated reaction database were improved because various discrepancies with lipid reactions were noticed during tests with the CoReCo-reconstructed models. Metabolites were identified as lipids if they were part of the KEGG BRITE lipid database BRITE08002. Hence, all reactions containing these lipids were removed from the database, unless the source was yeast (version 7) or KEGG.

The yeast community model v. 7 was selected as the 'gold standard,' since the creators of the model [33] describe the lipid metabolism in detail, and provide scripts to update the lipid reactions from the yeast community model v. 6 to v. 7. The lipid reactions from other whole-genome metabolic models were removed from the updated reaction database. Only reactions having lipids with more than 20 carbon atoms needed to be updated.

All reactions entering the updated reaction database were named using identifiers from the source database. When possible, the KEGG reaction identifier was coupled to the reaction name from the source metabolic model; for example, r0226-YCM606-R00086 denotes the reaction r\_0226 from the yeast community model (YCM606) corresponding to KEGG reaction R00086. KEGG and CHEBI identifiers were used as metabolite identifiers. In cases where the metabolite was not found from either source, a new identifier was composed. These CoReCo internal identifiers were called 'Cluster' followed by a number, for example, Cluster7157 (*N*-carbamoylglycine) corresponds to entry 6988657 in PubChem and *N*-CARBAMOYLGLYCINE in MetaCyc.

#### Updates to CoReCo algorithm

As described in the "Results" section, the improvements to the CoReCo algorithm include the introduction of a better reaction database, update of the reaction scoring scheme, and introduction of reaction directionalities in the reconstruction step. In addition to the above changes, a cluster version of the CoReCo algorithm was created to allow the user to run the pipeline in a parallel mode.

The SBML output file of the CoReCo model now includes exchange reactions (based on the yeast community model [42]), reaction bounds (harvested from the whole-genome models and also manually curated in the reconstruction) and two biomass equations. One biomass equation is based on the yeast community model, the other is the experimentally determined *T. reesei* biomass composition. This update renders CoReCo created models immediately suitable for simulations. The models were simulated in R using the sybil package [50] and in Matlab environment using the Cobra toolbox [51].

P/O ratio has been established by coupling the Ferrocycytochrome-c: oxygen oxidoreductase (r0438-YCM606-R00081) and ATP phosphohydrolase (r0226-YCM606-R00086) reactions. The P/O ratio was set to 1.

#### Reference sequence data

The CoReCo reactions scoring scheme is dependent on a reliable sequence to E.C. information. Here, SwissProt E.C. number annotations and gene rules from the input metabolic models were used as the reference. To create a database for protein sequence BLAST, protein sequences of the genes annotated as enzymes in the included metabolic models were combined to the protein sequences from SwissProt. A file containing the relationship between the protein sequence name and the E.C. number for all the sequences was created by combining the E.C. annotations of sequences in UniProt to the E.C. numbers annotated in the included models. In cases where the E.C. number assignment in SwissProt and in a whole-genome model were conflicting, both annotations were used. In cases where the E.C. assignment in Uniprot was in TrEMBL and not in SwissProt, only the E.C. from the source model was used. In some cases, the source whole-genome models included gene annotations for reactions without an E.C. In these cases, a pseudo E.C. number of the same format, i.e., 7.xx.xx.xx, was used instead to provide a coupling between the reaction and the gene information.

#### Sequence data for *T. reesei* and other fungi

Instead of the actual genomes of the input organisms, CoReCo uses the full set of protein sequences from each organism. In each case, the protein sequence data for the species of interest were downloaded from JGI and the FASTA headers modified to have a clear identifier as the first word in the header. For 49 fungi, the FASTA sequences were already collected for the previous CoReCo run [21]. This set was complemented with the protein sets of *Trichoderma asperellum*, *Trichoderma atroviride*, *Trichoderma harzianum*, *Trichoderma longibrachiatum* and *Trichoderma virens*. Since the previous

CoReCo run, an updated genome sequence for *Komagataella (Pichia) pastoris* has become available [52] and was used to replace the previous data for this organism.

The phylogenetic tree of the organisms was computed using the CVtree algorithm [39]. CVtree is an alignment free composition vector tree based method, and hence does not require selection of specific genes for phylogeny reconstruction. The only parameter required by the method is the length K of the oligopeptides, which was set to 7. The K parameter controls the resolution of the method and it is recommended by authors of the method to be set to 6 or 7 for fungi. We used the fully predicted proteomes of 67 fungi, and the choanoflagellate *Monosiga brevicollis* as an outgroup [53] and extracted a subtree for the 56 fungi from this tree.

#### Biomass measurements in *T. reesei*

The biomass composition of *T. reesei* VTT D-00775  $\Delta$  mus53 was analyzed to create the biomass equation in the model of *T. reesei* (The deletion of the gene mus53 is done to help further construction of modified strains by enhancing homologous recombination in the strain construction process.) The strain was cultivated as follows: 400 ml of culture medium 7.6 g/l  $(\text{NH}_4)_2\text{SO}_4$ , 15.0 g/l  $\text{KH}_2\text{PO}_4$ , 2.4 mM  $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ , 4.1 mM  $\text{CaCl}_2 \cdot \text{H}_2\text{O}$ , 3.7 mg/l  $\text{CoCl}_2$ , 5 mg/l  $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ , 1.4 mg/l  $\text{ZnSO}_4 \cdot 7\text{H}_2\text{O}$ , 1.6 mg/l  $\text{MnSO}_4 \cdot 7\text{H}_2\text{O}$ , pH adjusted to 5.2 with KOH, and supplemented with 25 g/l cellobiose was inoculated with  $8 \cdot 10^7$  spores, and cultivated in shake flasks on a rotary shaker (250 rpm) at 28 °C for 3 days. 100 ml of the pre-culture was transferred to Sartorius Q plus bioreactors containing 900 ml of the medium (4.4 g/l  $(\text{NH}_4)_2\text{SO}_4$ , 15.0 g/l  $\text{KH}_2\text{PO}_4$ , 2.64 mM  $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ , 4.5 mM  $\text{CaCl}_2 \cdot \text{H}_2\text{O}$ , 4.1 mg/l  $\text{CoCl}_2$ , 5.5 mg/l  $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ , 1.54 mg/l  $\text{ZnSO}_4 \cdot 7\text{H}_2\text{O}$ , 1.76 mg/l  $\text{MnSO}_4 \cdot 7\text{H}_2\text{O}$ , 25 g/l cellobiose). Cultivation temperature was 28 °C. pH was adjusted to  $4.8 \pm 0.1$  by addition of 15% KOH or 15%  $\text{H}_3\text{PO}_4$ . The dissolved oxygen saturation level in the cultures was >30%, agitation 500–1200 rpm with a tip speed of 1.1 to 2.7 m/s, and a total aeration flow of 0.6 l/min. The cultures off-gas was monitored on-line for  $\text{CO}_2$  and  $\text{O}_2$ , and samples of the cultures were collected at 0, 16, 24, 40, 64, 88 and 112 h. The mycelial samples were separated from the culture supernatant by filtering through Whatmann GF/B filters and washing with an equal volume of water, frozen immediately in liquid nitrogen, and stored at  $-80$  °C for further analysis. Culture supernatant samples were stored at  $-20$  °C. For sugar analytics, 1.5 ml culture supernatant samples were acidified by the addition of 10  $\mu\text{l}$  of 97%  $\text{H}_2\text{SO}_4$  before storing. Analysis of carbon source consumption, growth and protein production in the cultures as well as transcriptome data is described in [17].

The samples collected at 40 h after inoculation of the bioreactors were analyzed for biomass composition. Three replicate cultures were analyzed.

Biomass dry weight of the cultures was measured by filtering and drying the mycelium at 105 °C to constant weight.

Analysis of DNA in the mycelium was analyzed essentially as described by [54]. 2.5 ml of ice cold 0.25 M  $\text{HClO}_4$  was added to the samples of mycelium (10–30 mg of dry weight per assay). The samples were kept in an ice water bath for 30 min with occasional shaking, and then centrifuged. The supernatant was discarded, and the pellet was extracted by resuspending in 1 ml of 0.5 M  $\text{HClO}_4$  by vortex, incubated at 70 °C for 15 min with occasional shaking, and finally the supernatant was separated by centrifugation. The extraction was repeated twice by resuspending the remaining pellet in 0.5 ml of 0.5 M  $\text{HClO}_4$ , vortexing, incubating at 70 °C for 15 min with occasional shaking. Supernatants from each extraction step were combined, the volume adjusted to 2.5 ml with 0.5 M  $\text{HClO}_4$ , and DNA measured using the diphenylamine method. 1–2 ml of sample was mixed with 2 ml of diphenylamine reagent containing acetaldehyde, incubated at 30 °C overnight, and the optical density measured at 600 nm. The result was compared with standards that were treated in the same way.

RNA content of fungal biomass was measured as described by [55]. Mycelium was washed three times by resuspending in 3 ml of cold 0.7M  $\text{HClO}_4$  and centrifuging, after which the mycelium was resuspended in 3 ml of 0.3M KOH and incubated at 37 °C for 60 min with occasional shaking. After cooling to room temperature, the samples were neutralized by adding 1.0 ml 3M  $\text{HClO}_4$ , and centrifuged. The supernatant was collected, and the pellet was washed twice with 4 ml of cold 0.5M  $\text{HClO}_4$ . The supernatants were combined, and the volume was adjusted to 15 ml with 0.5M  $\text{HClO}_4$ . Finally, the samples were cleared by centrifugation, and the absorbance  $A_{260\text{nm}}$  was measured.

Carbohydrate amount in the fungal biomass samples was measured using the phenol method, essentially as described by [54]. Mycelium samples were ground using a mortar and pestle under liquid nitrogen and lyophilized. 20–200  $\mu\text{g}$  of lyophilized cells were resuspended in water. 1 ml of 5% phenol was added to the samples, the standards prepared from glucose, and to the reagent blanks (1 ml of water). 5 ml of concentrated sulphuric acid was added as a stream to the surface of all the tubes, while shaking the tubes simultaneously. The tubes were allowed to stand for 10 min at room temperature, shaken, and placed in a water bath at 25–30 °C for 10–20 min before measuring the absorbance at 488 nm.

The amount of cellular protein was measured as described by [54]. Mycelium samples were ground using a mortar and pestle under liquid nitrogen and lyophilized. Lyophilised mycelium, corresponding to 1–5 mg dry weight, was resuspended in 2 ml of water. 1 ml of 3M NaOH was added, and the samples were transferred to a boiling water bath for 5 min, and cooled in cold (+4 °C) water bath for 5 min. 1 ml 2.5% CuSO<sub>4</sub> was added, the samples were shaken thoroughly, let stand for 5 min, after which the samples were centrifuged, and the absorbance A<sub>555nm</sub> was measured. A reagent blank containing 2 ml of distilled water instead of cell suspension, and a set of standard protein solutions were treated in the same way, including the heating step.

For extraction of lipids, freeze-dried mycelium (5 mg) was resuspended in 200 µl of 15 mM NaCl and spiked with internal standards [triheptadecanoate (50 µg) and heptadecanoic acid (25 µg)]. Extraction of lipids for fatty acid and lipid class analyses was performed with chloroform:methanol (2:1, 1000 µl). After vortexing and 30 min extraction time at room temperature, the samples were centrifuged at 10,000 rpm for 3 min. The extract (lower layer) was separated and evaporated into dryness under nitrogen flow, and dissolved into 1000 µl of petroleum ether (fatty acid samples) or 100 µl of dichloromethane (lipid class samples).

For GC-MS analysis of fatty acids, lipids were transesterified with sodium methoxide by adding 500 µl of 0.5 N NaOMe in MeOH and a couple of boiling stones and incubating the mixture at 45 °C for 5 min. The samples were acidified with 15% NaHSO<sub>4</sub> and the methyl esters as well as free fatty acids were extracted with petroleum ether. The separated petroleum ether layer was evaporated and dissolved into 100 µl of hexane. Fatty acid methyl esters were analyzed on an Agilent 7890A GC combined with an Agilent 5975C mass selective detector. The column was an Agilent FFAP silica capillary column (25 m × 0.2 mm × 0.3 µm). Helium was used as carrier gas with a split ratio of 15:1. The oven temperature programme was from 70 °C (2 min) to 235 °C at a rate of 10 °C/min, total run time was 30 min. The temperatures of the injector and MS source were 220 and 230 °C, respectively. The samples (2 µl) were injected by a Gerstel MPS injection system and the data were collected in EI mode (70 eV) at a mass range of *m/z* 40–600. After analyzing fatty acid methyl esters by GC-MS, the same samples were trimethylsilylated (TMS) to determine free fatty acid (FFA) and sterol contents. Samples were evaporated, dissolved into 30 µl dichloromethane (DCM) and silylated with 25 µl of MSTFA [*N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide] at 80 °C for 20 min. Trimethylsilylated

samples were analyzed by GC-MS on a Restek Rtx<sup>®</sup>-5MS column (15 m × 0.25 mm × 0.25 µm). The split ratio was 20:1 and the oven temperature programme from 70 °C (1 min) to 270 °C at a rate of 10 °C/min, the total run time was 30 min.

The lipid class analyses (HPLC-ELSD) were performed on a Waters Alliance HPLC combined with a Cuno DDL21 evaporative light scattering detector (ELSD). Separation of the lipid classes was carried out on a Waters Spherisorb<sup>®</sup> silica column (5 µm, 150 × 4.6 mm I.D.). The gradient system consisted of (A) MTBE (methyl tert-butyl ether)-tetrahydrofuran (99:1), (B) 2-propanol-DCM (dichloromethane; 4:1) and (C) 2-propanol-water (1:1) containing triethylamine and formic acid (50 µmol). The temperature of the detector was 40 °C and air flow 27 psi. The multigradient system started from 100% A, the proportion of A decreased to 32%, that of B increased to 52% and simultaneously that of the water containing C increased to 16%. Keeping the cycle running continuously enabled stable retention times. The injection volume was 30 µl.

#### Construction of the biomass equation

The coefficients used in the biomass equation of the *T. reesei* model were calculated based on the measurement data on *T. reesei* biomass composition. The coefficients in the biomass equation indicate the concentration of the compound as mmol/g CDW. Additionally, a few coefficients were copied from the biomass equations of *S. cerevisiae* and *A. oryzae* as explained below.

The amount of cellular proteins [0.451 g/ (g CDW)] in *T. reesei* cultures was measured as described above. The molar ratio of amino acids in the proteins was calculated based on the codon abundance in the RNAseq data of transcriptome excluding 31 transcripts encoding secreted proteins that were identified based on 2D-gel analysis [48]. The RNAseq data was from *T. reesei* cultivated in similar conditions as the cultures for biomass measurements [17]. The molar ratios of the amino acids were converted to weight ratios using the formula weight of each amino acid subtracted by the formula weight of a water molecule released in peptide bond formation, and the weight ratios used for calculation of the amount of each amino acid in the cellular proteins, and subsequently in the fungal biomass (g/(g CDW)) (Table 2). The corresponding molar amino acid amounts were used as coefficients in the biomass equation (mmol/(g CDW)) (Table 3).

Esterified and free fatty acids were measured using GC-MS as described above. The amount of triacylglycerols, phosphatidylethanolamines or phosphatidylcholines in the cells was calculated based on the ratio of triacylglycerols : phosphatidylethanolamines :



phosphatidylcholines (0.52 : 0.16 : 0.32 (w/w/w)) determined using HPLC-ELSD and the measured amount of esterified fatty acids in the cells. The measured fatty acid residues were assumed to be distributed equally in the lipid classes, three fatty acid residues in triacylglycerols and two fatty acid residues in phosphatidylethanolamines and phosphatidylcholines. Total amount of the esterified fatty acids was 0.0995 mmol/(g CDW) and the average MW of the fatty acid esters in the pool was 274.2 g/mol, which was used for calculation of the average MW of triacylglycerols, phosphatidylethanolamines or phosphatidylcholines and subsequently for calculation of the total amount of lipids in the three classes. The amount of glycerol-3-*P*, phosphoethanolamine and phosphocholine needed for the synthesis of triacylglycerols, phosphatidylethanolamines and phosphatidylcholines was estimated based on the amount the lipid classes in the cells, assuming consumption of 1 mol of glycerol-3-*P*, phosphoethanolamine or phosphocholine per 1 mol of triacylglycerol, phosphatidylethanolamine or phosphatidylcholine, respectively.

Total carbohydrates were measured as described as above. The amount of chitin (g/(g CDW)) in the cells was estimated to be the same as the measured chitin content of *A. oryzae* [8], but corrected for the different ash content of the biomasses of the two species. In the biomass equation of the *T. reesei* model, the amount of chitin is represented as the amount of the monomeric unit, *N*-acetyl-*D*-glucosamine. The remaining carbohydrates, other than chitin, were presented as glucose units in the model. Furthermore, a trace amount (1e-06 mmol/(g CDW)) of C00096 GDP-mannose and C01083  $\alpha$ , $\alpha$ -trehalose were added for consistency.

The amount of DNA (0.91% (w/w)) was measured as described above. The amount of deoxyribonucleotides in DNA was calculated based on the published GC content of the genome, 52.0% [56].

The amount of RNA in the cells (6.12% (w/w)) was measured as described above. The amount of ribonucleotides in the RNA was estimated based on the nucleotide ratio in 18S-28S pre-rRNA region in the genome (Ensembl, supercontig:GCA\_000167675.2:GL985064:1035205:1040758:-1, reverse strand).

The coefficients for C00001 Water, C00002 Adenosine triphosphate, C00008 Adenosine diphosphate, C00009 Organic phosphorous, C00059 sulfate and C00255 riboflavin copied directly from the *S. cerevisiae* model IMM904 [42]. The coefficients were taken as is without any scaling.

Finally, trace amount (1e-06 mmol/(g CDW)) of typical cofactors were added to the biomass equation to ensure the model's capability to produce these compounds: C00003 Nicotinamide adenine dinucleotide (NAD), C00010 Coenzyme A, C00016 Flavin adenine

dinucleotide (FAD), C00051 Glutathione, C00059 sulfate, C00101 tetrahydrofolate, and C00575 3',5'-cyclic AMP.

### Metabolic model reconstruction

In addition to the reaction database, the CoReCo pipeline required as inputs: (1) genomic data input: genome of the organism of interest and genomes of related species as protein sequences, as well as a phylogenetic tree describing the relationships of the organisms, and (2) reconstruction constraints: source compound list of entities available for the organism to build up its metabolome and biomass through the reactions present and exchange reactions (i.e., metabolites that the organism can take up or secrete).

The basic steps for the reconstruction of the 56 fungi models using CoReCo pipeline were the following:

1. Protein sequences and information about the 56 fungi species were downloaded from JGI.
2. The preprocessing step of the pipeline included adding the protein sequences from the whole-genome models to the file downloaded from Uniprot creating an "augmented SwissProt" as described in "Reference sequence data" section.
3. Two-way BLAST was performed against the "augmented SwissProt" for each of the organisms involved in the reconstruction. The BLAST runs were carried out in parallel under our SGE VTT cluster.
4. GTG was run in parallel under VTT cluster for each of the organisms involved in the model reconstruction.
5. INTERPRO results were downloaded from JGI and reformatted adding the latest GO and E.C. ids. For 49 fungi the INTERPRO results were already computed in [21].
6. A phylogenetic tree was built with CVTree using the protein sequences of all the organisms involved in the model reconstruction.
7. A probabilistic model was built using the results of BLAST, GTG and the phylogenetic tree. Conditional probability distributions were estimated using the E.C. numbers identified with InterProScan in each species as a reference. A score was given to each E.C. for each organism involved in the model reconstruction.
8. Scores were assigned to reactions by copying the score of the E.C. assigned to that reaction. If several E.C.s were annotated to the same reaction, the reaction was assigned the highest score among the scores for the individual enzymes.
9. During the model reconstruction step, the high scoring reactions were added to each model, sequentially

filling the gaps from the source compound list. An atom graph was used to find the pathways from the sources to the added high score reaction. This subgraph contains preferably only high score reactions but low score reactions can be added when needed. This step uses two input parameters:  $\alpha$  (acceptance threshold), a reaction whose cost is below to this threshold will be added to the subgraph and  $\beta$  (rejection threshold), or maximum cost of the complete subgraph. Three different  $\alpha$  parameters were tested for this work: 0.3, 0.4 and 0.5. The last one was selected based on the production of the biomass components by the *T. reesei* model.

10. A SBML model was created including all the reactions chosen during the reconstruction step. In addition, two biomass reactions and exchange reactions were added in a post-processing step.
11. Models were tested to identify problems like unrealistic production of compounds and reaction bounds were curated manually. The pipeline was run again from step 5 until no more problems could be identified.

## Additional file

**Additional file 1.** Figure showing a carbon normalized version of the production of each individual biomass compound using FBA.

## Abbreviations

FBA: flux balance analysis; GTG: global trace graph; SBML: systems biology markup language; E.C. number: enzyme commission number; GO: gene ontology; CDW: cell dry weight; JGI: Joint Genome Institute; SGE: sun grid engine; HPLC-ELSD: high-pressure liquid chromatography-evaporative light scattering detector; GC-MS: gas chromatography-mass spectrometry.

## Authors' contributions

SC implemented the algorithmic improvement to include reaction directionalities in the CoReCo pipeline, made the curation of the models based on the literature and calculated the compound and biomass production using the metabolic models. DB, SC and MO created several input files needed to run CoReCo pipeline. MA curated the models by iteratively changing reactions bounds and re-running the CoReCo pipeline, tested the metabolic models and performed the model simulations. PB, DB and MO created the compound and reaction database used in the network reconstruction pipeline. TP supervised experimental work, collected cultivation data and calculated the biomass components. TSL carried out the lipid analysis. HN carried out the amino acids analysis. DS participated in planning and supervising the cultivations. EP implemented the algorithmic improvement for the reaction scoring in the CoReCo pipeline. MO, MA and MP conceived and designed the study. All authors read and approved the final manuscript.

## Author details

<sup>1</sup> VTT Technical Research Centre of Finland, Tietotie 2, P.O. Box FI-1000, 02044 Espoo, Finland. <sup>2</sup> Department of Computer Science, University of Helsinki, P.O. 68 (Gustaf Hällströmin katu 2b), 00014 Helsinki, Finland.

## Acknowledgements

Aili Grundström and Eila Leino are thanked for their extremely skillful technical assistance.

## Competing interests

The authors declare that they have no competing interests.

## Availability of supporting data

The metabolic models can be downloaded from BIOMODELS database. *T. reesei* metabolic model ID is MODEL1604280024. CoReCo pipeline code is freely available in the GitHub repository (<https://github.com/esaskar/CoReCo>)

## Funding

This work has been supported in part by the European Union FP7 Cooperation Work programme (Grant BIOLEDGE FP7-KBBE-289126 'BIO knowLEDGE Extractor and Modeller for Protein Production') and the Finnish Funding Agency for Innovation TEKES under the Large Strategic Opening project 'Living Factories' (decision number 40128/14).

Received: 9 June 2016 Accepted: 10 November 2016

Published online: 21 November 2016

## References

1. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, Haraldsdottir H, Mo ML, Rolfsson O, Stobbe MD, Thorleifsson SG, Agren R, Bölling C, Bordel S, Chavali AK, Dobson P, Dunn WB, Endler L, Hala D, Hucka M, Hull D, Jameson D, Jamshidi N, Jonsson JJ, Juty N, Keating S, Nookaew I, Le Novère N, Malys N, Mazein A, Papin JA, Price ND, Selkov Sr E, Sigurdsson MI, Simeonidis E, Sonnenschein N, Smallbone K, Sorokin A, van Beek JHGM, Weichart D, Goryanin I, Nielsen J, Westerhoff HV, Kell DB, Mendes P, Palsson BO. A community-driven global reconstruction of human metabolism. *Nat Biotechnol.* 2013;31(5):419–25.
2. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kürdar B, Penttilä M, Klipp E, Palsson BO, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol.* 2008;26(10):1155–60.
3. Caspeta L, Shoaie S, Agren R, Nookaew I, Nielsen J. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials. *BMC Syst Biol.* 2012;6(24):2016. doi:10.1186/1752-0509-6-24.
4. Poolman MG, Miguet L, Sweetlove LJ, Fell DA. A genome-scale metabolic model of arabidopsis and some of its properties. *Plant Physiol.* 2009;151(3):1570–81.
5. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. *Mol Syst Biol.* 2011;7:535.
6. Henry CS, Zinner JF, Cohoon MP, Stevens RL. iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biol.* 2009;10(6):69.
7. Andersen MR, Nielsen ML, Nielsen J. Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol Syst Biol.* 2008;4(178):2016. doi:10.1038/msb.2008.12.
8. Vongsangnak W, Olsen P, Hansen K, Krogsgaard S, Nielsen J. Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. *BMC Genome.* 2008;9:245.
9. David H, özçelik IC, Hofmann G, Nielsen J. Analysis of *Aspergillus nidulans* metabolism at the genome-scale. *BMC Genom.* 2008;9:163.
10. Yoshikawa K, Kojima Y, Nakajima T, Furusawa C, Hirasawa T, Shimizu H. Reconstruction and verification of a genome-scale metabolic model for *Synechocystis* sp. PCC6803. *Appl Microbiol Biotechnol.* 2011;92(2):347–58.
11. Juty N, Ali R, Glont M, Keating S, Rodriguez N, Swat MJ, Wimalaratne SM, Hermjakob H, Le Novère N, Laibe C, Chelliah V. BioModels: content, features, functionality and use. *CPT Pharmacometrics Syst Pharmacol.* 2015;4(2):e3.
12. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2016;44(D1):457–62.
13. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and

- the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2014;42(D1):459–71.
14. Nielsen J, Jewett MC. Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae*. *FEMS Yeast Res.* 2008;8(1):122–31.
  15. Jouhten P, Wiebe M, Penttilä M. Dynamic flux balance analysis of the metabolism of *Saccharomyces cerevisiae* during the shift from fully respirative or respirofermentative metabolic states to anaerobiosis. *FEBS J.* 2012;279(18):3338–54.
  16. Nocon J, Steiger MG, Pfeffer M, Sohn SB, Kim TY, Maurer M, Rußmayer H, Pflugl S, Ask M, Haberhauer-Troyer C, Ortmayr K, Hann S, Koellensperger G, Gasser B, Lee SY, Mattanovich D. Model based engineering of *Pichia pastoris* central metabolism enhances recombinant protein production. *Metab Eng.* 2014;24(129–138):2016. doi:10.1016/j.jymben.2014.05.011.
  17. Pakula TM, Nygren H, Barth D, Heinonen M, Castillo S, Penttilä M, Arvas M. Genome wide analysis of protein production load in *Trichoderma reesei*. *Biotechnol Biofuels.* 2016;9(132):2016. doi:10.1186/s13068-016-0547-5.
  18. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc.* 2010;5(1):93–121.
  19. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol.* 2010;28(9):977–82.
  20. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput Biol.* 2013;9(3):1002980.
  21. Pitkänen E, Jouhten P, Hou J, Syed MF, Blomberg P, Kludas J, Oja M, Holm L, Penttilä M, Rousu J, Arvas M. Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput Biol.* 2014;10(2):1003465.
  22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10.
  23. Heger A, Mallick S, Wilton C, Holm L. The global trace graph, a novel paradigm for searching protein sequence databases. *Bioinformatics (Oxford, England).* 2007;23(18):2361–7.
  24. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong S-Y, Bateman A, Punta M, Attwood TK, Sigrist CJA, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 2015;43(Database issue):213–21.
  25. Pitkänen E, Arvas M, Rousu J. Reconstructing gapless ancestral metabolic networks. In: Fred A, Filipe J, Gamboa H, editors. *Biomedical engineering systems and technologies. Communications in computer and information science.* Berlin: Springer; 2011. p. 126–40. doi:10.1007/978-3-642-29752-6\_10.
  26. Feist AM, Palsson BO. The biomass objective function. *Curr Opin Microbiol.* 2010;13(3):344–9.
  27. Förster J, Famili I, Fu P, Palsson BO, Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 2003;13(2):244–53.
  28. Balagurunathan B, Jonnalagadda S, Tan L, Srinivasan R. Reconstruction and analysis of a genome-scale metabolic model for *Scheffersomyces stipitis*. *Microb Cell Factories.* 2012;11:27.
  29. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, Lee J, Liu P, Mandal R, Krishnamurthy R, Sinelnikov I, Wilson M, Wishart DS. YMDB: the yeast metabolome database. *Nucleic Acids Res.* 2012;40(Database issue):815–20.
  30. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly M-A, Forsythe I, Tang P, Shrivastava S, Jeroncik K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. HMDB: the human metabolome database. *Nucleic Acids Res.* 2007;35(Database issue):521–6.
  31. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 2013;41(D1):456–63. <http://nar.oxfordjournals.org/content/41/D1/D456.full.pdf+html>
  32. Morgat A, Axelsen KB, Lombardot T, Alcántara R, Aimo L, Zerara M, Niknejad A, Belda E, Hyka-Nouspikel N, Coudert E, Redaschi N, Bougueleret L, Steinbeck C, Xenarios I, Bridge A. Updates in Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Res.* 2015;43(D1):459–64.
  33. Aung HW, Henry SA, Walker LP. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind Biotechnol (New Rochelle, NY).* 2013;9(4):215–28.
  34. Heinonen M, Lappalainen S, Mielikäinen T, Rousu J. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J Comput Biol J Comput Mol Cell Biol.* 2011;18(1):43–58.
  35. Mavrouniotis ML. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol Bioeng.* 1990;36(10):1070–82.
  36. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J.* 2008;95(3):1487–99.
  37. Noor E, Bar-Even A, Flamholz A, Lubling Y, Davidi D, Milo R. An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics.* 2012;28(15):2037–2044. <http://bioinformatics.oxfordjournals.org/content/28/15/2037.full.pdf+html>
  38. Noor E, Haraldsdóttir HS, Milo R, Fleming RMT. Consistent estimation of Gibbs energy using component contributions. *PLoS Comput Biol.* 2013;9(7):1–11.
  39. Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res.* 2009;37(Web Server issue):174–8.
  40. Pakula TM, Salonen K, Uusitalo J, Penttilä M. The effect of specific growth rate on protein synthesis and secretion in the filamentous fungus *Trichoderma reesei*. *Microbiology.* 2005;151(1):135–43.
  41. Caspeta L, Shoaie S, Agren R, Nookaew I, Nielsen J. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials. *BMC Syst Biol.* 2012;6(24):2016. doi:10.1186/1752-0509-6-24.
  42. Mo ML, Palsson BØ, Herrgård MJ. Connecting extracellular metabolic measurements to intracellular flux states in yeast. *BMC Syst Biol.* 2009;3(1):37.
  43. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Büthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Novère NL, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kürdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J, Kell DB. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol.* 2008;26(10):1155–60.
  44. Aung HW, Henry SA, Walker LP. Revising the representation of fatty acid glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind Biotechnol.* 2013;9(4):215–28.
  45. Harman GE, Kubicek CP. *Trichoderma* and *gliocladium*: basic biology, taxonomy and genetics. Boca Raton: CRC Press; 2002. p. 95–181.
  46. Mukherjee PK, Horwitz BA, Singh US, Mukherjee M, Schmolli M. *Trichoderma*: biology and applications. CABI. 2013.
  47. Bergès T, Barreau C, Peberdy JF, Boddy LM. Cloning of an *Aspergillus niger* invertase gene by expression in *Trichoderma reesei*. *Curr Genet.* 1993;24(1–2):53–9.
  48. Arvas M, Pakula T, Smit B, Rautio J, Koivistoinen H, Jouhten P, Lindfors E, Wiebe M, Penttilä M, Saloheimo M. Correlation of gene expression and protein production rate—a system wide study. *BMC Genom.* 2011;12(1):616.
  49. Pakula TM. The effect of specific growth rate on protein synthesis and secretion in the filamentous fungus *Trichoderma reesei*. *Microbiology.* 2005;151(1):135–43.
  50. Gelius-Dietrich G, Fritzscheier CJ, Desouki AA, Lercher MJ. sybil—efficient constraint-based modelling in R. *BMC Syst Biol.* 2013;7(1):125.
  51. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgård MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat Protoc.* 2007;2(3):727–38.
  52. De Schutter K, Lin YC, Tiels P, Van Hecke A, Glińska S, Weber-Lehmann J, Rouzé P, Van de Peer Y, Callewaert N. Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nat Biotechnol.* 2009;27(6):561–6.

53. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Sequencing J, Bork P, Lim WA, Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev IV, Rokhsar D. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*. 2008;451(7180):783–8. doi:[10.1038/nature06617](https://doi.org/10.1038/nature06617).
54. Herbert D, Phipps PJ, Strange RE. Chapter III chemical analysis of microbial cells. In: Ribbons JRNADW, editor. *Methods in microbiology*, vol 5, Part B. Cambridge: Academic Press; 1971. p. 209–344. <http://www.sciencedirect.com/science/article/pii/S058095170870641X>. Accessed 2 Feb 2016.
55. Benthin S, Nielsen J, Villadsen J. A simple and reliable method for the determination of cellular RNA content. *Biotechnol Tech*. 1991;5(1):39–42.
56. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, Chapman J, Chertkov O, Coutinho PM, Cullen D, Danchin EGJ, Grigoriev IV, Harris P, Jackson M, Kubicek CP, Han CS, Ho I, Larrondo LF, de Leon AL, Magnuson JK, Merino S, Misra M, Nelson B, Putnam N, Robbertse B, Salamov AA, Schmoll M, Terry A, Thayer N, Westerholm-Parvinen A, Schoch CL, Yao J, Barabote R, Nelson MA, Detter C, Bruce D, Kuske CR, Xie G, Richardson P, Rokhsar DS, Lucas SM, Rubin EM, Dunn-Coleman N, Ward M, Brettin TS. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat Biotechnol*. 2008;26(5):553–60.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

