Genome-scale reconstruction and metabolic modelling of the fast-growing thermophile *Geobacillus* sp. LC300

Emil Ljungqvist, Martin Gustavsson *  
AlbaNova University Center, KTH Royal Institute of Technology, Department of Industrial Biotechnology, SE-106 91, Stockholm, Sweden

**A R T I C L E   I N F O**

**Keywords:**
Genome-scale metabolic modelling  
Flux sampling  
Metabolic engineering  
*Geobacillus*  
Thermophile

**A B S T R A C T**

Thermophilic microorganisms show high potential for use as biorefinery cell factories. Their high growth temperatures provide fast conversion rates, lower risk of contaminations, and facilitated purification of volatile products. To date, only a few thermophilic species have been utilized for microbial production purposes, and the development of production strains is impeded by the lack of metabolic engineering tools. In this study, we constructed a genome-scale metabolic model, an important part of the metabolic engineering pipeline, of the fast-growing thermophile *Geobacillus* sp. LC300. The model (iGEL604) contains 604 genes, 1249 reactions and 1311 metabolites, and the reaction reversibility is based on thermodynamics at the optimum growth temperature. The growth phenotype is analyzed by batch cultivations on two carbon sources, further closing balances in carbon and degree-of-reduction. The predictive ability of the model is benchmarked against experimentally determined growth characteristics and internal flux distributions, showing high similarity to experimental phenotypes.

1. Introduction

Thermophilic microorganisms show great potential as hosts in biorefinery processes producing various biochemicals and proteins. Their high growth temperatures can yield process advantages in large-scale, including increased reaction rates, lowered cooling costs, decreased risk of contaminations and facilitated purification of volatile products (Kananavičiute and Citavičius, 2015). *Geobacillus* is a genus of aerobic, thermophilic, Gram-positive bacteria with typical optimal growth temperatures ranging between 55 and 60 °C, and in some cases up to 75 °C (Ilssein et al., 2015). Due to their capabilities of consuming a diverse range of substrates including hexoses, pentoses, sugar polymers and fatty acids (Hussein et al., 2015; De Maayer et al., 2014), *Geobacillus* show potential as microbial biorefinery hosts. Despite this potential, only a few examples of metabolic engineering of *Geobacillus* for the production of industrial biochemicals have been reported (Cripps et al., 2009; Lin et al., 2014; Bashir et al., 2019; Yang et al., 2020; Daas et al., 2018). The lack of metabolic engineering efforts could be explained by the limited availability of engineering tools for *Geobacillus* and thermophilic organisms in general (Kananavičiute and Citavičius, 2015). Several recent publications report an expansion of this engineering toolbox, such as improving expression plasmids (Reeve et al., 2016; Marcano-Velazquez et al., 2019), discovery and adaptation of native *Geobacillus* CRISPR systems (Harrington et al., 2017; Mougakos et al., 2017) and synthetic promoter and ribosomal binding site libraries for tuning expression and translation (Reeve et al., 2016). In light of the increased availability of metabolic engineering tools, the prospect of implementing *Geobacillus* in biorefineries draws closer.

Genome-scale metabolic models (GEMs) facilitate prediction of a cell’s metabolic fluxes and can as such be used for prediction of different metabolic scenarios. These could be the effect of a certain drug on a human cell or the effect of a gene knock-out in a bacterium. Thus, GEMs can be powerful tools when combined with metabolic engineering efforts to guide the researcher to, for example, the optimal gene target to knock out for an optimized production of a certain product (Rocha et al., 2008). To create a GEM, the proposed genes of an annotated genome are assigned a function based on their homology to a database of potential orthologs. Compared to mesophilic organisms, thermophiles often have small genomes and amino acid sequences are usually short (Panja et al., 2008). Furthermore, thermophilic proteins tend to differ in amino acid composition from their mesophilic counterparts, for example favoring charged residues over polar ones to increase stability through salt bridges (Pica and Graziano, 2016; Argos et al., 1979). This difference in amino acid chain length and composition complicates homology-based
metabolic modeling of thermophilic organisms, since this relies on using similarity of the protein sequence to assign it to a group of orthologs to predict its function. Since the majority of characterized enzymes stem from mesophilic organisms, assigning thermophilic protein sequences can thus be problematic. Despite these difficulties several GEMs of thermophilic organisms exist, including *Clostridium thermocellum* (Thompson et al., 2016; Garcia et al., 2020), *Parageobacillus thermogluco- cosidans* (Ahmad et al., 2017; Mol et al., 2021) and *Geobacillus igicanus* (Kulyashov et al., 2020). *Geobacillus* sp. LC300 (hereafter LC300) is a recently discovered thermophobe showing promise for biorefining, with reported fast growth (doubling times under 30 min) on both xylose and glucose, the two main sugars of lignocellulosic biomass, in defined media (Cordova et al., 2015). Due to high substrate utilization rates, LC300 constitutes an interesting subject for metabolic engineering to rapidly convert substrates to renewable products of interest. Although a central carbon metabolism model of LC300 has previously been published (Cordova et al., 2015), and flux analysis of the central metabolism has been performed (Cordova and Antoniewicz, 2016; Cordova et al., 2017), further insight in the central metabolism is required. For example, the core metabolism models overestimate biomass formation rates by 30%, and analysis of data from these previous reports reveal that this results from that the carbon and degree of reduction balances are still to be closed (Supplementary Table 6). Such metabolic knowledge is key to guiding rational metabolic engineering approaches, and genome-scale metabolic models are important tools in such endeavors (Lee and Kim, 2015).

To this end, we present a homology-based genome-scale metabolic model of LC300. The model has reaction reversibilities set using thermodynamics data at LC300’s reported optimal growth temperature (72 °C), and gene-reaction connections allowing for phenotype predictions based on genomic changes. Additional insight into the LC300 metabolism is provided through batch cultivations on two carbon sources, where gaps in the carbon and electron balances seen in earlier reports are further closed. The predictive ability of the model is benchmarked against experimentally determined growth characteristics and internal flux distributions, showing high similarity to experimental phenotypes.  

2. Materials & methods

2.1. Growth media

For growth in shake-flasks and bioreactors, a previously described LC300 growth medium was used (Swarup et al., 2014). The medium contained (per liter) 0.5 g K2HPO4, 0.5 g KH2PO4, 0.5 g NH4Cl, 0.5 g NaCl, 0.05 g Yeast Extract (Difco laboratories Art. No: 212720), 40 mL 1M Tris-HCl. After autoclaving, the following chemicals were added by sterile filtration: 0.24 g MgSO4 × 7H2O, 0.033 g CaCl2 × 2H2O, 5 mL Wolfe’s Minerals solution, 5 mL Wolfe’s Vitamins solution, and 10 g D-Xylose and D-Glucose as indicated. Wolfe’s Minerals contained (per liter): 0.5 g EDTA, 3.0 g MgSO4 × 7H2O, 0.5 g MnSO4 × H2O, 1.0 g NaCl, 0.1 g FeSO4 × 7H2O, 0.1 g Co(NO3)2 × 6H2O, 0.1 g CaCl2 × 2H2O, 0.1 g ZnSO4 × 7H2O, 0.01 g CuSO4 × 5H2O, 0.01 g AlK(SO4)2 × 2H2O, 0.01 g Na2MoO4 × 2H2O, 0.001 g Na2SeO3, 0.01 g Na2WO4 × 2H2O, 0.022 g NiSO4 × 6H2O. Wolfe’s vitamins contained (per liter): 2.0 mg Folic acid, 10 mg Pyridoxine hydrochloride, 5.0 mg Riboflavin, 2.0 mg Biotin, 5.0 Thiamine hydrochloride, 5.0 mg Nicotinic acid, 5.0 mg Calcium pantothenic acid, 0.1 mg Vitamin B12, 5.0 mg p-Aminobenzoic acid, 5.0 mg Thiotic acid, 900 mg KH2PO4.

2.2. Strains and growth conditions

*Geobacillus* sp. LC300 was a kind gift from Professor Maciek Antoniewicz. The received sample was grown to exponential phase and stored in 1 mL aliquots containing 20% v/v glycerol in –80 °C. Seed cultivations were performed by inoculation of these glycerol working stocks into 5 mL of the aforementioned medium in 50 mL screw cap tubes (Sarstedt, Art. No: 62.547.254), and incubated at 68 °C and 200 rpm in an orbital shaker (ES-80, GrantBio, Shepreth, United Kingdom). When an optical density at 600 nm (OD600) of 0.1 was reached, the cells were transferred into 1 L Erlenmeyer flasks containing 100 mL medium and cultivated to an OD600 of approximately 0.1 before being placed at 4 °C overnight. The next morning, these flasks were transferred back to the incubator (68 °C, 200 rpm), and the bacteria were grown until an OD600 of 0.8 was reached. This procedure allowed for reproducible inoculum growth with minimal lag time after transfer from 4 °C. The exponentially growing cultures were subsequently inoculated to four parallel 1.5 L stainless steel stirred-tank bioreactors (Greta, Belach bioteknik AB, Skogås, Sweden) containing 800 mL cultivation medium to a starting OD600 of approximately 0.05. The temperature of the growth medium was regulated at 68±0.1 °C through automatically recirculating warm water (85 °C) in the bioreactor jacket, and the pH was maintained at 7.0 through automatic titration using 24.5% NH2OH. To ensure sufficient oxygenation, the dissolved oxygen tension was kept above 30% saturation by automatically increasing the stirrer speed (200–700 rpm). Aeration was increased in one step from 0.2 L min−1 to 0.5 L min−1 when the stirrer speed reached its maximum. When required, foaming was reduced by manual addition of antifoam.

2.3. Genome sequencing

Genomic DNA from *Geobacillus* LC300 was isolated using the ThermoFisher genomic DNA purification kit (cat. No. K0512) according to the manufacturer’s protocol. The sample was sequenced by Eurofins genomics INVIE resequencing service. Protein coding sequences containing alterations compared to the LC300 reference sequence (Supplementary Table 8) were manually updated and the resulting modified protein sequences were aligned to the previously published ones using the EMBOSS Needle web server (Madeira et al., 2022) and selected modified protein sequences where analyzed using NCBI BLAST (Johnson et al., 2008) to predict their function (Supplementary Table 8). Updated protein coding sequences based on the variants identified during sequencing can be found in the GitHub repository associated with this article.

2.4. Analyses

Samples of 2 mL were collected from bioreactors every 30 min to follow substrate consumption, metabolite excretion, and biomass formation. Biomass accumulation was monitored through OD600 and cell dry weight measurements. OD600 measurements were performed in a spectrophotometer (Genesys 20, Thermo Scientific) after diluting to OD600 of 0.06–0.25 in saline solution (0.9% w/v NaCl). Cell dry weight (CDW) was determined in triplicates at 5 time-points throughout the cultivations. The CDW samples were collected by centrifugation of 10 mL culture sample at 4500 rpm (Z206A, Hermle, Gosheim, Germany), followed by drying of the cell pellet at 110 °C over night. The OD600: CDW conversion ratio was determined to 1.9, which was then used to convert remaining OD600-values to CDW. Samples for determining metabolite and substrate concentration were centrifuged at 13000 rpm for 5 min in a tablopet microcentrifuge (Biofuge A, Heraeus Sepatech), transferred to new microcentrifuge tubes and then centrifuged again at 13000 rpm for 5 min. The resulting supernatants were analyzed by HPLC (Waters, Milford, MA, USA) with an Aminex HPX-87H column (Bio-Rad, Hercules, CA, USA). The separation was performed at 60 °C with 5 mM H2SO4 as a mobile phase at a flow rate of 0.6 mL min−1. Sugars were quantified using a refractive index detector (Waters) at 410 nm, and organic acids were quantified through absorbance at 210 nm using a diode array detector (Waters). In addition, the degree of cell lysis was estimated by analyzing the total protein content of the supernatants by Bradford assay (Sigma Aldrich, B6916) according to the manufacturer’s protocol.
CO₂ and O₂ concentrations in the off-gas of two of the reactors in each quadruplicate culture experiment were monitored using on-line gas analyzers (1313 Fermentation Monitor, INNOVA AirTech Instruments, Ballerup, Denmark, and Industrial Emissions Monitor 1311, Brüel & Kjær, Nærum, Denmark).

2.5. Calculation of balances

Calculations on carbon and degree of reduction (Heijnen, 1994) balances were performed on the data published by Cordova et al. (Cordova and Antoniewicz, 2016; Cordova et al., 2017), and data from the cultivations described above. The degree of reduction is a measure of the electron-donating capabilities of each compound, where each compound is assigned a value based on its chemical formula (represented by γ), and the sum of the degree of reduction of the in-going compounds must match the sum of the outgoing compounds in any system. For example, the degree of reduction of carbon dioxide is calculated in equation (1). Calculations of degree of reduction and carbon balances for LC300 growing on glucose and xylose are shown in Supplementary Table 6.

\[ γ_{\text{Carbon}} = 4, \quad γ_{\text{Oxygen}} = -2, \quad γ_{\text{CO}_2} = 4 + 2 \times (-2) = 0 \] (1)

2.6. Model generation

A draft model of LC300 was generated using the RAVEN (Reconstruction, Analysis and Visualization of Metabolic Networks) toolbox 2 (v. 2.3.1) (Wang et al., 2018) for MATLAB R2019b, which constructs a model based on homology of protein sequences to orthologs in the KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000) database. The predicted protein coding sequences of LC300 (see GitHub repository) were downloaded from the NCBI genome database (Coordinators, 2016) (NCBI Accession number: ASM119162v1). These sequences were then used as input in RAVEN, which parsed them against the KEGG database lacks information about metabolite charge and uses each metabolite was collected from MetaCyc (Caspi et al., 2013), since metabolites were found in the MetaCyc database (145 metabolites), a script developed by Kam et al. (2017) was used to calculate metabolite formulas and charges automatically. This script was also used to confirm the molecular weight of the biomass component to 1 g mol⁻¹, which results in the flux of the biomass reaction correctly corresponding to the specific growth rate \( \mu \) (h⁻¹) (Chan et al., 2017). One issue in creating models of thermophiles is that reaction reversibilities in common databases are typically annotated for mesophilic growth temperatures. Thus, these are not necessarily valid at the optimum growth temperature of LC300. To alleviate this issue, the reversibility of reactions was determined based on thermodynamic calculations using the eQuilibrator API (Flamholz et al., 2012), using a script developed by Knežek et al. (2019). The reaction conditions were set to pH 7, ionic strength 0.25 M and temperature 345 K. To evaluate reversibility, the oxidation of malate to oxaloacetate was used as a reference. This reaction has a positive \( ∆\Gamma^\text{in} \) of 30 kJ/mol but is still driven in the forward direction by differences in metabolite concentration (Bar-Even et al., 2012). Since intracellular metabolite concentrations are unknown, this value of ±30 kJ/mol was selected as a cut-off for reversibility of each reaction (Bar-Even et al., 2012). Using eQuilibrator, \( ∆\Gamma^\text{in} \) predictions were obtained for 1074 of the 1174 non-exchange reactions in the model (Supplementary Table 4). This reversibility prediction was incorrect for some reactions (Supplementary Table 5), which had to be corrected manually. The remaining reactions not predicted by eQuilibrator were left at their default reversibilities as annotated in the KEGG database.

2.8. Biomass objective function

The biomass objective function (BOF) was mainly based on data from the \( ^{13}\text{C} \) analysis of the LC300 central metabolism performed by Cordova et al. (2017). The BOF was constructed as a lumped reaction of protein, RNA, DNA, cell membrane, cell wall, glycerogen, and metabolic cofactors, along with the estimated necessary growth-associated ATP cost of 40 mmol gDW⁻¹. The protein part of the BOF was created by all 20 amino acids along with the hydrolyzation of 1 mol ATP to 1 mol AMP, and 2 mol GTP to 2 mol GDP per mol of amino acid. The respective concentration of each amino acid was set based on the data from Cordova et al. (2017), with the addition of Arg, Cys, His and Trp which were excluded in the analysis of the befo-mentioned paper. The total amount of these amino acids was set to <1% of the total amount of amino acids, to deviate from the measured amino acid composition as little as possible. The RNA and DNA fractions of the BOF were constructed based on the biomass composition measured by Cordova et al., with RNA constructed from XMPs and DNA from dXMPs. RNA and DNA compositions were based on the GC-content of the LC300 genome. The cell membrane part of the BOF was constructed from fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids. The cell wall part of the BOF was constructed of fatty acids (tetrade-canoic acid, palmitic acid and octadecanoic acid) and 0.5 mol gycerol phosphate per mol of fatty acids.
and the Gurobi optimizer (v. 8.11). Two pFBA runs were performed, using glucose and xylose as carbon sources respectively. For both runs, biomass formation was set as the objective function. Flux sampling was performed using the artificial centering hit-and-run (ACHR) sampler in the COBRA toolbox 3.0 (Heirendt et al., 2019). Prior to executing the sampling algorithm, the maximum biomass formation rate was determined using pFBA, and all blocked reactions in the pFBA solution were removed to simplify the sampled model. The sampling options were set to: 200 steps per point, 1000 points returned, 5000 warmup points.

3. Results

3.1. Model generation

A draft genome-scale metabolic model of LC300 was generated automatically based on protein homology using the RAVEN toolbox. In this draft model, 20 non-spontaneous reactions were found to be missing (Supplementary Table 2), leading to gaps preventing synthesis of essential biomass precursor metabolites. The amino acid sequences of the *G. stearothermophilus* genes responsible for these reactions were manually aligned against the LC300 genome. All these genes either resulted in hits with an E-value of less than $10^{-20}$ or were found to be already annotated in the LC300 genome. Based on this, it was considered likely that LC300 contained the genes necessary for these reactions, which is further supported by previous reports of growth in defined media (Cordova et al., 2015). Thus, the 20 missing reactions were manually added to the model. In line with earlier findings (Kulyashov et al., 2020; Cordova et al., 2015), no gene for 6-phosphogluconolactonase, catalyzing a necessary reaction in the oxidative branch of the pentose phosphate pathway, was identified in the LC300 genome. Nevertheless, its catalyzed reaction has been shown to be spontaneous at room temperature in *vitro* (Hussein et al., 2015), and it was therefore included in the model. Similar to earlier findings (Cordova et al., 2015), candidate biosynthesis pathways for all 20 amino acids were found and added to the model automatically. Furthermore, likely biosynthesis pathways for the estimated essential cofactors (Xavier et al., 2017) were also identified. Many of the common vitamins such as riboflavin and biotin also had complete synthesis pathways. Like *Bacillus subtilis*, riboflavin biosynthesis is presumed to be carried out by enzymes encoded in the *rib* operon (Vitreschak et al., 2002), and biotin biosynthesis appears to be divided in two parts: synthesis of pimelic acid through the fatty acid biosynthesis, and reconstruction of the free fatty acid to a heterocyclic ring (Manandhar and Cronan, 2017). Contrary to earlier findings (Cordova et al., 2015), a complete biosynthesis pathway of vitamin B12 could not be found. While the anaerobic biosynthesis pathway is almost complete, lacking only CblF and CblJ, the aerobic biosynthesis pathway is missing six essential genes (CobG, CobJ, CobF, CobK, CobN and CobQ), suggesting a possible vitamin B12 autotrophy in LC300. Teichoic acid synthesis was also found to lack an essential gene, glycerol-3-phosphate cytidylyltransferase. However, all genes necessary for teichuronic acid could be found, and it was chosen as the major cell wall component in the model instead.

After establishing the initial reaction network from the LC300 genome, exchange reactions and import reactions (Supplementary Table 3) for medium metabolites were added manually to allow simulating the uptake of nutrients necessary for growth. These include exchange and import reactions of the four experimentally validated carbon sources (glucose, xylose, galactose, and mannose), as well as seven additional potential substrates (arabinose, glycerol, sucrose, cellobiose, acetate, palmitate and butyrate). Furthermore, reactions for the electron transport chain (ETC) and ATP synthase were manually added to the model, with the composition of the ETC detailed in the Materials and Methods section. Finally, the biomass objective function was constructed as a lumped reaction of macromolecules consisting of protein, RNA, DNA, cell membrane, cell wall and glycogen, as well as necessary cofactors and ATP. The amount of each component in the biomass function (Fig. 1) was based primarily on the biomass composition reported previously (Cordova et al., 2015), with additional assumptions detailed in the Materials and Methods. The resulting elemental composition of the biomass component was verified to a molecular weight of 1 g mmol⁻¹, ensuring that the flux of the biomass reaction correctly corresponds to the specific growth rate $\mu$ (h⁻¹) (Chan et al., 2017). The final model, *igel604*, contains 604 genes, 1249 reactions and 1311 metabolites. The reaction distribution across the metabolism subsystems can be seen in Table 1.

With the reactions and metabolites of *igel604* established, the next step was to differentiate between reversible and irreversible reactions. This was performed based on thermodynamic calculations using the reported optimal growth temperature of LC300 (72 °C) as input temperature. As a result, the reversibility for 1074 of the 1155 non-exchange reactions in the model could be set according to the calculated thermodynamics. For the remaining 81 reactions, the default reversibilities obtained from KEGG were assumed.

3.2. Tuning the model with experimental data

As a final step of the model generation, the model was fine-tuned to give accurate predictions of the growth rate of LC300. To this end, two quadruplicate bioreactor batch cultivations were performed using glucose and xylose as carbon sources, respectively. The cultivations were performed at 68 °C, and biomass accumulation, substrate consumption and by-product formation were monitored over time (Fig. 2). The maximum biomass-specific consumption and production rates for substrate, acetate, oxygen, and carbon dioxide were determined during the logarithmic growth phase for all cultivations (Table 2). With glucose as carbon source, these rates resulted in gaps in carbon and degree of reduction balances by 17.8% and 6.1% respectively. During the logarithmic growth phase on glucose, acetate was the sole by-product. During the logarithmic growth phase on xylose no by-products were found, although acetate was produced in the early stationary phase. On xylose, these rates resulted in gaps in carbon and degree of reduction of 12.3% and 9.7% respectively. A potential explanation for these gaps could be the absence of cell lysis during the experiments. This hypothesis was tested by analysing the total protein content of culture supernatants. However, no extracellular protein was detected during the exponential phase on either of the sugars, indicating that these gaps could be caused by the presence of unidentified byproducts rather than cell lysis.

To tune the model, parsimonious flux balance analysis (pFBA) was performed while constraining sugar uptake, oxygen uptake and acetate production to the experimentally determined rates, with biomass
formation set as the objective function. Here it was verified that the model supports growth on all the four experimentally validated carbon sources (glucose, xylose, mannose and galactose). In addition, the model supported growth on arabinose, palmitate or butyrate. Of the experimentally validated carbon sources, detailed growth data was available only for glucose and xylose. This data was used to adjust the P/O-ratio of the ETC, the growth-associated maintenance, and the non-growth associated maintenance (q_m), until the model gave accurate estimations of biomass yield. As a result, the P/O ratio on NADH was set to 1.5, the P/O ratio on FADH_2 was set to 1, with the resulting electron transport chain reaction stoichiometries described in the methods section. The q_m was set to 10 mmol ATP g_{DW}^{-1} h^{-1}. An initial estimation of the growth-associated maintenance was set based on literature (Verduyn et al., 1991) and then adjusted as described above, to 40 mmol g_{DW}^{-1} h^{-1}. With these parameters, the predicted growth rate on both glucose and xylose was very close to the experimentally determined values (1.51 h^{-1} vs. 1.56 h^{-1} and 1.00 h^{-1} vs. 1.07 h^{-1} respectively). As a final validation step, iGEL604 was analyzed with the MEMOTE tool (Lieven et al., 2020), showing close to 100% charge and mass balance. The detailed MEMOTE report can be found on the GitHub repository associated with this article.

### 3.3. Comparison of model predictions to experimental flux data

After constructing iGEL604 and fitting it to predict experimental biomass yields, we proceeded to validate the predicted metabolic fluxes by comparison with published ^13^C metabolic flux data for LC300 grown at 72 °C (Cordova and Antoniewicz, 2016; Cordova et al., 2017). A common way of solving the linear equation system comprising GEMs such as iGEL604 and get approximate fluxes of each reaction is pFBA. However, due to the underdetermined nature of GEMs, the flux solution resulting from pFBA is but one of an infinite number of flux solutions inside the solution space. In comparison, flux sampling estimates the viable flux range of each reaction in the network simultaneously by sampling the flux of all reactions multiple times, thereby giving insight in both the range and probability distribution of fluxes, as well as the solution space size (Schellenberger and Palsson, 2009). Thus, flux sampling was performed on the iGEL604 solution space to get a more representative estimate of the predicted fluxes.

This sampling was done using glucose and xylose as carbon sources with the minimum biomass formation rate constrained to above 95% of the maximum obtained by pFBA on the different carbon sources. The predicted central carbon metabolism reaction fluxes (see GitHub repository for data) were normalized to a substrate uptake rate of 100 mmol g_{DW}^{-1} h^{-1}, and compared to the previously reported experimentally determined maximum biomass-specific production and consumption rates for carbon sources glucose and xylose. Growth rates (μ), substrate consumption rates and acetate production rates are the means of quadruplicate cultivations (± standard deviation). Oxygen consumption rates and carbon dioxide consumption rates are the mean of duplicate measurements (± mean deviation).

| Carbon source | μ (h^{-1}) | Maximum biomass-specific consumption/production rate (mmol g_{DW}^{-1} h^{-1}) |
|---------------|------------|--------------------------------------------------------------------------------|
| Source        | Substrate  | Acetate | Oxygen | Carbon dioxide |
| Glucose       | 1.56 (± 0.10) | 16.05 (± 0.70) | 7.11 (± 0.84) | 34.41 (± 3.85) | 24.33 (± 0.33) |
| Xylose        | 1.07 (± 0.06) | 13.61 (± 0.60) | – | 32.74 (± 1.21) | 31.81 (± 2.82) |

### Table 1

Reaction distribution across metabolism subsystems in the model, as annotated in the KEGG database.

| Metabolism subsystem                          | Number of reactions |
|-----------------------------------------------|---------------------|
| Amino acid metabolism                         | 299                 |
| Carbohydrate metabolism                       | 186                 |
| Metabolism of cofactors and vitamins           | 150                 |
| Nucleotide metabolism                         | 138                 |
| Lipid metabolism                              | 97                  |
| Xenobiotics degradation and metabolism         | 86                  |
| Metabolism of terpenoids and polyketides      | 55                  |
| Exchange reactions                            | 55                  |
| Energy metabolism                             | 37                  |
| Biosynthesis of other secondary metabolites   | 36                  |
| Import reactions                              | 20                  |
| Glycan biosynthesis and metabolism            | 7                   |
| Other                                         | 83                  |

### Table 2

Experimentally determined maximum biomass-specific production and consumption rates for carbon sources glucose and xylose. Growth rates (μ), substrate consumption rates and acetate production rates are the means of quadruplicate cultivations (± standard deviation). Oxygen consumption rates and carbon dioxide consumption rates are the mean of duplicate measurements (± mean deviation).

**Fig. 2.** Growth, substrate consumption, and acetate production of LC300 at 68 °C, with glucose (A) or xylose (B) as carbon source. The x-axis represents time from inoculation. The primary y-axis represents concentrations of substrate and acetate, and the secondary y-axis represents cell mass on a log_{10} scale. Data points show the mean of quadruplicate bioreactor cultures, with error bars indicating standard deviation.
determined fluxes (Cordova and Antoniewicz, 2016; Cordova et al., 2017), as visualized in Fig. 3. Overall, the reactions have small standard deviations, indicating a highly constrained solution space under the simulated conditions. One exception to this is the phosphoenolpyruvate to pyruvate conversion, where three separate reactions (utilizing ADP, dADP and dGTP respectively as cofactors) can carry flux. The flux of these reactions and their standard deviation have been combined to correctly represent the total flux going from PEP to PYR, and the standard deviation is increased as a result. Another exception is the conversion of succinyl-CoA to succinate, which can proceed both through succinyl-CoA ligase in the TCA cycle or alternatively using reactions in the lysine biosynthesis pathway. Finally, observed differences between iGEL604 and previous reports (Cordova and Antoniewicz, 2016; Cordova et al., 2017) illustrate the sensitivity of the anaplerotic reactions, and neighbouring fluxes at the pyruvate node, to differences in experimental and model-predicted biomass yields. Nevertheless, for both carbon sources the overall fit is good with $R^2$-values above 0.9. This indicates that iGEL604 is able to predict overall intracellular fluxes well using only external measurements as input.

4. Discussion

In this study we constructed a genome scale metabolic model, iGEL604, of Geobacillus LC300. The biomass composition was constructed based on experimental data, and the reaction kinetics were adapted to match the thermodynamics at the optimum growth temperature of the organism. The final model contains 604 genes, covering roughly 25% of the predicted protein coding fraction of the LC300 genome. Compared to earlier published central carbon metabolism models, iGEL604 offers insight into the biosynthesis pathways of many essential macromolecules. Furthermore, since each enzymatic reaction is coupled to its’ corresponding gene, possible phenotypic effects from gene perturbations can be investigated, allowing for usage of this model in metabolic engineering efforts. The biomass objective function is additionally more complex, consisting of the experimentally determined (Cordova et al., 2015) macromolecule composition as compared to a lumped reaction of central carbon metabolites.

When constrained with experimentally determined metabolite consumption rates, iGEL604 generates good biomass yield predictions with glucose and xylose as carbon sources. In addition, when compared to previously reported $^{13}$C-MFA data, iGEL604 shows great similarity throughout the central carbon metabolism (Fig. 3) and respiration pathways of oxygen and carbon dioxide. Based on this, we conclude that the modelling assumptions made above seem reasonable, and consequently iGEL604 could be used as a realistic estimate of the LC300 metabolism. When glucose is used as carbon source, flux through the pentose phosphate pathway is especially high with 40% of the catabolic flux, compared to Escherichia coli where approximately 25% of catabolic flux goes through this pathway (Gonzalez et al., 2017). It is likely that a high flux through here is needed when growing at the high maximum growth rate of LC300 due to the high demand for NADPH. Previous reports (Cordova et al., 2017) and experimental data from this study shows acetate to be the major overflow metabolite of LC300 when grown on glucose, which is also predicted by iGEL604.

Several assumptions have been made when constructing this genome scale model. For example, we chose teichuronic acid as a major part of the cell wall rather than teichoic acid as is the case for B. subtilis. The reason for this choice is the lack of an identified glycerol-3-phosphate cytidylyltransferase in LC300, an enzyme necessary for the synthesis of the teichoic acid backbone. Previous studies have shown that the cell wall of some G. stearothermophilus strains lacks teichoic acid (Sutow and Welker, 1967). Furthermore, four amino acids (Arg, Cys, His, Trp) were missing from the amino acid composition analysis of LC300 (Cordova et al., 2015). These amino acids are obviously still necessary for growth, and they were included in the protein fraction of the biomass component, with their combined amount corresponding to 1% of the total amino acid content. Branched chain fatty acids were not included in the biomass composition analysis (Cordova et al., 2015) and as a result, were also excluded in the biomass composition of iGEL604, despite being shown to generally be part of the fatty acid composition of Geobacilli and related thermophiles (Cho and Salton, 1966; Ljungdahl, 1979). This simplification should not have a major impact on the growth.

Fig. 3. Comparison of estimated fluxes in the central carbon metabolism by Cordova et al. (Cordova and Antoniewicz, 2016; Cordova et al., 2017) and predicted fluxes by flux sampling of iGEL604. Each point represents a reaction with its experimentally determined flux (Cordova and Antoniewicz, 2016; Cordova et al., 2017) on the x-axis and the mean of the flux sampling of iGEL604 on the y-axis. The standard deviation of each reaction is visualized as error bars. Fluxes are scaled to a flat distribution along the y-axis. (A) Flux sampling with glucose as carbon source, constraining glucose uptake and oxygen uptake to experimentally determined values. Red dots indicate reactions of metabolites at branch-points in the metabolism (Glucose-6-phosphate, Pyruvate, Acetyl-CoA and Isocitrate). (B) Flux sampling with xylose as carbon source, constraining xylose and oxygen uptake to experimentally determined values. Purple dots indicate reactions of metabolites at branch-points in the metabolism. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
predictions of the model, but is an important aspect to consider if performing engineering for e.g. fatty acid-derived biofuels. It is also unclear which cofactors are necessary for bacterial growth, but we consider the ones presented by Xavier et al. to be a good estimate based on their consensus with earlier reports (Xavier et al., 2017). The non-growth and growth associated maintenance and P/O-ratio are additional model uncertainties. These were estimated by aligning model predictions with experimental biomass yields, and further experimental characterization of LC300 using e.g. chemostat conditions would be required to validate these parameters and enable accurate predictions of growth phenotypes outside batch conditions at maximum substrate uptake rates. Upon analysis of biosynthesis pathways in iGEL604, a possible vitamin B12 auxotrophy was discovered. In Parageobacillus thermoglucosidasius NCIMB 11955 a hybrid pathway of vitamin B12 biosynthesis has been suggested (Mol et al., 2021), where enzymes from the anaerobic pathway can perform the reaction steps where the aerobic enzymes are missing and vice versa. However, the enzymes converting precorrin 4 to precorrin 5 (ChlF/CebM), precorrin 6A to precorrin 6B (ChlJ/CobK), and adenosyl cobyrinate diamide to adenosyl cobyrinate hexamidine (ChlP/CobQ) are either missing or found wdisrupted in both the aerobic and anaerobic branch in the LC300 genome. Thus, a hybrid pathway is unlikely to rescue vitamin B12 biosynthesis in LC300, but experimental validation is needed to confirm this hypothesis.

For experimental validation of the model, bioreactor cultivations were initially attempted at the previously reported optimum temperature of 72 °C (Cordova et al., 2015). However, this reproducibly produced no growth in temperature-calibrated bioreactors (data not shown). Upon reexamining the flask pre-cultures it was discovered that an incubator air temperature set point of 72 °C resulted in a temperature of 69 °C in the cultivation medium. Upon increasing the incubator set point to achieve 72 °C in the shake flask medium, no growth was again observed. Present, the cause of this growth temperature discrepancy compared to previous reports is unknown, but a temperature set point of 68 °C was chosen for model validation to ensure reproducible growth. Due to the observed phenotypic difference, a genomic DNA sample was resequenced to verify that the cultivated bacteria was indeed LC300 and not a different thermophilic contaminant. The resulting assembly covered 96% of the published genome with a coverage of ≥120x (98% coverage ≥2x), identifying 31 intragenic frameshifts with a mutation frequency of ≥95% (Supplementary Table 8). Most frameshifts were found in regions of nucleotide repeats, commonly causing mis-reads in genome sequencing and error in DNA replication. Upon analysis of the affected coding sequences, possible functional changes were difficult to determine since most affected proteins are of unknown function. Detailed information can be found in Supplementary Table 8 and in the GitHub repository for iGEL604. Considering the high similarity of the genomic sequences, we conclude that the studied bacterium is indeed Geobacillus LC300.

Previously reported cultivation data (Cordova and Antoniewicz, 2016; Cordova et al., 2017) has left a gap in the carbon and degree of reduction balances, and the core metabolism models consequently over-estimate the biomass yield with approximately 30% (Supplementary Table 6). In our data from cultivating LC300 at 68 °C, the missing carbon and degree of reduction in the observed products is significantly reduced, but the balances are still not completely closed. On both carbon sources, the carbon balance has a larger gap than the degree of reduction balance, which is likely caused by inaccurate off-gas measurements. The remaining gaps could indicate the presence of currently unidentified byproduct(s) from the LC300 metabolism, for example volatile compounds. If these byproducts are produced in higher quantities at increased temperatures, they could contribute to the larger balance gaps in previous reports.

5. Conclusions

In conclusion, we have constructed a genome scale model of the thermophilic bacterium Geobacillus LC300, with reaction reversibilities following thermodynamics at the reported optimum growth temperature. The model follows the LC300 metabolism closely and allows for reliable predictions of growth, metabolite consumption and production corresponding to experimental data on both glucose and xylose. This model presents a tool to further characterize the metabolism and accelerate the metabolic engineering efforts of the Geobacillus genus and shorten the route for the implementation of Geobacilli as biorefinery hosts.

Author contributions

Emil Ljungqvist: Methodology, Software, Investigation, Data curation, Formal analysis, Visualization, Writing - Original draft; Martin Gustavsson: Conceptualization, Software, Supervision, Writing - review and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors thank Professor Maciek Antoniewicz for kindly giving a sample of Geobacillus sp. LC300, Dr. Gustav Sjöberg for valuable discussions and advice on the modeling and software development, and Prof. Dr. Antonius van Maris for input during revision and writing. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mec.2022.e00212.

References

Ahmad, A., Hartman, H.B., Krishnakumar, S., Fell, D.A., Poolman, M.G., Srivastava, S., 2017. A Genome Scale Model of Geobacillus thermoglucosidasius (CG5-Y993) reveals its biotechnological potential on rice straw hydrolysate. J. Biotechnol. 251 https://doi.org/10.1016/j.jbiotec.2017.03.031.

Argons, P., Rossmann, M.G., Grus, U.M., Zubler, H., Frank, G., Tratschin, J.D., 1979. Thermal stability and protein structure. Biochemistry 18 (25), 5698-5703. http://doi.org/10.1021/bi00592a028.

Bar-Even, A., Flamholz, A., Noor, E., Milo, R., 2012. Thermodynamic constraints shape the structure of carbon fixation pathways. Biochem. Biophys. Acta Bioenerg. 1817 (9), 1646-1659. https://doi.org/10.1016/j.bbabio.2012.05.002.

Bashir, Z., Sheng, L., Anil, A., Lalı, A., Minton, N.P., Zhang, Y., 2019. Engineering Geobacillus thermoglucosidasius for direct utilisation of holocellulose from wheat straw. Biotechnol. Biofuels 12, 199. https://doi.org/10.1186/s13068-019-1540-6.

Caspi, R., Altman, T., Billington, R., et al., 2013. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 42 (D1), D459-D471. https://doi.org/10.1093/nar/gkt1103.

Chan, S.H.J., Cai, J., Wang, L., Simons-Senftle, M.N., Maranas, C.D., 2017. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. Bioinformatics 33 (22), 3603-3609. https://doi.org/10.1093/bioinformatics/btx453.

Cho, K.Y., Salton, M.R.J., 1966. Fatty acid composition of bacterial membrane and wall lipids. Biochim. Biophys. Acta Lipids Lipid. Metabol. 116 (1), 73-79. https://doi.org/10.1016/0005-2766(66)90093-2.

Coordinators, N.R., 2016. Database resources of the national center for biotechnology information. Nucleic Acids Res. 44 (D1), D7-D19. https://doi.org/10.1093/nar/gkv1296.

Cordova, L.T., Antoniewicz, M.R., 2016. 13C metabolic flux analysis of the extremely thermophilic, fast growing, xylose-utilizing Geobacillus strain LC300. Metab. Eng. 33, 148-157. https://doi.org/10.1016/j.men.2015.06.004.
