gapseq: Informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models

Johannes Zimmermann[1], Christoph Kaleta[1], and Silvio Waschina[1,2]

[1]Research Group Medical Systems Biology, Institute of Experimental Medicine, Christian-Albrechts-University Kiel (UKSH Campus)
[2]Nutriinformatics, Institute for Human Nutrition and Food Science, Christian-Albrechts-University Kiel

March 20, 2020

Abstract

Microbial metabolic processes greatly impact ecosystem functioning and the physiology of multi-cellular host organisms. The inference of metabolic capabilities and phenotypes from genome sequences with the help of prior biomolecular knowledge stored in online databases remains a major challenge in systems biology. Here, we present `gapseq`: a novel tool for automated pathway prediction and metabolic network reconstruction from microbial genome sequences. `gapseq` combines databases of reference protein sequences (UniProt, TCDB), in tandem with pathway and reaction databases (MetaCyc, KEGG, ModelSEED). This enables the statistical prediction of an organism's metabolic capabilities from sequence homology and pathway topology criteria. By incorporating a novel LP-based gap-filling algorithm, `gapseq` facilitates the construction of genome-scale metabolic models that are suitable for metabolic phenotype predictions by using constraint-based flux analysis. We validated `gapseq` by comparing predictions to experimental data for more than $1,000$ bacterial organisms comprising over $10,000$ phenotypic traits that include enzyme activity, energy sources, fermentation products, and gene essentiality. This large-scale phenotypic trait prediction test showed, that gapseq yields an overall accuracy of 80% and thereby outperforming other commonly used reconstruction tools. Furthermore, we illustrate the application of `gapseq`-reconstructed models to simulate biochemical interactions between microorganisms in multi-species communities. Altogether, `gapseq` (`https://github.com/jotech/gapseq`) is a new method that improves the predictive potential of automated metabolic network reconstructions and further increases their applicability in biotechnological, ecological, and medical research.

1 Introduction

> Anything you have to do repeatedly may be ripe for automation.
>
> — Doug McIlroy

Metabolism is central for organismal life. It provides metabolites and energy for all cellular processes. A majority of metabolic reactions are catalysed by enzymes, which are encoded in the genome of the respective organism. Those catalysed reactions form a complex metabolic network of numerous biochemical transformations, which the organism is presumably able to perform [21].

In systems biology, the reconstruction of metabolic networks plays an essential role, as the network represents an organism's capabilities to interact with its biotic and abiotic environment and to transform nutrients into biomass. Mathematical analysis has shown great potential for dissecting the functioning of metabolic networks on the level of topological, stoichiometric, and kinetic models [79], which together provide a wide array of methods [47]. Although different microbial metabolic modelling approaches exist, they can be summarised by a theoretical framework that provides a unifying view on microbial growth [38]. Metabolic models not only have demonstrated their ability to predict phenotypes on the level of cellular growth and gene knockouts, but also provide potential molecular mechanisms in form of gene and reaction activities, which can be validated experimentally [87]. Due to this predictive potential, genome-scale metabolic models have been applied to identify metabolic interactions between different organisms [1, 32, 44, 80, 96], to study host-microbiome interactions [33, 64, 95], to predict novel drug targets to fight microbial pathogens [55, 85], and for the rational design of microbial genotypes and growth-media conditions for the industrial production or degradation of biochemicals [59, 66]. Furthermore, recent advances in DNA-sequencing technologies have led to a vast increase in available genomic- and metagenomic sequences in databases [48], which further expands the applicability of genome-scale metabolic network reconstructions.

The reconstruction of metabolic networks links genomic content with biochemical reactions and therefore depends on sequence annotations and reaction databases, which are both crucial for overall network quality [83, 92]. A general problem in reconstructing metabolic networks occurs by an incorrect representation of the organism's physiology. First, inconsistencies in databases can lead to an incorporation of imbalanced reactions into the metabolic network, which may become responsible for incorrect energy production by futile cycles [83]. Second, many genes are lacking a functional annotation due to a lack of knowledge [7] and, thus, also the gene products cannot be integrated into the metabolic networks, which potentially lead to gaps in pathways. Third, the gap-filling of metabolic networks is frequently done by adding a minimum number of reactions from a reference database that facilitate growth under a chemically defined growth medium [34, 63, 84]. Such approaches miss further evidences potentially hidden in sequences and are biased towards the growth medium used for gap-filling. And fourth, the validation of predictions made by metabolic networks is so far only performed with smaller experimental data sets from model laboratory strains such as *Escherichia coli* K12 or *Bacillus subtilis* 168 and therefore the overall performance of many metabolic models is insufficiently assured.

Genome-scale metabolic network reconstructions are increasingly applied to simulate complex metabolic processes in microbial communities [45]. Such simulations are highly sensitive to the quality of the individual metabolic networks of the community members.

This is because the accurate prediction of fermentation products and carbon source utilisation is crucial for the correct prediction of metabolic interactions since the substances produced by one organism may serve as resource for others [61]. Thus in multi-species communities, the metabolic fluxes of organisms are intrinsically connected, which can lead to error propagation when one defective model affects otherwise correctly working models.
In this work, we present `gapseq` a novel method for pathway analysis and metabolic network reconstruction. The pathway prediction is based on multiple biochemistry databases that comprise information on pathway structures, the pathways' key enzymes, and reaction stoichiometries. Moreover, `gapseq` constructs genome-scale metabolic models that enable metabolic phenotype predictions as well as the application in simulations of community metabolism. Models are constructed using a manually curated reaction database that is free of thermodynamically infeasible reaction cycles. As input, `gapseq` takes the organism's genome sequence in FASTA format, without the need for an additional annotation file. Topology as well as sequence homology to reference proteins inform the filling of network gaps, and the screening for potential carbon sources and metabolic products is done in a way that reduces the impact of growth medium definitions. Finally, we used large-scale experimental data sets to validate enzyme activity, carbon source utilisation, fermentation products, and gene essentially.

2 Material and methods

2.1 Program overview & source code availability

The source code is accessible and maintained at `https://github.com/jotech/gapseq`. The program is called by `./gapseq`, which is a wrapper script for the main modules. Important program calls are `./gapseq find` (pathway and reaction finder), `./gapseq find-transport` (transporter detection), `./gapseq draft` (draft model creation), `./gapseq fill` (gap-filling), or `./gapseq doall` to perform all in line. When ever necessary, method sections directly refer to config, data and source code files from the gapseq package, which contains the main subfolders `src/` with source code files and `dat/`, which contains databases and also the sequence files in `dat/seq/`. Figure 1 shows an overview of the different `gapseq` modules.

2.2 Pathway and sequence databases

Pathways are considered as a list of reactions with enzyme names and EC numbers. Pathway definition were obtained from MetaCyc [13], KEGG [39], and ModelSEED [34]. For MetaCyc, PathwayTools [40] was used in combination with PythonCyc to obtain pathway definitions [78] (`src/meta2pwy.py`). Information on Kegg pathways were retrieved directly from the KEGG homepage: reactions (`http://rest.kegg.jp/list/reaction`), and EC numbers (`http://rest.kegg.jp/link/pathway/ec`) and further processed (`src/kegg_pwy.R`). In case of ModelSEED, subsystem definition were obtained from the homepage: `http://modelseed.org/genomes/Annotations` (`src/seed_pwy.R`).
In addition, manual defined and revised pathways are stored in the file `dat/custom_pwy.tbl`.
Sequence data needed for pathway prediction were downloaded from UniProt [82] for each reaction identified by EC number, enzyme name, or gene name. Both reviewed and unreviewed sequences are considered and stored as clustered UniPac sequences (`src/uniprot.sh`).

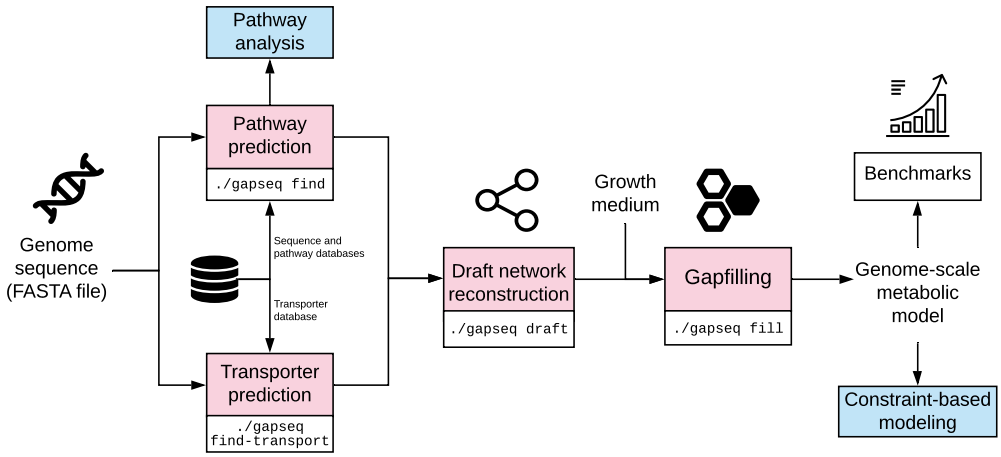

Figure 1: Chart showing the main components and workflow of `gapseq`. Free icons were used from `https://www.flaticon.com` (creators: Freepik, Gregor Cresnar, Freepik, Smashicons).

To increase the sequence pool for a given reaction, alternative EC numbers from BRENDA [37] and from the Enzyme Nomenclature Committee `https://www.qmul.ac.uk/sbcs/iubmb/enzyme/` are integrated (`src/altec.R`, `dat/brenda_ec.csv`).

2.3 Pathway prediction

For each pathway selected from a pathway database (MetaCyc, KEGG, ModelSEED, custom), `gapseq` searches for sequence evidence and a pathway is defined as present if enough of its reactions were found to have sequence evidence. In more detail, sequence data (section 2.2) is used for homology search by *tblastn* [12] with the protein sequence as query and the genome as database. By default, a bitscore $\geq$ 200 and a coverage of at least 75% is needed to make a hit. For certain reactions, the user can define additional criteria, for example an identity of $\geq$ 75% (`dat/exception.tbl`). In case of protein complexes with subunits, a more complex procedure is followed (section 2.4). Spontaneous reactions, which do not need an enzyme, were set to be present in any case. In general, a pathway is considered to be present if at least 80% of the reactions are found (`completenessCutoffNoHints` threshold). This pathway completeness threshold is lowered for pathways in following cases:

1. If the pathway contains key reactions, as it is defined for some pathways in MetaCyc, and all key reactions are found, then `completenessCutoff` of the total reactions needed to be found. We used a value of 2/3 for this threshold.

2. In the cases in which no sequence data is available for specific reactions, the status of the reactions is set to "vague" and these reactions do not count as missing if they account for less than `vagueCutoff` of the total reactions of a pathway. We used a value of 1/3 for this threshold.

The pathway prediction algorithm is implemented in the bash shell script `src/gapseq_find.sh`, which uses GNU parallel [81] and fastaindex/fastafetch from exonerate [75].

2.4 Protein complex prediction

A problem with automatic sequence download for reactions (as FASTA files) comes with protein complexes, for which a simple blast hit may be not sufficient to predict enzyme presence. In `gapseq`, subunits are detected by text matching in the FASTA headers. Search terms are: "subunit", "chain", "polypeptide", "component", and different numbering systems (roman, arabic, greek) are homogenised. To avoid artifacts in text matching, subunits that occur less than five times in the sequence file are not considered, and in cases in which a subunit occurs almost exclusively ($\geq 66\%$) the other entries are not taken into account. All FASTA entries, which could not matched by text mining, or which are excluded because of the coverage, are labeled 'undefined subunit' and do not add to the total amount of subunits. For each recognised subunit, a blast search is done. A protein complex counts as present if more than 50% of the subunits could be found, whereby the presence of 'undefined subunits' tip the balance if exactly 50% of the subunits were found. The text matching with regular expressions is done with R's stringr [90] and biostrings [57] as defined `src/complex_detection.R`. The script is called from within the shell script `src/gapseq_find.sh`.

2.5 Transporter prediction

For transporter search, sequence data from the Transporter Classification Database is employed [70]. In addition, manual defined sequences can be defined in `dat/seq/transporter.fasta`. The sequence set is reduced to a subset of transporters that involve metabolites known to be produced or consumed by microorganisms (`dat/sub2pwy.csv`). Subsequently, the genome is queried by the reduced sequences using *tblastn* [12]. For each hit (default cutoffs: bitscore ≥ 200 and coverage $\geq 75\%$), the transporter type (1. Channels and pores, 2. Electrochemical potential-driven transporter, 3. Primary active transporters, 4. Group translocators) is determined using the TC number mentioned in the FASTA header. A suitable candidate reaction is searched in the reaction database. If there is a hit for a transporter of a substance but no candidate reaction for the respective transporter type can be found, then other transporter types are considered. The transporter search is done by the shell script `src/transporter.sh` that uses GNU parallel [81] and fastaindex/fastafetch from exonerate [75].
Candidate transporters are selected from the reaction database by transporter type and substance name. This is done by text search and is currently implemented only for the ModelSEED namespace. From the ModelSEED reaction database all reaction with the flag $is\_transport = 1$ are taken and the transporter type is predicted by keywords: "channel", "pore" (1. Channels and pores); "uniport", "symport", "antiport", "permease", "gradient" (2. Electrochemical potential-driven transporters); "ABC", "ATPase", "ATP"(3. Primary active transporters); "PTS" (4. Group translocators). If no transporter type could be identified by keywords, additional string matching is done for ATPases, proton/sodium antiporter, and PTS by considering the stoichiometry of the involved metabolites. The transported substance is identified as the substance that occurs on both sides of the reaction. In addition, reactions from the reaction database can be linked manually to substances and transporter types (`dat/seed_transporter_custom.tbl`).
The text matching with regular expressions is done with stringr [90] (`src/seed_transporter.R`).

2.6 Biochemistry database curation and construction of universal metabolic model

For the construction of genome-scale metabolic network models, `gapseq` uses a reactions and metabolite database that is derived from the ModelSEED database [34] as from January 2018. In addition, 30 new reactions and 2 new metabolites were introduced to the `gapseq` biochemistry database (see suppl. table S1). All reactions and metabolites from the database were included for the construction of a full universal metabolic network model; an approach that is also used in CarveMe [50]. We curated the underlying biochemistry database in order to correct inconsistencies in reaction stoichiometries and reversibilities. Inconsistencies were identified by optimising the universal network model for ATP-production without any nutritional input to the model using flux balance analysis. In case of ATP-production, the flux distributions of such thermodynamically infeasible reaction cycles were investigated by cross-checking the involved reactions with literature information, the BRENDA database for enzymes [37] and the MetaCyc database [13]. Stochiometries and reversibilities of erroneous reactions were corrected accordingly. This curation procedure was repeated until no theromodynamically infeasible and ATP-generating reaction cycles were observed.
Hits from the pathway prediction (2.3) and transporter prediction (2.5) are mapped to the `gapseq` reaction database using different common identifiers. A majority of reactions are directly matched via their corresponding Enzyme Commission (EC) system identifier [88] and Transporter Classification (TC) system identifier [70], respectively. For this mapping, also alternative EC-numbers for enzymatic reactions as defined in the BRENDA database [37] are considered. Moreover, the databases used for pathway and transporter predictions often provide cross-links to the reaction's KEGG ID, which is also assigned to most reactions in the `gapseq` database and used to match reactions. Additionally, the MNXref database [6] provides cross links between several biochemistry databases, which `gapseq` also utilises to translate hits from the pathway predictions to model reactions. Finally, a manual translation of enzyme names to model reactions is done for some reactions, which we identified as important reactions but which failed to match between the pathway databases (2.3) and the `gapseq` model reactions using other reaction identifies (`dat/seed_Enzyme_Name_Reactions_Aliases.tsv`). The overall mapping is done by the function `getDBhit()` as defined in `./src/gapseq_find.sh`.

2.7 Model draft generation

A draft genome-scale metabolic model is constructed based on the results from the pathway and transporter predictions (see above). A reaction is added to the draft model if the corresponding enzyme/transporter was directly found or if the a pathway was predicted to be present (i.e. due to pathway completeness and key enzymes) in which the reaction participates. Additionally, spontaneous reactions as defined in the MetaCyc database as well as transport reaction of compounds, which are know to be able to cross cell membranes by means of diffusion, are directly added to every draft model. As part of the draft model construction `gapseq` adds a biomass reaction to the network that aims to describe the composition of molecular constituents that the organism needs to produce in order to form 1 g dry weight (1 gDW) of bacterial biomass. `gapseq` uses the biomass composition definition from the ModelSEED database for Gram-positive (`dat/seed_biomass.DT_gramPos.tsv`) and Gram-negative

bacteria (`dat/seed_biomass.DT_gramNeg.tsv`). If no Gram-staining property is specified by the user, `gapseq` predicts the Gram-staining-dependent biomass reactions by finding the closest 16S-rRNA-gene neighbor using a `blastn` search against reference 16S-rRNA gene sequences from 4647 bacterial species with known Gram-staining properties that are obtained from the PROTRAITS database [10]. The model draft generation is done by the R script `src/generate_GSdraft.R`. Currently, only reactions are added

2.8 Gap-filling algorithm

`gapseq` provides a gap-filling algorithm that adds reactions to the model in order to enable biomass production (i.e. growth) and likely anabolic and catabolic capabilities. The algorithm uses the alignment statistics (i.e. the bitscore) from the pathway- and transporter prediction steps of `gapseq` (see above) to preferentially add reactions to the network, which have the highest genetic evidence. This approach is especially relevant in cases where the sequence similarity to known enzyme-coding genes was close to but did not reach the cutoff value b, which is required for a reaction to be included directly into the draft network. In contrast to the gap-filling algorithms described in previous works [4] and [50], which also use genetic evidence-weighted gap-filling, the gap-filling problem in `gapseq` is not formulated as Mixed Integer Linear Program (MILP) but as Linear Program (LP), and is derived from the parsimonious enzyme usage Flux Balance Analysis (pFBA) algorithm developed by Lewis et al., 2010 [47]. Therefore, the alignment statistics (i.e. bitscore) are translated into weights for the corresponding model reactions and incorporated into the problem formulation:

$$\text{max: } v_j - c \sum_{i \epsilon R_{all}} w_i |v_i| \quad , \tag{1}$$

$$w_i = \begin{cases} w_{min} & b_i \geq u \quad | \quad i \, \epsilon \, R_{draft} \\ (b_i - u) \left(\frac{w_{min} - w_{max}}{u - l} \right) + w_{min} & l \leq b_i < u \\ w_{max} & b_i < l \end{cases}$$

s.t.

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0}$$

$$\boldsymbol{lb} \leq \boldsymbol{v} \leq \boldsymbol{ub}$$

Where R_{all} is the set of all reaction in the universal model, R_{draft} are the reactions, which are already part of the draft network before gap-filling, v_j is the flux through the objective reactions (e.g. biomass production), v_i the flux through reaction i, w_i the weight for reaction i, v the flux vector for all reactions, and c a scalar factor that determines the contribution of the absolute reduction of weighted fluxes to the overall FBA solution (default: $c = 0.001$). Moreover, a maximum weight value w_{max} (default: 100) is assigned if the reaction's highest bitscore is smaller than a threshold l (default: 50). A minimum reaction weight w_{min} (default: 0.005) is assigned to reactions with a bitscore higher than u (default: 200) or if the reactions are already part of the draft model. S is the stoichiometric matrix and lb and ub the lower and upper flux bound vectors. Following the the solution of the LP (1), reactions carrying a flux and which are not part of the draft model are added to the network model. The algorithm is implemented in `src/gapfill4.R`.

2.9 Gap-filling of biomass, carbon sources, and fermentation products

Gap-filling of a draft model in `gapseq` requires only for the first step a user-defined growth medium that is ideally known to support growth of the organism of interest *in vivo*. A set of generic growth media (e.g. LB, TSB, M9) is provided in the folder `dat/medium/`. In addition, the user can provide also a custom-made growth medium definition. The above described gap-filling algorithm is used to improve the generated draft model in four steps.

1. **Biomass production**: To ensure that the model is able to produce biomass under the given nutritional input (medium) the gap-filling algorithm is applied while the objective is defined as the flux through the biomass reaction. This step will add all missing reactions that are essential for *in silico* growth.

2. **Individual biomass components**: It is checked whether the model supports the biosynthesis of biomass components. Therefore, the objective function is set to the production of one biomass component at a time and the gap-fill algorithm is performed. This gap-filling step is repeated for each biomass component metabolite twice, with and without oxygen to potentially allow aerobic and anaerobic growth for facultative anaerobe species.

3. **Alternative energy sources**: `gapseq` tries to gap-fill likely metabolic pathways, which enable the utilisation of alternative energy sources, which might not be part of the defined growth medium. To this end, the model is re-constrained to a M9-like minimal medium containing a single carbon source of interest at the time. As objective function, the summed flux of artificial reactions that accept electrons from the electron carriers ubiquinol, menaquinol, or NADH is defined. This test can be considered as an *in silico* simulation of the commonly used BIOLOG carbon source utilisation test arrays [76] in which the colometric effect is coupled to a dehydrogenase [8]. This gap-filling step is performed for all metabolites defined in `dat/sub2pwy.csv`.

4. **Metabolic products**: Finally, the same list of compounds as for step 3, is used to check whether the network can be gap-filled to allow the formation of these metabolites. For each compound the gap-filling algorithm is applied with the production of the focal compound as objective function.

While step 1 considers all reaction from the universal model as potential candidate reactions for gap-filling, steps 2-4 allow only the addition of candidate reactions to the model with a corresponding bitscore from the pathway prediction (2.3) higher than a threshold value b (default: 50). Thus, these so-termed "core reactions" represent only reactions, for which `gapseq` has found genomic sequence or pathway evidence. This approach for steps 2-4 is chosen to avoid the addition of biosynthetic capabilities to the model, which the organism presumably does not possess.

2.10 Validation with enzymatic data (BacDive)

The Bacterial Diversity Metadatabase (BacDive) [67] was used to obtain enzymatic activity data. For this purpose, a list of type strains IDs where downloaded using the

advanced search. Afterwards the IDs were used to query the database via the R package BacDiveR (0.9.1) to obtain the data [46]. If the stored data contained non-zero entries for enzymatic activity and if a genome assembly was available on NCBI, the type strain was considered for the validation analysis and. The respective genome assemblies were downloaded with `ncbi-genome-download` (`https://github.com/kblin/ncbi-genome-download`). If multiple genomes were available for one type strain, then 'representative' and 'complete' (NCBI tags) genomes were preferred and possibly the most complete genome selected. Genome completeness was estimated by employing the software BUSCO (3.0.2) [74]. In total, 3017 type strain genomes were taken as input for ModelSEED (2.5.1), CarveMe (1.2.2), and `gapseq` to create metabolic models. The gap-filling was set to default for each program, i.e. a complete medium was assumed. The final test whether a reaction activity is covered by a model was done by checking if the corresponding reaction is present in the model. This was done by matching enzymes and reactions via EC numbers. For CarveMe the vmh (`https://www.vmh.life`) and for ModelSEED and `gapseq` the ModelSEED (`http://modelseed.org`) reaction database was used to association reactions and EC numbers. For the EC numbers 3.1.3.1, 3.1.3.2, the corresponding reactions were the same, and thus unspecific, so that both EC numbers were not considered for the validation analysis. In general, the enzyme activities in the BacDive database have the form active ("+") or not active ("-") but some entries were ambiguous (e.g.: "+/-"). The ambiguous entries were not taken into account.

2.11 Validation with carbon sources data (ProTraits)

Data for the validation of carbon source utilisation was obtained from the "atlas of prokaryotic traits" database (ProTraits) [10]. A tab-separated table with binarised predictions with a stringent threshold of precision of ≥ 0.95 were downloaded from `http://protraits.irb.hr/data.html`. For organisms which had at least one carbon source prediction, a genome was searched on NCBI RefSeq [71]. In cases where a genome assembly was found, it was taken as input for ModelSEED, CarveMe, and `gapseq` to create metabolic models. The number of potential carbon sources was reduced to a subset for which a mapping from substance name to ModelSEED and CarveMe model namespace existed (`dat/sub2pwy.csv`). The tests for D-lyxose were removed because it was listed as all negative in ProTraits and also all compared pipelines predicted no utilisation. The main test whether a carbon source can be used by a model was done in a BIOLOG-like manner as described above (see 2.9). To this end, temporary reactions to recycle reduced electron carriers as carbon source utilisation indicators were added to the respective model. The objective for optimisation was set to maximise the flux through these recycling reactions. The exchange reactions were limited to a minimal medium with minerals and the focal potential carbon source. This theoretical approach tests, whether the model is able to pass electrons from the potential carbon source to electron carrier metabolites. A carbon source was predicted to be able to serve as energy source if the recycle reactions carried a positive flux.

2.12 Prediction of gene essentiality

To predict the essentiality of genes we performed *in silico* single gene deletion phenotype analysis for the network reconstructions of *Escherichia coli* str. K-12 substr. MG1655 (RefSeq assembly accession: GCF_000005845.2), *Bacillus subtilis* substr. *sub-*

tilis str. 168 (GCF_000789275.1), *Shewanella oneidensis* MR-1 (GCF_000146165.2), *Pseudomonas aeruginosa* PAO1 (GCF_000006765.1), and *Mycoplasma genitalium* G37 (GCF_000027325.1). The analysis was performed on the basis of the models' Gene-Protein-Reaction (GPR) mappings and according to the protocol by Thiele and Palsson, 2010 [83]. To this end, the contingency tables of predicted growth/no growth phenotypes from the network models and experimentally determined growth phenotypes of gene deletion mutants were constructed. Genes were predicted to be conditionally essential under the given growth environment if the predicted growth rates of the models were below 0.01 hr^{-1}. The growth media compositions for growth predictions were defined as M9 with glucose as carbon- and energy source for *E. coli*, lysogeny broth (LB) for *B. subtilis* and *S. oneidensis*, M9 with succinate as carbon and energy-source for *P. aeruginosa*, and a complete medium (all external metabolites available for uptake) for *M. genitalium*. Experimental data for gene essentiality was obtained from [29, 54, 62, 86, 93].

2.13 Fermentation product tests

The release of by-products from anaerobic metabolism was predicted using Flux Balance Analysis (FBA) coupled with a minimisation of total flux [35] to avoid fluxes that do not contribute to the objective function of the biomass production. In addition, Flux-Variability Analysis (FVA) [52] was applied to predict the maximum fermentation product release of individual metabolites across all possible FBA solutions. Metabolites with a positive exchange flux (i.e. outflow) were considered as fermentation products. The analysis was performed for 18 different bacterial organisms, which (1) have a genome assembly available in the RefSeq database [71], (2) are known to grow in anaerobic environments, and (3) for which the fermentation products have been described in the literature based on anaerobic cultivation experiments (suppl. table S2). The gap-filling of the network models using `gapseq`, CarveMe, and ModelSEED as well as the simulations of anaerobic growth were all performed assuming the same growth medium that comprised several organic compounds (i.e. carbohydrates, polyols, nucleotides, amino acids, organic acids) as potential energy sources and nutrients for growth (see media file `dat/media/FT.csv`). Since the amount of fermentation product release depends on the organism's growth rate, we normalised the outflow of the individual fermentation products, which has the unit $mmol * gDW^{-1} * hr^{-1}$, by the predicted growth rate of the respective organism which has the unit hr^{-1}. Thus, we report the amount of fermentation product production in the quantity of the metabolite that is produced per unit of biomass: $mmol * gDW^{-1}$.

2.14 Pathway prediction of soil and gut microorganisms

The pathway analysis was done by comparing predicted pathways of soil and gut microorganisms. For this means, genomes were downloaded from a resource of reference soil organisms [14] and gut microbes [51]. The default parameter of `gapseq` were used for pathway prediction. The principal component analysis was done in R using the factoextra package [41]. For predicted pathways for soil and gut microorganisms, it was checked if samples belong to different distributions using a bootstrap version of the Kolmogorov-Smirnov test [72].

2.15 Anaerobic food web of the human gut microbiome

Representative bacterial organisms known to be relevant in the human intestinal cross-feeding of metabolites were selected based on the proposed food webs by Louis *et al.*, 2014 [49] and Rivera-Chavez *et al.*, 2015 [69]. The genomes of organisms were obtained from NCBI RefSeq [71] and metabolic models reconstructed using `gapseq`. A medium containing minerals, vitamins, amino acids, fermentation- and metabolic by-products (acetate, formate, lactate, butyrate, propionate, H_2, CH_4, ethanol, H_2S, succinate), and carbohydrates (glucose, fructose, arabinose, ribose, fucose, rhamnose, lactose) was used for gap-filling. Furthermore, a published model of *Methanosarcina barkeri* was added to the community [20] to represent archaea that are also known to be part of anaerobic food webs [73]. All organisms of the modeled community and their respective genome assembly accession numbers are listed in supplementary table S3. All metabolic models were then simulated with BacArena [3] by using the described medium but without the fermentation and by-products, plus sulfite and 4-aminobenzoate which were needed for growth by the *M. barkeri* model. The community was simulated for five time steps (corresponding to 5 hours simulated time). The analysis of metabolite uptake and production were done after the third time step, for which all organisms were still growing exponentially.

2.16 Technical details

The pathway prediction part of `gapseq` is implemented as Bash shell script and the metabolic model generation part is written in R. Linear optimisation can be performed with a different solvers (GLPK or CPLEX). Other requirements are exonerate, bedtools, and barrnap. In addition, the following R packages are needed: data.table [19], stringr [91], sybil [28], getopt [17], reshape2 [89], doParallel [16], foreach [53], R.utils [5], stringi [23], glpkAPI [27], and BioStrings [58]. Models can be exported as SBML [9] file using sybilSBML [28] or R data format (RDS) for further analysis in R, for example with sybil [28] or BacArena [3].

3 Results

3.1 Biochemistry database and universal model

The pathway-, transporter, and complex prediction is based on a protein sequence database that is derived from UniProt as well as TCDB and consists in total of 130,671 unique sequences (111,542 reviewed unipac 0.9 clusters and 19,129 TCDB transporter) and also 1,131,132 unreviewed unipac 0.5 cluster that can be included optionally. In addition, the protein sequence database in `gapseq` can be updated to include new sequences from Uniprot and TCDB. For the construction of genome-scale metabolic network models we have built a biochemistry database, that is derived from the ModelSEED biochemistry database. In total, the resulting curated `gapseq` metabolism database comprises 14.287 reactions (including transporters) and 7.570 metabolites. All metabolites and reactions from the biochemistry database are incorporated in the universal model that `gapseq` utilises for the gap-filling algorithm. When removing all dead-end metabolites and corresponding reactions, the universal model comprises 10.194 reactions and 3.337 metabolites. It needs to be noted, that the current biochemistry database and the derived universal model represents bacterial metabolic functions and that, at the current version

of gapseq, the database does not include light-dependent and archaea reactions. However, those reactions and, thus, also the possibility to use gapseq for the reconstruction of photosynthetic bacteria and archaea will be included in an later version of the software.

3.2 Agreement with enzymatic data (BacDive)

We used experimental data of active metabolic enzymes to compare the accuracy of model generation pipelines. In total, we compared 10,538 enzyme activities, 30 unique enzymes, in 3,017 organisms. gapseq models had much less false negatives, 6% compared with 32% (CarveMe) and 28% (ModelSEED), and correspondingly also more true positives, 53% compared with 27% (CarveMe) and 30% (ModelSEED), whereas results for false positives and true negatives were comparable (Figure 2A). For this test, the most prominent EC numbers were the catalase, 1.11.1.6, accounting for 26% of the comparisons and the cytochrome oxidase, 1.9.3.1, accounting for 22%.

3.3 Validation of carbon source usage (ProTraits)

Growth predictions are essential for metabolic models. We checked the quality of model generation pipelines to predict the growth on different carbon sources. In summary, we compared 1,795 different growth prediction for 526 organism and 48 carbon sources (Figure 2B). gapseq outperformed the other methods in terms of false negatives (14% compared with 29% ModelSEED and 37% CarveMe) and true positives (45% compared with 31% ModelSEED and 23% CarveMe). ModelSEED showed slightly fewer false positives (5% compared with 10% gapseq and 11% CarveMe) and more true negatives (35% compared with 30% gapseq and 30% CarveMe). gapseq, predicted most false positives for formate (29 times). This overestimate of formate as potential carbon source is likely due to the fact that we tested carbon source utilisation on the basis of electron transfer from the source to electron carriers (i.e. ubiquinol, menaquinol, or NADH), which is analogous to the experiemental carbon source test of BIOLOG plates [76]. However, while it is known that formate can serve in fact as electron donor in a number of different bacteria [15], the role as source of carbon atoms for the synthesis of biomass components is limited to a few known methylotrophs [30].

Across all methods, the best predicted carbon sources, with more than 100 tested organisms, were fructose (91% correct predictions), mannose (89%), or arginine (84%), whereby less good predictions were obtained for arabinose (29% correct predictions), dextrin (40%), or acetate (42%).

3.4 Gene essentiality

We compared the ability of gapseq models to predict the essentially of genes with predictions from ModelSEED and CarveMe reconstructions as well as with curated models for the same organisms (Figure 3). As expected, the curated models outperform all three automated reconstruction tools for most species and prediction metrics. Interestingly, for *P. aeruginosa* the gapseq model shows better gene essentiality predictions in terms of sensitivity, accuracy, and F1-score than the curated model (Figure 3D). Compared to CarveMe, gapseq shows generally a higher sensitivity in essentiality predictions but, at the same time, a lower precision rate. This pattern is attributed to the fact, that gapseq models tend to predict more genes as essential than CarveMe, leading to a higher number

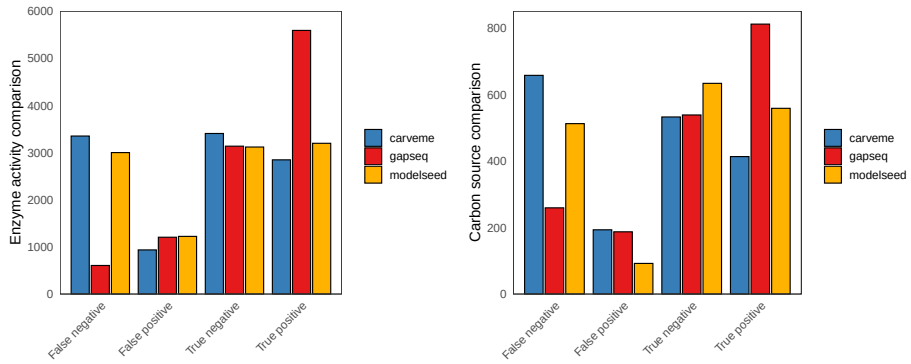

Figure 2: Results from enzyme activity and carbon source validations. A) In total 10,538 enzyme activities (30 enzymes and 3,017 organisms) from experimental standardised experiments from the DSMZ BacDive database were compared for three different methods. B) The predictions of 1,795 carbon sources (48 unique carbon sources and 526 organisms) were validated with data from the ProTraits database.

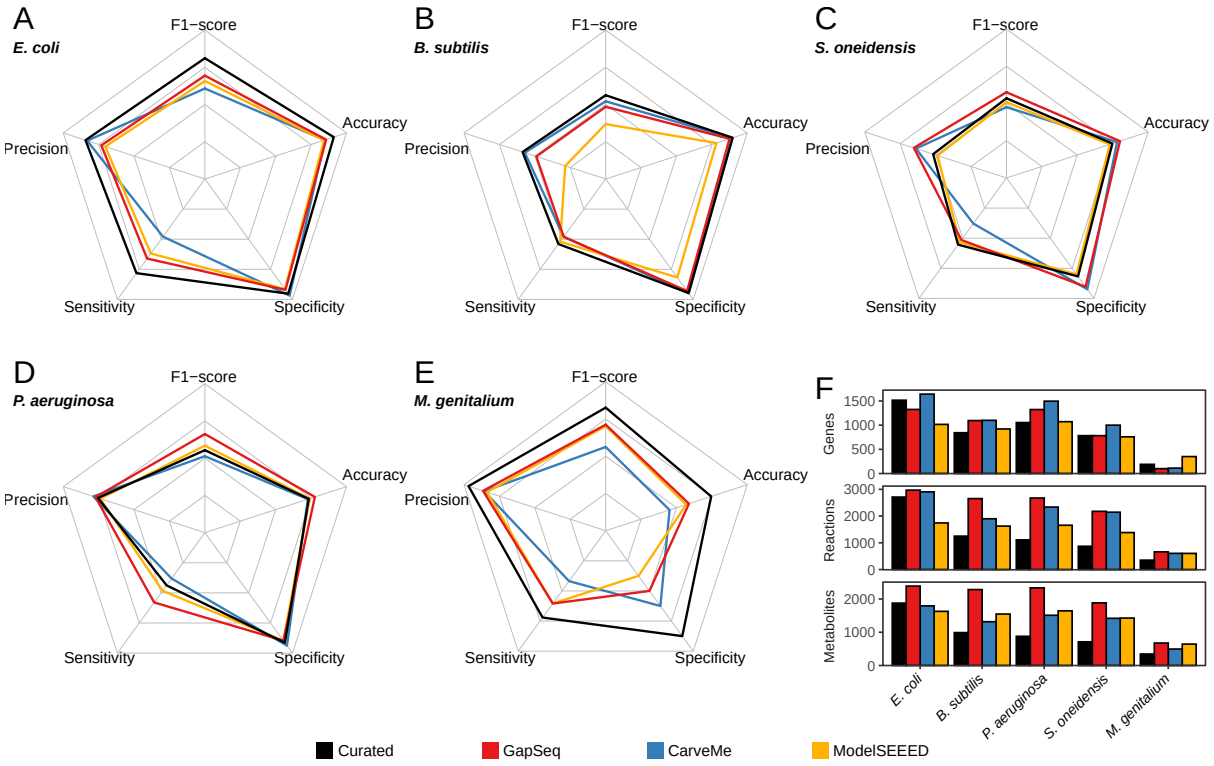

Figure 3: Results from model gene essentiality tests for five bacterial species: (A) *Escherichia coli*, (B) *Bacillus subtilis*, (C) *Shewanella oneidensis*, (D) *Pseudomonas aeruginosa*, and (E) *Mycoplasma genitalium*. Results from `gapseq` models (red) are compared to CarveMe- (blue) and ModelSEED (yellow) models, as well as to published curated genome-scale metabolic models (black) of the respective organisms. (F) Counts of genes, reactions (including exchanges and transporters), and metabolites in each reconstruction.

of true positive (TP) predictions and more false positives (FP). For most organisms and on the basis of most prediction metrics, `gapseq` outperforms network models that were reconstructed using ModelSEED.

3.5 Fermentation products

Anaerobic or facultative anaerobic bacteria utilise different fermentation pathways in order to extract energy from the oxidation of organic compounds. We tested if the identity of fermentation products can be predicted by metabolic network model constructions obtained from `gapseq`, CarveMe, and ModelSEED for 18 different bacterial organisms (Figure 3). The organisms were selected based on following criteria: (1) the organisms have a published RefSeq genome sequence [71], (2) are known anaerobic or facultative anaerobic organisms, and (3) the identity of fermentation products has been experimentally described and reported in primary literature (Suppl. table S1). Overall, `gapseq`

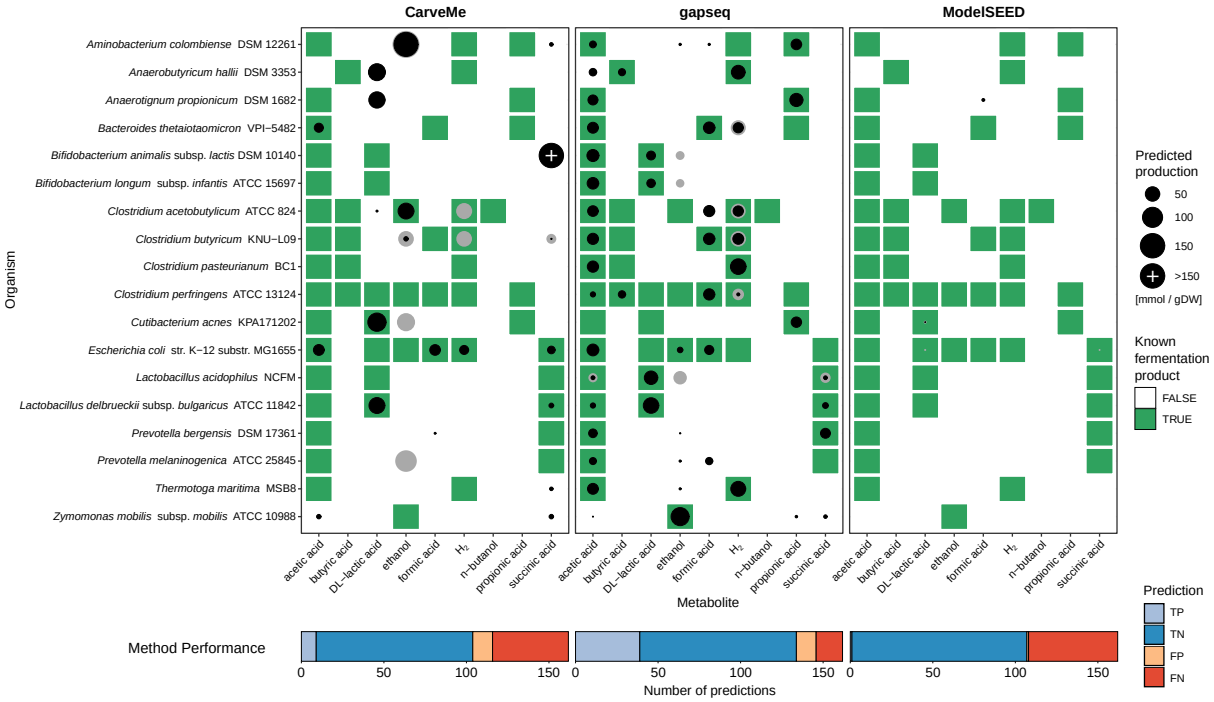

Figure 4: Results of the fermentation product test of 18 bacterial organisms under anaerobic growth with models generated using `gapseq`, CarveMe, and ModelSEED. Point sizes indicate the predicted production of a fermentation product metabolite (columns) by the corresponding organism (row). Predictions (black) are based on Minimize-Total-Flux (MTF) flux balance analyses. Grey circles indicate the upper production limit obtained from Flux-Variability-Analysis (FVA). Metabolite-organism-combinations highlighted in green denote known fermentation products, which have been reported in literature based on experimental measures of the metabolite in anaerobic cultures.

showed the highest number of true positive predictions (TP) with 36 TP predicted with the MTF solution and 37 TP predicted with FVA which is substantially higher compared to CarveMe (8 TP with MTF, 10 TP with FVA) and modelSEED (1 TP, 3 TP). The production of all short-chain-fatty-acids acetate, butyrate, and propionate was correctly predicted by `gapseq` in 78% of cases and thereby outcompetes CarveMe (9%) and modelSEED (0%), which did not predict butyrate or propionate production for any organism tested. Moreover, `gapseq` correctly predicted homolactic fermentation by *Lactobacillus delbrueckii* and *Lactobacillus acidophilus*, which is dominated by lactic acid as well as heterolactic fermentation by *Bifidobacterium longum*. However, `gapseq` failed to predict lactic acid production of organisms that utilise different fermentation strategies, which

also yield lactic acid (e.g. mixed-acid fermentation by *E. coli*).

Interestingly, the predicted quantities of fermentation product release is higher for true positive than for false negative predictions (Figure 3). This further suggests, that `gapseq` is able to predict the main fermentation products of bacterial organisms during anaerobic growth based on the organism's genome sequence.

3.6 Validatiton test summary

The performance of different approaches for metabolic model creation is summarised in Figure 5. The overall accuracy (proportion all correct prediction in relation to all

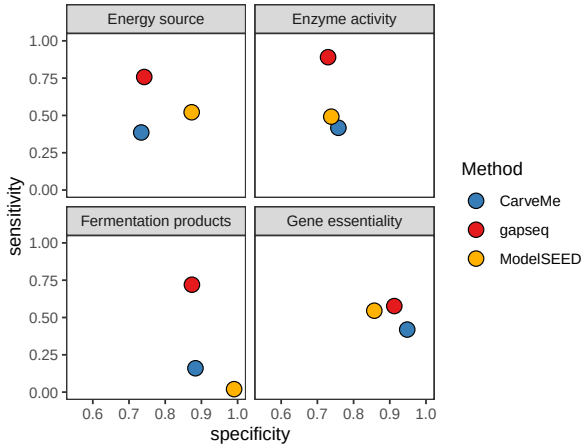

Figure 5: Summary of different validation tests. The specificity and sensitivity for all compared methods are shown. This includes results from benchmarks concerning enzyme activities, energy sources, fermentation products, and gene essentiality.

predictions made) of model predictions with experimental and literature data was 64% (CarveMe), 68% (ModelSEED), and 80% (`gapseq`). In addition, predictions and experimental or literature data were compared to calculate the specificity and sensitivity for each approach. Sensitivity measures the proportion of correctly predicted positives, whereas specificity accounts for the accurate prediction of negatives. All approaches showed a high specificity > 0.7 with highest values for fermentation product and gene essentiality tests. Notably, `gapseq` showed the highest sensitivity over all tests. In summary, `gapseq` outperformed other methods in terms of accuracy and sensitivity while showing similar specificity at the same time.

3.7 Sample application I - Anaerobic food web of gut microbiome

The prediction of metabolic interactions between microbial organisms is of special interest in ecology, medicine, and biotechnology. So far, we showed the capacity of `gapseq` on the level of individual models. In a next step, we simulated several individual models together as a multi-species community to validate the potential of `gapseq` in community modelling. As sample application we selected representative members of the human gut microbiome that are known to form an anaerobic food web [49, 69]. Altogether, we employed 20 organisms and simulated the combined growth in a shared environment for several time steps. On the community level, simulations captured all important substances, which are known to be produced in the context of the food web (Figure 6).

This included the production of short chain fatty acids (acetate, propionate, butyrate), lactate, hydrogen, hydrogen sulfide (H_2S), methane, formate, and succinate. The formation of acetate, formate, and hydrogen was most prevalent, which are also common end-products of fermentation. Lactate, succinate, acetate, hydrogen, formate, and H_2S were further metabolised by some community members (Figure 6). The predicted identity of fermentation end-products and other by-products of metabolism was found to be in line with literature information [49, 56, 69]. For example, the formation of lactate was observed for *Lactobacillus acidophilus* and *Bifidobacterium longum*, and butyrate was released by known butyrate producers, i.e. *Faecalibacterium prausnitzii*, *Anaerobutyricum hallii*, *Clostridium perfringens*, and *Coprococcus* spp.. Especially the main products of mixed acid fermentation (acetate, formate, hydrogen, ethanol) were predicted for most members of the community which is in agreement with what is known about common metabolic end products of many gut-dwelling microorganisms [56]. Interestingly, for *Faecalibacterium prausnitzii* no acetate production is reported [56], which was also observed in simulations. Moreover, H_2S was correctly predicted to be produced by *Desulfovibrio desulfuricans*. In general, the anaerobic oxidation of fatty acids is not favored by the gut environment because the host competes for the uptake of butyrate, propionate, and acetate, which serve as energy source for colonic epithelial cells and are involved in many host functions [68]. Therefore, the gut community lacks syntrophic organisms which are able to anaerobically degrade butyrate but are slow growing and therefore not favored by the gut environment [94]. In agreement with this, we found no microbial uptake of butyrate in the community simulation. In contrast, lactate was predicted to produced and consumed by distinct community members. We found utilisation of lactate by *Coprococcus comes*, *Megasphaera elsdenii*, and *Veillonella dispar*, which is a known feature of these organisms [49]. In addition, succinate was correctly predicted to be used by *Bacteroides* species [56]. The formation of methane is known to be limited to methanogenic archaea, and thus *Methanosarcina barkeri* produced methane from acetate and hydrogen during our simulations. In summary, `gapseq` models were able to recapitulate the major interactions, which are described for microbial communities in the human gut. The overall consumption pattern and individual microbial contributions were found to be in agreement with literature data. Taken together, the community simulation results illustrate the capacity of `gapseq` to construct predictive models for complex metabolic interaction networks comprising several different species.

3.8 Sample application II - Pathway prediction of soil and gut microorganisms

To demonstrate the pathway prediction capabilities of `gapseq`, we analysed two communities of soil and gut microorganisms. The energy metabolism of both communities were characterised, whereas both communities comprised a similar number of organisms (soil 922 organisms, gut 822 organisms). The distribution of pathways was found to be habitat-specific and not determined by phylogeny. In a principal component analysis, most variance could be explained by subsystems of pathways that are involved in chemoautotrophic, respiratory and fermentative processes including hydrogen production (Figure 7A). Out of 128 energy pathways, 40 differed significantly (Kolmogorov-Smirnov test, $P < 0.05$) between soil and gut microorganisms and could be categorised to 12 subsystems (Figure 7B). In total, gut microorganisms showed less variety in energy pathways than soil microorganisms. Only pathways relevant for the formation of acetate, hydrogen,

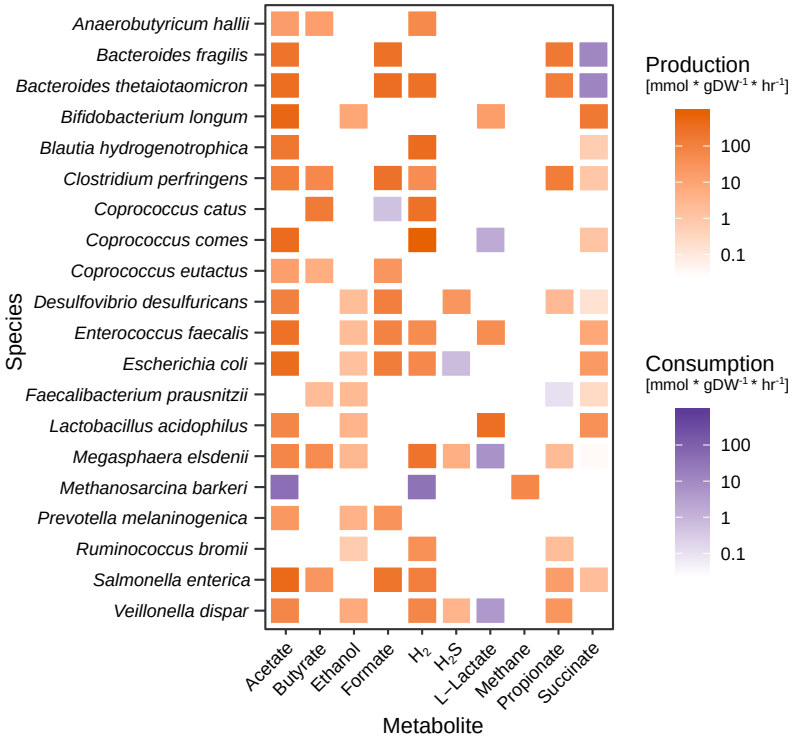

Figure 6: Predicted human gut microbial community metabolism using `gapseq` models. The metabolism of a community consisting of 19 bacterial species and one archaeon (*Methanosarcina barkeri*) was predicted using BacArena [3]. All bacterial models were reconstructed using `gapseq` and a published manually curated model was used for *M. barkeri*.

and lactate were predicted to be enriched. In the case of all other energy subsystems, more pathways were predicted for soil organisms, most prominently pathways relevant for aerobic and anaerobic respiration as well as the tricarboxylic acid cycle (TCA). In summary, members of the soil community showed a more versatile energy metabolisms, which potentially indicates a higher energetic specialisation of gut microbes. This sample application demonstrates how `gapseq` can facilitate the characterisation and comparison of microbial communities based on the analysis of the presence and absence of specific metabolic pathways.

4 Discussion

Here, we introduced `gapseq` - a new tool for metabolic pathway analysis and genome-scale metabolic network reconstruction. The novelty of `gapseq` lies in the combination of (i) a novel reaction prediction that is based both on genomic sequence homology as well as pathway topology, (ii) a profound curation of the reaction database to prevent thermodynamically infeasible reaction cycles, and (iii) a reaction evidence score-oriented gap-filling algorithm. In order to scrutinise `gapseq` metabolic models, we compared the models' network structures and predictions with large-scale experimental data sets, which were retrieved from publicly available databases. Furthermore, the ability of `gapseq` to predict bacterial phenotypes was compared to two other commonly used automatic reconstruction methods, namely, CarveMe [50] and ModelSEED [34] (Table 1). ModelSEED

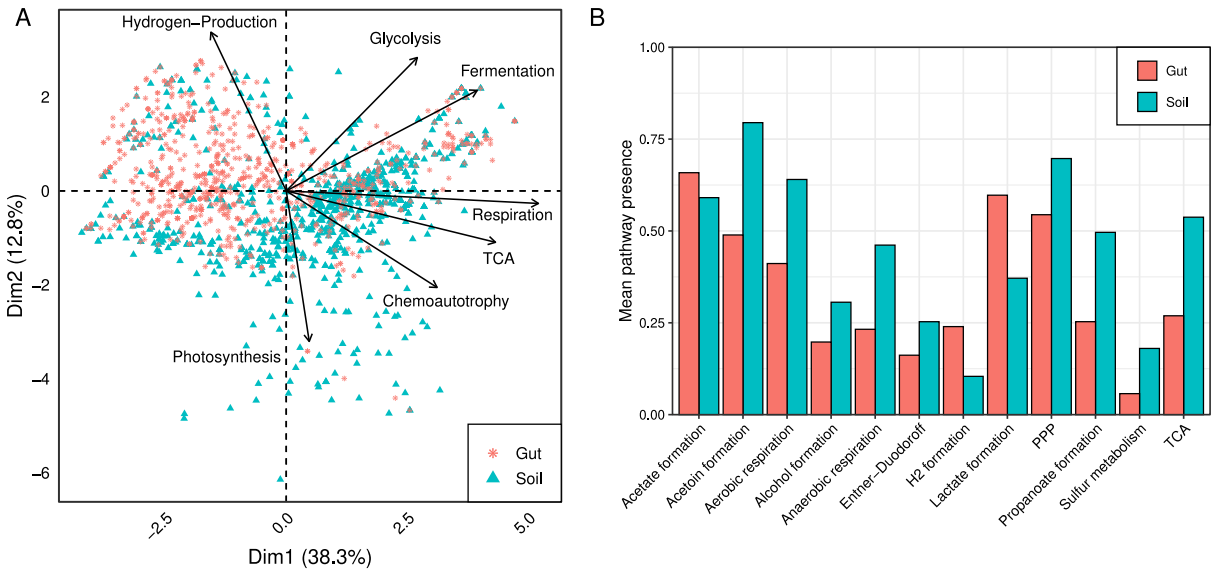

Figure 7: Sample applications of `gapseq`. Comparison of energy metabolism between soil and gut community. A) A PCA plot with the first two dimension explaining more than 50% of the variance. Selection of subsystems from energy metabolism with highest quality and impact are shown. B) List of subsystems of energy metabolism that differ significantly between members of the soil and gut community (TCA: tricarboxylic acid cycle, PPP: Pentose phosphate pathway).

is also implemented in the KBASE online software platform [42].

Large-scale benchmarking of metabolic models is crucial

The quality of genome-scale metabolic networks can be assessed by comparing model predictions with experimental physiological data. The protocol by Thiele and Palsson (2010) for the reconstruction of genome-scale metabolic networks recommends the quality assessment and manual network curation using data for (i) known secretion products (e.g. fermentation end-products), (ii) single-gene deletion mutant growth phenotypes (i.e. gene essentiality), and (iii) the utilisation of carbon/energy sources [83]. Tools for the automatic reconstruction of metabolic networks should also make use of such physiological data whenever available for benchmarking. Here, we tested our `gapseq` approach on the basis of all three recommended phenotypic data and compared the performance with CarveMe and ModelSEED. Additionally, we included a fourth (iv) and novel benchmarking test where experimental information on the activity of specific enzymes [67] was compared to reconstructed networks. Across all four benchmark tests, we could show that `gapseq` outperformed CarveMe and ModelSEED in terms of sensitivity while achieving specificity scores that are comparable to the other two tools (Figure 5). Publicly available genome sequences of microorganisms, which can be subject for automated metabolic network reconstruction are massively increasing in number due to continuing advances in high-quality and high-throughput sequencing technologies [48]. This development is further fueled by the the increasing number of genome assemblies from metagenomic material [65]. In contrast, standardised phenotypic data for microorganisms remains a bottleneck for the validation of automated metabolic network reconstruction pipelines such as `gapseq`. As consequence, it is crucial for the future development of automated network reconstruction software to include possibly all available phenotypic data for benchmark-

Table 1: Summary of different methods that were compared in this work.

Metric	**CarveMe**	**gapseq**	**ModelSEED**
Implementation			
Infrastructure	local	local	web service
Input (FASTA file)	protein	nucleotide	nucleotide
Programming languages	python	shell script, R	perl/javascript
Gap-fill solver	CPLEX	GLPK/CPLEX	not needed*
Gap-fill problem formulation	MILP	LP	MILP
Performance			
Accuracy	0.64	0.80	0.68
Sensitivity	0.35	0.74	0.39
Specificity	0.83	0.81	0.86

* Solver runs on ModelSEED server. No local solver is required.

ing, especially data from non-model organisms. To benchmark `gapseq` in in relation to CarveMe and ModelSEED using phenotypic data from mainly non-model organisms, we retrieved phenotypic data of enzyme activity for more than 3,000 organisms and carbon source utilisation for more than 500 organisms from online databases, which is, to our knowledge, the yet largest phenotypic data set used for validation of automatically reconstructed metabolic networks. In this validation approach `gapseq` achieved the highest prediction accuracy among all three tools tested (Figure 2).
Hence, those results suggest that `gapseq` is a powerful new tool for the automated reconstruction of genome-scale metabolic network models. Moreover, the underlying reference protein sequences as well as the pathway database can readily updated using online resources, which makes `gapseq` flexible to include future developments and findings in microbial metabolic physiology.

Automated network reconstructions for community modelling

While single organisms can be considered as the building blocks of microbial communities, individual metabolic models of organisms are the building blocks of *in silico* microbial community simulations. Therefore, genome-scale metabolic models are increasingly applied to predict the function of multi-species microbial communities [31, 43, 51]. To correctly infer metabolic interaction networks between different organisms, it is important that individual models accurately predict nutrient utilisation (e.g. carbon source) and metabolic end-products (e.g. fermentation products). In this study, the benchmarks for carbon source utilisation and fermentation end-product identity indicated that `gapseq` has the highest prediction performance compared to other reconstruction tools (Figure 2 and Figure 4).
To illustrate the applicability of `gapseq`-reconstructed metabolic models for the simulation of multi-species community metabolism, we generated models for bacterial strains from the human gut microbiota and simulated their growth in a shared environment. Without further curation, the community simulation reproduced all important hallmarks of intestinal anaerobic food webs [49, 56]. Above all, short chain fatty acids (SCFA) were predicted to be the primary end products of fermentation. This prediction is important to

represent intestinal metabolism, because SCFA are crucially involved in host physiology by affecting regulatory response in intestinal and immune cells [11, 77]. Furthermore, the simulation predicted the exchange of metabolites between different members of the microbial community (Figure 6). Cross-feeding of metabolites and the formation of anaerobic food chains have been associated with a healthy microbiome [1, 60]. For instance, the cross-feeding of lactate has been reported to be vital for the early establishment of a healthy gut microbiota in infants [60]. In this study, the exchange of lactate between different bacterial species was also observed in the community simulations (Figure 6) and involved known lactate producers (e.g. *Enterococcus faecalis*) and consumers (e.g. *Megasphaera elsdenii*). This example illustrates that we are able to predict key features of the anaerobic food-web within the gastrointestinal microbiota using `gapseq` models. In addition to the ability to accurately model metabolic processes within existing microbial communities, `gapseq` will further promote the potential of metabolic modelling to predict how complex microbial communities can be modulated by targeted interventions. Specific interventions, which could for instance be predicted, are the introduction of new species to the community (i.e. probiotics) or microbiome-modulating compounds (prebiotics) to the environment. Predictions of potential intervention strategies that target the microbiome are of vast relevance for biomedical research.
Taken together, the results obtained with `gapseq` suggest, that metabolic models which are reconstructed using `gapseq` are promising starting points to construct ecosystem-scale models of inter-species biochemical processes and to predict targeted strategies to modulate microbiome structure and function.

Pathway analysis of microbial communities

The construction of genome-scale metabolic models is based on metabolic networks that are inferred from genomic sequences in the context of biochemical databases [83]. Although, the reconstruction of metabolic networks is closely related to the prediction of metabolic pathways, metabolic modelling and pathway analysis are often treated separately [25]. In `gapseq`, the prediction of metabolic pathways is intrinsically tied to the reconstruction of metabolic networks and gap-filling. In addition, reaction, transporter, and pathway predictions can also be used to evaluate the functional capacities of microorganisms without the need of metabolic modelling. As an example for metabolic pathway analysis, we compared the predicted energy metabolism of two large microbial communities that occur in soil and the human gut. We could show that the predicted distribution of pathways differ between both communities based on the habitat, which usually accommodates the members of the respective community. Gut microorganisms showed a less versatile energy metabolism and a specialisation towards fermentation pathways, which lead to the formation of acetate, hydrogen, and lactate. Variations in pathways distributions between both communities may be explained by distinct evolutionary histories. The habitat of the diverse group of soil microorganisms more likely represents an open ecosystem, whereas the gut microbiome is directly constraint by a multi-cellular host that potentially controls microbial traits [22]. In general, metabolic modelling should be accompanied by the analysis of pathways based on statistical methods [25] to compensate for additional assumptions, which are introduced in constraint-based metabolic flux modelling [38].

Limitations and outlook

gapseq takes 1-2h for the reconstruction of a single model, whereas ModelSEED and CarveMe operate faster (10min) on a standard desktop computer. Nonetheless, CarveMe needs as input gene sequences (protein or nucleotide), which has to be predicted first, and ModelSEED works as a web service, which can complicate the handling of large-scale reconstruction projects. In gapseq, pathways were predicted based on topology and sequence homology searches. However, the assignment of enzymatic function from sequence comparisons has been shown to potentially miss protein domain structures and thus can cause false annotations [2, 24]. In addition, gapseq uses many resources to find potential sequences for reactions in pathway databases. Together this might explain why although gapseq performed better than other methods on predicting positive phenotypes (function present), it went head to head with regard to negative phenotype predictions (function not present). CarveMe takes a different approach when inferring function by taking care of functional regions (protein domains) [26], resulting in orthologous groups [36], which results in a slightly better specificity (true negative phenotype predictions) in benchmarks (Figure 5). Future developments of gapseq will address orthologous groups by using multiple inference methods. Furthermore, the integration of functional predictions coming from phylogenetic inference without the need of genomic sequences [18] might also be promising for further developments of gapseq.

Conclusion

We propose a new tool called gapseq that is suitable for metabolic network analysis and metabolic model reconstruction. To enhance phenotype predictions, gapseq employs various data sources and a novel gap-filling procedure that reduces the impact of arbitrary growth medium requirements. We further brought together the so far largest benchmarking of genome-scale metabolic models, in which gapseq performed on average better than comparable alternative tools. Altogether, we consider gapseq as important contribution to the modelling of microbial communities in the age of the microbiome.

Acknowledgements

We thank Martin Sperfeld for fruitful comments and discussions during the developmental phase. The software was thankfully tested by Georgios Marinos, Shan Zhang, and Lena Best.

Supplementary information

Table S1. New reactions and metabolites added to biochemistry database.
see file: github.com/jotech/gapseq/preprint/Table_S1.xlsx

Table S2. Organisms included in fermentation product validation test.
see file: github.com/jotech/gapseq/preprint/Table_S2.xlsx

Table S3. Organisms used in modelling of the anaerobic food web of the human gut microbiome.

RefSeq Assembly	Organism name	Reconstruction method
GCF_000173975.1	*Anaerobutyricum hallii DSM 3353*	gapseq
GCF_000025985.1	*Bacteroides fragilis NCTC 9343*	gapseq
GCF_001314975.1	*Bacteroides thetaiotaomicron*	gapseq
GCF_000196555.1	*Bifidobacterium longum subsp. longum JCM 1217*	gapseq
GCF_000157975.1	*Blautia hydrogenotrophica DSM 10507*	gapseq
GCF_000013285.1	*Clostridium perfringens ATCC 13124*	gapseq
GCF_003434235.1	*Coprococcus catus*	gapseq
GCF_000155875.1	*Coprococcus comes ATCC 27758*	gapseq
GCF_000154425.1	*Coprococcus eutactus ATCC 27759*	gapseq
GCF_000189295.2	*Desulfovibrio desulfuricans ND132*	gapseq
GCF_000391485.2	*Enterococcus faecalis EnGen0107*	gapseq
GCF_000005845.2	*Escherichia coli str. K-12 substr. MG1655*	gapseq
GCF_000162015.1	*Faecalibacterium prausnitzii A2-165*	gapseq
GCF_003047065.1	*Lactobacillus acidophilus*	gapseq
GCF_001304715.1	*Megasphaera elsdenii 14-14*	gapseq
GCF_000195895.1	*Methanosarcina barkeri str. Fusaro*	manually curated (BiGG-ID: iAF692)[20]
GCF_000144405.1	*Prevotella melaninogenica ATCC 25845*	gapseq
GCF_900101355.1	*Ruminococcus bromii*	gapseq
GCF_000006945.2	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	gapseq
GCF_900637515.1	*Veillonella dispar*	gapseq

References

[1] Konrad Aden et al. "Metabolic Functions of Gut Microbes Associate With Efficacy of Tumor Necrosis Factor Antagonists in Patients With Inflammatory Bowel Diseases". In: *Gastroenterology* (2019). URL: http://www.sciencedirect.com/science/article/pii/S0016508519411189.

[2] Alex Bateman et al. "The Pfam Protein Families Database". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 263–266. eprint: http://oup.prod.sis.lan/nar/article-pdf/28/1/263/9895152/280263.pdf. URL: https://doi.org/10.1093/nar/28.1.263.

[3] Eugen Bauer et al. "BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities". In: *PLOS Computational Biology* 13.5 (May 2017), pp. 1–22. URL: https://doi.org/10.1371/journal.pcbi.1005544.

[4] Matthew N. Benedict et al. "Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models". In: *PLOS Computational Biology* 10.10 (Oct. 2014), pp. 1–14. URL: https://doi.org/10.1371/journal.pcbi.1003882.

[5] Henrik Bengtsson. *R.utils: Various Programming Utilities*. R package version 2.9.2. 2019. URL: https://CRAN.R-project.org/package=R.utils.

[6] Thomas Bernard et al. "Reconciliation of metabolites and biochemical reactions for metabolic networks." eng. In: *Brief Bioinform* 15.1 (Jan. 2014), pp. 123–135. URL: http://dx.doi.org/10.1093/bib/bbs058.

[7] Crysten E. Blaby-Haas and Valérie de Crécy-Lagard. "Mining high-throughput experimental data to link gene and function". In: *Trends in Biotechnology* 29.4 (Apr. 2011), pp. 174–182. URL: https://doi.org/10.1016/j.tibtech.2011.01.001.

[8] Barry R. Bochner. "Global phenotypic characterization of bacteria". In: *FEMS Microbiology Reviews* 33.1 (2009), pp. 191–205. eprint: /oup/backfile/content_public/journal/femsre/33/1/10.1111_j.1574-6976.2008.00149.x/2/33-1-191.pdf. URL: http://dx.doi.org/10.1111/j.1574-6976.2008.00149.x.

[9] Benjamin J. Bornstein et al. "LibSBML: an API library for SBML." eng. In: *Bioinformatics* 24.6 (Mar. 2008), pp. 880–881. URL: http://dx.doi.org/10.1093/bioinformatics/btn051.

[10] Maria Brbić et al. "The landscape of microbial phenotypic traits and associated genes". In: *Nucleic Acids Research* (Oct. 2016), gkw964. URL: http://dx.doi.org/10.1093/nar/gkw964.

[11] Mariana X Byndloss et al. "Microbiota-activated PPAR-γ signaling inhibits dysbiotic Enterobacteriaceae expansion". In: *Science* 357.6351 (2017), pp. 570–575. URL: https://doi.org/10.1126/science.aam9949.

[12] Christiam Camacho et al. "BLAST+: architecture and applications." In: *BMC Bioinformatics* 10 (2009), p. 421. URL: https://doi.org/10.1186/1471-2105-10-421.

[13] Ron Caspi et al. "The MetaCyc database of metabolic pathways and enzymes". In: *Nucleic Acids Research* 46.D1 (2018), pp. D633–D639. eprint: /oup/backfile/content_public/journal/nar/46/d1/10.1093_nar_gkx935/2/gkx935.pdf. URL: http://dx.doi.org/10.1093/nar/gkx935.

[14] Jinlyung Choi et al. "Strategies to improve reference databases for soil microbiomes". In: *The ISME Journal* 11.4 (Dec. 2016), pp. 829–834. URL: https://doi.org/10.1038/ismej.2016.168.

[15] Gregory M. Cook et al. "Chapter One - Energetics of Pathogenic Bacteria and Opportunities for Drug Development". In: *Advances in Bacterial Pathogen Biology*. Ed. by Robert K. Poole. Vol. 65. Advances in Microbial Physiology. Academic Press, 2014, pp. 1–62. URL: http://www.sciencedirect.com/science/article/pii/S0065291114000022.

[16] Microsoft Corporation and Steve Weston. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.15. 2019. URL: https://CRAN.R-project.org/package=doParallel.

[17] Trevor L Davis and Allen Day. *getopt: C-Like 'getopt' Behavior*. R package version 1.20.3. 2019. URL: https://CRAN.R-project.org/package=getopt.

[18] Gavin M. Douglas et al. "PICRUSt2: An improved and extensible approach for metagenome inference". In: *bioRxiv* (June 2019). URL: https://doi.org/10.1101/672295.

[19] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*. R package version 1.12.6. 2019. URL: https://CRAN.R-project.org/package=data.table.

[20] Adam M. Feist et al. "Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri." eng. In: *Mol Syst Biol* 2 (2006), p. 2006.0004. URL: http://dx.doi.org/10.1038/msb4100046.

[21] David A Fell. "Systems properties of metabolic networks". In: *Unifying Themes In Complex Systems, Volume 1*. Ed. by Yaneer Bar-Yam. CRC Press, 2003, pp. 163–178.

[22] Kevin R. Foster et al. "The evolution of the host microbiome as an ecosystem on a leash". In: *Nature* 548.7665 (Aug. 2017), pp. 43–51. URL: https://doi.org/10.1038/nature23292.

[23] Marek Gagolewski. *R package stringi: Character string processing facilities*. 2019. URL: http://www.gagolewski.com/software/stringi/.

[24] Michael Y Galperin and Eugene V Koonin. "Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption". In: *In silico biology* 1.1 (1998), pp. 55–67.

[25] Miguel A. García-Campos, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. "Pathway analysis: state of the art". In: *Frontiers in physiology* 6 (2015), p. 383. URL: https://doi.org/10.3389/fphys.2015.00383.

[26] Sara El-Gebali et al. "The Pfam protein families database in 2019". In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D427–D432. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D427/27436497/gky995.pdf. URL: https://doi.org/10.1093/nar/gky995.

[27] Gabriel Gelius-Dietrich. *glpkAPI: R Interface to C API of GLPK*. R package version 1.3.1. 2018. URL: https://CRAN.R-project.org/package=glpkAPI.

[28] Gabriel Gelius-Dietrich et al. "Sybil–efficient constraint-based modelling in R." eng. In: *BMC Syst Biol* 7 (2013), p. 125. URL: http://dx.doi.org/10.1186/1752-0509-7-125.

[29] John I Glass et al. "Essential genes of a minimal bacterium". In: *Proceedings of the National Academy of Sciences* 103.2 (2006), pp. 425–430.

[30] I Goldberg et al. "Bacterial yields on methanol, methylamine, formaldehyde, and formate". In: *Biotechnology and bioengineering* 18.12 (1976), pp. 1657–1668.

[31] S Graspeuntner et al. "Gut Dysbiosis With Bacilli Dominance and Accumulation of Fermentation Products Precedes Late-onset Sepsis in Preterm Infants". In: *Clinical Infectious Diseases* 69.2 (Oct. 2018), pp. 268–277. eprint: https://academic.oup.com/cid/article-pdf/69/2/268/28893438/ciy882.pdf. URL: https://doi.org/10.1093/cid/ciy882.

[32] William R. Harcombe et al. "Metabolic Resource Allocation in Individual Microbes Determines Ecosystem Interactions and Spatial Dynamics". In: *Cell Reports* 7.4 (2014), pp. 1104–1115. URL: http://www.sciencedirect.com/science/article/pii/S2211124714002800.

[33] Almut Heinken and Ines Thiele. "Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework". In: *Gut Microbes* 6.2 (2015). PMID: 25901891, pp. 120–130. eprint: https://doi.org/10.1080/19490976.2015.1023494. URL: https://doi.org/10.1080/19490976.2015.1023494.

[34] Christopher S Henry et al. "High-throughput generation, optimization and analysis of genome-scale metabolic models". In: *Nature Biotechnology* 28.9 (Aug. 2010), pp. 977–982. URL: https://doi.org/10.1038/nbt.1672.

[35] Hermann-Georg Holzhütter. "The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks". In: *European journal of biochemistry* 271.14 (2004), pp. 2905–2922.

[36] Jaime Huerta-Cepas et al. "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." In: *Nucleic Acids Res.* 47 (2019), pp. D309–D314. URL: https://doi.org/10.1093/nar/gky1085.

[37] Lisa Jeske et al. "BRENDA in 2019: a European ELIXIR core data resource". In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. D542–D549. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D542/27437170/gky1048.pdf. URL: https://doi.org/10.1093/nar/gky1048.

[38] Hidde de Jong et al. "Mathematical Modeling of Microbes: Metabolism, Gene Expression, and Growth". In: *Journal of the Royal Society Interface* 14.20170502 (2017). URL: https://doi.org/10.1098/rsif.2017.0502.

[39] Minoru Kanehisa et al. "New approach for understanding genome variations in KEGG". In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D590–D595. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D590/27436321/gky962.pdf. URL: https://doi.org/10.1093/nar/gky962.

[40] Peter D. Karp et al. "Pathway Tools version 19.0: Integrated Software for Pathway/Genome Informatics and Systems Biology". In: (Oct. 2015). eprint: 1510.03964. URL: https://doi.org/10.1093/bib/bbv079.

[41] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.6. 2019. URL: https://CRAN.R-project.org/package=factoextra.

[42] "KBase: The United States Department of Energy Systems Biology Knowledgebase". In: *Nature Biotechnology* 36.7 (July 2018), pp. 566–569. URL: https://doi.org/10.1038/nbt.4163.

[43] Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee. "Current state and applications of microbial genome-scale metabolic models". In: *Current Opinion in Systems Biology* 2 (2017). Regulatory and metabolic networks • Cancer and systemic diseases, pp. 10–18. URL: http://www.sciencedirect.com/science/article/pii/S2452310017300483.

[44] Sabine Koch et al. "RedCom: A strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing experimental datasets from anaerobic digestion". In: *PLOS Computational Biology* 15.2 (Feb. 2019), pp. 1–32. URL: https://doi.org/10.1371/journal.pcbi.1006759.

[45] Manish Kumar et al. "Modelling approaches for studying the microbiome". In: *Nature Microbiology* 4.8 (July 2019), pp. 1253–1267. URL: https://doi.org/10.1038/s41564-019-0491-9.

[46] Katrin Leinweber. *BacDiveR: A Programmatic Interface For BacDive, The DSMZ's Bacterial Diversity Metadatabase*. R package version 0.6.0. 2019. URL: https://github.com/TIBHannover/BacDiveR.

[47] Nathan E Lewis et al. "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models". In: *Molecular Systems Biology* 6.1 (2010), p. 390. eprint: https://www.embopress.org/doi/pdf/10.1038/msb.2010.47. URL: https://www.embopress.org/doi/abs/10.1038/msb.2010.47.

[48] Nicholas J. Loman and Mark J. Pallen. "Twenty years of bacterial genome sequencing". In: *Nature Reviews Microbiology* 13.12 (Nov. 2015), pp. 787–794. URL: https://doi.org/10.1038/nrmicro3565.

[49] Petra Louis, Georgina L Hold, and Harry J Flint. "The gut microbiota, bacterial metabolites and colorectal cancer." In: *Nat. Rev. Microbiol.* 12 (2014), pp. 661–72. URL: https://doi.org/10.1038/nrmicro3344.

[50] Daniel Machado et al. "Fast automated reconstruction of genome-scale metabolic models for microbial species and communities". In: *Nucleic Acids Research* 46.15 (June 2018), pp. 7542–7553. eprint: http://oup.prod.sis.lan/nar/article-pdf/46/15/7542/25689981/gky537.pdf. URL: https://doi.org/10.1093/nar/gky537.

[51] Stefania Magnusdottir et al. "Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota". In: *Nature Biotechnology* (Nov. 2016). URL: http://dx.doi.org/10.1038/nbt.3703.

[52] R. Mahadevan and C. H. Schilling. "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models." eng. In: *Metab Eng* 5.4 (Oct. 2003), pp. 264–276. URL: https://doi.org/10.1016/j.ymben.2003.09.002.

[53] Microsoft and Steve Weston. *foreach: Provides Foreach Looping Construct*. R package version 1.4.7. 2019. URL: https://CRAN.R-project.org/package=foreach.

[54] Jonathan M Monk et al. "iML1515, a knowledgebase that computes Escherichia coli traits". In: *Nature biotechnology* 35.10 (2017), p. 904.

[55] Matthew A Oberhardt, Keren Yizhak, and Eytan Ruppin. "Metabolically re-modeling the drug pipeline". In: *Current Opinion in Pharmacology* 13.5 (2013). Anti-infectives • New technologies, pp. 778–785. URL: http://www.sciencedirect.com/science/article/pii/S1471489213000660.

[56] Kaitlyn Oliphant and Emma Allen-Vercoe. "Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health." In: *Microbiome* 7 (2019), p. 91. URL: https://doi.org/10.1186/s40168-019-0704-8.

[57] H. Pagès et al. *Biostrings: Efficient manipulation of biological strings*. R package version 2.50.2. 2019.

[58] H. Pagès et al. *Biostrings: Efficient manipulation of biological strings*. R package version 2.54.0. 2019.

[59] Jin Hwan Park and Sang Yup Lee. "Towards systems metabolic engineering of microorganisms for amino acid production". In: *Current Opinion in Biotechnology* 19.5 (2008). Tissue, cell and pathway engineering, pp. 454–460. URL: http://www.sciencedirect.com/science/article/pii/S0958166908001006.

[60] Van T. Pham et al. "Early colonization of functional groups of microbes in the infant gut". In: *Environmental Microbiology* 18.7 (May 2016), pp. 2246–2258. URL: https://doi.org/10.1111/1462-2920.13316.

[61] Vanessa V Phelan et al. "Microbial metabolic exchange–the chemotype-to-phenotype link." In: *Nat. Chem. Biol.* 8 (2012), pp. 26–35. URL: https://doi.org/10.1038/nchembio.739.

[62] Morgan N Price et al. "Mutant phenotypes for thousands of bacterial genes of unknown function". In: *Nature* 557.7706 (2018), p. 503.

[63] Sylvain Prigent et al. "Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks". In: *PLOS Computational Biology* 13.1 (Jan. 2017). Ed. by Christoph Kaleta, e1005276. URL: https://doi.org/10.1371/journal.pcbi.1005276.

[64] Rosina Pryor et al. "Host-Microbe-Drug-Nutrient Screen Identifies Bacterial Effectors of Metformin Therapy". In: *Cell* 178.6 (2019), 1299–1312.e29. URL: http://www.sciencedirect.com/science/article/pii/S0092867419308918.

[65] Christopher Quince et al. "Shotgun metagenomics, from sampling to analysis". In: *Nature Biotechnology* 35.9 (Sept. 2017), pp. 833–844. URL: https://doi.org/10.1038/nbt.3935.

[66] Martin H Rau and Ahmad A Zeidan. "Constraint-based modeling in microbial food biotechnology." In: *Biochem. Soc. Trans.* 46 (2018), pp. 249–260. URL: https://doi.org/10.1042/BST20170268.

[67] Lorenz Christian Reimer et al. "BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis". In: *Nucleic Acids Research* 47.D1 (Sept. 2018), pp. D631–D636. eprint: http://oup.prod.sis.lan/nar/article-pdf/47/D1/D631/27436018/gky879.pdf. URL: https://doi.org/10.1093/nar/gky879.

[68] David Ríos-Covián et al. "Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health." eng. In: *Front Microbiol* 7 (2016), p. 185. URL: http://dx.doi.org/10.3389/fmicb.2016.00185.

[69] Fabian Rivera-Chávez and Andreas J. Bäumler. "The Pyromaniac Inside You: Salmonella Metabolism in the Host Gut". In: *Annual Review of Microbiology* 69.1 (2015). PMID: 26002180, pp. 31–48. URL: https://doi.org/10.1146/annurev-micro-091014-104108.

[70] M. H. Saier et al. "The Transporter Classification Database". In: *Nucleic Acids Research* 42.D1 (Dec. 2013), pp. D251–D258. URL: http://dx.doi.org/10.1093/nar/gkt1097.

[71] Eric W Sayers et al. "Database resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 48.D1 (Oct. 2019), pp. D9–D16. URL: https://doi.org/10.1093/nar/gkz899.

[72] Jasjeet S. Sekhon. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R". In: *Journal of Statistical Software* 42.7 (2011), pp. 1–52. URL: http://www.jstatsoft.org/v42/i07/.

[73] Jessica R. Sieber, Michael J. McInerney, and Robert P. Gunsalus. "Genomic Insights into Syntrophy: The Paradigm for Anaerobic Metabolic Cooperation". In: *Annual Review of Microbiology* 66.1 (2012). PMID: 22803797, pp. 429–452. eprint: https://doi.org/10.1146/annurev-micro-090110-102844. URL: https://doi.org/10.1146/annurev-micro-090110-102844.

[74] Felipe A Simão et al. "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." In: *Bioinformatics* 31 (2015), pp. 3210–2. URL: https://doi.org/10.1093/bioinformatics/btv351.

[75] Guy St C Slater and Ewan Birney. "Automated generation of heuristics for biological sequence comparison." In: *BMC Bioinformatics* 6 (2005), p. 31. URL: https://doi.org/10.1186/1471-2105-6-31.

[76] Kornelia Smalla et al. "Analysis of BIOLOG GN Substrate Utilization Patterns by Microbial Communities". In: *Applied and Environmental Microbiology* 64.4 (1998), pp. 1220–1225. eprint: https://aem.asm.org/content/64/4/1220.full.pdf. URL: https://aem.asm.org/content/64/4/1220.

[77] Patrick M Smith et al. "The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis". In: *Science* 341.6145 (2013), pp. 569–573. URL: https://doi.org/10.1126/science.1241165.

[78] SRI International. *PythonCyc*. 2014. URL: https://github.com/latendre/PythonCyc.

[79] Ralf Steuer. "Computational approaches to the topology, stability and dynamics of metabolic networks". In: *Phytochemistry* 68.16 (2007). Dynamic Metabolic Networks, pp. 2139–2151. URL: http://www.sciencedirect.com/science/article/pii/S0031942207002919.

[80] Sergey Stolyar et al. "Metabolic modeling of a mutualistic microbial community". In: *Molecular Systems Biology* 3.1 (2007), p. 92. eprint: https://www.embopress.org/doi/pdf/10.1038/msb4100131. URL: https://www.embopress.org/doi/abs/10.1038/msb4100131.

[81] O. Tange. "GNU Parallel - The Command-Line Power Tool". In: *;login: The USENIX Magazine* (Feb. 2011), pp. 42–47.

[82] The UniProt Consortium. "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 45.D1 (Nov. 2016), pp. D158–D169. eprint: http://oup.prod.sis.lan/nar/article-pdf/45/D1/D158/23819877/gkw1099.pdf. URL: https://doi.org/10.1093/nar/gkw1099.

[83] Ines Thiele and Bernhard Ø. Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction." eng. In: *Nat Protoc* 5.1 (Jan. 2010), pp. 93–121. URL: http://dx.doi.org/10.1038/nprot.2009.203.

[84] Ines Thiele, Nikos Vlassis, and Ronan M T. Fleming. "fastGapFill: efficient gap filling in metabolic networks." eng. In: *Bioinformatics* 30.17 (Sept. 2014), pp. 2529–2531. URL: http://dx.doi.org/10.1093/bioinformatics/btu321.

[85] John D. Trawick and Christophe H. Schilling. "Use of constraint-based modeling for the prediction and validation of antimicrobial targets". In: *Biochemical Pharmacology* 71.7 (2006). Special Issue on Antibacterials, pp. 1026–1035. URL: http://www.sciencedirect.com/science/article/pii/S0006295205007136.

[86] Keith H Turner et al. "Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum". In: *Proceedings of the National Academy of Sciences* 112.13 (2015), pp. 4110–4115.

[87] Amit Varma and Bernhard O. Palsson. "Metabolic Capabilities of Escherichia coli II. Optimal Growth Patterns". In: *Journal of Theoretical Biology* 165.4 (Dec. 1993), pp. 503–522. URL: http://dx.doi.org/10.1006/jtbi.1993.1203.

[88] Edwin C Webb et al. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Ed. 6. Academic Press, 1992.

[89] Hadley Wickham. "Reshaping Data with the reshape Package". In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20. URL: http://www.jstatsoft.org/v21/i12/.

[90] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations.* R package version 1.4.0. 2019. URL: https://CRAN.R-project.org/package=stringr.

[91] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations.* R package version 1.4.0. 2019. URL: https://CRAN.R-project.org/package=stringr.

[92] Ulrike Wittig and Ann De Beuckelaer. "Analysis and comparison of metabolic pathway databases". In: *Briefings in bioinformatics* 2.2 (2001), pp. 126–142.

[93] Bingyao Zhu and Jörg Stülke. "SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis". In: *Nucleic Acids Research* 46.D1 (Oct. 2017), pp. D743–D748. eprint: https://academic.oup.com/nar/article-pdf/46/D1/D743/23162685/gkx908.pdf. URL: https://doi.org/10.1093/nar/gkx908.

[94] Ryan M. Ziels, Masaru K. Nobu, and Diana Z. Sousa. "Elucidating Syntrophic Butyrate-Degrading Populations in Anaerobic Digesters Using Stable-Isotope-Informed Genome-Resolved Metagenomics". In: *mSystems* 4.4 (2019). Ed. by Josh D. Neufeld. URL: https://msystems.asm.org/content/4/4/e00159-19.

[95] Johannes Zimmermann et al. "The functional repertoire contained within the native microbiota of the model nematode Caenorhabditis elegans". In: *The ISME Journal* 14.1 (Sept. 2019), pp. 26–38. URL: https://doi.org/10.1038/s41396-019-0504-y.

[96] Ali R. Zomorrodi, Mohammad Mazharul Islam, and Costas D. Maranas. "d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities". In: *ACS Synthetic Biology* 3.4 (2014). PMID: 24742179, pp. 247–257. eprint: https://doi.org/10.1021/sb4001307. URL: https://doi.org/10.1021/sb4001307.