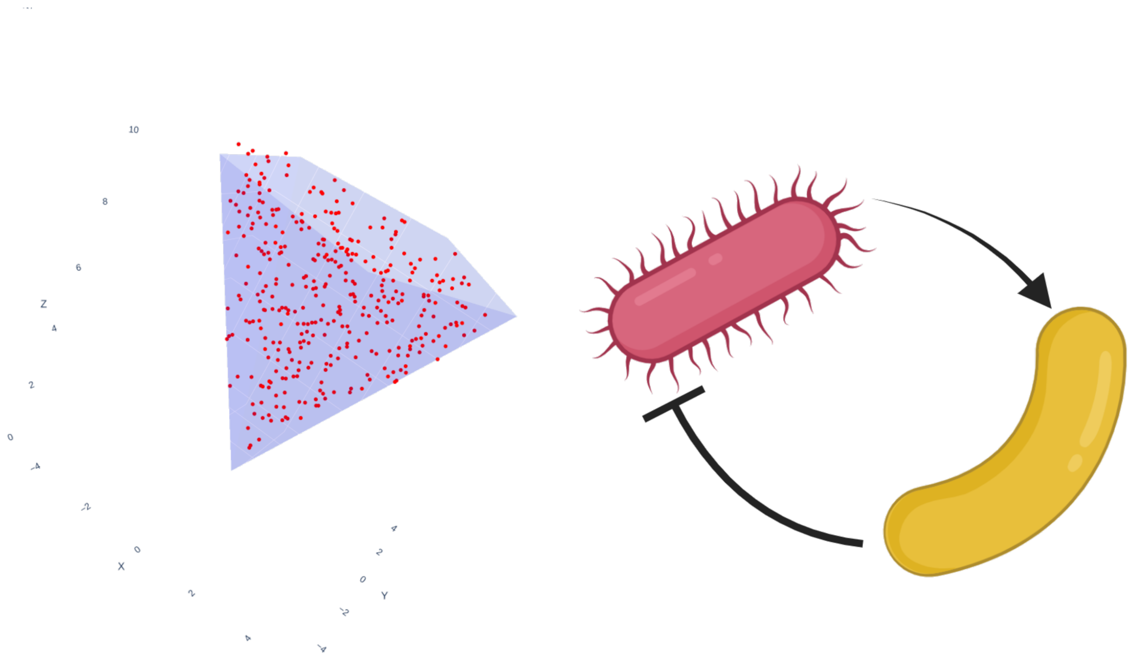
Notes on metabolic modelling analysis

Haris Zafeiropoulos

Dec 09, 2024

CONTENTS



STATUS IN PREPARATION

In this e-book, we will try to cover the basics of metabolic modelling with a special focus in microbial taxa and communities.

Our intention is **not** to come up with a textbook on the field. For this, we recommend a series of books you me fin

Textbooks

- A collaborative open-access textbook: **Economic Principles in Cell Biology** link
- Palsson, B. (2015). Systems biology. **Constraint-based Reconstruction and Analysis**. *Cambridge university press*. DOI
-

CHAPTER

ONE

INTRODUCTION

This is a *book* about microbial metabolic models, their reconstructions and analysis at the strain and the community level. It is intended to give only some insight from the user's perspective and not a thorough background on each analysis presented. Yet, the basics will do be shown but mostly *when* to use a type of analysis, *what* can we learn from it, *how* to interpret their results and what are the assumptions made.

The book contains numerous examples *as programs*, including implementations of many concepts. Each chapter is generated from a self-contained Jupyter Notebook. You can click on the "download" button at the top-right of the chapter, and then select ".ipynb" to download the notebook for that chapter, and you'll be able to execute the examples yourself. Many of the examples are generated by code that is hidden (for readability) in the chapters you'll see here. You can show this code by clicking the "Click to show" labels adjacent to these cells.

This *book* is open source, and the latest version will always be available online here. The source code is available on GitHub. If you would like to fix a typo, suggest an improvement, or report a bug, please open an issue on GitHub.

The techniques described in this book have developed out of the study of *data privacy*. For our purposes, we will define data privacy this way:

Definition 1 (M-models)

Genome-scale metabolic models (**M-models**) provide for a metabolic description of genotype–phenotype relationship without accounting explicitly for synthesis of enzymes. M-models employ Boolean logic statements relating genes, proteins, and reactions, or the Gene–Protein–Reaction associations, or Gene-Protein-Reactions (GPRs). A reaction can only carry a non-zero flux if its GPR statement evaluates to True [1].

integrated models of metabolism and expression (ME-Models) account explicitly for the genotype–phenotype relationship. Macromolecular expression is directly integrated with cellular metabolism [1].

CHAPTER

TWO

METABOLIC MODELS

2.1 Genome-scale models

2.1.1 Background

Definition 2 (flux)

(from [2]) The metabolic flux can be defined as the rate at which material is processed through a metabolic pathway. A reaction's flux refers to the **rate** at which the biochemical reaction proceeds in a biological system. It's a measure of *how quickly* reactants are being converted into products within a specific cellular context.

For insightful visualization that may help you with the concept of a flux, you may have a look here.

In a simplified picture of balanced growth, all metabolic processes are balanced: the rate at which material flows into the cell matches the rate at which it is converted, which again matches the production rate of macromolecule precursors. In addition, we assume that these fluxes are constant, such that the whole metabolic network is in a 'steady-state'. Taken together, we thus assume that the metabolic network can take up and produce external metabolites (e.g. extracellular metabolites and macromolecular precursors), but that all internal metabolites ("inside" the metabolic network) are mass-balanced, that is, for each of these metabolites, production and consumption cancel out.

Since each enzyme has a maximal catalytic rate (the k_{cat} value), a reaction flux will require a certain (minimal) amount of enzyme, which takes up cellular space; since cellular space is limited, fluxes cannot increase infinitely since there is always an upper bound on a weighted sum of reaction fluxes. This constraint implies compromises between different reaction fluxes: one flux can only be increased at the expense of others.

Definition 3 (Balanced growth)

Balanced growth is the average state of a cell in a cell bacterial population growing exponentially at the specific (constant) growth rate \$$\mu \geq 0$\$, i.e. the amount of produced biomass per biomass per cell per unit of time.

The mathematical model:

- *variables* to describe: the metabolic **fluxes** in steady-state metabolism,
- *constraints* to apply: the **balance** of production and consumption of all **internal** metabolites

Importantly, the model will be able to describe compromise: for example, with a given carbon influx and assuming mass balance, the carbon atoms can either be used to generate energy **or** biomass; if one function increases, the other one goes down.

To obtain realistic predictions, we may introduce additional constraints, for example known flux directions or experimentally measured uptake rates.

All this information will not suffice to predict metabolic fluxes precisely, but it allows us to narrow down the possible flux distributions.

$$N \times v = 0 = N \times v^{+} - N \times v^{-} = [N \; -N] \begin{bmatrix} v^{+} \\ v^{-} \end{bmatrix} \tag{2.1}$$

The mass-balance constraints in the previous equation, combined with the property that v_i^{+} , $v_i^{-} \geq 0$ can be expressed in the form

$$A \begin{bmatrix} v^{+} \\ v^{-} \end{bmatrix} \geq 0 \tag{2.2}$$

where:

$$A = \begin{bmatrix} N & -N \\ -N & N \\ I & 0 \\ 0 & I \end{bmatrix}$$

The set of constraints on $(v^{+}$, v^{-}) define a **polyhedral cone** and since they are non-negative, the cone is also pointed, meaning it contains no complete line and the zero vector is the only vertex (extreme point) of the cone.

The space of solutions that satisfies is called the **flux cone**.

2.1.2 Metabolic models with `cobrapy`

2.1.3 Boundary reactions

As described in the `cobrapy` ReadTheDocs, boundary reactions are **unbalanced pseudo reactions**, that means they fulfill a function for modeling by adding to or removing metabolites from the model system but are not based on real biology.

- An **exchange reaction** is a *reversible* reaction that adds to or removes an extracellular metabolite from the extracellular compartment.
- A **demand reaction** is an *irreversible* reaction that *consumes an intracellular metabolite*.
- A **sink** is similar to an exchange, but specifically for *intracellular metabolites*, i.e., a reversible reaction that adds or removes an intracellular metabolite.

An interesting issue on the topic.

Groups

By `group.kind` one may get the kind of group.

Members of a *classification* group should have an *is-a* relationship to the group (e.g. member is-a polar compound, or member is-a transporter).

Members of a *partonomy* group should have a *part-of* relationship (e.g. member is part-of glycolysis).

Members of a *collection* group do not have an implied relationship between the members, so use this value for kind when in doubt (e.g. member is a gap-filled reaction, or member is involved in a disease phenotype).

2.2 Kinetic models

Kinetic models are typically formulated as a set of deterministic **ordinary differential equations (ODEs)**.

Definition 4 (kinetic variables)

kinetic parameters:

- k_{cat}: It is the maximum rate at which an enzyme can catalyze a specific reaction when it is saturated with substrate. It indicates the number of substrate molecules converted into product per enzyme molecule per unit time under optimal conditions. In simpler terms, it reflects how fast an enzyme can convert substrate into product.
- K_M:
- $\frac{k_{cat}}{K_M}$:

Assumptions used in the formulation of biological network models

Assumption	Description
Continuum assumption	Do not deal with individual molecules, but treat medium as a continuum
Finer spatial structure ignored	Medium is homogeneous
Constant-volume assumption	V is time-invariant, $\frac{dV}{dt} = 0$
Constant temperature	Isothermal systems; Kinetic properties a constant
Ignore physico-chemical factors	Electroneutrality and osmotic pressure can be important factors, but are ignored

The **stoichiometric matrix** (S) represents the reaction topology of a network. For an overview on its characteristics see [3].

Definition 5 (gradient matrix)

(from [3]) Each link in a reaction map has kinetic properties with which it is associated. The reaction rates that describe the kinetic properties are found in the rate laws, $v(x; k)$, where the vector k contains all the kinetic constants that appear in the rate laws. Ultimately, these properties represent time constants that tell us how quickly a link in a network will respond to the concentrations that are involved in that link.

The *reciprocal* of these time constants is found in the gradient matrix G, whose elements are

$$g_{ij} = \frac{\partial v_i}{\partial x_j}$$

These constants may change from one member to the next in a biopop- ulation, given the natural sequence diversity that exists. Therefore, the gradient matrix is a *genetically determined* matrix. Two members of the population may have a different G matrix.

Mathematically speaking, G has several challenging features. Unlike the stoichiometric matrix, its numerical values vary over many orders of magnitude. Some links have very fast response times, while others have long response times. The entries of G are real numbers and, therefore, are not “knowable.” The values of G will always come with an error bar associated with the experimental method used to determine them. It has though, the same sparsity properties as the matrix S.

Definition 6 (Jacobian matrix)

S gives us network structure and G gives us kinetic parameters of the links in the network. Their product, the **Jacobian matrix** (J) gives us the network dynamics.

Observation 1

Fluxes are measured in moles per unit of time per cell.

Definition 7 (MASS model)

a metabolic network model that explicitly accounts for the regulatory enzymes, and all their bound states, as components in the network. The result is a data-driven process for constructing mass action stoichiometric simulation (MASS) models that are based on mapping top-down omics data onto bottom-up network reconstructions.

For more about MASS models you may check Palsson's book on dynamic models [4] and one of his papers [5]

2.3 Enzymatic constraints

adssa

2.4 Community models

Here we keep some insight on how to build such models.

From `cite`{wedmark_hierarchy_2024}

Quote!

The community model was made by **merging the individual models into a new model** and **connecting them through exchanged metabolites in a shared compartment**. Specifically, the exchange reactions of the individual models were connected to the shared metabolites and new exchange reactions between the shared compartment and the environment were created, allowing the inter microbial and environmental exchanges described in Fig 1A of Taffs et al. [26]. A **pseudo-metabolite consisting of equal shares of biomass from each of the individual microbes** was constructed and used as the substrate in a community biomass reaction, thus requiring balanced growth of all three microbes.

CHAPTER

THREE

CONSTRAINT BASED ANALYSIS

In this notebook, we are going to discuss some basics of CBA from their mathematical, theoretical and biological perspective.

3.1 Overflow metabolism is caused by two growth-limiting constraints

Literature

- de Groot et al (2020) DOI
- Basan et. al (2015) DOI

Overflow metabolism in Escherichia coli results from efficient proteome allocation.

The authors assume that the yield of energy per carbon molecule is higher for respiration than for fermentation: $n_r > n_f$, but that fermentation is more proteome-efficient: $\epsilon_f > \epsilon_r$.

3.2 Thermodynamically infeasible loops

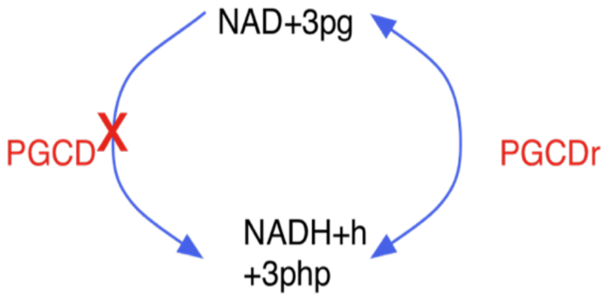
Quote from [6]

Quote!

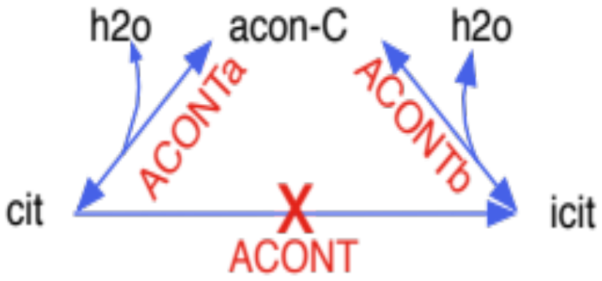
One of the limitations of constraint-based genome-scale models is that the mass balance constraints only describe the net accumulation or consumption of metabolites, without restricting the individual reaction fluxes […] there can be **cycles which do not consume or produce any metabolite**. Therefore, the overall thermodynamic driving force of these cycles are zero, implying that no net flux can flow around this cycle [7].

The authors had a 4-side approach to identify such cycles: they **turned off all the nutrient uptakes** to the cell and used FVA which maximizes and minimizes each of the reaction fluxes subject to mass balance constraints. Fluxes reaching either the lower bound or upper bound were defined as **unbounded reactions** and were grouped together as a **linear combination of the null basis** (lumped) of their stoichiometric matrix. To eliminate the cycles, they:

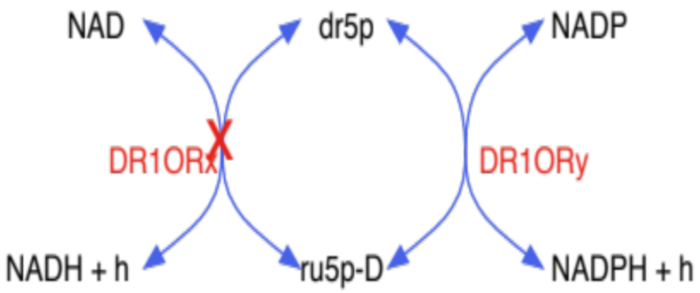
- **removed duplicate reactions:** same reaction occurs both as reversible and irreversible



- **turned off lumped reaction**: a model may include all steps of a process but also a lumped reaction of it:



- **selectively turned reactions on/off based on available cofactor specificity information**: the same biochemical conversion is carried out by different cofactors in the model, while in reality strain would use only one cofactor:



Applying random flux sampling to boundary reactions, we obtained probability distributions over metabolite exchange fluxes.

These probability distributions give an overview of the growth-supporting flux space, but they cannot be directly decomposed to the elementary metabolite conversions provided by pathways.

The issue here is that even in the smallest model, e_coli_core, none of the samples returned included **CO2 uptake**, which was feasible in the model and part of several pathways.

By sampling fluxes separately for each MP, we may find further uptake fluxes and that sampling could complement pathway analysis by allowing detailed analysis of individual pathways' flux spaces.

3.3 Unbiased methods

3.3.1 Minimal pathways

3.4 Dictionary

homogeneous constraint -> right-han-side is zero

CHAPTER

FOUR

HOW TO

In this notebook, we will keep track of handy implementations for metabolic modeling related tasks.

4.1 Linear Programming software

4.1.1 Get CPLEX

When a software says it has CPLEX as a prerequisite, they refer to the ILOG CPLEX Optimization Studio

Analytical decision support toolkit for rapid development and deployment of optimization models using mathematical and constraint programming. It combines an integrated development environment with the powerful Optimization Programming Language and high-performance CPLEX and CP Optimizer solvers.

https://hpc-community.unige.ch/t/guide-for-installing-ibm-cplex/2104

CPLEX has an academic license and once you get this, you can then set Python and other interfaces to use it. However, this process can be rather challenging and more often than not it's far from straightforward.

In this link I found some good enough notes from 2022.

https://academic.ibm.com/a2mt/email-auth#/

4.2 Media and environment setup

4.2.1 Extracellular reactions

COMETS includes the capability to simulate reactions happening in the extracellular environment, without association to a specific organism. Users can implement either elementary reactions of arbitrary order based on mass-action kinetics, or enzyme-catalyzed reactions obeying Michaelis–Menten kinetics, e.g., for the simulation of extracellular enzymes.

4.2.2 Get complete medium for ModelSEED

With the term *complete medium** we describe an *in silico* object where any compound that could be used as a nutrient, it is available for the model.

To build this object for the case of ModelSEED, we need to first get all the possible compounds. And we can do this by first, getting locally the ModelSEEDDatabase repo.

Then we can explore the `Biochemistry` folder of that to retrieve all possible nutrients that could be imported in our model.

From the Biochemistry folder of the dev branch of the ModelSEEDDatabase repository, run:

```
!awk -F"\t" '$6 != 1 && $18==0  {print $5}' reaction_*.tsv   > TRANSPORT_REACTIONS.tsv
```

```
awk: fatal: cannot open file `reaction_*.tsv' for reading: No such file or␣
↪directory
```

Now, with something like the following Python chunk, you can build the complete medium and export in a `.csv` file that with the applied format, could be used for gapfilling with the `fill` command of the `gapseq` tool.

```
def write_to_gapseq_format(all_compounds, cpd2name, output_file):
    """
    Write a 3-col csv file with the compound id, its name and a boundary flux of 1000
    """
    with open(output_file, "w") as f:
        counter = 0
        for compound in all_compounds:
            if compound in cpd2name:
                counter += 1
                f.write(f"{compound}\t{cpd2name[compound]}\t1000\n")
            else:
                print(f"Compound {compound} not found in cpd2name dictionary")

    print(f"Total compounds written: {counter}")


def process_transport_reactions(input_file, output_file=None):
    """
    Parse the TRANSPORT_REACTIONS.tsv file to export compounds that should be part of␣
↪the complete medium.
    """
    with open(input_file) as f:
        lines = f.readlines()

    ex = [line.strip() for line in lines if len(line.split(";")) == 2]

    cpd2name = {}
    all_compounds = set()

    for reaction in ex:
        compounds = reaction.split(";")
        c1 = compounds[0].split(":")[1]
        c2 = compounds[1].split(":")[1]

        if c1 == c2:
            name = compounds[0].split(":")[-1]
```

(continues on next page)

(continued from previous page)

```
            all_compounds.add(c2)
            if c2 not in cpd2name:
                cpd2name[c2] = name

    if output_file is not None:
        write_to_gapseq_format(all_compounds, cpd2name, output_file)

# Main execution
if __name__ == "__main__":
    process_transport_reactions("TRANSPORT_REACTIONS.tsv", "complete_modelseed_medium.
 ↪csv")
```

```
Total compounds written: 0
```

CHAPTER

FIVE

SOFTWARE AND RESOURCES

In this page we will keep a list of the software approaches we are aware of for the various metabolic modeling tasks. Apparently, this list can never be complete, but it can be improved with your contribution! So, feel free to make a PR adding something or contact us to do that for you.

Note: When a hyperlink is given in the *Tool* column, it points to further commentary on the tool within our e-book.

5.1 Reconstruction, gap-flling, validation

- DEMETER
- FROG analysis https://github.com/EBI-BioModels/frog-specification
-

5.1.1 Draft reconstructions

Tool	Description	Architecture	Repo	Documentation	DOI
carveme					
gapseq					
ModelSEEDpy					
PathwayTools					
metage2metabo					

5.1.2 Gap-fillers

Tool	Description	Architecture	Repo	Documentation	DOI
DNNGIOR					

`gapseq`, `ModelSEED` they come with their own gap-filling approaches; see table above for their details.

5.2 Further constraints

5.2.1 Thermodynamics

Tool	Description	Architecture	Repc	DOI
pyTF	implementations of the original thermodynamics-based Flux Analysis (TFA) paper. Specifically, they include explicit formulation of Gibbs energies and metabolite concentrations, which enables straightforward integration of metabolite concentration measurements.	stand alone	GitH	OA
Opt-MDF-pathway**	integration of thermodynamic information in metabolic models to assess the feasibility of flux distributions by thermodynamic driving forces. Extends the framework of Max-min Driving Force (MDF) for thermodynamic pathway analysis. Identifies both the optimal MDF for a desired phenotypic behavior and the respective pathway that supports the optimal driving force.	stand alone	GitH	OA

- matTFA is the MATLAB version of pyTFA ** part of the CNApy library

5.2.2 Enzymes

Tool	Description	Architecture	Repo	DOI
AutoPACMEN	Retrieves kcat data and adds protein allocation constraints to stoichiometric metabolic models according to the sMOMENT method.	standalone	GitHu	OA
ET-GEMs	Construction of enzymatic and thermodynamic constrained GEMs in a single Pyomo* modelling framework.	scripts	GitHu	OA

*Python-based, open-source optimization modeling language

CHAPTER

SIX

DATABASES

Tool	Description	Link	DOI
equilibrator	estimated thermodynamic constants	link	OA
SABIO-RK	curated information about biochemical reactions, their kinetic rate equations with parameters and experimental conditions.	link	OA

has also an API library `equilibrator-api` ReadTheDocs | GitLab

and a cache one `equilibrator-cache`: A database application for caching data used by eQuilibrator and related projects. Stored data includes compound names, structures, protonation state information, reaction and enzyme info, and cross-references to other databases. All compounds stored in equilibrator-cache are cross-referenced using InChIKey. GitLab

6.1 Transporter annotations

Tool	Description	Architecture	Repo	DOI
SPOT	machine learning model that can successfully predict specific substrates for arbitrary transport protein	standalone	GitHu	OA
TranSyT	Identifies transport systems and the compounds carried across membranes, based on the annotations of the Transporter Classification Database (TCDB). Generates the respective transport reactions while providing the respective Gene-Protein-Reaction associations.	web-sevice	GitHu	OA

- They showed that the majority of TCDB entries are of low UniProt-based annotation scores.

6.2 Topological approaches

- QFCA
- **FluxModeCalculator**

Efficient Elementary flux mode (EFM) calculation

paper | GitHub (part of the `MCS` repo)

6.3 Dynamic approaches

- **dfba** GitLab repo Documentation paper A recent approach for dynamic FBA that considers the solution non-uniqueness.

-**COMETS**

-**BacArena**

6.4 Community modelling

- mergem
- PyCoMo
-

6.5 Data integration

- FASTCROMICS
- TRFBA

tellurium

A Python Environment for Reproducible Dynamical Modeling of Biological Networks

6.6 Other resources and tools of interest

Tool	Description	Architecture	Repo	DOI
gRodon2	predicts maximal growth rates using genomic data	R-package	GitHub	OA

CHAPTER

SEVEN

BIOLOGICAL INSIGHT

Here I keep track of bio facts that can be of use in my *in silico* explorations.

Apparently, this notebook will make much less sense since it will include pieces of information that do not necessarily lead somewhere :sweat_smile:

Pseudomonas aeruginosa: May rely on glutamine synthetase and nitrate/nitrite reduction for glutamine production.

7.1 Growth-yield trade-off

Harvey et al. (2014)

They compared a syntrophic consortia to a monoculture with equivalent metabolic capability. They found that **consortia biomass is always lower** than a monoculture with the same metabolic dynamics.

Increasing the **growth rate** or substrate affinity **does not explain** the observed consortial advantage. Increased metabolic pathway efficiency (yield) provides the observed increase in productivity.

7.2 Metabolic strategies

Bacteria can **switch from sequential to co-utilization** of the available resources especially when the **concentrations** of those are **low** Okano et al. [8].

From D'Souza et al. (2014)

- the loss of essential biosynthetic genes was generally beneficial when the required metabolite was sufficiently present in the cells' growth environment,
- the metabolite concentration an auxotroph required to attain WT growth levels differed significantly depending on the metabolite as well as the species analyzed,
- the loss of different genes from the same metabolic pathway resulted in differential fitness consequences for the corresponding mutants, and
- auxotrophs of two species that lacked the same biosynthetic gene responded very differently when exposed to the same concentrations of the required amino acid

Mycobacterium tuberculosis is a shikimate pathway specialist (22); this pathway is strongly related to aromatic amino acid biosynthesis.

7.3 Degenerate pathways

E.coli uses can produce tryptophan through the central metabolic pathway, the shikimic acid (SK) pathway, and the chorismate (CHO) pathways.

CHAPTER

EIGHT

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Bernhard Palsson. *Systems biology*. Cambridge university press, 2015.

[2] Gregory Stephanopoulos. Metabolic fluxes and metabolic engineering. *Metabolic engineering*, 1(1):1–11, 1999.

[3] Bernhard Ø Palsson. *Systems biology: properties of reconstructed networks*. Cambridge university press, 2006.

[4] Bernhard Ø Palsson. *Systems biology: simulation of dynamic network states*. Cambridge University Press, 2011.

[5] Neema Jamshidi and Bernhard Ø Palsson. Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. *Biophysical journal*, 98(2):175–185, 2010.

[6] Mohammad Mazharul Islam, Samodha C Fernando, and Rajib Saha. Metabolic modeling elucidates the transactions in the rumen microbiome and the shifts upon virome interactions. *Frontiers in microbiology*, 10:2412, 2019.

[7] Jan Schellenberger, Nathan E Lewis, and Bernhard Ø Palsson. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical journal*, 100(3):544–553, 2011.

[8] Hiroyuki Okano, Rutger Hermsen, and Terence Hwa. Hierarchical and simultaneous utilization of carbon substrates: mechanistic insights, physiological roles, and ecological consequences. *Current opinion in microbiology*, 63:172–178, 2021.

PROOF INDEX