Construction and Curation Genome Scale Models

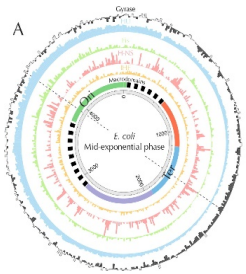
Trunil S. Desai

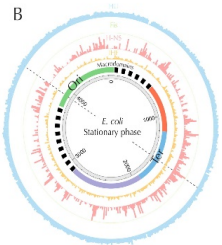
December 4, 2024

Genome-Scale Metabolic Model

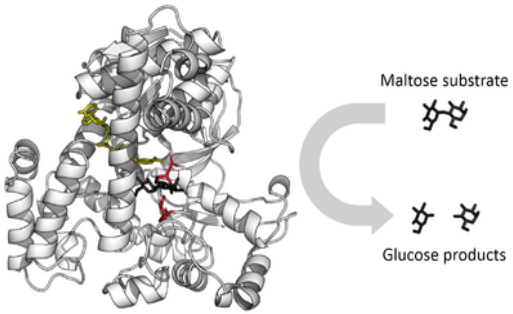
Large size models: usually with 100s to 1000s reactions and metabolites

GSM describes the metabolic interactions in a given organism based on reaction network predicted from enzymes encoded by the genome

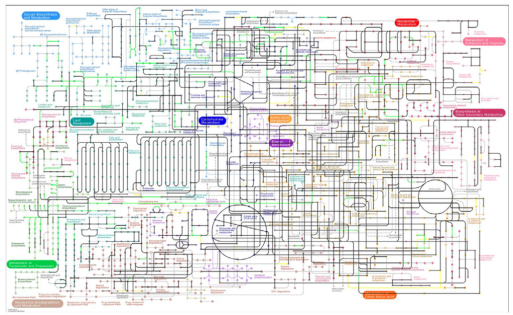




Genome Sequence



Gene prediction and annotation
(Prodigal and Prokka)

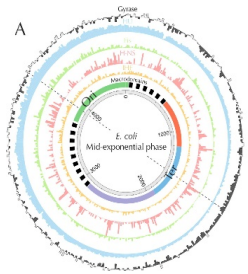


Metabolic Reconstruction
(Pathway Tools)

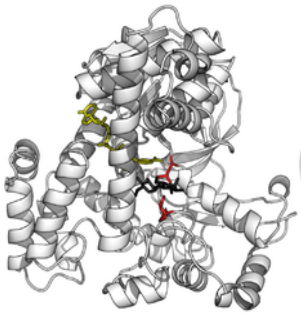
Genome-Scale Metabolic Model

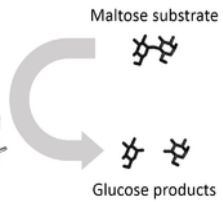
Large size models: usually with 100s to 1000s reactions and metabolites

GSM describes the metabolic interactions in a given organism based on reaction network predicted from enzymes encoded by the genome

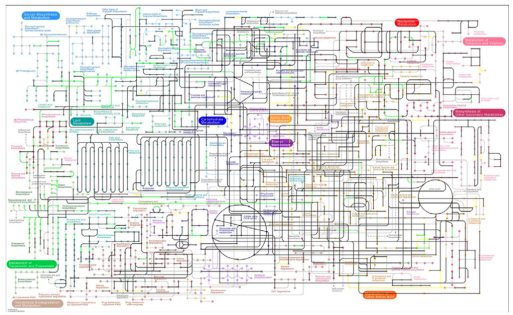


Genome Sequence





Gene prediction and annotation
(Prodigal and Prokka)



Metabolic Reconstruction
(Pathway Tools)

Constructing a genome scale model

Easy !!

- Choose your favourite organism
- Go to your favourite data base
- Save the reactions in a suitable format
- Job done ($\approx$ 1 minute with a local db)

Unfortunately, NO

Constructing a genome scale model

Easy !!

- Choose your favourite organism
- Go to your favourite data base
- Save the reactions in a suitable format
- Job done ($\approx$ 1 minute with a local db)

Unfortunately, NO

Constructing a genome scale model

Easy !!

- Choose your favourite organism
- Go to your favourite data base
- Save the reactions in a suitable format
- Job done ($\approx$ 1 minute with a local db)

Unfortunately, NO

Constructing a genome scale model

Any GSM reconstruction based solely upon data base is likely to exhibit a number of problems:

- inconsistent naming of metabolites and reactions identifiers
- incorrect reaction stoichiometries and reversibility
- missing gene-protein-reaction (GPR) associations

Therefore, in order to represent a realistic representation of a given organism, an initial draft model will require refinement and curation by the user

Constructing a genome scale model

Any GSM reconstruction based solely upon data base is likely to exhibit a number of problems:

- inconsistent naming of metabolites and reactions identifiers
- incorrect reaction stoichiometries and reversibility
- missing gene-protein-reaction (GPR) associations

Therefore, in order to represent a realistic representation of a given organism, an initial draft model will require refinement and curation by the user

Constructing a genome scale model

Any GSM reconstruction based solely upon data base is likely to exhibit a number of problems:

- inconsistent naming of metabolites and reactions identifiers
- incorrect reaction stoichiometries and reversibility
- missing gene-protein-reaction (GPR) associations

Therefore, in order to represent a realistic representation of a given organism, an initial draft model will require refinement and curation by the user

Problems in construction

Non-specific metabolites e.g. :

- “Some-tRNA”
- “Long-Chain-Fatty-Acids"
- “An alcohol”

Inconsistent metabolites identifiers e.g. :

- Multiple identifiers for same metabolite
- Example of Ribose: D-Ribofuranose, D-Ribopyranose, CPD0-1108, CPD0-1110 etc
- Artificially created network discontinuity

Incorrect stoichiometries, e.g.:

"3.2.1.58–RXN": NOTHING –> NOTHING

Problems in construction

Non-specific metabolites e.g. :

- “Some-tRNA”
- “Long-Chain-Fatty-Acids"
- “An alcohol”

Inconsistent metabolites identifiers e.g. :

- Multiple identifiers for same metabolite
- Example of Ribose: D-Ribofuranose, D-Ribopyranose, CPD0-1108, CPD0-1110 etc
- Artificially created network discontinuity

Incorrect stoichiometries, e.g.:

"3.2.1.58–RXN": NOTHING –> NOTHING

Problems in construction

Non-specific metabolites e.g. :

- “Some-tRNA”
- “Long-Chain-Fatty-Acids"
- “An alcohol”

Inconsistent metabolites identifiers e.g. :

- Multiple identifiers for same metabolite
- Example of Ribose: D-Ribofuranose, D-Ribopyranose, CPD0-1108, CPD0-1110 etc
- Artificially created network discontinuity

Incorrect stoichiometries, e.g.:

"3.2.1.58–RXN": NOTHING –> NOTHING

The problem with polymers

Polymeric species, consisting of an undefined number of monomeric units, can give rise to mass inconsistancies e.g

- BioCyc database reports a starch synthesis reaction as:

  “ ADP−D−GLUCOSE” −>“ADP” + “Starch”

- single glucose moiety is added to starch
- Amylase reaction as:

  "Starch" −> "MALTOSE" + "GLUCOSE"

- with an overall conversion of 1 glucose moiety into five glucose moieties

The problem with polymers

Polymeric species, consisting of an undefined number of monomeric units, can give rise to mass inconsistancies e.g

- BioCyc database reports a starch synthesis reaction as:

  “ ADP–D–GLUCOSE” –>“ADP” + “Starch”

- single glucose moiety is added to starch
- Amylase reaction as:

  "Starch" –> "MALTOSE" + "GLUCOSE"

- with an overall conversion of 1 glucose moiety into five glucose moieties

The problem with polymers

Polymeric species, consisting of an undefined number of monomeric units, can give rise to mass inconsistancies e.g

- BioCyc database reports a starch synthesis reaction as:

  ‘‘ ADP–D–GLUCOSE’’ –>‘‘ADP’’ + ‘‘Starch’’

- single glucose moiety is added to starch
- Amylase reaction as:

  "Starch" –> "MALTOSE" + "GLUCOSE"

- with an overall conversion of 1 glucose moiety into five glucose moieties

Stoichiometric inconsistencies

Clearly:

$$A \longrightarrow B$$
$$A \longrightarrow B + C$$

The reactions cannot both be true.
They violate mass conservation)

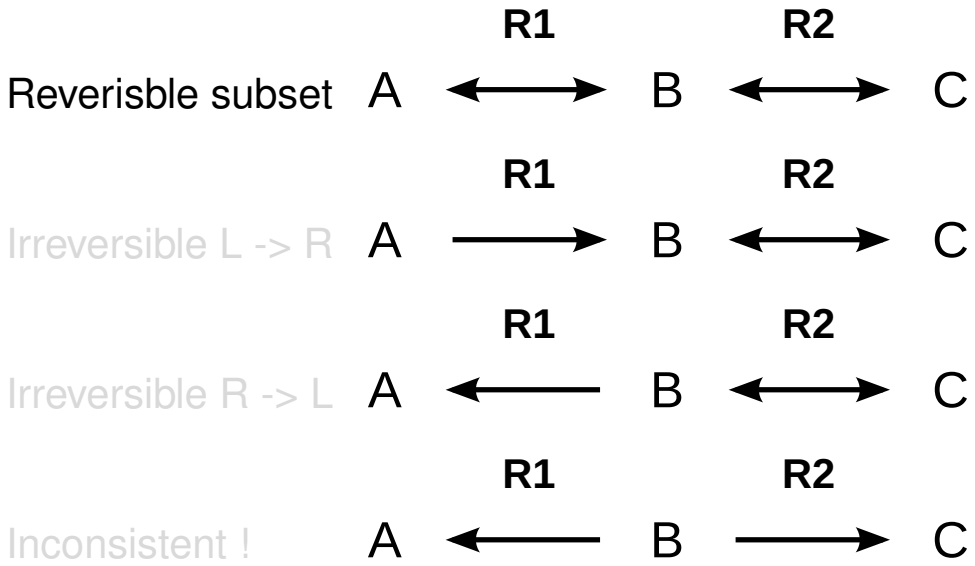
Stoichiometric inconsistencies

Even without empirical formulae such sets of reactions can be identified by a combination of:

- Analysis of left null-space
- Linear programming
- Provides an automatic method for identification of the polymer problem.
- See: Gevorgyan *et al*, 2008, Bioinformatics

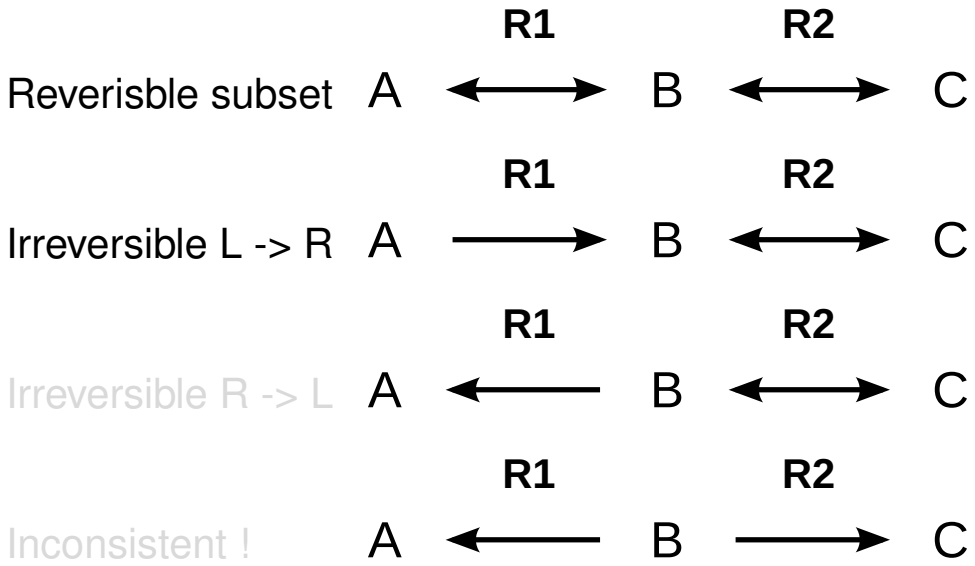
Reaction irreversibility/thermodynamics

Inconsistent subsets:



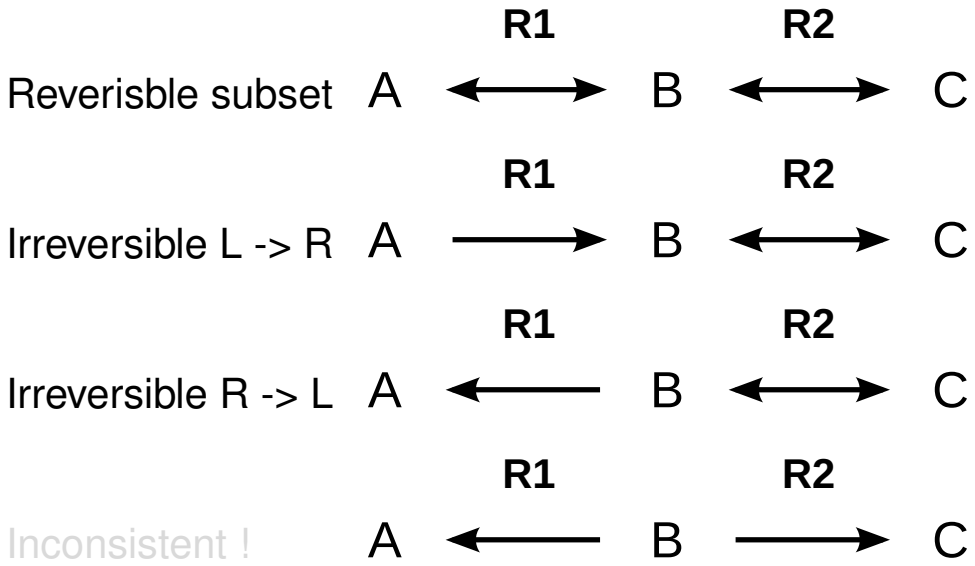
Reaction irreversibility/thermodynamics

Inconsistent subsets:



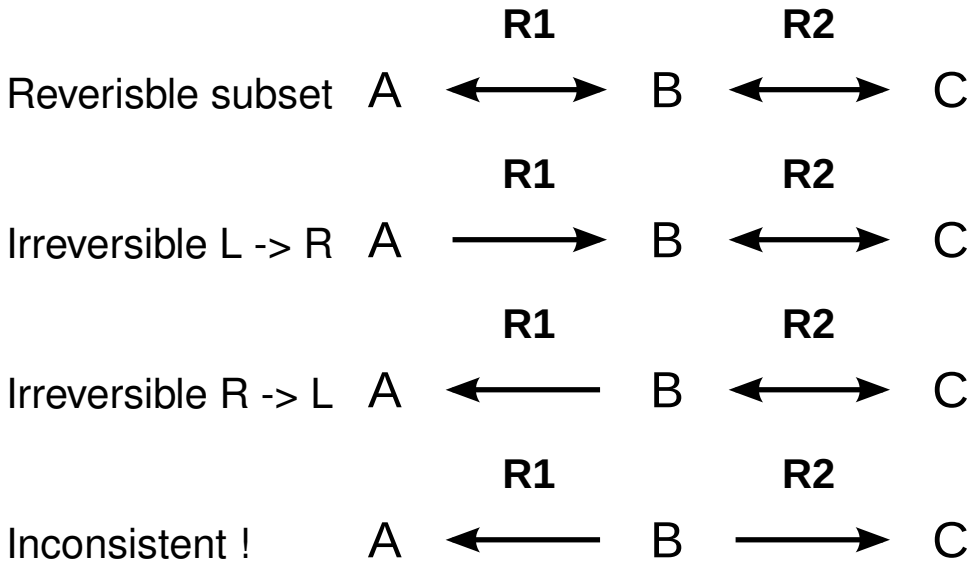
Reaction irreversibility/thermodynamics

Inconsistent subsets:



Reaction irreversibility/thermodynamics

Inconsistent subsets:



Reaction irreversibility/thermodynamics

Violations of conservation of energy:

- Identify from LP:
  1. Constrain all transporters to zero flux.
  2. Set a demand for ATP and/or NAD(P)H.
  3. If the LP has a viable solution an inconsistency exists.
  4. All reactions in the solution must now be examined.

Problems in construction

Mis-annotation:

- Reactions absent that should be present
- Example, Campylobacter model was not able to produce asparagine
- Due to missing reaction in the model
- Reactions present that should be absent
- Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
- This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

Problems in construction

Mis-annotation:

- Reactions absent that should be present
  - Example, Campylobacter model was not able to produce asparagine
  - Due to missing reaction in the model
- Reactions present that should be absent
  - Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
  - This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

Problems in construction

Mis-annotation:

- Reactions absent that should be present
- Example, Campylobacter model was not able to produce asparagine
- Due to missing reaction in the model
- Reactions present that should be absent
- Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
- This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

Problems in construction

Mis-annotation:

- Reactions absent that should be present
- Example, Campylobacter model was not able to produce asparagine
- Due to missing reaction in the model
- Reactions present that should be absent
- Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
- This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

Problems in construction

Mis-annotation:

- Reactions absent that should be present
- Example, Campylobacter model was not able to produce asparagine
- Due to missing reaction in the model
- Reactions present that should be absent
- Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
- This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

Problems in construction

Mis-annotation:

- Reactions absent that should be present
- Example, Campylobacter model was not able to produce asparagine
- Due to missing reaction in the model
- Reactions present that should be absent
- Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
- This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

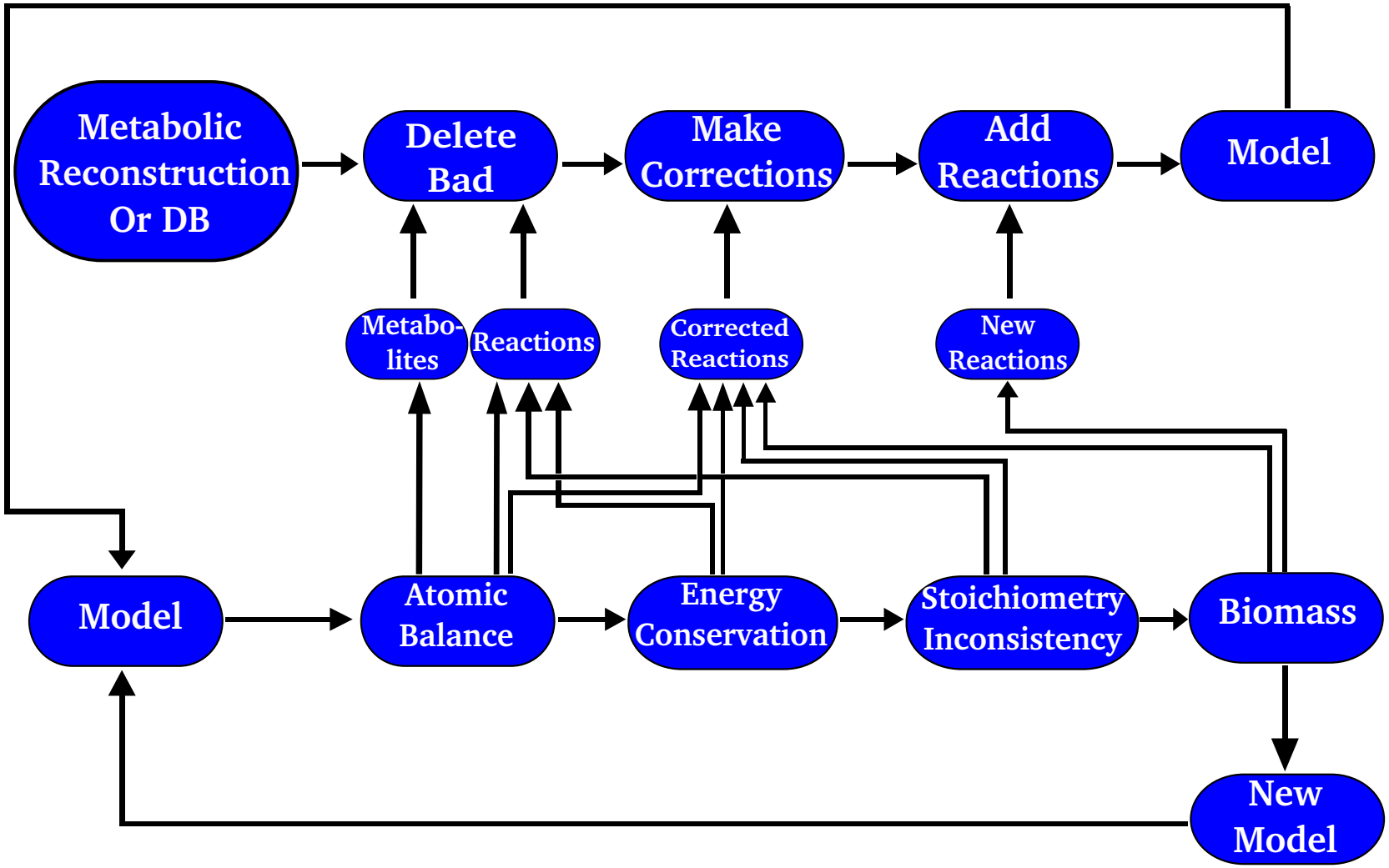
Problems in construction

Mis-annotation:

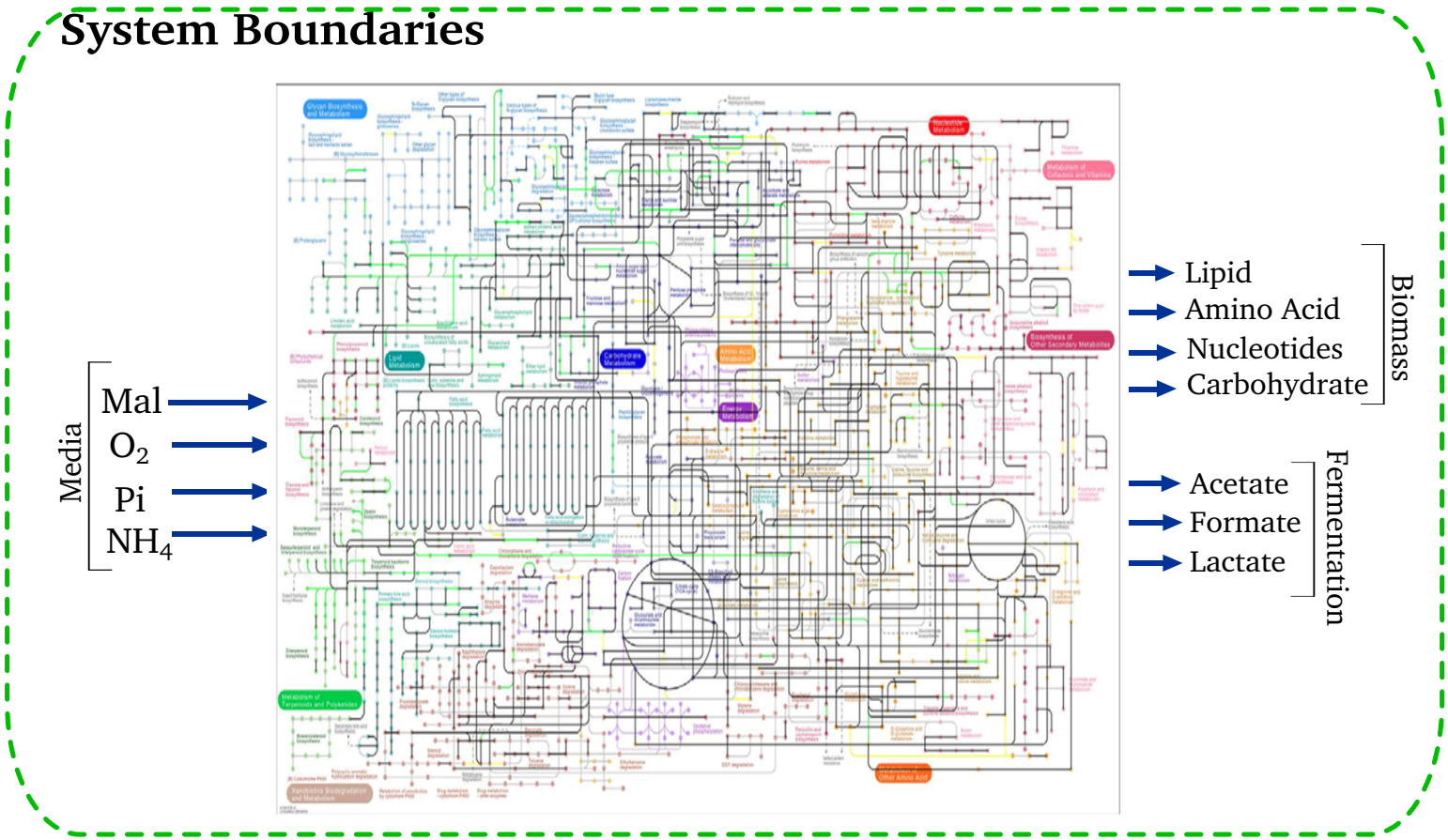
- Reactions absent that should be present
- Example, Campylobacter model was not able to produce asparagine
- Due to missing reaction in the model
- Reactions present that should be absent
- Alternatively, the model was able to synthesis niacinamide, though Campylobacter is auxotrophic to niacinamide
- This was due to over-predicted reaction from Bioinformatics tools

Bioinformatic tools along with experimental data can be used for the curation

Genome-Scale Metabolic Model Construction Pipeline



Genome-Scale Metabolic Model



Naming Convention: Transporters

For our convenience:

- Environmental/external metabolites are distinguished from internal metabolites by using prefix 'x_'
- Transport reactions exchange metabolites between the model and the environment.
  - Transport reactions are differentiated from other reactions by using suffix '_tx'
    - Biomass transporter: 'bm_tx'
    - Media transporter: 'amm_tx'
  - All transport reactions are defined with external species on the left side such that positive flux represents transport of metabolite into the system and negative flux represents transport of metabolites out of the system.

Naming Convention: Transporters

For our convenience:

- Environmental/external metabolites are distinguished from internal metabolites by using prefix 'x_'
- Transport reactions exchange metabolites between the model and the environment.
  - Transport reactions are differentiated from other reactions by using suffix '_tx'
    - Biomass transporters: '_bm_tx'
    - Media transporters: '_mm_tx'
  - All transport reactions are defined with external species on the left side such that positive flux represents transport of metabolite into the system and negative flux represents transport of metabolites out of the system.

Naming Convention: Transporters

For our convenience:

- Environmental/external metabolites are distinguished from internal metabolites by using prefix 'x_'
- Transport reactions exchange metabolites between the model and the environment.
  - Transport reactions are differentiated from other reactions by using suffix '_tx'
    - Biomass transporters: '_bm_tx'
    - Media transporters: '_mm_tx'
  - All transport reactions are defined with external species on the left side such that positive flux represents transport of metabolite into the system and negative flux represents transport of metabolites out of the system.

Naming Convention: Transporters

For our convenience:

- Environmental/external metabolites are distinguished from internal metabolites by using prefix 'x_'
- Transport reactions exchange metabolites between the model and the environment.
  - Transport reactions are differentiated from other reactions by using suffix '_tx'
    - Biomass transporters: '_bm_tx'
    - Media transporters: '_mm_tx'
  - All transport reactions are defined with external species on the left side such that positive flux represents transport of metabolite into the system and negative flux represents transport of metabolites out of the system.

Naming Convention: Transporters

For our convenience:

- Environmental/external metabolites are distinguished from internal metabolites by using prefix 'x_'
- Transport reactions exchange metabolites between the model and the environment.
  - Transport reactions are differentiated from other reactions by using suffix '_tx'
    - Biomass transporters: '_bm_tx'
    - Media transporters: '_mm_tx'
  - All transport reactions are defined with external species on the left side such that positive flux represents transport of metabolite into the system and negative flux represents transport of metabolites out of the system.

Naming Convention: Transporters

For our convenience:

- Environmental/external metabolites are distinguished from internal metabolites by using prefix 'x_'
- Transport reactions exchange metabolites between the model and the environment.
  - Transport reactions are differentiated from other reactions by using suffix '_tx'
    - Biomass transporters: '_bm_tx'
    - Media transporters: '_mm_tx'
  - All transport reactions are defined with external species on the left side such that positive flux represents transport of metabolite into the system and negative flux represents transport of metabolites out of the system.

Naming Convention: Compartments

Reactions between compartments can be differentiated using suffix. e.g. eukaryotic models

- Reactions in cytosol: '_Cyto'
- Reactions in mitochondria: '_Mito'
- Reactions in mitochondria: '_Plas'

Constructing models - Summary

- Constructing GSMs from a database is easy
- Constructing meaningful GSMs from databases require rigorous and methodical curation
- Lastly, not all the published models are correct. Do your quality checks before using them

Constructing models - Summary

- Constructing GSMs from a database is easy
- Constructing meaningful GSMs from databases require rigorous and methodical curation
- Lastly, not all the published models are correct. Do your quality checks before using them

Constructing models - Summary

- Constructing GSMs from a database is easy
- Constructing meaningful GSMs from databases require rigorous and methodical curation
- Lastly, not all the published models are correct. Do your quality checks before using them

Accessing BioCyc From ScrumPy

BioCyc (http://BioCyc.org) a suite of organism specific data-bases ('EcoCyc', 'AraCyc' etc.)
Three levels (tiers) of curation:

1. Exhaustively checked by experts in the field
2. Some manual checking
3. Purely automatically generated

Most are tier 3

Accessing BioCyc From ScrumPy

Organism specific data bases describe:

- Metabolites
- Reactions
- Enzymes
- Proteins
- Genes

This proides a convenient framework for constructing GSMs and establishing reaction/enzyme/gene relationships.

Transporters and compartmental information is relatively poor.

Accessing BioCyc From ScrumPy

A database consists of a set of *Records*.

Each record is identified by its uniqe identifier (UID)

Each record is comprised of a set of *fields* specific to the type of record.

Accessing BioCyc From ScrumPy

ScrumPy provides the 'PyoCyc' package to access local BioCyc data-bases (downloaded from the BioCyc web-site):

```
>>> from ScrumPy.Bioinf import PyoCyc
>>> db = PyoCyc.Organism('Ecoli_MG1655')
```

If no organism name is specified the *MetaCyc* database will be loaded.

Accessing BioCyc From ScrumPy

In ScrumPy, the database acts as a dictionary, the BioCyc UIDs are the keys:

```
>>> g = db['GLC']
```

g is the record for glucose (UID GLC)
This is also a dictionary, the keys refer to the fields of that record.

Accessing BioCyc From ScrumPy

```
>>> print(g)
UNIQUE-ID - GLC
TYPES - Glucopyranose
.
.
CHEMICAL-FORMULA - (C 6)
CHEMICAL-FORMULA - (H 12)
CHEMICAL-FORMULA - (O 6)
.
.
```

Accessing BioCyc From ScrumPy

Likewise for reactions

```
>>> fk = db['FRUCTOKINASE-RXN']
>>> type(fk)
<class 'ScrumPy.Bioinf.PyoCyc.Reaction.Record'>
```

These can also be accessed by EC number:

```
>>> rec = db['EC-2.7.1.4']
>>> rec
[FRUCTOKINASE-RXN]

>>> type(rec[0])
<class 'ScrumPy.Bioinf.PyoCyc.Reaction.Record'>
```

Accessing BioCyc From ScrumPy

Reaction records of a number of properties, including the ability to represent themselves in ScrumPy format:

```
>>> r = db['FRUCTOKINASE-RXN']
>>> print(r.AsScrumPy())

"FRUCTOKINASE-RXN":
      "BETA-D-FRUCTOSE" + "ATP" <> "PROTON" + "FRUCTOSE-6P" + "ADP"
      ~
```