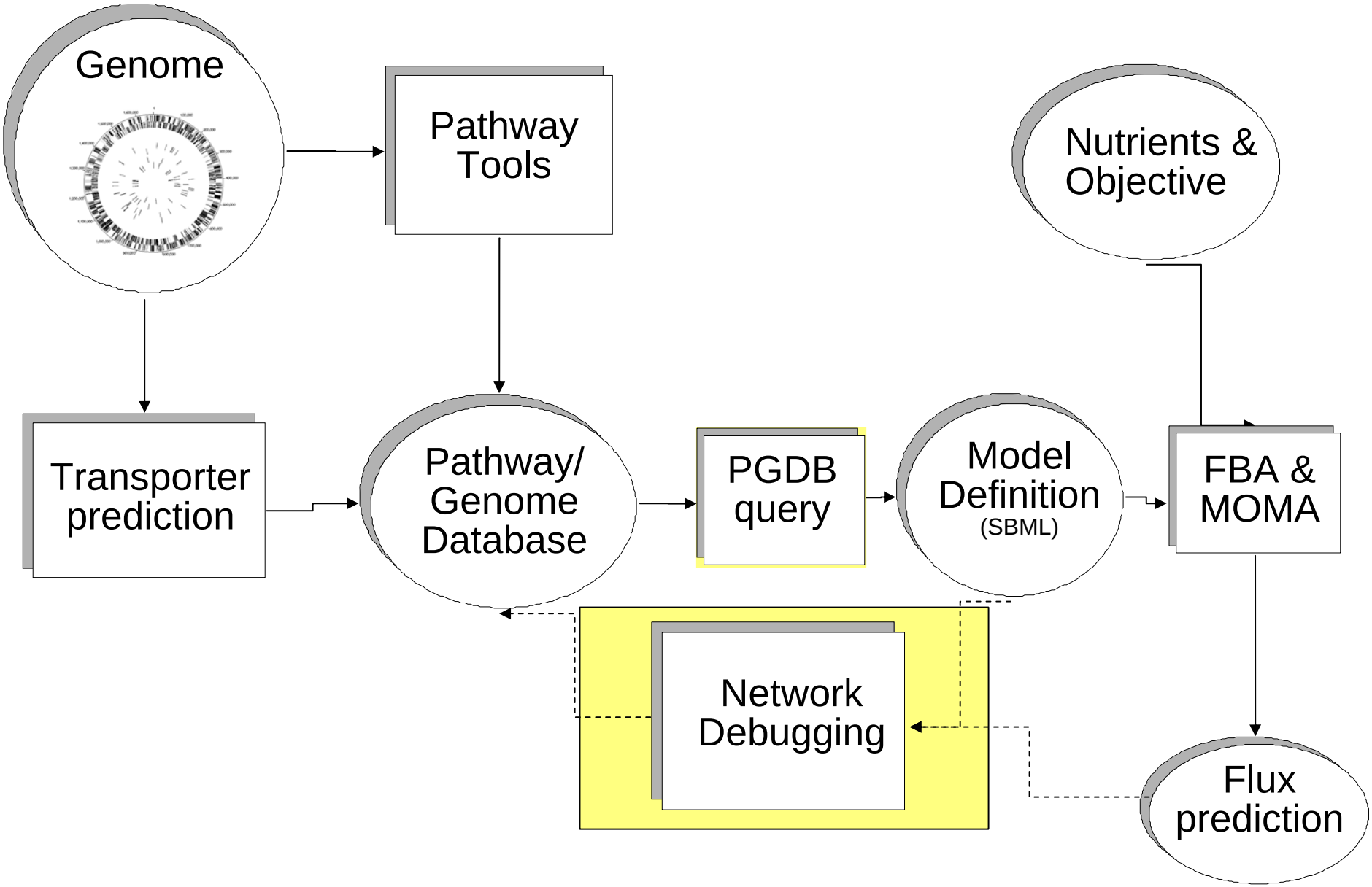
Debugging the Bug

Lessons learned from developing integrating EcoCyc and iJR904

Jeremy Zucker
Dana-Farber Cancer Institute
Harvard Medical School

Automated Metabolic Flux Analysis Pipeline



The project

- The team
- The system
- The questions
- The approach
- The lessons learned
- The results
- The next steps

The Team

- SRI: Markus and Ingrid
- Palsson lab: Adam Feist and Jennifer Reed
- Northwestern: Chris Henry
- York University: Gavin Thomas
- BioPAX: Alan Ruttenberg and Jeremy Zucker

The System: E. coli K12 MG1655

Model	Year(s)	No. of metabolic reactions	No. of metabolites
Majewski and Domach	1990	14	17
Varma and Palsson	1993-1995	146	118
Pramanik and Keasling	1997-1998	300 (317)	289 (305)
Edwards and Palsson	2000	720	436
Covert and Palssonc	2002	113	77
Reed and Palsson	2003	929	626

The System: *E. coli K12 MG1655*

- *KB: EcoCyc 9.5*
- *Goal: Representation of metabolic knowledge*
- *Maturity: 13+ years*
- *Size:*
  - *4480 genes*
  - *1075 enzyme rxns*
  - *977 compounds*

- Model: iJR904
- Goal: Constraint-based metabolic flux model
- Maturity: 13+ years
- Size:
  - 904 genes
  - 929 enzyme rxns
  - 626 compounds

The Goal

- Map iJR904 compounds to EcoCyc compounds
- Map iJR904 reactions to EcoCyc reactions
- Get a working Flux balance model for EcoCyc
- Create FBA enabled BioCyc PGDB's
- Create a disciplined approach to data integration.

The approach

- Automated mapping of compounds
- Semi-automated mapping of compounds
- Manual mapping of compounds
- Automated mapping of reactions
- Semi-automated mapping of reactions
- Manual mapping of reactions

Lessons Learned

- “Standard is better than best”
- “A good representation is the key to good problem solving”--Patrick Winston
- “1+1 is equal but not identical to 2”
- “When it comes to data cleaning, there is no such thing as a free lunch” --Sir Tim Berners-Lee
- “You don't have to do it all by yourself”—Matt Temple

“Standard is better than best”

- Utilize W3C Web Ontology Language (OWL)
- Utilize open source description logic reasoners (Pellet, FaCT++)
- Simplify DL syntax with lisp macros.
- Utilize BioPAX/Pathway tools schema where possible, extend when not possible.

What is OWL?

- 3 flavors: OWL-Lite, OWL-DL, OWL-Full
- Description Logic. Very similar to Frame-based systems
- Description logic reasoners are sound and complete
- A standard in flux (OWL 1.1 coming out soon)

“A good representation is the key to good problem solving”-- Patrick Winston

- CAS and KEGG ID's are supposed to be unique identifiers, so make the unificationXref property inverse functional
- Compounds are classes
- Reactions can be defined in terms of compounds

1+1is equal but not identical to 2

- When are two reactions equal?
  - Same EC number?
  - When does one reaction subsume another?
  - Same enzymes catalyze?
- When are two genes equal?
  - dut vs dfp
- When are two compounds equal?
  - Same CAS number? Same KEGG ID?
  - structural isomers: Glucose vs. Fructose
  - enantiomers: D-Glucose vs L-Glucose
  - Beta-D-glucose vs Alpha-D-glucose

When it comes to data integration, there's no such thing as a free lunch”

– Sir Tim Berners-Lee

- Bugs in GAS and KEGG ID xrefs
- Merged reactions hisD
- Reactions in comments
- Multiple functions of the same gene
- Missing compound mappings

How do I use a DL-reasoner?

- Pellet: Java, accessible from ABCL:

```
(define-ontology simple-class-
   hierarchy ()
  (class !A :partial)
  (class !B :partial)
  (class !C :partial)
  (sub-class-of !B !A)
  (sub-class-of !C !B)
  )
(describe-entity !C (kb simple-
   class-hierarchy))
Direct Superclasses: ex:B
All Superclasses: ex:B, ex:A
Axioms:
(:subclass-of "ex:C" "ex:B")
```

- FaCT++: C++, with lisp syntax interface

```
(equal_c |trans-rxn-157|
 (and (atmost 2 left)
      (atleast 1 left
       |phospho-enol-pyruvate|)
      (atleast 1 left |glc|)
      (atmost 2 right)
      (atleast 1 right |glc-6-p|)
      (atleast 1 right
    |pyruvate|)))
```

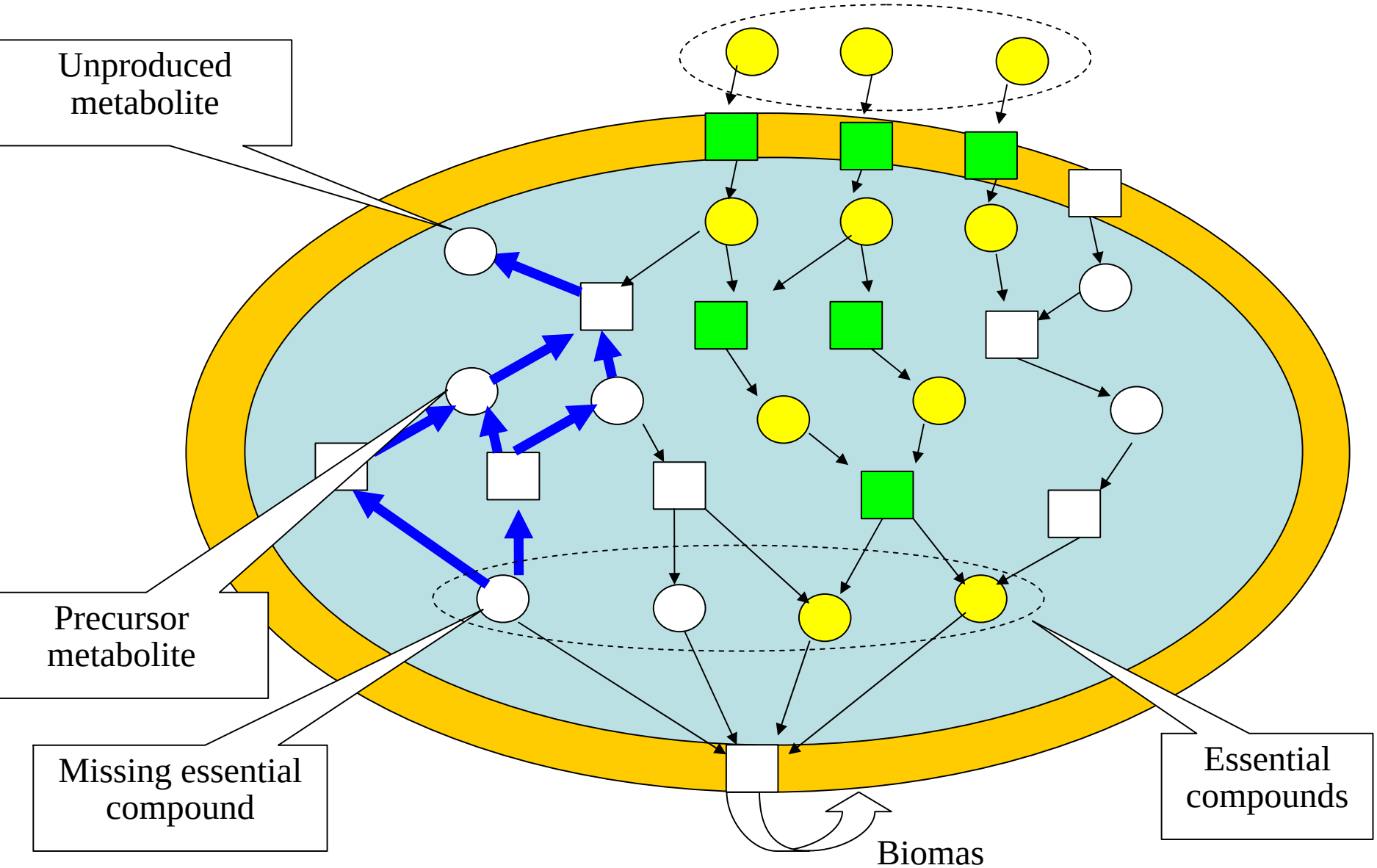
Reasoning with FaCT++

```
 |ecocyc:ribonucleoside-dip-reducti-rxn|
 (and (atmost 3 left)
      (atleast 1 left |ecocyc:ox-thioredoxin|)
      (atleast 1 left |ecocyc:water|)
      (atleast 1 left |ecocyc:deoxy-ribonucleoside-diphosphates|)
      (atmost 2 right)
      (atleast 1 right |ecocyc:red-thioredoxin|)
      (atleast 1 right |ecocyc:ribonucleoside-diphosphates|)))
(equal_c |ijr904:RNDR3|
   (and (atleast 1 right |ijr904:cdp|)
    (atleast 1 right |ijr904:trdrd|)
    (atmost 2 right)
    (atleast 1 left  |ijr904:dcdp|)
    (atleast 1 left  |ijr904:h2o|)
    (atleast 1 left  |ijr904:trdox|)
    (atmost 3 left)))
(equal_c |ijr904:trdrd| |ecocyc:red-thioredoxin|)
(equal_c |ijr904:trdox| |ecocyc:ox-thioredoxin|)
(equal_c |ijr904:dcdp| |ecocyc:dcdp|)
(equal_c |ijr904:h2o| |ecocyc:water|)
(implies |ecocyc:dcdp| |ecocyc:deoxy-ribonucleoside-diphosphates|)
(implies |ecocyc:cdp| |ecocyc:ribonucleoside-diphosphates|)
=> (:SUB "ijr904:RNDR3" :SUPER  ("ecocyc:ribonucleoside-dip-reducti-rxn")))
```

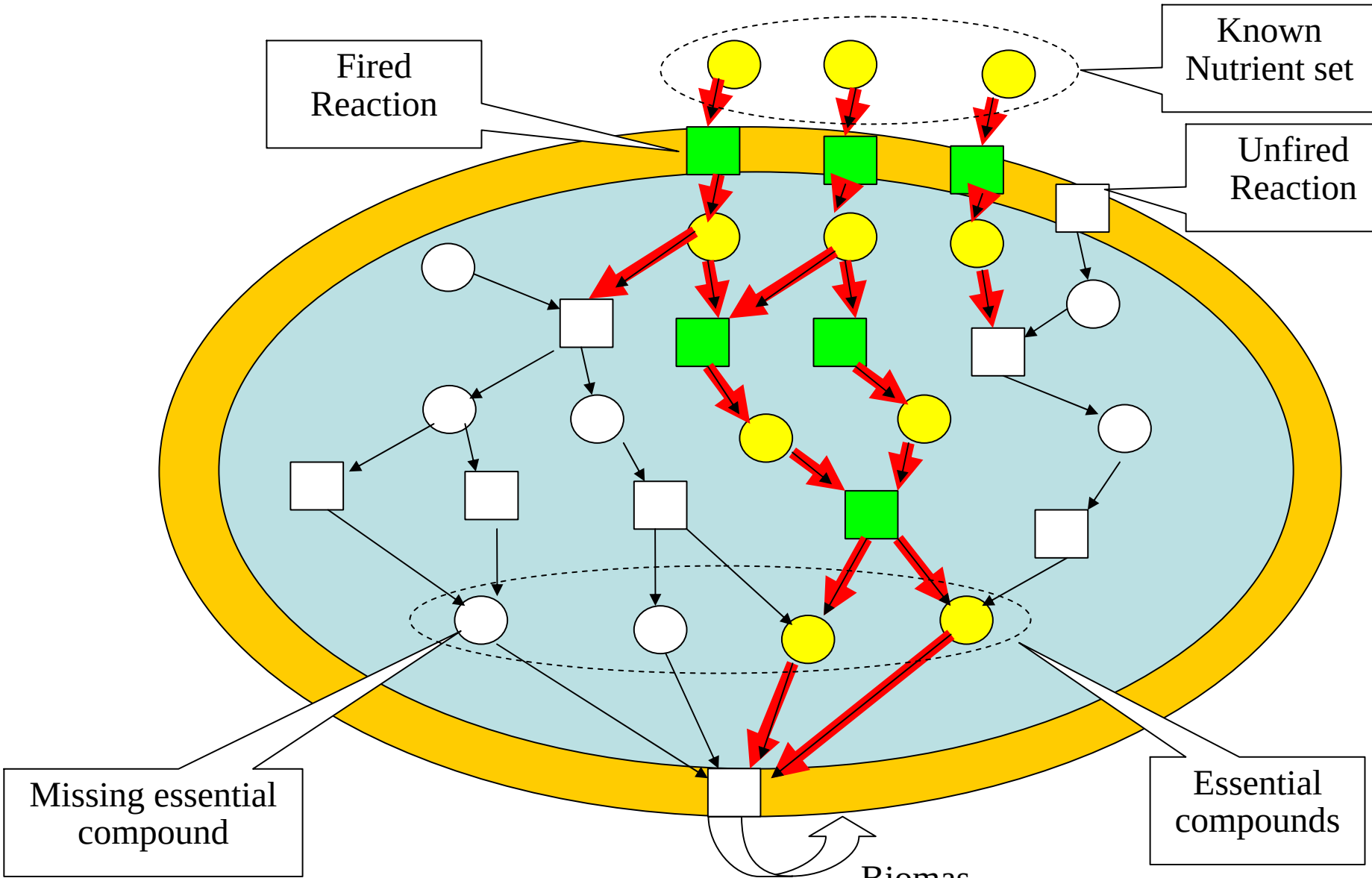
The results (so far)

- 589 out of 773 iJR904 compounds manually mapped to EcoCyc
- 41 errors in EcoCyc compound cross-references
- 80 errors in iJR904 compound cross-references
- 637 out of 937 iJR904 reactions automatically mapped to EcoCyc.
- 83 reaction subsumptions

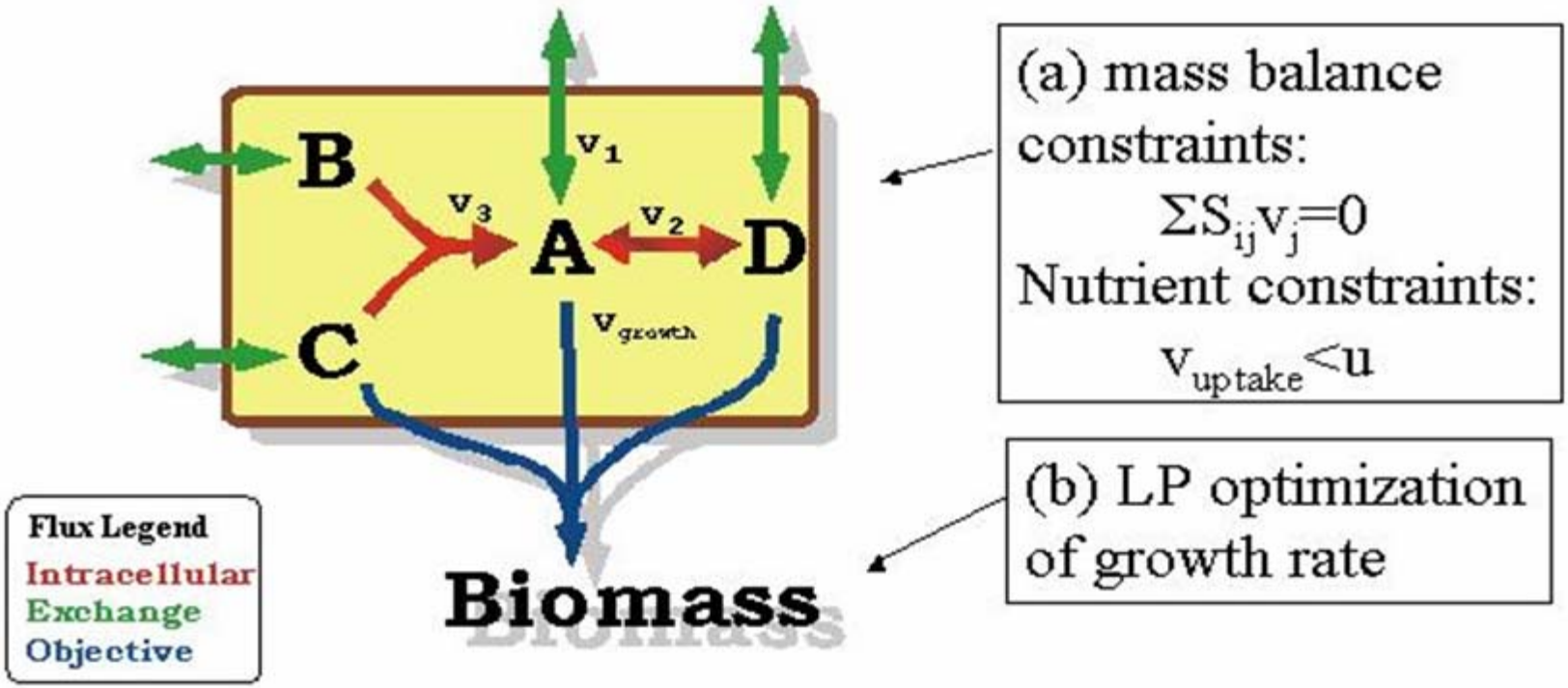
Bugs in Network structure revealed by Forward and Backward chaining



Nutrient analysis: Forward-chaining

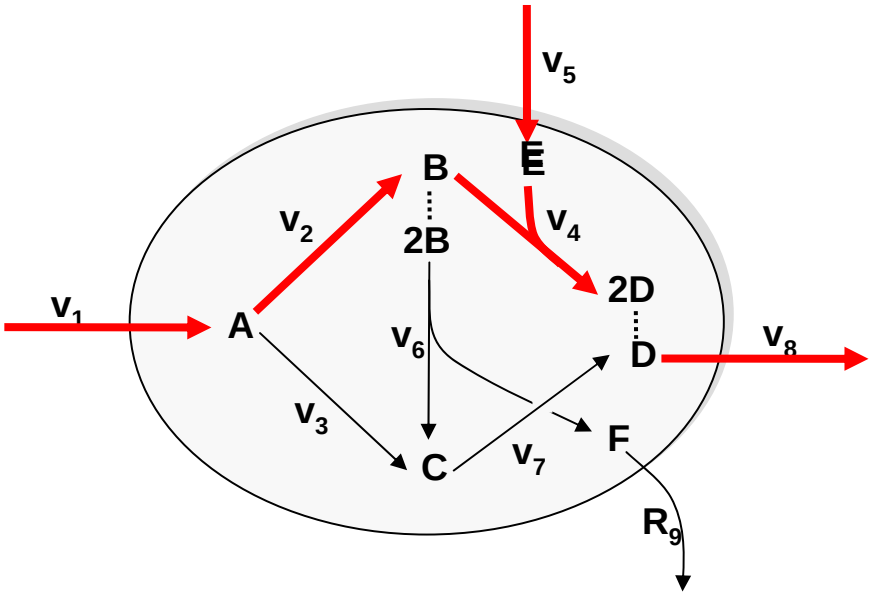


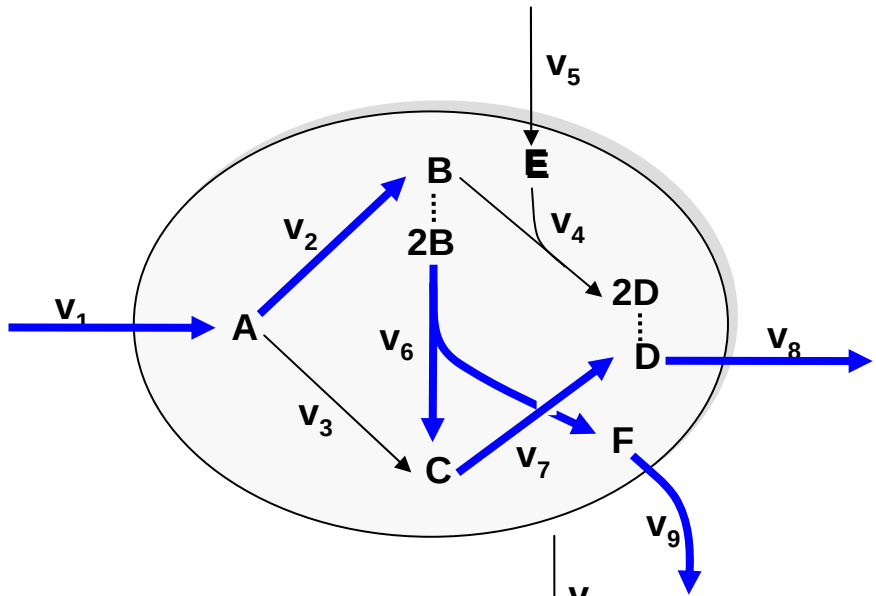
Metabolic flux analysis assumptions



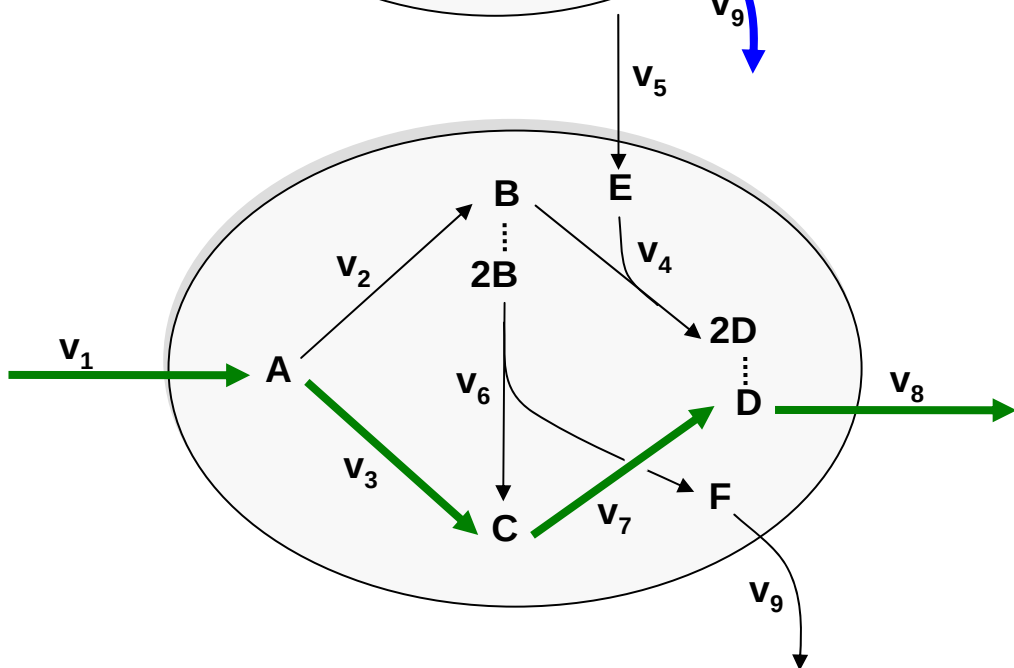
From annotated genomes to metabolic flux models and kinetic parameter fitting
Daniel Segrè et al., *Omics*, 7:301-16 (2003).

First, analyze the steady state-behavior

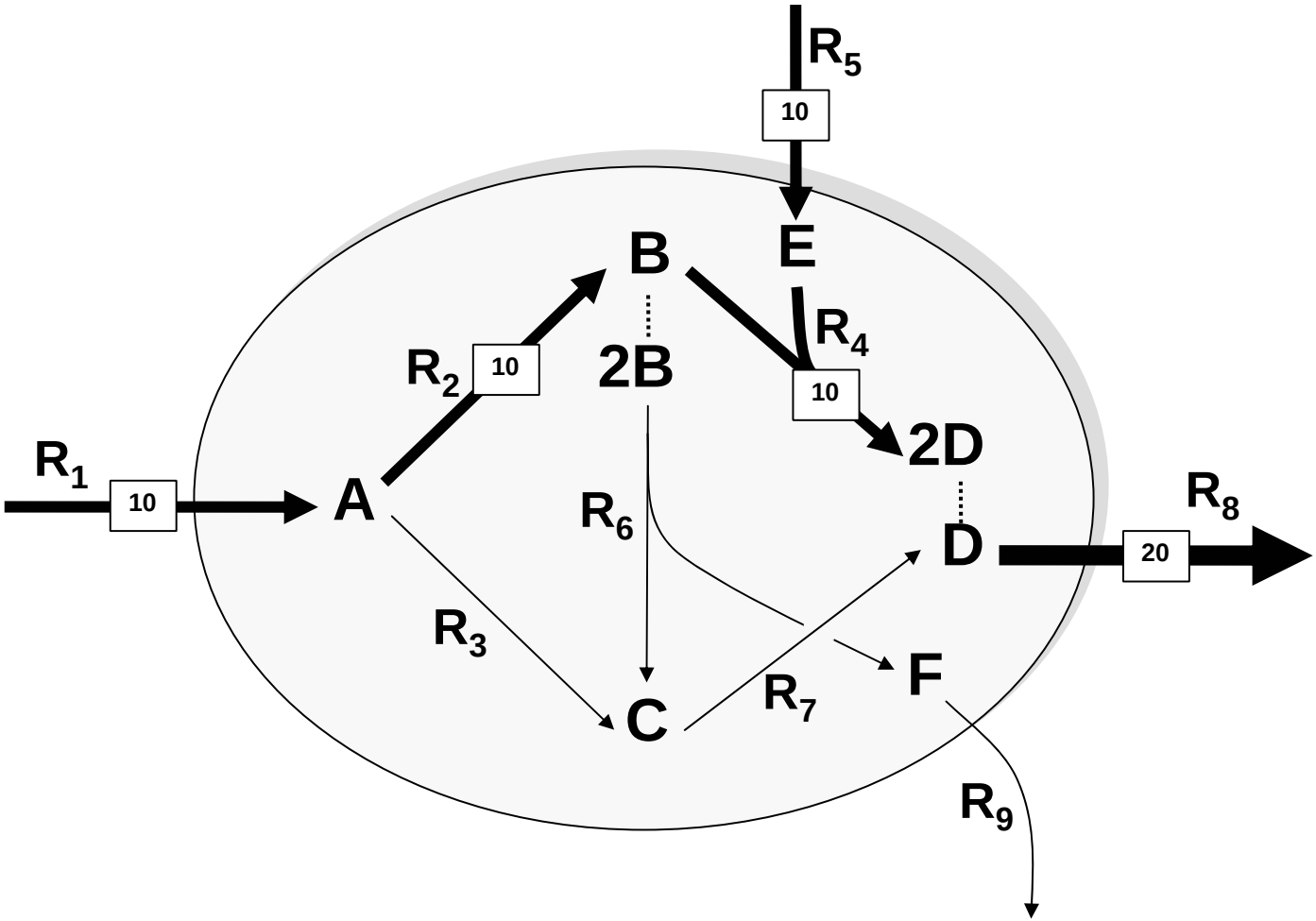




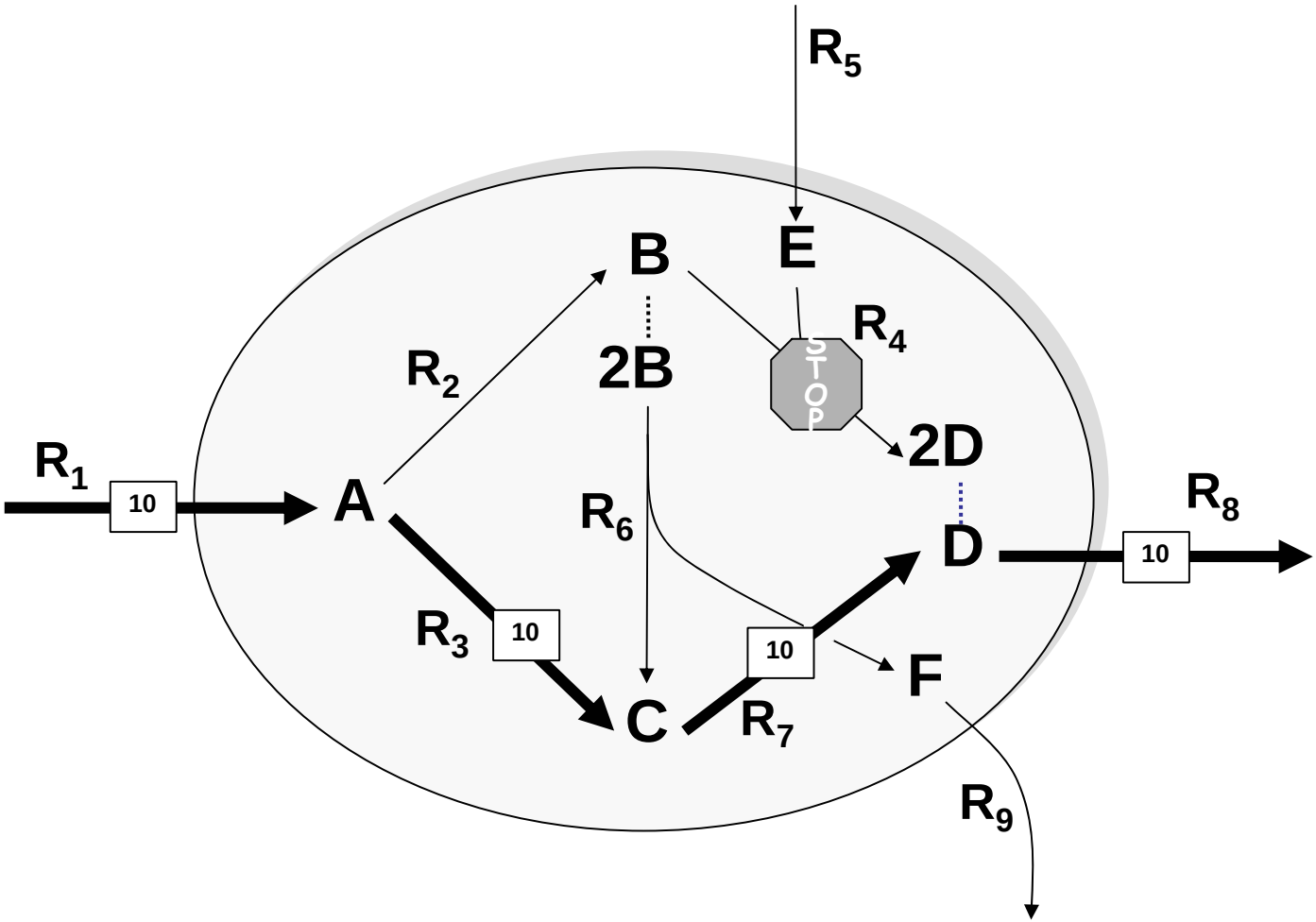
All metabolic capabilities in steady states are composed of elementary flux modes, which are the minimal sets of enzymes that can each generate valid steady states. **--Trends Biotechnol 1999 Feb;17(2):53-60**



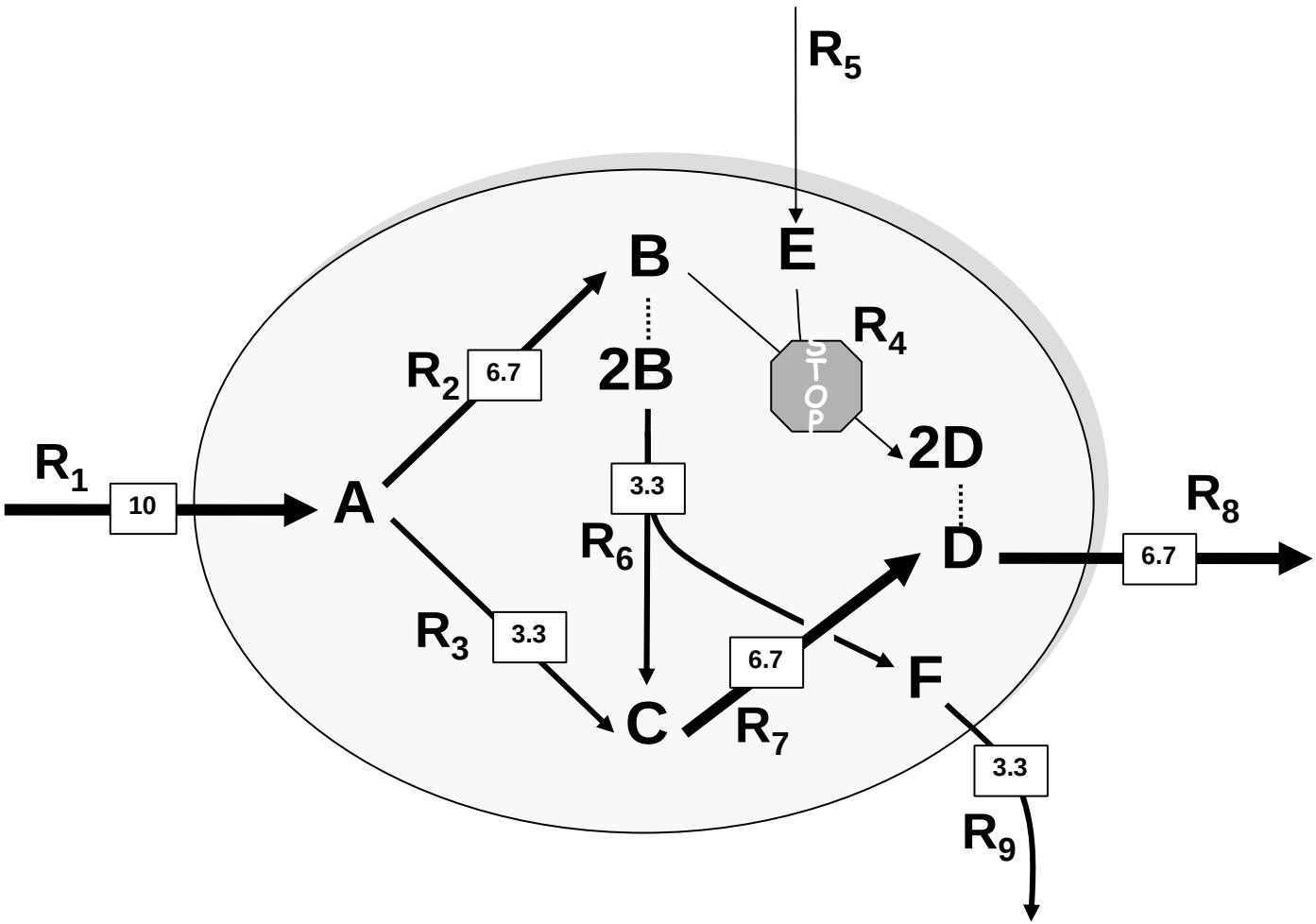
Wild-type FBA Prediction



Mutant FBA Prediction



Mutant MOMA prediction



Classifying the bugs

- bugs in the data (syntax)
  - Mass balance:
  - Thermodynamics (Gibbs)
  - Missing chemical formulas
  - Missing/incorrect gene (Errors in Genome annotation)
  - Missing/incorrect reactions (Orphaned enzymes)
  - Missing pathways (pathway hole filling)
- bugs in the data representation (semantics)
  - Identity problem: Palsson integration, Synonyms, fuzzy names
  - insufficient expressivity: Polymerization, generalized reactions, Transport protein database, SBML
  - Provenance (multiple source problem)
- Bugs in the model assumptions with respect to experiment
  - Growth: FBA vs. MOMA vs. ROOM
  - Internal fluxes: Isotope labelling
  - Epistasis
  - Minimal nutrient sets
  - Expression, Protein levels, and Metabolite concentrations

Diagnose and debug the bug

- Fixing bugs in the data (syntax)
  - Mass balance:
    - balance-p diagnosis uses atomic balances for C S N and P
    - Provenance used to fix the problem with stoichiometry
  - Thermodynamics: include Gibbs free energy
    - Diagnosis: Can be calculated empirically
  - Missing gene:
    - Diagnosis: Network--filling
    - Orphaned enzymes
  - Missing/incorrect gene: (Errors in Genome annotation)
    - Diagnosis: Gene-enzyme-reaction predicate in conjuction with knockout experiments
  - Missing/incorrect reactions (Orphaned enzymes)
    - Diagnosis: Nutrient analysis, backward chaining
  - Missing pathways
- Bugs in the data representation (semantics)
  - Identity problem: Palsson integration, Synonyms, fuzzy names
  - insufficient expressivity: Polymerization, generalized reactions, Transport protein database, SBML
  - Provenance (multiple source problem)
- Bugs in the model assumptions with respect to experiment
  - Growth: FBA vs. MOMA vs. ROOM
  - Internal fluxes: Isotope labelling
  - Epistasis
  - Minimal nutrient sets
  - Expression, Protein levels, and Metabolite concentrations

Debugging bugs

- bugs in the data (syntax)
  - Mass balance:
  - Thermodynamics (Gibbs)
  - Missing chemical formulas
  - Missing/incorrect gene (Errors in Genome annotation)
  - Missing/incorrect reactions (Orphaned enzymes)
  - Missing pathways (pathway hole filling)
- bugs in the data representation (semantics)
  - Identity problem: Palsson integration, Synonyms, fuzzy names
  - insufficient expressivity: Polymerization, generalized reactions, Transport protein database, SBML
  - Provenance (multiple source problem)
- Orphaned enzymes
  - Errors in genome annotation
  - Enzyme genomics project

Bug class	Diagnosis	Fix