Symbolic Systems Biology

Using Formal Logics to Model and Reason about Biological Systems

Carolyn Talcott
SRI International
August 2009

PLAN

- Symbolic systems biology
- Pathway Logic
  - Representation in PL
  - Computing with PL models
  - PL + BioCyc -- first steps
- Minimal nutrient set computation

Symbolic Systems Biology

Symbolic Systems Biology

- Symbolic -- represented in a logical framework
- Systems -- how things interact and work together, integration of multiple parts, viewpoints and levels of abstraction

- Specific Goals:
  - Develop formal models that are as close as possible to domain expert's mental models
  - Compute with, analyze and reason about these complex networks
  - New insights into / understanding of biological mechanisms

Logical Framework

- Making description and reasoning precise
- Language
  - for describing things and/or properties
  - given by a signature and rules for generating expressions (terms, formulas)
- Semantic model -- mathematical structure (meaning)
  - interpretation of terms
  - satisfaction of formulas: M |= wff
- Reasoning -- rules for inferring valid formulae
- Symbolic model -- theory (axioms) used to answer questions

Executable Symbolic Models

- Describe system states and rules for change
- From an initial state, derive a transition graph
  - nodes -- reachable states
  - edges -- rules connecting states
- Path -- sequence of nodes and edges in transition graph (computation / derivation)
- Execution strategy -- picks a path

Symbolic Analysis I

- Static Analysis
  - how are elements organized -- sort hierarchy
  - control flow / dependencies
  - detection of incompleteness
- Forward simulation from a given state (prototyping)
  - run model using a specific strategy
  - fast, first exploration of a model
- Forward collection
  - find potentially reachable states

Symbolic Analysis II

- Search transition graph from a given state S
  - Forward
    - find ALL possible outcomes
    - find only outcomes satisfying a given property
  - Backward
    - find initial states leading to S
- Backward collection
  - find transitions that contribute to reaching S

Symbolic Analysis III

- Model checking
  - determines if all pathways from a given state satisfy a given property, if not a counter example is returned
  - example property:
    - molecule X is never produced before Y
  - counter example:
    - pathway in which Y is produced after X

Symbolic Analysis IV

- Constraint solving
    - Find values for a set of variables satisfying given constraints -- x + y < 1, P or Q
    - MaxSat deals with conflicts
        - weight constraints
        - find solutions that maximize the weight of satisfied constraints
    - Finding possible steady state flows (flux) of information or chemicals through a system can be formulated as a constraint problem.

A Sampling of Formalisms

- Rule-based + Temporal logics
- Petri nets + Temporal logics
- Membrane calculi -- spatial process calculi / logics
- Statecharts + Live sequence charts
- Stochastic transitions systems and logics
- Hybrid Automata + Abstraction

Pathway Logic (PL) Representation of Signaling

http://pl.csl.sri.com/

About Pathway Logic

Pathway Logic (PL) is an approach to modeling biological processes as executable formal specifications (in Maude)

The resulting models can be queried

- using formal methods tools: given an initial state
  - execute --- find some pathway
  - search --- find all reachable states satisfying a given property
  - model-check --- find a pathway satisfying a temporal formula
- using reflection
  - find all rules that use / produce X (for example, activated Rac)
  - find rules down stream of a given rule or component

SIGNALING PATHWAYS

- Signaling pathways involve the modification and/or assembly of proteins and other molecules within cellular compartments into complexes that coordinate and regulate the flow of information.

- Signaling pathways are distributed in networks having stimulatory (positive) and inhibitory (negative) feedback loops, and other concurrent interactions to ensure that signals are propagated and interpreted appropriately in a particular cell or tissue.

- Signaling networks are robust and adaptive, in part because of combinatorial complex formation (several building blocks for forming the same type of complex), redundant pathways, and feedback loops.

About Rewriting Logic

- Rewriting Logic is a logical formalism that is based on two simple ideas
  - states of a system are represented as elements of an algebraic data type
  - the behavior of a system is given by local transitions between states described by rewrite rules
- Rewrite theory: (Signature, Labels, Rules)
  - Signature: (Sorts, Ops, Eqns) -- data, system state
  - Rules have the form label : t => t' if cond
- Rewriting operates modulo equations -- generates computations/pathways

Pathway Logic Organization

A Pathway Logic (PL) system has four parts

- Theops --- sorts and operations
- Components --- specific proteins, chemicals ...
- Rules --- signal transduction reactions
- Dishes --- candidate initial states

Knowledge base: Theops + Components + Rules

Equational part: Theops + Components

A PL cell signaling model is generated from

- a knowledge base
- an initial state (aka dish)

THEOPS

Specifies sorts and operations (data types) used to represent cells:

- Proteins and other compounds
- Complexes
- Soup --- mixtures / solutions / supernatant ...
- Post-translational modifications
- Locations --- cellular compartments refined
- Cells --- collection of locations
- Dishes --- for experiments, think Petri dish

Sample from components

```
sort ErbB1L . subsort ErbB1L < Protein . *** ErbB1 Ligand

op Egf : -> ErbB1L [metadata "(\
  (spname EGF_HUMAN)\
  (spnumber P01133)\
  (hugosym EGF)\
  (category Ligand)\
  (synonyms \"Pro-epidermal growth factor precursor, EGF\" \
            \"Contains: Epidermal growth factor, Urogastrone \"))"] .


op EgfR : -> Protein [metadata "(\
  (spname EGFR_HUMAN)\
  (spnumber P00533)\
  (hugosym EGFR)\
  (category Receptor)\
  (synonyms \"Epidermal growth factor receptor precursor\" \
            \"Receptor tyrosine-protein kinase ErbB-1, ERBB1 \"))"] .


op PIP2 : -> Chemical [metadata "(\
   (category Chemical)\
   (keggcpd C04569)\
   (synonyms \"Phosphatidylinositol-4,5P \" ))"] .
```

Example Rule

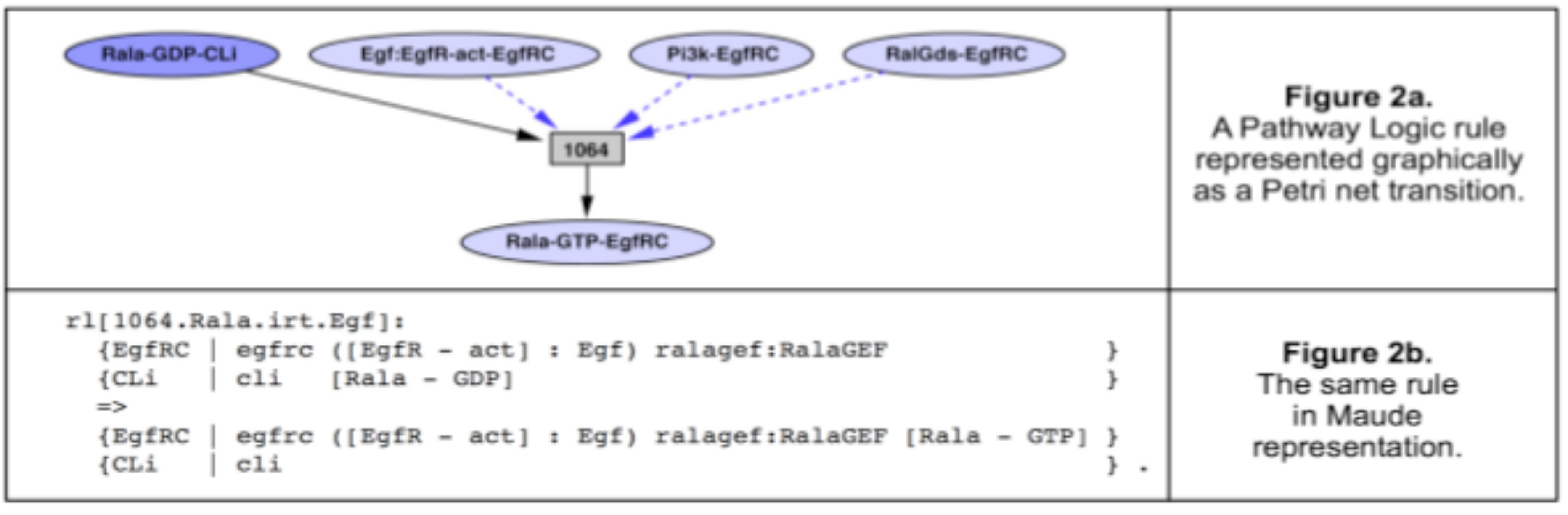


Figure 2a.
A Pathway Logic rule represented graphically as a Petri net transition.

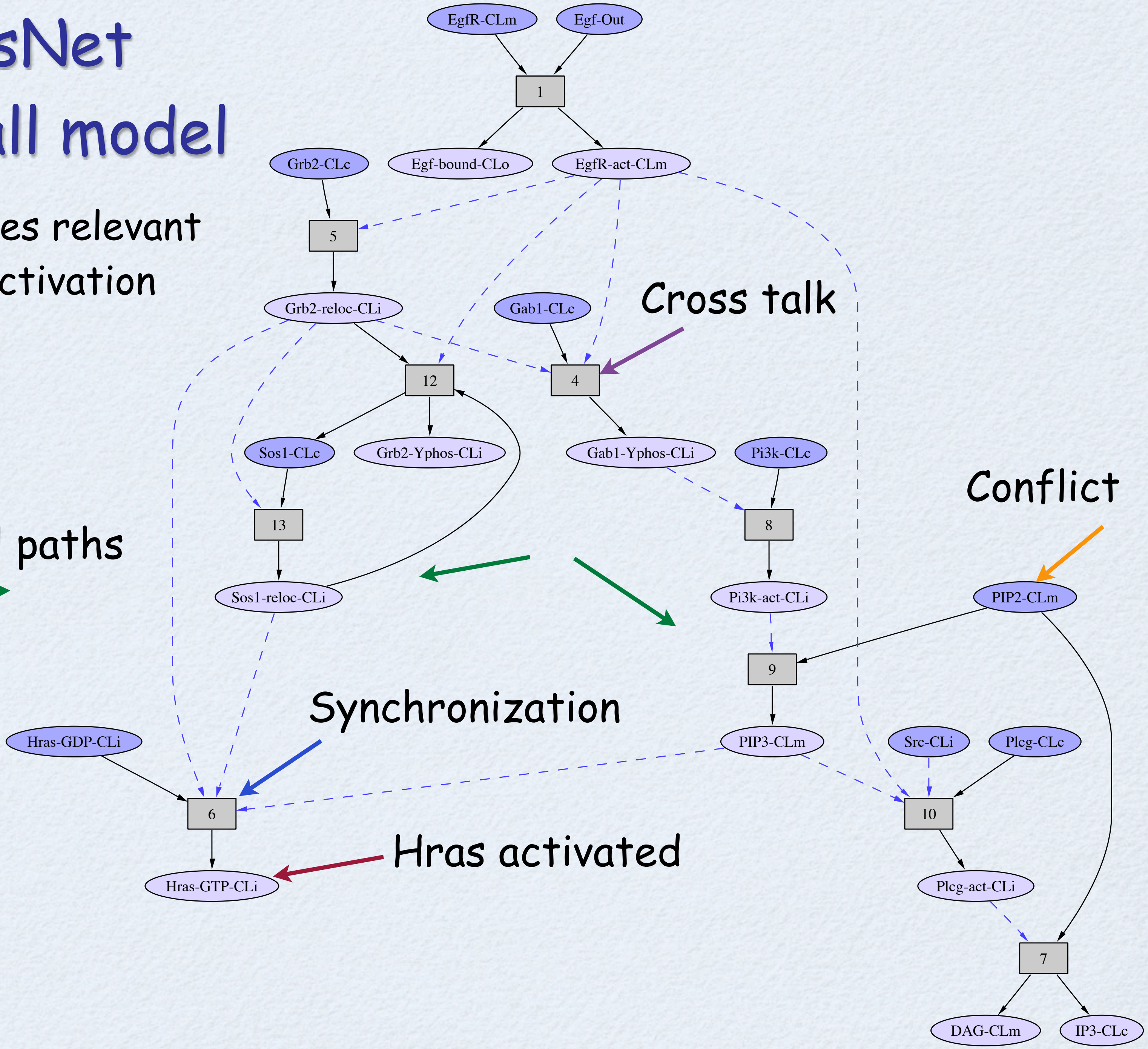
```
rl[1064.Rala.irt.Egf]:
  {EgfRC | egfrc ([EgfR - act] : Egf) ralagef:RalaGEF              }
  {CLi   | cli   [Rala - GDP]                                      }
  =>
  {EgfRC | egfrc ([EgfR - act] : Egf) ralagef:RalaGEF [Rala - GTP] }
  {CLi   | cli                                                     } .
```

Figure 2b.
The same rule in Maude representation.

rasNet
A small model

Rule instances relevant to Hras activation

Parallel paths



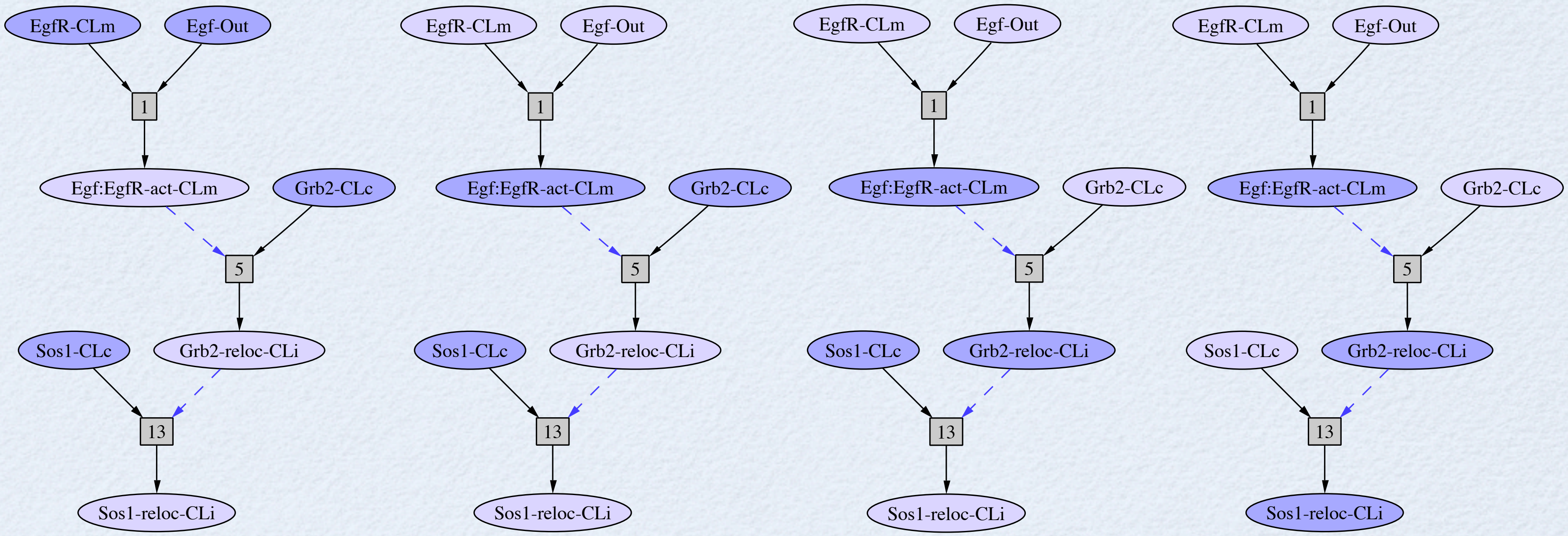
Cross talk

Conflict

Synchronization

Hras activated

Rule Execution as Petri Nets



rasDish =rule1=> rasDish1 =rule5=> rasDish2 =rule13=> rasDish3

Ovals are occurrences -- components in locations.

Dark ovals are present in the current state (marked).

Squares are rules.

Dashed edges connect components that are not changed.

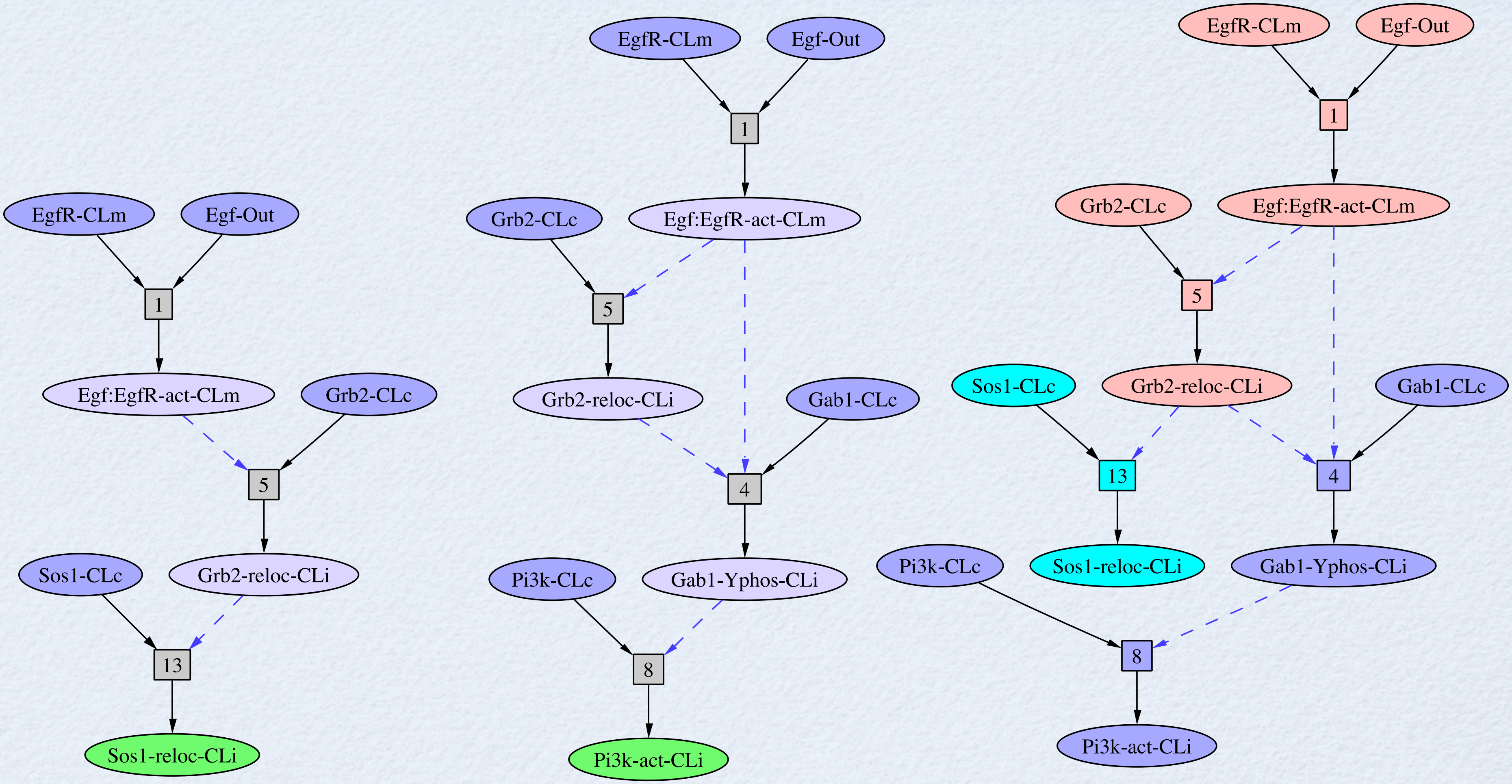
The Pathway Logic Assistant (PLA)

- Provides a means to interact with a PL model
- Manages multiple representations
  - Maude module (logical representation)
  - PetriNet (process representation for efficient query)
  - Graph (for interactive visualization)
- Exports Representations to other tools
  - Lola (and SAL model checkers)
  - Dot -- graph layout
  - JLambda (interactive visualization, Java side)
  - SBML (xml based standard for model exchange)

A Simple Query Language

- Given a Petri net with transitions P and initial marking O (for occurrences) there are two types of query
  - subnet
  - findPath - a computation / unfolding
- For each type there are three parameters
  - G: a goal set---occurrences required to be present at the end of a path
  - A: an avoid set---occurrences that must not appear in any transition fired
  - H: as list of identifiers of transitions that must not be fired
- findPath returns a pathway (transition list) generating a computation satisfying the requirements.
- subnet returns a subnet containing all (minimal) such pathways.

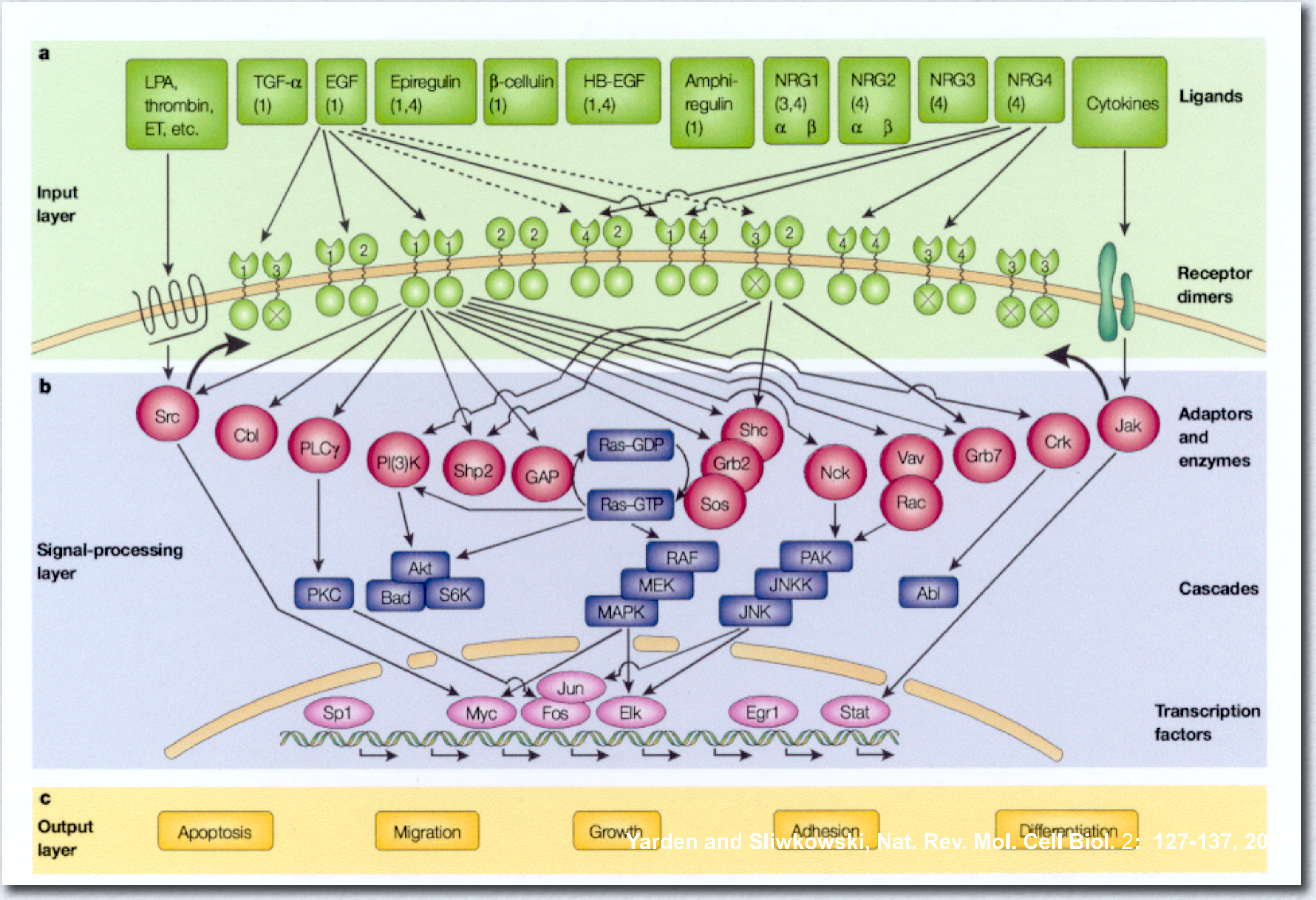
Pathway Examples



Full Model of EGF Stimulation

(by Merrill Knapp)

The ErbB network (cartoon form)

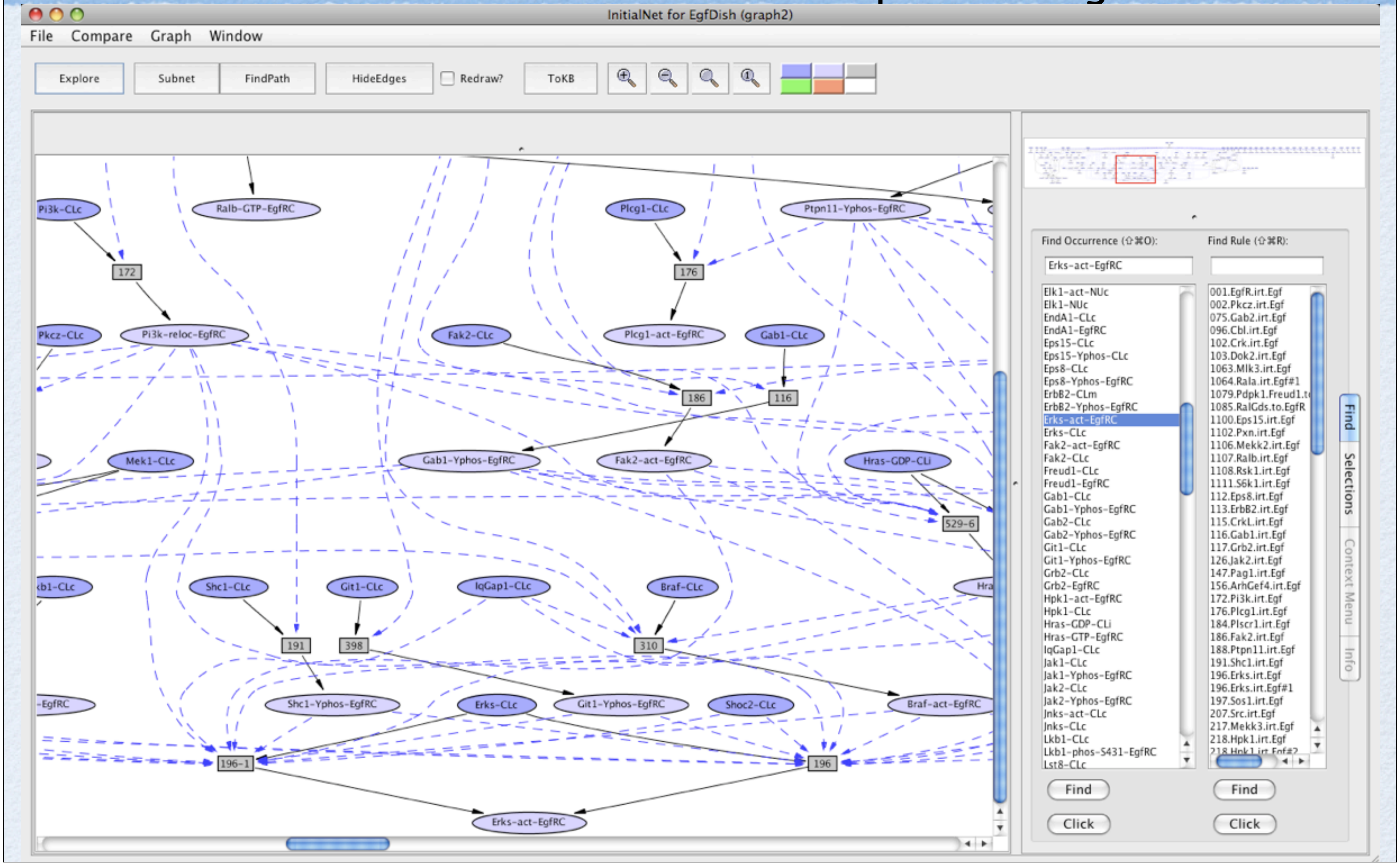


Yarden and Sliwkowski, Nat. Rev. Mol. Cell Biol. 2: 127-137, 20

PL EGF MODEL

Events that could occur in response to Egf

Curated by
Merrill Knapp



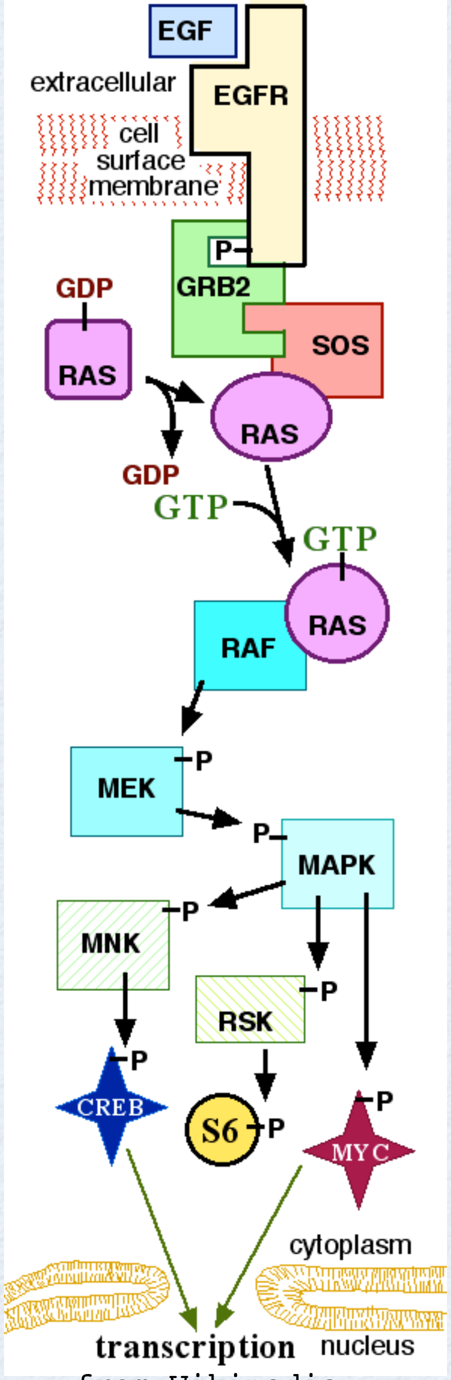
Find Occurrence (⇧⌘O):

Erks-act-EgfRC

Elk1-act-NUc
Elk1-NUc
EndA1-CLc
EndA1-EgfRC
Eps15-CLc
Eps15-Yphos-CLc
Eps8-CLc
Eps8-Yphos-EgfRC
ErbB2-CLm
ErbB2-Yphos-EgfRC
Erks-act-EgfRC
Erks-CLc
Fak2-act-EgfRC
Fak2-CLc
Freud1-CLc
Freud1-EgfRC
Gab1-CLc
Gab1-Yphos-EgfRC
Gab2-CLc
Gab2-Yphos-EgfRC
Git1-CLc
Git1-Yphos-EgfRC
Grb2-CLc
Grb2-EgfRC
Hpk1-act-EgfRC
Hpk1-CLc
Hras-GDP-CLi
Hras-GTP-EgfRC
IqGap1-CLc
Jak1-CLc
Jak1-Yphos-EgfRC
Jak2-CLc
Jak2-Yphos-EgfRC
Jnks-act-CLc
Jnks-CLc
Lkb1-CLc
Lkb1-phos-S431-EgfRC
Lst8-CLc

Find Rule (⇧⌘R):

001.EgfR.irt.Egf
002.Pkcz.irt.Egf
075.Gab2.irt.Egf
096.Cbl.irt.Egf
102.Crk.irt.Egf
103.Dok2.irt.Egf
1063.Mlk3.irt.Egf
1064.Rala.irt.Egf#1
1079.Pdpk1.Freud1.t
1085.RalGds.to.EgfR
1100.Eps15.irt.Egf
1102.Pxn.irt.Egf
1106.Mekk2.irt.Egf
1107.Ralb.irt.Egf
1108.Rsk1.irt.Egf
1111.S6k1.irt.Egf
112.Eps8.irt.Egf
113.ErbB2.irt.Egf
115.CrkL.irt.Egf
116.Gab1.irt.Egf
117.Grb2.irt.Egf
126.Jak2.irt.Egf
147.Pag1.irt.Egf
156.ArhGef4.irt.Egf
172.Pi3k.irt.Egf
176.Plcg1.irt.Egf
184.Plscr1.irt.Egf
186.Fak2.irt.Egf
188.Ptpn11.irt.Egf
191.Shc1.irt.Egf
196.Erks.irt.Egf
196.Erks.irt.Egf#1
197.Sos1.irt.Egf
207.Src.irt.Egf
217.Mekk3.irt.Egf
218.Hpk1.irt.Egf
218.Hpk1.irt.Egf#2

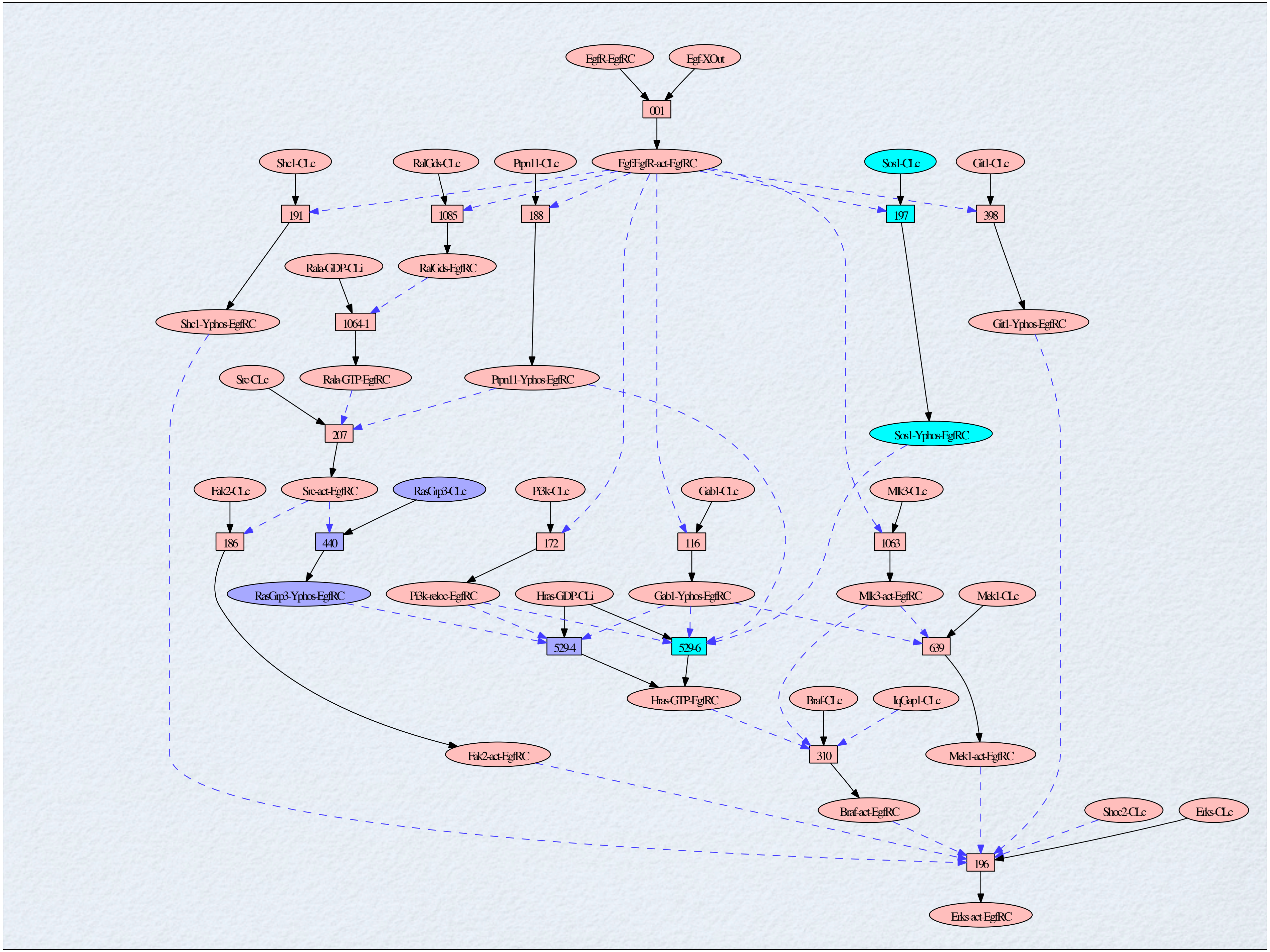


from Wikipedia

Egf stimulation of the Mitogen Activated Protein Kinase (MAPK) pathway.

Egf → EgfR → Grb2 → Sos1 → Ras → Raf1 → Mek → Erk

- Egf (EGF) binds to the Egf receptor (EgfR) and stimulates its protein tyrosine kinase activity to cause autophosphorylation, thus activating EgfR.
- The adaptor protein Grb2 (GRB2) and the guanine nucleotide exchange factor Sos1 (SOS) are recruited to the membrane, binding to EgfR.
- The EgfR complex activates a Ras family GTPase
- Activated Ras activates Raf1, a member of the RAF serine/threonine protein kinase family.
- Raf1 activates the protein kinase Mek (MEK), which then activates Erk (MAPK)
- ...


EgfR-EgfRC
Egf-XOut
001
Shc1-CLc
RalGds-CLc
Ptpn11-CLc
Egf:EgfR-act-EgfRC
Sos1-CLc
Git1-CLc
191
1085
188
197
398
Rala-GDP-CLi
RalGds-EgfRC
Shc1-Yphos-EgfRC
1064-1
Git1-Yphos-EgfRC
Src-CLc
Rala-GTP-EgfRC
Ptpn11-Yphos-EgfRC
207
Sos1-Yphos-EgfRC
Fak2-CLc
Src-act-EgfRC
RasGrp3-CLc
Pi3k-CLc
Gab1-CLc
Mlk3-CLc
186
440
172
116
1063
RasGrp3-Yphos-EgfRC
Pi3k-reloc-EgfRC
Hras-GDP-CLi
Gab1-Yphos-EgfRC
Mlk3-act-EgfRC
Mek1-CLc
529-4
529-6
639
Hras-GTP-EgfRC
Braf-CLc
IqGap1-CLc
Fak2-act-EgfRC
310
Mek1-act-EgfRC
Braf-act-EgfRC
Shoc2-CLc
Erks-CLc
196
Erks-act-EgfRC

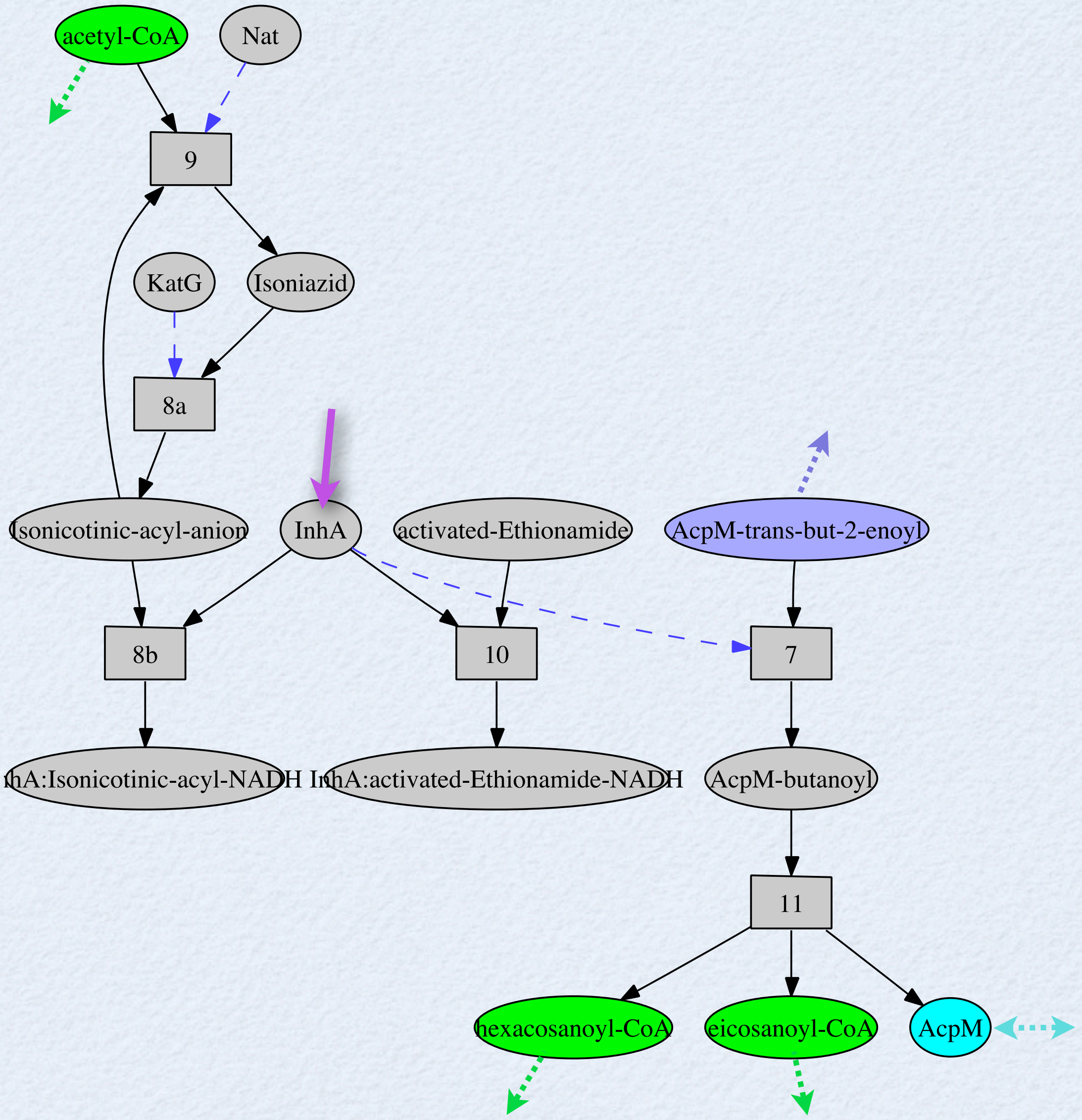
Modeling Metabolic Processes

(work of Malabika Sarker)

Model Action of Drugs

- Problen: Identify candidate drug targets in mycobacteria
- Idea: integrate screening data, molecular structure models, and metabolic models
- Case study
  - curation of PL model of mycolic acid synthesis (including drug action)
  - importing PGDBs into PL

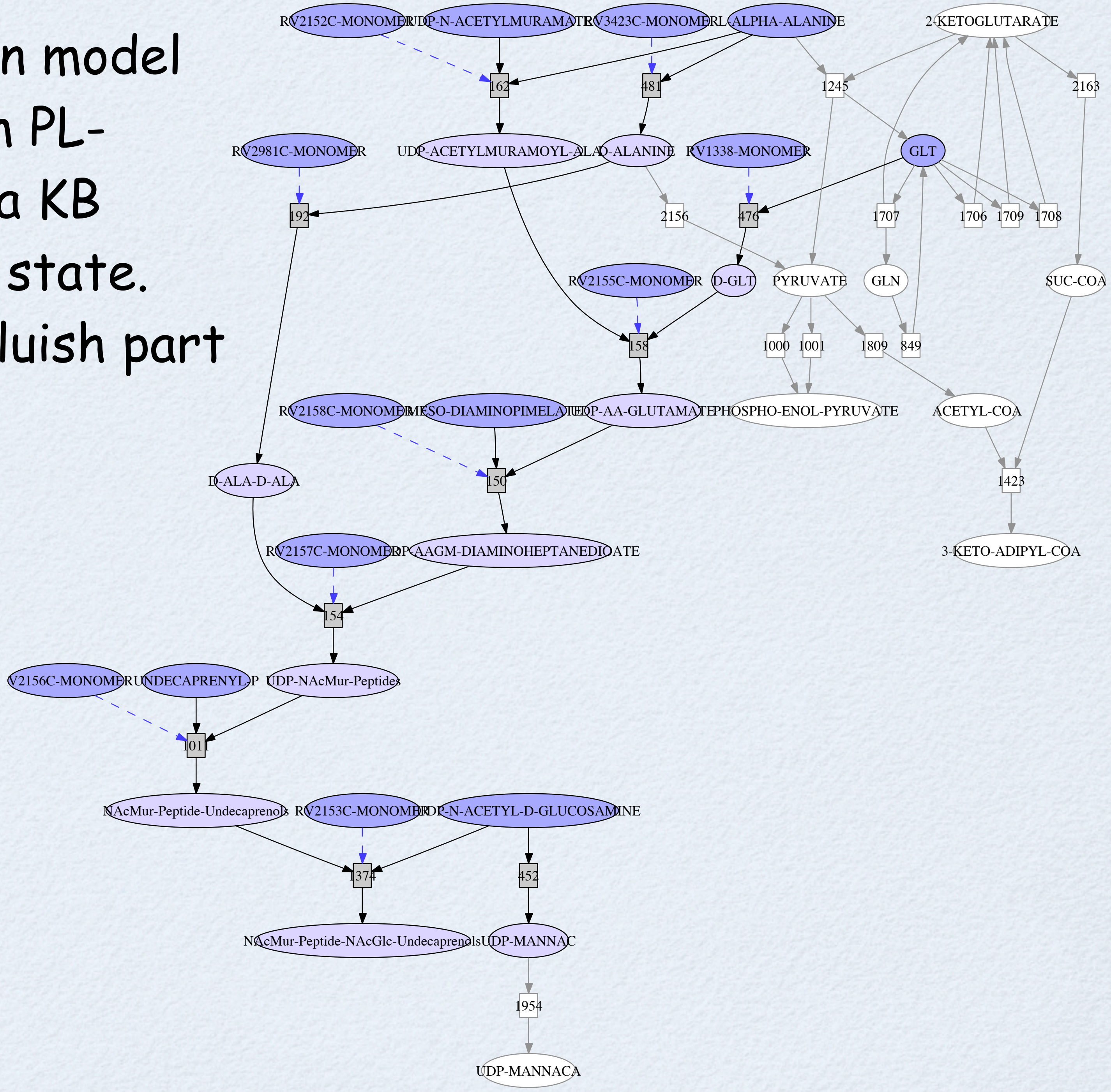
Mycolic acid fragment showing inhibition of InhA



Importing PGDBs into PL

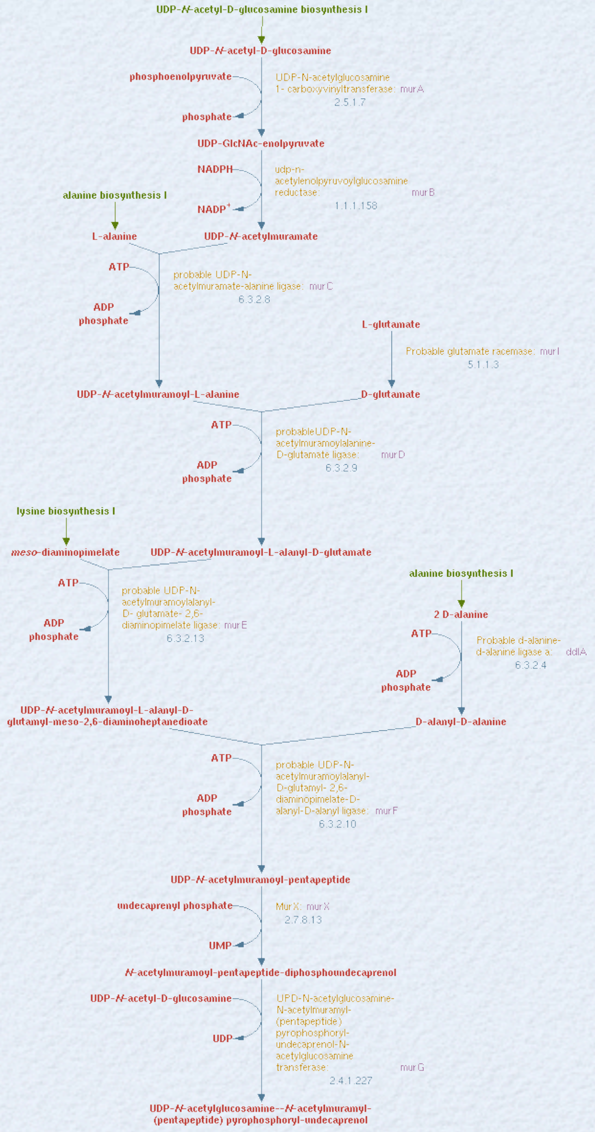
- Map compounds to PL components
- Start with reaction and enzrxn files
- Extract information for PL rules
  - lhs, rhs, enzyme
  - (determine direction)
- Convert to PL syntax
- Apply to M. tuberculosis H37Rv PGDB

Peptidoglycan model derived from PL-mycobacteria KB and starting state. Pathway is bluish part

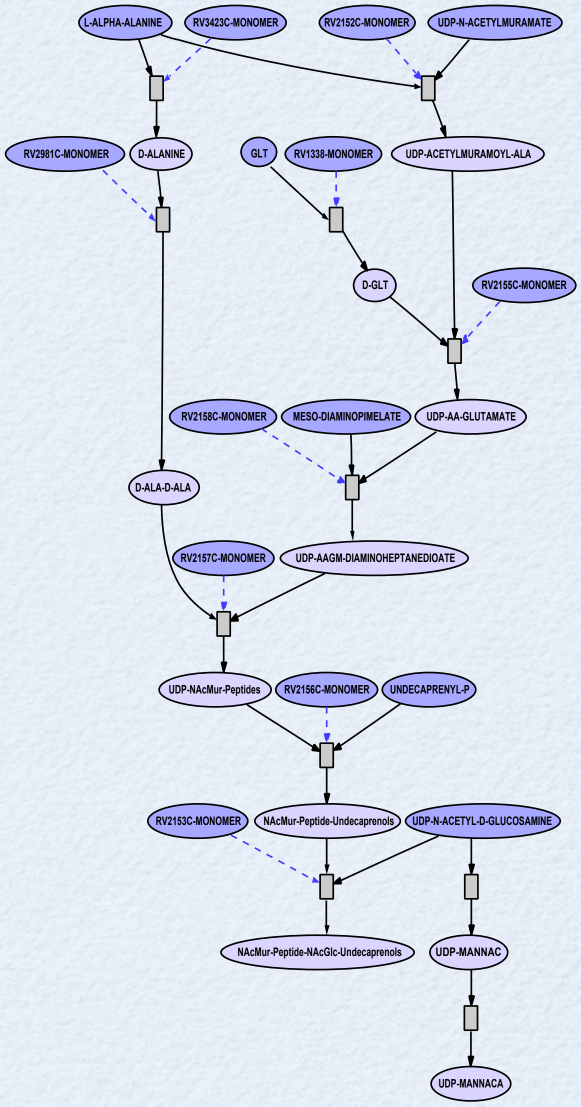


PEPTIDO-GLYCAN PATHWAY

From Biocyc


UDP-N-acetyl-D-glucosamine biosynthesis I
UDP-N-acetyl-D-glucosamine
phosphoenolpyruvate
phosphate
UDP-N-acetylglucosamine 1- carboxyvinyltransferase: murA
2.5.1.7
UDP-GlcNAc-enolpyruvate
NADPH
NADP+
udp-n-acetylenolpyruvoylglucosamine reductase: murB
1.1.1.158
alanine biosynthesis I
L-alanine
UDP-N-acetylmuramate
ATP
ADP
phosphate
probable UDP-N-acetylmuramate-alanine ligase: murC
6.3.2.8
L-glutamate
Probable glutamate racemase: murI
5.1.1.3
UDP-N-acetylmuramoyl-L-alanine
D-glutamate
ATP
ADP
phosphate
probableUDP-N-acetylmuramoylalanine-D-glutamate ligase: murD
6.3.2.9
lysine biosynthesis I
meso-diaminopimelate
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate
ATP
ADP
phosphate
probable UDP-N-acetylmuramoylalanyl-D- glutamate- 2,6-diaminopimelate ligase: murE
6.3.2.13
alanine biosynthesis I
2 D-alanine
ATP
ADP
phosphate
Probable d-alanine-d-alanine ligase a: ddlA
6.3.2.4
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminoheptanedioate
D-alanyl-D-alanine
ATP
ADP
phosphate
probable UDP-N-acetylmuramoylalanyl-D-glutamyl- 2,6-diaminopimelate-D-alanyl-D-alanyl ligase: murF
6.3.2.10
UDP-N-acetylmuramoyl-pentapeptide
undecaprenyl phosphate
UMP
MurX: murX
2.7.8.13
N-acetylmuramoyl-pentapeptide-diphosphoundecaprenol
UDP-N-acetyl-D-glucosamine
UDP
UPD-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol-N-acetylglucosamine transferase: murG
2.4.1.227
UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol

Assembled in PL



Minimal Nutrient Sets

Diet planning for Microbes

The Problem

- Given a model of metabolism for an organism (microbe), determine minimal sets of nutrients that will support growth.
  - Model -- network of metabolic reactions (R)
  - Nutrients -- transportables (T), compound that have transporter reactions
  - Growth -- production of essential compounds (E)
- A subset N of T is a <u>nutrient set</u> if E is R-producible from N
- N is <u>minimal</u> if no proper subset is a nutrient set

A Little Math

- S - stochiometric matrix for R S_{ij} coef of C_i in R_j
- **r** - a vector of relative firing rates, r_j the rate for R_j
- **p** = S **r** -- production $\mathbf{p}_i$ is the production rate of C_i
  - $\mathbf{p}_i = S_{i1}\,\mathbf{r}_1 + ... + S_{ik}\,\mathbf{r}_k$
- Basic constraints
  - $\mathbf{r}_i >= 0$ -- reactions run forward
  - $\mathbf{p}_i > 0$ if C_i in E
  - $\mathbf{p}_i >= 0$ if C_i not in E or N

Simple Example

- R_1: A + B -> C + D, R_2: C + F -> B + E
- E is the essential compound, A, F transportables
- S

	r1	r2
A	-1	0
B	-1	1
C	1	-1
D	1	0
E	0	1
F	-1	0

- Constraints
  - $\mathbf{r}_1, \mathbf{r}_2 >= 0$
  - B: $-\mathbf{r}_1 + \mathbf{r}_2 >= 0$ (> 0)
  - C: $\mathbf{r}_1 - \mathbf{r}_2 >= 0$ (> 0)
  - E: $\mathbf{r}_2 > 0$

- Stable growth: If a non-essential, non-transportable such as B or C is drained away, the system will fail to grow.
- Add constraint that says: if a compound C_j not in E or T is used (a reactant), it must be produced $(\mathbf{p}_j > 0)$.

Problem Simplification

- Impossibility elimination
  - drop reactions that have reactants that can not be produced (or transported)
  - (uses forward collection)
- Uselessness elimination
  - drop useless compounds and reactions whose products are all useless,
  - the useful compounds are found by backwards propagation from E
  - (uses backwards collection)

The Search for Minimal Nutrient Sets

- Define nutset(N) for N a subset of T by

  nutset(N) = true if the constraints for N are satisfiable

  = false owise

- Use a constraint solver to determine if there is a solution
- Find one minimal N: start with N = T and eliminate elements until no mare can be eliminated.
- Finding all minimal Ns requires some cleverness to do it feasibly. Our approach uses a representation of boolean functions called BDDs (binary decision diagrams) to search for extensions of a set of minimal solutions.

Equivalence and Reduced Solutions

- Problem: The system is highly underconstrained leading to a large number of minimal nutrient sets (over 1000).
- Solution: Define two nutrients A,B to be equivalent if whenever A appears in a minimal nutrient set then replacing A by B yields another nutrient set, and conversely.
- Reduced nutrient sets: equivalence class representatives
- Benefit:
    - Small number of solutions
    - Insights into the role of each nutrient

Diet Planning for E. coli

- Model (from EcoCyc version 13.5)
  - 160 transportables
  - 1378 compounds
  - 2251 reactions
  - 36 essentials
- Result
  - 1156 solutions
  - 9 reduced solutions

Ten equivalence classes

- 4 unitary
  - Na+ (?)
  - HPO4 (P)
  - nicotinamide mononucleotide (CNP)
  - 2,3-diketo-L-gulonate (C)
- 3 with two elements
  - sulfate/taurine (S)
  - L-methionine/glutathione (CNS)
  - beta-d-glucose-6-phosphate (CP)
- 1 with nine elements
  - L-valine/NH4+ .. (N)
- 2 very large
  - fumarate/malate ... (C)
  - cytidine/cyanate ... (CN)

Some reduced solutions

- # Reduced solution 7
  - (CCO-PERI-BAC@VAL "L-valine" "C5H11NO2")

    N source -- equivalent to ammonia, nitrite
  - (CCO-PERI-BAC@GLC-6-P "beta-D-glucose-6-phosphate" "C6H11O9P")
  - (CCO-PERI-BAC@SULFATE "sulfate" "O4S")

- # Reduced solution 1
  - (CCO-PERI-BAC@SULFATE "sulfate" "O4S")
  - (CCO-PERI-BAC@NICOTINAMIDE_NUCLEOTIDE "nicotinamide mononucleotide" "C11H14N2O8P")

    CPN source, singleton, too complex to be practical

Mystery Solutions

- # Reduced solution 5 --- mystery -- cytidine ~ cyanate
    - (CCO-PERI-BAC@CYTIDINE "cytidine" "C9H13N3O5")
    - (CCO-PERI-BAC@SULFATE "sulfate" "O4S")
    - (|CCO-PERI-BAC@Pi| "phosphate" "HO4P")
- # Reduced solution 9 --- what is the role of Na+?
    - (CCO-PERI-BAC@NA+ "Na+" "Na")
    - (CCO-PERI-BAC@VAL "L-valine" "C5H11NO2")
    - (CCO-PERI-BAC@SULFATE "sulfate" "O4S")
    - (CCO-PERI-BAC@2-3-DIKETO-L-GULONATE "2,3-diketo-L-gulonate" "C6H7O7")
    - (|CCO-PERI-BAC@Pi| "phosphate" "HO4P")

Lessons Learned

- Analysis is a great way to debug a knowledge base.
  - gaps in network
  - missing participants
  - wrong direction
- Explain unexpected growth conditions
  - Cross checks such as carbon balance
  - Witness information -- sample solution
- Some compounds have no known production pathway
  - Used fudge factors

THATS ALL FOLKS!