SYMBOLIC SYSTEMS BIOLOGY

USING FORMAL LOGICS TO MODEL AND REASON ABOUT BIOLOGICAL SYSTEMS

Carolyn Talcott
SRI International
August 2009
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: $M \models \text{wff}$
- 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

- 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
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 $S$
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
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/
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 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.
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.
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
sort ErbB1L. subsort ErbB1L < Protein. *** ErbB1 Ligand

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

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

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

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

Figure 2b.
The same rule in Maude representation.

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  }.
Rule instances relevant to Hras activation

Parallel paths

Cross talk

Synchronization

Conflict

Hras activated
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)
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
- $H$: a 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.
FULL MODEL OF EGF STIMULATION
(by Merrill Knapp)
THE ERBB NETWORK (CARTOON FORM)

PL EGF MODEL
Events that could occur in response to EGF
Curated by Merrill Knapp
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)
- ...
MODELING METABOLIC PROCESSES
(work of Malabika Sarker)
Problem: 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

acetyl-CoA → Nat

KatG → Isoniazid

Isonicotinic-acyl-anion → InhA

InhA:activated-Ethionamide → AcpM-trans-but-2-enoyl

AcpM-trans-but-2-enoyl → AcpM-butanoyl

AcpM-butanoyl → eicosanoyl-CoA

hexacosanoyl-CoA → AcpM
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.
From Biocyc

Assembled in PL
MINIMAL NUTRIENT SETS

Diet planning for Microbes
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 nutrient set if E is R-producible from N

N is minimal 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  \( p_i \) is the production rate of \( C_i \)
  - \( p_i = S_{i1} r_1 + \ldots + S_{ik} r_k \)
- Basic constraints
  - \( r_i \geq 0 \) -- reactions run forward
  - \( p_i > 0 \) if \( C_i \) in E
  - \( p_i \geq 0 \) if \( C_i \) not in E or N
**SIMPLE EXAMPLE**

- $R_1$: $A + B \rightarrow C + D$, \quad $R_2$: $C + F \rightarrow B + E$
- $E$ is the essential compound, $A$, $F$ transportables

<table>
<thead>
<tr>
<th></th>
<th>r1</th>
<th>r2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>$B$</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>$C$</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$D$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$E$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$F$</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Constraints**
  - $r_1, r_2 \geq 0$
  - $B$: $-r_1 + r_2 \geq 0$ (> 0)
  - $C$: $r_1 - r_2 \geq 0$ (> 0)
  - $E$: $r_2 > 0$

- **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 ($p_j > 0$).
**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)
Define nutset(N) for N a subset of T by

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

= false otherwise

Use a constraint solver to determine if there is a solution

Find one minimal N: start with N = T and eliminate elements until no more 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.
**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)
# 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
THAT'S ALL FOLKS!