Simulation and verification for computational modelling of signalling pathways

Marta Kwiatkowska, Gethin Norman, David Parker, Oksana Tymchyshyn, John Heath and Eamonn Gaffney

UNIVERSITY OF BIRMINGHAM

Winter Simulation 2006

Overview

- The context
  - Challenges of systems biology
  - Role of computational methods
- Modelling frameworks and formalisms
  - Continuous deterministic vs discrete stochastic approach
  - Process calculi for biology
  - Simulation vs verification
- First results
  - Case study: aspects of FGF signalling [CMSB'06]
  - Related projects
- Future challenges

Biological processes...

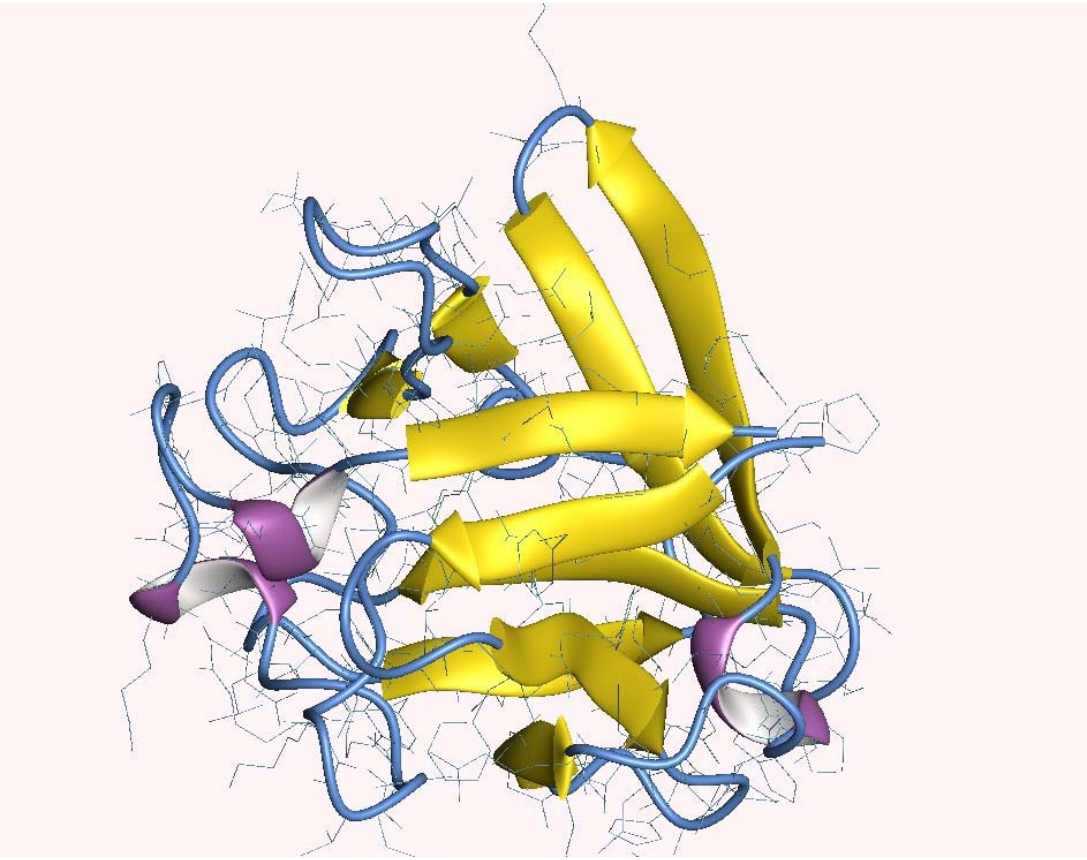
- Networks of subsystems
  - Organisms, cells, molecules, …
- Interaction
  - Governed by rules
  - Causes transformations
- Evolution
  - Continuous and discrete dynamics
- Mobility
  - Motion in space and time, re-configurability, …
- Stochastic behaviour
  - Unpredictability, noise, …

Not unlike computers, networks and the Internet…

and yet new challenges for systems biology

Modelling signalling pathways

- Focus on
  - networks of molecules
  - interaction
  - continuous & discrete dynamics
- Rather than
  - geometry
  - structure
  - sequence
- Qualitative and quantitative



Google images: Human FGF, http://160.114.99.91/astrojan/prot1t.htm

Modelling frameworks

- Assume wish to model mixture of molecules
  - N different molecular species, interact through reactions
  - Fixed volume V (spatially uniform), constant pressure and temperature

- Continuous deterministic approach
  - Approximate the number of molecules in V at time t by a continuous function
  - Obtain ODEs (ordinary differential equations)
  - Not for individual runs, but average
  - Assumptions must be respected, ensure large numbers of molecules

- Discrete stochastic approach
  - Discrete system evolution, via discrete events for reactions
  - Obtain discrete-state stochastic process

Discrete stochastic approach

- Work with states as vectors **x** of molecule counts for each species
  - Probability P(**x**,t) that at time t there will be x_A of species

$$\frac{\partial P(\mathbf{x},t)}{\partial t} = \sum_{i=1}^{2} \left(a_i(\mathbf{x}-\mathbf{v})P(\mathbf{x}-\mathbf{v},t) - a_i(\mathbf{x})P(\mathbf{x},t)\right)$$

- If constant state-dependent rates, obtain continuous time Markov chain [well studied]
- Admits
  - discrete event simulation
  - numerical solution (probabilistic model checking)
- Realistic for a single time course evolution
- Usually population based, but individual-based also possible (if cannot assume well stirred)

Typical modelling/analysis approach…

- Use biochemical reactions as basis
  - Write in SBML (Systems Biology Mark Up Language)
  - Share models on WWW, access a range of solvers
    - Analytical solutions rarely feasible
    - ODE and stochastic models generated automatically
    - e.g. Gillespie, for population-based discrete stochastic models
- Choose a formalism, model directly
  - Discrete models: graphical notations (Petri nets), textual (process calculi, rewrite systems), or their stochastic extension…
  - Continuous: ODEs, PDEs
- Choose an analysis method
  - Simulation, if focus is on time course trajectories
  - Otherwise, more powerful methods needed

Why process calculus?

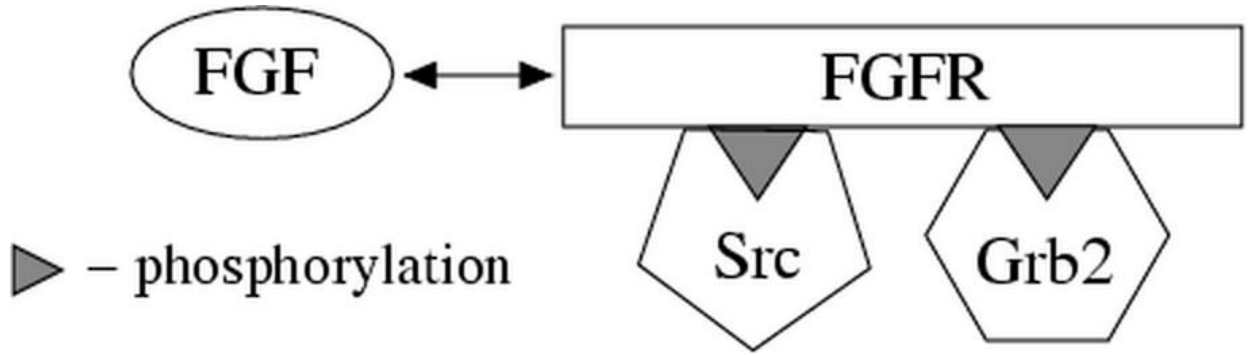
- Language for modelling networks of objects
  - Compact description for networks of interacting objects
  - Molecule-centred, but an be used at all levels, molecular, cellular, tissue
  - Ease of textual manipulation: add/remove/modify reaction

- Calculus and algebra for processes
  - Compositional models
  - Induce state transition systems
  - Stochastic variants generate discrete stochastic models
  - Proof rules for reasoning about process behaviour

- Support powerful analysis methods
  - Simulation, to obtain individual trajectories
  - Verification, reasoning about all possible behaviours

Fragment of FGF Pathway

- Fragment of Fibroblast Growth Factor (FGF) pathway
  - regulator of skeletal development, e.g. number of digits



- Biological challenges
  - compartments, mobility, unknown function of molecules
  - expensive experimental scenarios

- Aim to develop ODE and discrete stochastic models
  - ODE: use Cellarator & Mathematica
  - Discrete: simulation (BioSPI, SPiM), verification (PRISM)

The Reactions

1: FGF binds/releases FGFR

FGF + FGFR → FGFR:FGF	$k_1 = 5e+8\ M^{-1}s^{-1}$
FGF + FGFR ← FGFR:FGF	$k_2 = 0.002\ s^{-1}$

2: Phosphorylation of FGFR (whilst FGFR:FGF)

FGFR1 → FGFR1P	$k_3 = 0.1\ s^{-1}$
FGFR2 → FGFR2P	$k_4 = 0.1\ s^{-1}$

3: Dephosphorylation of FGFR

FGFR1P → FGFR1	$k_5 = 0.1 s^{-1}$
FGFR2P → FGFR2	$k_6 = 0.1 s^{-1}$

4: Effectors bind phosphorylated FGFR

SRC + FGFR1P → SRC:FGFR	$k_7 = 1e+6\ M^{-1}s^{-1}$
SRC + FGFR1P ← SRC:FGFR	$k_8 = 0.02\ s^{-1}$
GRB2 + FGFR2P → GRB2:FGFR	$k_9 = 1e+6\ M^{-1}s^{-1}$
GRB2 + FGFR2P ← GRB2:FGFR	$k_{10} = 0.02\ s^{-1}$

5: Relocation of FGFR (whilst SRC:FGFR)

SRC:FGFR → relocFGFR	$k_{11} = 1.1e-3\ s^{-1}$

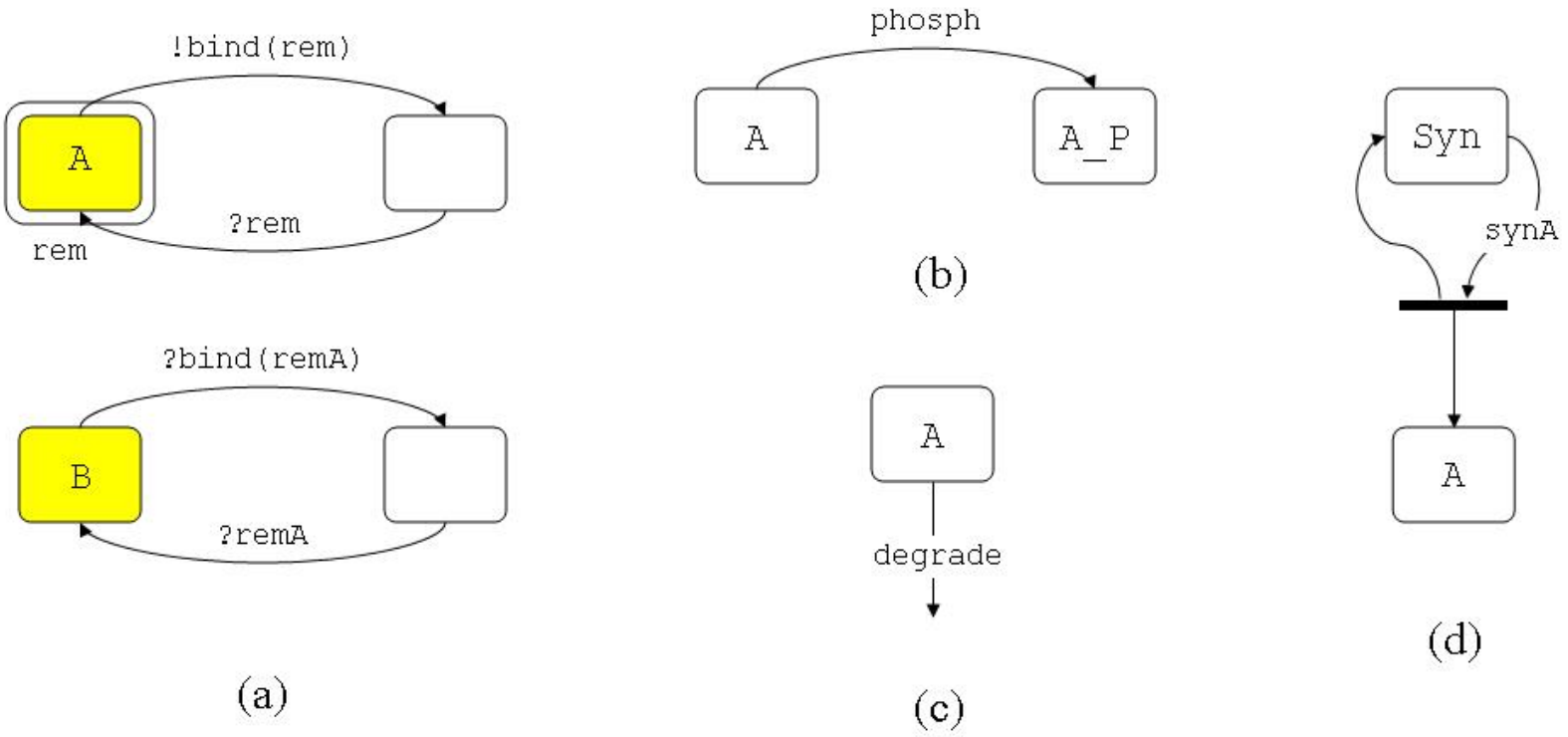
SBML Code Fragment

```
<listOfSpecies>
    <species id="FGFR_Ph1" initialConcentration="0" ... />
    <species id="SRC" initialConcentration="N" ... /> ...
</listOfSpecies>
<reaction id="Reaction1" reversible="false">
    <listOfReactants>
        <speciesReference species="FGFR_Ph1" />
        <speciesReference species="SRC" /> ...
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="FGFR_SRC" /> ...
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <apply> <times/>
                <ci>k7</ci> <ci>FGFR_Ph1</ci> <ci>SRC</ci>
            </apply>
        </math>
    </kineticLaw>
</reaction>
```

Stochastic π-calculus code fragment

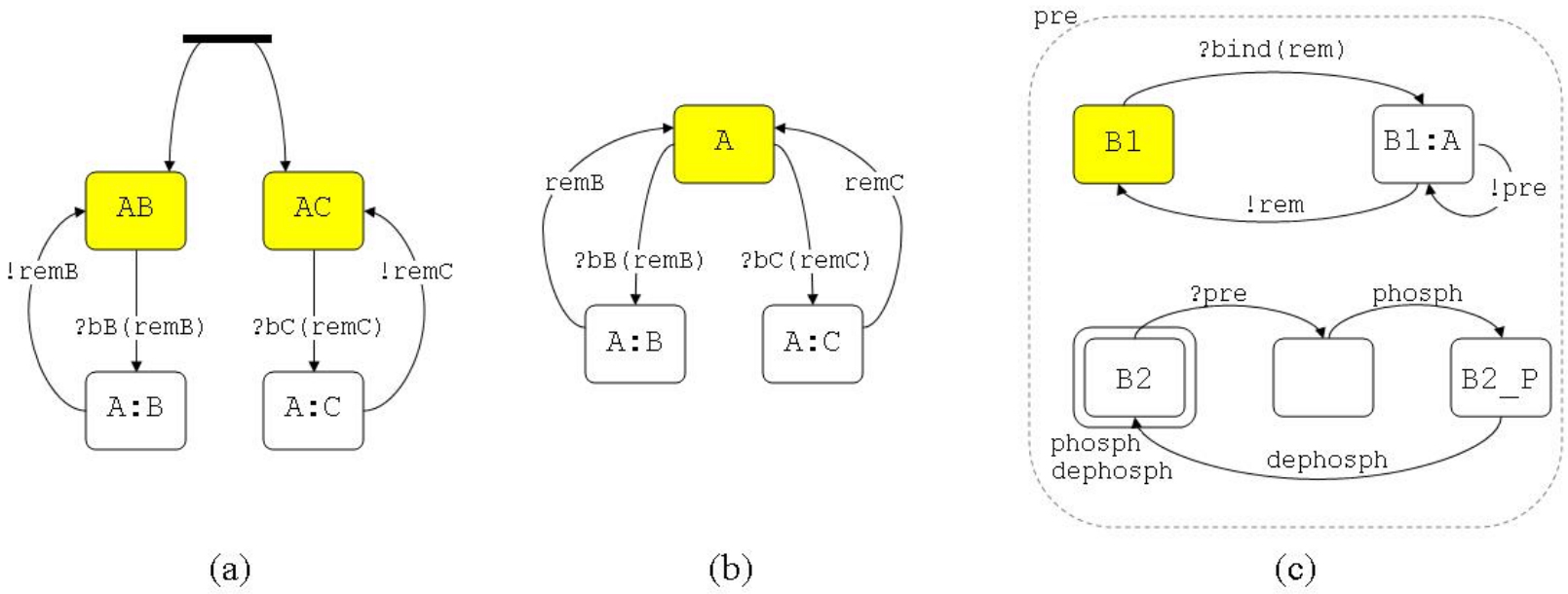
```
FGFR ::= FGFR_FGF_0 | FGFR_Ph1_0 | ...

FGFR_FGF_0 ::= reloc1?[], true ; % relocation
    bind_fgf!{ rel_fgf, reloc4 }, FGFR_FGF_1. % binding FGF

FGFR_FGF_1 ::=  rel_fgf?[] , FGFR_FGF_0 ; % releasing FGF
                ph1?[] , FGFR_FGF_1 ; % phosphorylation
                reloc1?[] , reloc4 ! [] , true ; % relocation ...

FGFR_Ph1_0 ::= ph1![] , FGFR_Ph1_1 . % phosphorylation
FGFR_Ph1_1 ::= dph1![] , FGFR_Ph1_1; % dephosphorylation
    bind_src!{rel_src1, rel_src2 } , FGFR_SRC. % binding Src

FGFR_SRC ::= rel_src1?[], FGFR_Ph1_1 ; % releasing Src
    dph1![], rel_src2![], FGFR_Ph1_0; % dephos (& release Src)
    reloc![], reloc1![], reloc2![] , true .  % relocation
```

Stochastic π-model visually



- A graphical π-calculus representation of complexation (a), phosphorylation (b), degradation (c) and synthesis (d) reactions
- See SPiM of Phillips & Cardelli [BIOCONCUR'05].

Stochastic π-model visually ctd



- Representation of parallel (a), competitive (b), and contextual (c) reactions.

PRISM language fragment

```
module fgfr
  fgfr_fgf : [0..1] init 0; // FGF bound
  fgfr_ph1 : [0..1] init 0; // state receptor 1 phosphorylated
  fgfr_src : [0..1] init 0; // Src bound
  reloc_fgfr : [0..1] init 0; // FGFR relocated ...

  // binding and release of FGF
  [bind_fgf] reloc_fgfr=0 ∧ fgfr_fgf=0 → k1 : (fgfr_fgf'=1);
  [rel_fgf]  reloc_fgfr=0 ∧ fgfr_fgf=1 → k2 : (fgfr_fgf'=0); ...
  // phosphorylation/dephosphorylation (release SRC under dephosphorylation)
  [] reloc_fgfr=0 ∧ fgfr_fgf=1 ∧ fgfr_ph1= 0 → k3 : (fgfr_ph1'=1);
  [] reloc_fgfr=0 ∧ fgfr_ph1=1 ∧ fgfr_src= 0 → k5 : (fgfr_ph1'=0);
  [rel_src] reloc_fgfr=0∧fgfr_ph1= 1∧fgfr_src=1 → k5 :(fgfr_ph1'=0)∧(fgfr_src'=0); ...
   // binding and release of Src
  [bind_src] reloc_fgfr=0 ∧ fgfr_ph1=1 ∧ fgfr_src=0 → k7 : (fgfr_src'=1);
  [rel_src]  reloc_fgfr=0 ∧ fgfr_src=1 → k8 : (fgfr_src1'=0); ...
  // relocation (caused by Src)
  [] reloc_fgfr=0 ∧ fgfr_src=1 → k11 : (reloc_fgfr'=1);
endmodule
```

Fragment of ODEs

$$
\begin{aligned}
Fgfr_{0,0}'(t) &= - bind\_fgf \cdot Fgf(t) \cdot Fgfr_{0,0}(t) + rel\_fgf \cdot Fgfr\_Fgf_{0,0}(t) \\
&\quad + dph1 \cdot Fgfr_{1,0}(t) + dph1 \cdot Fgfr_{2,0}(t) \ldots \\
Fgfr_{1,0}'(t) &= - bind\_fgf \cdot Fgf(t) \cdot Fgfr_{1,0}(t) + rel\_fgf \cdot Fgfr\_Fgf_{1,0}(t) \\
&\quad - dph1 \cdot Fgfr_{1,0}(t) - bind\_src \cdot Src(t) \cdot Fgfr_{1,0}(t) \\
&\quad + rel\_src \cdot Fgfr_{2,0}(t) \ldots \\
Fgfr_{0,1}'(t) &= - bind\_fgf \cdot Fgf(t) \cdot Fgfr_{0,1}(t) + rel\_fgf \cdot Fgfr\_Fgf_{0,1}(t) \\
&\quad + dph1 \cdot Fgfr_{1,1}(t) + dph1 \cdot Fgfr_{2,1}(t) \ldots \\
Fgfr_{1,1}'(t) &= - bind\_fgf \cdot Fgf(t) \cdot Fgfr_{1,1}(t) + rel\_fgf \cdot Fgfr\_Fgf_{1,1}(t) \\
&\quad - dph1 \cdot Fgfr_{1,1}(t) - bind\_src \cdot Src(t) \cdot Fgfr_{1,1}(t) \ldots \\
Fgfr_{2,0}'(t) &= - bind\_fgf \cdot Fgf(t) \cdot Fgfr_{2,0}(t) + bind\_src \cdot Src(t) \cdot Fgfr_{1,0}(t) \\
&\quad + rel\_fgf \cdot Fgfr\_Fgf_{2,0}(t) - rel\_src \cdot Fgfr_{2,0}(t) \\
&\quad - reloc \cdot Fgfr\_Fgf_{2,0}(t) - dph1 \cdot Fgfr\_Fgf_{2,0}(t) \\
Fgfr_{0,2}'(t) &= - bind\_fgf \cdot Fgf(t) \cdot Fgfr_{0,2}(t) + rel\_fgf \cdot Fgfr\_Fgf_{0,2}(t) \\
&\quad + dph1 \cdot Fgfr_{2,2}(t) + dph1 \cdot Fgfr_{1,2}(t) \ldots
\end{aligned}
$$

In the terms $Fgfr_{res1,res2}$ and $Fgfr\_Fgf_{res1,res2}$, res1 an res2 correspond to two independent residues of the protein: unphosphorylated, phosphorylated and bound to Src or Grb2.

Contrasting the approaches…

- SBML enables
  - Sharing of models
  - Automatic generation of models

- But
  - SBML supports a single discrete state per molecule
  - If parallel molecular state changes, e.g. formation of complexes, observe exponential growth in number of ODEs
  - Discrete approaches more amenable to parallel changes but can result in exponential state explosion

- Also
  - Averages can be misleading if numbers of molecules small
  - Simulation approximate, cannot deal with causality…

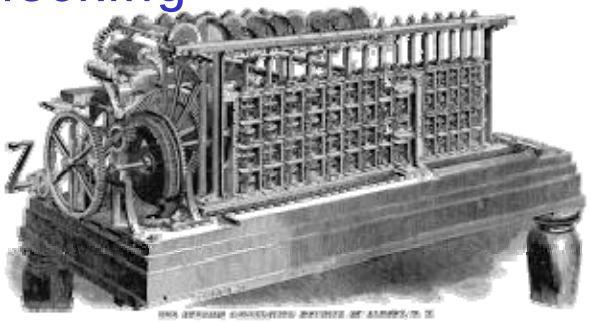
Our focus: verification

- Rather than follow conventional approaches…
  - ODEs poor for modelling discrete phenomena
  - Population models can mask the dynamics of individual cells
- … employ
  - Stochastic process calculi and models
  - Formal verification, especially probabilistic model checking
- … to complement conventional analysis by
  - Enable inspection of models against detailed queries: temporal, causality, probability, expectation
  - Calculate quantitative predictions for a range of possible scenarios and parameters

Formal verification

- Amounts to a mathematical proof that the system satisfies the specification
  - Theorem proving, manual or automated
  - Reasoning over all executions
  - User intervention necessary

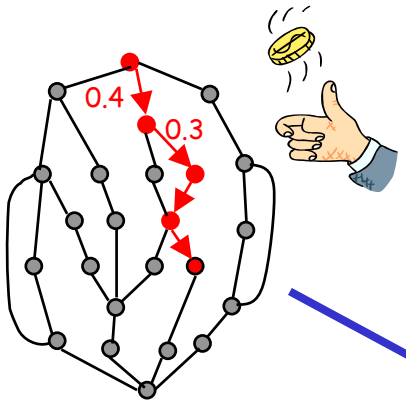
- Automatic verification, via model checking
  - For all possible executions
  - Mechanical, no user intervention
  - Finite state models only



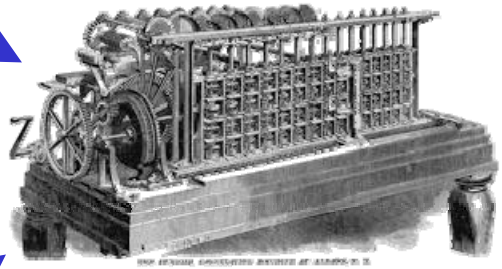
- Unlike simulation, verification is exhaustive

Probabilistic model checking...

in a nutshell



Probabilistic model

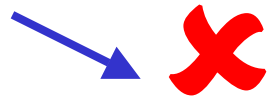


Probabilistic
Model Checker

send → $\mathcal{P}_{\geq 0.9}(\Diamond \text{deliver})$

Probabilistic temporal logic specification

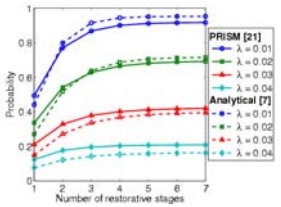
or



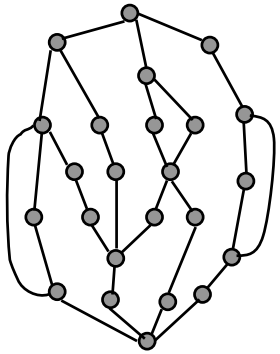
or

The probability

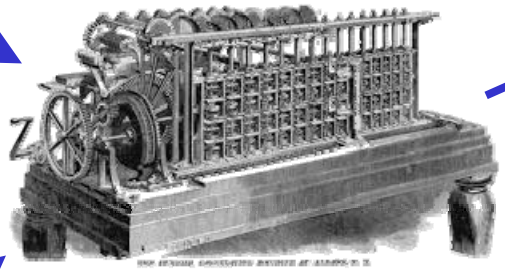
State 5: 0.6789
State 6: 0.9789
State 7: 1.0
...
State 12: 0
State 13: 0.1245



Verification via model checking



Finite-state model



Model Checker

or

✓

✗

Error trace

Line 5: …
Line 21: …
Line 15: …
…
Line 27: …
Line 45: …

$init \rightarrow \Diamond response$

Temporal logic specification

Exhaustive model checking

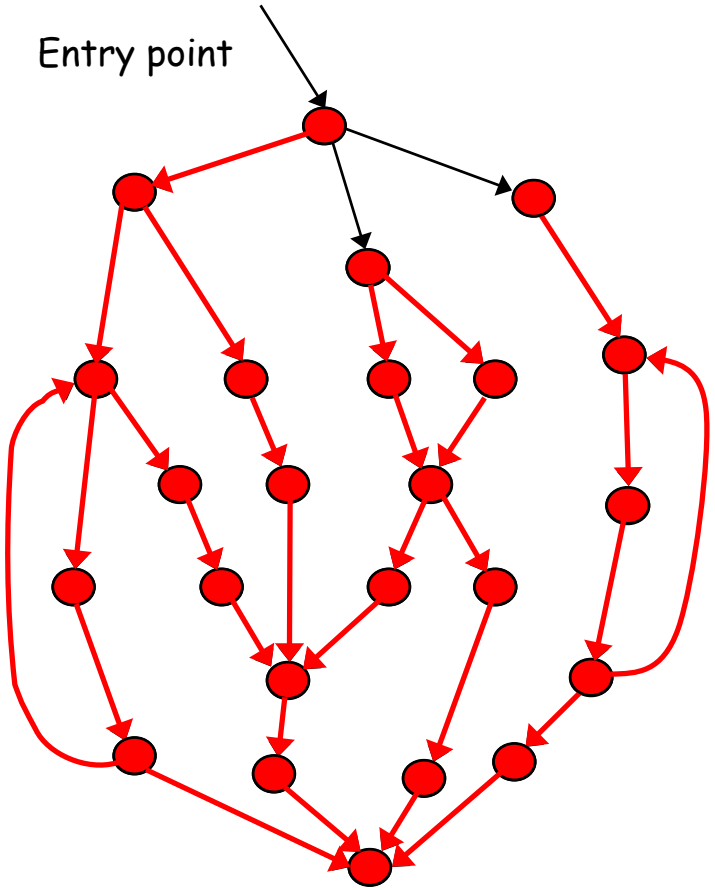
Is it possible to reach a danger state from the initial state(s)?

Does every execution from the initial state lead to danger state?

NB considers all executions.

Model finite-state, hence termination assured.

Extends to probabilistic systems and quantitative analysis.



Simulation vs Verification

- Verification exact and detailed
  - Wide range of quantitative properties
  - Compute for range of parameters: quantitative trends
  - Can definitively establish causal relationships
  - Able to identify best/worst case scenarios
  - but suffers from state explosion problems

- Must consider all possible executions – often not feasible mechanically!
  - [NB a challenging problem in computer science…]

- Simulation approximate
  - but OK for averages over large numbers of runs
  - Generally greater scalability

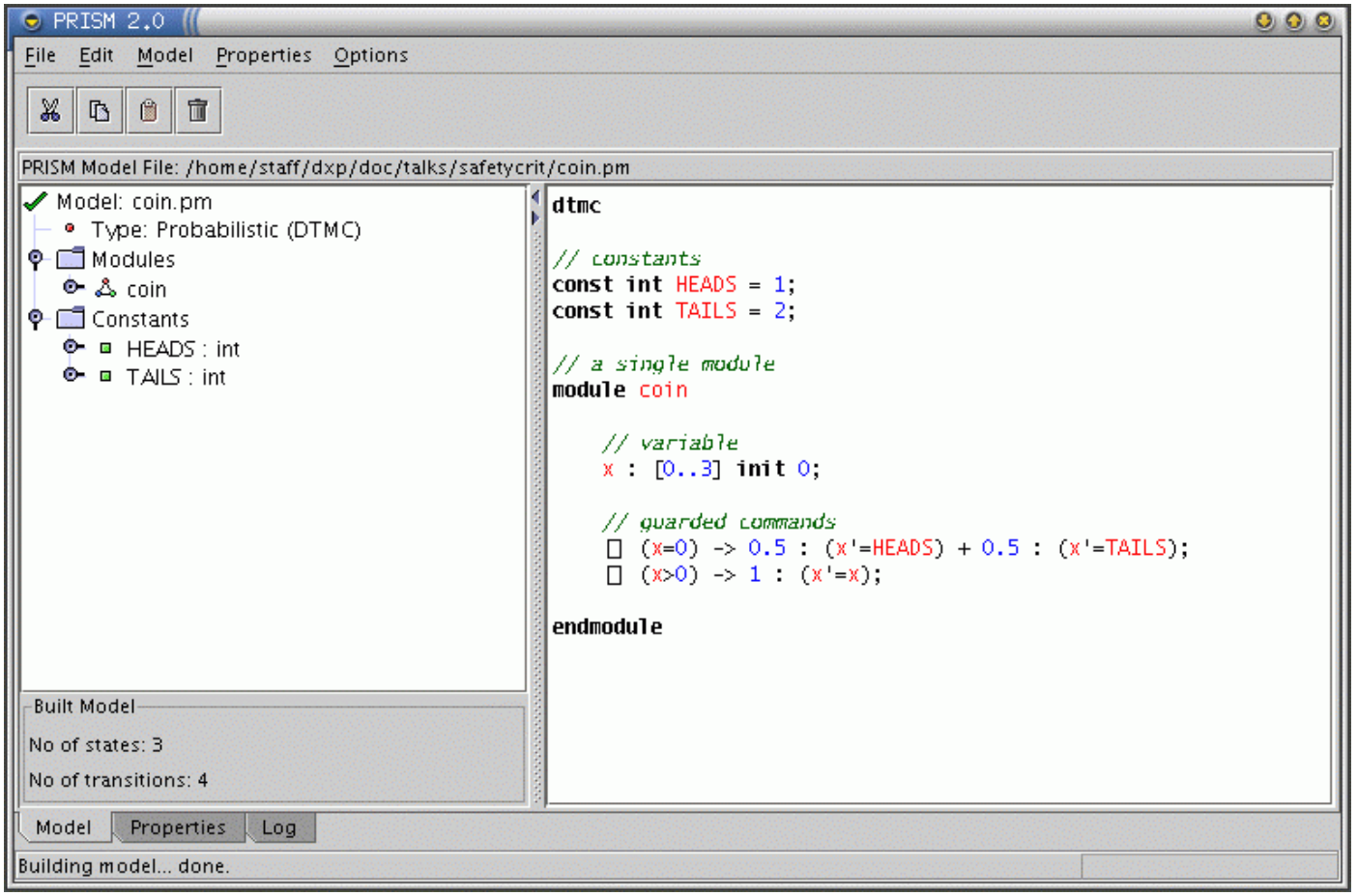
The PRISM model checker: Overview

- Simple, discrete stochastic modelling language
  - networks formed from interacting modules
  - interactions are associated with state-dependent rates
  - similar expressive power to variants of stochastic pi-calculus

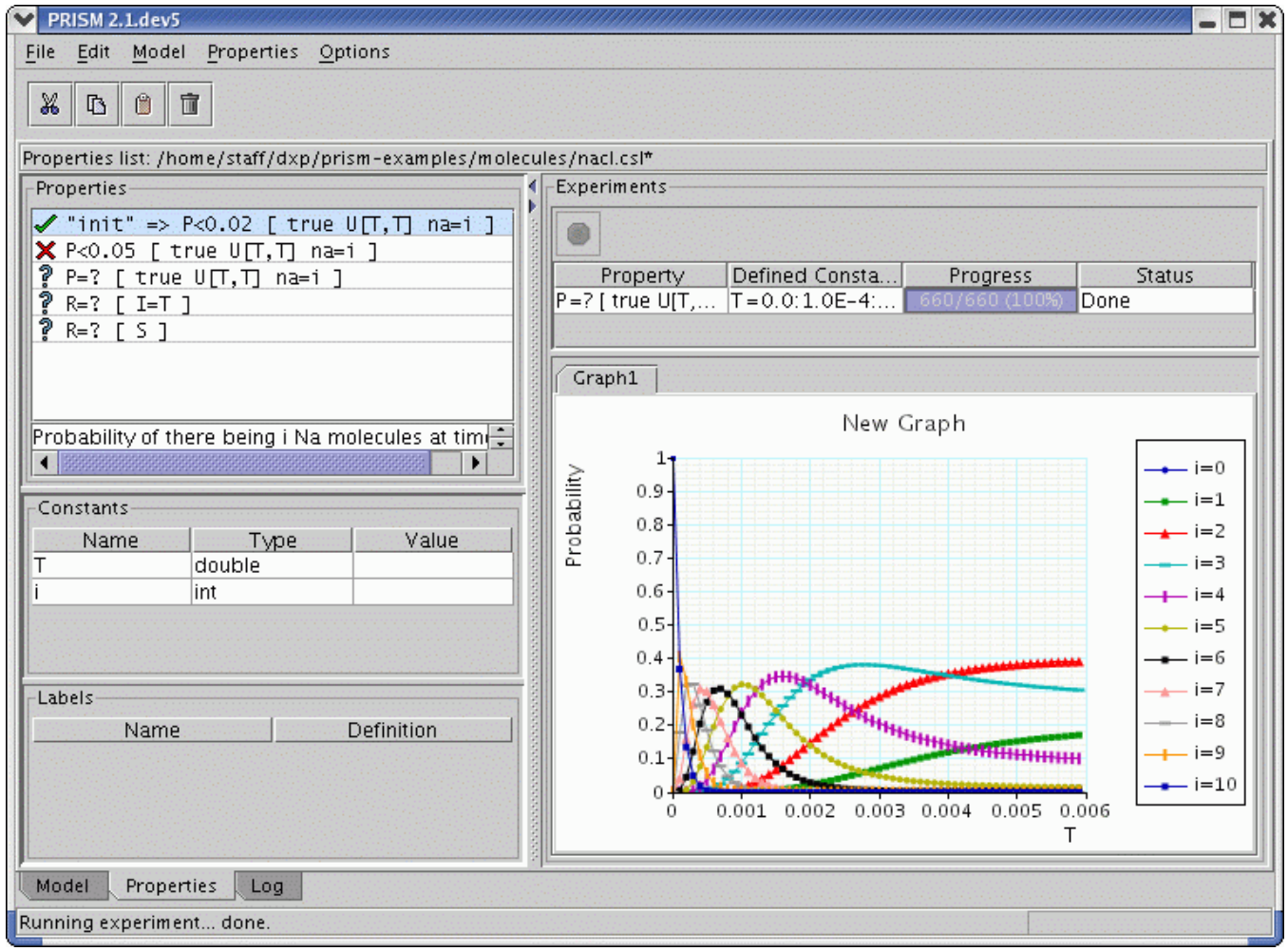
- Specifications given in temporal logic:
  - what is the probability that concentration is less than min at time 10? $\mathbf{P}_{=?}$ [true $\mathbf{U}^{[10,10]}$ c<min]
  - what is the probability the concentration reaches min?
    $\mathbf{P}_{=?}$ [true $\mathbf{U}$ c≥min]
  - in the long run, what is the probability that the concentration remains stable between min and max
    $\mathbf{S}_{=?}$ [(c≥min)∧(c≤max)]

- Numerous case studies and errors detected in computer network protocols

Screenshot: Text editor



Screenshot: Graphs



Simple PRISM Example

1. **A+B** $\leftrightarrow$ **A:B** (complexation/decomplexation rates r_1/r_2)
2. **A** $\rightarrow$ (degradation rate r_3)

```
module A
  a : [0..1] init 1
  [bind] a=1 → r1 : (a'=0);
  [rel] a=0 → r1 : (a'=1);
  [] a=1 → r1 : (a'=0);
endmodule
```

```
module B
  b : [0..1] init 1
  [bind] b=1 → (b'=0);
  [rel] b=0 → (b'=1);
endmodule
```

```
module AB
  ab : [0..1] init 0
  [bind] ab=0 → (ab'=1);
  [rel] ab=1 → (ab'=0);
endmodule
```

Reward structure 1:
time A is unbound

```
rewards
  a=1 : 1;
endrewards
```

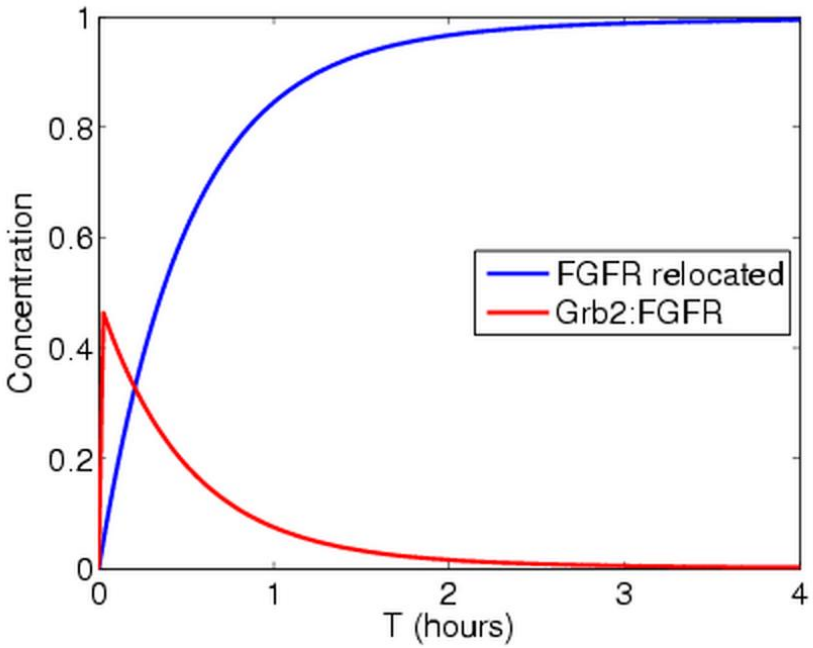
Reward structure 2:
binding of A & B

```
rewards
  [bind] true : 1;
endrewards
```

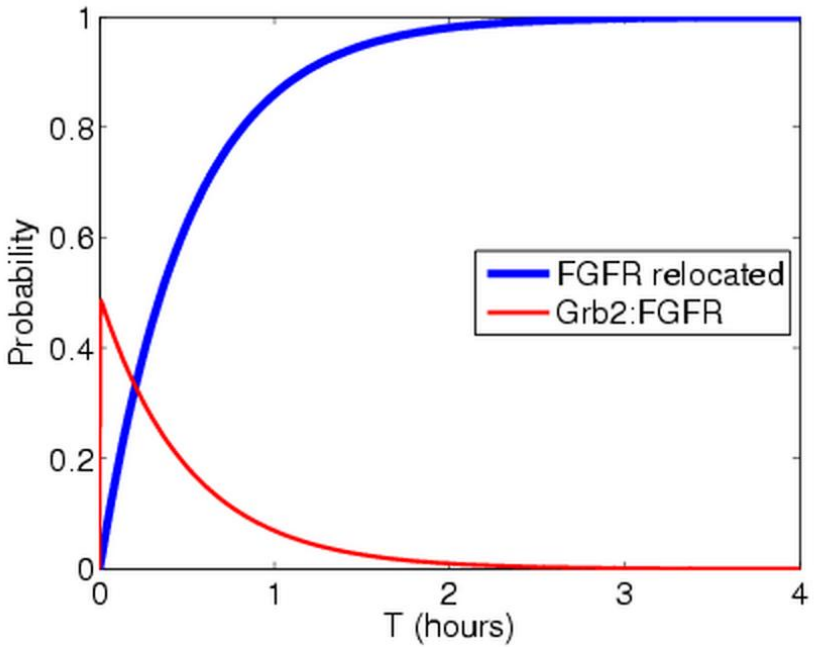
Results

Concentration/quantity of two forms of FGFR over time

ODEs

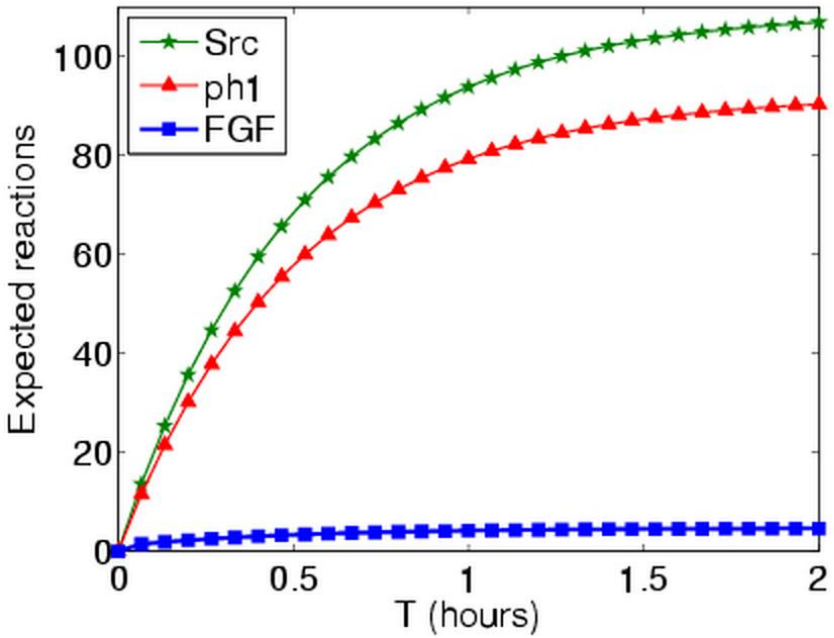


BioSPRI (30 runs)

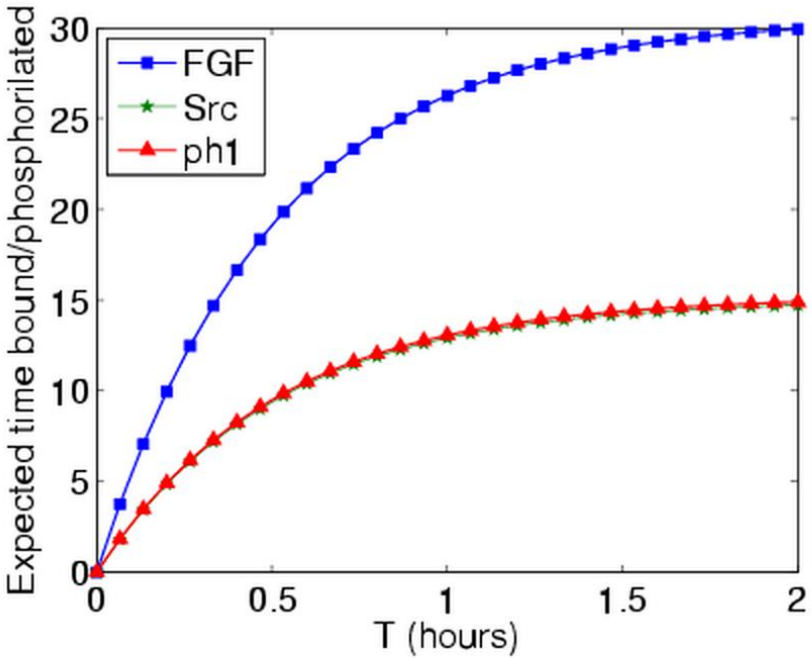


More PRISM results

Expected number of reactions by time T



Expected time complex spends bound by time T



A variant of the FGF fragment

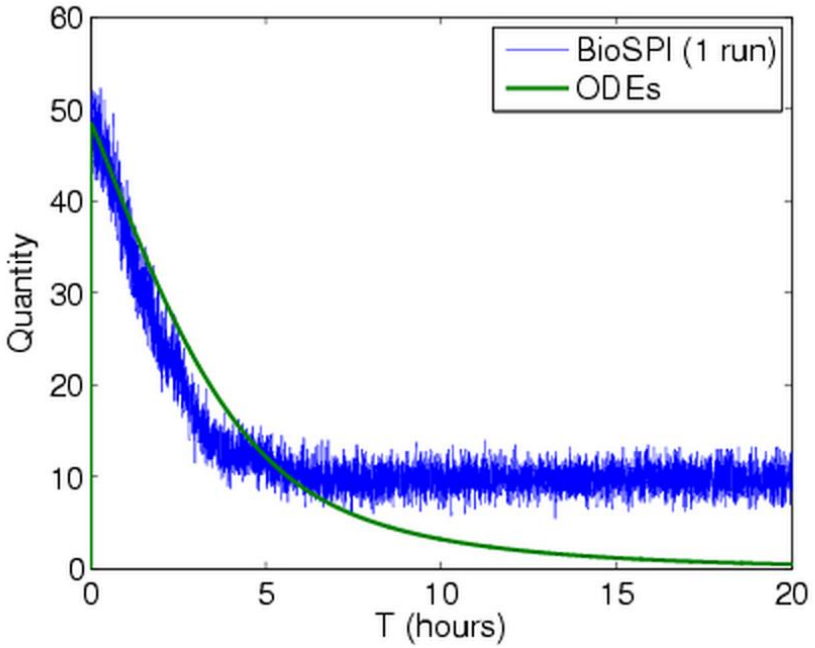
Src positively regulates FGFR signalling by recruiting non-activated FGFR to the membrane, add reaction:

$$FGFR{:}Src \rightarrow FGFR{:}Src + FGFR + Src$$

Change initial amount of Src from 100 to 10 molecules, and similarly for ODEs

Difference between ODE and BioSPI caused by stochastic approach more accurate when number of molecules small

i.e. Src cannot be totally degraded in ODE



PRISM model of full FGF pathway

- Biological Model
  - 12 elements
  - 14 phosphorylation sites
  - 14 sets of reaction rules (38 rules)

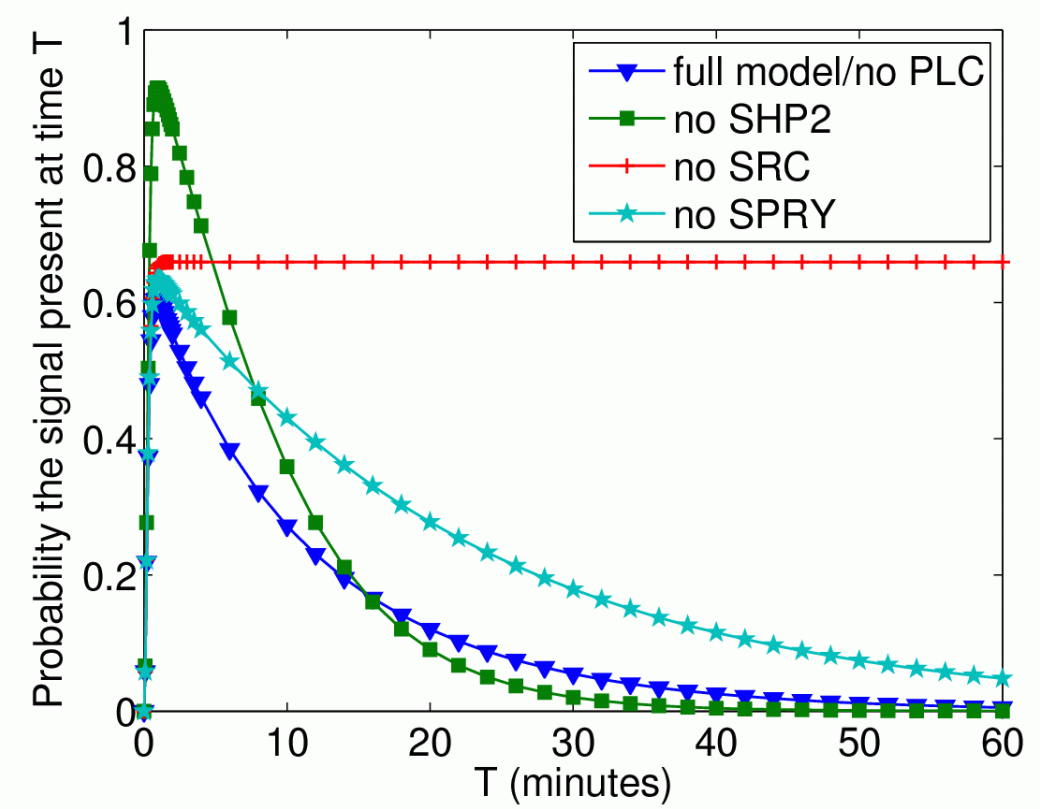
- PRISM model
  - Suppose one element of each type
  - 10 modules and 26 variables
  - 80,616 states and 560,520 transitions

- For full details see CSMB'06
  - Relatively small state space
  - However, highly complex: large number of interactions
  - ODE model > 300 equations

Model checking results

Probability signal present at time T

$$\mathbf{P}_{=?}[\text{true } \mathbf{U}^{[T,T]}\ a_{Grb2}]$$



No SRC: no relocation of FRS2, and hence the signal can remain active

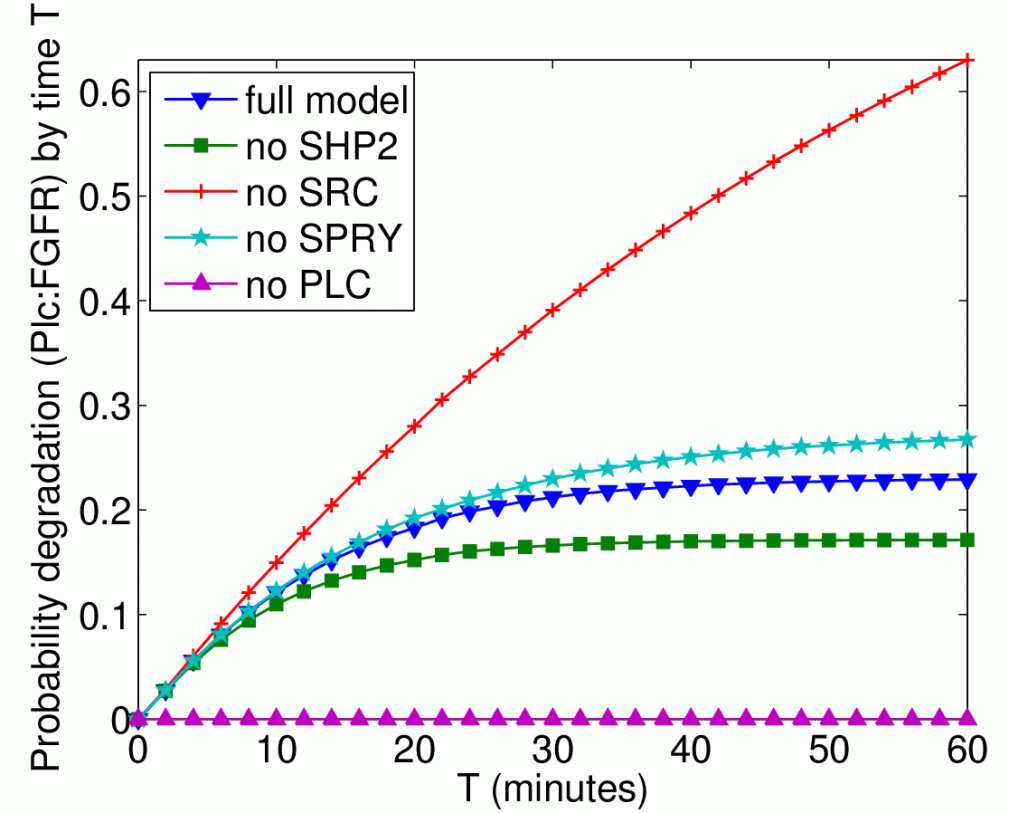
No SHP2: main cause of FRS2 dephosphorylation lost increasing the chance that:

- Grb2 bound to FRS
  faster increase in signal
- SRC bound to FRS2
  faster degradation in signal

Model checking results

Probability PLC causes degradation/relocation by T

$$\mathbf{P}_{=?}[\neg(a_{src} \vee a_{spry} \vee a_{plc})\ \mathbf{U}^{\leq T}\ a_{plc}]$$



No PLC: PLC cannot cause degradation

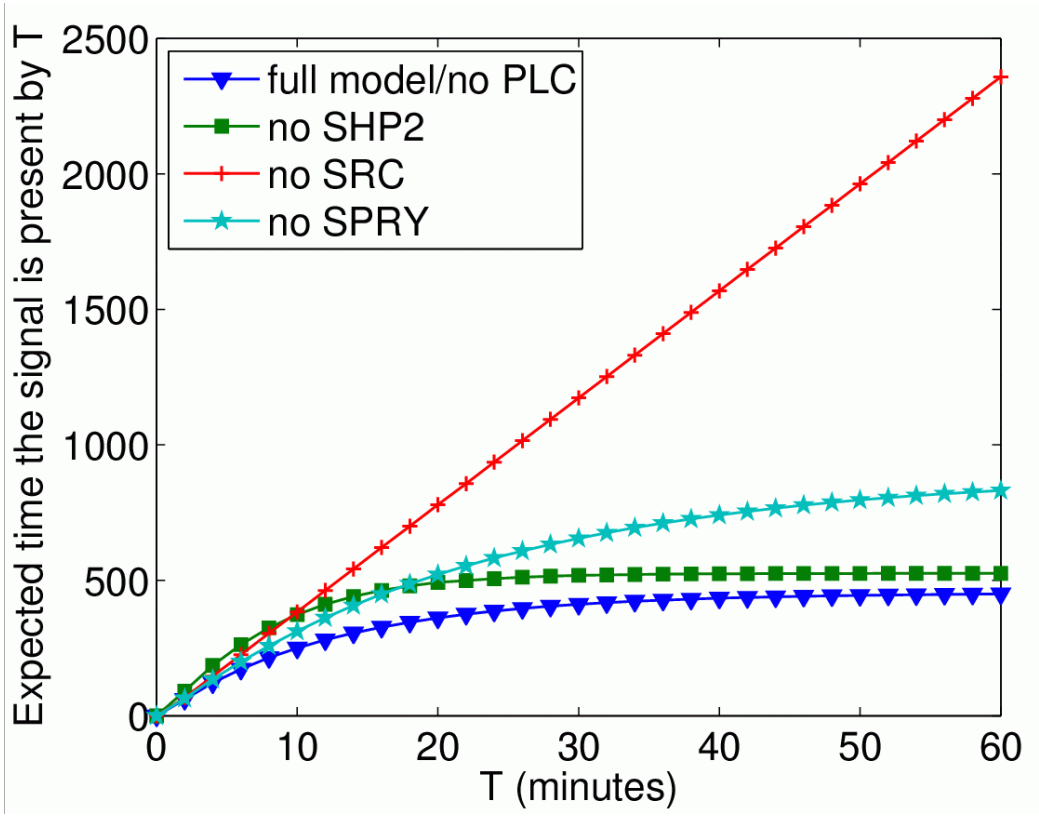
No SRC: FRS2 not relocated, more chance of degradation by PLC

No SHP2: greater chance SRC bound to FRS2, increasing the possibility of FRS2 causing relocation

Model checking results

Expected time GRB2 bound to FRS2 within time T

$\mathbf{R}_{=?}[\mathbf{C}^{\leq T}]$ (assign reward 1 to states where Grb2:FRS2)



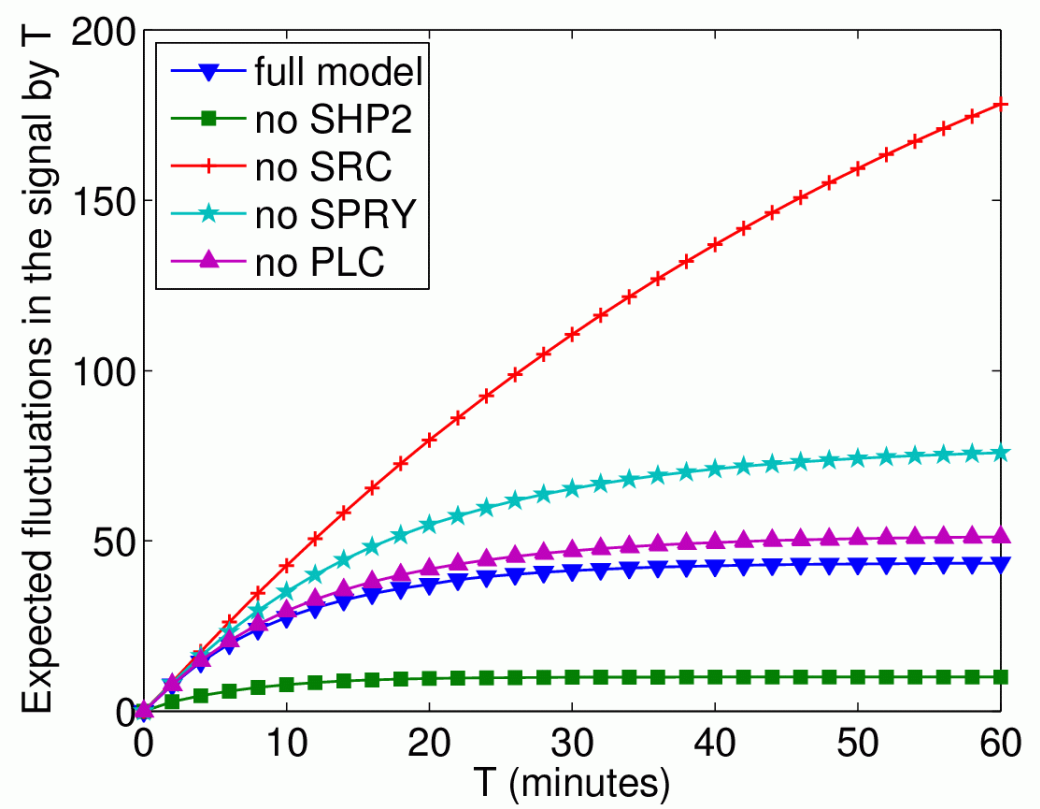
No SRC: no relocation of FRS2 and greater chance FRS2 remains active for longer, hence GRB2 and FRS2 spend more time bound

SPRY: no degradation of FRS2, again GRB2 and FRS2 spend more time bound (but SPRY has smaller influence than SRC)

Model checking results

Expected number of times GRB2 & FRS2 bind by T

$\mathbf{R}_{=?}[\mathbf{C}^{\leq T}]$ (assign reward 1 to transitions binding Grb2 & FRS2)



Cases when SRC and SPRY removed: increased chance that FRS2 remains active, and hence GRB2 and FRS2 can bind more often

No SHP2: decrease in the chance that GRB2:FRS2 unbind, therefore the chance that GRB2 and FRS2 are in a position to (re)bind decreases

Challenges for future

- Scalability
- Exploiting structure
  - abstraction/refinement
  - model reductions (symmetry, etc)
  - decomposition…
- Compositional reasoning
- Parametric probabilistic verification
- Approximation methods
- Connection to SBML
- Inter-translation between different methodologies
- Model extraction from data
- More real pathway case studies

Acknowledgements

- Joint (inter-disciplinary) work with
  - John Heath (Biosciences)
  - Oksana Tymchyshyn, Gethin Norman, Dave Parker (Computer Science)
  - Eamonn Gaffney (Mathematics)

- Funding
  - Microsoft Research Cambridge project on Predictive modelling of signalling pathways via probabilistic model checking with PRISM
  - EPSRC, especially Integrative Biology project, and CRUK

- More information
  - For publications, case studies, software, etc, see
    - www.cs.bham.ac.uk/~dxp/prism/
    - www.cs.bham.ac.uk/~oxt/fgfmap.html

Related projects

- Predictive analysis of signalling pathways with PRISM, Microsoft Research Cambridge
  - SBML standard for stochastic pi-calculus
  - SBML-enabled probabilistic model checking with PRISM
  - FGF as a case study

- Modelling cell cross-talk with statecharts
  - Use state-based models with real microarray data
  - Based on Harel's model of T-cell maturation in Thymus

- CancerGrid
  - Grid infrastructure for randomised cancer trials
  - Cambridge, Oxford, Birmingham, UCL and Belfast
  - MRC funded

Integrative Biology

Exploiting e-Science to combat fatal diseases

- Large e-Science project
  - Two components, computational modelling of human heart and modelling of cancer
  - Oxford, Nottingham, UCL, Leeds, Birmingham, Sheffield

- Cancer modelling
  - 3 institutions, 2 RFs & 10 PhDs
  - Multiscale
  - Tissue, cellular, subcellular, drug response
  - Colorectal cancer
  - Crypt formation
  - Cell migration in crypt, role of stem cells