Statistical Model Checking for Biological Systems

Paolo Zuliani

Department of Computer Science

Carnegie Mellon University

Joint work with
Edmund Clarke, James Faeder, Haijun Gong, Qinsi Wang

Verification of Rule-based Models

- Temporal properties over the stochastic evolution of the model
- Example: "does *p53* reach 4,000 within 20 minutes, with probability at least 0.99?"
- In our formalism, we write:

$$P_{\geq 0.99}\ (\mathbf{F}^{20}\ (p53 \geq 4{,}000))$$

- For a property Φ as above and a fixed $0<\vartheta<1$, we ask whether

$$P_{\geq\vartheta}\ (\Phi) \qquad \text{or} \qquad P_{<\vartheta}\ (\Phi)$$

Statistical Model Checking

Key idea

(Haakan Younes, 2001)

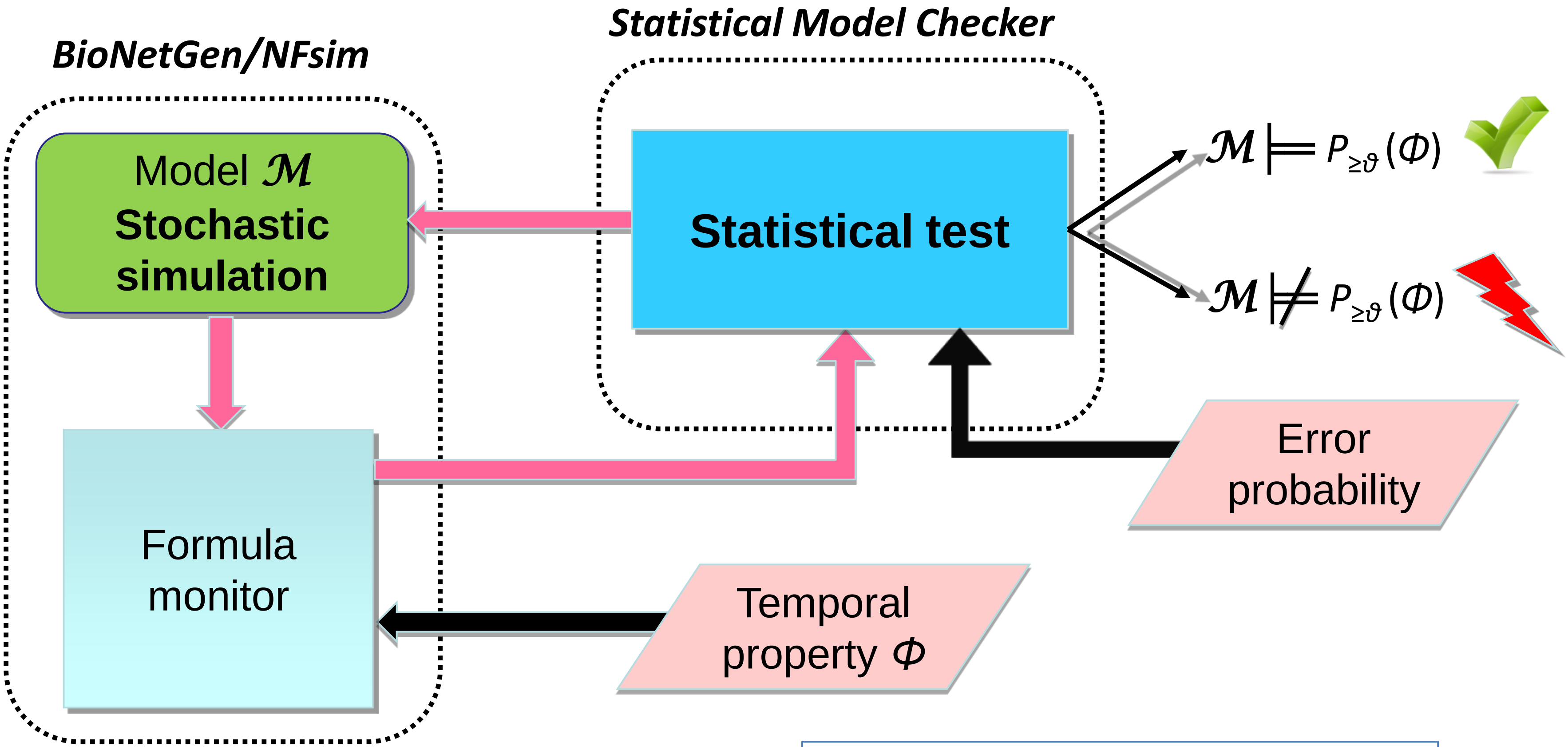
- Suppose system behavior w.r.t. a (fixed) property *Φ* can be modeled by a Bernoulli of parameter *p*:
  - System satisfies *Φ* with (unknown) probability *p*
- Questions: $P_{\geq\vartheta}$ (*Φ*)? (for a fixed $0<\vartheta<1$)
- Draw a sample of system simulations and use:
  - Statistical hypothesis testing: Null vs. Alternative hypothesis

    $$H_0 : \mathcal{M} \models P_{\geqslant\theta}(\phi) \qquad H_1 : \mathcal{M} \models P_{<\theta}(\phi)$$

  - Statistical estimation: returns "*p* in (a,b)" (and compare a with ϑ)

Our Approach

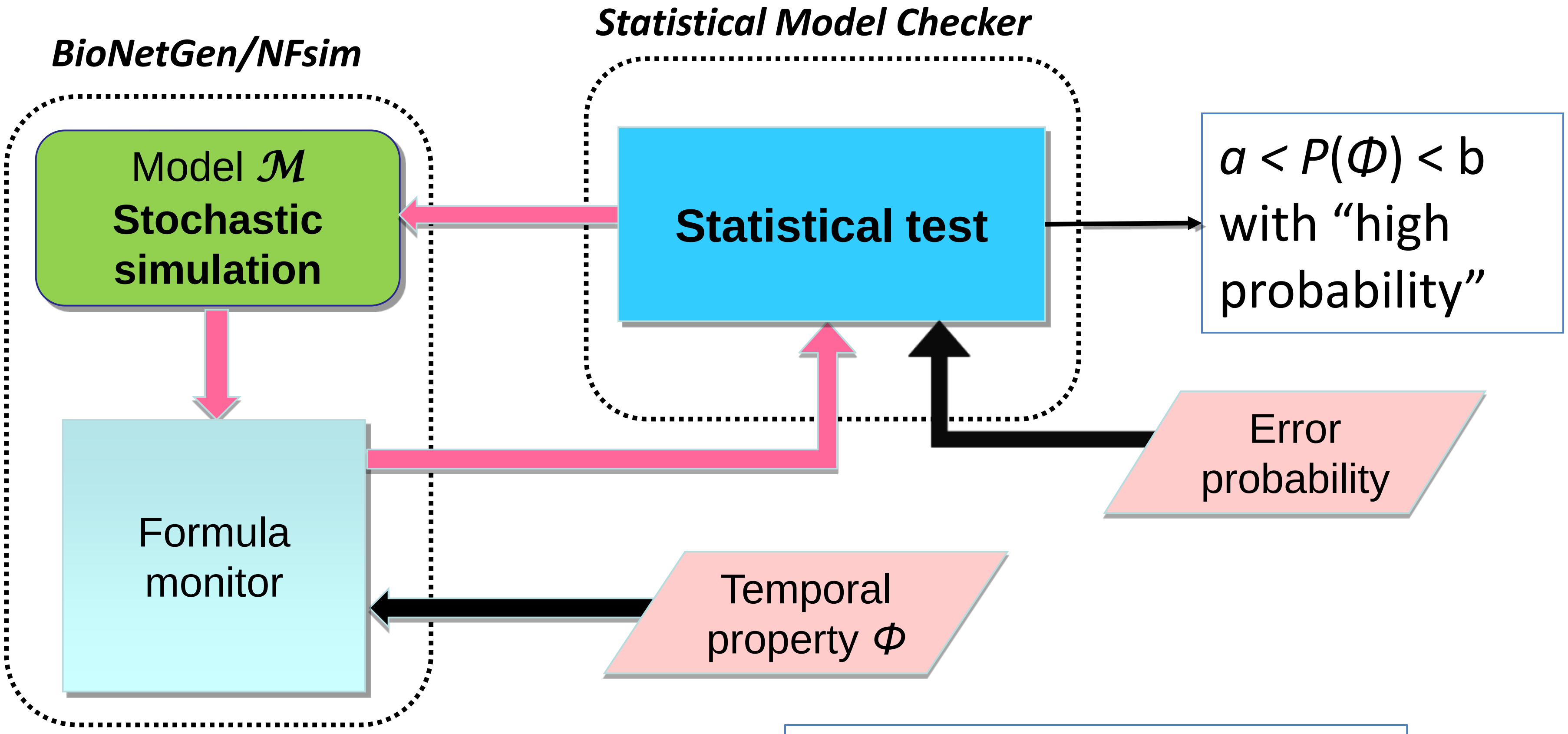
Statistical Model Checking: $\boxed{\mathcal{M} \models P_{\geq\vartheta}(\Phi)?}$



Zuliani, Platzer, Clarke. *HSCC 2010*.

Our Approach

Statistical Model Checking: what is $P(\Phi)$?



Zuliani, Platzer, Clarke. *HSCC 2010*.

Motivation

- State Space Exploration infeasible for large systems
  - Symbolic MC with OBDDs scales to 10^{300} states
  - Scalability depends on the structure of the system
  - Probabilistic symbolic MC (eg PRISM) scales to 10^{10} states
- Pros: simulation is feasible for **many more** systems
  - Often easier to simulate a complex system than to build the transition relation for it
- Pros: easier to parallelize
- Cons: answers may be **wrong**
  - But error probability can be bounded
- Cons: simulation is **incomplete** (continuous state spaces)

Bayesian Statistical Model Checking

- Sequential sampling
- Performs Hypothesis Testing (and Estimation)
- Based on Bayes Theorem
- Application to BioNetGen

Bounded Linear Temporal Logic

- Bounded Linear Temporal Logic (BLTL): A version of LTL with time bounds on temporal operators.

- Let $\sigma = (s_0, t_0), (s_1, t_1), \ldots$ be an execution of the model
  - along states $s_0, s_1, \ldots$
  - the system stays in state s_i *for time* t_i
  - divergence of time: $\Sigma_i\, t_i$ diverges (i.e., non-zeno)

- σ^i: Execution trace starting at state *i*

- A model for simulation traces (e.g. BioNetGen)

Semantics of BLTL

The semantics of BLTL for a trace σ^k:

- $\sigma^k \models ap$ iff atomic proposition *ap* true in state s_k
- $\sigma^k \models \Phi_1 \vee \Phi_2$ iff $\sigma^k \models \Phi_1$ or $\sigma^k \models \Phi_2$
- $\sigma^k \models \neg\Phi$ iff $\sigma^k \models \Phi$ does not hold
- $\sigma^k \models \Phi_1 \,\mathcal{U}^t\, \Phi_2$ iff there exists natural *i* such that
  1) $\sigma^{k+i} \models \Phi_2$
  2) $\Sigma_{j<i}\, t_{k+j} \leq t$
  3) for each $0 \leq j < i$, $\sigma^{k+j} \models \Phi_1$

  "within time t, Φ_2 will be true and Φ_1 will hold until then"

- In particular, $F^t\,\Phi = true \,\mathcal{U}^t\, \Phi$, $G^t\,\Phi = \neg F^t \neg\Phi$

Bayesian Statistics

Three ingredients:

1. Prior distribution
   - Models our initial (a priori) uncertainty/belief about parameters (what is P(H)?)

2. Likelihood function
   - Describes the distribution of data, given a specific parameter range: P($data$ | H)

3. Bayes Theorem
   - Posterior probability - Revises uncertainty upon experimental data

$$P(H \mid data) = [P(data \mid H) \cdot P(H)] \,/\, P(data)$$

Sequential Bayesian Statistical MC - I

- Model Checking $H_0 : \mathcal{M} \models P_{\geqslant\theta}(\phi)$ $H_1 : \mathcal{M} \models P_{<\theta}(\phi)$
- Suppose $\mathcal{M}$ satisfies ϕ with (unknown) probability *p*
  - *p* is given by a random variable *U* (defined on [0,1]) with density *g*
  - *g* represents our prior belief that $\mathcal{M}$ satisfies ϕ
- Generate independent and identically distributed (iid) sample traces.
- x_i: the i^{th} sample trace σ satisfies ϕ
  - $x_i = 1$ iff $\sigma_i \models \phi$
  - $x_i = 0$ iff $\sigma_i \not\models \phi$
- Then, x_i will be a Bernoulli trial with conditional density (likelihood function)

$$f(x_i | u) = u^{x_i}(1 - u)^{1-x_i}$$

Sequential Bayesian Statistical MC - II

- $X = (x_1, \ldots, x_n)$ a sample of Bernoulli random variables
- Prior probabilities *P(H₀), P(H₁)* strictly positive, sum to 1
- Posterior probability (Bayes Theorem [1763])

$$\boxed{P(H_0|X) = \frac{P(X|H_0)P(H_0)}{P(X)}}$$

for *P(X)* > 0

- Ratio of Posterior Probabilities:

$$\boxed{\frac{P(H_0|X)}{P(H_1|X)} = \underbrace{\frac{P(X|H_0)}{P(X|H_1)}}_{\text{Bayes Factor}} \cdot \frac{P(H_0)}{P(H_1)}}$$

Sequential Bayesian Statistical MC - III

```
Require: Property $P_{\geq\vartheta}(\Phi)$, Threshold $T \geq 1$, Prior density $g$
$n := 0$            {number of traces drawn so far}
$s := 0$            {number of traces satisfying $\Phi$ so far}
repeat
    $\sigma$ := draw a sample trace of the system (iid)
    $n := n + 1$
    if $\sigma \models \Phi$ then
        $s := s + 1$
    endif
    $\mathcal{B} := BayesFactor(n, s, g)$
until ($\mathcal{B} > T \vee \mathcal{B} < 1/T$)
if ($\mathcal{B} > T$) then
    return $H_0$ accepted
else
    return $H_0$ rejected
endif
```

Correctness

Theorem (Termination)

The Sequential Bayesian Statistical Hypothesis Testing algorithm terminates with probability one.

Theorem (Error bounds)

When the Bayesian algorithm using threshold T stops, the following holds:

$$\boxed{\begin{array}{l}\text{Prob (“accept } H_0\text{” } \mid H_1) \leq 1/T \\ \text{Prob (“reject } H_0\text{” } \mid H_0) \leq 1/T\end{array}}$$

Note: bounds independent from the prior distribution.

Bayesian Interval Estimation - I

- Estimating the (unknown) probability *p* that “system $\models \Phi$”
- Recall: system is modeled as a Bernoulli of parameter *p*
- *Bayes’ Theorem* (for conditional iid Bernoulli samples)

$$f(u \mid x_1, \ldots, x_n) = \frac{f(x_1 \mid u) \cdots f(x_n \mid u) g(u)}{\int_0^1 f(x_1 \mid v) \cdots f(x_n \mid v) g(v)\ dv}$$

- We thus have the posterior distribution
- So we can use the mean of the posterior to estimate *p*
  - mean is a posterior Bayes estimator for *p* (it minimizes the integrated risk over the parameter space, under a quadratic loss)

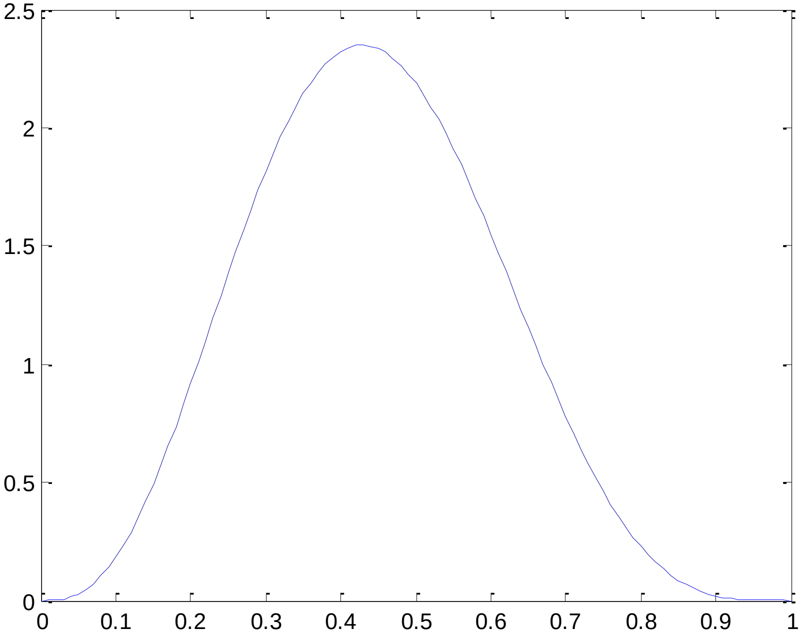
Bayesian Interval Estimation - II

- Bayesian interval for *p*: integrate the posterior
- Fix a coverage ½ < *c* < 1. Any interval (t_0, t_1) such that

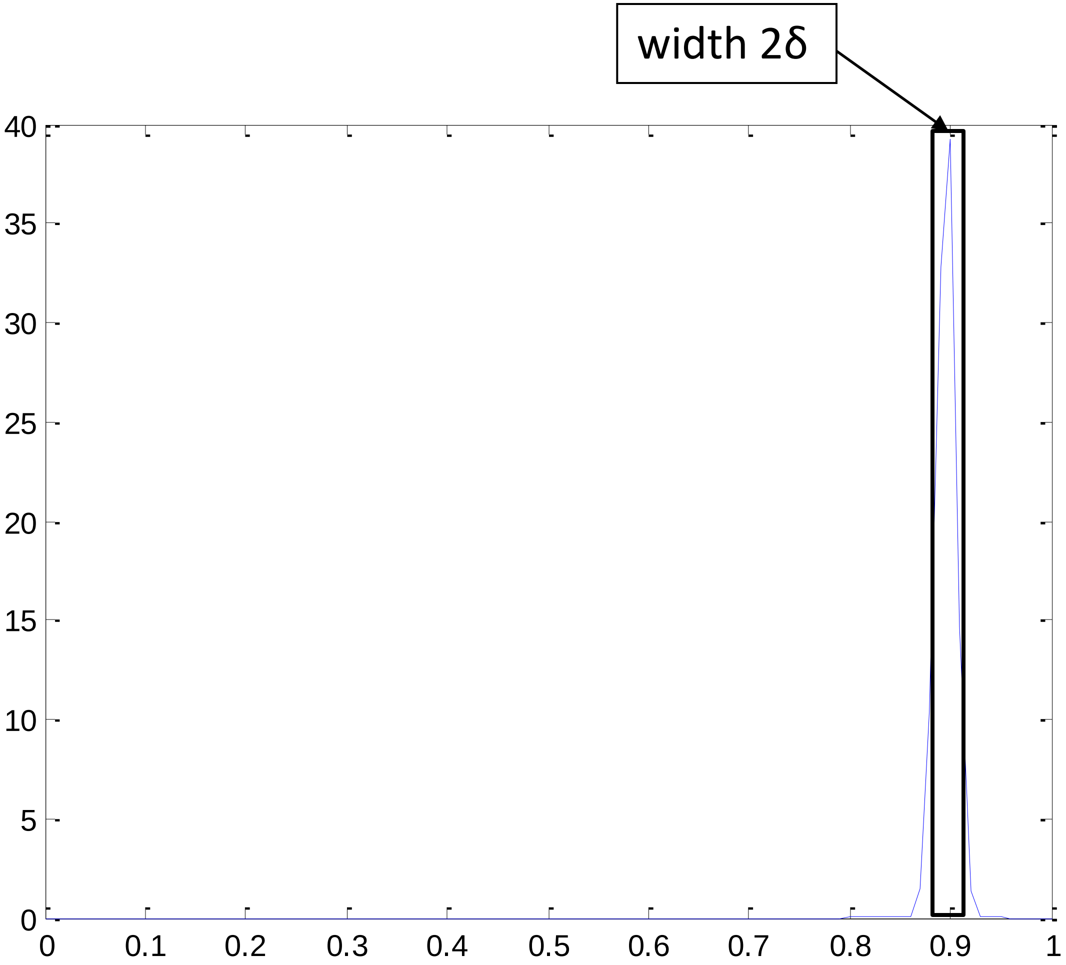
  $$\int_{t_0}^{t_1} f(u \mid x_1, \ldots, x_n)\ du = c$$

  is called a 100c percent Bayesian Interval Estimate of *p*
- *An optimal interval* minimizes t_1- t_0: difficult in general
- Our approach:
  - fix a half-interval width δ
  - Continue sampling until the posterior probability of an interval of width 2δ containing the posterior mean exceeds coverage c

Bayesian Interval Estimation - IV



prior is beta(α=4,β=5)



posterior density after 1000 samples and
900 “successes” is beta(α=904,β=105)
posterior mean = 0.8959

Bayesian Interval Estimation - V

Require: **BLTL** ***property*** *Φ*, ***interval-width*** *δ*, ***coverage*** *c*, ***prior*** *beta parameters α,β*

```
$n := 0$                {number of traces drawn so far}
$x := 0$                {number of traces satisfying so far}
repeat
        $\sigma$ := draw a sample trace of the system (iid)
        $n := n + 1$
        if $\sigma \models \Phi$ then
                $x := x + 1$
        endif
        mean = $(x+\alpha)/(n+\alpha+\beta)$
        $(t_0,t_1)$ = (mean-$\delta$, mean+$\delta$)
        $\mathcal{I}$ := PosteriorProbability $(t_0,t_1,n,x,\alpha,\beta)$
until $(\mathcal{I} > c)$
return $(t_0, t_1)$, mean
```

Bayesian Interval Estimation - VI

Theorem (Termination)

The Sequential Bayesian Estimation algorithm terminates with probability one.

- Recall the algorithm outputs the interval (t_0, t_1)
- Define the null hypothesis $H_0: t_0 < p < t_1$

Theorem (Error bound)

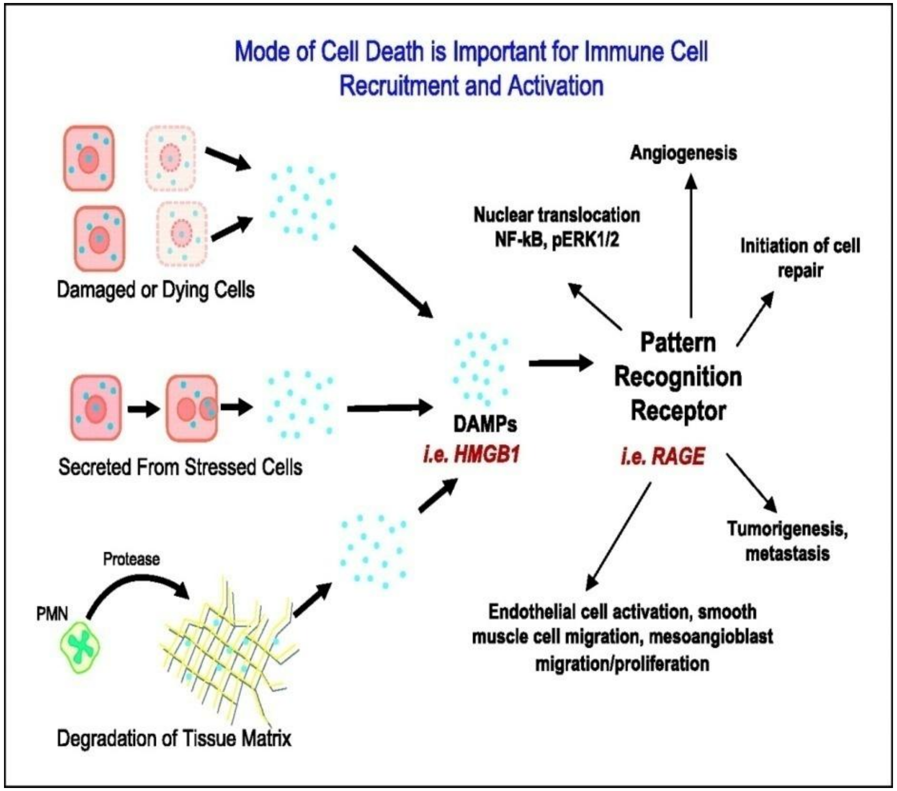
When the Bayesian estimation algorithm (using coverage $½ < c < 1$) stops – we have

$$\boxed{\begin{array}{l} \text{Prob (“accept } H_0\text{” } \mid H_1) \leq (1/c - 1)\pi_0/(1-\pi_0) \\ \text{Prob (“reject } H_0\text{” } \mid H_0) \leq (1/c - 1)\pi_0/(1-\pi_0) \end{array}}$$

π_0 is the prior probability of H_0

Verification of Biological Signaling Pathways in BioNetGen

The Protein HMGB1



Mode of Cell Death is Important for Immune Cell Recruitment and Activation
Damaged or Dying Cells
Secreted From Stressed Cells
Protease
PMN
Degradation of Tissue Matrix
DAMPs
i.e. HMGB1
Pattern Recognition Receptor
i.e. RAGE
Angiogenesis
Nuclear translocation NF-kB, pERK1/2
Initiation of cell repair
Tumorigenesis, metastasis
Endothelial cell activation, smooth muscle cell migration, mesoangioblast migration/proliferation

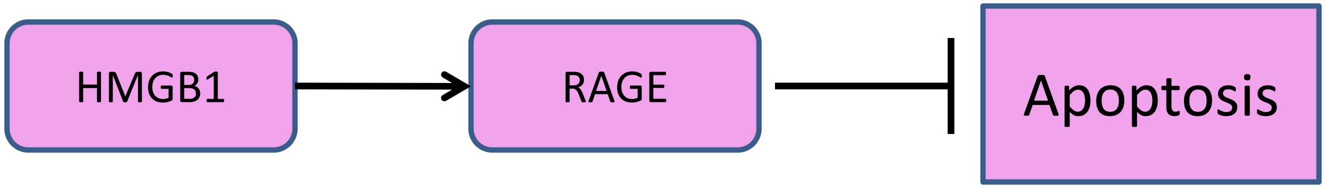
- High-Mobility Group Protein 1 (HMGB1):
  - DNA-binding protein and regulates gene transcription
  - released from damaged or stressed cells, etc.

- HMGB1 activates RAGE or TLR2/4
  - RAGE: Receptor for Advanced Glycation End products.
  - TLR: Toll-like receptor

- RAGE/TLR activation can activate NFkB and RAS signaling pathways which causes inflammation or tumorigenesis.

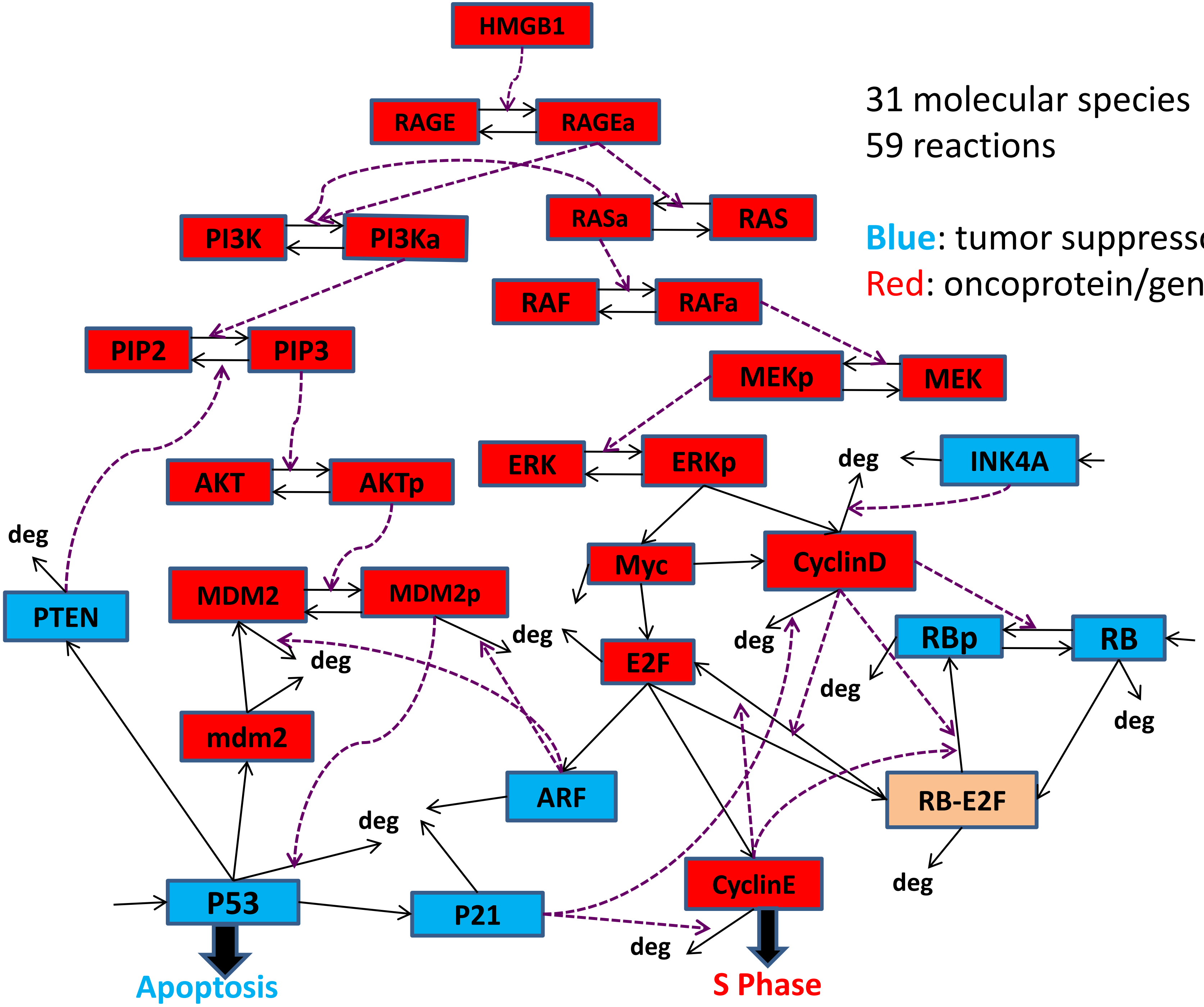
HMGB1 and Pancreatic Cancer

(Lotze *et al.*, UPMC)



Experiments with pancreatic cancer cells:

- Overexpression of HMGB1/RAGE is associated with diminished apoptosis, and longer cancer cell survival time.
- Knockout of HMGB1/RAGE leads to increased apoptosis, and decreased cancer cell survival.



31 molecular species
59 reactions

Blue: tumor suppressor
Red: oncoprotein/gene

BioNetGen.org

- Rule-based modeling for biochemical systems
- Ordinary Differential Equations and Stochastic simulation (Gillespie's algorithm: Continuous-Time Markov Chain)
- *Example*: AKT has a component named d which can be labeled as U (unphosphorylated) or p (phosphorylated)

```
begin species
   AKT(d~U)        1e5
   AKT(d~p)        0
end species
```

```
begin parameters
   k        1.2e-7
   d        1.2e-2
end parameters
```

Faeder JR, Blinov ML, Hlavacek WS **Rule-Based Modeling of Biochemical Systems with BioNetGen.** In *Methods in Molecular Biology: Systems Biology*, (2009).

BioNetGen.org

- Example:
  - PIP3 can phosphorylate AKT
  - dephosphorylation of AKT

begin reaction_rules

```
PIP(c~p)  +   AKT(d~U) ⟶  PIP(c~p)  + AKT(d~p)   k
AKT(d~p) ⟶  AKT(d~U)                              d
```

end reaction_rules

- The propensity functions for Gillespie’s algorithm are:

k·[PIP(c~p)]·[AKT(d~U)]

d·[AKT(d~p)]

Verification - I

- Overexpression of HMGB1 will induce the expression of the cell cycle regulatory protein CyclinE

$$P_{\geq 0.9}\ \mathbf{F}^{600}\ (\ \text{CyclinE} > 900\ )$$

"within 600 minutes, the number of CyclinE molecules will be greater than 900"

- error probability < 0.001

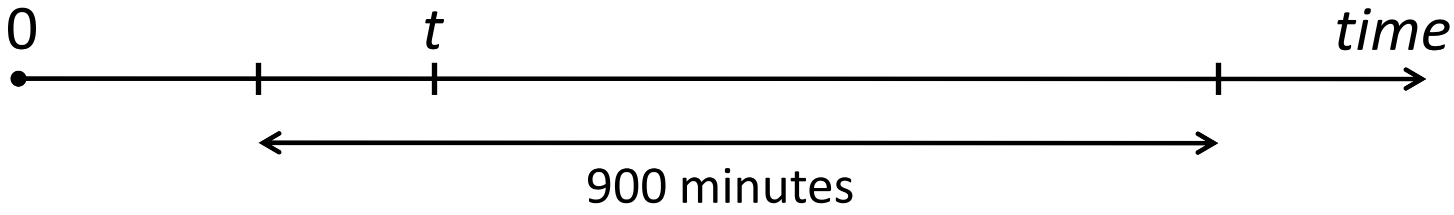
HMGB1	# samples	# Success	Result
10^2	9	0	False
10^3	55	16	False
10^6	22	22	True

Verification - II

- p53 is expressed at low levels in normal human cells

$$P_{\geq 0.9}\ \mathbf{F}^{t}\ (\ \mathbf{G}^{900}\ (\ p53 < 3.3 \times 10^{4}\ )\ )$$

"within t minutes, p53 will stay low for 900 minutes"



t (min)	# Samples	# Success	Result	Time (s)
400	53	49	True	597.59
500	23	22	True	271.76
600	22	22	True	263.79

Verification - III

- Expression level of HMGB1 influences the 1st peak of p53's level

$$P_{\geq 0.9}\ \mathbf{F}^{100}\ (\ p53 \geq a\ \&\ \mathbf{F}^{100}\ (\ p53 \leq 4 \times 10^{4}\ )\ )$$

"within 100 minutes, p53 will pass a, and in the next 100 minutes it will eventually be below 4x10⁴"

HMGB1	a (x 10^4)	# Samples	# Success	Result	Time (s)
10^3	5.5	20	3	False	29.02
10^2	5.5	22	22	True	19.65
10^2	6.0	45	12	False	56.27
10	6.0	38	37	True	41.50

Verification - IV

- Coding oscillations in temporal logic
- R is the fraction of NFkB molecules in the nucleus
- We model checked the formula

$$P_{\geq 0.9}\ \mathbf{F}^{t}\,(R \geq 0.65\ \&\ \mathbf{F}^{t}\,(R < 0.2\ \&\ \mathbf{F}^{t}\,(R \geq 0.2\ \&\ \mathbf{F}^{t}\,(R < 0.2))))$$

- The formula codes four changes of R that must happen in consecutive time intervals of maximum length t
- **Note**: the intervals need not be of the same length

Verification - IV

- Verifying oscillations of NFkB with statistical model checking

$$P_{\geq 0.9}\ \mathbf{F}^{t}\,(R \geq 0.65\ \&\ \mathbf{F}^{t}\,(R < 0.2\ \&\ \mathbf{F}^{t}\,(R \geq 0.2\ \&\ \mathbf{F}^{t}\,(R < 0.2))))$$

HMGB1	t (min)	# Samples	# Success	Result	Time (s)
10^2	45	13	1	False	76.77
10^2	60	22	22	True	111.76
10^2	75	104	98	True	728.65
10^5	30	4	0	False	5.76

Statistical MC: Weaknesses

- Rare events – too many samples needed
  - But there are ways to “solve” the problem
- Simulation is incomplete (continuous evolution)
  - OK for biological systems modeled as CTMCs

Statistical MC: Strengths

- Widely applicable!
  - Only need simulation
- Can address large (or infinite) system spaces
- Better scalability
- Can trivially exploit multi-core CPUs

The End

Questions?