Stochastic hybrid models for DNA replication

John Lygeros

Automatic Control Laboratory, ETH Zürich

www.control.ethz.ch

Outline

1. Hybrid and stochastic hybrid systems
2. Reachability & randomized methods
3. DNA replication
   - DNA replication in the cell cycle
   - A stochastic hybrid model
   - Simulation results
   - Analysis
4. Summary

Hybrid dynamics

- Both continuous and discrete state and input
- Interleaving of discrete and continuous
  - Evolve continuously
  - Then take a jump
  - Then evolve continuously again
  - Etc.
- Tight coupling
  - Discrete evolution depends on continuous state
  - Continuous evolution depends on discrete state

But what about uncertainty?

- Hybrid systems allow uncertainty in
  - Continuous evolution direction
  - Discrete & continuous state destinations
  - Choice between flowing and jumping
- "Traditionally" uncertainty worst case
  - "Non-deterministic"
  - Yes/No type questions
  - Robust control
  - Pursuit evasion game theory
- May be too coarse for some applications

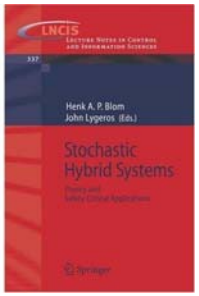
Stochastic hybrid systems

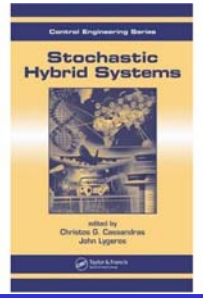
- Richer models to allow probabilities
  - Continuous evolution (e.g. SDE)
  - Discrete transition timing (Markovian, forced)
  - Discrete transition destination (transition kernel)
- Stochastic hybrid systems

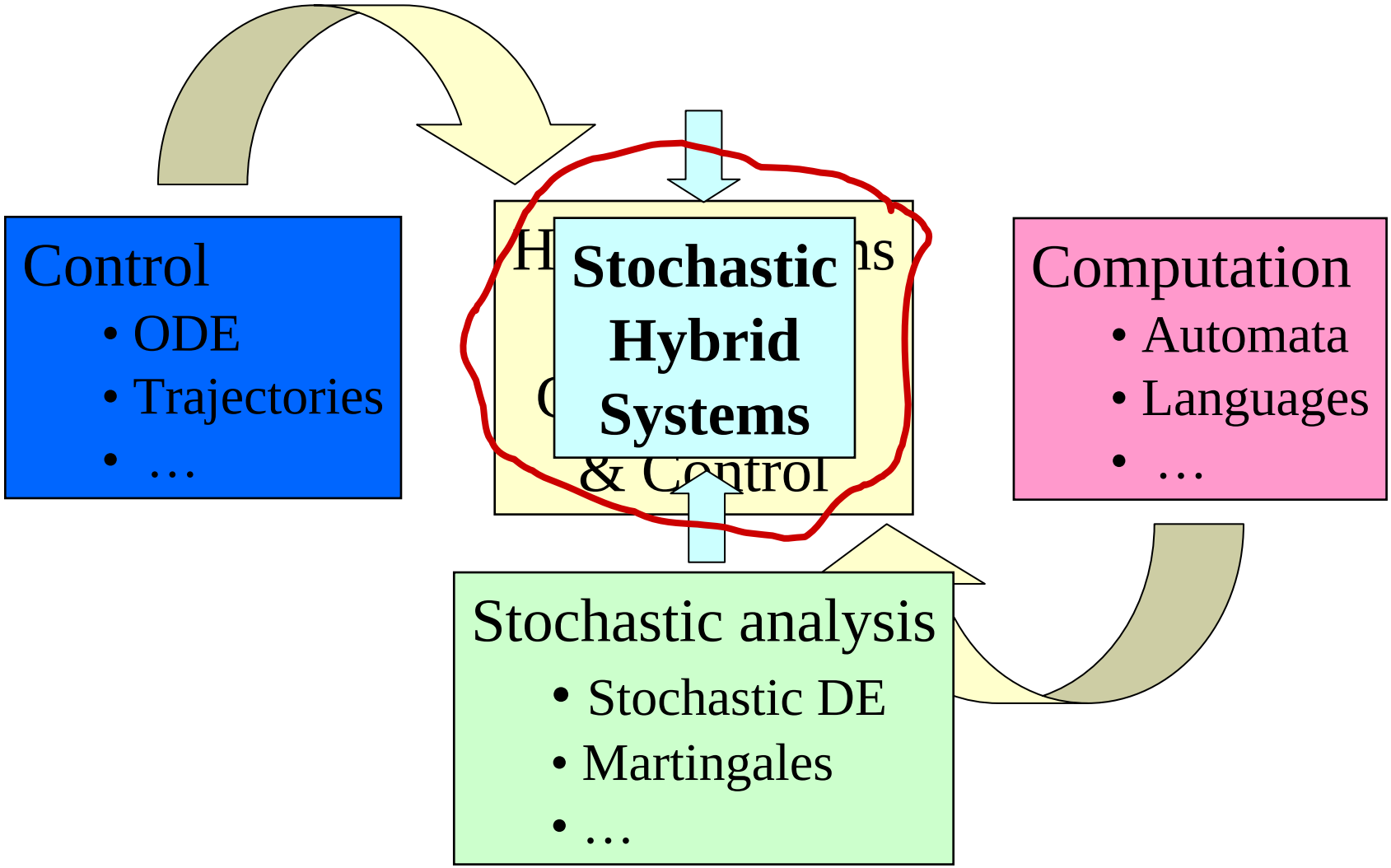
Shameless plug:

H.A.P. Blom and J. Lygeros (eds.), "*Stochastic hybrid systems: Theory and safety critical applications*", Springer-Verlag, 2006

C.G. Cassandras and J. Lygeros (eds.), "*Stochastic hybrid systems*", CRC Press, 2006



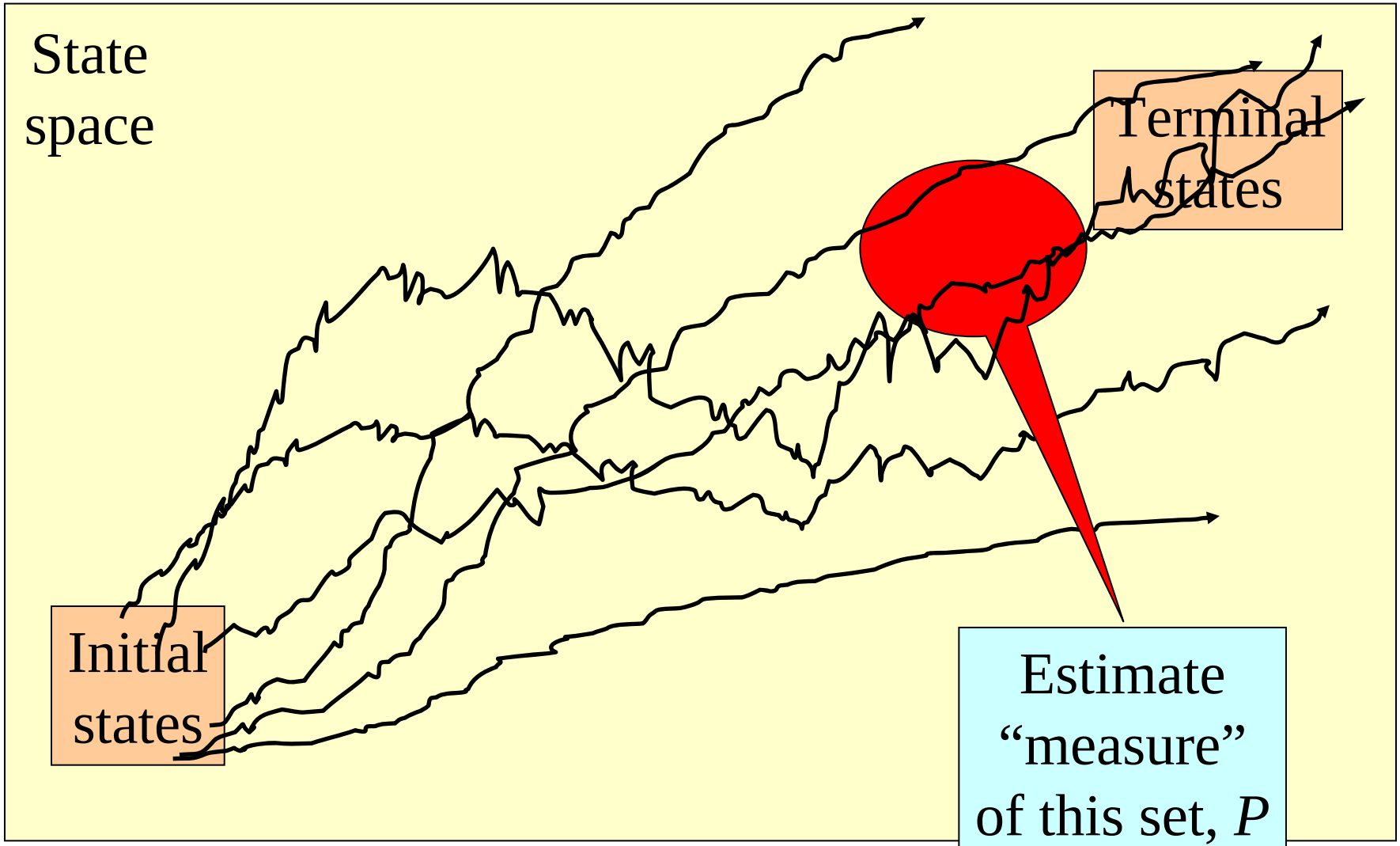



Control
• ODE
• Trajectories
• …
Stochastic Hybrid Systems
& Control
Computation
• Automata
• Languages
• …
Stochastic analysis
• Stochastic DE
• Martingales
• …

Outline

1. Hybrid and stochastic hybrid systems
2. Reachability & randomized methods
3. DNA replication
   - DNA replication in the cell cycle
   - A stochastic hybrid model
   - Simulation results
   - Analysis
4. Summary

Reachability: Stochastic HS



Monte-Carlo simulation

- Exact solutions impossible
- Numerical solutions computationally intensive
- Assume we have a simulator for the system
  - Can generate trajectories of the system
  - With the right probability distribution
- "Algorithm"
  - Simulate the system N times
  - Count number of times terminal states reached (M)
  - Estimate reach probability P by $\hat{P} = \frac{M}{N}$

Convergence

- It can be shown that $\hat{P} \to P$ as $N \to \infty$
- Moreover …

Probability that $\left|\hat{P} - P\right| \geq \varepsilon$ is at most δ as long as

$$N \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{2}{\delta}\right)$$

- Simulating more we get as close as we like
- "Fast" growth with ε slow growth with δ
- No. of simulations independent of state size
- Time needed for each simulation dependent on it
- Have to give up certainty

Not as naïve as it sounds

- Efficient implementations
  - Interacting particle systems, parallelism
- With control inputs
  - Expected value cost
  - Randomized optimization problem
  - Asymptotic convergence
  - Finite sample bounds
- Parameter identification
  - Randomized optimization problem
- Can randomize deterministic problems

Outline

1. Hybrid and stochastic hybrid systems
2. Reachability & randomized methods
3. DNA replication
   - DNA replication in the cell cycle
   - A stochastic hybrid model
   - Simulation results
   - Analysis
4. Summary

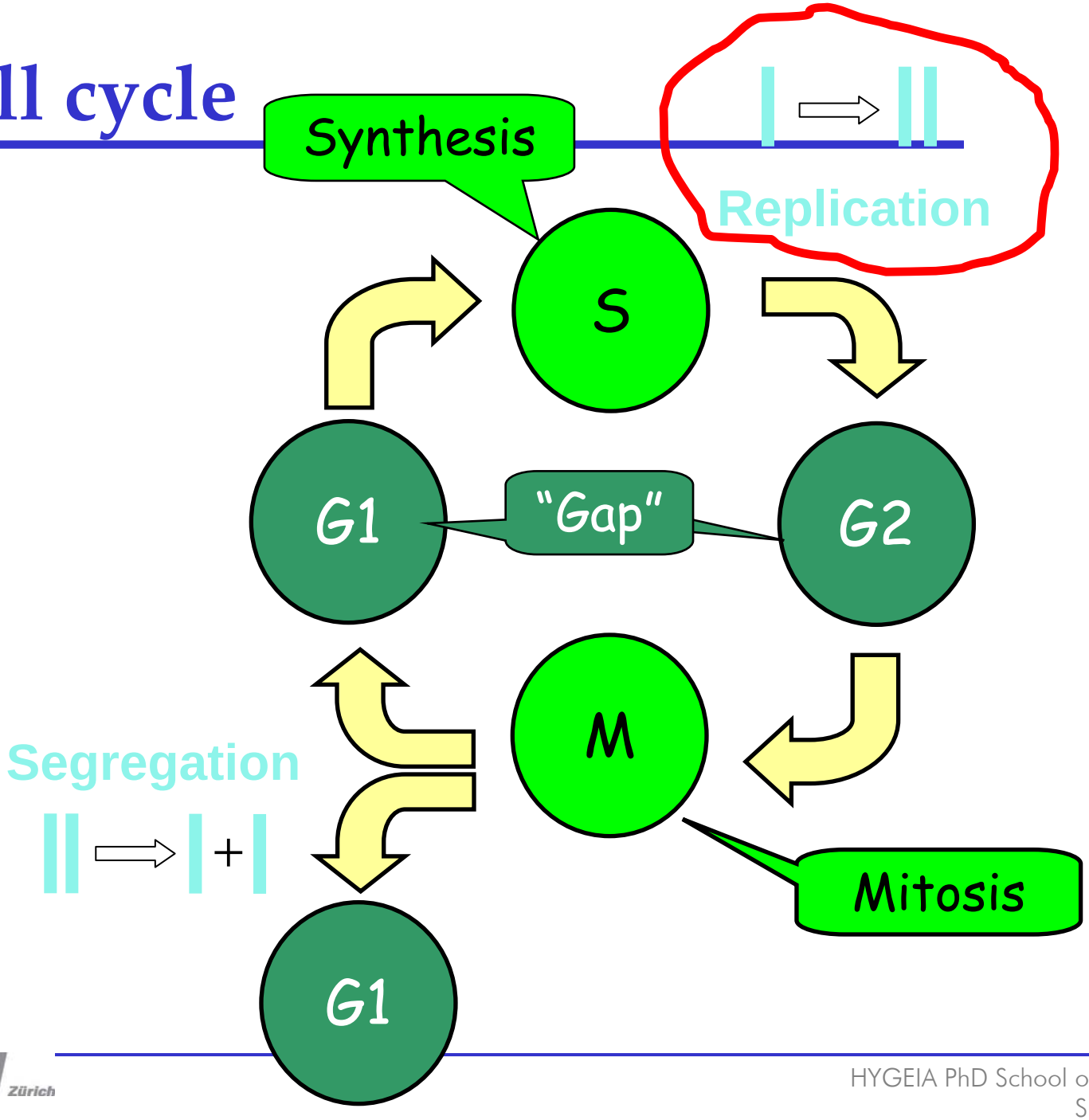
Credits

- ETH Zurich:
  - John Lygeros
  - K. Koutroumpas
- U. of Patras:
  - Zoe Lygerou
  - S. Dimopoulos
  - P. Kouretas
  - I. Legouras
- Rockefeller U.:
  - Paul Nurse
  - C. Heichinger
  - J. Wu

HYGEIA
FP6-NEST-04995

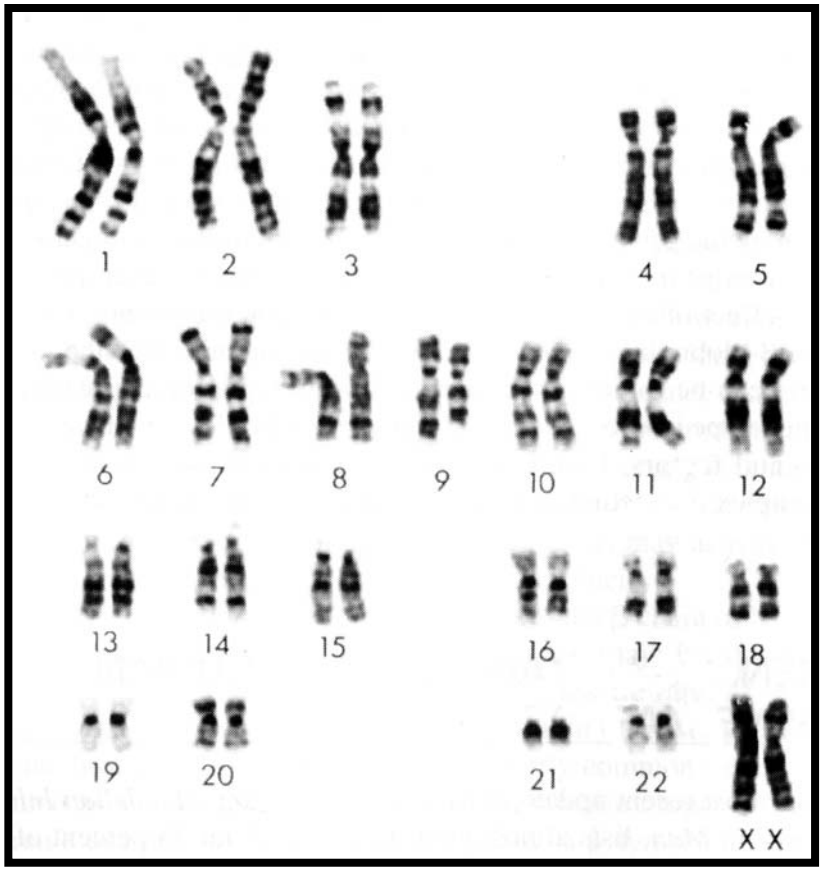
www.hygeiaweb.gr

Cell cycle

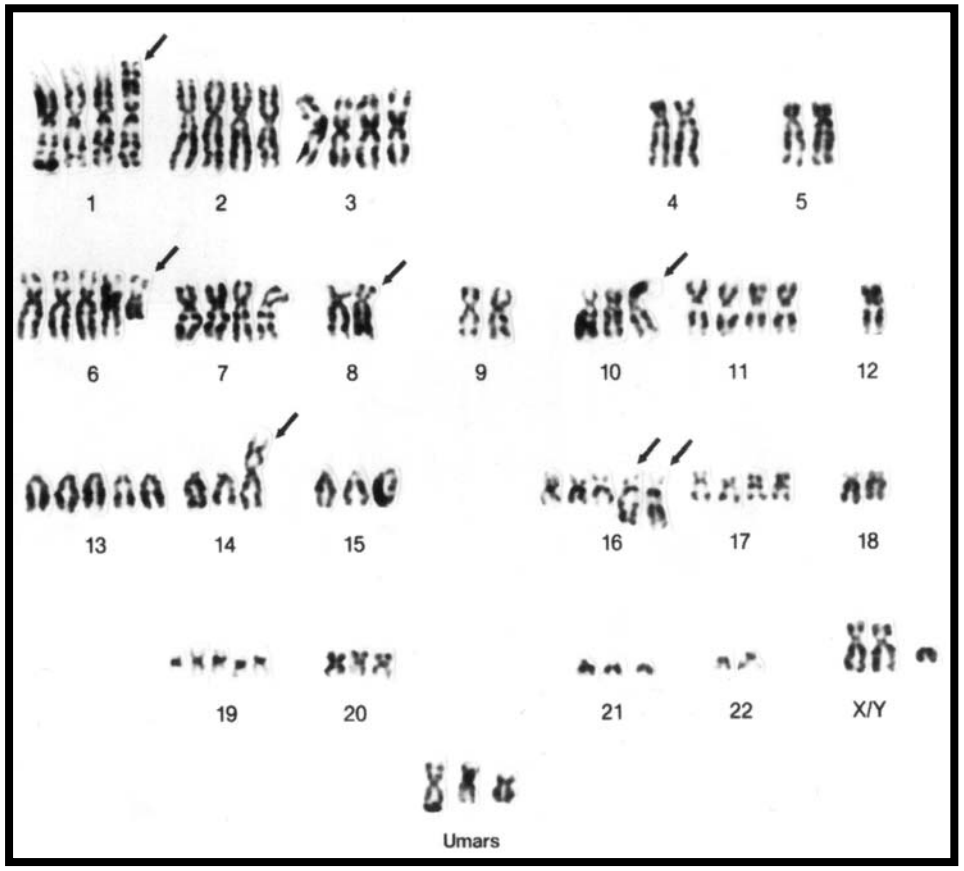


Process needs to be tightly regulated

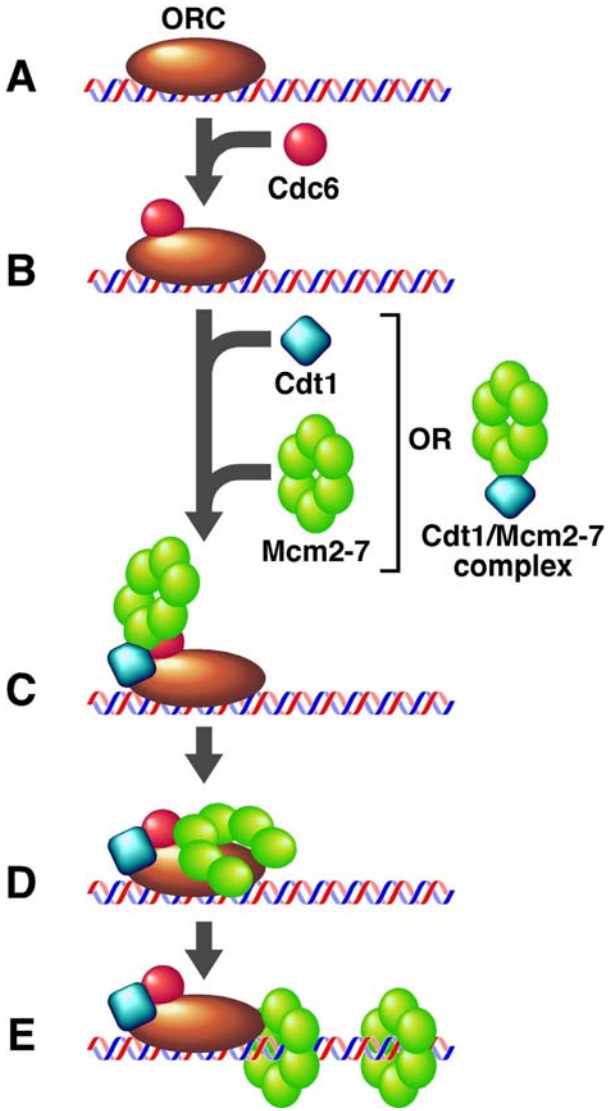
Normal cell

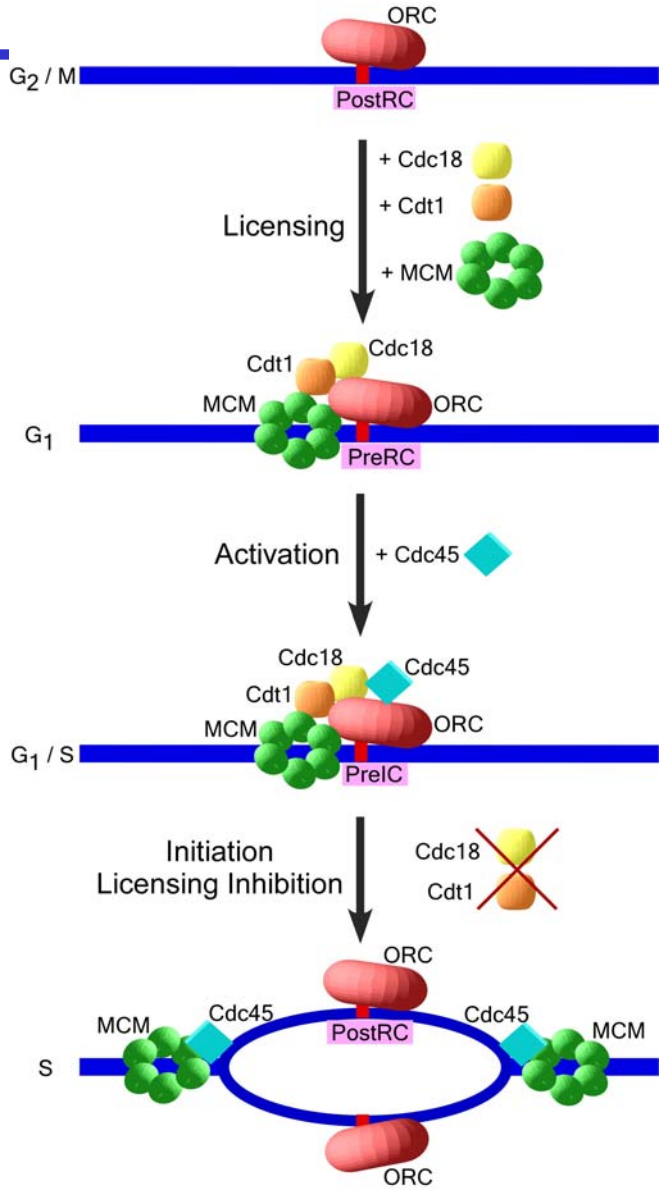


Metastatic colon cancer



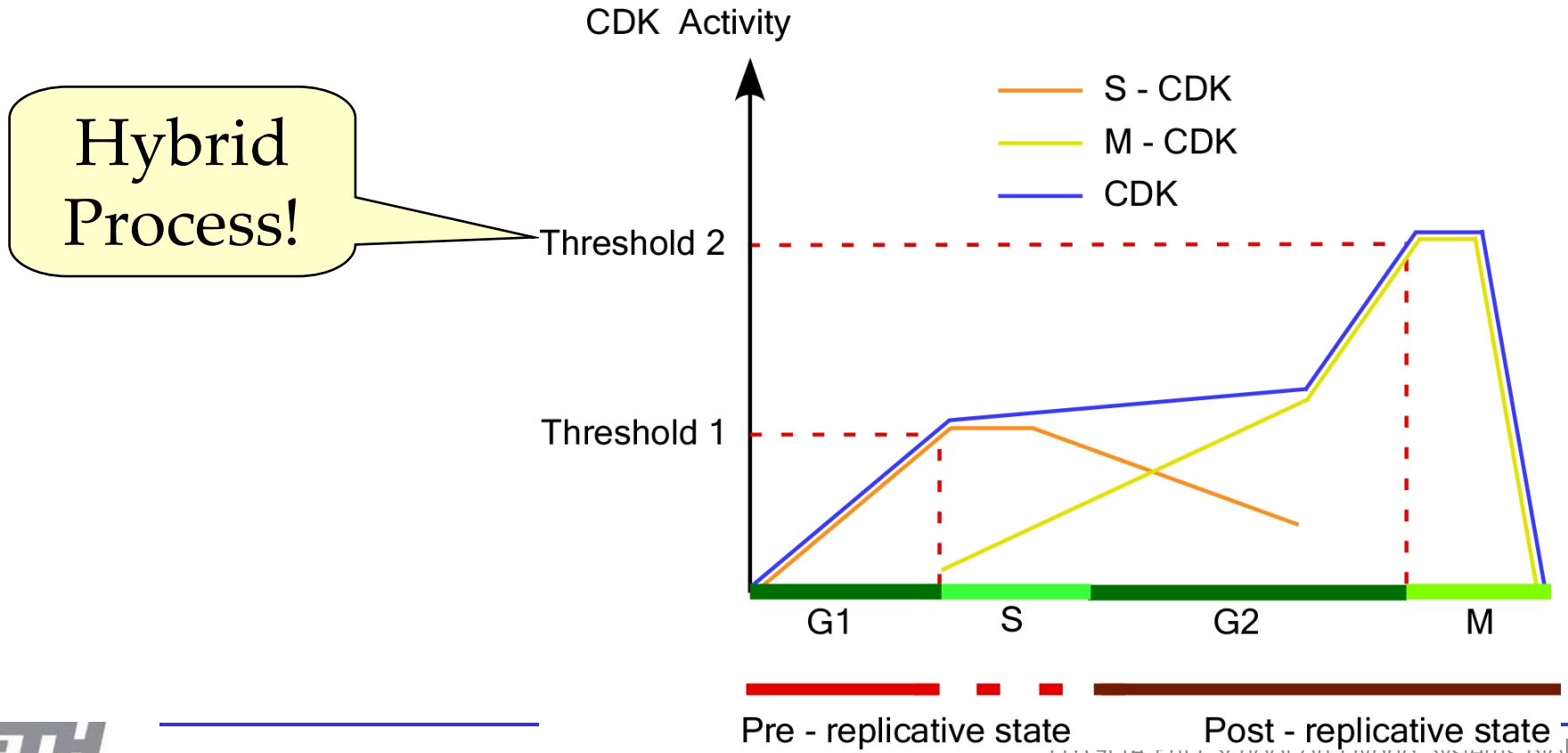
Origins of replication





Regulatory biochemical network

- CDK activity sets cell cycle pace [Nurse et.al.]
- Complex biochemical network, ~12 proteins, nonlinear dynamics [Novak et.al.]



Process "mechanics"

- Discrete
  - Firing of origins
  - Passive replication by adjacent origin
- Continuous
  - Forking: replication movement along genome
  - Speed depends on location along genome
- Stochastic
  - Location of origins (where?)
  - Firing of origins (when?)

Different organisms, different strategies

- Bacteria and budding yeast
  - Specific sequences that act as origins
  - With very high efficiency (>95%)
  - Process very deterministic
- Frog and fly embryos
  - Any position along genome can act as an origin
  - Random number of origins fire
  - Random patterns of replication
- Most eukaryots (incl. humans and *S. pombe*)
  - Origin sequences have certain characteristics
  - Fire randomly with some "efficiency"

N. Rind, "DNA replication timing: random thoughts about origin firing", *Nature cell biology*, 8(12), pp. 1313-1316, December 2006

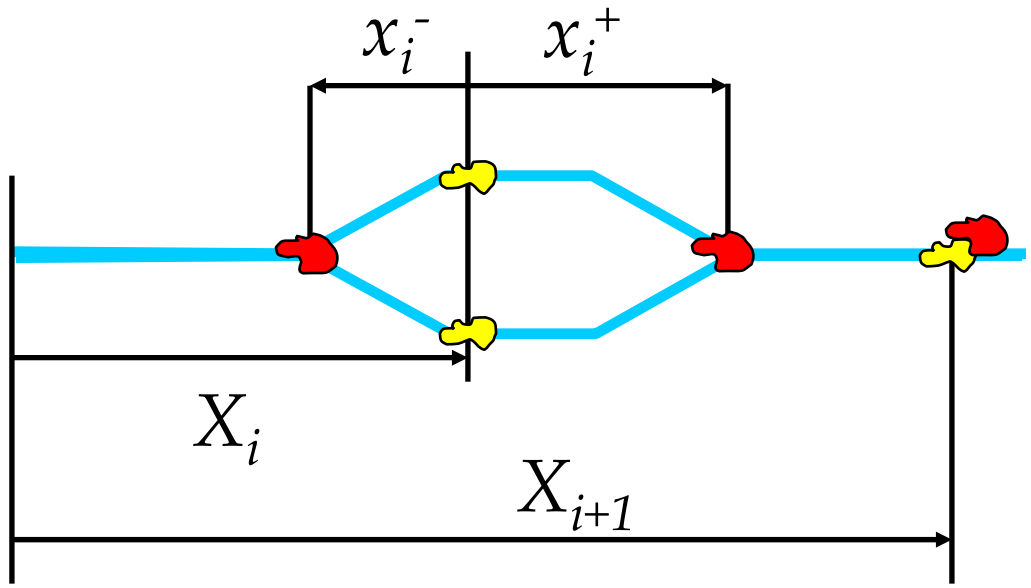
Model data

- Split genome into pieces
  - Chromosomes
  - May have to split further
- For each piece need:
  - Length in bases
  - # of potential origins of replication (n)
  - $p(x)$ p.d.f. of origin positions on genome
  - $\lambda(x)$ firing rate of origin at position x
  - $v(x)$ forking speed at position x

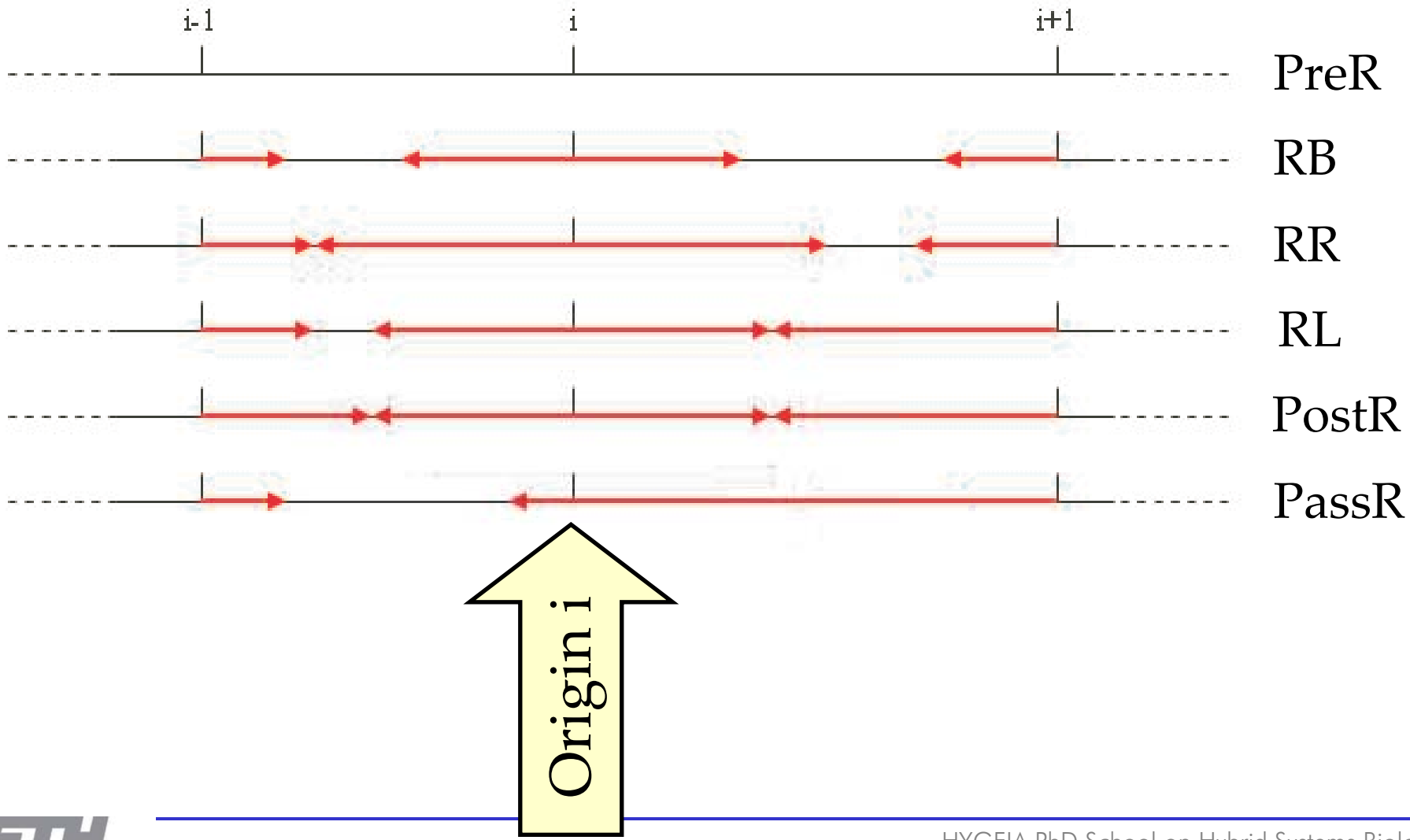
Stochastic terms

- Extract origin positions $X_i \sim p(x),\ i = 1, \ldots, n$
- Extract firing time, T_i, of origin i

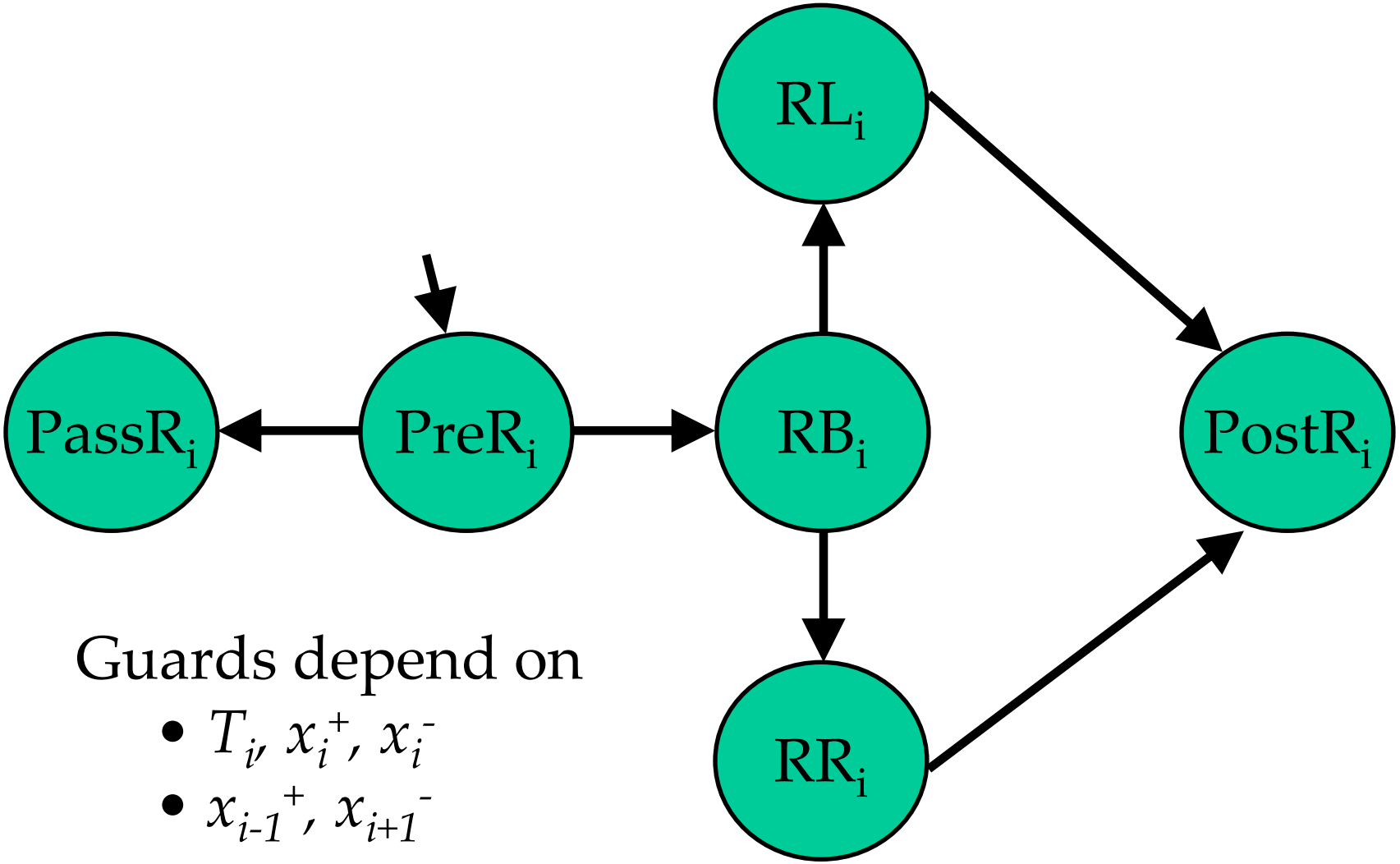
$$P\{T_i > t\} = e^{-\lambda(X_i)t}$$



Different "modes"



Discrete dynamics (origin i)



Guards depend on

- T_i, x_i^+, x_i^-
- x_{i-1}^+, x_{i+1}^-

Continuous dynamics (origin i)

- Progress of forking process

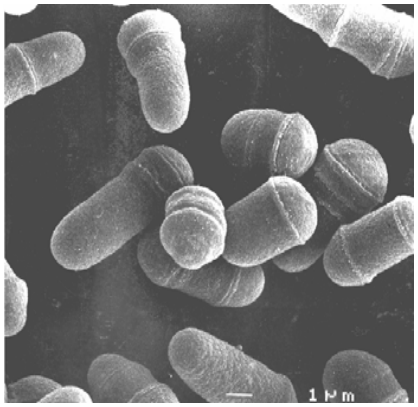
$$\dot{x}_i^+ = \begin{cases} v(X_i + x_i^+) & \text{if } q(i) \in \{RB, RR\} \\ 0 & \text{otherwise} \end{cases}$$

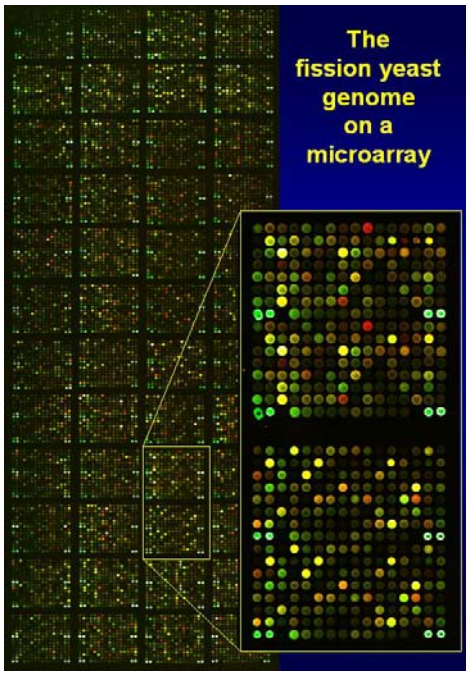
$$\dot{x}_i^- = \begin{cases} v(X_i - x_i^-) & \text{if } q(i) \in \{RB, RL\} \\ 0 & \text{otherwise} \end{cases}$$

P. Kouretas, K. Koutroumpas, J. Lygeros, and Z. Lygerou, "Stochastic hybrid modeling of biochemical processes," in *Stochastic Hybrid Systems* (C. Cassandras and J. Lygeros, eds.), no. 24 in Control Engineering, pp. 221–248, Boca Raton: CRC Press, 2006

Fission yeast model

- Instantiate: *Schizzosacharomyces pombe*
  - Fully sequenced [Bahler et.al.]
  - ~12 Mbases, in 3 chromosomes
  - Exclude
    - Telomeric regions of all chromosomes
    - Centromeres of chromosomes 2 & 3
  - 5 DNA segments to model
- Remaining data from experiments
  - C. Heichinger & P. Nurse





C. Heichinger, C.J. Penkett, J. Bahler, P. Nurse, "Genome wide characterization of fission yeast DNA replication origins", *EMBO Journal*, vol. 25, pp. 5171-5179, 2006

Experimental data input

- 863 origins
- Potential origin locations known, $p(x)$ trivial
- "Efficiency", FP_i, for each origin, i
  - Fraction of cells where origin observed to fire
  - Firing probability
  - Assuming 20 minute nominal S-phase

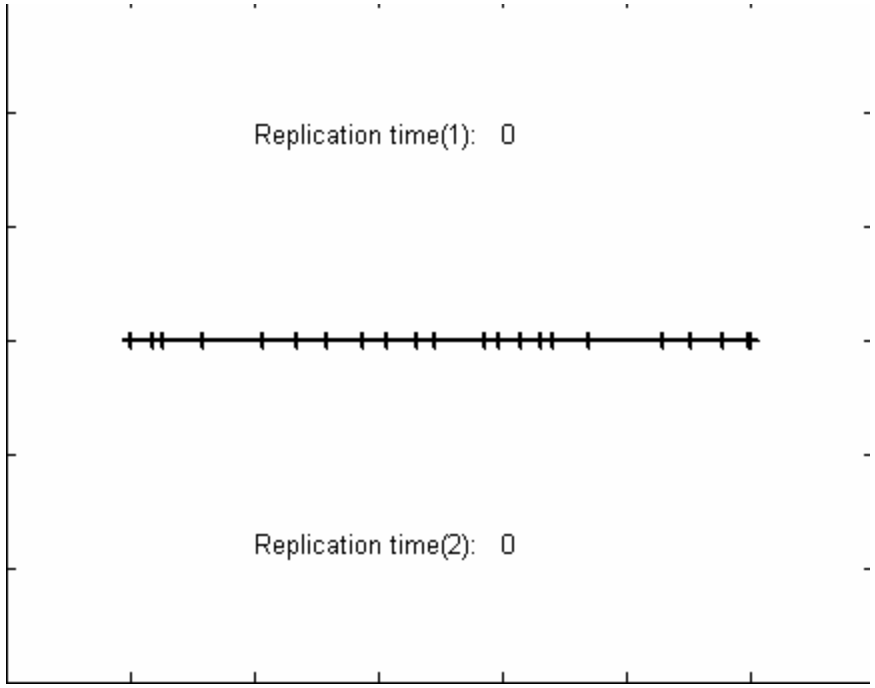
$$FP_i = \int_0^{20} \lambda_i e^{-\lambda_i t} dt \Rightarrow \lambda_i = -\frac{\ln(1-FP_i)}{20}$$

- Fork speed constant, $v(x)$=3kbases/minute

Simulation

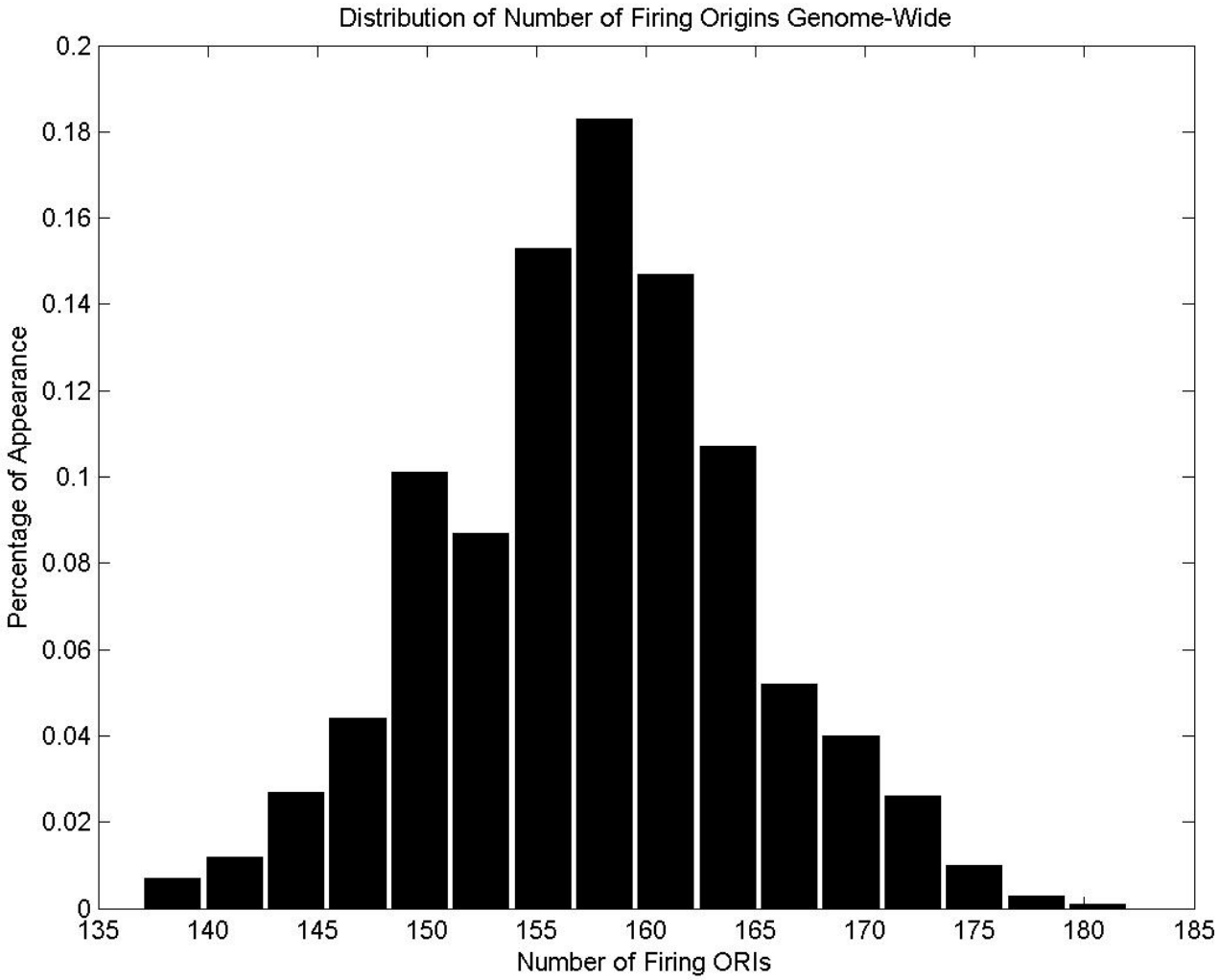
- Piecewise Deterministic Process [Davis]
- Model size formidable
  - Up to 1726 continuous states
  - Up to 6^{863} discrete states
- Monte-Carlo simulation in Matlab
  - Model probabilistic, each simulation different
  - Run 1000 simulations, collect statistics
- Check statistical model predictions against independent experimental evidence
  - S. phase duration
  - Number of firing origins

Example runs



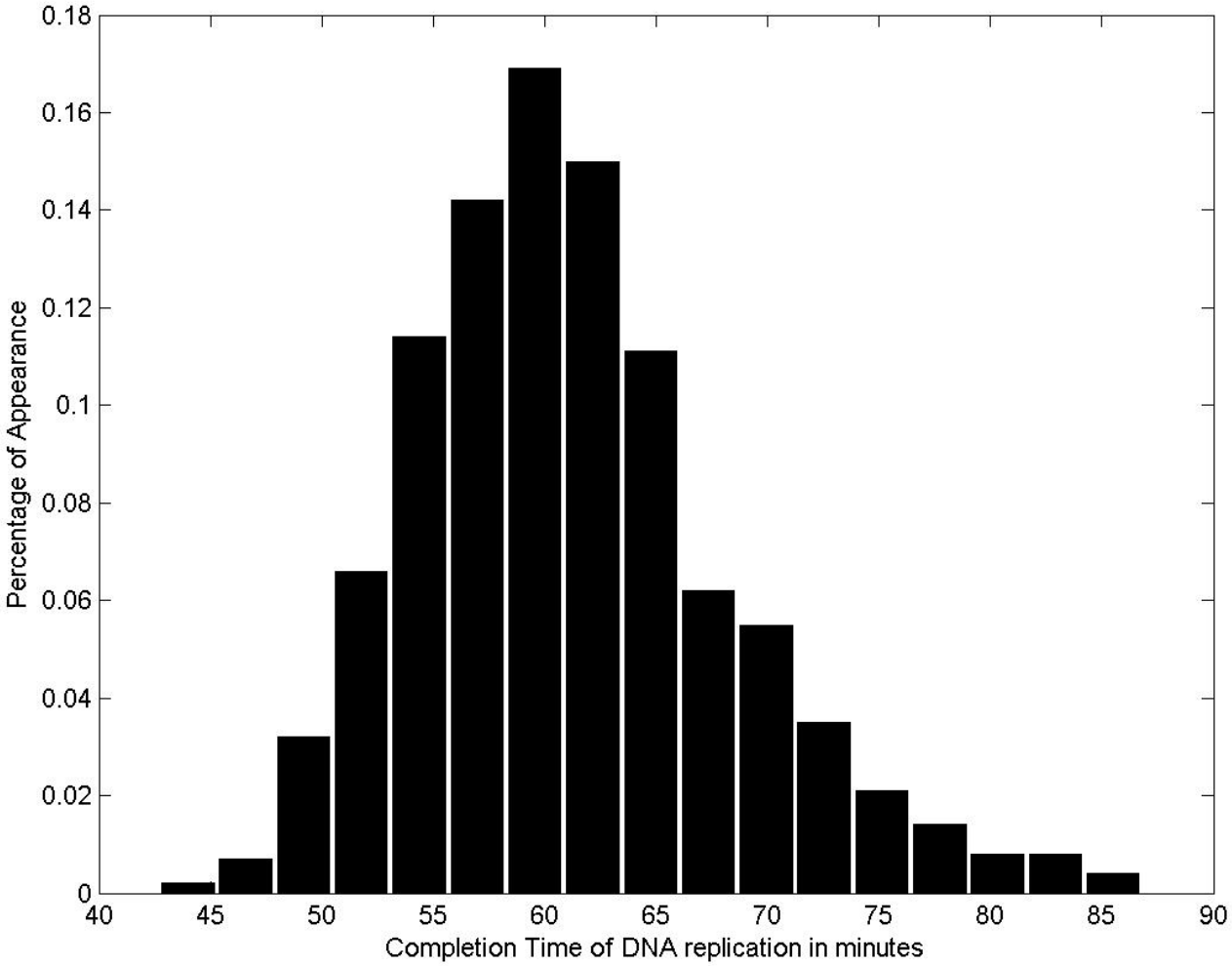
Created by
K. Koutroumpas

MC estimate: efficiency



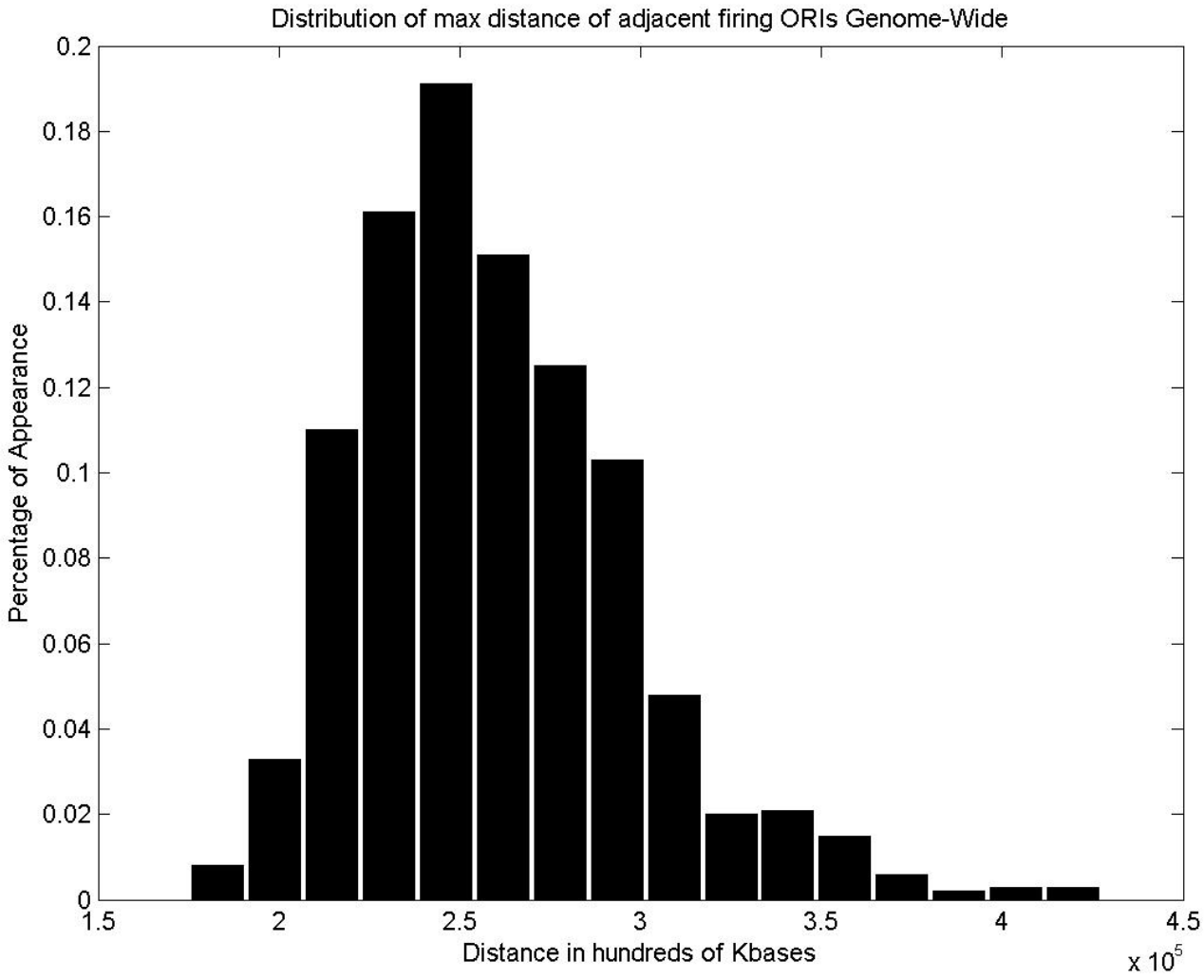
Close to experimental

MC estimate: S-phase duration



Empirical:
19 minutes!

MC estimate: Max inter-origin dist.



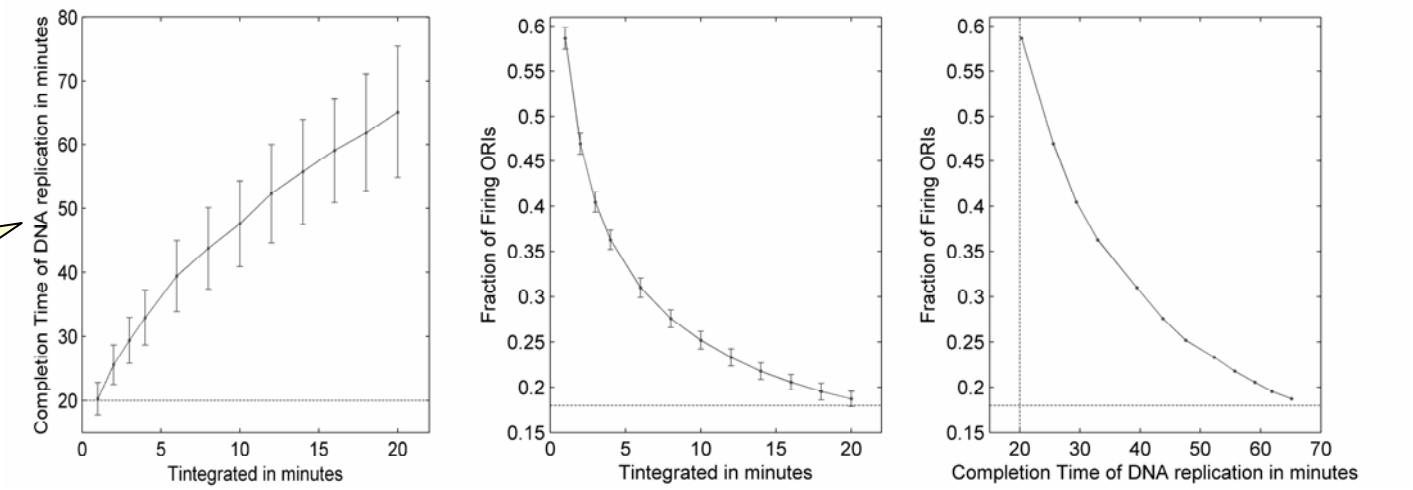
Random gap problem

Possible explanations

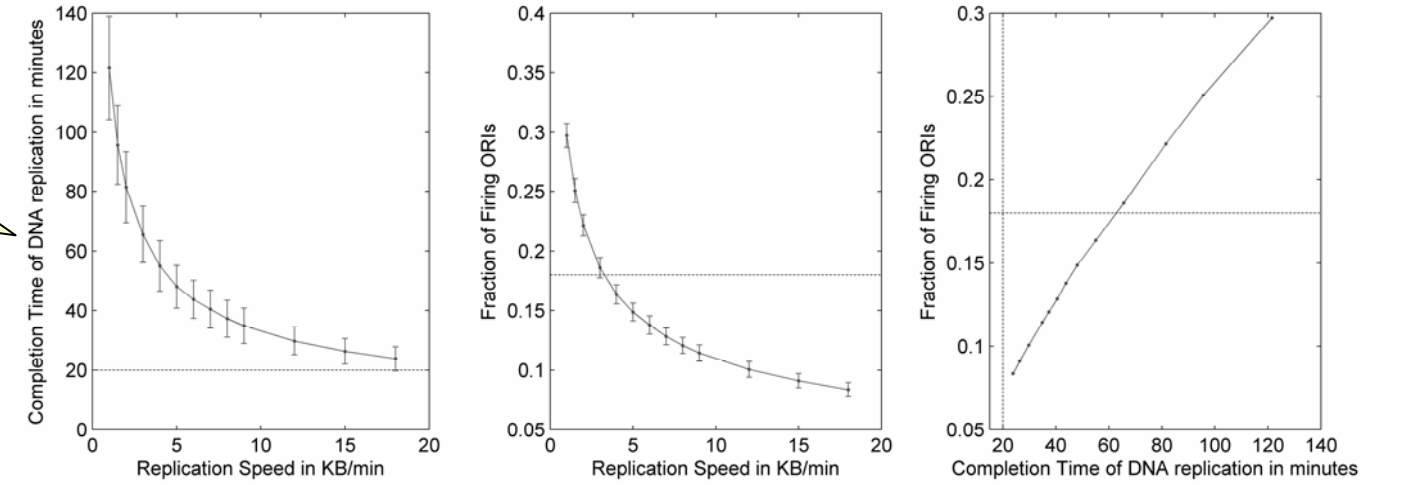
- Efficiencies used in model are wrong
  - System identification to match efficiencies
  - Not a solution, something will not fit
- Speed approximation inaccurate
  - "Filtering" of raw experimental data
  - Not a solution, something will not fit
- Inefficient origins play important role
  - Motivation for bioinformatic study
  - AT content, asymmetry, inter-gene, …
  - Also chromatin structure
  - Not a solution

Possible explanations (not!)

Increasing efficiency

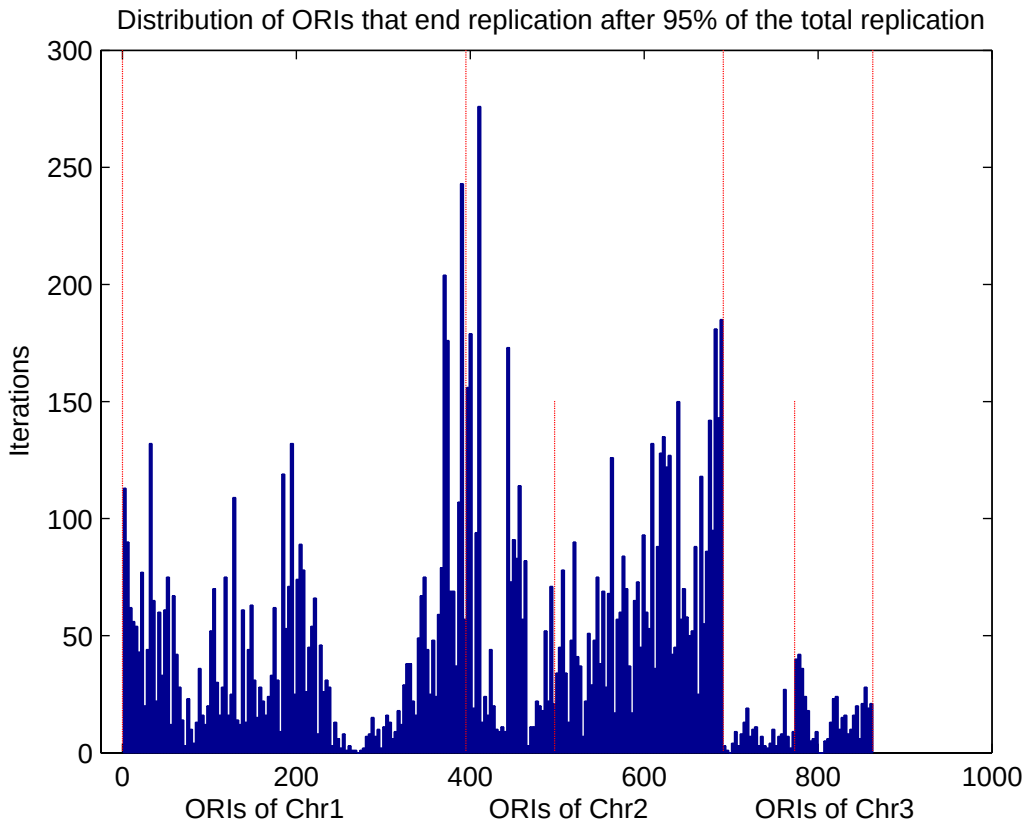


Increasing fork speed



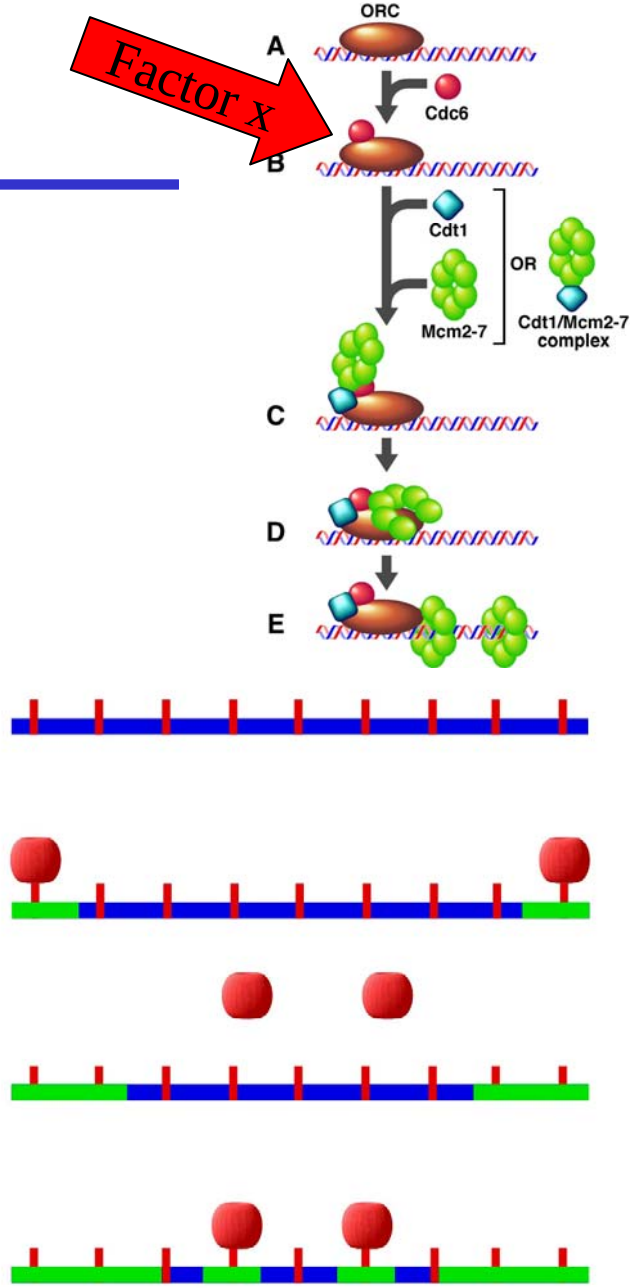
Possible explanations

- DNA replication continues into G2 phase
  - Circumstantial evidence S phase may be longer
  - Use model to guide DNA combing experiments

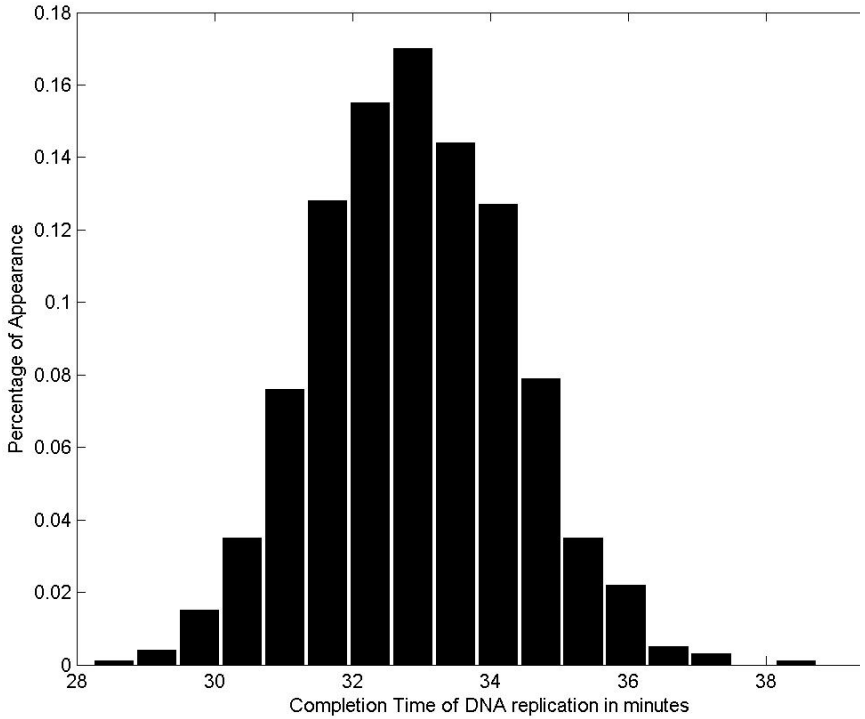


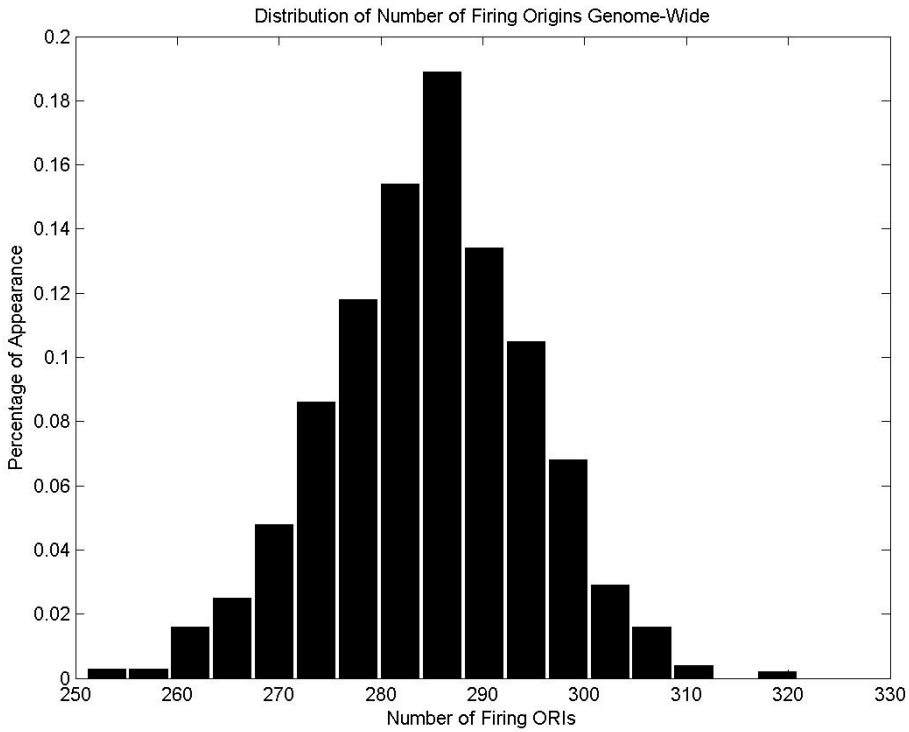
Possible explanations

- Firing propensity redistribution
  - Limiting "factor" binding to potential origins
  - Factor released on firing or passive replication
  - Can bind to pre-replicating origins
  - Propensity to fire increases in time

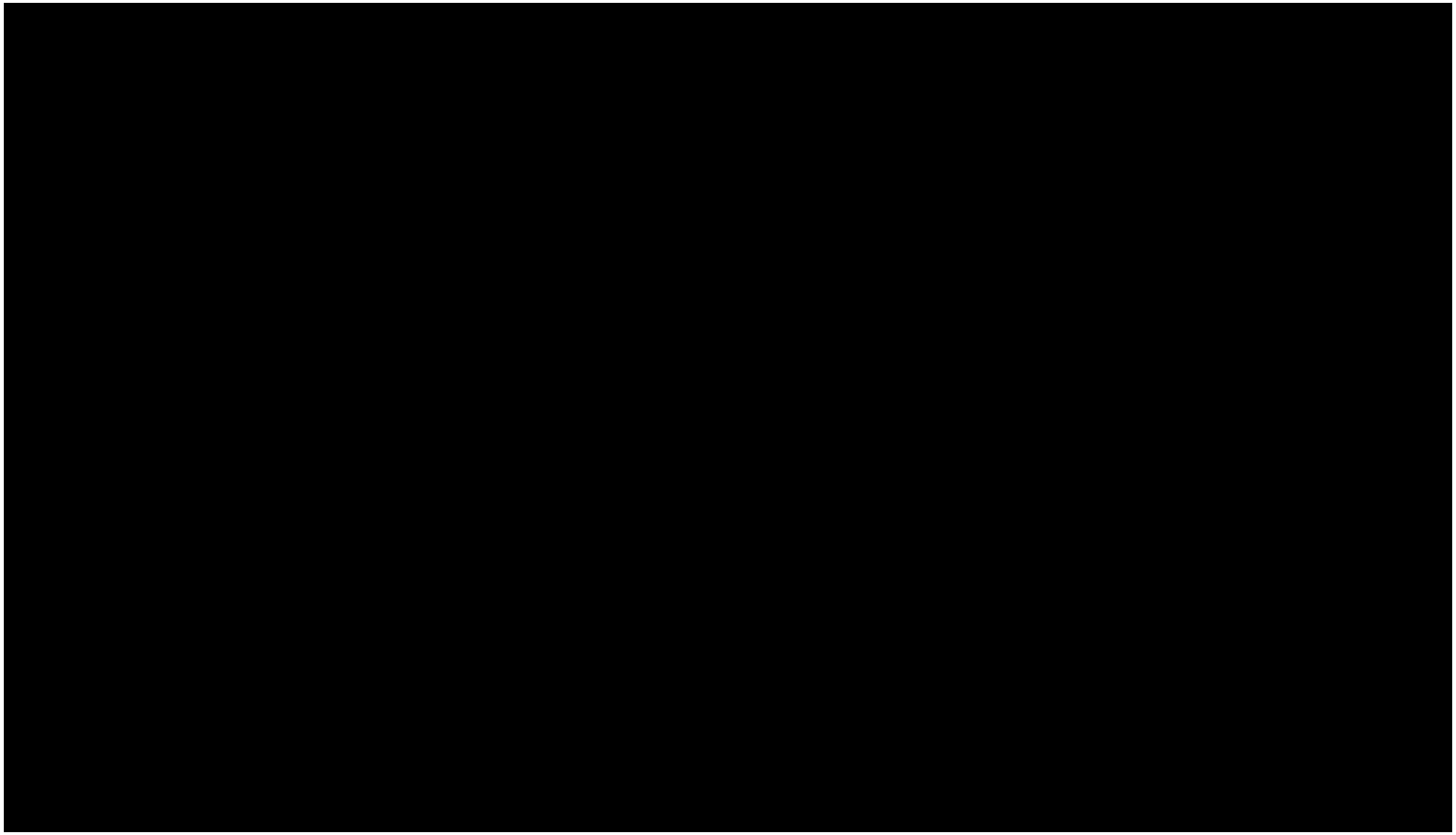


Firing propensity redistribution





Re-replication



Created by
K. Koutroumpas

Outline

1. Hybrid and stochastic hybrid systems
2. Reachability & randomized methods
3. DNA replication
   - DNA replication in the cell cycle
   - A stochastic hybrid model
   - Simulation results
   - Analysis
4. Summary

Concluding remarks

- DNA replication in cell cycle
  - Develop SHS model based on biological intuition & experimental data
  - Code model and simulate
  - Exposed gaps in intuition
  - Suggested new questions and experiments
- Simple model gave rise to many studies
  - System identification for efficiencies, filtering for fork speed estimation, bioinformatics origin selection criteria
  - DNA combing to detect G2 replication
  - Theoretical analysis
  - Extensions: re-replication
- Promote understanding, e.g.
  - Why do some organisms prefer deterministic origin positions?