Stochastic Biomodelling

Andrzej Mizera (andrzej.mizera@uni.lu)
Faculty of Science, Technology and Communication
University of Luxembourg

Course content

1. Introduction
   - Stochasticity in biological processes
   - Deterministic vs stochastic biomodelling
2. Prerequisite:
   - Crash course on probability theory
3. Stochastic modelling of chemical kinetics: the chemical master equation (CME)
4. Stochastic simulation of the CME – Gillespie's direct method algorithm
5. Practicals:
   - Implementing the Gillespie's algorithm in MATLAB and investigating its characteristics on various biochemical systems
   - Comparing the obtained simulation results with the solutions in the deterministic formulation

Some words on what is the role of mathematical modelling in Systems Biology

Course webpage

http://combio.abo.fi/teaching/tucs-short-courses/stochastic-biomodelling/

Stochastic modelling in Systems Biology

Stochastic modelling of chemical kinetics

- Discrete-time Markov chains (DTMC)
- Continuous-time Markov chains (CTMC)
- Chemical Master Equation (CME)

Stochastic processes

A **stochastic process** describes the behaviour of a ***random variable*** which ***evolves over time***.

- Instead of dealing with only one possible "reality" of how the process might evolve under time (e.g., ODEs), in a stochastic process there is some *indeterminacy* in its future evolution *described by probability distributions*: even if the initial condition is known, there are many possibilities the process might go to, but some paths are more probable and others less.

- The **state** may be continuous or discrete and it can also evolve over time in a discrete or continuous way.

- *In the context of modelling in systems biology, the random variable of a stochastic process is usually the state of a biochemical network.*

Acknowledgments

The continuation follows the materials from Ion Petre – "Advanced computational modeling", Åbo Akademi University (2012-2013), Lectures 5 & 7

http://www.users.abo.fi/ipetre/advcompmod/

Stochastic processes

In every stage of the model there are several options of how to continue

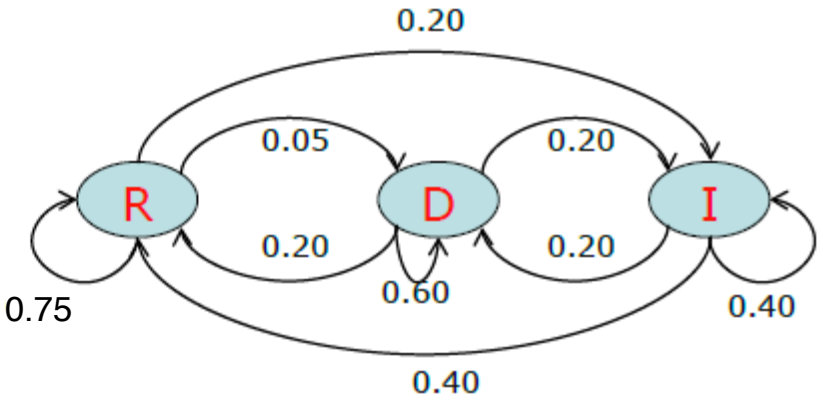
- evaluate which options are available
- what is the probability of each option
- choose one according to the probability distribution
- the system ceases having only one future; several evolutions possible
- the probability of each option might change as the system advances

Stochastic processes

Example: voting tendencies

- Assume 3 parties: Republicans, Democrats, Independents.
- Problem: identify the long-term behaviour of voters in a presidential election
- Assumptions
  - data collected over the last 10 years shows the following average trends in voting:

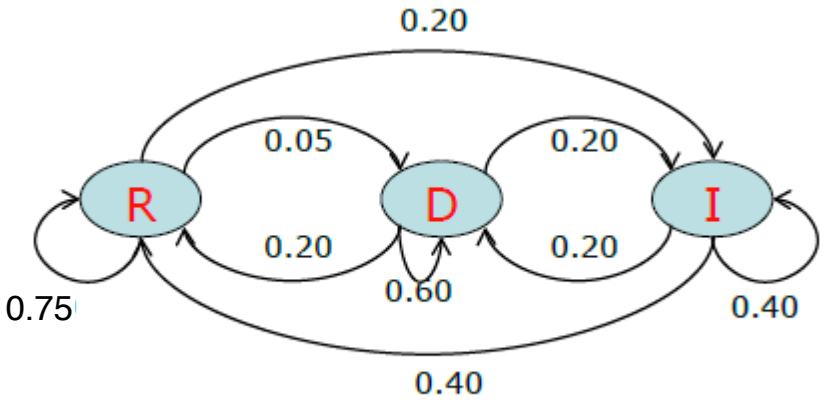
Last election \ Next election	Republican	Democrat	Independent
Republican	0.75	0.05	0.20
Democrat	0.20	0.60	0.20
Independent	0.40	0.20	0.40



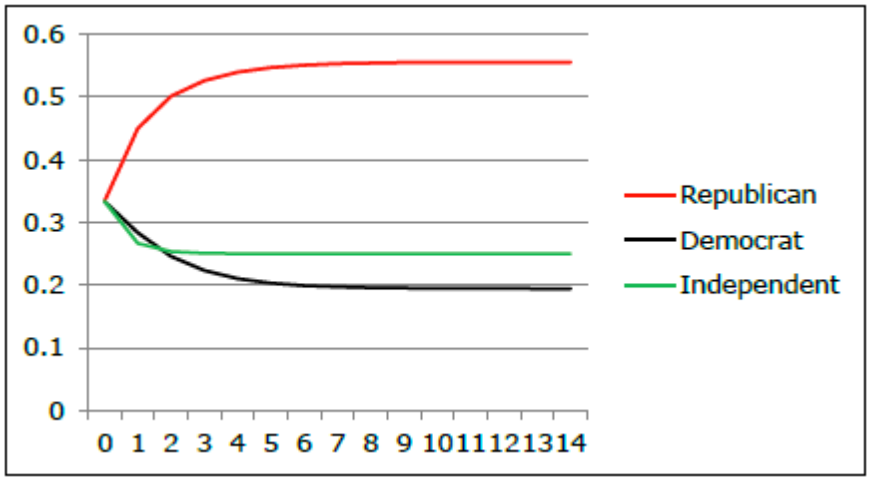
Stochastic processes

- Model formulation
  - R_n=percentage of voters to vote Republican in election n
  - D_n=percentage of voters to vote Democratic in election n
  - I_n=percentage of voters to vote Independent in election n

$$R_{n+1} = 0.75\,R_n + 0.20\,D_n + 0.40\,I_n$$
$$D_{n+1} = 0.05\,R_n + 0.60\,D_n + 0.20\,I_n$$
$$I_{n+1} = 0.20\,R_n + 0.20\,D_n + 0.40\,I_n$$



- Numerical solution: start from an initial distribution of voters and calculate the model predictions for future elections and for long-term (asymptotic) behaviour



Discrete-time Markov chains

Discrete-time Markov chain (DTMC)

- A discrete stochastic process – the state-space is discrete
- Satisfies the Markov property: the next state is only dependent on the present state, not on the past states
  - in other words: the process has no memory
- Formally: a sequence of random variables X_1, X_2, … taking values in a countable set S, such that
  - $P(X_{n+1}=x \mid X_1=x_1, X_2=x_2, \ldots, X_n=x_n) = P(X_{n+1}=x \mid X_n=x_n)$,
  - $P(X_{n+1}=x \mid X_n=x_n)$ is called the (1-step) transition probability from state x_n to state x,
  - S is called the state space of the chain,
  - the Markov chain can be described by a labelled directed graph with the elements of S as the vertices and the transition probabilities are the labels of the edges.
- Note: discrete time, discrete state space

DTMC

- We only consider in this course **time-homogenous** Markov chains (the transition probability matrix does not change in time)

- The transition probability matrix P is a **stochastic matrix**, i.e.:
  - quadratic number of elements with respect to the size of S
  - $0 \leq P(r,s) \leq 1$, for all states $r, s \in S$
  - $\Sigma_s P(r,s)=1$, for any state $r \in S$

- Easy to see that P^n is a stochastic matrix, for any $n \geq 1$

Stochastic modelling of biological systems

- We follow each species through its number of particles
- Result: a discrete system with a continuous-time dynamics
  - The state of the system: the number of particles of each species
  - A reaction taking place triggers a change in the number of particles in several species, according to its stoichiometric coefficients
    - $A+B \rightarrow 2C+3D$
    - This is modelled through changing the state to that state where the number of particles of type A is 1 less, that of type B is 1 less, that of type C is 2 more and that of that of type D is 3 more

Stochastic modelling of biological systems

- Question: When will the next reaction occur given the current state of a system?
  - If there is more than one enabled reaction, which one will be the first to occur?
    - The first reaction to occur will change the state of the system and thus the set of enabled reactions.
  - More than just one possible run of the system! How do we cope with having more than one possible future?

Continuous-time Markov chains (CTMC)

Main difference with respect to discrete-time Markov chains (discrete- vs. continuous-time)

- **DTMC:**
  - in each step, the system changes to a new (possibly the same) state
  - time is not modelled: only the sequence of states (events) is considered
- **CTMC:**
  - the system will be in the current state for some random period of time, it then changes to a new (possibly the same) state
  - the real time is explicitly considered and modelled

CTMC

- A stochastic process $\{X(t), t \geq 0\}$ is a **continuous-time Markov chain (CTMC)** if for all s, $t \geq 0$ and nonnegative integers $0 \leq u \leq s+t$,

$$P\{X(s+t)= x_{s+t} \mid X(s)= x_s, X(u)= x_u, 0 \leq u \leq s\} =$$

$$= P\{X(s+t)= x_{s+t} \mid X(s)= x_s\}$$

- In other words:
  - the process has no memory;
  - its future is only determined by the present state;
  - the future behaviour of the process depends on the past behaviour of the process only via the current state

CTMC

- The CTMC is called time-homogenous if

$$P\{X(s+t)=j \mid X(s)=i\} = P\{X(t)=j \mid X(0)=i\}$$

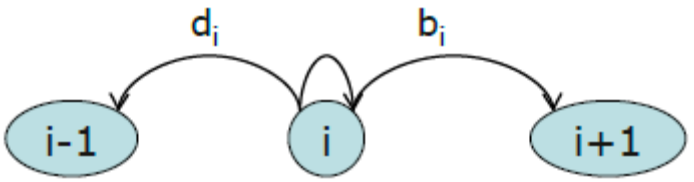
  for all $s \leq t$ and all states i, j.
  - We only consider time-homogenous systems in this course
- For such systems, every time the system enters a state i, it behaves as if it starts from that state at time 0
  - In particular, the holding time in state i is distributed in the same way any time the system enters state i.
  - Also, the states where it can possibly jump to from state i is always the same any time the system enters state i.
  - The probability of jumping from state i to state j is the same any time the system enters state i.

CTMC

- If we denote by p_{ij} the probability that a system in state i will enter state j at the next jump, then the resulting matrix yields a discrete time Markov chain (DTMC): the **embedded Markov chain**
  - In other words: if we forget the waiting time, then we obtain a DTMC

Example: birth/death processes

- We have i elements in the system: state i
  - Birth: we get one more, i.e., we move to state i +1
  - Death: one less, i.e., we move to state i-1
  - Birth is exponentially distributed with rate b_i
    - $p(X(s+t)=i+1 \mid X(s)=i) = b_i \exp(-b_i t)$
  - Death is also exponentially distributed with rate d_i
    - $p(X(s+t)=i-1 \mid X(s)=i) = d_i \exp(-d_i t)$
  - In state i we have two clocks
    - The one that rings first will determine the next state
    - The time when it rings determines the time of the next jump



Memory-less probability distributions

- X a continuous random variable with values in $[0, \infty)$
- The probability distribution of X is **memory-less** if for any $t, s \geq 0$, we have

$$P(X > t+s \mid X > t) = P(X > s)$$

- **Theorem.** The only memory-less (continuous) probability distribution is the exponential distribution: $P(X > t) = e^{-\lambda t}$, for some $\lambda > 0$.

- Note: similar concept and result hold for discrete probability distributions (the geometric distribution)

CTMC

- **Theorem.** Consider a time-homogenous CTMC. The holding time in any state is exponentially distributed.
- Proof:
  - Denote by T_i the holding time at state i, i.e., the time it takes until the system changes to a different state
  - Assume that the system has just entered state i. By time homogeneity this is the same as the system starting in state i.
  - Take $s,t \geq 0$. By definition of T_i, we have that the event "$T_i > s$" is the same as the event "$X(u)=i$, for all $0 \leq u \leq s$" (remark: by the above assumption $X(0)=i$).

Obvious
Markov
Time hom
Definition

$$P(T_i > s+t \mid T_i > s) = P(X(u) = i \text{ for } 0 \leq u \leq s+t \mid X(u) = i \text{ for } 0 \leq u \leq s)$$
$$= P(X(u) = i \text{ for } s < u \leq s+t \mid X(u) = i \text{ for } 0 \leq u \leq s)$$
$$= P(X(u) = i \text{ for } s < u \leq s+t \mid X(s) = i)$$
$$= P(X(u) = i \text{ for } 0 < u \leq t \mid X(0) = i)$$
$$= P(T_i > t).$$

  - In other words: the distribution of T_i is memory-less, i.e., it is exponential:

$$P(T_i > t) = e^{-\lambda t}, \text{ for some rate } \lambda > 0.$$

Stochastic biomodelling

- Discuss next the **chemical master equation**: the probabilistic counterpart of the mass-action principle

- Discuss **Gillespie's stochastic simulation algorithm** for the chemical master equation

The grand probability function

- $P(X_1,X_2,\ldots,X_n,t)$= the probability that at time t there are:
  - X_1 molecules of species S_1,
  - X_2 molecules of species S_2,
  - …,
  - X_n molecules of species S_n

  Remark: Not to be confused with the notation used for DTMC.
- Knowing this grand probability function, we may get for example:
  - the expected amount of molecules of species S_1 at time t:

$$E(X_1,t) = \sum_{X_1=0}^{\infty} \ldots \sum_{X_n=0}^{\infty} X_1 P(X_1,\ldots,X_n,t)$$

  - the standard deviation for the amount of molecules of species S_1 at time t: $(E(X_1^2,t)-E^2(X_1,t))^{1/2}$, where

$$E\left(X_1^{\,2},t\right) = \sum_{X_1=0}^{\infty} \ldots \sum_{X_n=0}^{\infty} X_1^{\,2} P(X_1,\ldots,X_n,t)$$

The chemical master equation approach

- The **chemical master equation** is describing the time evolution of the grand probability function
  - Write $P(X_1,\ldots,X_n,t+dt)$ as the sum of probabilities of all possible ways to be in state $(X_1,\ldots,X_n)$ at time t+dt, where **dt is infinitesimally small**
- We need a way to reason about the probabilities of various reactions to be triggered in the next infinitesimal time interval (t,t+dt)

Stochastic biochemical reactions

Consider as an example a reaction $S_1+S_2 \rightarrow S_3$

- Consider the probability that a *particular (not arbitrary!)* pair of molecules S_1-S_2 will collide in the next time interval of length *dt*
- Consider the molecules as balls of radius r_1, r_2
- Crucial assumption: *the system is well stirred and at thermal equilibrium*
  - In this case, the molecules are at all times randomly and uniformly distributed throughout the volume V of the system

Stochastic biochemical reactions

Consider as an example a reaction $S_1+S_2 \rightarrow S_3$

- Reason now about the average relative speed of that particular pair of molecules S_1-S_2 and the volume that one of them is spanning with that speed in the time interval $(t,t+dt)$
  - Collision probability: probability of the other molecule being in that volume
    - $P=V_{col}/V=\pi \cdot (r_1+r_2)^2 \cdot v_{12} \cdot dt/V$
    - For Maxwell-Boltzman velocity distributions: $v_{12}=(8 \cdot k \cdot T/(\pi \cdot m_{12}))^{1/2}$, where $m_{12}=m_1 \cdot m_2/(m_1+m_2)$ is the reduced mass and k is the Boltzman constant.
  - It follows that the probability of that particular pair of molecules reacting in the time interval $(t,t+dt)$ is $c \cdot dt$, where $c= \pi \cdot (r_1+r_2)^2 \cdot v_{12} /V$.
  - Consequently, since there are $X_1 \cdot X_2$ pairs, we have $X_1 \cdot X_2 \cdot c \cdot dt$ to be the probability that one such reaction will occur somewhere in the volume in the time interval $(t,t+dt)$.

Stochastic biochemical reactions

The fundamental hypothesis of the stochastic formulation of chemical kinetics:

the average probability that a particular combination of reactants will react according to a given reaction R in the time interval (t,t+dt) is $c_R \cdot dt$, for a certain constant c_R

- c_R is called the stochastic constant of the reaction R.
- The constant depends on the reaction (the properties of the reactants) and on the temperature of the system.

This is a reformulation of the principle of mass action!

- The probability of a reaction R taking place in the time interval *(t,t+dt)* is $N_R \cdot c_R \cdot dt$, where N_R is the number of all combinations of reactants in the current state
  - for a reaction $S_1+S_2 \rightarrow S_3$, $N_R=X_1 \cdot X_2$
  - for a reaction $2S_1 \rightarrow S_4$, $N_R=X_1 \cdot (X_1-1)/2$

Stochastic biochemical reactions

- The probability of a reaction R taking place in the time interval $(t,t+dt)$ is $\mathcal{N}_R \cdot c_R \cdot dt$, where N_R is the number of all combinations of reactants in the current state

- NOTE: Having an infinitesimally small time interval implies that at most one reaction takes place in that interval
  - In fact, the probability that two or more reactions take place in that interval is at least quadratic in dt, i.e., vanishingly small even after division by dt

Writing the chemical master equation

- m reactions $R_1, R_2, \ldots, R_m$ and n molecular species $S_1, S_2, \ldots, S_n$
- The chemical master equation:
  - Write $P(X_1, \ldots, X_n, t+dt)$ as the sum of probabilities of all possible ways to be in state $(X_1, \ldots X_n)$ at time t+dt, where dt is infinitesimally small

Writing the chemical master equation

- $P(X_1,\ldots,X_n,t+dt)$: probability of state $(X_1,\ldots,X_n)$ at time t+dt, dt infinitesimally small
- **Question:** how did we get to be in that state at t+dt?
- **Answer:** there are two possibilities
  1) we were in state $(X_1,\ldots,X_n)$ at time t and no reaction took place

     OR

  2) we arrived in state $(X_1,\ldots,X_n)$ as a result of one single reaction
     - for each reaction R_k, let $a_k dt$ be the probability of reaction R_k occurring in the interval *(t,t+dt)*, given the state $(X_1,\ldots,X_n)$ at time t
     - for each reaction R_k, let $B_k dt$ be the probability that reaction R_k occurs in (t,t+dt), resulting in the state $(X_1,\ldots,X_n)$

$$P(X_1,\ldots X_n, t+dt) = P(X_1,\ldots X_n, t)(1-\sum_{k=1}^{m} a_k dt) + \sum_{k=1}^{m} B_k dt$$

$$\Rightarrow P(X_1,\ldots X_n, t+dt) - P(X_1,\ldots X_n, t) = -\sum_{k=1}^{m} a_k P(X_1,\ldots X_n, t)dt + \sum_{k=1}^{m} B_k dt$$

$$\Rightarrow \frac{\partial}{\partial t} P(X_1,\ldots X_n, t) = \sum_{k=1}^{m} (B_k - a_k P(X_1,\ldots X_n, t))$$

Writing the CME – Examples

Example 1: A→

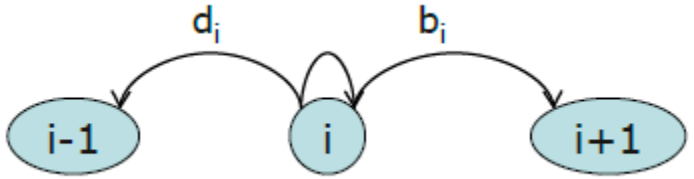
- Initial amount of A molecules: N_0
- Let c be the stochastic constant associated to this reaction
  - If we have m molecules, the probability of a reaction to take place in the next interval dt is c·m·dt. The probability that no reaction takes place is 1 − c·m·dt.
- $P(X_1,\ldots,X_n,t+dt)$ is the probability that we are in state $(X_1,\ldots,X_n)$ at time t+dt:
  - we were in state $(X_1,\ldots,X_n)$ at time t AND no reaction took place in *(t,t+dt)*, plus
  - the probability of having arrived in state $(X_1,\ldots,X_n)$ after one reaction occurred
- $P(n,t+dt) = P(n,t)(1-c\cdot n\cdot dt) + P(n+1,t)c\cdot(n+1)\cdot dt$, for $n < N_0$
  Note that $P(N_0,t+dt)=P(N_0,t)(1-c\cdot N_0\cdot dt)$
- $P(n,t+dt)-P(n,t) = -c\cdot n\cdot P(n,t)\cdot dt + P(n+1,t)c\cdot(n+1)\cdot dt$, for $n < N_0$
  $P(N_0,t+dt)-P(N_0,t)=-c\cdot N_0\cdot P(N_0,t)\cdot dt$
- $dP(n,t)/dt=c(-n\ P(n,t) + (n+1)\ P(n+1,t))$, for $n < N_0$
  $dP(N_0,t)/dt= -\ c\ N_0\ P(N_0,t)$, which can be solved: $P(N_0,t)= N_0\ e^{-c\ N0\ t}$

Writing the CME – Examples

Example 2: Birth-death process

- Consider the following two reactions
  - birth: $X \rightarrow 2X$
  - death: $X \rightarrow$
  - The stochastic rate constants for birth and death are λ and μ, respectively.
- Writing the CME:
  - $P(k,t+dt) = P(k,t)(1-\lambda\cdot k\cdot dt-\mu\cdot k\cdot dt) + P(k+1,t)\cdot \mu\cdot(k+1)\cdot dt + P(k-1,t)\cdot \lambda\cdot(k-1)\cdot dt$, for $k > 1$
    Note that $P(0,t+dt)=P(0,t) + P(1,t)\cdot\mu\cdot dt$
  - $P(k,t+dt)-P(k,t) = P(k,t)\cdot(-\lambda\cdot k-\mu\cdot k)\cdot dt + P(k+1,t)\cdot \mu\cdot(k+1)\cdot dt + P(k-1,t)\cdot \lambda\cdot(k-1)\cdot dt$, for $k > 1$
    $P(0,t+dt)-P(0,t) = P(1,t)\cdot\mu\cdot dt$
  - $dP(k,t)/dt = P(k,t)\cdot(-\lambda\cdot k-\mu\cdot k) + P(k+1,t)\cdot \mu\cdot(k+1) + P(k-1,t)\cdot \lambda\cdot(k-1)$
    $dP(0,t)/dt = P(1,t)\cdot\mu \Rightarrow P(0,t)=\mu\cdot\int P(1,\tau)d\tau$
- $d_i=\mu\cdot i$, $b_i=\lambda\cdot i$

From 0 to t



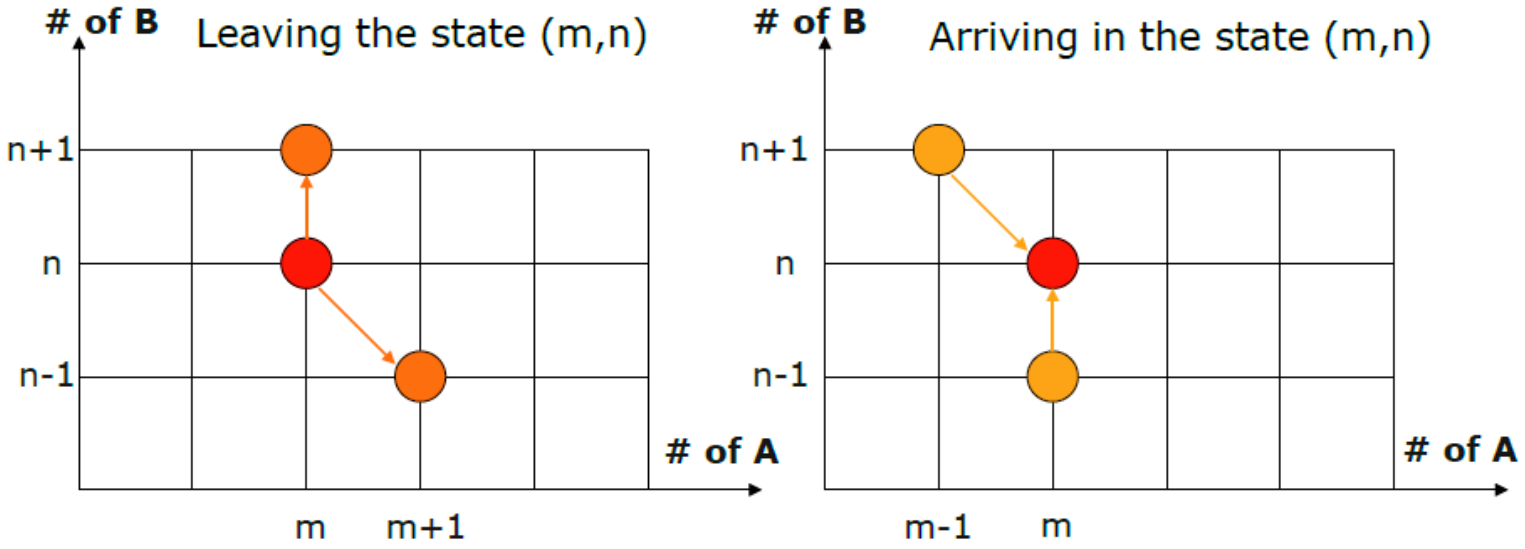
Kinetic constant, stochastic constant

(For a more careful discussion of the relationship we refer to Gillespie's 1976 paper)

- Let k be the kinetic constant of a reaction, c its stochastic constant
- Example: reaction $X+Y\rightarrow\ldots$
  - Kinetic reaction rate: $R=k\cdot[X]\cdot[Y]\cdot V$ ([X] and [Y] are the **concentrations** of X and Y, respectively, e.g., in mol/L. V is the volume of the reacting system, e.g., in litres)
  - Stochastic reaction rate: $R=c\cdot[x\text{-}y]$, where [x-y] is the number of X-Y pairs
  - by default we assume no correlations between X and Y and so, $[x\text{-}y]=[x]\cdot[y]=[X]\cdot[Y]\cdot V^2$ ([x] and [y] denote the **numbers** of molecules of X and Y, respectively).
  - In other words: $\mathbf{k = c\cdot V}$
- Example: reaction $X+X\rightarrow\ldots$
  - Kinetic reaction rate: $R=k\cdot[X]\cdot[X]\cdot V$
  - Stochastic reaction rate: $R=c\cdot[x\text{-}x]$, where [x-x] is the number of X-X pairs
  - $[x\text{-}x] = [x]\cdot([x] - 1)/2 \approx [x]^2/2=[X]^2\cdot V^2/2$
  - In other words: $\mathbf{k = c\cdot V/2}$
- Example: For reaction $X \rightarrow \ldots$ we have that $\mathbf{k = c}$ and for reaction $\varnothing \rightarrow X$ we have that $\mathbf{k = c/V}$.

Writing the CME – more examples

- Reactions:
  - A+B→A+2B, let c_1 be the stochastic constant for this reaction
  - B→A, let c_2 be the stochastic constant for this reaction



- Write the CME

Writing the CME – more examples

- The reactions
  - $A+B \rightarrow A+2B$, let c_1 be the stochastic constant for this reaction
  - $B \rightarrow A$, let c_2 be the stochastic constant for this reaction

- Writing the CME:
  - $P(m,n,t+dt)=P(m,n,t)\cdot(1-c_1\cdot m\cdot n\cdot dt-c_2\cdot n\cdot dt)+ P(m,n-1,t)\cdot c_1\cdot m\cdot(n-1)\cdot dt+ P(m-1,n+1,t)\cdot c_2\cdot(n+1)\cdot dt$
  - $P(m,n,t+dt)-P(m,n,t)= (-c_1\cdot m\cdot n-c_2\cdot n)\cdot P(m,n,t)\cdot dt+ P(m,n-1,t)\cdot c_1\cdot m\cdot(n-1)\cdot dt+ P(m-1,n+1,t)\cdot c_2\cdot(n+1)\cdot dt$
  - $dP(m,n,t)/dt= -(c_1\cdot m\cdot n + c_2\cdot n)\, P(m,n,t) + c_1\cdot m\cdot(n-1)\cdot P(m,n-1,t) + c_2\cdot(n+1)\cdot P(m-1,n+1,t)$

Writing the CME – more examples

- Consider the following reactions where M is an mRNA species and P is the corresponding protein species
  - mRNA production: $\rightarrow M$
  - mRNA degradation: $M \rightarrow$
  - protein synthesis: $M \rightarrow P$
  - protein degradation: $P \rightarrow$
- The stochastic constants associated to these 4 reactions are c_1, c_2, c_3, c_4, respectively
- Write the CME: each reaction contributes one positive term (gain) and one negative term (loss)

$$dP(m,p,t)/dt = -c_1P(m,p,t) - c_2mP(m,p,t) - c_3mP(m,p,t) - c_4pP(m,p,t)$$
$$+c_1P(m-1,p,t)+c_2(m+1)P(m+1,p,t)+c_3(m+1)P(m+1,p-1,t)$$
$$+c_4(p+1)P(m,p+1,t)$$

Writing the CME – more examples

- Reactions:
  - $A+B \rightarrow A+2B$, let c_1 be the stochastic constant for this reaction
  - $2B \rightarrow A$, let c_2 be the stochastic constant for this reaction

Solution

$$dP(n,m,t)/dt = P(n,m,t)\,(-c_1 \cdot n \cdot m - c_2 \cdot m \cdot (m-1)/2) + P(n,m-1,t) \cdot c_1 \cdot n \cdot (m-1) + P(n-1,m+2,t) \cdot c_2 \cdot (m+2) \cdot (m+1)/2$$

Observe that

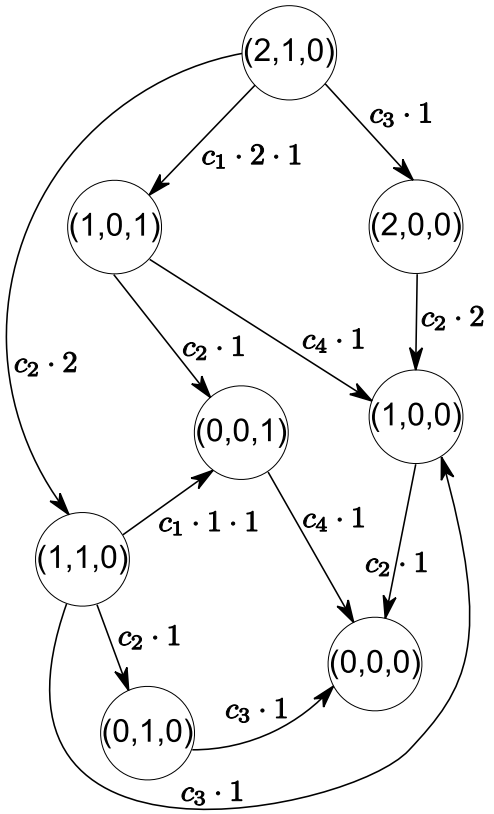
$$\binom{m}{2} = \frac{m!}{2! \cdot (m-2)!} = \frac{m \cdot (m-1)}{2}$$

$$\binom{m+2}{2} = \frac{(m+2)!}{2! \cdot (m+2-2)!} = \frac{(m+2) \cdot (m+1)}{2}$$

Underlying CTMC - example

Draw the graphical representation (in the form of a graph with transitions labelled with reaction rates) of the continuous-time Markov chain underlying the following system of reactions starting in the state (#A = 2, #B = 1, #C = 0):

1. $A + B \rightarrow C$ (stochastic rate constant: c_1)
2. $A \rightarrow$ (stochastic rate constant: c_2)
3. $B \rightarrow$ (stochastic rate constant: c_3)
4. $C \rightarrow$ (stochastic rate constant: c_4)
5. $C + B \rightarrow 2A$ (stochastic rate constant: c_5)



Explanation: In state (2,1,0) only reactions 1, 2, and 3 are enabled. In this state reaction 1 takes place with rate $c_1 \cdot 2 \cdot 1$, since there are 2 molecules of type A and 1 molecule of type B in the considered state. Similarly, the rates for reactions 2 and 3 are $c_2 \cdot 2$ and $c_3 \cdot 1$, respectively. Reactions 4 and 5 are not enabled, because there are no molecules of type C in state (2,1,0).