Ion PETRE

Research Fellow of Academy of Finland
Åbo Akademi University & TUCS, Turku, Finland

Tutorial on Computational Modeling in Systems Biology

Turku Centre for Computer Science (TUCS)
Computational Biomodeling Laboratory
http://combio.abo.fi/

Turku Centre for Computer Science

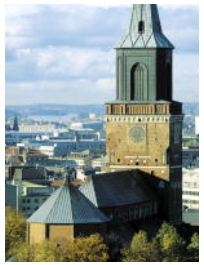
Tutorial on Computational Systems Biology

- Goal
  - give a crash course on molecular biology
  - give an account of several mathematical modeling techniques for sysbio:
    - differential equations,
    - stochastic processes (and Gillespie's algorithm),
    - petri nets,
    - process algebra
  - give a tutorial on a computer-based modeling environment: COPASI (Complex Pathway Simulator)
  - give a quick view on a sysbio project: the heat shock response
- Program
  - 14.30-15.15: A crash course on molecular biology
  - 15.15-15.45: Modeling with differential equations
  - 15.45-15.55 Break
  - 15.55-17.00 Modeling with stochastic processes. Gillespie's stochastic simulation algorithm
  - 17.00-17.30 The heat shock response

About Turku

- Turku
  - Former capital of Finland
  - Some 170000 people (5th largest in Finland, very close to 2nd)
  - Most beautiful archipelago in the world
  - Medieval castle
  - Cathedral
  - Fibonacci numbers
  - One of the best developed biotech sectors in Finland






Academic life in Turku

- Academic life in Turku
  - Three universities
  - We are with Åbo Akademi University, Department of IT
  - Turku Centre for Computer Science (TUCS): the graduate school in computer science, coordinating the PhD education in CS in all three universities
  - Located in the new ICT building, part of the Turku Technology Centre



My own research group

- The Computational Biomodeling Laboratory at Turku Centre for Computer Science and Åbo Akademi University
  - http://combio.abo.fi/
- Part of the Systems Biology national program of Academy of Finland
- Part of the Systems Biology research program of Turku Centre for Biotechnology
- We are a group of mathematicians and computer scientists
- Our projects are interdisciplinary, run in cooperation with biologists and biochemists from Finland and abroad
- Some recent projects
  - Gene assembly in ciliates
  - The hest shock response
  - The self-assembly of intermediary filaments

Ion PETRE

Research Fellow of Academy of Finland
Åbo Akademi University & TUCS, Turku, Finland

A crash course on molecular biology

Turku Centre for Computer Science (TUCS)
Computational Biomodeling Laboratory
http://combio.abo.fi/

Turku Centre for Computer Science

Life inside a cell

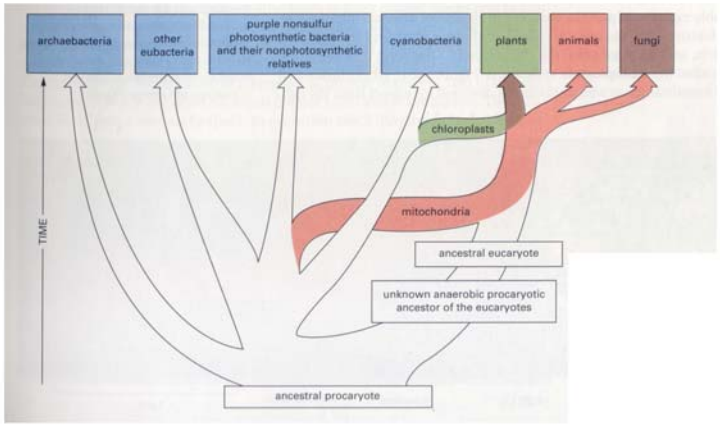
- Simplifications often made by biomodelers
  - Cell is "like a bag of chemicals floating in water"
  - Metabolites flow around chaotically
  - Metabolites are uniformly distributed
  - Proteins are just like balls (or cubes), DNA is just like a rope
  - In a DNA sequence, A is always matched with T, C always with G
  - Processes are isolated from each other and from the environment
  - ...
- The reality is surprisingly complex
  - The cell has a skeleton, gives it flexibility
  - Many intracellular boundaries, many specialized organelles
  - Highly specific metabolites
  - Very precise recognition of one's target
  - Energy efficiency optimized
  - Exquisite regulation, synchronization, signal propagation, cooperation
  - Some particles do move chaotically, but some others are transported
  - Some aspects are discrete (on/off), some others are continuous-like (always on, variable speed)
  - Huge pressure, crowded

A view on "The Inner Life of a Cell" (Harvard University, 2006):
http://aimediaserver.com/studiodaily/videoplayer/?src=harvard/harvard.swf&width=640&height=520
Beautiful representation of metabolite transportation, protein-protein binding, DNA replication, DNA ligase, microtubule formation/dissipation, protein synthesis, ...

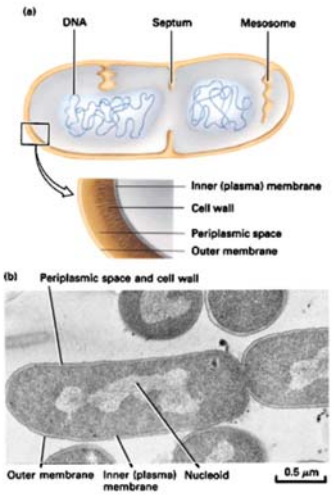
A crash course on Molecular Biology

- Cellular principles
  - Eukaryotes and prokaryotes; viruses and plasmids
  - Cellular tissues and colonies
  - Life cycle
  - Pathways
  - Energy
  - Individual interactions
  - Amplifications
  - Locality
- Biological macromolecules
  - DNA
  - Genes
  - Proteins
  - Enzymes
  - Chaperons
- Some of the slides are from Prof. Jyrki Heino (University of Turku)

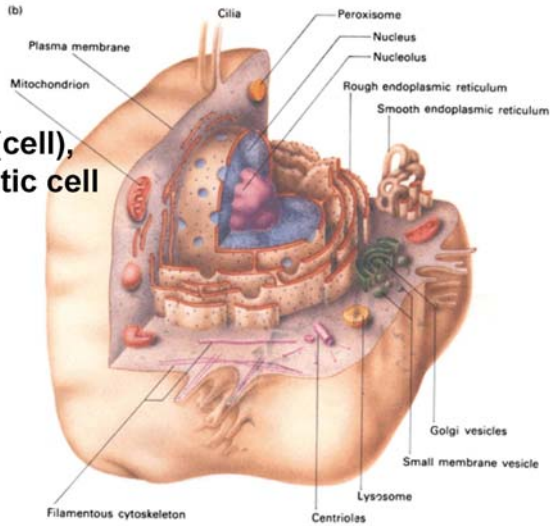
Evolution of cells from LUCA (last universal common ancestor) to modern cells



Bacterium, a prokaryotic cell



Animal (cell), a eukaryotic cell



Simpler things: viruses

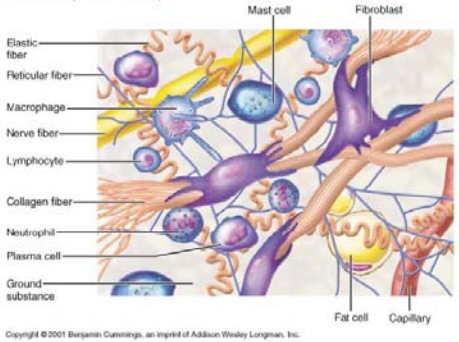
- Viruses are essentially just a protein coat hosting some DNA
  - In particular they do not have the machinery to replicate themselves
  - Well-studied example: lambda-phage
  - The protein coat attaches to the membrane of a cell and inserts the viral DNA into the cell
  - Once in, the viral DNA loops on itself forming a circular molecule
  - The cell's own transcription machinery will transcribe the viral DNA as if it were its own
  - In the case of the lambda-phage, the result is a protein called *lambda integrase* that inserts the viral DNA in the host's chromosomal DNA
  - The cell and all its descendants are from now on carriers of the viral DNA
  - Some external event may trigger the virus to become active: excise its DNA from the host's chromosome, multiply itself, create protein coats, assemble many copies of the virus, destroy the cell's membrane and release the new lambda phage to the intercellular environment

Plasmids

- There is nothing special about the viral DNA that makes the cell transcribe it as if it were its own
  - The same machinery will recognize any plasmid (circular DNA) and transcribe it as well
  - The basis for bioengineering (synthetic biology): encode into DNA the "instructions" and have the cell execute the code

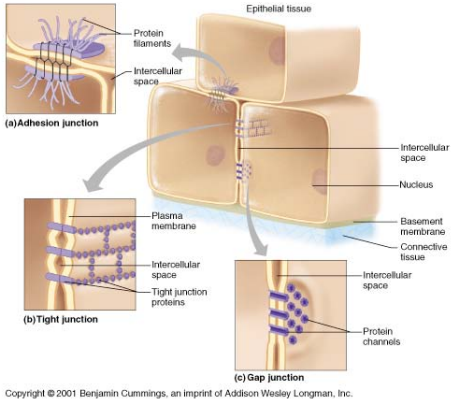
The extracellular environment

Cells form tissue: some tissues have only few cells and a lot of extracellular matrix (fibrillar proteins and macromolecules produced by the cells)



Copyright © 2001 Benjamin Cummings, an imprint of Addison Wesley Longman, Inc.

Some tissues are mainly formed by the cells that bind directly on each other



Copyright © 2001 Benjamin Cummings, an imprint of Addison Wesley Longman, Inc.

Cells cooperate

- At the molecular level, the cells in multicellular organisms are similar to unicellular organisms
  - Multicellular organisms have however specialized cells: they express a specific set of genes and perform only certain activities
    - Question: How can cells express only certain genes in the presence of exactly the same set of genes?
  - The inter-cellular communication is very important
    - Cells exchange signals (e.g., in the form of proteins), that are received by receptor sites on the plasma membrane
    - Signals may then be amplified and activate certain pathways

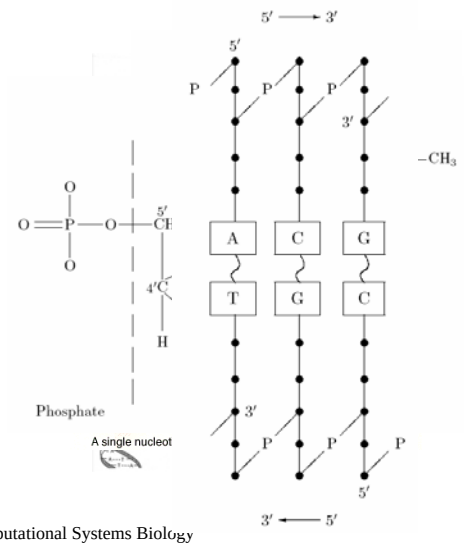
The intracellular environment

- The eukaryotic cell is a very crowded environment
- High pressure
- Many organelles, relatively little empty space (water)
- The cytoskeleton gives the cell
  - its shape
  - a degree of flexibility
  - ability to move
  - a “railway network” for protein transportation

Biological macromolecules

- Cells and organelles are formed by biological macromolecules
  - DNA is a (passive) storage of information
  - RNA are intermediates towards proteins, also role in regulation
  - Proteins are almost everything: building blocks, motors, regulators, enzymes, etc.
  - Lipids contribute to forming the membranes

- DNA
  - Nucleotide: consists of a deoxiribose sugar (5 atoms of carbon), a phosphate group and one of the four possible bases: adenine, cytosine, guanine, thymine
  - Phosphate attached to carbon 5, carbon 3 free for attachment
  - Single strands: sequences of nucleotides
  - Watson-Crick complementarity: A-T, C-G
  - Double strands: two single strands with complementary nucleotides bind together forming a double helix
  - Contains the blue print of the organism, each cell has a complete copy
  - Humans: some 3 billion base pairs in every single cell
  - DNA transcribed to RNA
  - RNA translated to proteins



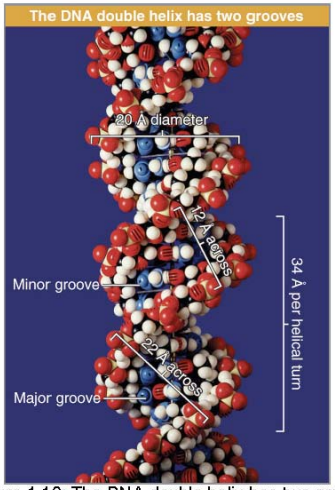
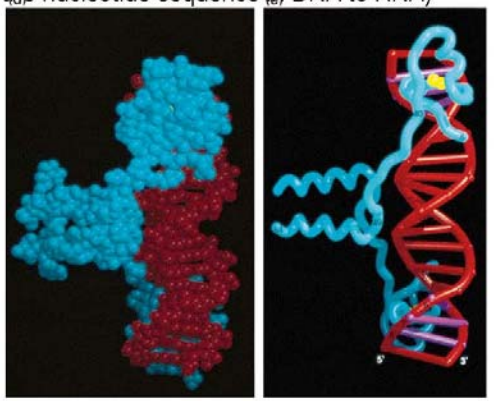


Photo © Photodisc

Figure 1.10: The DNA double helix has two grooves.

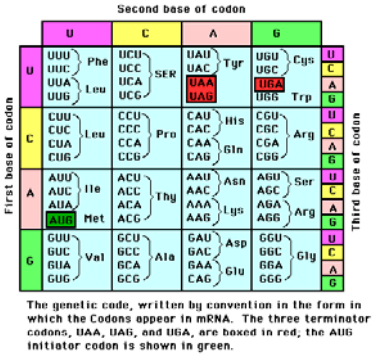
Transcription factors (proteins) recognize specific nucleotide sequences (when activated) in DNA and regulate reading of the genes (transcription of the nucleotide sequence in DNA to RNA)



Biological macromolecules

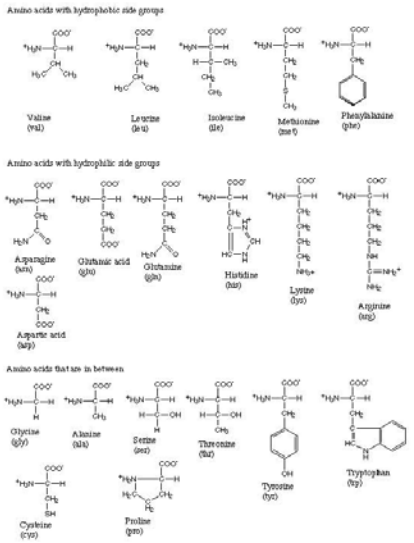
- Genes
  - DNA has coding blocks (genes) and non-coding blocks
  - Humans: some 20 000 – 30 000 genes (**in every cell!**)
  - Genes are transcribed into RNA that is then translated into proteins
  - RNA: similar structure as DNA, T replaced with U, mostly single stranded
  - Not all genes transcribed in all cells
  - Controllers: some non-coding blocks upstream of the gene – promoter regions
  - The RNA polymerase enzyme cannot bind to DNA on itself – helped by other enzymes that bind to the promoter region
  - Promoter region may be inhibited by other regions
  - **A robust computer science-like system: "if-then-else"**

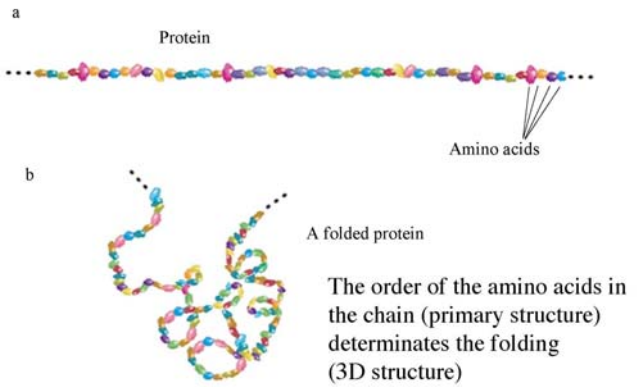
- Proteins
  - Sequences of amino-acids (20 possible)
  - Translated from RNA based on a universal code
  - 3 nucleotides (codon) code for one amino acid, some amino acids correspond to several codons
    - Only one start codon, 3 stop codons
  - Form a 3D fold – determines the function of the protein
  - The fold is determined by the sequence and the outside conditions
  - "Holy grail" of Bioinformatics: the protein folding problem – predict the 3D fold based on the (linear) amino acid sequence



The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

20 amino acids form all proteins





The order of the amino acids in the chain (primary structure) determinates the folding (3D structure)

Proteins

- Involved in molecular recognition
  - Recognize and bind to specific molecules (DNA, RNA, proteins).
  - In the case of DNA they may recognize a specific sequence of nucleotides, or even a specific pattern
- Their function depends on the 3D structure
  - May be turned active and inactive
  - Protein conformation may change after binding to other molecules
- Molecular motors
  - Protein may act as molecular motors through repeated changes in their 3D structure
  - Used for particle transportation or for cell locomotion
- Self-assembly
  - By binding to another protein, some new binding sites may be unveiled, for other proteins to bind, etc.

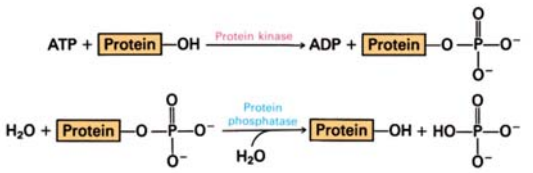
- Enzymes
  - Special type of proteins, specialize in recognizing very specific blocks of DNA (or protein) and binding to it
    - Some of them may then cut the DNA in a precise way, others may copy or repair DNA, etc.
  - Others may catalyze biochemical reactions, thus enabling reactions that would otherwise would be too slow
    - The speed-up may be of 3 orders of magnitude
  - They may be regulated by other enzymes, e.g., switched active/inactive
  - Crucial also in biotechnology

- Chaperons
  - Proteins assisting other proteins in achieving proper folding.
  - Many chaperones are heat shock proteins: proteins expressed in response to elevated temperatures.
    - Protein folding is severely affected by heat, and therefore chaperones act to counteract the potential damage.
  - Chaperones recognize unfolded proteins by the hydrophobic residues they expose to the solvent.
  - Incompletely folded proteins or misfolded proteins with exposed hydrophobic groups have a tendency to aggregate.
    - This aggregation is extremely detrimental to the cell: see Alzheimer's and Creutzfeld-Jacob's (human version of mad cow disease)
    - Chaperones help to prevent this by providing encapsulated hydrophobic environments that allow the protein to properly fold.

Protein phosphorylation

Protein kinases are enzymes that link a phosphate group into a protein

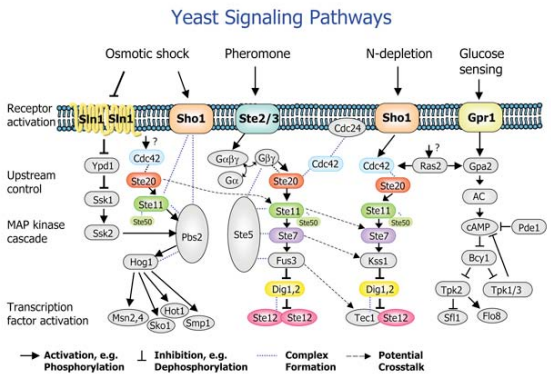
- Phosphate may change the 3D structure of a proteins or create a new binding site for other proteins



Apoptosis

- Apoptosis is the programmed cell death
  - Cells may also die a “violent” death, essentially exploding and hurting other neighboring cells
  - Through apoptosis cells commit to an “organized” death in which they step by step shut down their processes and eventually dissolve their membrane, releasing their intracellular content for other cells to reuse
  - Apoptosis may be triggered as a consequence of a series of events → the cell may trigger its own (organized) death if conditions are deemed that survival is impossible or at least detrimental for the whole organism
  - Apoptosis may also be triggered by signals from the extra-cellular environment

Signaling pathways



Amplification

- Sometimes an initial weak signal needs to be amplified
- The usual amplifying structure:
  - Initial (weak) signal activates a receptor protein
  - That receptor protein activates another protein
  - The receptor remains active and may activate many other proteins (as many as thousands per second)

Biological interactions are local

- Most effects of biological interactions are local
  - Unlike in computer science, viz. global variables, etc.
- Diffusion plays a role in spreading the effects of some local interactions
  - Some metabolites move chaotically, others are transported around the cell
- Membranes offer a physical implementation of locality
  - Their purpose is to give physical boundaries to the metabolites inside and to filtrate the inwards and outwards flux

Cells divide and multiply

- Prokaryotes: DNA is amplified, then attaches to different parts of the membrane and the cell divides
- Eukaryotes: more complex process because the DNA is organized on chromosomes
  - Cells must ensure that both daughter cells have the required number of chromosomes
  - In the case of sexual reproduction the process is even more complex, including a preliminary stage of exchanging haploid cells (only one copy of each chromosome)

Life inside a cell

A view on "The Inner Life of a Cell" (Harvard University, 2006):
http://aimediaserver.com/studiodaily/videoplayer/?src=harvard/harvard.swf&width=640&height=520
Beautiful representation of metabolite transportation, protein-protein binding, DNA replication, DNA ligase, microtubule formation/dissipation, protein synthesis, ...

Ion PETRE

Academy of Finland
Computational Biomodeling Laboratory, Turku, Finland

Modeling with differential equations

Turku Centre for Computer Science

Modeling

- What is a model?
  - A (partial) view of the reality
  - An abstraction of the reality
  - A representation of the (supposedly) main features of the reality, including the connections among them
  - For a given object of study, many models may be given, possibly focusing on different features of the object
- We focus in this tutorial on mathematical (and computational) models
  - Many other types of models exist
  - "Model" is indeed a very overloaded word
  - In this way, we also answer that a model is a *mathematical* representation of the reality
- What a model is not
  - A model is not the reality
  - A model is not certain!

Modeling

- Why mathematical modeling?
  - It allows for a precise formulation of the chosen aspects of the reality
  - It allows for a precise formulation of the current knowledge of the reality
  - It allows for precise reasoning about the reality
  - It allows for some types of analysis that would be impossible to perform on the reality
    - Model checking: verify all possible behaviors of the model in time
    - Scenario analysis: verify the behavior of the model in some well-defined extreme scenarios (e.g., disaster scenarios)
  - It allows for predictions

Modeling

- Model validation
  - Any model must always be subjected to experimental validation against the reality
  - A model may be invalidated by experimental data
  - No set of experimental data can confirm the "truthfulness" of a model

Modeling in science and engineering

- Great traditions of mathematical modeling in science and engineering
  - Physics
  - Chemistry
  - Engineering
  - Computer science
- Some mathematical modeling in biology also exists, but only expanding in scope in recent years

Modeling in biology

- Biology is in many ways transforming as a science
  - It has been by excellence an experimental science
    - Its "modus operandi"
      - hypothesis
      - experiment
      - data
      - suggestion for facts/principles/laws
    - Supported by many imaging techniques (some low-tech, some hi-tech)
  - In the last 20 years many technological advances
    - DNA microarrays, genome sequencers, mass spectrometry, advanced microscopy
    - All of these generate *numerical data*
    - Able to track inter-connections among many players, all contributing to the cell functions

Modeling in biology

- New modus operandi for biology
  - Hypothesis, inter-connections -> model
  - Experiment -> quantitative (as well as qualitative) data
  - Model fitting
  - Model analysis, predictions
  - Experiment -> testing the predictions
  - *Deducing* facts/principles/laws based on the model

Modeling approaches

- Mathematical models
  - Continuous vs. discrete mathematics
  - Deterministic vs. stochastic mathematics
- Computational (computer science) models
  - Boolean networks
  - Petri nets
  - Process calculi
  - Membrane systems
  - ...
- Hybrid models

Mathematical vs. computational models

- Mathematical modeling
  - The de facto standard in physics, chemistry, engineering
  - Basic paradigm
    - Identify the main actors, they become the (numerical) variables of the model
    - Identify the transfer function: it relates the numerical quantities to each other, expressing how they are to be updated based on the current values
    - Transfer functions may be composed yielding large, complex networks of inter-related variables
    - The end result: a mathematical object (equations) that can be numerically approximated (or solved analytically)
    - Quantitative models!
  - Several types of modeling approaches
- Computational modeling
  - Widely used in computer science
  - Basic paradigm
    - Identify the main actors, their possible (discrete) configurations make up the states of the model
    - Write a state machine that defines how, given certain events, the model changes state
    - State machines may be composed yielding complex reactive systems
    - The end result: an algorithm that can be executed
    - Most often qualitative models!
  - Several types of modeling approaches

Mathematical models

Type \ Time	*Continuous*	*Discrete*
Deterministic	Differential equations	Difference equations
Stochastic	Stochastic differential equations	Stochastic processes

Modeling with differential equations

- Modeling paradigm
  - The objects
    - the concentrations of all metabolites of interest
      - ❖ Do not consider the individual instances of each metabolite
      - ❖ Depending on the model, it may also be translated in terms of number molecules, by multiplying with the volume
    - the rates of all reactions
  - Main assumptions
    - The system is well-stirred
    - The system is at thermodynamical equilibrium

The law of mass action

- Waage, Guldberg 1864, Guldberg, Waage 1867, 1879
  - The reaction rate is proportional to the probability of a collision of the reactants
  - The probability of the collision is proportional to the concentration of reactants to the power of the molecularity
- Examples
  - For a reaction A->, the reaction rate is $v(t)=kA(t)$
  - For a reaction A+B→C, the reactions rate is $v(t)=kA(t)B(t)$, for some constant k
  - For a reaction A+B<->C, the reaction rate is $v(t)=k_+A(t)B(t) - k_-C(t)$, for some constants k_+, k_-
  - For a reaction 2A+3B<->4C+D, the reaction rate is $v(t)=k_+A^2(t)B^3(t) - k_-C^4(t)D(t)$, for some constants k_+, k_-

The differential equations

- The reaction rate gives the change per unit of time of the concentration of every metabolite involved in the reaction
  - For a consumed metabolite, the change will be $-v(t)$
  - For a produced metabolite, the change will be $v(t)$
- Example
  - For a reaction A->, the reaction rate is $v(t)=-kA(t)$
    - $dA/dt=-kA(T)$, solution $A(t)=A_0e^{-kt}$
  - For a reaction A+B→C, the reactions rate is $v(t)=kA(t)B(t)$, for some constant k
    - $dA/dt=-kA(t)B(t)$, $dB/dt=-kA(t)B(t)$, $dC/dt=kA(t)B(t)$

Coupled reactions

- Assume we have a set of reactions
  - A+B->C
  - A+2C<->B
  - C->2A
- Write the rates of all reactions
  - $V_1=k_1AB$
  - $V_2=k_2^+AC^2-k_2^-B$
  - $V_3=k_3C$
- Write the differentials: for each metabolite, consider all reactions where it participates
  - $dA/dt=-v1-v2+2v3=-k_1AB-k_2^+AC^2-k_2^-B+2k_3C$
  - $dB/dt=-v1+v2=-k_1AB+k_2^+AC^2-k_2^-B$
  - $dC/dt=v1-2v2-v3=k_1AB-2k_2^+AC^2+2k_2^-B-k_3C$

Coupled reactions

- The resulting system of differential equations may also be written in a matrix form:
  - **$dX/dt=Sv$**,
  - where X is the vector of m reactants, S is the (m x r)- *stoichiometric matrix* and v is the vector of r reaction fluxes
    - The (i,j) component of the *stoichiometric matrix* tells how the number of copies of the i-th reactant is changed as a result of the j-th reaction taking place
    - Writing v depends on the chosen modeling paradigm (e.g., mass action) and accounts for both directions of a reversible reaction
- Example: 2A->B, B->A, A+B->2B

- The stoichiometric matrix:

$$\begin{vmatrix} -2 & 1 & -1 \\ 1 & -1 & 1 \end{vmatrix}$$

- $X=(A\ B)^t$
- $v=(v_1\ v_2\ v_3)^t$,
  - Where $v_1=k_1A^2(t)$, $v_2=k_2B(t)$, $v_3=k_3A(t)B(t)$
- The system of differential equations is then $dX/dt=Sv$:
  - $dA/dt=-2k_1A^2(t)+k_2B(t)-k_3A(t)B(t)$
  - $dB/dt=k_1A^2(t)-k_2B(t)+k_3A(t)B(t)$

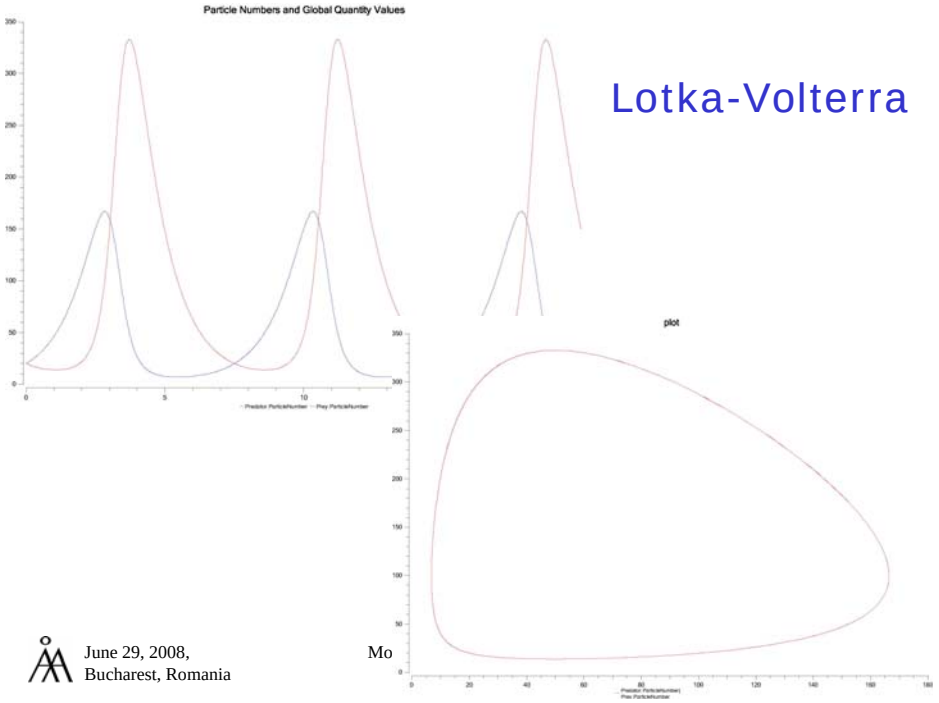
Differential equations

- Analytic solutions only for very simple equations (linear systems)
- For the other types
  - Numerical approximations of the solution, depending on the initial state
  - Analysis of the steady state: existence, uniqueness, stability
    - To compute the steady state, one must solve the algebraic system of equations where all differentials are equal to 0 and all unknowns are scalars (not functions of time)
    - This comes to solving the *algebraic equation $Sv=0$,* where S is the stoichiometric matrix

An example: the Lotka-Volterra model

- Two populations: predator (X) and prey (Y)
- An ecological system where the predator feeds on prey, multiplies when prey is available, and eventually dies
- The prey multiplies (food assumed to be always available) and is killed by the predator
- Many models exist. Here is one variant
  1. Consumption of prey: X+Y→X
  2. Growth of predators: X+Y→2*X+Y
  3. Growth of preys: Y→2*Y
  4. Death of predators: X→
- Mathematical model associated to it:
  - Kinetic rate constants k1, k2, k3, k4 corresponding to reactions 1-4 respectively
  - $dX/dt=k_2X(t)Y(t)-k_4X(t)$
  - $dY/dt=-k_1X(t)Y(t)+k_3Y(t)$

Lotka-Volterra



Michaelis-Menten kinetics

- Other modeling approaches than ***mass action*** exist
  - Michaelis-Menten
  - Hill
  - ...
- Michaelis-Menten kinetics have to do with the modeling of enzymatic reactions in some special conditions
  - E+S <-> E:S -> E+P
  - E is an enzyme
  - S is a substrate
  - P is a product

Enzymatic reactions

- E+S <-> E:S -> E+P
- Mass action formulation:
  1. $dS/dt=-k_1ES+k_{-1}(E:S)$
  2. $d(E:S)/dt=k_1ES-(k_{-1}+k_2)(E:S)$
  3. $dE/dt=-k_1ES+(k_{-1}+k_2)(E:S)$
  4. $dP/dt=k_2(E:S)$
- Briggs, Haldane 1925: in some conditions, it may be assumed that E:S reaches quickly a steady state
  - This is the case if S(0)>>E
  - Also if the binding of E and S is a much faster reaction than the production of P, k_1, k_{-1} >> k_2, Michaelis, Menten 1913
  5. $d(E:S)/dt=0$
- It follows from equations 2 and 3 that E+E:S is constant, say $E+E:S=E_{tot}$
- Then $E=E_{tot}-E:S$
- Steady state: $d(E:S)=0$:
  - $k_1(E_{tot}-E:S)S=(k_{-1}+k_2)(E:S)$
  - $E:S=(E_{tot}S)/(S+(k_{-1}+k_2)/k_1)$
- Thus, $dS/dt=-v_{max}S/(S+K_m)$, $dP/dt=v_{max}S/(S+K_m)$
  - Where v_{max} is the maximal rate (velocity) that can be obtained for reaction 2 (when the enzyme is completely saturated with substrate)
  - $vmax=k_2E_{tot}$
  - K_m is the Michaelis constant
  - $K_m=(k_{-1}+k_2)/k_1$, equal to the substrate concentration that yields the half-maximal reaction rate

TUCS

Ion PETRE

Academy of Finland
Computational Biomodeling Laboratory, ÅA

Modeling with stochastic processes

The chemical master equation approach
Gillespie's stochastic simulation algorithm

Turku Centre for Computer Science

TUCS

Modeling with differential equations: some physical difficulties

- Assumes that the time evolution of a chemically reacting system is both continuous and deterministic
- Difficulties with this assumption
  - the time evolution is **NOT** continuous: molecular population levels increase and decrease only with discrete amounts
  - the time evolution is **NOT** deterministic (even when ignoring the quantum effects and assuming classical mechanics for the molecular kinetics)
    - it is only deterministic in the full position-momentum phase space (knowing the positions and velocities of all molecules)
    - it is not deterministic in the N-dimensional space of the species population numbers
- However:
  - in many cases the time evolution of a chemical system can be treated as continuous and deterministic
  - the difficulties come when some species populations are small, or in conditions of chemical instability
  - Solution in these cases: stochastic models!

TUCS

Mathematical models

- Stochastic model
  - Given the current state of the system, many possible future behavior are possible
  - Probability distributions dictate the behavior of the system
  - Well-suited to model **individual**, rather than average behavior
  - Typical
    - Number of molecules are modeled
    - Reactions are taking place following "collisions" among the reactants
    - Markov processes
- Deterministic model
  - Given the current state of the system, all future behavior of the system is uniquely defined
  - Usually the model reflects the **average** behavior of the observed system
  - Typical methods used: differential or difference equations
  - Typical:
    - Concentrations of molecules are modeled
    - Reactions are taking place diffusion-like (gradient-like)
    - Differential equations

TUCS

Modeling with stochastic processes

- Modeling paradigm
  - The objects
    - **the number of copies of all metabolites of interest**
    - the rates of all reactions
  - Main assumptions
    - The system is well-stirred
    - The system is at thermodynamical equilibrium
  - Methods
    - Those of probability theory
    - Not part of "classical math": "only" about 200 years old
    - Some expertise from modeling in physics, especially in quantum physics
- *Versus differential equations*
  - *The objects*
    - *the concentrations of all metabolites of interest*
    - *the rates of all reactions*
  - *Main assumptions*
    - *The system is well-stirred*
    - *The system is at thermodynamical equilibrium*
  - *Methods*
    - *Those of mathematical analysis (continuous mathematics)*
    - *Arguably the most developed part of mathematics*
    - *Great expertise from modeling in physics, chemistry, engineering*

Writing the model

- Stochastic model
  - It is the description of a continuous time, discrete state Markov process
  - *Grand probability function*: $P(X_1,X_2,...,X_n,t)$ is the probability that at time t there are X_1 molecules of species S_1, …, X_n molecules of species S_n
  - The *grand probability function* may be obtained through a differential equation: the *chemical master equation*
    - Reason what is the probability of being in a certain state after one step
- *Versus differential equations*
  - *The reaction rate gives the amount with which the concentration of every metabolite involved in the reaction changes per unit of time*
    - *For a consumed metabolite, the change will be –v(t)*
    - *For a produced metabolite, the change will be v(t)*

The grand probability function

- $P(X_1,X_2,...,X_n,t)$= 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
- Knowing this grand probability function, we may get for example:
  - the expected amount of molecules of species S_1 at time t:

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

  - 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

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

The chemical master equation approach

- The chemical master equation is describing the time evolution of the grand probability function
  - Write $P(X_1,\dots X_n,t+dt)$ as the sum of probabilities of all possible ways to be in state $(X_1,\dots X_n)$ 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 interval (t,t+dt)

Stochastic 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 vanishingly small time interval dt
- Crucial assumption: *the system is well stirred and at thermal equilibrium*
  - as such, the molecules are at all times randomly and uniformly distributed throughout the volume
  - reason now about the average relative speed of that pair of molecules and the volume that one of them is spanning with that speed in the time interval (t,t+dt) and consider the probability of the other molecule being in that volume
    - $P=V_{col}/V=\pi(r_1+r_2)^2 v_{12} dt/V$
    - For Maxwell-Boltzman velocity distributions: $v_{12}=(8kT/\pi m_{12})^{1/2}$, where $m_{12}=m_1m_2/(m_1+m_2)$ is the reduced mass and k is the Boltzman constant
  - It follows that the probability of that particular pair of molecules reacting in the next infinitesimal time interval *(t,t+dt)* is *c·dt*
  - Consequently, since there are X1·X2 pairs, we have X1·X2·c·dt the probability that one such reaction will occur somewhere in the volume in the next infinitesimal time interval (t,t+dt)

Stochastic 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 next infinitesimal time interval dt is $c_R \cdot dt$, for a certain constant c_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 next infinitesimal time interval (t,t+dt) is $N_R \cdot c \cdot dt$, 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 $2S_1 \rightarrow S_4$, $N_R=X_1(X_1-1)/2$

Writing the chemical master equation

- Assume we have 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
  - Having an infinitesimally small time interval implies that at most one reaction takes place in that interval
  - $P(X_1,\ldots X_n,t+dt)$ is the probability that
    - we were in state $(X_1,\ldots X_n)$ at time t and no reaction took place, plus
    - the probability of having arrived in state $(X_1,\ldots X_n)$ after one reaction occurred
      - 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 $(X_1,\ldots X_n)$

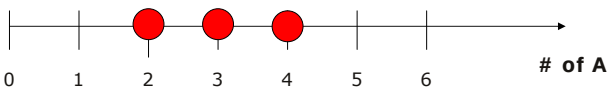
$$P(X_1,\ldots X_n, t+dt) = P(X_1,\ldots X_n, t)(1-\sum_{k=1}^{m} a_k dt) + \sum_{k=1}^{m} B_k dt$$

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

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

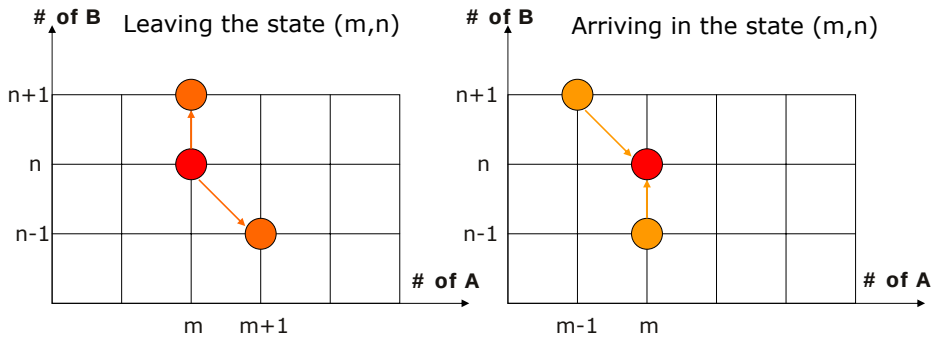
Example 1

- Examples
  - A->
  - Initial amount of A molecules: N_0
  - Let c be the stochastic constant associated to this reaction
  - $P(X_1,\ldots X_n,t+dt)$ is the probability that
    - 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
  - $P(n,t+dt) = P(n,t)(1-c \cdot n \cdot dt) + P(n+1,t)c \cdot (n+1) \cdot dt$
  - Note that $P(N_0,t+dt)=P(N_0,t)(1-c \cdot N_0 \cdot dt)$
  - $P(n,t+dt)-P(n,t) = -c \cdot n \cdot P(n,t) \cdot dt + P(n+1,t)c \cdot (n+1) \cdot dt$
  - $P(N_0,t+dt)-P(N_0,t)=-c \cdot N_0 \cdot P(N_0,t) \cdot dt$
  - $dP(n,t)/dt=c(-n\,P(n,t) + (n+1)\,P(n+1,t))$, for $n<N_0$
  - $dP(N_0,t)/dt= - k\,N_0\,P(N_0,t)$, which can be solved: $P(N_0,t)=e^{-kN_0 t}$



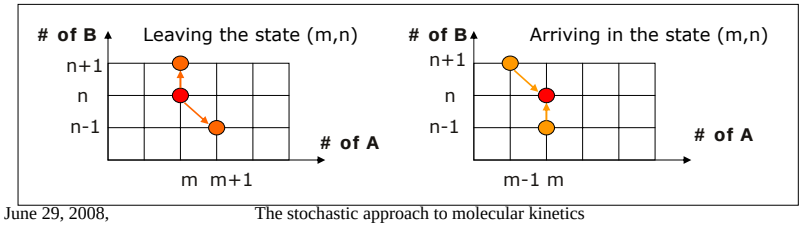
Example 2

- Example 2:
  - A+B->A+2B
  - B->A



Example 2 continued

- The reactions
  - A+B->A+2B, let k_1 be the stochastic constant for this reaction
  - B->A, let k_2 be the stochastic constant for this reaction
- Writing the CME:
  - $P(m,n,t+dt)=P(m,n,t)(1-k_1 mndt-k_2 ndt)+$
    $P(m,n-1,t)k_1 m(n-1)dt+$
    $P(m-1,n+1,t)k_2 (n+1)dt$
  - $P(m,n,t+dt)-P(m,n,t)= -(k_1 mn+k_2 n)\ P(m,n,t)dt+P(m,n-1,t)\ k_1 m(n-1)dt+P(m-1,n+1,t)k_2(n+1)dt$
  - $dP(m,n,t)/dt= -(k_1 mn + k_2 n)\ P(m,n,t) + k_1 m(n-1)\ P(m,n-1,t) + k_2\ (n+1)\ P(m-1,n+1,t)$



Example 3

- Consider the following reactions where M is an mRNA species and P is the corresponding protein species
  - mRNA production: → M
  - mRNA degradation: M→
  - protein synthesis: M→P
  - protein degradation: P→
- The stochastic constants associated to these 4 reactions are k_1, k_2, k_3, k_4, respectively
- Write the CME: each reaction contributes one positive term (gain) and one negative term (loss)

$dP(m,p,t)/dt=-k_1 P(m,p,t) - k_2 mP(m,p,t) - k_3 mP(m,p,t) - k_4 pP(m,p,t)$
$+k_1 P(m-1,p,t)+k_2(m+1)P(m+1,p,t)+k_3(m+1)P(m+1,p-1,t)+k_4(p+1)P(m,p+1,t)$

Towards numerical simulations

- The chemical master equation is exact and elegant
- Difficult to use it for numerical simulations
  - it can be analytically solved only for the simplest reactions
  - it describes the evolution of the probability of all states in time
    - it does not give directly the transitions from state to state
  - the differential equations for the time evolution of the molecular populations $X_i(t)$ may be written, but they involve the expected values of higher powers X_i^n and thus lead to infinite systems of ODEs
- Solution: Gillespie's stochastic simulation algorithm (SSA), 1976, 1977

Gillespie's SSA: foundations

- Assume we have m reactions $R_1, R_2, \ldots, R_m$ and n molecular species $S_1, S_2, \ldots, S_n$
- Given that the system is in state $(X_1,\ldots,X_n)$ at time t, we need to answer two questions in order to simulate the evolution of the system
  - when will the next reaction occur?
  - which reaction will it be?
- We combine the answers to these 2 questions in the following joint probability distribution:
  - $P(\tau,\mu)d\tau$ = the probability that, given the state $(X_1,\ldots,X_n)$ at time t, the next reaction will occur in the infinitesimal time interval $(t+\tau,t+\tau+d\tau)$ AND it will be reaction R_μ
  - note that if we thought about the probability of a reaction occurring exactly at time $t+\tau$, then the probability would be 0
- Strategy: based on CME, deduce the analytical expression of $P(\tau,\mu)$

Gillespie's SSA: foundations

- Given the state $(X_1,\dots,X_n)$ at time t, we need to compute
  - $P(\tau,\mu)d\tau$ = the probability that the next reaction will occur in the infinitesimal time interval $(t+\tau,t+\tau+d\tau)$ AND it will be reaction R_μ
- Let h_μ be the number of distinct combinations of reactants for reaction R_μ in the state at time $t+\tau$ (same as at time t!)
  - then, as observed for the CME, the probability that reaction R_μ will occur in the infinitesimal time interval $(t+\tau,t+\tau+d\tau)$ is $h_\mu c_\mu d\tau$
- Let $P_0(\tau)$ be the probability that no reaction occurs in the time interval $(t,t+\tau)$
  - Then $P(\tau,\mu)d\tau=P_0(\tau)\cdot h_\mu c_\mu d\tau$

Gillespie's SSA: foundations

- Given the state $(X_1,\dots,X_n)$ at time t, we need to compute
  - $P(\tau,\mu)d\tau$ = the probability that the next reaction will occur in the infinitesimal time interval $(t+\tau,t+\tau+d\tau)$ AND it will be reaction R_μ
  - $P(\tau,\mu)d\tau=P_0(\tau)\cdot h_\mu c_\mu d\tau$
- We need to compute $P_0(\tau)$, the probability that no reaction occurs in the time interval $(t,t+\tau)$
  - careful because the time interval $(t,t+\tau)$ may not necessarily be infinitesimal!
- Consider first $P_0(\tau+d\tau)$: no reaction occurs in the interval $(t,t+\tau+d\tau)$ if and only if no reaction occurs in $(t,t+\tau)$ AND no reaction occurs in the infinitesimal interval $(t+\tau,t+\tau+d\tau)$
  - Thus, $P_0(\tau+d\tau)=P_0(\tau)(1-\sum_\mu h_\mu c_\mu d\tau)$, i.e. $dP_0(\tau)/d\tau=-P_0(\tau)\sum_\mu h_\mu c_\mu$
  - It follows that $P_0(\tau)=\exp(-\sum_\mu h_\mu c_\mu \tau)$
- Finally, we obtain that

$$\mathbf{P(\tau,\mu) = h_\mu c_\mu e^{-\alpha\tau},}$$

for all $\tau\geq 0$ and $\mu=1,\dots,n$, where $\alpha=\sum_\mu h_\mu c_\mu$

Gillespie's SSA

- To simulate numerically the time evolution of our system starting from the given initial state:
  - Generate a pair (τ,μ) according to the probability density function $P(\tau,\mu)$
  - Adjust the molecular levels according to reaction $R\mu$ (decrease the level of reactants, increase the level of the output species)
  - Advance time to $t+\tau$
  - Iterate the procedure

Gillespie's SSA

- We need to generate a pair (τ,μ) according to the probability density function $P(\tau,\mu)=h_\mu c_\mu e^{\alpha\tau}$, where $\alpha=\sum_\mu h_\mu c_\mu \tau$
  - we first generate the time point τ such that the next reaction (any kind of reaction!) occurs in the infinitesimal time interval $(t+\tau,t+\tau+d\tau)$
  - the corresponding probability function is $P(\tau)=\sum_\mu P(\tau,\mu)=\alpha e^{-\alpha\tau}$
    - To do this, generate a random number r_1 in (0,1) and let τ_0 be such that $P(\tau<\tau_0)=r_1$:

$$P(\tau<\tau_0)=\int_{-\infty}^{\tau_0} P(\tau)d\tau = e^{-\alpha\tau_0}$$

    - Thus, $\tau_0=1/\alpha \ln(1/r1)$ is the time point we will consider
  - we then select the reaction $R\mu$ according to their relative probabilities of being triggered in the current step: $P(\mu)= P(\tau,\mu)/\sum_\nu P(\tau,\nu)= h_\mu c_\mu/\alpha$
    - To do this, generate a random number r_2 in (0,1) and let μ_0 be such that $P(\mu\leq\mu_0)=r_2$
    - We consider the distribution $F(m)=\sum_{i\leq m}P(i)$ and choose μ_0 such that $F(\mu_0-1)<r_2\leq F(\mu_0)$:

$$\frac{1}{\alpha}\sum_{\nu=1}^{\mu_0-1} h_\nu c_\nu < r_2 \leq \frac{1}{\alpha}\sum_{\nu=1}^{\mu_0} h_\nu c_\nu$$

TUCS

Gillespie's SSA: summary

- This is the only exact simulation algorithm of the chemical master equation
  - it is essentially just a reformulation of CME
  - the crucial point is that there is no time slicing (as in the numerical simulation of ODEs): jump to the next time point according to the correct probability distribution
- Many variants of Gillespie's SSA exist
  - some offer speedups
  - some are reformulations for various special cases, such as for hybrid models, involving both continuous and discrete variables

TUCS

The deterministic and the stochastic formulations: conclusions

- Deterministic approach
  1. based on the concept of diffusion-like reactions
  2. the time evolution of the system is a continuous, entirely predictable process
  3. governed by a set of ODEs
  4. The system of ODEs is often impossible to solve
  5. it models the average behavior of the system
  6. assumes that the system is well-stirred and at thermodynamical equilibrium
  7. conceptual difficulties when small populations are involved
  8. numerical simulations are straightforward and fast
  9. impossible to reason about individual runs rather than the average
- Stochastic approach
  1. based on the concept of reactive molecular collisions
  2. the time evolution of the system is a random-walk process through the possible states
  3. governed by a single differential equation: the chemical master equation
  4. the CME is often impossible to solve
  5. it models individual runs of the system
  6. assumes that the system is well-stirred and at thermodynamical equilibrium
  7. no difficulties with small populations
  8. numerical simulations via Gillespie's SSA are slow
  9. only gives individual runs; estimate the average through many runs

Ion PETRE

Academy of Finland
Computational Biomodeling Laboratory, ÅA

A mathematical model for the heat shock response in eukaryotes

Turku Centre for Computer Science

The group

- Senior members
  - IP (Turku, IT)
  - Ralph Back (Turku, IT)
  - John Eriksson (Turku, Biology)
  - Lea Sistonen (Turku, Biochemistry)
  - Andrey Mikhailov (Turku, Biochemistry)
- Graduate members
  - Claire Hyder (Turku, Biochemistry)
  - Andrzej Mizera (Turku, IT)
  - Diana Preoteasa (Helsinki, Mathematics)
- Undergraduates
  - Henrik Rönnholm (IT)
- Former members
  - Andreas Pada (IT)
  - Stefan Saxen (IT)
  - Kristian Nylund (IT)
  - Henry Ato Ogoe (IT)
  - Cristina Seceleanu (IT)

Outline

- The heat shock response: a new kinetic model
- Model building
- Parameter estimation
- Model validation
- Model analysis
- Predictions
- Perspectives

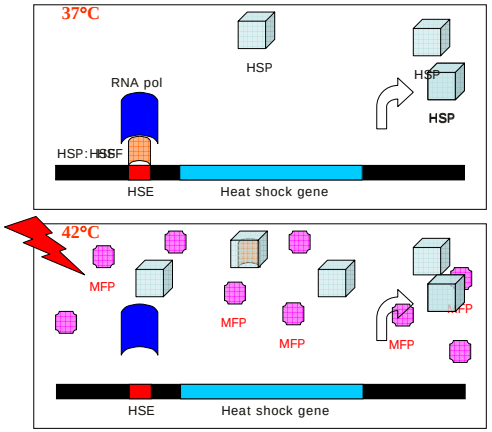
The heat shock response

- Cell's response to elevated temperatures
- Intense research on HSR in the last years
  - HSR is a very well-conserved regulatory network across all eukaryotes; bacteria have a similar mechanism
    - Good candidate for deciphering the engineering principles of regulatory networks
  - Heat shock proteins are very potent chaperones (sometimes called the "**master proteins**" of the cell)
    - Involved in a large number of regulatory processes
    - Also in anti-inflammatory processes
    - Found in extra-cellular environment, which may suggest they are used for signaling
    - Major role in the resilience of cancer cells; attractive as targets for cancer treatment
  - Tempting for a biomodeling, SysBio project because it involves relatively few main actors (at least in a first, simplified presentation)

Heat shock response: main actors

- Heat shock proteins (HSP)
  - Very potent chaperones
  - Main task: assist the refolding of misfolded proteins
  - Several types of them, we treat them all uniformly in our model with hsp70 as base denominator
- Heat shock elements (HSE)
  - Several copies found upstream of the HSP-encoding gene, used for the transactivation of the HSP-encoding genes
  - Treat uniformly all HSEs of all HSP-encoding genes
- Heat shock factors (HSF)
  - Proteins acting as transcription factors for the HSP-encoding gene
  - Trimerize, then bind to HSE to promote gene transcription
- Generic proteins
  - Consider them in two states: correctly folded and misfolded
  - Under elevated temperatures, proteins tend to misfold, exhibit their hydrophobic cores, form aggregates, lead eventually to cell death (see Alzheimer, vCJ, and other diseases)
- Various bonds between these metabolites

The molecular model for HSR



Our new molecular model

- **Transcription**

1. $HSF + HSF <-> HSF_2$
2. $HSF + HSF_2 <-> HSF_3$
3. $HSF_3 + HSE <-> HSF_3:HSE$
4. $HSF_3:HSE -> HSF_3:HSE + HSP$

- **Backregulation**

5. $HSP + HSF <-> HSP:HSF$
6. $HSP + HSF_2 -> HSP:HSF + HSF$
7. $HSP + HSF_3 -> HSP:HSF + 2HSF$
8. $HSP + HSF_3:HSE -> HSP:HSF + 2HSF + HSE$

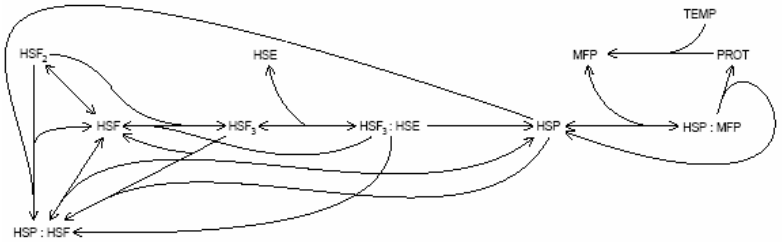
- **Response to stress**

9. $PROT -> MFP$
10. $HSP + MFP <-> HSP:MFP$
11. $HSP:MFP -> HSP + PROT$

- **Protein degradation**

12. $HSP \rightarrow 0$

The flux diagram of the model



The mathematical model

Table 1. The associated mathematical model

$$d[\mathrm{hsf}]/dt = -2k_1^+[\mathrm{hsf}]^2 + 2k_1^-[\mathrm{hsf}_2] - k_2^+[\mathrm{hsf}][\mathrm{hsf}_2] + k_2^-[\mathrm{hsf}_3] - k_5^+[\mathrm{hsf}][\mathrm{hsp}] + k_5^-[\mathrm{hsp}:\mathrm{hsf}] + k_6[\mathrm{hsf}_2][\mathrm{hsp}] + 2k_7[\mathrm{hsf}_3][\mathrm{hsp}] + 2k_8(\mathrm{hsf}_3:\mathrm{hse})\,\mathrm{hsp} \quad [13]$$

$$d[\mathrm{hsf}_2]/dt = k_1^+[\mathrm{hsf}]^2 - k_1^-[\mathrm{hsf}_2] - k_2^+[\mathrm{hsf}][\mathrm{hsf}_2] + k_2^-[\mathrm{hsf}_3] - k_6[\mathrm{hsf}_2][\mathrm{hsp}] \quad [14]$$

$$d[\mathrm{hsf}_3]/dt = k_2^+[\mathrm{hsf}][\mathrm{hsf}_2] - k_2^-[\mathrm{hsf}_3] - k_3^+[\mathrm{hsf}_3][\mathrm{hse}] + k_3^-[\mathrm{hsf}_3:\mathrm{hse}] - k_7[\mathrm{hsf}_3][\mathrm{hsp}] \quad [15]$$

$$d[\mathrm{hse}]/dt = -k_3^+[\mathrm{hsf}_3][\mathrm{hse}] + k_3^-[\mathrm{hsf}_3:\mathrm{hse}] + k_8[\mathrm{hsf}_3:\mathrm{hse}][\mathrm{hsp}] \quad [16]$$

$$d[\mathrm{hsf}_3:\mathrm{hse}]/dt = k_3^+[\mathrm{hsf}_3][\mathrm{hse}] - k_3^-[\mathrm{hsf}_3:\mathrm{hse}] - k_8[\mathrm{hsf}_3:\mathrm{hse}][\mathrm{hsp}] \quad [17]$$

$$d[\mathrm{hsp}]/dt = k_4[\mathrm{hsf}_3:\mathrm{hse}] - k_5^+[\mathrm{hsf}][\mathrm{hsp}] + k_5^-[\mathrm{hsp}:\mathrm{hsf}] - k_6[\mathrm{hsf}_2][\mathrm{hsp}] - k_7[\mathrm{hsf}_3][\mathrm{hsp}] - k_8[\mathrm{hsf}_3:\mathrm{hse}][\mathrm{hsp}] - k_{11}^+[\mathrm{hsp}][\mathrm{mfp}] + (k_{11}^- + k_{12})[\mathrm{hsp}:\mathrm{mfp}] - k_9[\mathrm{hsp}] \quad [18]$$

$$d[\mathrm{hsp}:\mathrm{hsf}]/dt = k_5^+[\mathrm{hsf}][\mathrm{hsp}] - k_5^-[\mathrm{hsp}:\mathrm{hsf}] + k_6[\mathrm{hsf}_2][\mathrm{hsp}] + k_7[\mathrm{hsf}_3][\mathrm{hsp}] + k_8[\mathrm{hsf}_3:\mathrm{hse}][\mathrm{hsp}] \quad [19]$$

$$d[\mathrm{mfp}]/dt = \phi(T)[\mathrm{prot}] - k_{11}^+[\mathrm{hsp}][\mathrm{mfp}] + k_{11}^-[\mathrm{hsp}:\mathrm{mfp}] \quad [20]$$

$$d[\mathrm{hsp}:\mathrm{mfp}]/dt = k_{11}^+[\mathrm{hsp}][\mathrm{mfp}] - (k_{11}^- + k_{12})[\mathrm{hsp}:\mathrm{mfp}] \quad [21]$$

$$d[\mathrm{prot}]/dt = -\phi(t)[\mathrm{prot}] + k_{12}[\mathrm{hsp}:\mathrm{mfp}] \quad [22]$$

The mathematical model

- Derived based on the *principle of mass action* (Guldberg, Waage, 1864, 1879)
  - *The flux of each reaction is proportional to the concentration of reactants*
  - Other principles exist, leading to different formulations of the model
- We consider a continuous formulation, based on ODEs
  - A stochastic formulation through CME also possible
- Given the reactions, writing the differential equations is easy
  - **dS/dt=Nv**, where S is the vector of m reactants, N is the (m x r)-stoichiometric matrix and v is the vector of r reaction fluxes
    - Writing v depends on the chosen modeling paradigm (e.g., mass action) and accounts for both directions of a reversible reaction
    - The (i,j) component of the stoichiometric matrix tells how the number of copies of the i-th reactant is changed as a result of the j-th reaction taking place

      Example: A+B->A+C: A's coefficient in this reaction is 0, B's is -1, C's is 1

Modeling of the heat-induced misfolding

- Question: how do we model the heat-induced misfolding?
  - What is the temperature-dependant protein misfolding rate per second?
- Adapted from Pepper et al (1997), based on studies of Lepock (1989, 1992) on differential calorimetry

$$\varphi(T) = (1 - 0.4/e^{T-37}) \times 0.00001448471257 \times 1.4^{T-37}$$

- Formula valid for temperatures between 37 and 45, gives a generic protein misfolding rate per second

Parameter estimation

- Data readily available for the goal: Kline, Morimoto (1997) – heat shock of HeLa cells at 42C for up to 4 hours, data on DNA binding (HSF_3:HSE)
- Requirements for the model:
  - 17 independent parameters, 10 initial values to estimate
  - 3 conservation relations available
  - The model must be in steady state at 37C, which gives 7 more algebraic equations (each of them quadratic)
  - Altogether: 17 independent values
  - Other conditions: total HSF somewhat low, refolding a fast reaction, HSPs long-lived proteins

A good modeling/simulation environment

- Our choice: **COPASI** (www.copasi.org)
  - Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). COPASI — a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067-74.
  - User-friendly
  - Stochastic and deterministic time course simulation
  - Steady state analysis
  - Metabolic control analysis
  - Mass conservation analysis
  - Optimization of arbitrary objective functions
  - SBML-based
  - **Excellent for parameter estimation**
  - FREE!

Parameter estimation

- Standard estimation procedure in COPASI (and not only)
  - Give the data and the target function
  - Give the list of parameters
  - The program scans the range of parameters and makes choices; for each choice it evaluates the target function against the experimental data (least mean squares)
    - The way it scans the space of parameter values depends on the chosen method
    - Many sophisticated methods currently available
    - All are local-optimization methods
  - It reports the best set of values
- Estimation repeated over and over again, with various methods for scanning the parameter space, to improve on the score of the fit

Parameter estimation

- Ideal approach:
  - Solve analytically the steady state equations at 37C
  - Use the solution to decrease the number of independent parameters and initial values
  - Do parameter estimation on the remaining independent variables to fit the model based on the data at 42C
- Problem: The steady state (37C) equations cannot be solved because they have degree 14 (overall)

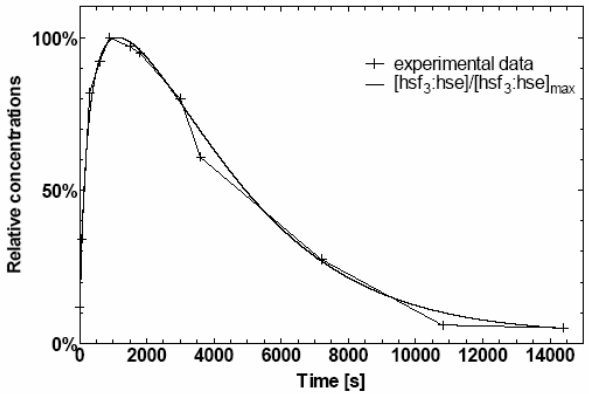
Parameter estimation

- Finding values for parameters and initial values so that the behavior at 42 is good is not difficult
- Problem: a good fit at 42C may not necessarily be in the steady state at 37C
  - Idea: Change the initial values so that we start in the steady state at 37
  - Outcome: the behavior at 42C is not satisfactory anymore
  - Idea: Iterate the procedure, estimating parameters and starting in the steady state
  - Outcome: the procedure does not converge to a good fit
  - Explanation: Changing the parameters will change the steady state, starting in the new steady state will modify the old behavior

Strategy for parameter estimation

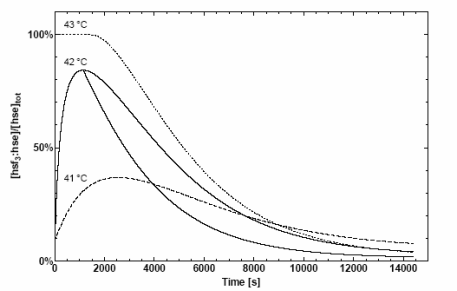
- Use the Kline-Morimoto experimental data to fit parameters
- *In the fit ask also that the fluctuations at 37C are (close to) 0*
  - Duplicate the model and run both at the same time (37&42C)
- The outcome of (countless rounds of) automated parameter estimation:
  - OK, but not good enough
  - The model is overfit: the HSR is shown to kick-in eventually even at 37C, albeit in a very mild form
  - Why?
  - Answer: we do not start close enough to the steady state!
- Idea
  - Set the initial values to be equal to the steady state values at 37C
    - We remain in the steady state at 37C
  - The difference is rather small in absolute terms, because the model was already fit to be close to the steady state at 37C
  - Test the behavior at 42C
- Result
  - **Excellent: agreement with the experimental data at 42C, steady state at 37C**

Parameter fit



Predictions and validation

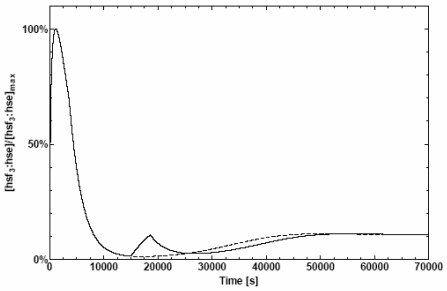
1. Higher the temperature, higher the response
2. Prolonged transcription at 43C confirmed
   - Unlike previous models
3. Heat shock removed at the peak of the response confirms a more rapid attenuation phase



All data is in relative terms with respect to the highest value in the graph so that it can be easily compared

Predictions and validation

- Experiment: two waves of heat shock, the second applied after the level of HSP has peaked
- Observation: the second heat shock response much milder than the first
  - The reason is that the cell is better prepared to deal with the second heat shock
  - Therapeutic consequences have been suggested: "train" the cell for heat shock by an initial milder heat shock
- The model prediction is in line with the experimental observation
  - Dotted line: heat shock at 42C for two hours, behavior followed up to 20 hours
  - Continuous line: heat shock at 42C for two hours, followed by a second wave of heat shock after the level of HSP has peaked



Skip

A-posteriori model validation

- We had a set of experiments done at Turku Biotech Centre
  - Idea: measure the levels of HSP at various time points, compare them with the model predictions
  - Difficult (expensive) to measure HSP levels directly
  - Cheap way around the problem: use an indicator for HSP
    - Use a cell line (human cancer cells) transfected with YFP-encoding genes, regulated by the same HSE
    - For each time point (15 of them from 0 up to 36 hours of heat shock at 42C), the fluorescence intensities of each of 10.000 cells are recorded
    - Do 3 biological repeats

Skip

The BTK data

- For each time point (up to 36 hours of heat shock at 42C), the fluorescence intensity of 10000 cell is recorded
  - That should give a measure of the number of YFPs in the cell (how?)
  - That in turn, should give an indication of the number of HSPs in the cell (how?)
- Assumptions:
  - The fluorescence intensity is proportional to the level of YFPs
    - not many aggregates!
  - The level of synthesized YFP is proportional to the level of synthesized HSP
    - Same regulation mechanism
- [HSP] is a prediction of the model
- Test it against the measurements on fluorescence intensity
- **HOW?**

Validation against the BTK data

- **Idea:** we test the hypothesis that the stochastic variable

  YFP(t)-E(YFP(0))

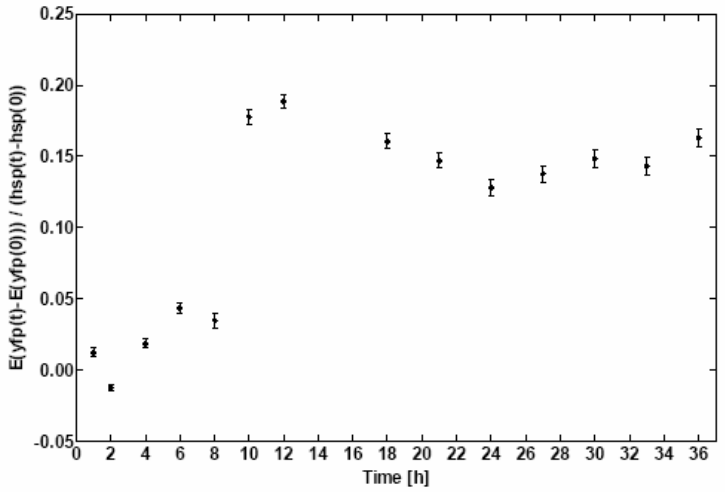
  is "proportional" to

  HSP(t)-HSP(0)
- Formally, we test if the stochastic variables

  (YFP(t)-E(YFP(0))) / (HSP(t)-HSP(0))

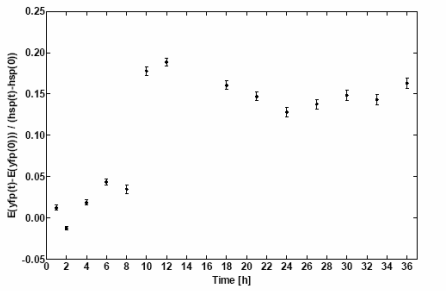
  have the same expected value for all 15 time points t
  - The standard way of estimating the expected value of a stochastic variable based on a sample is through a confidence interval
  - Approach:
    - Take 95% confidence intervals for the expected value of each of the 15 stochastic variables
    - See if they have a non-empty intersection.

95% confidence intervals



How do we read the results

- The intersection of the confidence intervals is surely **empty**
- Still, they suggest that the expected values may assume "somewhat" constant values, with a different regime on the intervals 0-8 hours and 8-36 hours
- **WHY?**
  - **Short answer:** we do not know!
  - **Suggestions:**
    - higher death rate for non-responsive cells
    - longer half-life for YFP under sustained heat shock (consistent with previous reports that under sustained heat shock, the cell "shuts down" all "unnecessary" processes)
    - *YFP is a poor quantitative reporter!!*



HSR and a system-based approach to cancer

- Previous research proposed HSP as a drug target in cancer treatment.
  - **B.Vastag (2006) Nature Biotechnology**
  - **Idea:** lower the level of HSP so that the cancerous cell cannot cope with the high level of MFP and becomes apoptotic
  - **Difficulty:** the observation was that even if the level of HSP is artificially lowered, even more HSP is being produced eventually
    - HSR kicks in

HSR and a system-based approach to cancer

- **Better idea:** inhibit the whole heat shock response in cancer cells
  - Approach: find which reactions in our network influences the most the level of MFP and target those reaction.
  - Careful that the effects are only triggered for the excited cells and not for the normal cells: only cancer cells would be lead to apoptosis
- We can select the most suitable reactions by comparing the scaled sensitivity coefficients at 37C and 42C
- Our model suggest 3 different possible reactions to target
  - The result is supported by intuition: make the bond HSP:HSF more stable, or the bond HSP:MFP more unstable
  - Not tested in the lab yet
- More complex analysis may be performed in terms of multi-dimensional sensitivity analysis