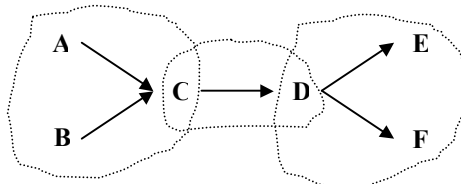


# Identifiability of a simple biological system

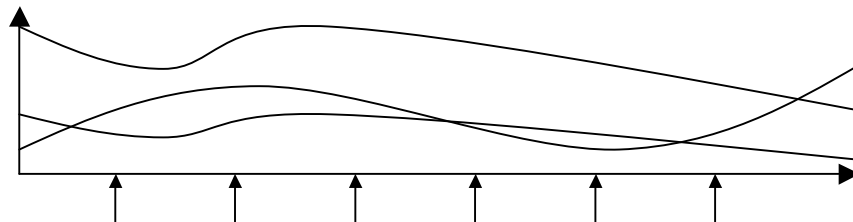
Jotun Hein 1.1.08

**Motivation.** High throughput technologies allow the observation of the genome, expression levels and more of biological systems. Systems biology attempts to infer properties of the underlying systems that normally aren't observable in full. This project explores simple models (that allows simulations to be performed) and then investigates the ability to infer the underlying system.

- *Quantity of data.* This can be varied by observing the system at very many, very dense time points, starting from different initial conditions and possibly different sets of variables could be observed.
- *Quality of data.* Each observation of a variable is observed with noise. This can be added in many ways: independent random variables added to each observation, the noise could be correlated over time and among variables, the time of a variable could be observed with noise, the noise could correlate with the size of the variable observed, there could be bias in the noise. We will for simplicity assume that we know when a variable was observed and that each noise is  $N(0, \sigma^2)$  and all are independent.
- *Biological knowledge of structure of the system.* Defining "the system" is far from straight forward and in often assumptions about what is relevant and what is not, is tacitly introduced by the biologist. It will typically be a set of  $k$  variables and we will presently assume that this is correct. The dynamics of many biochemical systems are described by mass action equations. Often all terms are 0<sup>th</sup>, 1<sup>st</sup> or 2<sup>nd</sup> order. 0<sup>th</sup> order corresponds to constant addition or removal, 1<sup>st</sup> order corresponds to constant decay or growth of a variable and 2<sup>nd</sup> order corresponds to creation/destruction by collision of molecules. Knowledge of the system will often enter as which molecules interacts/reacts with which (or not) and can be described as a network. Knowledge could also state which reactions are present/absent. Or it could enter in a more global fashion a constraint that the reaction graph must be connected. Such knowledge will constrain, which equations will be considered as legal description of the system. An alternative to deterministic constraints, would be the use of a Bayesian prior.
- Distinguishing *two alternative models* of the same system. This can involve both equation structure and continuous parameters. The more similar models, the more data will be needed to distinguish them.
- *Knowledge of a closely related system.* Often several related biological systems are being studied and they should thus be modelled in combination. From a practical viewpoint this allows transfer of knowledge from system **A** to system **B** if they are closely related. For simplicity, we will assume that we are studying system **A**, but have perfect knowledge of system **B** and additionally knows how far away it is evolutionary (t).



$$\frac{dA}{dt} = \frac{dB}{dt} = -k_{A,B}[A][B], \frac{dC}{dt} = k_{A,B}[A][B] - k_C[C], \frac{dD}{dt} = k_C[C] - k_D[D], \frac{dE}{dt} = \frac{dF}{dt} = k_D[D]$$



The graph describes the reaction graph between 6 components. A and B will create C, that decays in D that splits into E and F. A-F are the nodes of the graph. Each reaction defines a hyperedge. There are three hyperedges each encircled by dashed lines A,B→C; C→D and C→E,F. The law of mass action defines the dynamics of the system as described the equation system. The system could be modified to incorporate product inhibition, addition of A, B and removal of E, F from the system. Some of the components (here three) can be observed at different times {vertical arrows} – fictional curves.

The **dynamic model** representing the biological system is a set of equations. Biochemical systems without spatial structure are often described with a special class of ordinary differential equations called mass action equations and we will only consider

these here. Such equations can be illustrated by a directed hypergraph, where the nodes are labelled and represent molecules. It is a hypergraph because each edge can involve more than two nodes. The in-set will create the out-set in the reaction represented by the edge. In principle in- and out-sets can be multi-sets as reactants and products could be used/created in multiple copies. There will be real world restrictions for the graphs useful in biology as pairs of molecules collide and create one product or a single molecule decomposes into two products. In other we need rarely consider edge involving more than 3 nodes. It is possible that two nodes in in-set or out-set are identical. These considerations shows that enumeration of mass action equation can be done by enumerating the corresponding graphs.

The *evolutionary model* describing the evolution of one model into another is standard and simple. I.e. we know the full equation system for **B**. This will define a prior on parameters describing **A**:  $p_A = p_B + X$ , where  $X \sim N(0, t)$  and  $p$  is a parameter value. The structure of the dynamical system can evolve by adding/deletion components. When components are added, then decisions must be made about how they interact and how the complete equations systems must be modified. Again, we strive toward simplicity. If  $k$  is the index of the new variable, then it will interact an already existing variable with probability  $p$ . The interaction will be creation of a new product with rate  $r$  ( $\sim N(r_0, \sigma_r^2)$ ), where this normal distribution is the equilibrium distribution of the Ornstein-Uhlenbeck process according to which the continuous parameters in the model evolve. Brownian motion confined to a  $k$ -dimensional cube would also give a useable equilibrium distribution. Clearly, simple chemical considerations (like mass conservations) can rule out many reactions.

In silico *experiments* and *observations* can be made by generating a trajectory from the dynamic system and allowing key variables to be observed with noise.

### Method of inference.

**Curve fitting:** A problem is that trajectories are observed with noise and this might lead to use a complicated model in an attempt to over-fit towards the data. This can be compensated by giving a penalty for the use of extra parameters. Two obvious choices are: Mean Least Square distance  $\sum_i [x(t_i) - obs(t_i)]^2$  and MLS with parameter penalty:

$\sum_i [x(t_i) - obs(t_i)]^2 + n$ . Being clever about the search algorithm is clearly crucial given the size of the search space. It will

involve searching equation structures and then fitting continuous parameters. Several approaches should be tried and we will just sketch one here. Equation Structure Search – greedy recursive search: The simplest equation set has zeros on the right hand side corresponding to no change. Best fit to this would correspond to find constant functions fitting the observations the best. To this equation system can be added a term of the form  $k_{X,Y}[X][Y]$  or  $k_X[X]$ . If there were  $n$  components, that would give  $(n+1)n^2/2 + n^2$  choices with 1 continuous parameter to fit. From real applications many can be ignored as they would postulate impossible reactions. This will be essential in making a realistic algorithm. In the simulation scenario, one approach of this could be to predefine a set of possible reactions to be a fraction  $f$  of the possible set, where  $f$  could be .05 - .001.

Parameter Fitting: There are a variety of methods for doing this (Press et al., 2007). Most methods use some sort of local search for local optimum and it is essential to have a good starting guess to avoid finding a local optimum. The greedy equation search algorithm adds parameters one by one, so in each new step can use the previous estimates of all but the new parameter as starting points, almost reducing the parameter search problem to a 1-dimensional problem.

The greedy algorithm adds one component at a time. If computational power permits, it would be possible to add 2 or more at time, resulting in a more exhaustive search.

**Bayesian Inference:** Prior distributions will have to be defined on the dynamical system, its parameters and the noise with which variables are observed. Only parameters in mass action equations have not been discussed and a normal distribution would here be reasonable. We are now interested in the posteriors on the dynamical systems after observing a trajectory. Doing a Bayesian analysis would bring major benefits: It would quantify the value of evolutionary information, the value of biological constraint and would allow testing alternative hypotheses of the same data.

### Project Plan

1. Count and recursively generate mass action equations for up to 10 components.
2. Chose 5 systems of increasing complexity and write simulator that can generate dynamic trajectories from these.
3. Infer the continuous parameter from the system from perfect observations and given knowledge of the equation structure.
4. Same as 3, but no knowledge of equation structure.
5. Same as 3, but with noise added to each observation. Then also add noise to the time the observation has been made.
6. Same as 3, but only add partial knowledge of the equation system.
7. Same as 3, but assume only a subset of the variables have been observed.
8. If time permits: Assume a related system is perfectly known and it is  $t$  time units away from the one under observation.
9. If time permits: Bayesian Analysis

### Comments.

- i. This project will take good abilities in programming and data analysis to do the project well. The project is very self-contained and would need very little reading to start on.
- ii. A series of projects addressing exactly the same questions for more complex systems and involving real data (written with Dagmar Iber) can be found at <http://mathgen.stats.ox.ac.uk/bioinformatics/projects/> and could be suited for collaborative projects between several DTC students.
- iii. There are 3 other major classes of networks in biology and the analogous models could be pursued for these as well: signal transduction, protein interaction networks and regulatory networks. Finally, different types of networks can interact and should be modelling in combination.

### References

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery (2007) "Numerical Recipes 3rd Edition: The Art of Scientific Computing"

### **Further comments.**

i. The number of equation structures can be counted by simple combinatorics: If there is  $k$  components, whose concentrations are  $x_i(t)$  at time  $t$ . In the equation describing the change in  $x_i(t)$ , there can be zero order, first order and second order terms. For each of these a binary choice must be made. In the equation for  $x_i(t)$  there are 1 possible zero order term. There are  $k$  possible 1<sup>st</sup> order terms. Second order terms can be any subset of size 3 including  $i$ . Additionally, it must be decided if there are 2 *in-nodes*/1 *out-node* or 1 *in-node*/2 *out-nodes*. Subsets of size 3 with different elements is  $k(k-1)(k-2)/6$ , with 2 identical elements  $k(k-1)$ , with 3 identical elements  $k$ . The total number of second order reactions is  $k(k-1)(k-2)/3 + 2k(k-1) + k$  and total number comes to  $k(k-1)(k-2)/3 + 2k(k-1) + k + k + 1 = k(k-1)(k-2)/3 + k^2 + 1$ .

ii. As described above there are no biological restrictions on which reactions are possible and which not. Even without biological knowledge it could be relevant to restrict the number of reactions allowed in the system. Three levels of biological restrictions could be imposed: Firstly, conservations of atom numbers. This seems an absolute requirement that could be not deviate from. Secondly, single chemical reactions cannot be too complex and one could forbid changes in molecules that needed more than one transfer of a sub-molecule. Lastly, one could only have allows reactions that has been observed in some organism.

iii. Illustration of evolution on the basic example. There are two aspects: continuous change of the parameters and change in the structure of the equations system. In the basic example, we had three parameters ( $k_{AB}$ ,  $k_C$ ,  $k_D$ ) that could evolve according the Ornstein-Uhlenbeck process or Brownian Motion in a box. The structure of the equation system would change by toggling the existence/non-existence of terms in the equation system. For a system of 6 variables this would come to 117 binary choices. The number of neighbor equations systems that would differ by one reaction would be 117, the number of systems differing by 2 would be  $117 * 116 / 2$  etc. To delete an existing term is straightforward, but if a new term came into existence, then one would have to choose the appropriate parameter. It would here be natural to choose in the equilibrium distribution determined by the chosen evolutionary model.

iv. In the project description we have very few references. This is not because there aren't any, but rather that there are a flurry of approaches and many of the questions in the project seems to be new, like an evolutionary model of equations. The following text will refer to some of the relevant literature in systems identification.

### **More references:**

- Craciun et al. (2007) "Toric Dynamical Systems"
- Gatermann (2002) "A family of Sparse Polynomial Systems arising in Chemical Reaction Systems" J. Symbolic Computation 33:275-305.
- Ljung, L (1999) "Systems Identification : Theory for the User" Prentice-Hall
- Ljung, L (2007) "Systems Identification Toolbox 7"
- Srinivasan, A. et al. (2008) "Incremental Identification of Qualitative Models of Biological Systems using inductive Logic Programming" IEEE Transactions on Automatic Control (2005) vol. 50.10 special issue on systems identification.
- Todorovski and Dzeroski (2007) "Integrating Domain knowledge in Equation Discovery" LNAI 4660 69-97.
- Vysheirsky and Girolami (2007) "Bayesian Ranking of Biochemical System Models" Bioinformatics