

MCMC Integration over Evolutionary Histories of Metabolic Networks

4.5.08

Supervisors: Jotun Hein and Tom Snijders

Goal: To devise methods that could explore stochastic models that could calculate the likelihood of two homologous metabolic pathways by summing over all evolutionary histories that could relate them. Such a method would open the way for important data analysis, that is presently not possible or at least done by flawed methods: Parameter estimation in the evolutionary process of metabolic pathways, testing of hypotheses about the evolution of metabolic pathways and making statements about the path of evolution (ancestral analysis).

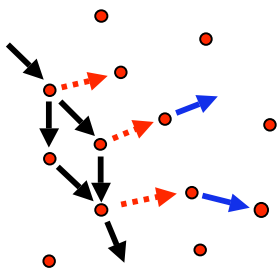
Motivation: Comparative Biology has recently experienced a major boom in the form of comparative genomics, where the interpretation of a genome is strongly augmented by the comparison with other genomes. This trend is now moving into networks (Sharan and Ideker, 2006), but comparison is ubiquitous in biology and as data and models accumulate at higher biological levels and structures, the application of evolutionary modelling will expand accordingly.

Background and basic concepts: There are key classes of networks in biology: Metabolic, Regulatory, Signalling and Protein Interaction Networks. The concept of network is so general (set of objects with pairwise relationships) that it will appear everywhere in science. The different kinds of network in biology describe different kinds of dynamics and the reasonable model of evolution will vary from class to class. Mathematical modelling of networks is a well-explored area (Dorogovtsev and Mendes, 2003), but evolutionary models are only now being undertaken (Wiuf et al., 2006). A network will typically have a discrete component – nodes and edges - and possibly a continuous component that labels edges (for instance flux) or nodes (for instance concentration). The discrete components are associated different kinds of models. For instance, metabolic pathways will have reactions (edges) deleted/added. Protein Interaction Networks (PIN) will have nodes duplicated and simultaneously the edges of the duplicated node copied as well.

We will focus on metabolic pathways. It is probably not ideal to represent metabolic pathways by a standard graph, as some of the basic events are: **A** and **B** collide and creates **C** or the reverse, that **C** decomposes into **A** and **B**. These are not standard edges as they involve three nodes (the first two input nodes and one output node) and thus a hyper-edge. Hyper-edges can be decomposed into standard edge (above a new node (**N**) could be defined that **A** and **B** both pointed to, and **N** could point to **C**). We will assume that a metabolism can be represented by a standard graph. We will also assume that the process of evolution is Markovian. A useful addition assumption in most evolutionary models is that the process is time reversible, ie that $P(N_1)P(N_1 \rightarrow N_2) = P(N_2)P(N_2 \rightarrow N_1)$. This is useful as this allows one network to be viewed as the ancestor of the other, while they in reality have descended from a common ancestor. This should be checked for models investigated.

To calculate $P(N_1, N_2)$, two quantities have to be calculated: $P(N_1)$ and $P(N_1 \rightarrow N_2)$. In principle both these quantities can be calculated by exponentiation: Let **Q** be the rate matrix for the continuous time Markov process on networks. $P(N_1 \rightarrow N_2)$ can be found as an entry in $P(t)$, that can then be found as $e^{tQ} = I + tQ + t^2Q^2/2 + t^3Q^3/3! + \dots$. $P(N_1)$ can be found as entry in $P(\text{infinity})$, where in column N_1 all entries will be identical to this quantity. If the set of nodes were fixed and the pathway evolved by independent turning off and on of reactions (edges), these calculations would be easy as the probability for the complete pathways would be the product of what happened to individual edges. However, pathways are subject to constraints that involve many edges simultaneously thus undermining independence. One example of such a constraint could be if the graph had to be connected. Exponentiation could in principle still solve the problem, but would involve square matrices with dimensions of cardinality equal the number of possible metabolisms, which grows prohibitively fast in the number of nodes.

A model on metabolisms represented by an undirected graph with a core metabolism that must be present in all organisms, could be illustrated as follows.



$$\frac{dP(M)}{dt} = \lambda \sum_{M' \in D(M)} P(M') + \mu \sum_{M'' \in A(M)} P(M'') - P(M)[\lambda|A(M)| + \mu|D(M)|]$$

<i>n</i>	Number of all graphs with <i>n</i> nodes	Number of states
1	1	1
2	2	2
3	8	8
4	64	61
5	1024	969
6	32768	31738
7	2097152	2069964
8	268435456	267270033
9	68719476736	68629753641
10	35184372088832	35171000942698

To the left is a metabolism, black edges constitute the core, red edges present in this metabolism. The metabolism evolves by addition and deletion of edges subject to the constraint that the metabolism remains connected. The blue edges are addable and the red are deletable. If edges are inserted with rate λ and deleted with rate μ , then metabolisms will obey the middle equation that balances the rate that the metabolism can be created by insertions and deletions minus the rate with which it could be destroyed by insertions and deletions. On the right is tabulated the number of graphs and the number of connected graphs as function of nodes.

An example of an evolutionary history of a toy metabolism could be:



We only observed the start and end points of this sequence and all possible paths from the first to the second should be explored. In a probabilistic setting this would mean summed over (both time points and events). Additionally, the probability of start point should also be calculated. This in principle implies summing over all paths from an arbitrary starting point arbitrary far back in time.

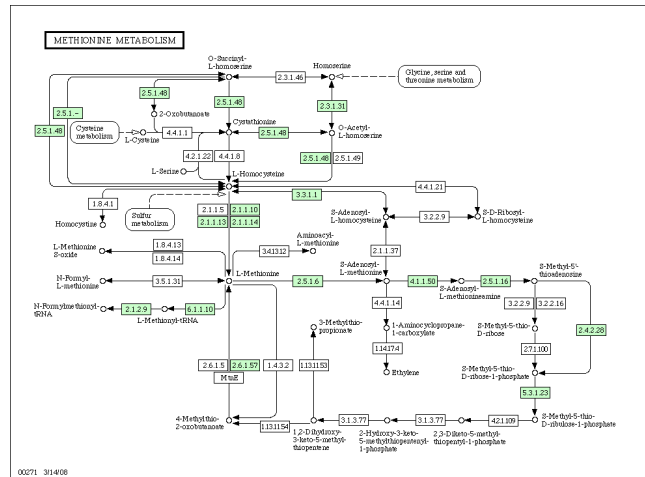
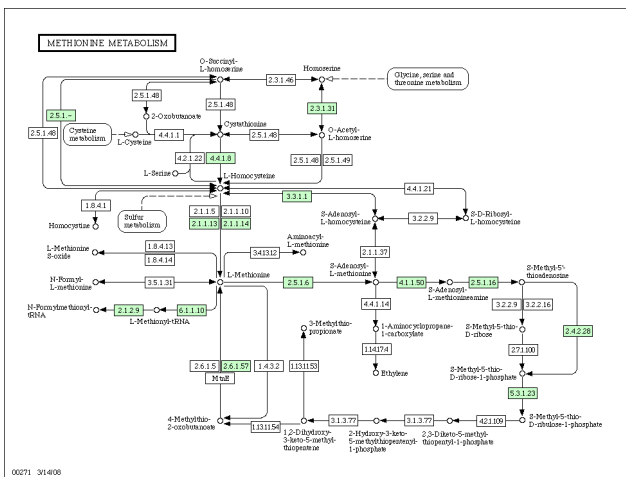
Earlier work (Giannoulatou and Mithani) used exponentiation to analyze transition probabilities for networks up to 6 nodes, corner cutting allowed this to be extended to 10-11 nodes dependent on μT and λT . Mithani implemented MCMC that could integrate over pathways from N_1 to N_2 satisfactory for up to 20 nodes. The real valued waiting times, can be integrated out using the method of for instance Miklos et al. (2003), so the MCMC can operate on discrete valued paths. The models investigated were simple either assuming total independence among reactions or obeying the constraints that non-isolated nodes constitute a connected set. The problem of calculating $P(N)$ under some model has not been considered in this project, but should at some point. When calculating $P(N_1 \rightarrow N_2)$ by exponentiation, then all entries are calculated, but it could be advantageous to focus on the entry of question - $P_{ij}(N_1 \rightarrow N_2)$. At present there seems to have been no publications on stochastic models for metabolism change.

Proposed Research:

1. An MCMC procedure has to be set up for simulation of a drastically simplified model. This can be done following some of the ideas in Snijders, Koskinen, and Schweinberger (2008). A candidate for such a simplified model is one in which a digraph is given of "all" possible metabolic reactions, nodes being reactions and arcs being compounds. The evolution allows only changes between subgraphs of this graph, where a subset of nodes is taken out together with all their connections. The likelihood function can be based on some simplified fitness function. The algorithm will be based on sequential deletion and insertion of nodes.
2. Next an algorithm has to be developed and programmed for parameter estimation, either Bayesian (cf. Koskinen & Snijders 2007) or frequentist (cf. Snijders et al. 2007)
3. Finally proposals of model formulations must be made for models that are less drastically simplified, and still obeying basic biochemical rules such as conservation of atoms. MCMC procedures for these models fall outside the scope of this 2-month project.

Work Plan:

- Read key articles (Eleni Giannoulatou's DTC report; chapter of 5 of Aziz Mithani's transfer report, Sharon and Ideker; Tanaka et al.,...) and expand these 2 pages into a 5 page detailed work plan.
- Propose more realistic models and simulate to investigate their behaviour. Examples of realism could be: Allowing a few connected components. Introducing a fitness function that was dependent on the metabolic capability, which would lead to selective trend toward larger networks. It could be possible to turn such selection on and off.
- Analyze these two homologous versions of methionine metabolism below. It could here be natural to investigate a model, that created a small set of relevant metabolites (nodes) and only edges/hyperedges (reactions) was considered that had been observed in some specie. This would reduce the edge set substantially.



Consider the Methionine metabolism pathway map from KEGG consisting of a total of 41 reactions represented by boxes connecting 77 different metabolites represented by small circles (only 30 shown in the figure). To the right we find Methionine Metabolism in *P. fluorescens PfO-1* and to the left Methionine Metabolism in *P. syringae pv. tomato DC3000*. Out of these 41 reactions, 14 are reversible reactions. The methionine metabolism in *Pseudomonas syringae pv. tomato DC3000* and *P. fluorescens PfO-1* where the active reactions are shown in green. Out of the total 41 reactions, only 14 reactions (4 reversible) are known to be present in *P. syringae pv. tomato DC3000* whereas *P. fluorescens* methionine metabolism is known to consist of 19 reactions (7 reversible) with 13 reactions in common.

Comments:

- i. This project is demanding and would need good programming skills and knowledge of MCMC, basic combinatorics and probability theory.
- ii. This project could be expanded to a full PhD devoted to MCMC on biological networks and extended from two observed networks to k networks.
- iii. Besides Hein and Snijders, Mithani and Reinert are also interested in this area.

References:

- Babu et al. (2005) "Structure and Evolution of Transcriptional Regulatory Networks" *Curr.Opin.* 14.283-91
- S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks, From Biological Nets to the Internet and WWW*, Oxford University Press (2003).
- Giannoulatou's DTC report on <http://mathgen.stats.ox.ac.uk/bioinformatics/projects/>
- P.J. Ingram, M.P.H. Stumpf, J. Stark, *Network motifs: structure does not determine function*, *BMC Genomics* 7, 108 (2006).
- S. A. Kauffman (1969) Metabolic stability and epigenesis in randomly constructed genetic nets *J. Theoretical Biology, Volume 22, Issue 3.437-467*
- Koskinen, J. (2004) Bayesian Inference for Longitudinal Social Networks. Research Report, number 2004:4, Stockholm University, Department of Statistics.
- Koskinen, J. and Snijders, T. (2007) Bayesian inference for dynamic social network data, *Journal of Statistical Planning and Inference*, 137, 3930-3938.
- Lynch, M. (2007) "The Evolution of Genetic Networks by Non-Adaptive Processes" *Nature Genetics* 8.803-
- R. Sharan, T. Ideker, *Modeling cellular machinery through biological network comparison*, *Nature Biotechnology*, 24, 427 (2006).
- I. Miklos, G.A. Lunter and I. Holmes (2004) A "long indel" model for evolutionary sequence alignment. *Mol. Biol. Evol.* 21(3):529-540. Appendix A
- Quayle and Bullock (2006) "Modelling the evolution of genetic regulatory networks" *J.Theor.Biol.* 238.737-753.
- R. Sharan, T. Ideker, *Modeling cellular machinery through biological network comparison*, *Nature Biotechnology*, 24, 427 (2006).
- T. Shlomi et al. (2007) [A genome scale computational study of the interplay between transcriptional regulation and metabolism](#). *Molecular Systems Biology (MSB)*,
- Torsten Reil: Dynamics of Gene Expression in an Artificial Genome - Implications for Biological and Artificial Ontogeny. *ECAL 1999: 457-466*
- Snijders, T. (2001) "Statistical evaluation of social networks dynamics" in *Sociological Methodology* By Michael Sobel
- Snijders, T. et al. (2008) "Maximum Likelihood Evaluation for Social Network Dynamics" In press
- R.Somogyi & CA Sniegoski (1996) Modelling the Complexity of Genetic Networks *Complexity* 1.6.45-64.
- Soyer, Pfeiffer and Bonhoeffer (2006) "Simulating the Evolution of Signal Transduction Networks" *J. Theor. Biol.* 241.223-232.
- Tanaka, Ikeo and Gojobori (2006) "Evolution of metabolic networks by gain and loss of enzymatic reaction eukaryotes" *Gene* 365.88-94.
- C. Wiuf, M. Brameier, O. Hagberg, M.P.H. Stumpf, (2006) *A likelihood approach to analysis of network data*, *PNAS*, 103.20.7566-70
- Chen-Hsiang Yeang and Martin Vingron, "A joint model of regulatory and metabolic networks" (2006). *BMC Bioinformatics*. 7, pp. 332-33.

More information in the two metabolisms to be analyzed:

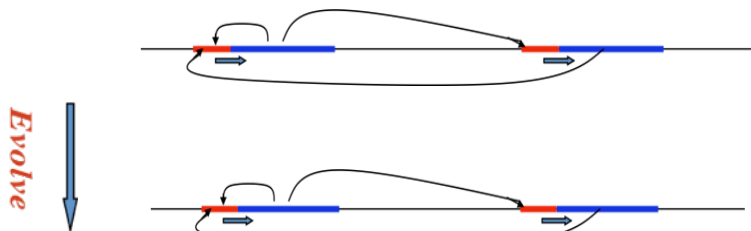
http://www.kegg.com/dbget-bin/get_pathway?org_name=pfo&mapno=00271 <http://www.kegg.com/dbget-bin/get_pathway?org_name=pfo&mapno=00271>

http://www.kegg.com/dbget-bin/get_pathway?org_name=pst&mapno=00271 <http://www.kegg.com/dbget-bin/get_pathway?org_name=pst&mapno=00271>

Further Comments: There are 3 other major classes of biological networks: signal transduction (SD), protein interaction networks (PINs) and transcription regulatory networks (TRN). Each class needs its own model of evolution and inference from data.

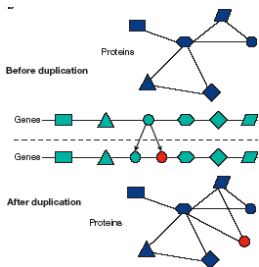
Evolution of these Networks:

TRN: The first models for the TRP was Boolean Networks as described in Kauffman (1969). Reil (1999) introduced a model call Artificial Genome (AG) where proteins could interact with promoters and promoters determined the expression of proteins. Both Kauffman and Reil investigated the dynamics of the resulting pathways. Quayle and Bullock (2006) has an explicit model based on an AG where proteins interact with promoters and promotor evolution.



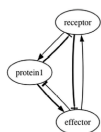
Promoters can appear-disappear. Protein (TF) and signal interactions can change and the rules from the status of the TFBS to transcription can change.

PIN: These has been modelled by duplication and attachment. There is a set of proteins and a pair can stick to each other or not. When a gene for a protein duplicates the two identical proteins will attach to the same proteins, but over time they can lose or gain attachments.



A gene can duplicate creating two identical copies that naturally will have the same neighbor relations as the mother gene. The edge can also change over time.

SD: Soyer, Pfeiffer and Bonhoeffer (2006) have a simple model, where the number of proteins didn't evolve, but the strength of their interactions did. The quantity of interest was the how it the network transmitted a signal.



$$\frac{d[P_i]}{dt} = [P_i^* \sum_j l_{ij}[P_j^*] - [P_j](\delta_{il}[L] \sum_j k_{ij}[P_j^*])]$$

One protein is receptor and one is effector. All proteins can switch between activated (*) and not activated and the dynamics of is described by the equation above. Parameters of the equation can change and proteins can be gained/lost at get new relationships to other proteins. The performance of the network is evaluated by the fraction of effectors activated as function of receptor activated.

Network Inference will make statements about the architecture and parameters of a network based on the observables that it describes. Network inference goes back 60-70 years for MP and originally belonged to pure biochemistry. Since late 1960s the first dynamical models for RN was presented, but it is only in the last decade that investigations of how to infer networks from observations of dynamical paths has been investigated seriously - clearly motivated by the rise of relevant types of high throughput data.

Integration of Networks: Since enzymes are switched on/off by regulatory networks and genes are regulated by the presence/absence of metabolites, it is natural to attempt an integrated description of these two classes of networks. Yeang and Vingron (2006) and Shlomi et al. (2007) have both considered this.