MaPhySto
The Danish National Research Foundation:
Network in Mathematical Physics and Stochastics

Miscellanea
no. 26 March 2004

Marianne Huebner and Michael Sørensen (eds.):
Workshop on Dynamical Stochastic Modeling in Biology

ISSN 1398-5957

Workshop on

Dynamical Stochastic Modeling in Biology

MaPhySto - Centre for Mathematical Physics and Stochastics

DYNSTOCH - Statistical Methods for Dynamical Stochastic Models

8 – 10 January 2003

Organized by

Marianne Huebner (Michigan State University)

and

Michael Sørensen (University of Copenhagen)

Introduction

The workshop "Dynamical Stochastic Modeling in Biology" was held in the days 8 – 10 January 2003 at the Department of Applied Mathematics and Statistics, University of Copenhagen. It was organized jointly by the Centre for Mathematical Physics and Stochastics funded by the Danish National Research Foundation and by the research training network DYNSTOCH under the Human Potential Programme (contract no. HPRN-CT-2000-00100).

The main aim of the workshop was to discuss dynamical stochastic models for biological problems and to identify new areas where such models might be useful. The emphasis of the workshop was on ecology, gene regulatory networks and topics in bioinformatics.

This booklet contains extended abstracts of the talks given at the workshop followed by the list of participants.

Extended Abstracts

List of extended abstracts:

Small world networks

Andrew D. Barbour
University of Zürich

Gesine Reinert
University of Oxford

1 Small Worlds

It happens to most of us that we meet a stranger, and in conversation discover that we have a joint acquaintance. "It is a small world", we might then say. The small world phenomenon has been studied in the 1960's by Milgram [11], who sent a number of packets to people in Nebraska and Kansas with instructions to deliver these packets to one of two specific persons in Massachusetts as promptly as possible. The constraint was that the packets could be sent only to persons whom the sender knew on a first-name basis. Milgram determined a median of only about five intermediary recipients to be required to get such a packet to the final destination.

About thirty years later, Watts and Strogatz [17] suggested a mathematical model for social networks that was able to mimic this *small-world phenomenon*. The original model was soon modified, (see Newman, Moore and Watts [14]), to make it more amenable for mathematical analysis. In its simplest form, L vertices are put on a one-dimensional ring lattice. Each vertex is connected to its neighbours at distance at most k away. Distance here is lattice distance, each bond on the lattice has length one. To this deterministic graph, random shortcuts are added. With probability ϕ per connection in the deterministic graph, two points are connected by a shortcut. Thus there are $Lk\phi$ shortcuts on average. We are now faced with a random graph that is quite different to the Bernoulli random graphs introduced by Erdös and Rényi [9], in that small worlds display a higher degree of clustering for a given diameter; see also Bollobas [6].

To describe such a network, typically the following summary statistics are used (e.g. Dorogovtsev and Mendes [7], [8]). First, to measure the diameter, the average shortest path length ℓ is introduced. Pick two vertices, calculate their shortest path (using lattice distance, 1 unit per connection), and take the average over all these pairs. To measure clustering, the clustering coefficient C is introduced: let C_i be the fraction of existing connections between nearest neighbours of the node i, then C is the average over all C_i.

The small world phenomenon can intuitively be described as follows: if ℓ is approximately like that for a (Bernoulli) random graph, then C is much larger than for that random graph. As a Bernoulli random graph need not be connected, this formulation does not make rigorous sense, but an intuitive understanding can be obtained observing that, to a very crude order (see [7]),

$$\ell_{random} \approx \frac{\ln L}{\ln(\phi L)}, \quad C_{random} \approx \phi. \tag{1}$$

The argument for this is that, the average number of neighbours of a node is $z = L\phi$, so about z^ℓ nodes of the network are at distance ℓ or closer to it. With $L \sim z^{\mathbf{E}\ell}$, we obtain $\mathbf{E}\ell_{random} \approx \frac{\ln L}{\ln(\phi L)}$. The clustering coefficient is approximately $\frac{L\phi/2}{\binom{L}{2}} = \frac{L\phi}{L+1} \approx \phi$.

Examples where this small-world phenomenon has been assessed empirically include the neural network of `C. elegans`, the metabolic network for `E. coli`, and the power grid of the Western United States. The following table gives the summary statistic, and the comparison with random graphs in the sense of (1); see [7].

	ℓ_{actual}	ℓ_{random}	C_{actual}	C_{random}	L
C. elegans	2.65	2.65	0.26	0.05	282
E.coli	2.9	2.9	≈ 0.3	0.025	282
Power grid	18.7	18.7	0.08	0.0005	4941

Further examples include social networks, the world wide web, rumor propagation (see Zanette [18]), the spread of epidemics (see Ball *et al.* [2], scientific collaboration networks (such as described by Erdös-numbers), metabolic networks (Fell [10]; Ravasz *et al.* [15]), and many more. Indeed, for $k = 1$ the small-world model was first suggested in [2], where it is called the *great circle model*, to study the spread of disease. It can be shown that, to control a disease, movement restrictions (elimination of shortcuts) slow down the spread considerably (see Brian Grenfell's contribution to this conference). From a theoretical physics viewpoint, scaling and percolation are some of the features of interest (e.g. Newman and Watts [12]; Newman *et al.* [13]). In general, small-world networks serve as models for networks that do not appear to be "purely random". More examples, more references as well as more details can be found in the recent books by Watts [16], Barabasi [3] and Dorogovtsev and Mendes [8] as well as in the survey papers by Albert and Barabasi [1] and by Dorogovtsev and Mendes [7].

For a small world network, ℓ and C will both be random quantities, and thus their distribution needs assessing. The presented work concerned approximating the distribution of ℓ, with a bound on the approximation error.

2 The distribution of the shortest path length

To study the distribution of ℓ, Newman, Moore and Watts [14] introduced the continuous circle model. Instead of the ring lattice, study a circle C of circumference L, to which a Poisson $(L\rho/2)$ number of shortcuts are added uniformly over the circle. With neighbourhood collapsed by dividing distances by k, ρ corresponds to $2k\phi$. In this continuous circle model, chords between points have length zero.

Using mean-field heuristics, [14] derive the approximate distribution of the shortest distance ℓ; in particular they state that

$$\mathbf{E}\ell = \frac{L}{k} f(Lk\phi),$$

where

$$\begin{aligned} f(z) &= \frac{1}{2\sqrt{z^2+2z}} tanh^{-1}\sqrt{\frac{z}{z+2}} \\ &\sim \begin{cases} \frac{1}{4} & \text{for } z \ll 1 \\ \frac{\log(2z)}{4z} & \text{for } z \gg 1. \end{cases} \end{aligned}$$

Let us restrict attention to the case that $L\rho > 1$ – if the probability is high that there are no shortcuts, then the shortest distance between two points will mostly just be the distance on the deterministic graph.

Barbour and Reinert [5] show that, uniformly in $|x| \leq \frac{1}{4}\log(L\rho)$,

$$\begin{aligned} &\mathbf{P}\left(\mathcal{D} > \frac{1}{\rho}\left(\frac{1}{2}\log(L\rho) + x\right)\right) \\ &= \int_0^\infty \frac{e^{-y}}{1+e^{2x}y}\, dy + O\left((L\rho)^{-\frac{1}{5}}\log^2(L\rho)\right). \end{aligned} \tag{2}$$

Also an exact expression for the bound on the distance is given.

The corresponding approximating probability from [14] is

$$\mathbf{P}\left(\mathcal{D} > \frac{1}{\rho}\left(\frac{1}{2}\log(L\rho) + x\right)\right) \approx \frac{1}{1+e^{2x}}\left(1 + O\left((L\rho)^{-\frac{1}{2}}\right).$$

The difference to (2) can be explained by the mean-field approximation in [14] being of rather crude order. We will return to this point once the result (2) is explained in more detail.

To derive the limiting distribution, the following heuristic may be useful. For the rigorous argument, see [5]. Pick a point P at random from C, and denote by $R(t)$ the set of points that can be reached from P within time t. Here we assume that the process walks from P at the same speed 2ρ in all possible directions, taking any shortcut that it can find. Thus it will grow at rate 2ρ from P along the circle. Whenever it encounters a shortcut, it will take it, creating new intervals on the circle that are covered by the process. This process will in due time meet some areas that it has covered before. This introduces dependence in the intervals. We compare this process to a pure growth process $S(t)$ starting at P with growth rate $2p$, which ignores overlap. For small times t, we expect that $R(t) \approx S(t)$.

Now pick another point P' at random from C, and let an independent pure growth process run from that point. The time at which the two independent pure growth processes will meet will be approximately $\frac{1}{2}\mathcal{D}$, where $\mathcal{D}$ is the length of the shortest path between P and P'.

To make the above heuristic more precise, denote that for the pure growth process $S(t)$ started at P (which is a Yule process) by $M(t)$ the number of intervals at time t, and by $s(t)$ the total length of the circle covered at time t. Then

$$\mathbf{E}M(t) = e^{2\rho t}, \quad \mathbf{E}s(t) = \frac{1}{\rho}\left(e^{2\rho t} - 1\right).$$

Denote by $N(t)$ and $u(t)$ the corresponding quantities for the pure growth process started at the point P'. Running both pure growth processes from time 0, at time t there are approximately $e^{4\rho t}$ pairs of intervals, and each has approximately length $\frac{1}{\rho}$. If V_t denotes the number of intersecting pairs of intervals at time t, one from the process started at P, the other from the process started at P', then

$$V_t \approx \frac{2}{L\rho} e^{4\rho t}.$$

The time scale at which the first encounter of the two processes will happen should be such that V_t is a small but visible number. Thus we put

$$\tau_x = \frac{1}{2\rho} \left\{ \frac{1}{2} \log(L\rho) + x \right\};$$

then

$$V_{\tau_x} \approx 2e^{2x}.$$

Indeed V_t is random, and a mixed Poisson approximation for V_t can be derived. Given that $M(\tau_x) = m$, with interval lengths $s_1, \ldots, s_m$, and $N(\tau_x) = n$, with interval lengths $u_1, \ldots, u_n$, we have that

$$V_{\tau_x} \approx \text{ Poisson} \left(\frac{2}{L} \sum_{i=1}^{m} \sum_{j=1}^{n} \min(s_i, u_j) \right).$$

If $\hat{V}_t$ denotes the number of intersections at time t in the original process $R(t)$, started from P and P', then $\hat{V}_{\tau_x} \approx V_{\tau_x}$. Intuitively this stems from τ_x being rather a small time; for later times the process may differ considerably. As

$$\{V_{\tau_x} = 0\} \approx \{\hat{V}_{\tau_x} = 0\} = \{\mathcal{D} > 2\tau_x\},$$

we thus obtain

$$\begin{aligned} \mathbf{P}\{\mathcal{D} > 2\tau_x\} &\approx \mathbf{E} e^{-\frac{2}{L} \sum_{i=1}^{M(\tau_x)} \sum_{j=1}^{N(\tau_x)} \min(s_i, u_j)} \\ &= \mathbf{E} e^{-\frac{4}{L} \int_0^{\tau_x} M(v) N(v) dv}. \end{aligned}$$

To derive the final result, a martingale argument is employed. We know that

$$e^{-2\rho t} M(t) \to W \quad a.s.,$$

where W is exponentially distributed with parameter 1. As

$$\begin{aligned} e^{-\frac{4}{L} \int_0^{\tau_x} M(v) N(v) dv} &\approx e^{-\frac{4}{L} WW' \int_0^{\tau_x} e^{4\rho v} dv} \\ &\approx e^{-e^{2x} WW'}, \end{aligned}$$

where W and W' are independent, exponential random variables with parameter 1. Noting that

$$\mathbf{E} e^{-e^{2x} WW'} = \int_0^{\infty} \frac{e^{-y}}{1 + e^{2x} y} \, dy$$

yields the stated result (2). Moreover, bounds on these approximations are given.

It might be intuitive to compare the above result (2) to that obtained by [14]. From (2) we have that

$$\mathbf{P}\left(\rho\mathcal{D} > \left(\frac{1}{2}\log(L\rho) + x\right)\right) \approx \int_0^\infty \frac{e^{-y}}{1+e^{2x}y}\,dy. \tag{3}$$

Note that

$$\begin{aligned}\mathbf{E}\left\{e^{-e^{2x}WW'}|W,W'\right\} &= e^{-e^{2x}WW'} \\ &= e^{-\exp\{2x+\log W+\log W'\}} \\ &= e^{-\exp\{2x-G_1-G_2\}},\end{aligned}$$

where $G_1 := -\log W$ and $G_2 := -\log W'$ both have the Gumbel distribution. Let T be a random variable such that $\mathbf{P}(T > x)$ is given by the right-hand side of (3), then with this construction,

$$\mathbf{P}[2T - \{G_1 + G_2\} > x \,|\, W, W'] = e^{-e^x},$$

whatever the values of W and W', and hence of G_1 and G_2, implying that, in distribution,

$$2T = G_1 + G_2 - G_3,$$

where G_1, G_2 and G_3 are independent random variables with the Gumbel distribution. In contrast, the limiting distribution in [14] can be written as the distribution of $G_1 - G_3$, thus ignoring some of the initial branching variation.

The generalization of the above to higher-dimensional lattices derived in [5] shows that the reduction in shortest distance as a result of introducing shortcuts decreases with increasing dimension.

In forthcoming work, Barbour and Reinert study discrete small worlds, first a discrete circle with continuous time evolution, secondly the discrete circle with discrete time. The latter covers the original small world model. where shortcuts have length one, and neighbourhood sizes come in explicitly.

3 Conclusion

The above work is but a start on studying the statistical properties of small world networks. Yet there are already more complicated models suggested that demand treatment. Possible extensions of the above work include the following.

Often real networks are found to display a hierarchical structure, such as many small, highly connected substructures linked by a larger structure, see, for example, Ravasz *et al.* [15]. The above does not incorporate such hierarchical networks, but might be adaptable to do so.

Another summary statistic to describe networks is the degree $\langle k \rangle$, the average number of neighbours for a vertex (see for example [7]). In many real networks, it has been

observed that there are some vertices that have a very large number of connections to other vertices; indeed, often a scaling law for the degree is postulated. This is modelled using so-called *scale-free networks*, where the probability of a shortcut is biased towards vertices that already have shortcuts. Due to the uniformity of the shortcut construction, small-world networks do not display this scale-free property. It would be interesting to explore this further.

References

[1] Albert, R. and Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47–97.

[2] Ball, F., Mollison, D. and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89.

[3] Barabasi, A.-L. (2002). *Linked: the new science of networks.* Perseus, Cambridge, Massachusetts.

[4] Barbour, A.D., Holst, L., and Janson, S. (1992). *Poisson Approximation.* Oxford Science Publications, Oxford.

[5] Barbour, A.D. and Reinert, G. (2001, 2003). Small Worlds. *Random Structures and Algorithms* **19**, 54–74. Correction: submitted to Random Structures and Algorithms, 2003.

[6] Bollobas, B. (1985) *Random Graphs.* Academic Press, London.

[7] Dorogovtsev, S.N. and Mendes, J.F.F. (2002). Evolution of random networks. *Adv. Phys.* **51**, 1079–1187.

[8] Dorogovtsev, S.N. and Mendes, J.F.F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW.* Oxford University Press, Oxford.

[9] Erdös, P. and Rényi, A. (1959). On random graphs. *I. Publicationes Mathematicae (Debrecen)* **6**, 290–297.

[10] Fell, D.A. and Wagner, A. (2000). The small world of metabolism. *Nature Biotech.* **189**, 1121–1122.

[11] Milgram, S. (1967). The small world problem. *Psychol. Today* **2**, 60–67.

[12] Newman, M.E.J. and Watts, D.J. (1999). Scaling and percolation in the small-world network model. *Phys. Rev. E* **60**, 7332–7344.

[13] Newman, M.E.J., Jensen, I., and Ziff, R.M. (2002). Percolation and epidemics in a two-dimensional small world network. *Phys. Rev. E* **65**, 021904.

[14] Newman, M.E.J., Moore, C. and Watts, D.J. (2000). Mean-field solution of the small-world network model. *Phys. Rev. Lett.* **84**, 3201–3204.

[15] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabasi, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1554.

[16] Watts, D.J. (1999). *Small Worlds*. Princeton University Press, Princeton.

[17] Watts, D.J. and Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. *Nature* **393**, 440–442.

[18] Zanette, D.H. (2002). Dynamics of rumor propagation on small-world networks. *Phys. Rev. E* **65**, 041908.

Modelling lipid oxidation

Bo Martin Bibby, Ib M. Skovgaard, Lise R. Nissen, Grete Bertelsen, and Leif Skibsted

The Royal Veterinary and Agricultural University, Denmark

Abstract

A new approach for evaluating lipid oxidation was developed by modelling data obtained by the oxygen consumption method. Based on the generalized scheme for lipid autoxidation, a compartment model involving the concentration of the four oxidation specimens of the unsaturated fatty acid, RH, $R\cdot$, $ROO\cdot$, and $ROOH$ as well as the concentration of oxygen and the rate constants for initiation (a), formation of peroxyl radicals (b), and formation of alkyl radicals (c) was constructed. As all rates of reaction were considered to be of second order the dynamic part of the model could be described by five coupled differential equations expressing the overall reaction rate for both the lag phase and the propagation phase of lipid oxidation.

4 The Problem

Eventually storage of food leads to that it becomes unfit for human consumption. It is believed that this is mainly due to oxidation of lipids in particular unsaturated free fatty acids. Lipid oxidation is considered one of the most serious quality reducing factors with a negative effect on

- Taste
- Smell
- Texture
- Nutritional value

The problem is to control (delay) this deterioration. In this project the focus has mainly been on trying to understand and describe/model the primary chemical reactions in the oxidation process.

5 The Data

For this purpose an experiment was carried out involving six linoleic acid solutions prepared for measuring lipid peroxidation using the oxygen consumption method as described by [3] and [1]. Three of the solutions were from a new batch of linoleic acid and the remaining three were from an old batch. The oxygen concentration was recorded at time intervals of 5 seconds for 20 minutes or until the concentration had reached zero. The data is shown in Figure 1.

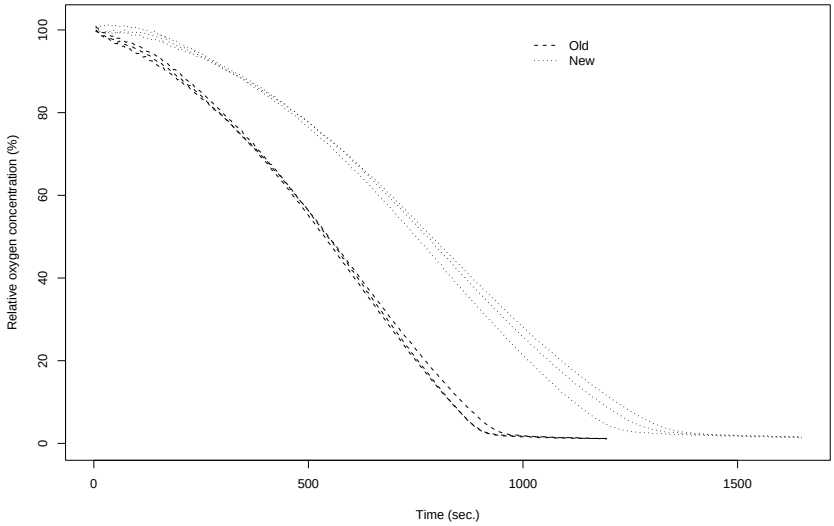


Figure 1: Relative oxygen concentration as a function of time (sec.) for the six linoleic acid solutions.

6 The Model

The primary chemical processes in the autoxidation of lipids are discussed in [2] from which the scheme in Figure 2 is taken (slightly modified), see also [4]. The following notation is used,

$$
\begin{aligned}
RH &\quad — \quad \text{Unsaturated Lipid} \\
ROOH &\quad — \quad \text{Hydroperoxide} \\
ROO\cdot &\quad — \quad \text{Peroxyl radical} \\
R\cdot &\quad — \quad \text{Alkyl radical} \\
O_2 &\quad — \quad \text{Oxygen}
\end{aligned}
$$

The aim is to translate the understanding of the chemical processes into a model for the autoxidation with a focus on the processes closest to the observed oxygen consumption. Our proposal is the compartment model depicted in Figure 3.

In Figure 3 boxes indicate compartments and arrows with the same symbol correspond to a chemical reaction between the compounds associated with the compartments from which the arrows come. The compartment at the receiving end of two arrows, with the same symbol, corresponds to the end product of that chemical reaction. The symbol represents the rate constant of the reaction.

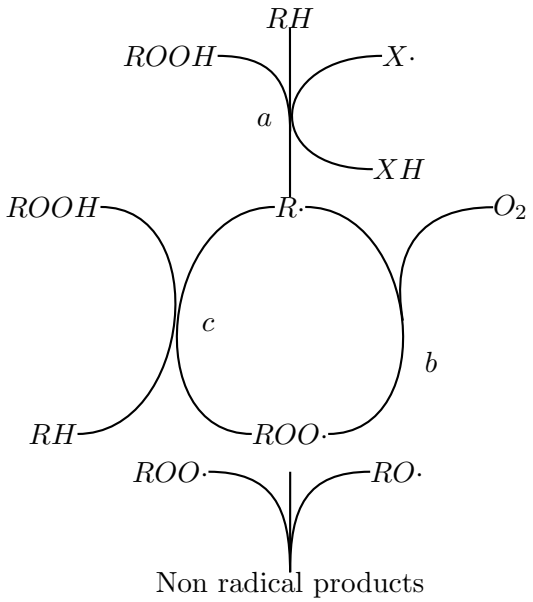


Figure 2: Scheme for autoxidation of lipids.

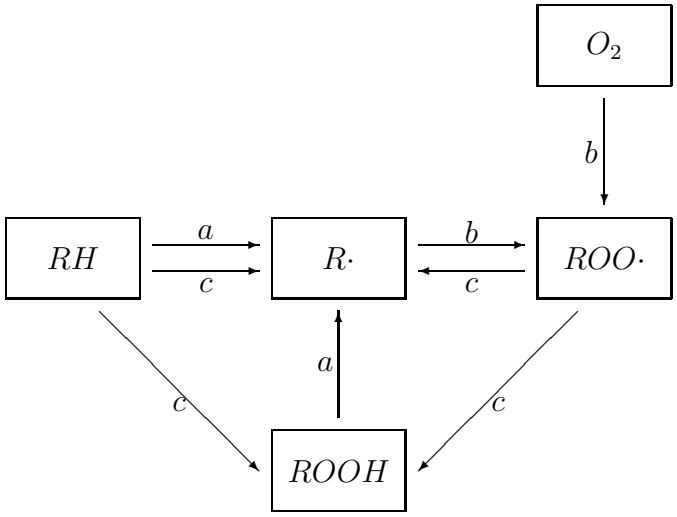


Figure 3: A graphical representation of the compartment model.

Based on the assumption of second order kinetics the dynamical part of the model can be described by a system of coupled differential equations, one for each compartment. The five coupled differential equations associated with the model are given

below.

$$
\begin{aligned}
\frac{dRH_t}{dt} &= -\ aRH_tROOH_t & & & -\ cRH_tROO_t, & \quad RH_0 = rh_0,\\
\frac{dR_t}{dt} &= \ aRH_tROOH_t & -\ bR_tO_{2t} & +\ cRH_tROO_t, & & \quad R_0 = 0,\\
\frac{dO_{2t}}{dt} &= & -\ bR_tO_{2t}, & & & \quad O_{20} = o_{20},\\
\frac{dROO_t}{dt} &= & & bR_tO_{2t} & -\ cRH_tROO_t, & \quad ROO_0 = 0,\\
\frac{dROOH_t}{dt} &= -\ aRH_tROOH_t & & & +\ cRH_tROO_t, & \quad ROOH_0 = rooh_0.
\end{aligned}
$$

7 The Fit

The statistical analysis of the data is based on maximum likelihood estimation in an $AR(2)$-model for the residuals, see [5]. The residuals were obtained by subtracting numerical solutions of the coupled differential equation system at the observation time-points from the observations. The parameters of interest were the rate constants

a — Initiation,
b — Formation of peroxyl radicals,
c — Formation of alkyl radicals,

and the initial concentrations of RH, O_2, and $ROOH$. The result was almost a perfect fit as can be seen in Figure 4.

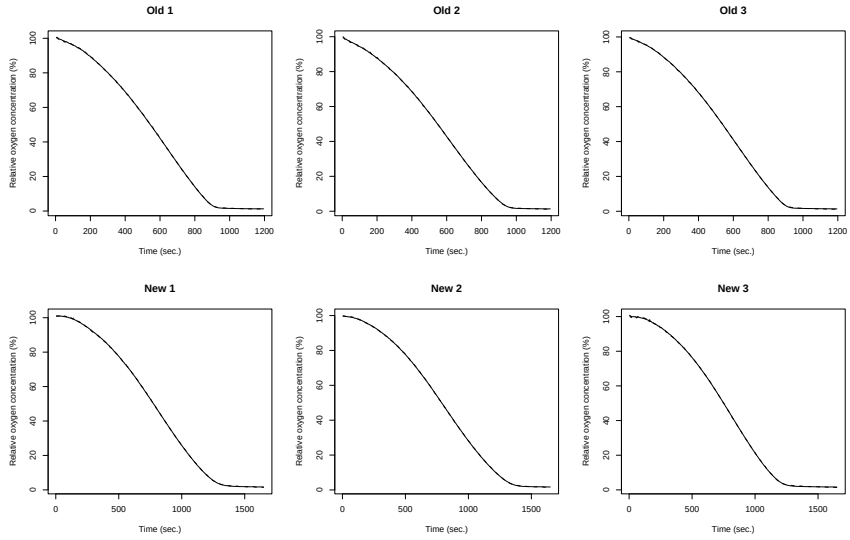


Figure 4: Observed (dotted curve) and fitted (solid curve) relative oxygen concentration for the six linoleic acid solutions.

In Figure 5 the five compartment processes are shown along with the $AR(2)$-residuals for the first curve from the old batch. The five processes look more or less as you would expect, but the residuals show a clear pattern particularly very early and very late in the experimental period. This might suggest room for improvement but the practical implications of this would be minimal because of the good overall fit.

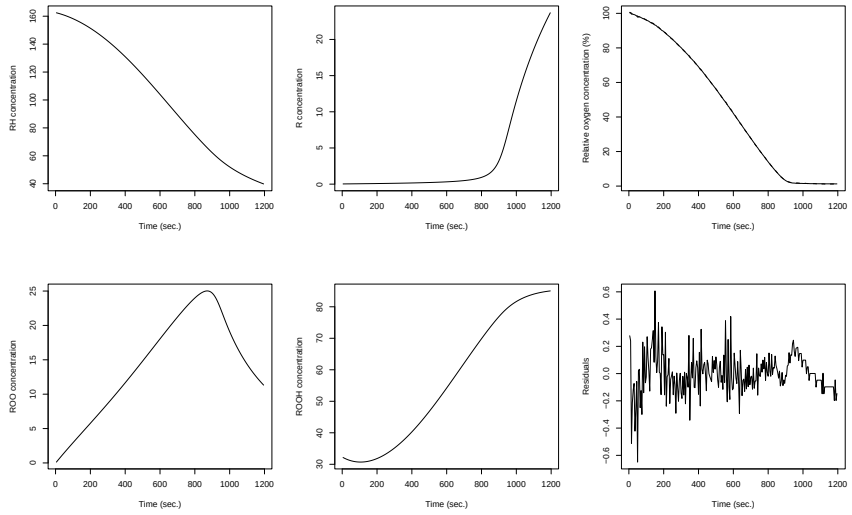


Figure 5: The five compartment processes and the $AR(2)$-residuals for the first curve from the old batch. The data points are given by the dotted curve in the oxygen plot.

Some of the rate constants and initial pool sizes did not match chemically founded expectations but in light of the good fit to the data and with only observations from a single compartment there is not much hope of supporting a refinement of the model in the data analysis.

8 Incorporating Antioxidants

As already mentioned, a main goal of the modelling was to get an understanding of the mechanisms behind the lipid oxidation. Taking this a step further it is also of interest to investigate possible ways of prolonging the lag-phase and/or inhibiting the propagation phase of the autoxidation. This could be described as an anti-oxidative effect, and many spices and natural extracts are known to work as antioxidants.

There are many possibilities for including an antioxidant in the model introduced here, but the belief was that the primary reaction of the antioxidant was with the peroxyl radicals. Based on this observation, a natural extension of the model is given in Figure 6, where S denotes a stable radical, which is not of interest, and A is the antioxidant.

Based on oxidation curves corresponding to several concentrations of the antioxidant under study, the idea is to estimate a common value of the rate constants a, b, c, and d along with the initial concentrations of RH and $ROOH$ and then to relate the

estimates of the initial concentrations of the antioxidant to actual concentrations used in the experiment.

In Figure 7 the result of this estimating procedure is shown for an example involving grape fruit extract as an antioxidant.

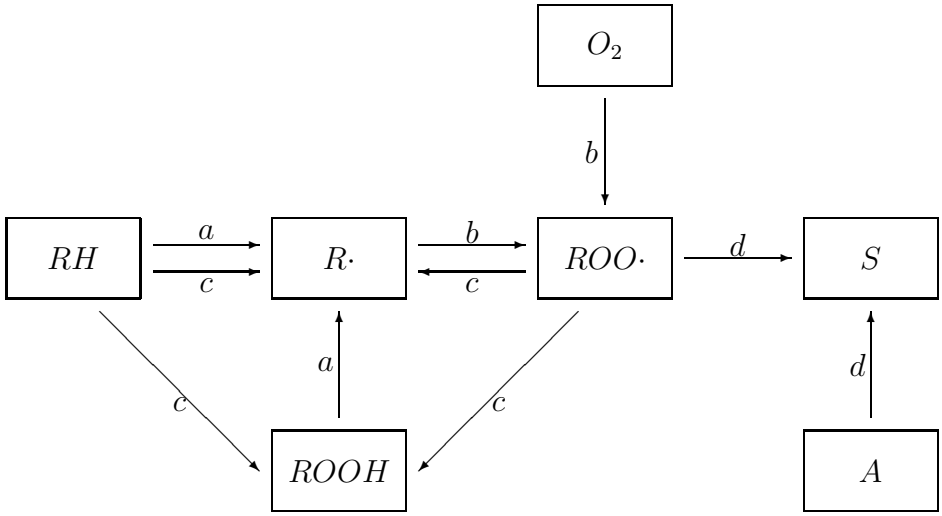


Figure 6: A graphical representation of the compartment model including an antioxidant.

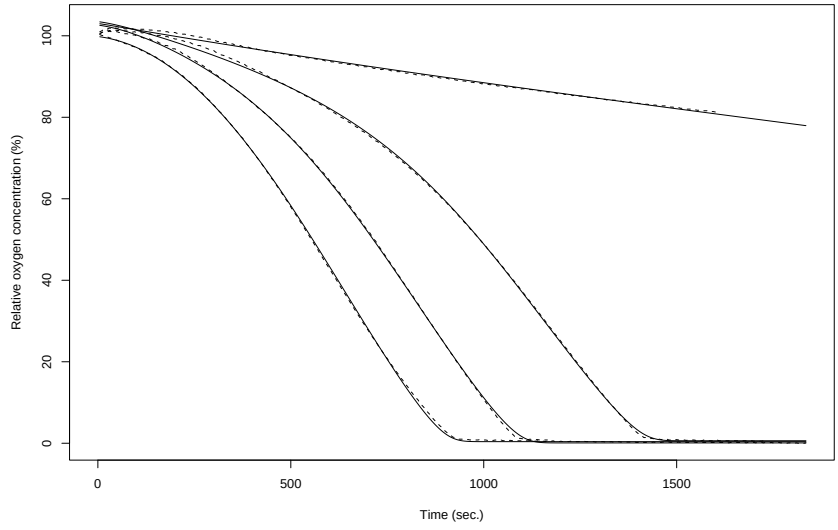


Figure 7: Observed and fitted relative oxygen concentrations for the antioxidant grape fruit. From top to bottom the grape fruit concentrations are 0.20%, 0.10%, 0.05%, and 0%.

The fit is not as good as for the single curves but it seems that the model has captured essential features of the oxidation process. Again a refinement of the model would probably require observations from other parts of the system.

References

[1] Andersen, H. J. & Skibsted, L. H. (1992). "Kinetics and Mechanism of Thermal Oxidation and Photooxidation of Nitrosylmyoglobin in Aqueous Solution". *J. Agri. Food Chem.*, 40:1741–1750.

[2] Madsen, H. L.; Bertelsen, G. & Skibsted, L. H. (1997). "Antioxidative Activity of Spices and Spice Extracts". In Risch, S. J. & Ho, C.-T., editors, *Spices. Flavor Chemistry and Antioxidative Properties*, chapter 14, pages 176–187. American Chemical Society, Washington DC.

[3] Mikkelsen, A.; Sosnieki, L. & Skibsted, L. H. (1992). "Myoglobin Catalysis in Lipid Oxidation". *Z. Lebensm. Unters. Forsch.*, 195:228–234.

[4] Nawar, W. W. (1985). "Lipids". In Fennema, O. R., editor, *Food Chemistry, 2nd edition*, chapter 4, pages 139–244. Marcel Dekker Inc, New York.

[5] Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear Regression*. John Wiley & Sons, Inc.

Physical networks in cell signalling

Dennis Bray
University of Cambridge

The interior of living cells is a strange environment - very different to anything usually considered by physical chemists. Molecules are present in a slurry rather than in solution and there is a great deal of organisation and inhomogeneity. Macromolecules are densely packed together and their chemical reactions are strongly influenced by their location in the cell and their mechanical effects. Many processes are driven by numbers of molecules small enough that the thermal fluctuations in reaction rates become significant. In order to understand and make predictions about events in this strange domain we believe we must take quantitative data at many different scales, obtained by biological, chemical and physical techniques, and integrate them into large-scale computer models.

A system that has emerged in recent years as a testing ground for computational cell biology is the chemotactic response of Escherichia coli - arguably the best understood form of cell behaviour (Bren and Eisenbach 2000). All of the intracellular proteins involved in the chain of responses from receptor to flagellum have been purified and all have been sequenced at the DNA level. Structural information is now available for all of the proteins, and the enzymatic reactions they catalyze have been analysed kinetically. Many mutants lacking identified proteins, singly or in combination, have been isolated and their chemotactic responses documented. Computer-based analyses have been used to test the consistency of this large body of data and to check its integrative properties. Computer models have also been developed to simulate the behaviour of flagellar motors, especially the stochastics of switching.

Our own work in this area features a close collaboration between computer simulation and experiment, especially work carried out by Robert Bourret and his colleague at the University of North Carolina. This productive interaction led to the development of a computer-based model BCT (bacterial chemotaxis), using a deterministic rate-equation approach similar to those used in metabolic models (Bray et al. 1993). Development and refinement of BCT continues, and it now provides a detailed account of the stimulus response and adaptation of cells to aspartate, and correctly predicts the phenotype of over 60 mutants with altered chemotactic genes (results and program for download are at www.zoo.cam.ac.uk/comp-cell). The validity of the BCT program has been confirmed on numerous occasions, for example by correctly predicting the phenotype of a set of "gutted" mutants.

However, deterministic computer models are unable to match the increasing resolution of experimental techniques used to study bacterial swimming. We were therefore led to explore a novel type of computer simulation capable of much higher resolution in which individual molecules are represented as software objects rather than as concentrations(Morton-Firth 1998).Encounters between molecules are represented not by rate equations (as in the usual type of biochemical simulation) but by weighted probabilities. The program StochSim has been developed into an accurate and robust computational tool, and used to explore aspects of signal transduction in bacterial chemotaxis (available for download from ftp://ftp.cds.caltech.edu/pub/dbray).

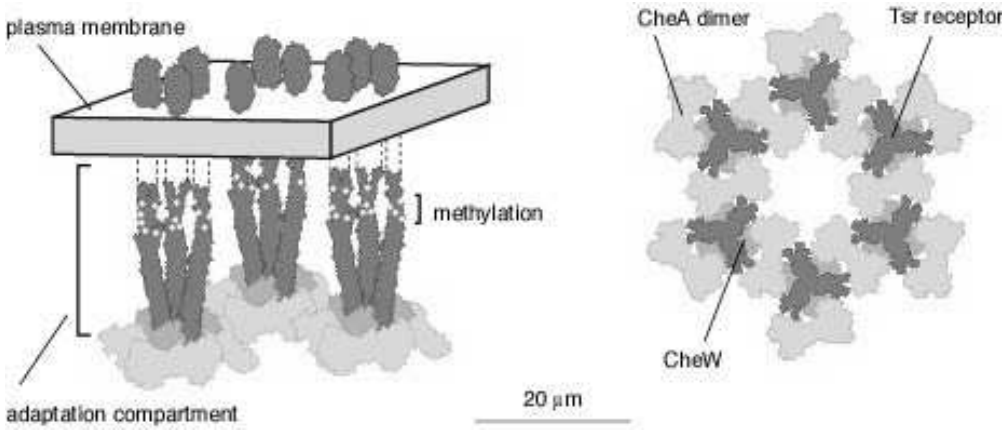


Figure 1: Lattice of chemotactic receptors. A portion of the lattice viewed from the side (left) shows extracellular domains of the chemotactic receptors on one side of the plasma membrane and the long α-helical cytoplasmic domains on the cytoplasmic side. A lattice of CheA and CheW molecules, attached to the ends of the receptor tails about 30 nm from the plasma membrane, forms a closed volume of cytoplasm, termed the "adaptation compartment". The plan view(right) shows the hexagonal lattice as though viewed from the plasma membrane looking into the cell.

It has a number of advantages over more conventional programs such as the Gillespie algorithm, including an improved capacity to handle very large numbers of similar reactions (the "combinatorial explosion" often encountered by programmers) and an ability to represent protein conformational transitions during the transmission of signals (Morton-Firth et al. 1999).

Over the past several years our level of understanding of the chemotactic pathway has increased greatly due to the elucidation of the structures of key proteins in the pathway. Structures of the extracellular domain of one receptor, and most of the cytoplasmic domain of another, have been published. So too have the structures of the histidine kinase CheA and those of the methylation and demethylating enzymes CheR and CheB. The atomic coordinates of both CheW and CheZ have recently been published. The availability of these structures led us to use conventional molecular graphics programs, in conjunction with plastic models generated by 3-D printer technology, to predict how these proteins were arranged in relation to the plasma membrane (Shimizu et al. 2000). The structure we proposed is a regular two-dimensional lattice in which the cytoplasmic ends of chemotactic receptor dimers inserted into a hexagonal array of CheA and CheW molecules (Figure 1). The array creates separate compartments for adaptation and downstream signaling and suggests a possible basis for the spread of activity within the cluster. This model is consistent with a large body of biochemical, mutational and protein structural data, including recent mutagenesis studies from the Parkinson laboratory (Ames et al. 2002).

This lattice of receptors and associated proteins is the basis by which signals are generated during the chemotactic response. Conformational change appears to be the

most likely mechanism by which extracellular stimuli can be transmitted across the plasma membrane, since the receptors in this case are permanently dimeric and receptor dimerization is not part of the rapid detection response. Consistent with this view, experimental evidence has been obtained for a small shift in protein conformation accompanying binding of the ligand aspartate and for the existence of two distinct conformations of the histidine kinase Che A (Falke and Hazelbauer 2001). The widely accepted view is that an unoccupied receptor favors the "active" conformation of CheA in which it triggers the catalytic formation of phosphoryl groups. Binding of an attractant such as aspartate to the receptor is thought to cause the receptor to change its conformation to an "inactive" conformation which terminates, or inhibits, phosphoryl generation. Recent FRET analysis confirms the remarkably high amplification, or gain, shown by the receptor complex (Bray 2002; Sourjik and Berg 2001).

We have suggested that the remarkable sensitivity and range of response of bacterial chemotaxis might depend on the clustering of chemotactic receptors on the surface of the bacterium (Bray et al. 1998). Specifically, we hypothesized that when a ligand binds to a receptor, the resulting change in activity might propagate to neighboring receptors in a cluster. We calculated that if the size of this "infective" spread was changed by adaptation, then the system could readily reproduce the chemotactic response of actual bacteria. The molecular mechanism of this effect, however, was still arbitrary and not easily expressed in terms of quantitative physical chemistry. We therefore sought the help of Tom Duke, a physicist at the Cavendish Laboratory, who suggested that the energetic exchanges between proteins in a two-dimensional lattice was in many ways analogous to the interactions of magnetic dipoles in a spin glass. Statistical mechanical analysis of a simplified array, in the form of an Ising model, then showed that the inclusion of a single free energy term due to cooperative interactions between adjacent receptors could integrate their activities over an extended lattice (Duke and Bray 1999). Selection of suitable interaction energy in this "Duke/Bray" model then predicted a lowered threshold and a greatly increased range of detection for the array of receptors.

The broad outcome of this excursion into physics was the idea that protein conformations might propagate, in domino fashion through an extended multimolecular complex, a mechanism we call conformational spread. We recently analyzed this postulated mechanism in greater detail by applying it to the one-dimensional, unbounded case of a closed ring of proteins (Duke et al. 2001). The simple geometry of this new situation allowed us to define rigorously conditions under which a ring will show cooperative switching and to relate the physics of conformational spread to classical models of allostery (the canonical MWC and KNF models emerge naturally as limiting cases of conformational spread). This study also revealed the interesting fact that the time taken for a large multiprotein assembly to change its state may be much greater than that of individual allosteric transitions - a notion we think is likely will be of great significance in the living cell.

All of the above analyses were simplified in various ways to facilitate mathematical analysis. Proteins, arranged in a closed ring or infinite square lattice, were assigned properties that were as simple as could be. Thus, each protein, or "protomer", was able to adopt one of two possible conformations and had a single site of ligand occupancy - in the Duke/Bray model there was an additional site of adaptational modification

(such as a single site of methylation). Free energy changes associated with activation, ligand binding and methylation were made as symmetric as possible. We have now extended this analysis to a more realistic model of conformational changes in Escherichia coli chemotactic receptors using the StochSim program (Shimizu, Aksenov and Bray , submitted for publication).

Current simulations are of a hexagonal lattice of receptors,with finite (as opposed to infinite) boundaries. The receptors are methylated at up to 4 methyl groups, there are significant dwell times for ligand and conformationally-sensitive binding affinities for CheR and CheB. These analyses have so far shown that the enhanced sensitivity and range of response shown by the Duke/Bray model due to coupling between receptors is retained in the more complete description, but changed in significant ways. The presence of a range of methylation states means that there is not a single critical value of the coupling strength but rather a range of values over which effective performance is enhanced. The increase in sensitivity (chemotactic gain) obtained from a StochSim simulation is less dramatic than that in the idealized single methyl group model, but less "brittle" in the sense that a significant improvement in performance is obtained over a range of possible energy values. Most intriguingly, we have found that the dynamics of the situation lead to the spontaneous emergence of order in the receptor lattice, such that receptors with 4 methyl groups (fully methylated) and receptors with 0 methyl groups (fully unmethylated) tend to lie next to each other in the array. This is only a minimal degree of order, and could be an artefact of the simulation. However the spontaneous emergence of order within a stochastically fluctuating field of allosteric proteins is an intriguing and potentially important phenomenon and we intend to pay attention to it in the future.

Literature cited

Ames, P., Studdert, C. A., Reiser, R. H., and Parkinson, J. S. (2002). "Collaborative signaling by mixed chemoreceptor teams in *Escherichia coli*." *Proc. Natl. Acad. Sci. USA*, 99, 7060–7065.

Bray, D. (2002). "Bacterial chemotaxis and the question of gain." *Proc. Natl. Acad. Sci. USA*, 99, 7–9.

Bray, D., Bourret, R. B., and Simon, M. I. (1993). "Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis." *Mol. Biol. Cell*, 4(May), 469–482.

Bray, D., Levin, M. D., and Morton-Firth, C. J. (1998). "Receptor clustering as a cellular mechanism to control sensitivity." *Nature*, 393, 85–88.

Bren, A., and Eisenbach, M. (2000). "How signals are heard during bacterial chemotaxis: protein-protein interactions in sensory signal propagation." *J. Bacteriol.*, 182, 6865–6873.

Duke, T. A. J., and Bray, D. (1999). "Heightened sensitivity of a lattice of mem-

brane receptors." *Proc. Natl. Acad. Sci. USA*, 96, 10104–10108.

Duke, T. A. J., Le Novère, N., and Bray, D. (2001). "Conformational spread in a ring of proteins: a stochastic view of allostery." *J. Mol. Biol.*, submitted for publication.

Falke, J. J., and Hazelbauer, G. L. (2001). "Transmembrane signaling in bacterial chemoreceptors." *Trends Biochem. Sci.*, 26, 257–265.

Morton-Firth, C. J. (1998). "Stochastic simulation of cell signalling pathways.," Ph.D., Cambridge.

Morton-Firth, C. J., Shimizu, T. S., and Bray, D. (1999). "A free-energy-based stochastic simulation of the Tar receptor complex." *J. Mol. Biol.*, 286, 1059–1074.

Shimizu, T. S., Le Novère, N., Levin, M. D., Beavil, A. J., Sutton, B. J., and Bray, D. (2000). "Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis." *Nature Cell Biol.*, 23, 792–796.

Sourjik, V., and Berg, H. C. (2001). "Receptor sensitivity in bacterial chemotaxis." *Proc. Natl. Acad. Sci. USA*, 99, 12669–12674.

An integrated probabilistic model for functional prediction of proteins

Minghua Deng, Ting Chen and Fengzhu Sun
University of Southern California

Protein function prediction is an important problem in molecular biology. The most widely used method for protein function prediction is by database search using programs such as PSI-BLAST [1] and FASTA [11], and then predict functions based on sequence homologies. However, a large fraction of protein sequences are not similar to proteins with known functions. For example, about a third of yeast proteins (one of the most studied model organisms) do not have defined functions. The development of high-throughput bio-techniques and their applications in many areas of biology have generated a large amount of data that are useful for the study of protein functions, for example, protein physical interactions [8, 9, 13], genetic interactions [10, 12], protein complexes [5, 7] and protein co-expression from gene expression data. Individual protein features, such as their domain content, also contain information about their functions. A challenging task that lies ahead is to discover the functional roles of the unknown proteins combining different sources of information.

Methods based on chi-square statistics [6] and on frequencies of interaction partners having certain functions of interest [4] have been used to assign functions to unknown proteins based on protein interactions. However, these methods have serious limitations in predicting protein functions. We developed an integrated probabilistic model to combine protein physical interactions, genetic interactions, highly correlated gene expression network, protein complex data and domain structures of individual proteins together to prediction protein functions based on Markovian random field theory [2, 3]. The model is flexible that other protein pairwise relationship information and features of individual proteins can be easily incorporated. We applied our integrated approach to predict functions of yeast proteins based on MIPS protein function classifications and the interaction networks based on MIPS physical and genetic interactions, gene expression profiles, and Tandem Affinity Purification (TAP) protein complex data, and protein domain information. We study the sensitivity and specificity of the integrated approach using different sources of information by the leave-one-out approach. As more data are incorporated into the model, the accuracy of the approach increases. Compared to using MIPS physical interactions only, the integrated approach combining all the information increases the sensitivity from 57% to 87% when the specificity is set at 57%, an increase of 30%. It should also be noted that by enlarging the interaction network, the number of proteins whose functions can be predicted is also greatly increased.

Key words: Protein function prediction, Protein-Protein Interaction, Markov Random Field, Gibbs Sampler.

References

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389-3402.

[2] Deng, M.H., Zhang, K., Mehta, S., Chen, T., and Sun, F.Z. (2003) Prediction of protein function using protein-protein interaction data. *J. Comp. Biol.*, in press.

[3] Deng, M.H., Chen, T., and Sun, F.Z. (2003) An Integrated Probabilistic Model for Functional Prediction of Proteins. *Proc. RECOMB2003.*

[4] Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology* **18:** 1257 − 1261.

[5] Gavin, A., Böche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141−147.

[6] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* **18:** 523 − 531.

[7] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K. et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415:** 180 − 183.

[8] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto,K., Kuhara, S., Sakaki, Y. (2000) Toward a Protein-protein Interaction Map of the Budding Yeast: a Comprehensive System to Examine Two-hybrid Interactions in All Possible Combinations Between the Yeast Proteins. *Proc. Natl. Acad. Sci. USA* **97:** 1143 − 1147.

[9] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A Comprehensive Two Hybrid Analysis to Explore the Yeast Protein Interactome. *Proc. Natl. Acad. Sci. USA* **98:** 4569 − 4574.

[10] Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**: 31-34.

[11] Pearson,W.R., and Lipman,D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85:** 2444 - 2448.

[12] Tong, A.H.Y., Drees, B., Nardelli, G., Bader G.D., Brannetti, B., Castagnoli, L. Evangelista, M., Paoluzi, S., Quondam, M., Zucconim A, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**: 321 − 324.

[13] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, et al. (2000) A Comprehensive Analysis of Protein-protein Interactions in *Saccharomyces Cerevisiae*. *Nature* **403**: 623 − 627.

A model of the uptake of alternative fatty acids by isolated rat liver based on stochastic differential equations

Susanne Ditlevsen
University of Copenhagen

Andrea De Gaetano
Università Cattolica del Sacro Cuore, Rome

Deterministic and stochastic differential equations models of the uptake of dodecanedioic acid (C12) in isolated rat livers are considered. The main focus is on including spontaneous erratic variations in the model of the metabolic processes, see also [5].

Mathematical models capable of reproducing observed characteristics in dynamical systems are powerful tools to understand physiological mechanisms. Often a mathematical model of a physiological system is based on ordinary differential equations that describe the dynamics of some state variables under ideal and theoretical conditions [4, 10, 11, 12]. However, such a model is an idealization and does not account for deficiencies of assumed ideal physical conditions, or for the accumulated effect of neglected factors, which often occur in physiological descriptions as the systems can rarely be isolated from influences from the surroundings. A natural extension of the deterministic model is given by a statistical model of stochastic differential equations or diffusions, where relevant parameters are randomized or a term of dynamic noise is added [5, 15, 22, 17]. These stochastic models are also in continuous time, which is important for the physiological interpretation.

The uptake of C12 was studied in the isolated perfused rat liver. A bolus of C12 was injected into the perfusing liver solution, and measurements of the concentration of C12 in perfusate samples were taken over a period of two hours after the injection of the bolus in nine experimental subjects. These data were modelled with a mono-exponential decay for each perfused rat liver [9]. However, the observed decays in the perfusate seem steeper than exponential shortly after injection of bolus, and slower at later time-points. Therefore a two-compartment model seems more appropriate to explain the data, where diffusion between perfusate and liver cells is considered. The model is illustrated in figure 1, and in figure 2 the curves fitted by least squares and data for two randomly chosen rats are shown.

A deterministic model of the elimination kinetics assumes two things: one, that the process actually follows a smooth course (continuous and continuously differentiable), which may be mathematically described as a function of estimable parameters; and, two, that the variability of the actual measurements is due to observation error, which does not influence the course of the underlying process. An alternative, stochastic, approach would result from the hypothesis that the underlying elimination process itself is not smooth. The metabolizing organs and tissues are in fact subject to a variety of internal and external influences, which change over time (e.g. blood flow, energy requirements, hormone levels, the cellular metabolism of the tissues themselves) and which may affect the minute-to-minute rate at which tissues dispose of the load of substrate. This second approach maintains that some degree of noise is already

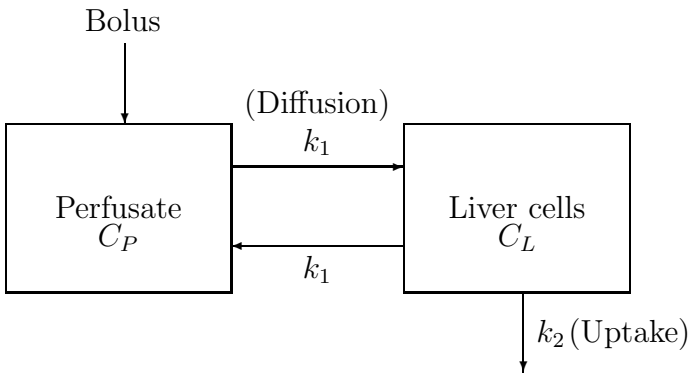


Figure 1: Two compartment model of the uptake of dodecaneioic acid in isolated rat liver.

present in the disposition process itself, and that observational noise may additionally be present. A generalization of the deterministic model based on a stochastic extension is achieved by randomizing the elimination rate constant from the liver cells. This approach defines unambiguously two noise sources: a dynamical noise term that is a part of the process, such that the process at time t depends on this noise process up to time t, and a measurement noise term, that does not affect the process, but only the observations of the process. We end up with a non-gaussian two-dimensional stochastic Itô integral with state-dependent noise. Our main concern is to estimate the model parameters.

Estimating parameters in this kind of model is not straightforward except for simple cases. A natural approach would be maximum likelihood inference, but the transition densities are rarely known, and thus it is not usually possible to write the likelihood function explicitly. A variety of methods for statistical inference for discretely observed diffusion processes has been developed during the past decades, see e.g. [1, 2, 3, 6, 7, 8, 13, 14, 16, 18, 19, 20, 21].

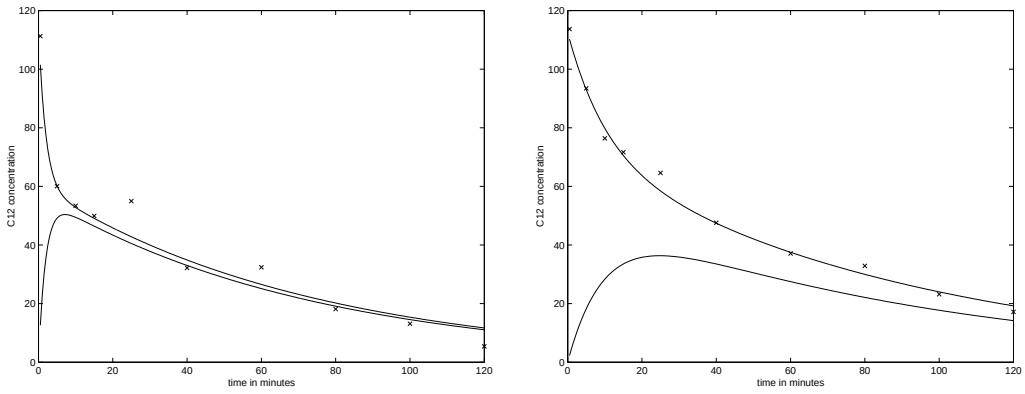


Figure 2: Observed data and the best least square fit curves for the deterministic, two-compartment model for two randomly chosen rats. The curves starting at a positive value represent the C12 concentrations in perfusate, which are measured, and the curves starting at zero represent a scaling of the non-observed concentrations in the liver cells.

Here we are not dealing with a stationary process. Moreover, we only observe one state variable out of two, namely the C12 concentration in perfusate, whereas the C12 amount in the liver cells is unobserved; and measurement errors should also be considered. We can explicitly solve the ordinary differential equations in the drift part of the diffusion, which represent the mean of the process. We first use this to estimate parameters entering in the drift by least squares, and afterwards approximate the unknown likelihood function through Monte-Carlo simulations [15] to estimate the parameter in the diffusion part, as proposed by Pedersen [18]. The standard deviation of the measurement errors is considered fixed, but estimation has been carried out for different values in order to check the robustness of estimates. In figure 3 a simulated trajectory from the fitted model for both coordinates of the process is plotted for the same rats as in figure 2, together with the data, the mean curve and empirical 95% confidence intervals. The estimation procedure and results are described in more detail in [5].

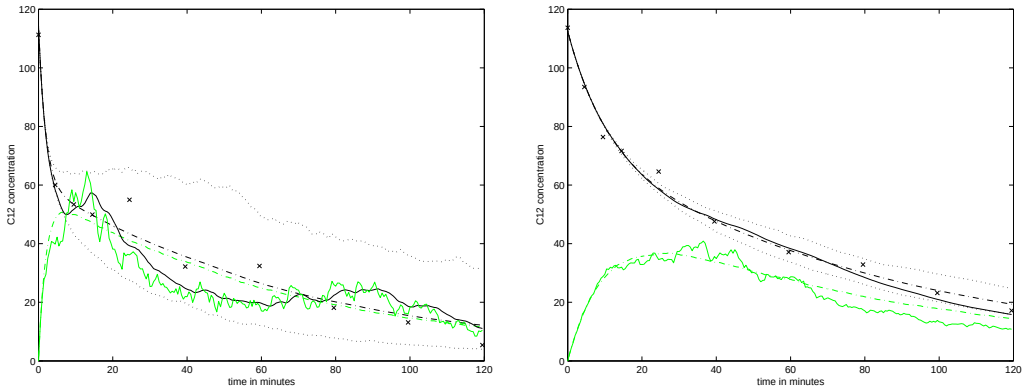


Figure 3: A simulated trajectory (solid lines), mean curve (dash-dot lines) with 95% confidence limits (dotted lines) and observations (X). Trajectories starting at positive values represent concentration in perfusate (black lines), which are measured, trajectories starting at 0 represent a scaling of the non-observed concentration in liver cells (grey lines). Estimation of structural parameters by least squares, diffusion parameter estimated from approximate maximum likelihood.

References

[1] Y. Ait-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70:223–262, 2002.

[2] B.M. Bibby, M. Jacobsen, and M. Sørensen. Estimating functions for discretely sampled diffusion-type models. In Y. Ait-Shalia and L.P. Hansen, editors, *Handbook of Financial Economics*. Amsterdam: North Holland, 2003. Forthcoming.

[3] B.M. Bibby and M. Sørensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1(1/2):017–039, 1995.

[4] A. De Gaetano and O. Arino. Mathematical modelling of the intravenous glucose tolerance test. *J. Math. Biol.*, 40:136–168, 2000.

[5] S. Ditlevsen and A. De Gaetano. Stochastic vs. deterministic uptake of dodecanedioic acid by isolated rat livers. Technical Report 03/11, Department of Biostatistics, University of Copenhagen, 2003.

[6] S. Ditlevsen and M. Sørensen. Inference for observations of integrated diffusion processes. *Scand. J. Statist*, 2003. To appear.

[7] O. Elerain, S. Chib, and N. Shepard. Likelihood inference for discretely observed non-linear diffusions. *Econometrica*, 69:959–993, 2001.

[8] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Stochastics*, 20(4):547–557, 1989.

[9] A.V. Greco et al. Uptake of dodecanedioic acid by isolated rat liver. *Clinica Chimica Acta*, 258:209–218, 1997.

[10] N-H. Holstein-Rathlou and P.P. Leyssac. Oscillations in the proximal intratubular pressure: a mathematical model. *Am. J. Physiol.*, 252:F560–F572, 1987.

[11] N-H. Holstein-Rathlou and D.J. Marsh. A dynamic model of the tubuloglomerular feedback mechanism. *Am. J. Physiol.*, 258:F1448–F1459, 1990.

[12] N-H. Holstein-Rathlou and D.J. Marsh. A dynamic model of renal blood flow autoregulation. *Bull. Mat. Bio.*, 56:411–429, 1994.

[13] M. Kessler. Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist*, 24:211–229, 1997.

[14] M. Kessler and M. Sørensen. Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(2):299–314, 1999.

[15] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Verlag, 1999.

[16] A.R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, 22:1, 55–71, 1995.

[17] A.R. Pedersen. Estimating the nitrous oxide emission rate from the soil surface by means of a diffusion model. *Scand. J. Statist.*, 27:3, 385–403, 2000.

[18] A.R. Pedersen. Likelihood inference by monte carlo methods for incompletely discretely observed diffusion processes. Technical Report 1, Department of Biostatistics, Uni. of Aarhus, 2001.

[19] R. Poulsen. Approximate maximum likelihood estimation of discretely observed diffusion processes. Technical Report 29, University of Aarhus, Centre for Analytical Finance, 1999.

[20] B.L.S. Prakasa Rao. *Statistical Inference for Diffusion Type Processes*. Arnold Publishers, 1999.

[21] M. Sørensen. Prediction-based estimating functions. *Econometrics Journal*, 3:123–147, 2000.

[22] B. Øksendal. *Stochastic Differential Equations*. Springer Verlag, 1998.

Pattern and process in the spatio-temporal dynamics of childhood infections

Bryan T. Grenfell
University of Cambridge

1. Introduction

Epidemiological time series are characteristically non-stationary, exhibiting marked variations through time and often space in mean, variance and the period and extent of oscillations. Traditional frequency domain time series analyses, such as Fourier spectra, cannot deal efficiently with these – often important -- temporal patterns. Recently, these problems have been tackled in the physical sciences and physiology by the development of wavelet spectra, which track the power of oscillations and other features of time series through time, as well as frequency. Here, wavelet spectra are introduced, then applied to explore non-stationarity in the well known measles notification data set for England and Wales.

This paper focuses on an introduction to wavelet methods for epidemiological time series. In the oral presentation, we shall first demonstrate non-stationary variations in epidemic period and explain these patterns using models. We then refine the methods to explore spatio-temporal dynamics, revealing dramatic travelling waves in measles abundance.

2. Methods

Applying 'windowed' Fourier spectra to parts of the series is a partial solution to cyclic non-stationarity; but this method is inefficient and brings its own problems with choice of window width (Torrence & Compo 1998). Recently, this problem has been addressed by the development **of wavelet analysis** of time series, which adds a time axis to the traditional Fourier spectrum, so that variation at different frequencies can be tracked through the time series (Daubechies 1992; Ivanov et al. 1996; Lau & Weng 1995; Nason & von Sachs 1999; Torrence & Compo 1998). These methods have been widely applied to time series and image analysis in meteorology and other physical sciences, as well as increasingly in physiology. However, wavelets have only recently been generally adopted in time series ecology and epidemiology (Bradshaw & Spies 1992; Grenfell et al. 2001).

(Details are summarised from (Torrence & Compo 1998)). Assume that we have a time series, x_n, n=0,…,N-1, with time step δt. We analyse temporal changes in the distribution of power at different frequencies using a ***wavelet function,*** $\psi_0(\eta)$, of a scaled (non-dimensional) time parameter, η. The wavelet function must have both zero mean and be localised in time and frequency space.

Choice of wavelets Wavelets can be either discrete or continuous, real or imaginary. To detect smooth changes in cycle period and phase, we use a continuous, non-orthogonal imaginary **Morlet** wavelet (Fig. 1):

$$\psi_0(\eta) = \pi^{-1/4} \exp(i\omega_o\eta)\exp(-\eta^2/2) \quad |\omega_o \equiv 6|$$

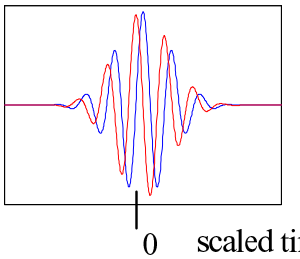


Fig. 1

The ***Continuous Wavelet Transform (CWT)*** of x_n is the convolution of x_n with a scaled and translated version of $\psi_0(\eta)$, ψ^*

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi^* \left[\frac{(n'-n)\partial t}{s} \right].$$

the wavelet transform is normalised to have unit energy as the scale, s, varies. In practice, the CWT is calculated as the Inverse Fourier Transform of the product of the Fourier Transforms of x_n and ψ,

$$W_n(s) = IFFT\left[FFT(x_n) FFT\left(\psi^*\right) \right]$$

at a range of scales, s, generally up to half the length of the series. For the Morlet, scale is almost equal to wavelength, so that the lowest scale, s_0 roughly corresponds to the maximum (Nyquist) frequency of 0.5 cycles per time step. The local wavelet power spectrum at time point *n* and and scale *s* is then given by $|W_n(s)|^2$

Reconstruction of components of the series

Given a CWT, we can reconstruct the original series; in terms of the contribution of variation at different frequencies. This provides a powerful way to isolate, possibly non-stationary, frequency components of interest. Specifically, the full reconstruction is

$$x_n = \frac{\delta j \delta t^{1/2}}{C_\delta \psi_0(0)} \sum_{j=0}^{J} \frac{\Re\{W_n(s_j)\}}{s_j^{1/2}},$$

where $\Re\{W_n(s_j)\}$ is the real part of W, δj is the fraction of the time step, δt, at which we choose to calculate the wavelet transform and C_δ and $\psi_0(0)$ are wavelet-specific constants. Similarly, the contribution to the reconstructed series of variation between scales a and b is

$$x_n = \frac{\delta j \delta t^{1/2}}{C_\delta \psi_0(0)} \sum_{j=a}^{b} \frac{\Re\{W_n(s_j)\}}{s_j^{1/2}}$$

Spatial patterns: phase relationships between time series

Phase differences between time series provide a useful means of discerning evidence for spatio-temporal patterns such as travelling waves (Grenfell et al. 2001). A complex wavelet, such as the Morlet has a phase angle, defined by

$$\theta = \arctan\left[\Im\{W_n\} \Big/ \Re\{W_n\} \right]$$ where $\Re\{W_n\}$ is the imaginary part of W_n

We can use this to plot absolute phases or phase differences (where the phase difference is restricted to the range $+/-180°$, to avoid spurious 'jumps' in phase difference). The phase

difference can also be calculated from the *Cross Wavelet Spectrum* between two series (Torrence & Compo 1998).

3. An epidemiological illustration – pre-vaccination measles in New York

The wavelet spectrum (Fig. 2) for this famous monthly series (Schaffer & Kot 1985) clearly shows the annual power associated with seasonal forcing by schooling patterns, as well as major epidemic variation. The latter is highly non-stationary, showing biennial variation post-1941 and more irregular 2-3 year cycles in the 1930s. As discussed in the talk, these variations are predominantly due to secular variations in birth rate (Earn et al. 2000).

In the presented paper, we then discuss how wavelet spectra can be used to discern spatio-temporal patterns, particularly dramatic travelling waves in the measles series for England and Wales (Grenfell et al. 2001). Finally, we use simple epidemiological models to interpret these patterns.

4. Acknowledgements

This work was supported financially by the Wellcome Trust

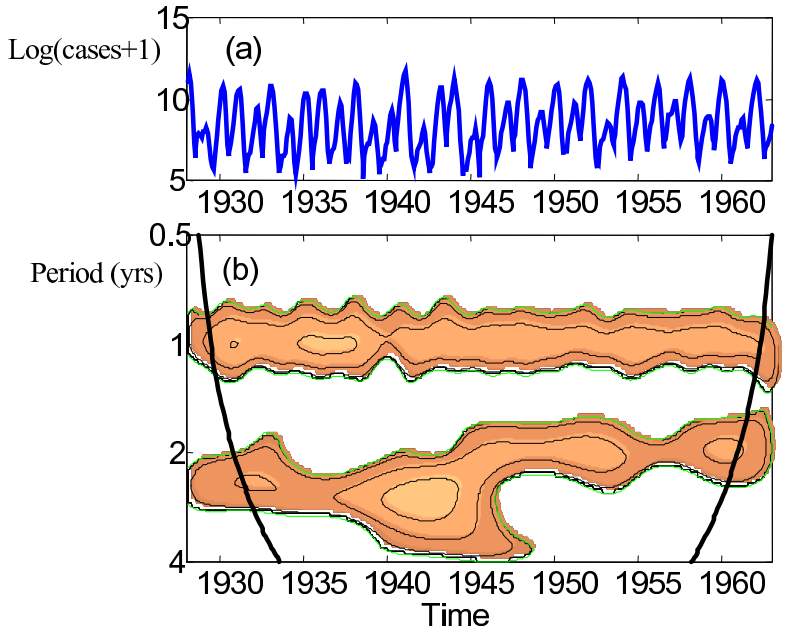


Fig. 2 (a) Log transformed pre-vaccination time series for measles in New York. (b) Wavelet spectrum for this series, as described in the methods. Colour shows wavelet power (orange=low to tan=high); white denotes no significant variation

REFERENCES

Bradshaw, G. A. & Spies, T. A. 1992 Characterizing Canopy Gap Structure in Forests Using Wavelet Analysis. *Journal of Ecology* **80**, 205-215.

Daubechies, I. 1992 *Ten lectures on wavelets*: Society for Industrial and Applied Mathematics.

Earn, D. J. D., Rohani, P., Bolker, B. M. & Grenfell, B. T. 2000 A simple model for complex dynamical transitions in epidemics. *Science* **287**, 667-670.

Grenfell, B. T., Bjørnstad, O. N. & Kappey, J. 2001 Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716-723.

Ivanov, P. C., Rosenblum, M. G., Peng, C. K., Mietus, J., Havlin, S., Stanley, H. E. & Goldberger, A. L. 1996 Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis. *Nature* **383**, 323-327.

Lau, K. M. & Weng, H. 1995 Climate signal detection using wavelet transform: How to make a time series sing. *Bulletin of the American Meteorological Society* **76**, 2391-2402.

Nason, G. P. & von Sachs, R. 1999 Wavelets in time series analysis. *Philos. Trans. R. Soc. Lond A* **357**, 2511-2526.
Schaffer, W. M. & Kot, M. 1985 Nearly one dimensional dynamics in an epidemic. *Journal of Theoretical Biology* **112**, 403-427.
Torrence, C. & Compo, G. P. 1998 A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* **79**, 61-78.

RÉSUMÉ
Les séries temporelles en épidémiologie ne sont pas en générale stationaires, mais possèdent des variations marquées temporelles ou spatialles en moyenne, en variance, et dans la periode et grandeur des oscillations. Les analyses spectroscopiques traditionelles, telles que celle de Fourier, ne se conviennent pas tres bien a l'étude de ces patternes temporeaux. Dans les domaines de physiologie et les sciences physiques, on a récemment abordé ces problèmes en développant des techniques en spectre d'ondelette, qui suivent la puissance et d'autres propriétés des oscillations dans le temps.

Ce papier se concentre sur une présentation de l'application des methods d'ondelette dans les séries temporelles épidémiologiques, en particulier pour l'ensemble bien connu de données de notification de la rougeole en Angleterre et le Pays de Gales. Dans la présentation orale, nous démontrons d'abord des variations non-stationaires dans la période entre les épidémies, et expliquons ces patternes en utilisant des modèles. Nous raffinons alors les méthodes pour explorer la dynamique spatio-temporelle, et trouvons des ondes importantes de déplacement dans l'abondance de la rougeole.

Extending the stochastic susceptible-infected -removed epidemic model to pig-production applications

Michael Höhle
Royal Veterinary and Agricultural University, Denmark

With an annual production of 23 million pigs and pig meat constituting 6.8% of the total Danish export[2, 1], surveillance and management of diseases plays a substantial role in Danish pig production [15]. In the light of the recent Foot and Mouth epidemics in England [12, 21] or classical swine fever in the Netherlands [11, 28] it becomes apparent that understanding the dynamics of an epidemic is important in order to predict the spread and evaluate the effect of control policies. On the individual farm, problems with infectious disease are although also of a more daily nature. Endemic presence of flue, pneumonia, porcine reproduction and respiratory syndrome (PRRS) cause reduced growth and increased mortality having a notable affect on production economy [24, 23]. As an example, Figure 1 shows an attempt to use visualizations of pneumonia treatments at a Danish pig production site to improve on-site health management [16]. Insight about disease dynamics is again a prerequisite for providing more advanced decision support.

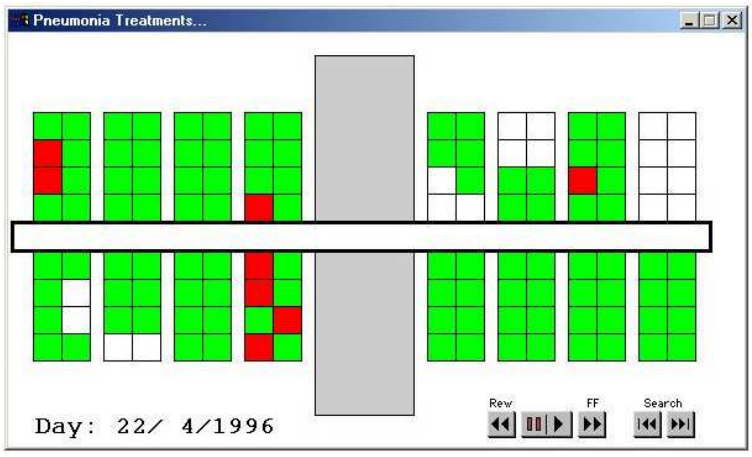


Figure 1: Screenshot from software used to visualize daily pneumonia treatments at a Danish production site. Green pens indicate presence of pigs in the pen, while red pens have one or more treatments that given day.

A tool to gain this insight is to perform a *disease transmission experiment*: In a controlled environment, one or more individuals are inoculated with virus strains after which observations about virological, serological and clinical findings are made at regular intervals for the entire population. Aim of such an experiment ranges from quantifying disease transmission [22, 29] to determining the effect of a vaccine [9, 8, 25, 10]. Data are analyzed using stochastic epidemic models like the susceptible-infected-recovered (SIR) model or the more biological plausible SEIR model, also taking disease incubation into account [7, 4]. Upon disease transmission (exposure), initially susceptible

individuals go through phases of incubation and infectiousness before recovering again. A recovery occurs once an individual is cured, dies, or in other ways cannot contribute to the spread of the disease anymore. Disease transmission is modeled by a set of Poisson contact processes and it assumed that no re-infection can occur.

Veterinary literature has focused on using the transmission experiments to estimate the so called basic reproduction ratio, R_0, a summarizing quantity of the SIR/SEIR model, telling whether the epidemic can result in large outbreaks [9, 4]. Partial observability of the epidemic, infrequent observations, etc. prohibited a more detailed analysis. Recently, a greater awareness about frequent observations, improvements in the availability and affordability of virologic test [20] have improved the data material achieved by the experiments. In parallel, methodology advances in the statistical analysis of partially observed epidemics by computer intensive methods have been made [13, 26, 14, 27]. Together this provides the opportunity to perform a more detailed analysis of the transmission experiments than before.

To apply the above statistical advances to the context of disease transmission experiments the following extensions need to be taken care of.

- Spatial layout of the confinement units (e.g. pens and sections), age, vaccination, etc. provides heterogeneity.
- Multiple data sources should be fused, e.g. virologic, serologic, and clinical observations.
- Missing information occurs both due to unobservable events, but also due to specific disease characteristics, protocol errors, or censoring.

By extending the SEIR model to a multi-type epidemic [17, 18] shows how these extension can be made, while still operating within the domain of well-investigated SEIR model. This means quantities like final size distribution [5, 3], basic reproduction number [6, 4] are readily computable for the formulated models. Estimation of model parameters by Markov Chain Monte Carlo methods in a Bayesian framework described by [26, 27] are extended to the multi-type setup and additional diagnostic tests are added by exploiting survival analysis [19]. The methods are exemplified in [18] by analyzing data from a Belgian disease transmission experiment with classical swine fever virus [10]. Obtained parameter estimates allow to answer questions about the effect of vaccination and design of interior walls for the specific test environment.

9 Acknowledgments

Thanks to Erik Jørgensen, Biometric Research Unit, Danish Institute of Agricultural Sciences, Denmark and Philip O'Neill, School of Mathematical Sciences, University of Nottingham, UK, for valuable discussions. Furthermore, Jeroen Dewulf, Faculty of Veterinary Medicine, Ghent University, Belgium, is thanked for providing additional information about the dataset in [10].

References

[1] *Annual Report 2001.* National Comittee of Pig Production, 2001. Available from `http://www.danskeslagterier.dk/`.

[2] *Statistik 2001.* Danske Slagterier, 2001. Available from `http://www.danskeslagterier.dk/`.

[3] C.L. Addy, I.M. Longini, and M. Haber. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, 47:961–974, 1991.

[4] H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*. Number 151 in Lecture notes in statistics. Springer, 2000.

[5] F. Ball. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Prob.*, 18:289–310, 1986.

[6] F. Ball and D. Clancy. The final size and severity of a generalised stochastic multitype epidemic model. *Adv. Appl. Prob.*, 25, 1993.

[7] N. G. Becker. *Analysis of Infectious Disease Data*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1989.

[8] A. Bouma, A.J. De Smit, M.C.M. De Jong, E.P. Kluijver, and R.J.M. Moormann. Determination of the onset of the herd-immunity induced by the E2 sub-unit vaccine against classical swine fever virus. *Vaccine*, 18:1374–1381, 2000.

[9] M.C.M. De Jong and T. G. Kinman. Experimental quantification of vaccine-induced reduction in virus transmission. *Vaccine*, 12:761–766, 1994.

[10] J. Dewulf, H. Laevens, F. Koenen, H. Vanderhallen, K. Mintiens, H. Deluyker, and A. de Kruif. An experimental infection with classical swine fever in E2 sub-unit marker-vaccine vaccinated and in non-vaccinated pigs. *Vaccine*, 19:475–482, 2001.

[11] A.R.W. Elbers, A. Stegeman, H. Moser, H.M. Ekker, J.A. Smak, and F.H. Pluimers and. The classical swine fever epidemic 1997-1998 in the Netherlands: descriptive epidemiology. *Preventive Veterinary Medicine*, 42(3–4):157–184, 1999.

[12] N. M. Ferguson, C. A. Donnelly, and R. M. Anderson. The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, 292:1155–1160, 11th May 2001.

[13] G. J. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics applied in Medicine & Biology*, 15:19–40, 1998.

[14] Y. Hayakawa, D. Upton, P.S.F. Yip, and P.D. O'Neill. Inference for a multitype epidemic model using Markov chain Monte Carlo methods. Technical Report 00-03, Statistics Division, School of Mathematical Sciences, University of Nottingham, 2000. Available from `http://www.maths.nottingham.ac.uk/statsdiv/research/reports.html`.

[15] Hans Houe, editor. *Annual Report 1999*. Research Centre for the Management of Animal Production and Health, 1999.

[16] M. Höhle. Decision support for pneumonia management in pig production. In *Proceedings of the 1st European Workshop on Sequential Decisions under Uncertainty in Agriculture and Natural Resources*, pages 51–56, september 2002.

[17] M. Höhle. Estimating parameters for stochastic epidemics. Research Report 102, DINA, Available from `http://www.dina.dk/~hoehle/pubs/dina102.pdf`, 2002.

[18] M. Höhle, P.D. O'Neill, and E. Jørgensen. Analysis of disease transmission experiments using stochastic epidemic models. To appear.

[19] J. G. Ibrahim, M-H. Chen, and D. Sinha. *Bayesian Survival Analysis*. Springer Series in Statistics. Springer, 2001.

[20] C.A. Janeway, P. Travers, M. Walport, and M. Shlomchik. *Immunobiology*. Garland Publishing, 5th edition, 2001.

[21] M. J. Keeling, M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell. Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294:813–817, 2001.

[22] H. Laevens, F. Koenen, H. Deluyker, and A. de Kruif. Experimental infection of slaughter pigs with classical swine fever virus: transmission of the virus, course of the disease and antibody response. *Veterinary Record*, 145:243–248, 1999.

[23] L.P. Larsen and P. Bækbo. *Sundhed og Sygdom hos Svin*. Landbrugsforlaget, 3rd edition, 1997.

[24] J.P. Nielsen, C. S. Jensen, P. Bækbo, S. E. Jorsal, V. Sørensen, and H. Houe. Mycoplasma hyopneumoniae infection - a review of pathogenesis, diagnosis, risk factors and biological effects as input for economical modelling. Technical report, Research Centre for the Management of Animal Production and Health, 2000.

[25] G. Nodelijk, M.C.M. De Jong, L.A.M.G. van Leengoed, G. Wensvoort, J.M.A. Pol, P.J.G.M Steverink, and J.H.M. Verheijden. A quantitative assessment of the effectiveness of PRRSV vaccination in pigs under experimental conditions. *Vaccine*, 19:3636–3644, 2001.

[26] P. D. O'Neill and G. O. Roberts. Bayesian inference for partially observed stochastic epidemics. *J. R. Statist. Soc.*, 162:121–129, 1999.

[27] P.D. O'Neill and N.G. Becker. Inference for an epidemic when susceptibility varies. *Biostatistics*, 2(1):99–108, 2001.

[28] A. Stegeman, A.R.W Elbers, A. Bouma, H. de Smit, and M. C. M. de Jong. Transmission of classical swine fever virus within herds during the 1997-1998 epidemic in the Netherlands. *Preventive Veterinary Medicine*, 42(3–4):201–218, 1999.

[29] K.D.C. Strk, D.U. Pfeiffer, and R.S. Morris. Within-farm spread of classical swine fever virus - a blueprint for a stochastic simulation model. *Veterinary Quarterly*, 22(1):36–43, jan 2000.

Daphnia, parasites and lake bottom dynamics

Marianne Huebner and Alan Tessier
Michigan State University

1 Introduction

To encounter planktonic hosts, many parasites rely on physical mixing to remain suspended in the water column. We consider the situation of *Daphnia dentifera* and its fungal parasite, *Metschnikowia bicuspidata* in small lakes. *Daphnia* populations establish during spring and reach highest density in mid-summer, but disapppear from the water column in winter. *Daphnia* are infected by ingestion of the fungal asci from the water. Infections are fatal and upon host death parasite are liberated. High prevalence of infection does not occur until September, and the magnitude of epidemics varies greatly (from less than 10 % to 80 %) among lakes.

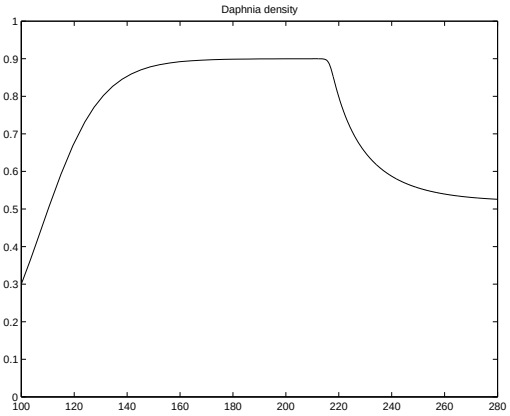


Figure 1: Model: Density of Daphnia during the season

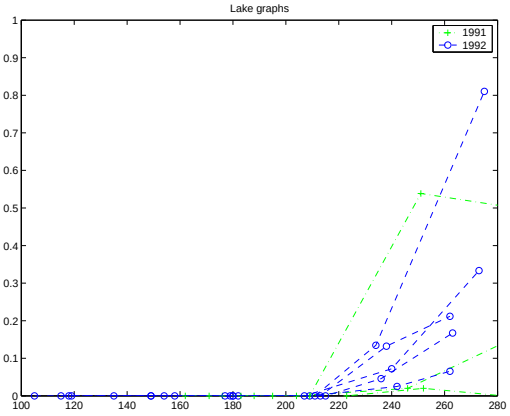


Figure 2: Data: Incidence of infection

2 The Model

The SI model describes the dynamics of susceptible (S) and infected (I) hosts. Susceptible hosts are born from both susceptible and infected Daphnia, and the population grows logistically with carrying capacity K. The rate at which susceptible host become infected is assumed to be proportional to the number of encounters between susceptible hosts and suspended parasites. It is It is important to explicitly consider the dynamics of the parasite spores. Initially, spores enter the water column via resuspension from the sediment, which is a function of the number of spores in the sediment, $G(Z_s)$. After infection, more spores are added to the water column as the infected host dies (cIf). Spores are lost from the water column through sedimentation (R). Parameters involved in the model are the transmission rate β, the birth rate b, and the mortality rate of infected hosts m_I and of susceptible hosts m_S. The mortality of the infected hosts is much higher than the mortality of the healthy Daphnia.

$$\frac{dS}{dt} = b(S+I)\left(1-\frac{S+I}{K}\right) - \beta S Z_w - m_S S \tag{1}$$

$$\frac{dI}{dt} = \beta S Z_w - m_I I \tag{2}$$

$$\frac{dZ_w}{dt} = cIf + G(Z_s) - \frac{Z_w}{R} \tag{3}$$

We can model the resuspension function $G(Z_s)$ with a compound Poisson process that adds spores to the water column in the Fall.

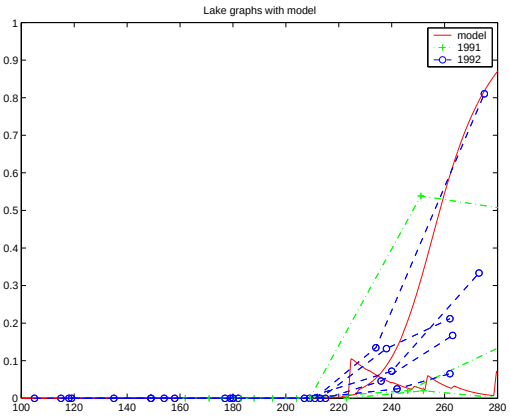


Figure 3: Incidence of infection and Simulation results

Lake characteristics determine the size of an epidemic in that particular lake. The exact process that induces resuspension in the fall is unknown. To account for the high variability of disease prevalence it is necessary to study convective motions, and temperature gradients throughout the season.

References

[1] Anderson, R.M. and May, R.M. (1981). The population dynamics of microparasites and their invertebrate hosts. *Philosophical Transacations of the Royal Society London B.* **291**: 451-524.

[2] Cacéres, C.E. (1997). Temporal variation, dormancy and coexistence: a field test of the storage effect. *Proc. Natl. Acad. Sci.* **94**: 9171-9175.

[3] Ebert, D., Payne, J.H., and and Weisser, W.W. (1997). The epidemiology of parasitic diseases in Daphnia. 91-111. *In:* K. Dettner, G. Bauer, W. Völkl (eds.). *Vertical food web interactions: Evolutionary patterns and driving forces.* Springer-Verlag, Berlin.

[4] Grenfell, B.T. and Dobson, A.P. (1995). *Ecology of infectious disease in natural populations.* Cambridge University Press.

[5] Gubbins, S. and Gilligan, C.A. (1997). Persistence of host-parasite interactions in a disturbed environment. *Journal of Theoretical Biology* **188**: 241-258.

[6] Gubbins, S., Gilligan, C.A. and Kleczkowski, A. (2000). Population dynamics of plant-parasite interactions: thresholds for invasion. *Theoretical Populations Biology* **57**: 219-233.

[7] Hakanson, L. (1982). Bottom dynamics in lakes. *Hydrobiologia* **91**: 9-22. .

[8] MacIntyre, S. (1993). Vertical mixing in a shallow eutrophic lake: possible consequences for the light climate of phytoplankton. *Limnology adn Oceanography* **38**: 798-817.

[9] Tessier, A.J., Bizina, E.V., and Geedey, C.K. (2001). Grazer-resource interactions in plankton: Are all daphniids alike? *Limnology and Oceanography* **46**: 1585-1595.

The spread of macroparasites: The effects of spatial scale and spatial clumping in the infection process

Valerie Isham
University College London

Extended abstract of talk given to the Workshop on Dynamical Stochastic Modeling in Biology, Copenhagen, January 2003.

Acknowledgements: In this talk, an overview is given of some research carried out in collaboration with Stephen Cornell, Bryan Grenfell and Julian Herbert. The considerable contributions of my co-authors are acknowledged with thanks, as is support from the Engineering and Physical Sciences Research Council and the Biotechnology and Biological Sciences Research Council in the UK, and from the European Commission (through the Research Training Network DYNSTOCH).

1 Introduction

For mathematical modellers, understanding the effects of spatial structure on the transmission dynamics of infectious diseases, and making appropriate allowance for this structure in their models, represent important challenges. This talk focuses on macroparasitic infections and is illustrated by considering models appropriate to a managed animal population and, specifically, to gastrointestinal nematodes in a herd of sheep. The occurrence of multiple infections, the need for the parasite to mate within the host in order to reproduce, and the host density, all have significant effects on the persistence/extinction of a parasite population and the invasion of parasite strains with particular genetic traits such as treatment resistance. Two particular problems are addressed: first, the effect of spatial clumping of the infection process on the dynamics of genetic resistance to anti-parasitic drugs; second, the effect of spatial scale (represented by the size of the host population) on parasite persistence. These issues are investigated by means of a mechanistic, stochastic model representing the physical processes involved, and by two simplified generic metapopulation models that seek to focus on particular aspects of the process, using a combination of analytic techniques and simulation. Further details of the models can be found in the cited references, which also contain more details of the context of the research and appropriate references to the wider literature.

In standard models for microparasite (*e.g.* bacterial, viral) infections, hosts are simply allocated to a number of classes: susceptible, latent, infected *etc.* . The justification for this is that once a host is infected, the parasites multiply rapidly within the host to reach an equilibrium level. In contrast, macroparasite infections are more complicated to model, because macroparasites have relatively long lifetimes with no direct reproduction in the host. This means that the host parasite load only increases by reinfection, and models must allow for the parasite life cycle and the parasite load of each host. The host's parasite load determines its immune reaction, which affects

its resistance to reinfection and the fertility and mortality of its parasites, as well as its ability (indirectly) to infect other hosts.

The ideal is a fully stochastic model of host and parasite dynamics. However, this is far from straightforward due to the complicated dependencies of the dynamics on the numbers of parasites in each host, and the need to model reinfection via the part of the parasite lifecycle external to the hosts. In practice, models divide into two classes: hybrid models, in which appropriate parts of the process are replaced by deterministic mechanisms, and fully stochastic models which concentrate on specific aspects of interest and make strong simplifying assumptions about the rest. Examples of the latter are to ignore the part of the parasite lifecycle external to the host and thus assume that the infection is transmitted directly from one host to another, or to ignore the feedback element of the infection process and thus assume that the infections do not depend on the current levels of hosts' infections.

2 Stochastic models for host-parasite dynamics

2.1 A mechanistic model of host-parasite dynamics

This is a model for parasite dynamics in a cohort of hosts. There is no host population dynamics,and host age and time are synonymous. It is assumed that infections are caused by a contaminated environment, and specifically that they occur in a (possibly non-homogeneous) Poisson process, independently of parasite loads. Since there is no feedback in the infection process, the parasite loads in the hosts evolve independently. An example where such an assumption is appropriate is of a cohort of new lambs put out to pasture at the start of the season, where there is some residual infection of the pasture left from the previous season. The model allows for compound infections (spatial clumping of the infection process), parasite stages with non-exponentially distributed durations, host heterogeneities, parasite-induced host mortality, and parasite-induced host immunity allowing increased mortality of parasites, reduced parasite fertility and increased resistance of the host to new infections. This is a very tractable model, and a wide range of properties can be determined analytically. Further details of the model and its properties can be found in [5] and [6].

2.2 Effects of compound infections on genetic variation

It is well-known that there is widespread resistance of gastrointestinal nematodes in sheep to antiparasitic drugs. This raises the question of how the grouping of parasites within hosts, and the multiple infection process, interact to affect the persistence or extinction of rare (eg treatment-resistant) genotypes. The model described in the previous section is used to examine this question in [1], where, again, the focus is on the early-season dynamics of a cohort of hosts. For the genetics, we assume that the trait of interest (*e.g.* susceptibility to control treatments) is represented by two alleles (s, S) at a single locus, where the rare homozygote (ss) has a selective advantage (*e.g.* there is a differential susceptibility of strains to control treatments). We need to assume a suitable distribution for the size and genetic make-up of the infecting clumps of parasites and choose a range of alternatives to illustrate possible effects. We assume

promiscuous mating *i.e.* that female parasites mate (at a fixed rate) with males chosen at random from those in same host, and determine properties such as the (random) rates of production of each type of parasite offspring.

As is to be expected, because of the need for parasites to mate within hosts, the rates of parasite offspring production are lower for this model than would be the case for corresponding deterministic (mass action) models. More importantly, it is shown that compound infections can favour persistence of rare genotypes, where the extent of this effect depends on the mix of genotypes in the infecting clumps. These results are for the parasite dynamics early in the season, when infections result from an infected environment, but it is intuitively obvious that such effects will be amplified when feedback in the infection process is included in the model. Simulation results for a model with feedback, showing the promotion of rare, recessive traits, are discussed. In this model, hosts generate infecting clumps of parasites, where the genetic mix of a particular clump reflects that of the parasites within the generating host. In such a model, there are complicated effects due to the compound infections *per se*, to inbreeding because parasites within a clump are related, and to inbreeding due to self-infection (because parasites in an infecting clump may be related to those currently within the host). The latter effect is particularly important when host populations are small. Some simple toy models are investigated (see below) that aim to distinguish these effects.

Simulations of scenarios appropriate to successive cohorts of naive lambs put on initially infected pasture, and allowing for overwintering of the parasite on the pasture, show great variation between realisations. In some, the resistant parasites saturate the host population while in others they become extinct. A strong inbreeding effect on the probabilities of these outcomes is demonstrated, by comparisons with simulations of a similar model with clumped infections but in which the genotype mix within a clump reflects that of the whole parasite population rather than that of a particular host.

A simple continous-time branching process.

The first toy model retains the clumped infections and the genetics of the mechanistic model described above, but assumes an infinite host population, so that there is no self-infection of the host. Interest focuses on the initial growth of the parasite population, so parasite mortality is ignored, and the model assumes hermaphrodite parasites with all hosts initially having two parasites so that all parasites are mated. The rare recessive genotype has no selective advantage, so that the mean allele frequencies (q, $1-q$ say) do not change and any effect on persistence of the rare homozygote is conservative. It is found that the density of these homozygotes is enhanced relative to a similar model with clumped infections but where the genotypic mix within a clump reflects that of the whole parasite population rather than that of a particular host. In particular, the proportion of these homozygotes is greater than q^2, and can be shown to be proportional to q in the limit as $q \to 0$. This enhancement increases with the clump size. Further details of this model and its properties can be found in [3].

2.3 The effects of spatial scale (cohort size).

In the mechanistic model, hosts are assumed to mix homogeneously within the cohort and, since the host density is kept fixed, the size of the cohort reflects the spatial scale of social interaction. We consider the effects of scale on the fluctuations and persistence or extinction of parasite populations and, in this section, revert to the original form of the model without genetics. Intuitively, for a small herd we expect the feedback effect of any parasite birth/death/infection event to be greater than for a large herd, extinctions to be more common (there is more scope for a 'rescue' effect in a larger host population), and parasite mating probabilities initially to be higher due to self-infection. These effects can be confirmed by simulation, where complicated interactions of stochasticity, nonlinearity and spatial scale can be seen.

A promiscuous bisexual Galton-Watson metapopulation model

The second toy model has non-overlapping discrete parasite generations, with the offspring of the mated female parasites in one generation forming the next generation of parasites. These offspring are randomly distributed among the n hosts. Again we assume promiscuous mating. The model is an extension to a metapopulation of the bisexual Galton-Watson process first proposed by [4] . The parasite population becomes extinct when there are no mated females in a generation, the number of mated females being a Markov chain. As the size of the host population (n) goes to infinity, the parasite population approaches a deterministic limit. In this limit, there is a critical threshold for the initial level of infection, above which the parasite population grows without limit and below which it goes extinct. When n is finite and the initial level of infection is above this threshold, the probability of an epidemic increases with n, as a result of the rescue effect. However, below the threshold there is a trade-off between the rescue effect and the decreased chance of parasite mating due to host self-infection. Asymptotic (large n, small n) approximations to the probability of an epidemic can be combined to give a good approximation to this probability for all n. Further details of the model and its properties can be found in [2].

3 Summary

A general mechanistic stochastic model for a cohort of hosts a) without cross-infection, and b) with feedback, has been used to explore the effects of spatial clumping of infections on the persistence of rare, recessive, genetic traits, and of spatial scale (cohort size). Complicated interactions between stochasticity, nonlinearity and spatial scale are observed, and two toy models (a continuous-time branching process model, and a discrete-time Galton-Watson process model) have been developed to investigate these further with interesting results. In particular, the strength of spatial clumping in the infection process, and the genetic mix of parasites within these clumps, is a key to the epidemic outcome, which is also strongly affected by the amount of mixing within the host population.

References

[1] Cornell, S. J., Isham, V. S. and Grenfell, B. T. (2000) Drug resistant parasites and aggregated infection—early season dynamics. *J. Math. Biol.* **41,** 341–360.

[2] Cornell, S. J. and Isham, V. S. (2004) Ultimate extinction of the promiscuous bisexual Galton-Watson metapopulation. *Austr. NZ J. Statist.* to appear.

[3] Cornell, S. J., Isham, V. S., Smith, G. and Grenfell, B. T. (2003) Spatial parasite transmission, drug resistance and the spread of rare genes. *Proc. Natl. Acac. Sci. USA,* to appear.

[4] Daley, D.J. (1968) Extinction probabilities for certain bisexual Galton-Watson branching processes. *Z. Wahr. v. Geb.* **9,** 315–322.

[5] Herbert, J. and Isham, V. (2000) On stochastic host-parasite interaction models. *J. Math. Biol.* **40**, 343–371.

[6] Isham, V. (1995) Stochastic models of host-macroparasite interaction. *Ann. Appl. Prob.* **5**, 720–740.

Sustained oscillations and time delays in gene expression of protein Hes

M. H. Jensen and K. Sneppen
University of Copenhagen

G. Tiana
University of Milano and INFN

Keywords: oscillations, Hes1, time delay

A number of genes change their expression pattern dynamically by displaying oscillations. In a few important cases these oscillations are sustained and can work as molecular clocks, as in the well known cases of the circadian clock [1] and the cell cycle [2]. In other cases the oscillations in protein expression are connected with the response to external stimuli, as reported for protein p53 after induction by DNA damage [3] or as reported in association to specificity in gene expression [4]. Recently oscillations have been observed for the Hes1 system studied in the very interesting paper [5]. The Hes1 system is particularly interesting because it is connected with cell differentiation, and the temporal oscillations of the Hes1 system may thus be associated with the formation of spatial patterns in development.

Oscillations may be obtained by a closed loop of inhibitory couplings, provided that there are at least 3 different elements [5,6]. Alternatively, it was noted in the study of the p53 network [7] that a time delay in one of the components can give rise to oscillations also in a system composed of only two species (in this case, p53 and mdm2).

We suggest that time delay can be a general mechanism which produces oscillatory responses in a more economical way than 3–species inhibitory networks do. A delay in a biological system can typically be related to transcription and translation times, and to transport between cellular compartments. An example is the Hes1 system recently examined in Ref. [5]. In this system the protein Hes1 represses the transcription of its own mRNA, and the system displays oscillations in both the concentration of the protein and of its mRNA. To explain this behavior, the authors of Ref. [5] suggest a third, hidden factor which would complete a 3–species inhibitory networks of the kind discussed in Ref. [6]. There is however no direct evidence for such a factor. Furthermore, since there is a non negligible time for transport between the cell nucleus, where the protein controls mRNA transcription and the cytoplasm, where mRNA is translated into the protein. we feel compelled to suggest a simpler scenario.

We want to test the hypothesis that Hes1 and its mRNA are sufficient ingredients to produce oscillations in the system. The equations for the concentrations $[mRNA]$

and $[Hes1]$ read

$$\begin{aligned} \frac{d[mRNA]}{dt} &= \frac{\alpha k^h}{k^h + [Hes1(t-\tau)]^h} - \frac{[mRNA(t)]}{\tau_{rna}} \\ \frac{d[Hes1]}{dt} &= \beta[mRNA(t)] - \frac{[Hes1(t)]}{\tau_{hes1}}. \end{aligned} \tag{1}$$

The meaning of these equations is that mRNA is produced at rate α when Hes1 is bound to the DNA. The probability that Hes1 is bound to DNA is $k^h/(k^h + [Hes1]^h)$, where k is a characteristic concentration for dissociation of Hes1 from the DNA, and h is the Hill coefficient that takes into account the cooperative character of the binding process. Moreover, Eqs. (1) say that mRNA undergoes degradation with characteristic time τ_{rna}, that the production

rate of Hes1 is proportional to the concentration of mRNA and that Hes1 is degraded on the time scale τ_{hes1}. Note that the terms associated with degradation in Eqs. (1) not only describe the spontaneous degradation of the protein, but also the outflow caused by the protein going to interact with other parts of the cell.

The key point is that the production of mRNA is delayed by a time τ, which takes into account the lengthy molecular processes involved in the system (translation, transcription, etc.). If one inserts the delay in the production of Hes1 (the second of Eqs. 1), instead that of mRNA, the results remains very similar to the ones reported here.

An important factor which determines the cooperativity in the production of mRNA is the fact that Hes1 is a dimer, and consequently we expect that the Hill coefficient h is of the order of 2. On the other hand, its precise value is not known. We have repeated our calculations for different values of h (i.e., $h = 1.5, 2, 4$) and found that the system displays oscillations in all cases analyzed, although the detailed features of these oscillations (e.g., those displayed in Table 1) depend on the particular choice of h. This result agrees with the fact that the physical reason which causes oscillations is not the nonlinearity of the equations but the delay. In the following we analyze in detail the case $h = 2$.

From Ref. [5], τ_{rna} and τ_{hes1} are of the order of 25 minutes. The value of the time delay is difficult to assess, since it is determined by a variety of molecular processes. One can guess that its order of magnitude is tens of minutes.

The solution of Eqs. (1) is displayed in Fig. 1. For the chosen set of parameters, the system displays damped oscillations with period $\Delta\tau \approx 170$ min and damping time $\tau_{damp} \sim 9500$ min. The dependence of $\Delta\tau$ and τ_{damp} on the delay τ is listed in Table 1. The oscillation period stays constant for low value of the delay and increases as $\tau \gg \tau_{rna}$. Also the damping time increases with τ, the oscillations becoming sustained for $\tau > 80$.

For any delay in the range $10 < \tau < 50$ min, the oscillation period is consistent with that found experimentally, and also the time difference between the peaks in Hes1 and mRNA is 18 min, similar to the experimental findings. For $\tau < 10$ min, the system shows no oscillations. To check the robustness of the results, we have varied α, β and k over 5 orders of magnitude around the basal values listed in the caption to Fig. 1, and observed no qualitative difference with the oscillatory behaviour described above. On the other hand, an increase of τ_{hes1} and τ_{rna} disrupts the oscillatory mechanism.

This is because these two quantities set the time scale of the system, with which τ has to be compared. Increasing such time scales at constant τ is equivalent to decreasing τ for a given time scale, putting the system in the low–delay part of Table 1 where no oscillations are detected.

The time delay picture gives a natural description of the Hes1 network, without the need of additional unknown factors. This is the minimal model which, nevertheless, provides a very detailed agreement with the experimental findings. The delay summarizes a number of molecular processes, such as, the time between transcription and final protein, the intracellular transport, or the time associated with the involvement of additional intermediates in the system. A striking overall result of our simulations is that the oscillatory period remains unchanged over a wide variety of values of the delay. Thus the observed time behaviour mostly depends on the degradation times, whereas it is robust to variations in other parameters. This is functionally meaningful, since degradation times can be directly controlled by changing protease activities. In general, we speculate that, more than giving a description of the system, the time delay mechanism is a tool adopted by the cell to display oscillatory behaviors in an economic and robust way, making use of as few factors as possible.

We are grateful to Eric Siggia who directed our attention to the study of the Hes1 system.

References:

1. M.P. Antoch, et al. *Cell*, **89**: 655-67 (1997).

2. B. Novak & J.J. Tyson, J. Theor. Biol. **173**, 283-305 (1995).

3. Y. Haupt, R. Maya, A. Kazaz and M. Oren, Nature **387** 296 (1997).

4. A. Hoffmann, A. Levchenko, M.L. Scott & D. Baltimore, Science **298** 1241-1245 (2002).

5. H. Hirata et al. *Science* **298** 840 (2002).

6. M.B. Elowitz, S. Leibler, *Nature* **403**, 335 (2002).

7. G. Tiana, M. H. Jensen, K. Sneppen, Europ. Phys. J. B **29**, 135 (2002).

$\tau[min]$	$\tau_{damp}[min]$	$\Delta\tau$ [min]	$\Delta\tau_{peaks}$ [min]
0	0	0	
10	450	170	18
20	500	170	18
30	870	170	18
40	1900	170	18
50	9500	170	18
80	∞	280	18
100	∞	360	18

Table 1: The damping time τ_{damp}, the oscillation period $\Delta\tau$ and the time difference between the peaks in hes1 and mRNA, as function of the delay τ. Infinite damping time means that oscillations are sustained. For $\tau = 0$ the system shows no oscillations.

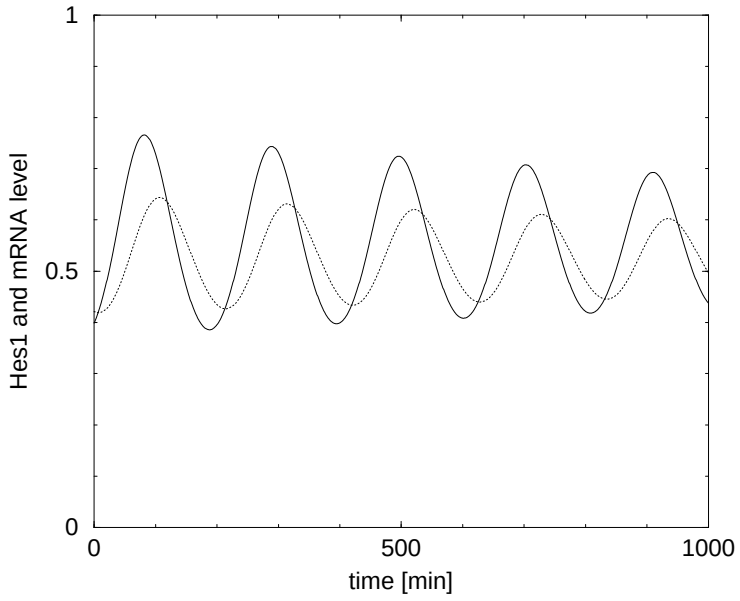


Figure 1: The oscillatory behavior of the concentration $[Hes1]$ of the protein Hes1 (dashed curve) and mRNA (solid curve), as calculated from Eq. (1). The following parameters are used: $\tau_{rna} = 24.1$ min, $\tau_{hes1} = 22.3$ min, $\alpha = 1$ $[R]_0$/min, $\beta = 0.1$ min^{-1}, $k = 0.1[R]_0$, $h = 2$, $\tau = 50$ min, and the plot show concentrations in units of $[R]_0$.

Qualitative simulation of the initiation of sporulation in *bacillus subtilis*

Hidde de Jong
Institut National de Recherche en Informatique et en Automatique (INRIA), France

1 Introduction

It is now commonly accepted that most interesting properties of an organism emerge from the interactions between its genes, proteins, metabolites, and other constituents. This implies that, in order to understand the functioning of an organism, we need to elucidate the networks of interactions involved in gene regulation, metabolism, signal transduction, and other cellular and intercellular processes.

The study of *genetic regulatory networks* has taken a qualitative leap through the use of modern genomic techniques that allow simultaneous measurement of the expression levels of all genes of an organism. In addition to experimental tools, computer tools for the *modeling* and *simulation* of gene regulation processes will be indispensable. As most networks of interest involve many genes connected through interlocking positive and negative feedback loops, an intuitive understanding of their dynamics is difficult to obtain and may lead to erroneous conclusions. Modeling and simulation tools, with a solid foundation in mathematics and computer science, allow the behavior of large and complex systems to be predicted in a systematic way [2].

Several computer tools for the simulation of biochemical reaction networks by means of differential equations are currently available. These tools can be used to simulate genetic, metabolic, and signal transduction networks described by differential equations. In addition, they allow the user to perform tasks like the analysis of steady states and the estimation of parameter values. The currently-available tools are essentially restricted to *quantitative* models of reaction networks, in the sense that numerical values for the kinetic parameters and molecular concentrations need to be specified. However, since this information is usually absent, especially in the case of systems that are not well-understood, the above-mentioned tools may be difficult to apply.

This abstract describes *Genetic Network Analyzer (GNA)*, a computer tool for the qualitative simulation of genetic regulatory networks. GNA employs piecewise-linear (PL) differential equation models that have been well-studied in mathematical biology [6, 10, 11]. While abstracting from the precise molecular mechanisms involved, the PL models capture essential aspects of gene regulation. Their simple mathematical form permits a qualitative analysis of the dynamics of the genetic regulatory systems to be carried out. Instead of numerical values for parameters and initial conditions, GNA asks the user to specify qualitative constraints on these values in the form of algebraic inequalities. Unlike precise numerical values, these constraints can usually be inferred from the experimental literature.

The use of GNA will be illustrated in the context of a regulatory network of biological interest, consisting of the genes and interactions regulating the initiation of

sporulation in the Gram-positive soil bacterium *Bacillus subtilis* [1, 7, 8]. Under conditions of nutrient deprivation, *B. subtilis* can decide not to divide and form a dormant, environmentally-resistant spore instead. The decision to either divide or sporulate is controlled by a regulatory network integrating various environmental, cell-cyle, and metabolic signals. The aim of the example is to show that GNA is able to reproduce experimental findings in the case of a large and complex network that is well-understood by molecular biologists.

2 Qualitative simulation of genetic regulatory networks

The dynamics of genetic regulatory networks can be modeled by a class of piecewise-linear differential equations of the following general form [6, 10, 11]:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x})\,\boldsymbol{x},\ \boldsymbol{x} \geq \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)'$ is a vector of cellular protein concentrations, and $\boldsymbol{f} = (f_1, \ldots, f_n)'$, $\boldsymbol{g} = \mathrm{diag}(g_1, \ldots, g_n)$. The rate of change of each concentration x_i, $1 \leq i \leq n$, is defined as the difference of the rate of synthesis $f_i(\boldsymbol{x})$ and the rate of degradation $g_i(\boldsymbol{x})\,x_i$ of the protein. The function $f_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ consists of a sum of step function expressions, each weighted by a rate parameter, which expresses the logic of gene regulation [10, 12]. The function $g_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{>0}$ is defined analogously. On a formal level, the PL models are related to a class of asynchronous logical models proposed by Thomas and colleagues [12].

Figure 1 gives an example of a simple genetic regulatory network. Genes a and b, transcribed from separate promoters, encode proteins A and B, each of which controls the expression of both genes. More specifically, proteins A and B repress gene a as well as gene b at different concentrations. Repression of the genes is achieved by binding of the proteins to regulatory sites overlapping with the promoters.

The network in figure 1 can be described by means of the following pair of state equations:

$$\dot{x}_a = \kappa_a\, s^-(x_a, \theta_a^2)\, s^-(x_b, \theta_b^1) - \gamma_a\, x_a \tag{2}$$

$$\dot{x}_b = \kappa_b\, s^-(x_a, \theta_a^1)\, s^-(x_b, \theta_b^2) - \gamma_b\, x_b. \tag{3}$$

Gene a is expressed at a rate $\kappa_a > 0$, if the concentration of protein A is below its threshold θ_a^2 and the concentration of protein B below its threshold θ_b^1, that is, if $s^-(x_a, \theta_a^2)\, s^-(x_b, \theta_b^1) = 1$. Recall that $s^-(x, \theta)$ is a step function evaluating to 1, if $x < \theta$, and to 0, if $x > \theta$. Protein A is spontaneously degraded at a rate proportional to its own concentration ($\gamma_a > 0$ is a rate constant). The state equation of gene b is interpreted analogously.

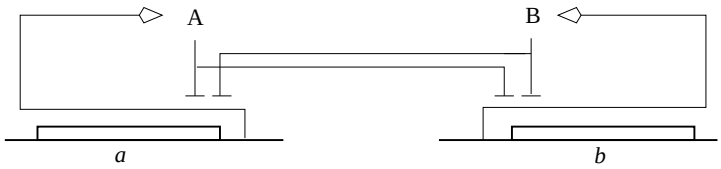


Figure 1: Example of a genetic regulatory network of two genes (a and b), each coding for a regulatory protein (A and B). The notation follows, in a somewhat simplified form, the graphical conventions proposed by Kohn [9].

Most of the time, precise numerical values for the threshold and rate parameters in the differential equations are not available. Rather than numerical values, we specify qualitative constraints on the parameter values. These constraints, having the form of algebraic inequalities, can usually be inferred from biological data. The first type of constraint is obtained by ordering the p_i threshold concentrations of gene i, yielding the *threshold inequalities*. The second type of constraint, the *equilibrium inequalities*, are obtained by ordering the quotients of production and degradation parameters with respect to the thresholds. In the example, we specify the constraints:

$$0 < \theta_a^1 < \theta_a^2 < max_a, \qquad \theta_a^2 < \kappa_a/\gamma_a < max_a, \tag{4}$$

$$0 < \theta_b^1 < \theta_b^2 < max_b, \qquad \theta_b^2 < \kappa_b/\gamma_b < max_b. \tag{5}$$

On the one hand, the parameter inequalities divide the phase space into regions where the systems behaves in a qualitatively distinct way. These regions correspond to *qualitative states* of the system. On the other hand, the parameter inequalities allow possible transitions between qualitative states to be determined by exploiting the mathematical properties of the PL models. A *qualitative simulation* consists of the generation of all qualitative states reachable through one or more transitions from a given initial qualitative state. A qualitative simulation results in a transition graph, consisting of qualitative states and transitions between qualitative states. The paths in the transition graph represent the possible qualitative behaviors predicted by the simulator [5] (figure 2).

The qualitative simulation method has been implemented in Java 1.3 in the program *Genetic Network Analyzer (GNA)* [4]. GNA is available for non-profit academic research purposes at `http://www-helix.inrialpes.fr/gna`. The core of the system is formed by the simulator, which generates a transition graph from a PL model, parameter inequalities, and an initial qualitative state. The input of the simulator is obtained by reading and parsing text files specified by the user. A graphical user interface (GUI), named *VisualGNA*, assists the user in specifying the model of a genetic regulatory network as well as in interpreting the simulation results.

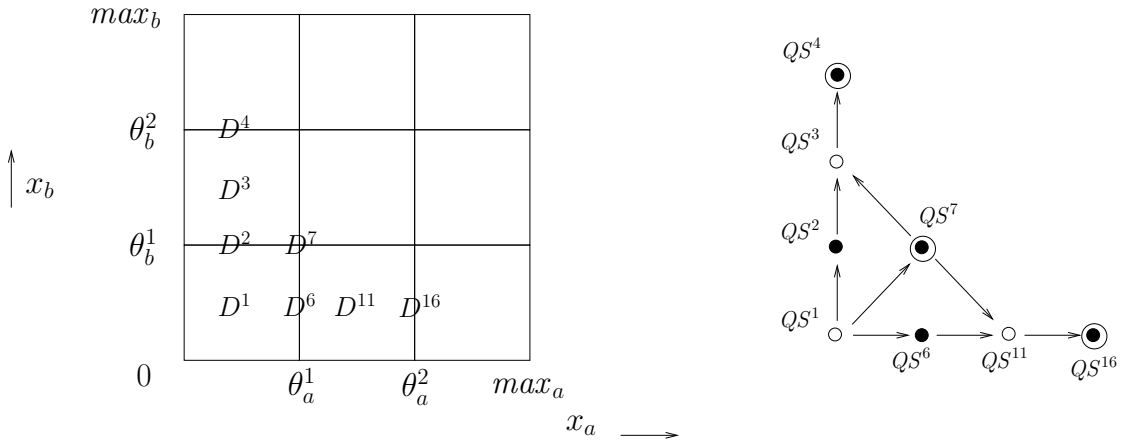


Figure 2: The left figure shows the subdivision of the phase space into regions corresponding to qualitative states, for the example network in figure 1. The right figure shows the transition graph resulting from a simulation of the network starting in the qualitative state QS^1, corresponding to the domain D^1. Qualitative states corresponding to an equilibrium of the differential equations are circled [5].

3 Initiation of sporulation in *B. subtilis*

The use of GNA can be illustrated in the context of a large and complex regulatory network of biological interest, consisting of the genes and interactions regulating the initiation of sporulation in the Gram-positive soil bacterium *Bacillus subtilis* [1, 7, 8]. Under conditions of nutrient deprivation, *B. subtilis* cells may not divide and form a dormant, environmentally-resistant spore instead. The decision to either divide or sporulate is controlled by a regulatory network integrating various environmental, cell-cyle, and metabolic signals. A graphical representation of the network is shown in figure 1, displaying key genes and their promoters, proteins encoded by the genes, and the regulatory action of the proteins (see [3] for details and references to the experimental literature).

The graphical representation of the network can be translated into a PL model supplemented by qualitative constraints on the parameters. The resulting model consists of nine state variables and two input variables. The 49 parameters are constrained by 58 parameter inequalities, the choice of which is largely determined by biological data. Simulation of the sporulation network by means of GNA reveals that essential features of the initiation of sporulation in wild-type and mutant strains of *B. subtilis* can be reproduced by means of the model [3]. In particular, the choice between vegetative growth and sporulation is seen to be determined by competing positive and negative feedback loops influencing the accumulation of the phosphorylated transcription factor Spo0A. Above a certain threshold, Spo0A∼P activates various genes whose expression commits the bacterium to sporulation, such as genes coding for sigma factors that control the alternative developmental fates of the mother cell and the spore.

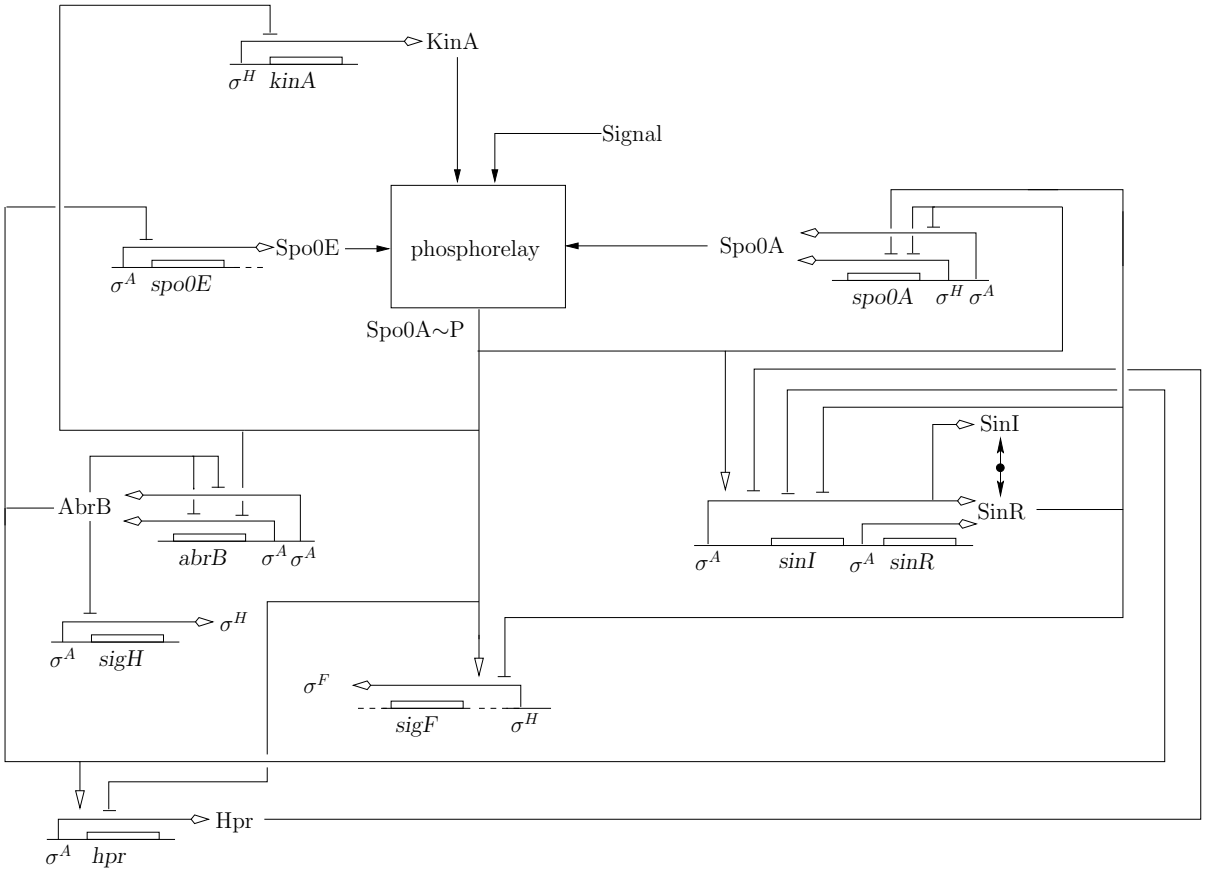


Figure 3: Key genes, proteins, and regulatory interactions making up the network involved in *B. subtilis* sporulation. In order to improve the legibility of the figure, the control of transcription by the sigma factors σ^A and σ^H has been represented implicitly, by annotating the promoter with the sigma factor in question.

4 Conclusions

We have presented the computer tool GNA for the qualitative simulation of genetic regulatory networks and illustrated its use in the analysis of the network of interactions controlling the initiation of sporulation in *B. subtilis*. GNA implements a simulation method that is based on a class of piecewise-linear (PL) differential equation models described in mathematical biology [5]. Instead of giving numerical values to the parameters and initial conditions, which are usually not available, we use qualitative constraints in the form of algebraic inequalities. These are obtained by directly translating biological data into a mathematical formalism.

References

[1] W.F. Burkholder and A.D. Grossman. Regulation of the initiation of endospore formation in *Bacillus subtilis*. In Y.V. Brun and L.J. Shimkets, editors, *Prokaryotic Development*, chapter 7, pages 151–166. ASM, 2000.

[2] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.*, 9(1):69–105, 2002.

[3] H. de Jong, J. Geiselmann, G. Batt, C. Hernandez, and M. Page. Qualitative simulation of the initiation of sporulation in *B. subtilis*. Technical Report RR-4527, INRIA, 2002.

[4] H. de Jong, J. Geiselmann, C. Hernandez, and M. Page. Genetic Network Analyzer: Qualitative simulation of genetic regulatory networks. *Bioinformatics*, 19(3):336–344, 2003.

[5] H. de Jong, J.-L. Gouzé, C. Hernandez, M. Page, T. Sari, and H. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. Technical Report RR-4407, INRIA, 2002.

[6] L. Glass and S.A. Kauffman. The logical analysis of continuous non-linear biochemical control networks. *J. Theor. Biol.*, 39:103–129, 1973.

[7] A.D. Grossman. Genetic networks controlling the initiation of sporulation and the development of genetic competence in *Bacillus subtilis*. *Ann. Rev. Genet.*, 29:477–508, 1995.

[8] J.A. Hoch. Regulation of the phosphorelay and the initiation of sporulation in *Bacillus subtilis*. *Ann. Rev. Microbiol.*, 47:441–465, 1993.

[9] K.W. Kohn. Molecular interaction maps as information organizers and simulation guides. *Chaos*, 11(1):1–14, 2001.

[10] T. Mestl, E. Plahte, and S.W. Omholt. A mathematical framework for describing and analysing gene regulatory networks. *J. Theor. Biol.*, 176:291–300, 1995.

[11] E.H. Snoussi. Qualitative dynamics of piecewise-linear differential equations: A discrete mapping approach. *Dyn. Stabil. Syst.*, 4(3-4):189–207, 1989.

[12] R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behaviour of biological regulatory networks: I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, 57(2):247–276, 1995.

Understanding the mutation mechanisms during polymerase chain reaction

Yinglei Lai and Fengzhu Sun
University of Southern California

Abstract

The polymerase chain reaction (PCR) is an important laboratory technique that uses test tubes (*in vitro*) for producing large amount of copies of a specific gene from small amount of complex molecules. Under ideal conditions, the DNA molecules generated from PCR experiments should be the same as the original molecules. However, mutations often occur during PCR. We developed mathematical models for the generation of template molecules, for point mutations of molecular templates without any repeats, and for expansion/contraction mutations for molecular templates with repeats (microsatellites). Based on the models, we developed methods to estimate the mutation rates during PCR and applied the methods to real data from PCR experiments. Important insights about the mutation mechanisms during PCR can be obtained.

Key words: branching processes, PCR, mathematical modelling, estimation methods

1 Introduction

The polymerase chain reaction (PCR) is one of the most important biotechnologies for generating a large amount of DNA molecules from a small number of, or even single, molecules [Saiki et al. 1985, Saiki et al. 1988, Scharf et al. 1986]. PCR uses the mechanism of DNA replication. In order to perform a PCR experiment, a DNA region of interest (called target) is first selected and short sequences (usually 20-25 base pairs) flanking the target must be known. The nucleotide bases at the flanking regions are used to design primers used during PCR. There are three steps in a PCR cycle. In the first step, the double-stranded DNA molecules are heated to near boiling temperature so that the double-stranded DNA molecules are separated completely into two single-stranded sequences. This process is called *denaturing.* The single-stranded sequences generated by denaturing are used as templates for the primers and the DNA polymerase. In the second step, the temperature is lowered such that the primers anneal to the templates. This process is called *annealing.* In the third step, the temperature is raised again to the temperature that is optimum for the polymerase to react. The DNA polymerases use the single-stranded sequences as templates to extend the primers that have been annealed to the templates. This process is called *polymerase extension.* The three steps form a PCR cycle. The experiment is repeated for many cycles. In ideal situations, the number of molecules containing the target doubles in every PCR cycle. However, due to a variety of reasons, such as incomplete denaturation, primer

annealing, and polymerase extension, not all the templates can generate a new copy. Suppose that in each PCR cycles, a fraction λ of templates make a complete copy. λ is called the efficiency of PCR. A standard branching process can model the generation of the templates.

PCR is not a perfect process and mutations occur during PCR. Only when a new template sequence is generated (with probability λ), mutations can occur along the newly generated sequence. When amplifying DNA molecules without any repeats, point mutations are the major source of variation. However, when amplifying DNA molecules with repeats (microsatellites), expansion/contraction mutations of one or more repeat units dominate over point mutations. The problem is to understand the mutation mechanisms for both point mutations and microsatellite mutations during PCR.

In this paper, we review the literature on the modelling and analysis of mutations during PCR and provide prospectives for future research.

2 Point mutations during PCR

Sun (1995) and Weiss and von Haeseler (1995) independently developed the first stochastic model for PCR with point mutations incorporating the randomness in the generation of template molecules and point mutations when new template sequences are generated. The model can be briefly described as follows.

2.1 Modelling the generation of template molecules

We first consider the generation of template molecules without considering mutations. In each PCR cycle, every template molecule generates a new template with probability λ independent of other templates and the template itself always stays in the pool once it is generated. igure 1 shows the mechanism of generating template molecules during PCR. The process generates a random binary tree.

It is obvious that the number of template molecules after every PCR cycle form a branching process. The expected number of template molecules generated from one template is

$$m = 2\lambda + (1 - \lambda) = 1 + \lambda.$$

Therefore, the expected number of template molecules after n PCR cycles S_n, is

$$S_n = S_0(1 + \lambda)^n,$$

where S_0 is the number of templates when the PCR experiment started.

When we consider mutations in the following sections, the templates cannot be considered as identical. For example, templates generated from the original molecules through two replications are more likely to have more mutations than template molecules generated from the original molecules through one replication. Based on this observation, Sun (1995) introduced a novel concept referred as *generation number.*

Definition: The original templates are called 0-th generation templates; the templates generated directly from 0-th generation templates are called first generation templates; the templates generated directly from the first generation templates are called

second generation templates; $\cdots$; the templates generated directly from k-th generation templates are called $k+1$-st templates, $\cdots$.

From the model for generating template molecules, it can be seen that that the expected number of k-th generation template molecules after n PCR cycles satisfy the following recursive equation,

$$S_{n+1}(k) = S_n(k) + \lambda S_n(k-1).$$

Using induction, Sun (1995) showed that $S_n(k) = S_0 \binom{n}{k} \lambda^k, \quad k = 0, 1, 2, \cdots, n.$

Sun (1995) proposed the following approximation for the probability distribution for the generation number K of a random chosen template after n PCR cycles when S_0 is large based on strong law of large numbers,

$$\Pr\{K = k\} \approx \frac{S_n(k)}{S_n} = \frac{\binom{n}{k} \lambda^k}{(1+\lambda)^n}, \quad k = 0, 1, 2, \cdots, n. \tag{1}$$

Piau (2002) recently provided an upper bound for the approximation error. From this upper bound, this approximation is still valid as long as the number of PCR cycles is large, say at least 20.

2.2 Modelling point mutations during PCR

Here we consider point mutations when amplifying template molecules without repeat units. Point mutations are superimposed onto the random binary tree as follows according to different assumptions about the mutation mechanism. Sun (1995) and Weiss and von Haeseler (1995) considered a rare mutation model assuming that point mutations occur according to a Poisson process with parameter μ when a new template is generated. That is, the number of point mutations in a target sequence of G bases is a Poisson random variable with mean μG per PCR replication. They also assumed that whenever new mutations occur, they occur at new positions (no back mutations). Let M be the number of mutations in a randomly chosen sequence after n PCR cycles. Under the above assumptions, we have

$$\{M|K = k\} \sim Poisson(k\mu G),$$

that is,

$$P\{M = m|K = k\} = \frac{(k\mu G)^m}{m!} \exp(-k\mu G), m = 0, 1, 2, \cdots. \tag{2}$$

From Equations (1) and (2), the following theorem was obtained by Sun (1995).

Theorem 1. *Let M be the number of mutations of a randomly chosen sequence after n PCR cycles. Then*

i). For any $0 \le m \le G$

$$\Pr\{M = m\} = \frac{(\mu G)^m (1 + \lambda e^{-\mu G})^n}{m!(1+\lambda)^n} E\left(Bin\left(n, \frac{\lambda e^{-\mu G}}{\lambda e^{-\mu G} + 1}\right)\right)^m.$$

$$EM = \frac{n\lambda\mu G}{1+\lambda}, \qquad Var(M) = \frac{n(\lambda\mu G)}{(1+\lambda)^2}(\mu G + 1 + \lambda).$$

ii). Suppose μ and G change with n, denoted by μ_n and G_n, such that $\lim_{n\to\infty} n\mu_n G_n = \nu$. *Then M is approximately Poisson$(\lambda\nu/(1+\lambda))$ as n tends to infinity.*

It is also of great interest to study the distribution of pairwise differences H between two randomly chosen sequences from the PCR products. In order to study the distribution of H, we first studied the distribution for the number of replications D between two randomly chosen sequences. The expectation and variance of D can be obtained [Sun 1995].

$$ED = \frac{2n\lambda}{1+\lambda} - \frac{2}{(1+\lambda)S_0 + 1 - \lambda} + O(\frac{1}{S_0(1+\lambda)^n}).$$

$$Var(D) = \frac{2n\lambda}{(1+\lambda)^2} - \frac{2(3+\lambda)}{((1+\lambda)S_0 + (1-\lambda))(1+\lambda)}$$
$$-\frac{2}{((1+\lambda)S_0 + 1 - \lambda)^2} + O(\frac{1}{S_0(1+\lambda)^n}).$$

Sun (1995) proposed the following approximate distribution of H when the number of initial molecules is large. Similar to Piau (2002), this approximation should also hold when the number of PCR cycles is relatively large, say 20, and the number of initial molecules is small.

Theorem 2. *Let H be the number of pairwise differences between two randomly chosen sequences after n PCR cycles. Let G be the target length and μ be the mutation rate per base per PCR cycle. Then*

i). The probability generating function $\varphi_H(s)$ of H is

$$\varphi_H(s) = \varphi_D(\exp(\mu G(s-1))).$$

ii). The expectation and variance of H are

$$EH = (\mu G)ED, \qquad Var(H) = (\mu G)ED + (\mu G)^2 Var(D).$$

iii). For $0 < \lambda \le 1$, $\frac{(1+\lambda)H - 2\lambda n\mu G}{\sqrt{2\lambda n\mu G(1+\lambda+\mu G)}}$ is asymptotically normal N(0,1) as $n \to \infty$.

iv). If μ and G change with n, denoted by μ_n and G_n, such that $\lim_{n\to\infty} n\mu_n G_n = \nu$, *then H is approximately Poisson$(\frac{2\lambda}{1+\lambda}\nu)$.*

Wang et al. (2000) extended the above results to situations with relatively high point mutation rates. Moore and Maranas (2000) considered different mutation rates at different positions. These extensions are very useful when we study *in vitro* evolution (Sun et al. 1996) using error prone PCR and DNA shuffling (Sun 1999).

2.3 Estimating point mutation rates during PCR

After a PCR experiment, s sequences from the PCR products are sampled and sequenced. If we know the nucleotide bases of the original molecules to be amplified, we can count the number of mutations of the sampled PCR products. Let $M_1, M_2, \cdots, M_s$

be the number of mutations of the sampled sequences. From Theorem 1, the moment estimator of the mutation rate μ is given by

$$\hat{\mu_1} = \frac{(1+\lambda)\sum_{i=1}^{s} M_i}{n\lambda G s}.$$

We also studied the variance of the above estimator. In particular, we have

Theorem 3. *Let S_0 be the initial number of sequences. Then for $\lambda = 1$,*

$$Var(\sum_{i=1}^{s} M_i) = \frac{sn\mu G}{4}(\mu G + 2) + \binom{s}{2}\frac{\mu G}{S_0}(1 - 2^{-n}).$$

For $0 < \lambda < 1$,

$$\lim_{S_0 \to \infty} S_0 \left\{ Var(\sum_{i=1}^{s} M_i) - \frac{sn\lambda\mu G}{(1+\lambda)^2}(\mu G + 1 + \lambda) \right\} = sA + 2\binom{s}{2}B,$$

where

$$\begin{aligned} A &= -\frac{(1-\lambda)\mu G}{(1+\lambda)^2}\left(1 + \frac{(1-\lambda)\mu G}{1+\lambda}\right)(1 - (1+\lambda)^{-n}), \\ B &= \frac{n\lambda^2\mu G}{(1+\lambda)^{n+2}} + \frac{\mu G}{(1+\lambda)^2}\left(2 - \frac{n\lambda + 2}{(1+\lambda)^n}\right)\left(1 + \frac{(1-\lambda)\mu G}{1+\lambda}\right). \end{aligned}$$

If the nucleotide bases of the target molecules to be amplified are not known, we can compare the pairwise differences among the sampled molecules. Let $H_{i,j}$ be the pairwise Hamming distance—the number of different bases between sequence i and sequence j. We proposed to estimate the mutation rate using

$$\hat{\mu_2} = \frac{\sum_{i \neq j,\ i,j=1}^{s} H_{i,j}}{\binom{s}{2} ED \times G},$$

where $ED = \frac{2n\lambda}{1+\lambda} - \frac{2}{(1+\lambda)S_0 + 1 - \lambda} + O(\frac{1}{S_0(1+\lambda)^n})$, S_0 is the initial number of molecules.

Weiss and von Haeseler (1997) proposed a maximum likelihood approach using extensive simulations to maximize the probability of observing the total number of mutations. Wang et al. (2000) extended the above two moment based estimators to situations with relatively high mutation rate. We expected that this maximum likelihood based approach should perform significantly better than the simple moment estimation method given above. However, extensive simulation studies by Wang et al. (2000) showed the following results. (1) When the mutation rate μ is relatively low, say less than 10^{-3} per base per PCR cycle, the four methods gave roughly the same results when the initial number of molecules is relatively large, say at least 100. (2) When the number of initial sequences is small (≤ 10), MLE does not perform as well as the other three methods. (3) When the mutation rate is relatively high, such as greater than 5×10^{-3} per base per PCR cycle, the moment method of Sun (1995) and the MLE method of Weiss and von Haeseler (1997) underestimate the mutation rate, while the two methods developed in Wang et al. (2000) approximate the true mutation

rate. (4) The moment method based on the total number of mutations in the sampled sequences and the moment method based on pairwise differences in Wang et al. (2000) have roughly the same accuracy in all the situations considered indicating we do not need to know the exact nucleotide bases of the original molecules to accurately estimate the mutation rate.

The surprisingly good performance of the moment based estimation methods prompted Piau (2002) to look into the theoretical issues related to the moments and he provided theoretical bases for the observed simulation results. Based on our simulation results, we suggest the use of the modified moment based estimation methods in practice.

Another development in the modelling of point mutations during PCR was given by Moore and Maranas (2000) where they considered the situation that mutation rates might be different at different positions.

3 Microsatellite mutations during PCR

Microsatellites are tandem repeats of DNA sequences. For example, $(CA)_6$ indicates *CACACACACACA*, the motif CA repeated six times. Microsatellite markers are very common and highly polymorphic in the human genome as well as in genomes of other organisms. They are widely used in many genetic studies including population genetics, forensics, linkage and association studies for human diseases. Weber and May (1993) observed that the final PCR products of microsatellites starting from molecules with the same number of repeat units can have different number of repeat units. This indicates that, during each PCR cycle, one or more repeat units can be inserted (expansion) or deleted (contraction), referred as slippage mutations. During PCR amplification of microsatellites, slippage mutations dominate over point mutations. Due to slippage mutations, in addition to the main band, several minor bands are often observed after PCR amplification, referred as stutter patterns or stutter profiles. The presence of stutter patterns in PCR products can cause problems in assigning alleles for genotyping. Miller and Yuan (1997) first studied the mutation mechanisms of microsatellites during PCR. In their study, they assumed that microsatellites with different numbers of repeat units have the same mutation rate, an assumption that were clearly not supported by experimental data.

In an effort to understand the mutation mechanisms of microsatellites during PCR, Shinde et al. (2003) PCR amplified single molecules with different numbers of repeat units for poly-A and poly-CA. For details of experimental conditions, see Shinde et al. (2003). The complete data can be downloaded from www-hto.usc.edu/~ fsun. The experimental data clearly indicate that the range of the number of repeat units in the final PCR products increases with the number of repeat units of the original single molecule. Thus, we expect that the mutation rate of a template per PCR cycle increases with the number of repeat units. To better understand the relationship between mutation rate and the number of repeat units, Lai et al. (2003) and Lai and Sun (2003) developed mathematical models and estimation methods for the mutation mechanisms of microsatellites during PCR and applied to the data in Shinde et al. (2003). Here, we summarize the main results from these studies.

3.1 Modelling microsatellite mutations during PCR

The model for template generation is the same as discussed in Section 2 except we assumed that PCR efficiency depends on the cycle number. For microsatellite mutations, we assumed that when a new template is generated from a parent template with j repeat units, it has a probability μ_j of being mutated. Given a mutation occurs, the probability that a repeat unit is inserted (expansion) is e and the probability that a repeat unit is deleted is $1-e$.

Let λ_n be the efficiency during the n-th PCR cycle and $S(n)$ be the expected number of template molecules after n PCR cycles. Then we have the following recursive equation.

$$S(n) = (1+\lambda_n)S(n-1). \tag{3}$$

The expected number of template molecules with j repeat units after n PCR cycles, $S_j(n), j, n = 1, 2, \cdots$, satisfy the following recursive equation

$$S_j(n) = S_j(n-1)+S_j(n-1)\lambda_n(1-\mu_j)+S_{j-1}(n-1)\lambda_n\mu_{j-1}e+S_{j+1}(n-1)\lambda_n\mu_{j+1}(1-e). \tag{4}$$

The above equation can be understood as follows. The first term is the number of templates with j repeat units after the $n-1$-st cycle. The second term is the expected number of newly generated templates from parent templates of j repeat units with no mutations. The third term is the expected number of templates newly generated from parent templates of $j-1$ repeat units with one repeat unit inserted during the n-th cycle. The last term is the expected number of templates newly generated from parent templates of $j+1$ repeat units with one repeat unit deleted during the n-th cycle.

Using the general theory of mean field approximation [Lai et al. 2003, Lai and Sun 2003], we showed that when the number of PCR cycles is relatively large, the fraction of molecules with j repeat units after n PCR cycles can be approximated by $f_j(n) = S_j(n)/S(n)$. From Equations (3, 4), we can find a recursive equation for $f_j(n)$,

$$f_j(n) = f_j(n-1)\left(1-\frac{\lambda_n\mu_j}{1+\lambda_n}\right)+f_{j-1}(n-1)\frac{\lambda_n\mu_{j-1}e}{1+\lambda_n}+f_{j+1}(n-1)\frac{\lambda_n\mu_{j+1}(1-e)}{1+\lambda_n}. \tag{5}$$

3.2 Estimating microsatellite mutation rates

We proposed a quasi-likelihood approach for estimating the mutation rates and expansion rate. Let I indicate the set of experiments. For each $i \in I$, let $o_j^{(i)}$ be the observed fraction of molecules with j repeat units in the i-th experiment. Let $f_j^{(i)}$ be the theoretical value of the observed fraction of molecules with j repeat units in the i-th experiment calculated from Equation (5). The quasi-likelihood of the data is defined as

$$L(\mu, e) = \prod_{i\in I}\prod_{j\in J}(f_j^{(i)})^{o_j^{(i)}}, \tag{6}$$

where J is the range of repeat units of interest. The maximization of the above equation is achieved using the Kiefer-Wolfowitz stochastic approximation algorithm (Kiefer and Wolfowitz 1952).

Lai et al. (2003) first studied the accuracy of the proposed quasi-likelihood approach under various mutation models using simulations. In the simulations, we simulated as

closely as possible to their experimental conditions: starting from single molecules with the same number of repeat units, the same number of PCR cycles, and the same PCR efficiency as estimated from real PCR experiments. The differences between the simulations are the different function forms for the mutation rate and the number of repeat units. We showed that the quasi-likelihood approach proposed above can accurately recover the relationship between the mutation rate and the number of repeat units. We then applied the quasi-likelihood approach to the real data and an approximate linear relationship between the mutation rate and the number of repeat units was observed. We then fitted a linear model for the mutation rate. For poly-CA, we obtained

$$\mu_j = 3.60 \times 10^{-3} \times (j - 4) - 4.09 \times 10^{-4}, \tag{7}$$

and for poly-A, we obtained

$$\mu_j = 1.52 \times 10^{-2} \times (j - 8) - 2.30 \times 10^{-3}. \tag{8}$$

The probability of expansion was estimated at 0.068 and 0.158 for poly-CA and poly-A, respectively. We noted that there might be a threshold effect such that the mutation rate is very small when the number of repeat units is small. The threshold for poly-CA was estimated at 4 repeat units (8 based) and the threshold for poly-A was estimated at 8 (also 8 bases). The biological implications of our findings can be found in Shinde et al. (2003).

4 Future research

Despite the extensive research on the modelling and estimation of point mutations during PCR, many problems remain to be studied. The model for the generation of template molecules described above is similar to the coalescent model in population genetics [Griffiths and Tavaré 1994] although the two models are different. In coalescent theory, a sample of individuals are traced back to their most recent common ancestor forming a random binary tree. In every coalescence, two individuals from the current sample are selected and they coalescence into one individual. The time for the coalescence depend on the size of the current sample and the population histories. Once a random tree is generated, mutations are then superimposed onto the random tree. One of the most important problems in population genetics is to estimate the population mutation rate during evolution. Significant research work has been carried out to estimate the mutation rates during population evolution based on Markov Chain Monte-Carlo (MCMC) approaches. The idea of the MLE approach of Weiss and von Haeseler (1997) is similar to the idea in coalescence approach. However, it is not a complete maximum likelihood of the all the data. Instead, it is the maximum likelihood of the total number of observed mutations in all the sample. Is it possible to design a complete maximum likelihood approach for all the data, not just for the total number of mutations?

Another commonly used methods for estimating mutation rate is based on the famous Luria-Delbrück distribution [Luria and Delbrück 1943]. However, no connections have been made between the Luria-Delbrück distribution with mutations during PCR.

For point mutations during PCR, the mutation rate during PCR is above the range of mutation rates that the famous theory of Luria-Delbrück distribution is practical. For microsatellite mutations during PCR, no corresponding theories exist. One of the important questions is to develop a type of analysis similar to the Luria-Delbrück approach.

Acknowledgements

The work summarized here was inspired by Drs. Michael Waterman and Norman Arnheim of the University of Southern California. Their consistent encouragements and ideas into problems related to PCR are greatly appreciated.

REFERENCES

Griffiths, R. C. and Tavaré, S. (1994) Ancestral inference in population genetics. *Statistical Science*, **9** 307-319.

Kiefer, J and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.*, **23** 462-466.

Lai, Y. L., Shinde, D., Sun, F.Z. and Arnheim, N. (2003) The mutation process of microsatellites during the polymerase chain reaction. *J. Comp. Biol.* **10** 143-155.

Lai, Y. L. and Sun, F. Z. (2003) Microsatellite mutations during the polymerase chain reaction: mean field approximations and their applications. *J. Theor. Biol.*, in press.

Luria, S. E. and Delbrück, M. (1943) Mutations of bacteria from virus sensitivity to virus resistance, *Genetics*, **28** 491-511.

Miller, M. J. and Yuan, B. Z. (1997). Semiautomatic resolution of overlapping stutter patterns in genotyping microsatellite analysis. *Anaytical Biochemistry*, **251** 50-56.

Moore, G. L. and Maranas, C. D. (2000) Modeling DNA mutation and recombination for directed evolution experiments, *J. Theor. Biol.*, **205** 483-503.

Piau, D. (2002) Mutation-replication statistics of polymerase chain reactions. *J. Comp. Biol.*, **9** 831-847.

Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230** 1350-1354.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., et al. (1988) Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239** 487-491.

Scharf, S. J., Horn, G. T. and Erlich, H. A. (1986) Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*, **223** 1076-1078.

Shinde, D., Lai, Y., Sun, F.Z. and Arnheim N. (2003) *Taq* DNA Polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: $(CA/GT)_n$ and $(A/T)_n$ microsatellites. *Nucleic. Acid. Res.*, **31** 974-980.

Sun, F. Z. (1995) The polymerase chain reaction and branching processes. *J. Comp. Biol.*, **2** 63-85.

Sun, F. Z. (1999) Modeling DNA shuffling. *J. Comp. Biol.*, **6** 77-90.

Sun, F. Z., Galas, D. and Waterman, M. S. (1996) A mathematical analysis of in vitro molecular selection-amplification. *J. Mol. Biol.*, **258** 650-660.

Wang, D., Zhao, C., Cheng, R. and Sun, F. Z. (2000) Estimation of the mutation rate during error-prone polymerase chain reaction. *J. Comp. Biol.*, **7** 143-158.

Weber, J and May, P (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, **44** 388-396.

Weiss, G and von Haeseler, A (1995). Modeling the polymerase chain reaction. *J. Comp. Biol.*, **2** 49-61.

Weiss, G. and von Haeseler, A. (1997) A coalescent approach to the polymerase chain reaction. *Nucleic. Acid. Res.*, **25** 3082-3087.

A non-linear deconvolution problem coming from corn pollen dispersal estimation

Catherine Larédo
Universités Paris 6-Paris 7 et INRA, France

Etienne Klein
Institut National Agronomique (INRA), France

There is an interest in studying pollen dispersal in many biological fields: population genetics and ecology (to study gene flows and structure of populations), pollination biology (to study the reproductive systems and hybridization), agronomy (for the genetic purity of crops and seeds), paloecology (reconstruction of past vegetation patterns). Recently, studies concerning pollen dispersals have been strengthened by the development of Genetically Modified organisms (G.M.O.) and debates related to their large scale cultivation; an important question is for instance the escape of transgenes by pollen.

We first present a global framework to study corn pollen dispersal together with some general questions concerning wind dispersed models (Section 1). We detail in Section 2 the construction and the various assumptions we need in order to build mechanistic models of pollen dispersal. Section 3 is devoted to the presentation of the experiments, statistical modeling and analysis of the data. Indeed, the statistical model is a non linear deconvolution problem, and thus it is difficult to use classical approaches to solve it. This difficulty is overcome here by building parametric models that we have named "Quasi-mechanistic models" involving various approaches. We estimate the parameters on two field experiments. We discuss in Section 4 the results, and some perspectives for future studies

This short paper is based on a joint work with C. Lavigne and P.H. Gouyon (Laboratoire d'Ecologie et Systematique, Université Paris-Sud, Orsay, France), and X. Foueillassar (A.G.P.M., Maize producer Association) that is to appear in Ecological Monographs(Klein et al, 2003).

1 Wind dispersed Models

Corn pollen is dispersed by wind. Therefore, a pollen grain may be assimilated to a particle, and its study arises similar questions and methods than spore dispersal, seed dispersal by wind or pollutant dispersal. There are two general concerns about this dispersal:

- What is the global shape of dispersal curves over long distances?
- How is it possible to get precise quantitative knowledge to make predictions about the levels of pollen flows between two fields, a field and feral populations, a field and wild relatives.

We address here the second question. Without data over long distances, it is just possible to interpolate dispersion curves without being able to assess the goodness-of-fit of the statistical data analysis.

There are two basic approaches to measure pollen dispersal: The first one studies the physical dispersal of pollen grains by setting pollen traps at various distances; the second one consists in studying pollen dispersal by measuring presence of a genetic marker in progeny (this is the case of many experiments). We have followed this second approach.

The pollen source consists in a patch of plants homozygous for a monogenic dominant marker; the pollen receptors consist in a larger patch containing plants homozygous for the recessive allele. Therefore, the presence of the marker in offspring means an occurence of efficient pollen dispersal and pollination. Now, there are two different pollination functions to be considered.

- **The backward dispersal function** i.e. the proportion of ovules at a given distance that are fertilized by the source (marked). This describes the pollen cloud composition above a plant. This is directly observed from the experiments. It is very sensitive to the experimental design (size, shape, position).

- **The forward dispersal function**, i.e. the proportion of the source plants that fertilize ovules at a given distance. This is not directly observable (since it is impossible to track pollen emitted by one plant). But it is easier to model by mechanistic approaches. Moreover, it provides robust measures of dispersal since it is useful to predict pollen clouds and gene movements.

Now, there are usually two kinds of approaches. The first one consists in **Empirical models** used to fit experimental data and chosen for their mathematical simplicity. The second one consists in **Mechanistic models**, that operate at the scale of a pollen grain. They are obtained by modeling physical phenomena (air flow, emission conditions,transport deposition), and thus include numerous parameters. There are only used to achieve predictions with physical measures of parameters. There are not used to fit dispersal data. They all present this drawback: they present too many parameters to be estimated and the measures of these parameters are impractical in natural conditions

Here, we adopt an intermediate approach that we name **Quasi-mechanistic Models**, where we consider only a few major phenomena. Models are simple enough to be fitted to experimental data but sophisticated enough to include parameters having a physical meaning.

2 Quasi-Mechanistic Models

Our concern is to model forward pollination. It is defined as the probability that a pollen grain emitted at point $(0, 0)$ falls and fecundates a target plant located at (x, y). We denote it by $\{\gamma(x, y), (x, y) \in R^2\}$. This is a two-dimensional probability distribution density, that models the **efficient pollen dispersal function**.

We will see later that backward pollination (i.e. what is observed) can be obtained from the forward pollination by means of noisy observations of a non linear convolution. These models include the following parameters: the difference in height between male and female flowers, the settling velocity, the mean wind intensity, the turbulence (simplified). These models are an extension of Tufto et al (1997). They are related to models for turbulence data (Barndorff-Nielsen et al (1978), and used later on for stochastic volatility modelling (Barndorff-Nielsen, 1997). Two main phenomena are considered:

- Paths of pollen grains $\{(X_t, Y_t, Z_t), t > 0\}$
- Pollination times (random variables on R^+ denoted by T.

The forward pollination is then derived as follows.

Proposition: *Assume that T is almost surely finite. Then, the forward pollination function $\gamma(x, y)dxdy$ is obtained as the marginal distribution on R^2 of the random process (X_T, Y_T).*

2.1 Models for individual pollen dispersal

Let us now detail the various models for pollination.

PATHS OF POLLEN GRAINS. The simplest model is to assume that the path (X_t, Y_t, Z_t) of a pollen grain is well enough approximated by a 3-dimensional Brownian motion with drift,

$dX_t = \mu_x dt + \sigma_x dB^1{}_t$, $X_0 = 0$
$dY_t = \mu_y dt + \sigma_y dB^2{}_t$, $Y_0 = 0$
$dZ_t = \mu_z dt + \sigma_z dB^3{}_t$, $Z_0 = b > 0$.

We take into account the mean wind velocity (represented by parameters (μ_x, μ_y)), and gravity $\mu_z < 0$. The atmospheric turbulence is roughly modeled by $\sigma_x, \sigma_y, \sigma_z$. Here, all these drift and diffusions coefficients are assumed to be constant, and thus are parameters to be estimated. Moreover, the three Brownian motions B_t^1, B_t^2, B_t^3 are assumed independent. In particular, it implies that (Z_t) and (X_t, Y_t) are independent.

POLLINATION TIMES. This is the time where the pollen grain stops its course on a female flower. It is a random variable denoted by T. We assume that female flowers are located at a height **b'** strictly smaller than the height **b** of male plant. We denote by **h = b-b'** this difference (note that $h > 0$).

Three cases can be distinguished:

(i) Vegetation dominance and Exponential hitting times.
The vegetation is the main factor to stop pollen. In Tufto et al(1997), T is an exponential distribution with parameter λ, independent of (X_t, Y_t, Z_t). We consider here a

more realistic model, taking into account only pollen grains that participate to pollination. The density of pollination time T_1 is obtained as the conditional density of T on R^+ with respect to the event $Z_T = h$. It is equal, setting $\lambda'_z = 2 + \mu_z^2/\sigma_z^2$, for positive t,

$$f_1(t) = \frac{\lambda'_z}{\sqrt{2\pi}} \exp(\lambda'_z h/\sigma_z)\, t^{-1/2} \exp -(\lambda_z^2\, t + \frac{h^2}{2\sigma_z^2}\frac{1}{t}).$$

(ii) Ground dominance and Inverse Gaussian hitting times.
This is another model for pollination times, assuming that a pollen grain fertilizes an ovule when ,starting from $Z_0 = b$, it first reaches the height $Z = b'$. The pollination time T is a stopping time defined by

$$T_2 = \inf\{t > 0, Z_t = b'\} = \inf\{t > 0, Z_t - Z_0 = -h\}\ .$$

Its density is the Inverse Gaussian Distribution, defined for positive t,

$$f_2(t) = \frac{1}{\sigma_z\sqrt{2\pi}}\, h \exp(\frac{h|\mu_z|}{\sigma_z^2})\, t^{-3/2}\ \exp -(\frac{\mu_z^2}{2\sigma_z^2}\, t + \frac{h^2}{2\sigma_z^2}\frac{1}{t}).$$

(iii) Intermediate position and generalized Inverse Gaussian Distribution.
The two densities $f_1(t)$, $f_2(t)$ appear very similar except for the factor $-\alpha$ in front of t in the exponential term ($\alpha = 1/2$ in f_1 and $3/2$ in f_2). This leads us to propose a more general model including the two above, which allows to moderate but not eliminate the influence of vegetation in pollination times,

$$f_3(t) = \frac{1}{I(\alpha)}\, t^{-\alpha}\ \exp -(\frac{\mu_z^2}{2\sigma_z^2}\, t + \frac{h^2}{2\sigma_z^2}\frac{1}{t}).$$

The normalizing constant I depends on the modified Bessel functions of the third kind K_ν . It is equal to,

$$I(\alpha) = 2|\frac{h}{\mu_z}|^{1-\alpha}\ K_{1-\alpha}(\frac{h|\mu_z|}{\sigma_z^2}).$$

Joining now these models for pollen grains paths and pollination times leads to parametric families for the forward dispersal function. Pollination times are independent of the 2-dimensional process (X_t, Y_t) (this is an assumption in cases 1 and 3; it is a consequence in case 2) since $B_3(t)$ is independent of (X_t, Y_t). Thus, we can derive explicitely the probability density on R^2 of (X_T, Y_T).

Theorem (see Barndorff-Nielsen et al., 1978). *If T follows a Generalized Inverse Gaussian distribution (G.I.G.) with $1/2 \leq \alpha \leq 3/2$ and (X_t, Y_t) a Brownian motion with drift independent of T , then the marginal distribution of (X_T, Y_T) is a Generalized Hyperbolic Distribution (G.H.D.).*

Therefore, the forward dispersal function is a G.H.D distribution, whose parameters depends on $(\mu_x, \mu_y, \mu_z, \sigma_x, \sigma_y, \sigma_z)$, the difference in height $h = b - b'$ and α. Clearly, all these parameters are defined up to a scaling, as it appears in the analytic expression of the G.H.D. distribution,

$$\gamma_{GHD}(x, y) = \frac{\lambda_z^{1-\alpha}\delta_x\delta_y}{2\pi} \times \frac{p^{\alpha/2}}{q(x, y)^{\alpha/2}} \times \frac{K_\alpha(\sqrt{p}q(x, y))}{K_{1-\alpha}(\lambda_z)}\ \exp(\delta_x\lambda_x x + \delta_y\lambda_y y),$$

with
$\lambda_z = \frac{|\mu_z|h}{\sigma_z^2},\ \lambda_x = \frac{\mu_x h}{\sigma_x \sigma_y},\ \lambda_y = \frac{\mu_y h}{\sigma_y \sigma_z},\ \delta_x = \frac{\sigma_z}{h\sigma_x}\ \delta_y = \frac{\sigma_z}{h\sigma_y},\ p = \lambda_x^2 + \lambda_y^2 + \lambda_z^2$

$$q(x,y) = 1 + \delta_x^2 x^2 + \delta_y^2 y^2.$$

Denote by $\theta \in \Theta \subset R^6$ the parameters $\theta = (\alpha, \lambda_x, \lambda_y, \lambda_z, \delta_x, \delta_y)$, where $\Theta = [1/2, 3/2] \times R \times R \times R^{+*} \times R^{+*} \times R^{+*}$. The problem reduces now in estimating these parameters from the experimental data.

3 Statistical estimation from the observations

3.1 The experiments

They have been performed with Maize producers (AGPM, X. Fouillessar). Two experiments have been studied.

The first one is a corn production for grains in Montargis (located near Orleans, in the center of France). A corn field 120mx120m contained a square patch (20m x20m) of plants homozygous for a dominant marker (blue coloured maize) which was surrounded by a field of plants homozygous for the absence of marker (yellow-coloured grains). After the pollination period, data were collected. According to their position with respect to the blue path, yellow maize was pollinated by pollen coming from both yellow and blue maize. Roughly speaking, above a given plant, there is a pollen cloud, whose composition determines its pollination: Each plant yields an ear. On each ear located at (x,y) in a finite 2-dimensional grid (i.e. the locations of cultivated yellow maize) the number of blue grains was counted, together with the total number of grains. The number of ears that have been analyzed is K=3063. The second experiment is a corn production for seeds, located in Messanges, near Biarritz which is a coastal area in the south west of France. A central plot of 20mx20m of blue maize was surrounded by a field 135mx135m being sown with yellow maize. The number of analyzed ears is K= 1860. Thus, these two experiments are quite similar, although meteorogical conditions are different (very windy in the second one).

Now, the observations are here related to the backward pollination function, since the data describe the proportion of pollen originated from the marked source at point (x,y) in the pollen cloud.

3.2 Relation between forward and backward pollination

Denote by $\mu(x,y)$ the result at (x,y) of individual pollen dispersal from many plants (marked and non-marked). It describes the composition of the pollen cloud above a plant located at (x,y) (marked pollen $\mu(x,y)$, nonmarked pollen $(1-\mu(x,y))$. This is the backward dispersal function. It takes into account numbers and respective positions of all the plants). We assume that

- all the plants possess the same individual pollen dispersal function;
- both marked and non-marked plants produce the same amount of pollen;

- both pollen types are equally efficient.

These are quite natural assumptions for these experiments.

Let B denote the set of indexes for locations of the central plot with blue grains maize and J the set of indexes for yellow plants locations. Then the function μ has the expression

$$\mu(x,y) = \frac{\Sigma_{k\in B}\ \gamma(x-x_k, y-y_k)}{\Sigma_{k\in B}\gamma(x-x_k,y-y_k) + \Sigma_{k\in J}\gamma(x-x_k,y-y_k)}$$

An equivalent formula with convolution products is

$$\mu(x,y) = \frac{(\gamma \star 1_B)(x,y)}{(\gamma \star 1_F)(x,y)} \quad with \ \ F = B \cup J$$

3.3 The statistical model

We have now to retrieve the individual dispersal function γ from the noisy observations of the backward dispersal function μ.

Set $z = (x,y)$. Maize ears are cropped on a non regular sampling grid ($z_k = (x_k, y_k) \in R^2, k \in K$). Let N_z denote the total number of grains on an ear located at z, and n_z the number of blue grains on this ear.

Since we are studying count data, it is natural to assume that n_z follows a binomial distribution $Bin(N_z, \mu(z))$. Thus, the statistical model for the observations is

$$n_z = N_z\mu(z) + \epsilon_z\ ,$$

where the random variables ϵ_z are independent, centered with variance

$$Var(\epsilon_z) = Var(n_z) = N_z\mu(z)(1-\mu(z)).$$

Therefore, recovering the forward dispersal function from the observations is indeed a non linear deconvolution problem with a specific structure for the errors. It is a non standard, ill-posed problem, quite difficult to solve using non parametric approaches.

Here, we have been able, in this case of wind pollination, to greatly reduce the problem building parametric models and so, we just have now to solve a parametric inference problem.

3.4 Parametric inference for $(\gamma_\theta(x,y)\ , \theta \in \Theta)$

The number of blue grains n_z on an ear located at z follows a binomial distribution $Bin(N_z, \mu(z))$, the exact likelihood $L(\theta; \dots)$ writes down, given the K observations,

$$Log\ L(\theta; n_1, n_2, \dots, n_K) = \sum_{z=1\dots K} Log\ P_\theta(n_z)$$

$$P_\theta(n_z = i) = C_{N_z}^{n_z}\ \mu(\theta;z)^i\ (1-\mu(\theta;z))^{N_z-i}.$$

Given the data, one has to find $\theta \in \Theta$ maximizing $Log\ L$,

$$\hat{\theta} = argsup\ \{\sum_{z=1\dots K} n_z Log\ \mu(\theta;z) + (N_z - n_z) Log\ (1-\mu(\theta;z)),\ \theta \in \Theta\}$$

When the variance of the observations is underestimated by the binomial model, it is usual to introduce an overdispersion parameter ϕ (see Collett 1991). Indeed,the underlying assumption is that the individual binary observations that make up the observed proportions are independent. This is not entirely verified here. Then, $\mu(z)$ is no longer deterministic, but is a random variable $A(z)$ on $[0,1]$ with mean $\mu(z)$ and variance $\phi\mu(z)(1-\mu(z)$. We then use a quasilikelihood method, associated with observations having mean $N_z\ \mu(\theta;z)$ and variance

$$Var\ n_z = N_z\ \mu(\theta;z)(1-\mu(\theta;z))(1+\phi(N_z-1)).$$

The criterium becomes

$$0 = \sum_{z=1\ldots K} \frac{N_z}{Var\ n_z}\frac{\partial \mu_\theta}{\partial \theta}(z)(n_z - N_z\mu(\theta;z))$$

Remark: We had not at our disposal the observations of the total number of grains on each ear ($N_z, z \in K$). We had only an average number that we substituted in the above expressions.

We have estimated the parameters using the M.L.E. and the Least Square Estimation criteria. We have not yet used the quasilikelihood method. This is work in progress since we have now evidence for the existence of correlations between data.

3.5 The results

Observed dispersal patterns are shaped primarily by the major wind dispersion. In both experiments, the wind blew in almost only one direction, resulting in elliptic dispersal patterns. Long-distance dispersal events were not rare and blue seeds were observed at the maximal distances sampled in the downwind direction. Despite the hundreds of kilometers between both experiments, the G.H.D. models gave the better fit for both criteria, which was expected since it contains one more parameter and includes the two models. However, the likelihood ratio test found no significant difference between N.I.G. and G.H.D. Therefore, we have chosen the simplest model (N.I.G.), which implies that the assumptions of a Brownian with drift for pollen grains sample paths coupled with hitting times are quite satisfactory. However, model validation and testing of hypothesis have to be studied more carefully (for instance the NIG model corresponds to a parameter α located at a bound of the estimation interval, ...)

Another way to assess the performances of these models together with the statistical analysis is to predict backward dispersal functions using the estimated forward dispersal functions, using computer simulations. Such a presentation of the results is given in Figure 1. This is the pollen dispersal in the first experiment (Montargis). In (a) the observed dispersal : each gray rectangle represents a sampled ear and its shade of gray represents the proportion of blue grains of the total number of grains on this ear. Axes are measured in term of distance from one corner of the field. Predictions of the proportion of marked pollen at each point of the field are given using (b) the G.H.D. model;(c) the NIG model; (d) the GTM model.

The comparison between the predicted/observed plots showed that no bias occurred in any part of the dispersal functions. The fit was quite good using NIG or GIG models

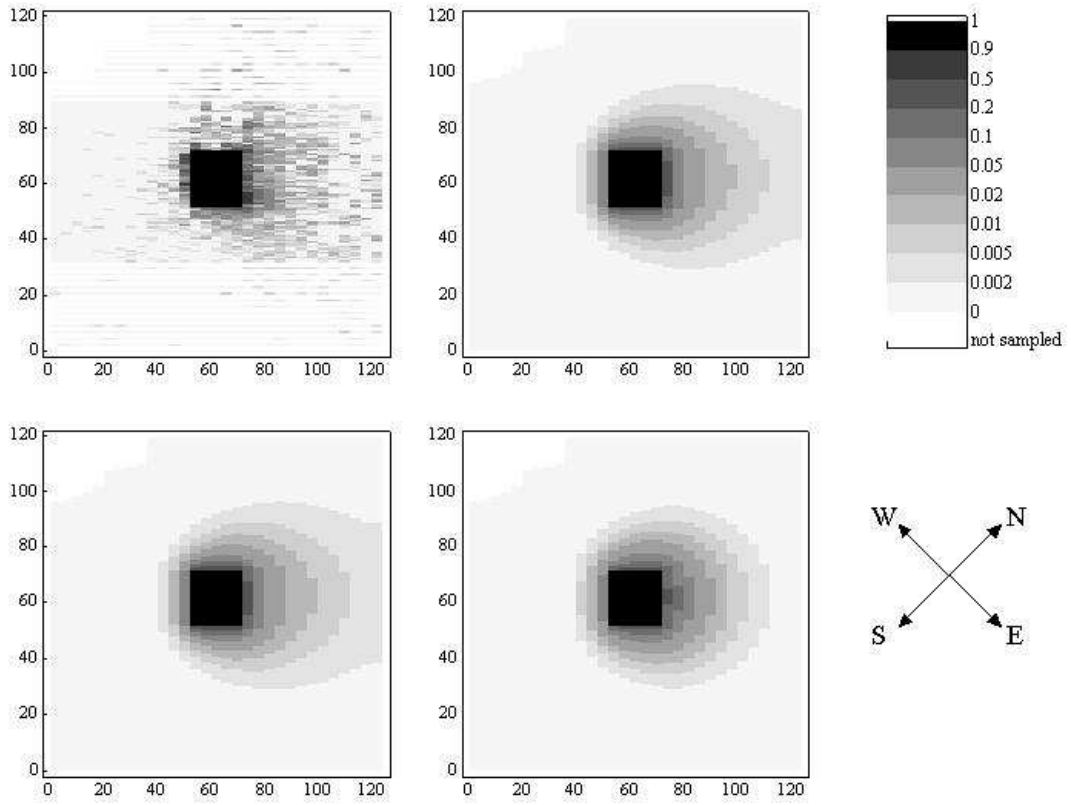


Figure 1: The pollen dispersal in the first experiment (Montargis). In (a) the observed dispersal : each gray rectangle represents a sampled ear and its shade of gray represents the proportion of blue grains of the total number of grains on this ear. Axes are measured in term of distance from one corner of the field. Predictions of the proportion of marked pollen at each point of the field are given using (b) the G.H.D. model;(c) the NIG model; (d) the GTM model.

and MLE criterium. In downwind directions, the NIG leads to quite satisfactory heavy tails, which was expected, and in the upwind direction, it decreases quite quickly. It appears from both experiments to be the most adequate model for describing the individual dispersal function of corn pollen.

To conclude, the results of this study are quite satisfactory, as illutrates the simulated and observed paths of dispersion. We have obtained a non isotropic probability distribution on R^2 which fits the data quite well. This is a real progress compared to the dispersion functions usually used (isotropic with geometric or exponential decay w.r.t. distances. Moreover, this modelisation provides a useful tool in practice to predict the levels of pollutions between for instance a transgenic and a non transgenic fields and, in particular,it shows that these levels depends on the relative sizes and positions of fields. The results are consistent with the apriori knowledges of the pollinations in the two areas.

Many questions arise now after this first study of wind pollen dispersal. We have partly neglected the existence of a wind threshold for pollen emission. The compari-

son between parameters values estimated from dispersal patterns and calculated from independent informations suggest that this should be done: there is work in progress in that direction. All the estimations on the individual dispersion function are in fact valid for a continuous canopy. Studies have to be done in non-homogeneous (or discontinuous) landscapes. Discontinuities can be roads, edges, bare soil or populations of other plant species,..). There is also work in progress in that direction.

References

Barndorff-Nielsen, O., Blaesild, P. and Halgreen, C. (1978). First hitting times models for the generalized inverse Gaussian distribution. *Stoch. Proc. Appl.*, **7**, 49–54.

Barndorff-Nielsen, O. (1997) Normal inverse distributions and stochastic volatility modeling. *Scand. Journal of Statistics*, **24**, 1–13.

Collett, D. (1991). Modelling Binary Data. Chapman & Hall/CRC.

Jørgensen, B. (1982). Statistical properties of the Generalized Inverse Gaussian Distribution. *Lecture Notes in Statistics*, **9**, Springer-Verlag, Berlin.

Klein, E.K., Lavigne, C., Foueillassar, X., Gouyon, P.H. and Larédo, C. (2003). Corn Pollen Dispersal: Quasi-mechanistic Models ansd Field Experiments. *Ecological Monographs*, **73**, 131–150.

Tufto, J., Engen, S. and Hindar, K. (1997). Stochastic dispersal processes in plant populations. *Theoretical Population Biology*, **52**, 16–26.

Stochastic effects in enzymatic biomolecular systems: framework, fast & slow species and quasi-steady state approximations

Michael Samoilov
University of California, Berkeley

1 Introduction

Enzymatic reactions represent a ubiquitous class of biochemical mechanisms. Their dynamics within broader biomolecular networks provide the chemical basis for many types of cellular behaviors, while subnetworks of enzymatic reactions often form recognizable control motif topologies making better understanding of these mechanisms an increasingly important subject. The characteristic feature of many such systems is a type of mesoscopic property, whereby typically high concentrations of reaction substrates are contrasted with frequently low concentrations of enzymes driving and controlling these processes, which could be present in quantities as low as single digit molecular copy numbers. (This is true both in the more extremal cases, such as that of DNA, where the number of copies typically varies in single digits for normal growth and division genomic — to teens and higher if we consider plasmid, organelle, viral, etc.; as well as the more conventional protein ones, such as β-galactosidase, which in its role as a lactose sensor is present in fewer than 10 copies for *E. coli* grown on other carbon sources (Stryer 1988).) While this feature of enzymatic biomolecular systems has been extensively studied within the scope of classical deterministic chemistry, e.g. to obtain the various Michaelis-Menten (MM) type approximations to such systems' mass-action kinetics description (Segel and Slemrod 1989; Murray 1993), its stochastic properties generally have not received as much consideration. This is often the case in spite of the fact that the low molecular enzyme counts make stochastic treatment of such mechanisms essential for accurate modeling of real biological processes (Arkin, Ross et al. 1998; Srivastava, Peterson et al. 2001; Ozbudak, Thattai et al. 2002; Rao, Wolf et al. 2002; Blake, Kærn et al. 2003). The main reasons behind this situation remain grounded in the combination of the relative complexity of the stochastic treatment of such processes analytically and the still high computational intensity of simulating them numerically, which is further exacerbated by the substantially low insight such methods contribute to our overall understanding of these processes. Towards such problems, this work considers alternative approaches that could potentially contribute to helping rectify the situation as well as considers examples demonstrating their uses.

2 Stochastic (Bio) Chemistry

Today the most widely accepted view of stochastic phenomena in chemical processes is based on the Master Equation formalism (Gillespie 1992), which results in a differential-

difference equation with origins in statistical mechanics and kinetic theory, whereby the ensemble distribution of a chemical system is described by

$$\frac{\partial}{\partial t}P(\vec{X},t) = \sum_{\vec{r}} \Big[W(\vec{X}-\vec{r},\vec{r})P(\vec{X}-\vec{r},t) - W(\vec{X},\vec{r})P(\vec{X},t) \Big]. \tag{1}$$

Here $P(\vec{X},t) \equiv P(\vec{X},t|\vec{X}_0,t_0)$ is, loosely speaking, the probability that a thermally equilibrated chemical system in a perfect CSTR (continuously stirred reaction tank) of volume Ω will be found — pending the initial conditions — to have $\vec{X}$ molecules at time t. Quantities $W(\vec{X},\vec{r})$ for individual reactions with molecule-change vectors $\vec{r} = \vec{\nu}^p - \vec{\nu}^s$ (product minus substrate) are referred to as "propensity functions" and represent the transition rates $\vec{X} \longrightarrow \vec{X}+\vec{r}$, typically in a polynomial form (Gillespie 1992),

$$W(\vec{X},\vec{r}) = k_{\vec{r}} \prod_i \frac{X_i!}{(X_i-\nu_i^s)!} = k_{\vec{r}} \prod_i X_i^{\nu_i^s} + O\Big(\frac{r}{X}\Big). \tag{2}$$

That is, a solution of equation (1) in the given sense completely describes the behavior of a classical spatially homogeneous (bio) chemical system, including such stochastic properties as rate of fluctuations, relative stability of multiple steady states, etc. The connection to mass-action chemical kinetics is usually provided by considering the average behavior of the species, which from (1)–(2) is:

$$\frac{d\big\langle \vec{X} \big\rangle}{dt} = \sum_{\vec{r}} \vec{r} \Big\langle W(\vec{X},\vec{r}) \Big\rangle = \sum_{\vec{r}} \vec{r} k_{\vec{r}} \Big\langle \prod_i X_i^{\nu_i^s} \Big\rangle \neq \sum_{\vec{r}} \vec{r} k_{\vec{r}} \prod_i \Big\langle X_i \Big\rangle^{\nu_i^s}. \tag{3}$$

Notice that the full stochastic behavior does not generally conform to the deterministic kinetics (given on the right side of the formula). And although for large classes of reactions the difference is negligible, it could be shown that it is by no means always true, which is where the deterministic chemical kinetics description begins to fail.

Regretfully, due to its semi-discrete structure as well as physical restriction that $X_i \geq 0$, equation (1) is very difficult to solve directly — whether exactly or approximately — even for the simplest of reaction mechanisms (Gardiner 1990; Van Kampen 1992; Dykman, Mori et al. 1994; Gillespie 2000), and the few existing solutions are rather obscure and difficult to work with (Leonard and Reichl 1990; Robertson, Shushin et al. 1993; Samoilov and Ross 1995; Laurenzi 2000).

3 Simulating Stochastic Effects in Enzymatic Systems

While obtaining general solutions to (1) is difficult if not impossible for most classes of biochemical reactions, it is nonetheless necessary in certain cases — such as enzymatic biomolecular reactions — to recognize the characteristic effects such treatment might add to our understanding of their behavior. It is thus desirable to consider approaches that might yield answers to the more specific questions one could ask of a particular system under study, without the need to explicitly solve the master equation.

3.1 *Gillespie Algorithm*

By far the most successful and useful of such general methods is the exact simulation approach, often referred to as the "Gillespie Algorithm" or "GA" — see, for example, (Gillespie 1992). GA is called "exact", because it allows user to simulate individual system paths *exactly* according to the distribution described in (1), i.e. starting with some initial condition one constructs a realization of a single trajectory system might stochastically take over time under $P(\vec{X}, t)$. That is GA yields examples of how a system defined by (1) might actually behave as well as allows for numerical computations of certain system quantities, such as variances and higher moments at steady state, but not the distribution itself. It is particularly useful if one wants to "visualize" the behavior of the system for a given set of parameters in order to get a sense of its stochastic characteristics, since it could often be implemented quite efficiently for even substantial size systems, for example, as was shown by simulation and outcome analysis of λ-phage decision switch operation in *E. coli* (Arkin, Ross et al. 1998).

3.2 *Approach Limitations*

Unfortunately, GA suffers from three fundamental shortcomings, which serve to substantially restrict its usefulness in the biochemical setting. First, it is a "random time-step" algorithm, i.e. the time period between two successive points on the simulated trajectory is a random variable, which means that neither the question of simulation time nor the state of the system at a particular time point could be answered with great fidelity. Second, GA does not provide any *a priori* evidence as to what the effects of parametric changes on the system might be or where in time they might occur. Thus, if looking for a spatially resonant or temporary rare feature — such as a bifurcation or large fluctuation — one has to be relying on "blind search" (or pure luck) to find any evidence with GA-only strategy. Third and most problematic issue is in the nature of the exact simulation approach itself. That is, GA faithfully simulates all reactions comprising the system under consideration without regard for their relative rates, since it is "exact" by definition. Thus, if the system happened to have some reactions that are a lot faster than the rest — most of the time would be spent running and updating those, while the overall behavior the system would remain obscured if at all observable. This is especially true for biomolecular systems with enzymatic reactions, since biological organisms often rely on high turnover enzymes maintained in what is termed "quasi-steady state" (which could be loosely thought of as "rapid equilibrium" and in the simplest deterministic case is equivalent to the MM approximation), i.e. where enzyme-substrate complex is being synthesized/degraded much faster than other species, thus rendering many such systems outside the realm of efficient GA applications. There have been a number of attempts — including by Gillespie himself — to augment GA for quasi-steady state conditions in order to make it more useful for dealing with biological problems. However, to date none of those methods could be considered truly successful, which — along with other factors discussed earlier — provides an impetus to consider alternative formulations. We will do so next in the framework of the simplest biomolecular enzymatic system: the Michaelis-Menten enzyme reaction.

4 Alternative Modeling Approaches: Michaelis-Menten Enzyme

While MM reaction — Figure 1 with additional constraint equations (4) — is, perhaps, the simplest example of enzymatic biomolecular system,

$$\boldsymbol{S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \xrightarrow{k_2} P + E}$$

Figure 1: Michaelis-Menten enzyme reaction.

$$\begin{aligned} E_0 &= E + C = Const, \\ S_0 &= S + C + P = Const, \\ \text{with} \quad K_m &= (k_{-1} + k_2)/k_1, \end{aligned} \tag{4}$$

it nonetheless exhibits all of its characteristic qualities, such as net zero change, nonlinearity, etc. A notable property of this reaction is the existence in the deterministic case of a singular perturbation solution, known as a "quasi-steady state approximation"or "QSSA", (Segel and Slemrod 1989; Murray 1993), which is a general analytical approach that in the specific case of MM essentially validates the limit:

$$\frac{dC}{dt} = k_1(E_0 - C)S - (k_{-1} + k_2)C = 0 + O(\varepsilon) \quad \Longleftrightarrow \quad C(t) = \frac{E_0 S}{S + K_m} + O(\varepsilon) \tag{5}$$

with

$$\varepsilon = \frac{E_0}{S_0 + K_m} \ll 1, \tag{6}$$

by allowing one to separate away the "fast" variables that equilibrate quickly, such as E, from the "slow" ones that are of ultimate interest in this biomolecular system, such as P, thus greatly simplifying its analysis.

As was noted before, it has been a challenge to properly define an equivalent "quasi-steady state" criterion and/or variable separation methods in the case of a general stochastic system. We thus look next to explicitly analyze such properties only for the MM system as described directly via the master equation (1), which in this case takes form [1]

$$\begin{aligned} \frac{dPr(S,E,C,P;t)}{dt} &= k_1(S+1)(E+1)Pr(S+1,E+1,C-1,P;t) \\ &+ k_{-1}(C+1)Pr(S-1,E-1,C+1,P;t) \\ &+ k_2(C+1)Pr(S,E-1,C+1,P-1;t) \\ &- [k_1 SE + (k_{-1}+k_2)C]\, Pr(S,E,C,P;t), \end{aligned} \tag{7}$$

[1]Notation for probability was changed from P to Pr in order to avoid confusion with reaction product label.

and — using the mass conservation conditions (4) — could be reduced to

$$\begin{aligned}\frac{dPr(C,P;t)}{dt} &= k_1(S_0-P-C+1)(E_0-C+1)Pr(C-1,P;t) \\ &+ k_{-1}(C+1)Pr(C+1,P;t)+k_2(C+1)Pr(C+1,P-1;t) \\ &- [k_1(S_0-P-C)(E_0-C)+(k_{-1}+k_2)C]Pr(C,P;t),\end{aligned} \tag{8}$$

in order to see if we could demonstrate some useful approaches to dealing with stochastic biomolecular systems on this relatively simple example.

4.1 *Direct Evaluation*

We begin by looking to compare the master equation results with equation (5), to see if the stochastic predictions do indeed match the deterministic ones. [2] Thus, summing over P in equation (8), we can deduce the evolution formula for the probability of C only:

$$\begin{aligned}\frac{dPr(C;t)}{dt} &= k_1(S_0-\langle P|C-1;t\rangle-C+1)(E_0-C+1)Pr(C-1;t) \\ &+ (k_{-1}+k_2)(C+1)Pr(C+1;t) \\ &- [k_1(S_0-\langle P|C;t\rangle-C)(E_0-C)+(k_{-1}+k_2)C]Pr(C;t),\end{aligned} \tag{9}$$

while, similarly, summing over C we get a formula for the probability of P:

$$\frac{dPr(P;t)}{dt} = k_2\langle C|P-1;t\rangle Pr(P-1;t) - k_2\langle C|P;t\rangle Pr(P;t), \tag{10}$$

where $\langle X|Y;t\rangle \equiv \sum_X P(X|Y;t)$ denotes the conditional average of X at time t.

From (3) we can see that the one case where the stochastic and deterministic results are guaranteed to match is the case of *linear* reactions. The only specie that has a purely linear form for the evolution of its average in MM reaction is the product, $\langle P\rangle$, which could be seen to follow the same equation in both instances:

$$\frac{d\langle P\rangle}{dt} = k_2\langle C\rangle\,. \tag{11}$$

Now consider evolution the *conditional* average of the product, $\langle P|C;t\rangle$.

i. Initially at $t=0$ the system has deterministic inputs: $P(0)$, $C(0)$, etc. — that is, $\langle P\rangle(0) = \langle P|C\rangle(0)$.

[2] As could be observed from (3), the equivalence of the stochastic model and its deterministic limit is not at all guaranteed and *in general is not true*.

Also, from (2)–(4) it is easy to see that $\langle P|C;t\rangle \in [\langle P|\Im\{0\};t\rangle, \langle P|\Im\{E_0\};t\rangle]$, – where the limits are over sets of extremal trajectories, along which there is either no C or there is maximal amount *all the time*. After direct evaluation they yield:

ii. For the *lower* bound – there is no production of P whatsoever, so $\langle P|\Im\{0\};t\rangle = P(0)$.

iii. For the *upper* bound – there is always (fixed) maximal amount of complex, so the reaction always proceeds at the maximal rate, where (1)–(2) & (10) give: $\frac{d}{dt}\langle P|\Im\{E_0\};t\rangle = k_2 E_0$.

Since concentrations are always positive, we observe that $\langle C\rangle \in [0, E_0]$ via (3), which combined with (11) & i)-iii) shows that,

$$\langle P\rangle - \langle P|C\rangle = O(\varepsilon), \tag{12}$$

where ε is defined in equation (6).

With the substitution of (12) into equation (9),

$$\begin{aligned}\frac{dPr(C;t)}{dt} &= k_1\langle S\rangle (E_0 - C + 1)Pr(C-1;t) + (k_{-1}+k_2)(C+1)Pr(C+1;t) \\ &- [k_1\langle S\rangle (E_0 - C) + (k_{-1}+k_2)C]Pr(C;t) + O(\varepsilon),\end{aligned} \tag{13}$$

we can compute the kinetic equation for the evolution of the "fast" enzyme-complex averages in this limit:

$$\frac{d\langle C\rangle}{dt} = k_1\langle S\rangle (E_0 - \langle C\rangle) - (k_{-1}+k_2)\langle C\rangle + O(\varepsilon), \tag{14}$$

which is the same as predicted deterministically via (3)–(5).

Now, since the MM approximation itself is also order ε we can conclude that the stochastic results do indeed correspond to the deterministic approximation ones given in equation (5) in this case, i.e.

$$\langle C\rangle = \frac{E_0\langle S\rangle}{\langle S\rangle + K_m} + O(\varepsilon). \tag{15}$$

Finally, utilizing the same techniques as in (12) with the help of equation (15) we get

$$\langle C\rangle - \langle C|P\rangle = O(\varepsilon), \tag{16}$$

i.e. equation (10) similarly becomes in "slow" variable limit:

$$\frac{dPr(C;t)}{dt} = k_2\langle C\rangle Pr(P-1;t) - k_2\langle C\rangle Pr(P;t) + O(\varepsilon). \tag{17}$$

Thus, from (11) & (13)–(14) we can conclude that, at least for the Michaelis-Menten reaction, the stochastic predictions for the evolution of the species averages indeed match the classical deterministic ones to order ε.

4.2 *Implication for Stochastic QSSA and Fast–Slow Variable Separability*

It is remarkable to note that in the course of trying to validate the deterministically predicted behavior of the system we ended up with the same criterion as the quasi-steady state one for MM approximation. That is, if $\varepsilon \ll 1$ — not only *both* the deterministic limit, (3), and the further Michaelis-Menten approximation, (4), remain valid, but the *stochastic* quasi-steady state approximation system exists and its meaning is clear. Furthermore, the reaction explicitly separates into "fast" and "slow" variables — naturally so as given by (13) and (17).

Fast variables, $\{E,\, C\}$, to order ε evolve according to the reduced master equation (13), which is equivalent to a *linearized* reaction given in Figure 2,

$$E \underset{k_{-1}+k_2}{\overset{k_1\langle S\rangle}{\rightleftarrows}} C$$

Figure 2: Stochastic evolution of "fast" species in Michaelis-Menten reaction for $\varepsilon \ll 1$.

thus allowing us to, for example, easily calculate the explicit stationary distribution of the fast species reaction within the same order, as done by many authors, e.g. (Samoilov and Ross 1995).

Slow variables, $\{S,\, P\}$, to order ε also evolve according to the reduced master equation (17), which is equivalent to as a linearized reaction given in Figure 3,

$$\langle C\rangle \xrightarrow{k_2} P$$

Figure 3: Stochastic evolution of "slow" species in Michaelis-Menten reaction for $\varepsilon \ll 1$.

where S could then be obtained via the conservation conditions (4) and equations (14)–(16).

Thus, we can conclude that (13) and (17) provide the desired separation of variables into the fast and slow ones, which — among other things — demonstrates an example of approach that could be further used to generally allow for more efficient simulations of enzymatic biomolecular reactions, as was discussed earlier in the context of the GA.

5 Discussion

This note has attempted to broadly outline some of the issues associated with stochastic effects in enzymatic biomolecular systems and exemplify them by considering a couple of modeling approaches, such as GA simulations and QSSA-type approximations. While such methods have been investigated previously in a much more extensive framework for improving enzymatic system simulations' speed, (Haseltine and Rawlings 2002; Rao and Arkin 2003), a notable thing about the results presented herein — albeit applied only to the Michaelis-Menten reaction — is that they not only provide a consistent stochastic QSSA and fast/slow variable separation picture for the considered system, but also do so under a rather mild criterion of $\varepsilon \ll 1$ only. In addition to being the same condition as the deterministic one, equation (6), — which provides the connection between the two approaches and guarantees the consistency of results — it avoids a lot more stringent and somewhat arbitrary conditions: $\frac{dP(Slow|Fast;t)}{dt} = 0$ or $\frac{dP(Fast|Slow;t)}{dt} = 0$ authors have previously had to, respectively, impose as fundamental assumptions justifying their intuitive reasoning. In addition to being somewhat contradictory, these conditions do need broad validation and might generally not be true. The reasoning provided here helps to alleviate some of these concerns, at least in the case of the MM reaction, which — although a single mechanism — does have certain universal properties common to many biomolecular enzymatic mechanisms, as discussed earlier.

Overall, the presented analysis might be viewed as the stochastic analogue to the situation encountered in the deterministic case, where the *ad hoc* method for obtaining the deterministic version of the Michaelis-Menten approximation is to set $\dot{C} = 0$ (Segel 1975) — which has correct "intuitive meaning", but gives no applicability or error criteria. The full justification (and meaning) for the result is provided only via a singular perturbation differential equation analysis (Segel and Slemrod 1989), which analytically establishes the criteria for both approximation validity and error estimation. The analysis outlined herein provides a similar justification for the case of the stochastic description of the Michaelis-Menten reaction as well as suggests a unified treatment for giving meaning to and defining applicability of the fast/slow variable separation and stochastic quasi-steady state approximation in enzymatic reactions.

6 Acknowledgments

I would like to thank Prof. Adam Arkin, Sergey Plyasunov and Dr. Chris Rao for prompting many stimulating discussions and work on the subject discussed herein.

7 References

Arkin, A., J. Ross, et al. (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *E. coli* cells." Genetics 149(4): 1633-1648.

Blake, W. J., M. Kærn, et al. (2003). "Noise in eukaryotic gene expression." Nature 422(6932): 633-637.

Dykman, M. I., E. Mori, et al. (1994). "Large fluctuations and optimal paths in chemical reaction." J. Chem. Phys 100: 5735-5750.

Gardiner, C. W. (1990). Handbook of stochastic methods for physics, chemistry, and the natural sciences. Berlin ; New York, Springer-Verlag.

Gillespie, D. T. (1992). Markov processes : an introduction for physical scientists. Boston, Academic Press.

Gillespie, D. T. (1992). "A rigorous derivation of the chemical master equation." Physica A 188: 404-425.

Gillespie, D. T. (2000). "The chemical Langevin equation." J. Chem. Phys 113(1): 297-306.

Haseltine, E. L. and J. B. Rawlings (2002). "Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics." J. Chem. Phys. 117(15): 6959-6969.

Laurenzi, I. J. (2000). "An analytical solution of the stochastic master equation for reversible bimolecular reaction kinetics." J. Chem. Phys 113(8): 3315-3322.

Leonard, D. and L. E. Reichl (1990). "Stochastic analysis of a driven chemical reaction." J. Chem. Phys 92(10): 6004-6010.

Murray, J. D. (1993). Mathematical Biology. New York, Springer-Verlag.

Ozbudak, E. M., M. Thattai, et al. (2002). "Regulation of noise in the expression of a single gene." Nat Genet 31(1): 69-73.

Rao, C. V. and A. Arkin (2003). "Stochastic chemical kinetics and the quasi-steady state assumption: Application to the Gillespie algorithm." J. Chem. Phys 118(11): 4999-5010.

Rao, C. V., D. M. Wolf, et al. (2002). "Control, exploitation and tolerance of intracellular noise." Nature 420(6912): 231-7.

Robertson, S. H., A. I. Shushin, et al. (1993). "Reduction of the two-dimensional master equation to a Smoluchowsky type differential equation with application to $CH_4 \rightarrow CH_3 + H$." J. Chem. Phys 98(11): 8673-8679.

Samoilov, M. and J. Ross (1995). "One-dimensional chemical master equation: Uniqueness and analytical form of certain solutions." J. Chem. Phys. 102: 7983-7987.

Segel, I. H. (1975). Enzyme kinetics : behavior and analysis of rapid equilibrium and steady state enzyme systems. New York, Wiley.

Segel, L. A. and M. Slemrod (1989). "The quasi-steady-state assumption: a case study in perturbation." SIAM Rev. 31(3): 446-477.

Srivastava, R., M. S. Peterson, et al. (2001). "Stochastic kinetic analysis of the *Escherichia coli* stress circuit using σ^{32}-targeted antisense." Biotechnol Bioeng 75(1): 120-9.

Stryer, L. (1988). Biochemistry New York W. H. Freeman and Company

Van Kampen, N. G. (1992). Stochastic Processes in Physics and Chemistry. Amsterdam, North-Holland.

Genetic network modeling

E. P. van Someren, E. Backer and M. J. T. Reinders
Delft University of Technology

1 Introduction

In pharmacogenomics and related areas, a lot of research is directed towards discovering, understanding and/or controlling the outcome of some particular biological pathway. Numerous examples exist where the manipulation of a key enzyme in such a pathway did not lead to the desired effect [5]. This usually happens because the intended effect was compensated for by the genetic regulation of enzyme levels. Such examples illustrate the importance of accounting for genetic regulation.

We know that the structure of complex genetic and biochemical networks lies hidden in the sequence information of our DNA but it is far from trivial to predict gene expression from the sequence code alone. The current availability of microarray measurements of thousands of gene expression levels during the course of an experiment or after the knockout of a gene provides a wealth of complementary information that may be exploited to unravel the complex interplay between genes. It now becomes possible to start answering some of the truly challenging questions in systems biology. For example, is it possible to model these genetic interactions as a large network of interacting elements and can these interactions be effectively learned from measured expression data?

Since Kauffman [21] introduced the concept of mathematical modeling of complex systems, the reverse engineering of genetic networks has triggered the imagination of many molecular biologists. Somogyi [31][1] also investigated some of the properties of Boolean networks in relation to biological systems. These researchers showed that Boolean networks possess properties like global complex behavior, self-organization, stability, redundancy and periodicity. Analogies between basins of attraction and different tissue types, as well as cyclic attractors and cell cycles have also been discussed by many other researchers.

The inference of genetic interactions from measured expression data is one of the most challenging tasks of modern functional genomics. When successful, the learned network of regulatory interactions yields a wealth of useful information. An inferred genetic network contains information about the pathway to which a gene belongs and which genes it interacts with. Furthermore, it explains the genes function in terms of how it influences other genes and indicates which genes are pathway initiators and therefore potential drug targets.

Obviously, such wealth comes at a price and that of genetic network modeling is that it is an extremely complex task. Although the behavior and properties of artificial networks match the observations made in real biological systems well, the field of genetic network modeling has yet to reach its full maturity. The automatic discovery

[1]For reasons of brevity, the authors consistently refer only to the first author of each reference.

of genetic networks from expression data alone is far from trivial because of the combinatorial nature of the problem and the poor information content of the data. First, to model genetic regulation, one needs to take into account the fact that gene expression levels are regulated by the combined action of multiple gene products [17]. Second, the number of measurements (arrays) is relatively small compared to the number of measured objects (genes) and the data are corrupted with a substantial amount of measurement noise. Together, these two complicating factors make the construction of genetic networks from empirical observations extremely difficult. In addition, results are further complicated by the presence of inherent noise caused by, for example, variations between different individuals, small numbers of molecules available in a given cell, variations between tissues in a given individual, variations caused by effects that are not measured etc.

The dimensionality problem (many objects and few measurements) plays a fundamental role in genetic network modeling causing the straightforward estimation of model parameters to become extremely unreliable (many equally good solutions). The common approach to avoid this problem is to either reduce the models complexity or to apply constraints on the parameters. Consequently, the relatively young field of genetic network modeling has been governed by the introduction of a plethora of different models and learning strategies.

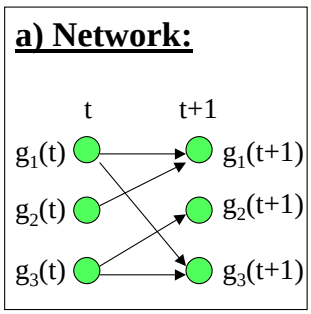
This abstract provides an small overview of genetic network modeling approaches that employ expression data to automatically discover genetic interactions. Reviews on genetic network models have also appeared recently. In a recent review [42], models are placed in an historical context and the qualitative properties of the models and their learning strategies are compared. De Jong [10] focuses in his review more on the mathematical properties of the models. An experimental comparison of a limited number of genetic network models is presented in [48, 39].

2 Reverse Engineering of Genetic Network Models

The introduction of microarray technology made it possible to measure the gene-expression levels of thousands of genes simultaneously. This introduced a new impulse to genetic network modeling, namely the reverse engineering of large-scale genetic networks based on measured expression data. Starting from microarray data and a general model of genetic interactions, the parameters of this general network model are learned from the data. Here we will describe only dynamical models, i.e., models that are learned on time course gene expression data.

2.1 Boolean networks

In 1998, Liang [24] started off by introducing REVEAL, an algorithm that automatically constructs a large-scale Boolean network from data. In a general Boolean network model, all gene expression levels are discretized into binary expression levels; a gene is either on or off. The binary expression levels of all genes in the system at a certain point in time define the state of the network at that time instant. A state transition table defines, for each possible network state, which network state will be next (see



b) State-transition table:

$g_1(t)$	$g_2(t)$	$g_3(t)$	$g_1(t+1)$	$g_2(t+1)$	$g_3(t+1)$
0	0	0	0	1	0
0	0	1	0	0	1
0	1	0	0	1	0
0	1	1	0	0	1
1	0	0	0	1	1
1	0	1	0	0	1
1	1	0	1	1	1
1	1	1	1	0	1

c) Rules:

$g_1(t+1) = g_1(t)$ AND $g_2(t)$
$g_2(t+1) =$ NOT $g_3(t)$
$g_3(t+1) = g_1(t)$ OR $g_3(t)$

Figure 1: Example of: a) a Boolean network of three genes with corresponding b) state-transition table and c) Boolean rules.

Figure 1b). From this table, a Boolean rule can be determined for each gene that describes how its expression level at the next time instant depends on some combination of the gene expression levels at the current time instant.

Typical Boolean rules contain logical operators such as AND, OR and NOT (see Figure 1c). By placing connections between each of the input genes in the rule and the output gene, the structure of the network can be determined, which expresses the interactions among all genes (see Figure 1a). A typical gene expression dataset, after discretization, represents an incomplete state-transition table, since not all possible states will have been measured.

REVEAL constructs the rule for a target gene from this incomplete table by considering the mutual information between the input states of each single gene (k = 1) and the output state of the target gene. If the output can be perfectly determined by one of the inputs, the corresponding rules and connections are extracted. If not, all combinations of two genes (k = 2) are considered as input and it is examined whether this pair can perfectly predict the target. If not, the procedure repeats for k = k + 1 etc. In other words, the structure is learned using a forward exhaustive search procedure that stops as soon as a perfect reconstruction is possible.

A year later, Akutsu [1] proved, using a conceptually simpler approach, that $O(log_2 N)$ random measurements are sufficient to identify a network of N genes with bounded connectivity K but this algorithm takes $O(NK + 1Q)$ time, with Q the number of state transitions. This implies that for a typical gene expression dataset with 1000 genes and connectivity $K = 2$, in the order of 10 independent measurements are sufficient but in that case $O(10^1 0)$ time is required! The algorithm learns a Boolean model by performing an exhaustive search not only for each possible combination of inputs but also for each possible configuration of Boolean functions (using only AND or and NOT operators) that are consistent with the given state transitions. Unfortunately, this algorithm was not suited for noisy conditions but a year later Akutsu presented an algorithm that is robust to noise [2, 3].

2.2 Continuous models

Although Boolean networks provide a good starting point, they are generally criticized because only two discrete expression levels are allowed. Many examples exist where genes are regulated in a continuous manner rather than just turned on or off [29, 20, 19]. This inspired the introduction of models with a continuous representation of gene expression.

DHaeseleer [12] learned a linear model on data from the rat central nervous system (CNS), during development and injury after kainate injection [47]. He coupled two partly overlapping datasets, to utilize as much information as possible, resulting in a dataset of 65 genes and 28 time points. Even this simple linear model (with a single parameter per gene) contains more parameters than the number of measurements. This so-called dimensionality problem makes it possible to find many parameter sets that perfectly reconstruct the data. As a result, the parameter estimations become unreliable. To accommodate the fact that the datasets were differently sampled, DHaeseleer employed a nonlinear interpolation method (resulting in 68 time points). By employing a nonlinear interpolation scheme, he enforces smoothness and tries to avoid the dimensionality problem.

Weaver [46] also employed the linear model but augmented it with a biologically inspired, non-linear doseresponse curve. Although nonlinear, this model is essentially a recurrent neural network without a hidden layer. By de-squashing the doseresponse curve, the model can be solved by simple linear algebra. To handle the dimensionality problem, Weaver proposed the use of the Moore-Penrose pseudo-inverse. This special matrix inverse produces a solution for under-determined problems that minimizes the sum of the squared weights but still perfectly fits the data. To introduce limited connectivity, he proposed a greedy backward search that iteratively sets the smallest weight to zero and then recomputes the pseudo-inverse on the now slightly less under-determined problem. Unfortunately, the de-squashing step is quite sensitive to small changes in the data. Rather than a discrete-time model, Wahde [43] employed a continuous-time recurrent neural network. A genetic algorithm (GA) was employed to find the parameters of small networks (four genes) learned on the average profiles of clustered data. A genetic algorithm [26] is an optimization technique based on natural selection in which a set of possible solutions, called a population, is evaluated in parallel. New populations of potentially better solutions are generated and evaluated by combining (crossover) and modifying (mutation) the best solutions in the current population. After learning the parameters with a GA, a qualitative description of the parameters is given. Wahde showed results on artificial data as well as on the CNS dataset presented by Wen [47]. Using artificial data he showed that it is better to have multiple shorter time series than one long series. In later work [44, 45], he suggested a procedure that forced parameters that were not significant to zero. Repeated elimination of the most unreliable parameters can also be viewed as a form of backward search.

Chen [9] proposed an even more realistic model based on a system of differential equations that models both mRNA and protein levels, including degradation. Chen showed that, provided that both mRNA and protein levels are given, solving this model is similar to the problem of finding minimum weight solutions to linear equations

(MWSLE). Unfortunately, this problem is known to be NP complete. However, for a constant connectivity, K, the problem can be solved in O(QNK+1) time (using a dataset of N genes and Q time points) by just checking all NK possible structures. Chen also reasoned that, as many genes showed periodic expression, the Fourier transform for stable systems (FTSS) might be employed as an alternative approach.

A year earlier, Spirov [32] had also suggested the use of a system of differential equations but for a smaller network and with more data points. For learning the parameters, he suggested first using a genetic algorithm to come up with an initial population of globally optimal solutions, which is then used as seeds for a parallel simulated annealing (SA) search. Simulated annealing is a sequential optimization technique that is based on evaluating random changes to the current solution. Better solutions are always accepted, whereas worse solutions are accepted with a probability that decreases during optimization. As a result, SA moves consistently to better solutions but is able to jump out of local optima. When these runs have almost converged, a local gradient descent (GD) approach is employed.

2.3 Modeling concerns

Apart from many papers that introduced a new reverse engineering approach based on yet another model, gradually more papers emerged that addressed the issues associated with genetic network modeling itself. With the reductionists approach, the combinatorial nature of genetic regulation had largely been ignored [35]. Therefore, it took some time before researchers realized the immense complexity that learning genetic networks from expression data involved and the early enthusiasm subdued. Szallasi [35] claimed that there are four factors inherent in biological systems that influence the reverse engineering of genetic networks from expression data. First, the nature of genetic networks is undoubtedly stochastic but microarray measurements are population averaged, which may mask the real individual regulatory interactions. Also, a faster sampling rate is not always possible because the measurement error determines a lower bound on the sampling interval, i.e., the expected difference in expression within one sampling interval should be larger than the measurement noise. Secondly, there are also many regulatory factors that are not modeled, such as (de-)stabilization of mRNA, translocation, phosphorylization etc. Thirdly, he reasons that the information content of the data is not as large as its size would suggest (12 orders of magnitude smaller), as only a few genes cycle and even fewer show frequent changes during cell cycle. On the other hand, a property that is favorable for network analysis is that networks are believed to exhibit a high level of compartmentalization.

Spirtes [33] also discussed some of the complicating issues of data acquisition in relation to construction of genetic networks. Apart from the above- mentioned issues of small sample sizes (dimensionality problem), the substantial measurement error and the masking effect of population averaged measurements, he also points to the fact that the final results can be influenced by hidden (e.g., not modeled) effects and the loss of synchronization of cells.

Erb [14] experimentally examined the influence of measurement noise. He performed Khalils sensitivity analysis on a complex non-linear model proposed by Mjolsness [13], employing a fully connected network of only three genes. Already with such

a small network, the parameters turned out to be very sensitive to noise in the data.

A comparative study done by Wessels [48, 39] proposed a set of mathematical properties that genetic network models should possess and by means of which they can be compared. In a small experimental study of continuous models, in which the models were learned on data generated by the other models, he reported disappointing results in terms of how well models can reveal the underlying interactions when faced with noise and limited data. The results favor simple, i.e., linear or pair-wise, models that are less sensitive to unfavorable data conditions [2].

2.4 Pairwise models

One way to overcome the dimensionality problem is to restrict the complexity of the model, for example, by only considering pair-wise relationships. Arkin [4] was the first to suggest the construction of biochemical pathways by means of timeshifted pair-wise correlations. First, the position and magnitude at which the maximal timeshifted cross-correlation occurs is computed in a pair-wise fashion. From this, a distance measure is constructed and single linkage hierarchical clustering is employed, resulting in a singly linked tree that connects associated genes. Augmented with directional and time-lag information this association diagram reveals temporal interactions. Arkin suggested that his approach could also be used to learn genetic networks.

Later, Chen [8] proposed a similar scheme, based on matching peaks in the signals rather than using correlation. After thresholding and clustering, the remaining profiles are represented as a set of peaks. Then peaks in the profiles are compared in a pair-wise fashion to determine the causal activation scores. Similarly, inhibition scores are determined. From these scores a putative regulation network is constructed using simulated annealing.

Woolf [50] was the first to describe a fuzzy model for learning genetic interactions. He searched for all possible triplets of an activator and a repressor (two inputs) that influence a target gene (one output). All triplets are scored and ordered on how well they fit the expression data and on whether the inputs showed enough variation. Unfortunately, these pair-wise (triple-wise) models are fundamentally limited to considering only singly (doubly) connected networks.

2.5 Qualitative models

A different way to cope with the limitations of the data is to learn qualitative models, thus avoiding the necessity to estimate model parameters precisely. Akutsu [2, 3] described a collection of algorithms that are an intermediate solution, somewhere between Boolean models and continuous differential models. These qualitative models are based on linear differential equations but instead of trying to learn the exact parameters, the researcher derives qualitative abstractions of the parameters. For instance, it is only relevant whether the differences are positive, negative or zero. In this case, a solution can be found by solving a set of inequality relations. Provided that a lot of data are available, these inequalities can be solved using linear programming (LP).

[2] The GA of the Wahde model converged slowly and was therefore stopped early, not allowing the model to converge completely.

Alternatively, the parameters of a non-linear S-system (power-law) can be found using linear algebra by taking the logarithm on both sides of the equations. An S-system is a set of non-linear differential equations of a special form belonging to the power-law formalism (products of exponentially weighted inputs). If the logarithm is taken, the obtained parameter values only portray a relative meaning. But this was exactly the goal: to obtain a qualitative description.

Because of the multitude of detailed biological information acquired over the years, a qualitative model provides an excellent tool to describe the working hypothesis of researchers. Shrager [30] proposed an automatic scheme to revise an initial qualitative model such that it better matches the expression data. This scheme is based on comparing the expected pair-wise correlations of all pairs in the initial scheme with the correlations in the expression data. This measure of data fit is used to construct a fitness function, which is augmented with terms to reduce the number of variables and links in the model. With this fitness function a simple greedy search is performed based on considering single changes in the model. Unfortunately, the employed pairwise correlation measure does not fully capture the combinatorial nature of the qualitative model.

2.6 Modeling revisited

A better understanding of the consequences of the dimensionality problem resulted in modeling approaches that were better adapted to handle the limitations of the data. For example, strategies started to focus on first reducing the problem (e.g., taking a smaller network, using clustering or structure determination) such that the resulting parameters are estimated more reliably. As a result, the boundaries between the analytic and synthetic approaches gradually became blurred.

Van Someren suggested a number of general approaches to reduce the dimensionality problem by incorporating biologically motivated constraints and showed results from artificial data generated with linear networks. The reduction of the number of genes by clustering gene expression profiles was considered by many [12, 43, 44, 45, 32, 8, 38, 27, 11]. However, Van Someren [38] studied the relationship between clustering and its effect on the dimensionality problem when learning linear genetic network models. In [40, 42], he showed that genetic network models could be made robust to noise by minimizing the first-order derivative of the models output with respect to its input. For non-linear models, robustness is imposed by learning the model on a set of noisy profiles. To impose limited connectivity of the models, Van Someren [41] compared a number of search algorithms that search for structures with limited connectivity. In this comparison, a forward beam search approach proved to be the best. Mjolsness [27] also suggested the use of clustered data and learned a system of non-linear differential equations using simulated annealing. Apart from minimizing the prediction error, he included a weight-decay term to minimize the weight values and an exponential term that keeps the parameters bounded in the cost function. Koza [22] employed genetic programming to determine the structure and rate constants of small metabolic pathways. He showed that it was possible to automatically create a metabolic pathway involved in the phospholipid cycle using 270 time points of E-CELL simulations of a 4-enzyme network where all enzymes were perturbed. Unfortunately, a large amount of

data were required. Maki [25] proposed a two-step approach in which first the structure of a pair-wise Boolean network is learned from the steady-state expressions obtained after perturbation of each gene in the network. The resulting network structure is used to define smaller networks modeled by S-systems. The parameters of these systems are then learned using a GA applied on dynamic data. Unfortunately, this approach still needs a lot of measurements, i.e., at least perturbation experiments of all genes.

2.7 Trend towards Integrated Approaches

Ideker [18] presented a fully integrated approach on large-scale data in which four main steps were taken:

- define an initial model of a pathway
- perturb components in the pathway and measure the responses in mRNA and protein levels
- check the responses with the model
- refine the model to explain the unpredicted responses

He was the first to present mRNA expression data (microarrays) as well as protein abundance data, using isotope-coded affinity tag (ICAT) reagents and tandem mass spectrometry (MS/ MS) and to integrate this with information from databases of known physical interactions of the galactose pathway.

Clearly the integration of different information sources is playing an essential part in modern approaches towards genetic network modeling. The modeling trend that is revealed by this quick review is the use of a larger variety of information for learning genetic network models, be it in terms of other types of measurements, information stored in databases or desired properties of networks. Therefore, we might expect that, in the near future, results from pathway scoring [16, 23, 28, 34, 49] and promoter analysis [7, 6, 36, 37, 15] approaches will become integrated within the learning algorithms of genetic network models. Advancements like these, will unlock and exploit the full potential that genetic network modeling has to offer.

References

[1] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing '99*, volume 4, pages 17–28. World Scientific Publishing Co., 1999.

[2] T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for inferring qualitative models of biological networks. In *Pacific Symposium on Biocomputing 2000*, volume 5, pages 290–301. World Scientific Publishing Co., 2000.

[3] T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16:727–734, 2000.

[4] A. Arkin, P. Shen, and J. Ross. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277:1275–1279, 1997.

[5] J.E. Bailey. Lessons from metabolic engineering for functional genomics and drug discovery. *Nature Biotechnology*, 17(7):616–618, July 1999.

[6] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.

[7] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

[8] T. Chen, V. Filkov, and S. Skiena. Identifying gene regulatory networks from experimental data. In *Proceedings of the third annual international conference on Computational molecular biology (RECOMB99)*, pages 94–103. Association for Computing Machinery, 1999.

[9] T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. In R.B. Altman, K. Lauderdale, A.K. Dunker, L. Hunter, and T.E. Klein, editors, *Pacific Symposium on Biocomputing '99*, volume 4, pages 29–40, Singapore, 1999. World Scientific Publishing Co.

[10] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[11] P. D'Haeseleer, S. Liang, and R. Somogyi. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[12] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mrna expression levels during cns development and injury. In *Pacific Symposium on Biocomputing '99*, volume 4, pages 41–52. World Scientific Publishing Co., 1999.

[13] D. Sharp E. Mjolsness and J. Reinitz. A connectionist model of development. *Journal of Theoretical Biology*, 152(4):429–454, 1991.

[14] R.S. Erb and G.S. Michaels. Sensitivity of biological models to errors in parameter estimates. In *Pacific Symposium on Biocomputing '99*, volume 4, pages 53–64. World Scientific Publishing Co., 1999.

[15] D. Guhathakurta, L.A. Schriefer, M.C. Hresko, R.H. Waterston, and G.D. Stormo. Identifying muscle regulatory elements and genes in the nematode caenorhabditis elegans. In *Pacific Symposium on Biocomputing 2002*, volume 7, pages 425–436. World Scientific Publishing Co., January 2002.

[16] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. In *Pacific Symposium on Biocomputing 2001*, volume 6, pages 422–433, Hawai, January 2001. World Scientific Publishing Co.

[17] F.C. Holstege, E.G. Jennings, J.J. Wyrick, T.I. Lee, C.J. Hengartner, M.R. Green, T.R. Golub, E.S. Lander, and R.A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–728, 1998.

[18] T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and L. Hood. Itegrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292:929–934, May 2001.

[19] K. Kalthoff. *Analysis of Biological Development*. 1996.

[20] W.S. Katz, R.J. Hill, T.R. Clandini, and P.W. Sternberg. Different levels of the c. elegans growth factor lin-3 promote distinct vulval precursor fates. *Cell*, 82:297–307, 1995.

[21] S.A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993.

[22] J.R. Koza, W. Mydlowec, G. Lanza, J. Yu, and M.A. Keane. Reverse engineering of metabolic pathways from observed data using genetic programming. In *Pacific Symposium on Biocomputing 2001*, volume 6, pages 434–445, Hawai, January 2001. World Scientific Publishing Co.

[23] M.P. Kurhekar, S. Adak, S. Jhunhunwala, and K. Raghupathy. Genome-wide pathway analysis and visualization using gene expression data. In *Pacific Symposium on Biocomputing 2002*, volume 7, pages 462–473. World Scientific Publishing Co., January 2002.

[24] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing '98*, volume 3, pages 18–29. World Scientific Publishing Co., 1998.

[25] Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe, and Y. Eguchi. Development of a system for the inference of large scale genetic networks. In *Pacific Symposium on Biocomputing 2001*, volume 6, pages 446–458, Hawai, January 2001. World Scientific Publishing Co.

[26] M. Mitchell. *An Introduction to Genetic Algortihms*. Cambridge, 1996.

[27] E. Mjolsness, T. Mann, R. Castao, and B. Wold. From coexpression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data. In *Advances in Neural Information Processing Systems*, volume 12, pages 928–934. MIT Press., 2000.

[28] P. Pavlidis, D.P. Lewis, and W.S. Noble. Exploring gene expression data with class scores. In *Pacific Symposium on Biocomputing 2002*, volume 7, pages 474–485. World Scientific Publishing Co., January 2002.

[29] G. Ruvkun and J. Guisto. The caenorhabditis elegans heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. *Nature*, 338:313–319, 1989.

[30] J. Shrager, P. Langley, and A. Pohorille. Guiding revision of regulatory models with expression data. In *Pacific Symposium on Biocomputing 2002*, volume 7, pages 486–497. World Scientific Publishing Co., January 2002.

[31] R. Somogyi and C.A. Sniegoski. Modeling the complexity of genetic networks: Understanding multigene and pleiotropic regulation. *Complexity*, 1:45–63, 1996.

[32] A.V. Spirov and et al. The inverse problem for ode models of genetic networks: The determination of parameters of the system from experimentally observed data. In *http://academic.mssm.edu/molbio/tmp/circuits/hox1circ.html*, 1997.

[33] P. Spirtes, C. Glymour, and R. Scheines. Constructing bayesian network models of gene expression networks from microarray data. In *Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems and Technology*, 2000.

[34] M. Stephens, M. Palakal, S. Mukhopadhyay, R. Raje, and J. Mostafa. Detecting gene relations from medline abstracts. In *Pacific Symposium on Biocomputing 2001*, volume 6, pages 483–496, Hawai, January 2001. World Scientific Publishing Co.

[35] Z. Szallasi. Genetic network analysis in light of massively parallel biological data acquisitions. In *Pacific Symposium on Biocomputing '99*, volume 4, pages 5–16. World Scientific Publishing Co., 1999.

[36] S. Tavazoie, J.D. Hughes, M. J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, July 1999.

[37] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281:827–842, 1998.

[38] E.P. van Someren, L.F.A. Wessels, and M.J.T. Reinders. Linear modeling of genetic networks from experimental data. In R. Altman, T.L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I.N. Shindyalov, L.F. Ten Eyck, and H. Weissig, editors, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 355–366, La Jolla, California, August 2000. AAAI.

[39] E.P. van Someren, L.F.A. Wessels, and M.J.T. Reinders. Genetic network models: A comparative study. In *Proceedings of SPIE, Micro-arrays: Optical Technologies and Informatics*, volume 4266, pages 236–247, San Jose, California, January 2001.

[40] E.P. van Someren, L.F.A. Wessels, M.J.T. Reinders, and E. Backer. Robust genetic network modeling by adding noisy data. In *Proceedings of the 2001 IEEE - EURASIP Workshop on Nonlinear Signal and Image Processing*, Baltimore, Maryland, June 2001.

[41] E.P. van Someren, L.F.A. Wessels, M.J.T. Reinders, and E. Backer. Searching for limited connectivity in genetic network models. In *Proceedings of the Second International Conference on Systems Biology*, pages 222–230, Pasadena, California, November 2001.

[42] E.P. van Someren, L.F.A. Wessels, M.J.T. Reinders, and E. Backer. Regularization and noise injection for improving genetic network models. In *Chapter 12 in Computational and Statistical Approaches to Genomics*. Kluwer, 2002.

[43] M. Wahde and J. Hertz. Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, 55(1 - 3):129–136, February 2000.

[44] M. Wahde and J. Hertz. Modeling genetic regulatory dynamics in neural development. *Journal of Computational Biology*, 2000.

[45] M. Wahde, J. Hertz, and M.L. Andersson. Reverse engineering of sparsely connected genetic regulatory networks. In *Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology*, 2001.

[46] D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. In *Pacific Symposium on Biocomputing '99*, volume 4, pages 112–123, Hawai, January 1999. World Scientific Publishing Co.

[47] X. Wen, S. Fuhrman, G.S. Michaels, D.B. Carr, S. Smith, J.L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. In *Proceedings of the National Academy of Sciences of the USA*, volume 95 of *1*, pages 334–339. National Academy of Sciences, 1998.

[48] L.F.A Wessels, E.P. van Someren, and M.J.T. Reinders. A comparison of genetic network models. In *Pacific Symposium on Biocomputing 2001*, volume 6, pages 508–519, Hawai, January 2001. World Scientific Publishing Co.

[49] L. Wong. Pies, a protein interaction extraction system. In *Pacific Symposium on Biocomputing 2001*, volume 6, pages 520–531, Hawai, January 2001. World Scientific Publishing Co.

[50] P.J. Woolf and Y. Wang. A fuzzy logic approach to analyzing gene expression data. *Physiological Genomics*, 3:9–15, 2000.

Stochastic models of aging and mortality

David Steinsaltz
University of Califonia, Berkeley

1 Introduction

One of the most remarkable empirical discoveries of 19th century statistics is the Gompertz mortality curve[Gom25]: over a broad range of ages, human mortality rates increase exponentially with age to an exceptional degree of precision — even the rates attributed to many individual causes of death. The succeeding nearly two centuries have largely confirmed and extended Gompertz's observation[OC97], to a great many other creatures[Fin90], from the large to the small, under proper interpretation even all the way down to the humble *Saccharomyces cerivisiae*, budding yeast [JEC89]. The one major exception — already suggested by W. Makeham [Mak67] in 1867 — is a deceleration of mortality-rate increase in extreme old age. (For an overview of the history of the search for demographic laws of mortality, with emphasis on the Gompertz curve, see [OC97].)We discuss some of the many mathematical models that have been proposed to explain this pattern. While considerable ingenuity and insight have gone into these models, it is remarkable how often they are undermined by errors of mathematics or of logic, abetted by wishful thinking.

2 Markov mortality models

Many of the models are of a type sometimes referred to as a *changing-vitality mortality model*. These represent the state of senescence (or, reversing the valence, the "vitality") of an organism as a Markov process (or, occasionally, a deterministic process), with death as a random stopping time for the process, defined either by a state-dependent killing rate, or by instantaneous killing when the process hits a lethal set. The most famous of these models goes back to 1960, and is due to B. Strehler and A. Mildvan.[SM60] Here, vitality is a fixed line with negative slope. The organism is confronts "challenges to homeostasis," which arise at a constant rate. The challenges are assumed to be denominated in units of energy, and have the Maxwell-Boltzmann distribution, which gives them a distribution $P\{X > x\} = e^{-kx}$. Thus, the probability of exceeding the vitality $-at + b$ is $e^{-kb+kat}$ at age t. The defect of this theory is its arbitrariness. The Maxwell-Boltzmann distribution is offered with no justification, while the explanation of the linear decline in vitality — itself an undefined quantity — is perfunctory, referring merely to certain measures of physiologic capacities whose population average declines with age. The arbitrariness is compounded by obviousness: It is hardly a surprise that combining a declining line with a negative exponential would yield an increasing exponential.Similar criticisms could be leveled against a Markovian model proposed by H. Le Bras [Bra76], and also discussed by Gavrilov and Gavrilova [GG91]. Le Bras imagines the possible senescence states of an organism being denominated by the natural numbers. Motion is only upward, by a continuous time Markovian

birth process. The transition rate from state n to $n+1$ is λn. The process is killed at a random time, with a rate given by μn when in state n. It is easy to compute that the hazard rate for the killing time is

$$\frac{\mu(\mu+\lambda)e^{(\mu+\lambda)t}}{\lambda+\mu e^{(\mu+\lambda)t}}.$$

Again, the Gompertz curve arises only under very particular assumptions (here, the linear growth both of the transition rate and of the killing rate), which are also arbitrary. And again, it is obvious that something like an exponential should come out when the rate of increase is proportional to the current state, by analogy to the differential equation $dx/dt = kx$. On the other hand, the smooth interpolation between the exponential and a constant asymptotic rate is not entirely obvious. The arbitrariness also seems less crass here, where the entire approach is more abstract and diagrammatic, rather than purporting to be a true description of the aging process.J. D. Abernethy [Abe79] proposes that organisms may be viewed as assemblages of independent identical components, such that the failure of any one implies the failure of the whole system. If the failure time distribution for each component is $F(t) = P\{T_i > t\}$, then the time of "death" has distribution

$$P\{T > t\} = P\{\min T_i > t\} = F(t)^n,$$

where n is the number of components. It is then shown that as n goes to ∞, and the time is properly rescaled, the Gompertz curve is one possible limit form for the hazard rate. This is hardly surprising, and is essentially a well known theorem in extreme-value theory. On the other hand, in the nonmathematical introduction and conclusion, it is claimed that the Gompertz curve is the only possible limit, which is simply untrue. In fact, only very extreme component hazard rates, such as the Gompertz itself, will produce a Gompertz curve in the limit.The idea of reproducing exponentially increasing hazard rates from the composition of independent devices with simpler failure times, in particular from constant hazard rates, has attracted many modelling attempts. One of the earlier efforts was M. Witten's 1985 paper [Wit85], which presented a model quite similar to Abernethy's, except that the components are redundant, so that death is identified with the maximum of the individual failure times. With an appropriate approximation scheme, it then derives a hazard rate of the form $ke^{\alpha t}$. Unfortunately, as L. Gavrilov and N. Gavrilova pointed out in [GG91], this α must be negative, which means that the hazard rate must be exponentially *decreasing*, not increasing. Having recognized the error of Witten's ways, Gavrilov and Gavrilova set out in the same book to repair it with a new model of their own, which is most completely presented in their later paper [GG01]. In doing so, they stumble nearly as badly. They combine the reliability models of Abernethy and Witten, proposing a structure with m essential "blocks", each comprising n redundant "components". A block fails when all of its components fail, and the organism dies when the first block fails. They approximate the hazard rate for small times t to be a power of t, the so-called Weibull hazard rate. They then propose that the number of components in each block should be random, and should have a Poisson distribution. The idea is that organisms are composed of mainly faulty components, and they argue that this explains the essential difference

between technical devices (which tend to Weibull failure rates) and organisms (which generally have approximately Gompertz rates). The main argument in favor of this theory is that they do derive Gompertzian hazard rates from it. Weighing against it, aside from the absence of plausible biological evidence, or even a coherent biological model, is the fact that the derivation is incorrect. In computing the hazard rate — the logarithmic derivative of the distribution function — for the random starting state, where they should mix the distribution functions and then compute the logarithmic derivative, they instead mix the logarithmic derivatives, which gives a very different, and incorrect, answer. The correct answer does not look significantly like a Gompertz curve.A different approach has been taken by J. Anderson [And00], and independently by J. Weitz and H. Fraser [WF01]. They propose to model vitality by a diffusion, namely, a Brownian motion with constant downward drift, killed when it hits 0. Weitz and Fraser use the well-known formula for the hitting time of drifting Brownian motion to show that this process produces mortality plateaus. This computation offers little insight into the fundamental reasons, and no indication of whether the result extends beyond this arbitrary special case. D. Steinsaltz and S. Evans [SE] have embedded the mortality plateaus for this and other Markov mortality models in the general theory of quasistationary distributions for Markov processes. In particular, they point out that the Tweedie R-theory (see especially [Twe74] and [TT79]) allows us to identify mortality plateaus with maximum eigenvalues of the Markov generator. This offers an intuitive explanation for the mortality plateaus, distinct from the standard ones, that either the population is initially heterogeneous (so that the progressive selection for more robust individuals produces an apparent levelling off of mortality rates) or that the aging process itself slows down with age. The alternative suggested by Markov models is that the state of health of the long-time survivors is conditioned by their survival to have a fixed distribution, which cannot be arbitrarily close to death. It is shown that this convergence can be guaranteed either by a compact space of vitalities, by strong inward drift (as in the Anderson-Weitz-Fraser model), or by killing rates that are sufficiently large at infinity, forcing the conditioned process to lurk near the origin (as in the Le Bras model).

3 Evolutionary models

Whereas the models discussed in section 2 are intendedprimarily to illustrate the mechanics of aging within individuals,there are also population-level models, based on evolutionary theoriesof senescence. These theories began in earnest with work byP. Medawar [Med57] and G. Williams [Wil57], who proposed thataging results from the accumulation of alleles which reduce theorganism's vitality at late ages, and either produce benefits("antagonistic pleiotropy") or simply no harm ("mutationaccumulation") earlier. The idea of mutation accumulation is that naturalselection has little power to remove damaging alleles from thepopulation if the damage comes late in life, after most of theorganism's likely reproduction has already been accomplished. If theallele in question actually produces benefits at earlier ages, then itmay in fact be positively selected.B. Charlesworth has produced an extensive corpus of work on populationgenetics in general, and the theory of senescence in particular. Inhis recent paper [Cha01], he

attempts to show that the mutationaccumulation model predicts Gompertz mortality. The idea is toimagine there to be a high level λ of extrinsic mortality, which isindependent of age. The probability of surviving to age x in thisregime will be $e^{-\lambda x}$. If we suppose that the organismsreproduce at a constant rate throughout their lifetimes, then thefitness cost to a marginal increase in mortality is also proportionalto $e^{-\lambda x}$. A simple calculation shows then that we can approximate the equilibrium prevalance of a mutant allele which produces a small unitincrease in mortality at age x to be $\nu e^{\lambda x}$, whereν is the rate at which the mutations are being generated. If the effects of a mutation have an additional component causing small harm at all ages, we find instead a mortality increment of the form $Ae^{\lambda x}/(B + Ce^{\lambda x})$, which is initially exponential, but converges to a plateau. A limitation of this model is the assumption that the mutants harm onlya single age (or all ages above a given age), rather than having morecomplex patterns of effects. Equally questionable is the reliance on the vaguely defined notion of "extrinsic mortality". With little empirical justification, we must assume that there is a natural age-independent mortality which swamps the accretion of senescence. The Gompertz pattern appears then as the result of artificially suppressing the natural background mortality. This would seem to contradict the observation of Gompertz-like mortality rates even in field studies of other species. It also suggests that the doubling time of mortality in modern humans (about 8 years) should be about the same as the life expectancy of our Pleistocene ancestors, an inference which is vastly at odds with current thinking about prehistoric demography (cf. [Lan90]). Charlesworth does try to address these objections by iterating the model: the computed mortality increment becomes part of the background for the next round of computation. This variant is carried out only in simulations. The results are sketchy, and hard to interpret. More work along these lines would be helpful. L. Mueller and M. Rose [MR96] have used a related model to study the effects of antagonistic pleiotropy. They consider a Markov chain whose states are sequences of mortality rates at 101 different ages. One move of the chain involves picking a random mutation, and allowing it to either become fixed in the population or disappear, with probability depending on the fitness benefit relative to the current mortality rates. A mutation raises mortality in one randomly chosen window of ages, and reduces it in another window. One would expect that mortality would rise without limit, since there is a positive feedback: the higher mortality rises at late ages, the less selective pressure there is against further increases. The claim in this paper, though, on the basis of simulations, is that the mortality rates develop a plateau at late ages. As K. Wachter points out in [Wac99], the states achieved in these simulations are not stationary states. If the simulations had run longer, the plateaus would have disappeared. While one might argue that evolution in the real world never achieves its "stationary states", it is hard to say what it would mean to associate a transient state of the Markov model with a general condition achieved by evolution. Not only is the time-scale arbitrary, but there is no credible interpretation of the starting state.

References

[Abe79] J. D. Abernethy. The exponential increase in mortality rate with age attributed to wearing-out of biological components. *Journal of Theoretical Biology*, 80:333–54, 1979.

[And00] James J. Anderson. A vitality-based model relating stressors and environmental properties to organism survival. *Ecological monographs*, 70(3):445–70, 2000.

[Bra76] H. Le Bras. Lois de mortalité et age limite. *Population*, 33(3):655–91, May-June 1976.

[Cha01] Brian Charlesworth. Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation-accumulation theory of ageing. *Journal of Theoretical Biology*, 210(1):47–65, 2001.

[Fin90] Caleb Finch. *Longevity, Senescence, and the Genome*. University of Chicago Press, Chicago, 1990.

[GG91] Leonid A. Gavrilov and Natalia S. Gavrilova. *The biology of lifespan: a quantitative approach*. Harwood Academic Publishers, Chur, Switzerland, 1991.

[GG01] Leonid A. Gavrilov and Natalia S. Gavrilova. The reliability theory of aging and longevity. *Journal of Theoretical Biology*, 213:527–45, 2001.

[Gom25] Benjamin Gompertz. On the nature of the function expressive of the law of human mortality and on a new mode of determining life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–85, 1825.

[JEC89] S. M. Jazwinski, N. K. Egilmez, and J. B. Chen. Replication control and cellular life span. *Experimental Gerontology*, 24:423–436, 1989.

[Lan90] H. O. Lancaster. *Expectations of Life: A study in the demography, statistics, and history of world mortality.* Springer-Verlag, New York, Heidelberg, Berlin, 1990.

[Mak67] William Makeham. On the law of mortality. *Journal of the Institute of Actuaries*, 13:325–58, 1867.

[Med57] Peter Medawar. An unsolved problem in biology. In *The uniqueness of the individual*. Basic Books, 1957.

[MR96] Laurence D. Mueller and Michael R. Rose. Evolutionary theory predicts late-life mortality plateaus. *Proc. Natl. Acad. Sci. USA*, 93(26):15249–53, December 24 1996.

[OC97] S. Jay Olshansky and Bruce A. Carnes. Ever since Gompertz. *Demography*, 34(1):1–15, February 1997.

[SE] David Steinsaltz and Steven N. Evans. Markov mortality models: implications of quasistationarity and varying initial conditions. Preprint. Available at http://www.demog.berkeley.edu/ dstein/agingpage.html.

[SM60] B. Strehler and A. Mildvan. General theory of mortality and aging. *Science*, 132(3418):14–21, 1960.

[TT79] Pekka Tuominen and Richard L. Tweedie. Exponential decay and ergodicity of general Markov processes and their discrete skeletons. *Adv. in Appl. Probab.*, 11(4):784–803, 1979.

[Twe74] Richard Tweedie. r-theory for Markov chains on a general state space I: solidarity properties and r-recurrent chains. *The Annals of Probability*, 2:840–64, 1974.

[Wac99] Kenneth W. Wachter. Evolutionary demographic models for mortality plateaus. *Proc. Natl. Acad. Sci. USA*, 96:10544–7, August 1999.

[WF01] Joshua Weitz and Hunter Fraser. Explaining mortality rate plateaus. *Proc. Natl. Acad. Sci. USA*, 98(26):15383–6, December 18 2001.

[Wil57] George C. Williams. Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, 11:398–411, December 1957.

[Wit85] M. Witten. A return to time, cells, systems, and aging: III. Gompertzian models of biological aging and some possible roles for critical elements. *Mechanisms of Ageing and Development*, 32:141–77, 1985.

List of Participants

Jørn Attermann
Deptartment of Biostatistics
University of Aarhus
Vennelyst Boulevard 6
DK-8000 Århus C
Denmark

jorn@biostat.au.dk

Bo Martin Bibby
Department of Mathematics and Physics
The Royal Veterinary and Agricultural University
Thorvaldsensvej 40
DK-1871 Frederiksberg C
Denmark

bibby@dina.kvl.dk
http://www.dina.kvl.dk/∼bibby/

Dennis Bray
Department of Anatomy
University of Cambridge
Downing Street
Cambridge CB2 3DY
United Kingdom

db10009@cam.ac.uk
http://info.anat.cam.ac.uk/groups/comp-cell/DennisBray.html

Sune Danø
Kemisk Laboratorium
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark

sdd@ki.ku.dk

Susanne Ditlevsen
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3
DK-2200 København N
Danmark

s.ditlevsen@biostat.ku.dk
http://www.pubhealth.ku.dk/∼sudi/

Patrik Eden
Deptartment of Theoretical Physics
Lund University
Sölvegatan 14A
S-223 62 Lund
Sweden

patrik@thep.lu.se

Gino Favero
Università degli Studi di Padova
Via Belzoni 7
I-35131 Padova, Italy
Italy

favero@math.unipd.it

Raouf Ghomrasni
Department of Mathematical Sciences
University of Aarhus
Ny Munkegade
DK-8000 Århus C
Denmark

raouf@imf.ai.dk

Bryan T. Grenfell
Department of Zoology
University of Cambridge
Downing Street
Cambridge CB2 3EJ
United Kingdom

bryan@zoo.cam.ac.uk
http://www.zoo.cam.ac.uk/zoostaff/grenfell/people/bryan.htm

Margaret Gutowska
Edith Cowan University
10 Harness Street
Kingsley
6026 Perth, Western Australia
Australia

Mgutowska@aol.com

Ernst Hansen
Department of Applied Mathermatics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark

erhansen@math.ku.dk
http://www.math.ku.dk/∼erhansen

Jesper Schmidt Hansen
Deptartment of Life Science and Chemistry
Roskilde University
Denmark

schmidt@koala.ruc.dk

Michael Höhle
Department of Animals Science and Animals Health
The Royal Veterinary and Agricultural University
Denmark

hoehle@dina.kvl.dk

Marianne Huebner
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824
U.S.A.

http://www.stt.msu.edu/∼huebner
huebner@stt.msu.edu

Valerie Isham
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
United Kingdom

valerie@stats.ucl.ac.uk

Martin Jacobsen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark

martin@math.ku.dk

Mogens Høgh Jensen
The Niels Bohr Institute
University of Copenhagen
Blegdamsvej 17
DK-2100 Copenhagen
Denmark

mhjensen@nbi.dk

Hidde de Jong
Institut National de Recherche en Informatique et en Automatique (INRIA)
Unité de recherche Rhône-Alpes
655 avenue de l'Europe
Montbonnot
38334 Saint Ismier Cedex
France

Hidde.de-Jong@inrialpes.fr

Niels Keiding
Department of Biostatistics
University of Copenhagen
Blegdamsvej 3
DK-2200 København N
Danmark

N.Keiding@biostat.ku.dk
http://www.pubhealth.ku.dk/bsa/staff/nk-e.htm

Timo Koski
Department of Mathematics
University of Linköping
S-58183 Linköping
Sweden

tikos@mai.liu.se
http://www.mai.liu.se/∼tikos/Welcome_e.html

Anders Krogh
Bioinformatics Centre
University of Copenhagen
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark

krogh@binf.ku.dk
www.binf.ku.dk/users/krogh

Catherine Larédo
Institut National de la Recherche Agronomique (INRA)
Lab. de Biométrie
Centre de Recherche de Jouy-en-Josas
F-78352 JOUY-EN-JOSAS
France

cl@banian.jouy.inra.fr

Morten Lindow
Bioinformatics Centre
University of Copenhagen
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark

morten@binf.ku.dk

Mogens Sandø Lund
Danish Institute of Agricultural Sciences
Research Centre Foulum
DK-8830 Tjele
Denmark

Mogens.Lund@agrsci.dk

Leonardo De Maria
Novozymes A/S
Smrmosevej 25, 2C S44
DK-2880 Bagsvrd Denmark

LeDM@novozymes.com

Rosa Maria Mininni
Dept. Interuniversitario de Matematica
Università di Bari
Via Orabona, 4
I-70125 Bari
Italy

mininni@dm.uniba.it

Rune Viig Overgaard
Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads - Building 321
DK-2800 Lyngby
Denmark

rvo@imm.dtu.dk
http://www.imm.dtu.dk/∼rvo/

Thomas Poulsen
Novozymes A/S
Smrmosevej 25, 2C S44
DK-2880 Bagsvrd
Denmark

tapo@novozymes.com

Gesine Reinert
Deptartment of Statistics
University of Oxford
1 South Parks Road
Oxford OX1 3TG
United Kingdom

reinert@stats.ox.ac.uk
http://www.stats.ox.ac.uk/people/reinert/index.htm

Michael S. Samoilov
Engineering Systems Research Center
University of California, Berkeley
Physical Biosciences Division
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720
U.S.A.

MSSamoilov@lbl.gov

Ib Michael Skovgaard
Department of Mathematics and Physics
The Royal Veterinary and Agricultural University
Thorvaldsensvej 40
DK-1871 Frederiksberg C
Denmark

Ib.M.Skovgaard@imf.kvl.dk
http://www.dina.kvl.dk/∼ib

Eugene van Someren
Information and Communication Theory Group
Faculty of Information Technology Systems
Delft University of Technology
Mekelweg 4
2628 CD Delft
The Netherlands

E.P.vanSomeren@its.tudelft.nl

Helle Sørensen
Department of Mathematics and Physics
The Royal Veterinary and Agricultural University
Thorvaldsensvej 40
DK-1871 Frederiksberg C
Denmark

helle@dina.kvl.dk
http://www.matfys.kvl.dk/∼helle/

Michael Sørensen
Department of Applied Mathematics and Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen Ø
Denmark

michael@math.ku.dk
http://www.math.ku.dk/∼michael

David Steinsaltz
Department of Demography
2232 Piedmont Ave.
University of Califonia
Berkeley, CA 94720-2120
U.S.A.

dstein@demog.berkeley.edu

Fengzhu Sun
Department of Biological Sciences
University of Southern California
1042 West 36th Place, DRB 288
Los Angeles, CA 90089-1113
U.S.A.

fsun@usc.edu
http://www-hto.usc.edu/∼fsun

Kasper Munch Terkelsen
Bioinformatics Centre
University of Copenhagen
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark

kasper@binf.ku.dk
www.binf.ku.dk/users/kasper

Uffe Høgsbro Thygesen
Danish Institute for Fisheries Research
Charlottenlund Slot
Jægersborg All 1
DK-2920 Charlottenlund
Denmark

uht@dfu.min.dk
http://www.dfu.min.dk

Christoffer Wenzel Tornøe
Clinical Pharmacology & Kinetics Department
Ferring Pharmaceuticals
Kay Fiskers Pl.11
DK-2300 Copenhagen S
Denmark

christoffer.tornoe@ferring.com

Carl Troein
Department of Theoretical Physics
Lund University
Sölvegatan 14A
S-223 62 Lund
Sweden

carl@thep.lu.se

Henrik Wachmann
Danske Slagterier
Axelborg
Axeltorv 3
DK-1609 Copenhagen V
Denmark

hew@danskeslagterier.dk