Ecology and Evolution: Populations

701-2413-00L WS or 701-1415-00L

Sebastian Bonhoeffer
Theoretical Biology
Institute of Integrative Biology
ETH Zürich

September 27, 2011

Contents

List of Figures

List of Tables

Foreword

Interactions between populations and interactions of the populations with their environment typically are highly nonlinear and thus their dynamical behaviour frequently defies intuition. Therefore mathematical models are important tools to establish the factors underlying the temporal changes in the abundance of natural populations. Traditionally the models and methods developed in population biology have been used to describe populations of animals or plants. In this context models play a crucial role, for example, in helping to establish how natural populations respond to habitat fragmentation and are thus relevant for issues regarding conservation biology. More recently the methods and models have been adopted to describe the population biology of infectious diseases both within a host population and within an individual host. In this context, mathematical models have made important contributions to our understanding of the control of infectious diseases, the consequences of vaccination, and the pathogenicity of infectious agents.

In contrast to most other areas of biology mathematical models play a central role in ecology and evolution. Therefore it is important to be familiar with the underlying concepts and methods. The degree of mathematical sophistication in the population genetic and population biological literature can be high. However, in a course such as this one, it will not be possible to go into great mathematical detail, nor is it possible to give complete derivations of all mathematical concepts introduced. The emphasis of this course will therefore be on developing an intuitive understanding and familiarity with the concepts used in population biology.

Chapter 1

Toolbox: Population dynamics

1.1 Continuous time models of a single species

The rate at which a population increases or decreases typically depends on its current value. In mathematical terms this can be expressed as

$$dn(t)/dt = F(n(t)) \tag{1.1}$$

where F is a function that relates the population's rate of change $dn(t)/dt$ to its size $n(t)$ at the time t. An equation that relates the derivatives of a variable (here $dn(t)/dt$) to the value of the variable itself (here $n(t)$) as a continuous function of a single parameter (here t) is called a *ordinary differential equation*. A differential equation is called linear, if F is a linear function of n.

Often it is easier to formulate the dynamics of a population in terms of the *per capita growth rate*, $C(n)$, i.e. the growth rate per individual in the population. In these terms the dynamics are given by

$$dn/dt = C(n)n$$

where $C(n) = F(n)/n$. Note, that from here on I drop the explicit time dependence of n for notational convenience.

1.1.1 Exponential growth

The simplest model of population growth is obtained if one assumes that the per capita growth rate is constant over time, i.e.

$$dn/dt = rn,$$

where the per capita growth rate is $C(n) = r$. (Note, that here n is still a function of time, while r is a constant.)

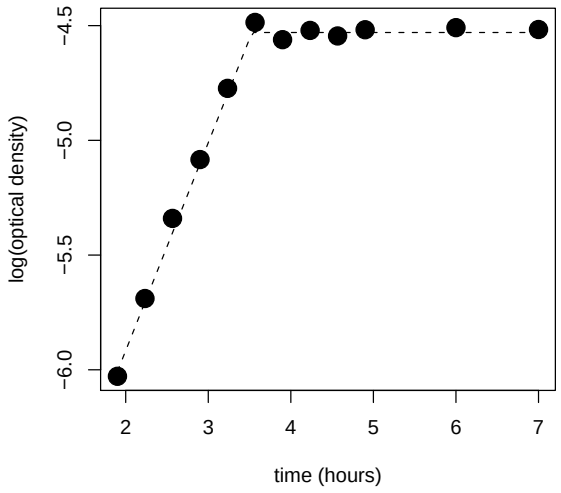

Figure 1.1: Exponential growth in a bacterial culture of E.coli in minimal glucose medium. The population size is plotted as the logarithm of the optical density in a spectrophotometer. The growth is to a very good approximation exponential (i.e. linear in a log plot) until the bacteria run out of glucose and enter stationary phase.

Today differential equations (in as far as they can be solved analytically at all) can often be solved using standard mathematical software (such as Mathematica or Matlab). The solution to this differential equation, however, is very simple and can be obtained by integration.

$$\int_{n_0}^{n} \frac{1}{n'} dn' = \int_0^t r dt' \tag{1.2}$$

$$log(\frac{n}{n_0}) = rt \tag{1.3}$$

$$n = n_0 e^{rt} \tag{1.4}$$

where n_0 is the population density at time $t = 0$. Thus, the solution is the well-known exponential function. (You can easily verify that the exponential function is a solution to the above differential equation by differentiation.)

The exponential growth is a often a reasonable description for the initial growth of biological populations in the absence of limiting factors (see for the examples of bacterial growth in culture or of the HIV epidemic in South Africa in figs. 1.1 and 1.2) . However, the exponential function fails as a description for the long term since no population can ever grow unboundedly. The problem is that the assumption of a constant per capita growth rate is unrealistic. Often factors such as resource depletion or competition for space (such as territories) lead to a decrease of the per capita growth rate with increasing population size.

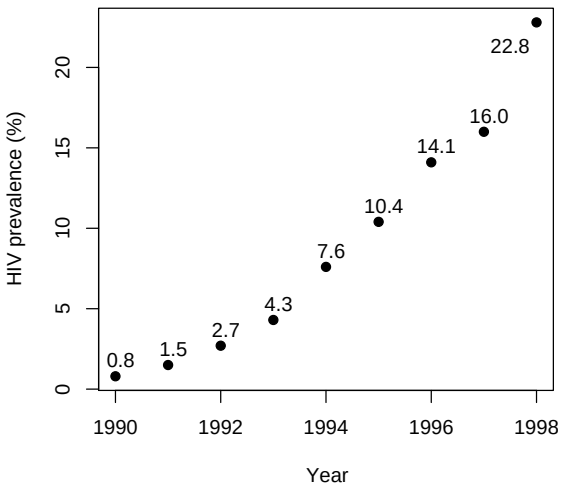

Figure 1.2: Prevalence of HIV in South Africa from 1990 to 1998. It can be seen that in the first few years the prevalence roughly doubles each year. In later years the rate of increase begins to decrease, in part because of increased public awareness of the risks of HIV infection, but also because of the decrease of susceptible individuals.

1.1.2 Logistic growth (in continuous time)

The simplest way of introducing a density dependent regulation is to assume that the per capita growth rate decreases linearly with population size, i.e. $C(n) = r(1 - n/K)$, where K is the *carrying capacity*. The corresponding differential equation, the *logistic differential equation*, is thus given by

$$dn/dt = rn(1 - \frac{n}{K}) \tag{1.5}$$

Here, the parameter r describes the maximal per capita growth rate (i.e. the per capita growth rate for when the population is vanishingly small). The solution of the logistic differential equation can be obtained either analytically or using mathematical software such as Mathematica or Maple. It is given by

$$n(t) = n_0 \frac{K}{n_0 + (K - n_0)e^{-rt}} \tag{1.6}$$

where n_0 is the population size at time $t = 0$[a]. The interpretation of the carrying capacity K becomes clear by inspecting the behavior of this equation for large times. The carrying capacity represents the value that the population attains as the time t goes to infinity.

Starting from a small population n_0 at time $t = 0$, the time $t^\star$ taken until the population

[a]Exercise: Verify (i) that n_0 is indeed the solution of this equation for time $t = 0$ and (ii) that this equation is a solution of the logistic differential equation by differentiation and substitution into the differential equation 1.5.

reaches half its maximal value $n = K/2$ can be calculated as

$$K/2 = n_0 \frac{K}{n_0 + (K - n_0)e^{-rt^\star}} \tag{1.7}$$

or

$$t^\star = -\frac{1}{r}\log(\frac{n_0}{K - n_0}) \approx -\frac{1}{r}\log(\rho_0) \tag{1.8}$$

where $\rho_0 = n_0/K$. Thus the time to reach the half of the carrying capacity depends inversely on r but logarithmically on n_0.

1.1.3 Stability analysis of the logistic differential equation

If $n_0 = K$ then the population will remain at this value forever (see eq. 1.6). Thus $n_0 = K$ represents an equilibrium of the logistic differential equation. Note that depending on whether the population size is initially larger or smaller than the carrying capacity (i.e. $n_0 > K$ or $n_0 < K$), the population increases or decreases monotonically towards the carrying capacity. This implies that K is a *globally stable* equilibrium, since the population will eventually go to K for any value of n_0 (except $n_0 = 0$).

Clearly, $n_0 = 0$ is another equilibrium, since $n(t) = 0$ for all t, if $n_0 = 0$. However, this equilibrium is *unstable*, because any small deviation from $n_0 = 0$ will always lead the population to diverge from this equilibrium. To illustrate the concept of an unstable equilibrium, think of a pendulum with its center of gravity placed exactly vertically above its point of fixation. Any small deviation from that exact position will make the pendulum fall. When the center of gravity is exactly below the point of fixation the pendulum is in a stable equilibrium, since the pendulum will always return to that equilibrium for any small perturbation of its position.

In many cases it is not possible to derive a closed analytical expression for the solution to a differential equation, and therefore the dynamical behavior of the system and the stability characteristics of the equilibria cannot be simply deduced from the solution. However, the equilibria can be determined in the absence of an explicit solution of the differential equation. A population is in equilibrium if its size does not change over time. In mathematical terms this means that $dn/dt = 0$. Inspection of the logistic differential equation (see eq 1.5) shows that $dn/dt = 0$ if either $n = 0$ or $n = K$. The stability of these equilibria can be determined graphically. This approach is illustrated in figure 1.3. The graphical analysis can be summarized as follows: The equilibria are given as the values for which $F(n) = 0$. These equilibrium values are called stable if the derivative of F with regard to n (i.e. dF/dn) is negative at the equilibrium value. This implies that the population increases when it is smaller than the equilibrium value and decreases if the population is larger than the equilibrium value. Conversely, if $dF/dn > 0$ at an equilibrium value, then this implies that the equilibrium is unstable. Strictly speaking, the sign of dF/dn at an equilibrium value only determines its local stability, since the stability analysis is based on a linearization around the equilibrium value. Hence, it only determines the behavior of the dynamical system for small perturbations of the equilibrium value.

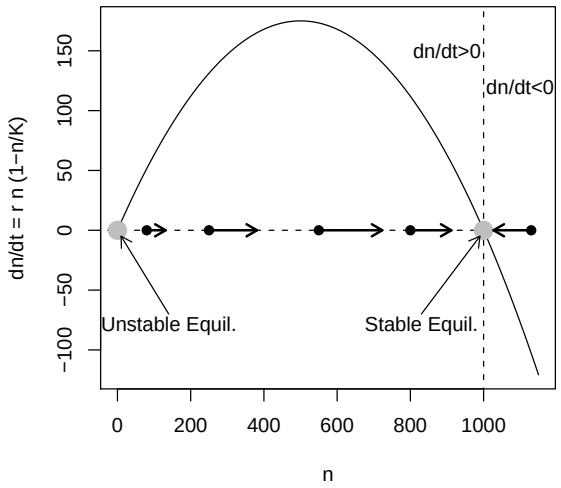

Figure 1.3: Graphical stability analysis of the logistic differential equation 1.5 using $r = 0.7$ and $K = 1000$. Here, we plot the $dn/dt = F(n) = rn(1 - n/k)$ (solid line) against the population size n. Whenever $F(n) > 0$ the population size will increase (see arrows on dashed line) since the population growth rate $dn/dt > 0$). $F(n) = 0$ for $n = K$(vertical dashed line). The length of the arrows on the dashed line indicate the growth rate of the population at the given value of n. For $n < K$ the population grows (since $dn/dt > 0$), while for $n > K$ the population declines (since $dn/dt < 0$). Thus, this graphical analysis indicates why the equilibrium $n = 0$ is unstable, while the equilibrium $n = K$ is stable.

1.1.4 Characteristics and limitations of the logistic differential equation

The logistic differential equation is a very simple description of a population dynamical process. It is therefore not too surprising that it frequently fails to reproduce important features of the growth of a population even under very simple lab conditions. Figure 1.4 shows the growth of a population of waterfleas (*Daphnia*) in the laboratory. A characteristic behavior is the that population overshoots its equilibrium value. This behavior cannot be accounted for by logistic growth[b].

One reason why the population may overshoot the carrying capacity is that there may be a time delay between reproduction and resource depletion. Such time delays can lead to oscillatory behaviour. One could thus attempt to make the logistic differential equation more realistic by assuming that the growth rate dn/dt does not depend on the population size at time t, but on the population size at time $t - \tau$, (where τ represents such a time delay). The decline of the per capita growth rate with increasing population size in the logistic differential equation is a heuristic description of the underlying biological processes. (This means, that we are just assuming that there must be a decline of the per capita growth rate with increasing population size, but we are actually not deriving how exactly the per capita growth rate should decline based on a specific biological process.) An alternative approach to describe the population dynamical process more mechanistically and model the dynamics of the population in parallel with the dynamics of the resource. This would imply that we have instead of just one differential equation two coupled differential equations (since the rate of change of the population depends on the resource level and the rate of change of the resource level depends on the population size).

1.2 Continuous time models of two species

1.2.1 Interactions between populations

Often the dynamics of a population are not only determined by the interactions between the individuals of the population, but also by the interactions with other populations. The most common type of interactions are competition (for resources, territory, etc) or predation. The fundamental difference between these two types of interactions is that for competition each populations has a negative impact on the other's growth, while for predation the predator has a negative effect on the prey population, but the prey has a positive effect on the predator population[c].

[b] Question: Why can't logistic growth overshoot the carrying capacity?. (Inspect eq. 1.5, or eq. 1.6, or figure 1.3)

[c] Question: What other type of interactions can you think of? How would you call an interaction, where both populations have a positive effect on each other?

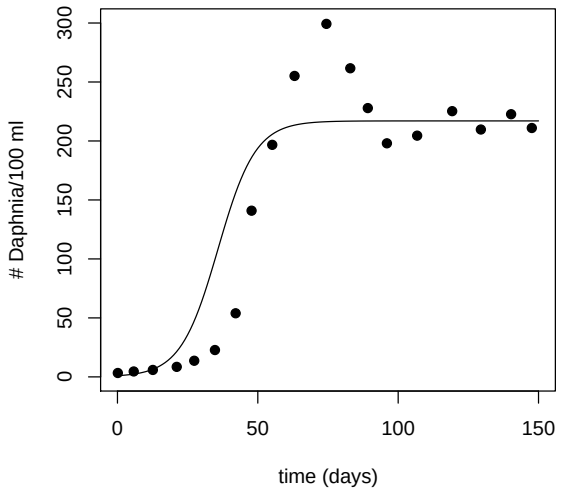

Figure 1.4: Growth of a *Daphnia* population in laboratory conditions. The black dots represent measured data, while the solid line represents the logistic growth function (eq. 1.6). The logistic growth function does not only fail to give a good quantitative description of the population dynamics, but also fails to reproduce some important qualitative aspects. In particular the population overshoots the carrying capacity, which is a phenomenon that cannot be captured by logistic growth.

1.2.2 Lotka-Volterra model

The Lotka-Volterra model[d] is the archetypal model for the competition between several populations. It represents a natural extension of the logistic differential equation. For two competing species the model is given by

$$dn_1/dt = r_1[1 - (n_1 + \gamma_{12}n_2)/K_1]n_1 \quad (1.9)$$
$$dn_2/dt = r_2[1 - (n_2 + \gamma_{21}n_1)/K_2]n_2 \quad (1.10)$$

where n_1 and n_2 are the densities of populations 1 and 2. Note, that if either n_1 or n_2 are zero, then the equation for the growth of the other population is identical to the logistic differential equation (eq 1.5). Thus, the interpretation of the parameters $r_{1,2}$ and $K_{1,2}$ are equivalent to the logistic equation. The interpretation parameters γ_{12} and γ_{21} becomes clear from inspecting the above equations. They describe the effects of competition between the two populations. If $\gamma_{12} > 1$, this implies that the negative effect of species 2 on species 1 is stronger than the negative effect of species 1 on itself. Hence, in this case for species 1 intraspecific competition is weaker than interspecific competition. Conversely, if $\gamma_{12} < 1$ then intraspecific competition is stronger than interspecific competition for species 1.

The populations are in equilibrium if both dn_1/dt and $dn_2/dt = 0$. There are four equilibria: (i) the trivial equilibrium $n_1 = n_2 = 0$ with both species absent, (ii) an equilibrium $n_1 = K_1$ and $n_2 = 0$, where species 1 is at carrying capacity and species 2 is absent, (iii) an equilibrium $n_1 = 0$ and $n_2 = K_2$, where species 2 is at carrying capacity and species 1 is absent, and (iv) an equilbrium

$$n_1 = \frac{K_1 - \gamma_{12}K_2}{1 - \gamma_{12}\gamma_{21}} \quad (1.11)$$
$$n_2 = \frac{K_2 - \gamma_{21}K_1}{1 - \gamma_{12}\gamma_{21}} \quad (1.12)$$

where both populations 1 and 2 coexist.

As is often the case for more complex differential equations, an analytical solution cannot be obtained. A numerical solution, however, can be easily obtained by numerical integration of the differential equation. Many mathematical packages offer sophisticated routines to integrate differential equations. A simple (yet often imprecise or inefficient) algorithm to integrate differ-

[d](i) Lotka, Alfred James (1880 - 1949), USA, chemist, demographer, ecologist and mathematician, was born in Lviv (Lemberg), at that time situated in Austria, now in Ukraine. He came to the United States in 1902 and wrote a number of theoretical articles on chemical oscillations during the early decades of the twentieth century, and authored a book on theoretical biology (1925). He then left (academic) science and spent the majority of his working life at an insurance company (Metropolitan Life). (ii) Vito Volterra (May 3, 1860 - October 11, 1940) was an Italian mathematician and physicist, best known for his contributions to mathematical biology. Born in Ancona, then part of the Papal States, into a very poor family. Volterra was professor of mathematics in Pisa, Turin, and Rome. After World War I, Volterra turned his attention to the application of his mathematical ideas to biology. The most famous outcome of this period is the Volterra-Lotka equations. In 1922, he joined the opposition to the Fascist regime of Benito Mussolini and in 1931 refused to take a mandatory oath of loyalty. He was compelled to resign his university post and membership of scientific academies, and, during the following years, he lived largely abroad, returning to Rome just before his death.

ential equations is given by the Euler method[e]. Here, the value of n_1 and n_2 at a time $t + \Delta t$ is obtained as

$$n_1(t + \Delta t) = n_1(t) + \Delta t \; dn_1(t)/dt \tag{1.13}$$
$$n_2(t + \Delta t) = n_2(t) + \Delta t \; dn_2(t)/dt \tag{1.14}$$

where Δt is a small time interval and $dn_1(t)/dt$ and $dn_2(t)/dt$ are given by eqs. 1.9 and 1.10. Iterating the above relation one obtains a numerical solution for the Lotka-Volterra model.

1.2.3 Dynamics of the Lotka-Volterra model

The dynamical behavior can be visualized in a variety of ways. The conventional method is to plot n_1 and n_2 as a function of time (see figure 1.5). Another way to plot the dynamics is a *phase diagram*, where one plots $n_1(t)$ against $n_2(t)$ (see figure 1.6). Finally, one can infer the dynamics from a *vector field diagram* (see figure 1.7), in which one plots arrows from selected points (n_1, n_2) to $(n_1 + \Delta t \; dn_1/dt, n_2 + \Delta t \; dn_2/dt)$ for a freely chosen small time step Δt. The size of the vectors in the vector field diagram are proportional to the velocity of change in the population sizes (n_1, n_2).

[e]Leonhard Euler (1707-1783) was arguably the greatest mathematician of the eighteenth century and one of the most prolific of all time; his publication list consists of 886 papers and books. Euler's complete works fill about 90 volumes. Remarkably, much of this output dates from the the last two decades of his life, when he was totally blind. Though born and educated in Basel, Switzerland, Euler spent most of his career in St. Petersburg and Berlin. He joined the St. Petersburg Academy of Sciences in 1727. In 1741 he went to Berlin at the invitation of Frederick the Great, but he and Frederick never got on well and in 1766 he returned to St. Petersburg, where he remained until his death. Euler's prolific output caused a tremendous problem of backlog: the St. Petersburg Academy continued publishing his work posthumously for more than 30 years. Euler married twice and had 13 children, though all but five of them died young. Euler's powers of memory and concentration were legendary. He could recite the entire Aeneid word-for-word. He was not troubled by interruptions or distractions; in fact, he did much of his work with his young children playing at his feet.

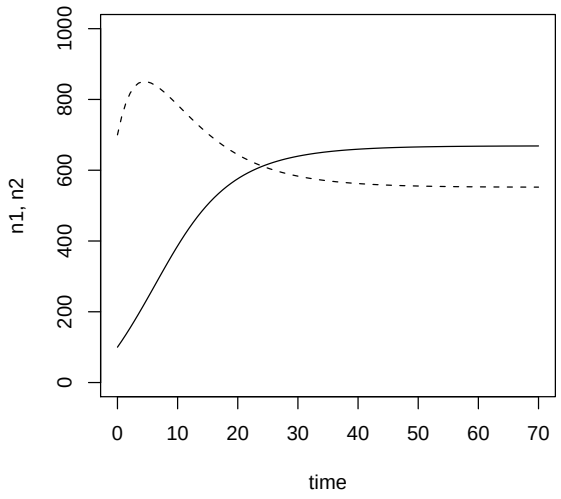

Figure 1.5: Time plot of the numerical solution of the Lotka-Volterra model. Population 1 is indicated by a solid line, population 2 is indicated by a dashed line. The parameters are $K_1 = K_2 = 1000, r_1 = r_2 = 0.5, \gamma_{12} = 0.6$ and $\gamma_{21} = 0.67$. The initial conditions (i.e. starting values at time $t = 0$) are $n_1 = 100$ and $n_2 = 700$.

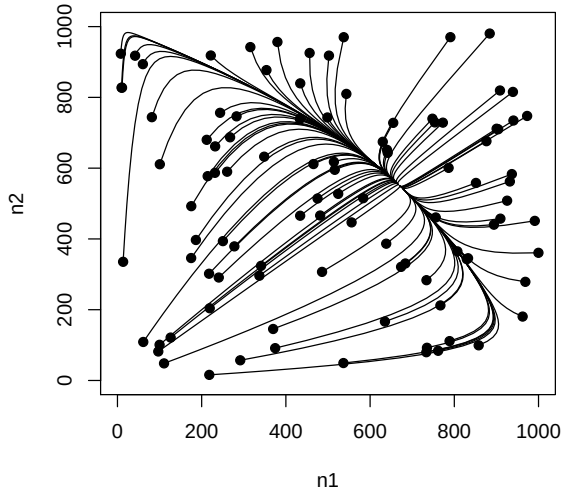

Figure 1.6: Phase diagram of the numerical solution of the Lotka-Volterra model. The parameters are identical to figure 1.5. The plot shows 100 trajectories starting from random initial conditions for n_1 and n_2 at time $t = 0$. The initial starting points are indicated as full circles. From these starting points the solution of the Lotka-Volterra model is then plotted until $t = 20$.

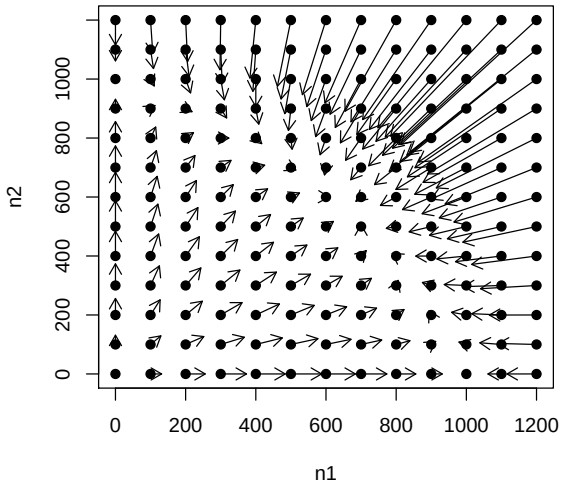

Figure 1.7: Vector field diagram of the Lotka-Volterra model. The parameters are identical to figure 1.5. The arrows connect the points (n_1, n_2) and $(n_1 + \Delta t\ dn_1/dt, n_2 + \Delta t\ dn_2/dt)$ for a freely chosen small time step Δt. The size of the vectors are thus proportional to the velocity of change. The relation between the vector field diagram and the phase diagram is that the velocity vectors represent the tangent to any point of a trajectory in the phase diagram.

1.2.4 Graphical stability analysis of the Lotka-Volterra model

In section 1.2.2 we have derived four steady states (equilibria) for the two-species Lotka-Volterra model (eq. 1.9 and 1.10). To determine the stability characteristics of these equilibria it is often useful to plot the *nullclines*[f] in a phase diagram. The nullclines are defined as the sets of points in the phase diagram where one of the variables (here n_1 or n_2) does not change over time (i.e. here $dn_1/dt = 0$ or $dn_2/dt = 0$). Hence, for the Lotka-Volterra model (eqs. 1.9 and 1.10) we have two n_1-nullclines for which $dn_1/dt = 0$. These are given by $n_1 = 0$ and $n_2 = (K_1 - n_1)/\gamma_{12}$. Analogously, we have two n_2-nullclines given by $n_2 = 0$ and $n_2 = K_2 - \gamma_{21} n1$. Figure 1.8 shows the n_1- and n_2-nullclines in the phase diagram and illustrates the graphical stability analysis of the Lotka-Volterra model.

1.2.5 Mathematical stability analysis

When working with more than two interacting species a graphical stability analysis is often less straightforward. A more general approach is therefore a local stability analysis. Here, I will illustrate the mathematical stability analysis first for the simpler one-dimensional case. (Here, one-dimensional refers to the fact that we consider only one population). Then, I will illustrate how the method is extended for the case of multidimensional systems without going into great mathematical detail. The idea is to introduce important concepts such as eigenvalues and the Jacobi matrix, since they appear frequently in the population biological literature. However, the goal is not to give a complete mathematical derivation of these concepts. To fully appreciate the mathematical basis of the procedures discussed in this chapter it is necessary to be familiar with such topics as matrix diagonalization and Taylor expansion which are discussed in detail in many textbooks. The interested reader is referred for example to Michael Bulmer's book entitled "Theoretical Evolutionary Ecology" published by Sinauer Press or Sarah Otto's and Troy Day's book entitled "A Biologist's Guide to Mathematical Modeling in Ecology and Evolution" published by Princeton University Press).

Let us consider the simple one-dimensional differential equation

$$dn/dt = F(n) \tag{1.15}$$

Let $\hat{n}$ be the equilibrium of the above differential equation (i.e. $F(\hat{n}) = 0$). In the vicinity of a chosen point any function can always be approximated by its value at that point and its first derivative. In particular, in the vicinity of $\hat{n}$ we can write

$$F(n) \approx F(\hat{n}) + (n - \hat{n}) \frac{dF}{dn}|_{n=\hat{n}} \tag{1.16}$$

where the vertical bar denotes that the derivative is evaluated at the point $\hat{n}$. Mathematically, this is called a Taylor expansion (in which all terms higher than first order are neglected).

Let us now study the dynamical behaviour of small perturbations from the equilibrium value so that

$$n = \hat{n} + p \tag{1.17}$$

[f] Nullclines are sometimes also referred to as isoclines. The definition of an isocline is $dn/dt = const.$

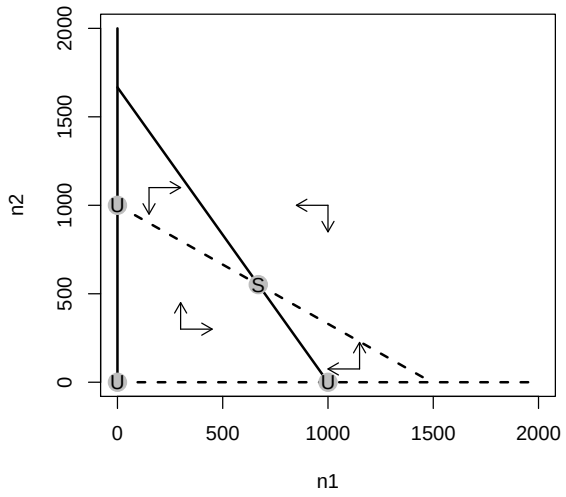

Figure 1.8: Nullclines of the Lotka-Volterra model in the n_1, n_2 phase diagram. The solid lines indicate the n_1 nullclines and the dashed lines indicate the n_2-nullclines. The four equilibrium points (grey dots, U for unstable, S for stable) are given by the intersections of the solid and the dashed lines. The parameters are $K_1 = K_2 = 100, \gamma_{12} = 0.6$ and $\gamma_{21} = 0.67$. Since both γ_{ij} are smaller than one, we are depicting here the situation in which the intraspecific is stronger than the interspecific competition in both populations. The n_1, n_2 nullclines separate the phase space into regions where $dn_1/dt > 0$ and $dn_2/dt > 0$, respectively. Hence, above (below) the n_1-nullcline, population 1 decreases (increases) as indicated by the arrows pointing in the direction of increasing (decreasing) n_1. Similarly, above (below) the n_2-nullcline, population 2 decreases (increases) as indicated by the arrows pointing in the direction of increasing (decreasing) n_2. The arrows thus indicated the stability characteristics of the 4 equilibria. (Note, that this figure only represents the case that both $\gamma_{ij} < 1$. There are three more cases depending on the values of K_1, K_2, γ_{12} andγ_{21}: The dotted and the dashed line may not intersect (which is the case one of the γ_{ij} is larger and the other smaller than one), or they may intersect but the n_2 nullcline is steeper than the n_1-nullcline (which is the case if both γ_{ij} are larger than one). For these cases the stability of the equilibria changes.

Substituting this in the above differential equation we obtain

$$d(\hat{n}+p)/dt = F(\hat{n}+p) \tag{1.18}$$

$$\approx F(\hat{n}) + p\frac{dF}{dn}|_{n=\hat{n}} \tag{1.19}$$

Note that $d\hat{n}/dt = 0$ since $\hat{n}$ is a constant and that $F(\hat{n}) = 0$ since $\hat{n}$ is the equilibrium of the differential equation. Hence we obtain

$$dp/dt = \lambda p \tag{1.20}$$

where

$$\lambda = \frac{dF}{dn}|_{n=\hat{n}}$$

From section 1.1.1 we know that the solution to this differential equation is the exponential function. Hence we obtain that

$$p(t) = p_0 e^{\lambda t} \tag{1.21}$$

where p_0 is the value of p at time $t = 0$. Importantly, this equation is only valid in the vicinity of the equilibrium $\hat{n}$ (i.e. as long as the right hand side of 1.16 is a good approximation to $F(n)$).

Returning back to the issue of stability, eq. 1.21 shows p is decreasing towards zero if $\lambda < 0$, that is, if the derivative of F at the equilibrium $\hat{n}$ is negative. This implies that for small perturbations p the system returns back to the steady state $\hat{n}$. The equilibrium is thus called *locally stable*. If $\lambda > 0$, (i.e. if the derivative of F at the equilibrium $\hat{n}$ is positive) then p increases always. Thus the system is said to be *locally unstable*, since any small perturbation makes the population diverge from the equilibrium $\hat{n}$. This is in agreement with the graphical stability analysis of the logistic differential equation as shown in figure 1.3.

The above procedure can be applied in an analogous manner to differential equations of higher dimensionality. Consider a general multi-species model with k populations:

$$dn_j/dt = f_j(n_1, n_2, \ldots, n_k) \tag{1.22}$$

for $j = 1, \ldots, k$. Let $\hat{n}_1, \ldots, \hat{n}_k$ be an equilibrium (i.e. $f_j(\hat{n}_1, \ldots, \hat{n}_k) = 0$ for all $j = 1, \ldots, k$). As above we write

$$n_j = \hat{n}_j + p_j \tag{1.23}$$

where $p_1, \ldots, p_k$ are small perturbations from the equilibrium. Linearizing the differential equation around the equilibrium we obtain

$$dp_i/dt = \sum_{j=1}^{k} \frac{\partial f_i}{\partial n_j} p_j \tag{1.24}$$

where the symbol $\partial/\partial n_j$ denotes the partial derivatives with regard to the variables n_j. The above equation can be written in matrix notation as

$$d\mathbf{p}/dt = \mathbf{J}\mathbf{p} \tag{1.25}$$

where J is the so-called *Jacobian matrix* of the partial derivatives.

It can be shown (but I don't show here) that the solution for the small perturbation $\mathbf{p}$ is given by

$$\mathbf{p}(t) = \mathbf{c_1} e^{\lambda_1 t} + \mathbf{c_2} e^{\lambda_2 t} + \ldots + \mathbf{c_k} e^{\lambda_k t} \tag{1.26}$$

where the $\mathbf{c_i}$ are constant vectors and λ_i are the so-called *eigenvalues* of the Jacobian. The eigenvalues of a matrix can best be calculated using standard mathematical software packages. The eigenvalues of the Jacobian determine the stability characteristics of the equilibrium much like the derivative of F determines the stability in the case of the one-dimensional system. Hence, if only one of the λ_i is larger than zero, than the equilibrium is locally unstable, because for any small perturbation the system diverges from the equilibrium. Conversely if all eigenvalues are negative then the equilibrium is stable, since $\mathbf{p}(t)$ tends to zero.

To illustrate this procedure we use the two-dimensional Lotka-Volterra model (eqs. 1.9 and 1.10) discussed further above. In the notation used above we $f_1 = r_1 n_1 (1 - (n_1 + \gamma_{12} n_2)/K_1)$ and $f_2 = r_2 n_2 (1 - (n_2 + \gamma_{21} n_1)/K_2)$. The Jacobian is then given by the partial derivatives of f_1 and f_2 with regard to n_1 and n_2. Thus we have

$$J = \begin{pmatrix} r_1(K_1 - 2n_1 - \gamma_{12} n_2)/K_1 & -r_1 \gamma_{12} n_1 / K_1 \\ -r_2 \gamma_{21} n_2 / K_2 & r_2(K_2 - 2n_2 - \gamma_{21} n_1)/K_2 \end{pmatrix}_{n=\hat{n}} \tag{1.27}$$

Let us now consider an explicit numerical example with the following parameters $K_1 = K_2 = 1000, r_1 = r_2 = \gamma_{12} = \gamma_{21} = 1/2$. Thus the four equilibria of the Lotka-Volterra model are given by (i) $n_1 = n_2 = 0$, (ii) $n_1 = 1000, n_2 = 0$, (iii) $n_1 = 0, n_2 = 1000$ and (iv) $n_1 = n_2 = 2000/3$. With these parameters the Jacobians for the four equilibria are given by

$$\begin{array}{cc} \text{(i)} \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} & \text{(ii)} \begin{pmatrix} -1/2 & -1/4 \\ 0 & 1/4 \end{pmatrix} \\ \text{(iii)} \begin{pmatrix} 1/4 & 0 \\ -1/4 & -1/2 \end{pmatrix} & \text{(iv)} \begin{pmatrix} -1/3 & -1/6 \\ -1/6 & -1/3 \end{pmatrix} \end{array} \tag{1.28}$$

For a 2x2 matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{1.29}$$

there is an explicit formula for the eigenvalues given by

$$\lambda_{1,2} = \frac{1}{2}(a + d) \pm \sqrt{(a + d)^2 - 4(ad - bc)} \tag{1.30}$$

Hence we have for the following eigenvalues for the four equilibria

$$\begin{array}{cc} \text{(i) } \lambda_{1,2} = (1/2, 1/2) & \text{(ii) } \lambda_{1,2} = (-1/2, 1/4) \\ \text{(iii) } \lambda_{1,2} = (-1/2, 1/4) & \text{(iv) } \lambda_{1,2} = (-1/2, -1/6) \end{array} \tag{1.31}$$

We have pointed out further above that an equilibrium is stable if all eigenvalues are smaller than zero. Thus, for the chosen parameter values we find that only equilibrium (iv) is stable

while all other equilibria are unstable[g]. These findings are in agreement with the graphical stability analysis shown in figure 1.8.

1.2.6 Predator-prey models

Another classical model of interactions between two species are the so-called *predator-prey models*[h]. In contrast to the Lotka-Volterra model, the predator has a negative influence on the growth of the prey population, while the prey has a positive influence on the growth of the predator population. The simplest predator-prey model is given by

$$dn_1/dt = r_1 n_1 - \alpha_1 n_1 n_2 \quad (1.32)$$

$$dn_2/dt = -r_2 n_2 + \alpha_2 n_1 n_2 \quad (1.33)$$

where n_1 is the density of the prey population and n_2 is the density of the predator population. The parameters are r_1 for the birth rate of the prey, α_1 for the killing rate of the prey by the predator, r_2 for the death rate of the predator, and α_2 for the reproduction rate of the predator (per unit of prey).

Because of its simplicity this model is frequently used for didactic purposes. However, it is important to note that it is based on some unrealistic assumptions. First, in absence of the predator ($n_1 = 0$) the prey population is assumed to grow unboundedly. Second, the model assumes that the number of prey eaten by the predator is proportional to the prey density and increases without limits as the prey density increases. Clearly a lion cannot eat a hundred gazelles a day even if they were around!

The equilibria of the predator prey model are given by (i) $n_1 = n_2 = 0$ and (ii) $n_1 = r_2/\alpha_2$ and $n_2 = r_1/\alpha_1$. Note, that the equilibrium density of the predator depends only on the parameters of the prey population (and vice versa). Thus, doubling the birth rate of the prey does not affect the prey population (as one might expect intuitively), but only increases the predator population. This phenomenon is called the *paradox of enrichment*.

To study the stability of the predator-prey model, we employ the mathematical tools for stability analysis that we developed in the previous section. Using $r_1 = r_2 = 1$ and $\alpha_1 = \alpha_2 =$

[g]One of the most important and most difficult questions of ecology is whether the stability (i.e stability towards invasion of new species or towards tolerating indignities imposed by man and climate) of an ecosystem increases with its complexity (i.e. the number of species in the ecosystem). While it seems intuitively plausible that larger ecosystems may be more stable, Robert May showed in his seminal work on "Stability and complexity in model ecosystems" (Princeton University Press) that stability is not a straightforward mathematical consequence of increasing multispecies complexity. This work was largely based on studying the properties of the eigenvalues of interaction matrices with increasing number of species. A module on "stability and complexity in ecosystems" can be investigated in the in the "Modelling course in population and evolutionary biology" that takes place every summer term

[h]In this script I differentiate between Lotka-Volterra models and predator prey models on the basis of the type of interaction between the species. However, in the literature some people also refer to predator-prey models as Lotka-Volterra models

0.01, we obtain for the Jacobian (at equilbrium (ii))

$$J = \begin{pmatrix} r_1 - \alpha_1 n_2 & -\alpha_1 n_1 \\ \alpha_2 n_2 & \alpha_2 n_1 - r_2 \end{pmatrix}_{n=(r_2/\alpha_2, r_1/\alpha_1)} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{1.34}$$

Using formula 1.30 to determine the eigenvalues we find that the expression in front of the square root is zero and the expression under the square root is negative. The square root of a negative number is an imaginary number (denoted by the symbol $\imath = \sqrt{-1}$). Thus we obtain for the eigenvalues

$$\lambda_{1,2} = 0 \pm \imath \tag{1.35}$$

If all the real part of all eigenvalues is zero than we have a so-called *neutrally stable* equilibrium. Hence, the population neither converges to nor diverges from the equilibrium. If the imaginary part of the eigenvalues is not zero (as is the case here), than this implies that the system oscillates around the equilibrium. This behavior can be verified in the time plots and phase diagram of the predator-prey model (see figure 1.9).

To summarise the important results how the eigenvalues of the Jacobian determine the local stability, we reiterate our findings:

An equilibrium is locally stable only if the real part of all eigenvalues are smaller than zero. If the real part of one eigenvalue is zero, than the equilibrium is (likely) locally neutrally stable[a]. If one eigenvalue has a real part larger than zero, then the equilibrium is locally unstable. If the imaginary part of an eigenvalue is non-zero, this indicates that the solution oscillates around the equilibrium. These oscillations can be damped (if the real part of the eigenvalue is negative), neutral (if the real part of the eigenvalue is zero) or increasing (if the real part of the eigenvalue is positive).

[a] Strictly speaking, whether the equilibrium is neutral or not depends in this case on higher terms in the Taylor expansion

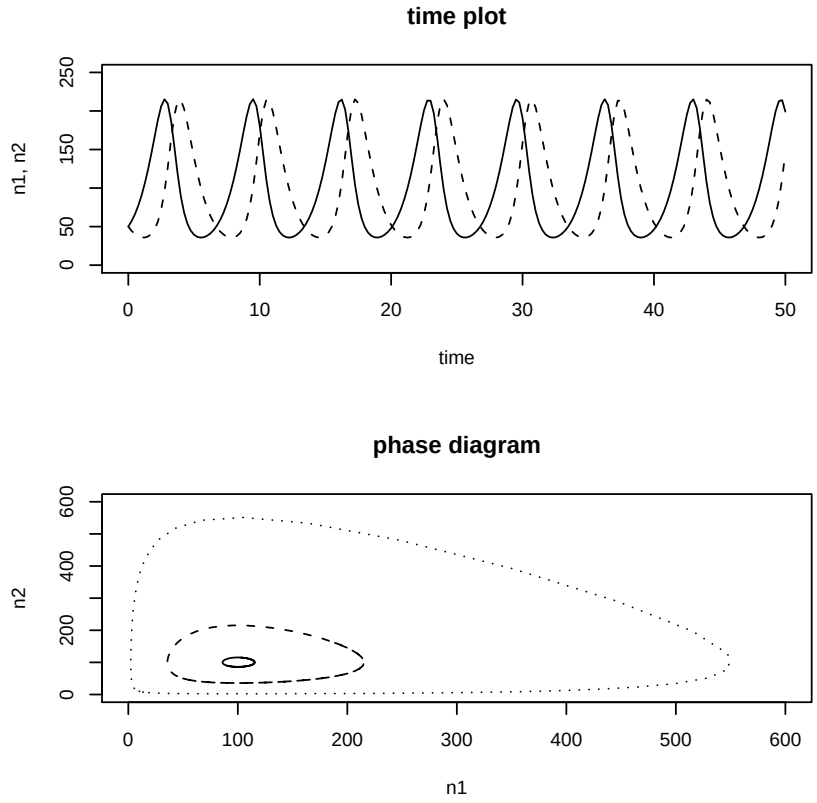

Figure 1.9: The upper plot shows the time course of the predator-prey model (eqs. 1.32 and 1.33). The solid line line depicts population 1 and the dashed line depicts population 2. The starting values of population 1 and 2 where 50 individuals. The lower plot shows the phase diagram of the model for three initial condictions: (i) $n_1 = n_2 = 90$ at $t = 0$ (solid line), (ii) $n_1 = n_2 = 50$ at $t = 0$ (dashed line), and (iii) $n_1 = n_2 = 10$ at $t = 0$ (dotted line). Both plots show neutral oscillations (i.e oscillations that neither increase nor decrease in amplitude) around the equilibrium value given by $n_1 = r_2/\alpha_2$ and $n_2 = r_1/\alpha_1$ in agreement with the mathematical stability analysis of the predator-prey model. In both plots the parameters are $r_1 = r_2 = 1$ and $\alpha_1 = \alpha_2 = 0.01$.

1.3 Difference equations

So far we have considered population dynamical models in continuous time. Models in continuous time are usually appropriate for organisms that have overlapping generations. On the other hand, many biological populations are more accurately described by non-overlapping generations (such as insect populations with one generation per year or annual plants). The dynamics of these populations often are more appropriately expressed by so-called *difference equations*. Here, the population size of the next generation $t + 1$ is expressed as a function of the population size in the current generation t.

In general a single species difference equation can be expressed as

$$n_{t+1} = C(n_t)n_t$$

In analogy to the exponential equation (see section 1.1.1) the simplest difference equation is obtained if $C(n_t)$ is a constant:

$$n_{t+1} = Rn_t \tag{1.36}$$

It is easy to see (by iteration of the above equation) that the solution is given by

$$n_t = n_0 R^t \tag{1.37}$$

where n_0 is the population size in generation $t = 0$. This equation is the analogue of the exponential function (eq. 1.4) in discrete time.

1.3.1 Logistic difference equation

Comparing equations 1.4 and 1.37, we see that $C(n_t) = R$ in the difference equation corresponds to e^r in the differential equation. By analogy we thus obtain for the logistic difference equation

$$n_{t+1} = n_t e^{r(1-n_t/K)} \tag{1.38}$$

Dividing the equation by K on both sides and substituting $x_t = n_t/K$ we obtain

$$x_{t+1} = x_t e^{r(1-x_t)} \tag{1.39}$$

Figure 1.10 shows the time development of the logistic equation for 3 values of r. For $r < 2$ the dynamics are similar to the logistic differential equation in that the population always converges towards the steady state $x_t = 1$ for large t (which implies that the population n_t always is at its carrying capacity K). However for $r > 2$ the behavior becomes more complex. While for $r = 2.2$ the solution oscillates between two values, for $r = 3.5$ the solution exhibits seemingly random behavior.

Figure 1.11 shows a *bifurcation diagram*. This plot is obtained in the following way. For each value of r a random starting value is chosen and the logistic differential equation is iterated for 500 generations (as in figure 1.10). Then the value after the 500th generation $x[500]$ is

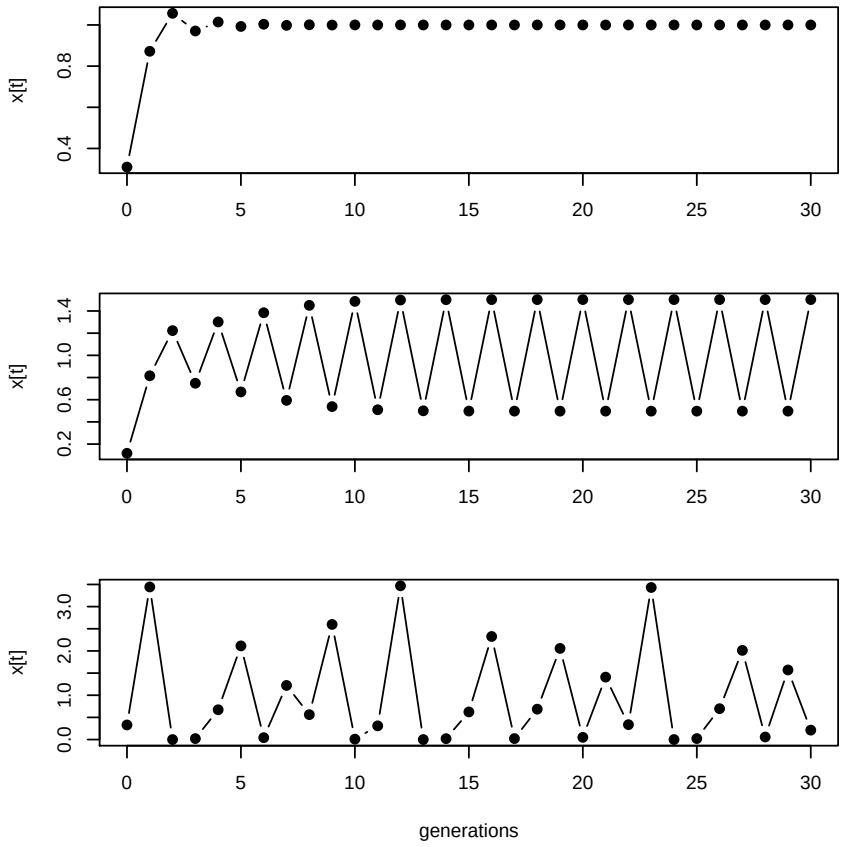

Figure 1.10: Time course of the logistic difference equation (eq. 1.39) for $r = 1.5$ (top) $r = 2.2$ (middle) and $r = 3.5$ (bottom).

plotted into the graph for the given value of r. This procedure is repeated 200 times until r is incremented by a small amount. For $r < 2$ the the logistic difference equation converges to 1. At $r = 2$ there is the first bifurcation (period doubling) as the solution begins to oscillate between two values for this value of r. As r increases further there are further period doublings. Eventually for even larger values of r the logistic differential equation shows chaotic behavior, which means that the population behavior cannot be predicted accurately for long times. Two trajectories starting from nearly identical values will diverge further and further away from each other.

The important message is that simple models can show complex behavior. Hence, if we observe a complex dynamic behavior of a natural population in the field, this does not necessarily imply that the underlying rules determining the dynamics have to be complex too[i]. Moreover, the complex behavior of the logistic difference equation is in strong contrast with the comparatively simple behavior of the logistic differential equation. As a rule of thumb difference equations show more complicated behavior than the corresponding models in continuous time[j].

[i]The discovery of the role of deterministic chaos in biology largely goes back to Robert May. Being a native of Australia, Robert May was first a theoretical physicist. Becoming the successor to Robert McArthur as a theoretical ecologist at Princeton University, Robert May moved on later to Oxford University. As one of very few scientist Robert May was made a Lord by the British Queen in 2001. Robert May has also received an honorary doctorate from the ETH Zurich.

[j]The dynamical behaviour of the Logistic Difference Equation can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

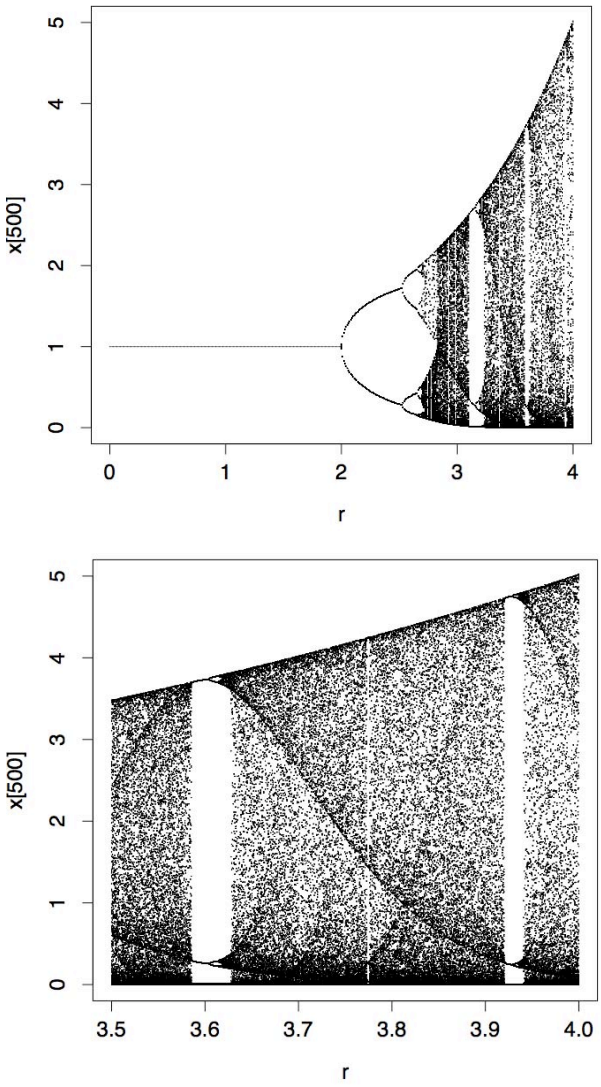

Figure 1.11: The bifurcation diagram of the logistic difference equation for $0 < r < 4$ (top) and $3.5 < r < 4$ (bottom).

1.3.2 Nicholson-Bailey Model

In the 1930s Nicholson and Bailey[k] proposed a simple discrete-time model for population dynamics of insect hosts and their parasitoids that has since become one of the classical models of population biology. Parasitoids, such as parasitic wasps, lay eggs into their hosts and thus the completion of the parasitoid life cycle requires that their hosts be killed. Parasitoids resemble parasites as they grow inside a host, but also resemble predators in that they are obligate killers of their host. The Nicholson-Bailey model is a two-dimensional system of difference equations given by

$$H_{t+1} = RH_t e^{-aP_t} \tag{1.40}$$
$$P_{t+1} = cH_t(1 - e^{-aP_t}) \tag{1.41}$$

Here H_t and P_t represent the densities of the host and parasite population at year t. R is the number of offsprings of an unparasitized hosts surviving to the next year. Assuming random encounter between hosts and parasites the probability that a hosts escapes parasitism can be approximated by e^{-aP_t}, where a is a proportionality constant. Similarly, the probability to become infected is then given by $(1 - e^{-aP_t})$. Finally, the parameter c describes the number of parasitoids that hatch from an infected host.

The equilibrium of the Nicholson-Bailey model is obtained by setting $H_{t+1} = H_t$ and $P_{t+1} = P_t$ and is given by

$$P = \frac{\log(R)}{a} \qquad \text{and} \qquad H = \frac{R}{R-1}\frac{\log(R)}{ac} \tag{1.42}$$

However, this equilibrium is unstable[l]. The simulation of the Nicholson-Bailey model (see figure 1.12) shows that that the dynamics are characterized by oscillations of increasing amplitude until the population crashes.

1.3.3 Spatial Nicholson-Bailey model

Although host-parasitoid interactions are often characterized by very strong fluctuations from year to year, unlike in the above Nicholson-Bailey model they typically do not lead to the complete extinction of either the host or the parasite. One important factor neglected in the

[k](i) Bailey, Victor Albert, born Alexandria, Egypt, 18 December 1895; died Geneva, Switzerland, 7 December 1964. Education Queen's College Oxford (B.A. 1919; D.Philosophy 1923). Demonstrator in physics at Oxford University 1924-25. Associate Professor of physics at Sydney University 1926-36; Professor of experimental physics 1936-53; Research Professor 1952-60. Worked for the WHO at the end of his career. (ii) Nicholson, Alexander John (England-Australia 1895-1969, population biologist, entomologist, and science administrator. Worked on mimicry and animal population regulation. President of the Royal Society of Australia.

[l]The stability of difference equations can be determined mathematically in a procedure similar to the mathematical stability analysis for differential equations as discussed in section 1.2.5. As for differential equations one needs to determine the eigenvalues of the Jacobian (which in discrete time is often called the "next-generation matrix"). In contrast to differential equations, where the stability depends on whether the largest eigenvalue is larger or smaller than zero, for difference equations the stability depends on whether the largest eigenvalue is larger or smaller than 1.

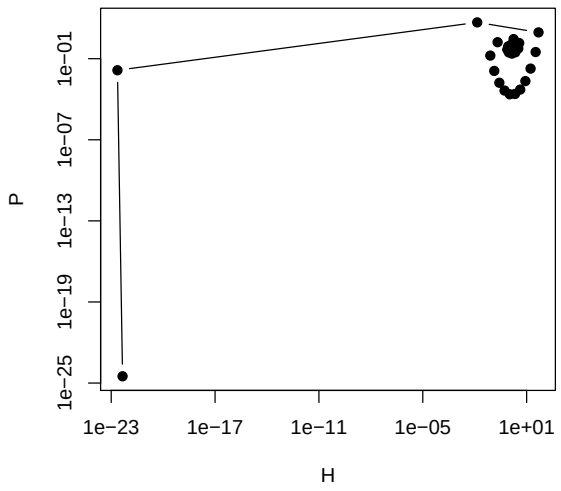

Figure 1.12: Phase diagram of the Nicholson-Bailey model. Starting from a value near the equilibrium the trajectory spirals outwards until the population crashes (i.e. reaches extremely low values, such that realistically all individuals in the population die). Thus the Nicholson-Bailey model is an example for an unstable oscillation.

Nicholson-Bailey model above is space. (Note, that all models discussed so far tacitly neglected space for the sake of mathematical simplicity). Space can be neglected if the populations can be considered to be well mixed. However, often this may often not be justified. Both the interactions between species as well as the dispersal of offspring may be local.

The above Nicholson-Bailey model can be extended to incorporate space. To this end we consider a spatial grid on which the Nicholson-Bailey dynamics take place. At each site (i, j) in the grid the dynamics are (more or less) given by the simple Nicholson-Bailey model of the previous section, but in addition we allow that hosts and parasitoids disperse to all immediately neighboring sites (with dispersal rates d_h and d_p). Mathematically, the spatial model is thus given by the following equations

$$H_{i,j}(t+1) = RH^\star_{i,j}(t)e^{-aP^\star_{i,j}(t)} \quad (1.43)$$

$$P_{i,j}(t+1) = cH^\star_{i,j}(t)(1-e^{-aP^\star_{i,j}(t)}) \quad (1.44)$$

(where the time dependence is no longer indicated by a subscript but by brackets.) Here

$$H^\star_{i,j}(t) = (1-d_h)H_{i,j} + d_h/8\sum H_{k,l}(t) \quad (1.45)$$

$$P^\star_{i,j}(t) = (1-d_p)P_{i,j} + d_p/8\sum P_{k,l}(t) \quad (1.46)$$

where the sum is over all 8 neighboring fields (i.e $(i-1, j-1), \ldots, (i+1, j+1)$).

Importantly, it can be shown (see figure 1.13) that in the spatial Nicholson-Bailey model

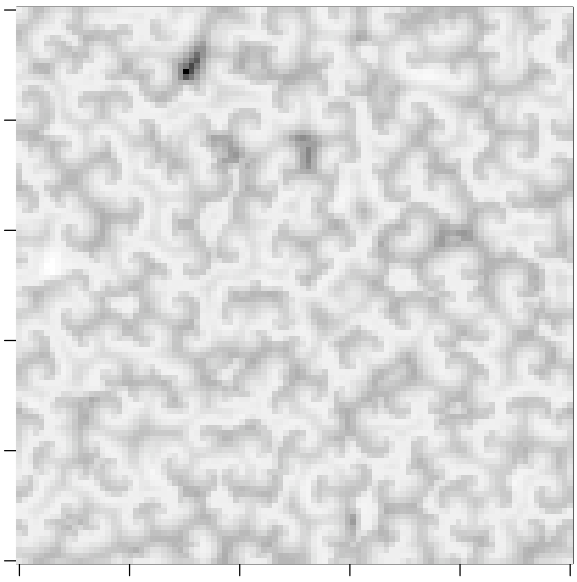

Figure 1.13: Simulation of the spatial Nicholson-Bailey model. Depending on the dispersal rates of hosts and parasites d_p and d_h parasitoids and hosts can coexist indefinitely. The plot shows the density of hosts on quadratic grid. Dark shades becoming paler represent patches with increasing host densities.

both host and parasitoids can coexist indefinitely (in contrast to the non-spatial model)[m].

[m]The dynamical behaviour of the homogeneous and spatial Nicholson-Bailey Model can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

Chapter 2

Population biology of infectious diseases

2.1 Epidemiology of infectious diseases

Infectious diseases are a major public health problem worldwide. In the developing world many childhood infections still cause substantial mortality and morbidity in children. Moreover HIV has accumulated a death toll of over 40 million people in the last twenty to thirty years. More than 90% of deaths from infectious diseases are caused by a just a handful of diseases. Table 2.1 shows the leading causes of death due to infectious diseases.

In this section we will show how the mathematical tools developed in chapter 1 can be used to study the epidemiology of infectious diseases and investigate methods for their efficient control.

2.1.1 SIR Models

In general the population dynamics of infectious diseases can often be usefully described in terms of so called *SIR models*. Here the population is subdivided into susceptible (i.e non-infected, non-immune) hosts (S), infected hosts (I), and recovered (immune) hosts (R). A graphical scheme of an SIR model is given in figure 2.1. Mathematically this scheme translates into the following equations

$$\begin{aligned} dS/dt &= B(S,I,R) - \delta S - \beta SI + qR && (2.1)\\ dI/dt &= \beta SI - aI - rI && (2.2)\\ dR/dt &= rI - \delta R - qR && (2.3) \end{aligned}$$

The model is based on the following assumptions. (i) Susceptible individuals are born at a rate $B(S,I,R)$ which is assumed to be a function of the densities of the susceptible, infected and

Disease	Death	DALYs[a]
Lower respiratory infections	4.2 $\times 10^6$	94.5 $\times 10^6$
Diarrheal diseases	2.2 $\times 10^6$	72.8 $\times 10^6$
HIV/AIDS	1.8 $\times 10^6$	58.5 $\times 10^6$
Tuberculosis	1.3 $\times 10^6$	34.2 $\times 10^6$
Malaria	0.8 $\times 10^6$	34.0 $\times 10^6$
Measles	0.16 $\times 10^6$	14.8 $\times 10^6$
Neglected Diseases[b]	2-5 $\times 10^5$	18-57 $\times 10^6$
Sexually Transmitted Diseases[c]	1.3 $\times 10^5$	10.4 $\times 10^6$
Polio	1 $\times 10^3$	34 $\times 10^3$
Other infectious diseases[d]	1.9 $\times 10^6$	69 $\times 10^6$

Table 2.1: Leading causes of death due to infectious diseases (taken from http:www.globalhealth.org/infectious_diseasesmortality_morbidity). [a]DALY = Disease adjusted life years. [b]Includes: African trypanosomiasis, Chagas disease, schistosomiasis, leishmaniasis, lymphatic filariasis, onchocerciasis, dengue, ascariasis, trichuriasis, and hookworm. [c]Excludes HIV/AIDS. [d]Includes: pertussis, diphtheria, tetanus, meningitis, hepatitis B, hepatitis C, Japanese encephalitis, maternal sepsis, and neonatal infections.

recovered hosts. (ii) Susceptibles are infected at a rate given by the product of the densities of susceptible and infected hosts, times a proportionality constant β describing the infectivity rate per contact between the two types of host. The assumption to model the infection rate proportional to SI is justified if both hosts types are well mixed and the encounter between the two host types is random. This assumption is often called *mass-action* kinetics and derives from chemical kinetics. (iii) Susceptible and recovered hosts die at a rate δ which describes the natural death rate due to causes unrelated to infection. (iv) Infected hosts die at a rate a which includes both the natural death rate plus the disease induced death rate. (iv) Infected hosts recover at a rate r. (v) Recovered hosts lose immunity at a rate q and become susceptible again.

The above model may describe qualitatively the dynamics of infectious diseases with contact dependent transmission (such as the common cold). However, it is important to note, that the above model still ignores many possibly relevant biological aspects. For example, age structure of the population is ignored. The infectivity, death rate, and rate of recovery may all depend on the age of the infected individual. Moreover, spatial structure is ignored. Often transmission to cohabiting individuals (i.e. members of the family) occurs with increased probability. All of these issues have been addressed in detail using more complex models. We ignore these higher levels of complexity and focus on the above simple model for didactic reasons.

Often the total population size remains roughly constant over the period of interest (such as the time for an epidemic to occur). In terms of the above model this implies that $N = S + I + R$ is constant and therefore $dN/dt = dS/dt + dI/dt + dR/dt = 0$. Summing up equations 2.1-2.3 we thus obtain for the birth rate:

$$B(S, I, R) = \delta(S + R) + aI \tag{2.4}$$

If the total population size is assumed to be constant, we can drop one of the three popu-

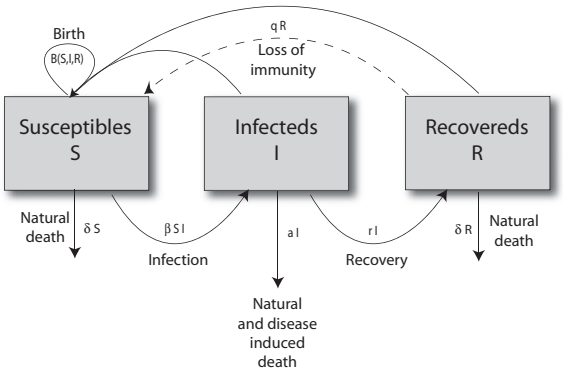

Figure 2.1: Graphical scheme of the SIR model.

lation variables, say R, since R is given by $N - S - I$. We thus obtain

$$dS/dt = (\delta + q)(N - S - I) + aI - \beta SI \quad (2.5)$$
$$dI/dt = \beta SI - aI - rI \quad (2.6)$$

The model has two equilibrium solutions

$$\text{(i) } S = N, I = 0 \quad \text{and} \quad \text{(ii) } S = \frac{a+r}{\beta}, I = \frac{(\beta N - a - r)(\delta + q)}{\beta(\delta + q + r)} \quad (2.7)$$

The first equilibrium represents the case where none of the individuals are infected (uninfected equilibrium). The second equilibrium represents the case where a fraction of the individuals are infected (infected equilibrium). Note, that there is no equilibrium with all of the individuals in the population being infected[a].

2.1.2 Basic reproductive rate

We now address under which conditions the SIR model (eqs. 2.5 and 2.6) converges to the uninfected or the infected equilibrium. To this end we need to determine whether the growth rate of the infecteds, dI/dt, is larger or smaller than 0. From equation 2.6 we derive that the population of infecteds grows if (and only if)

$$\frac{\beta S}{a + r} > 1$$

We want to determine under which conditions an epidemic can spread in the population. Thus we consider a scenario in which a (infinitesimally) small number of infected individuals is placed

[a]SIR models can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

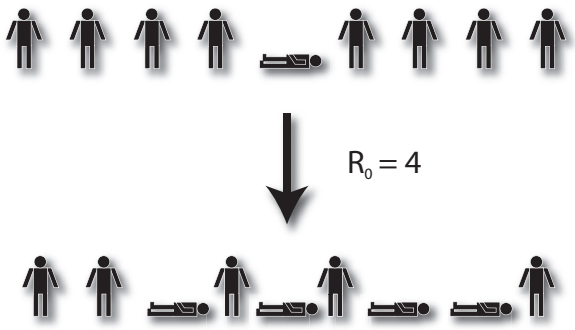

Figure 2.2: Graphical illustration of the basic reproductive rate R_0. Here a sick individual is placed into an otherwise susceptible population of 9 individuals. Over the entire duration that this individual is infected, she or he infects 4 other individuals. Thus the basic reproductive rate here is 4.

into a population in which all other individuals are susceptible. Mathematically this implies substituting $S = N$ for the population density of susceptibles (i.e. the equilibrium population density when all individuals are uninfected) in the above equation. We obtain

$$R_0 = \frac{\beta N}{a + r} > 1 \tag{2.8}$$

R_0 is the so-called *basic reproductive rate*[b]. Although most people in the field use the term basic reproductive rate, some prefer the term basic reproductive number, since strictly speaking R_0 is not a rate. (The quantities βN and $a + r$ both have the dimension time^{-1} and therefore the unit of time cancels out in the expression of R_0. R_0 is thus a number and not a rate).

Above we have determined the basic reproductive rate for the given SIR model (eqs. 2.5 and 2.6). More generally, the definition is:

> *The basic reproductive rate is the number of secondary infected hosts when one primary infected host is placed into a wholly susceptible population.*

A graphical illustration is given in figure 2.2. An infection can spread in the population if $R_0 > 1$. Conversely an infection will die out if $R_0 < 1$. Hence if $R_0 > 1$ the population will converge to the infected equilibrium and if $R_0 < 1$ the population will converge to the uninfected equilibrium.

An intuitive explanation for the basic reproductive rate as given in eq. 2.8 can be obtained as follows: N is the population size of susceptibles (when all individuals in the population are

[b]The concept of R_0 goes back to Klaus Dietz (University of Tübingen), Roy Anderson (Imperial College, London) and Robert May (University of Oxford)

susceptible). The rate at which individuals leave the infected compartment is given by $a + r$, describing the combined loss of infecteds through death and recovery. Hence the inverse of $a+r$ is the average duration of an infection. Thus over the entire duration of the infection an infected individual will infect $\beta N/(a + r)$ individuals.

We will now derive that for childhood infections with life-long immunity, the basic reproductive rate is approximately given by

$$R_0 = \frac{L}{A} \tag{2.9}$$

where L is the life expectancy of individuals in the population and A is the average age of infection. Life-long immunity implies that $q = 0$ in the SIR model (eqs. 2.5 and 2.6). We now do the following elementary calculations

$$\begin{aligned} R_0 &= \frac{\beta N}{a+r} \\ &= \frac{\beta(S+I+R)}{a+r} \\ &= \frac{a+r+\beta I+\beta R}{a+r} \\ &= 1+\frac{\beta I(1+r/\delta)}{a+r} \\ &= 1+\frac{\beta I}{\delta}\frac{\delta+r}{a+r} \end{aligned} \tag{2.10}$$

Note, that we are using here the population sizes at the infected equilibrium, since we want to express R_0 as a function of these values. In particular we have used in this calculation that at the infected equilibrium $S = (a + r)/\beta$ and $R = (r/\delta)I$ if $q = 0$.

Let us first discuss the expression $(\delta+r)/(a+r)$. Clearly, we have $\delta < a$ since the parameter a contains both the natural and the disease induced mortality. Hence $(\delta + r)/(a + r) < 1$. Moreover, for childhood infections it is reasonable to assume that the rate of recovery r is much larger than the natural death rate δ, but also than the mortality rate of infecteds a. Therefore $(\delta + r)/(a + r) \approx 1$. Thus we obtain

$$R_0 \approx 1 + \frac{\beta I}{\delta} \tag{2.11}$$

Note, that δ is the (average) natural death rate of uninfecteds. Hence the reciprocal $1/\delta$ is the average life expectancy L of uninfecteds. On the other hand βI is the rate at which susceptibles become infected. Hence its reciprocal $1/(\beta I)$ is the average age at which susceptibles first get infected. Altogether we obtain

$$R_0 \approx 1 + \frac{L}{A} \approx \frac{L}{A} \qquad \text{if } L >> A \tag{2.12}$$

It is important to spell out the assumptions underlying this calculation. As mentioned already above this result was derived assuming life long immunity (i.e $q = 0$) which is a reasonable assumption for childhood diseases, but not for other types of infections[c]. More importantly, the

[c]It is straightforward to repeat the above calculation assuming that $q > 0$. If $q >> \delta$ then essentially $R_0 \approx L/(Aq)$.

above calculation assumes that all parameters such as infectivity, recovery, natural death rate and disease associated death rate do not depend on the age of the individual, which is clearly a biologically unjustified assumption. The SIR model (eqs. 2.5 and 2.6) upon which our calculation is based ignores age structure in the population. Basically, it assumes what ecologists refer to as Type II mortality, namely that the death rate is constant over age. In the developed world the death rate of humans is fairly low until age 60 or higher and then begins to increase markedly with age. In the developing world child mortality is high, followed by a moderate mortality in intermediate ages (which is actually currently changing because of AIDS), and a high mortality at old age. Thus a more realistic mortality (at least for the developed world) is the so called Type I mortality, which assumes a zero mortality rate until age L and an infinitely high mortality afterwards. Figure 2.3 illustrates these two types of mortality functions.

Age structure can be incorporated into population dynamical models by various techniques. The Leslie matrix approach subdivides the population into discrete classes with regard to age. Partial differential equations can be used to describe the population dynamics with continuous age distributions. Importantly, it can be shown with these approaches that the basic reproductive rate R_0 is also given by L/A for the more realistic Type I mortality.

For some infections it is also necessary to incorporate that the probability of infection depends on age. For example, the risk of infection by sexual transmitted diseases is high at intermediate ages, but vanishes at young age. Moreover, the model (eqs. 2.5 and 2.6) assumes that uninfected and infected individuals always give birth to uninfected offspring, which is only justified for horizontally transmitted infections[d].

Table 2.2 lists the average age of infection for various childhood diseases in different countries. The relation $R_0 = L/A$ makes clear that childhood diseases must have high basic reproductive rates, because of the early average age of infection. The data in table 2.2 also makes clear that the basic reproductive rate may be different in different localities because of differences in the age of first infection. (Note that the expected life-span will also differ between countries.)

[d]Horizontal transmission refers to transmission between individuals independent of whether they are related or not. In contrast, vertical transmission refers to transmission from mother to child.

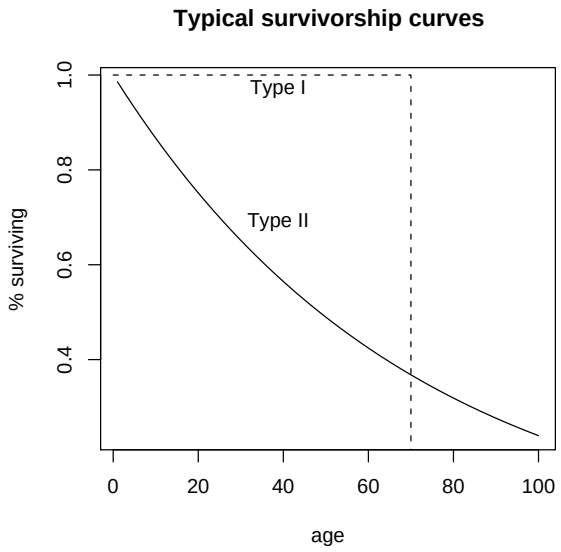

Figure 2.3: Type I (dashed line) and Type II (solid line) mortality. Type II mortality assumes a constant death rate independent of age, while Type I mortality assume a vanishing death rate until age L and infinitely high death rate afterwards. Type II mortality is a better approximation for the survivorship of human populations in the developed world. Type III mortality (not shown here) assumes a high mortality at young ages that decreases markedly with increasing age. Type III mortality may for example describe survivorship in fish populations where most individuals never make it to adulthood.

Infectious disease	A	Location and time period
Measles	5-6	USA, 1955-1958
	4-5	England/Wales, 1948-1968
	2-3	Morocco, 1963
	2-3	Ghana, 1960-1968
	1-2	Senegal, 1964
Rubella	9-10	Sweden, 1965
	9-10	USA 1966-1968
	6-7	Poland, 1970-1977
	2-3	Gambia, 1976
Chicken Pox	6-8	USA, 1921-1928
Polio	12-17	USA 1955
Pertussis	4-5	England/Wales, 1948-1968
Mumps	6-7	England/Wales, 1975-1977

Table 2.2: Average age at infection, A, for various childhood diseases in different geographical localities and time periods (Source, Anderson & May, Infectious Diseases of Humans, Oxford University Press)

2.1.3 Vaccination

Next we want to determine the effect of vaccination. In particular, we want to ask what fraction of the population needs to be vaccinated in order to eradicate the disease. In the previous section we derived that $R_0 = \beta N/(a+r)$. Vaccination has the effect of transferring individuals directly from the susceptible to the immune class. So we ask to what level $S^\star$ do we have to reduce the susceptible population such that the disease can no longer spread through the population. The threshold condition is given by

$$R_0 = 1 = \frac{\beta S^\star}{a+r} \tag{2.13}$$

Solving for $S^\star$ we obtain

$$S^\star = \frac{a+r}{\beta} \tag{2.14}$$

which is exactly the population of susceptible individuals at the infected equilibrium. Thus the critical proportion of individuals that need to be vaccinated is given by

$$p_c = \frac{N - S^\star}{N} = 1 - \frac{a+r}{\beta N} = 1 - \frac{1}{R_0} \tag{2.15}$$

Hence, the higher the basic reproductive rate, the higher is the fraction of individuals that need to be vaccinated in order to eradicate the disease. Note, however, that in no case it is necessary to vaccinate the entire population. Once the population of susceptibles falls below a critical level (given by $p_c N$) the entire population is protected against infection. This concept is called *herd immunity*. Importantly, this type of immunity is not a property of the individual, but a property of the population only. Table 2.3 gives the critical proportion for several important infectious diseases.

Infectious disease	p_c
Malaria (*P. falciparum* in hyperendemic regions)	99%
Measles	90-95 %
Pertussis	90-95 %
Human parvovirus infections	90-95 %
Chicken pox	85-90 %
Mumps	85-90 %
Rubella	82-87 %
Polio	82-87 %
Diptheria	82-87 %
Scarlet fever	82-87 %
Small pox	70-80 %

Table 2.3: Critical proportion p_c of individuals that need to be vaccinated in order to eradicate a disease through herd immunity. (Source: Anderson & May, Infectious Diseases of Humans, Oxford University Press)

These considerations have a number of important implications:

(i) First of all the data in table 2.3 shows why it is so hard to eradicate diseases such as malaria or certain childhood diseases. The vaccination needs to cover a very large fraction of the population. Moreover, the above calculation assumes that a vaccine offers 100% protection, which is not the case for many vaccines.

(ii) There is a conflict between the interests of the individual and the interests of the community. Because vaccines can have side-effects[e] parents frequently decide not to have their children vaccinated, essentially because they estimate that the risk of infection by a childhood disease (and its complications) may be lower than the risk of side-effects through vaccination. However, the risk of infection is only low because so many children are vaccinated. Hence, from the perspective of the individual it may be rational not to vaccinate one's children. However, on the basis of the population it is clear that it is necessary to maintain a high coverage of vaccination, since childhood infection can often lead to very severe or even fatal outcomes (see the 600,000 deaths by measles in 2002 in table 2.1).

(iii) Vaccination essentially has the effect of lowering the basic reproductive rate by decreasing the number of susceptibles in the population. However, since the basic reproductive rate is given by the ratio of the life expectancy, L, and the average age of infection, A, lowering the basic reproductive rate has the effect of increasing the average age of infection. Thus while the protection though vaccines leads to a lower number of infections, an unwanted side effect of vaccination can be that some childhood infections have more severe effects when they occur in older individuals. Rubella, for example, can lead to severe complications during pregnancy. Measles can lead to male sterility in adults. It is often suggested that the history of poliomyelitis in developed countries represents such a case of unwanted side-effects of vaccination. When contracted at an early age polio is less likely to have serious consequences. However, increased standards of hygiene led to a drop of poliovirus infections (which are transmitted through the oral-fecal route of infection), which in turn increased the average age of infection. However, in older individuals the infection is more likely to lead to paralysis and other complications and therefore the epidemic became more troublesome as the incidence declined. These considerations caution that population biological effects of vaccination policies need to be carefully investigated for new vaccination policies.

2.1.4 Stochastic effects

It is important to note that so far the models that we have considered assume that all processes are deterministic. This assumption is often justified when populations are large. However, when an epidemic starts growing, the population size of infecteds is initially small and is thus affected by stochastic events.

To illustrate the importance of stochastic effects, consider a simple *branching process* in which each infected individual infects n other individuals with a probability $P(n)$ (see figure 2.4). Thus instead of assuming that (at the beginning of an infection) an infected individual spreads the infection to a fixed number of other individuals (given by R_0), we assume the

[e]Most side-effects of vaccines against childhood infections are harmless (such as a transient fever). Severe complications are extremely rare.

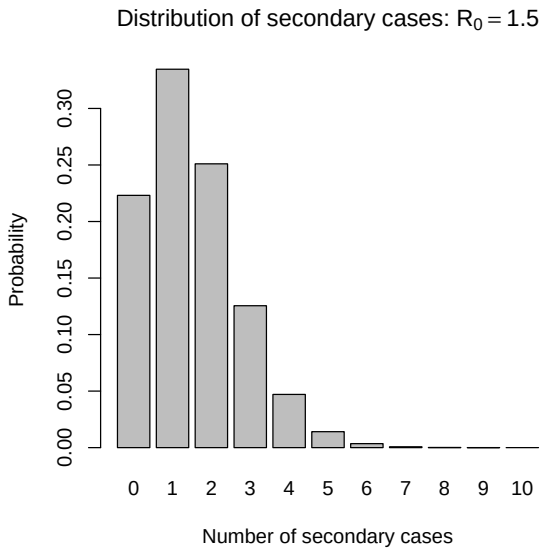

Figure 2.4: Probability distribution that one infected individual will spread the infection to n other individuals (at the beginning of the epidemic, when most individuals are uninfected). The probability distribution shown here is a Poisson distribution with mean $R_0 = 1.5$. A Poisson distribution for the number of infected individuals is obtained if we assume a well-mixed population in which there is a small constant probability of transmission per contact between susceptible and infected individual.

transmission by an individual is characterized by a probability distribution with mean R_0. In figure 2.5 we show a computer simulation of this process. This simulation shows that even with $R_0 = 1.5$ a large number of epidemics actually die out. The probability that an epidemic goes extinct decreases rapidly with the initial number of infecteds (see figure 2.6). This demonstrates how important it is to react early to control an epidemic, when the number of infecteds is still small[f].

[f]Stochastic epidemic models can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

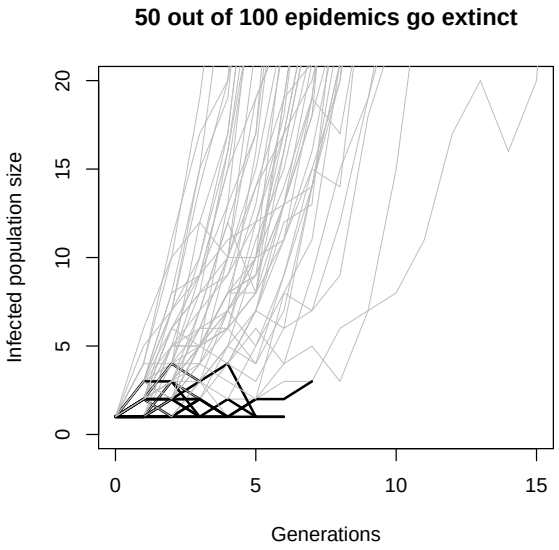

Figure 2.5: Simulation of the branching process with the probability distribution given in figure 2.4. Starting from one infected individual approximately 50% of the epidemics go extinct despite the fact that the basic reproductive rate is larger than one ($R_0 = 1.5$).

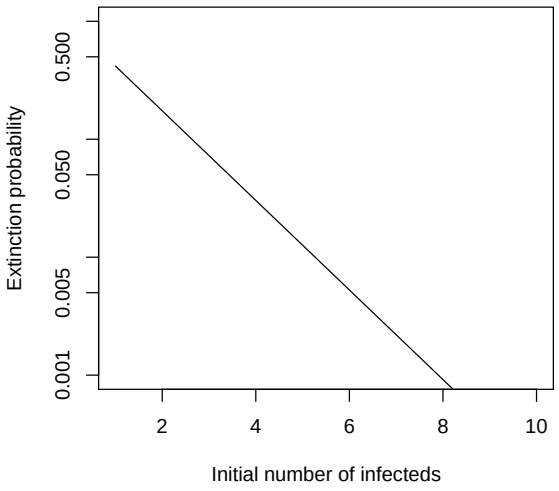

Figure 2.6: Dependence of the probability of extinction on the initial number of infecteds. These results are obtained for the probability distribution shown in figure 2.4. The extinction probability decreases exponentially with increasing initial number of infecteds. Hence, once an epidemic has grown to 20 infected individuals, then the probability of extinction due to stochastic effects is neglible. Note also, that since the epidemic grows fast, epidemics that eventually die out do so after only few generations. In the graph here, all epidemics that died out, did so after less than 10 generations.

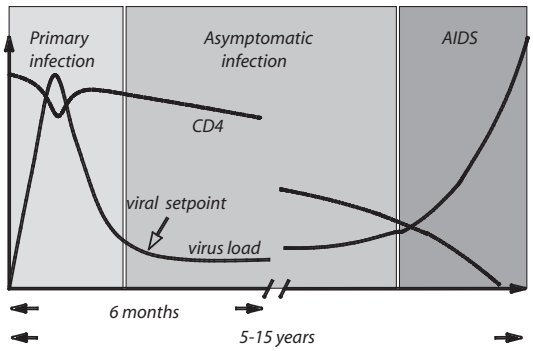

Figure 2.7: The typical course of HIV-1 infection in an untreated individual. After the individual becomes infected the virus load rapidly rises to very high levels. This transient viremia is typically accompanied by flu like symptoms and fever. Then the virus load declines over several orders of magnitude until it settles at a level that remains more or less stable for several years (the so-called *viral set-point*). The initial phase of characterized by the rapid increase and decrease of the virus load is called *primary infection* and typically last only a few months. During primary infection there is typically also a significant (although often transient) decrease in the the CD4 cell population, which are the major target cells of HIV. These target cells fulfill an important role in orchestrating the patient's immune response. After primary infection the patient enters the *asymptomatic phase*, because there are no obvious symptoms that accompany the disease at this stage. During this stage the virus load remains stable over prolonged periods of time, while the number of CD4 cells continue to decline steadily. Eventually the CD4 cells fall below a critical level (<200 CD4 per μl of blood) which defines the onset of AIDS and below which the CD4 cells can no longer maintain a functional immune response. At this stage the AIDS-defining opportunistic infections typically arise and the virus load increases sharply.

2.2 Intrahost dynamics of infectious diseases

The application of population biological models to the epidemiology of infectious disease has a long tradition that can be traced back to the beginning of last century. More recently, the methods have also been applied to study the population dynamical processes of the replication of infectious pathogens within individual hosts, in order to develop a better understanding of the pathogenesis of infectious diseases and the development of resistance to chemotherapeutic agents. This approach has been adopted for a variety of microbial infections. Here I will focus on HIV infection.

Figure 2.7 gives a schematic illustration of the course of an HIV infection in an untreated individual. Population biological models have been used to describe the initial phase of the infection, to address the decline of virus load during primary infection, and to develop an understanding of why the duration of the infection is so long and variable. Here we will focus on another important aspect of population biological models, namely their application to estimate crucial kinetic parameters that are not accessible through direct experimentation.

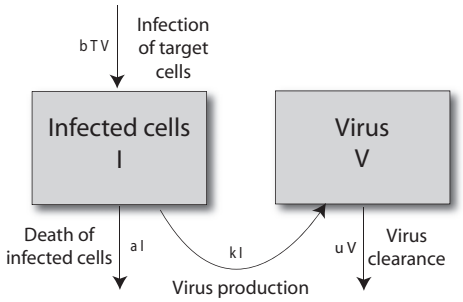

Figure 2.8: Schematic illustration of the virus dynamics model given by eqs 2.16 and 2.17.

As can be seen from figure 2.7 both the virus load and the target cell population change only slowly during the asymptomatic phase. Thus over shorter time spans of weeks to a few months, one can safely assume that the population is in a so-called *quasi steady state*. The dynamics of the virus population is thus given by

$$dI/dt = bTV - aI \tag{2.16}$$
$$dV/dt = kI - uV \tag{2.17}$$

where I is the population density of infected target cells and V is the population density of free virus. T represents the population density of susceptible target cells, which is assumed constant over the time span of interest. The parameters are b for the infectivity rate of virus (per contact between virus and susceptible target cell), a for the death rate of infected cells, k for the rate of virus production by infected cells, and u for the clearance rate of infectious viruses (for example by the immune system). A schematic illustration for this model is given in figure 2.8.

As the virus load is approximately constant during the asymptomatic phase, the death of infected cells must be approximately balanced by the infection of target cells by virus or mathematically $bTV \approx aI$. Now we consider what happens when a patient is treated with antiretroviral drugs. We model this by setting $b = 0$ as these drug potently block viral infectivity. The first differential equation (eq. 2.16) is then given by $dI/dt = -aI$. We know from section 1.1.1 that the solution to this differential equation is the exponential function. Thus we have

$$I(t) = I_0 e^{-at} \tag{2.18}$$

where I_0 is the density of infected cells at the start of therapy. The differential equation for the virus load (eq. 2.17) is then given by

$$dV/dt = kI_0 e^{-at} - uV \tag{2.19}$$

We can obtain a solution for the explicit time dependence of the virus load by integration

$$V(t) = V_0 \frac{ae^{-ut} - ue^{-at}}{a - u} \tag{2.20}$$

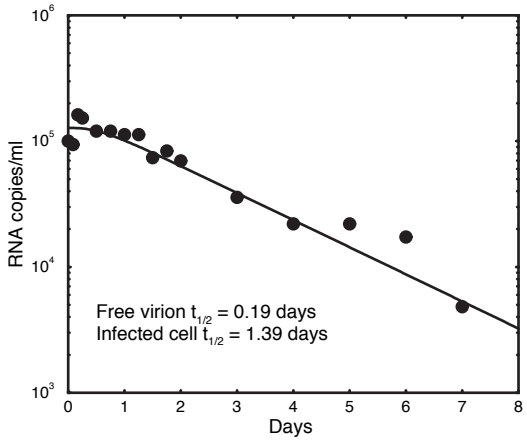

Figure 2.9: The decline of virus load in a patient treated with a antiretroviral drugs. The virus load is measured as copies of viral RNA per ml of plasma. Note that the virus load decreases more than an order of magnitude in only seven days of treatment. (Data taken from Perelson et al, Science, 1996)

where $V_0 = kI/u$ (assuming that free virus and infected cells are in equilibrium prior to treatment).

Thus after a transient phase the decline rate of virus is dominated by the exponential function with the smaller exponent (in absolute value) [g].

Virus load in HIV infected patients is commonly measured as the number of RNA copies per ml of plasma (the cell free fraction of the blood). This measure reflects the abundance of free virus (here denoted by V) and not of the load of infected cells (here denoted by I). Comparison of clinical data with the explicit solution for the virus load as a function of the time after treatment can thus be used to estimate two important parameters, namely a the death rate of infected cells and u the clearance rate of free virus. Figure 2.9 shows the optimal fit of the virus load function (eq. 2.20 to data of a treated patient. The estimates for the parameters a and u can then be used to calculate the so-called *half-life* of free virus and infected cells. The half-life describes the time that an exponential decaying population requires to fall to half its initial value. The relation between the half-life and the (exponential) rate is thus given by $t_{1/2} = ln(2)/a$ for the infected cell population and $t_{1/2} = ln(2)/u$ for the free virus population.

Let us first discuss the biological implications of this finding, before turning to the more technical point how these parameters were estimated. In the mid 1990's significant progress was made in terms of the development of highly potent inhibitors of retroviral replication. Clinical trials with these drugs indicated that the virus load declined rapidly over several orders of magnitude within very short time frames (as for shown in figure 2.9). Fitting of simple mathematical models as the one discussed above then allowed kinetic parameters of viral replication

[g]The time taken until the first term is n-fold smaller then the second term in this equation is given by $ln(na/u)/(u-a)$.

in patients to be estimated that could not have been obtained by direct experimentation. It turned out that the half-life of infected cells in essentially all patients studied was around 1-2 days. This implies that 50% of the infected cell population dies in only 1-2 days, which in turn implies that an equal number of infected cells must have been produced in the same time interval before therapy was started. This finding was in strong contrast to the then prevailing view in HIV research that the asymptomatic phase was characterised by viral latency (i.e. the virus was assumed to remain quiescent during this phase and it was believed that little turnover occurs during this phase of the infection). HIV researchers were mislead by the fact that the virus load is low during the asymptomatic phase. However, with the benefit of hindsight this was almost a trivial mistake, since an equilibrium can be low even when there is a high level of production of infected cells provided that there is also high levels of death of infected cells. The discovery that HIV is replicating fast throughout the entire course of the infection has had important consequences for our current view of HIV pathogenesis[h].

Let us now turn to the technical question how parameters can be estimated by fitting a model to data. Assume we have a set of n measurements $m_1, \ldots, m_n$ at times $t_1, \ldots t_n$. Assume further we have a model M which we want to fit to the data. In general the model will depend on a number of parameters and the time. Then the estimation of the parameters is done by minimizing the following function

$$\mathcal{M} = \sum_{i=1}^{n} (m_i - M(t_i))^2 \tag{2.21}$$

This means that we are changing the model parameters to minimize the sum of the squared distances between the measured data and the value predicted by the model. The best fit parameters are given by that combination of parameters for which $\mathcal{M}$ is minimal. If the model is linear the best fit parameters can be obtained analytically by performing a *linear regression*. If the model depends nonlinearly than it is more difficult to find the best fit parameters, and it can no longer be guaranteed that the best fit parameters are actually obtained. However, standard statistical software usually has sophisticated routines for fitting of nonlinear models.

The fact that it is the square of the distances and not the absolute value which are minimised, has to do with the assumption that the errors are normally distributed. If estimates for the variance of the measured data are available then individual points can be weighted by the inverse of the variance (i.e. $\mathcal{M} = \sum (m_i - M(t_i))^2/\sigma_i$, where σ_i is the variance of the measurement i).

2.3 Intrahost dynamics of resistance

Over the last 10 years combination therapy with different classes of antiretroviral inhibitors has greatly reduced AIDS mortality. However, the emergence of drug resistant HIV-1 strains jeopardizes successful long term treatment with antiretroviral drugs. Thus, to optimize the use

[h]The within host dynamics of HIV and the estimation of kinetic parameters from clinical data can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

of currently available and future drugs, a detailed understanding of the emergence HIV-1 drug resistance is of utmost importance for providing optimal care to HIV-infected individuals.

Population biological models have also played an important role in developing a better understanding of the emergence of resistance and the development of strategies to prevent or delay treatment failure due to resistance. Again, we will consider here only the simplest models to highlight some central issues. To address these issues we need to modify the above model in several ways. First, the finding that the half-life of free virus is around 6 hours and that of infected cells is around 1-2 days (see previous section and figure 2.9), implies that the dynamics of the free virus are fast in comparison to the infected cells[i]. Thus we can assume that at all times the free virus is to a good approximation proportional the infected cells (i.e. $V = kI/u$). Second, we incorporate explicit dynamics for the population of target cells. The following model is often referred to as the basic model of virus dynamics[j].

$$dT/dt = \lambda - \delta T - \beta IT \tag{2.22}$$

$$dI/dt = \beta TI - aI \tag{2.23}$$

Here λ represents the rate at which target cells are produced from a pool of progenitor cells, δ is the natural death rate of uninfected target cells and $\beta = bk/u$ is the cell infectivity rate (i.e. the infectivity of free virus times the ratio of free virus to infected cells). In analogy to the basic reproductive rate for an epidemic, we can calculate the basic reproductive rate for virus replication. The growth rate of the infected cell population is positive if $\beta T/a > 1$ (see eq.2.23. To determine the basic reproductive rate, we need to substitute the density of susceptible target cells when $I = 0$ for T. From equation 2.22 we derive that the equilibrium of susceptible cells in the absence of virus is given by $T = \lambda/\delta$. Thus we obtain for the basic reproductive rate

$$R_0 = \frac{\lambda\beta}{a\delta} \tag{2.24}$$

If $R_0 > 1$ the virus infection can spread in an individual and the system goes to the equilibrium

$$T = \frac{a}{\beta} \quad \text{and} \quad I = \frac{\lambda}{a} - \frac{\delta}{\beta} = \frac{\lambda}{a}(1 - \frac{1}{R_0}) \tag{2.25}$$

As we have pointed out above, the effect of drug therapy is to reduce the parameter β and thus reduce R_0. Figure 2.10 shows the dependence of the equilibrium load of infected cells (see eq.2.25) as a function of the basic reproductive rate R_0. The figure makes clear that the equilibrium virus load is only expected to decline significantly if the treatment is strong enough to reduce the basic reproductive rate close to 1, which also implies that the virus is close to eradication[k]. The main effect of a reduction of the parameter β is an increase of the equilibrium level of the uninfected cells, but not a decrease of the infected cells. This is reminiscent of the paradox of enrichment observed in predator-prey models.

[i] More recent estimates suggest that the half-life of free virus is even shorter (i.e. $t_{1/2} = 1 - 2$ hours).

[j] See for example Nowak and May, Virus dynamics, Oxford University Press

[k] Note, that we are considering here the equilibrium virus load. A treatment-induced reduction of the infectivity parameter β reduces the infected cell load temporarily, but may not have a strong effect on the equilibrium infected cell load unless it reduces R_0 to a value close to 1.

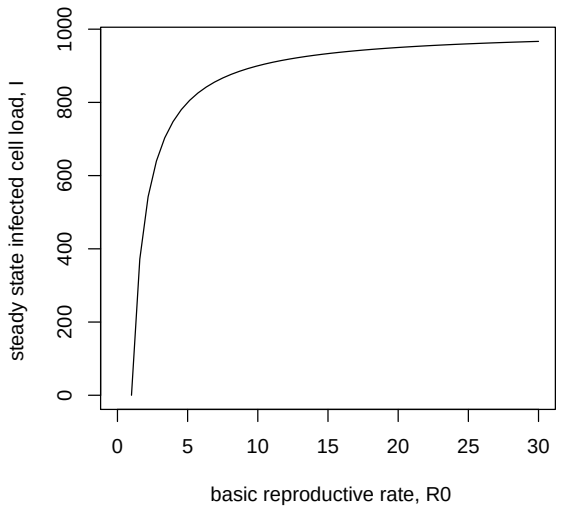

Figure 2.10: The equilibrium load of infected cells (eq. 2.25) of the basic model of virus dynamics as a function of the basic reproductive rate R_0. The figure illustrates that a significant change in the infected cell load is only expected if drug therapy reduces the basic reproductive rate to values close to one, and thus close to eradication.

Let us now subdivide the infected cells into two populations, one infected with drug sensitive wildtype virus, I_s, and one infected with drug resistant mutant virus, I_r. Our model is now given by

$$\begin{aligned} dT/dt &= \lambda - \delta T - ((1-\epsilon)b_s I_s + b_r I_r)T & (2.26)\\ dI_s/dt &= (1-\epsilon)(1-\mu)b_s I_s T - aI_s & (2.27)\\ dI_r/dt &= b_r T I_r - aI_r + (1-\epsilon)\mu b_s I_s T & (2.28) \end{aligned}$$

Here b_s and b_r are the infectivities of sensitive and resistant virus, respectively, ϵ is the efficacy of the drug in inhibiting the replication of sensitive virus and μ is the mutation rate at which sensitive wildtype virus mutates into resistant mutant virus. (Note, that we ignore the back mutation rate from resistant to wildtype virus here, which is justified as long as the sensitive virus is much more abundant than the resistant mutant).

In the absence of drugs (i.e. $\epsilon = 0$) the equilibrium fraction of cells infected with resistant virus can be calculated as

$$f = \frac{\mu}{1 - b_r/b_s} = \frac{\mu}{s} \qquad (2.29)$$

where s is the so-called *selection coefficient* of the resistant strain (i.e. $1-s$ is the relative fitness of resistant to the sensitive strain in absence of drugs, $\epsilon = 0$). Thus, the frequency of resistant virus in an untreated patient increases with increasing mutation rate and decreasing selection coefficient.

Typically the selection coefficients for drug resistance mutations in HIV-1 (in the absence

of drugs) are fairly small ($s = 0.01 - 0.1$). Given that the population size of infected cells in an untreated HIV-1 infected patient during the asymptomatic phase is of order $10^7 - 10^8$ and the mutation rate (per nucleotide and replication cycle) is around 3.5×10^{-5}, we expect that all mutants that differ only by a single point mutation are present in an untreated individual. Since single point mutations (at the relevant loci) often confer a considerable degree of resistance (against treatment with a single drug), we conclude that many drug resistant variants exist in a patient even prior to drug therapy. Very roughly the probability that a mutant with n mutations is present prior to therapy scales with μ^n. Therefore, HIV infected patients are treated with combinations of 3 or more drugs in order to decrease the likelihood of treatment failure due to pre-existing resistance.

Finally, we will address how the rate at which resistant virus increases during therapy depends on the efficacy of inhibition, ϵ. For mathematical simplicity, we neglect the production of resistant virus by mutation (i.e. $\mu = 0$). We rewrite the above equation in terms of the total load of infected cells, $I = I_s + I_r$ and the fraction of resistant virus $\rho = I_r/I$. The derivatives of these two new variables are given by

$$dI/dt = dI_s/dt + dI_r/dt \tag{2.30}$$

$$d\rho/dt = \frac{dI_r/dt}{I} - \frac{I_r}{I}\frac{dI/dt}{I} = \frac{dI_r/dt}{I_r} - \rho\frac{dI/dt}{I} \tag{2.31}$$

Thus we have

$$dT/dt = \lambda - \delta T - ((1-\epsilon)b_s(1-\rho) + b_r\rho)TI \tag{2.32}$$

$$dI/dt = ((1-\epsilon)b_s(1-\rho) + b_r\rho)TI - aI \tag{2.33}$$

$$d\rho/dt = b_rT\rho - a\rho - \rho[(1-\epsilon)b_s(1-\rho) + b_r\rho)T - a] \tag{2.34}$$

$$= (b_r - (1-\epsilon)b_s)T\rho(1-\rho) \tag{2.35}$$

We have encountered eq. 2.35 before. For constant T, this equation is identical to the logistic differential equation (see eq. 1.5) with $r = (b_r - (1-\epsilon)b_s)T$ and $K = 1$. To approximate the behavior we thus assume that T is indeed constant. In particular we assume that the susceptible cell density is given by the equilibrium value of susceptible cells during treatment $T = a/b_r$. This is the value that the susceptible cell density will eventually attain (provided treatment is strong enough to reduce the infectivity of the sensitive virus below that of the resistant virus, i.e. $(1-\epsilon)b_s < b_r$)[1]. Using $T = a/b_r$ and substituting into eq. 1.8, which in the context here describes the time until 50% of the cells are infected with resistant virus, we obtain

$$t^\star \approx -\frac{1}{a(1-(1-\epsilon)b_s/b_r)}\log(\rho_0) = -\frac{1}{as'}\log(\rho_0) \tag{2.36}$$

where s' is the selection coefficient of the sensitive relative to the resistant virus during therapy and ρ_0 is the frequency of resistant virus at the start of therapy. Thus, the time until the population is dominated by resistant virus depends more strongly on the selection coefficient

[1]Prior to the start of therapy the density of susceptible cells is approximately given by $T = a/b_s$. Thus by assuming $T = a/b_r$ we transiently overestimate the density of susceptible cells in the initial phase of treatment. Consequently we overestimate $t^\star$.

than on the initial frequency, since the time depends inversely on s' but logarithmically on ρ_0. To give a numerical example, a two-fold difference in the selection coefficient doubles (or halves) the time to resistance. To have a similar effect on the time to resistance the initial concentration of resistant virus needs to be changed approximately by a factor 10 (since $\log(10) = 2.3$)). The above equation also shows that the stronger the efficacy of the drug (i.e. the larger ϵ), the faster is the rise of resistance. This implies that stronger drugs may select more rapidly for resistance, provided resistant mutants are present already at the start of therapy. This result is supported by the finding that highly potent antiretroviral inhibitors are often associated with a more rapid rise of resistance[m].

[m]The emergence of drug resistance in HIV can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

Chapter 3

Evolution of parasite virulence

3.1 Why should parasites be harmful?

Infectious diseases vary greatly in the effect on their hosts. While some infectious pathogens (such as Ebola virus, HIV-1, avian influenza, and many others) induce very high host mortality rates, many other infectious pathogens cause mild diseases or even go completely unnoticed. Moreover, some pathogens can cause death a few days after infection, whereas others only cause death after years of infection. The pathogenic effects, however, do not only differ greatly between different pathogen species, but can also vary substantially between different strains of a single pathogen species. Since there is likely a heritable component to variation in parasite virulence, it is a challenge to evolutionary biologists to establish the factors determine the evolution of parasite virulence. In this chapter, we will employ population biological methods as a useful tool to address how different factors may affect the evolution of parasite virulence. Moreover, we will discuss some key experimental findings.

Until recently, medical doctors and biologists where brought up on the notion that in the long term parasites[a] should evolve to be harmless to their hosts. This so called *conventional view* is based on the notion, that parasites typically require a living host for their transmission. Since the death of the host usually implies the end of a parasite's opportunities for transmission, it was argued that, evolutionarily speaking, it is in the parasite's interest to keep its host alive. Harmful parasites were thus presumed to be recent parasites that have not yet evolved to avirulence.

Evolutionary biologists have questioned the conventional view both on theoretical and on experimental grounds. First, there are examples of diseases that are known to have a long association with their hosts, but have remained virulent until today. It is believed, for example, that measles has become endemic in the human populations around 10,000 years, when the widespread use of agriculture lead to increased population densities and larger communities[b].

[a] We use the term parasite here in a broad sense to include viruses, bacteria, protists and macroparasites.

[b] Comment: We have learned in the previous chapter that the basic reproductive rate increases with the (susceptible) population density. It is believed that measles was unable to persist in the human population prior to the population increase associated with the agricultural revolution.

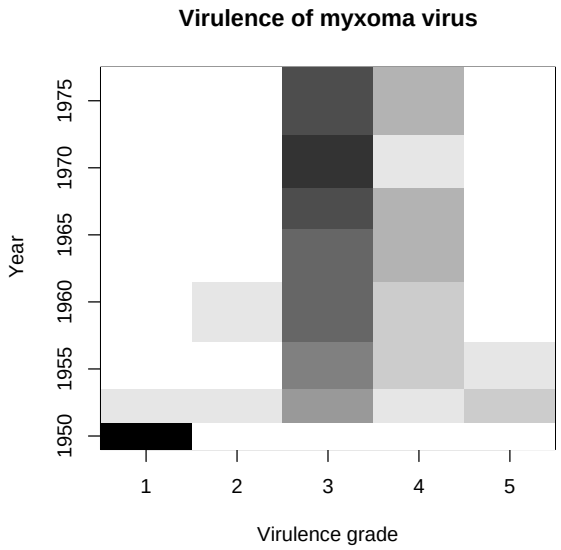

Figure 3.1: Evolution of virulence of myxoma virus after its introduction into the population of wild rabbits in Australia. Virulence grade 1 is the highest and grade 5 is the lowest. The proportion of individuals in a given virulence grade are indicated by increasing grey levels. Thus virulence decreased rapidly over the first few years, but then remained stably at an intermediate level after 1964. Thus contrary to the conventional view, the virus did not evolve to avirulence. (Source: Fenner, Proc. Roy . Soc., 1983)

Measles is transmitted exclusively from humans to humans and despite 10,000 years of coevolution with its host, measles has remained highly virulent. Another striking example is that fig wasps together with their parasitic nematodes have been found in million year old amber. This fig wasp-nematode association is known to be highly virulent today. Hence, either virulence has not declined over a million years of evolution, or alternatively virulence must have increased over time. Both types of explanation are in contrast to the conventional view.

One of the most widely known experimental studies of virulence was done in the 1950's by the introduction of a highly virulent strain of myxoma virus (a pox virus) into the Australian rabbit population as a means to control the population explosion of rabbits (see figure 3.1). In the first years after the introduction the virulence of myxoma virus declined rapidly and eventually settled at an intermediate level rather than evolving towards zero. Thus this study provides strong experimental evidence against the conventional view[c].

Before progressing further it is necessary to provide a concise definition of the term virulence. In evolutionary biology the virulence of a parasite is defined as the fitness costs to the hosts that are induced by the parasite[d]. Parasites can affect host fitness in various ways such as increasing

[c]The decrease in virulence could also be explained by an increase in resistance in the rabbit population. But also this interpretation of the data argues against the conventional view.

[d]Note that in other fields such as microbiology and plant pathology the term virulence is used to refer to the ability of a parasite to infect a certain tissue or a certain host.

mortality, increasing morbidity, reducing host reproductivity[e] or by modulating host behavior[f]. In what follows we will use the term virulence to denote the parasite induced host mortality unless stated otherwise.

3.2 Maximization of the basic reproductive rate

Consider two strains of a parasite competing for transmission in host population as described by the following model

$$\begin{aligned} dS/dt &= \lambda - \delta S - (b_1 I_1 + b_2 I_2) S + q_1 R_1 + q_2 R_2 && (3.1)\\ dI_1/dt &= b_1 S I_1 - (\delta + v_1 + r_1) I_1 && (3.2)\\ dI_2/dt &= b_2 S I_2 - (\delta + v_2 + r_2) I_2 && (3.3)\\ dR_1/dt &= r_1 I - \delta R_1 - q_1 R_1 && (3.4)\\ dR_2/dt &= r_2 I - \delta R_2 - q_2 R_2 && (3.5) \end{aligned}$$

This is a straightforward extension of the SIR model (eqs. 2.1 - 2.3) discussed in the previous chapter. Here, I_1 and I_2 denote the densities of hosts infected with parasite strain 1 and 2, respectively. We assume that susceptible hosts, S, immigrate at a constant rate λ and die with a natural death rate δ as this represents the simplest dynamical equation that leads to a stable density of susceptible hosts ($S = \lambda/\delta$) in absence of infection (i.e. $I_1 = I_2 = 0$). This assumption is made for the sake of mathematical simplicity, but has no consequence for the results derived further below. The parameters b_1 and b_2 describe the infectivities of strain 1 and 2, respectively. In the SIR model (eqs. 2.1 - 2.3) we used the parameter a to describe the combined mortality of infected hosts arising from natural causes and disease-associated causes of death. Here, we explicitly account for the virulences v_1 and v_2, that describe the mortality rates associated with infection by parasite strain 1 or 2.

In the previous chapter we derived the basic reproductive rate in the context of the SIR model. Substituting $N = S = \lambda/\delta$ for the density of susceptible individuals in the absence of infection into eq. 2.8. We thus obtain for the basic reproductive rates of the two parasite strains

$$R_0^{(1)} = \frac{\lambda b_1}{\delta(\delta + v_1 + r_1)} \qquad \text{and} \qquad R_0^{(2)} = \frac{\lambda b_2}{\delta(\delta + v_2 + r_2)}$$

Let us assume first that only parasite strain 1 circulates in the population. Provided $R_1^0 > 1$ the equilibrium density of susceptible cells is the given by

$$S^\star = \frac{\delta + v_1 + r_1}{b_1}$$

[e] Jukka Jokela, for example, works on trematode parasites that infect the fresh water snail *Potamopyrgus antipodarum*. These parasites consume the reproductive tissue of their hosts and therefore sterilize the host.

[f] Helminths often manipulate the behavior of their intermediate hosts to make them more susceptible to predation and thereby to increase their own transmission to the definitive host.

(see eq. 2.7). Consider now that a small amount of hosts infected with parasite strain 2 are placed into the population (in which parasite strain 1 is already at equilibrium). The population of hosts infected with strain 2 can increase if

$$dI_2/dt > 0 \qquad \Rightarrow \qquad \frac{b_2}{\delta + v_2 + r_2} > \frac{1}{S^\star} \qquad \Rightarrow \qquad R_0^{(2)} > R_0^{(1)}$$

Hence, strain 2 can invade into the host population infected by strain 1, if the basic reproductive rate of strain 2 is larger than that of strain 1. It is obvious that the converse is also true, i.e. that strain 1 can only invade strain 2, if its basic reproductive rate is larger than that of strain 2. Taken together, we can thus say that the parasite with the larger basic reproductive rate can invade and replace the other strain. The model does not allow for coexistence of both strains in equilibrium. More generally, we have thus derived that the parasite with maximal basic reproductive rate outcompetes all competitors. Hence, natural selection is expected to maximise the basic reproductive rate.

3.3 Trade-offs between infectivity and virulence

If there are no constraints that couple virulence to infectivity rate or recovery rate, we expect that natural selection to lead to increasing rates of infectivity (b) and decreasing virulence (v) and decreasing recovery rate (v) since this simultaneously maximises the duration of the infection and the transmission per contact with a susceptible hosts. Thus this simple model shows that we expect evolution towards avirulence as postulated by the conventional view, if there are no constraints between the parameters. In general terms, however, it is likely that these parameters are intrinsically coupled. For example, both the infectivity rate per contact and the virulence may increase with increasing parasite load. Similarly, slow recovery may require escape from the immune responses and thus may be related to increasing virulence.

To illustrate this situation let us consider different functional forms that may represent possible trade-offs between the infectivity parameter b and the virulence parameter v. In particular, we consider three cases:

(i) The infectivity rate increases linearly with virulence (i.e. $b = \alpha v$, where α is a proportionality constant).

(ii) The infectivity rate increases faster than linearly with virulence (i.e. $b = \alpha v^2$ for example).

(iii) The infectivity rate increases slower than linearly with virulence (i.e. $b = \alpha\sqrt{v}$ for example).

These trade-offs between virulence and infectivity are shown in figure 3.2. Substituting these trade-offs into the basic reproductive rate we obtain

$$\text{(i) } R_0 = \frac{\lambda\alpha v}{\delta(\delta + r + v)} \qquad \text{(ii) } R_0 = \frac{\lambda\alpha v^2}{\delta(\delta + r + v)} \qquad \text{(iii) } R_0 = \frac{\lambda\alpha\sqrt{v}}{\delta(\delta + r + v)}$$

For case (i) and (ii) R_0 is maximal if for maximal virulence (i.e. $v = \infty$). The optimal level of virulence for case (iii) can be obtained by differentiation:

$$dR_0/dv = 0 \qquad \Rightarrow \qquad \frac{\alpha\lambda}{2\delta}\frac{\delta + r - v}{\sqrt{v}(\delta + r + v)^2} = 0$$

Hence dR_0/dv is zero, if the enumerator of the second fraction equals zero. We thus obtain for the optimal virulence

$$v_{opt} = \delta + r$$

In summary, the analysis of the simple two-strain SIR model suggests that natural selection should lead to ever increasing levels of virulence, if the infectivity rate increases linearly or faster with increasing virulence. If, however, the infectivity rate increases slower than linearly, then natural selection should lead to an intermediate level of virulence. If there are no constraints between infectivity rate and virulence, then natural selection should lead to avirulence. Perhaps, the biologically most plausible scenario is that there are diminishing returns in terms of infectivity rate for higher and higher levels of virulence (i.e. case (iii)). For many parasites both the harm done to the host and the probability of transmission increase with increasing parasite load[g]. At very high parasite loads, however, the harm to the hosts continues to increase, while the probability of transmission is expected to saturate. For this case, the above model thus predicts the evolution towards intermediate levels of virulence (as was for example observed for myxoma virus, see figure 3.1).

3.4 Key assumptions of the virulence model

The discussion of the SIR model in the previous section provides a formal basis to investigate how different factors may affect the evolution of virulence. In particular, the model helped to delineate the conditions under which we expect natural selection to lead to avirulence. Before moving ahead it is important to consider the assumptions that underlie the analysis above:

1. Infection is assumed to be proportional to the product of S and I. As discussed previously, this term assumes that the population is well-mixed (i.e. that the population has no spatial structure). We will discuss the effects of spatial structure in section 3.9.

2. The model describes horizontal, contact-dependent transmission of the pathogen, because the infection proportional to the product of S and I. Many diseases, however, are not directly transmitted via contact but instead have water-borne (e.g. Cholera), food-borne (e.g. Hepatitis A), or vector-borne (e.g. malaria) transmission. We will discuss below how the transmission mode may affect the evolution of virulence. In particular, we will discuss in section 3.7 how vertical (i.e. parent to offspring) versus horizontal transmission affects the evolution of virulence.

[g] A correlation between virulence and transmission has been confirmed for a variety of pathogens, but may not be common to all parasites. An example for a correlation between transmission and virulence for a bacteriophage is shown in figure 3.5.

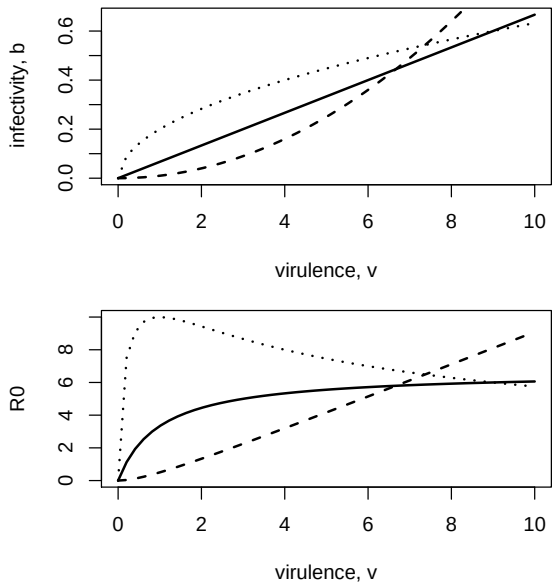

Figure 3.2: Examples for three tradeoffs between virulence and infectivity are shown in the upper graph. The solid line assumes that the infectivity rate is proportional to the virulence (i.e. $b \propto v$). The dashed line assumes that infectivity increases more than linearly with virulence (i.e. $b \propto v^2$). The dotted line assumes that the infectivity increases slower than linearly with increasing virulence (i.e. $b \propto \sqrt{v}$). The basic reproductive rates corresponding to these three cases are shown in the bottom graph. Hence, if the infectivity rate increases less than linearly with increasing virulence, then the basic reproductive rate has a maximum at intermediate values of virulence. Otherwise, the basic reproductive rate increases with increasing virulence.

3. We have considered only trade-offs between the rate of infectivity, b, and virulence, v. It is also possible that there are trade-offs between the rate of recovery, r, and virulence (or all three parameters b, r and v may be depend on each other).

4. The model assumes that all infected hosts are infected by a single parasite strain only. Hence, infection provides immunity against superinfection by other strains of the same parasite (or coinfection by other parasites). In the model there is no competition between distinct strains within an individual hosts. We will discuss the consequences of intra-host competition in section 3.8 below. In particular we will show, that under these condition it is no longer the basic reproductive rate that determines the success in competition.

5. To derive the result that natural selection maximizes the basic reproductive rate we used that the host population is infected by strain 1 only (and is in the corresponding infected equilibrium) and then studied the conditions under which strain 2 can invade. The important assumption underlying the derivation of this result is that the density of susceptibles is determined by the presence of the parasite. While this is certainly the case in the simple SIR model, this may not generally be the case. Moreover, the density of susceptible hosts may not be in equilibrium, but could for example increase. Hence, if there is no feedback of the parasite on population density, then there is strictly speaking no competition between different parasite strains. To study the evolution of virulence under non-equilibrium conditions is beyond the scope of this script, but it is important to add a cautionary note that the evolution of virulence may depend on whether the total host population is in equilibrium[h].

3.5 Host density and the evolution of virulence

It is frequently stated that virulence should increase with increasing host density, since increased host density reduces the costs of virulence. The intuition for this claim is that the increased opportunities of transmission relieve the constraints to keep the host alive for long times. We will now investigate this claim first on the basis of the results from the models. In our model (eqs. 3.1 - 3.5) the host density can increase either because of an increased birth (or immigration) rate λ or because of a decrease in the natural mortality rate δ. Assuming a diminishing returns trade-off between infectivity rate and virulence (i.e. case (iii) in section 3.3) we have derived that the optimal virulence is given by $v_{opt} = \delta + r$ (see eq. 3.3). This implies, that the optimal level of virulence depends on δ but not on λ. Hence, the model suggests that increasing the birth rate has no effect on the optimal level of virulence, while decreasing the death rate δ is expected to lead to a decrease in virulence. This is the opposite of what was claimed further above on intuitive grounds, and highlights the use of formal models to delineate the factors affecting the evolution of virulence. On second thought, the result of the model does also make intuitive sense. The shorter the natural life span of a host, the faster the parasite needs to exploit the host for its own transmission. Thus in short living hosts parasites should evolve to become more virulent.

[h]The effect of non-equilibrium conditions is discussed for example in Lenski & May, J. Theor. Biol., 1994 or Bonhoeffer et al., Proc. Roy. Soc. 1996

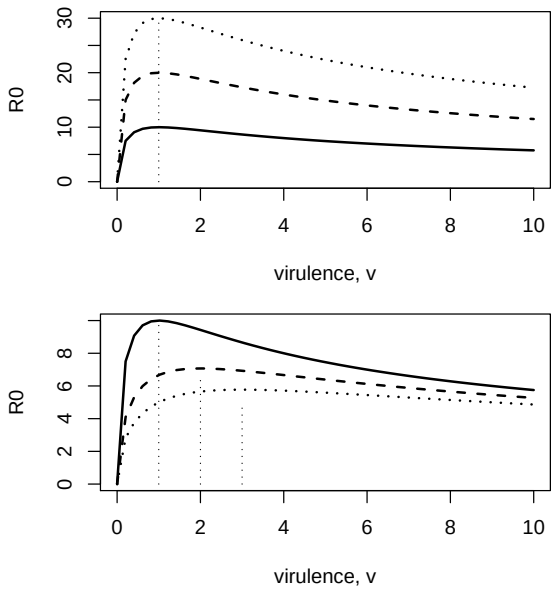

Figure 3.3: The top plot shows the basic reproductive rate for a diminishing returns trade-off for three increasing values of the birth rate b (in order solid line, dashed line and dotted line). The bottom plot shows the basic reproductive rate for three increasing values of the natural mortality rate δ (again in order solid line, dashed line, dotted line). Hence, the optimal level of virulence (indicated by the thin dotted line) is not affected by changes in the birth rate, but increases with increasing death rate.

The problem with the argument at the outset of the previous paragraph is that while increasing host density does allow more virulent pathogens to exist, it does not necessarily affect the evolutionary optimal level of virulence. This can be seen as follows. The basic reproductive rate is proportional to the birth rate λ (see eq. 3.2). Thus increasing the birth rate changes the magnitude of R_0, but does not shift its maximum with regard to virulence (see fig. 3.3). Note, however, that parasites that previously were not capable of spreading in the population (because of a basic reproductive rate smaller than one) may be capable of spreading after an increase in b, because of their increased basic reproductive rate. In other words, an increased population density may allow parasites to spread that were previously too virulent. However, the increased population density (if due to an increase in λ) does not have an effect on the evolutionarily optimal level of virulence that the parasite should eventually attain.

The effect of an increasing extrinsic mortality rate on the evolution of virulence has been tested with water fleas (*Daphnia magna*) infected by microsporidian gut parasites (*Glugoides intestinalis*) by Dieter Ebert and Katrina Mangin[i]. The water fleas were grown together with the parasites in small beakers under conditions of low and high host mortality. Since from the point of view of the parasite it does not matter whether hosts are killed or simply removed from the system, host mortality can be easily manipulated by replacing regularly a certain fraction of the hosts in the beakers. Contrary, to the prediction of the simple SIR model, they found that

[i]Source: Ebert & Mangin, Evolution, 1997.

parasites that were kept under conditions of high host mortality did not evolve high virulence. The discrepancy between the model and the experiment turned out to be most likely due to the fact that multiplicity of infection rather than host mortality determined virulence in this system. It turned out that the average duration of infection was significantly longer in hosts under low mortality conditions, presumably leading to increased probability of multiple infection. We will discuss the consequences of multiplicity of infection for the evolution of virulence further below.

3.6 Treatment and the evolution of virulence

Infected hosts can recover from infection naturally or due to treatment. Both cases are reflected in the recovery rate parameter r in the model 3.1-3.5. Assuming again a diminishing returns trade-off with between virulence and infectivity rate the optimal level of virulence (case (iii) in section 3.3) depends on r (see eq. 3.3). Thus the model suggest that increasing the rate of recovery by medical intervention may lead to the unwanted side effect of an increased optimal level of virulence. The intuition is the same as for increased host mortality, since from the perspective of the parasite transmission opportunities end irrespective of whether the patient dies or recovers. One needs to emphasise, however, that the main effect of treatment is to reduce the duration of an infection and thereby decrease the prevalence of the disease. Thus treatment will in most cases be beneficial in terms of the overall reduction of the burden of disease in a population, even if it is conceivable that medical interventions could also lead to an increased level of virulence[j]. In this context, it is interesting to note, that intensive antibiotic therapy as is common for example in intensive care units in hospitals does indeed select for highly virulent and highly resistant bacteria (although it needs to be pointed out the the evolution of virulence and the evolution of antibiotic resistance are not equivalent processes).

In summary, the model suggests that treatment could lead to increased levels of virulence. While this is indeed a conceivable side effect of therapy, it is necessary to point out that many other factors (some of which we will discuss below) may also affect virulence. Hence, before you convince your friends to throw away their medications, continue to read. Generally, the current scientific understanding of the factors that may influence the evolution of virulence is too poor to make reliable medical recommendations.

3.7 Horizontal versus vertical transmission

One of the most straightforward predictions regarding the evolution of virulence is that parasites that transmit only vertically from parent to offspring should evolve to become avirulent, since in a loose analogy exclusively vertically transmitted, virulent parasites can be regarded as a dominant deleterious gene and should thus be eliminated by natural selection. Mathematical models have confirmed this basic prediction, but have also predicted more counterintuitive outcomes provided

[j]This is analogous to the situation discussed in the section on vaccination (section 2.1.3) where it was argued that vaccination against certain childhood infections can have the unwanted side-effect of increasing the average severity of a disease, but its overall effect is to reduce the burden of disease.

there is a trade-off for efficiency of vertical transmission and virulence[k]. However, instead of focussing on the mathematical intricacies of the effect of vertical versus horizontal transmission we will discuss here two experiments that addressed the evolution of virulence in the context of varying degrees of horizontal and vertical transmission.

The first experimental test of the hypothesis that increasing levels of vertical transmission correlate with decreasing levels of virulence was based on a field study of fig-pollinating wasps and their nematode parasites that was carried out in Panama by Allen Herre[l]. The life cycle of fig-wasps begins when one or more females enter the so-called syconium, which is the enclosed inflorescence that eventually ripens to become the fig fruit. Inside the syconium, the pollen bearing fig wasps pollinate the flowers, lay eggs and die. As the fig fruit ripens the newly hatched fig wasps mature and mate within the fig, before they leave and the cycle begins anew. The bodies of the dead mother wasps remain in the fig and can thus be counted.

Fig wasps are generally highly specific to distinct species of fig trees. Each fig wasp species in turn is associated with a specific species of nematodes. The nematode's life cycle is thus intimately connected with its host species. The nematodes are carried with their host species into the fig. The nematodes reside in the body cavity of the fig wasps, where they consume their hosts' tissue. Once the nematodes have matured inside the fig wasps around 6-7 adult nematodes emerge from the body of a dead wasp, mate and lay eggs. The nematodes of the next generation then crawl onto to the newly hatching fig wasps that mature inside the ripening fruit.

The opportunities for horizontal transmission of nematodes depends on the number of fig wasps that enter a syconium. If only a single fig wasps enters the syconium, then transmission is purely vertical. With increasing number of fig wasps inside a fig the opportunities for horizontal transmission rise.

Allen Herre collected field data for 11 different species of fig wasps that are each specific for a single fig tree species and have all a specific nematode parasite. He scored the proportion of broods founded by a single wasp against the virulence of the nematode (measured as the life time reproductive success of nematode infected wasps relative to uninfected wasps). In agreement with the hypothesis he found that the virulence increases with decreasing proportion of single foundress broods (see figure 3.4). However, it should be kept in mind that there are also alternative explanations for these observations. In particular, the increasing level of virulence with decreasing proportion of single foundress broods could also be due to higher multiplicity of infection and thus due to increased intra-host competition (see section 3.8).

Another test of the hypothesis that increasing levels of vertical transmission should be associated with lower virulence has been performed by Sharon Messenger and coworkers[m] using *E. coli* as the host and the filamentous bacteriophage f1 as the parasite. The bacteriophage was engineered to carry a gene that confers resistance to an antibiotic. Thus bacteria infected with a phage are resistant to that antibiotic. Filamentous phages (in contrast to most other known phages) establish a permanent infection, during which they constantly reproduce without killing

[k] See: Lipsitch et al, Evolution, 1996

[l] Source: E.A. Herre, Science 1993

[m] Source: S.L. Messenger, I.J. Molineux, and J.J. Bull, Proc. Roy.Soc, 1999

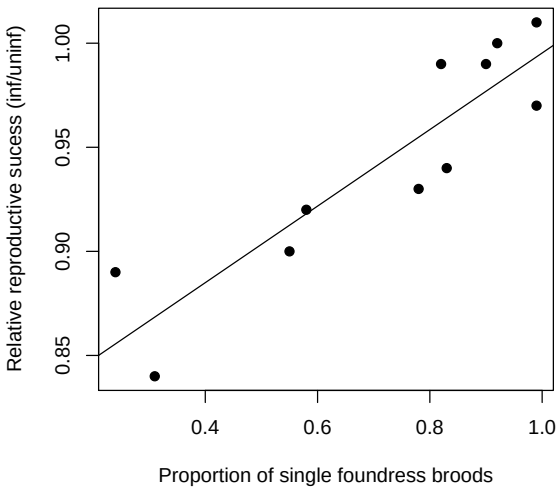

Figure 3.4: Virulence as a function of the proportion of single foundress broods in 11 species of fig wasps each infected with a host species specific nematode. If broods are founded by a single wasp, then nematodes are transmitted purely vertically. With decreasing proportion of single foundress broods the reproductive success of infected relative to uninfected wasps decreases. Hence, virulence increases with increasing opportunities for horizontal transmission. The linear regression is highly significant ($R^2 = 0.81, p = 0.00016$).

the host.

Infection of a bacterium via horizontal transmission requires the presence of F-pili on their host cell (i.e. requires cells carrying the conjugal F-plasmid). Upon entering a cell the single-stranded phage DNA is converted by host enzymes into a double stranded replicative form. The pool of replicative forms increases rapidly to 30-50 copies in the first hour of infection and settles at an equilibrium of 5-15 copies per cell 2 hours after infection. Vertical transmission occurs when an infected bacterium divides. The replicative forms are presumably distributed randomly over the dividing cells. Due to the high copy number of replicative forms vertical transmission in this system is close to perfect (i.e. more than 99% of cells remain infected after a cell division).

In this system the degree of horizontal and vertical transmission can be controlled experimentally. How this works in detail will become clear further below. In essence, to ensure exclusively horizontal transmission free phage is separated from the cells and used to infect a new population of uninfected bacteria. To ensure exclusive vertical transmission bacteria and phage are cultured in the presence of the antibiotic. Under these conditions all cells that are not infected are killed by the antibiotic.

In the experimental protocol the bacteria where grown for 24 days. Each day the bacteria were grown for 24 h in the presence of the antibiotic and then diluted 1000-fold and placed into fresh growth medium (containing antibiotics). The experiment design consists of two treatment

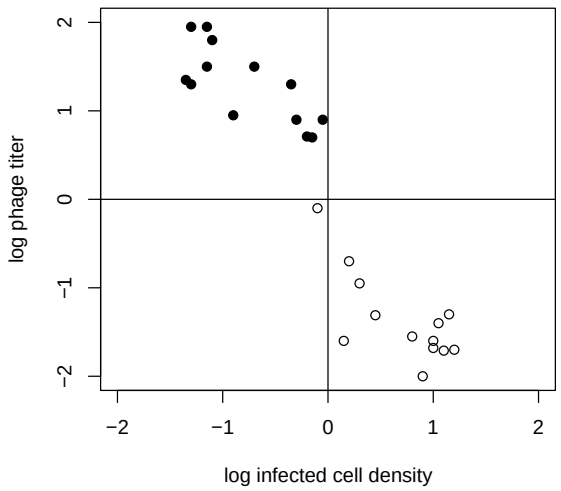

Figure 3.5: Log phage titer versus log infected cell density in L_1 lines (filled circles) and L_8 lines (open circles). For a detailed description of the experiment see main text. Log phage titer is used here as a proxy for fecundity. Virulence increases with decreasing infected cell density after 24 hours of growth. The figure provides evidence for a correlation between fecundity (i.e. horizontal transmissibility) and virulence. Moreover, the virulence of the L_8 lines is consistently lower than that of the L_1 lines, demonstrating that virulence decreases with decreasing opportunity of horizontal transmission.

arms that differ in the opportunities for horizontal transmission. The free phages were either harvested daily (L_1 lines) or every eighths day (L_8 lines) and were transfered to a new uninfected population of bacteria. Thus there were 24 steps of horizontal transmission in the L_1 lines and only three in the L_8 lines. After 24 days the virulence of the bacteriophages in both treatment arms was measured using the logarithm of the bacterial density as a marker for virulence. (Note, that increasing log cell density corresponds to increasing avirulence.) Figure 3.5 shows that the L_8 lines had consistently lower virulence (i.e. higher log infected cell densities) than the L_1 lines. Moreover the L_1 lines had consistently higher fecundity as measured by the log phage titer produced during one hour of growth in fresh medium. The figure also provides evidence for a correlation between fecundity (i.e. horizontal transmissibility) and virulence.

A relevant feature of the experimental system is that superinfection cannot occur, since a consequence of phage gene expression is that the F-pili are permanently disassembled on the host cell. Hence, shortly after infection the phage actively prevents superinfection. Moreover, given the relatively low copy number of replicative forms per cell and the high fidelity of virus replication (which is performed by the replication machinery of *E. coli*, there is presumably limited intra-host diversity. Thus in this experimetal setup it can be ruled out that the increased virulence that is associated with higher horizontal transmission is caused by increased intra-host competition (see section 3.8).

Both experimental tests lend support to the hypothesis that natural selection should act to decrease virulence the more a parasite's transmission is vertical as opposed to horizontal (although an alternative interpretation is possible for the fig wasp example). However, purely vertically transmitted parasites can also be virulent if they possess the ability to increase their representation in the next generation sufficiently to offset the effects of the selection against infected hosts. *Wolbachia*, for example, are such an exception to the general rule. These intracellular bacteria can induce feminization in their insect hosts. Since they are exclusively transmitted from mothers to their female offspring, the advantage gained by a distortion of the sex ratio can compensate for the cost in terms of virulence.

3.8 Intra-host competition

So far we considered that competition between distinct parasite strains occurs exclusively at the level of transmission between hosts. In particular, the simple SIR model (eqs. 3.1 - 3.5) did not allow for the superinfection of an already infected host by a second parasite. Many parasites, however, face strong competition within an individual infected hosts. Intra-host competition arises whenever there is phenotypic variation between parasites in the same host. Such phenotypic variation can arise due to several processes:

Mutation: Many parasites (in particular viruses) have high mutation rates resulting in a high phenotypic diversity within an individual infected host

Superinfection: Infected individuals may become re-infected by a related pathogen thus generating intra-host diversity.

Coinfection: Host may be infected by several unrelated parasites simultaneously, which can reside in the same or in different tissues. Unrelated parasites residing in different tissues can nevertheless be in competition, since ultimately both depend on the survival of the host.

Intrahost competition is expected to be strong if

(i) there is high intra-host diversity (as can be generated by mutation, superinfection, or coinfection).

(ii) different strains differ markedly in their fitness within the host. This can be for example the case if some mutants can escape from the control by the host's immune responses.

(iii) there are many rounds of replication between the time of infection and the time of transmission to the next host. This is the case if the turnover rate of the pathogen is high and there is a long interval between infection and transmission.

Within-host competition can lead to increased virulence. This has been documented for many pathogens in so called *serial transfer experiments*, in which pathogens are transferred

experimentally between infected hosts, thus releasing the selection for efficient natural transmission. When the host population is genetically sufficiently homogeneous such serial transfer experiments typically result in the selection of more virulent pathogens (see fig. 3.6). Hence, there is a correlation between virulence and intra-host competitive ability.

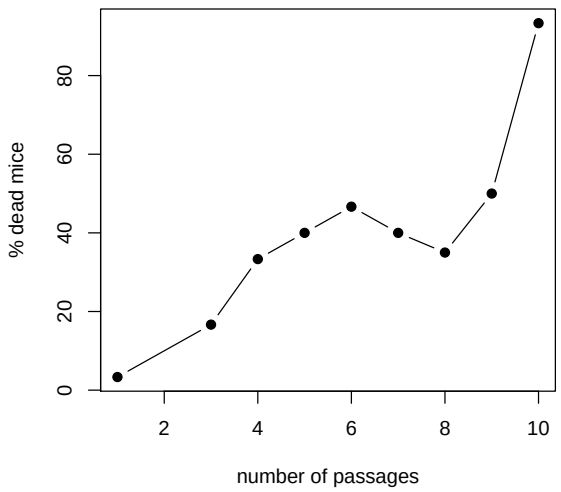

Figure 3.6: Serial transfer of *Salmonella typhimurium* in mice. Repeated experimental passage of the pathogen to new mice leads to increasing virulence. Hence, there is a correlation between virulence and intra-host competitive ability. (Source: I.A. Sutherland, Exp. Parasitol., 1996)

Intra-host competition is expected to select typically for increased virulence. More generally, however, intra-host competition will select a competitively superior variant irrespective of its eventual effect on host survival. Hence, competition for survival within a host and competition for transmission between hosts, may have opposing effects on the evolution of virulence. While inter-host competition may (often, but not generally) select for a low or intermediate level of virulence, intra-host competition may select for high levels of virulence. Importantly, when taking intra-host competition into consideration the overall optimal level of virulence is no longer determined by the maximization of the basic reproductive rate, since the basic reproductive rate only takes into account the competition for transmission between hosts.

This can be illustrated by the following simple model of superinfection:

$$dS/dt = \lambda - \delta S - b_1 S I_1 - b_2 S I_2 \tag{3.6}$$

$$dI_1/dt = b_1 S I_1 - (\delta + v_1) I_1 - \sigma I_1 I_2 \tag{3.7}$$

$$dI_2/dt = b_2 S I_2 - (\delta + v_2) I_2 + \sigma I_1 I_2 \tag{3.8}$$

The parameters are as in the SIR model (eqs. 3.1 - 3.5). For simplicity, we neglect that infecteds can recover. The model makes the additional assumption that strain 2 can superinfect hosts that are already infected with strain 1. Hence, we assume that strain 2 has a competitive advantage within a superinfected host resulting in a net rate, σ of production of host infected by strain 2

per contact between both types of infected hosts. Provided the basic reproductive rate of strain 1 is larger than 1, the equilibrium values in absence of parasite strain 2 (i.e. $I_2 = 0$) are given by $S^\star = (\delta + v_1)/b_1$ and $I_1^\star = \lambda/(\delta + v_1) - \delta/b_1$. Substituting these values into eq. 3.8, we obtain that strain 2 can invade (i.e. its growth rate is positive for small I_2) if

$$R_0^{(2)} > R_0^{(1)} - \frac{\lambda\sigma(R_0^{(1)} - 1)}{(\delta + v_2)(\delta + v_1)}$$

where $R_0^{(1)} = \lambda b_1/(\delta(\delta + v_1))$ and $R_0^{(2)} = \lambda b_2/(\delta(\delta + v_2))$ are the basic reproductive rates of the respective strains. Since $R_0^{(1)}$ is by definition larger than one, the second term in the above equation is always negative. Hence, the basic reproductive rate of strain 2 can be smaller than that of strain 1 provided its competitive advantage within the host compensates for its smaller basic reproductive rate.

In summary, intra-host competition can lead to levels of virulence that are higher than optimal for maximal transmission between hosts. Further above we derived that selection for maximal transmission (i.e. maximal R_0) is expected to lead to avirulence if there is no constraint between the infectivity rate and virulence. Taking into account intra-host competition, one would expect evolution towards intermediate levels of virulence even in this scenario, since intra-host competition would select for higher virulence (due to a correlation between intra-host competitive ability and virulence), while inter-host competition would select for lower virulence (in order to keep the host alive longer).

As pointed out further above intra-host competition leads to the selection of properties that confer a fitness advantage within the host irrespective of its consequences for the transmission of the parasites to new hosts. This phenomenon has been termed *short sighted evolution*, although it should be pointed out that natural selection never acts with foresight. Such short-sighted evolution can for example explain the virulence of polio virus. Polio virus normally replicates in the gut of infected hosts. When the infection is confined to the gut, the disease is typically harmless. However, occasionally during an infection the virus also infects nervous tissue, which leads to the severe neurological consequences that are associated with poliomyelitis. Clearly, infection of the nervous tissue is not of advantage for transmission to new hosts, since the virus can not be transmitted directly between the nervous tissue of two hosts. Instead, the infection of nervous tissue can be better understood selective advantage within a host due to the ability colonise of a new niche in a host[n].

3.9 Local adaption, host heterogeneity, and cross-species transmission

It is often believed that parasites that cross a species border are highly virulent. Indeed, diseases such as HIV, SARS, avian influenza, West Nile virus, and Ebola virus are characterized by high virulence. However, this belief is likely due to an observational bias. Clearly, only those infections

[n]The concept of short-sighted evolution was proposed in B. R. Levin & J. J. Bull, Trends in Microbiology, 1994.

that cause virulent effects will be noticed, while many cross-species transmission with mild or no effects may go unnoticed.

A similar argument is frequently made regarding the virulence of parasites that have moved between different populations of the same species. In this context it is often stated that parasites are more virulent when entering a novel population that has not previously encountered the parasite. Many examples can be given that seem to support this view. Introduction of measles into the population of native American Indians led to devastating epidemics. Similarly the introduction of small pox into the Aztec population by the Spanish conquistadores may have actually contributed more to the victory of Hernando Cortez[o], then his much acclaimed skills as a conqueror. A further example is the introduction of chestnut blight fungus into the population of American chestnuts. The epidemic was started by the introduction of the fungus to the botanical garden in Brooklyn, New York around 1900 . By the 1940 the fungus had essentially wiped out the American Chestnut and has since reduced a formerly magnificent tree on which a large timber industry relied to the existence as a shrub. Although these examples are impressive, they should not obscure our view on whether we generally expect new host-pathogen associations to be more virulent than long-lasting ones.

One study that attempted to address this question[p] was based on strains of water fleas (*Daphnia magna*) that were isolated from natural ponds which were up to 3000 km apart. In addition three strains of the horizontally transmitted parasite *Pleistophora intestinalis* were isolated from three ponds that were in close proximity to each other. In contrast to the common belief that new host-pathogen associations should be more virulent, this experiment supports the view that locally co-adapted host parasite system tend to be characterized by higher virulence than novel host parasite associations (see figure 3.7).

There are also strong arguments against the view that cross-species transmissions should generally be associated with high virulence. It has been frequently shown that increasing adaptation towards one host species lowers virulence in other host species. Figure 3.8, for example, shows the serial passage of a fungus on barley that was originally adapted for growth on wheat. Here increases in virulence on barley were associated with loss of virulence on wheat.

Serial passage in a homogeneous host population typically leads to increasing virulence (see for example fig. 3.6). However, under natural circumstance host populations will typically be heterogeneous. The observation that increases in virulence in one host type may decrease virulence in other host types suggest that host heterogeneity may be an important factor preventing the evolution of high virulence[q].

The fact, that passage in a new host results in attenuated virulence in the original host has in fact been employed for the development of life attenuated vaccines[r] . The development of the

[o]Hernando Cortez (1485-1547) was a Spanish explorer who is famous mainly for his march across Mexico and his conquering of the Aztec Empire in Mexico

[p]Source: D. Ebert, Science, 1994.

[q]For detailed investigation of the effects of host heterogeneity see R.R. Regoes, M. A. Nowak & S. Bonhoeffer, Evolution, 2000.

[r]Life attenuated vaccines typically induce the most protective immune responses, but in contrast to killed vaccines, life attenuated vaccine involve the risk that a replication competent pathogen is infected into patients, which can potentially revert to restore its virulence

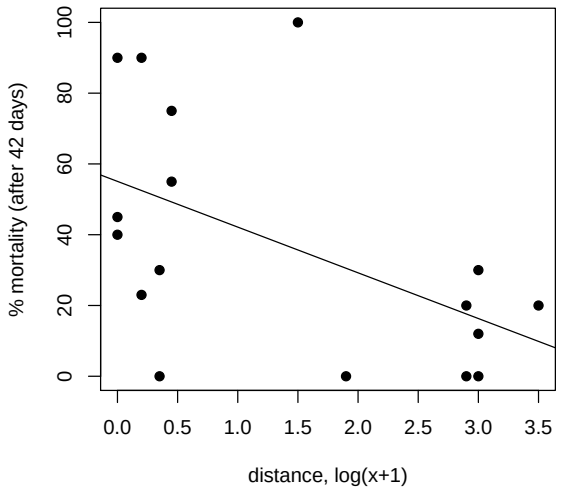

Figure 3.7: Local adaptation in of the cytoplasmic parasite of *Pleistophora intestinalis* to its host *Daphnia magna*. Three parasite strains isolated from ponds in close proximity to each other were tested for their effect on hosts isolated from ponds with increasing distance from the the pond from which the parasites were sampled. Increasing geographic distance is correlated with decreasing virulence, implying that locally adapted host-parasite combinations tend to be more virulent. Therefore this data is evidence against the conventional view that co-adapted host parasite associations should be less virulent.

life-attenuated Sabin[s] vaccine against polio virus is a good example. The Sabin vaccine used to immunize against the disease poliomyelitis is a live poliovirus, which does not lead to disease (except to about 1-2 in a million people vaccinated) because it is genetically weakened so that the immune response can control it. The attenuated vaccine strains came from wild, virulent strains of poliovirus, but they were evolved by Albert Sabin to become attenuated. Essentially, he grew the viruses outside of humans, and as the viruses became adapted to those non-human conditions, they lost their ability to cause disease in people.

[s]Albert Sabin, was born in Poland in 1906. He and his family emigrated to the United States in 1921 to escape anti-Semitism. Sabin's research included work on pneumonia, encephalitis, toxoplasmosis, viruses, sandfly fever, dengue and cancer. He began to work on poliomyelitis after World War II. Sabin scoured the world looking for weak strains of polio virus, found three, and began to develop his oral, "live" vaccine, administered at first on a lump of sugar or in a teaspoonful of syrup. In 1957 the World Health Organization (WHO) decided Sabin's vaccine deserved world-wide testing. But at home in the United States, Sabin had a hard time convincing the U.S. Public Health Service that his method was any better than Jonas Salk's "killed" vaccine method. An advantage of Dr. Sabin's oral vaccine, especially in less developed countries, is ease of administration: no shots. But two other pluses are even more important. First, the live vaccine gives both intestinal and bodily immunity; the killed vaccine gives only bodily immunity and allows the immune person to still serve as a carrier or transmitter. Second, the Sabin vaccine produces lifelong immunity without the need for a booster shot or vaccination.

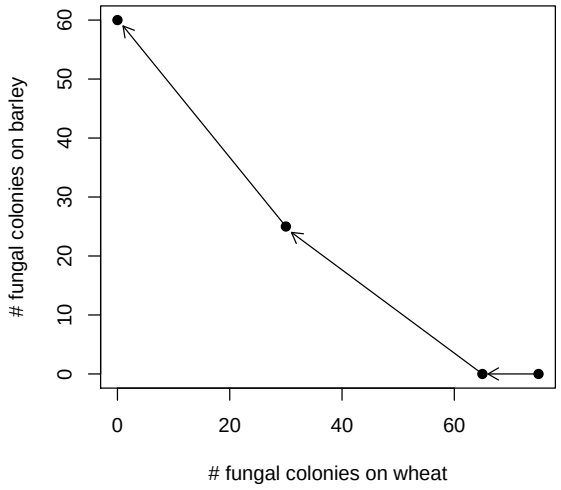

Figure 3.8: Change of the number of colonies of a wheat adapted strain of the fungus *Septoria nodorum* after serial passage on barley. The data show that increased virulence (as measured by the number of fungal colonies) on a new host is accompanied by changes in virulence on the original host. The arrows indicate the direction of adaptation in successive rounds of serial passage starting from strain that was adapted for growth on wheat.

Chapter 4

Toolbox: Evolutionary game theory

4.1 Historical introduction

Frequently the consequences of making a choice depends on what other individuals chose to do. For example, think of the decision to take the train or the car to go for the weekend to Ticino. Clearly this decision can in part be made based on external factors, such as the price of the train ticket versus the costs of fuel. However, your decision may also depend on how many other people decide to take the car, because this affects your chance of getting stuck in a traffic jam at the Gotthard Tunnel. Often, after having made your decision, you realize that many others have made the same decision as you and you find yourself in an overcrowded train with no traffic at the Gotthard Tunnel or your are stuck at the Gotthard Tunnel and you see one train after the other go by.

Game theory is a mathematical approach to study the reward of an individual's choice of strategy as a function of the other strategies found in the population. Not surprisingly, game theory has been developed and applied widely in economics, since many quantities of economical interest such as the value of a currency or the price of options depends on what other people are willing to pay for it. However, in evolutionary biology game theory has also become an important tool to study adaptation. For the application of game theory to economics individuals are typically assumed to behave rationally according to some criterion of self-interest (such as a maximization of their money). In evolutionary game theory, however, the assumption is not one of rational decision making but rather that the payoff of a particular strategy is assumed to be related to Darwinian fitness. Hence better strategies are assumed to leave more offspring in the next generation, and natural selection will thus select for the best strategy. Game theory is a tool to derive the optimal strategy for an individual in a population. Its application to economy is thus based on the assumption that all individuals participating in a "game" behave rationally, an assumption that is frequently highly questionable for human behavior. Thus, one may argue that the assumptions underlying game theory are better fulfilled in evolutionary biology,. The reason is that provided the payoff of a strategy is related to Darwinian fitness, natural selection will do the job of selecting for individuals adopting the optimal strategy. Thus the application of

game theory to evolutionary biology does not require the capacity of rational behavior. Instead it is necessary to assume that the payoff of a strategy is linked to fitness and that the propensity to adopt a particular strategy is genetically heritable.

Game theory was developed by the great US/Hungarian mathematician John von Neumann[a] and the German-born economist Oskar Morgenstern[b]. In 1967 the British evolutionary biologist Bill Hamilton[c] published a seminal paper on an unbeatable strategy of sex ratio allocation that used game-theoretical concepts. The British evolutionary biologist John Maynard Smith[d] and the American population geneticist George Price[e] extended these ideas and developed the concept of an evolutionary stable strategy (ESS). This idea is closely related to the game-theoretic concept of the Nash Equilibrium, which goes back to the American mathematician John Nash[f].

[a]John von Neumann (1903-1957) was a child prodigy, born into a banking family is Budapest, Hungary. When only six years old he could divide eight-digit numbers in his head. After simultaneously earning a doctorate in mathematics from the University of Budapest and a doctorate in chemistry from the University of Zurich, he joined the faculty of the University of Berlin in 1927. He quickly gained a reputation in set theory, algebra, and quantum mechanics. At a time of political unrest in central Europe, he was invited to visit Princeton University in 1930. When the Institute for Advanced Studies was founded there in 1933, he was appointed to be one of the original six Professors of Mathematics, a position which he retained for the remainder of his life. Being undoubtedly one of the most creative mathematicians of the last century, he was heavily involved among other things in the development of the first computer. John von Neumann's 1944 book with Oskar Morgenstern, "Theory of Games and Economic Behavior" was a landmark of twentieth century social science. In case you are worried about the mathematical content of these lectures, then a quote by John von Neumann might help you: "If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is."

[b]Oskar Morgenstern (1902-1977) was born in Germany. After earning his doctorate, he became a professor at the University of Vienna in 1935. He was on leave in the United States in 1938 when the Nazis occupied Vienna. He was dismissed from the university because he was considered "politically unbearable." So Morgenstern became a professor at Princeton University, where he remained until his retirement in 1970.

[c]William Donald "Bill" Hamilton, (1936 - 2000) was a British evolutionary biologist. Hamilton was born in Cairo, Egypt, the second eldest of six children. His father, A. M. Hamilton was a New Zealand-born engineer, and his mother, B. M. Hamilton was a medical doctor. Hamilton is generally considered as one of the most important biologist of the last century. He became famous for his theoretical work expounding a rigorous genetic basis for the existence of kin selection. Hamilton also published important work on sex ratios and the evolution of sex.

[d]John Maynard Smith (1920 - 2004) was a British evolutionary biologist and geneticist. John Maynard Smith was born in London, the son of a surgeon but following his father's death in 1928 the family moved to Exmoor, where he became interested in natural history. He went to Eton College but was very unhappy there. At a young age he developed an interest in Darwinism and mathematics, having read the work of old Etonian J.B.S. Haldane, whose books were in the school's library despite the bad reputation Haldane had at Eton for his affiliation to communism. On leaving school, Maynard Smith joined the Communist Party of Great Britain and started studying engineering at Trinity College Cambridge. Originally an aeronautical engineer during the Second World War, he then took a second degree in genetics under the great J.B.S. Haldane. Maynard Smith was instrumental in the application of game theory to evolution and theorised on other problems such as the evolution of sex and signalling theory.

[e]George R. Price (1922 - 1975) was a American population geneticist. Originally a physical chemist and later a science journalist, he moved to London in 1967, where he worked in theoretical biology at the Galton Laboratory, making three important contributions: Firstly, rederiving W.D. Hamilton's work on kin selection with a new Price equation; secondly, introducing (with John Maynard Smith) the concept of the evolutionarily stable strategy (ESS), a central concept in game theory; and thirdly, formalising Fisher's fundamental theorem of natural selection. A troubled man, Price converted from atheism to christianity, and after giving all his possessions to the poor, committed suicide.

[f]John Forbes Nash (born 1928) is an American mathematician who works in game theory and differential geometry. He shared the 1994 Nobel Prize in Economics with two other game theorists, Reinhard Selten and John Harsanyi. From Pittsburgh he went to Princeton University where he worked on his equilibrium theory.

4.2 Basic concepts

In the third semester you have had a brief introduction into the hawk-dove game and the prisoner's dilemma as simple metaphors for animal contests and the evolution of cooperation. Here, we will briefly review some of the basic concepts, such as the payoff matrix and evolutionary stable strategies.

The hawk-dove game is a simple, two-player game, which despite is oversimplification, has become a paradigm for the study of animal contests. Imagine, that two individuals from the same species compete for a resource that gives a gain G in fitness. Suppose, that the individuals in such a contest adopt one of two *strategies*:

Hawk: Fight until you lose or win

Dove: Display, but give in if opponent escalates

If both opponents fight, it is assumed that one of them is injured and that the fitness cost to injury is C.

The *payoff matrix* for the hawk-dove game is given by

$$\begin{array}{c} \\ \text{Hawk} \\ \text{Dove} \end{array} \begin{array}{c} \begin{array}{cc} \text{Hawk} & \text{Dove} \end{array} \\ \begin{pmatrix} \frac{1}{2}(G-C) & G \\ 0 & \frac{1}{2}G \end{pmatrix} \end{array} \tag{4.1}$$

Here we assume that a hawk playing against another hawk has an even chance of getting harmed or winning the contest. When a hawk meets a dove, the hawk gains the full resource, while the dove gets no reward. If two doves meet, the assumption is both have an even chance of winning the resource, but neither of them gets injured. The payoff matrix is read in following way: The headers of the rows refer to player A and the headers of the columns refer to player B. The payoff matrix then gives the reward of player A playing against player B.

The game depends critically on whether the value of the resource is greater or less than the cost of injury. We consider two numerical examples: (i) $G = 4$ and $C = 2$ and (ii) $G = 2$ and $C = 4$. The corresponding payoff matrices are given by

$$\text{(i)} \quad \begin{pmatrix} 1 & 4 \\ 0 & 2 \end{pmatrix} \qquad \text{and} \qquad \text{(ii)} \quad \begin{pmatrix} -1 & 2 \\ 0 & 1 \end{pmatrix} \tag{4.2}$$

Consider now a population consisting only of doves. A rare hawk mutant can invade into a dove population for both cases (i) and (ii), since the second columns of both matrices show that hawks are at an advantage when playing against doves. Consider now a population consisting

He received a Ph.D. in 1950 with a dissertation on non-cooperative games, which led to the concept that was later called the Nash equilibrium. Having been an outstanding mathematician in his early years, his career was hampered by mental illness. John Nash is known to suffer from schizophrenia. His life was subject of the Hollywood movie "a beautiful mind".

only of hawks. Whether a dove can invade into the hawk population depends on whether the cost of injury is greater or smaller than the fitness gain associated with the resource. The first column of the payoff matrices shows that the doves are at a disadvantage in case (i), since the payoff of hawks playing against hawks is higher than the payoff of doves playing against hawks. Thus in this case the hawk strategy is evolutionarily stable since it cannot be invaded by a mutant strategy. However, in case (ii) the doves can invade into a population consisting of hawks, since the first column shows that doves play better against hawks then they do against themselves. Hence, in this case neither dove nor hawk are evolutionary stable strategies. Although the hawk-dove game may be naive in its assumptions, it is nevertheless useful for developing an understanding of why animals that can potentially cause serious harm to each other often tend not to escalate fights.

Let us derive a more formal (i.e. more generally applicable) definition of an *evolutionary stable strategy*, or ESS for short. Let p and q be strategies from a predefined set of strategies and let $W(p, q)$ denote the payoff of p against q. A strategy p is defined as an ESS if no mutant strategy q can invade into a population playing strategy p. Mathematically, we thus assume that an arbitrarily small fraction ϵ of the population plays strategy q, while the rest, i.e $1 - \epsilon$, play strategy p. Irrespective of what strategy an individual plays it will encounter individuals playing q with probability ϵ and individuals playing p with probability $1 - \epsilon$. Hence, on average an individual playing q gets the payoff $\epsilon W(q, q) + (1 - \epsilon)W(q, p)$ and an individual playing p gets the average payoff $\epsilon W(p, q) + (1 - \epsilon)W(p, q)$. Hence, no strategy q can invade p if

$$\epsilon W(q, q) + (1 - \epsilon)W(q, p) < \epsilon W(p, q) + (1 - \epsilon)W(p, p) \tag{4.3}$$

for all possible strategies $q(\neq p)$ and for sufficiently small ϵ. Hence we obtain two conditions

$$W(q, p) < W(p, p) \tag{4.4}$$

or if $W(q, p) = W(p, p)$ then

$$W(q, q) < W(p, q) \tag{4.5}$$

These two conditions together define an ESS. The first condition is the definition of a (strict) Nash equilibrium. It implies that a strategy is the unique best reply to itself, since no other strategy plays as well against p as p plays against itself. The second criterion states that if another strategy q exists, that plays as good against p as p plays against itself, then p has to play better against q than q plays against itself.

Using this definition we see that or the strategy set "hawk" and "dove ", hawk is an ESS if the cost of injury C is smaller than the fitness gain G associated with the resource. However, if $C > G$, then neither hawk nor dove are ESS. These two strategies are called *pure* strategies. An alternative to these two pure strategies, is a so called *mixed* strategy, where an individual plays "hawk" with probability ρ and "dove" with probability $1 - \rho$. There are obviously infinitely many mixed strategies, since ρ can vary between 0 and 1.

The notion of evolutionary stability relies implicitly upon dynamical considerations, much like the invasibility considerations of parasite strains discussed in section 3.2. To study the dynamical behavior of games such as the hawk-dove game it is frequently convenient to express them in terms of systems of differential equations. Let us assume that we have n types of

strategies in the population with frequencies x_1 to x_n. Generally, the growth rate of the frequency of a strategy i in the population is given by

$$dx_i/dt = (w_i - \bar{w})x_i \tag{4.6}$$

where w_i is the fitness of strategy i and $\bar{w}$ is the average fitness of the entire population. This dynamical representation of a game is called the *replicator equation* and is really nothing more than the basic Darwinian principle that the growth rate of a phenotype (here a strategy) in the population is given by the difference between the fitness of this phenotype and the average fitness of all phenotypes in the population.

Let us derive the replicator equation for the hawk-dove game allowing only for the pure strategies of hawk and dove. If x denotes the frequency of hawks, then the frequency of doves is $1 - x$. Thus, it is sufficient to write down the dynamics for only one of the two strategies, since the dynamics of the other strategy will then be given too. Since in evolutionary game theory we generally equate the fitness with the payoff, the fitness of the hawk strategy is given by $w_H = x(G - C)/2 + (1 - x)G$. The fitness of doves is given by $w_D = (1 - x)G/2$. Thus the average fitness is given by

$$\begin{aligned} \bar{w} &= xw_H + (1 - x)w_D && (4.7)\\ &= x^2(G - C)/2 + x(1 - x)G + (1 - x)^2G/2 && (4.8)\\ &= 1/2(G - Cx^2) && (4.9) \end{aligned}$$

Thus we obtain for the replicator equation

$$\begin{aligned} dx/dt &= (w_H - \bar{w})x && (4.10)\\ &= (x(G - C)/2 + (1 - x)G - 1/2(G - Cx^2))x && (4.11)\\ &= 1/2x(1 - x)(G - Cx) && (4.12) \end{aligned}$$

Hence, we have three equilibrium solutions: (i) $x = 0$, (ii) $x = 1$, and (iii) $x = G/C$. The graphical stability analysis shown in figure 4.1 demonstrates that $x = 1$ is stable if the costs of injury (C) are smaller than the fitness gain (G) associated with securing the resource. If $G < C$, then the only stable equilibrium is a mixed population where hawks are present with a frequency G/C in the population. In fact, it can be shown that a mixed strategy $\mathcal{M}$, which consists of playing hawk with a probability $\rho = G/C$ (assuming that $G < C$) and dove otherwise, is an ESS, since no strategy that plays hawk with a lower or higher frequency can invade into the population consisting only of individuals playing $\mathcal{M}$.

We still need to introduce a few more basic concepts of game theory. The first concept regards the distinction between symmetric and asymmetric games. Asymmetric games are games in which there is an asymmetry between the roles to the participants in the game. An example for an asymmetric game is chess. The role of white and black is different, since by convention white moves first. The hawk-dove game is an example for a symmetric game, since their is no distinction between to role of the two players (although they may chose to employ different tactics). In many biological "games" there are inherent asymmetries. For example, your willingness to escalate a fight over a territory may depend on whether you are the current occupant of the territory of whether you are an intruder.

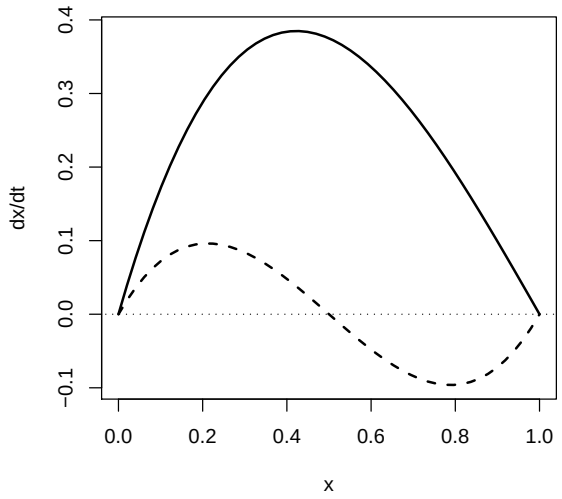

Figure 4.1: Graphical stability analysis of the replicator equation for the hawk-dove game. The growth rate dx/dt given by eq. 4.12 is plotted against the frequency of hawk strategists in the population. The solid line is represents the case where the costs of injury C are smaller than the fitness gain G associated with the resource ($G = 4$ and $C = 2$). The dashed line represents the reverse situation with $G = 2$ and $C = 4$. In case $G > C$ we see that $x = 1$ is the only stable equilibrium point since $dx/dt > 0$ for when $x > 0$. However, if $G < C$ then the only stable equilibrium point is $x = G/C$, since for values $x > G/C$ we have $dx/dt < 0$ and for values $x < G/C$ we have that $dx/dt > 0$.

Many games have the property that one individual's gain is exactly the other individuals loss. Such games are called *zero-sum* games. One example for a zero-sum game is the children's game "Rock, paper, scissors". Here players chose one of three objects, a rock, a pair of scissors, or a piece of paper. The rule is that stone blunts scissors, scissors cut paper, and paper wraps stone. The gain of the winner corresponds to the loss of the loser. The pay-off matrix for this game is

$$\begin{array}{r} \\ \text{Rock} \\ \text{Scissors} \\ \text{Paper} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Rock} & \text{Scissors} & \text{Paper} \end{array} \\ \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix} \end{array} \tag{4.13}$$

The rock-scissors-paper game is not only a children's game. It has for example been observed to govern the frequency of three mating strategies in the male side-blotched lizard *Uta stansburnia*[g]. Orange-throated males O maintain territories large enough to contain several females. Blue-throated males B maintain territories that contain a single female. Finally, yellow-throated males Y have no territory. If the population is dominated by the O type, then the Y type has an advantage, because the O males are busy defending their territory, while the Y males sneak in and mate with the females. If Y males predominate, then the B males have an advantage, since they can defend their territories against the Y males. Finally, if the B males predominate, then the O males have an advantage, since they have more females, but need not spend time to defend their territory [h].

The hawk-dove game or the rock-scissors-paper game are two player games. This focus on two player games is partially for convenience, but many interactions between animals do indeed involve only two players (such as animal contests). However, it is important to keep in mind that many interactions are more appropriately addressed as multiplayer games.

4.3 The prisoner's dilemma as a metaphor for the evolution of cooperation

4.3.1 The prisoner's dilemma

Just as much as the hawk-dove game has become the game-theoretic paradigm of animal contests, the prisoner's dilemma has become the paradigm for the evolution of cooperation. Virtually thousands of articles have been published on the prisoner's dilemma. The prisoner's dilemma is a two-player game and goes as follows: The players have to choose between two options, namely to cooperate with the other player or to defect. If both players cooperate they both get a payoff R (reward). If both players defect they get a payoff P (punishment). If one player defects and the other cooperates, then the defector obtains a payoff T (temptation), while the cooperator obtains the sucker's payoff S. The prisoner's dilemma is defined by $T > R > P \geq S$. An

[g]Source: B. Sinervo and C.M. Lively, Nature, 1996.

[h]The Rock-paper-scissors game can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

numerical example for the payoff matrix of the prisoner's dilemma is

$$\begin{array}{c} \\ \text{Cooperate} \\ \text{Defect} \end{array} \begin{array}{c} \begin{array}{cc} \text{Cooperate} & \text{Defect} \end{array} \\ \begin{pmatrix} 3 & 0 \\ 5 & 1 \end{pmatrix} \end{array} \tag{4.14}$$

The original tale behind the prisoner's dilemma goes as follows: The police have caught two mafiosos and interrogate both criminals separately. They police offer to reduce the expected prison term if they are willing to sing. Both prisoners thus have the choice to sing or to remain silent. If both prisoners decide to remain silent, then the police will not be able to find out about the full list of crimes committed and thus they can both expect a shorter prison sentence. If one of them sings, then he will receive a very short prison sentence, while the other has to go behind bars for all of the crimes committed together. However, if both sing they will have to go to prison for a long time, with little shortening of the prison sentence. Notably, not singing (i.e. not cooperating with the police) is regarded as cooperation, while cooperating with the police is regarded as defection (because the prisoners betray each other).

What should the prisoners do? Actually, the first observation about the prisoner's dilemma is that it actually isn't a dilemma at all. The rational behavior is to defect, since whatever the other player is doing, defecting is the better response than cooperating. Indeed, in the numerical example given above when defecting against a cooperating players you earn 5 points, while when cooperating you earn only 3 points. Similarly, when defecting against a defector you earn 1 point, while when cooperating against a defector you earn no points. The problem is that if your opponent is a rational player too, then you and your opponent predictably will earn only 1 point each.

You may not find it easy to accept this conclusion, and indeed experimental test show that many people chose to cooperate rather than defect. But this does not undercut the argument that this behavior is nonetheless irrational. So, why are we inclined to cooperate, and more generally, why does cooperative behavior evolve?

4.3.2 The iterated prisoner's dilemma

Let us move away from the example of prisoners. The prisoners dilemma is actually much more general. Despite being simplistic, it is nevertheless characteristic for a lot of interactions, where both parties expect an advantage from an interaction, but one party can maximize its payoff by cheating on the other party. One way to insure that defection is not too rewarding is to impose laws, as is the case in human societies. But why should animals cooperate in the absence of a law enforcement agency?

One of the reasons why cooperation may actually be a sensible strategy in a prisoner's dilemma situation is that there is the prospect that the game will be played again. Cooperating in this round may be reciprocated in future games. Although this idea is compelling it is not quite as easy as this. Assume that two-players are chosen to play the game repeatedly against each other. If the players know in advance how many games will be played, then defection

again emerges as the only rational strategy. As a classic saying in game theory states, the last movement is always to defect. Indeed, the last game is equivalent to the simple "one-shot" prisoners dilemma. The decision to cooperate in this last game is not confounded by expectations of future returns. So, the rational strategy in this last game is to defect. Given that you defect in the last game, what should you do in the second to last game? Clearly, the same argument holds, and thus you can see that the only rational behavior is to defect right from the beginning!

What promotes cooperative behavior is the expectation of future games. If you are certain that this is your last game with a partner, then you have to defect. However, if there is a chance of future interactions, then the argument of the last game does not apply, and this makes cooperation enticing. However, what would the optimal strategy be in such a game where there is always a possibility of future encounters? This question is far from easy to answer. In 1979 Robert Axelrod[i] had the idea to make a computer tournament. He invited researchers in the field to submit their favorite strategies. Fifteen strategies were submitted and Axelrod measured their performance against each other and against themselves. It turned out that the simplest strategy actually outperformed all others. This strategy is called *tit for tat*[j] and is defined by cooperating in the first round and then playing whatever the opponent has played in the previous round. The fact that the simplest strategy won the tournament is perhaps made less surprising if one knows that tit-for-tat was submitted by one of the grand old men of game theory, Anatol Rapoport[k]. What is surprising though, is that tit-for-tat turned out to be the winner without winning a single game. In fact, tit-for-tat can at best earn as many points as its opponent, but never more. Tit-for tat is friendly in the sense that it never defects first and tit-for-tat is never the first to break a series of rounds of cooperation. On the other hand, tit-for-tat cannot be easily exploited, since it strikes back immediately after the first defection. Thus its strength is that is fairs well against strategies that tend to be more cooperative, while preventing to be exploited by strategies that tend to defect.

Of course, there could be strategies that are still better than tit-for-tat. After Axelrod published his analysis of the 15 submitted strategies he called for another tournament. This time 62 strategies were submitted. It seemed clear that the strategy to beat was tit-for-tat. John Maynard Smith, for example, submitted tit-for-two-tat, which defects only if the opponent has defected twice in a row. Having done his home-work properly, this strategy would have won the first tournament. However, it finished only twenty first in the second tournament. The surprising winner of the second tournament was again tit-for-tat. If all competitors are known, it is certainly possible to find a strategy that outperforms tit-for-tat. However, the strength of tit-for-tat is that it performs well also against unknown strategies. Tit-for-tat is a generalist that can deal well with many strategies.

[i]Robert Axelrod is professor of political science at the University of Michigan. As an undergraduate he studied mathematics at the University of Chicago and then did his Masters and PhD work in political science at Yale.

[j]German translation: "Wie du mir so ich Dir".

[k]Anatol Rapoport (born 1911) is a Russian-born American mathematician. Among other things, he worked on mathematical biology , the mathematical modeling of social interaction and stochastic models of contagion. He applied his mathematical expertise with psychological insights to the study of game theory. He also studied piano, conducting and composition, at the Staatsakademie für Musik und darstellende Kunst, in Vienna between the years (1929-1934). However due to the rise of Nazism he found it impossible to make a career as a pianist. He shifted his career into mathematics and obtained a Ph.D. degree in mathematics. His work on game theory was motivated by a commitment to the peace movement and conflict resolution.

Note that tit-for-tat (TFT) is nevertheless not an ESS. Consider the strategy $AllC$, which always cooperates. We have that $W(TFT, AllC) = W(TFT, TFT)$ and $W(AllC, AllC) = W(TFT, AllC)$. Hence, tit-for-tat does not fulfil the conditions for an ESS given by inequalities 4.4 and 4.5. Instead, $AllC$ can invade a population consisting only of TFT by neutral drift. If you allow for a third strategy, $AllD$, which unconditionally defects, then one observes cyclical dynamics, since $AllC$ can invade TFT, $AllD$ invades $AllC$, but TFT again rises in the presence of $AllC$ and $AllD$, because it can prevent exploitation by $AllD$ while cooperating with $AllC$.

4.3.3 The prisoner's dilemma in spatial structure

Cooperation between related individuals can be explained by kin selection[l]. The prisoner's dilemma is interesting, because it offers a potential explanation for cooperation between unrelated individuals based on reciprocation. (The idea of evolution of cooperation by reciprocation goes back to Bob Trivers[m].) However, the evolution of altruistic behavior by reciprocation requires substantial cognitive capacities among the players. In particular, it requires the ability to remember the outcome of past encounters. While we like to think that humans would generally possess such cognitive skills, one only needs to think of how difficult it is for some of us (including me) to remember faces. Hence, remembering the outcome of all interactions when living in a large group is indeed a taxing cognitive task.

However, cooperative behavior is also found in simple organisms that do not posses sufficient cognitive skills[n]. In many instances cooperation may be explained by kin selection, but nevertheless it is important to show that there are also circumstances under which cooperative behavior can evolve in the absence of cognitive skills. An important step in this direction was made by Martin Nowak[o] and Bob May, by studying the spread of simple strategies, when the prisoner's dilemma is played on a spatial grid[p].

Let us consider only two simple strategies: C, which denotes unconditional cooperation and D, which denotes unconditional defection. Note, that these strategies do not require any cognitive abilities, since their strategy is not based on memory of past encounters. The strategies C and D could be hardwired genetically. Individuals playing either C or D are placed on a 100×100 spatial grid. Each site on the grid is thus occupied by either a C- or a D-type individual. In each round of the game the individuals plays with all of its immediate neighbors. The score for each individual is the sum of the payoffs that result from the interactions with all

[l] I assume that you are familiar with the concept of kin selection. Kin selection was first proposed by Bill Hamilton in 1964. The original reference is WD Hamilton, J. theor. Biol., 1964.

[m] Robert Trivers, born 1943, is professor of anthropology at Rutgers University, NJ. In the early 70s he made several seminal contributions to evolutionary biology. Among other things, he published ground-breaking papers on parental investment, sexual selection, parent-offspring conflicts, and reciprocal altruims. Because of its implications for human behavior, his work has become politically highly controversial, although he is generally regarded as one of the most creative thinkers in evolutionary biology.

[n] For example, Paul Turner and Lin Chao showed that competitive interactions in an RNA virus can be viewed as a prisoner's dilemma. Source: PE Turner and L. Chao, Nature, 1999.

[o] Martin Nowak, born 1965, is an Austrian theoretical evolutionary biologist. After having studied biochemistry, he did his PhD thesis is mathematics at the University of Vienna with Karl Sigmund. Currently he is a professor at Harvard and is the director of the program for evolutionary dynamics.

[p] Source: M.A. Nowak and R.M.May, Nature, 1992

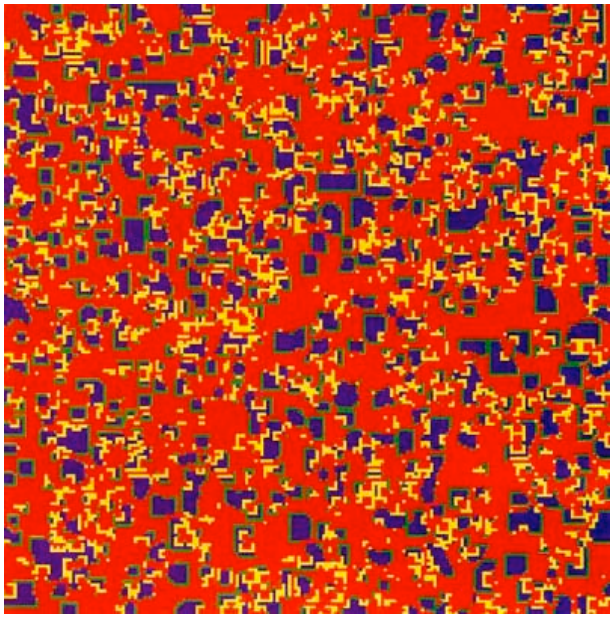

Figure 4.2: Snap shot of a computer simulation of the prisoner's dilemma on a 100×100 spatial grid. The color code is as follows: Red (blue) are those sites that have been occupied by defectors (cooperators) for the last two generations or more. Yellow are sites where a cooperator in the previous generation is replaced by a defector in the current generation. Green are sites where a defector in the previous generation is replaced by a cooperator in the current generation. Thus yellow and green represent the recent changes on the grid.

of its neighbors. In the next generation the grid is then updated and each grid site is given to the individual with highest payoff among the individuals in the immediate neighborhood and the previous owner of the grid site[q]. The payoffs are chosen such that $R = 1$, $T = b$ (with $b > 1$) and $S = P = 0$. This parametrization was done in order to reduce the number of variables, but it preserves the essential characteristics of the prisoner's dilemma. An example for the behavior of the spatial game is shown in figure 4.2. The dynamical behavior of the spatial game depends on the parameter b. The important finding, however, is that cooperators can persist indefinitely in a spatial simulation. The intuitive reason for this observation is that in a spatially structured environment, like tends to be surrounded by like, because individuals place their offspring in their immediate neighborhood. Hence cooperators tend to be surrounded by cooperators, and defectors tend to be surrounded by defectors. While cooperators benefit from being surrounded by other cooperators, defectors suffer from being surrounded by other defectors. Hence, spatial structure can facilitate the evolution of cooperation[r].

[q]The take-over rule need not be deterministic, but can also be probabilistic. This does not change the qualitative behavior of the spatial game. Source. M.A. Nowak, S. Bonhoeffer and R.M. May, Proc. Natl. Acad. Sci. 1994.

[r]While it is often the case that spatial structure facilitates cooperation, this need not generally be true. Hauert and Doebeli recently published a paper, in which they showed that for some evolutionary games (in particular the so called snow-drift game which is similar to the hawk-dove game) space actually reduces the frequency of cooperators. Source: C. Hauert and M. Doebeli, Nature, 2004.The spatial prisoners dilemma can be explored as a module in the "Modelling course in population and evolutionary biology" that takes place every summer term

4.4 Public goods games

4.4.1 The tragedy of the commons

In a seminal article in 1968 Garret Hardin[s] coined the phrase "tragedy of the commons" to describe situations where the selfish interests of individuals lead to an over-exploitation of a common good. This phenomenon has been termed the tragedy of the commons in reference the commons[t], a pasture on which all farmers in the village are allowed to let their cows graze. The use of the commons exemplifies the conflict between the interests of the individual and the community. When a farmer decides to put another one of his cows on the pasture, he alone gets the benefits (e.g. the extra milk from the cow), while the costs (e.g. overgrazing) are shared equally among all farmers. It is thus in the interest of each farmer to put more of his own cows on the commons. The tragedy lies in the fact, that the selfish interest of the individual can lead to an over-exploitation of the resource, such eventually each individual in the population is worse off. The world is full of examples for the tragedy of the commons, ranging from every day nuisance problems such as spam mails to some of most pressing global problems of our society such as global warming due to greenhouse gases or the depletion of biodiversity.

The tragedy of the commons also plays an important role in ecology and evolution[u]. One example for the tragedy of the commons in biology is the evolution of the sex ratio. Clearly, a population could reproduce faster if the sex ratio in a population was female biased, since it takes only few males to fertilize all females. However, if the sex ratio is female biased, an individual that produces off-spring with a more male biased sex ratio, will have a higher relative representation in the next generation. Thus, a 50/50 sex ratio evolves as a consequence of the "selfish" interests of the individual, although it involves a cost for the population as a whole.

4.4.2 The tragedy of the commons in heterotrophic energy metabolism

Another example for the tragedy of the commons has been identified in the production of adenosine triphosphate (ATP) in heterotrophic organisms[v]. ATP is a key compound in energy metabolism and is required to drive vital biochemical reactions. Its degradation into adenosine diphospate (ADP) and phosphate is generally used to drive thermodynamically unfavorable reactions, such as active transport and biosynthesis. Thus ATP has to be continuously regenerated for cellular growth. In heterotrophic organisms, the production of ATP is coupled to the degradation of energy-rich organic compounds. Interestingly, there is a trade-off between rate (i.e. number of ATP molecules produced per unit of time) and yield (i.e. number of ATP

[s] Garrett Hardin (1915 2003) was a ecologist from Dallas, Texas who is most known for his 1968 Science paper, "The Tragedy of the commons. He is also known for Hardin's First Law of Ecology, which states "You cannot do only one thing." Hardin received a B.S. in zoology from the University of Chicago in 1936 and a PhD in microbiology from Stanford University in 1941. He served as Professor of Human Ecology at the University of California, Santa Barbara from 1963 until his retirement in 1978.

[t] German translation: Allmende

[u] The tragedy of the commons is related to the prisoner's dilemma, but in the TOC individuals play against a group rather than an individual.

[v] Source: T. Pfeiffer, S. Schuster, and S. Bonhoeffer, Science, 2001.

molecules produced per unit of resource) in heterotrophic ATP production. This trade-off can, for example, be observed for the degradation of sugar in many heterotrophic organisms, since alternative ATP-producing pathways with opposing properties in yield and rate, such as respiration and respiro-fermentation can be utilized. The rate of ATP production can be increased by using fermentation in addition to respiration (i.e. using respiro-fermentation instead of respiration). However, using fermentation comes at a cost in terms of yield, since fermentation has a much lower yield than respiration (2 molecules of ATP per molecule of glucose for fermentation compared 32 molecules of ATP per molecule of glucose for respiration). On a more fundamental level, such a trade-off arises from thermodynamic constraints because the free energy difference between substrate and product in an ATP-producing pathway is divided into one part that is used to phosphorylate ADP to ATP and another part that is used to drive the pathway. This division causes a trade-off between yield and rate of ATP production, because the larger the part conserved by producing ATP, the slower the pathway is.

Among other things, ATP is required in large amounts for biosynthesis and thus for growth. Since energetic limitation is likely an important factor for organisms in their natural environment, we expect that the properties of ATP producing pathways have been under strong selection pressure during evolution. Thus the existence of a trade-off between rate and yield of ATP production raises the question: Under which conditions is it favorable to use a pathway with high yield but low rate? Conversely, under which conditions is it favorable to use a pathway with high rate but low yield? In other words, under which conditions does it pay to be fast and inefficient versus slow and efficient in regards to the conversion of an energy resource.

This question calls for a game theoretical approach, since the choice of strategy of (here fermentation or respiro-fermentation) depends on the strategy of other individuals in the population. Imagine that a single individual has a pile of sugar just to itself. In this case, it would make sense to convert the sugar with maximal efficiency into ATP, and hence use respiration rather than respiro-fermentation. However, the situation may change if several individuals share the resource. As we will show below, under these conditions the rate of resource consumption and ATP production is what matters and not the yield. Hence, under these circumstances, it is better to use respiro-fermentation rather than respiration.

We will now address this question with a game-theoretical approach. But instead of deriving an explicit payoff matrix for this evolutionary game, we focus directly on a dynamic situation that corresponds to the replicator equation approach. We consider three variables: S for the density of the sugar resource, N_R for the density of individuals using the "respiration" strategy, and N_{RF} for the density of individuals using the "respiro-fermentation" strategy. The dynamics can then be described by the following system of differential equations:

$$dS/dt = -N_R J_R^S(S) - N_{RF} J_{RF}^S(S) \quad (4.15)$$

$$dN_R/dt = cJ_R^{ATP}(S) N_R \quad (4.16)$$

$$dN_{RF}/dt = cJ_{RF}^{ATP}(S) N_{RF} \quad (4.17)$$

where $J_R^S(S)$ and $J_{RF}^S(S)$ are the per capita rate at which sugar is consumed by respirators and respiro-fermentors (as a function of the sugar density S), and $J_R^{ATP}(S)$ and $J_{RF}^{ATP}(S)$ are the per capita rates of ATP production by respirators and respiro-fermenters, and c is proportionality constant that accounts for the number of ATP molecules required to produce an new individ-

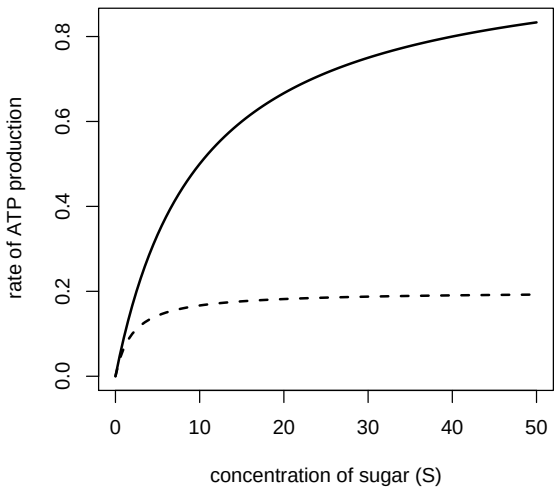

Figure 4.3: Schematic illustration of the rate of ATP production in respirators (dashed line) and respiro-fermentors (solid line) as a function of the sugar concentration. For all sugar concentrations the rate of ATP production is higher for respiro-fermentors than that of respirators. The rate of ATP production by respiration saturates faster for high sugar levels. Note that this is just a schematic illustration and is not based on actual values for any real organism. The functions plotted are given by eq. 4.18 and the chosen parameter values are $a_{RF} = 1, a_R = 0.2, b_{RF} = 10$ and $b_R = 2$. Hence at high rates of sugar the rate of ATP production by respiro-fermentors is five-fold higher than that of respirators.

ual. The model describes the growth of both types of populations when starting with a sugar concentration $S = S_0$ at time $t = 0$.

The rate of ATP production of both the respirator and the respiro-fermentor saturate for high levels of sugar. At low levels of sugar both types have a similar rate of ATP production, but at high sugar levels the rate of ATP production of the respiro-fermentor is considerably larger than that of the respirator. Thus we model the rate of ATP production as

$$J_R^{ATP} = \frac{a_R S}{b_R + S} \qquad \text{and} \qquad J_{RF}^{ATP} = \frac{a_{RF} S}{b_{RF} + S} \tag{4.18}$$

where $a_R < a_{RF}$ and $a_R/b_R = a_{RF}/b_{RF}$. A graphical illustration of the rates of ATP production of both types is shown in figure 4.3.

The yield of ATP is given by the ratio of the rate of ATP production and the rate of sugar consumption. Hence the yield of respirators is given by $y_R = J_R^{ATP}(S)/J_R^S(S)$ and the yield of respiro-fermentors is $y_{RF} = J_{RF}^{ATP}(S)/J_{RF}^S(S)$. The tragedy of commons in heterotrophic ATP production results from the fact that the yield of respirators is larger than the of respiro-fermentors (i.e. $y_{RF} < y_R$), while the rate of ATP production of respiro-fermentors exceeds that of respirators (i.e. $J_{RF}^{ATP} > J_R^{ATP}$). Figure 4.4 illustrates this phenomenon with numerical

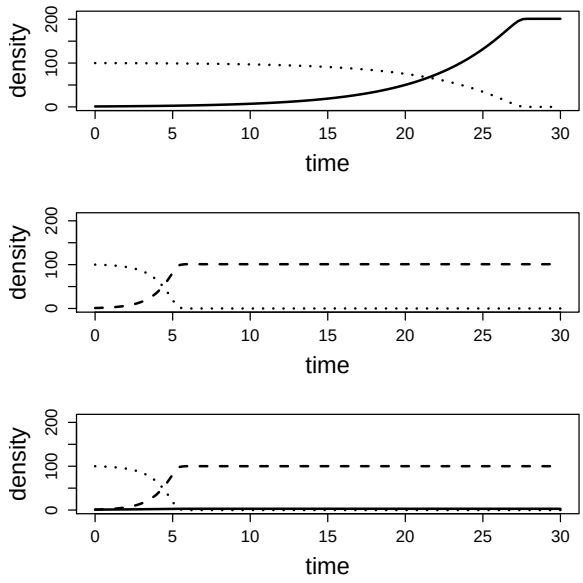

Figure 4.4: These three plots illustrate the tragedy of the commons that occurs for heterotrophic ATP production. The simulations are based on the model 4.15-4.17. In all simulations we start with an initial sugar concentration of $S_0 = 100$. The ATP fluxes are shown in figure 4.3. The top plot shows the growth of a pure respirator population (i.e. we start the simulation with the initial condition $S_0 = 100, N_R(t = 0) = 1$ and $N_{RF}(t = 0) = 0$). The middle plot shows the growth of a pure respiro-fermentor population (i.e. initial conditions $S_0 = 100, N_R(t = 0) = 0$ and $N_{RF}(t = 0) = 1$). Finally, the bottom plot shows the situation when both types are grown in competition (i.e. initial conditions $S_0 = 100, N_R(t = 0) = 1$ and $N_{RF}(t = 0) = 1$). The simulations illustrate that the respirators grow to higher population density for a given amount of sugar, but they lose out when grown in competition with respiro-fermentors, although they eventually end up at a lower population density. This illustrates the tragedy of the commons.

simulations. While the a population consisting exclusively of respirators will eventually establish a larger population size after the consumption of a resource, the respiro-fermentors outcompete the respirators when they grow in competition (although they eventually end up at a smaller population size).

The above line of reasoning suggests that heterotrophic organisms that compete for shared food should have evolved to utilize pathways with high rates of ATP production, even if this comes at a cost in terms of yield. Indeed, sugar bacteria and fungi such as *Lactobacilli* , *Mucor*, and *Saccharomyces*, which are present in the early phase of degradation of organic material, indeed use respiro-fermentation for ATP production even in the presence of oxygen, i.e also under conditions in which they could, in principle, use respiration only. In contrast, organisms metabolizing internal energy resources are expected to utilize respiration. Indeed, higher animals that derive their energy from ingested food items and thus have exclusive access to the sugar resource, do indeed use respiration as the main mode of energy metabolism.

Utilizing a shared resource via a slow but efficient ATP pathway can thus be viewed as co-

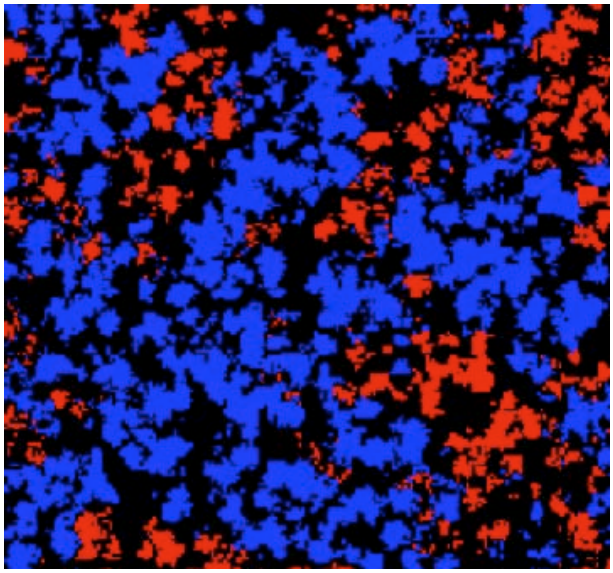

Figure 4.5: Snapshot of a spatial simulation of competition of respirators (blue) and respiro-fermentors (red). Black sites are empty sites. Depending on the rate of resource diffusion in space and the rate of cell movement, respirators can outcompete respiro-fermentors (in contrast to competition in a well-mixed environment as is described by eqs. 4.15 - 4.17). For high rates of cell movement or resource diffusion, respiro-fermentors outcompete respirators (since it resembles the well-mixed situation), while for low rates, respirators can outcompete the respiro-fermentors. (Source: Pfeiffer, Schuster & Bonhoeffer, Science, 2001)

operation[w]. As we have discussed in section 4.3.3, spatial structure can facilitate the evolution of cooperative behavior. Analogously, if resource competition between respirators and respiro-fermentors is simulated in a spatial setting one does find that cooperators (i.e. respirators) can outcompete defectors (i.e. respiro-fermentors) depending on the rates of resources diffusion and cell movement (see fig. 4.5). For high rates of cell movement or high rates of resource diffusion the effects of spatial competition vanish and thus the respiro-fermentors outcompete the respirators. However, for low rates of cell movement or resource diffusion respirators can outcompete the respiro-fermentors. The reasoning for the predominance of cooperative behavior under these conditions is analogous to that for the spatial prisoner's dilemma simulations. Limited cell movement in space has as a consequence that individuals tend to be surrounded by other individuals of the same metabolic type. Respiro-fermentors suffer more from being surrounded by their own type because they suffer from the consequences of fast local resource depletion.

The correlation between the use of respiration and spatial aggregation that is observed in spatial simulations (see fig. 4.5) finds an interesting parallel in the metabolism of in the dimorphic fungus *Mucor racemosus*. These fungus is facultatively multicellular. When unicellular it

[w]An interesting observation in this context is that cancer cells use respiro-fermentation for ATP production, which can be interpreted as opting out of the cooperation, since all other cells in a multicellular use respiration as their main mode of energy metabolism. The observation of the association of cancer cells with fermentation goes back to the German Nobel laureate Otto Warburg (Science, 1956).

uses respiro-fermentation, but when switching to a multicellular mycelial stage, the cells switch to respiration. Thus, one may speculate that the use of respiration may be a reward to multicellularity, and may even have been implicated in the evolutionary transition from unicellular to simple undifferentiated multicellular heterotrophs[x].

[x]Source: Pfeiffer & Bonhoeffer, Proc. Natl. Acad. Sci. (USA), 2004.