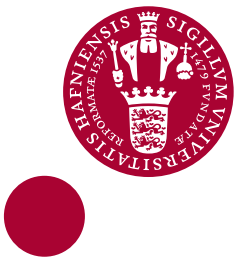

Emergent Percolating Nets in Evolution. Self-organizing networks of molecules and species, and their relation to critical systems

Engelhardt, Robin

Publication date:
1998

Document version
Early version, also known as pre-print

Emergent Percolating Nets in Evolution

Self-organizing networks of molecules and species, and their relation to critical systems

Robin Engelhardt
Center for Chaos and Turbulence Studies
Chemistry Laboratory III, H.C.Ørsted Institute
University of Copenhagen

Dissertation submitted for obtaining the Ph.D. degree at the University of Copenhagen, Denmark. September 1998.

Principal Advisor:

Associate Professor Axel Hunding
Chemistry Laboratory III, H.C.Ørsted Institute
University of Copenhagen

Evalutation Committee:

Assistent Professor Kim Sneppen,
Nordita, Niels Bohr Institute, Denmark

Associate Professor Ricard Solé,
Complex Systems Research Group,
Dept. de Fisica i Enginyeria Nuclear,
Campus Nord, Modul B4,
Universitat Politècnica de Catalunia,
Barcelona, Spain

Professor Alexander S. Mikhailov,
Complex Systems Group,
Dept. of Physical Chemistry,
Fritz-Haber-Institut der Max-Planck-Gesellschaft,
Berlin, Germany

Typeset in LaTeX

ISBN 87-7834-304-6 HCØ Tryk

Front illustration: Isis and Orisis and the origin of life (Anselm Kiefer: *Kontraste*).

Contents

1
Origins

As a former passionate chess player I often wondered about how the perfect game of chess would be. Would white win because of the advantage of making the first move, or would it not be enough? Maybe it was the interest in nonlinear phenomena in nature and their physical description which led me to the belief that there is not only *one* ultimate and perfect game of chess - but many. And they will all lead to a draw.

The idea was that maybe, if one could make a diagram of all the possible perfect games of chess, and trace them move by move on a sheet of paper, the resulting structure would have a "fractal" appearance. Every time a player makes just the smallest error, the game will inevitably fall off the drawish path and end into a disastrous defeat. Nevertheless, there will be an innumerable number of games leading to a draw and these innumerable games will have innumerable different histories and endings. A pat, an ending with a King versus a King, or a perpetual check. Only the extrinsic rules of human impatience - like the "50 moves rule", or the "three times perpetual check rule" - will let the games end. If these rules were not invented, the perfect games would still be played if started at the beginning of the universe, and they would contain the most strange appearances and the most simple oscillatory patterns.

Equally, if one could make a diagram of all the possible molecular combinations which would not end in some kind of thermal equilibrium or energetic death, their resulting patterns (if they were drawable!) would also have a self-similar and fractal shape and show the most strange and peculiar structures - including, well, yes, life and pulsars.

How can it be that carbon atoms are necessary for everything we know

as life? Maybe it is just a coincidence; a particular local realization of a yet drawish game on the universal chess board. Or, maybe, it is the only possible variant, the perfect game with only one outcome, whatever that will be. But in any case, the sole necessity of carbon atoms in the architecture of living organisms contains information: the information that they work best. Neither iron-, silicium- nor sulfur atoms can combine the way carbon atoms can; the way nucleotides and amino acids can; the way cells can.

However, just because these materials are of particular interest for the evolutionary process, it is not automatically given that life has started with them. The initial accumulation of complex structures which could overcome a defeat in time, could equally well consist of other materials than proteins, sugars and heterocyclic bases. They could have been sulfur structures, or, as in the theory of Cairns-Smith, simple clay crystals that grow and break continuously and carry along defects with them, which, in turn, would survive and be successful.

But the very mechanisms that washes away all these possibilities which could have worked biochemically - sulfur structures or clay crystals - must contain some criteria accessible for a structural scientific investigation, since this "shaking" or "natural selection" only leaves behind the atoms and molecules that are the most flexible and exchangeable. One immediate thought is that, if exchangeability and flexibility are the decisive criteria for a winning design, it is so because the all-important property is not the substance, but the form. And form is information. One of the main dogmas, the *sine qua non* of evolution, is the continuity of information between the generation, regardless of the material which carries the message.

How should we characterize this intricate drawish path called evolution? In the modern history of theories for the origin of life, the emphasis has been almost exclusively on "mechanistic" viewpoints where the investigations have focused on the material only. Experiments and chemical plausibility were the main weapons for their formulation. This led to the formulation of evolutionary theories of "the primordial soup" due to Stanley Miller, or to the theory of the "prebiotic pizza" due to Günther Wächtershäuser or to the complex mechanics of an "RNA world".

But in the last few decades, more "structuralistic" approaches have been developed, where the emphasis not so much has been on the actual chemical setup, but on the architectonic principles, the mathematical distribution functions, the self-organization and the emergent networks of interactions upon

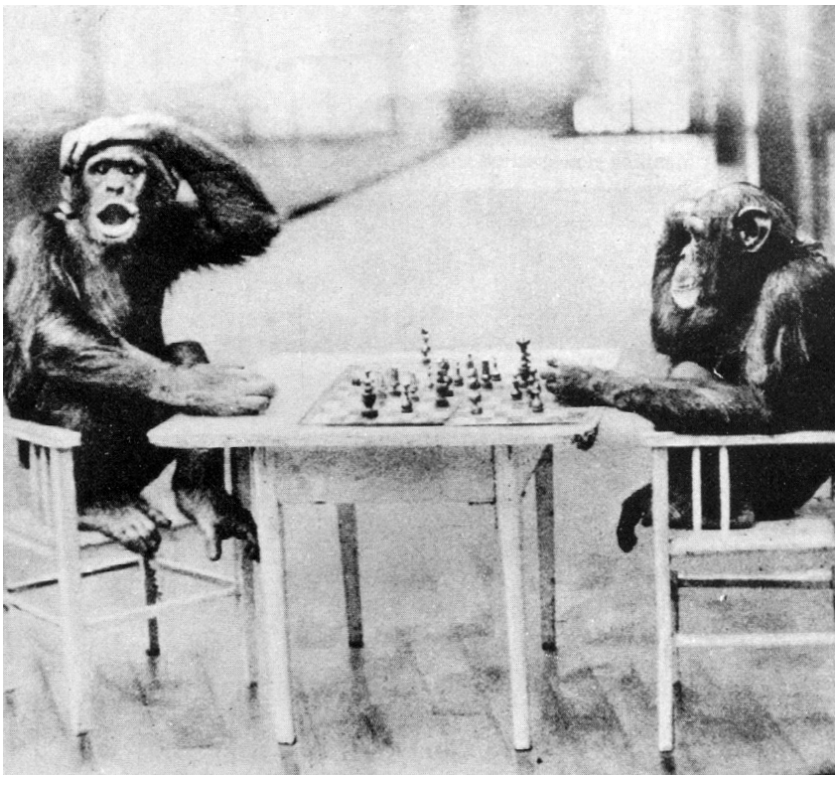

Figure 1.1: *"What! You are giving up your Queen? Cheer madness!" "Hmm..., I thought it was a draw."*

which natural selection acts.

1.1 Is the world an infinite soup?

The still dominating scientific theory for the origin of life was formulated in the middle of the 20th century. It was a time, where the chemical and biochemical sciences went from one triumph to another, and it was therefore only natural also to reach for an answer of the most ultimate question a scientist from these fields can ask: How did it all begin? What defines life and under which circumstances can it arise?

As every true chemist would do, the answers were sought to be found from the knowledge and methods a chemist has and uses: Huge flasks were filled with diverse chemicals and gases one thought were typical ingredients in the

atmosphere on earth several billions of years ago (such as methane, ammonia, water and carbon dioxide) and they became then exposed to intense outer influences such as violent changes in the temperature, shaking and lightning and thunder. When the subsequent identifications of the remaining chemicals showed traces of organic material, one felt to be on safe ground: the theory of a *Prebiotic Soup* was born. Of course, there were lots of loose ends and unanswered questions, but these were seen as secondary problems which, in time, could be explained without any further troubles.

But time has shown that these loose ends were so loose that they instead became used as starting points for completely new and different theories for the origin of life. The primary problems were the following:

- The amount and type of organic material in the lightning generated slimy pond was highly limited. Although several different amino acids were formed (in small concentrations), many "vital" molecules such as ribose, nucleotides (especially pyrimidines) and lipids refused any acquaintance with the "broth".

- Even though there were formed many organic substances, subsequent reactions destroyed them again. Especially the ultraviolet light from the sun works as an effective killer of most organic material.

- Every system of organic molecules, which in a most superficial sense could be defined as an autonomous life form, needs to have a high degree of internal dependencies where the molecules presuppose each other. This "hens and eggs"- problem show itself most clearly in the mutual inter-dependency of nucleic acids and proteins. The first can store hereditary informations, while the last reads and expresses this information through enzymatic activity. None of them can function without the other.[1]

- This last fact points towards theories which include a kind of co-evolution of molecules, but even that is utmost problematic according to several investigations. These investigations, which necessarily had to be more abstract in nature and have been done with the help of computer models and mathematical calculations, tell that the biggest problem is stability

[1] The idea of an "RNA world", where RNA functions as both an information carrier and enzyme, has been suggested as a road away from this problem, and, in fact, the RNA-world theory is gaining more and more advocates.

(at least when the starting point is common mass action reaction kinetics, where concentration changes are described by differential equations). The smallest perturbation can tip the balance towards the ordinary heat death.

1.2 Or is it a giant chess board?

Since this result seems to be in contradiction with our empirical knowledge, the error might be our way of looking at the problem or maybe the way we make our models.

First of all there is this problem with the differential equations. Chemists like to use differential equations because their molecules normally are in large concentrations, so that the mathematical description of infinitesimal concentration changes is appropriate. But in molecular evolution, the *variation* in architecture among the similar molecules is often much higher than their individual concentrations. This difference becomes even more significant under the operation of natural selection since the differences in success of different molecules might be easier describable (at least on a computer) on a discrete landscape of possibilities, where every single "molecule", or "species" acts on a set of discrete rules and is traced separately through the phase space. The recent explosion in the use of cellular automata and genetic algorithms on this kind of problems supports this suspicion. The *discretization* of the infinite soup makes it easier to focus on the combinatorics - like on a giant chess board.

Such a more structural-dynamical approach has brought together scientists from many different fields, such as game theorists, chaos theorists, neurologists, people in population dynamics, economy and also physical chemists with interest in phase transitions and dissipative structures. This new, cross-disciplinary field has by some people been called the *science of complexity* because it tries to describe so many different and complex phenomena - from evolution to the stock market - from the same point of view. If such "general" models exist, and they have been shown to exist, they are maybe some good candidates for a new understanding of the structural-dynamical principles of the living world. Of course, the approach is very general, but still, there are common trends. Especially when looking at some of the statistical distributions of such phenomena, one often finds similar behavior in very different systems - normally expressed as *power law relationships*, which are the proper and characteristic mathematical equations for self-similar structures and fractals in nature. How-

ever, the question is still open, whether these similar statistical properties - the power laws - of such a large and heterogeneous group of real phenomena are a result of a common, fundamental and yet unknown mechanism (we could call it an URO - an *unidentified random object*), or they just are some peculiar stochastic effects without any common mechanisms, and maybe describable by some (yet unknown) kind of (generalized) central limit theorem.

The present thesis tries to continue on this path. It develops some new models which are general enough to encompass previously used models in evolution such as the NK model due to Kauffman and Johnson and the Lotka-Volterra equations in ecological theory. The emphasis is on one side on the actual behavior of the models, but on the other side, the emphasis is also on the resulting *distribution functions* of these models and their relation to critical systems and percolation theory (So, between the lines there will be an ongoing discussion *for and against* the appropriateness of describing natural phenomena through distribution functions only).

1.3 Outline of thesis

The thesis consists of three parts. The first part deals with the generation of networks of molecules which are selectively neutral. We will investigate a model, which was developed by Mark Newman and myself during my stay at the Santa Fe Institute in New Mexico in the summer 1997. Without the bright ideas of Professor Mark Newman, this model would newer have been a part of the present thesis and I thank him for the opportunity to participate in this (still ongoing) project.

The second part consists of a new model of self-organizing food webs in species communities. Although many stimuli from friends and colleagues (Mark Newman and Axel Hunding) have moulded the model into the present form and sharpened my ideas about it, the main core of the model is homemade. It is never easy to learn about and get absorbed into an entirely new discipline of science, but in the case of ecological modeling and food web theory, it has been an exiting and fruitful experience, and to my surprise (but still only partial knowledge), there seems not to exist any previous model which lets food webs develop themselves from totally random initial conditions as this present model does.

The third part discusses the properties of the two models in terms of critical phenomena, percolation theory and the notion of self-organized criticality. We

will make some analytical investigations of a percolation threshold in the model for neutral evolution, and we will also discuss whether the food web model can be said to be self-organized critical. There will emerge a partial yes to this question, but the last chapter will caution the reader not automatically to identify any power law relationships in the models with criticality nor self-organization. Formulated in terms of a simple model of evolution, we will find that such self-similar distribution functions are also a result of simple tabulation effects and stochastic regularities.

Part I

Neutrality in Evolution

Introduction

When browsing through shelves of popular books on evolution, one can be quite sure that *the neutral theory of molecular evolution* is not mentioned at all. And if it's mentioned anyway in one of the books, then only as a dim fact from the category of technicalities, not worth to bother about any further.

But history has shown that this is not true, and the real reason for the missing popularizations of the neutral theory is to be found in its mathematical and abstract nature. When an author would do the work and explain it to the reader, he might fear that the only result would be a severe headache - for him and the reader. Instead, the shelves are filled with the traditional neo-Darwinian body of theory, and the reader sees them explained as if they were the sole truth about matters of gene substitutions, advantageous mutations, and adaptive processes. All these phenomena are all regarded as a consequence of one force: Natural Selection, which inevitably will improve the fitness of the organism or the population.

In reality, this picture is too naive, at least in two points: The first observation from the neutral theory due to Motoo Kimura says that a population with many different genetic configurations among its members will be much more susceptible to *random genetic drift* (followed by extinction or fixation) rather than to adaptive mutations: In the next chapter in section 2.6 on page 23 we will define random genetic drift and show that even slightly deleterious mutations can have a finite probability of becoming fixed in the population. The other major point of the *neutral theory* is the problem of *polymorphisms*, that is, the fact that a typical population of one species consists of many individuals with different genetic makeups.

This fact is not solely explainable by some kind of "balancing" or "stabilizing" selection (this concept will be explained in appendix B), as maintained by the neo-Darwinian (or selectionist) point of view. Polymorphism is on the contrary a purely stochastic effect, representing a transient ensemble of an ongoing dynamical process of mutational input and a concomitant random extinction or fixation of genes. Rather than describing gene substitution and polymorphisms by two different mechanisms (successful mutations and balancing selection), the insights from the neutral theory explains that they are but two facets of the same phenomenon - namely neutrality (Kimura and Otha, 1971).

The consequences for the overall perception of evolution and evolutionary theory are profound: the deterministic slant in the neo-Darwinian theory is replaced by a general recognition that chance plays a major part in evolution, and that random genetic drift might be equally important as is natural selection for the continuous evolution of new and diversified species.

There are other important consequences of neutrality: one is the fact that neutrality can help populations to become more robust against environmental changes, an idea already mentioned by Kimura (Kimura, 1983), although the argument was without any support form his equations. *Neutral mutations* in this context are definied as mutations in genes which do not change the fitness of the organims.

In effect, neutral mutations have the tendency to create a broad ensemble of equally fit mutants in the population, which in turn can function as an effective buffer against changes in living conditions, because there is a greater probability that some of the mutants are immune to such changes. Or even better: if one or several more advantageous mutants are found, the population amplifies quickly around the best mutant by normal adaptive selection.

In essence, this argument relies on the idea that neutrality *induces quicker and better evolutionary optimization* so that it can *help* populations to attain higher fitnesses. In this sense, neutrality and random genetic drift might be a much more edifying principle for evolution than expected. It is precisely this idea, which will be the central theme of the following two chapters. We will develop a new model for neutral evolution, which is able to substantiate these ideas in terms of simulations and explicit relations between fitness and neutrality (and other parameters) of an evolving population.

A natural questions would be: The idea of selective neutrality has a long history, so why hasn't this been done before? The primary reason is that, despite many decades of hard work, we still have a rather poor understanding of the way in which genomic sequences map onto molecular structures, and hence onto a fitness measure (the "fitness" of a molecule is seen as a function of its "form": its ability to perform certain functions in the working organism in order to reproduce successfully). This is not only the case for molecules, but also for entire organisms, where the fundamental problem is that of calculating the mapping from a genotype to a phenotype. Although most of the new mutations are mainly deleterious, most changes in an organisms genome have no immediate effect on its reproductive success, because the mutations are quickly removed from the population. The bad mutations contribute neither to the rate of substitution nor to the amount of polymorphisms within

populations. But the problem remains: how to assign fitness values to the different mutants [2]?

Our solution is that of abstraction: We will construct a mathematical model of neutral evolution which sidesteps the problem of incorporating the chemistry of real molecules in our calculations, so that the properties of the system can be investigated more quickly and in better detail than it is possible with, for example, RNA structure calculations. We will use a variant of the NK model proposed by Kauffman (Kauffman and Johnsen, 1991; Kauffman, 1993), which has the pleasant property of generating infinite complex combinations of fitness values, so that we can expect it to mimic real biological systems in some regimes of the model parameters.

In the model, structures (our phenotypes) appear as contiguous sets or "neutral networks" of sequences possessing the same fitness. We will analyze the properties of the generated fitness landscape, and show that for appropriate choices of the parameters, this model can be used to mimic real biological systems such as RNA's and proteins. We also find that evolving populations can reach beyond low local fitness peaks by utilizing the properties of these percolating nets. All this is done in chapter 3.

Before this however, chapter 2 will be a short repetition of the standard catechism in neo-Darwinian theory, that is, a brief discussion of some of the most fundamental mathematical aspects of natural selection. It is divided in two major categories, namely the deterministic approach to population dynamics through differential equations, and the second (more realistic) stochastic approach of random genetic drift and neutrality. Also two models representing more recent development of these two complementary discoveries, are shortly discussed: The quasi species model by Manfred Eigen and the NK model due to Stuart Kauffman.

[2] in the literature, genetic mutants are called *alleles* when one talks about organisms; nucleotides when talking about RNA or DNA, and amino acids in the case of proteins.

2 Law and Chance in Molecular Evolution

2.1 Natural Selection

Natural selection is defined as the *differential reproduction of structurally distinct individuals within a population.* The individuals might be species, organisms, molecules, clays or other agents in the universe (or even the universes themselves), and the differences between the individuals might be caused by variable genetical, conformational or some other (variable) structural units - as long as they carry the information needed for the individual to survive. If the individuals of a population would not differ from one another in such traits, they would not be subject to natural selection.

Reproduction is the ability of these information carriers to make copies of themselves by creating some kind of auto-catalytic growth. In molecular evolution one normally talks about "replication" instead of reproduction, since the information contained in these molecules (such as RNA or viruses) is directly used for the synthesis of a copy of themselves.

2.2 Fitness

The "differential reproduction", can then be quantified by a number w, called "fitness", which tells us something about the ability of an individual to survive and reproduce. Of course, this quantification can only be measured as

a *relative* increase or decrease of abundance of this individual compared to other individuals. It can never be an absolute measure of evolutionary success. Nevertheless, it is a nice trick, because it makes it possible for us to formulate simple theories, which are useful for the understanding of the dynamical change in genetic configurations within populations undergoing natural selection.

2.3 Selection equations

We can look at the simplest class of such deterministic models derivated from these assumptions: We assume that the fitness of an individual i with n_i copies can be expressed by a single number w_i. The growth of this individual in time can then be expressed by the auto-catalytic differential equation

$$\frac{dn_i(t)}{dt} = w_i n_i(t) \tag{2.1}$$

so that the *Malthusian growth rate* w_i can be viewed equivalent to the fitness of the individual (as a first order approximation). But since mutations have created many different information carriers, we have a population of such individuals with different fitnesses. What is the relative success of individual i compared to all the others? First we define the relative population variables[1]

$$x_i(t) = \frac{n_i(t)}{\sum_k n_k(t)}; \quad 0 \le x_i \le 1; \quad \sum_k x_k(t) \equiv 1$$

so that the change of x_i in time is determined by

$$\begin{aligned} \frac{dx_i}{dt} &= \frac{d\left[\frac{n_i}{\sum n_k}\right]}{dt} = \frac{\sum n_k \left(\frac{dn_i}{dt}\right) - n_i \left(\frac{d\sum n_k}{dt}\right)}{\left(\sum n_k\right)^2} \\ &= \frac{n_i}{\sum n_k} w_i - \frac{n_i}{\left(\sum n_k\right)^2} \sum w_k n_k \\ &= x_i \left[w_i - \sum_k w_k x_k \right] \end{aligned} \tag{2.2}$$

[1] see also appendix C

where the sum within the brackets is the (time dependent) average fitness of the population

$$\bar{w}(t) = \sum_{k}^{N} w_k x_k(t) \tag{2.3}$$

The equation 2.2 is called the *first order continuous time selection equation*, or first order *replicator equation* and is equivalent to the *Lotka-Volterra equation* used in ecological population dynamics (see part II).

The *second order continuous time selection equation* or *second order replicator equation* is analogous to equation 2.2, except that the fitness w_i of an individual is defined as the linear addition of all contributions in the population vector with which the individual interacts:

$$w_i(\mathbf{x}) = \sum_{j}^{N} a_{ij} x_j \tag{2.4}$$

where the a_{ij}'s are the elements of an interaction matrix which defines the strength and type of interactions among the individuals (remember that the "individuals" not need to be organisms; they could also be different alleles on a chromosomal locus or other enities subject to natural selection).

In many cases this definition of an individual fitness is seemingly a more reasonable assumption, since the fitness of an individual not only may be dependent on its own growth rate, but also on the gain of interaction with other individuals in the population (for instance sexually reproducing organisms have diploid genomes with two interacting alleles at each locus. The fitness is then a function of the kind of interaction of the two alleles at that chromosomal locus, see appendix B for examples of that.). In this case, the second order replicator equation becomes

$$\frac{dx_i}{dt} = x_i \left(\sum_{j}^{N} a_{ij} x_j - \sum_{ij}^{N} a_{ij} x_i x_j \right) \tag{2.5}$$

The replicator equation 2.5 has also been used to model prebiotic evolution of primitive RNA replicators. The theory, known as the hypercycle equation, was due to Manfred Eigen and Peter Schuster (Eigen and Schuster, 1997; Eigen et al., 1981). The hypercycle equation is a special variant of 2.5, where the indices $i = 1, 2, ..., n$ are counted *modulus* n, so that $x_0 = x_n$. The x_i's are

RNA fragments that catalyse each other's replication in a closed feedback loop in order to accumulate molecular information.

The two selection equations 2.2 and 2.5 go back to the pioneers of population genetics, Robert A. Fisher (Fisher, 1930), J. B. S. Haldane (Haldane, 1932) and Sewall Wright (Wright, 1967; Provine, 1986). They used their discrete counterparts as the fundamental equations for the gene frequencies in a population, so that, when the selective differences were small, and the generations were overlapping, the continuous versions could be used as a simplifying approximation. In appendix B these rather general equations are exemplified for the case of one locus and two alleles, leading to the important concept of *Hardy-Weinberg equilibrium*. The different modes of selection; codominance, overdominance (that is "balancing" seletion) and complete dominance are also discussed there.

This kind of "population thinking" has been the backbone of the neo-Darwinian synthesis, and it has been extremely successful. It has cleared the understanding of the conditions under which slightly advantageous mutant alleles at some locus in the genetic makeup of an organism might invade a population. The mathematical analysis has concerned the effects of population size, homozygotic or heterozygotic selection, linkage of genes, sex, effects of recombination, etc.. The success has led to the general belief that the fundamental unit of selection are the genes, and that the resulting organisms, the phenotypes, just are necessary vehicles for the genetic information to be transmitted through time.

2.4 The fundamental theorem

In addition, Robert A. Fisher could show (Fisher, 1930) that the mean fitness of a population subject to natural selection always is increasing - so that the population always is "in progress" so to speak. From equation 2.5 we can see[2]

[2]This is maybe not so easy to "see" anyway, but in appendix C there is another formulation of, how to arrive at a general replicator equation: the rate of increase of one species is given by the difference in its own fitness and the average fitness of the whole population. Therefore is the right hand side sum within the brackets of both the equations 2.2 and 2.5 always representing the average fitness of the population.

that the mean fitness $\bar{w}(t)$ at time t is

$$\bar{w}(t) = \sum_{ij}^{N} a_{ij} x_i x_j \tag{2.6}$$

Differentiating the mean fitness with respect to time yields

$$\frac{d\bar{w}}{dt} = \frac{d}{dt}\left(\sum_{ij}^{N} a_{ij} x_i x_j\right) = \sum a_{ij} \frac{dx_i}{dt} x_j + \sum a_{ij} x_i \frac{dx_j}{dt} \tag{2.7}$$

We can simplify the above equation by introducing a matrix notation, so that $\sum_j a_{ij} x_j = (\mathbf{Mx})_i$ and $\sum_{ij} a_{ij} x_i x_j = \mathbf{x} \cdot \mathbf{Mx}$. If we now assume that $a_{ij} = a_{ji}$ (which is physically resonable most of the time for alleles at a diploid chromosomal locus: there is no preferred position in this case. See also appendix B), then both sums in equation 2.7 are equal and, following (Hofbauer and Sigmund, 1988), we obtain

$$\begin{aligned} \frac{1}{2}\frac{d\bar{w}}{dt} &= \sum_i \frac{dx_i}{dt}(\mathbf{Mx})_i \\ &= \sum_i x_i \left[(\mathbf{Mx})_i - \mathbf{x} \cdot \mathbf{Mx}\right](\mathbf{Mx})_i \\ &= \sum_i x_i (\mathbf{Mx})_i^2 - \left(\sum_i x_i (\mathbf{Mx})_i\right)^2 \\ &= \sum_i x_i \left[(\mathbf{Mx})_i - \mathbf{x} \cdot \mathbf{Mx}\right]^2 \geq 0 \end{aligned} \tag{2.8}$$

since the last equation never can be negative. But looking at the terms in 2.8, we see they represent the variance of the fitness $w_i = \sum_j a_{ij} x_j$ of the individual i. So, Fisher could not only conclude that the average fitness of a population increases for each generation, but also that the increase is directly proportional to the additive genetic variance in fitness. This he termed "the fundamental theorem of natural selection"[3].

[3]In the history of science, only the first fact, namely the increase in average fitness (the everlasting progress of humanity...) has become canonical, even though it rests on the second fact, namely the increase in *variation*, meaning that many species in the population very well might have a constant, or even decreasing fitness. It is only the *increase* in the number of "races" which is evolutionary important. However, racist arguments have always found

Although the different versions of equations 2.2 and 2.5 have had a profound impact on the perception of evolution as a primarily deterministic process, their derivations lie on a shaky ground (Hofbauer and Sigmund, 1998). First of all, the ansatz'es in the equations 2.1 and 2.4 would imply exponential growth for the whole population. Second, for diploid organisms the second order replicator equation relies on the concept of Hardy-Weinberg equilibrium, which in turn relies on the implicit assumption of random sampling, equal birth and death rates and on an infinite population of the genes. All of these assumptions are fairly idealized (if not wrong) and therefore many population geneticists refuse to use eq. 2.5 in their work. Fourth, the fundamental theorem relies on constant fitness values of the genes - which additionally have to be independent of their frequency. For many traits such constancy does not hold (Smith and Szathmáry, 1995).

2.5 Selection-mutation equations

Yet another argument against the selection equations is that there until now not has been introduced any kind of mutational input. In the classical population genetics, mutations (such as point mutations, insertion, deletions or recombinations) were regarded as subordinate to the force of selection and the evolution of populations. Although mutations were seen necessary for the continuous variation upon which the evolutionary "progress" could operate, their role was that of "noise" with only inferior and unimportant consequences for the selection of species.

However, today we know that mutation is a property which is regulated to the smallest possible detail, and that it is of fundamental importance for the structure and evolution of populations. Sequence analysis of viral genomes has shown that the target of selection is not a singular wild type sequence, but a cloud of equally fit mutants. These mutants have shown to be so abundant that even normal diluted test tube fractions of viral genomes shows that their wild type is below experimental detectability (Domingo et al., 1978). An average RNA virus has a genome of size 10^4. Still, a population solely consisting

awkward formulations and justification for the superiority of the white man, and even Fisher himself, in his classical book *The genetical theory of natural selection* (Fisher, 1930) (which is the all-important bible in population genetics) used a modified racist scheme for the improvement of the Anglo-Saxon pale-face. For an illuminating essay on this subject, see *The Smoking Gun of Eugenics* by Stephen Jay Gould in (Gould, 1995).

of (equally abundant) mutants of this RNA virus would yield the wild type sequence at each position with an accuracy of 0.9999. This is illustrated in figure 2.1 on page 23. The lesson is that mutants are produced not only as error copies of the wild type, but also through self-replication. And selection will bias the distribution of mutants accordingly. The "unit of selection" is therefore not the individual, but the whole cloud of mutants moving erratically on a complex fitness landscape.

The first experiments in evolution were carried out by Spiegelman in the late 1960's (Mills et al., 1967; Mills et al., 1973). The replication system of the RNA bacteriophage Qβ was isolated, whereafter large concentrations of nucleotide triphosphates were mixed with the 4200 bases long single stranded Qβ genome. The newly copied RNA strands were equally infectious as the old ones, but when this process was repeated many times, the RNA strands lost their infectious property in favor of growing quicker (and their length became drastically reduced). This happened because of the artificial selective pressure on the RNA strands through the seriel procedure.

But the Qβ replicase showed also to be able to produce totally *new* strands of RNA without the help of the genome (Sumper and Luce, 1975; Biebricher et al., 1981a; Biebricher et al., 1981b; Biebricher, 1986). This evolution *in novo* could happen because the enzyme has a strong affinity towards certain tetranucleotides, which randomly gathered on the surface of the enzyme in the right order. Thus, oligomerization was possible.

2.5.1 The quasi species

One deterministic implementation of these findings is the theory of the *Molecular quasi species* due to Manfred Eigen and co-workers (Eigen et al., 1988; Eigen et al., 1989), which primarily is formulated in the language of RNA viruses such as Qβ replicase. The starting point is the replicator equation 2.2 with the additional assumption that the production of RNA virus x_i also depends on the sum of frequencies of erroneous copying from (and to) all the other viruses

$$\frac{dx_i}{dt} = x_i\left[w_{ii} - \bar{w}(t)\right] + \sum_{k\neq i} w_{ik}x_k \tag{2.9}$$

so that the w_{ik} are the off-diagonal elements of the mutation matrix W and the w_{ii} are the diagonal elements representing perfect self-replication.

Solving this set of coupled differential equations is quite involved. But one simple example is still very illuminating: Assume that the wild type is ten times fitter than all the other mutants, which in turn are selectively neutral compared to each other. Then, for an increasing mutation probability, computer simulations (Swetina and Schuster, 1982) show that the population encounters an *error threshold* beyond which the cloud of mutants looses the information about their wild type. This is of course an irrecoverable situation for the whole viral quasi species, because the wild type population will drop off to zero and the virus will no longer be able to infect its host and survive. The situation is comparable with a phase transition, a melting point through an accumulation of errors. But just close below the error threshold the virus has the best conditions for evolution: Here the wild type is stable, and at the same time there is a maximal number of mutants which can adapt quickly to a change in its environment. Experiments with $Q\beta$ replicase (Mills et al., 1967; Biebricher, 1986) show that natural mutants indeed operate just below the error threshold. In many cases selection may even favor a mutant with a lower fitness than the best adapted one, if the surrounding mutants have a comparably higher selection value than it is the case for the best adapted individual (Eigen, 1992). This interesting behavior is of course an important complement to the classical picture of evolution and a strong modification of the neo-Darwinian dictum of "survival of the fittest".

Although the theory of the molecular quasi species represents an important step towards a greater appreciation of chance effects and neutrality in molecular evolution, it is still using the classical apparatus of deterministic mass action kinetics. It is however questionable, and this question has even been raised by Manfred Eigen himself (Eigen et al., 1989), how far conventional chemical kinetics can be used for problems in evolutionary theory. The main doubt is that while normal chemical reaction kinetics only involves a small number of different molecules each of them present in a very large amount, the situation for molecular evolution is quite the opposite: the number of different RNA mutant sequences (or protein mutant sequences) is enormous, while the amount of each of them often is only one or just a few[4]. The fundamental assumption of the use of continuous differential equations is that of an infinitesimal difference in concentration change in time, an assumption only usable when there is a very large number of such identical mutants (preferably somewhere around

[4]Actually, the number of possible mutants in an ordinary sequence of RNA or protein is far greater than the number of molecules in the whole universe (see appendix A).

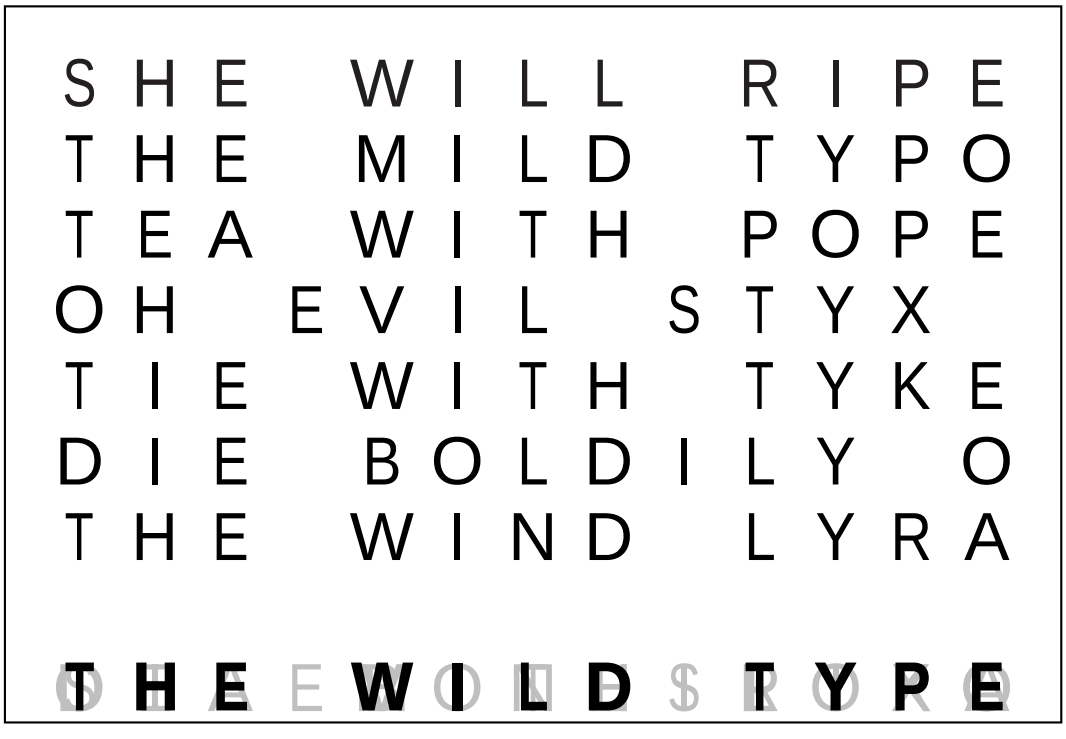

Figure 2.1: *The definition of a wild type is the consensus sequence of a mutant distribution (one could also say that the wild type is the* average sequence *of a population of many different (polymorphic) sequences, but when there only are discrete symbols, the average is not allways well-defined). So, the wild type represents the center of the quasi species, even though it might not exist itself. Inspired after (Eigen, 1992).*

Avogadros number). This is seldomly the case.

When this is not the case, it might be better to model such systems by discrete cellular automata-type of models, where each individual is assigned its fitness value and is traced separately through the dynamics. This of course implies an enormous amount of computational capacity, but modern computers make it possible.

2.6 Random Genetic Drift

One simple example of the effect of randomness in finite populations is that of *random genetic drift*. It was Sewall Wright (Wright, 1932) who first devel-

oped the concept of random genetic drift in small separated populations as a source of variability upon which normal selection can act. Later on Kimura and Otha attributed a much higher significance to random genetic drift, even for the mechanism of evolution, and for the fundamental aspects of genetic architecture in finite populations. Kimura's conclusion (Kimura, 1968) was that the concept of neutrality and random drift was a real alternative to the prevailing selectionist point of view at that time.

Imagine a situation in which all members of a (finite) population of N diploid organisms have the same fitness so that they are selectively neutral (as in the case of the quasi species model without the wild type). At any given locus in the genome of the organisms there are two alleles A_1 and A_2 with the frequency p and $(1-p)$ respectively. For simplicity, imagine further that there is only one locus in each individual, so that there are $2N$ genes in total. The probability P_{n_1} that the sample contains exactly n_1 genes of type A_1 is given by the binomial probability distribution

$$P_{n_1} = \binom{2N}{n_1} p^{n_1} (1-p)^{2N-n_1} \tag{2.10}$$

If there initially is an equal amount of the two alleles ($p = \frac{1}{2}$ and $n_1 = n_2$), then, using eq. 2.10, there is a high probability that the allele frequency has changed in the next generation (for $N = 5$, the probability to have the same frequency of alleles in the next generation is $P_5 = 0.25$). For an increasing number of generations, the probability to have both alleles coexisting in the population, decreases therefore rapidly, and once the frequency of an alleles reaches 0 or 1, it is either *extinct* or *fixed*. For long enough time, this is inevitable. Figure 2.2 on the facing page shows that for an increasing population, this fate can be postponed, but never escaped.

2.6.1 Fixation

It is now interesting to ask how important this effect is - compared to the force of selection. Lets assume that the relative fitnesses of the genotypes A_1A_1, A_1A_2 and A_2A_2 are $1, 1+s$ and $1+2s$ respectively. The number $s > 0$ represents a selective advantage of allele A_2 over A_1 (the co-dominant mode of selection, see appendix B). Classical Darwinian theory would expect that the advantageous allele A_2 overtakes the population inevitably. But Kimura

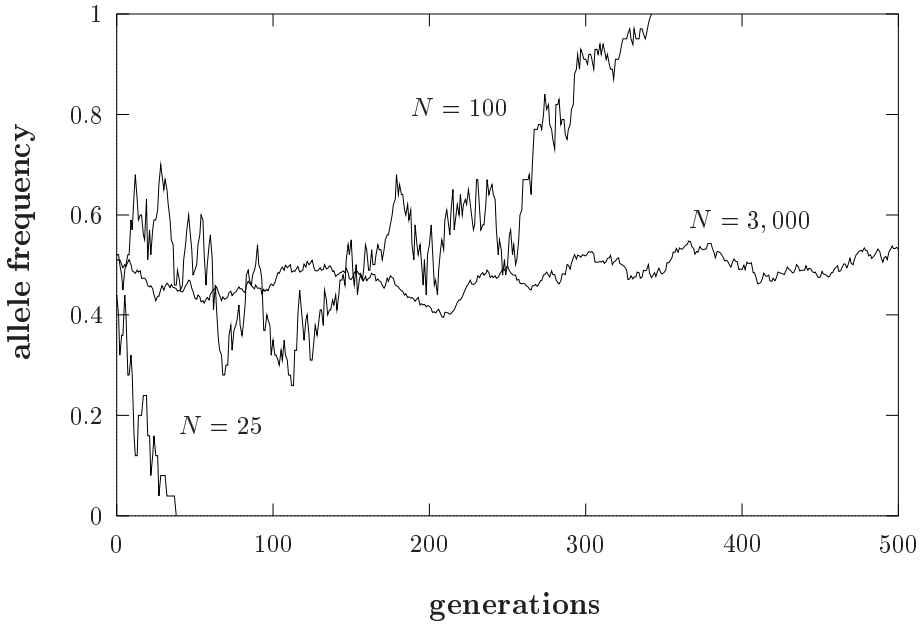

Figure 2.2: *Changes in allele frequencies by random genetic drift in three populations with different sizes. The smallest population reached fixation of one alleles - and extinction of the other (shown) - after 38 generations. The largest population was not fixed yet after 500 generations.*

(Kimura, 1962) has shown that the probability of fixation is "only"

$$P = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}} \tag{2.11}$$

where N is the effective population size and p the initial frequency of allele A_2. When the selective advantage $s \to 0$, we can set $e^{-x} \approx 1 - x$ and the equation reduces to $P \simeq p$. This means that for neutral alleles, the fixation probability equals its frequency in the population. Since random genetic drift in non-directional, this result it understandable, because a low frequency of a neutral gene gives a lower probability of fixation.

But now to the interesting part: Let us now assume that a new, selectively advantageous, mutant has appeared in the population. Initially, it has a frequency of $p = \frac{1}{2N}$. Inserting this in eq. 2.11, we obtain

$$P = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}$$

If we assume that s is small, this becomes

$$P = \frac{2s}{1 - e^{-4Ns}}$$

so that, for a large population and $s > 0$, we can approximate

$$P \approx 2s \tag{2.12}$$

This is a highly significant result (Li and Graur, 1991). For a selective advantage of $s = 0.01$ of one emerging mutant in a population, its probability to survive (and push the population towards higher fitness) is only 2%. In fact, 98% *of all advantageous mutations (with a relative selective advantage of* 1%*) will be lost by chance.* Even slightly deleterious mutants have a finite probability of becoming fixed in the population (Otha, 1972).

In conclusion, random genetic drift has a very strong impact on the evolution of populations, so strong that it in many cases makes natural selection helplessly insignificant. For well adapted highly fit species, the probability to find a new advantageous mutation is small. In such a case, one might expect that if the species population finds a good mutant anyway, this mutant only has a small selective advantage compared to the wild type. As we have seen in eq. 2.12, this does not necessarily do any good. Therefore, highly adapted populations with a high fitness need to employ the neutral mutants much more, in order to search and find a *really good mutant* somewhere in the fitness landscape, so that the probability of fixation of such a "really good mutant" is maximized. This will be the main subject of the next chapter.

2.7 The fitness landscape

The metaphor of a *fitness landscape* has now been used several times without any further explanations. Normal biologists would expect a fitness landscape to be a complex parameter space where all components defining the fitness of an organisms, such as fertility, fecundity, birth and death rates, etc., can be varied upon. In the present case, however, the term "fitness landscape" is used in a much more restricted way, as first proposed by (Wright, 1967; Wright, 1982).

As an example, we can look at the evolution of RNA's. All RNA molecules consist of the four nucleotides adenine, uracil, guanine and cytosine, which constitutes the size of an alphabet A with four letters. The number of combinations in which these four components can align to a molecule, often of

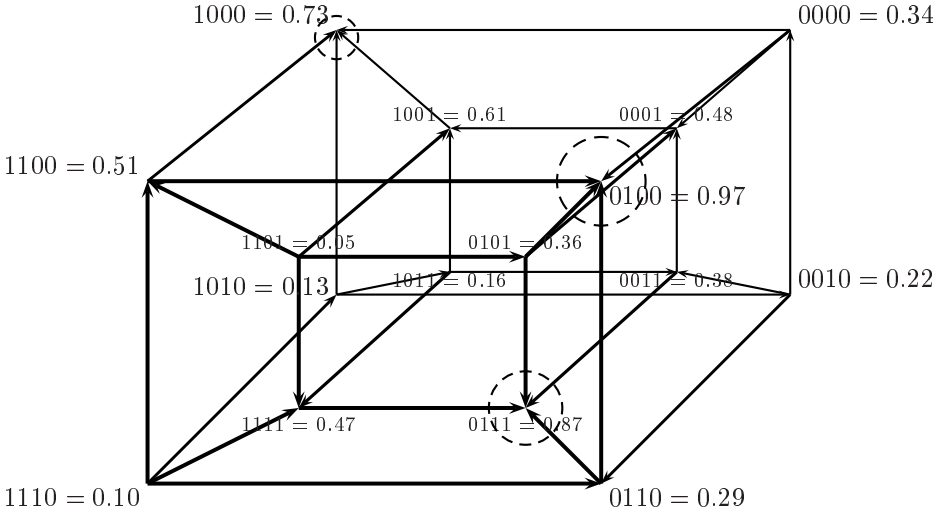

Figure 2.3: *Diagram of a sequence space for a string of length $N = 4$ an alphabet $A = 2$. Each combination of 0's and 1's is assigned a random fitness value, so that a population moving adaptively in such a space, either can be trapped in a local fitness peak (indicated with circles), or find the global (most fit) peak - in this case it is positioned in the front upper right. Modified after (Kauffman, 1993).*

considerable length, such as $N = 1,000$, is enormous. This set of all possible RNA molecules of length N constitutes an ensemble, called the *sequence space*, having a size of A^N. Each point in this high-dimensional sequence space represents one RNA molecule. At each site of one of the molecules there could be three other nucleotides, meaning that this molecule has $(A-1)N$ possible one-point mutants. The distance measured in point-mutations from the original sequence is called the *Hamming distance*, and is explained further in appendix A. Each of the mutants has assigned a specific random fitness value, so that a population represented by a unique wild type with a unique place in the sequence space can be regarded to make an adaptive walk in fitness landscape searching for the optimal combination of nucleotides - the peaks in the landscape.

A particular simple example is shown in figure 2.3. In this case the alphabet only consists of two possibilities 0 or 1, which could represent purines and pyrimidines, respectively. The topology of this sequence space is very different

from the alps. It is a four dimensional boolean hypercube, where each point is connected to four neighbors, or mutants, each with a different fitness value.

2.8 The NK model

The definition of NK-systems was initially used as a simplified model for the genetic regulatory system acting in cell differentiation (Kauffman, 1969). But the NK model can also be viewed as a generalization of the spin glass models used in statistical physics (Fischer and Hertz, 1991). It was also thought that the generality of the model could reflect evolution of a species population, where random mutations of individual genes make the species walk on their underlying random fitness landscape (Kauffman and Johnsen, 1991; Kauffman, 1993).

In this setup, an entire species of an unspecified organism is represented by a string of N genes (obviously a very simple one: haploid, with only a single copy of chromosomes) with a given configuration. Since evolutionary selection mainly works on a spectrum of many mutants within a population (so that diversity is maintained) this assumption of representing a whole species in just one sequence seems at a first glance quite unreasonable. But (Gillespie, 1984) has shown that if selection proceeds much faster than mutations occur, then one may represent a population by a dominant type, and ignore diversity. This dominant type, in our case the string of N genes, which is both a genotype and a phenotype, will then perform an adaptive walk on the fitness landscape, whose ruggedness is dependent on the way we choose the fitnesses of the individual genes.

The definition of the NK model is as follows: Consider a sequence of N loci, which could represent genes (alleles) in an organism, the number of nucleotides within a RNA molecule, or the number of positions in the primary sequence of proteins. Each locus i can be occupied by different sorts of such entities. This constitutes an alphabet of size A. In the case of RNA, A becomes the four possible nucleotides {A,C,G,U}, and in the case of proteins, A could be the 20 amino acids that can occupy each position in a protein sequence.

Also, each locus i interacts with K other loci in the sequence (called *epistatic interactions*)- they could be the neighbors or else-wise chosen, see figure 2.4 on the facing page. In RNA's, the bases are mostly only interacting with one other base, forming a Watson-Crick base pair, but they can also interact with other bases in order to generate a tertiary structure. Proteins have normally more complex types of interactions, so that K may be higher than

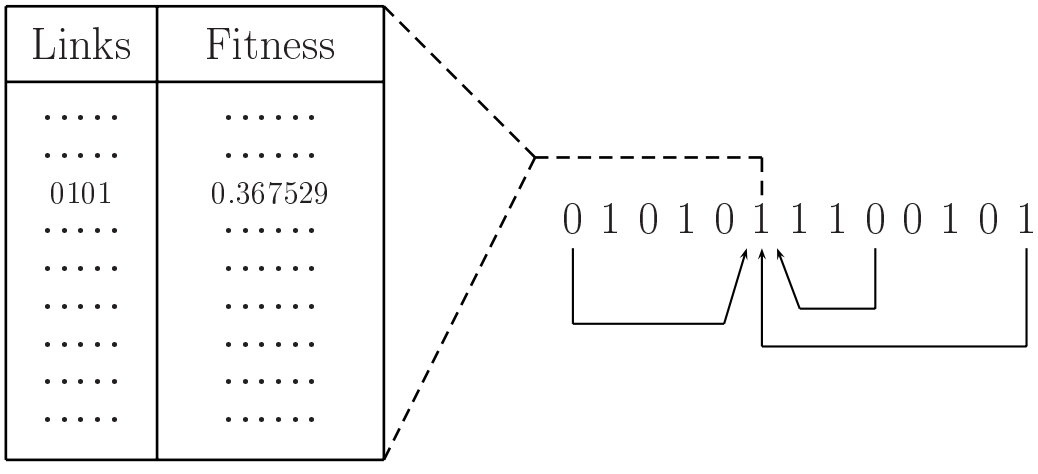

Figure 2.4: *Schematical representation of the assignment of random fitness values to each of the A^{K+1} possible configurations of links for one gene. For each place in the row of N genes, there is one unique lookup table like the one shown for the gene at position 6. Here $A = 2, K = 3$ and $N = 13$. Modified after (Barnett, 1997).*

unity.

Thus, the entire sequence can exist in A^N different configurations, and we call this ensemble the *sequence space*. In order to find a definite fitness W of the whole sequence, each locus i contributes with the fitness w_i, which of course is dependent on the K other loci with which it interacts. This means that each locus can have A^{K+1} different values of w_i. These contribute additively to W, but with how much? Since it is a hopeless enterprise to know the exact strength of epsitatic interactions - how good they are and how bad - we can as well chooses the values in the column "FITNESS" of figure 2.4 at random in the interval $0 \leq w_i < 1$, so that

$$W = \frac{1}{N} \sum_i w_i \tag{2.13}$$

It is now possible to picture the dynamics of the system: For $K = 0$, the species population will climb to the global maximum in the fitness landscape by successive point mutations in the genes which maximize w_i and thus W. For larger K though, the internal constraints on w_i, due to the K partners, create unbridgeable valleys so that the species might get trapped on a local peak. For $K = N - 1$ the fitness landscape is maximally rugged with uncorrelated

fitness values at each point in the landscape (as in fig. 2.3 on page 27). In this case the traditional NK system becomes equivalent with Derrida's random energy model (Derrida, 1981; Gardner and Derrida, 1989) which was used to analyze an A-state Potts model. We will return to this case later on in part III, when we investigate the percolation properties in such an uncorrelated fitness landscape.

2.9 Summary and Conclusion

In order to appreciate the coming chapter better we have made a short introduction of some of the most fundamental aspects of population biology and described some of the most well-known mathematical models in evolutionary theory:

After a verbal definition of natural selection and fitness, we have quantified these ideas in different versions of a replicator equation, which are the main tools for a modeling approach to natural selection, and fundamental for the fundamental theorem. The introduction of mutation in the selection equations led to the formulation of the *quasi species* which showed that the unit of selection not so much is the individual gene, but the *network of mutants* created through the mutational input, and therefore one can say that this emergent network of quasi species is selected as a whole.

This, in turn, led to the investigation of neutral mutations and the important force of random drift, which in many cases showed to be much stronger than natural selection itself. Finally, we looked at the topology of a possible fitness landscape upon which molecular species might evolve, and also on the NK model as a possible modeling candidate for the evolution of populations. In the next chapter we will extend and combine these efforts into a new model.

3 Effects of selective neutrality

3.1 Introduction

The idea of Kimura (Kimura, 1955; Kimura, 1983) was that if a mutation in a gene does not change the functionality, eg. its *functional form*, and leaves the viability of itself and its host unchanged, the mutation can be regarded as *neutral*. In the presence of a large number of different genetic architectures, Darwinian selection is more or less indifferent, as long as the resultant phenotypes are good for the organisms survival and reproduction. In fact, there are believed to be many such neutral mutations, which get pumped constantly into the gene pool, where after random genetic drift purges the unlucky alleles from it. Kimuras idea was that this highly dynamical process, this constant flux of alleles, causes neutral mutations to accumulate. Compared to that, the forms - the three-dimensional structures - remain remarkably stable over millions of years.

It has however been difficult to model this effect in a simple way, mainly because it is not obvious how to assign fitness values to the overwhelming number of possible mutants. How should we know, which mutation is neutral and which not? First the *phenotype*, that is, the body and legs of an animal, or the three-dimensional conformation of a RNA-molecule, and their reproductive success, can give us a clue about a definite fitness value.

Thus, the fitness expresses itself as a function of the phenotypes, but when

neutrality expresses itself in the *absence* of a physiological effect, how can we distinguish neutrality from nothing? Somehow, this problem of telling what is the effect of what, and the lack of a unique scheme for fitness assignment, has been the major obstacle for the mathematical modeling of neutrality.

The pondering about this problem in the last decades or two, mainly by Peter Schuster and his group (Schuster, 1994; Schuster et al., 1994), has led to the general recognition that we first need to understand the mapping procedure from a genotype towards a structure. The process leading from a genetic configuration to a functional form has to be investigated carefully before we can tell something about the evolutionary effects of neutrality.

There have been several approaches to this problem. One simple case in which neutral evolution has been investigated in some detail is that of RNA structures (Schuster et al., 1994; Grüner et al., 1996a; Grüner et al., 1996b; Huynen et al., 1996; Reidys et al., 1997). However, the calculations are so far limited to the secondary structures of RNA, and even these are relying on many restrictive approximations, so that the studies have to be taken more as a qualitative guide to the behavior of systems undergoing neutral evolution than as an accurate representation of the real world.

Also, simple models for protein folding (Li et al., 1996) and experiments for protein evolution (Dean, 1998; Wilks et al., 1988) have show that selective neutrality and the problem of fitness assignment in genotype-structure mappings are increasingly important concepts, both in evolutionary theory and in biotechnology, and that there is a strong need for a general and simple mathematical model which is capable to explain such effects on a more generic basis.

This does not mean that there not has been done anything to develop mathematical models of neutral evolution. On the contrary, there have been developed several genetic algorithms (Prügel-Bennett and Sharpio, 1994; Mitchell, 1996) which give some insight into the type of effects one may expect neutral evolution to produce. Especially the papers by (van Nimwegen et al., 1997a; van Nimwegen et al., 1997b) have developed some promising tools to investigate "metastability" and the "epochal" nature of evolution in the presence of selective neutrality. In these papers the neutrality is investigated by a reduction in dimensionality, where the "genes" need to be aligned in blocks, in order to obtain neutral regions (although there are no *percolating networks*), so that it is possible analytically to calculate important properties such as average fitnesses, fitness fluctuations, population distributions and the length of the epochs. But still, these investigations are not sufficiently general in order to assure that they apply to other biological systems as well.

In the next section we will develop a simple model system, which we believe is sufficiently general to mimic most biological systems under the influence of selective neutrality.

3.2 Model for neutral evolution

The model presented here and in (Newman and Engelhardt, 1998) is a generalization of the NK model from section 2.8 on page 28.

Again, we consider a sequence of N loci, representing the building blocks for a genetic code. The alphabet A of the letters at each position is variable - depending on what system we want to investigate ($A = 4$ for RNA's and $A = 20$ for proteins). And again, like in section 2.8, each locus i interacts with K other loci in the sequence, so that there to each locus belongs a "lookup table" like in figure 2.4 on page 29 in order to find the fitness value w_i.

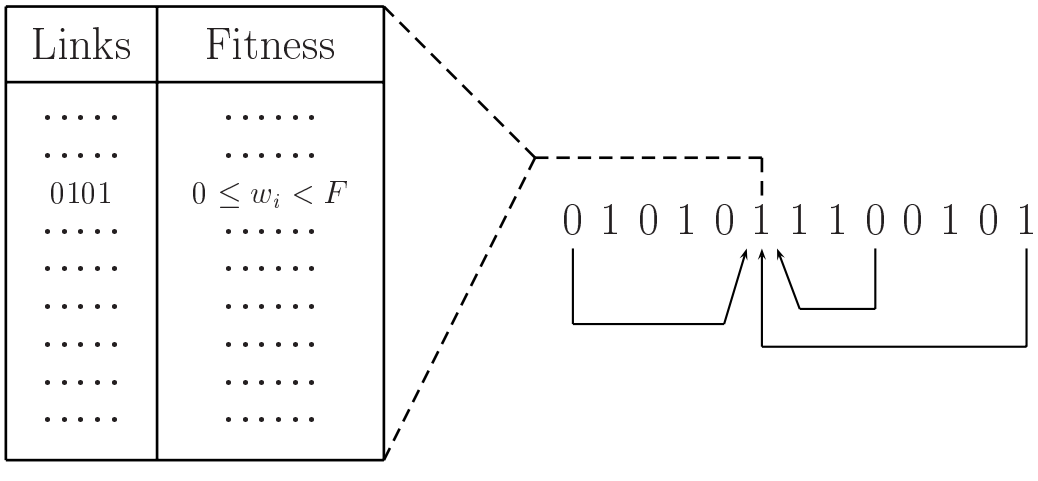

Figure 3.1: *A repetition of figure 2.4 on page 29 with the modification that w_i is now an* integer value, *so that it suddenly becomes much more probable that two different configurations have the same w_i and thus the same fitness W.*

However, in the case of the present model for neutral evolution, we choose the w_i's to be *integers* in the range $0 \leq w_i < F$ (so that there in the column "Fitness" of figure 2.4 are integers from zero to F, see figure 3.1. Thus, if $F = 2$ for example, each contribution w_i is either zero or one. The fitness of

the entire sequence is therefore

$$W = \frac{1}{N(F-1)} \sum_i w_i \tag{3.1}$$

so that the fitness of all sequences fall in the range from zero to one, and there are $NF - N + 1$ possible fitness values in this range.

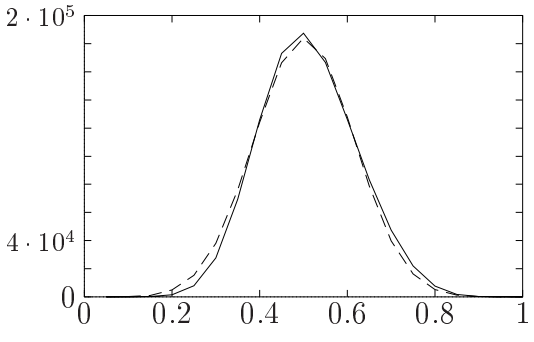

Figure 3.2: *The number of sequences as a function of fitness. $K = 1$ (full line), and for $K = N - 1$ (dashed line). $N = 20, A = 2$ and $F = 2$.*

For $F \to \infty$ the system degenerates into the normal NK model, and the NK model is therefore a special case of this model. The parameter F increases neutrality when it is decreased. The parameter K increases the ruggedness of the fitness landscape when it is increased.

In this formulation we have a model with a tunable degree of neutrality and a unique assignment of a fitness value to each possible sequence in the enormous ensemble of sequences. Since we have defined the total fitness of the genotype W to be within the range $0 \leq W < 1$, most sequences have a fitness around $W = 0.5$, see figure 3.2. In the case of $K = N - 1$ all possibles fitness values are uncorrelated: a mutation in one gene means that the configuration of all other genes also changes - there is no correlation between fitnesses due to successive point mutations. In this case the central limit theorem applies, and the fitnesses are thus Gaussian distributed (in the limit of large N). For $K \ll N$ the distributions vary a lot, depending on the random initial assignments of w_i, but in figure 3.2 we show that an average of many fitness landscapes for $K = 1$ with different initial conditions shows that the distribution also approaches the Gaussian.

The idea is now that if there exist two sequences with the same fitness, they are regarded to be equivalent to molecules that fold into the same structure and perform the same function, or to molecules with different structures, but still give the approximative same contribution to the reproductive success of the host organism. With this simple modification, we have introduced the effects of genotype-structure mapping into a NK-like fitness landscape. Into the model we have incorporated the fact that selective neutrality arises as a result of the many-to-one nature of the sequence-to-structure maps found in real biological systems.

In the next section we will analyze the properties of this discretized fitness landscape, and show that for appropriate choices of the parameters N, K, A and F this model can be used to mimic real biological systems such as RNA's and proteins.

3.3 The structure of neutral fitness landscapes

We can now look how these neutral networks appear in the present model. In biological evolution the most common types of mutations are *point mutations* - eg. a mutation of a single symbol at a locus. We therefore define a *neutral network* as *the set of sequences that all have the same fitness and that are connected together via such point mutations.*

For large molecular structures, it is a well known fact that single digit mutations almost never result in a conformational change of the molecule, and that most point mutations therefore are effectively neutral (Dean, 1998; Wilks et al., 1988; Schuster et al., 1994; Schuster, 1994). Thus, our neutral networks correspond to the tertiary structures of such biological molecules, or, in the organismal case, they correspond to the phenotypes.

Because most of the sequences have a fitness around $W = 0.5$, we would expect that the largest networks have a fitness close to $W = 0.5$, and this is indeed the case. Typically, the is a large number of small networks with low or high fitness values, and then there is a small number of very large networks with intermediate fitness values.

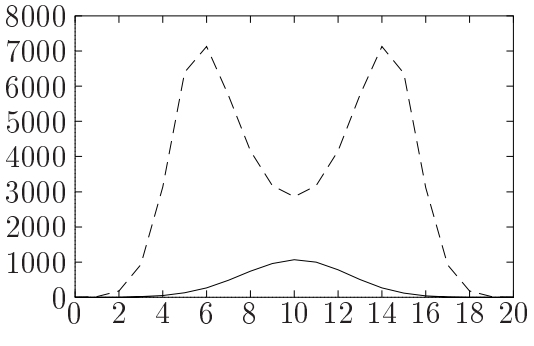

Figure 3.3: *The number of neutral networks as a function of fitness. $K = 1$ (full line), and for $K = 19$ (dashed line).*

Figure 3.3 shows this effect. In contrast to the previous figure 3.2 on the facing page, where we have plotted the number of sequences as a function of fitness, we here plot the number of *neutral networks*, connected by one-point mutations, as a function of fitness for two extreme values of K. For $K = 1$ the number of networks is again approximately Gaussian distributed, but for $K = 19$ the number of networks decreases drastically for intermediate fitness values, because it is here the giant clusters of percolating networks form. Also for intermediate values of K we see that

the largest networks appear naturally where the largest fraction of sequences has the same fitness value. Thus, the topology of the neutral fitness landscape depends strongly on K - the degree of "epistatic" interactions among the individual components. In addition, for larger K's there appear to be larger networks than for low K. This might be due to the better "mixing" of the networks in the sequences space, because the fitness landscape becomes more and more uncorrelated.

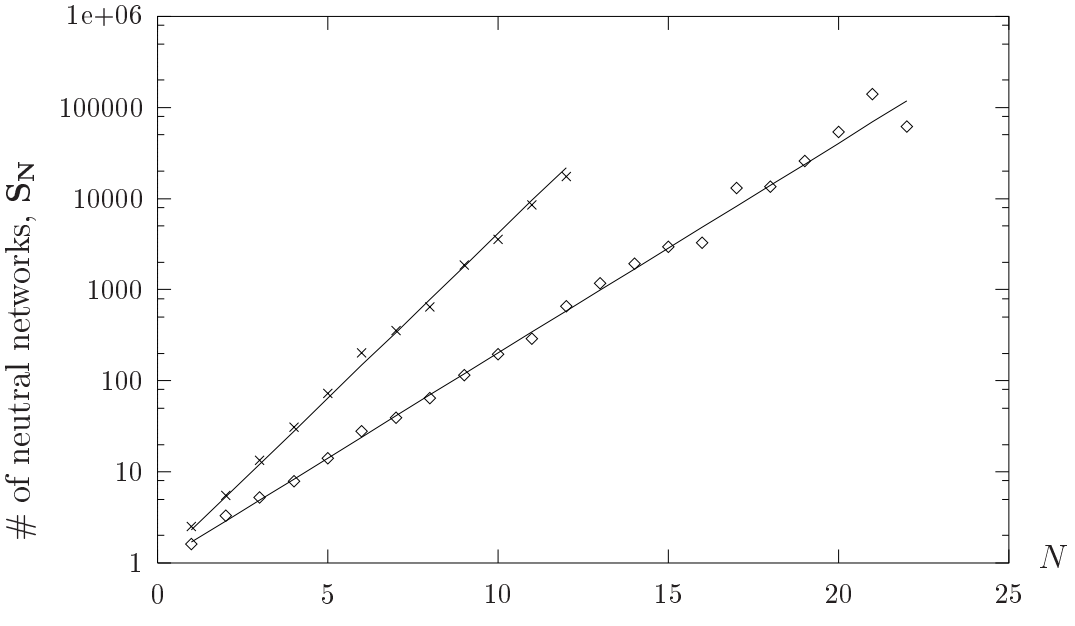

Figure 3.4: *The number of neutral networks as a function of N for $K = 1, F = 3$ and $A = 2$ (diamonds) and $A = 4$ (crosses). The best straight line fits are $S_N = 1.7^N$ and $S_N = 2.3^N$ respectively (the data are averages of ten runs).*

occurrence

We find the total number of neutral networks S_N to grow exponentially as a^N with increasing N. In figure 3.4 we show this dependency for $K = 1$ in the case of a binary alphabet and a four-letter alphabet. For a neutral landscape with $F = 3$, we find $a \simeq 1.7$ in the two letter case and $a \simeq 2.3$ in the four letter case. Interestingly, this has also been observed in the RNA studies by (Hofacker, 1994; Schuster et al., 1994; Baskaran et al., 1996; Schuster and Stadler, 1997) with the values $a = 1.6$ and $a = 2.35$ respectively.

Figure 3.5 shows the histogram of the sizes of neutral networks for $N = 20$ and various values of K. If we want to compare this result with the known

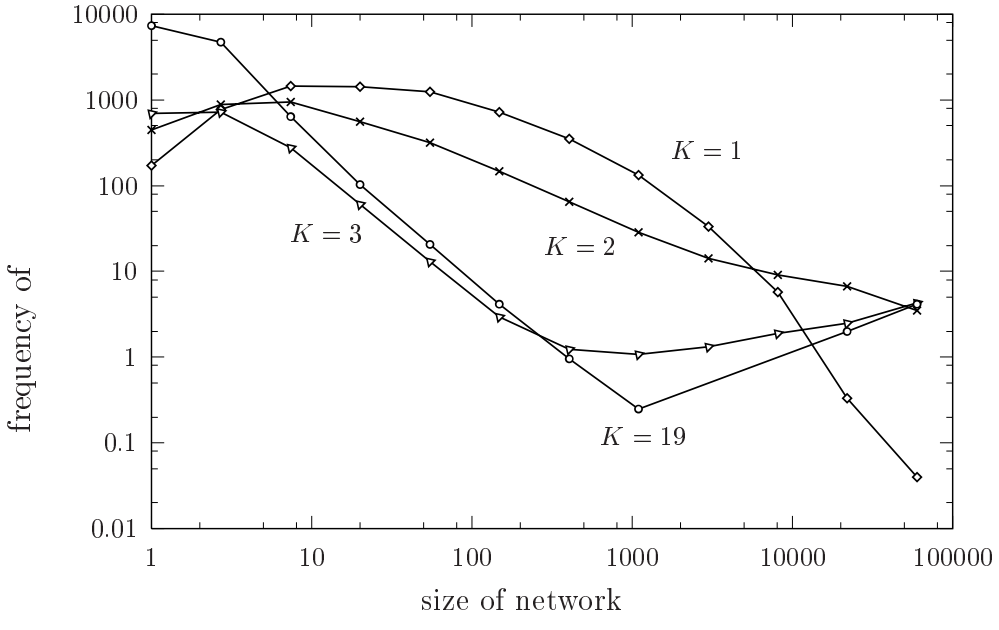

Figure 3.5: *The frequency of occurrence of neutral networks as a function of their size. The curves have $K = 1$ (diamonds), $K = 2$ (crosses), $K = 3$ (triangles) and $K = 19$ (circles). The parameters for this fitness landscape are $N = 20, A = 2, F = 2$, and the curves are averages of one hundred runs. The line for $K = 2$ has an approximate exponent of -0.7, and for $K = 19$ the exponent for the values below a size of 1000 is -1.5.*

data from RNA secondary structures, we have to look at the line corresponding to $K = 1$. The nucleotides in RNA's interact on average with only one other nucleotide: many of them form a Watson-Crick base pair with their complementary base; a few do not form a Watson-Crick base pairs since they are placed in the loops of the hairpins; and another few nucleotides interact with more than one nucleotide in order to generate the proper three dimensional structure. So, on average, we can expect K to be unity for RNA's. In figure 3.5 the curve for $K = 1$ appears to be convex, so that the distribution of network sizes falls of faster than a power law. Again, this has also been observed in the RNA studies by (Grüner et al., 1996a). But as the other lines in figure 3.5 indicate, this RNA-behavior is not generic. For larger K the curve flattens and becomes quickly concave. For $K = 2$ the curve is almost a straight line, indicating a power law decay. This might be due to some divergence of the scale parameter governing the distribution, a resemblance to critical phenomena, which will be

discussed in part III[1].

Since our investigations show excellent agreement with previous investigations of mapping of genotypes into fitness values for RNA secondary and/or tertiary structures, it is very suggestive to expect the properties of our neutral networks to be a general analogue to the mechanism of genotype-structure mapping, and applicable for many systems undergoing neutral evolution.

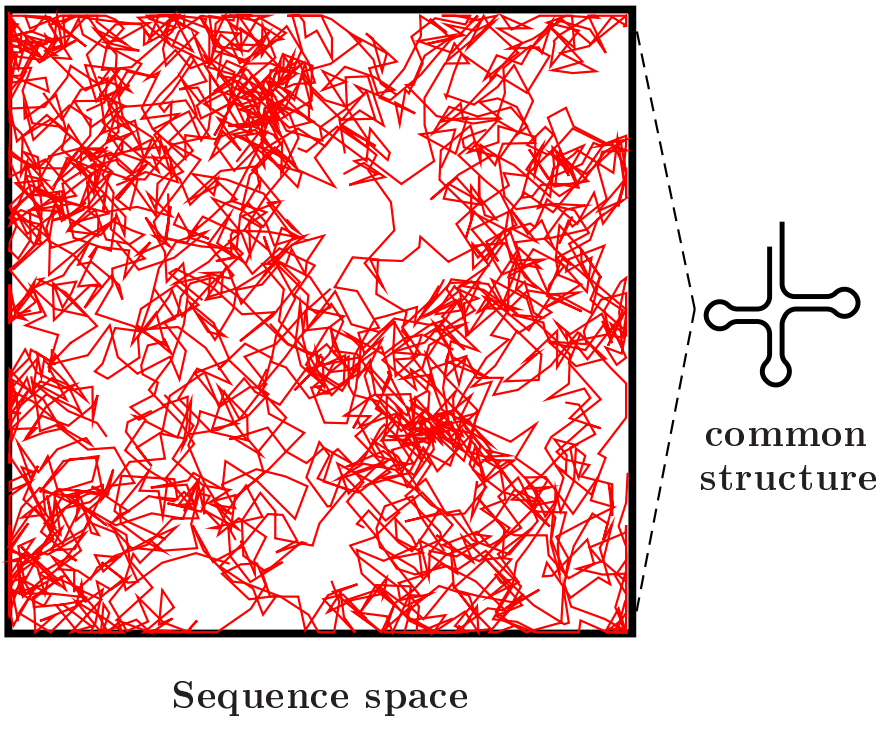

Figure 3.6: *A schematical two-dimensional representation of the high-dimensional sequence space upon which a common structure (here shown as tRNA) percolates. The lines are thought to correspond to sequences which are connected through point mutations and have the same fitness.*

3.4 Percolating neutral networks

The largest neutral networks percolate: They fill the sequence space more or less uniformly in such a way that every sequence in the sequence space is

[1]It is interesting to note that Kauffman in (Kauffman, 1993) links the situation for K close to two with the point where a phase transition from disorder to order is found (that is: a drastic decrease in the median cycle length of a normal NK attractor).

only a few point mutations away from the percolating network. We can also define the percolating neutral networks as "common structures" because they contain far more sequences than the networks of an average size. This definition of a *common structure* is due to (Grüner et al., 1996a), and we will also use it here (Newman and Engelhardt, 1998). In figure 3.6 we show a schematical representation of how to imagine the structure and size of such percolating, common neutral networks.

The number of common structures covers only a fraction of all networks, but they contain by far the most sequences. This tendency increases for increasing N, so that the fraction of sequences in common networks tends to one in the limit of large N. This finding has profound evolutionary implications: Large molecular species, such as proteins, will tend to fold into just a few forms, because these are the most probable to find. Almost any mutation away from a non-percolating an statistically improbable structure will result in the conformational rearrangement to a percolating, common structure. Natural selection will not be able to retain any other structures than the small fraction of the common ones. This has also been observed in simple models of protein folding (Li et al., 1996).

In turn, many mutations in the sequences of such percolating networks will be selectively neutral. The evolutionary fate of such macromolecules is therefore largely determined by random drift. If we for instance would return to a biological macromolecule (it could be a human protein) after one million years, it will be likely that the exact configuration of the aligned amino acids has changed dramatically, while the form and function of the molecule has remained the same. Waiting another million years, the configuration has changed again, while the form only very unlikely has been improved. Then, assuming that the rate of amino acid substitutions will be approximately constant over time for proteins of comparable length, we arrive at the interesting theoretical reiteration of the well known *molecular clock hypothesis* (Zuckerkandl and Pauling, 1965) due to the experimental findings of (Zuckerkandl and Pauling, 1962; Margoliash, 1963) in proteins among various mammalian lineages.

In any case, the results here suggest that our conclusions about the emergence of biological networks and common molecular structures might be applicable for other systems undergoing neutral evolution as well.

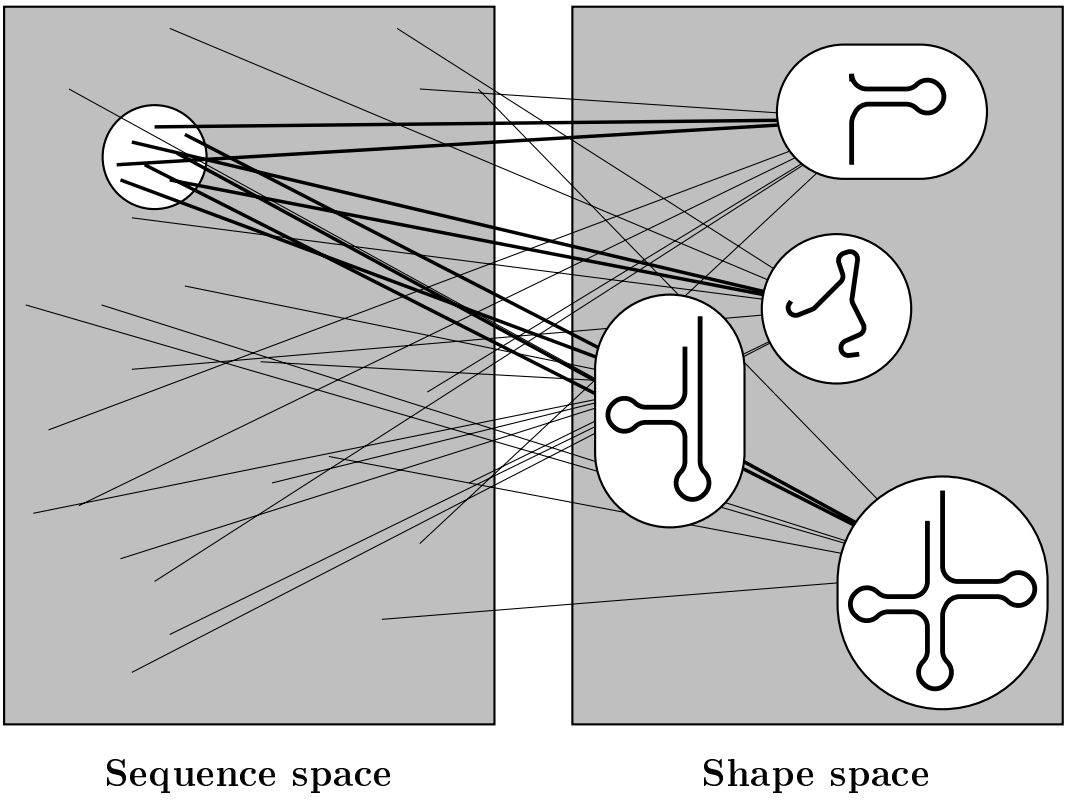

Figure 3.7: *Schematical representation of the mapping from sequences to structures and vice versa. In order to search for the most common structures, one only needs to move a small Hamming distance away from an arbitrarily chosen reference sequence, because the percolating common structures are abundant everywhere in the sequence space. The thick lines represent this correspondence. The thinner lines show that we could have started anywhere and still find the same behavior. Inspired after (Schuster, 1996).*

3.5 Covering radius

In addition to the concept of percolating neutral networks, there is another important measure for the problem of genotype to structure mapping, and this is the finding of a very small *covering radius*, leading to an effective *shape space covering* (Schuster, 1996). Starting from an arbitrarily chosen reference sequence, one only needs to screen the nearest neighboring mutants in order to find any of the common structures. The needed Hamming distances to do so is very small compared to the entire sequence space. The process of evolutionary searching towards a (common) target structure is therefore much easier than

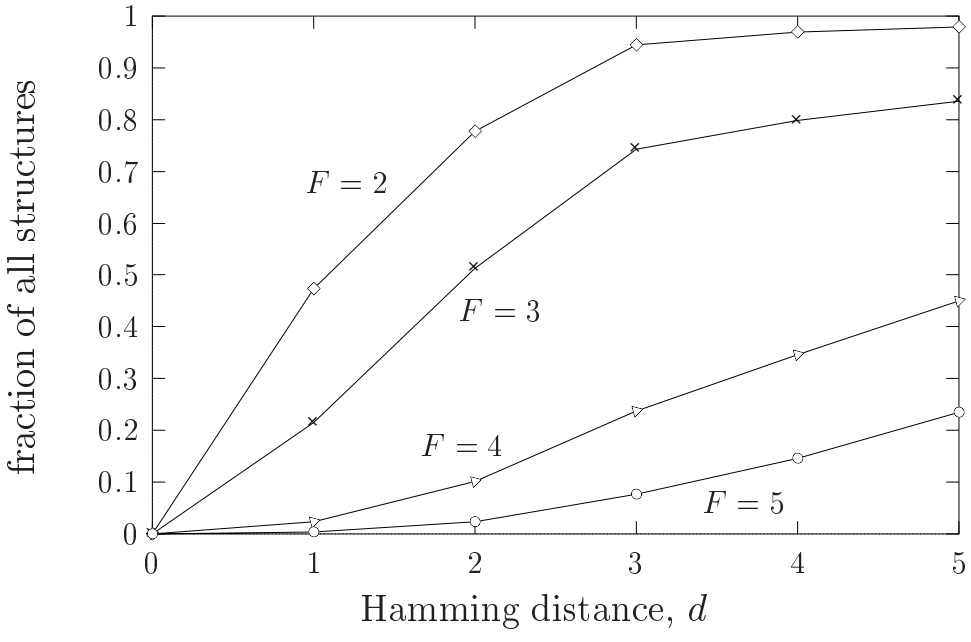

Figure 3.8: *The fraction of all structures that can be reached by neutral diffusion and d point mutations as a function of the Hamming distance d. The relevant parameters are $N = 10, A = 4$ and $K = 1$, and the four lines correspond to: $F = 2$ (diamonds); $F = 3$ (crosses); $F = 4$ (triangles) and $F = 5$ (circles). Averaged over ten simulations.*

previously expected. In figure 3.7 we have drawn a schematical picture of this phenomenon.

To test this phenomenon we also have made some simulation, and some of the results are shown in figure 3.8. The lines represent how easy it is for a given random starting sequence to reach any existing structure through the effect of percolation and a small number of point mutations. So, here the y-axis shows not only the fraction of common structures reachable, but the fraction of all possible structures reachable. We can therefore define a covering radius r_{cov} as the Hamming distance traveled from an initial random sequence in order to reach more than 90% of all structures. For instance, the upper line represents maximal neutrality, and shows that with only three mutational steps away from any initial sequence - of course they need to be the right ones - it is potentially possible to reach 95% of all structures! The important optimizational process, used both in natural evolution and in biotechnology, is heavily facilitated by this kind of neutrality.

3.6 Evolution of populations

Next, we have studied the dynamics of populations evolving on our fitness landscape. The investigations are partitioned in two steps: first we look at the dynamics of a *random hill climber*, which approximates the ensemble of sequences of a whole population to be just one sequence - the wild type.

Second, we investigate the evolution of true populations on our fitness landscape. Technically, we make use of *fitness proportionate selection*, so that each sequence (or organism) in the population is replicated into the next generation with a probability proportional to its fitness, so that the total population size remains constant. Introducing a mutation rate q per locus, we find very interesting behavior of "epochal" evolution.

3.6.1 The random hill climber

The "random hill climber" approximation due to (Kauffman and Johnsen, 1991) is - as previously mentioned in section 2.8 on page 28 - only valid when the time-scale for mutations is much longer than the time-scale on which selection operates. Thus, a random hill climber is a representation of a population by one single dominant strain. It tries point mutations consecutively, and if the new fitness is higher than before the mutation is accepted, and if not, its rejected. So, a random hill climber can never decrease in fitness (a hackneyed version of the fundamental theorem, see section 2.4 on page 18), and it therefore performs an adaptive walk in sequence space until it reaches a local fitness optimum.

In the case of the neutral landscape however, the random hill climber also accepts a neutral step which does not increase its fitness. As a result, the "population" will move diffusively on a neutral network until it find a mutation which takes it onto a network of higher fitness. This process continues until it reaches a non-percolating network, at which point it is confined to the surrounding of that network only, and the hill climber can then only climb to the local maximum within that region.

Figure 3.9 shows how the evolution of a random hill climber proceeding in time. The time needed to reach a new network of higher fitness increases exponentially for increasing fitness, and ultimately it reaches a maximal value determined by the fitness of the highest local peak where it cannot come any higher.

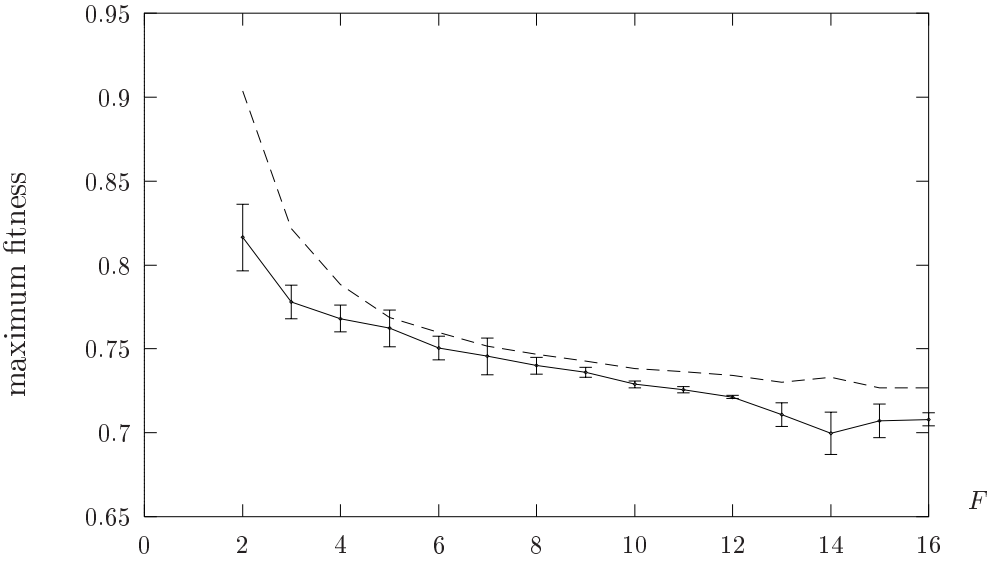

Figure 3.10: *Upper dashed curve: The maximum fitness attain of a random hill climber as a function of F. Lower solid curve: The fitness of the most fit percolating network in the system as a function of F. The percolating networks is in this case defined as the networks, or common structures, with a more than average number of sequences in them (excluded all networks of size one). The other parameters are $N = 20, A = 2$ and $K = 4$.*

This step-like behavior - or, as dubbed by (van Nimwegen et al., 1997a), "epochal" behavior - of an evolving population has also been seen in laboratory experiments on the evolution of bacteria (Lenski and Travisano, 1994; Sniegowski et al., 1997), although their explanation for this phenomenon is another. One could also draw similarities of the situation in figure 3.9 with the notion of "punctuated equilibrium", described by Gould and Eldredge in their paper about their observations of a similar kind in the fossil record (Eldredge and Gould, 1972; Gould and Eldredge, 1993). It is in the apparent periods of stasis where populations "diffuse" around in the neutral networks, and it is only because they are able to utilize these networks in a productive way, they can reach such high fitness values. Thus, neutrality *helps* populations to attain higher fitnesses.

In figure 3.10 it is investigated how high the random hill climber can climb on average (upper curve), and this is then compared to the average fitness of the most fit percolating net (lower curve). It shows quite clearly that the two curves follow each other closely in form. The climber has typically as little better fitness than the networks, because when the networks stop to percolate,

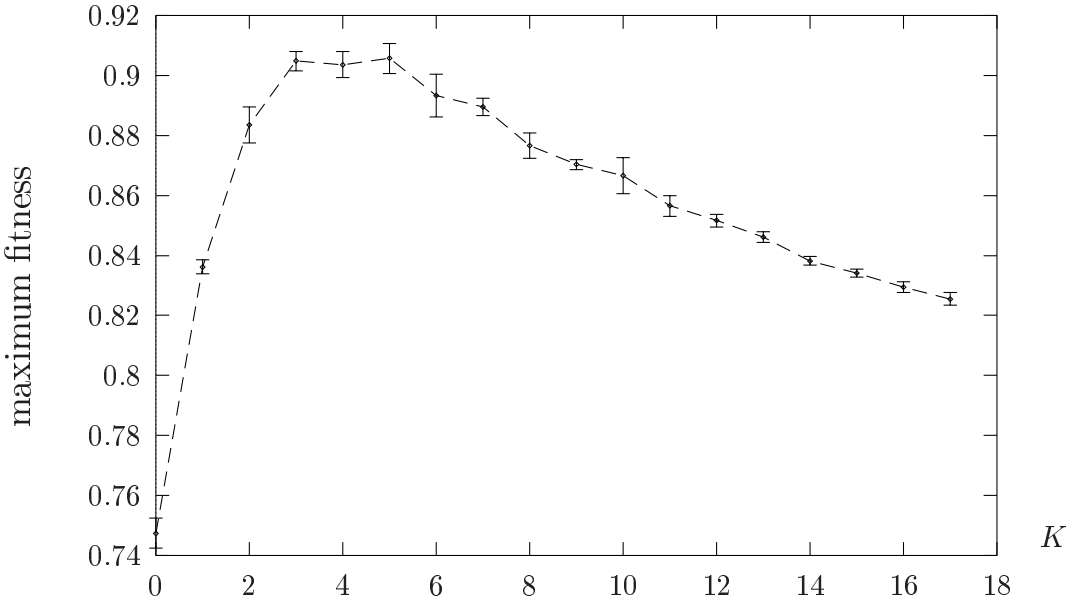

Figure 3.11: *In a landscape with a maximal degree of neutrality, we show the highest fitness reached of a random hill climber as a function of the number of epistatic interactions, K. The other parameters are $N = 20, A = 2$ and $F = 2$.*

the hill climber still can climb onto the highest local peak in that region.

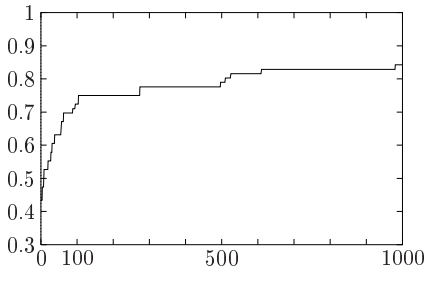

Figure 3.9: *The fitness of a random hill climber as a function of time. The parameters are $N = 76, K = 10, A = 2$ and $F = 2$.*

This tell us that the highest fitness attainable on a landscape with neutrality depends directly on the fitness of the most fit percolating networks. The higher degree of neutrality, the fitter percolating nets we find, and the better is the chance for the random hill climber to find a good peak in the landscape. Again, this idea has been mentioned in the literature before, but seeing it so clearly in a simple model like this, makes the argument that neutrality aids populations to find high peaks in the fitness landscape much more convincing. Another interesting observation is done in figure 3.11, where it is shown that the increase in epistatic interaction not necessarily means that the random hill climber finds even better fitness peaks. For K above $4 - 5$ the internal constrains of the "genes" make the landscape

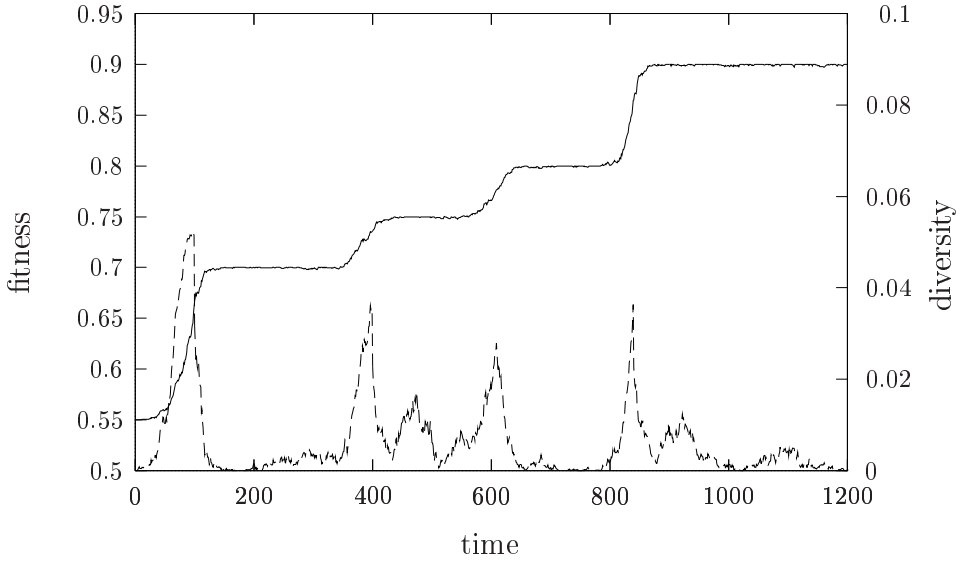

Figure 3.12: *The solid line (corresponding to the left hand side y-axis) shows the evolution of a population of* 500 *sequences in time with a mutation rate of* $q = 0.0001$ *and the parameter space:* $N = 20, A = 2, K = 4$ *and* $F = 2$. *The dotted line (right hand side y-axis) represents the diversity of the population at each stage. The diversity is defined as the fraction of non-aligned sequences compared to the wild type.*

too rugged and the relative differences between peaks and valleys decreases. This is also the case for the percolating networks (simulations not shown here): for large K there exist only percolating networks with fitness values relatively close to the mean of the distribution.

3.6.2 Fitness proportionate selection

In this section we show the results of simulations with true populations evolving on the neutral fitness landscape. A population of M sequences (all with the same initial configuration) is put into the system, and each sequence replicates itself into the next generation proportional to its fitness. The replication process is prone to errors at a rate q per locus. This means that each locus has a probability q to change the symbol it contains (the allele) into a new randomly chosen one.

In figure 3.12 we have shown the temporal evolution of such a population. Initially all members have the same sequence configuration, but in time, mu-

tants of all different kinds will emerge: an ensemble of quasi species is created.

In the periods of stasis, the population then performs a diffusive search in the landscape, where many mutants with the same fitness get produces. Of course, bad mutants with lower fitness values will also be produced, but they will be selected against in order to maintain the wild type. The average fitness value for such a population will therefore typically be a little lower than the fitness of the wild type, around which the mutant sequences - the quasi species - are distributed.

But also beneficial mutations will arise. If such a fitter mutant is lucky enough to replicate itself successfully in the beginning (when there is only one or a few of such favorable strains in the population), and not goes extinct again by random drift (which actually is the most probable outcome, see section 2.6 on page 23), it will quickly overtake the whole population because the selective force will pull the whole population onto the better fitness level (and a better structure) represented by that favorable mutant. The result is a stepwise improvement in the average fitness of the population like it was the case for the random hill climber. The speed at which such a population evolves is very dependent on the rate of diffusion on the neutral networks, which in turn is dependent on the value of the mutation rate q: for larger q than in figure 3.12, the probability to find fitter percolating networks increases, which in turn results in a more smooth curve for the average fitness, but still, for reasonable values of q, evolution is characterized by a *punctuated equilibrium* where long periods of apparent stasis are interspersed with rapid changes of the three-dimensional structure of the whole species population.

The dotted line in figure 3.12 is a crude measure of the *diversity* of the population. At each time step the wild type is calculated as the consensus sequence of the mutant distribution (see figure 2.1 on page 23 for a definition), where after the fraction of non-aligned symbols compared to that wild type is defined as the diversity of the population. When the population finds a fitter neutral networks, the figure 3.12 shows very clearly how the sequences are spread out like a rubber band getting dragged over an obstacle: The diversity increases drastically in the short periods of change but beyond the transition point it decreases quickly again. In the periods of diffusion, however, the diversity increases and decreases again, depending on the structure of the underlying fitness landscape.

For large average fitness values, the probability to find mutants with lower fitnesses becomes so high that the selective force which suppresses them is unable to do so anymore: at this point, further improvement in fitness becomes

impossible, and there might even be error threshold effects like in the quasi species model by Manfred Eigen as shown in (Swetina and Schuster, 1982).

Simulations on a population of RNA fragments by (Fontana and Schuster, 1987; Huynen et al., 1996) have show similar results. In this case, Fontana and coworkers solved the problem of fitness assignment by choosing a target structure (tRNA), which enabled them to defined a fitness measure as the distance from that target. The evolution of the population was then defined as the artificial selection towards their target and showed similar epochal behavior as in our case.

3.7 Summary

By studying this simple model of neutral evolution, we have been able to find a lot of properties, which we believe to be general enough to apply for many different systems undergoing evolution with selective neutrality. The main arguments and conjectures in this chapter were the following:

- There exists *neutral networks*, defined as sequences with the same fitness and connected through one-point mutations, which are believed to represent the set of possible three-dimensional structures such sequences can fold into.

- The distribution of fitness values among the A^N sequences is approximately Gaussian with the mean in the middle of the $N(F-1)$ fitness levels.

- The distribution of neutral networks is also approximately Gaussian for low epistatic interactions. But for larger K, the number of neutral networks with intermediate fitness values decreases quickly, because of the formation of giant percolating clusters which contain a large fraction of all sequences.

- The overall number of neutral networks S_N grows exponentially according to $S_N = a^N$, where a is some constant, and with appropriate choices of F and A, the findings are in excellent agreement with previous RNA secondary structure calculations and believed also to be valid for other evolving systems as well (such as proteins and maybe even organisms).

- Typically there are many networks containing only very few sequences and a few networks which contain a large fraction of all sequences. This finding becomes even more pronounced with increasing K. For $K = 2$ the distribution follows a nontrivial power law (with exponent -0.7), and for $K = N - 1$ the distribution follows another power law with an exponent around -1.5 (only for the small networks).

- The large clusters of percolating networks are defined as *common structures*, and fill the sequences space more or less uniformly, so that any common structure can be reached within a small Hamming distance from any random sequence. In addition, the covering radius, defined as the Hamming distance traveled in order to find more than 90% of all networks, is very low in a fitness landscape with maximal neutrality (for $A = 4, N = 10$ and $K = 3$ it is below $d = 3$), and it increases with increasing F (corresponding to decreasing neutrality).

- The evolution of a random hill climber is characterized by a step-like behavior, and the maximal fitness attained by such a random hill climber is directly dependent on the fitness of the most fit percolating networks, which in turn is dependent on the degree of neutrality. The lower F, the higher is the peak found by the random hill climber. Neutrality *helps* populations to attain higher fitnesses.

- Real population evolve similar to the random hill climber. A cloud of mutants searches on the neutral networks diffusively, until a beneficial mutant on a network of higher fitness is created, where after it eventually is able to drag the whole population onto this level by normal Darwinian selection. This process continues in this manner of "punctuated equilibrium" with long periods of stasis interspersed by rapid changes. This behavior has also been observed in experiments of bacterial evolution,(Lenski and Travisano, 1994; Sniegowski et al., 1997) and in genetic algorithms (van Nimwegen et al., 1997a). The fitness of the population is however limited by error threshold effects, because the ratio between neutral and deleterious mutants decreases exponentially with increasing fitness, so that the mutation rate puts an effective limit on the ability of selection to suppress the bad mutants.

There are a lot of additional properties which not have been investigated yet, including the detailed structure and size of neutral networks as a function of the variable topologies of the fitness landscape depending on the degree

of neutrality and the epistatic interactions. Better percolating measures and investigations of covering radii could be developed, and also details of the population dynamics, especially in the high-fitness regime; the length of epochs, error threshold quantifications, general entropy meassures instead of a crude diversity number, etc...all of which hopefully will be addressed in a forthcoming work. However, in part III, there is an analytical investigation of a percolation threshold of the system described here in the case of a maximally rugged landscape.

3.8 Conclusion

We have introduced a new model of evolution on a fitness landscape possessing a tunable degree of neutrality. We have investigated the general static and dynamic properties of this landscape and have found a number of phenomena also seen elsewhere: in RNA sequence-to-structure mapping; in simple protein models; in experimental work on evolving bacteria; in protein evolution, and in genetic algorithms designed as to exhibit neutral behavior. The phenomena repeatedly observed in all these studies are well described in this general and abstract model and include the existence - or the generation of percolating networks, the formation of "common structures", the easiness of finding such common structures from an arbitrary starting point, the utilization of neutrality in order to find high peaks in the fitness landscape, and the epochal nature of evolving populations in such settings.

The overall picture of these investigations is very much in harmony with the neutral theory suggested by Kimura and Otha. But in addition, we see the importance of random drift even more clearly in the view of these emerging molecular networks. The percolating networks create a new dynamics where they form a superimposed structured space of flows, canals and intricately complex streams of neutral mutants, through which the process of adaptation and optimization is speeded up and highly improved. The vast number of possible combinations of symbols, all equally good for the organism, form such a dynamical network, which like a stream penetrates the fitness landscape , or in Kauffmans words "searches for the adjacent possible" (Kauffman, 1996) and thereby increased the chance to survive and seems to push the unit of selection on a higher, more complex level.

In the next part we will move to such a higher level of natural selection, namely to the level of species interactions. Again we will find that the interac-

tions create interconnected and percolating networks, the so-called food webs, which are essential for the survival of the species.

Part II

Complexity in Evolution

Introduction

Natural selection operates on all levels of organization. But it also *creates* new levels of organization. In the preceeding part we have seen how the emergence of percolating neutral networks can shift the unit of selection from a molecular level to a morphological level. So, there seems to be a trend towards higher complexity.

Some of the most complex interactions in biology are the interactions between species: their competition, their spatial distribution, their diverse genetic configurations, their exchange of matter, and so on. Anyway, also at this level there will inevitably emerge some percolating networks: the so-called food webs. Obviously, these food webs will have to percolate, because the main purpose of the generation of food webs is their transfer of energy among the species (what is called the *food chains*). Species need something to eat in order to survive. Only a small fraction of these species are autotrophs, or *basal species* - that is - species which are able to transform inorganic material to usefull energy and use only that for their survival. All the other species, including us, depend on the consumption of organic material already available through these autotrophs.

This second part of the thesis deals with the generation and structure of such emerging food webs. We will introduce the first model which is able to simulate the *spontaneous self-organization of food webs*. At least to the knowledge of the present author, there has not been any model in the literature of ecology and food web theory which let's food webs generate themselves from scratch. The model presented in the second half of the next chapter and investigated in the following chapter will do the job. In fact, the model is a type of replicator equation (a Lotka-Volterra system) with only some minor but crucial changes in the governing equations.

But in order to keep the pedagogical line, we first introduce the kind of approach ecological theories have taken in history and discuss their problems and successes.

4
Ecological complexity

This chapter will start with a short account of the ideas, models and problems encountered in history of ecological modeling. The Lotka-Volterra equation (which, in fact, is a first order replicator equation which the reader already has met in section 2.3 on page 16) is introduced and its stability problems are explained. This will give us an idea of what kind of dynamical behaviour we can expect for randomly wired ecosystems.

Then, before introducing the new food web model, we have compiled and listed the most important properties observed in real ecological communities and also properties observed in the statistical distributions from the fossil record (these lists include lots of references to the literature). After that, the model is presented and some preliminary investigations are made in order to get an idea of how the system behaves qualitatively, before we in the next chapter investigate its more quantitative behaviour.

4.1 A short history of ecological modeling

In the midst of this century, the ecologists Elton (Elton, 1958) and MacArthur (MacArthur, 1955) tried to draw conclusions about the stability of ecosystems from the new knowledge obtained from the field of population dynamics. Simple models of predators and their prey had already been formulated by Lotka[1]

[1]Initially A. J. Lotka formulated a hypothetical chemical reaction, which he said could exhibit periodic behavior in the chemical concentrations (Lotka, 1910). In 1921 Bray (Bray, 1921) found temporal oscillations in his experiments with the hydrogen peroxide-iodate ion reaction, wherein he made explicit reference to Lotka's work. But this historically

and Volterra (Lotka, 1925; Volterra, 1926), suggesting highly fluctuating but stable communities. Also some experimental work in the 1930's (Gause, 1934) manifested the initial dogma: that small and simple communities were less stable than large and complex ones (although there was no evidence for the last assumption (Pimm, 1991)). On the basis of that, Elton's seminal book (Elton, 1958) formulated some of the most important issues which ecologists should start to study: population stability and variability, population recovery, invasion and its consequences, and general community structures such as food webs (Elton, 1927; Elton, 1958).

But most of the discussion was concerned with the problem of stability, and when Gardner and Ashby published a short paper in 1970 (Gardner and Ashby, 1970), which later was rigorously formalized and extended by May (May, 1971; May, 1973), it came as a surprise to the community of ecologists that a general linear stability analysis of variably connected components in an ecosystem showed decreasing stability for larger and more densely connected systems than for small and loosely connected systems. The old wisdom was turned on its head.

4.2 Concerns about stability

May (May, 1973) suggested that many attributes of ecological communities could be examined through an analysis of the type and strength of interaction between species. Thus, what had to be investigated, was the species connectance as given by a random interaction matrix[2].

In particular, people before May based their investigations on the Lotka-Volterra equations, which in the two-species form can be written:

$$\begin{aligned} \frac{dN}{dt} &= N(a - bP) \\ \frac{dP}{dt} &= P(cN - d) \end{aligned} \tag{4.1}$$

where a, b, c and d are positive constants. Here, N stands for the *prey* population, and P for their *predators*.

important work was disbelieved and dismissed because it was thought to violate the second law of thermodynamics.

[2]In ecological literature a "random interaction matrix" for Lotka-Volterra type equations normally refers to the *community matrix*, which, in fact, is the Jacobian, evaluated at the equilibrium point.

One of the nice things about equation (4.1) is the fact that the per capita growth rate (the factors in the brackets) is a linear function of the population variable, and thus corresponds to a first approximation of a Taylor series expansion around the equilibrium point in a broad class of more general models (May, 1973)[3].

4.2.1 Local stability

But equation 4.1 has also some major drawbacks. One point is the unbounded growth of the prey population in the absence of predator, as expressed by the first equation in 4.1. We will return to this later on, when we instead introduce a logistic growth rate for the species.

The second point is the most important. These kind of Lotka-Volterra equations are structurally unstable: When making a linear stability analysis around the non-trivial steady state, we find that the oscillatory solution of eq. 4.1 is not a stable limit cycle where any small perturbation tends to zero asymptotically with time. Rather, the oscillatory behavior is *neutrally stable* (implying purely imaginary eigenvalues), and therefore any small perturbation away from the neutrally stable solution results in large-amplitude displacements of all species in the entire system[4]. Elton (Elton, 1958) just assumed this to be consistent with the observed fluctuations in real populations, without testing this result for a larger number of species, and presumably without giving the problem of perturbations much thought.

Now how, then, could May conjecture that larger and more complex systems were less stable than small systems if he knew that, in fact, such Lotka-Volterra type equations already are structurally unstable?

May looked at the general n-species version of the Lotka-Volterra equations:

$$\frac{dN_i}{dt} = N_i \left(\alpha_i - \sum_{j=1}^{N} \beta_{ij} N_j \right) \tag{4.2}$$

[3]May states further that it is for this reason that the *competitive exclusion principle*, which forbids the stable coexistence of two or more species making their livings in identical ways (limits on *niche overlap*), has been so successful in the ecological literature. See also the discussion below on neutral stability.

[4]Such systems are called *conservative systems* because it is possible to construct a potential function for them. The explicit mathematical consequences are discussed in (Engelhardt, 1994; Murray, 1989; Hofbauer and Sigmund, 1998)

where the α's and β's are constants defining the type and strength of the interactions. In particular, if some species N_i have positive α_i and β_{ij} values, they are again termed *preys*, and if the constants are negative, the species are *predators* feeding on the prey. He then performed the standard linear stability analysis of the non-trivial steady state by evaluating the community matrix A, and found that the roots of the characteristic equation satisfy:

$$\sum_{i}^{N} \lambda_i = trace(A) = 0 \tag{4.3}$$

But asymptotic stability requires $trace(A) < 0$, which only can mean two things: either all eigenvalues are purely imaginary, which implies neutral stability as in the two-component case, or at least one of the eigenvalues has a positive real part, in which case the system is not just structurally unstable but *really*, that is, exponentially unstable.

Even more general, May argued that a general system of differential equations in the form:

$$\frac{dN_i}{dt} = F_i(N_i)\left[G_i(N_1, N_2, ..., N_j, ..., N_m; j \neq i)\right] \tag{4.4}$$

with $i = 1...m$, where the function $F(N_i)$ always can be factorized out, and where the function G_i is independent of N_i, will imply that the diagonal elements in the community matrix a_{ii} are zero, leading (again) to either structural or exponential instability.

4.2.2 Randomly wired ecosystems

Additional statistical investigations on the community matrix with a random number of entries had been done numerically by Gardner and Ashby (Gardner and Ashby, 1970). Their conclusions, based on computer simulations with $4, 7$ and 10 variables, where that such systems could be expected to be stable up to some critical level of connectance, and beyond this point go suddenly unstable (because of a high probability to find positive eigenvalues which diverge exponentially). In appendix D we give a nice little analytical example of, why it is so probable to find some positive feed back loops (and thus positive eigenvalues for the governing differential equations) in such interconnected systems.

On the basis of that, May (May, 1972) could show analytically that for a given connectance C, defined as the fraction of non-zero entries in the community matrix A, and a variance σ^2 of such matrix elements (with zero mean),

the probability $P(N, C, \sigma)$ that the system is stable will be such that

$$P(N, C, \sigma) \to 1 \qquad \text{if} \quad \sigma\sqrt{NC} < 1 \tag{4.5}$$

$$P(N, C, \sigma) \to 0 \qquad \text{if} \quad \sigma\sqrt{NC} > 1 \tag{4.6}$$

Thus, assuming a connectivity independent of N, there will be a sharp phase transition for $N \to \infty$.

The standard deviation σ is expressing the *average interaction strength* common to all interactions. Typically, computer simulations work with a uniform distribution of random numbers between -1 and 1, which means that the variance $\sigma^2 = 1/3$, and $\sigma = 1/\sqrt{3}$ measures the normal deviation of two interacting entries[5].

In conclusion, the investigations on local stability around a hypothetical steady state for the entire ecosystem have proved insufficient in order to understand the relative stability of self-organizing ecologies observed in nature. Lately, it has also been questioned, whether local stability is an adequate account for systems exposed to perpetual extinctions, invasions, time-dependent interaction strength and an ever changing environment (Solé and Manrubia, 1997; Brown, 1994; Hall and Raffaelli, 1993).

4.2.3 Permanence

In the case of of the two dimensional Lotka-Volterra equation 4.1, it was possible for Volterra to derive a clever potential (Liapunov) function and thus settle the question of global stability.

But it soon became clear that a construction of a Liapunov function is inappropriate when dealing with a large number of interacting species, or just when dealing with another mathematical model. Surely, from a dynamic point of view, the behavior of population frequencies is not only limited to asymptotically stable steady states defined by Lotka-Volterra models. Rather, there exists a broad range of other models with other complex behaviors - from oscillatory movements (for $N \geq 2$) to chaotic or strange attractors (for $N \geq 3$). Of

[5]In experimental situations σ is mostly unknown, and people have assumed it to be constant (Paine, 1988). Then, assuming that an ecosystem is stable and that such general Lotka-Volterra equations describe them well, the product NC will be constant. The result is a hyperbolic relationship between N and C. We will later touch upon this in section 4.3 and 5.1.1.

course, biological considerations imply that solutions to 4.1 on page 56 or 4.2 only are meaningful in the positive cone $\mathbb{R}^n_+$, and that crossing the boundary $\delta\mathbb{R}^n_+$ corresponds to a situation where at least one species becomes extinct. Thus, it is reasonable to expect that there exist some non-trivial *permanent* trajectories in the phase space, even though the local equilibrium solution is linearly unstable.

One possible requirement, which has become popular under the name *permanence* (Jansen, 1987; Law and Blackford, 1992), would be to have repelling boundaries $\delta\mathbb{R}^n_+$ in the phase space, so that each orbit of the species trajectories never leaves the system. When one species comes dangerously close to $\delta\mathbb{R}^n_+$ - that is to extinction, the concept of *permanence* expects a change to positive growth rates for that species. To be realistic, one also requires that no orbit should tend to an infinite population density[6].

If the conditions for permanence are fulfilled, no species will go extinct in the presence of selection and conversely, if some originally missing species is introduced through mutation or invasion, it will spread. In contrast to the notion of local stability, the condition of permanence is a global criterion because it applies to orbits starting from every point at which species are present, and not just for species which are inside the boundaries. Even though local stability is lost, the global condition of permanence might still be valid.

But still, extinctions and unsuccessful invasions are frequent in real ecosystems, so why should such systems bother to fulfill the criteria of permanence? Virtually all plant and animal species that have ever lived on the earth are extinct (Raup, 1986). For this reason alone, extinction must play an important role in the evolution of life. Every model, which tries to explain species evolution, must therefore incorporate extinctions. This we will do in section (4.4) - but first we will look more carefully at the properties of real ecosystems.

4.3 Observable properties of species communities

The complexity and openness of ecosystems is hardly rivaled by any other natural occurring system (apart form the nervous and immune system, maybe). They consist of many individual parts (from hundreds to billions); they maintain themselves far from thermodynamic equilibrium by uptake and transformation of energy; they exchange matter across arbitrary complex spatial

[6]For a more explicit mathematical treatment of the concept of permanence, see (Hofbauer and Sigmund, 1988; Hutson and Schmitt, 1992).

boundaries; they are adaptive and have an irreversible history; they converge and diverge in genetic constitution, and they form hierarchic patterns and their populations exhibit an enormous variety of nonlinear dynamical properties.

Experimental field work tends to use only small-scale, short-term investigations of relatively simple systems, maybe reflecting the modern time-scale of scientific impatience and demands for publication. What is often missed by such studies is the long term effect of altered feeding links, slow population changes as a result of reorganizations, extinction modes, and a general appreciation of the dynamical, non-static nature of ecological communities as observed through the fossil record.

4.3.1 Ecological time scale

Therefore, in order to characterize ecosystems qualitatively, it is important to look at both the small and the large time scale. One the smallest time scale one primarily has to focus on the most prominent features such as the relationship of individual organisms with their environment, the structure of interactions and diversity of species, and the fluxes of energy (information on who eats whom). The following table summarizes the most common features, which have been cataloged empirically (Hall and Raffaelli, 1993).

Table 4.1: **Observable properties on the ecological time scale.**

Definition	*Description*	*Data*
Food web size, N	Total number of elements in the web.	Typically between $50 - 100$
Trophic link, L	A trophic link is established when there is an interaction between two species.	See *linkage density*.
Connectance, C	The connectance is the proportion of realized trophic links. Thus, $C = L/N^2$, where L is the total number of trophic links.	Hyperbolic relationship between C and N, but maybe decreasing for larger webs.

Linkage density, $\mathcal{L}$	Instead of measuring C, people have started to favor plots showing the number of trophic links L vs. N.	Numbers from $2-5$ are all consistent with the data. Since the relationship between C and N was thought to be hyperbolic, $\mathcal{L} = NC$ was thought to be constant. Now $\mathcal{L}$ is believed to increase for large webs (Cohen et al., 1990).
Basal species, N_b	Those species that have predators but no prey.	Independent of N but with high variance in the data (Briand and Cohen, 1984; Cohen and Briand, 1984).
Top species, N_t	Those species that have no predators.	High variance, and the ratio $N_t/N_b \simeq constant$ (Evans and Murdoch, 1968; Jeffries and Lawton, 1985).
Number of omnivores, N_o	Omnivores are organisms which feed on more than one trophic level. An unequivocal definition is still missing, so in the following investigations we have defined omnivores as predators also having a Malthusian growth rate, see section (4.5.2).	Less common in real webs than in randomly generated webs (Pimm, 1982).
Food chain length, ℓ	Food chains run from each top predator through the intermediate species down to the basal species. The food chain length is defined as the number of links in this path and is one less than the number of species in the chain.	Typically short, from 3 to 6. Very few feeding loops (Pimm, 1982; Cohen et al., 1986).

4.3.2 Time scales of speciation and extinction

On the larger time scale of speciation and extinction, as defined by (Stenseth and Maynard Smith, 1
one has made the following observations by looking primarily at the fossil record
(data mainly after (Solé and Manrubia, 1997)):

Table 4.2: Observable properties on the speciation-extinction scale.

Definition	*Description*	*Data*
Extinction patterns, $P(f)$	The time series of the number of extinction of a taxonomic unit (typically family or genera) as observed in the fossil record. $P(f)$ is the corresponding power spectrum (calculated as the Fourier transform of the autocorrelation function)	Typically $P(f) \propto f^{-\beta}$, where $\beta \simeq 1$ (Solé and Manrubia, 1997)
Extinction distributions, $S(m)$	The distribution of extinctions of size m	Follows a power law decay: $S(m) \propto m^{-\alpha}$ with an exponent $\alpha \simeq 2$ (Newman, 1996; Solé and Bascompte, 1996).
Lifetime distributions, $S(t)$	Distribution of lifetimes of families or genera	Again a power law decay $S(t) \propto t^{-\kappa}$, where the exponent $\kappa \simeq 2$ (Raup, 1991).
Static structures, $S_g(N)$	The number of genera, formed by N species, shows fractal properties	Power law distribution $S_g(N) \propto N^{-\alpha_b}$, $\alpha_b \simeq 2$ (Burlando, 1990). The first investigations of this came from Willis (Willis, 1922), who found an exponent ~ 1.5.
Van Valens constant extinction law, S_t	States that a species might disappear at any time, irrespective of how long it already has existed. Thus, there is an exponential decay in the number of surviving species throughout time (Van Valen, 1973)	Van Valen measured $N(t) = N(0)e^{-qt}$, with $q = 0.13$, but when narrowing down the sampling window, and looking at the *survivorship curves*, one finds pronounced stasis interspersed with major extinction events (Raup, 1986).

All these are fairly many observations, and it has over the years seemed as a Sisyphean labor, if not impossible, to construct a single model system, which was able to reproduce and interpret all these properties. No set of equations has been able to span the range of complex behaviors observed, while simultaneously being numerically tractable (not to speak of analytical

tractability).

One approach to cut through this pandemonium of difficulties has been the idea of *self-organized criticality* (SOC). In essence, it refers to the tendency of large dynamical systems to evolve *spontaneously* towards a critical state characterized by spatial and temporal self-similarity, which then expresses itself in a variety of *power law*-relationships of the variables. From a conceptual point of view, this approach is very imaginative and inspiring. But from a modeling point of view, one can have an insecure suspicion that SOC just might be a helpless etiquette, which tries to explain, but in fact just scratches the surface of overt facts, leaving us in the dark with the true dynamical mechanisms responsible for such a self-organization[7]. However, many characteristics which define SOC systems will also show up the the model we are going to define in the following section.

4.4 A model for self-organizing food webs

This model is a relatively simple model which has a rich variety of dynamical properties of the kind described in the previous section. In particular, it is the first model which lets many ecological properties, such as food webs, preys and predators, top and basal species, etc. *organize themselves*.

The key features are the following:

1. Most important, the strength and type of interactions among species changes in time. Thus, from a modeling point of view, one needs to let the constants, defining the interactions in equation 4.2 on page 57, be dependent of time:

$$\frac{dN_i}{dt} \propto \sum_{j=1}^{N} \beta_{ij}(t) N_i \tag{4.7}$$

   Moreover, the matrix elements $\beta_{ij}(t)$ incorporate two distinct properties which we can separate in two:

   a An interaction matrix $A_{ij}(t)$, defining who is interacting with whom, so that, when there is an interaction, for instance species 3 feeding

[7] We will discuss self-organized criticality more extensively in part III.

on species 7 (A_{37}), it is turned *on* ($A_{37} = 1$)[8], and, if there is no interaction, it remains zero.

b An extra matrix $B_{ij}(t)$ for the *interaction strength* defines how strong this interaction is. So, the values of $B_{ij}(t)$ are defined as real numbers between $[0; 1]$.

The interactions will then be defined as the products of the two matrix elements:

$$\frac{dN_i}{dt} \propto \sum_{j=1}^{N} A_{ij}(t)B_{ij}(t)N_i \tag{4.8}$$

2. Traditionally, the growth rate of a species is thought to be dependent on its own population size, as expressed by the factor N_i outside the brackets in equation 4.2, reflecting standard Malthusian growth (the $N_i\alpha_i$ term), and population-proportional depletion (the $N_i \sum_j \beta_{ij}N_j$ term). This is quite unrealistic (Murray, 1989) because it implies an unbounded growth of the prey population in the absence of predators. Normally, theoretical ecologists replace such a linear growth rate with a more realistic *logistic growth rate*:

$$\frac{dN_i}{dt} \propto \sum_{j=1}^{N} A_{ij}(t)B_{ij}(t)N_i(\kappa - N_j) \tag{4.9}$$

where κ is the *carrying capacity*, which says that the populations are more or less limited by the capacity of the environment to carry them, so that when the population reaches κ, the growth and death rates are about equal[9].

3. The final idea to introduce, apart from time-dependent interactions and logistic growth, is the fact that all species need to optimize their environmental interactions and therefore need to adapt to different, often

[8]Note that A_{73} is not affected, because we assume no causal connection between one species feeding on an other and that other being affected of the first. This might seem contra-intuitive, but in real ecosystems many species are not affected by the exploitation of other species on higher trophic levels. And also, the interactions might not always be competitive or beneficial, but more or less neutral.

[9]One could of course argue that the carrying capacity also changes when some species go extinct or other get introduced in the ecosystem competing for the same niche. Nevertheless, we will restrict ourselves to a constant κ.

conflicting demands in order to survive. Each favorable interaction (it could be tasty food) has some adaptational costs (good legs). Thus, we introduce an *adaptational load* or *foraging load*, which in the simplest approximation can be set inversely proportional to the number of species with which a species interacts.

Now we can extend eq. (4.9) to become

$$\frac{dN_i}{dt} = \frac{1}{K_i} \sum_{j=1}^{N} A_{ij}(t) B_{ij}(t) N_i(\kappa - N_j) \tag{4.10}$$

where

$$K_i = \sum_{j=1}^{N} A_{ij}, \qquad 1 \leq K_i \leq N \tag{4.11}$$

is the total number of interactions of the species i, that is, K_i is the specific *linkage density* $\mathcal{L}$ from the previous section.

This concept of a *foraging load*, $1/K_i$, is something new. It makes the assumption that a species - in a stable environment where there is plenty of all kind of food - will specialize for only one or a few food sources in order to minimize the adaptational cost[10]. But the crux here is that environments are not stable. In an ecosystem, where old species go extinct and new species get introduced; where interactions change and external perturbations are frequent, a one-eyed specializing strategy might prove fatal on a larger time scale. In the presence of extinctions and re-population by new species, a balanced strategy might instead prove beneficial.

4. How do we remove and introduce new species? The following updating rules are easily implemented:

   a When in eq. (4.10), at a time t, a population goes extinct such that $N_i(t) \leq 0$, all non-zero entries in the interaction matrix A are set to zero, that is $A_{ij} = 0$, and $A_{ji} = 0$ (so that all species which have interacted with species i are affected).

[10] This is of course an idealization because there often are many food sources which do not need further adaptational investments, see the discussion.

b All entries B_{ij} and B_{ji} are refilled with new random numbers between $[0;1]$ (so that a new species, which will be introduced here, will have new kind of interaction strengths with the other species).

c Introduce a random number (uniform between 1 and N) of 1's in the column of A_i (A_{i1}, A_{i5}, A_{i8} and A_{i9} for instance). The other species in the system will first encounter it, when they themselves change their interaction (the entries A_{1j}, A_{5j}... are not affected).

d random mutations in the entries A_{ij} and B_{ij} with values p_a and p_b respectively are introduced in order to reflecting the time dependence of the interactions.

It is easily seen that a system with this setup will be affected by at least two different time scales: one is the ecological time scale, where specialization, building of percolating nets for food transfer and permanent trajectories in phase space are important. The other time scale appears, when the criteria of permanence is not fulfilled: extinctions, introduction of new species and external perturbations dominate the dynamics, and diverse distributions of extinction patterns etc. can be investigated.

4.5 Preliminary qualitative investigations

However, the above model is still very complicated, when one wants to make exhaustive statistics on the emerging properties as discussed in section 4.3. This is of course due to the intrinsic stiffness of the coupled differential equations, which calls for sophisticated and time consuming integrators[11], and the still limited computer power when N becomes large.

4.5.1 Time-development and stability

When integrating eq. (4.10), with random numbers occupying the matrices A and B, one encounters the following picture (see figure (4.1)): On a large time scale the overall population dynamics becomes quite fluctuating, where small changes in the interactions result in abrupt changes in the growth rate. Some species are able to find a constant population value, often around the carrying capacity, but the most trajectories are dominated by large-amplitude

[11] The integrator used is a general stiff $4th$ order Runge-Kutta method, called Rosenbrock-method, see (Press et al., 1992).

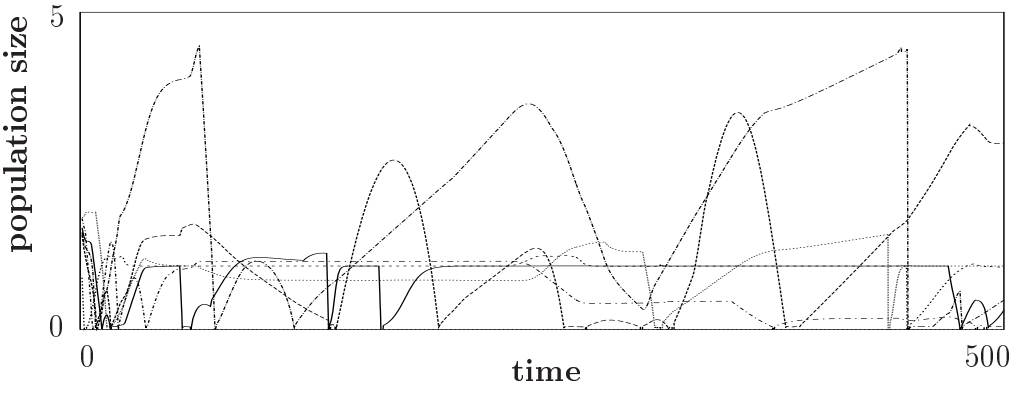

Figure 4.1: *Qualitative behavior of eq. (4.10) with* $N = 16, \kappa = 1$ *and a probability* $p_a = 0.001$ *for a mutation at a site in interaction matrix* A *and* $p_b = 0.002$ *for a mutation in* B *per unit time. Six of the sixteen population trajectories are plotted, and one can see that the dynamics is dominated by large-scale large-amplitude fluctuations. Some of the species go extinct, other find a temporary equilibrium around the carrying capacity* κ*, and still other fluctuate unpredictably in phase space. In general, the larger probability for change in the matrix elements* A *and* B*, the more fluctuations there will be.*

oscillations, and nobody can be safe from extinction. The trajectories with a negative growth rate
indexgrowth rate!negative will eventually reach extinction, where after a new species is introduced with different interactions and interaction strength, and an initially low population. With good fortune these will grow, depending on the survival of their food sources, defined as the interactions in matrix A. This means that extinctions often occur in bursts. When some key species go extinct, they might pull many of the other species along with them, leading to a cascade of extinction events. After that, the system needs some resilence time in order to become stable again, if possible.

The steady state $\mathbf{N}^* = 1$ for $\kappa = 1$ for all i seems for the most cases to be structurally unstable, but often the system settles in another, nontrivial quasi steady state, which, in turn, can go unstable again, as shown in figure (4.2). In this figure there are not introduced any *external* perturbations. Rather, it is an intrinsic instability which amplifies itself throughout the whole system. Without the requirement of repelling boundaries, the notion of *permanence* would be a reasonable description of the dynamics of equation 4.10. However, on a macro-evolutionary scale, global stability is *non possumus*, and that's also the reason, why there is such thing as evolution. The step from short-term population dynamics to long-term evolutionary dynamics can only be

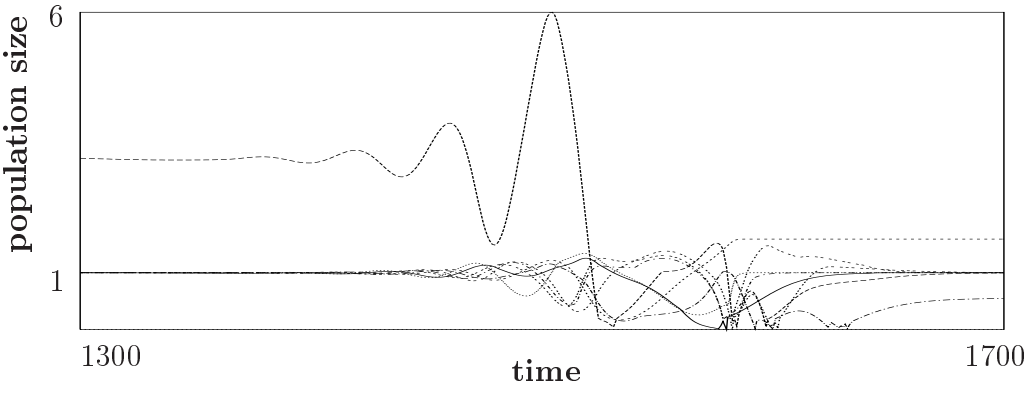

Figure 4.2: *Another situation. The system has been stable for a while, but suddenly an instability amplifies itself, and the trajectories form large-amplitude oscillations. However, after some extinctions, the system settles itself again in a quasi-stable state.*

introduced, when diverse stability criteria are violated *permanently*. It is also a well known fact (Pimm, 1991) that real species communities have long-term, large-amplitude changes in abundance, all though they seem quite stable on a shorter time scale.

4.5.2 Self-organizing food webs

The connectance C (or linkage density $\mathcal{L}$) of the system is here, in contrast to the general Lotka-Volterra models, not constant, and not even a variable. It is *self-organized* in a true sense. The specific linkage density K_i, defined in eq. (4.11), is initially set to a random number between 1 and N[12].

Because of the described dynamics, the system will then quickly *disconnect*, so that K on average is low, eg. - the individual species which have few but good food sources (few 1's in A but high values in B) will have larger growth rates than their partners with many food sources[13]. This can be seen in the insert of figure 4.4 on page 71. However, since everybody in the system competes for an optimal foraging strategy, the structure of the interactions have the tendency to become *very ordered*. Thus, this is the first model, which lets food webs *organize themselves*. We have not imposed any limitations or require-

[12]It is reasonable to expect this number to be fairly below N, since most species initially only would be capable of interacting with a few other ones - at least with far less species than there are existing in the whole ecosystem. See section (5.1.2) for further details.

[13]This is due to the larger variance of dn_i/dt when K_i is low

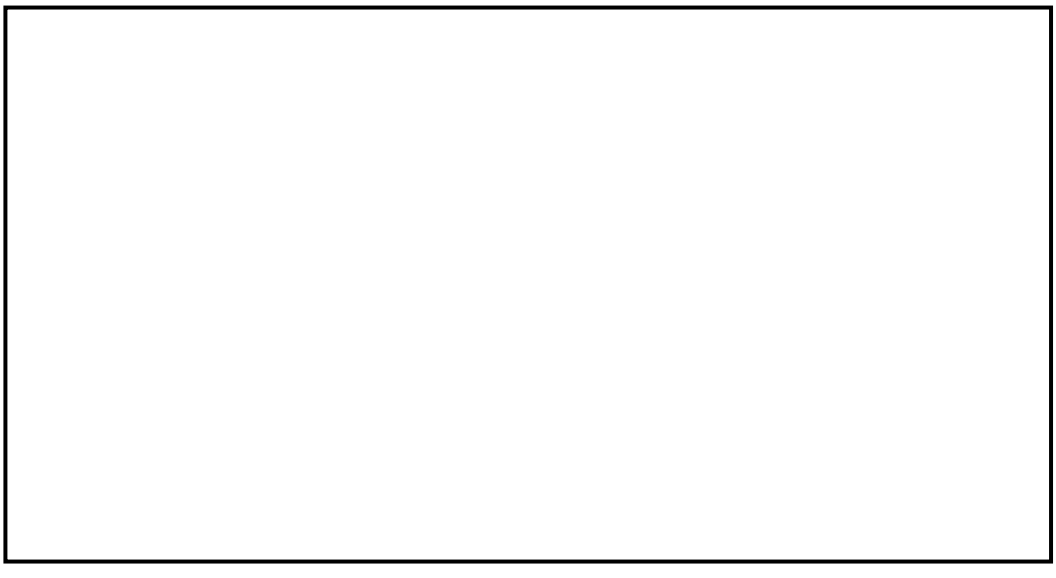

Figure 4.3: The self-organized food web. The arrows indicate the flow of energy and the circles represent autotrophs. See text for further explanation.

ments on how the species should interact. In the beginning of a simulation the matrices A and B are filled randomly, which normally results in heavily wired structures with lots of loops and circular food chains. But after only a short time of integration, the system undergoes a sudden transition towards well-ordered interactions where basal species and top species appear accordingly. The system self-organizes towards a percolating net of species community.

As an example, look at figure 4.3. It shows the resulting food web (the structure of the matrix A) of the simulation in figure 4.1 on page 68 at time $t = 500$.

What is seen in this figure, is the inter-dependencies of all sixteen species from the previous example, structuring themselves in optimal percolating food webs in order to maintain their growth rates and thus escape extinction. Out of the sixteen species, six are self-activating (having a 1 in the trace of A), which is indicated by the loops in the figure. The arrows show the flow of energy, that is: they show who is eating whom: For instance will an arrow from 9 to 1 indicate that 9 is the food source of 1, in the same way as 6 is the food source of 9 (and 8, 14 and 10). Seven of the sixteen species are not fed upon (species $3, 7, 15, 16, 1, 11$ and 2), and consequently they are termed *top species* as defined in section (4.3). Four species (numbers $5, 6, 4$ and 12) are not feeding upon any other species, and are only existing because of their self-

activation. Instead of defining them solely as cannibals, we define them as the autotrophic *basal species* feeding on external sources like the sun and soil. At a first cast this might seem a misinterpretation, but it is actually equivalent with the introduction of a positive (Malthusian) birth rate (Na) into the normal Lotka-Volterra equation (4.1). Besides the top and basal species, one can see all kind of intermediate species: species with different numbers of predators and preys or species with a positive growth rate in addition to predation (species 8 and 10). These last two species we define as *omnivores* (note that species 11 is not an omnivore).

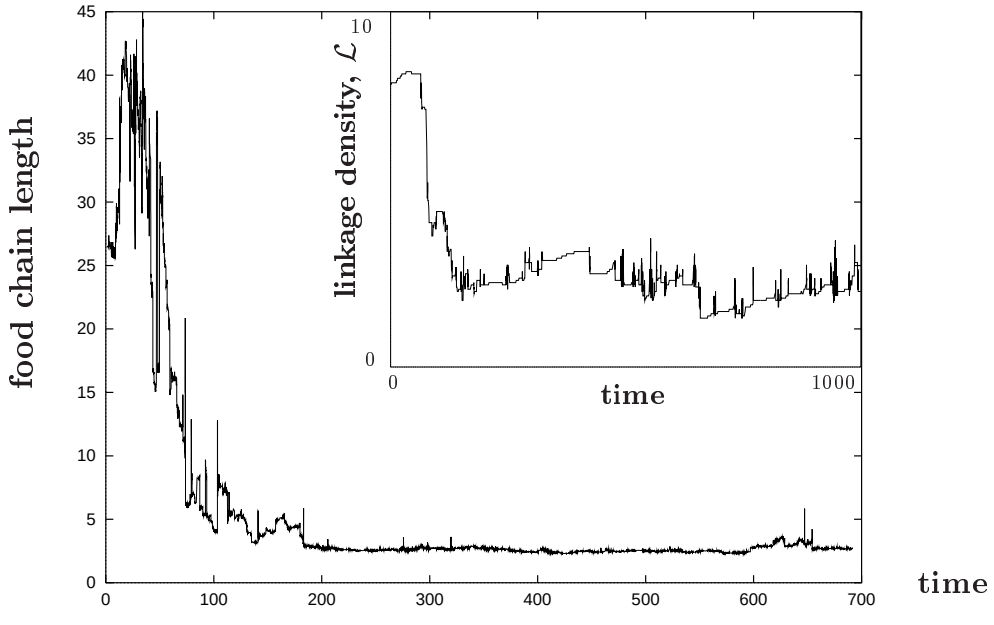

Figure 4.4: *Self-organization of the food chains. The average food chain length is plotted against time. The system undergoes a rapid transition from randomly wired structures to ordered, percolating webs.* $N = 130, p_i = 0.002$ *and* $p_f = 0.004$. *Insert: The time development of the average linkage density shows that for most of the species it is favorable to have a low number of trophic links in order to optimize their growth rate. Although the dynamical changes of the populations keep on to make new links due to introduction of new species and mutations, the tendency is clear: Randomly wired ecosystems disconnect quickly in order to maximize individual advantages. Note again the sudden transition from a randomly wired system (large* K*) to an ordered system (low* K*) at time* $\simeq 50$. $N = 16, p_i = 0.001$ *and* $p_f = 0.002$.

Another interesting feature of the model is the decoupling into different sub-webs: in our example there is one big web with 10 species, and two small

webs. This is quite plausible and also observed in real species communities as a result of specialization.

The food chains are normally defined as all the possible paths from a top species down to a basal species. In our example there are seven top species which have a total of 11 trophic links until they reach the basal species. Although crude, one approximate measure of the average food chain length is thus

$$\ell = \frac{L_{tot} - N_b - L_o}{N_t} \tag{4.12}$$

where L_{tot} is the total number of links in the system, N_b and N_t the number of basal and top species respectively, and L_o is the number of links which belong to omnivores. This definition of the food chain length is a lower limit, since the trophic interactions of intermediate species, which feed more than one top species, but themselves are feeding on less, not are counted multiple times (for instance: the total number of trophic links counted for both top species 1 and 11 linked to the intermediate species 9 is 3 and not 4). Using eq. (4.12) on figure (4.3), we find $\ell = 11/7$.

When filling the interaction matrix A with random entries in the beginning of a simulation, the food chain length is typically very large (rather it's not even a chain; it is a heavily entangled structure without any direction). But due to the dynamics, the structure decouples and becomes ordered very quickly. The main part of figure (4.4) shows an example of the development of the food web structure in terms of the food chain length versus time.

It is now easy to see the result of a possible perturbation of the system. If one species, say species 9, encounters a change in interaction strength (mutation in B), or a change/removal of a food source (mutation in A), or it just drops down to a dangerous low population level by the intrinsic dynamics, it will, after an eventual extinction, also remove the basis of food supply of the species 1 and 11. Maybe 11 will survive because of its connection to species 4, but number 1 will surely die. In this situation, the species population dies instantly, because it is removed by the algorithm and substituted with a new species. One could, in principle, modify this stringent rule with more reasonable assumptions such as different transition times, population-dependent mutations, and introduction of additional features (spatial, temporal, genetical, etc), but the fundamental aspects of ecological self-organization as suggested by this model would be blurred by a wealth of conflicting mechanisms. So, for these initial investigations, we will try to keep the model simple.

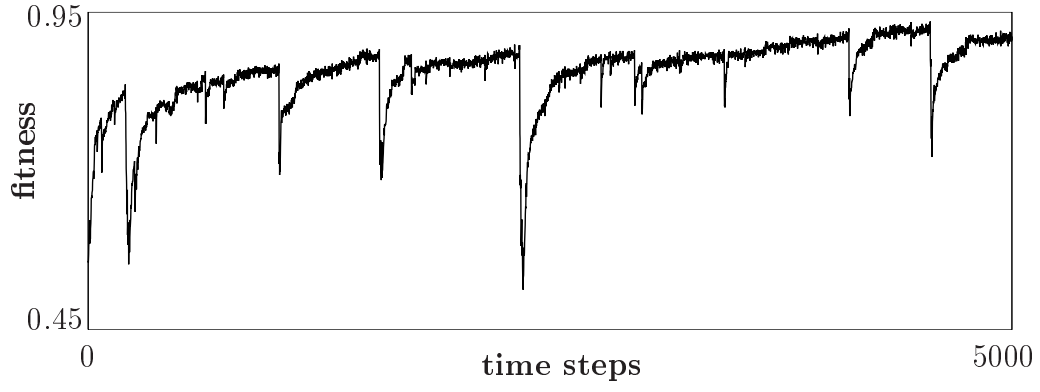

Figure 4.5: *Evolution of the average fitness in the discretized system. The number of species in this simulation is* $N = 32$, *and* $p_a = p_b = 0$.

4.6 A toy model of the model

An even simpler model - one without the troublesome coupled differential equations - could be desirable. This could then allow us to make even more extensive statistical measurements of the different properties in the model. In order to do so, we can replace the differential dn_i/dt with a unique *fitness function* f_i, which defines the fitness of the whole species, and is calculated as:

$$f_i = \frac{1}{K_i} \sum_{j=1}^{N} A_{ij} B_{ij} \tag{4.13}$$

This simplified version of equation 4.10 on page 66 is updated in discrete time steps, where the species with the lowest fitness value is removed in accordance to the rules given above in section (4.4,4)[14]. In this description the model is very easy to implement on a computer. Note that there is no population here. Every species is solely characterized by one fitness value. There is no such thing as a time scale at which one species population decreases towards an eventual extinction. Instead we remove the lowest fit element at every single discrete time step.

The dynamics of this system is very much alike the dynamics of the integrated version. Moreover, we can see here, how the system optimizes towards

[14] Actually, this version of the model was the first investigated by the author - but that's the way science goes: you get ideas from colleagues (Mark Newman, Axel Hunding), and then you get ambitious.

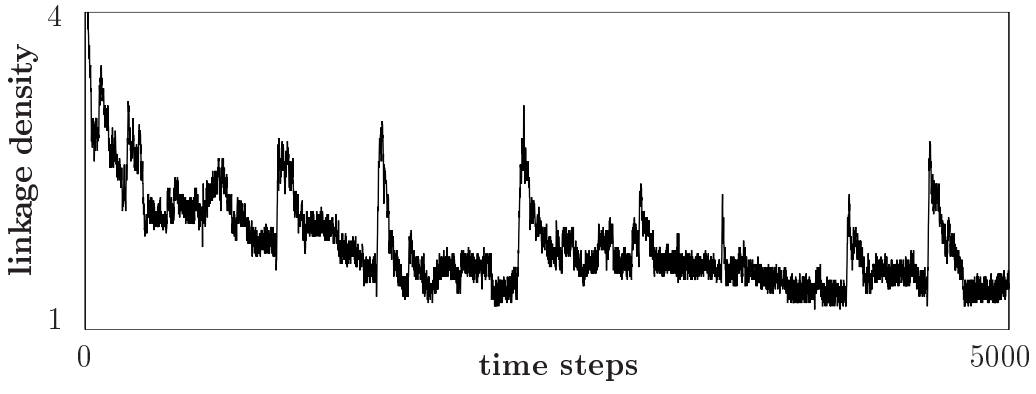

Figure 4.6: *The average linkage density in the same simulation as in figure 4.5*

larger fitness values in connection to the self-organization of the food webs figure 4.5. The average fitness increases rapidly because the dynamics always removes the lowest fit species. After a build-up period, sudden catastrophes of extinctions destruct the self-organized structure (a basal species goes extinct, for instance, resulting in a cascade of extinction of the species linked to it), and the system will rebuild itself again towards a new, ordered, food web structure. The average linkage density, $\mathcal{L}$, follows the average fitness closely, as seen in figure (4.6). When there is an extinction of an important basal species (because of an invasion of a fitter variant) with lots of trophic links feeding on it, it generally triggers many extinctions. The ecosystem becomes disordered and many of the most specialized species disappear, until the system has reorganized itself again.

In summary, this toy model, which replaces the population densities with an unique fitness function, has many of the same characteristics as the full system, formulated in eq. 4.10. The main new idea in these two models is the inclusion of the $1/K_i$-factor. Although the toy model - equation 4.13 - in many case is in quantitative agreement with the full system, we will in the following investigations concentrate on the differential version - equation 4.10 - because it's more realistic and also because it is more in harmony with the tradition of Lotka-Volterra type equations for ecological modeling and the investigations of replicator dynamics in general.

4.7 Discussion

Compared to the complexity of ecological system, the presented model is kept as simple as possible. Still, the qualitative behavior of the population densities is very unpredictable and complex. There is no parameter tuning in the equations, and it is therefore possible to say that the emerging structures are truly self-organizing. The system drives itself towards a critical state, characterized by a self-organized connectivity, directional percolating food webs with top and basal species, punctuated extinction patterns, etc. - all in order to keep the individual species populations in the game of evolution.

It should be noted that the model from equation 4.10 on page 66 can be rewritten, so that it with some modifications becomes the Lotka-Volterra equation:

$$\begin{aligned} \frac{dN_i}{dt} &= \frac{1}{K_i}\sum_{j=1}^{N} A_{ij}B_{ij}N_i(\kappa - N_j) \\ &= \frac{N_i}{K_i}\left(\sum_{j=1}^{N} A_{ij}B_{ij}\kappa - \sum_{j=1}^{N} A_{ij}B_{ij}N_j\right) \end{aligned}$$

so that, if we remember that the general N species Lotka-Volterra equation 4.2 from page 57 is

$$\frac{dN_i}{dt} = N_i\left(\alpha_i - \sum_{j=1}^{N}\beta_{ij}N_j\right),$$

then, the constants α and β become equal to

$$\begin{aligned} \alpha_i &= \frac{\kappa}{K_i}\sum_{j=1}^{N} A_{ij}B_{ij} \\ \beta_{ij} &= \frac{A_{ij}B_{ij}}{K_i} \end{aligned}$$

From this point of view, the present model is a special version of the general Lotka-Volterra equation, although with the important difference of an introduction of the factor $1/K$.

The introduction of the $1/K$-factor does maybe not need an explanation in terms of a foraging load (Mark Newman, private communication). It might

be enough to argue that if individual species depend on a number of other species, such as herbivores on several plants, or predators on many preys, they experience each individual contribution to their diet as comparatively less important - inversely proportional to the number of utilized food sources - $1/K$. For instance: if a predator eats ten preys for its livelihood, and one of the preys goes extinct, the predator only looses 1/10th of its fitness, since the individual contributions to the predators fitness are 1/10. But if a predator only eats one prey, it is very vulnerable, because it dies if that prey becomes extinct. In the literature, the normal modeling approach has been to expect an additive fitness. The more sources you can find, the larger fitness you get. But individual species only need a finite amount of energy input in order to live. This finite amount of energy can be obtained in different ways: either you specialize for only one or a few food sources, or you generalize for many food sources where each source contributes only partially to the total energy required. Both strategies have their advantages: In the short term specializers might have an advantage if their resource is very abundant (lessons from *optimal foraging theory*) (Emlen, 1966; MacArthur and Pianka, 1966; Schoener, 1971), or if their resource provides not only food, but also protection, mating possibilities, etc. (Colwell, 1986). Generalizers do better in the long run, because extinctions of their resources do not affect their survival in the same degree. Thus, one would expect a longer average lifetime of generalizers, although there might be less of them. In section 5.4 we will show that this is indeed the case. Also, one would expect that species, within a web, which interact with many others should do so weakly (small b_{ij}'s), while those which interact strongly should do so with but a few species. Also this will be shown to be true.

The present model is of course an extremely idealized version of the possible dynamical interactions in real ecological communities. Individual species surely develop many additional characteristics that cannot and will not be included in the algorithm. The tension between specialization and generalization could for instance be relaxed by specific choices of the mutations probabilities p_a (reflecting observations of "partial preference", see (Futuyma, 1986, chap.9)), or the $1/K$-factor could be weighted with the population densities, changeable carrying capacities, etc..

Since every simulation is different due to the randomness of the initial configuration and the randomness of the invading species, this model could at a first glance seem as a paradise for people who do not want their results reproduced. However, the emerging properties have of course a statistical significance which then can be compared to the experimental data. This will be

the subject of the next chapter.

5
A closer look on the data

In the following chapter we will analyze more deeply, how the N coupled differential equations construct their own network of interacting and evolving populations, and we will try to find some statistical results of the emerging properties of this model. Let us recapitulate that we investigate the coupled differential system

$$\frac{dN_i}{dt} = \frac{1}{K_i} \sum_{j=1}^{N} A_{ij} B_{ij} N_i (\kappa - N_j), \qquad K_i = \sum_{j=1}^{N} A_{ij} \tag{5.1}$$

where K_i is the total number of interactions of species i.

First we will look at the short time scale and investigate properties such as the connectivity, the linkage density, the food chains and the number of species on different trophic levels. Then we go on and investigate the extinction dynamics, the resulting distribution functions of lifetimes, survivorship curves, etc. Finally, after a summary, we compare our results with the experimental data obtained from the investigations of real food webs and from the fossil record.

5.1 Food web statistics

H

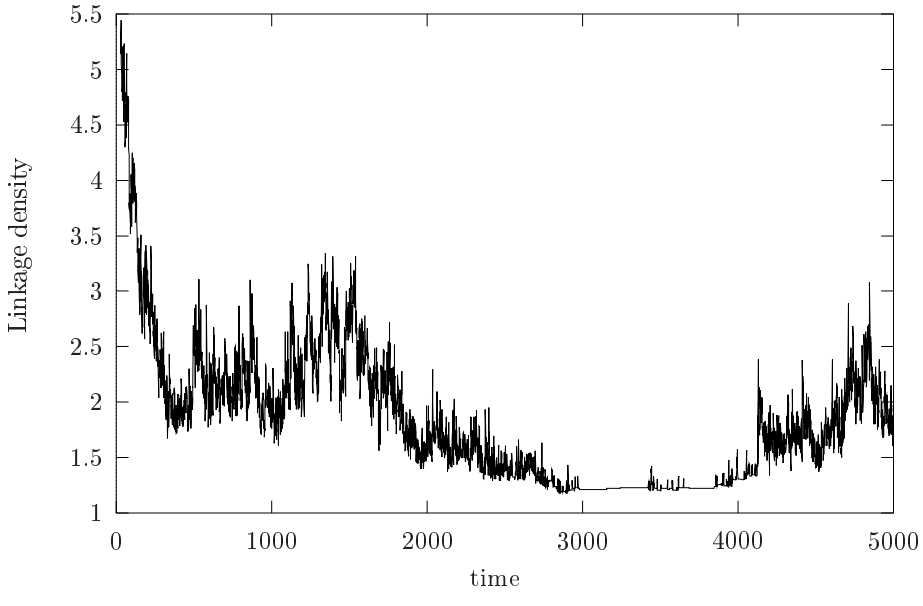

Figure 5.1: *The linkage density ℓ as a function of time. The system reaches a very stable period at $t \simeq 3000$, but looses it ultimately again. $N = 150$*

5.1.1 Linkage density

As noted in section 4.2.2 on page 58, May (May, 1972) obtained some analytical expressions for the relation between a given connectivity C and system size N. Assuming that Lotka-Volterra type equations describe ecosystems well, one expects a constant linkage density $\mathcal{L} = NC$, and thus a hyperbolic relationship between C and N. As can be seen in figure 5.1 the linkage density fluctuates from around 1.2 to 3 for a quite large system with $N = 150$ species. And this range in which the linkage density fluctuates does not change for increasing or decreasing N. Thus, as expected from the normal Lotka-Volterra equations. this model also has a constant linkage density (and therefore a hyperbolic relationship between C and N).

For the present model, the linkage density is also more or less constant, although there seems to be a small trend towards lower numbers for increasing system size, see the lower curve of figure 5.2 on the next page. But this trend is so small, especially when considering the large variance in the data[1], that is

[1]The number of samples used in order to obtain just one point in the figure (and also the

seems plausible to interpret the average linkage density, $\mathcal{L}$, as constant around 1.5 for all N.

The average food chain length, ℓ, as shown in the upper curve of figure 5.2, follows the linkage density quite closely. The experimental literature has an average linkage density of $\sim 2-5$ and an average food chain length of $\sim 3-6$. Considering that these data are obtained for ecosystems of sizes 30 to 120, where the largest variance is found, the predictions of the present model are quite fair.

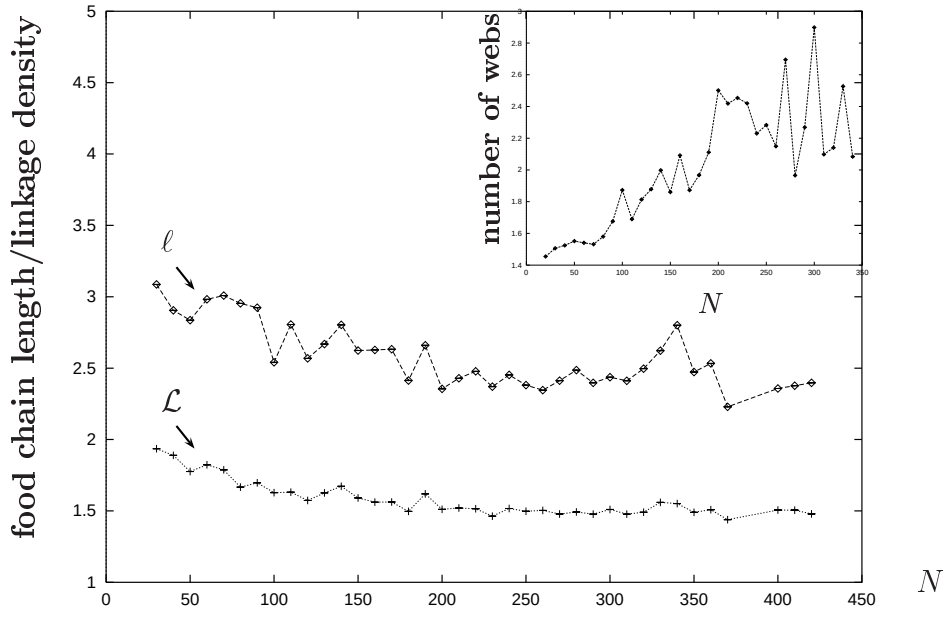

Figure 5.2: *The linkage density (lower curve) and average food chain length (upper curve). The simulations show that larger systems does not mean longer food chain lengths and larger linkage densities. On the contrary, there is all small trend towards lower values for both for large N. Insert: the number of isolated webs in the system grows with N.*

5.1.2 Top species

The top species are defined as species which feed on other species, but are not themselves fed upon. I the model this only means that one has to count

following figures) is *at least* 25.000. For a lower value the results become even worse than it is the case already. Also, in all following simulations we have set the probabilities of mutations in interaction strength (matrix elements in B), p_b, and interaction partner (matrix elements in A), p_a, to 0.0002 and 0.0001 respectively.

the number of species which have all zeroes in their column of matrix A, see eq. 4.10. The newly introduced species after an extinction event are necessarily top species since no other species in the community has yet had the possibility to utilize them as a food source. First after some time, mutations and later introduced species can eventually change these top species to intermediate or basal species. Figure 5.3 shows how the system develops though a transient period of an increasing number of top species until it reaches saturation.

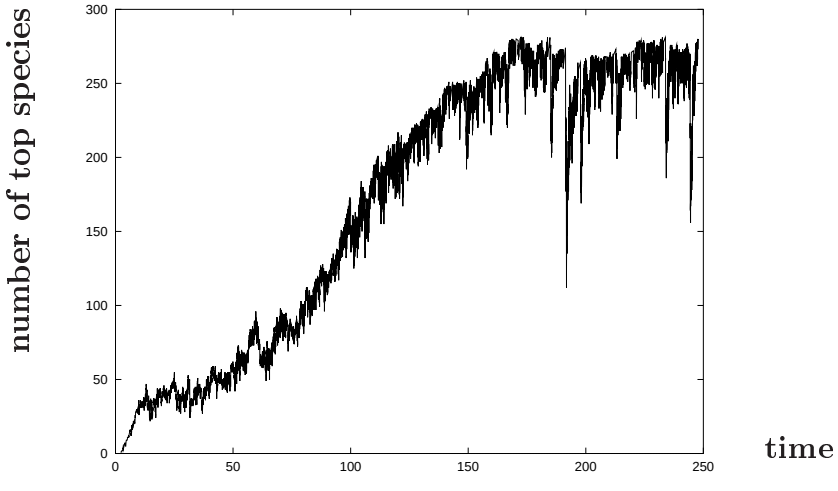

Figure 5.3: *The build-up of top species in the beginning.* $N = 400$, $p_f = 0.001$ *and* $p_i = 0.0002$.

But how does the number of top species depend on system size? The main part of figure 5.4 on the next page shows that the number of top species on average is proportional to system size in such a way that at least $N/2$ of all species are top species most of the time. Although the governing equation does not give any direct indications of the fact that having other species dependent on you is unfavorable for your growth rate (it's only a sum of the species *you self* depend upon), the dynamics shows nevertheless strong preference towards the strategy of being on top of the others. This also makes the structure of the food web very "bushy" in the sense that there are many "twigs" (links) going outwards from relatively few autotrophic basal species.

5.1.3 Basal species

The basal species however accumulate much slower through time. One reason of course is that new invaders have a large probability of having many interactions

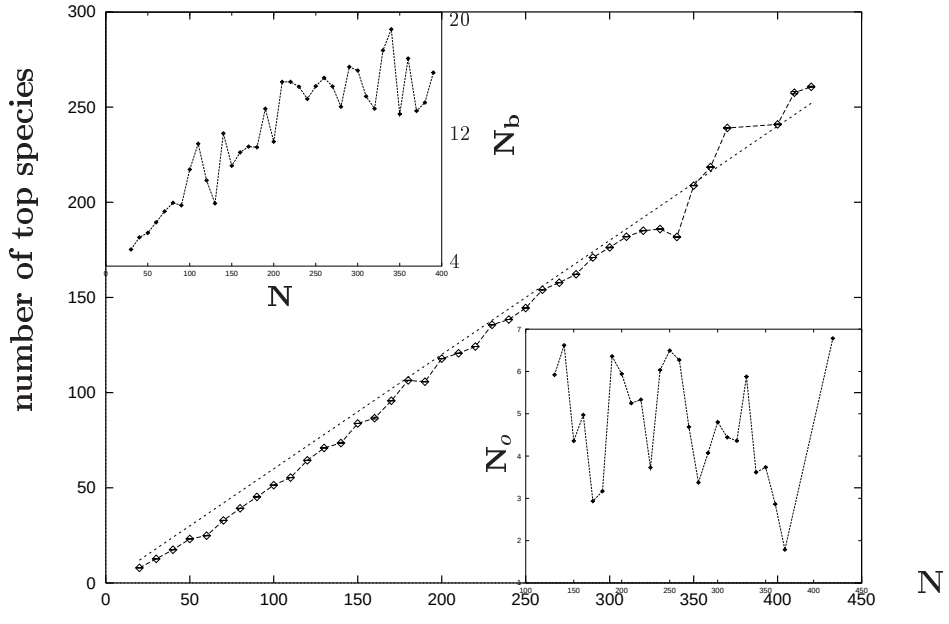

Figure 5.4: *Number top species as a function system size N. Best line fit for top species: $N_t \sim 0.6N$. Insert upper left: average number of basal species. Insert lower right: the average number of omnivores is highly variable but low for all system sizes.*

in the beginning. A more realistic assumption would be that the number of feeding links for a invader is a uniform distributed random number between one and M, where $M \ll N$. Thus, we actually have changed eq. 4.11 on page 66 to

$$K_i = \sum_{j=1}^{N} A_{ij}, \qquad 1 \leq K_i \leq M \tag{5.2}$$

where $M = 20$ for the simulations in this chapter. This is reasonable because in most cases new species only can interact, or feed upon, a small fraction of the whole community. The exact setting of M is however a matter of taste or convention, but the lower we set M, the shorter transient periods there are in the system - mainly because the basal species with only one interaction (in the trace of matrix A) are emerging more frequently the lower M is.

All the simulations show that the number of basal species is much lower on average than the number of top species. This is also known from experimental data.

Allthough it is difficult to obtain high quality data, as can be seen in the

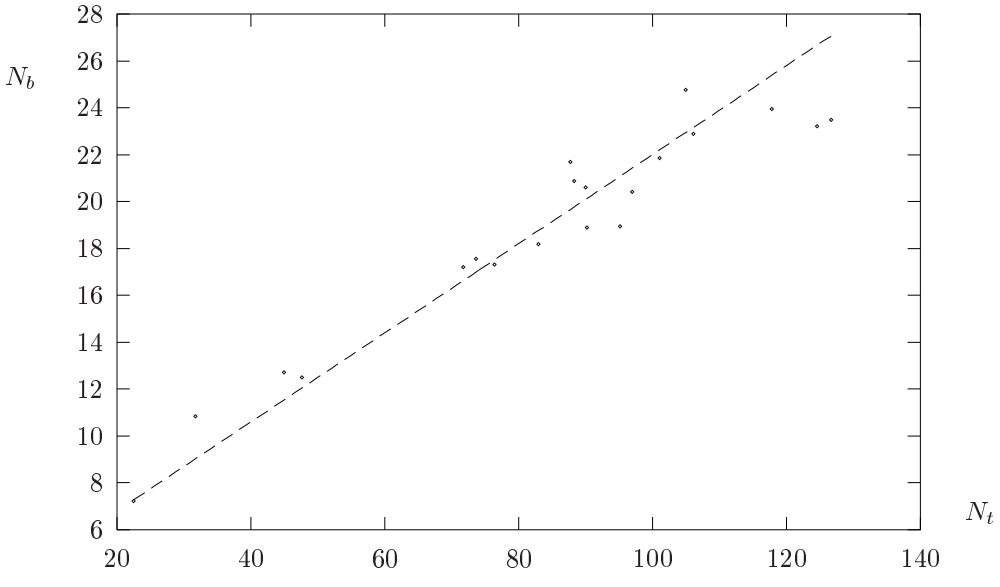

Figure 5.5: *The number of basal species as a function of the number of top species. The fit satisfies* $N_b = 0.19N_t + 3$.

upper left of figure 5.4. The reason is that the transient time for the average number of basal species increases rapidly for large N, but when plotting the number of top species against the number of basal species, as in fig. 5.5, one can adjust for the bias in transient times for different system sizes. The figure show that the number of top and basal species follow each other roughly linearly for increasing system sizes. This means that for each basal species there are approximately a constant number of top species associated with it, on average. It follows that the other properties, such as chain length and linkage density remain roughly constant for increasing N (as we have seen already).

5.1.4 Omnivores

The average number of omnivores, as defined in sec. 4.5.2 on page 69, remains low for all N. Although it still needs some more investigations, the simulations indicate that if the number of omnivores is high, the systems becomes more prone to extinctions. This has also been a subject in the literature (Pimm and Lawton, 1977; Pimm and Lawton, 1978), where it has been argued that systems with many trophic levels (more than three) get strongly

destabilized in the presence of omnivory. Later Yodiz (Yodiz, 1984) showed that the rarity of omnivore links in real webs could be accounted for by the lack of animals that feed on both plant and animal tissues. In any case, the model predictions agree with the experimental data that omnivores are rare, but (again) quite variable (Hall and Raffaelli, 1993).

5.1.5 Linkage strength

One of the motivations behind the splitting up of the matrix elements of section 4.4 on page 64 was that we then were able to investigate and quantify two distict properties immanent to species interactions: the A'_{ij}-elements tell us who interacts with whom and with how many, and the B_{ij}-elements tell us about the *strength* of these interactions. We expected that this strength changes, depending on which species we look at. Basal species which only "interact" in one link with themselves (rather: the autotrophic basal species which utilize external sources such as the sun and the soil) and species with only a few links are expected to have strong interactions (large values in B), against which the top species - normally with many links - are expected to have weaker connections.

All simulations show that this is correct. On average the basal species have an interaction strength $S_b = 0.60$, while the top species, on average, have an interaction strength $S_t = 0.47$ (remember that the random numbers thrown into B are uniformly distributed between zero and one). This means that there is a relatively strong selection process on the kind of interactions on the species, especially the basal species.

5.2 Extinction statistics

In this section we first investigate the survival of the species in the model in terms of survivorship curves and power spectra of extinction patterns. Then, we make a histogram of the extinction distributions, which yields an exponential dacay, but with the important recognition that such distributions are a result of an intrinsic *self-organization*.

5.2.1 Van Valens constant extinction law

If one measures the rate at which the whole group of species in the system present at a given time disappears over the subsequent time, one can obtain a feeling of the episodic character of the extinction events.

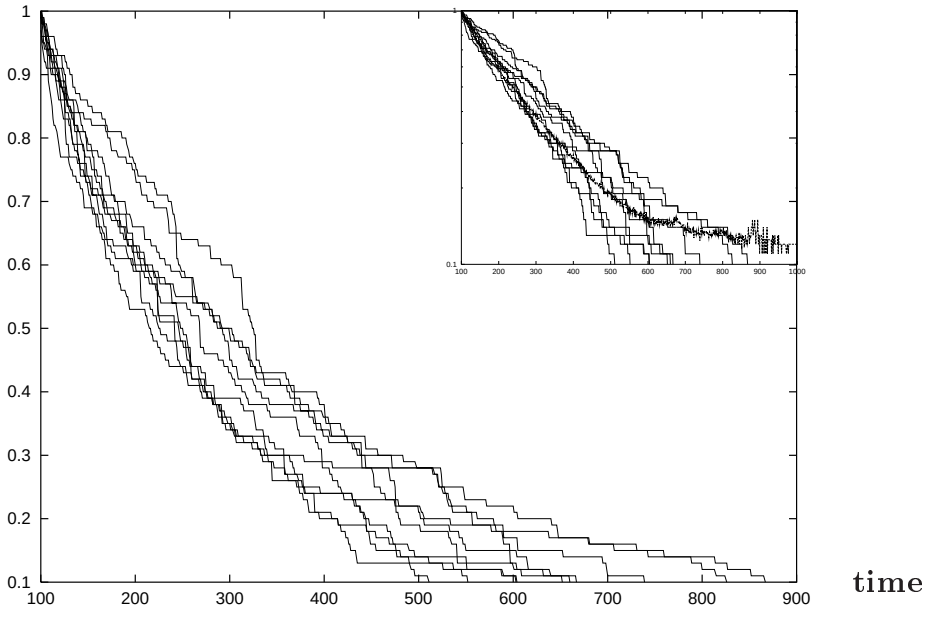

Figure 5.6: *Ten individual survivorship curves showing the percentage of survived species within a given time interval.* $N = 100$. *The insert shows the same figure in semi-logarithms indicating an exponential decay. The smooth line is an average over* 100 *runs.*

It has for some time been believed that there exists a continuous rate of "background" extinctions (Van Valen, 1973), known as *Van Valens constant extinction law*. Empirical samplings showed that this law could be expressed by an exponential decay with

$$N(t) = N_0 e^{-qt} \tag{5.3}$$

where N_0 and $N(t)$ are the number of survivors at some initial time and at time t, respectively, and q (with a best fit value of $q = 0.13$ (Raup, 1986)) is the probability of extinction of a species per million years. But when one looks at the data through smaller time intervals it has been possible to identify the apparent continuous background extinction rate as an episodic pattern (Raup, 1986), where long periods of stability without extinctions are interrupted by extinction events on all scales. The graphical representation of this process is called

a *survivorship curve*, and it is shown for the present model in figure 5.6. If one uses longer time intervals for the empirical sampling points (Van Valen, 1973), the episodic character disappears and the linear interpretation of eq. 5.3 becomes valid.

Our model is consistent with all these experimental investigations. Extinctions happen in bursts, separated by periods of stasis. Averaging over many simulation (just with different seeds for the random number generator), it is also possible to obtain an approximate exponential decay consistent with eq. 5.3, and with $q \simeq -0.004$, where q is the probability of extinction within one time unit. In the initial transient periods, where the system still is in an unorganized structure, the frequency of extinctions is much higher, although the episodic pattern remains. The exponential decay is in this case given by the value $q \simeq -0.01$.

5.2.2 Extinction patterns

The extinction pattern in the present model exhibits what is called "1/f-noise", see figure 5.7 on the following page. $1/f$-noise is a commonly used label for a certain type of time correlations found in many different real-world time series such as the flow in the river Nile, pressure variations in them air caused by music, sunspot activity, fluctuations in the electrical resistance of a conductor, etc... in spatially extended systems, this behaviour leads to self-similar fractal structures (Bak et al., 1987; Jensen, 1998).

"$1/f$" refers to the fact that low frequency power spectra of such systems display a power law behaviour $P(f) \propto f^{-\beta}$ over a long range of time scales. The "special" thing with $1/f$-noise is that there exists no general theory that explains its occurence (although *self-organized criticality* has tried to be a candidate, see part III).

Figure 5.7 shows the Fourier transform of the autocorrelation function of a time series of extinction events for a particular simulation of the model with mutation probabilities in the matrix elements A_{ij} and B_{ij} of $p_a = 0.0001$ and $p_b = 0.0002$ respectively. In this double logarithmic plot, the straight line has a slope of $\beta = 1$, consistent with the "1/f"-label. We therefore expect long time correlations in the extinction events, probably because of long time stability of certain food web structures leading to some scale free cascades of extinction when a basal species or an intermediate species population drops to zero.

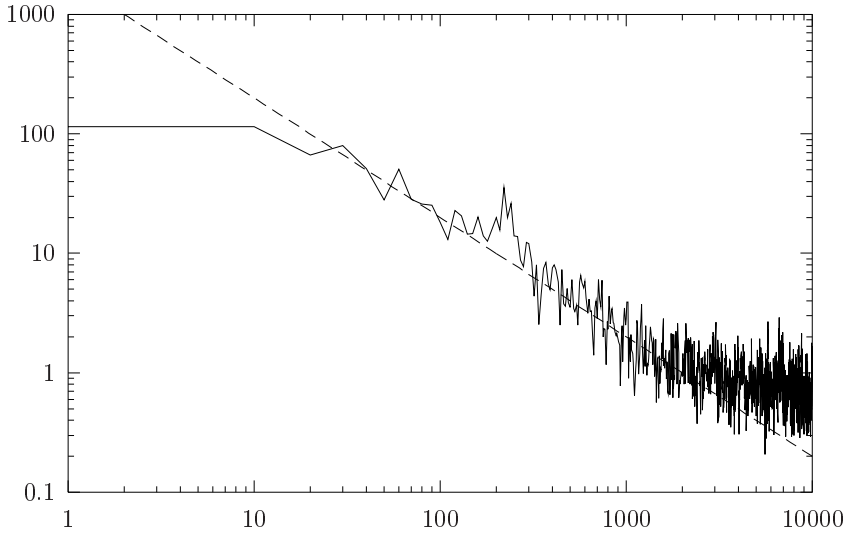

Figure 5.7: *The extinction pattern shows a power law decay with an exponent $\beta = 1$. It is calculated as the Fourier transform of the autocorrelation function, and the data point are collapsed in windows of ten. $N = 50$*

5.2.3 Extinction distributions

The extinction distributions are, in contrast, exponentially decaying. In figure 5.8 we have shown extinction distributions for various system sizes, and for $N = 350$ the distribution is markedly different from all the others. The reason for this is that the simulation still hasn't left its transient phase (which typically is much longer for large systems). This means that the extinctions still are so frequent at every time step that there are almost no small size extinctions[2]. The food web is still "under construction".

But from this observation, it is possible to conjecture that the food webs (and the thereof resulting extinction patterns) *self-organize to a critical state characterized by an exponential decay in the distribution functions and a power law decay in the power spectrum of the extinction patterns.*

[2]For the other simulations in the figure with other N, the same effect appears when the simulations aren't run long enough

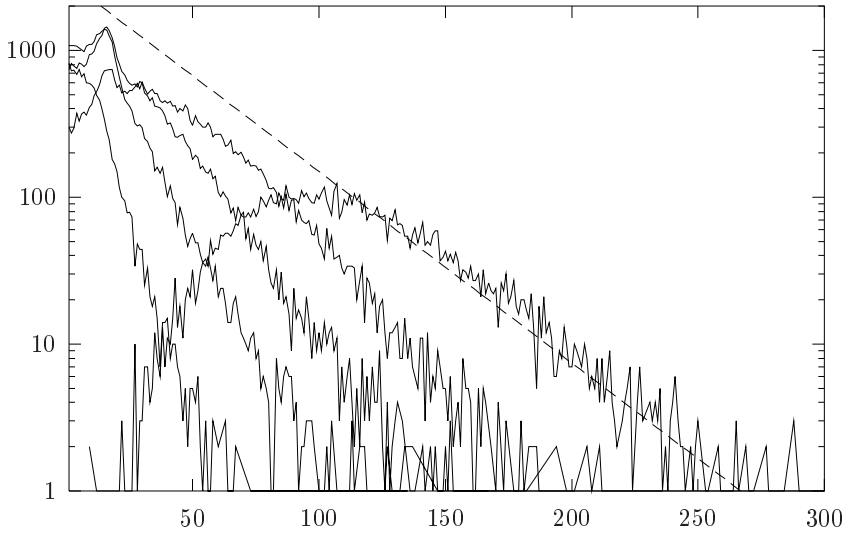

Figure 5.8: *The extinction distributions for various system sizes. From the fastest exponentially decaying curve one the most left to the right we have:* $N = 50, N = 100, N = 150, N = 200$ *and* 350. *The situation for* $N = 350$, *which not yet has reached its self-organized critical state is described in the text. The straight dashed line indicates an exponential decay with* $q = 0.03$.

5.3 Generalizers vs. specializers

In the conclusion of the last chapter it was mentioned that we expect a difference in average lifetimes between specializers and generalizers. It was argued that since specializers are more vulnerable to perturbations and extinctions in the system, they are exchanged with new species more frequently than generalizers. When looking at the data this is shows to be true. but there are also some minor surprises.

First of all, evaluating the average lifetimes, it becomes evident that invading species with a low number of links (the specializers) are far more abundant than species with many links[3]. This is clear, because we already know that the average linkage density $\mathcal{L}$ is normally quite low. But for how long, on average, do the individual species survive when they are classified by their number of connections?

The results are shown in the histograms of figure 5.9. Remember that an in-

[3]The typical scenario is of course that most of the invaders will get kicked out of the system at once, and their lifetimes are therefore equal zero. Only species with a lifetime longer than $\delta t = 1$ will count in the statistics.

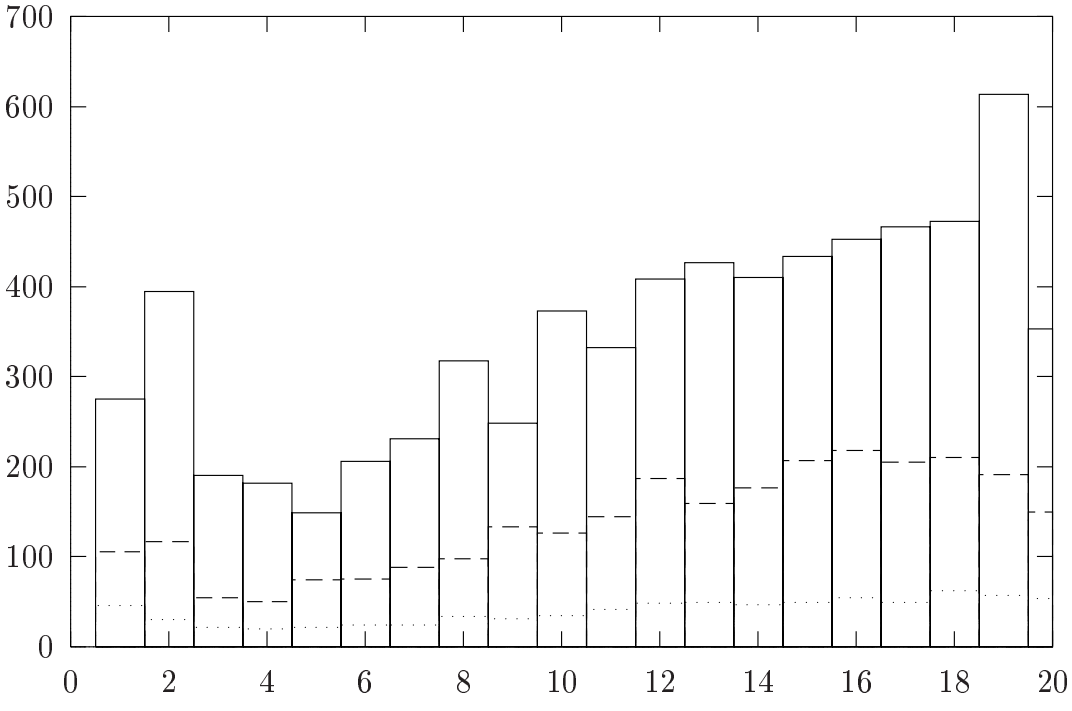

Figure 5.9: *The average life time of species as a function of their number of links. Species with a low number of links can be classified as specializers, and species with a high number of links as generalizers. Three histrograms are shown corresponding to three populations with* $N = 150$ *(full line),* $N = 200$ *(dashed line) and* $N = 250$ *(dotted line).*

vading species never can have more than 20 links in the beginning (a restriction imposed by the creator in order to minimize the transients in the beginning of a simulation). The figure shows three different histograms representing a simulation with the system sizes $N = 150, N = 200$ and $N = 250$ respectively. The first clear observation is therefore that smaller systems have longer average lifetimes, indicating greater stability. Thus, for systems with $N = 50$ the normal lifetimes are several thousand timesteps, while for $N = 400$ or more they are around ten or less. Smaller ecosystems imply longer lifetimes.

The next observation from figure 5.9 is that generalizers in fact do live longer that specializers. There cannot be very many of them, but when are successfull, they do great. The reason is that extinctions of their resources does not affect their survival in the same degree as it does for the specializers. But one surprise is constantly present in the model: species with exactly two trophic links do comparably better than both species with only one trophic link

or species with three. Moreover, for smaller systems, this trend becomes even more prominent, so that for $N = 100$ or less, the average lifetime of species with two links is longer than all other species. This is a peculiarity, for which we not yet have found a good explanation.

5.4 Lifetime distributions

However, evaluating average lifetimes for a process which contains a distribution of lifetimes on all scales might seem as a vulgar thing to do. Instead, it would be more sensible to plot the total lifetime distributions on a log-log graph and extract information from that. The result is shown in figure 5.10 where we have plotted the overall distribution of lifetimes of all species for a system of size $N = 150$. The figure shows a very clear power law decay $S(t) = t^{-\kappa}$ with an exponent $\kappa = 1.5$.

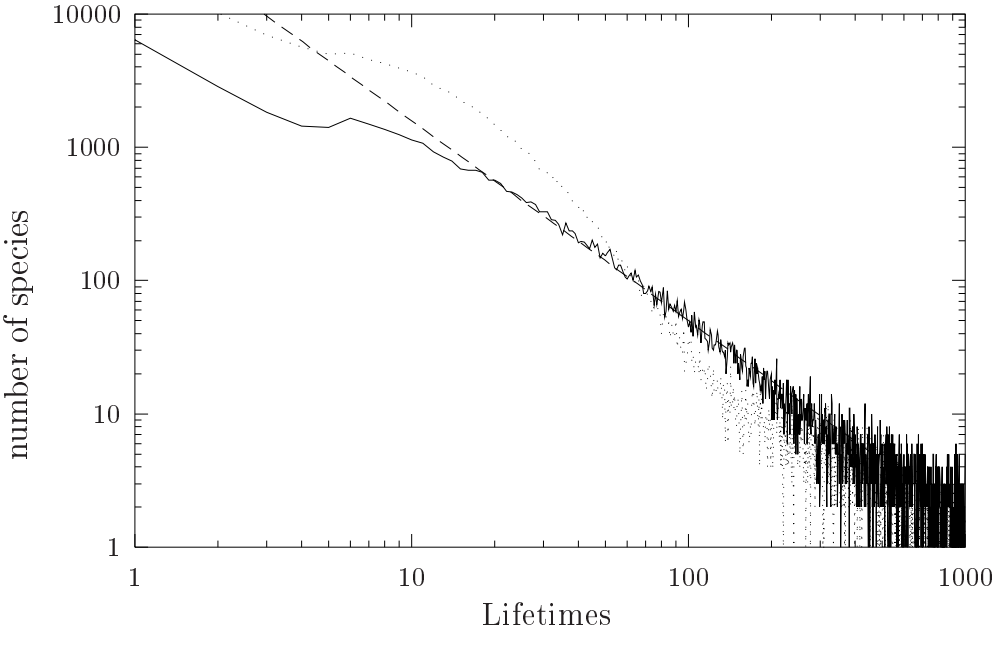

Figure 5.10: *The full line shows a histogram of lifetime distribution of the species in a population of size $N = 150$. The dashed fit has an exponent $\kappa \simeq 1.5$. The dotted line shows another distribution with $N = 150$ and an approximate exponent of 3.*

This is an interesting result, first of all because it indicates scale free be-

havior: the duration of life appears on all scales, from the geological second of human life to the seemingly eternal life spans of cockroaches. But it is also interesting because it is compatible with the experimental results by (Raup, 1991), where the exponent κ was measured to be around 2. Furthermore, it turns out that the size of the exponent κ in our simulations is dependent on N. For a systems size of $N = 100$, κ becomes $\simeq 1.3$, and for a system size $N = 200$, we have $\kappa \sim 2$ and for $N = 350$ $\kappa \simeq 3$ (of which only the last case is shown in fig. 5.10).

5.5 Summary

In this chapter we have investigated the statistical properties of the model defined in the previous chapter. The results are the following:

- The system undergoes a *spontaneous* reorganization from an initially randomly wired ecosystem to a well-defined food web.
- Loops become very rare so that the links which connect the species becomes *directional*. This directional path from an autotrophic basal species to the top species, the *food chain length* as well as the linkage density is typically small.
- The linkage density of the species and the average food chain length correlated very little with system size. Instead, if N is large, the system will typically split up in separate subsystems, so that the relative stability is maintained within that part of the community. This means that the initial N coupled differential equations become a set of coupled differential equations each of them containing only a fraction of the N species.
- The number of species feeding on more than one trophic level, the so-called omnivores, is small.
- The interaction strength defined as the absolute value of the matrix elements B_{ij} is, on average, much stronger for basal species (0.60) and species with only a few trophic links, than for top species (0.47) which typically have more links to fed on.
- The extinction statistics follows very well the *constant extinction law* of Van Valen, showing an episodic pattern in the survivorship curves with

an average that has an exponential decay. Extinctions happen in bursts, separated by periods of statis.

- The *temporal* extinction pattern follows another power law, called "$1/f$"-noise with an exponent β very close to unity. Thus, the temporal extinction pattern is self-similar and long time correlated.
- The extinction distribution follows an exponential decay. However, as long as the system still is in its transient period, the extinction distribution deviates strongly from an exponential form.
- The lifetime distributions follow a power law shape, and the exponent changes dependent on the system size. For $N = 150$ the exponent is $\kappa \simeq 1.5$ and for $N = 250$ it is $\kappa \simeq 3$.
- The model ecosystem develops different classes of species, which can be classified in *specializers* and *generalizers*. As expected, the generalizers live longer on average, because they are more stable against changes in the food web structure: If a generalizer looses one of its food sources, it still can survive by feeding on the other, but a specializer, which normally only has one or a few species to feed upon, will encounter a huge risk of getting extinct in this case.

5.6 Comparison with experimental data

Now we finally come to the interesting question: how sensible are the predictions of the model? Generally speaking, the properties for the food web statistics are in very good harmony with the experimental data. We have compiled a similar table as before containing the best numbers from the experimental data and the results from the present simulations.

Table 5.1: **Comparison of model an real life, part I.**

Property	*Model predictions*	*Data*
Linkage density, $\mathcal{L}$	Constant between 1.5 and 2 for all N with moderate variation. From this follows that the connectance $C = \mathcal{L}/N$ is a hyperbolic function for increasing N.	Numbers from $2 - 5$ are all consistent with the data. The constancy of the linkage density has been a major believe, but some studies (Cohen et al., 1990) show a slight increase for large webs.

Basal species, N_b	Remains comparatively low - around 5% of the community, but highly variable. Not very dependent on N.	Independent of N and with high variance in the data (Briand and Cohen, 1984; Cohen and Briand, 1984).
Top species, N_t	Generally more than 50% of the species are top species, and $N_t/N_b \simeq constant$.	High variance, and the ratio $N_t/N_b \simeq constant$ (Evans and Murdoch, 1968; Jeffries and Lawton, 1985).
Number of omnivores, N_o	Surprisingly few. Probably destabilizing.	Less common in real webs than in randomly generated webs (Pimm, 1982). Destabilizing (Pimm and Lawton, 1977; Pimm and Lawton, 1978).
Food chain length, ℓ (lower bound)	Average food chain lengths are relatively independent of N (all thought with a small trend towards smaller ℓ for increasing size), with average values between $1.5 - 3$, but very variable between $1.3 - 10$.	Typically short, from 3 to 6. Very few feeding loops (Pimm, 1982; Cohen et al., 1986).
Van Valens constant extinction law, $N(t)$	$N(t) = N(0)e^{-qt}$ for average values (or large sampling windows), but when narrowing down the sampling window, and looking at the survivorship curves, one finds pronounced stasis interspersed with major extinction events	Same effect in the observations (Van Valen, 1973; Raup, 1986).
Extinction patterns, $P(f)$	$P(f) \propto f^{-\beta}$, where $\beta \simeq 1$	$P(f) \propto f^{-\beta}$, where $\beta \simeq 1$ (Solé and Manrubia, 1997)
Extinction distributions, $S(m)$	The distribution of extinctions of size m follows an exponential decay and is dependent on system size.	Follows a power law decay: $S(m) \propto m^{-\alpha}$ with an exponent $\alpha \simeq 2$ (Newman, 1996; Solé and Bascompte, 1996).
Lifetime distributions, $S(t)$	Follows a power law decay with varying exponent dependend on N.	Power law decay $S(t) \propto t^{-\kappa}$, where the exponent $\kappa \simeq 2$ (Raup, 1991).

One of the most important points is the low average value of $\mathcal{L}$. Many data sets support the prediction of a low linkage density (average values of $\mathcal{L} = 1.9$ in the catalogues of (Cohen et al., 1986) and (Briand, 1983); 2.2 in insect webs (Schoenly et al., 1991), and 3.5 in pelagic webs (Havens, 1992)). But it has frequently been argued that the detectability of feeding links in real ecologies is very low, so that the actual value of $\mathcal{L}$ might be substantially higher (Cohen and Newman, 1988; Kenny and Loehle, 1991). One study (Goldwasser and Roughgarden, 1997) has even tried to estimate the effect of sampling intensity on varying food web properties, with the conclusion that the detection of trophic links (among other food web properties) is in strong correlation with the sampling intensity. However, it is not clear whether these infrequent links are ecologically relevant, or whether they just express partial preference of resource utilization without vital importance (Futuyma, 1986; Pimm, 1991).

Regarding the high number of top species, there actually exist no good estimates of an average number. The numbers of top species obtained in some of the newer data sets show ranges from zero top predators for the data from the Coachella Valley (Polis, 1991) to 65% in some insect food webs (Schoenly et al., 1991). The number of basal species is always lower than the number of top species. In most cases their percentual share is low (4% in the Ythan estuary (Hall and Raffaelli, 1991), 9% in the Yorkshire pond (Warren, 1989), 13% in Little Rock Lake (Martinez, 1991)), but it also can be high (50% in pelagic webs (Havens, 1992)) or highly variable ($2-31\%$ in insect webs (Schoenly et al., 1991)). Most of these data are taken from (Hall and Raffaelli, 1993).

5.7 Discussion and conclusion

Unfortunately, the experimental data on food web properties are very sketchy. Moreover, many ecologists express a big distrust towards the observed patterns - either because they think that the drawing up of webs is biased by subjective and casual sampling procedures (Paine, 1988), or simply because limitations in the data might result in artifacts (Polis, 1991; Lawton and Warren, 1988; Winemiller, 1990; Kenny and Loehle, 1991). This is quite a big problem. If the observed patterns were not very sensitive to incompleteness of the data, then the data could be used with some confidence, but if the incompleteness causes misinterpretations, then it might be that several existing theories may be attempting to explain properties that do not exist at all (Lawton and Warren, 1988;

Goldwasser and Roughgarden, 1997). Many of the clever theoretical constructions, including the one in this thesis, would hang like skyhooks in the air, without any roots in nature.

However, in relation to the present model formulation, it is also possible to give this problem a positive slant: if the incompleteness of the data mainly is a result of high actual variations, and not because of subjective or casual experimentalists, then it might be that the present model has captured an essential feature of real food web - namely the large actual variations. From all simulations in this chapter, the most striking feature has been the need for *very many sampling points* in order to obtain reliable average numbers. So the model actually *reflects* the data in this respect.

In conclusion, the food web model presented here looks as if it could become a good canditate for dynamical understanding of the spontaneous self-generation of food webs. Most of the here sampled data points to the conclusion that there might be some reality behind the approach. Naturally, there are a huge range of problems and properties which not have been investigated here, but they will hopefully be addressed in some forthcomming works.

Part III

Criticality in Evolution

6

Percolation, Fractals and Criticality

In this final part, we try to give a heuristic account of why the two models discussed so far might be linked to the emerging theories of percolation theory, critical phenomena and the idea of self-organized criticality. After introducing the main concepts of percolation theory and the theory of critical phenomena (regrettably most superficially), we make an analytical calculation of a percolation threshold in the model for neutral evolution and discuss the potential of the model to exhibit self-similar patterns. Then we discuss the food web model in terms of self-organized criticality an find it to be a good candidate for such a notion.

In the final chapter we lead the attention to possible problems with the uncritical identification of power laws to self-organized phenomena. By formulating a simple model for evolution - with some quite restrictive assumption - we find power law behavior without any connection to neither self-organization nor criticality. It is argued that one has to be very careful with identifying scale invariance with only one fundamental physical process, because random stochastic effects also are able to do the job.

6.1 Percolation theory

Percolation theory is becoming increasingly important for the understanding of non-equilibrium dynamics and critical phenomena in nature. As soon as many components are present in a random system and the richness of the interconnections of these components can be varied, we have a kind of percolation problem.

But what is percolation, and what is a percolation threshold? The basic idea can be explained by a simple example of *bond percolation* on a square lattice.

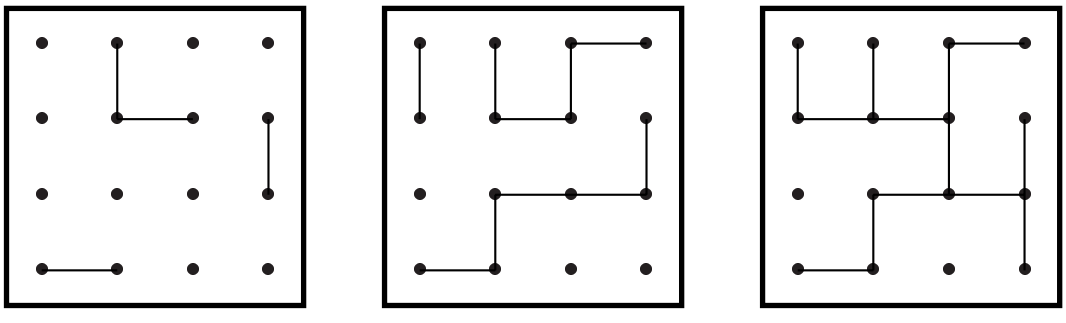

Figure 6.1: *Illustration of bond percolation on a square lattice. If the density p of bonds is below p_c, many small clusters are formed. If $p > p_c$ it becomes very probable that a giant cluster occupying a large fraction of all sites forms. At $p = p_c$ however, there exist clusters of all sizes.*

Consider an infinitely large lattice (of which there are shown three in figure 6.1) upon which an unidentified random object (URO) has thrown bonds, or bridges, between the sites, so that neighboring sites which are connected through these bridges form what is called *clusters*. All sites within one cluster are thus connected to each other by one unbroken chain of bridges which go from neighbor to neighbor to neighbor to etc...

For a low density p of bonds - they could represent electrical wires - only small, finite and localized clusters are formed with no chance to lead any current from one border to the other. The system is an insulator. But for a critical number of bonds, suddenly a *percolating* cluster forms which can transport the current throughout the system, and suddenly you have a conductor (in figure 6.1 there are shown three clusters for $p = p_c$ of which one is infinitely

large and spans the whole system). For even larger p the system typically has a only a few but very large clusters, called giant clusters or "common networks" as they were defined in the model of neutral evolution in part I.

So, the basic idea of percolation is the existence of a sharp transition at which the long-range connectivity of the system appears (or disappears), and this transition occurs abruptly when some generalized density in this system reaches a critical value p_c, the *percolation threshold*. The importance of percolation theory emerges from the fact that the change from having a finite cluster to having an infinite clusters is the *geometric analog of a phase transition*, and thus, many thermodynamical properties can be explained by this geometrical viewpoint of individual molecules connecting and dispersing. What is often visible from the outside is a sharp change in the properties of a substance: The transition from a liquid to a solid, gelation, transition from a conductor to a superconductor, from a paramagnet to a ferromagnet, etc.. (Yeomans, 1992). As soon as there is a singularity in the free energy of a system or one of its derivates, the system might encounter a percolation threshold at the microscopic level of molecules within a certain geometry.

It is not the purpose of this thesis to give a detailed description of percolation and its exact solutions for several lattice types (The reader should refer to (Stauffer and Aharony, 1994) for a good introduction to the subject). It is however interesting to note that many important quantities (such as the mean cluster size or the probability for an arbitrary site to belong to an infinite cluster) decay as *power laws* close to the percolation threshold.

6.2 Critical phenomena

These power law decays are examples of *critical phenomena* also seen in thermal phase transitions where quantities of interest go to zero or infinity by simple power law decay.

Historically, scientists have focused mainly on equilibrium systems were many-particle system have shown that correlation functions and distribution functions typically decay exponentially at distances or times longer than a finite correlation length ξ or correlation time τ. Only at critical points they have observed infinite correlation length and times, often resulting in "scale invariance" and "self-similarity" as a side-effect to a power law decay (which refers to the fact that power laws, unlike exponentials, lack a characteristic length that sets the scale for the decay (Grinstein, 1995)).

Since chemists and physicists are so familiar with equilibrium systems, they have a sense of an abnormal and fragile situation when confronted with critical phenomena. This is maybe so because under normal laboratory conditions (where equilibrium is seen as a kind of "ground state") phase transitions only can be achieved by fine-tuning some thermodynamic variables (for example the temperature) to their critical value. But when looking at the natural world, the opposite is true: scale invariance is recognized to be truly abundant, or even *ubiquitous*. When looking at trees, rivers, mountains and costal lines, or at the less poetic resistors, cheese, luminosity of stars and extinction patterns in evolution, they all show some kind of power law behavior, sometimes termed as "fractals" (Mandelbrot, 1977) or "1/f-noise", and always scale invariant and self-similar in nature.

How can it be that scale invariance is so abundant in the real world, while almost all our theoretical machinery on equilibrium systems (which satisfy detailed balance for an underlying Hamiltonian) predict exponential decays, at least when there is no parameter tuning? As you might have guessed, recent investigations answer this question with the growing interest in *non-equilibrium systems*, that is, open systems which are driven externally so that they never achieve equilibrium. The equations representing such non-equilibrium systems, of which the *Langevin equation* is a much analyzed member, are much more capable of scale invariant behavior, see an excellent survey in (Grinstein, 1995).

The recent interest in scale invariance and the huge effort of scientists in many diverse fields to find power law scalings in their data must be seen in the light of this ubiquitousness of scale invariance in the real world. But, as it will become clear soon, the sole existence of power law distributions in various data does not necessarily tell that there is *a generic reason* for them.

Sometimes it is possible to get a feeling that all these scale invariant power laws *mean something*. It has at least happened quite often for the present author that a lecture on some interesting phenomena ends in the triumphant revelation of a power law. Either it is found in simulations, in some experiments or in an analytical work, but in any case, you sit there, somewhere in the back of the auditorium, and marvel about why this should be so significant and wonder about this inscrutable mystery of life.

Physicists seek for unification of their theories, and it is therefore quite understandable to ask if there might be any fundamental "force" which produces all these power laws by nothing but one (physical) process. If it were possible to have a glimpse of this one-and-only force, this URO, and maybe even name it by its own name, you would be guaranteed a lot of attention from the

scientific community.

6.3 Self-organized criticality

The idea of *Self-organized Criticality* (SOC) is one of such candidates. It must clearly be recognized as one of the more interesting conceptual ideas by which to understand the observed scaling phenomena in complex systems.

Initially self-organized criticality was introduced by Bak, Tang and Wiesenfeld (Bak et al., 1988; Bak et al., 1987) in order to understand 1/f-noise, but it became quickly the main hypothesis which also tried to explain the general ubiquity of scale-invariant phenomena in nature. So, self-organized criticality was applied to a wide set of dynamical systems far from thermodynamic equilibrium such as sandpiles (Frette et al., 1996; Feder, 1995), earthquakes, solar flares (Luand and Hamilton, 1991), vortex-creep in type-II superconductors, traffic jams, biological evolution (Bak and Sneppen, 1993; Flyvbjerg et al., 1993; Bak et al., 1994; Sneppen et al., 1995; Sneppen, 1995), extinctions in the fossil record (Solé et al., 1997) and recently for neuronal assemblies (Jung, 1997) in the human brain.

But what is the philosophy behind SOC? The initial description (Bak et al., 1987; Bak and Sneppen, 1993) was somewhat vague in its reference to the "tendency of large dynamical systems to organize themselves into a *poised* state far out of equilibrium with propagating avalanches of activity of all sizes". This poised, or critical, state is an attractor for the dynamics so that the scaling properties of the attractor are independent of the parameters of the model.

In recent years, however, it has become more and more clear that the true difference between the mechanism of SOC and other "ordinary" model systems for scale invariance (such as the Langevin equations) is a *fundamental separation of time scales*. Conventional non-equilibrium systems relax on a time scale comparable to or longer than they are perturbed. By contrast, self-organized critical systems are imagined to relax far more rapidly than they are perturbed (or driven). In fact, as pointed out by (Grinstein, 1995), complete scale invariance over all times (or lengths) is only obtained in the limit where the ratio between the perturbation time to the relaxation time goes to infinity. For all situations, where this ratio is large but finite, there will exist a correlation length ξ beyond which the decay becomes exponential.

When pondering about the fractal structure of a mountain, it may be difficult to believe that such a seemingly eternal massive is far away from equi-

librium. But never the less, this is the case. The infinitely small perturbations through the tectonic motions and other geological forces will eventually lead to a abrupt disaster of unknown size. This separation of time scales is characteristic for all systems described by SOC[1]

With these characteristics in mind, we can go on and pose a hard question: Are the two models investigated in the previous parts self-organized critical? We will start to look at the neutral networks and make some analytical investigations of the percolation threshold. In section 6.6 on page 115 we turn to the question whether the food web model from part II can be called self-organized critical.

6.4 Critical neutral networks?

In the case of the model for neutral evolution we have found two power law distributions of the size of the networks: In figure 3.5 on page 37 the exponent for the frequency of networks of a certain size in the case of $K = 2$ was equal to -0.7, while for the totally uncorrelated system with $K = N - 1$ it was around -1.5 (but only for the networks below a size of 1000 sequences). Why is there such a discrepancy between these two values for the exponents? Are there two different mechanisms at work? At present, we have no answer to this question. However, there are some important points to note:

- As already mentioned in a footnote in section 3.3 on page 35, the normal NK model exhibits a distinct change in behavior of the attractor when changing K from three (chaotic motion) to two (ordered motion) (Kauffman, 1993). This indicates that there might exist a true phase transition close to $K = 2$.

- For the uncorrelated system with $K = N - 1$ every fitness value is completely random compared to all the other, and it is therefore more suggestive to attribute the scale free behavior of network sizes to a *stochastic effect*. In fact, since in this case the number of sequences with a certain fitness follows the Gauss distribution of figure 3.2 on page 34 and the number of networks with a certain fitness follows an "inflated" Gaussian of figure 3.3 on page 35, it should be a manageable job analytically to

[1]The reader of modern classical literature may now be able to recognize Self-organized Criticality as a known figure, or doctrine, in the Zen-trophic tradition. Brought down to its most simple form it says: "Shit happens", (and that in all sizes!).

verify the power law shape. Whether this really is possible remains to be seen in a forthcoming work, but it is important to note that power laws also are easily produced by random processes, and in the next chapter we will show a particular example of that.

- The all-important separation of time scales is not relevant for the static structure of the neutral fitness landscape. But for populations moving on such landscapes there is a clear separation of time scales, as seen in figure 3.12 on page 45. First, the population diffuses slowly on the neutral networks, and when it encounters a successful mutant, the whole population moves quickly onto that fitter network by the normal Darwinian selection process. But it is questionable whether any critical behavior emerges from this, because the length of the epochs correlates with the number of reachable networks which in turn is an exponentially decreasing function for increasing fitness.

6.5 Percolation threshold for the neutral evolution model

We now turn to some analytical results for the present model for neutral evolution. The idea is to define a percolation threshold at a critical fitness W_c, where the networks stop to be these giant clusters and instead become confined to local points in the fitness landscape (so that a population no longer can use them for diffusive searching). This is from an evolutionary point of view an important transition, because it puts a general limit on the positive effects of neutrality. Maybe it is easier to make this point clear when the reader (again) has a look on the figures 3.2 and 3.3 on the pages 34 and 35 respectively. These were the figures of the distribution of *sequences* as a function of their fitness and the distribution of the *networks* as a function of their fitness. The last figure has this "inflated" appearance because of the formation of giant clusters which span the whole sequence space, but at a certain point on the right hand side of this "inflated" Gaussian, the distribution of networks again follows an exponential tail. This suggests that there somewhere in this region is a sharp phase transition in the same way as for bond percolation.

We utilize the fact that in the case of $K = N - 1$ all fitness values are independent of each other: a mutation in one gene means that the configuration of all other genes also changes (see figure 3.1 on page 33). There is no

correlation between fitnesses due to successive point mutations, and in such an uncorrelated landscape all configurations are independent and the fitnesses are thus Gaussian distributed (in the limit of large N). Therefore, we can try to calculate the mean and variance of the general distribution function.

6.5.1 Probability distribution for the uncorrelated system

Let us initially assume that the fitness of the whole sequence is

$$W_u = \sum_i w_i \tag{6.1}$$

where W_u means that we have not normalized with the number of possible fitnesses $N(F-1)$. Now, since the neutrality parameter F is an uniformly distributed discrete number between zero and $(F-1)$ on each w_i, we can apply the central limit theorem and find the mean μ as a mean of the progression of natural numbers - since they are chosen independently at random (like from a dice with F faces):

$$\mu = \frac{1}{F}\sum_{n=0}^{F-1} n \tag{6.2}$$

where n is just an integer. This sum is a known series for natural numbers, namely:

$$1+2+3+4+...+n = \frac{n(n+1)}{2}$$

which evaluates to

$$\mu = \langle W_u \rangle = \frac{1}{F}\frac{F(F-1)}{2} = \frac{(F-1)}{2} \tag{6.3}$$

This is an expected result: the mean of $(F-1)$ independent equidistant natural numbers is just in the middle. The variance can be found by

$$\sigma^2 = \langle W_u^2 \rangle - \langle W_u \rangle^2 = \frac{1}{F}\sum_{n=0}^{F-1} n^2 - \left(\frac{1}{F}\sum_{n=0}^{F-1} n\right)^2 \tag{6.4}$$

where the first sum on the right hand side is another series for natural numbers[2], namely:

$$1^2 + 2^2 + 3^2 + 4^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$$

so that, if we set $n = F - 1$, the variance becomes

$$\begin{aligned} \sigma^2 &= \frac{F(F-1)(2F-1)}{6F} - \left(\frac{1}{2}(F-1)\right)^2 \\ &= \frac{(F^2-1)}{12} \end{aligned} \tag{6.7}$$

[2]If we could not read this nice formula from a book, we should do the following: a) Use the solvable *generating function* $S = \frac{1}{F}\sum_{n=0}^{F-1} e^{\beta n}$; b) which has an appropriate second derivate $\frac{\partial^2 S}{\partial \beta^2} = \frac{1}{F}\sum n^2 e^{\beta n}$, so that, c) when we let $\beta \to 0$, the second derivate becomes the sum in which we are interested in: $\frac{\partial^2 S}{\partial \beta^2} = \frac{1}{F}\sum n^2$. Now: how to solve S and find the second derivative? First we recognize $S = \frac{1}{F}\sum_{n=0}^{F-1} e^{\beta n}$ to be the geometric progression, so that

$$S = \frac{1}{F}\frac{e^{\beta F} - 1}{e^{\beta} - 1} \tag{6.5}$$

Instead of differentiating this expression directely, we can expand it as a Taylor series:

$$S = \frac{1 + \frac{1}{2}\beta F + \frac{1}{6}\beta^2 F^2 + ...}{1 + \frac{1}{2}\beta + \frac{1}{6}\beta^2 + ...}$$

Before taking the second derivative, we expand the bottom $1 + \frac{1}{2}\beta + \frac{1}{6}\beta^2 = \epsilon$ in $(1+\epsilon)^{-1} = 1 - \epsilon + \epsilon^2 + \mathcal{O}(\epsilon^3)$, and collecting the parts, we obtain something like

$$S = \mathcal{O}(1) + \mathcal{O}(\beta) + \frac{1}{12}\beta^2 - \frac{1}{4}\beta^2 F + \frac{1}{6}\beta^2 F^2 + \mathcal{O}(\beta^3)$$

where only parts in the second order of β are important. Finally we can differentiate S with respect to β twice and take the limit

$$\lim_{\beta \to 0} \frac{\partial^2 S}{\partial \beta^2} = \frac{1}{6} - \frac{1}{2}F + \frac{1}{3}F^2 = \frac{1}{F}\sum_{n=0}^{F-1} n^2 \tag{6.6}$$

which is equal to our desired sum.

The mean and variance are enough to describe the whole probability distribution because it can be approximated by the Gaussian:

$$p(W_u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(W_u-\mu)^2}{2\sigma^2}}$$

But since we only have calculated the mean and variance in the case of equation 6.1 and not in the normalized version:

$$W = \frac{1}{N(F-1)} \sum_i w_i$$

we multiply the mean μ and the variance σ^2 with N in order to keep the normalization

$$\mu = \frac{1}{2} N(F-1) \tag{6.8}$$

$$\sigma^2 = \frac{1}{12} N(F^2-1) \tag{6.9}$$

and replace W_u with W

$$W = \frac{1}{N(F-1)} W_u$$

so that we with a little algebra arrive at

$$p(W) = \sqrt{\frac{6}{\pi N(F^2-1)}} e^{-(W-\frac{1}{2})^2 \left[\frac{6N(F-1)}{F+1}\right]} \tag{6.10}$$

which is the general density function in the case of the uncorrelated fitness landscape $K = N - 1$[3].Note that the density function has a little untraditional normalization, so that the integral under the bell curve is not unity but $\frac{1}{N}$, see figure 6.2.

Note also, that this equation is independent of the number of symbol classes, A, from which we choose a given letter. A is only relevant for size of the underlying configuration space from which to choose. It does not play a part in the actual assignment of a certain fitness on a certain sequence. It is therefore not involved in the probability $p(w)$.

[3] We can try to check this result for $W = 1/2$ and $F = 2$, which is the height of the Gaussian in the middle (in the case of maximal neutrality): Since there are $\binom{N}{W}$ ways to

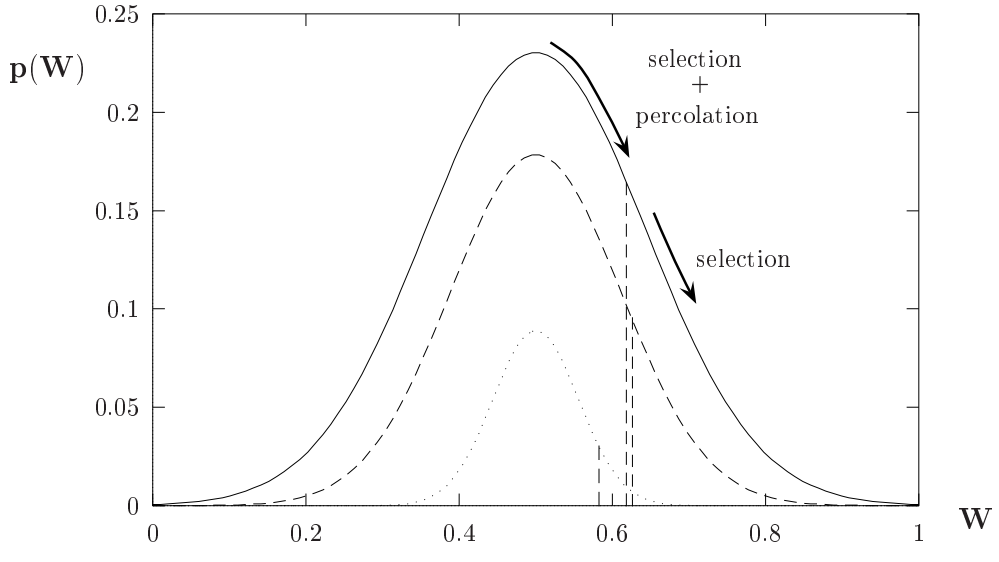

Figure 6.2: *The density function of equation 6.10 for three values of N - full line $N = 12$; dashed line $N = 20$ and dotted line $N = 80$ ($A = 2$ and $F = 2$). Also indicated are the thresholds where percolation disappears.*

---

have the fitness W, the probability to be in the middle is

$$p(w = 1/2) = \frac{1}{2^N}\binom{N}{N/2}$$

where we have divided by the number of possible states. Using Stirling's approximation $N! \sim \sqrt{2\pi N}2N^N e^{-N}$, we obtain:

$$p(w = 1/2) = \frac{\sqrt{2\pi N}N^N e^{-N}}{2^N(\sqrt{\pi N}(\frac{N}{2})^{N/2}e^{-\frac{1}{2}N})^2} = \frac{\sqrt{2}}{\sqrt{\pi N}}$$

which is equivalent to eq.6.10 with $F = 2$. In general there are $\binom{N}{w}$ ways to have the fitness w. The probability that two neighbors in the sequence space x_i and x_j have the same fitness for $F = 2$ is thus

$$\begin{aligned} p(\text{two neighbors same } w) = p &= \sum_{w=0}^{N} p(x_i = w)p(x_j = w) \\ &= \frac{1}{2^{2N}}\sum_{w=0}^{N}\binom{N}{w}^2 = \frac{1}{2^{2N}}\binom{2N}{N} \end{aligned}$$

6.5.2 Defining a percolation threshold

The biologically most interesting thing to know, is the threshold at which percolation stops. Figure 6.2 shows the Gaussian of equation 6.10 in the case of maximal neutrality. Selection pushes the population towards the right exponential tail, but somewhere on that path the percolating networks disappear, simply because at high fitnesses there is a low probability to find a one-point mutant neighbor with the same high fitness.

In order to have percolation, one would assume that there at least has to be one neighbor from where the networks enters, and another neighbor from where it leaves. So, on average, there has to be at least two neighbor mutants with the same fitness. But this argument does not assume that we already are on a percolating network in the beginning: it just hits randomly in configuration space including the sites which are not on the neutral network. Instead we will employ the Bethe lattice approximation (see (Stauffer and Aharony, 1994) for an introduction): Assume that we are on a neutral network, which we can call the origin. Then, there are $N(A-1)-1$ one point mutant neighbors, which themselves again have $N(A-1)-1$ new neighbors emanating. Since each of these belongs to the network with probability p, there will, on average, be $(N(A-1)-1)p$ occupied neighbors to which the percolating path can be continued. If the number $(N(A-1)-1)p$ is less than unity, the probability of finding an infinite path from the origin decreases exponentially. But the threshold, where percolation remains is thus

$$p(W_c) \simeq \frac{1}{N(A-1)-1} \tag{6.11}$$

Below this threshold value there will be compact clusters of neutral networks which will enable populations to search higher fitness levels diffusively. Above this critical point however, percolation will break down and positive selection is confined to climb only towards one of the closest local peaks in the fitness landscape.

Using Stirling's approximation, we get

$$p = \frac{1}{2^{2N}} \left[\frac{\sqrt{4\pi N}(2N)^{2N} e^{-2N}}{(\sqrt{2\pi N} N^N e^{-N})^2} \right] = \frac{1}{\sqrt{\pi N}}$$

So, for increasing N the probability of finding a neutral neighbor decreases slowly with the square root of the system size.

The equality sign in equation 6.11 would hold only if there were no loops in the network lattice (like in Bethe lattices). However, since the dimension of the system is large, the number of possible loops is very small, and we will therefore neglect this correction in the following.

6.5.3 Solution for the critical fitness

We can now find an expression for the critical fitness: Inserting eq. 6.10 on page 108 in 6.11 one obtains:

$$e^{-(W_c-\frac{1}{2})^2\left[\frac{6N(F-1)}{F+1}\right]} = \frac{1}{N(A-1)-1}\sqrt{\frac{\pi N(F^2-1)}{6}} \tag{6.12}$$

which can be solved for W_c in order to obtain an expression for the critical fitness on the right hand side of the Gaussian:

$$W_c = \sqrt{\frac{(F+1)\ln\left[\frac{6(N(A-1)-1)^2}{\pi(F^2-1)N}\right]}{12N(F-1)}} + \frac{1}{2} \tag{6.13}$$

Figure 6.3 shows three solutions of this equation, where the critical fitness is plotted as a function of N in the cases of $A = 2, 3$ and 4.

We can make a qualitative examination of this somewhat involved relation between the critical fitness where neutrality disappears and the other parameters in the model: When looking at the relation between W_c and N in eq. 6.12 one obtains something like

$$(W_c - 1/2)^2 \;\sim\; \frac{1}{N}\ln\sqrt{N} \tag{6.14}$$

which tells us that the factor $\frac{1}{N}$ dominates for large N and small F. So, we should expect that percolation stops very quickly for large sequences, but this only appears to be so, because the actual shape of the Gaussian is also inversely proportional to N (as mentioned before, the bell curve shrinks with larger N because of the normalization: the integral $\sqrt{\frac{6}{\pi N(F^2-1)}}\int e^{-(W-\frac{1}{2})^2\left[\frac{6N(F-1)}{F+1}\right]}dW$ has to be multiplied by N to give unity). So, in reality, the relation between W and N is

$$W_c \;\sim\; \sqrt{\ln\sqrt{N}}$$

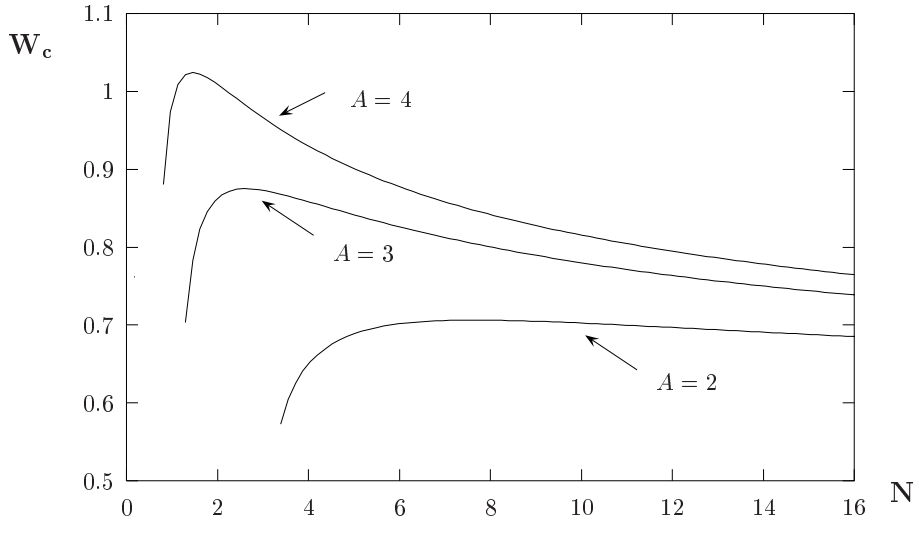

Figure 6.3: *Relation between the critical fitness W_c and sequence size N for three different sizes of the alphabet ($F = 2$).*

One could be tempted to attribute an evolutionary significance to this (positive) dependency: The most fit percolating network depends positively on the system size, so that evolution imposes a selective force on increasing sequence sizes (or increasing genomes) in order to attain fitter percolating networks. But a simple calculation tells us that it actually is *not significant*: organisms have already had around three billion years to increase their genome, but the effective gain through neutrality would only be ~ 3, which cannot be regarded as a really significant selective force. In addition, although long sequences are good for ensuring neutral percolating clusters, their maximal fitness is typically lower (relatively) than the corresponding fitnesses for the shorter sequences - at least in the situations where F is low. This can be seen in the next figure 6.4, where we have plotted the critical fitness from equation 6.13 as a function of neutrality for three different values of N. This figure shows clearly that, although percolation reaches higher fitness values for low N, it breaks down when F increases. Only for very long sequences percolation remains possible. In any case, a high degree of neutrality helps populations to obtain higher fitness values.

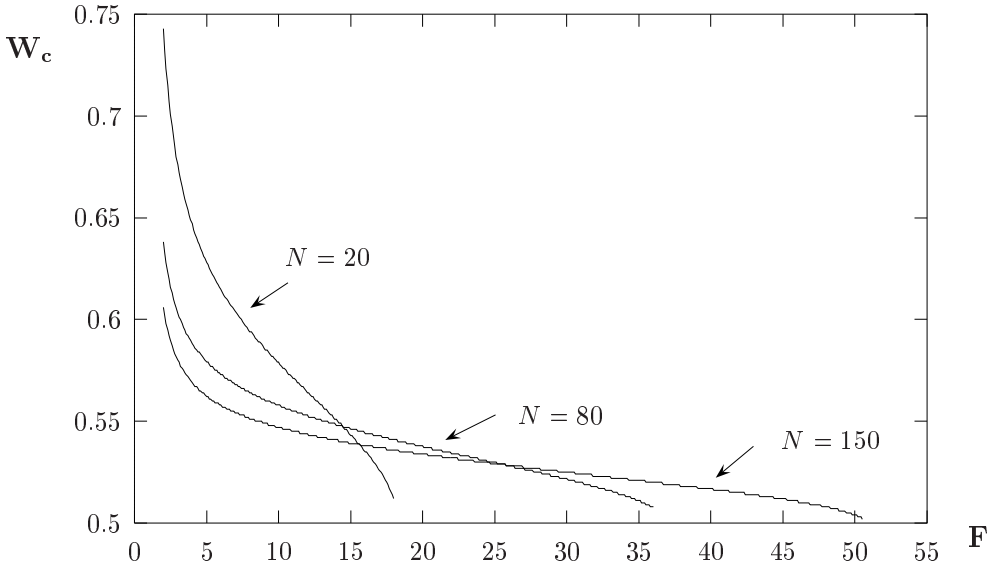

Figure 6.4: *Relation between the critical fitness W_c and neutrality F for three different sizes of the sequences ($A = 4$). The full line corresponds to $N = 20$; the dashed line corresponds to $N = 80$, and the dotted line to $N = 150$.*

6.5.4 Comparison with simulations

In section 3.6.1 on page 42 we investigated how the fitness of the most fit percolating network behaved with decreasing neutrality, and as a result, we found in figure 3.10 that the maximal fitness reached by a random hill climber follows quite closely the fitness of the most fit percolating networks. In order to obtain the figure, we defined a percolating network as a network which is larger than the average network size; e.g. it needs to contain more sequences than a network of average size. These we called "common structures".

However, in the case of an uncorrelated fitness landscape with $K = N - 1$, we know from figure 3.5 on page 37 that the frequency of occurring networks with a certain size follows a power law - at least for the small networks. Unfortunately this means that an average network size typically is very low - often only containing $3 - 5$ sequences. If we would use this definition we would call networks with only, say, 10 sequences percolating. This is obviously not true. Therefore, we will instead define a percolating common network as a network which contains *more than a thousand sequences* (which, in fact, is exactly the point at which the power law of the distribution function from figure 3.5 on

page 37 breaks off).

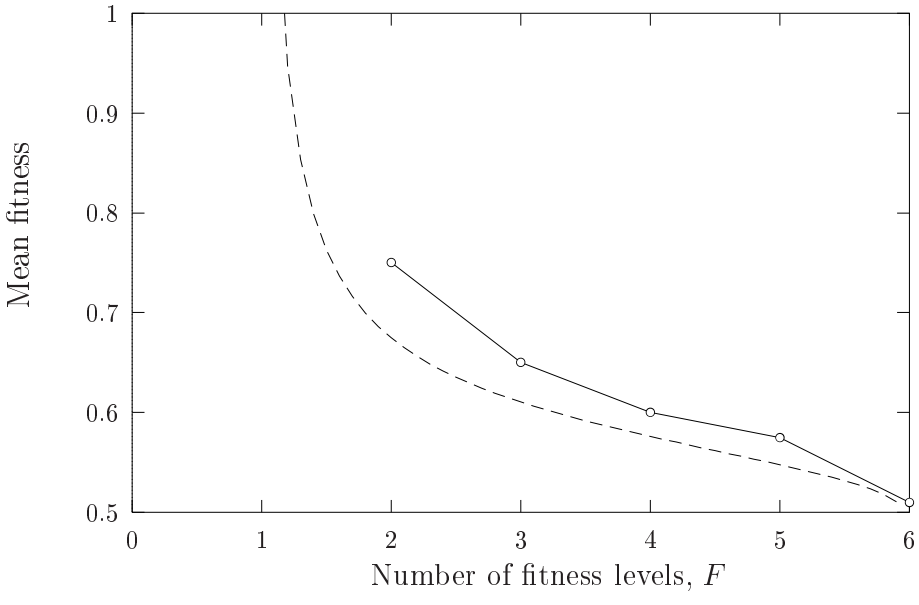

Figure 6.5: *The maximum fitness of the most fit percolating network with $N = 20$, $K = 17$ and $A = 2$, as a function of the neutrality parameter F (circles). The dashed line is the analytical approximation (going to infinity for $F = 1$ since in this case all fitnesses are equal). W_c is undefined for $F > 6$.*

With this redefinition we find in figure 6.5 a verification of the analytical calculations: In the figure we have plotted the analytical solution of eq. 6.13 as a function of F with $A = 2$, $N = 20$ and $K = 17$[4] (full dashed line; same as previous figure, except for the parameter A which now is equal 2) and the simulation for the fittest percolating network (solid line and circles). The simulation is in surprisingly good agreement with the analytical approximation. It shows clearly that there in fact exists a percolation threshold at which the common networks stop to percolate through the sequence space, and it shows that this threshold defines a critical fitness W_c which increases for increasing neutrality. Of course, for $F = 1$ the analytical solution goes to infinity because all fitness values are the same: the landscape is completely flat. For F larger than 6, there

[4]Note that it is not $K = N - 1 = 19$ because the needed number of stored random numbers for the fitness assignment in the computer simulations reaches overflow when $> 2^{17}$. There exists a trick to overcome this problem, and in a forthcoming work (Newman and Engelhardt, 1999) it will be done correctly. However, the fitness landscape for $K = 17$ is not much different from a landscape with $K = 19$.

is no percolation at all, and the model becomes more or less equal to the standard rugged NK landscape investigated earlier (Flyvbjerg and Lautrup, 1992; Bak et al., 1992).

6.5.5 Conclusion

All in all, the analytical investigations presented here suggest that the initial idea of figure 6.2 on page 109 is true: The random hill climber moves diffusively on a neutral network until it finds a network of higher fitness, at which point it shifts to that network. This process continues until there no longer are any percolating networks to diffuse upon. From that point on, the climber is confined to the local region occupied by the non-percolating network, so that it only can get as high as is possible by the local maximum within the region. This represents the threshold where percolation stops and only adaptive selection remains. Figuratively, this process can be imagined by a car trying to reach the top of the Alps: as long as there exist asphalted roads, the car can move "diffusively". But the higher it gets, the more impassable trails and dead ends emerge. Eventually, only one thing remains possible: get out and climb as good as you can.

6.6 Critical food webs?

The extinction patterns and life time distributions of species embedded in the self-organizing food webs of part II have all the characteristics of self-organized criticality:

- As we have seen, the intrinsic dynamics of food webs can create large avalanches of extinction: if a basal species goes extinct because of some unknown misfortune, it generally drags a large portion of species which feed upon it with it. This might even generate a domino effect, because the introduction of the new species which replaces the niche from the extinct ones, may destabilize the system as a whole - at least for a while.

- There exists a clear separation of time scales in the food web dynamics: for long periods nothing happens because the system is in a quasi stable steady state and the noise is too small to perturb the balance significantly. But as soon as one species goes extinct - typically has its population density been diminished only very slowly over a long period of time - it

creates chaos for many other species linked to it. The separation of time scales is clear in this case.

- The food web self-organizes critically in order to maximize the probability for survival of the individual species. This can also be seen in the relatively constant ratio of basal species and top species and the linkages between the species which are totally different from randomly generated webs.

- There is no parameter tuning expect for an arbitrarily set mutation rate, which is the source for destabilizations and sudden extinctions. The mutation rate has thus the function of a slow driving which keeps the system far away from equilibrium, just like standard SOC models.

In conclusion, the evidence so far suggests that the food web model is a good candidate for a self-organized critical phenomenon. However, in the next final chapter we will look at some probabilistic caveats in criticality.

7

Probabilistic Caveats in Criticality

It has been mentioned before that the sole existence of power law distributions in data of taxon life times and fossil extinction events and other enumerational facts in related biological data might be caused by a very simple random mechanism, rather than the more sophisticated mechanisms of self-organized criticality and coevolutionary avalanches. For instance, we have found two different power laws in the distribution functions of the size of neutral networks in the model from part I: one of them containing true correlations in the fitness values among neighboring sequences, and another without any such correlation (the case of $K = N - 1$).

In the following we will investigate such caveats in criticality and in the notion of self-organization with the formulation of a simple model. It is formulated in the context of a model for large-scale evolution with a resulting power law distribution for taxon life times and an exponential distribution for the extinction events. The reader might remember that it is precisely these distributions we found in the ecological food web model from part II.

We discuss whether this result implies some unknown *fundamental mechanism* (a mystical URO) or whether it is a rather natural *tabulation effect* of independent random numbers (a "non-mystical URO") without any criticality and with only a minimal relation to the notion of self-organization. If so, such random processes might equally well account for the observed data for life times and extinction events in the fossil record.

7.1 Background

The need to quantify various biological data in terms of clear and simple enumerational relations seems to be a necessary requisite if one seeks for possible physical principles that could underlie the processes under consideration. The more fundamental a physical mechanism of this kind is claimed to be, the wider a range of complex biological phenomena is expected to behave in accordance with it.

In particular biological large-scale phenomena like extinction events in the fossil record and distributions of taxon life times have been subject to much recent debate whether there might be some fundamental (and even essential) principles at work leading to the observed regular distribution functions. Some of the first ideas in this direction came from taxonomic analysis by Willis (Willis, 1922), who observed that the distribution of the number of genera, by numbers of species, follows a power law with a varying exponent, mostly around a value of 1.5.

Also the investigations by Zipf (Zipf, 1949) on distributions of words in human texts and language, showed inverse power law dependencies (also called *Zipf's law*), which Zipf himself attributed to the principle of least effort, because languages, from his point of view, are means of transmitting information, and therefore should have an optimal structure. Even though this view has been challenged by numerous studies, showing that random texts equally well give rise to these inverse power law distributions (Yule, 1924; Simon, 1955; Mandelbrot, 1961), the urge remains to ascribe fundamental and even *meaningful* principles to these enumerational facts (Mantegna et al., 1994; Tsonis et al., 1997). For instance have the oligonucleotide frequency distributions in non-coding DNA recently been subject to speculations whether their zipfian character (which in fact not is zipfian but log-normal (Perline, 1996; Borodovsky and Gusein-Zade, 1998)) in some way is significant for the functioning of the genetic code (Flam, 1994; Yam, 1995).

It seems indeed desirable to apply a conceptual framework to the apparent structured data of a broad range of biological distributions. But when one is confronted with the choice between a purely probabilistic argument and an argument which goes beyond the probabilistic approach, even if it not seems necessary, it should be self-evident that one should refrain from using such auxiliary assumptions and concepts as much as possible. That does of course not mean that these new ideas, which often are introduced in a broader sense, are irrelevant to such kind of phenomena. It only means that one should be

careful when identifying apparently structured data with fundamental mechanisms, because the structure in the data actually might be the result of some (maybe unknown) statistical regularities of essentially random processes.

One of the more interesting conceptual ideas by which to understand the observed scaling phenomena in complex systems, has been the mechanism of *self-organized criticality* introduced in section 6.3 on page 103.

The main *practical* argument for validating SOC has been the identification of power laws with a well defined exponent, ranging typically between 1 and 2. As mentioned, such scaling laws are characteristic for systems at a critical point, but unlike normal phase transitions, SOC-systems do not need the tuning of an external parameter in order to become critical.

But it became clear that the argument of scale invariance and hence power laws not was immanent to critical systems only. Various models (Newman and Roberts, 1995; Roberts and Newman, 1996; Solé and Manrubia, 1996; Solé and Bascompte, 1996; Solé et al., 1997), written in the context of extinctions events in the fossil record, but surely wider applicable, showed that power laws equally well could be obtained from a competition between exponential functions. Such systems are therefore not critical. It was concluded (Newman, 1996; Newman, 1997) that there was no evidence in the distributions of fossil extinction events to support the notion of self-organized critical behavior in evolution.

In the following we might argue even worse that the emergence of power laws not only can be independent of criticality, but also can be independent of self-organization. In fact, we will show that a power law with an exponent 2 can be generated by a particular simple way of counting independent random numbers. It follows that one has to think about, whether the observation of scale invariance in various empirical distributions equally well might be caused by the particular counting procedure of the observer. For the sake of argument, we will embed our model in a particular simple picture of understanding evolving agents in an ecosystem.

7.2 A very simple model

Imagine an ecosystem consisting of one agent (it could be the first living organism on earth). The survival of this agent is defined by one fitness value, or stress level, which is chosen randomly. During time, there will emerge another organisms with another random stress level value, but if this new value is greater than the first, it will replace it. If not, it will coexist with the first and

try to get the best out of the situation. This process continues for many time steps, and each time, a new agent finds a way to get a higher stress level, than some previous agent (or agents) have managed to survive with, it will replace them in the ecosystem. This means that every time a new agent improves the lowest stress level(s) considerably, it will also be the cause of a major extinction event. It might even be so that it wipes out the rest of the population, but then, history will just start all over again (with the following organisms having better chances to survive, though).

This already defines the model. We can picture the process as follows:

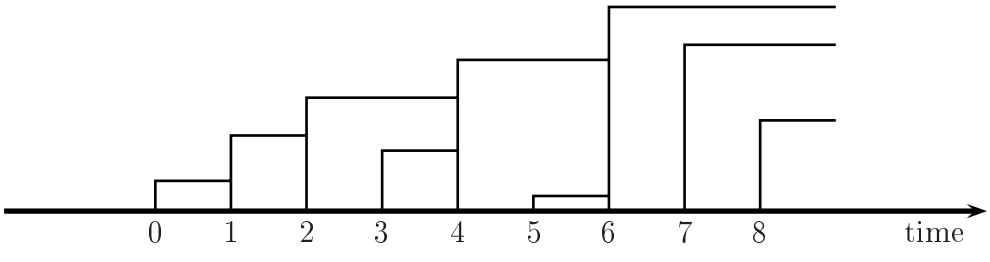

Figure 7.1: Schematical figure of the process. At time $t = 0$ a random number emerges and survives as long as there are no new numbers with a higher value. For this particular example it only survives one time step.

The idea is that the threshold at which a species goes extinct is equal to the stress level

$$p_{stress}(i) \approx p_{threshold}(i) \tag{7.1}$$

In order to formalize this process, we assume an array of random numbers $r_t, t = 0...t_{end}$, associated to each agent emerging at every new time step. This number then represent the stress levels of the agent in the ecosystem. The time $t = 0$ represents our divine starting point, which we just assume will happen. Since the best samplings of life times and extinction events in the fossil record is found for families and not for species, we could identify out agents with them instead.

7.3 Analytical results

Since there is no memory in the system and since random numbers are uncorrelated by definition, it is possible to make a simple mathematical analysis of this "random evolution model".

We can start to ask for the life time of a particular agent i, expressed by the stress levels, or fitness value r_t^i. For a start, we assume that the probabilities r are uniform distributed between $[0;1]$, which expresses that the closer r_t^i gets to unity, the higher possibility there is for a long life time of the particular agent i associated to r. Likewise the probability for this agent to get extinct due to an introduction of a new agent with higher $r^j > r^i$ is $(1 - r^i)$. As an example, the frequency of an agent with a life time of size one is the product of a start r^i, and an extinction $(1 - r^i)$.

Now we can ask for the overall frequency P of life times with size s, which is precisely

$$P = \int_0^1 (1 - r) r^{s-1} dr \tag{7.2}$$

since we have to integrate over the whole distribution of possible starts. After straight forward integration this gives

$$\begin{aligned} P &= \frac{1}{s} - \frac{1}{s+1} \\ &= \frac{1}{s(s+1)} \end{aligned}$$

which already for $s > 10$ is indistinguishable from a power law with exponent -2. Figure 7.2 shows a computer simulation together with the analytical solution.

From equation 7.2 we see that the frequency P actually is equal to a particular Beta function:

$$P = B(s, 2) \tag{7.3}$$

which converts to the gamma function via

$$B(s, 2) = \frac{\Gamma(s)\Gamma(2)}{\Gamma(s+2)}$$

It is a well known property of the gamma function (Simon, 1955) that as $m \to \infty$, and for any constant k,

$$\frac{\Gamma(s)}{\Gamma(s+k)} \sim s^{-k}$$

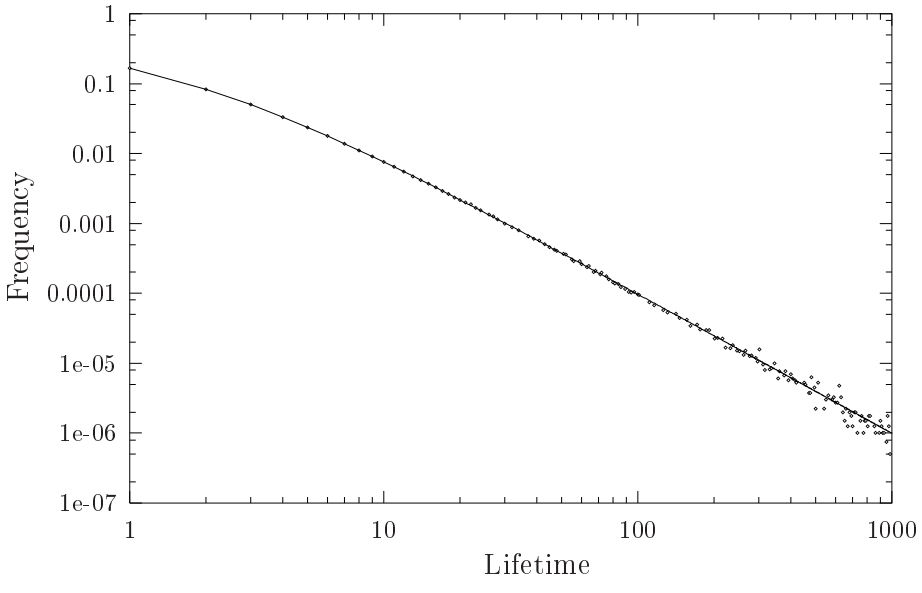

Figure 7.2: *Histogram of the frequency of lifetimes. The full line is the analytical solution.*

and therefore we have for the specific system above, as $s \to \infty$:

$$P = \Gamma(2)s^{-2}$$

The equation 7.3 (including a constant) has been coined a Yule distribution (Yule, 1924; Simon, 1955) after G. Udny Yule who constructed a probability model to explain the distribution of biological genera by number of species which was sampled by Willis (Willis, 1922). However, it is important to note that the emergence of the Yule distribution is a result of the stochastic regularities of a specific random process without any auxiliary assumptions of self-organization or criticality. The existence of power laws in systems driven by some kind of self-organization with or without criticality does therefore not necessarily allow conclusions about such an underlying mechanism to be the epitome of complex biological systems. On the contrary, the existing empirical data and their fitting to frequency distributions, might suggest more random effects than it is hoped for.

One might speculate whether the above analysis holds for other than uniform deviates of random numbers, for instance exponential or Gaussian deviates. In order to see that, we can generalize the above by defining the probability $p(r')$ of a general random number r' different from a uniform distribution

to become higher than r as:

$$\int_r^\infty p(r')dr'$$

and the probability to go below r as:

$$\int_{-\infty}^r p(r')dr' = 1 - \int_r^\infty p(r')dr'$$

which holds by normalization. Then, the transformation for equation 7.2 becomes:

$$P(s) = \int_{-\infty}^\infty \left[\int_r^\infty p(r')dr' \left(1 - \int_r^\infty p(r')dr'\right)^{s-1} \right] p(r)dr \qquad (7.4)$$

But by realizing that the differential

$$d\left[\int_r^\infty p(r')dr'\right] = -p(r)dr$$

we can rewrite eq. 7.4 as

$$P(s) = -\int_1^0 \left[\int_r^\infty p(r')dr' \left(1 - \int_r^\infty p(r')dr'\right)^{s-1} \right] d\left[\int_r^\infty p(r')dr'\right]$$

and substitute $x = \int_r^\infty p(r')dr'$ to get

$$P(s) = \int_0^1 x(1-x)^{s-1}dx$$

giving the result from equations 7.2 and 7.3 since $B(2,s) = B(s,2)$. This means that it does not matter how the random numbers are distributed. The power law remains stable for all situations.

7.3.1 Expectation values and lifetimes

In order to find an equation for the expected average number $\langle k \rangle_t$ of extinctions at a given time t in the array, we can start to observe that at $t = 0$ (the divine point) there can be no extinction:

$$\langle k \rangle_{t=0} = 0$$

At time $t = 1$ we can use equation 7.2 on page 121 for the expected life time s of one agent, to find the average number of extinctions at $t = 1$:

$$\langle k \rangle_{t=1} = \langle k \rangle_{t=0} + \frac{1}{t(t+1)} = \frac{1}{2}$$

At time $t = 2$ we obtain

$$\langle k \rangle_{t=2} = \langle k \rangle_{t=1} + \frac{1}{2(2+1)} = \frac{2}{3}$$

and so on. This means that the *average* (non-cumulative) number of extinctions at time $t = T$ becomes

$$\langle k \rangle_{t=T} = \sum_{t=0}^{T} \frac{1}{t(t+1)} = \frac{T}{T+1} \tag{7.5}$$

so, for $t \to \infty$ the average number of extinctions per time is 1. This means, that most of the agents will disappear within a very short time, and only the very most fit agents, with a stress value close to 1, will survive. However, it is important to note that these average values for the expected number of extinctions is different from the probabilities $p(k, t)$ of having an extinction of size k at time t.

In the same way, we can calculate the expected number of surviving agents $\langle s \rangle_t$ at each time step. Initially, the first is always expected to survive:

$$\langle s \rangle_{t=0} = 1 \tag{7.6}$$

but for the next step, the average number of surviving agents is found by using equation 7.5:

$$\langle s \rangle_{t=1} = 1 - \frac{t}{t+1} = \frac{1}{2}$$

from which it follows that the average *cumulative* number of surviving agents $\langle S \rangle$ at time $t = T$ is:

$$\begin{aligned} \langle S \rangle_{t=T} &= 1 + \sum_{t=1}^{T} \left(1 - \frac{t}{t+1} \right) \\ &= \sum_{t=0}^{T} \frac{1}{t+1} \\ &\approx \ln T + \gamma \end{aligned} \tag{7.7}$$

for large T, where $\gamma \approx 0.5772$ is Eulers constant.

Very naturally, the size of the surviving population will grow infinitely even though the extinctions might wipe out the rest of the population. Diversification is only limited by the requirement of well adapted, high stress values, and these will inevitably emerge in a sufficient long time interval.

7.3.2 Extinction distributions

We have found the distribution of lifetimes $P(s)$ to be power law distributed (equation 7.2 on page 121), and we have found the average number of surviving agents until an arbitrary time T (equation 7.7). We have also found the expectation value for the average number of extinctions per time in equation 7.5, which now allows us a shortcut argument in order to find the frequency of extinctions of different sizes k at a definite sampling time.

In the limit $t \to \infty$, we have all surviving species with a fitness value very close to one, $r_s \to 1$. This means that we for the moment can disregard these when calculating the probabilities for extinctions of size k at time $t \to \infty$.

As an example, we can assume that the probability $p(k, \infty)$ of an extinction of size $k = 1$, *only* is dependent on the previous fitness number r_{t-1}. Then the probability becomes

$$p(k = 1, t = \infty) = \frac{1}{2}$$

since the probability for one random number to be below another is $\frac{1}{2}$. The probability of an extinction of size $k = 2$ will then become the product of the last two fitness numbers r_{t-1} and r_{t-2} to be below r_t:

$$p(k = 2, t = \infty) = \left(\frac{1}{2}\right)^2$$

Thus, in the limit $t \to \infty$, which is the limit we want, we find the frequencies of extinctions distributed as

$$P(k) = \frac{1}{2^k} = e^{-k \ln 2} \tag{7.8}$$

In figure 7.3 we have plotted the result: is shows the relative frequency of extinction as a function of their size for different sampling times (measured as the distance from the starting point) together with the analytical result. This

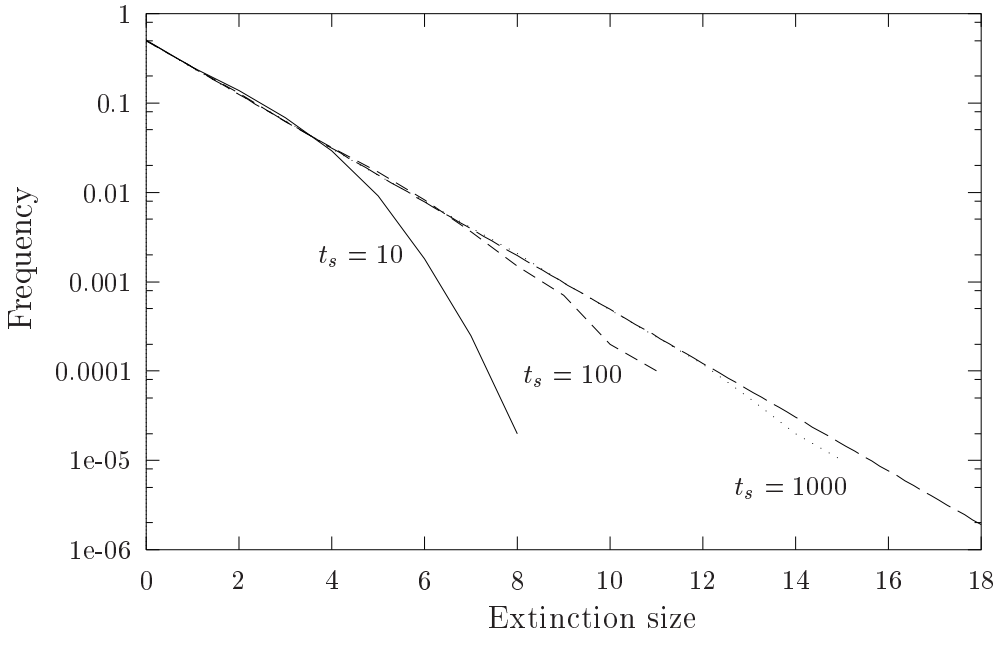

Figure 7.3: *Semi-log plot of the extinction frequencies for sampling times* $t_s = 10, t_s = 100$ *and* $t_s = 1000$ *all converging towards the the analytical solution (straight line with long dashes).*

means that there is a natural transient period before the distribution becomes exponential.

So, in summary, our model predicts a power law distribution of taxonomic life times with an exponent -2, and for the distribution of extinction sizes, the model predicts an exponential decay with a characteristic factor of $\ln 2$.

7.3.3 Extinction patterns

Unfortunately, the world is not that simple anyway. Although we have found equivalent distribution functions as in the food web model from part II, we find different extinction pattern statistics. While in the food web model the power spectrum showed 1/f-noise [1], the same analysis applied to this present random model reveals "f-noise", that is, not correlation, but *anti-correlation*. Intuitively this is understandable: in the food web model there was strong correlation in the sense that if an important basal species became extinct, it

[1] If the extinctions were white noise, the power spectrum would be flat.

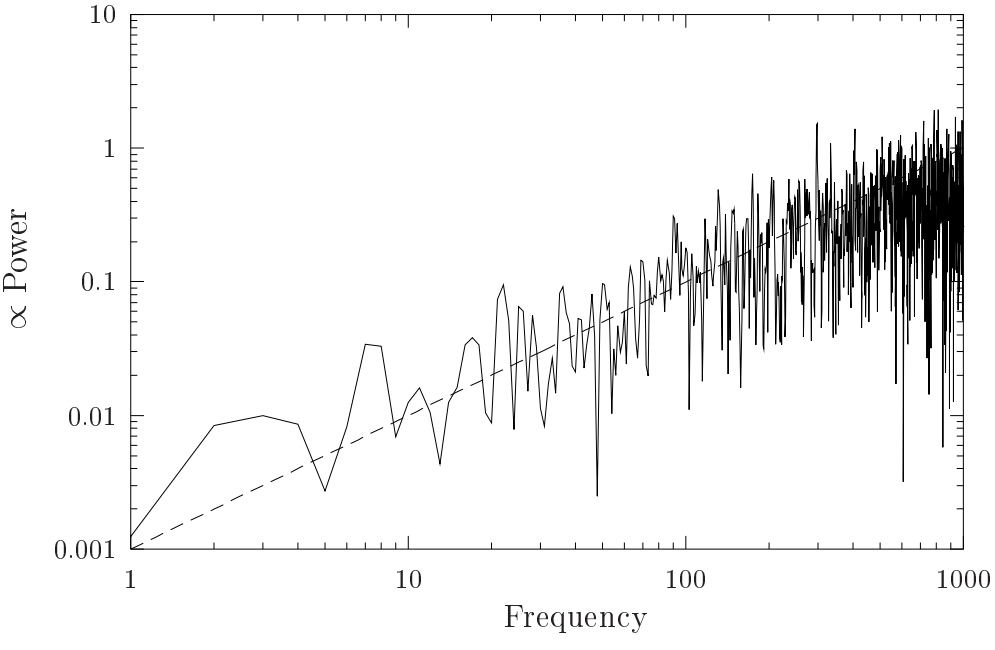

Figure 7.4: *The Fourier transform of the autocorrelation function shows increasing correlations for large frequencies, thus,* anti-correlation. *The straight dashed line follows* $P(f) \propto f$.

pulled also a large number of other species into extinction because they were feeding on it. In the present model however, a large extinction results in a strong reduction of existing species and a relatively peaceful subsequent time with low threshold values.

7.4 Conclusion

For a particular simple random model of evolution, we have shown that it is possible to obtain very reasonable results when compared to the data of the fossil record. We do not claim that this model is especially more realistic than other models of evolution, but we do claim that the empirical evidence not yet allows us to discriminate between co-evolutionary effects, as supposed by self-organizing critical models, and statistical regularities of essentially independent random numbers.

It is a subtle question whether our described process actually reflects some sort of self-organization. At first glance, it is obvious that, since the extinc-

tions thrive the stress levels towards higher and higher values, there is some sort of organizational principle of the agents. It is a selective pressure which forces all living agents to have high fitness values, which , in time, will become more and more difficult to beat for the newcomers. In this sense, the probability of an agent to go extinct, will decrease as its life time increases. So, the cockroaches indeed have a high probability of outliving human beings. But this organizational principle has nothing to do with possible interactions between cockroaches and humans. Actually, they both live quite independent of each other, and will (hopefully) die independent of each other. Likewise in our model, the agents are totally independent of each other. The apparent organization comes about only when tabulated and measured by an observer. In this sense, there is no self-organization, but only statistics.

In order to stress this point even further, we could formulate a model as follows: Imagine a multi-faced dice (say, N faces). Roll it, and write down the number (between 1 and N). Roll the same dice as many times, until you get a number which is greater that the one you wrote down. Tabulate the waiting time and begin again. This process gives immediately the desired power law distribution of the waiting times and the exponential distribution of extinction events, since it is equivalent to our model when $N \to \infty$.

We have shown that it is possible to produce power laws with an exponent of -2 by random numbers put together in a random fashion. The emergence of a power law only depends on some instantaneous properties of the system, which (in a quite challenging sense) only describes the way we look at the data. But this then means that scale-free phenomena observed in nature not necessarily calls attention towards some self-organizing or critical mechanism at work on a fundamental level. And finally, when this is so, the discussion of power laws in models of complex biological phenomena seems bound to be transferred to the examination of the possible discrepancies between the actually observed facts and the predictions of the theory. Some might put this in a rather popular way and say: what is measured are the crumbs in the bakery and not the cookies.

As mentioned in the beginning of the thesis, the question is still open, whether these similar statistical properties - the power laws - of such a large and heterogeneous group of real phenomena are a result of a common, fundamental and yet unknown mechanism, or just a statistical peculiarity. In any case: in order to understand the principles of evolving systems, it is important to investigate from *every possible point of view*. Although the general approach to the understanding of evolutionary dynamics through distribution functions is exiting, the most fruitful approach is probably still the formulation

of more specific models on more specific problems. As we have seen in the first two parts, the general aspects of criticality and self-organization can emerge anyway.

Appendix

A
Hamming Distances

The Hamming distance $d(i,k)$ of two sequences x_i and x_k counts the number of digits (or symbols) in which they differ. So, for a given reference sequence x_i, the number of possible mutants $S_{d(i,k)} = S_d$ with only one symbol different - the one-point mutants - is given by the sequence length N times the number of other symbols $(A-1)$ which are able to occupy the loci in the sequence. In general, the number of mutant sequences is therefore given by

$$S_d = \binom{N}{d}(A-1)^d$$

where N is the sequence length and A the number of possible symbol classes at each locus. In normal NK-systems one has binary sequences $A = 2$, and in the case of RNA one has $A = 4$ corresponding to the four different nucleotides. Summing over all possible Hamming distances $0 \leq d \leq N$

$$S = \sum_{d=0}^{N} \binom{N}{d}(A-1)^d = A^N$$

gives the total number of possible sequences.

Haemoglobin, for instance, is a very modestly sized protein with 141 amino acids, corresponding to a sequence of 423 nucleotides (disregarding introns within the genetic sequence). Still, this sequence has 4^{423} possible configurations and allready $1,269$ different one-point mutant neighbors.

B
Hardy-Weinberg Law

Darwin had a problem with his concept of "blending inheritance": He believed that offsprings blend the traits inherited from their parents into each other, so that distinct features like colors or special forms blend into an average feature, resulting in the loss of variability. The rediscovery of Mendel's work showed that hereditary traits are a kind of immortal atoms, which are reshuffled in each new generation, permitting variability to be maintained. The subsequent work of Hardy and Weinberg showed that the frequencies of these "atoms", the alleles sitting at the chromosomal loci, are maintained through the generations.

The simplest situation is that of two different genes sitting at a chromosomal locus in a diploid organism, so that the fitness of the genotype is determined by the type of interaction between the genes (alleles). For instance, the two alleles A_1 and A_2 could code for brown and blue eyes respectively. Then, the genotype is called *homozygous* if the same allele appears twice (A_1, A_1) or (A_2, A_2) and *heterozygous* when both are present (A_1, A_2). By convention the genotypes (A_1, A_2) and (A_2, A_1) cannot be distinguished, so it does not matter which of the two genes stems from the father and which from the mother. However, one allele may suppres the effect of the other, in which case it's called *dominant* and the other *recessive*.

In general, we can assume that the frequencies of the alleles $A_1, A_2, ..., A_n$

are given by $p_1, p_2, ..., p_n$, and the frequencies of the genotypes, eg. the gene pairs (A_i, A_j), by p_{ij}. Then, the frequency of allele p_i can be found as the sum of the contributions from the father and the mother -

$$p_i = \frac{1}{2}\sum_j p_{ij} + \frac{1}{2}\sum_j p_{ji} \tag{B.1}$$

where, the factor $\frac{1}{2}$ expects that there has to be one contribution from each parent. If we then assume random mating in the population, we can define the frequency of the genotype p'_{ij} in the next generation as

$$p'_{ij} = p_i p_j \tag{B.2}$$

saying that the frequency of combination p'_{ij} equals the product of the probabilty p_i to have A_i and p_j to have A_j. Using eq.(B.1), we can find the frequency of allele i in the next generation (p_i') as

$$p'_i = \frac{1}{2}\left(\sum_j p'_{ij} + \sum_j p'_{ji}\right) = \sum_j p_i p_j = p_i \tag{B.3}$$

This is the *Hardy-Weinberg law* saying that the frequency of the genes remains unchanged thoughout the generations.

B.1 Modes of selection

In the following, we shall use examples of discrete dynamics in the case of one locus with two alleles. Let us assume that the frequency of allele A_1 is p, and the frequency of allele A_2 is $q = 1 - p$. We attribute a definite fitness value to each of the three genotypes, so that

genotype	A_1A_1	A_1A_2	A_2A_2
fitness	w_{11}	w_{12}	w_{22}
frequency	p^2	$2pq$	q^2

Assuming that the genes are in Hardy-Weinberg equilibrium, we find the frequency of allele A_2 in the next generation as

$$q' = \frac{pqw_{12} + q^2 w_{22}}{p^2 w_{11} + 2qpw_{12} + q^2 w_{22}} \tag{B.4}$$

The extend of change in the frequency of allele A_2 per generation is defined as $\Delta q = q' - q$. Then, the rate of change becomes [1]

$$\Delta q = \frac{pq[p(w_{12} - w_{11}) + q(w_{22} - w_{12})]}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}} \tag{B.5}$$

Depending on how we choose the fitnesses w_{11}, w_{12} and w_{22}, we find quite different modes of how selection operates.

B.1.1 Directional selection

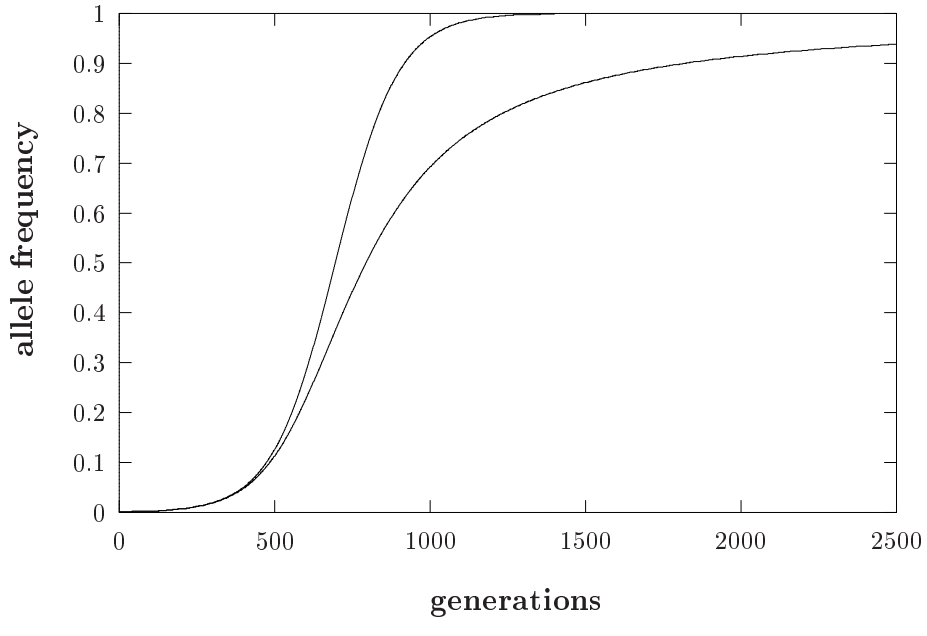

Figure B.1: *Frequency of an advantageous allele with $s = 0.01$. Upper curve shows th codominant mode, and the lower cruve show the complete dominant mode. Note that in both cases the initial growth rate is very small and thus succeptible to chance effects like random genetic drift.*

In the simplest case of directional selection, or *codominant selection*, it is expected that the heterozygote A_1A_2 has a fitness which is in-between the

[1] use that $p' - p = -(q' - q)$, in order to realize that $\Delta q = pq' - qp'$

fitnesses of the two homozygotes. Assuming a selective advantage s of allele A_2, this can be written as

genotype	A_1A_1	A_1A_2	A_2A_2
fitness	1	$1+s$	$1+2s$

and eq. (B.5) becomes

$$\Delta q = \frac{spq}{1 + 2spq + 2sq^2}$$

which is illustrated in the upper curve of figure (B.1) for $s = 0.01$.

B.1.2 Complete dominance

If we instead assume that the presence of an allele is dominanting completely, then the heterozygote is as fit as the most fit homozygote:

genotype	A_1A_1	A_1A_2	A_2A_2
fitness	1	$1+s$	$1+s$

Using eq. (B.5) we find

$$\Delta q = \frac{sqp^2}{1 + sq^2}$$

which is shown in the lower curve of figure (B.1) for $s = 0.01$. Note that complete dominant selection is slower than the codominant mode, since the inferior allele can be carried along in the heterozygote without affecting the fitness of the genotype. Ultimately, however, the inferior allele is lost.

B.1.3 Balancing selection

A more interesting situation appears when the most fit gamete is the heterozygote. Defining $t < s$, we have

genotype	A_1A_1	A_1A_2	A_2A_2
fitness	1	$1+s$	$1+t$

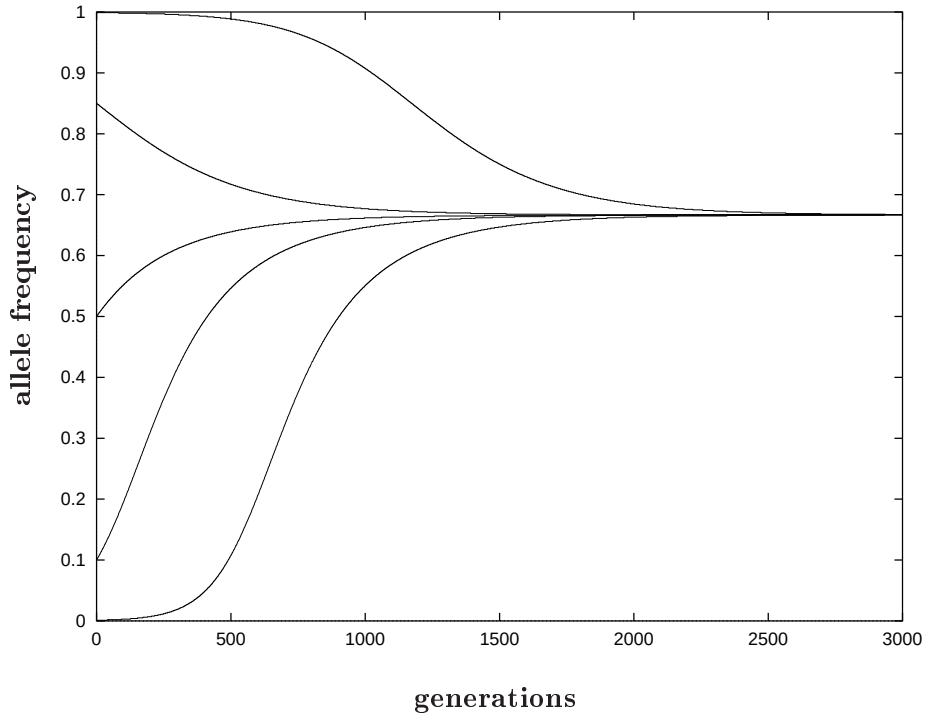

Figure B.2: *When the heterozygote is the most fit genotype, selection operates in the overdominant mode and finds a stable equilibrium where both alleles coexist.* $s = 0.01, t = 0.005$. *This type of balancing selection has been the main argument for explaining the high degree of polymorphisms found in real populations, see the discussion in the beginning of part I.*

in which case the change in frequency of A_2 becomes

$$\Delta q = \frac{pq(2sq - tq - s)}{1 + 2spq + tq^2}$$

In contrast to the two previous situations, in which one allele always will become eliminated, we here find a coexistence between the two alleles, see figure. This kind of selection, called *overdominant* belongs to a class of stabilizing or *balancing* selection. The equilibrium point q_e for allele A_2 can be solved to be

$$q_e = \frac{s}{2s - t} \tag{B.6}$$

C

The Replicator Equation and Lotka-Volterra systems

The general replicator equation is formulated as follows: the rate of increase of a species population $\frac{\dot{x}_i}{x_i}$, which, in a darwinian sense, is succesfull - compared to other species populations - is given by the difference in it's fitness $f_i(\mathbf{x})$ and the average fitness of the population $\bar{f}_i(\mathbf{x}) = \sum_i x_i f_i(\mathbf{x})$ so that:

$$\dot{x}_i = x_i \left(f_i(\mathbf{x}) - \bar{f}_i(\mathbf{x}) \right) \tag{C.1}$$

(Note that, implicitely, it has been assumed here that the degree of succes of species i is proportional to it's own increase.)

One particular interesting and much analyzed situation is, when the fitness of a species i is expected to be a *linear* function of the population vector such that

$$f_i(\mathbf{x}) = \sum_j^N a_{ij} x_j \tag{C.2}$$

so that equation C.1 becomes

$$\dot{x}_i = x_i \left(\sum_j^N a_{ij} x_j - \sum_{ij}^N a_{ij} x_i x_j \right)$$

or in matrix notation

$$\dot{x}_i = x_i \left[(\mathbf{Ax})_i - \mathbf{x} \cdot \mathbf{Ax} \right] \tag{C.3}$$

This replicator equation (or *continuous time selection equation*) C.3 is a cubic function in $\mathbf{x}$, but it is possible to show that equation C.3 in n variables is equivalent to the quadratic Lotka-Volterra equation in $n-1$ variables:

Consider the transformation

$$x_i = \frac{y_i}{\sum_j^N y_j}, \qquad\qquad i = 1, .., n$$

with $y_n \equiv 1$ so that

$$y_i = \frac{y_i}{y_n} = \frac{x_i}{x_n}, \qquad\qquad i = 1, .., n$$

It is possible (Hofbauer and Sigmund, 1998) to add a constant c_j to the j-th column of matrix $\mathbf{A}$ without changing equation C.3. This means that we can change the last row of the $n \times n$ matrix $\mathbf{A}$ to solely consist of zeroes without loss of generality. Then the transformation reads:

$$\begin{aligned} \dot{y}_i = \dot{\left(\frac{x_i}{x_n} \right)} &= \frac{\dot{x}_i}{x_n} - \frac{\dot{x}_n x_i}{x_n^2} \\ &= \frac{x_i}{x_n} \left[(\mathbf{Ax})_i - \mathbf{x} \cdot \mathbf{Ax} \right] - \frac{x_i}{x_n} \left[(\mathbf{Ax})_n - \mathbf{x} \cdot \mathbf{Ax} \right] \\ &= \left(\frac{x_i}{x_n} \right) \left[(\mathbf{Ax})_i - (\mathbf{Ax})_n \right] \end{aligned}$$

Since we appropriately have set $(\mathbf{Ax})_n = 0$, we obtain:

$$\dot{y}_i = y_i (\mathbf{Ax})_i = y_i \left(\sum_{j=1}^{n} a_{ij} x_j \right) = y_i \left(\sum_{j=1}^{n} a_{ij} y_j \right) x_n$$

And since x_n allways is strictly positive, we can change the time scale of integration and thereby remove the term x_n. Then, using $y_n = 1$, we arrive at the general Lotka-Volterra equation in $n-1$ dimensions:

$$\dot{y}_i = y_i \left(a_{in} + \sum_{j}^{n-1} a_{ij} y_j \right)$$

D

The Probability to Find an Autocatalytic System

Imagine a soup of N different molecules. [1] The probability that the molecule i increases the growth rate of an other molecule j can be defined as $p(i \to j)$. Then, the probability that j does the same to i at the same time, that is, creating a positive feedback loop, is $p(i \to j)^2$ (since for our simple case it's enough to say that $p(i \leftarrow j) = p(i \to j) = p$).
Now, how many of such pairs can we find? This must be a number like

$$\frac{N(N-1)}{2} \cong \frac{N^2}{2}$$

for large N. Thus, the chance to find such a minimal autocatalytic loop is

$$\frac{1}{2}(Np)^2(1-p)^{N-2}$$

where the last part says that the rest of the system doen't catalyse anyting. Equally, the chances for finding 3-loops, 4-loops and so on, are

$$\frac{1}{3!}(Np)^3((1-p)^{N-3}, \quad \frac{1}{4!}(Np)^4(1-p)^{N-4} \quad \ldots\ldots\ldots \quad \frac{1}{N!}(Np)^N$$

[1] The intial title to this small note was "Solution to the mystery of life, independent of Gödel".

The over all chance to find some kind of Münchausen-system must then be

$$p_{tot} = \sum_{j=1}^{N} \frac{Np^j}{j!}(1-p)^{N-j}$$

For $N \gg 1$ and $p \ll 1$ we can expand the last term in this sum according to

$$\begin{aligned}(1-p)^N &= 1 - Np + \frac{N(N-1)}{2}p^2 - \frac{N(N-1)(N-2)}{3!}p^3 ..... \\ &\simeq 1 - Np + \frac{N^2}{2!} - \frac{N^3}{3!}p^3 ..... \\ &= e^{-Np}\end{aligned}$$

Setting $x = Np$ we obtain

$$\begin{aligned}p_{tot} &\simeq \sum_{j=1}^{\infty} \frac{x^j}{j!}e^{-x} \\ &= e^{-x}(e^x - 1) \\ &= 1 - e^{-x}\end{aligned}$$

So, the total probability of finding a seemingly most improbable Münchausen-effect is given by this simple expression. For $p \ll 1/N$, that is $x \ll 1$, we have $p_{tot} \simeq 1 - e^{-1} \simeq x$. However, if p is large enough, or conversely, if the soup is large enough so that

$$p \geq \frac{1}{N},$$

then we have $Np \geq 1$, and we can now see that

$$p_{tot} \sim 1 \tag{D.1}$$

In words it says: since for a large enough system, the number of possible combinations increases so fast that even a most improbable event will find it's way to existence by the brute force of probabilistics. In it's philosophical variant it says: If the universe is infinite, everything will be in it, somewhere.

E
Publications

Effects of neutral selection on the evolution of molecular species

M. E. J. Newman
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. U.S.A.

Robin Engelhardt
Center for Chaos And Turbulence Studies, Dept. of Chemistry,
University of Copenhagen, Universitetsparken 5, Copenhagen Ø, Denmark

Abstract

We introduce a new model of evolution on a fitness landscape possessing a tunable degree of neutrality. The model allows us to study the general properties of molecular species undergoing neutral evolution. We find that a number of phenomena seen in RNA sequence-structure maps are present also in our general model. Examples are the occurrence of "common" structures which occupy a fraction of the genotype space which tends to unity as the length of the genotype increases, and the formation of percolating neutral networks which cover the genotype space in such a way that a member of such a network can be found within a small radius of any point in the space. We also describe a number of new phenomena which appear to be general properties of neutrally evolving systems. In particular, we show that the maximum fitness attained during the adaptive walk of a population evolving on such a fitness landscape increases with increasing degree of neutrality, and is directly related to the fitness of the most fit percolating network.

Published in Proc. R. Soc. London B (1998),**265** (1402), pp. 1333-1338.

Model for neutral evolution

M. E. J. Newman
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. U.S.A.

Robin Engelhardt
Center for Chaos And Turbulence Studies, Dept. of Chemistry,
University of Copenhagen, Universitetsparken 5, Copenhagen Ø, Denmark

Abstract

We make an analytical investigation of the model for neutral evolution (Newman and Engelhardt, 1998) in the case of a maximally rugged fitness landscape. While the existence of neutral networks, called common structures, helps populations to attain higher fitness values through diffusive searching, there exists, at a certain critical fitness, a percolation threshold at which the population no longer can utilize these neutral networks. We investigate this threshold in terms of the neutrality and other parameters of the system, and compare the results to simulations.

In preparation.

A new model for food webs

Robin Engelhardt
Center for Chaos And Turbulence Studies, Dept. of Chemistry,
University of Copenhagen, Universitetsparken 5, Copenhagen Ø, Denmark

Axel Hunding
Chemical Laboratory III
Dept. of Chemistry,
University of Copenhagen, Universitetsparken 5, Copenhagen Ø, Denmark

M. E. J. Newman
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. U.S.A.

Abstract

We introduce a new model for the spontaneous generation of food webs in ecological communities. Using a modified Lotka-Volterra scheme including extinctions and introductions of new species, we find a spontaneous self-organization of the most important properties in food web models such as well-defined directional food chains; top species, intermediate species and autotroph basal species; relatively constant, but fluctuating linkage densities, omnivores, species-specific interaction strength and a decoupling of the system into sub-webs. We also find a difference in the average lifetime of generalizers which feed upon many species (preys) and specializers which feed upon only a few preys. While the specializers are far more abundant than the generalizers, the latter have a longer lifetime, on average.

Also the extinction dynamics is investigated in terms of Van Valens constant extinction law, extinction patterns exhibiting "1/f-noise", extinction distributions and lifetime distributions which show exponential and power law decay, respectively. The data from the model is compared to existing data from experimental food webs and to the fossil record. They all show good agreement with this new model for the spontaneous organization of species communities.

Two papers. In preparation.

Stochastic effects and Extinction- and lifetime distributions

Robin Engelhardt
Center for Chaos And Turbulence Studies, Dept. of Chemistry,
University of Copenhagen, Universitetsparken 5, Copenhagen Ø, Denmark

Abstract

The sole existence of power law distributions in data of taxon life times and fossil extinction events and other enumerational facts in related biological data might be caused by a very simple random mechanism, rather than the more sophisticated mechanisms of self-organized criticality and coevolutionary avalanches. In order to discuss such caveats, We formulate a model for random large-scale evolution with a resulting power law distribution for taxon life times and an exponential distribution for the extinction events. We discuss whether this result is a rather natural *tabulation effect* of independent random numbers without criticality with only minimal relation to the notion of self-organization. We argue that this process equally well might account for the observed data for life times and extinction events in the fossil record.

In preparation.

Patterns Selection in Bistable Systems

S. Métens,
G. Dewel,
P. Borckmans,
Cervice de Chimie-Physique,
Center of Nonlinear Phenomena and Complex Systems,
and Nonlinear Chemistry Unit,
C.P. 231, université Libre de Bruxelles,
1050 Bruxelles, Belgium

Robin Engelhardt
Center for Chaos And Turbulence Studies, Dept. of Chemistry,
University of Copenhagen, Universitetsparken 5, Copenhagen Ø, Denmark

Abstract

Pattern selection in reaction-diffusion systems exhibiting bistability of homogeneous steady states is discussed. In agreement with recent experimental results, we obtain new bifurcation diagrams involving large amplitude structures that arise from the coupling of the spatial modes with a quasi-neutral homogeneous mode.

Published in Europhys. Lett.(1997), **37**(2),pp.109-114

Bibliography

Bak, K., Flyvbjerg, H., and Lautrup, B. (1992). Coevolution in a rugged fitness landscape. *Phys. Rev. A*, 46(10):6724–6731.

Bak, P., Flyvbjerg, H., and Sneppen, K. (1994). Can we model darwin. *New Scientist*, pages 36–39.

Bak, P. and Sneppen, K. (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Lett.*, 71:4083–4086.

Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: An explanation of 1/f noise. *Phys. Rev. Lett.*, 59(4):381–384.

Bak, P., Tang, C., and Wiesenfeld, K. (1988). Self-organized criticality. *Phys. Rev. A*, 38(1):364–374.

Barnett, L. (1997). Tangled webs: Evolutionary dynamics on fitness landscapes with neutrality. Master's thesis, School of Cognitive Sciences, University of East Sussex, Brighton.

Baskaran, S., Stadler, P., and Schuster, P. (1996). Approximate scaling properties of rna free energy landscapes. *J. theor. Biol.*, 181:299–310.

Biebricher, C. K. (1986). Darwinian evolution of self-replicating RNA. *Chemica Scripta*, 26B:51–57.

Biebricher, C. K., Eigen, M., and Luce, R. (1981a). ?? *J. Mol. Biol.*, 148:361.

Biebricher, C. K., Eigen, M., and Luce, R. (1981b). ?? *J. Mol. Biol.*, 148:391.

Borodovsky, M. Y. and Gusein-Zade, S. M. (1998). A general rule for ranged series of codon frequencies in different genomes. *Journal of Biomolecular Structures and Dynamics*, 6(5):1001–1012.

Bray, W. C. (1921). A periodic reaction in homogeneous solution and its relation to catalysis. *J. Amer. Chem. Soc.*, 43:1262–1267.

Briand, F. (1983). Environmental control of food web structure. *Ecology*, 64:253–263.

Briand, F. and Cohen, J. E. (1984). Community food webs have scale-invariant structure. *Nature (London)*, 307:264–266.

Brown, J. H. (1994). Complex ecological systems. In Cowan, G. A., Pines, D., and Meltzer, D., editors, *Complexity: Metaphors, Models and Reality*, pages 419–451. Addison-Wesley, Reading, Massachusetts, 3rd edition.

Burlando, B. (1990). ?? *Journal of Theoretical Biology*, 146:99.

Cohen, J. E. and Briand, F. (1984). Throphic links of community food webs. *Proc. Nat. Acad. Sci. USA*, 81:4105–4109.

Cohen, J. E., Briand, F., and Newman, C. M. (1986). A stochastic theory of community food webs. iii. predicted and observed lengths of food chains. *Proc. R. Soc. London Ser. B*, 228:317–353.

Cohen, J. E., Briand, F., and Newman, C. M. (1990). *Community Food Webs*. Springer Verlag, Berlin.

Cohen, J. E. and Newman, C. M. (1988). Dynamic basis of food web organization. *Ecology*, 69:1655–1664.

Colwell, R. K. (1986). Community biology and sexual selection: lessons from hummingbird flower mites. In Diamond, J. and Case, T., editors, *Community Ecology*, pages 406–424. Harper and Row, New York.

Dean, A. M. (1998). The molecular anatomy of an ancient adaptive event. *American Scientist*, 86(1).

Derrida, B. (1981). ?? *Physical Review B*, 24:2613.

Domingo, E., Sabo, D., Taniguchi, T., and Weissmann, C. (1978). ?? *Cell*, 13:735–744.

Eigen, M. (1992). *Steps towards Life*. Oxford University Press, Oxford.

Eigen, M., Gardiner, W., Schuster, P., and Winkler-Oswatitsch (1981). The origin of genetic information. *Scientific American*, 244:78–95.

Eigen, M., McCaskill, J., and Schuster, P. (1988). The molecular quasi species. *J. Phys. Chem.*, 92:6881.

Eigen, M., McCaskill, J., and Schuster, P. (1989). The molecular quasi species. *Advances in Chemical Physics*, 75:149.

Eigen, M. and Schuster, P. (1997). *The Hypercycle: A principle of natural self-organization.* Springer, New York.

Eldredge, N. and Gould, S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism. In Schopf, T. J. M., editor, *Models in Paleobiology.* Freeman, San Francisco.

Elton, C. (1927). *Animal Ecology.* Sidgwick and Jackson, London.

Elton, C. (1958). *The ecology of invasions by animals and plants.* Chapman and Hall, London.

Emlen, J. M. (1966). The role of time and energy in food preference. *Amer. Natur.*, 100:611–617.

Engelhardt, R. (1994). Modelling pattern formation in reaction-diffusion systems. Master's thesis, University of Copenhagen.

Evans, F. C. and Murdoch, W. W. (1968). Taxonomic composition, trophic structure and seasonal occurrence in a grassland insect community. *J. Animal Ecol.*, 37:259–273.

Feder, J. (1995). The evidence for self-organized criticality in sandpile dynamics. *Fractals*, 3(3):431–443.

Fischer, K. H. and Hertz, J. A. (1991). *Spin Glasses.* Cambridge University Press, Cambridge.

Fisher, R. A. (1930). *The genetical theory of natural selection.* Clarendon Press, Oxford.

Flam, F. (1994). Hints of a language in junk dna. *Science*, 266:1320.

Flyvbjerg, H. and Lautrup, B. (1992). Evolution in a rugged fitness landscape. *Phys. Rev. A*, 46(10):6714–6723.

Flyvbjerg, H., Sneppen, K., and Bak, P. (1993). Mean field theory for a simple model of evolution. *Phys. Rev. Lett.*, 71:4087–4090.

Fontana, W. and Schuster, P. (1987). A computer model of evolutionary optimization. *Biophys. Chem.*, 26:123.

Frette, V., Christensen, K., Malthe-Sørensen, A., Feder, J., Jøssang, T., and Meakin, P. (1996). Avalanche dynamics in a pile of rice. *Nature*, 379:49–52.

Futuyma, D. J. (1986). *Evolutionary Biology*. Sinauer Associates, Sunderland, Massachusetts, 2nd edition.

Gardner, E. and Derrida, B. (1989). ?? *J. Phys. A*, 22:1975.

Gardner, M. R. and Ashby, W. R. (1970). Connectance of large, dynamical (cybernetic) systems: critical values for stability. *Nature*, 228:784.

Gause, G. F. (1934). *The struggle for existence*. Williams and Wiklins, Baltimore.

Gillespie, J. (1984). ?? *Evolution*, 38:1116.

Goldwasser, L. and Roughgarden, J. (1997). Sampling effects and the estimation of food web properties. *Ecology*, 78(1):41–54.

Gould, S. J. (1995). *Dinosaur in a Haystack*. Random House, New York.

Gould, S. J. and Eldredge, N. (1993). Punctuated equilibrium comes of age. *Nature*, 366:223–227.

Grinstein, G. (1995). Generic scale inveriance and self-organized criticality. In McKane, A., Many, M., Other, O., and People, P., editors, *Scale invariance, Interfaces and Nonequilibrium Dynamics*. Nato advanced study series B: physics vol. 344, Plenum, New York.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., and Schuster, P. (1996a). Analysis of RNA sequence structure maps by exhaustive enumeration: neutral networks. *Mh. Chem.*, 127:355.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I. L., Stadler, P. F., and Schuster, P. (1996b). Analysis of RNA sequence structure maps by exhaustive enumeration: structure of neutral networks and shape space covering. *Mh. Chem.*, 127:375.

Haldane, J. B. S. (1932). *The Causes of Evolution*. Longmans and Green, London.

Hall, S. J. and Raffaelli, D. G. (1991). Food web patterns: lessons from a species rich web. *J. Animal Ecology*, 60:823–841.

Hall, S. J. and Raffaelli, D. G. (1993). Food webs: Theory and reality. *Advances in Ecological Research*, 24:187–239.

Havens, K. (1992). Scale and structure in natural food webs. *Science*, 257:1107. also pages 11 and 13.

Hofacker, I. L. (1994). A statistical characterization of the sequence to structure mapping in RNA. Master's thesis, Institut für Theoretische Chemie, University of Vienna.

Hofbauer, J. and Sigmund, K. (1988). *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, Cambridge.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1nd edition.

Hutson, V. and Schmitt, K. (1992). Permanence and the dynamics of biological systems. *Mathematical Biosciences*, 111(1):1–71.

Huynen, M. A., Stadler, P. F., and Fontana, W. (1996). Smoothness within ruggedness: The role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401.

Jansen, W. (1987). A permanence theorem for replicator and lotka-volterra systems. *Journal of Mathematical Biology*, 25:411–422.

Jeffries, M. L. and Lawton, J. H. (1985). Predator-prey ratios in communities of freshwater invertebrates: the role of enemy-free space. *Freshwater Biol.*, 15:105–112.

Jensen, H. J. (1998). *Self-Organized Criticality.* Cambridge lecture notes in physics, Cambrigde.

Jung, P. (1997). Noise induced spatio-temporal coherence and self organized criticality in neuronal assemblies. *Preprint, Georgia Institute of Technology*, R 2106.

Kauffman, S. (1996). Investigations: The nature of autonomous agents and the worlds they mutually create. *Santa Fe Preprint*, 96-08-072.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, 22:437–467.

Kauffman, S. A. (1993). *The Origins of Order.* Oxford University Press, New York.

Kauffman, S. A. and Johnsen, S. (1991). Coevolution to the edge of chaos: Coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.*, 149:467.

Kenny, D. and Loehle, C. (1991). Are food webs randomly connected? *Ecology*, 72:1794–1799.

Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA*, 41:144.

Kimura, M. (1962). On the probability of fixation of mutant genes in populations. *Genetics*, 47:713–719.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–626.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Kimura, M. and Otha, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature*, 229:467–469.

Law, R. and Blackford, J. C. (1992). Self-assembling food webs: A global viewpoint of coexistence of species in lotka-volterra communities. *Ecology*, 73(2):567–578.

Lawton, J. H. and Warren, P. H. (1988). Static and dynamic explanations for patterns in food webs. *Trends in Ecology and Evolution*, 3:242–245.

Lenski, R. E. and Travisano, M. (1994). Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl. Acad. Sci.*, 91:6808.

Li, H., Helling, R., Tang, C., and Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science*, 273:666–669.

Li, W.-H. and Graur, D. (1991). *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Lotka, A. J. (1910). Contribution to the theory of periodic reactions. *J. Phys. Chem.*, 14:271–274.

Lotka, A. J. (1925). *Elements of Physical Biology*. Williams and Wilkins, Baltimore.

Luand, E. and Hamilton, R. (1991).

MacArthur, R. H. (1955). Fluctuations of animal populations and a meassure of community stability. *Ecology*, 36:533–536.

MacArthur, R. H. and Pianka, E. R. (1966). On optimal use of a patchy environment. *Amer. Natur.*, 100:603–609.

Mandelbrot, B. (1961). On the theory of word frequencies and on related markovian models of discourse. In Jacobson, R., editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. Amer. Math. Soc.

Mandelbrot, B. B. (1977). *The Fractal Geometry of Nature*. W. H. Freeman and Company, New York, 1st edition.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M., and Stanley, H. E. (1994). Linguistic features of noncoding dna sequences. *Phys. Rev. Lett.*, 73(23):3169–3171.

Margoliash, E. (1963). Primary structure and evolution of cytochrome C. *Proc. Natl. Acad. Sci. USA*, 50:672–679.

Martinez, N. D. (1991). Artifacts or attributes? effects of resolution on the litte rock lake food web. *Ecol. Monogr.*, 61:367–392.

May, R. M. (1971). Stability in multi-species community models. *Mathematical Biosciences*, 12:59–79.

May, R. M. (1972). Will a large complex system be stable? *Nature*, 238:413–414.

May, R. M. (1973). *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton, New Jersey.

Mills, D. R., Kramer, F. R., and Spiegelman, S. (1973). ?? *Science*, 180:916.

Mills, D. R., Peterson, R. L., and Spiegelman, S. (1967). An extracellular darwinian experiment with a self-dublicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA*, 58:217.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, Mass.

Murray, J. D. (1989). *Mathematical Biology*. Springer-Verlag, New York, 2nd edition.

Newman, M. E. J. (1996). Self-organized criticality, evolution and the fossil extinction record. *Proc. R. Soc. London B*, 263:1605.

Newman, M. E. J. (1997). A model of mass extinction. *J. Theor. Biol.*, 189:235–252.

Newman, M. E. J. and Engelhardt, R. (1998). Effects of selective neutrality on the evolution of molecular species. *Proc. R. Soc. London B*, 265(1402):1333–1338.

Newman, M. E. J. and Engelhardt, R. (1999). A model of neutral molecular evolution. *In Preparation*.

Newman, M. E. J. and Roberts, B. W. (1995). Mass extinction: Evolution and the effects of external influences on unfit species. *Proc. Roy. Soc. B*, 260:31.

Otha, T. (1972). Population size and rate of evolution. *J. Mol. Evol.*, 1:305.

Paine, R. T. (1988). Food webs: Road maps of interactions or grist for theoretical development? *Ecology*, 69(6):1648–1654.

Perline, R. (1996). Zipf's law, the central limit theorem, and the random division of the unit interval. *Phys. Rev. E*, 54(1):220–223.

Pimm, S. L. (1982). *Food Webs*. Chapman and Hall, London.

Pimm, S. L. (1991). *The Balance of Nature?: Ecological Issues in the Conservation of Species and Communities*. The University of Chicago Press, Chicago.

Pimm, S. L. and Lawton, J. H. (1977). The number of trophic levels in ecological communities. *Nature (London)*, 268:329–331.

Pimm, S. L. and Lawton, J. H. (1978). On feeding at more than one trophic level. *Nature (London)*, 275:542–544.

Polis, G. A. (1991). Complex trophic interactions in derserts: an empirical critique of food web theory. *Am. Nat.*, 138:123–155.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge, second edition.

Provine, W. B. (1986). *Sewall Wright and Evolutionary Biology*. University of Chicago Press, Chicago.

Prügel-Bennett, A. and Sharpio, J. (1994). An analysis of genetic algorithms using statistical mechanics. *Phys. Rev. Lett.*, 72:1305–1309.

Raup, D. (1986). Biological extinction in earth history. *Science*, 231:1528–1533.

Raup, D. M. (1991). *Extinction: Bad Genes or Bad Luck?* W W Norton and Company, New York.

Reidys, C. M., Stadler, P. F., and Schuster, P. (1997). Generic properties of combinatory maps and neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397.

Roberts, B. W. and Newman, M. E. (1996). Model for evolution and extinction. *J. Theor. Biol.*, 180:39–54.

Schoener, T. W. (1971). Theory of feeding strategies. *Annu. Rev. Ecol. Syst.*, 22:103–114.

Schoenly, K., Beaver, R. A., and Heumier, T. A. (1991). On the trophic relations of insects: a food web approach. *Am. Nat.*, 137:597–632.

Schuster, P. (1994). Extended molecular evolutionary biology: Artificial life bridging the gap between chemistry and biology. *Artificial Life*, 1:39–60.

Schuster, P. (1996). How does complexity arise in evolution. *Complexity*, 2:22–30.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). From sequences to shapes and back: A case study in rna secondary structures. *Proc. R. Soc. London B*, 255:397.

Schuster, P. and Stadler, P. F. (1997). Sequence redundancy in biopolymers: A study on RNA and protein structures. *Santa Fe Institute Preprint*, 97-07-067.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrica*, 42:425.

Smith, J. M. and Szathmáry, E. (1995). *The Major Transitions in Evolution.* W. H. Freeman, Oxford.

Sneppen, K. (1995). Extremal dynamics and punctuated co-evolution. *Physica A*, 221:168–179.

Sneppen, K., Bak, P., Flyvbjerg, H., and Jensen, M. H. (1995). Evolution as a self-organized phenomenon. *Proc. Natl. Acad. Sci. USA*, 92(11):5209–5213.

Sniegowski, P. D., Gerrish, P. J., and Lenski, R. (1997). Evolution of high mutation-rates in experimental populations of escherichia coli. *Nature*, 387:703.

Solé, R. V. and Bascompte, J. (1996). Are critical phenomena relevant to large-scale evolution? *Proc. R. Soc. Lond. B*, 263:161–168.

Solé, R. V. and Manrubia, S. C. (1996). Extinction and self-organization in a model of large-scale evolution. *Phys. Rev. E*, 54(1):R42–R45.

Solé, R. V. and Manrubia, S. C. (1997). Criticality and unpredictability in macroevolution. *Phys. Rev. E*, 55(4):4500–4507.

Solé, R. V., Manrubia, S. C., Benton, M., and Bak, P. (1997). Self-similarity of extinction statistics in the fossil record. *Nature*, 388:764–767.

Stauffer, D. and Aharony, A. (1994). *Introduction to Percolation Theory*. Taylor and Francis, London, revised 2nd edition.

Stenseth, N. C. and Maynard Smith, J. (1984). Coevolution in ecosystems: Red queen evolution or stasis? *Evolution*, 38(4):870–880.

Sumper, M. and Luce, R. (1975). ?? *Proc. Natl. Acad. Sci. USA*, 72:162.

Swetina, J. and Schuster, P. (1982). Self-replication with error: A model for polynucleotide replication. *Biophys. Chem.*, 16:329.

Tsonis, A. A., Shultz, C., and Tsonis, P. A. (1997). Zipf's law and the structure and evolution of languages. *Complexity*, 2(5):12.

van Nimwegen, E., Crutchfield, J. P., and Mitchell, M. (1997a). Finite populations induce metastability in evolutionary search. *Phys. Lett. A*, 229:144.

van Nimwegen, E., Crutchfield, J. P., and Mitchell, M. (1997b). Statistical dynamics of the royal road genetic algorithm. *Theor. Comp. Sci. (in press)*.

Van Valen, L. (1973). ?? *Evol. Theory*, 1:1.

Volterra, V. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. Acad. Lincei.*, 2:31–113. Translation can be found in an appendix to Chapman's *Animal ecology*, New York, 1931.

Warren, P. H. (1989). Spatial and temporal variation in the structure of freshwater food web. *Oikos*, 55:299–311.

Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R., and Holbrook, J. J. (1988). A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science*, 242:1541–1544.

Willis, J. C. (1922). *Age and Area*. Cambridge University Press, London.

Winemiller, K. O. (1990). Spatial and temporal variation in tropical fish trophic networks. *Ecol. Monogr.*, 60:331–367.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In Jones, D. F., editor, *Int. Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366.

Wright, S. (1967). Surfaces of selective value. *Proc. Nat. Acad. Sci.*, 58:165.

Wright, S. (1982). Character change, speciation and the higher taxa. *Evolution*, 36(3):427–443.

Yam, P. (1995). Talking trash. *Scientific American*, 273(3):24.

Yeomans, J. M. (1992). *Statistical Mechanics of Phase Transitions*. Oxford University Press, Oxford.

Yodiz, P. (1984). How rare is omnivory. *Ecology*, 65:321–323.

Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f. r. s. *Philosophical Transactions of the Royal Society of London B*, 213:21–87.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Reading, Massachusetts.

Zuckerkandl, E. and Pauling, L. (1962). Molecular disease, evolution and genic heterogenity. In Kash, M. and Pullman, B., editors, *Horizons in Biochemistry*, pages 189–225. Academic Press, New York.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In Bryson, V. and Vogel, J., editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York.

Glossary

1/f-noise "One-over-f noise", occasionally called "flicker noise" or "pink noise", is a type of noise whose **power spectrum** P(f) as a function of the frequency f behaves like: $P(f) = 1/f^a$, where the exponent a is very close to 1 (that's where the name "1/f noise" comes from). If we mix visible light with different frequencies according to 1/f distribution, the resulting light may be pinkish. Mixtures using other distributions should have different colors. For example, if the distribution is flat, the resulting light is white (so noise with $P(f)$ =constant **power spectrum** is called "white noise")

allele An alternative form of a **gene** at a **locus**.

advantageous mutation A mutation that increases the **fitness** of the organism or structure which carries the mutation.

amino acid Organic molecule which is the subunit building block for **proteins**.

auto-catalysis The ability of substances to catalyze the production of themselves. So, when the product of a chemical reaction also is a catalyst of the reaction, auto-catalysis takes place.

bacteriophage A **virus** that multiplies in bacteria.

balancing selection A mode of selection, also called overdominance, or stabilizing selection, where the **heterozygote** has a higher **fitness** than either **homozygote**.

base see **nucleotide**.

base pair A base pair is a pairing of two **nucleotides** through a hydrogen-bond according to the Watson-Crick base pairing rules between a **purine** and a **pyrimidine**.

bifurcation When a non-linear dynamic system develops twice the possible solutions that it had before it passed its **critical level**. A bifurcation

cascade is often called the period doubling route to **chaos** because the transition from an orderly system to a chaotic system often occurs when the number of possible solutions begins increasing, doubling each time.

carrying capacity The maximum number of individuals in a population of a given **species** that can be sustained in a specified habitat.

central limit theorem This is "The Law of Large Numbers", and states that as a sample of independent, identically distributed random numbers approaches infinity, its probability density function approaches the **normal distribution**.

chaos A deterministic non-linear dynamic system that can produce random looking results. A chaotic system must have a **fractal** dimension, and exhibit sensitive dependence on initial conditions.

chromosome In prokaryotes, the **DNA** molecule containing the **genome**. In eukaryotes, a linear **DNA** molecule complexed with **proteins** forming a thread-like structure containing the genetic information.

codominance The "normal" mode of selection, also called positive -, or directional -, or genic selection, where the **fitness** of one **allele** is larger than the other so that the **fitness** of the **heterozygote** is the mean of the two **homozygotes**.

complexity theory The theory that processes with a large number of seemingly independent agents can spontaneously organize themselves into a coherent system.

consensus sequence A sequence that represents the most prevalent **nucleotides** or **amino acids** at each site in a population of sequences.

correlation : The degree to which factors influence each other.

critical state A situation where the values of control parameters reach a certain level where the nature of a non-linear dynamic system changes. The system can **bifurcate**, or make the transition from stable to turbulent behavior.

deleterious mutation A mutation that lowers the **fitness** of the organism or structure which carries the mutation.

diploid A **chromosomal** complement that contains two copies of each **chromosome**.

DNA (deoxyribonucleic acid): The molecule that encodes genetic information. DNA is a double- stranded molecule held together by weak bonds between **base pairs** of **nucleotides**. The four **nucleotides** in DNA contain the bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, **base pairs** form only between A and T and between G and C; thus the base sequence of each single strand can be deduced from that of its partner.

dominance The property of an **allele** to manifest its entire **phenotypic** effect in the **heterozygote**.

enzyme A **protein** that catalyzes a specific chemical reaction.

error threshold A **critical value** of the mutation rate, above which errors accumulate leading to the complete loss of information.

exponential law A mathematical expression in which a quantity is multiplied or divided by the same factor in equal intervals. For example, a population grows exponentially when it doubles every n years.

fitness A measure of the relative survival and **reproductive** success of an individual or a **genotype**.

fixation The situation achieved when an **allele** reaches a frequency of 100% in a population.

Fourier transform The mathematical technique for changing the time-domain representation of a signal (its waveform) into a frequency-domain representation (its spectrum).

fractal An object in which the parts are in some way related to the whole. That is, the individual components are "**self-similar**". An example is the branching network in a tree. While each branch, and each successive smaller branching is different, they are qualitatively similar to the structure of the whole tree.

fractal distribution A probability density function that is statistically **self-similar**. That is, in different increments of time, the statistical characteristics remain the same.

Gaussian distribution A system whose probabilities are well described by the **normal distribution**, or bell shaped curve.

gene A sequence of genomic **DNA** or **RNA** that is essential for a specific function.

genetic algorithms Models that optimize rules by mimicking the Darwinian Law of survival of the fittest. A set of rules are chosen by those that work the best. The weakest are discarded. In addition, two successful rules can be combined (the equivalent to genetic cross-overs) to produce offspring rules. The offspring can replace the parents, or they will be discarded if less successful than the parents. Mutation is also accomplished by randomly changing elements. Mutation and cross-over occur with low probability, as in nature.

genetic drift The fluctuation of **allele** frequencies from generation to generation caused by chance events.

genotype The sum of all the genetic information present in an organism.

haploid A cell or organism having a single set of unpaired **chromosomes**.

Hardy-Weinberg-equilibrium A condition under which the overall **genotypic** frequencies in a **diploid** population are constant in time.

heterozygote A **diploid** individual with different **alleles** at the **locus** in question.

homozygote A **diploid** individual with identical **alleles** at the **locus** in question.

hypercycle equations Set of equations governing a cyclic coupling pattern relating individual **reproduction** cycles.

inverse power law A special kind of **power law** with an exponent close to -1, also called **Zipf's law**.

limit cycle An attractor for non-linear dynamic systems which has periodic cycles or orbits in phase space. An example is an undamped pendulum which will have a closed circle orbit equal to the amplitude of the pendulum's swing.

locus The site on a **chromosome** where a particular **gene** is located.

mass action kinetics A thermodynamic relationship that lays down the quantities of different reaction partners that can coexist with one another in equilibrium.

molecular clock The belief that, in any given **gene** or **DNA** segment, mutations accumulate in an approximately constant rate in all evolutionary lineages as long as the **gene** or **DNA** segment retains its original function.

natural selection The differential **reproduction** of structurally distinct individuals within a population.

neo-Darwinism The theory of evolution according to Darwin and expanded by the laws of genetics and population biology (such as the **Hardy-Weinberg law** and the **fundamental theorem**) in the first half of the 20th century.

neutral mutation A mutation that does not change the **fitness** of the organism, and thus has no selective advantage.

neutral networks A network of genetic or **nucleotide**-configurations connected my one **point mutation**. In the case of **RNA** we identify one neutral network in the space of possible sequences as one particular realization of a structure and thus of a molecular **phenotype**.

neutral theory The idea that evolution at the molecular level is primarily determined by mutational input and random **genetic drift**, rather than by **natural selection**.

nucleic acid **DNA** or **RNA**

nucleotide A molecule composed of a nitrogen base, a sugar and a phosphate group. The basic building blocks of **nucleic acids**.

phenotype The morphological result of the expression of the **genes** in an organism. So it is the organism itself with all its characteristics. When we talk about "**RNA-phenotypes**", we just mean the spatial structure of the **RNA**'s.

point mutation A mutation of just one **base pair** in **RNA** or **DNA**.

polymorphism The coexistence of two or more **alleles** at one **locus**.

power law A mathematical expression of the form $f(x) = x^a$ characteristical for **self-similar** and **fractal** phenomena. Power law relationships have no characteristic scale for their decay (or increase) - they are scale-free. This means that the resulting "patterns" looks similar at all scales.

power spectrum The presentation of the square of the amplitudes of the signal of a time series as a function of the frequency of the signal. The power spectrum is proportional to the **Fourier transform** of the autocorrelation function.

protein A molecule composed of one or more polypeptide chains.

purine A type of nitrogen base present in **nucleotides** and composed of two joined ring structures, one five-membered and the other six-membered. The purines in **DNA** or **RNA** are adenine and guanine.

pyrimidine A type of nitrogen base present in **nucleotides** and composed of a single six-membered ring. The pyrimidines in **DNA** or **RNA** are cytosine and thymine and cytosine and uracil respectively.

Qβ virus An **RNA bacteriophage** with a single stranded **genome**, whose host is the bacterium E.coli.

recessive The lack of **phenotypic** expression of an **allele** in the **heterozygote**.

replication Molecular biologists use the word replication to describe the process by which a new strand of **nucleic acid** is synthesized on a template, where he speaks of **DNA** or **RNA** replication. When speaking about larger structures like a cell, one talks about **reproduction**, not replication. Intermediate structures, like **viruses** or even **proteins**, do only replicate when the information contained in them is directly used for the synthesis of a copy of themselves.

reproduction Good old sex used by larger, multicellular organisms, in contrast to **replication**.

RNA Ribonucleic acid. A macromolecular polymer of linked **nucleotides** in which the sugar residue is ribose. Single stranded.

scaling How the characteristics of an object change as you change the size of your measuring device. For a three dimensional object, it could be the volume of an object covered as you increase the radius of a covering sphere. In a times series, it could be the change in the amplitude of the time series as you increase the increment of time.

secondary structure In **proteins** and **nucleic acids**, the structure of the molecule brought about by the formation of hydrogen bonds between **amino acids** or **nucleotides**, respectively.

self-similarity When small parts of an object are qualitatively the same, or similar to the whole object. In certain deterministic **fractals**, like the Sierpinski Triangle, small pieces look the same as the entire object. In random **fractals**, small increments of time will be statistically similar to larger increments of time.

self-replicating RNA Self-replicating **RNA**'s, are the most promising **autocatalytic** entities which have been discovered experimentally. Joyce, Szostak and others have used a technique of molecular evolution to produce an **RNA** molecule which functions as a polymerase and moves along a sequence of **RNA** and add the appropriate **nucleic acids** on the right places. The resulting amplification cycle is carried out in a plastic test tube that provides a kind of membrane to keep it in one place and has a channel through which energy and nutrients are introduced. The process can even undergo Darwinian selection, but it ultimately runs down, because the **enzyme** catalysts in the system are not themselves **replicated**.

sequence space A multidimensional hypercube used for the theoretical representation of all possible variants of a sequence.

species A basic taxonomic category for which there are several definitions. The most use definition is that of a group of actually (or potentially) inter-breeding individuals that is **reproductively** isolated from other such groups.

strange attractor An attractor in phase space, where the points never repeat themselves, and orbits never intersect, but they stay within the same region of phase space. Unlike **limit cycles** or point attractors, strange

attractors are non-periodic, and generally have a **fractal dimension**. They are a picture of a non-linear, **chaotic** system.

taxon A taxonomic group of any rank (e.g. **species**, genus, kingdom).

tertiary structure In **proteins** and **nucleic acids**, the three-dimensional structure of the molecule brought about by its folding upon itself.

virus A non-cellular biological entity that can reproduce only within a host cell. Viruses consist of **nucleic acid** covered by **protein**; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus.

wild type The best adapted **genotype** of a **species** and representing the majority of the individuals of the **species**. In the molecular **quasi species**, the wild type is equivalent to the **consensus sequence**, that is, The sequence of **nucleotides** or **proteins** containing the most prevalent **allele** or **amino acid** at each site.

Zipf's law Zipf's law, named after the Harvard linguistic professor George Kingsley Zipf (1902-1950), is the observation that frequency of occurrence of some event P, as a function of the rank i when the rank is determined by the above frequency of occurrence, is a **power law** function $P_i \sim 1/i^a$ with the exponent a close to unity.

Index