PHASE TRANSITIONS IN FLUIDS AND BIOLOGICAL SYSTEMS

BY

MAKSIM SIPOS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Yoshitsugu Oono, Chair
Professor Nigel Goldenfeld, Director of Research
Professor Richard L Weaver
Assistant Professor Yann R Chemla

Abstract

In this thesis, I consider systems from two seemingly different fields: fluid dynamics and microbial ecology. In these systems, the unifying features are the existences of global non-equilibrium steady states. I consider generic and statistical models for transitions between these global states, and I relate the model results with experimental data. A theme of this thesis is that these rather simple, minimal models are able to capture a lot of functional detail about complex dynamical systems.

In Part I, I consider the transition between laminar and turbulent flow. I find that quantitative and qualitative features of pipe flow experiments, the superexponential lifetime and the splitting of turbulent puffs, and the growth rate of turbulent slugs, can all be explained by a coarse-grained, phenomenological model in the directed percolation universality class. To relate this critical phenomena approach closer to the fluid dynamics, I consider the transition to turbulence in the Burgers equation, a simplified model for Navier-Stokes equations. Via a transformation to a model of directed polymers in a random medium, I find that the transition to Burgers turbulence may also be in the directed percolation universality class. This evidence implies that the turbulent-to-laminar transition is statistical in nature and does not depend on details of the Navier-Stokes equations describing the fluid flow.

In Part II, I consider the disparate subject of microbial ecology where the complex interactions within microbial ecosystems produce observable patterns in microbe abundance, diversity and genotype. In order to be able to study these patterns, I develop a bioinformatics pipeline to multiply align and quickly cluster large microbial metagenomics datasets. I also develop a novel metric that quantifies the degree of interactions underlying the assembly of a microbial ecosystem, particularly the transition between neutral (random) and niche (deterministic) assembly. I apply this metric to 16S rRNA metagenomic studies of 6 vertebrate gastrointestinal microbiomes and find that they assembled through a highly non-neutral process. I then consider a phase transition that may occur in nutrient-poor environments such as ocean surface waters. In these systems, I find that the experimentally observed genome streamlining, specialization and opportunism may well be generic statistical phenomena.

Ad astra per aspera.

Acknowledgments

I acknowledge my advisor, Nigel Goldenfeld, first and foremost. Nigel has spent an astounding amount of time working with me in research. His many inventive ideas form the basis of this entire thesis, and this thesis would not have been possible without them. He has put a lot of hard work into many of the research topics presented here: he worked out calculations, he discussed and met with me at all times of the day, he analyzed data and reanalyzed it, he drafted write ups and he patiently edited mine. He charted the objectives, focused (and refocused) the research and provided an always-helping hand for various hurdles along the way.

Nigel is brilliant and holds true to his reputation of being interested in everything. It was a great honor to witness his *modus operandi* and to work with him. I hope I will be able to one day match, in just one field, Nigel's wide contributions and perspectives in Everything. Nigel's help went far beyond the research projects we have collaborated on. He has also taken a genuine interest in my well-being after the PhD. He obviously cares deeply that I should have a strong and successful career and that I should find personal happiness. I am truly honored to know Nigel and he will stay my lifelong friend.

I owe my academic career to my family. My parents Ljiljana and Djurica Šipoš have shown tremendous forward-thinking, faith and optimism in raising me. Their decision to buy a family computer when I was very young made a profound influence on my life. I am deeply thankful that they trusted me to have an independent and far-away education. Without my parents and their support, I would not have been able to study physics in the US. I am also indebted to my sister Marijana (*et* Serge) and my brother Vladimir (*i* Silvija) for their profound role in my education. In particular, my brother taught me the ropes of DOS and Basic a long-long time ago and he deserves significant credit for all code I write in my career. I owe my American family, Cathy, Eric, Ryan and Evan Haines for unselfishly hosting me and making my days at Ithaca blissful. Finally, Orsolya Haja deserves my deepest gratitude for being the sole supporting pillar of my life: she is the primary source of my motivation and optimism.

I am also indebted to the staff at Ithaca College for providing me with a caring and flexible undergraduate education. I am especially thankful to Bruce G. Thompson, David A. Brown, Luke D. Keller and Ali Erkan

for their hard work and patience in advising and teaching me. I have also learned a great deal from the countless other passionate faculty in the Physics, Mathematics and Computer Science departments. I am also thankful for the practical computer programming skills I learned while interning at Grammatech, Inc.

My close collaborators (who are also friends) deserve special mention: Bryan A. White, Patricio Jeraldo, Nicholas Chia and Hong-Yan Shih. Without them, some portions of this work would not have been possible. I am also especially thankful that I learned a great deal from colleagues and friends Andreas Menzel, Nicholas Guttenberg and Tom Butler. I am grateful for the wonderful and conducive environment in the Goldenfeld group and its other present and former members Luiza Angheluta, Michael Assaf, Zhenyu Wang, Farshid Jafarpour, Vikyath Deviprasad Rao and K. Michael Martini. Former members of the Institute for Genomic Biology Suleyman Yildrim and Carl Yeoman also deserve my thanks. Finally, I must thank my various other friends for the fun experiences, intelligent discussions and caring advice: Andrija, Andrada, Nitin, Alex, Anusha, Chris, Hamood, Marina, Jasmina, Yun, Onyeama and Rito.

This work was partially supported by the National Science Foundation and the L. S. Edelheit Fellowship.

Table of Contents

List of Abbreviations

CH	Calinski-Harabasz clustering quality
DP	Directed percolation
DPRM	Directed polymer in random medium
OTU	Operational taxonomic unit
PCA	Principal component analysis
TORNADO	Taxon Organization from RNA Dataset Operations

Chapter 1

Introduction

My work has been seemingly in different fields: microbial ecology and fluid dynamics. In particular, I have considered phase transitions in these systems. The underlying question that I asked in this thesis is: how do complex, spatially-extended and non-equilibrium systems transition between global states? My interest has been in answering aspects of this question in the context of quantitative biology and microbial ecology, as well as in fluid dynamics.

1.1 Transition to Turbulence

In fluid dynamics, my research has focused on the study of the transition to turbulence. Some background information about transition to turbulence is given in Chapter 2, but we give a brief introduction here. The phase of fluid flow depends upon the Reynolds number, Re, the dimensionless ratio of the inertial forces to viscous dissipation. In pipe flows, even at high Re, laminar flow is possible but the slightest disturbance nucleates the state of turbulence. The regions of turbulence (turbulent slugs) can expand but they can also spontaneously decay back to laminar flow (turbulent puffs). In physics, the abstract process of directed percolation has these same properties: sites on a lattice that are active can expand or spontaneously become inactive (absorbed).

Directed percolation is a ubiquitous process in nature: it occurs in a wide range of problems such as fluid percolation through porous rock, species expansion in ecology and spread of infections and immunity. Could the transition to turbulence problem in fact be in the directed percolation universality class? Yves Pomeau had this intuition in 1986 [1] and I set out to quantitatively test its validity. The results of this work are presented in Chapter 3. I simulated the directed percolation process in 3+1 dimensions in a pipe geometry. Under the critical threshold, I found that the characteristic lifetimes of the active states are superexponential in percolation probability. The same functional form for lifetimes of puffs in pipe flow experiments has been observed experimentally [2]. Also, the front growth rates of active regions in supercritical directed percolation match those observed in turbulent slug experiments [3]. Without using any details of the Navier-Stokes equation of fluid flow, I find that Directed Percolation model fits the functional form of many generic properties of pipe flow turbulence, with no adjustable parameters.

Pomeau gave intuitive arguments but no proof to support his conjecture. It is important to be able to show in more detail how the mapping between Navier-Stokes turbulence and directed percolation might arise. To answer this, I considered the transition to turbulence in Burgers Equation. This work is described in Chapter 4. Burgers' equation is a simplified form of Navier-Stokes equation, with the pressure term dropped. The pressure term is important for creation of vorticity in turbulence. However, it may be the case that at the directed percolation transition point the pressure is an irrelevant variable in the sense of Wilson [4]. The Burgers equation has a mapping to the KPZ equation, which in turn has a mapping to a statistical model of directed polymers in a random external potential. In this directed polymers model, it is known that scaling of the polymer fluctuations can undergo a directed percolation transition [5].

Taken together, this quantitative evidence suggests that the transition to turbulence occurs as a statistical phenomenon, a non-equilibrium phase transition in the directed percolation universality class. The details of Navier-Stokes equations governing the flow are not necessary to explain the universal features of observed puff lifetimes and turbulent front growth rates in pipe flow experiments. More generally, the transition to turbulence in any fluid described by Navier-Stokes equation appears to be in the Directed Percolation universality class.

1.2 Microbial Ecology

My study of microbial ecology began with my interest in the forces that structure communities of microbes, such as those inhabiting the gastrointestinal tract or waters in the oceans. Because culturing bacteria in the environment is essentially impossible in such cases, for many reasons, research focuses on sampling cellular DNA/RNA from these communities, using techniques known as metagenomics. In metagenomic studies, one sequences the living matter within an environment in order to ascertain the composition of the microbiome, the community formed by the microbes living within that environment. One can observe interesting ecological patterns in metagenomic datasets, such as functional forms of the species abundance curves and the extent of their diversity. Metagenomics and ecology is introduced in more detail in Chapter 5.

The questions one naturally asks are: what species are present? In what abundance? And how does this community drive the function and evolution of the ecosystem? In the case of microbes, these questions are hard to answer because the species concept is not strictly applicable, and one has to resort to genomics to identify, classify and count the organisms.

In order to be able to study these patterns, I developed a bioinformatics pipeline to process microbial metagenomics datasets. This pipeline is described in Chapter 6. While developing this pipeline, I addressed the problem of multiply aligning large metagenomic 16S rRNA datasets in such a way that analysis artifacts are kept at a minimum. Secondary structure (base pairing structure) is important for RNA sequences, and hence one should use a secondary-structure aware aligner for regions with strong secondary structure. For regions with no fixed secondary structure (hyper-variable regions), template based aligners are appropriate. I showed that the procedure of combining the two alignment methods using our pipeline outperforms each tool used separately.

To correctly study the ecology of a microbiome using 16S rRNA sequencing, it is also necessary to be able to cluster the sequenced 16S tags into OTUs (Operational Taxonomic Units), a proxy for species that is widely used by microbiologists. I developed a tool that can perform a good clustering procedure in a way that minimizes the artifacts due to overlapping clusters. The clustering algorithm is also sufficiently fast that it is applicable to larger datasets. I also investigated more heuristic clustering algorithms that may be applicable to even larger, contemporary datasets. The clustering work is described in Chapter 7.

With the appropriate bioinformatics tools developed, I turned my attention to the problem of understanding the diversity, composition and ecological community structure of the gastrointestinal microbiomes. The perspective that I acquired is a physical one: the OTUs are considered as points in a high-dimensional sequence space. With this view of a metagenomic dataset, the next question was: how do microbial communities assemble? When capacity in the environment becomes available for an additional bacterium, where does this additional bacterium come from? Does it enter the community through a cell division, or immigration? I implemented fast large dataset dimensional reduction methods, as well as physical models and digital life simulations to study this problem. As a result, I have formulated a novel metric that quantifies the degree by which interactions underlie the assembly of a microbial ecosystem.

I applied this metric to 16S rRNA metagenomic studies of 6 vertebrate gastrointestinal microbiomes and found that they assembled through a highly non-neutral process. This is despite the fact that OTU abundance curves in these datasets perfectly match the results of Hubbell's Neutral Theory which assumes that species (OTUs here) behave as if they were functionally equivalent. This indicates that abundance approaches alone are not sufficient in studying microbial communities; an approach that combines abundance and genomic data is necessary. This work is presented in Chapter 8.

I then turned my attention to the study of ocean picoplankton microbiome. This work is described in Chapter 9. Again, a fusion of genomic and abundance data shows that, in these energy-starved conditions, the picoplankton populations split into two different subpopulations. One population is abundant, specialized and adapted to living in such conditions. The other population is rare, generalist and opportunistic. I modeled this system with a rather general, energy-centric agent model. I found that, in steady state conditions, the organisms in the model simulations undergo an interesting symmetry-breaking transition towards a specialized state. They also streamline their genome shedding unnecessary functions, much like the picoplankton microbes do. However, in fluctuating conditions, the opportunistic plankton thrive. In the theme of this thesis, there is a phase transition of the evolution of the genomes of the microbial plankton as a function of the environmental conditions.

1.3 My contributions

All calculations and numerical simulations presented in this work have been performed by me in close collaboration with Nigel Goldenfeld. All the figures and plots have been generated by me apart from the following exceptions. The theoretical work in Chapter 4 has been in close collaboration with Hong-Yan Shih. Work in Chapters 6 and 8 has been in close collaboration and discussion with Patricio Jeraldo and Nicholas Chia. In particular, Patricio Jeraldo developed the merging script described in Section 6.2.1, and cleanup, quality filtering and taxonomy steps in the TORNADO pipeline. He also originally aligned, clustered and analyzed the chicken caecum datasets, and he created the Figure 8.9. Vikyath Deviprasad Rao created the video tutorials for the TORNADO pipeline. Work in Chapter 7 will be published in collaboration with Carl Yeoman. Carl Yeoman generated Figures 7.1 and 7.2 and conducted the study in Section 7.3.1.

1.4 List of publications

Three peer-reviewed publications make their appearance in this thesis. The bulk of Chapter 3 has previously been published in [6]. The chapter is slightly extended here with additional text and figures that could not fit into a PRE(R) publication. Chapter 6 has been slightly extended from the original publication in [7]. Finally, the bulk of Chapter 8 has been published originally in [8], but extended here with some further discussion. Further publications are planned for work described in chapters 4, 7 and 9.

Part I

Transition to Turbulence in Fluids

Chapter 2

Introduction to Fluid Dynamics, Transition to Turbulence and Directed Percolation

In its simplest classification, a fluid flow can be said to be laminar (smooth, regular) or turbulent (erratic, irregular). For Poiseuille flow [9, 10] (pipe flow), in the laminar case the fluid velocity along the axis of the pipe has a parabolic profile: the fluid flows the fastest along the central axis of the pipe and its velocity is zero along the walls of the pipe. There is virtually no flow along the other components of the velocity field. Flow is smoothly confined to the concentric circular layers, laminas, and hence it is termed laminar. In the case of turbulent flow, the flow field has non-zero fluctuating components perpendicular to the pipe axis. Eddies form in the pipe as result of fluctuations in the pressure. In this Part, we will investigate how the transition between these two types of flow occurs.

We will not study the precise mechanisms that trigger this transition. Generally, any sufficiently large fluctuation will initiate a transition from a laminar flow to a turbulent one, at sufficiently high fluid velocity [11]. Rather, we are interested in what we can say about the generic properties of this transition (those properties not dependent upon the specifics of the experimental setup). How does the turbulent state, once developed, expand into the laminar state? How long does the turbulent state persist? What mathematical analogues can be used to tie in this grossly statistical perspective to the actual fluid dynamics as embodied in the Navier-Stokes equations?

In this Chapter, we begin this study by recounting a brief history and phenomenology of the subject in Section 2.1. Section 2.2 introduces the Navier-Stokes equations of fluid flow and the statistical laws of Turbulence. In Section 2.3 we introduce the relevant statistical theory: the universality class of Directed Percolation and we described its experimental verification.

2.1 Brief History and Introduction

Pipe flow was first systematically studied by Poiseuille in the mid 19th century [9], who discovered the laminar parabolic velocity profile and dependence of pressure, pipe length and radius on flow velocity. In parallel, Hagen studied the same relationships and noted the laminar-turbulent transition and the perturbative effects of the entrance of the pipe [10]. The transition was first studied in detail by Reynolds in the late 19th century [12]. To characterize the flow state, he introduced the dimensionless parameter bearing his name: $\mathrm{Re} \equiv UL/\nu$. Here U and L are characteristic flow velocity and length scales, and ν is the kinematic viscosity.

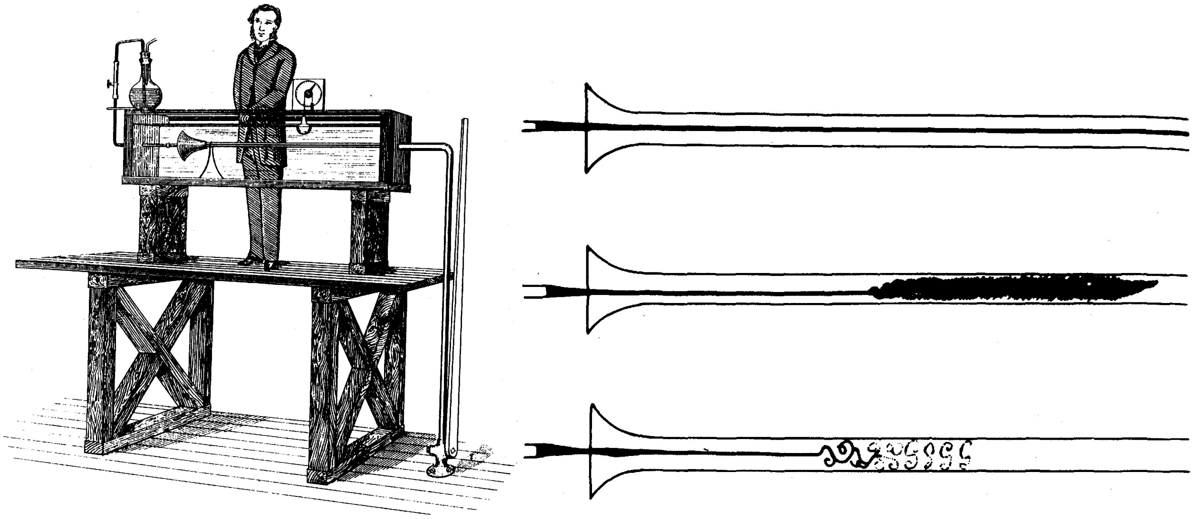


Figure 2.1: Sketches of the Reynolds experiment (left) and his dye observations (right) from [12]. Images scanned from [12] by Emmanuel S. Boss.

Reynolds' experimental setup was a large glass pipe with a flared entrance that carried water through it (See Figure 2.1). A thin pipe carrying dye was placed near the entrance of the pipe. Reynolds observed the dye as it advected down the pipe carrying water. In some cases the dye advected in straightforward jet fashion, but when fluid velocity was large (or the pipe wide enough), the dye stream would sometimes mix rapidly at points along the pipe in puff-like motion. Reynolds noted the spontaneousness and the abruptness of this transition (it could occur anywhere along the pipe) and he termed the resulting motion sinuous. This was an important first observation of the sensitive laminar-turbulent transition in pipes, as well as the first observation of the enhanced mixing effect of turbulence. An interesting recollection of Reynolds' experiments is given in [13] in context of the period and his peers.

Many studies have attempted to elucidate the transition to turbulence in a pipe since the time of Reynolds. In the 1970's, Wygnanski *et al.* [14] used hot wire anemometers and electronics to systematically describe the phase diagram of the transition as function of Re. They had a long pipe (500 diameters), carrying air, and a tunable inlet obstruction (serving as the source of the disturbance). At sufficiently high inlet disturbance and at Re $\sim$ 2000 they observed that the turbulent regions in laminar pipe flow appear as metastable "puffs" that can split and decay [15]. These puffs are localized regions of turbulence that fill the cross-section of the pipe. They extend along the pipe and have wide, noisy and rough leading and trailing edges. Wygnanski *et* *al.* also noticed, at higher Re ~ 2700 and smaller inlet disturbance that turbulence can appear in the form of stable regions of turbulent flow called "slugs". Unlike puffs, slugs grow with time and have clearly defined, sharp turbulent-laminar interfaces.

This interesting phenomenology observed by Wygnanski *et al.* is in seeming contrast to the fact that laminar pipe flow is linearly stable to perturbations. Whereas the calculations of linear stability have to be performed numerically [16], they have been extended to very large Re $= 10,000,000$ [17] and they show that laminar Poiseuille profile is linearly stable to all perturbations. Since finite disturbances can trigger a transition to the turbulent state, as Wygnanski *et al.*'s experiments showed, this indicates a subcritical pitchfork bifurcation in chaos theory [18]. In fact, it is known that near the transition the dynamics can be decomposed into the interaction of a small number of nonlinear modes, that have been predicted based on direct numerical simulations [19, 20, 21], and observed in experiments using high-speed particle image velocimetry [22, 23].

The phase diagram of the transition from laminar to turbulent flow in a pipe is still an active area of investigation today [24, 25, 26, 27]. The estimations for the values of Re at transitions between phases have varied widely. It is known that for sufficiently low Re the fluid flow is always laminar and any turbulent disturbances decay immediately. Traveling wave solutions are possible at Re < 1650 [19, 20, 21, 22]. When $1650 <$ Re < 2050, turbulent puffs appear. They are metastable and memoryless–their survival probability is exponentially distributed:

$$P(\mathrm{Re}, t) = \exp\left(-\frac{t - t_0}{\tau(\mathrm{Re})}\right), \tag{2.1}$$

where $\tau(\mathrm{Re})$ is the average lifetime which was found to grow superexponentially with Re [2]:

$$\tau(\mathrm{Re}) = \tau_0 \exp[\exp(c_1 \mathrm{Re} + c_2)] \tag{2.2}$$

where τ_0, c_1 and c_2 were fit from the empirical data. For larger values of Re, the characteristic lifetime of puffs grows, and they begin to split and show complex spatiotemporal behavior [28, 29]. In this regime, the puffs interact with a characteristic distance: if puffs come too close to each other, they may spontaneously annihilate [30]. The average random time between two puff split events reduces in superexponential manner as Re increases from 2000 to 2400. At Re $= 2040$, the time between two puff split events is roughly the same as the lifetime of the puff at Re $= 2040$, thereby leading some authors to term Re $= 2040$ as the critical point for sustained turbulence in pipe flows [27, 31]. The splitting process and spatiotemporal intermittency continue as Re is increased, and turbulence is easier to induce. It is known that the threshold to induce turbulence scales as Re^{-1} [11]. When Re exceeds a critical value $\mathrm{Re}_c \sim 2500$, for a sufficiently large inlet disturbance, a slug grows with clearly defined, sharp, turbulent-laminar interface and a growth rate that scales approximately with $\sqrt{\mathrm{Re} - \mathrm{Re}_c}$ [3, 32].

2.2 Fluid Dynamics

In this Part, we will use some concepts from fields of Fluid Dynamics and Turbulence in motivating and analyzing our models. Hence, in Section 2.2.1 we give some elementary introduction to the key equations of fluid flow, the Navier-Stokes equations. Then we introduce some relevant background on what is meant by Turbulence in Section 2.2.2.

2.2.1 Navier-Stokes Equation

When one considers the stresses acting on an element of a fluid, and one writes the Newton's law of motion for that element, one obtains the Navier-Stokes equations [33]

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \nu \nabla^2 \mathbf{u} + \xi, \tag{2.3}$$

where the incompressibility of the flow is expressed via

$$\nabla \cdot \mathbf{u} = 0. \tag{2.4}$$

Here $\mathbf{u}(\mathbf{x}, t)$ is the fluid velocity at position $\mathbf{x}$ and time t, ν is the kinematic viscosity, ξ is the external driving force, which may include a stochastic component, P is the pressure and ρ is the fluid density. The pressure is implicitly defined by (2.4). In particular, the term $-(1/\rho)\nabla P$ has the effect that the evolution of the $\mathbf{u}$ field does not violate (2.4). One can show that P satisfies a Poisson equation at time t explicitly in terms of $\mathbf{u}$ via [34]

$$\nabla^2 P = -\rho \frac{\partial U_i}{\partial x_j}\frac{\partial U_j}{\partial x_i}. \tag{2.5}$$

Setting a length scale L for $\mathbf{x}$, a velocity scale U for $\mathbf{u}$, and a time scale L/U for t, yields

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\nabla P + \frac{1}{\mathrm{Re}}\nabla^2 \mathbf{u} + \xi, \tag{2.6}$$

where $\mathrm{Re} = UL/\nu$. Here, we can see that the Re is the ratio of the strength of the non-linear inertial force (how much the fluid resists motion, UL) and the viscous dissipation (ν). Turbulence occurs when Re is large and the nonlinear effects are large in comparison to the viscous Laplacian "smoothing" of the velocity field.

The fluid vorticity,

$$\boldsymbol{\omega} \equiv \nabla \times \mathbf{u} \tag{2.7}$$

satisfies the equation of motion

$$\partial_t \boldsymbol{\omega} + (\mathbf{u} \cdot \nabla)\boldsymbol{\omega} = \nu \nabla^2 \boldsymbol{\omega} + \boldsymbol{\omega} \cdot \nabla \mathbf{u} + \nabla \times \xi. \tag{2.8}$$

(2.8) can be obtained by taking the curl of the Navier-Stokes equations. The pressure term dropped out since the curl of a gradient is always zero. Here, the $\boldsymbol{\omega} \cdot \nabla \mathbf{u}$ is termed the vortex stretching term, and it only exists in 3-dimensional flows. In 2-dimensional flows, this term is missing, and this is a reason why 2-dimensional turbulence is qualitatively different from 3-dimensional turbulence. One of the consequences of this difference is that 2D turbulence has two types of cascade, each of which is characterized by a different energy spectrum in the inertial range (see the following section).

2.2.2 State of Turbulence

In this section we briefly describe the relevant theoretical results on transition to turbulence, and the state of developed turbulence.

Once the transition occurs, we must ask a question: what does it mean for a fluid to be turbulent? One aspect of turbulence is the randomness, namely the seemingly random fluctuations in the velocity field. The slight perturbations in the initial conditions and boundary conditions of the turbulent flow, and the extreme sensitivity to those perturbations mean that turbulent flows are chaotic in nature. Consider, for example, the motion $x(t)$ of a particle embedded in a fluctuating velocity field $\mathbf{u}(x,t)$. This path $x(t)$ cannot be predicted for large t in a real laboratory setting because it is impossible to predict the fluctuations of $\mathbf{u}$ field at sufficient precision. In terms of chaos theory, we would characterize the motion $x(t)$ of this particle as having a positive Lyapunov coefficient [18]. However, one important distinction between fluid turbulence and low-dimensional chaos is that the dynamics in the former involves an infinite number of degrees of freedom, due to the spatial extension of the turbulent velocity field $\mathbf{u}(\mathbf{x},t)$.

Another aspect of turbulence is the energy cascade, introduced by Richardson in 1922 [35]. The idea is that we define a turbulent eddy, which is a roughly coherent dynamic structure with length scale l. The eddy is unstable though, and can break apart into smaller eddies, and through this breakup process kinetic energy is transferred to smaller length scales without dissipation. This Hamiltonian process continues until the smallest possible eddies form which are stable, and dissipate energy through molecular viscosity. This point was illustrated poetically by Richardson:

Big whorls have little whorls,
Which feed on their velocity;
And little whorls have lesser whorls,
And so on to viscosity.

This idea is illustrated in Figure 2.2. At high Re, the energy transfer rate ϵ in a dynamical steady state is given by the energy entering the system which is u_0/τ_0 where u_0 is the characteristic velocity of the large eddies and $\tau_0 = l_0/u_0$ is their turnover time. Hence, we have that

$$\epsilon = \frac{u_0^3}{l_0} \tag{2.9}$$

is independent of the viscosity ν. The length scale at which the energy dissipates, η, was worked out by Kolmogorov in 1941 [36] based on dimensional grounds. Since η can only depend upon ν and ϵ, then the only length scale possible is given by

$$\eta = (\nu^3/\epsilon)^{1/4} = \mathrm{Re}^{-3/4} l_0. \tag{2.10}$$

The scales η and l_0 define the *inertial range* of the turbulent energy spectrum. We associate the wavenumber $k = 2\pi/l$ with length scale l. Again based on dimensional grounds, Kolmogorov found that the energy spectrum should only depend upon k and ϵ, thereby deducing that:

$$E(k) \sim \epsilon^{2/3} k^{-5/3} \tag{2.11}$$

where $E(k)$ is the energy in the turbulent modes of wavelength k. More specifically, $E(k)$ is defined as the Fourier transform (integrated over all directions $\mathbf{k}'$ with $|\mathbf{k}'| = k$) of the correlation function $\langle \mathbf{u}'(\mathbf{x})\mathbf{u}'(\mathbf{x}+\mathbf{r})\rangle$ where $\mathbf{u}'(x,t) = \mathbf{u} - U$ are the velocity fluctuations around average flow rate U. The law (2.11) was verified in experiments [37] and has been a success of theory of turbulence. It is important to point out that in two dimensions, there is no vortex stretching and another cascade, in enstrophy, $\Omega = |\boldsymbol{\omega}|^2$ is possible, in the same direction as (2.11). Its energy spectrum is given by [38, 39]:

$$E(k) \sim \beta^{2/3} k^{-3} \tag{2.12}$$

where β is the constant enstrophy flux. This spectrum was observed in approximately two-dimensional turbulence in thin flowing soap-film experiments [40].

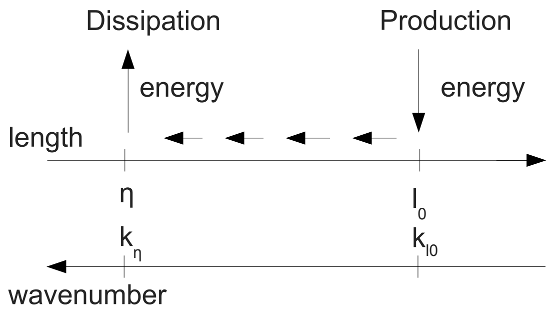


Figure 2.2: Illustration of the energy cascade in turbulence. The turbulence is driven at the length scale l_0, and the energy is transferred to progressively lower length scales (higher wave number). When turbulent energy reaches the Kolmogorov scale η it dissipates through molecular viscosity.

2.3 Directed Percolation

In this Part, we will argue that the transition to turbulence described above is in the directed percolation universality class. Hence, it is important that we first outline some of the relevant aspects of the directed percolation transition, and introduce a few concrete models that undergo this transition.

2.3.1 The model

Directed percolation is a generic non-equilibrium process [41] first introduced by Broadbent and Hammersley in 1957 [42]. Percolation describes the motion of particles (equally walkers or fluids) in a porous medium. Specifically, it describes motion that is primarily affected by the (random) properties of the medium. Whereas ordinary (isotropic) percolation allows particles to spread through the medium in any direction, directed percolation sets a preferred direction for the spreading. Percolation models (and more specifically directed percolation) may describe many phenomena ubiquitous in nature as varied as: forest fires [43], spreads of epidemics in gerbils [44], neural networks [45] and granular systems [46]. For a full review of possible experimental realizations of Directed Percolation, see [47].

We introduce directed percolation through a specific example of the bond percolation process. Consider the diagram in Figure 2.3. The diagram illustrates directed bond percolation on a diagonal lattice of nodes. Each two adjacent nodes in the lattice are connected by a bond that is open with probability p and closed with probability $1-p$. Imagine that a fluid is allowed to flow through the lattice. It is driven by gravity, hence it can only flow down-left or down-right (indicated by direction of arrows), and only through open bonds (indicated by solid lines). Then, the size of a percolating cluster is the number of nodes that are reached by the fluid starting from the nucleation point (labeled with a dark circle). We can take the direction of gravity

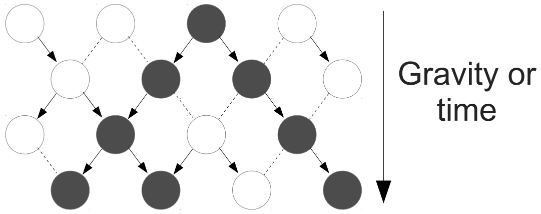


Figure 2.3: Illustration of a bond percolation process in a diagonal 1+1 dimensional diagonal lattice. The arrows indicate open bonds (that occur with probability p) through which the active sites can percolate. Dashed lines indicate bonds (occurring with probability $1-p$) that are impermeable. The active sites are indicated in dark color and are seeded with one active site at the top.

to be the direction of time to obtain the dynamical interpretation of the DP process. In this interpretation, each layer of the medium at height h is taken to be a snapshot of the state of the system at time $t = h$. The dimensionality of DP process is then commonly written $D + 1$, where $D \geq 1$ is the spatial dimension.

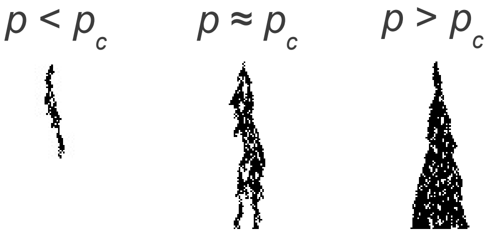


Figure 2.4: Cluster of percolating nodes (black) in diagonal bond percolation shown for 3 different values of p.

If the percolation probability p, the probability that a bond is open, is large then the percolating cluster will be large (on the order of the size of the system, i.e. the number of nodes). If p is small, we expect that the percolating cluster will be small, the percolation process will stop after a while (it will be absorbed), and it will not scale with the size of the system. Hence, there is some threshold p_c above which the fluid will percolate to any depth from the nucleation point. This is illustrated by the diagram in Figure 2.4. This transition (along p) is termed the directed percolation transition. The order parameter is the size of the percolating cluster or the density ρ of active (percolating) sites in the dynamical interpretation. The transition is continuous and is characterized by a number of universal critical exponents. Relating density of active sites ρ (the order parameter) in the stationary limit to the percolation probability p, we have

$$\rho(t \to \infty) \sim (p - p_c)^{\beta}. \tag{2.13}$$

One can define a correlation function for directed percolation, much like one can in equilibrium statistical

Exponent	Description	Dimensionality	Value
β	Density of active sites above p_c	1+1	0.276486(8)
		2+1	0.584(4)
		3+1	0.81(1)
$\nu_\parallel$	Correlation length in direction of time	1+1	1.733847(6)
		2+1	1.295(6)
		3+1	1.105(5)
$\nu_\perp$	Correlation length in direction of space	1+1	1.096854(4)
		2+1	0.734(4)
		3+1	0.581(5)
$\nu_\perp - \nu_\parallel$	Growth rate exponent	1+1	0.63699
		2+1	0.561
		3+1	0.52400

Table 2.1: Relevant universal Directed Percolation exponents.

mechanics. However, since DP is a dynamical process, there is two different correlation lengths, in time (along direction of gravity), $\xi_\parallel$, and in space (perpendicular to direction of gravity), $\xi_\perp$. These lengths also scale with percolation probability via:

$$\xi_\parallel \sim |p - p_c|^{\nu_\parallel}, \qquad \xi_\perp \sim |p - p_c|^{\nu_\perp} \tag{2.14}$$

The three exponents, β, $\nu_\parallel$ and $\nu_\perp$ are sufficient to define all of the other scaling exponents of DP through various scaling relations [41]. The difference $\nu_\perp - \nu_\parallel$ is important as it is the sole exponent that characterizes the growth of one phase into another. Another useful quantity that can be defined is the mean cluster lifetime T in subcritical percolation which has the scaling:

$$T \sim |p - p_c|^{-\tau} \qquad \text{where} \quad \tau = \nu_\parallel - \beta. \tag{2.15}$$

Table 2.1 gives a summary of the relevant exponents we will use in Chapters 3 and 4.

All of the above power laws hold when p is close to p_c. This leads to a practical difficulty in numerically observing these laws, as one needs to know a very precise value of p_c for the lattice in question. The value of p_c is not universal: Table 2.2 gives the values for some relevant systems. There are also other numerical problems associated with running directed percolation simulations. Since one is interested in measuring power laws, one needs to make measurements for very long times (to gather sufficient statistics and decades of scaling) and in large space (to minimize the finite-size effects). In spite of the difficulties, when statistical and finite-size effects are taken care of, measurements of directed percolation are very important because of their universality. They are known to be applicable to a wide variety of models, including site and bond

System	Dimensionality	p_c
Bond diagonal percolation	$1+1$	0.644700185(5) [50]
Site diagonal percolation	$1+1$	0.70548522(4) [50]

Table 2.2: Percolation transition values table for relevant systems.

diagonal percolation and contact processes. Directed percolation's wide applicability led to a conjecture that any system with the following rules is in the DP universality class [48, 49]:

1. There is a fluctuating active phase and a unique absorbing phase.
2. There is a positive, one-component order parameter.
3. Only short-range interactions are allowed.
4. There are no other special symmetries or quenched randomness.

2.3.2 Measurements of DP exponents in Liquid Crystal Experiments

Since we will be interested in observing signatures of directed percolation in real-life experiments, it is important that we describe the world's first measurement of directed percolation exponents in an experiment. This was performed for the first time by Takeuchi *et al.* in 2007 [51, 52]. In this section, we describe the methods that Takeuchi *et al.* used to measure the exponents as a point of comparison to our own measurements in Chapters 3 and 4. The experiment that they considered was a thin quasi-two-dimensional cell of a nematic liquid crystal subject to an external voltage strong enough to trigger a Carr-Helfrich instability. At the critical voltage, the tangled lines of the liquid crystal defects have two possible modes (so-called turbulent modes), labeled DSM (dynamic scattering modes) 1 and 2. The difference between DSM1 and DSM2 is in the density of the topological defects. In DSM1 these are present, but their density is low. In DSM2, these are at a high density, and they continuously elongate and split. Above the critical voltage, the regions of the crystal in DSM2 state can move about in DSM1 background. Since DSM2 regions have a greater density of topological defects, they are more opaque and can be readily imaged through an optical setup. The key measurements that Takeuchi *et al.* took were directly based on observing the time-dynamics of the field-of-view of their microscope. The steady-state density ρ of DSM2 patches in the DSM1 background was found to fit the DP value for 2+1 dimensions. The time and space correlations of DSM2 patches allowed them to measure the correlation length exponents $\xi_\perp$ and $\xi_\parallel$ as well, and these were also found to fit the DP values. Takeuchi *et al.* even found a way to nucleate DSM2 through laser pulses, and this allowed them to measure secondary exponents related to dynamics of the DP clusters nucleated from a single site.

Why did Takeuchi *et al.* succeed in measuring the DP exponents, while others failed to gather enough statistics? They had a large enough system (minimum sizes of regions of DSM2 were small enough that they had an effective resolution of 1650×1650), and the liquid crystal dynamics was fast enough to be able to gather a lot of statistics quickly. Finally, they were able to measure the exponents by directly observing the correlation lengths and density of states. In this sense, they used the turbulent liquid crystal as a sort of an analogue computer, simulating a DP process on a 2-dimensional space. Our own system of measuring the DP exponents in pipe flow given in Chapter 3 is quite direct, much like Takeuchi *et al.*'s. However, in Chapter 4, we consider a more subtle measurement of DP that arises as a consequence of the underlying lattice of the noise field. Not being able to directly observe the DP contact-process behavior is the big problem for our own data gathering efforts and will be described further in Chapter 4.

Chapter 3

Directed percolation describes lifetime and growth of turbulent puffs and slugs

In this chapter, we demonstrate that the phenomenology and quantitative details of many features of the laminar-turbulence transition (described in detail in Section 2.1) are consistent with the non-equilibrium phase transition in the universality class of directed percolation (DP). In Section 3.1 we introduce the background of the problem. In Section 3.2 we describe how we map DP to turbulence in a pipe. In Section 3.3.1, we compare our results on turbulence lifetime to those observed in puff experiments. In Section 3.3.2, we compare our results on turbulence growth rates to those observed in slug experiments, and we observe an interesting crossover from 3+1- to 1+1- dimensional DP. Finally, in Section 3.4, we relate our findings to known results from extreme value statistics and we overview other models of transition to turbulence.

3.1 Introduction

In this chapter, we argue that the transition to turbulence in a pipe is in the universality class of directed percolation. This logical argument was originally made by Pomeau [1] and subsequently by many authors (see, e.g. [53] and references therein). The argument proceeds as follows: The phenomenology of pipe flow turbulence, described in Section 2.1, demonstrates that an external fluctuation, or a boundary condition is necessary to nucleate turbulence from laminar flow. In this sense, laminar flow is an absorbing state that cannot be escaped without an external intervention. Similarly, Section 2.1 describes the conditions under which turbulence can spontaneously relaminarize. Hence, the active state (turbulence) can transition into the absorbing state. Finally, it is possible that laminar flow can be disrupted by a nearby turbulent patch. Eddies can advect from the turbulent patch into the laminar region and "infect" it with turbulence. These three conditions, Pomeau argued, may indicate that transition to turbulence is in DP universality class.

Our work measures the lifetime of active states in DP in a pipe geometry, finding agreement with the superexponential functional dependence measured by Hof *et al.* [2]. We also measure the growth rate of active DP clusters in the supercritical directed percolation and show that our scaling results are in good agreement with available experimental data on the growth rate of turbulent slugs [3]. These results show that dynamical phase transition phenomena in pipe flow turbulence may indeed be described by directed percolation.

3.2 Directed percolation model

The analogy between DP and turbulent-to-laminar transition is the following: Active states in a three dimensional lattice correspond to coarse-grained regions of size $\sim \eta$ where the turbulence intensity exceeds a threshold. Here η is the viscous scale given by (2.10). Inactive states correspond to patches of the fluid which are laminar. The dimension along the percolating direction is associated with time t in the usual interpretation of DP as a dynamical process, as described in Section 2.3.1. The percolating probability p is analogous to Reynolds number Re in the vicinity of the percolation transition, but the mapping need not be linear. For the metastable puffs, Re < 2050 region is mapped to $p < p_c$ whereas for the growing fronts, Re > 2500 is mapped to $p > p_c$. The critical region maps into the spatiotemporal regime, as summarized in Fig. 3.1, but this region and p_c is not strictly defined except in the limit of infinite system size.

We simulate DP in 3+1 dimensions (3D videos of the simulations performed in this chapter are available at http://guava.physics.uiuc.edu/projects/Turbulence/dp.html). The simulations are performed in the reference frame of the traveling puff, which travels slower than the laminar mean flow velocity U. DP can be simulated through various specific models. Consider a three dimensional Euclidean lattice $A^t(x, y, z)$. Here, $A^t(x, y, z) = 1$ (or 0) indicates an active (or inactive) site (x, y, z) at time t. The bond DP is simulated via the algorithm:

```
$A^t(x, y, z) = 0.$
For each nearest-neighbor site of $(x, y, z)$, $A^{t-1}_{nn}$ :
    If $A^{t-1}_{nn} = 1$, then:
        With probability $p$ set $A^t(x, y, z) = 1.$
```

On the other hand, site DP is simulated via:

```
$A^t(x, y, z) = 0.$
If any of nearest-neighbor sites of $(x, y, z)$, $A^{t-1}_{nn} = 1$, then:
    With probability $p$ set $A^t(x, y, z) = 1.$
```

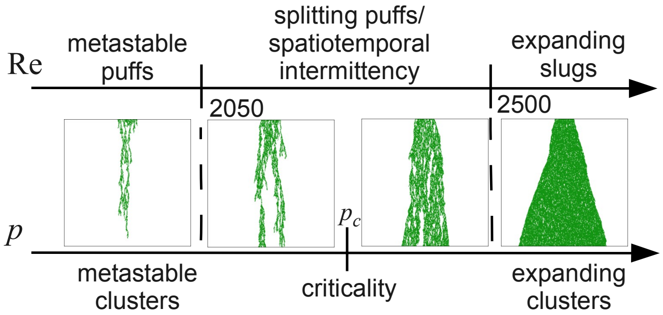


Figure 3.1: Comparison of the phenomenology of transitional turbulence as a function of Re with that of DP in 3+1 dimensions as a function of p, both in a pipe geometry.

Evidently, site percolation must have a higher percolation threshold p_c, since it has lower probability of being activated by a nearest-neighbor site. That's precisely the case, for site percolation $p_c \sim 0.705$, whereas for bond percolation $p_c \sim 0.645$. Here we use the bond percolation process but our results are not different if site percolation or contact process is used (except for moving everything to a new value of p_c). All of these models are in the universality class of DP. The inlet disturbance in pipe flow experiments is modeled as the initial region of active (turbulent) sites. Because the percolation process occurs on a diagonal lattice, the adjacent sites are chosen differently for odd and even time steps. For even time steps, adjacent sites are $(x+1, y, z)$, $(x, y+1, z)$ and $(x, y, z+1)$. For odd time steps, adjacent sites are $(x-1, y, z)$, $(x, y-1, z)$ and $(x, y, z-1)$. All updates are performed sequentially.

3.3 Comparison to experimental data

In this section we match our results in DP to relevant experiments in transition to turbulence in a pipe. First, we will compare the lifetime of active sites in DP to lifetime of turbulent puffs. Then we compare the expansion rate of active sites in DP to the growth rate of turbulent slugs.

3.3.1 Lifetime of turbulent puffs

The survival probability of turbulent puffs in pipe flow is known to be memoryless [54, 55],

$$P(\mathrm{Re}, t) = \exp\left(-\frac{t - t_0}{\tau(\mathrm{Re})}\right), \tag{3.1}$$

where $t > t_0$. Here, the survival probability $P(\mathrm{Re}, t)$ refers to the probability that the turbulent puff still exists after flowing for time t, t_0 is the formation time of the puff and $\tau(\mathrm{Re})$ is the characteristic lifetime. In

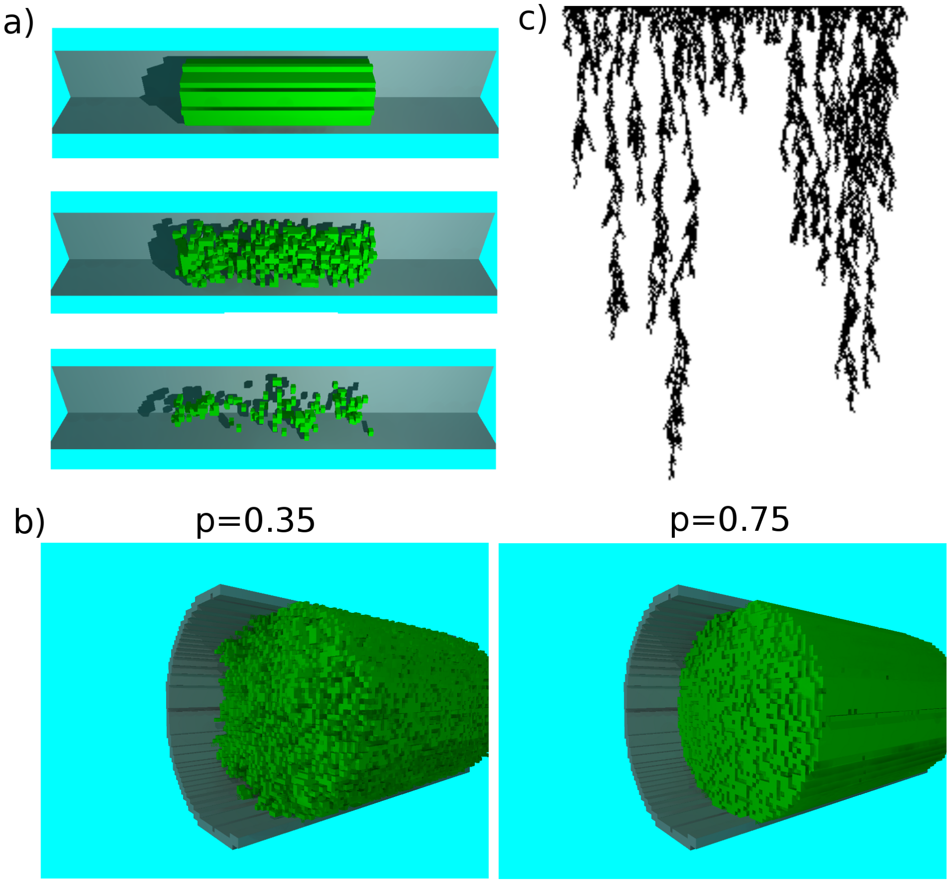


Figure 3.2: (a) Illustration of the decay of an active cluster in subcritical directed percolation in 3+1 dimensions in pipe geometry. In this figure the length of the pipe extends horizontally, and the active sites in percolation are displayed with green cubes. Inactive sites are not shown. (b) Different front shapes in the growing front bond DP model. Rough fronts occur for small $p - p_c$ whereas smoother fronts occur for large $p - p_c$. (c) Time evolution of an initially active 180 contiguous sites percolating downwards simulated via bond percolation for $p = 0.61 < p_c$. The active state (in black) fully decays into the absorbing state (white) after about 250 steps.

Hof *et al.*'s work t_0 was a constant ($70\ D/U$ where U is the mean flow velocity and D is the pipe diameter) [2]. By measuring $P(\text{Re}, t)$ for specific times t and Reynolds numbers Re, Hof *et al.* calculated $\tau(\text{Re})$ from (3.1). They discovered that τ scales superexponentially with Reynolds number [2], fitting well a parameterization of the form:

$$\tau(\text{Re}) = \tau_0 \exp[\exp(c_1\text{Re} + c_2)], \tag{3.2}$$

where $\tau_0 \sim D/U$.

The survival probability $P(p, t)$ in our DP measurements is the probability that there will be active sites left in the lattice after t DP steps. From $P(p, t)$, the lifetime of the disturbance τ can be measured just as in Hof *et al.*. This idea is illustrated in the snapshots of the simulation in Fig. 3.2(a). Here, DP is simulated in 3+1 dimensions in a pipe of radius of 5 lattice sites. In this simulation p is less than p_c, and so the puff eventually decays. One can measure the lifetime with a 3-dimensional lattice where two of the spatial dimensions span a disk of radius of R lattice points (corresponding to the pipe radius), with fixed boundary conditions. However, the measurement of lifetime in this way over many orders of magnitude is made difficult because of the system size. The lifetime measurements must be repeated many times to be able to obtain sufficient statistics. However, when p is close to the critical percolation threshold p_c, the correlation length $\xi_\perp \sim (p - p_c)^{\nu_\perp}$ along a space dimension becomes larger than R, and the nominally 3+1 dimensional DP is effectively 1+1 dimensional. Thus, to get sufficient statistics we simulate DP in a 1-dimensional lattice of length N that is initially made to be active in a subregion of length N_0. Below the DP critical point the active states will eventually decay into the absorbing state. In a finite-sized system, the decay can always occur, but we find that the characteristic lifetime τ of the decay grows super-exponentially with percolation threshold p, and beyond a certain percolation probability, the average lifetime of the active state is too large to be measurable on a computer.

Hof *et al.* were able to calculate τ via (3.1) by measuring $P(\text{Re}, t)$. Even though they could only extend t to $3450D/U$, they were able to resolve P to 100 ppm, giving them effective measurements of τ over 8 orders of magnitude. In the case of directed percolation, one cannot use this procedure, since t_0 is not constant, but instead depends on the percolation probability p. In directed percolation, t_0 is the time over which the initial state is remembered by the system. Hence, for each value p we must measure the survival lifetimes of many instantiations of directed percolation. The cumulative distribution function (CDF) of these survival lifetimes then approximates the survival probability $P(p, t)$, as long as sufficiently many instantiations have been performed. From the fit of the form (3.1) to the CDF data, one can read off $\tau(p)$ and $t_0(p)$. Our measurements of $\tau(p)$ for a lattice of size $N = 100$ and $N_0 = 20$ are given in Fig. 3.3(a). The line in the Figure is obtained by fitting τ to (3.2). The inset shows the linear fit to $\log_{10} \log_{10} \tau(p)/\tau_0$. Sufficiently far away from the critical point p_c, we find that the linear fit deviates, indicating that the superexponential behavior may somehow be related to the diverging correlation lengths at the critical point. Interestingly, the power-law scaling for average lifetime of a cluster seeded from a single active site (Equation 2.15) does not fit the data in Fig. 3.3(a) (it would appear as a horizontal line on the clearly upward sloping curve). This indicates that the super-exponential behavior of the lifetime is due to the system being with multiple degrees of freedom. This point is further motivated in Section 3.4.

From the CDF data, we can evaluate the survival probability functions analogously to Figure 2 of [2]. Figure 3.3(b) shows our numerical data $P(p,t)$ for 4 different times (S curves), along with the model $P(p,t) = \exp(-(t-t_0)/\tau(p))$. To evaluate the model fit, we use the value of $t_0(p_{0.2})$ where $p_{0.2}$ is found by finding where $P(p_{0.2},t) = 0.2$. Note that the fact that the S curves become steeper with p is a characteristic of the superexponential scaling of $\tau(p)$.

The numerical data presented so far in this chapter has been measured in a finite volume of size $N = 100$. Finite size effects in DP have been investigated thoroughly in the literature [56]. We ran our simulations in a volume bound only by the range of the integers on our computer (0 to $2^{63}-1$), and we didn't find any qualitative differences regarding the superexponential scaling of $\tau(p)$.

3.3.2 Growth rate of turbulent slugs

When $p > p_c$ active DP clusters grow in the pipe. We measured their growth rate and related it to the growth rate of turbulent slugs. The speed at which the front of the percolating clusters propagates into the neighboring inactive region is given by $G \sim \xi_\perp/\xi_\parallel \sim (p-p_c)^{\nu_\parallel - \nu_\perp}$, where $\xi \sim (p-p_c)^{-\nu}$ is the correlation length in direction of space (denoted by $\perp$) or in direction of time (denoted by $\parallel$). Using the above prescription and numerical values of DP critical exponents [56] one should expect that $G \sim (p-p_c)^\gamma$ where $\gamma = 0.524$ in 3+1 dimensional DP, whereas $\gamma = 0.637$ in 1+1 dimensions. These power laws are close to the exponent 0.5 first proposed in 1986 [3], as well as in modern experiments [32]. However, the data is not sufficient yet to differentiate between two such close power law exponents.

Measurements of the growth rate G of an initially active region in 3+1 DP in a pipe geometry are shown in Fig. 3.4. The measurements were made by simulating bond DP with $p > p_c$ and measuring the positions of the two fronts as functions of time. During the numerical simulation we also measure the correlation length $\xi_\perp$ by calculating the root-mean-square height (i.e. roughness) of the turbulent-laminar interface. The agreement with theoretical expectation is good, and we see numerical evidence for the crossover from 3+1 to 1+1 dimensions that arises as result of interplay of pipe geometry and correlation lengths. When $(p-p_c)/p_c \ll 1$, then $\xi_\perp > D$, where D is the pipe diameter, and the system is effectively 1+1 dimensional.

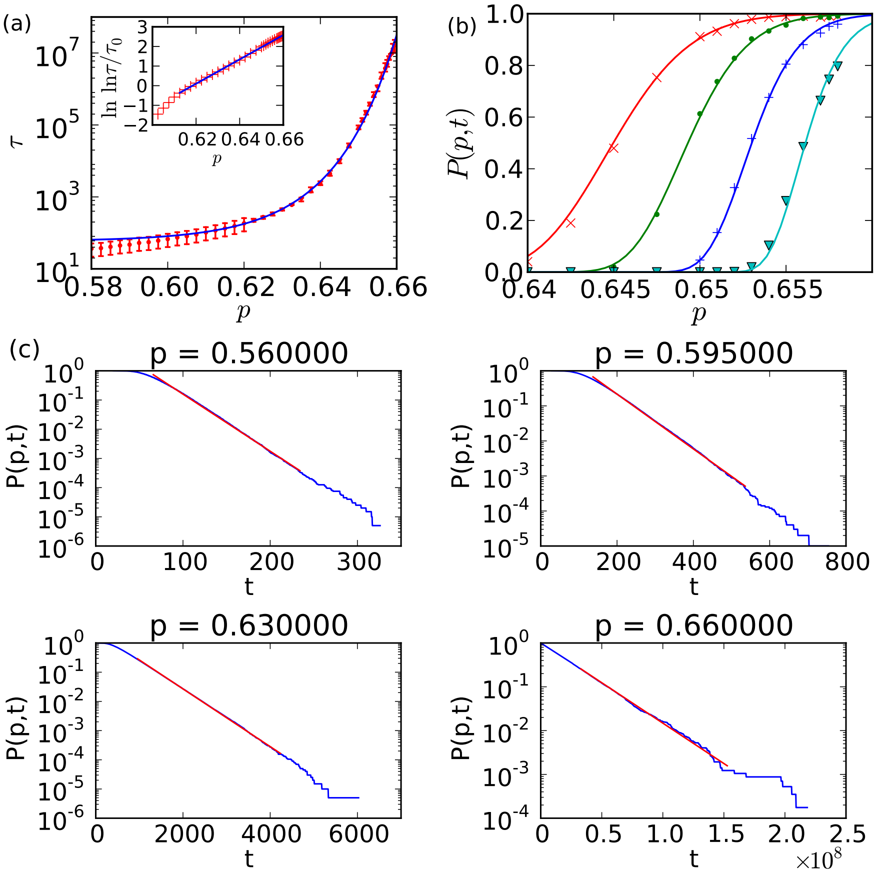


Figure 3.3: (a) Superexponential scaling of the characteristic lifetime τ. The line indicates the fit to (3.2). Error bars indicate 95% confidence intervals from Kolmogorov-Smirnov test (for $p > 0.62$) and 90% confidence intervals from χ^2 test (for $p \leq 0.62$). In the inset, $\tau_0 = 0.017$. (b) Numerical data for survival probabilities $P(p,t)$ (points) and a fit to (3.1) with τ given by (3.2) (solid lines). Data shown for $t = 300$ (red crosses), $t = 1000$ (green dots), $t = 6000$ (blue pluses) and $t = 80000$ (cyan triangles). The value of p_c observed corresponds to that of 1+1 dimensional bond percolation. (c) Measured survival probabilities as functions of t for 4 different values of percolation probability p. Blue line indicates measured data, whereas red line indicates a fit to the exponential distribution. Deviations from exponential distribution for small t are due to nonzero t_0.

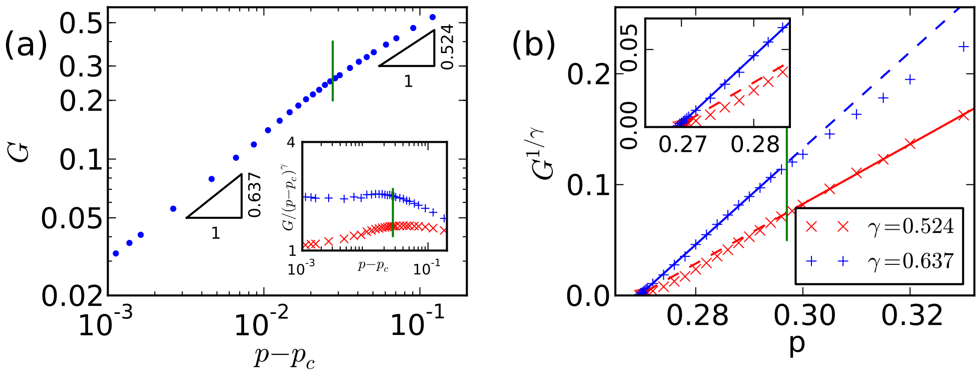


Figure 3.4: Measured values of front propagation velocity G (red crosses) compared to theoretical prediction for 3+1 dimensional and 1+1 dimensional DP. Green vertical line indicates the value of p for which $\xi_\perp$ exceeds $1.2R$. We call this value p_R. (a) Plot indicating the power law crossover of $G(p)$. Inset shows that $G/(p-p_c)^{0.637}$ is roughly constant for $p < p_R$ and similarly $G/(p-p_c)^{0.524}$ for $p > p_R$. (b) Both regimes of DP have the same critical point p_c (within the error of our measurements). The linear fits to $G^{1/\gamma}$ are shown in solid lines. Extrapolations are indicated with dashed lines, and they cross at the same p_c. This value of p_c is corresponds to that of 3+1 dimensional bond percolation.

Otherwise $\xi_\perp < D$, and the system is 3+1 dimensional. This argument is illustrated in a diagram in Figure 3.5.

In Fig. 3.4(b), we have plotted $G^{1/\gamma}$ versus p for different choices of γ corresponding to 1+1 and 3+1 dimensional DP. We see clearly the crossover between the expected regimes. Note that in this plot we did not need to guess p_c: both scaling regimes yield the same p_c. It is difficult however to extend the data for $G(p)$ close to p_c. Due to the finite size of the system, when $p-p_c$ is small, the active regions split and may decay into the absorbing state. This makes it difficult to clearly measure front propagation velocity. On the other hand, when $p-p_c$ is large, the scaling breaks down. Thus we expect the power law exponent of 0.524 to be observable only in an intermediate regime of $p-p_c$, sufficiently close to p_c but still such that $\xi < R$.

One may be motivated to believe that the interesting crossover phenomenon described above is the reason for the transition from puffs to slugs. In Figure 3.6, we plot the crossover of the density of active sites in the subset of the pipe bounded by active sites in 3+1-dimensional DP. We find numerical evidence for the crossover like for the case of $G(p)$. However, we also notice that the density is continuous at the point of the crossover. Hence, the crossover does not indicate a transition point for puffs (intermittent turbulence) and slugs (continuous, saturated turbulence density).

One other aspect of the phenomenology of pipe flow is captured by the DP model, namely that the fronts of active regions with $p - p_c \ll 1$ are much rougher than when $p-p_c$ is large. This is because the density of active states within the region is an increasing function of p. Furthermore, the width of the front is related to the spatial correlation length $\xi_\perp$ which becomes small when $p-p_c$ is large. Finally, the front

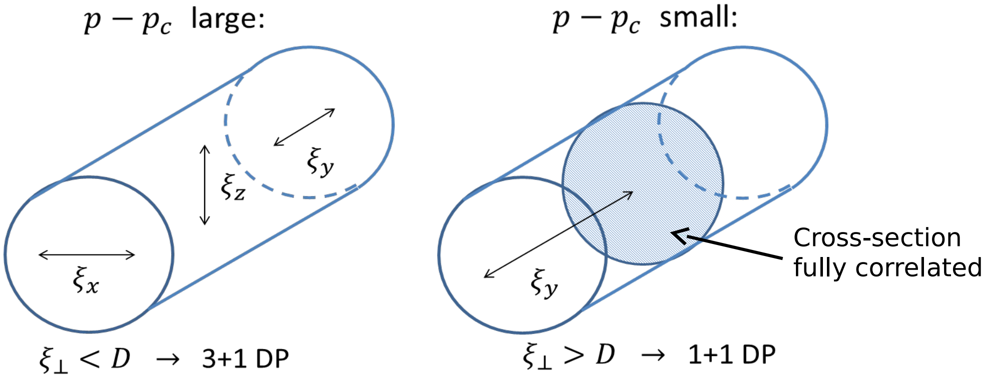


Figure 3.5: Dimensional crossover that arises as a result of pipe geometry and correlation lengths. *Left:* When correlation lengths are small and $p - p_c$ is large, then the system is effectively $3 + 1$ dimensional. *Right:* When $p - p_c$ is small, the correlation lengths become as large as the diameter of the pipe D. In that case, the cross-section of the pipe is effectively fully correlated and the DP process can proceed only along the axis of the pipe. Hence, the system becomes effectively $1 + 1$ dimensional.

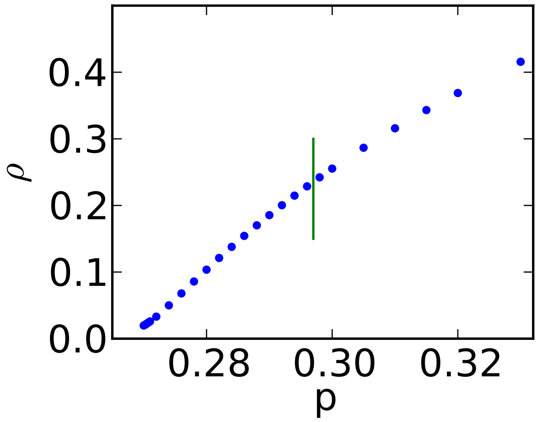


Figure 3.6: The density of sites ρ within an expanding cluster, shown as function of percolation probability p. Green vertical separator indicates the point at which the correlation length becomes on the order of the size of the pipe diameter. Notice that there is no discernible discontinuity in density ρ and hence there is no clear change from the regime of splitting puffs and that of growing slugs.

propagation velocity has an upper bound of 1 lattice space per unit time, which happens when p = 1, and this bound leads to a smoother interface. The difference between the rough and smooth front regimes is shown in Fig. 3.2(b). This is analogous to the results in pipe flow experiments hot wire measurements, where puff structures were found to have rough edges whereas slugs have clearly defined fronts [14, 57].

3.4 Conclusion

The DP simulations of characteristic lifetime presented in this chapter have been performed via the bond percolation algorithm. However, we found superexponential scaling of $\tau(p)$ for site percolation too. In fact, bond and site percolation are both special cases of the more general Domany-Kinzel (DK) model [58]. Compact directed percolation is also a special case of the DK model, but we find that $\tau(p)$ for compact DP scales as power law instead. A cluster at time t in 1+1 dimensional compact DP can be characterized by only a single degree of freedom (cluster width). Our interpretation is that the superexponential behavior doesn't occur because one needs many degrees of freedom for extremal value statistics to apply (see below).

The superexponential scaling of the lifetime is likely a universal characteristic of the directed percolation process. Goldenfeld *et al.* proposed that this superexponential character of the turbulent puff lifetime can be described by extreme value statistics [59], because puff decay occurs when turbulent energy fails to attain the required threshold at all points in the puff [60]. In the usual central limit theorem, under appropriate conditions [61] the distribution of a sum of random variables $\bar{X} = \sum_{i=1}^{N} X_i$ tends to a Gaussian distribution for large N. However, a maximum (or minimum) of N random variables $max(x_1, \ldots, x_N)$ is instead superexponentially distributed with 3 universality classes [59], selected by the underlying probability distribution of $\{x_i\}$. In Fig. 3.2(c) we show a time evolution of an initially active cluster percolating with $p < p_c$. The lifetime of the entire cluster is the lifetime of the longest active "strand" percolating downwards. Assuming that strand lifetimes are independent and identically (exponentially) distributed, then the lifetime of the longest strand is given by the type I Fisher-Tippett distribution $\exp(-\exp(-p))$. This argument has also been used to explain the superexponential distribution of size of largest connected cluster in ordinary (isotropic) percolation [62], and it was found there that correlations between cluster sizes (analogous to strand lifetimes) do not influence the superexponential scaling. The relationship between the percolation problem and extreme statistics was also investigated in [63]. It seems that a large number of degrees of freedom is important for obtaining the superexponential scaling of lifetime, which is why we do not see it in compact DP.

The DP model we proposed in this paper can account for the superexponential lifetime of the turbulent puffs, as well as the uniform growth rate of turbulent slugs. As shown in Fig. 3.1, the transition between these two regimes ($2050 < \mathrm{Re} < 2500$) occurs through the splitting and interactions of puffs. The spatio-temporal patterns of coarse-grained turbulent intensity obtained from a direct numerical simulation [29] bear similarities to those of directed percolation, but the data are not adequate to make a quantitative analysis.

3.4.1 Other Models for Transition to Turbulence in a Pipe

Following our publication, Allhof and Eckhardt [64] also studied DP in the context of the transition to turbulence. They studied a slightly different model of DP on an Euclidean lattice with different probabilities for sustaining an active site, and infecting nearby sites. However, their results show no marked difference from regular DP, thereby enforcing the universality of DP. In particular, their measured growth rate matches our results $G \sim \sqrt{Re - Re_c}$ well.

At about the same time that we performed our analysis described in this chapter, a more complicated, dynamical system model was proposed by Barkley [65]. In this section we briefly compare and contrast our model to Barkley's model. Barkley's motivation was to write down a model for turbulence in terms of two fields, turbulence intensity q and local flow rate u. Here, f_i^n indicates value of the field f at lattice point i and time step n. Barkley's equations of motion are:

$$q_i^{n+1} = F[q_i^n + d(q_{i-1}^n - 2q_i^n + q_{i+1}^n), \mathrm{Re}, u_i^n], \tag{3.3}$$

$$u_i^{n+1} = u_i^n + \epsilon_1(1 - u_i^n) - \epsilon_2 u_i^n q_i^n - c(u_i^n - u_{i-1}^n). \tag{3.4}$$

where F indicates a function given by two iterates of the chaotic tent map (not shown here) parametrized by further 3 parameters α, β, γ as well as Re. d indicates the rate of turbulence relaminarization, ϵ_1 the rate by which fluid velocity restores to mean flow velocity $U = 1$, ϵ_2 controls the reduction of flow velocity due to turbulence and c breaks left-right symmetry and controls advection of the flow field u. Evidently, Barkley's model is much more detailed, having 7 additional parameters, but it is able to model phenomena that our DP model is not able to. In our model, it is difficult (due to coarse-graining) to define puff boundaries and puff splitting events since, in general, any active cluster in directed percolation will be scattered with inactive sites inside it. Barkley's u field allows one to define a turbulent puff via the velocity profile. Puffs and slugs in Barkley's work have velocity profiles in character with hot wire measurements [57], and Barkley finds that puffs can annihilate if they are too close to each other [30]. He also finds the superexponential behavior of turbulence lifetime but also puff splitting rates (which we cannot measure). However, it is encouraging to note that his model seems also to be in the DP universality class. In fact, the turbulence fraction f

above $\mathrm{Re} = 2040$ (proportion of q field that in the chaotic portion of the tent map) has the functional form $f \sim (\mathrm{Re} - \mathrm{Re}_c)^{\beta}$ where β is the standard DP exponent for density of active sites. Therefore, this serves as another confirmation that the transition to turbulence in shear flows is in the universality class of DP.

Chapter 4

Directed Percolation Transition in Burgers Equation

4.1 Introduction

In Chapter 3, we considered a phenomenological model for the transition to turbulence. We motivated the idea that laminar-turbulent transitions are in the directed percolation universality class by taking a coarse-grained statistical mechanics perspective of the fluid velocity field. Statistical mechanics, and associated notions of phase transition universality, is of course well-understood in equilibrium. However, the transition to turbulence is a transition between two non-equilibrium steady states. Turbulent flows are usually considered from the dynamical systems perspective of partial differential equations (i.e. the Navier-Stokes equations), and it is something of an open question how stochastic flow phenomena arise from this type of description. The literature on the transition to turbulence has also been focused primarily on the idea that the transition to turbulence can be understood from the interaction of a small number of spatially-localized modes, leading to low-dimensional chaotic dynamics (see [66] and [25] and references therein). Such a description does not fully account for the detailed space-time phenomenology of turbulent pipe flow in the range $1500 < Re < 2500$, in particular, leaving out the details of how turbulence spreads, puffs split, and even a quantitative account of the lifetime statistics has not been given.

In order to make closer contact between a fluid description in terms of partial differential equations for a continuum flow field and Pomeau's proposal that the transition to turbulence can be in the universality class of a non-equilibrium statistical mechanical model, this chapter addresses explicitly the way in which a fluid velocity field equation can be mapped into statistical mechanics. We begin by considering the Navier-Stokes equation in Section 4.2, and its scaling of pressure under the directed percolation transition. We then spend the next three sections showing how one can map Burgers equation to a problem of directed polymers on a lattice. In this model, one can observe a directed-percolation transition as function of the structure of the noise field (this will be made specific later in the chapter). We then show in Section 4.6 how one can map the results on this transition back to the language of Burgers equation.

Burgers equation was originally proposed as a simplified model for the Navier-Stokes equation [67], and it arose in attempts to write down a statistical theory of turbulence. The essential feature of the turbulent spectrum arises from the interplay between external forcing, inertia and dissipation due to molecular viscosity. In order to distill the problem to its core, Burgers considered a simplest possible one-dimensional equation with diffusion and self-advection. The equation has no way to generate shear or vortex motion, and no way to generate new randomness from initial data – it can be integrated directly [68]. As such, Burgers abandoned considering the equation as a model for Turbulence. He writes that "the investigations must therefore be taken on their own merit, as a study of peculiar solutions of a primitive nonlinear problem" [69]. Despite Burgers' modesty, his equation has been widely used as a paradigm for how the generation of multi-scale cascades and the existence of turbulent phenomena of some sort can arise in physics [68, 70, 71, 72, 73]. In this chapter we will show how Burgers equation provides an explicit construction of the way in which a continuum field theory for a dynamical velocity field can be mapped into a non-equilibrium statistical model, and that the dynamical transition in the flow field is described by a non-equilibrium phase transition in the statistical model, which we show to be directed percolation.

4.1.1 Role of pressure

The pressure term ∇P in the incompressible Navier-Stokes equation enforces the incompressibility condition $\nabla \cdot \mathbf{u} = 0$, leading to creation of vorticity. The process happens in the following way: The advection and the viscosity terms lead to the production of divergence, which is negated, through the pressure term, producing non-zero curl of velocity (vorticity). Hence, pressure is important as it produces the vortices characteristic of turbulence and significant for the turbulent energy cascade. However, the implicit definition of the pressure field also makes the mathematical analysis difficult. The Burgers equation simply dispenses with the pressure term and hence the incompressibility condition, yielding:

$$\partial_t \mathbf{u} = -(\mathbf{u} \cdot \nabla)\mathbf{u} + \nu \nabla^2 \mathbf{u} + \xi, \tag{4.1}$$

where $\mathbf{u}$ is the fluid velocity, $(\mathbf{u} \cdot \nabla)\mathbf{u}$ is the fluid self-advection, $\nu \nabla^2 \mathbf{u}$ is the fluid diffusion due to viscosity ν and ξ is the external (usually random) force.

Recall from Section 2.2.1 that the fluid vorticity $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ in Navier-Stokes equation satisfies

$$\partial_t \boldsymbol{\omega} + (\mathbf{u} \cdot \nabla)\boldsymbol{\omega} = \nu \nabla^2 \boldsymbol{\omega} + \boldsymbol{\omega} \cdot \nabla \mathbf{u} + \nabla \times \xi. \tag{4.2}$$

Notice that (2.8) is missing the pressure term. However, it is an important point that the condition

$$\nabla \cdot \mathbf{u} = 0. \tag{4.3}$$

still holds. If one attempts to solve Navier-Stokes equation via (4.2), one evolves $\boldsymbol{\omega}$, which in turn defines the $\mathbf{u}$ field, which must satisfy (4.3). This is in contrast to the Burgers equation which lacks the pressure term but looks exactly like (4.2) in vorticity formulation. Yet, Burgers equation doesn't have a way of generating $\boldsymbol{\omega}$ outside of the $\nabla \times \xi$ term. Vorticity advects and diffuses like a passive scalar in Burgers turbulence.

One should note that the lack of pressure in Burgers equation is not the same as the compressible Navier-Stokes equation. Instead, it is the Navier-Stokes equation with no condition whatsoever on the divergence of $\mathbf{u}$. The pressure is simply removed from the equation. Fast-moving elements of the fluid can advect to "overtake" the slow-moving elements. There is no equation of state for fluid density and pressure to stop this from happening. In this way, the fast-moving fluid elements lead to shock fronts, a well known characteristic of the Burger equation. For this reason, Burgers equation is a useful toy model for shocks in sets of equations describing hyperbolic conservation laws, e.g. in gas dynamics. When the $\mathbf{u}$ field is plotted as function of position, in 1 dimension, these shocks look like trains of sawteeth (see Figure 4.1). Burgers equation and the shock front solution have since been used in a wide variety of problems, from statistical physics (through Burgers' equation's similarity to KPZ [73] and directed polymers [74]), and cosmology [68] (modeling the formation of large-scale structures in the Universe from primordial fluctuations), and even traffic flow [75]. It has also found its place in fluid dynamics mainly as a testbed of numerical methods, closure formulations and other mathematical methods [68].

Since the Fourier transform of a sawtooth function is $1/k$, the energy spectrum $E(k) = |v(k)|^2$ scales as k^{-2}. This range of scale-free energy spectrum [76] is a well known property of Burgers' equation, and it is similar to the scale-free energy spectrum of Navier-Stokes turbulence. One can even get the $k^{-5/3}$ Kolmogorov spectrum in randomly driven Burgers turbulence with a specific choice for the correlations of the random driving force [77]. Because of this, the solutions to the randomly forced Burgers equation, or Burgers equation with random initial conditions have been termed Burgers' turbulence, or Burgulence [68].

4.2 The relevance of pressure

In this section, we consider the relevance of pressure in the Navier-Stokes equations at the critical point of transition from laminar to turbulent flow. Here, we are interested in the scaling of the pressure term at the directed percolation critical point. If we can find evidence that the pressure is irrelevant then our results

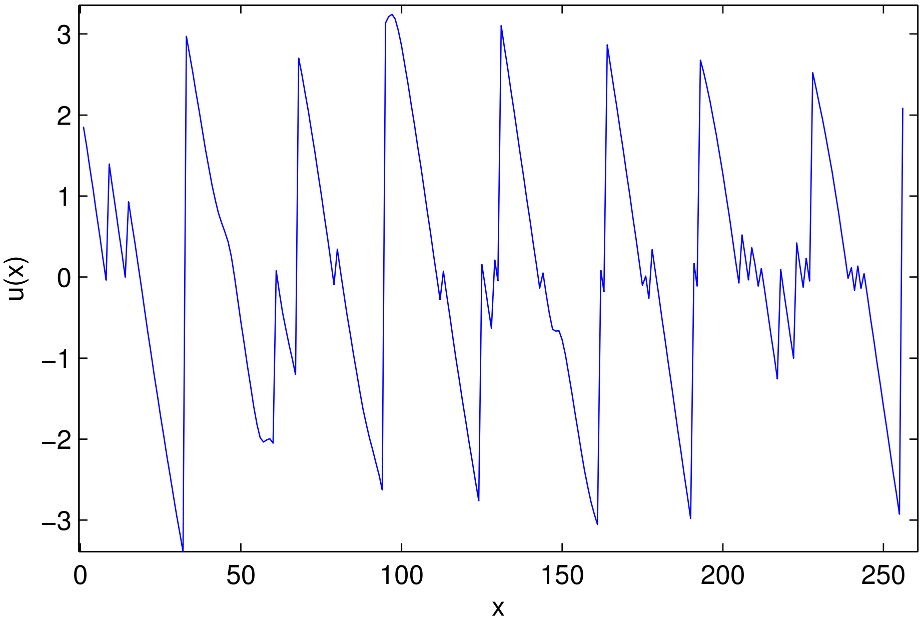


Figure 4.1: Train of sawteeth shocks in one-dimensional Burgers equation. The simulation was performed using a first-order finite-difference scheme for (4.1) with the added condition for upwind advection. The initial condition was a superposition of two sine waves with random Gaussian noise.

can be more directly mapped on the Navier-Stokes equations, and hence to real transition to turbulence.

Renormalization group techniques have been applied to Navier-Stokes equations and turbulence. Well known work is that by Forster *et al.* [71] and Yakhot and Orszag [78]. For the case of relevance of pressure the actual calculation is very difficult to do, so here we only give a scaling argument. However, we will see below that the scaling argument has some flaws. Therefore, this section should be taken as a short summary of some of the approaches and problems we have encountered. We are working on resolving these issues through a full-blown renormalization group calculation.

A scaling analysis such as the one shown here was performed by Medina *et al.* [72] for the case of the scaling of nonlinear term in the Kardar-Parisi-Zhang (KPZ) equation [73]

$$\partial_t h = \frac{\lambda}{2}(\nabla h)^2 + \nu\nabla^2 h + \eta, \tag{4.4}$$

where η has the correlation

$$\langle \eta(\mathbf{x},t)\eta(\mathbf{x}',t')\rangle = |\mathbf{x}-\mathbf{x}'|^{2\rho-d-1}|t-t'|^{2\theta-1}. \tag{4.5}$$

Medina *et al.* applied a change of scale $\mathbf{x} \to b\mathbf{x}$, $t \to b^z t$ and $h \to b^\chi h$ to (4.4) with $\lambda = 0$. The result is

$$b^{\chi - z}\partial_t h = \nu b^{\chi-2}\nabla^2 h + b^{(\rho - d/2+1/2)+(\theta-1/2)z}\eta. \tag{4.6}$$

Requiring that (4.6) should be scale-invariant, leads to

$$z = 2 \equiv z_0 \tag{4.7}$$

and

$$\chi = \rho + 2\theta + (3-d)/2 \equiv \chi_0. \tag{4.8}$$

If the nonlinear term is turned back on ($\lambda \neq 0$), then its scaling becomes $b^{z+\chi-2}$. Letting $z + \chi - 2 > 0$ and setting $z = z_0$ and $\chi = \chi_0$, gives the critical dimension $d_c = 3 + 2\rho + 4\theta$. If $d < d_c$, then the nonlinearity is relevant. When $d > d_c$ the nonlinearity in (4.4) is weak, and at large $\mathbf{x}$ and t the solution has Gaussian statistics.

Can one adapt the procedure used by Medina *et al.* to consider the scaling of the pressure term in Navier-Stokes equations? Consider a naive dimensional scaling of the form

$$\mathbf{x} \to b\mathbf{x} \qquad t \to b^z t \qquad \mathbf{u} \to b^\chi \mathbf{u} \tag{4.9}$$

in the Navier-Stokes equation

$$\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = \nu\nabla^2\mathbf{u} - \frac{1}{\rho}\nabla P \tag{4.10}$$

where χ is an unknown to be determined. Since we are interested in relevance of parameter $1/\rho$, we first set it to zero, and then rescale the equation yielding:

$$b^{-z+\chi}\partial_t \mathbf{u} + b^{2\chi-1}(\mathbf{u}\cdot\nabla)\mathbf{u} = b^{-2+\chi}\nu\nabla^2\mathbf{u}. \tag{4.11}$$

Here, two problems arise. One has to do with Galilean invariance. If we require the total derivative term to be scale invariant, we get

$$\chi = 1 - z. \tag{4.12}$$

But this yields just a result of dimensional analysis: velocity scales as length over time. Furthermore, if we require scale invariance for the diffusion term $\nu\nabla^2\mathbf{u}$, we get

$$z = 2 \tag{4.13}$$

just like Medina *et al.* did. However, the problem with this result is that (4.13) is inconsistent with directed percolation where we expect $z = 0.524$ to 0.637 (See Table 2.1). Equation 4.13 instead implies a Gaussian fixed point, which is the appropriate description of the non-equilibrium steady state of this model but is not a description of the phase transition we are interested in.

If we ignore the requirement that the critical point should correspond to the DP scaling, then we can use a result worked out by Kardar, Parisi and Zhang [73]. They find that in the case of the KPZ equation, the nonlinearity is irrelevant in 3 and higher dimensions. The $1/\rho\nabla^2 P$ term has the same dimensions as the nonlinearity, and if we require that it should scale in the same way, then it is also irrelevant in 3 and higher dimensions.

Neither of the arguments presented in this section is without flaw. To further complicate matters, in this chapter we will solve the Burgers equation in 1 dimension where pressure and Navier-Stokes equations have no physical meaning. As such, the definite conclusion on the relevance of the pressure term and the aptness of Burgers equation model for the transition to turbulence is left to further study.

4.3 From Burgers equation to Directed Polymers

Through a sequence of transformations, the Burgers equation can be mapped to the directed polymers model [70]. We will consider the case of a 1-dimensional Burgers equation here. One can use a change of variables

$$\mathbf{u} = -\nabla h, \tag{4.14}$$

to map the equation (4.1) to the KPZ equation

$$\partial_t h = \frac{1}{2}(\nabla h)^2 + \nu\nabla^2 h + \phi, \tag{4.15}$$

where $\xi = -\nabla\phi$. The transformation can be most easily seen by taking the derivative of (4.15) and using the product rule on the $1/2(\nabla h)^2$ term, thus obtaining (4.1). The KPZ equation is an embodiment of an important front-propagation universality class . The KPZ equation can further be transformed into a linear differential equation by using the Hopf-Cole transformation [79, 80]:

$$h = 2\nu \ln Z \tag{4.16}$$

When one substitutes (4.16) into (4.15), one finds that Z satisfies the equation

$$\partial_t Z = \nu \nabla^2 Z - \frac{1}{2\nu} \phi Z. \tag{4.17}$$

The equation (4.17) is an imaginary-time Schrödinger equation, i.e. a diffusion equation with a linear coupling to the disorder. It is the equation for a partition function of a directed polymer, embedded in a disordered potential $\frac{1}{2\nu}\phi$.

$Z(x, t)$ can be seen as the sum of all paths of the directed polymer that terminate at point x at time t. We let the probability distribution function of the position of the polymer at time $t = 0$ be $\rho(x)$. Then the probability that a polymer is in the region $(x, x + dx)$ is $\rho(x)dx$. The solution to (4.17) can be written as a sum of all paths weighted by the Boltzmann weight for each path (a path integral):

$$Z(x, t) = \int \rho(x_0) dx_0 \int_{x'(t=0)=x_0}^{x'(t)=x} Dx' \exp(-\mathbf{H}[x']). \tag{4.18}$$

Here, $\mathbf{H}$ is the Hamiltonian for the directed polymer which consists of 2 parts, an entropic contribution and a contribution from the cost of the disorder along the path of the polymer:

$$\mathbf{H}(x) = \int_0^t d\tau \left[\frac{T}{4\nu} \left(\frac{dx(\tau)}{d\tau} \right)^2 + \phi(x(\tau), \tau) \right] \tag{4.19}$$

where T is the temperature and $x(t)$ is the path of the polymer.

4.4 Directed Polymers

What is the object described by the Hamiltonian in (4.19)? It is a directed polymer, a polymer that is always extended along the time direction. Along the time direction, it is not allowed to loop back on itself: $x(t)$ in (4.19) is a single-valued function.

The properties of directed polymers were first studied in detail in 1985 [74] and especially after their relation to the KPZ and Burgers equations became evident [81]. An important measure is the wandering exponent, namely the scaling of the position of the polymer as function of time,

$$\langle [x(t+\tau) - x(t)]^2 \rangle \sim \tau^{2\zeta} \tag{4.20}$$

where $\zeta = 2/3$, a clearly non-diffusive result. This result was first derived for uncorrelated noise field ϕ at $T = 0$ [74] in the context of domain walls (with no "overhangs") in the Ising model. However, the result was proven to hold for $T > 0$ for the case of uncorrelated noise [82, 81]. This is an important point since later in this chapter we will conjecture that a certain related result known for $T = 0$ also holds at $T > 0$.

It is known that the result in (4.20) also depends on dimensionality [83]. The disorder term in the polymer Hamiltonian (4.19) is known to be relevant for directed polymers in $1 + 1$ dimensions. The disorder term was found marginal in $2 + 1$ dimension and irrelevant in $3 + 1$ and higher dimensions. These perturbative results imply that in $1 + 1$ and $2 + 1$ dimensions, the system is always in the pinned phase characterized by the value of $\zeta = 2/3$ in (4.20) (in the limit $t \to \infty$), regardless of the value of T. However, as the number of dimensions increases, there are more available regions with weak noise for the polymer to expand to at high temperatures. The polymer can expand randomly to any of these regions and hence ζ becomes the diffusive $1/2$. This is termed the free phase. But, at sufficiently small T or strong noise ϕ in $3+1$ dimensions or more, the polymer can still be in the pinned phase. To summarize, in $1 + 1$ and $2 + 1$ dimensions the polymer is always in the pinned phase, whereas in more dimensions, there is a transition at a non-zero T.

We started studying directed polymers since the pinning-depinning transition is reminiscent of the laminar-to-turbulent transition in fluid dynamics. Here temperatures (or viscosity ν) take the place of Re in the laminar-to-turbulent transition. A notable difference with turbulence is in $2 + 1$ dimensions, where there is no pinning-depinning transition. However, in fluid dynamics it is known that 2-dimensional turbulence is very-well possible and it is characterized by the unique direct enstrophy cascade.

There is another important difference between fluid dynamics and directed polymer transition. The transition in fluid dynamics is spatially extended and subcritical. Even at high Re, there can be regions that are laminar, and the turbulent-laminar fronts can move as function of $\mathrm{Re} - \mathrm{Re}_c$. This was the basis of our argument in Chapter 3 where we associated the turbulent-laminar fronts and lifetimes of the metastable turbulent regions to the Directed Percolation transition. The pinning-depinning transition in directed polymers is different, since it occurs as function of T, and manifests itself in the wandering exponent for a single non-interacting polymer. In other words, in the example of $1 + 1$ dimensions, the evolution of the directed percolation system is given by 2-variable function $D(x, t)$, where $D(x, t)$ can take value of 1 (active) or 0 (inactive). In directed polymers the evolution of a directed polymer is a single-variable path in time $x(t)$.

The above argument shows that the pinning-depinning transition cannot possibly be in the Directed Percolation universality class in $1+1$ dimensions. However, it does not necessarily rule out $2+1$ dimensional transition. The pinning-depinning transition does not exist for uncorrelated noise, but it may exist for correlated noise since correlations in the noise increase the fluctuations and the effect of the disorder. We will postulate in Section 4.7.1 one way in which this transition can be observed in $2+1$ dimensions directly through the Burgers equation. Now, we turn our attention to the $1+1$ dimensional case. Is there any possible way to observe the Directed Percolation transition in the Directed Polymers model but not through the pinning-depinning transition? In this Chapter, we will show that there is. It relies on the fact that in Directed Percolation one can define a unique percolating path in the thermodynamic limit at the percolation threshold. This percolating path is 1-dimensional function of time, just like the path that the polymer takes. The wandering exponents of the two paths can thus be directly compared. This is the subject of the following Section.

4.5 Transition in Directed Polymers for $T = 0$

So far, we have used a continuous description of the Directed Polymers. In this Section we will introduce a discrete model of Directed Polymers on a diagonal square lattice, in order to be able to run numerical simulations, as well as compare our results with directed percolation on equivalent lattices. We choose a bimodal form of the noise function $\phi(x,t)$. In particular, ϕ will be allowed to take values of 0 or 1. The bimodal form of the noise function ϕ is chosen for a reason, the points where ϕ is 1 are analogous to the missing nodes/bonds in the problem of percolation on a lattice.

Examples of polymers on a site-disorder and a bond-disorder lattice are shown in Figure 4.2. The paths shown in the figure are the optimal ones, which minimize the total energy cost along the path of the polymer $\sum_\tau \phi(x(\tau),\tau)$. Because ϕ are drawn from a discrete distribution, degeneracies of the paths are possible. We now turn our attention to how to find the optimum polymer path numerically.

4.5.1 Numerically finding the $T = 0$ path of the polymer

A directed polymer in the lattice described in Section 4.5 has 2^t possible conformations for a lattice of size t. It would seem that finding the path that optimizes the energy cost of the conformation is a problem that requires $O(\exp t)$ computations. However, the energy cost of a path can be expressed as a recursive function of the energy costs of shorter paths. Whenever an optimization problem has such a recursive formulation, one can create a polynomial time algorithm to perform the optimization. This procedure was termed “dynamic

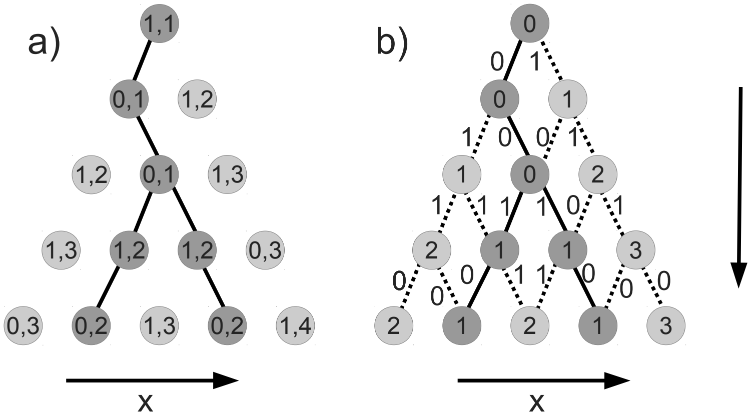


Figure 4.2: A representation of a directed polymer on a diagonal lattice with (a) site disorder, (b) bond disorder. The arrows indicates the direction of time and space. The solid lines indicate the conformation of the optimum (minimum energy) polymer path. Note the degeneracy of the optimum paths. (a) The numbers at the sites are displayed in the form (ϕ, E_{min}), where $\phi = \phi(x,t)$ is the value of the disorder at that site and E_{min} is the minimum energy of a polymer that ends at that site and starts at the origin. (b) The numbers displayed at the sites are E_{min}, whereas the numbers displayed at the edges are ϕ_l and ϕ_r, bond disorders along left- and right-moving bonds.

programming" and first developed by Bellman in the 1940's [84].

Huse and Henley [74] were interested in calculating domain wall (boundary) statistics in an Ising model with quenched disorder. The domain boundary in an Ising model with a rectilinear grid consists of horizontal and vertical bonds. Huse and Henley considered a system where one infinite domain is separated from another by a horizontal boundary (along x). The horizontal bonds incur a cost of disorder J_x (randomly distributed). To reduce this cost, the interface roughens but this incurs the additional constant cost per bond J_z of adding the vertical bonds (along z axis). The Hamiltonian for this system is:

$$H = \sum_x \left[J_z|z(x) - z(x+1)| + J_x(x, z(x))\right]. \tag{4.21}$$

Optimally, the energy of a domain wall that begins at (x_1, z_1) and ends at (x_2, z_2) then satisfies the recursive relation

$$E(x_1, z_1; x_2, z_2) = \min_{z'} \left[E(x_1, z_1; x', z') + E(x', z'; x_2, z_2)\right], \tag{4.22}$$

for any $x' \in (x_1, x_2)$. Equation (4.22) is the basis for the dynamic programming technique, also referred to as transfer matrix approach by Derrida and Vannimenus [85]. The approach is to use (4.22) to build up the function $E(x_1, z_1; x_2, z_2)$ step-wise starting from $E(x_1, z_1; x_1, z_1) = J_x(x_1, z_1)$, then using

$$E(x_1, z_1; x_2, z_2) = \min_{z'} \left[E(x_1, z_1; x_2 - 1, z') + E(x_2 - 1, z'; x_2, z_2)\right], \tag{4.23}$$

where, because of (4.21)

$$E(x_2-1,z';x_2,z_2) = J_z|z_2-z'| + J_x(x_2,z_2). \tag{4.24}$$

and $E(x_1,z_1;x_2-1,z')$ is stored from the previous step. Therefore to calculate the optimal position of the domain $z(x)$, we have to keep the computed $E(x-1,z)$ for all the possible z [86]. Typically these simulations are ran in a long strip geometry with z ranging from 0 to z_{max} and the strip is extended along the x direction, i.e. $x_{max} \gg z_{max}$. Thanks to the dynamic programming technique, the time-complexity of this calculation is $O(z_{max}x_{max})$, instead of $O(\exp x_{max})$.

Huse and Henley found the optimum domain walls, which minimized the combination of roughening costs and disorder. In the case of Directed polymers at $T=0$, there is no roughening cost since a polymer is only allowed to go left-down and right-down in the lattice shown in Figure 4.2. Therefore, the recursion relation is (using notation from the Figure)

$$E_{min}(x,t) = \min\left[E_{min}(x-\frac{1}{2},t-1), E_{min}(x+\frac{1}{2},t-1)\right] + \phi(x,t) \tag{4.25}$$

for site-disorder case (Figure 4.2a) and

$$E_{min}(x,t) = \min\left[E_{min}(x-\frac{1}{2},t-1) + \phi_r(x-\frac{1}{2},t-1), E_{min}(x+\frac{1}{2},t-1) + \phi_l(x+\frac{1}{2},t-1)\right] \tag{4.26}$$

for the bond-disorder case (Figure 4.2b). With these recursion relations it is sufficient to store $E_{min}(x,t-1)$ in order to compute $E_{min}(x,t)$. The dynamic programming technique illustrated in this section is powerful and has been adapted to many systems, including percolation (and directed percolation) in lattices [63].

4.5.2 Directed Percolation transition

Armed with the numerical method of Section 4.5.1, we can measure the distance of the average position of the polymer as function of time at $T=0$. In this specific case we will consider a bimodal distribution for the $\phi(x,t)$. We allow $\phi(x,t)$ to take values 0 or 1. The motivation for this distribution stems from the work of Balents and Kardar [87]. They considered the problem of a directed polymer on a diagonal lattice with bonds missing (with probability $1-p$). For $p < p_c$, where p_c is the directed percolation threshold for the lattice, there is no percolation cluster and the polymers cannot extend. It is interesting however, to see what happens at p_c. At the percolation threshold, the polymers are constrained to extend only along the percolation cluster wedge whose boundaries are defined by

$$x(t) \ \sim \xi_\perp/\xi_\| \sim t^{\nu_\perp}/t^{\nu_\|}. \tag{4.27}$$

(This is via the same argument as that in Chapter 3.) Thus, the polymer's wandering exponent ζ will be constrained by the smaller of $\nu_\perp - \nu_\|$ and 2/3 (the usual polymer wandering exponent). Since in $1+1$ dimensions $\nu_{perp} - \nu_\| \approx 0.63$, ζ will take that value when $p = p_c$. Above the percolation threshold ζ will restore to the usual value of 2/3.

The idea that Balents and Kardar considered was generalized by Lebedev and Zhang [5]. Instead of removing bonds from the lattice, they considered a full lattice but with a bimodal distribution for ϕ parametrized by p

$$\phi(x,t) = \begin{cases} 1 & \text{with probability } 1-p, \\ 0 & \text{otherwise.} \end{cases} \tag{4.28}$$

In this case the bonds with $\phi(x,t) = 1$ are energetically unfavorable, and are analogous to the missing bonds in Balents' and Kardar's model. However, in this case the model can be readily extended to the $p < p_c$ regime, since there is nothing to prevent the directed polymer from acquiring a conformation in the lattice. In this model, below p_c, the polymers acquire a conformation with $E \sim t$, whereas at and above p_c, they acquire a conformation with $E = 0$ since a percolating path consisting only of $\phi = 0$ links is available. Lebedev and Zhang found a surprising result that for $p < p_c$, the exponent $\zeta = 0.63$, the Directed Percolation value.

Perlsman and Havlin in 1999[88] challenged the results of Lebedev and Zhang by showing, in fact, that for $p < p_c$, $\zeta = 2/3$, the directed polymer value, but at the $p = p_c$, there is a transition to the 0.63 value. In particular they ran numerical simulations in geometry of Figure 4.2b, and identified $x_l(t)$ and $x_r(t)$, the positions of the left-most and right-most energy minima. The $x_l(t)$ and $x_r(t)$ may be the same for some t, but in general the optimum paths will be degenerate. They then defined the quantity

$$D(t) = \frac{x_l(t) + x_r(t)}{2}, \tag{4.29}$$

where $D(t)$ is the distance of the optimum end-point from the origin at time t. Evidently, $D(t)$ can be used as a proxy for computing the scaling of the wandering exponent (4.20). For the case of the bond percolation model (4.26) the transition in the wandering exponent is illustrated in Figure 4.3. It was originally computed by Perlsman and Havlin. I recomputed it, given the delicacy of the numerical issues, and was able to confirm their findings for myself.

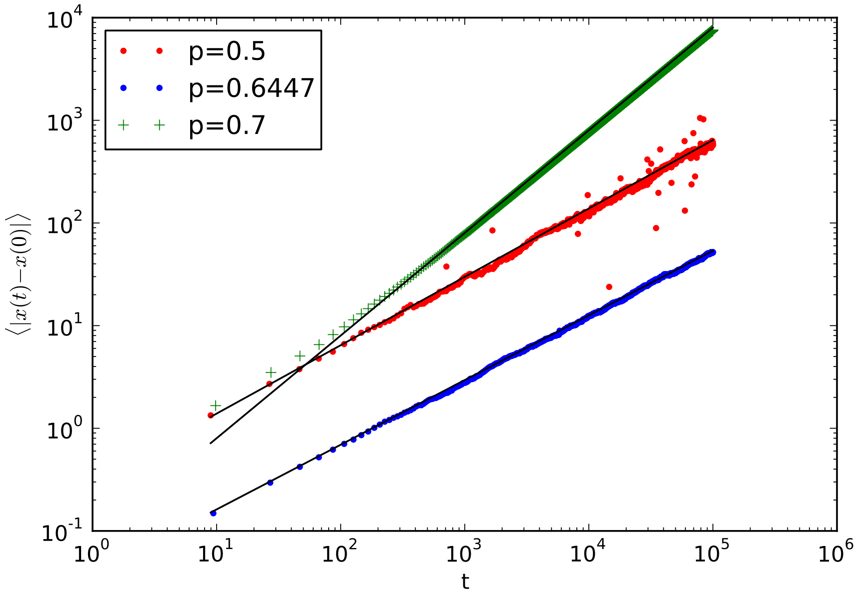


Figure 4.3: Calculations of the wandering exponent of a directed polymer with bond disorder, for $p < p_c$, $p \sim p_c$ and $p > p_c$. Solid lines are guides for the eye with powers 2/3 (for $p < p_c$), 0.63 (for $p \sim p_c$) and 1 for $p > p_c$.

Perlsman and Havlin also studied the scaling of the "cloud" of nearly optimal end-points of the polymer as function of time. They define the width of this cloud via

$$W(t) = x_r(t) - x_l(t). \tag{4.30}$$

It has been known that below the percolation threshold there is a region of points of width $\sim t^{1/3}$ around the global optimum, whose energies as calculated via (4.25) and (4.26) asymptotically approach the global energy minimum [89, 90]. Therefore at large t, and when $p < p_c$, we expect that $W(t) \sim t^{1/3}$. However, at the percolation threshold the percolating cluster width scales with the time as $t^{\nu_\perp - \nu_\parallel}$ and this scaling is observed for $W(t)$ too. Hence, at $p = p_c$, we have $W(t) \sim t^{0.63}$ for $1 + 1$ dimensions.

4.5.3 Transition in site polymer model

Results of Section 4.5.2 were worked out by various authors in 1990's for the case of bond percolation and equation (4.26). Would the same results hold for site percolation and (4.25)? On the one hand, the equations, the lattice, the structure of the noise field ϕ and the percolation thresholds given in Table 2.2 are all different. On the other hand, as described in Section 2.3.1, Directed Percolation is conjectured to be universal for a wide range of models, as long as basic conditions are met, and those don't seem to be different in this case. Here, we show, in fact, that the directed polymer to directed percolation transition happens for the case of site percolation too.

To observe the transition clearly, we plot the local exponent of the power law. Consider a function of one variable $f(t)$. The idea is to compute the local power-law exponent via

$$f'(t) = \frac{\log(f(t+\tau)) - \log(f(t))}{\log(t+\tau) - \log(t)}, \tag{4.31}$$

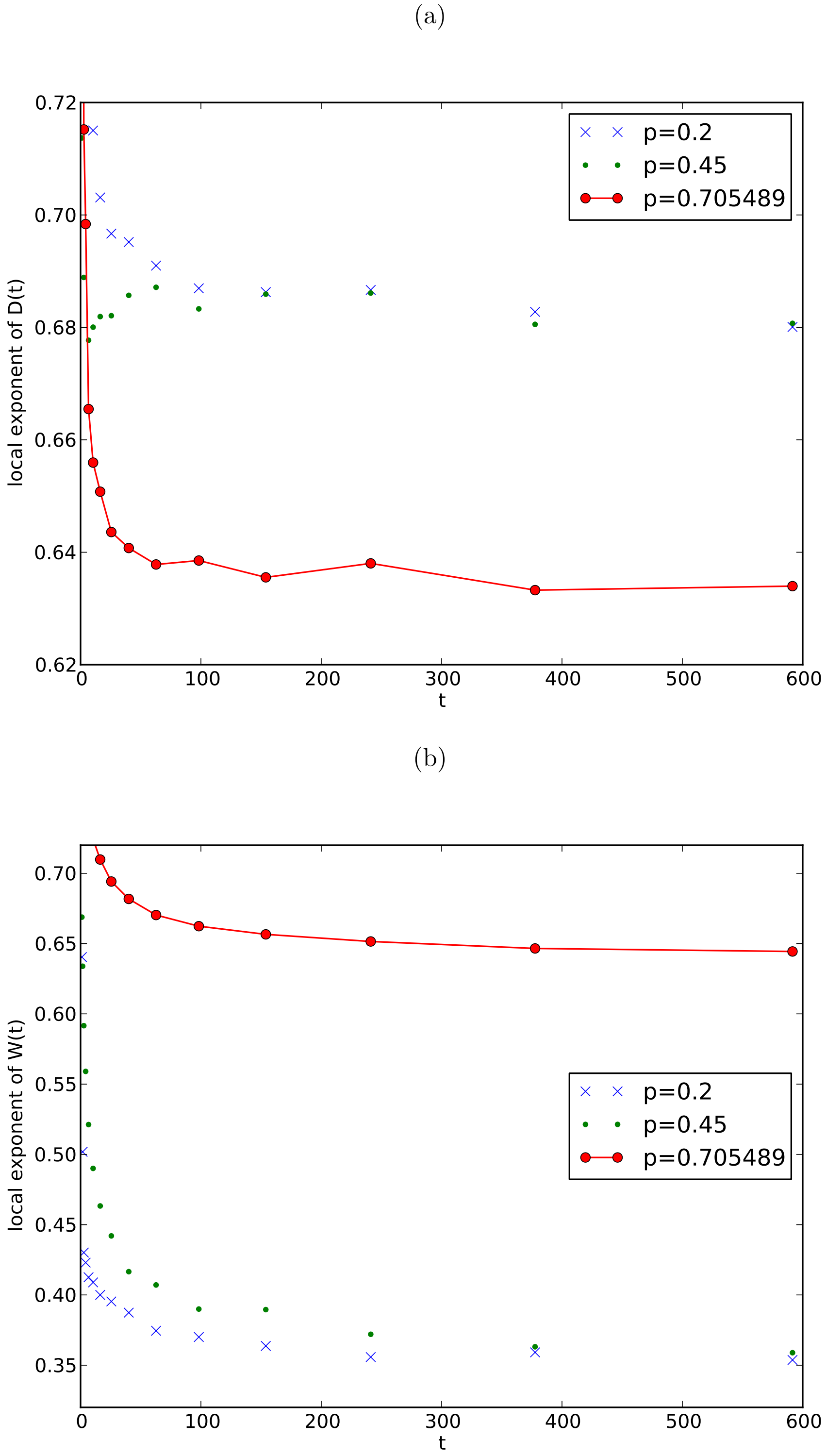
where τ is the step size for the calculation of slope. In practice, with the directed polymer datasets, this is tricky to do. It is convenient to choose step size $\tau = t$ so that

$$f'(t) = \frac{\log(f(2t)) - \log(f(t))}{\log(2)}, \tag{4.32}$$

and one only needs to compute $f(t)$ at times $t = 2^n$, $n = 0, \ldots$. However, one needs many measurements of $f(t)$ to have sufficient statistics to resolve small differences in the power law exponent. It would be useful to use values of $f(t)$ for $t \neq 2^n$ to have the computation converge faster – it is quite slowly converging otherwise. One approach would be to bin the values of $t \neq 2^n$ into the nearest $t = 2^n$ bin. This works well to reduce the noise in measurements of $f(t)$. However this sort of uncontrolled binning can subtly change the power law exponents if performed without any weighing. Another approach is to construct a cubic interpolation function over the $f(t)$ versus t data, so that the exponent can be calculated directly via (4.32). However, cubic interpolation uses too little of the $f(t)$ measurements for large t. In this chapter we take a hybrid approach, we perform a prebinning with logarithmic bins to reduce the number of points that define $f(t)$. Then we construct a cubic interpolation which is used to compute the exponent via (4.32).

Results of running the site polymer model, the local exponents of $D(t)$ and $W(t)$ are shown in Figure 4.4. In the figure, note the difference in the y-scales: the proxy for the wandering exponent, $D(t)$, undergoes a more subtle transition between exponents of 0.63 and 2/3, whereas $W(t)$ undergoes a much clearer transition between 1/3 and 2/3. In this case simulations have been run for t up to 1024 but the $t \sim 590$ data points indicate the average slope of the power law up to 1024 in accordance with the method outlined above.

We have just shown that site percolation on diagonal lattices also undergoes the directed-polymer, directed-percolation transition, at least in the $T = 0$ case. This is important since in Section 4.6.2 we will take the next logical step and go from this lattice to a continuum equation. Sites on a lattice can be directly interpreted as continuous, smooth regions of the noise field $\phi(x,t)$. This interpretation would be harder to motivate for a bond-percolating lattice. However, before we take the continuum limit, we first


(a)

(b)

Figure 4.4: The local exponents of (a) $D(t)$ and (b) $W(t)$ for the case of a site polymer model.

consider the $T \neq 0$ case.

4.6 Transition for $T \neq 0$

In the directed polymers field of study, researchers first considered the $T = 0$ paths numerically, due to the simplicity of generating the optimum paths (as shown in Section 4.5.1). For non-zero temperatures, numerically generating the polymers through a classical algorithm such as Metropolis-Hastings is difficult since the polymer can make large jumps in phase space. For instance, the polymer can change its conformation for small t thereby making the conformation for large t impossible. The research has hence focused on theoretical study of Directed Polymers at $T > 0$, most notably the pinning-depinning transition described in Section 4.4.

Directed polymers in random media described in Section 4.5 have a rugged phase space. In the same instantiation of noise field $\phi(x,t)$ it is possible for an optimal $T = 0$ polymer path with an endpoint at $t = t_2$ to have an entirely different conformation from that of a polymer with an endpoint at $t = t_1$. Hence, at each time step there are multiple competing local optimum paths that compete for the global optimum. This raises an interesting question: what are the statistical properties of these local optimum paths? Do the wandering exponent results (valid for global optimum paths) also hold for the local optimum neighborhood? To answer this question fully, we now turn our attention to the non-optimal polymer trajectories.

4.6.1 Non-optimal paths

Recall that one can use the dynamic programming technique from Section 4.5.1 to compute the minimum energy E_{min} of a directed polymer starting at $(x = 0, t = 0)$ and ending at $(x = x_f, t = t_f)$. For a particular t_f, there may be a few x_f for which E_{min} is smallest possible, but in general, for most x_f the polymer energy will be non-optimal. In this section we study these non-optimal polymer trajectories. To clarify, the trajectories we consider are still generated via the techniques of Section 4.5.1 but they are not necessarily globally optimal.

What do these trajectories look like? This question was first studied by Kardar and Zhang in 1987 [91], and they found that these paths have a tree-like structure, lending support to the universality of the wandering exponent scalings. Kardar and Zhang considered the case of Gaussian distribution of the $\phi(x,t)$. We find that in the case of bimodal DPRM, the paths are still tree-like, and do not depend much on p_c – apart from the in-filling when $p \gg p_c$. This is so, because when $p \gg p_c$, the percolating cluster itself fills space and there are many degenerate paths that merge. The examples of trajectories are shown in Figure 4.5.

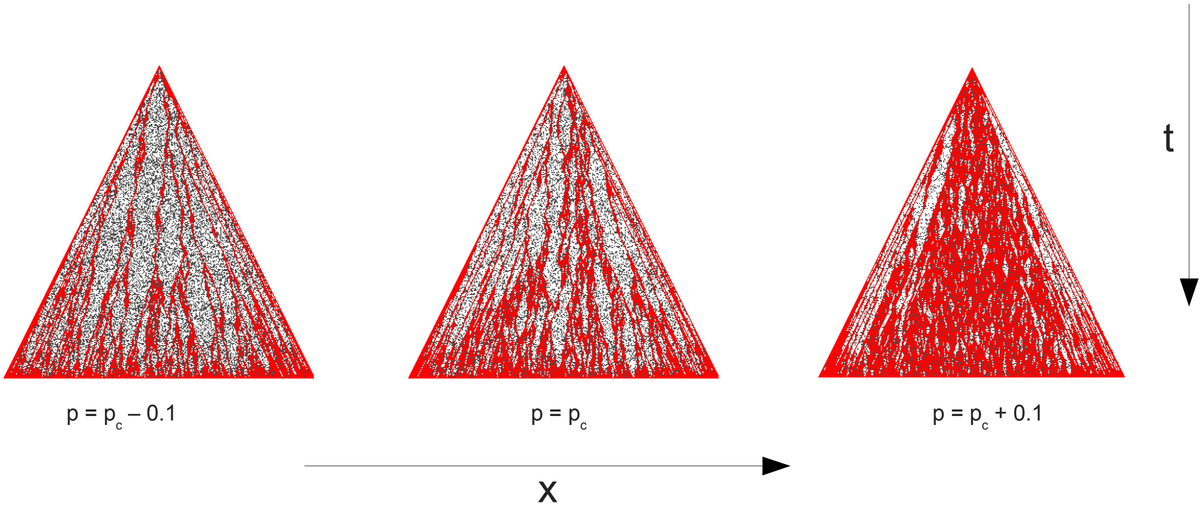


Figure 4.5: The trajectories of the optimum paths (red) to (x_f, t_f), starting from $(0, 0)$ shown here for each x_f. Gray areas indicate the locations of the disorder $\phi(x, t) = 1$. For $t < 5$, $\phi(x, t) = 0$ in order to reduce the (sizable) random effects of initial conditions.

We now make a few comments on the numerical methods of tracing the trajectory of the paths shown in Figure 4.5. Unlike in Section 4.5.1, one needs to keep $E_{min}(x, t)$ for all x, t in the wedge geometry. Once $E_{min}(x, t)$ is computed for all $t < t_f$, one starts from $t = t_f$ and "traces back" the trajectory. Each node in the $t - 1$ slice with $E_{min}(x, t - 1)$ is a part of a trajectory if one of the nodes directly below it with $E_{min}(x \pm 1/2, t)$ is a part of the trajectory and satisfies

$$E_{min}(x, t - 1) = E_{min}(x \pm \frac{1}{2}, t) - \phi(x \pm \frac{1}{2}, t). \tag{4.33}$$

The paths shown in Figure 4.5 are shown without consideration for the energy of the polymer. We now introduce temperature T in the following way. We weight the probability of each polymer that ends (optimally) at x_f, t_f with a Boltzmann factor

$$\exp\left(-\frac{1}{T}E_{min}(x_f, t_f)\right) \tag{4.34}$$

The dependence of the probability distributions for the polymer conformations can then be given a dependence upon T. This dependence is illustrated in Figure 4.6 where the density of the filling indicates the probability density of the polymer being at that point in space and time. We make a few observations about the Figure. First, as expected, for very large T, all end-points of the polymers are possible. Second, notice

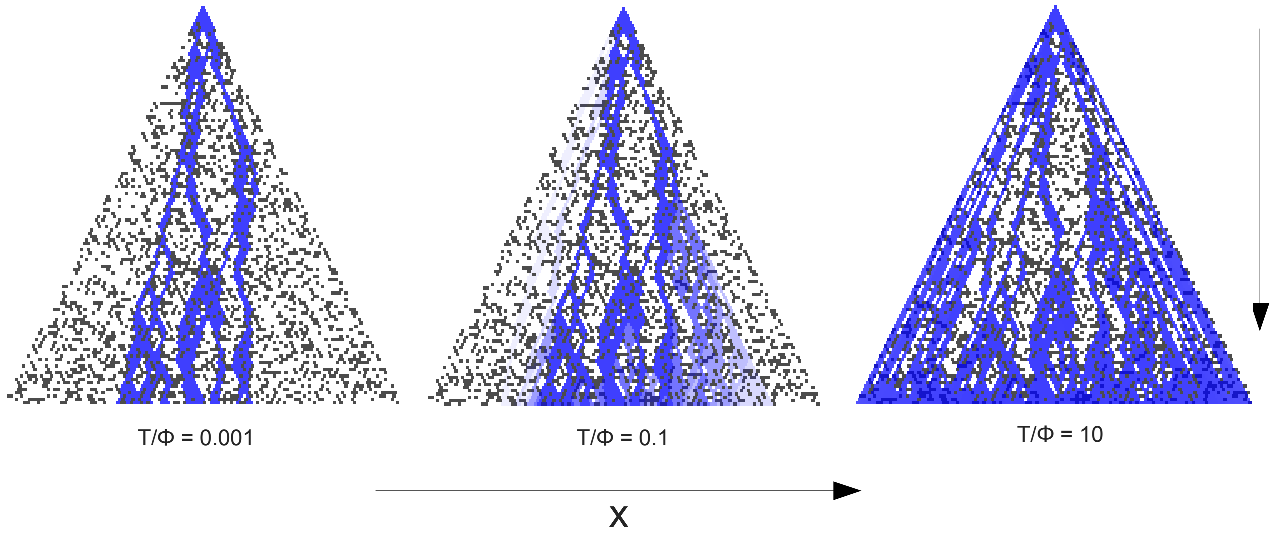


Figure 4.6: The probability distributions of the optimum paths (blue) of polymers starting from $(0, 0)$ with no constraint upon the endpoint of the polymer. The results shown here are for 3 temperatures for the same lattice disorder $\phi(x, t)$ with $p = p_c = 0.7055$ and amplitude of disorder Φ. Gray areas indicate the locations of the disorder $\phi(x, t) = \Phi$.

that the degeneracy of the optimal and nearly-optimal paths is high even for very small T. Third, increasing T slightly tends to blur the distribution towards the ends of the polymer (large t), leaving the probability distributions at small t largely unchanged.

These observations serve as evidence that, even when $T \neq 0$, there exists a directed-polymer directed percolation transition in p, as function of T/Φ where Φ is the magnitude of the disorder. The transition would be observable for small-time wandering exponent, but it would need a large running-time of the simulation in order to escape the effects of the shallow profile of E_{min} at $p = p_c$ [92]. This transition is important, since in the language of the Burgers equation it may exist at a finite Re. We now proceed to describe what this transition looks like in the context of the Burgers equation, first discussing it in context of the directed polymer partition function Z.

4.6.2 Directed percolation transition in the partition function

In Section 4.5, the Directed Percolation transition occurs in a discrete lattice. However, the Burgers equation is a partial differential equation that is defined on a continuous space. Can we generalize the results of Section 4.5 to the Burgers PDE? In this section, we will first consider how the transition occurs in the partition function for the polymers, $Z(x, t)$ undergoing a dynamics given by a partial differential equation. In the subsequent section, we will map these results further to the Burgers equation.

One can consider a mechanistic way of looking at the dynamics of Equation (4.17). Consider $Z(x,t)$ as the probability distribution of the polymer position at time t. The term $\nabla^2 Z(x,t)$ lets the probability density explore the lattice through a diffusive process. When the probability density enters regions with positive disorder, i.e. coarse-grained sites of the lattice where $\phi(x,t) > 0$, it is exponentially supressed through the $\phi(x,t)Z(x,t)$ term. In this way, the probability density $Z(x,t)$ undergoes an expansion through space, naturally "choosing" the paths where $\phi(x,t) = 0$.

How does the mechanistic dynamics of $Z(x,t)$ map to the optimum path of the polymer in a lattice described previously? At zero temperature, the dynamics are not exactly the same. For the optimum path, we have the position of the path as given by the $x(t)$ for which $E_{min}(x,t)$ is at the minimum (for a given time slice t). Recall the recurrence relation for E_{min}:

$$E_{min}(x,t) = \min\left[E_{min}(x-\frac{1}{2},t-1), E_{min}(x+\frac{1}{2},t-1)\right] + \phi(x,t), \tag{4.35}$$

where the true endpoint of the polymer is given by x for which $E_{min}(x,t)$ is minimum. Recall that the dynamics of $Z(x,t)$ is given by the PDE (4.17). The effective recurrence relation for $Z(x,t)$, based on the dynamical perspective described above (and straightforward finite difference Euler scheme) is given by

$$Z(x,t) = (1-\Phi\phi(x,t))Z(x,t-1) + D\left[Z(x-\frac{1}{2},t-1) - 2Z(x,t-1) + Z(x+\frac{1}{2},t-1)\right] \tag{4.36}$$

where Φ is the strength of the disorder and D is the effective diffusion coefficient. The "true" endpoint of the polymer in this case is the mode of Z which we will define in two ways. First is the simple maximum of $Z(x,t)$,

$$x_{max}(t) = x' \mid Z(x',t) = \max_{x'} Z(x,t). \tag{4.37}$$

The second is the first moment of $Z(x,t)^n$:

$$\overline{x}_n(t) = \int Z(x',t)^n x' \, dx'. \tag{4.38}$$

Here we will take $n = 1$, but $n > 1$ would define a "sharper mode" of $Z(x,t)$. In the remainder of this section, we will conjecture and attempt to motivate the above two metrics as proxies for Perlsman and Havlin's $D(t)$ given in (4.29). We propose that the above measures will also undergo the directed-polymer directed percolation transition in the wandering exponent just as $D(t)$ does. As a proxy for W (Equation

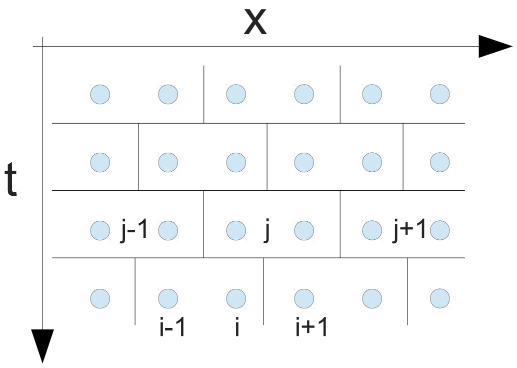


Figure 4.7: The lattice used for the simultaneous simulations of $Z(x,t)$ and optimum paths on diagonal lattice. The lines indicate boundaries between sites where the disorder $\phi(x,t)$ is specified. The disorder sites are labeled by j. The numerical evaluation of $Z(x,t)$ is performed using finite differences on the grid specified by indices labeled by i.

4.30), we propose the standard deviation of $Z(x,t)^n$,

$$\sigma^2_{Z,n} = \langle x^2 \rangle_Z - \langle x \rangle^2_Z = \int Z(x',t)^n x'^2 \, dx' - \left(\int Z(x',t)^n x' \, dx' \right)^2 . \tag{4.39}$$

We expect that (4.39) will transition from the exponent of 1/3 to $\nu_\perp - \nu_\parallel$ at p_c.

In order to evaluate (4.36), and preserve the diagonal lattice structure, we construct a rectilinear grid in the way illustrated in Figure 4.7. Sites with the disorder $\phi(x,t)$ are straddled in the same fashion like in a diagonal lattice in Figure 4.2. Each of these sites is composed of two sub-sites (or a multiple of 2) on which the $Z(x,t)$ dynamics occurs. This sort of a straddled lattice is easily generalized to higher dimensions.

The numerical construction in Figure 4.7 allows us to simultaneously run numerical simulations for both the optimum path in a lattice and $Z(x,t)$ on a rectilinear grid. In this chapter, we use (4.36), a first-order finite difference scheme for the evaluation of Z. Optimum paths are computed via (4.35). The comparison of the optimum paths and $Z(x,t)$ is shown in Figure 4.8. In the Figure, blue sites indicate sites along the optimum path, whereas the density of the green color indicates the magnitude of $Z(x,t)$. In Figure 4.8a, we verify that the numerical computation works as expected by turning off the strength of the noise and observing a diffusive process. In Figure 4.8b, we see that for $p < p_c$, a small diffusion $\nu = 0.2$ is sufficient for the mode of Z to track the optimum path. In Figure 4.8c, we see the scaling of the noise field at critical p. At critical percolation threshold the optimal path is highly degenerate. The mode of $Z(x,t)$ has problems keeping up. In Figure 4.8d we also see the behavior at critical p, but now with smaller viscosity. The mode of $Z(x,t)$ follows an optimum path near the origin, but it is unable to reach optimum paths far away from the origin.

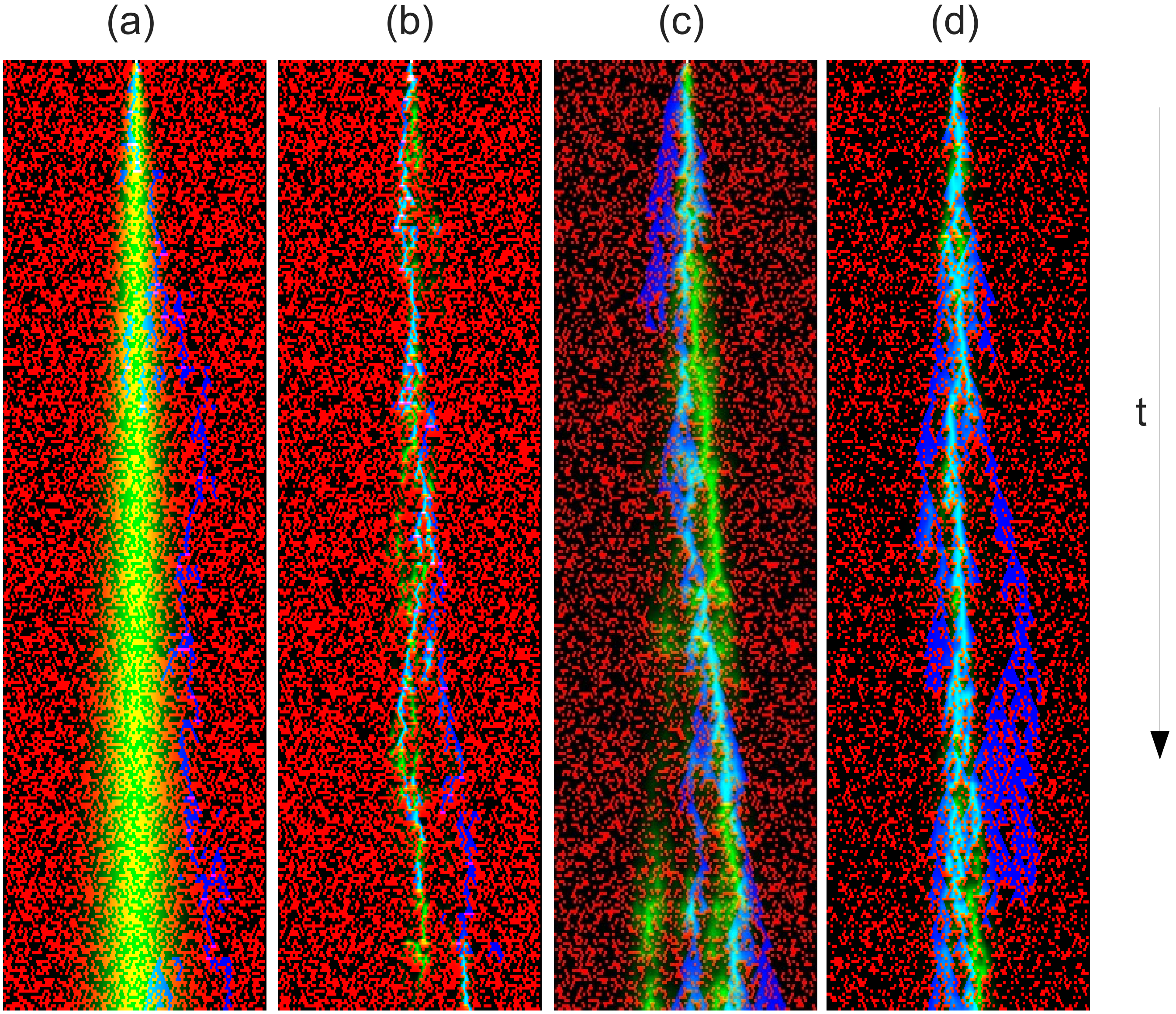


Figure 4.8: Four simultaneous numerical simulations of (4.35) and (4.36). Red indicates regions of $\phi(x,t) = 1$. Blue indicates the sites along the optimum path. The magnitude of green overlay indicates the strength of $Z(x,t)$ at that point. $Z(x,t)$ is renormalized at each time step to have unity mass. (a) With $\Phi = 0$, the dynamics for $Z(x,t)$ is simply diffusive. (b) $p = 0.5 < p_c, \nu = 0.2, \Phi = 0.2$. (c) $p = 0.7055 = p_c, \nu = 0.36, \Phi = 0.2$, (d) Same as (c) but with $\nu = 0.2$, diffusion is too slow to reach all parts of the optimum path.

Figure 4.8 gives an insight into how $Z(x,t)$ qualitatively behaves. Overall, choosing the appropriate values for ν and the noise strength Φ is key to having $Z(x,t)$ follow the optimum path. We ran more than 60,000 runs of simulations with 500,000 time steps, t ranging up to 250,000 and in a space of 16,000 grid points for cases of $p = 0.45$ and $p = 0.7055$, with finite-difference scheme of (4.17) in lattice given in Figure

4.7. We measured $D(t) \equiv \overline{x}_1(t)$ and $W(t) \equiv \sigma_{Z,1}$ and calculated the local exponents using the same method as in Section 4.5.3. We plot the results in Figure 4.9 and the local exponents in Figure 4.10. We see that the local exponents for $D(t)$ and $W(t)$ are approaching $2/3$ but we are unable to differentiate between the two phases of the transition.

The problem in observing the transition in $Z(x,t)$ may be because of the following reason: a careful tuning of the parameters is necessary. Consider our conjectured phase diagram in Figure 4.11a. From the directed polymer literature and the calculation in Section 4.6.1, we know that the transition has to occur at the critical probability p (for the lattice) and up to some (unknown) maximum viscosity ν (temperature, in polymer language). However, we see from Figure 4.8 that a careful tuning of parameters ν and Φ is necessary in order to have $Z(x,t)$ behave suitably enough to observe the slight change in local exponents. We find that one should keep the ratio of $\nu/\Phi = 1$ roughly, but with ν depending upon p and neither too small, nor too large. We hence propose that one needs to be in the suitable region of parameter space, qualitatively illustrated in Figure 4.11b, in order to observe the transition. However, we do not know if this region of parameter space overlaps with the Directed Percolation line in Figure 4.11a. That the phase diagram depends on Φ and p, properties of the noise field defined on the lattice, may be in contradiction to the statement of universality of Directed Percolation. In any case, the resolution of this seeming paradox and the observation of the transition in $Z(x,t)$ and the full understanding of the phase diagram will take a number of other numerical simulations and is left for future work.

4.6.3 Back to Burgers equation

In this section we use the reverse of the process of Section 4.3 to map the meaning of our results for $Z(x,t)$ to the velocity field $\mathbf{u}(x,t)$. In Section 4.6.2 we conjectured that

$$\overline{x}_n \equiv \int Z(x',t)^n x' \, dx' \sim t^{\zeta}. \tag{4.40}$$

What is the meaning of this statement in the language of the Burgers PDE? For the case of $n = 1$ in a system of size C, we have, after applying the Cole-Hopf map (4.16)

$$\int_0^C \exp \frac{h(x)}{2\nu} \, dx \sim t^{\zeta}. \tag{4.41}$$

Transforming h into $\mathbf{u}$ via (4.14) yields

$$\int_0^C \exp -\frac{1}{2\nu} \left(\int_K^x \mathbf{u}(x',t) \, dx' \right) dx \sim t^{\zeta}. \tag{4.42}$$

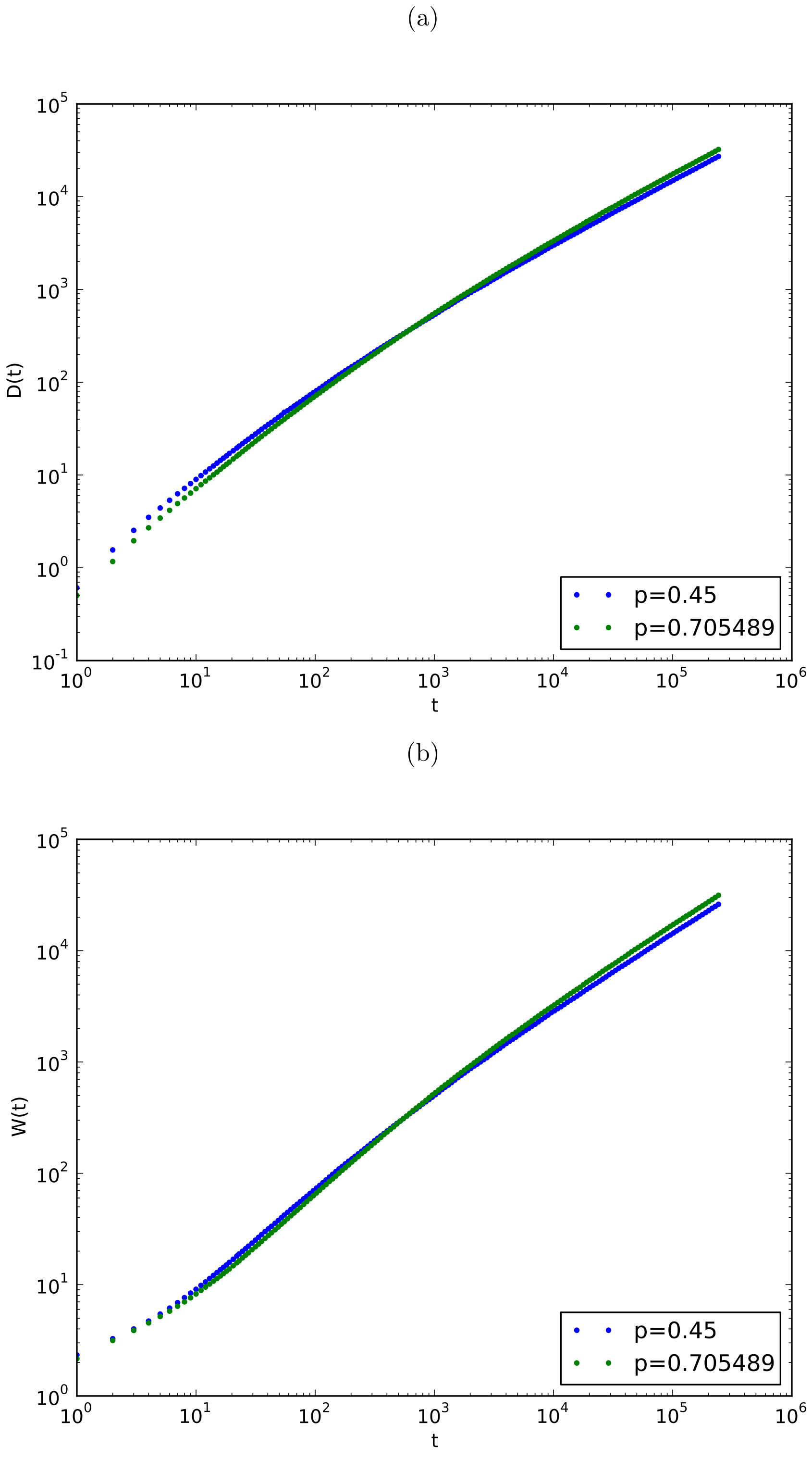
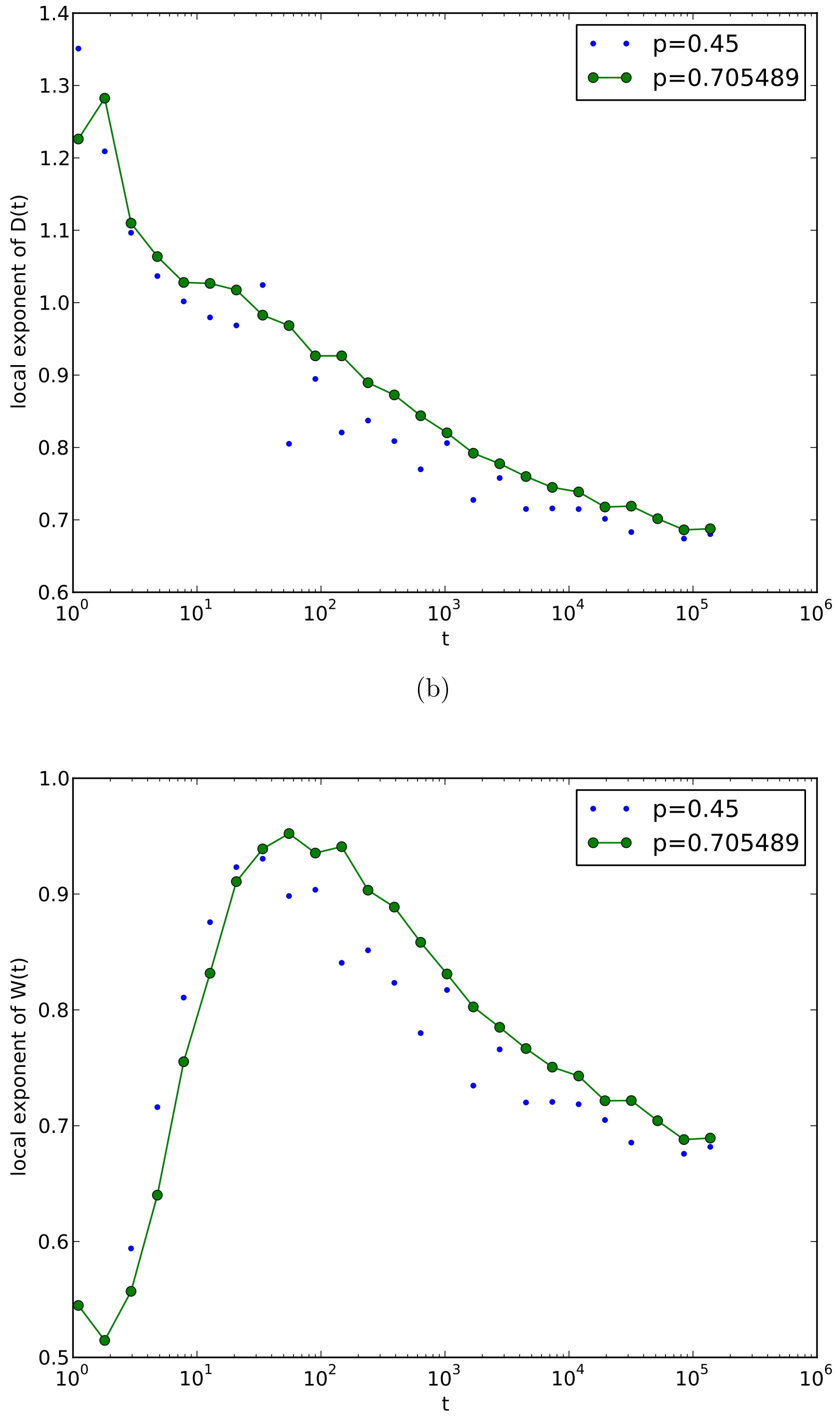


Figure 4.9: (a) $D(t)$ and (b) $W(t)$ for the case of an evolution of $Z(x,t)$ in a lattice given in Figure 4.7 with parameters $\Phi = 0.3$ and $\nu = 0.3$.


(a)

(b)

Figure 4.10: The local exponents of (a) $D(t)$ and (b) $W(t)$ for the case of an evolution of $Z(x,t)$ in a lattice given in Figure 4.7 with parameters $\Phi = 0.3$ and $\nu = 0.3$.

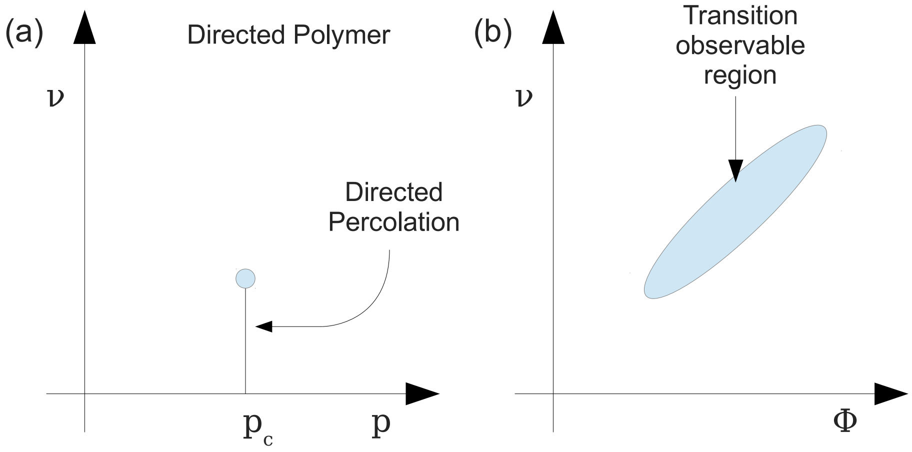


Figure 4.11: (a) Phase diagram of the directed polymer-directed percolation transition. (b) Conjectured observable region for the directed polymer-directed percolation transition in the partition function $Z(x, t)$.

The lower limit in the inner integral is arbitrary (as long as it is in the range $\{0, C\}$, since it will at most affect the constant multiplying the power law on the right. The outer integral in (4.42) is of course not surprising since we need to integrate out x, in order to observe the transition in the exponent of the power law in t. However, the inner integral is interesting since it gives a global character to the transition which stems from the global optimization process of the directed polymer in the random medium.

4.7 Conclusion

In this chapter we have given further evidence that the transition to turbulence is in the Directed Percolation universality class. Our goal was to go directly from a phase transition in a stochastic PDE to a DP statistical mechanical model, thus showing that it is in principle possible for a coarse-grained description such as DP to be an appropriate description of a fluid turbulence transition. In this we were partially successful.

In Section 4.2 we motivated our consideration of Burgers equation as the model system. In Section 4.3, we showed how one can map Burgers equation to a problem of Directed Polymers and in Section 4.4, we reviewed some relevant results for the Directed Polymer universality class. In Section 4.5, we reviewed a particular model of Directed Polymers that has a Directed Polymer-Directed Percolation transition at $T = 0$. In particular, in Section 4.5.3 we presented a novel extension of these results to the site polymer model. Then, in Section 4.6 we gave evidence for the existence of the transition at $T \neq 0$ (i.e. finite Re), in the polymers on a lattice (Section 4.6.1), the polymer partition function (Section 4.6.2) and we mapped these results to the Burgers equation (Section 4.6.3).

Our final result, (4.42) can be taken as a metric that a numerical or experimental fluid dynamicist can apply in order to try to understand the transition to turbulence. However, it is unlikely that this metric is of utility with present technology, because of the difficulty in observing the subtle transition in exponents (even in simulated systems).

In Part 1 of this thesis, we have provided evidence that the transition to turbulence is in directed percolation universality class. However, the evidence is far from conclusive, and more direct measures of the transition are needed. In this section we outline a few possible paths a future student or researcher may take in theoretically studying the transition to turbulence in hardened Burgers equation (Section 4.7.1) or in experimental/numerical two dimensional turbulence (Section 4.7.2). We also raise an interesting question in Section 4.7.3 about the relationship of extreme value statistics and spatially extended systems with a subcritical transition.

4.7.1 Direct observation of directed percolation in hardened Burgers equation

In section 4.4, we hinted at the possibility of directly measuring the directed percolation exponents through simulations of Burgers equation. We propose that one can coarse grain the solution of the Burgers equation, by "hardening", i.e. by setting a threshold for velocity $|u(x,t)|$. Regions where the velocity field crosses the threshold are analogous to active regions in DP. In that case, one can verify the DP exponents in the same way that Takeuchi *et al.* did in turbulent liquid crystal experiments (see Section 2.3.2), by measuring the active site density and correlation lengths. One can even directly track the shock fronts and measure the analogue to turbulence growth rate G.

However, this calculation is not trivial. We attempted to do this measurement in 1-dimensional Burgers equation with no success. The challenge is that in 1-dimensional Burgers equation driven by noise, left-moving and right-moving shocks can become "pinned" producing artifacts in the measurement. However, it is possible that in a 2-dimensional simulation, pinning is not an important effect. In any case, one needs to drive the Burgers equation with correlated noise, a task that is not numerically trivial, especially in more than 1 dimension.

4.7.2 Direct observation of directed percolation in turbulent fluids

The details of the phase diagram of the flow in 3D pipes described in Chapter 3 are still not completely understood around the puff-slug transition point. However, most of the phenomenology is understood through a number of experiments and direct numerical simulations. On the other hand, the laminar-to-turbulent transition is much less understood in 2D flows, such as those in gravity driven soap films. Unlike 3D turbulence, 2D turbulence is characterized by having two possible spectral exponents, the energy and the enstrophy cascades. A further study of the turbulence-laminar transition in 2D flows would be interesting for a couple of reasons.

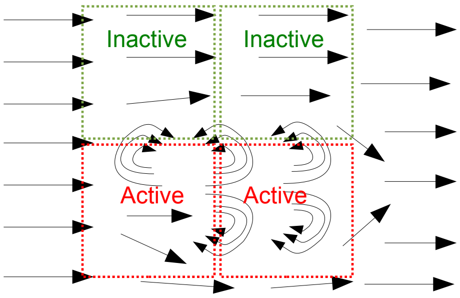


Figure 4.12: A sketch of how one can perform a coarse-graining procedure of a velocity field. For each coarse-grained site (indicated with dashed lines), average the turbulence intensity in that site. If it is above a certain threshold, then label the site as active, otherwise as inactive.

Direct numerical simulations are much easier and computationally cheaper to perform in 2 dimensions. This allows relaminarization simulations for long times and measurements of the scaling of turbulent lifetime. Furthermore, the comparison to experiments is easier. Namely, unlike in 3D flows, turbulence in a soap film flow can be readily visualized by filming the interference patterns formed by the differences in thickness [93]. These can be readily mapped into the vorticity field of the fluid [40]. From vorticities, one can solve for the velocity by solving a Poisson equation. One can also coarse-grain such measurements of the velocity field, by overlaying a larger lattice on top of the velocity field (see illustration in Figure 4.12). Those lattice sites that have a turbulence intensity larger than a certain threshold can be labeled as active. Otherwise sites are labeled as inactive. One can then consider the coarse-grained dynamics of the active and inactive sites and see if it corresponds to what one would expect in directed percolation (by calculating the density of active sites and correlation lengths as function of Re for instance).

Based on linear stability arguments, it is known that laminar-to-turbulent transition in 2D Poiseuille flow is subcritical [94]. Hence, there is hope to be able to measure puff lifetimes and slug front growth rates much like in pipe flow. The slug fronts should expand at a different rate than in 3D Poiseuille as shown in Chapter 3. We expect to see $G \sim (Re - Re_c)^{0.561}$ with a crossover to $G \sim (Re - Re_c)^{0.637}$ for $(Re - Re_c)$ small.

4.7.3 Superexponential scaling of systems with subcritical transition

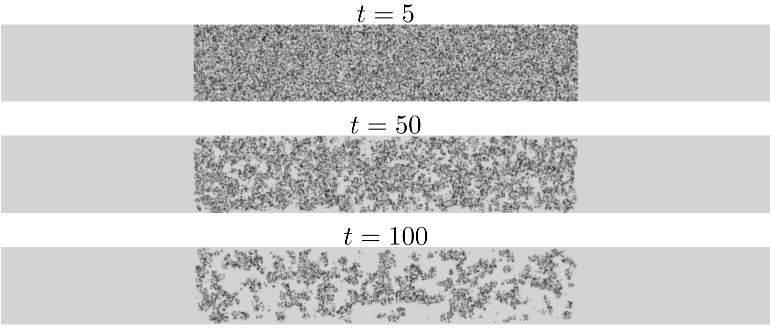


Figure 4.13: Decay of an active region in a coupled map lattice system [95] with parameter $\epsilon = 0.3$. Black sites indicate active regions, sites in the lattice where the dynamics is in the chaotic region. White sites indicate inactive (absorbing) regions, sites in the lattice where the dynamics is in the neighborhood of a stable, attracting fixed point. Systems such as this also seem to have superexponentially scaling lifetimes of active state.

Another interesting question relates the association of turbulence lifetimes and extremal statistics shown in Chapter 3. This seems to be a property of subcritical bifurcations. Subcritical bifurcations differ from supercritical bifurcations in that when a bifurcation threshold is attained, a previously stable state becomes metastable and the fixed point of the system discontinuously changes [18]. It is thought that turbulence is always in a metastable state, it is simply that when Re is sufficiently large, its lifetime becomes much longer than that of the age of the Universe [2]. This phenomenon also happens in many spatially extended systems [96]. In particular, subcritical transitions appear in Kuramoto-Sivashinsky PDE [97], as well as in at least one type of a reaction-diffusion system [98]. We've measured such dynamics in the case of Chate's coupled map lattice [95], graphically shown in Figure 4.13. It would be interesting to see if systems other than turbulence also show superexponential lifetime as a function of the bifurcation parameter, as we would expect from the extremal statistics arguments, and when and under what conditions does extremal statistics argument not apply.

Part II

Community Dynamics in Microbial Ecology

Chapter 5

Introduction to Metagenomics and Ecology

Scientific research in quantitative microbiology can be roughly divided according to scale:

1. single molecule experiments, such as measurements of physical properties of biological molecules;
2. cultured single cell experiments, such as measurements of genomics, transcriptomics or cell processes such as motility and division;
3. bulk experiments using cultured cells, such as measurements of biofilm formation, quorum sensing and chemotaxis in controlled experiments;
4. bulk experiments using uncultured cells, such as measurements of metagenomics and metatranscriptomics.

In this thesis we will consider some problems that are squarely in the last category. In particular, our approach will be a physical one rather than a traditionally biological one: we will not mention any specific taxa by name. Instead, we will focus on what we can state about biological systems in a generic and statistical way by focusing on the collective phenomena of the microbial communities.

Why study microbial communities? They are estimated to form as much biomass as all the plants on the planet [99], they oxygenated the atmosphere of the earth 2 billion years ago [100], and they may remain a key component of the global climate and carbon cycle [101]. A diverse community of microbes are a necessary part of a human body–there are 10 times as many microbial cells as there are human cells in a human body [102]. The importance of microbiomes in biomedicine is evident and it is exemplified by the Human Microbiome Project [102, 103].

It is believed that more than 99% of microbes in nature are not culturable in a lab setting. While they are usually viable, they do not form colonies on plates [104]. This necessitates *in situ* measurements such as metagenomics. Metagenomics [105] is the study of genes present in the organisms living in an environment. We briefly review methods of metagenomics in Section 5.1. On the modeling side, we use methods from theoretical ecology to characterize communities of microbes. A short review of the relevant theory of ecology is given in Section 5.2.

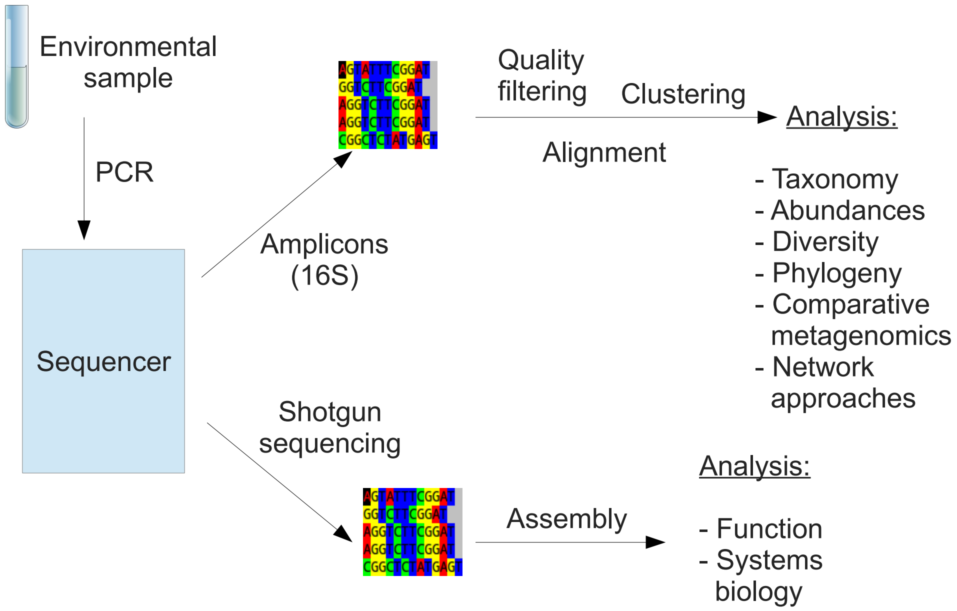


Figure 5.1: High-level overview of the metagenomics process. Depending on the study, one can perform a metagenomics of amplicons of a specific gene (such as, say, 16S), or a *de novo* assembly of contiguous sequences and even entire genomes.

5.1 Sampling a microbial environment

In Figure 5.1 we give a general overview of a metagenomic study. The DNA content of a sample from the environment is amplified and fed into a high-throughput sequencer. Depending on the study, a specific gene can be amplified, or an assembly of contiguous sequences can be made into entire genomes [106]. A typical choice for a specific gene study is the 16S gene, a necessary part of each bacterial ribosome's smaller subunit. Because 16S is necessary for the operation of the ribosome, it is strongly conserved and hence serves as a good "evolutionary clock". The phylogeny (tree of the historical evolution of the molecule) has enabled Woese to trace descent of life on Earth into three domains [107].

Once reads are produced by the sequencer, they are processed for quality. Each sequencing platform has different types of read errors that can occur including chimeric reads–errors where one read is actually a concatenation of two different reads [108]. In the case of *de novo* assemblies, short reads are algorithmically matched and assembled into contiguous regions [109]. In the case of 16S, the reads are typically aligned as the next step in the study. We now turn our attention to sequence alignment.

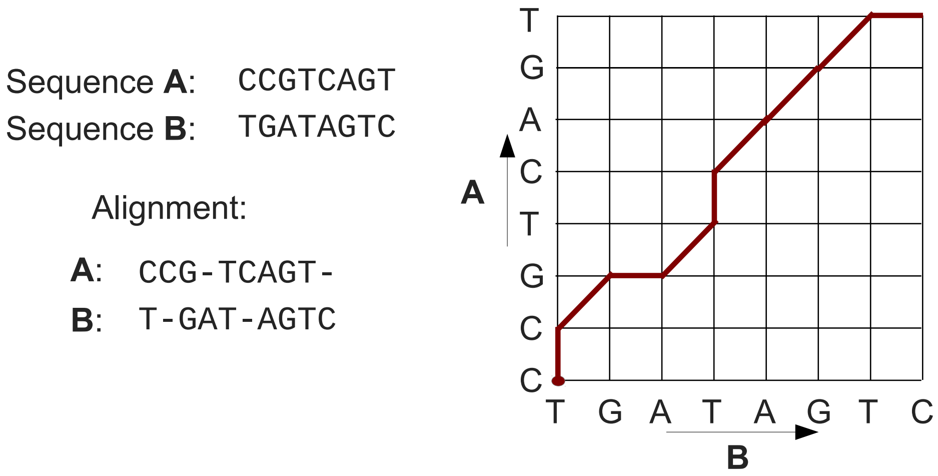


Figure 5.2: Overview of the Needleman-Wunsch algorithm. On the left, two sequences are given, and one of their possible alignments. On the right, a path through the lattice that corresponds to the alignment is shown in red.

5.1.1 Sequence alignment

In order to facilitate comparisons between the metagenomic sequences, an alignment of sequences to each other is necessary. This is called a multiple alignment, and the canonical algorithm for performing a multiple alignment was developed by Needleman and Wunsch [110]. Since the algorithm uses the same principle, dynamic programming, that we used in Chapter 4 to solve for directed polymer paths, a number of relationships have been discovered between the statistics of alignment scores for related sequences and that of scaling of directed polymers [111].

The Needleman-Wunsch algorithm operates on the basis of minimizing the cost of the alignment (or similarly, maximizing the similarity of aligned sequences). Consider the example of aligning two sequences: A and B. The alignment score $S_{00,xy}$ for aligning the first x letters of A to first y letters of B is simply computed via:

$$S_{00,xy} = G_{xy}g + M_{xy}m, \tag{5.1}$$

g and m are the costs for gaps and mismatches and G_{xy} and M_{xy} are numbers of gaps and mismatches, respectively. If necessary, one can also readily generalize the algorithm to use different costs for transition (A↔ G, C↔ T) and transversion (A↔ T, A↔ C, G↔ C, G↔ T) mutations. The algorithm operates by progressively filling in the lattice illustrated in Figure 5.2. In the Figure, the alignment at the left is represented by the path in the lattice shown on the right. Horizontal and vertical steps in the lattice indicate gaps. The alignment score can be expressed in terms of a recursion relation, involving a minimum of the score of the sub-sequences:

$$S_{00,xy} = \min_{x'y'} \left(S_{00,x'y'} + S_{x'y',xy} \right) \tag{5.2}$$

One can implement a dynamic programming algorithm to solve (5.2) much like the one we described in Chapter 4.

For aligning two sequences, the Needleman-Wunsch algorithm's operating time scales as $O(N^2)$ where N is the length of the sequences. The algorithm can be easily extended to multiple alignment, by changing the lattice in Figure 5.2 into a hypercube. However, in that case the operating time of the algorithm scales as $O(N^n)$ where n is the number of sequences. This exponential scaling of time is what makes this algorithm prohibitive for multiple alignments of large datasets. To go around this problem, we will introduce two alternative alignment methods and investigate their performance in Chapter 6.

5.1.2 Analyzing metagenomic reads

With alignment complete, an analysis of metagenomic reads can reveal a plethora of information about the underlying community. Whole genome assemblies can be used to detect presence of particular genes or instances of horizontal gene transfer. The genome can be used to construct a metabolic network, which can be studied through flux balance analysis to understand the metabolic pathways and functions of organisms in the community [112].

Amplicon reads such as 16S rDNA are typically clustered following the alignment into OTUs (operational taxonomic units). The OTUs are groups of reads that are very similar to each other. Typically 3% identity between sequences is used to define OTUs. This radius is a matter of convention, and may give different results for different subregions of 16S rDNA, or may correspond to different phylogenetic distance for different lineages. After clustering, the OTUs can be used for taxonomic identification (by looking up reads in a taxonomic database), and the number of reads in an OTU is used as a proxy for the relative abundance of the taxon in the community. The OTU assignments can be used to estimate the species richness (number of species in a community), to compare communities through Bray-Curtis ordination [113], create phylogenetic trees and generate OTU coexistence networks [114]. Importantly, the distribution of OTU abundances can be used as a proxy for a distribution of species abundances, an important ecological metric. We now turn our attention to how species abundance distributions are used in ecology.

5.2 Ecology

Ecology aims to explain the relationship between living organisms and the environment they live in. In particular, ecologists are interested in living organisms' abundance, diversity, spatial distribution and energy use. Of course, these questions are important in order to understand humans' impact on the environment and biodiversity. However, the factors governing how living organisms are distributed on large scales are still not fully understood, with conceptual gaps between our understanding of population dynamics, gene flow and energy flow.

An important question in ecology is the distribution of abundances of species in an ecosystem, and the total number of species, the species richness. What decides how abundant a species will be? Are the modal (highly abundant) species more abundant because of their better adaptation, or because of chance? What is the role in the ecosystem of the low abundance organisms? Ecologists have been trying to answer these questions since the 1940's [115]. Functional forms for species abundance, based on statistical assumptions rather than time-dependent mechanisms of community development, have been proposed. These include Fisher's log series and Preston's log normal, niche apportionment [116] or self-organized criticality models [117] have been used to fit the data. Alternative models for abundance distributions are mechanistic, ones that conjecture some sort of a meaningful process for evolution and assembly of species. The simplest of all models in the latter category is the ecosystem counterpart to the ideal gas, due to Hubbell, and we begin by describing this model.

5.2.1 Neutral theory

In recent years, Hubbell's Neutral model [118] has become popular as a null hypothesis for testing observed abundance and diversity patterns. The key assumption (and source of controversy) behind Hubbell's theory is the idea of functional equivalence [119]. Namely, all species are assumed to be the same for all purposes (i.e. in fitness, per capita death, growth and speciation rates). Heuristically, this is motivated by the notion that living species are all near optimally fit, within the range of fluctuations compatible with the selection pressure.

Hubbell's theory has two parts. The first part models the *metacommunity* dynamics. A metacommunity is a large ecosystem, considered over a timescale in which speciations can occur. The second part of Hubbell's theory models the smaller local community (e.g. an island) which is in contact with the metacommunity (e.g. a continent). The local community is modeled over a short enough timescale that speciation cannot occur, but migration from the metacommunity can take place. Notice that in its original form, Hubbell's local community model is not spatially extended (it only involves interaction between mean-field local community and mean-field metacommunity).

The metacommunity undergoes a dynamics that is a competition between the introduction of new sequences, and the zero-sum expansion of already abundant organisms. The expansion is zero-sum since that's the simplest way to limit the amount of available resources in the model. However, this assumption is not necessary [120]; what is important are the fluctuations in relative abundance of organisms. More abundant organisms in the neutral model tend to become more abundant. With pure speciation (essentially diffusion in genome space), the abundance patterns would tend to be roughly uniform with Gaussian fluctuations. Instead, the rank abundance patterns in Neutral Theory show an exponential tail for rare organisms, and significantly higher abundances for some modal organisms.

The dynamical rules in the metacommunity are simple. Consider a population of J organisms. Each organism $i \in 1, \ldots, J$ belongs to a species S_i, where S_i is a label for a certain species (e.g. an integer). At each time step with probability $1 - \nu$, we pick one organism i at random, and replace it with an organism of species S_j from the population, effectively setting $S_i = S_j$. Alternatively with probability ν, we pick one organism i at random and replace it with an organism of an entirely new species by setting S_i equal to a new label (e.g. an integer we haven't used before). Here ν is the speciation rate per generation. The metacommunity model, as defined here is a statistical process with detailed balance. The number of species with n organisms remains constant in the equilibrium since the probability of a reaction occuring that creates a new species of n organisms is the same as the probability of a reaction occuring that removes one species of n organisms.

The relative abundance P_s of species s is then simply the proportion of organisms that belong to it, $P_s = \frac{1}{J} \sum_{i=1}^{J} \delta(S_i - s)$, where δ is the Kronecker delta. The question is, what is the steady-state distribution of P_s? Traditionally, P_s is plotted as a function of the rank R of species s. Here $R = 1$ for the most abundant species, $R = 2$ for next most abundant, etc. Evidently, when $\nu = 1$, then $P_s = 1/J$ for each existing species s. When $\nu = 0$, then the zero-sum dynamics will proceed from initial conditions until only one species remains, with $P_s = 1$. When $0 < \nu < 1$, Hubbell found that the P_s have the form shown in Figure 6.4, which is asymptotically similar to the Fisher log-series in the $J \to \infty$ limit [118]. If one takes, in addition, Hubbell's local-community model then one can also obtain another empirical law, Preston's log-normal distribution. Analytic exact solutions and asymptotic forms are known for both the metacommunity and the local community models [121, 122, 123, 124].

Chapter 6

Robust computational analysis of rRNA hypervariable tag datasets

Next-generation DNA sequencing is increasingly being utilized to probe microbial communities, such as gastrointestinal microbiomes, where it is important to be able to quantify measures of abundance and diversity. The fragmented nature of the 16S rRNA datasets obtained, coupled with their unprecedented size, has led to the recognition that the results of such analyses are potentially contaminated by a variety of artifacts, both experimental and computational. Here we quantify how multiple alignment and clustering errors contribute to overestimates of abundance and diversity, reflected by incorrect OTU assignment, corrupted phylogenies, inaccurate species diversity estimators, and rank abundance distribution functions. We show that straightforward procedural optimizations, combining preexisting tools, are effective in handling large ($10^5 - 10^6$) 16S rRNA datasets, and we describe metrics to measure the effectiveness and quality of the estimators obtained. We introduce two metrics to ascertain the quality of clustering of pyrosequenced rRNA data, and show that complete linkage clustering greatly outperforms other widely-used methods.

6.1 Summary

Microbes constitute the majority of the living mass and genetic diversity on the Earth. They construct communities with elaborate patterns of abundance and diversity. Major current scientific interest is in quantifying these patterns in various microbial ecologies, from microbes living in oceans, to microbes in vertebrate intestines and on human palms. To ascertain the diversity and abundance of microorganisms in a sample, experimenters sequence organismal tags. However, it is important to carefully analyze these data sets in order not to overestimate the number of species living in an environment. This chapter describes how to make correct computational analyses of the next generation of deeply sequenced genome populations, where the unprecedented sample size and fragmented data sets pose special problems to conventional approaches. As more and more sequence data are collected, it becomes critical to have robust, reliable and fast computational methods for analyzing these data. We have developed and made available online the software TORNADO (Taxon Organization from RNA Dataset Operations) to perform such computational analysis, and to quantitatively measure its quality.

6.2 Introduction

There is a long history of using environmental 16S rRNA [125] to estimate microbial diversity [105]. While early techniques relied on using clone libraries [126], next-generation high-throughput sequencing technology, such as pyrosequencing, directly generates vast libraries of sequences [127]. Next-generation high-throughput sequencers are capable of producing large datasets of more than a million reads from a single plate [128]. As the size of these datasets grow, the ability to computationally manage and characterize such data becomes a larger and more critical component of microbial ecology.

The goal of analyzing these sequences is to quantify the diversity and abundance distributions of organisms present in the environment. As pyrosequencing technology advances, our ability to measure microbial diversity increases. Already, this technique has been used to study the diversity of microbiomes from a variety of environments [129, 130], resulting in reports of a so-called "rare biosphere" of low-abundance organisms [131].

In order to assess the microbial diversity present in any dataset, the ability to appropriately measure the distance between different sequences and to reliably group them into operational taxonomic units (OTUs) is paramount. Typically, the abundance of OTUs is plotted in a rank abundance plot. These plots have been used as a gold standard for ecological population modeling for many decades[115]. In addition, the OTU groupings are utilized by other metrics in determining relative species compositions, microbial diversity, and community comparisons[132, 133]. While much effort and controversy has been focused on measurements of the quality of next-generation sequences[131, 134, 135, 136, 137], or the interpretation of pyrosequencing flowgrams [138], less attention has been given to computational analysis of pyrosequenced 16S rRNA data after quality processing, despite the large discrepancies in OTU numbers and diversity when different analysis methods are used[139, 136, 140, 138].

There are two major components to the analysis of OTU abundance. The first is multiple alignment of 16S rRNA or fragments of 16S rRNA. The second is clustering the sequences based on a distance metric. The purpose of this chapter is to provide a careful discussion of the computational analysis of alignment and clustering of pyrosequencing datasets, identifying sources of error, and appropriate ways to handle the data to mitigate these artifacts. In particular, we use the Calinski-Harabasz (CH) index [141] to compare the quality of clustering. We find unexpectedly large differences in the performance of different algorithms. We show that data analysis is surprisingly sensitive to even small errors in multiple alignment and clustering, but that with relatively little difficulty, these artifacts can be substantially mitigated using a judicious combination of preexisting tools, and others that we have made available on a web site (http://tornado.igb.uiuc.edu). Following these procedures results in robust characterization of microbial ecosystems.

6.2.1 Multiple Alignment: NAST and Infernal

Multiple alignment is the starting point of almost all analyses performed on microbiome sequences. Most phylogeny [142, 143], community distance estimates [144, 133], and abundance distributions [145, 146, 147] ultimately rely on input from a multiple alignment to compute sequence distances within a consistent alignment template.

The goal of multiple alignment is to align sequences according to their evolutionary relationships. In order for a multiple alignment to be meaningful in this context, all sequences in the multiple alignment must have a common origin. The various match, mismatch, and indel events then represent possible reconstructions of the evolution of those related sequences. In contrast to pairwise alignment, multiple alignment leverages conserved features of an entire gene family to obtain a broader evolutionary picture. This picture can then be fed into various algorithms such as maximum-likelihood phylogeny[142, 148, 143] in order to reconstruct the evolutionary relationships between the individual sequences.

The use of 16S rRNA sequences for discerning evolutionary relationships has a long history. The very first studies that organized the Bacteria according to their evolutionary relationships and resulted in the discovery of the Archaea utilized this important ribosomal molecule as a molecular fossil [149] and it still remains the most widely used evolutionary marker in microbial ecology today [131, 150, 128]. As such, it is not surprising that a number of tools exist which are specifically tailored to 16S rRNA such as the NAST pipeline [151] or Ribosomal Database Project [152].

These specialized 16S rRNA alignment tools all incorporate information about the 16S rRNA secondary structure. The importance of the secondary structure is two-fold. First, the conservation of 16S rRNA sequences stems from the conserved structure. Second, unlike proteins which are built from up to 20 different amino acids, there are only 4 basic RNA bases. Randomly chosen RNA bases have a greater chance of aligning well with one another than randomly chosen protein sequences, making it more difficult to distinguish between evolutionary relationships and random matches. Secondary structure can, and should, be used to provide extra discriminatory power beyond that available from the one-dimensional sequence alone.

The NAST algorithm [151] tries to align new sequences against a precalculated multiple alignment template, and has been integrated into many commonly used 16S rRNA analysis tools such as Mothur [133] and

GreenGenes[153]. Typically, this template is hand-curated to include the appropriate secondary structure considerations. In this chapter we will use the SILVA SEED SSURef database version 102 [154] as the template and refer to this alignment method as NAST+SILVA. The weaknesses of this method are that errors in the hand-curated multiple alignment propagate and that alignment against a fixed-size template necessitates the inclusion of purposeful misalignments. Overall, this results in alignments that are sometimes inconsistent with alignments based on secondary structure. An example of this is shown in the alignments in Figure 6.1b. By contrast, Infernal [155], which has been integrated into the Ribosomal Database Project 16S rRNA Pipeline [152], aligns sequences against a predefined structure. However, even among the well-conserved structures of 16S rRNA, there exist hypervariable regions which vary in their secondary structure from taxon to taxon. These regions cannot be aligned to a fixed structural template, and are left unaligned by Infernal (leading to a multiple alignment whose length is not fixed but may be different in different datasets). An example of this Infernal's alignment is shown in Figure 6.1a. It is important to note that while both methods align to a seed model of some sort, the practical difference is that RDP+Infernal does better in regions of strong secondary structure whereas NAST+SILVA does better in hypervariable regions.

To exploit this distinction, one can merge the best alignments from each tool by combining the hypervariable regions aligned using the NAST algorithm with the regions of strong secondary structure aligned by Infernal. An example of sequences aligned using the merging method is given in Figure 6.1c. One can also make adjustment to the multiple alignment by hand. Done properly, this can produce a better quality alignment than automated methods alone. An example of the merged alignment in which the hypervariable region was further hand-curated is given in Figure 6.1d. For a brief description of the process and the tools that we developed to perform the merging and hand-curation, please refer to the Materials and Methods section.

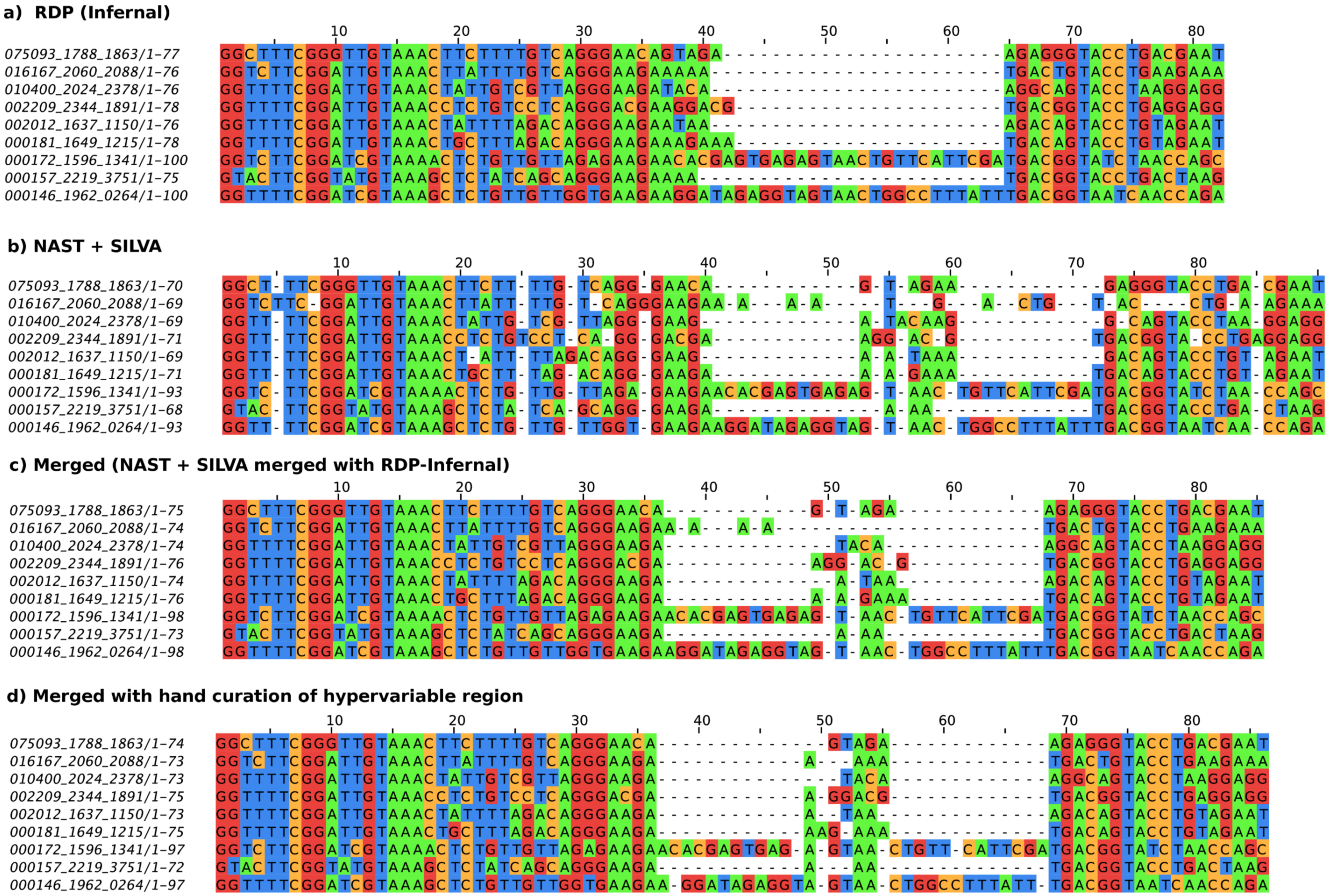


Figure 6.1: Snippets of 9 sequences aligned using the 4 different methods described in this chapter. The 9 sequences are in the V3 region of the 16S rRNA. **(a)** Sequences aligned via RDP [152] which uses the Infernal aligner [155]. Note that the hypervariable region is left unaligned (bases 36 through 64). **(b)** Sequences aligned via NAST [151] (as implemented by Mothur [133]) to the SILVA [154] database. Notice the inconsistencies in the alignment of the regions with strong secondary structure conservation (bases 5, 25, 29, 72 through 79, and 85). **(c)** Sequences aligned using the merge program in the tool we developed, TORNADO (`http://tornado.igb.uiuc.edu`). The merge process takes the unaligned, hypervariable parts of the sequence aligned by (a) and replaces them by the alignment in (b). **(d)** Sequences aligned like in (c), but with the final hand-curation step of the hypervariable regions.

With the availability of a variety of tools that perform multiple sequence alignment, it is imperative to have a way to assess its quality. One way to do that is through maximum-likelihood (ML) phylogeny. ML phylogeny tries to identify the set of relationships with best likelihood value. Conversely, ML scores can also be used to judge the likelihood of a multiple alignment reflecting sequence evolution. Indeed, similar measures have been used in the past [156] and tools such as SATÉ [157] already take advantage of this measure when iterating between multiple alignments and phylogeny to automate the search for the best alignment and tree. However, exploring enough multiple alignments for large datasets is prohibitively expensive and therefore remains impractical for now. Nonetheless, it is feasible to use the ML method in order to compare the quality of alignments by measuring their likelihood values, and we use this below to compare different alignment strategies.

6.2.2 Clustering Algorithms

Clustering algorithms, such as complete linkage [158], are essential for quantifying the diversity of microbial communities. The goal of clustering is to group sequences that are within some measure of evolutionary distance. Distances can be calculated using many different metrics such as percent sequence identity (PSI) or distance along the phylogenetic tree branches. Ideally, a clustering algorithm should identify the natural boundaries between the clusters without utilizing more clusters than necessary to account for the entire dataset. Ultimately, clustering should accurately reflect the underlying phylogenetic and taxonomic distribution of sequences.

Complete linkage clustering (also known as furthest neighbor) has become the most widely-used clustering algorithm in microbial ecology. It relies on input from a distance matrix that can be generated from the pairwise distances between sequences in a multiple alignment. When calculating sequence distances, it is important to clearly note how alignment gaps are dealt with. One can ignore the gaps (like Phylip DNADIST does [147]) or count them in a number of different ways [145]. Once pairwise distances are obtained, complete linkage operates by progressively merging smaller clusters into larger ones, as long as each element in a cluster is within a defined radius from other elements in the cluster [158]. Alternative linkage algorithms are also possible. One can merge clusters into larger ones as long as the average distance within a cluster is within a defined radius (this is average neighbor clustering, or average-linkage). Finally, one can use single linkage (or nearest neighbor), where one merges clusters as long as they share 1 sequence that is within a defined radius. Single linkage generally performs very badly as it creates very large, “snake-like” clusters.

The performance of linkage clustering algorithms, for example as implemented in Mothur [133], scales poorly (N^3, where N is the number of sequence reads) as the number of sequence reads generated per study increases. In the studies reported below, we reimplemented Mothur's clustering algorithm, achieving an improvement in computational complexity (scaling as $N^2 \log N$), better memory usage, and an overall speedup that is typically a factor of 5-10, leading to the ability to handle datasets with N up to about 30,000. Despite these improvements, it is understandable that heuristic, computationally efficient algorithms have been developed, such as FastGroup [159] and ESPRIT[160].

FastGroup does not order clustering in any particular way, but instead chooses a sequence at random, grouping everything within a defined PSI distance of that sequence. As an example of how that is different from the complete linkage clustering employed by Mothur, consider the clustering of a scatter of points in two dimensions, as shown in Figure 6.2. The two-dimensional space is a very simple example of sequence space, with position in the space corresponding to the particular sequence of an organism. A set of points in this space, if sufficiently close to one another, represents a set of sequences that can be considered to be grouped into a single equivalence class—in other words, an OTU. The largest allowable distance between points in a single equivalence class corresponds to the sequence similarity required for sequences to be included in the same OTU (typically 97% is used).

When the FastGroup algorithm is used to group these sequences, with a radius equal to the radius of the circle of points, the number of clustered OTUs can vary, depending on the order of chosen cluster centers. One example of FastGroup's clustering is given in Figure 6.2a. On the other hand, complete linkage clustering with the same diameter correctly identifies the existence of 1 cluster (Figure 6.2b), by progressively merging clusters as long as they are within a cluster diameter (see Figure 6.3 for the progress of the complete linkage algorithm).

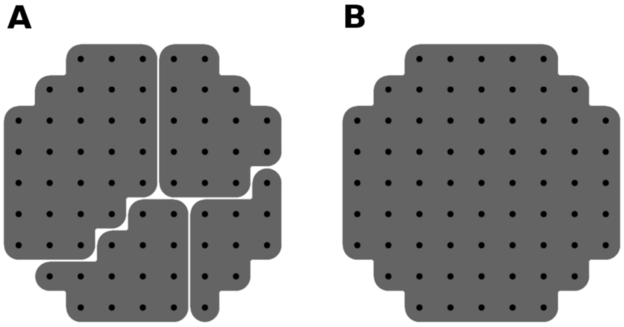


Figure 6.2: Calculation of clustering a set of points in a plane. (a) FastGroup's method. (b) complete linkage clustering. Both of these clusterings are performed with the same radius r equal to the radius of the set of points. FastGroup constructs 4 clusters whereas complete linkage finds 1.

ESPRIT[160] goes one step further and does away with multiple alignment entirely and processes the clusters in two steps: the first relying on a k-mer heuristic, similar to that used in BLAST[161], in order to

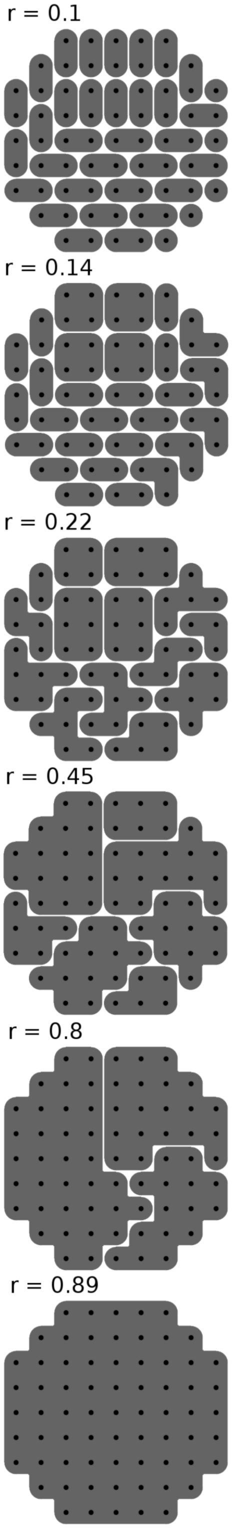


Figure 6.3: Illustration of the process of the complete linkage algorithm. Smaller clusters are progressively merged into larger ones as long as no two elements of a cluster are farther than r from each other.

group closely related sequences under one representative sequence; the second relying on pairwise distances between representatives in order to determine the final clusters. Both FastGroup and ESPRIT differ from the more controlled calculations of the complete linkage algorithm, but at the same time promise less computationally intensive results. Before pursuing such alternatives, it is important to understand the differences between the results produced by each of these algorithms.

In other words, do the heuristic algorithms produce natural cluster borders and correct cluster compositions? A natural cluster should have a representative sequence that is near the center of the cluster, i.e. the representative sequence should be one that shares the most similarity to all other sequences in the cluster. Natural clusters should not partition the dataset into more groups than necessary. One way to quantify this goodness of clustering is via the Calinski-Harabasz (CH) index [141, 162] that has been found to be the best in a comprehensive study of 30 different clustering quality indices [163]. CH is defined in the following way:

$$CH = \frac{n-k}{k-1} \frac{\sum_{K,L}(\bar{x}_K - \bar{x}_L)}{\sum_K \sum_{i,j \in K}(x_i - x_j)},$$

where n is the number of elements $(x_1, \ldots, x_n)$, k is the number of clusters (with means $\bar{x}_1, \ldots, \bar{x}_k$), K and L label clusters, i and j label elements within clusters and subtraction of the x indicates distance according to a specified metric. Qualitatively, CH is larger when clusters are compact (denominator becomes smaller), or when clusters are farther away from each other (numerator becomes larger). The CH index is also correctly normalized so as to be comparable for different number of OTUs. The implementation of the program that calculates the index is available at http://tornado.igb.uiuc.edu/.

6.3 Results

In this work, we demonstrate that different methodologies can lead to very different estimates of OTU abundances. We characterize these differences and deconstruct their two primary sources: multiple alignment and the clustering method used. We measure the performance of both components of this process, restricting ourselves to 16S rRNA based techniques. We also provide metrics to quantitatively evaluate the effectiveness of algorithms used. Our analysis includes an examination of the robustness of these algorithms on real biological data. We perform our analysis on a dataset of 22,911 bacterial 16S rRNA sequences (V3 region) with an average length of 205bp from a sample of a chicken caecum. We note however, that our methodology is also applicable to longer 16S rRNA reads.

Before further analysis, we treated our dataset in the following way. To handle length variation among sequences, we trimmed our sequences to only be between the first and last conserved columns in the NAST [151] alignment to SILVA database [154]. We further removed any sequences less than 100bp long and any sequences that contained an unknown nucleotide (N). After cleanup our dataset had 21,646 sequences.

6.3.1 Multiple Alignment: Performance

We compared the effectiveness of the different alignment algorithms by using the likelihood values returned by maximum-likelihood phylogeny. In alignment of nucleotide sequences with secondary structure the aligners that are aware of the secondary structure generally outperform those that rely on sequence data alone [164], such as ClustalW [165] and MUSCLE [166]. In addition, these aligners scale poorly with dataset size [167]. Thus, we test the two commonly used 16S rRNA alignment algorithms: RDP [152]+Infernal and NAST in conjunction with the SILVA database [154]. Using ML phylogeny, we find log-likelihood scores of $-17,012$ and $-17,322$ for RDP+Infernal and NAST+SILVA, respectively, as obtained from the FastTree ML algorithm [142]. The merged alignment of RDP+Infernal with NAST+SILVA, described in the introduction, has a log-likelihood value of $-16,262$, representing an improvement over either of the two algorithms alone. When we perform further hand-curation of hypervariable regions of the 16S V3 in the merged alignment, we obtain a log-likelihood score of $-15,036$—reflecting the misalignments that can occur in the other automated procedure.

We can also take these different multiple alignments and cluster them in order to see how the OTU abundance results depend on the multiple alignment procedure. The OTU numbers after complete linkage clustering with radii 3%, 5% and 7% on seven different alignments are shown in Table 6.1. Here, "merge" refers to the merging of RDP+Infernal with NAST+SILVA. Note that running the aligners on sequences after quality processing produces thousands of OTUs. However, performing hand-trimming of sequence tails reduces the number of OTUs by an order of magnitude. This suggests that poorly curated alignments may overestimate microbial diversity.

6.3.2 Clustering: Performance

We compared the three clustering algorithms (complete linkage, FastGroup and ESPRIT) by running them on the hand curated alignment described in the previous section. We can visualize the effect of the choice of clustering algorithm by comparing rank abundance curves and cluster compositions. Rank abundance curves for the chicken caecum dataset are compared in Figure 6.4, for the 3% sequence difference clustering distance (and 1.5% FastGroup). As demonstrated by the curve, complete linkage clustering, ESPRIT and FastGroup 1.5% obtain the same shape of the curve, but FastGroup with 3% finds a very different one. This is because complete linkage at distance r corresponds to clusters where every element is at distance r to

Alignment method	**Number of OTUs**		
	3%	5%	7%
NAST+SILVA (on raw)	1141	646	406
RDP+Infernal (on raw)	3588	2313	1743
Merged (on raw)	3647	2297	1682
NAST+SILVA (on trimmed)	425	251	187
RDP+Infernal (on trimmed)	406	234	169
Merged (on trimmed)	393	227	165
Hand-curated	354	207	153

Table 6.1: Dependence of the number of OTUs on the alignment method used. The percentages indicate clustering radius. Trimmed sequences refer to sequences in which elementary hand-curation was performed (see introductory paragraphs of Results for more information). Merged refers to the multiple alignment that is a merging of the hypervariable regions aligned by NAST+SILVA regions with strong secondary structure conservation aligned by RDP+Infernal. See Introduction for more information. Note that crude hand-curation can reduce numbers of OTUs by a whole order of magnitude.

every other element in the cluster. On the other hand, FastGroup guarantees that every element is only at distance r from the chosen center of the cluster. This means that there may be elements in the same cluster that are at a distance of $2r$ from each other. Hence, r for FastGroup denotes the “radius” of the cluster, whereas r for complete linkage denotes the “diameter” of the cluster. Thus, FastGroup at 1.5% sequence distance can be compared to complete linkage and ESPRIT at 3%.

We find that FastGroup at 1.5% overestimates the number of OTUs in the sample. The binning in Figure 6.4 hides the fact that the number of OTUs found by FastGroup 1.5% is much larger than that of ESPRIT and complete linkage. FastGroup 1.5% finds 834 OTUs compared to complete linkage (354) and ESPRIT (434). Most of these extra OTUs are singletons. Of the 834 FastGroup OTUs, 440 are singleton OTUs. In comparison, complete linkage has 103 OTUs that are singletons out of total of 354. ESPRIT has 122 OTUs that are singletons out of 434 total. This is in accordance to the idea that is sketched in Figure 6.2, that a clustering algorithm such as FastGroup overestimates the number of OTUs.

We now evaluate the clustering quality via the CH index. For 3% clustering distance, complete linkage has a CH index of 167,771, whereas ESPRIT clustering has a CH index of 244. FastGroup with 1.5% clustering radius has CH index of 94,696. We note that complete linkage significantly outperforms other linkage clustering algorithms: nearest neighbor linkage (single linkage) got a CH index of 14,042 and average neighbor linkage got 23,512. We can also compare CH indices for clustering assignments that have roughly the same number of OTUs, rather than the same clustering distance. We find that complete linkage has CH indices between 140,000 and 160,000 for a range of clustering assignments with 200 to 300 OTUs. ESPRIT produced two clustering assignments in this range: first with 235 OTUs has a CH score of 280, and second with 303 OTUs has a CH score of 286. Finally, FastGroup (with 3% distance) got a CH score of 16,000 for

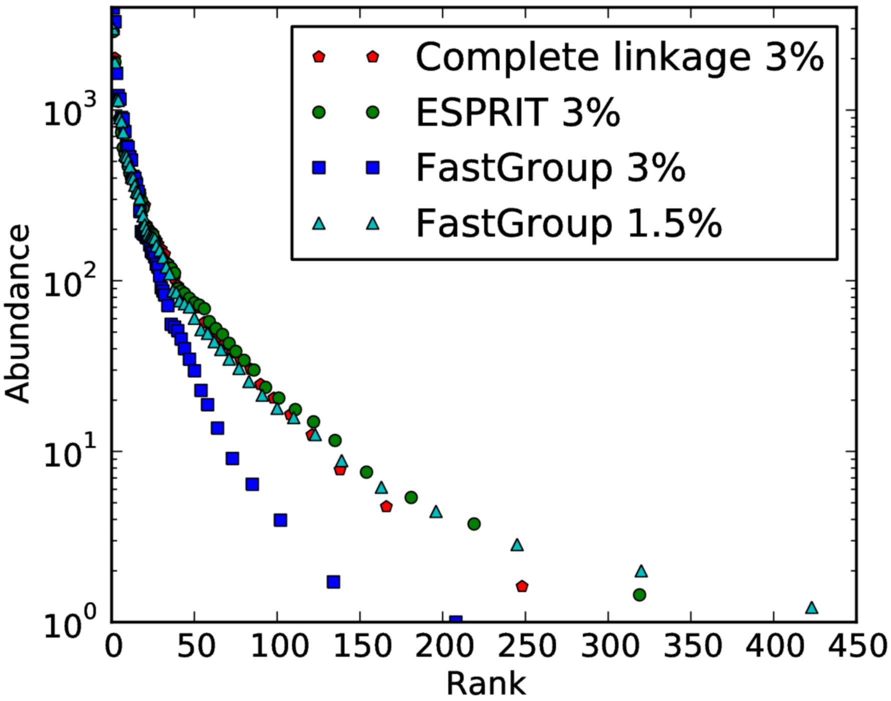


Figure 6.4: Rank abundance curves obtained with different algorithms and/or clustering distances. Notice that FastGroup with 1.5% sequence distance identifies a similar rank abundance curve to those of ESPRIT and complete linkage. However, it is not evident from the Figure that FastGroup identifies almost two times the number of OTUs than ESPRIT or complete linkage.

a clustering assignment with 251 OTUs.

Another way to quantitatively judge the goodness of clustering is by comparing the OTU assignments to the structure of the maximum likelihood phylogenetic tree. To do this, we count the number of clades in a phylogenetic tree that contain only sequences of the same OTU (as determined by the clustering algorithm). We expect that a good clustering assignment will have many such clades. Two examples of this calculation are sketched out in Figure 6.5. We ran this calculation on 2 phylogenetic trees, one made by FastTree [142] (FT) and one made by RAxML [143] (RX), both inferred from our dataset described above. We find that complete linkage clustering has the most clades with uniform OTUs: 863 in FT and 698 in RX. Clustering with FastGroup (with 1.5% distance), we find 427 clades in FT and 367 in RX, whereas ESPRIT performs the most poorly: 6 clades in FT and 7 in RX.

We also explored if the rank abundance curves depend upon the clustering distance metric used. We find that the complete linkage clustering with the hand curated multiple alignment is very robust with respect to the choice of distance metric. In Figure 6.6a we compare the rank abundance curves (made by complete

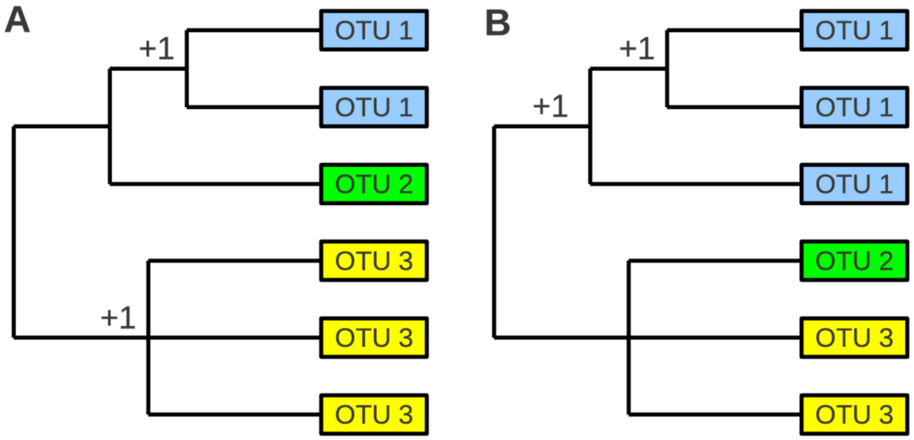


Figure 6.5: Sketch of the calculation of the number of clades with uniform OTUs. A phylogenetic tree with 2 different cluster (OTU) assignments is shown. The cluster assignment is indicated by OTU number and color. Both cluster assignments have 2 uniform clades (interior nodes indicated by +1). **(a)** The uniform clades are: one made up of two OTU 1 organisms, and one made up of three OTU 3 organisms. **(b)** The uniform clades are: one made up of two OTU 1 organisms and one made up of three OTU 1 organisms.

linkage) for three different distance metrics: Phylip DNADIST[147], percent sequence identity and distance along phylogenetic tree constructed by FastTree. We see that regardless of the choice of the distance metric, the shape of the rank abundance curve is conserved.

If we seek universal laws in the rank abundance data, we should expect that the shape of the rank abundance curve does not depend upon the particular clustering radius chosen. If instead rank abundance changes significantly with radius, that would imply that there is an interesting interplay between population dynamics and sequence distance. The complete linkage clustering with our hand curated multiple alignment is found to be robust with respect to choice of clustering radius. As an example, see Figure 6.6b for the rank abundance curves of our chicken caecum microbial sample clustered at three different distances. By rescaling the axis of the rank abundance curves, while keeping areas under them constant, we can compare the functional forms (i.e. shapes) of the rank abundance curves. The figure shows that the chicken caecum microbial sample rank abundance seems to obey a universal law over a range of clustering distances.

6.4 Discussion of the Results

In the literature, the quality of data from pyrosequencing has been called into question [139, 168], especially with regard to its use in surveys of OTU diversity. Concern has been directed mostly at the experimental process of acquiring DNA sequences with high quality. Sogin *et al.* [131] showed that a number of heuristics can guarantee that per-base error rate of pyrosequencing is lower than that of Sanger sequencing

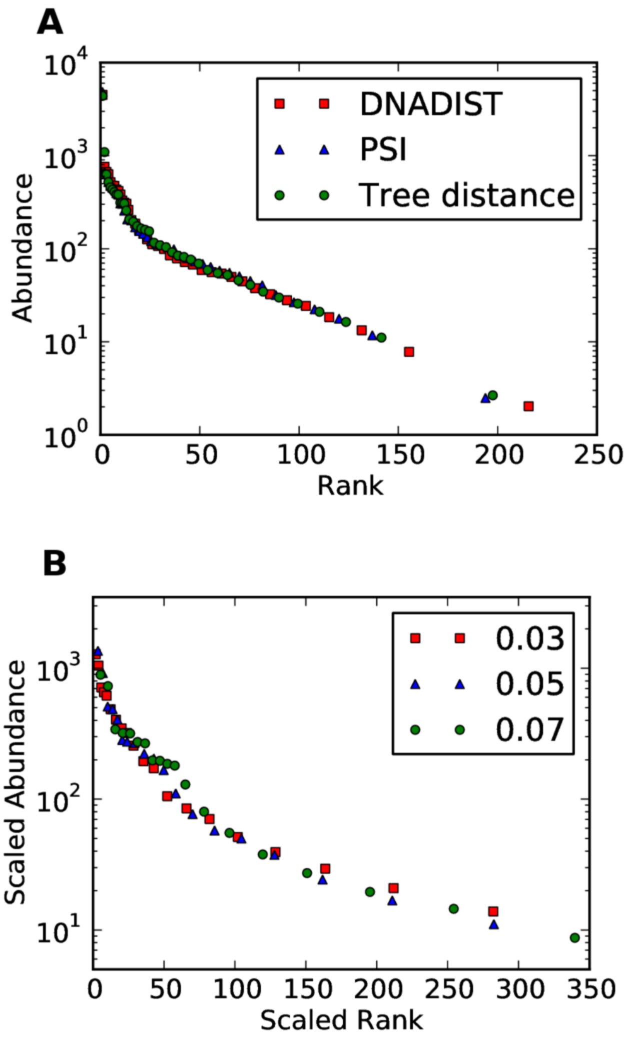


Figure 6.6: Two checks that should be used to verify quality of rank abundance curves. Both plots show rank abundance curves of the chicken caecum dataset. (a) Comparison of rank abundance curves for three clusterings using three different distance metrics. We compare the clusterings that produce 300 OTUs (which corresponds to different radii r for different metrics). (b) Rank abundance curve is robust if it does not change shape (functional form) when a different clustering radius is used. The rank abundance curves for different clustering radii all fall onto the same curve after rescaling the ranks to the same number of OTUs (while keeping area under the curve constant).

while retaining more than 90% of data. Other artifacts that raised concern came from the shortness of pyrosequenced reads [134, 135]. Quince *et al.* [138] showed that reinterpreting pyrosequencing flowgrams via a maximum-likelihood scheme can lead to fewer OTUs. In this chapter we showed that a significant part of the discrepancy may arise from different computational analyses employed. Recent work [140] that has been similarly motivated has been commensurate with the conclusion that clustering is an important step in OTU analysis. In particular, they suggest that a preclustering step can help fix problems where deep sequencing overestimates species richness. Our work presents more general quantitative metrics that can be used as a standard for clustering programs. In addition, we find that calculating the log-likelihood of a maximum-likelihood phylogenetic tree is a good way to compare the quality of nucleotide alignments. Clustering quality index such as Colinski-Harabasz can even be used to verify what clustering radius is appropriate for a particular dataset. Our results that the multiple alignment and distance metrics can have a large effect on OTU abundances are also in agreement with recent work by Schloss [169].

In general, we found that multiple alignments can have a large influence on OTU abundance information, and the automated 16S rRNA alignment tools should be subjected to hand curation. Fast clustering tools such as ESPRIT do not make use of a multiple alignment and rely on k-mer heuristics to calculate pairwise distances between ungapped sequences. Our results show that such tools, intended to improve upon complete linkage, actually perform significantly worse. Hence, even with increasing dataset sizes, it is important to verify that the clustering method used performs no worse than complete linkage. We developed tools that ease the burden of performing hand curation and complete linkage of large contemporary datasets. These are available as supplementary software and are described in more detail in Materials and Methods.

6.4.1 Discussion of the Recommended Analysis Pipeline

In this section, we summarize for the reader's convenience, step-by-step recommendations for handling a large 16S rRNA dataset, based on the analyses we have reported here. These are graphically illustrated in Figure 6.7.

1. *Quality Processing:* Remove short reads and sequences with unknown nucleotides (N). Make an alignment to the SILVA database [154], via NAST [151] as implemented by Mothur [133]. Trim sequences to be between the first and last strongly conserved columns in this alignment.

2. *Alignment:* From the trimmed dataset, produce another alignment through RDP pipeline's [152] front end to the Infernal aligner [155]. Merge the two alignments (NAST+SILVA with RDP+Infernal) using the tool that's a part of the TORNADO pipeline at `http://tornado.igb.uiuc.edu`. Further hand-optimize hypervariable regions of the reads by using the tool available on the website above.

3. *Cluster:* Cluster the dataset using the complete linkage tool available on the website above.

Further analysis can be performed by calculating estimators in Mothur [133], or by estimating phylogenetic trees via RAxML [143] or FastTree [142].

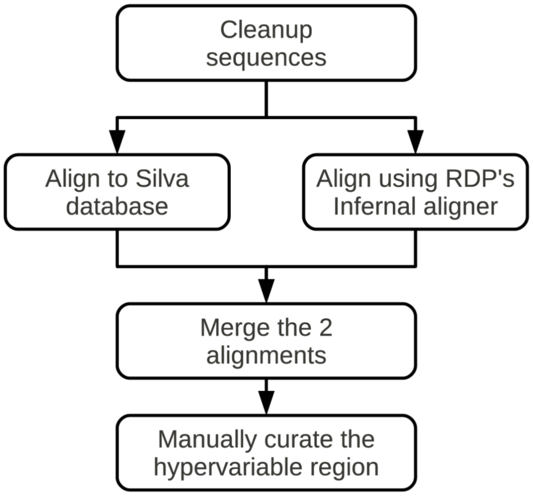

Cleanup sequences
Align to Silva database
Align using RDP's Infernal aligner
Merge the 2 alignments
Manually curate the hypervariable region

Figure 6.7: Diagram of our proposed 16S rRNA alignment pipeline, TORNADO. After the preliminary clean up step, we align the sequences in two different ways. First, we use Mothur[133] to align our sequences to the SILVA[154] database. Second, we align using Ribosomal Database Project's front end[152] to the Infernal aligner[155]. We then merge the two, using Infernal's secondary-structure-aware alignments and SILVA's alignment of hypervariable region. Finally, we manually curate the hypervariable regions, using a helper tool, `splicer`, we developed (see Fig. 6.8).

6.5 Materials and Methods

6.5.1 V3 rRNA amplicon sequencing.

We used the V3 rRNA sequences from the chicken caecum from batch B of a previous study [170]. PCR specific primers flanking the V3 hypervariable region of bacterial 16S rRNA were used to generate PCR products for pyrosequence analysis. The forward fusion primers for pyrosequencing included 454 Life Sciences A adapter, and barcode *A* fused to the 5′ end of the V3 primer 341F (5′ gcctccctcgcgccatcag-ACGAGTGCGT -CCTACGGAGGCAGCAG3′) or with barcode *B* (5′ gcctccctcgcgccatcag-ACGCTCGACA-CCTACGGA-GGCAGCAG3′). The reverse fusion primer included 454 Life Sciences B adapter fused to 5′ end of V3 primer 534R (5′ gccttgccagcccgctcag-ATTACCGCGGCTGCTGG3′). Cycling conditions (20 cycles) were; initial denaturation at 94 °C for 5 min; 20 cycles of 94 °C 30 s, 60 °C 30 s and 72 °C 30 s; then 72 °C 7 min for final extension. The amplicon products were cleaned using PCR purification clean-up kit and SPRI size exclusion beads. The quality of products was assessed using a Bioanalyzer using DNA1000 chip. The fragments in amplicon libraries were subjected to a single pyrosequence run using a 454 Life Science Genome Sequencer GS FLX (Roy J. Carver Biotechnology Center, University of Illinois). The resulting dataset had 22953 sequences of average length 204.7bp. Before further analysis was performed, we performed basic filtering. We removed all sequences that were shorter than 100bp reducing the number of sequences to 21646. The sequences have been uploaded to GenBank (accession numbers HQ293272-HQ315544).

6.5.2 Multiple Alignments

We compared 4 different alignment methods as illustrated in Figure 6.1. (1) We fed the sequences into Infernal [155] with bacterial secondary structure template as provided by RDP [152]. (2) We aligned the sequences to the SILVA database [154] using the NAST[151] algorithm as implemented by Mothur [133, 171] (align.seqs command). (3) The results of (1) and (2) were then fed into a merger script we have made available on the Web at `http://tornado.igb.uiuc.edu/`. (4) The merged data sets' hypervariable regions were then hand curated using `splicer`, a tool we developed and made available on the Web as part of our pipeline TORNADO at `http://tornado.igb.uiuc.edu/`. This tool allowed us to greatly reduce the number of unique snippets of the hypervariable region of V3 down from 21,646 to about 200, by cutting the longest hypervariable subregion from the alignment, and then dereplicating it. These snippets of sequences in the hypervariable subregion ranged from 1bp to about 30bp. This meant that we only needed to hand-curate 200 short snippets to handle the alignment of the hypervariable region. These snippets were separated into two groups according to their secondary structure: loop, and stem-loop-stem. We used RNAfold web server [172, 173, 174, 175] to verify the structure. The two groups were then hand curated and merged back into the complete multiple alignment using the `splicer merge` command. For clarity, the process of using `splicer` is described in Figure 6.8. All multiple alignments are available at `http://tornado.igb.uiuc.edu/`.

6.5.3 Likelihood Scores

Each data set described in the previous section was dereplicated producing 2215 clones each. Likelihood scores were then computed for each dataset using FastTree 2.1.1 with command line parameters `-gamma -nt -gtr`.

6.5.4 Distance metrics

We compared 3 different distance metrics to generate Figure 6.6a. (1) Phylip DNADIST 3.67 [147] with default model parameters. (2) Percent sequence difference calculated using a program we developed, `psi-distance`, available at `http://tornado.igb.uiuc.edu/`. The program constructs pairwise differences by calculating the number of letters that are different between every two sequence (gap is considered a letter). This number is then divided by the average of the ungapped lengths of the two sequences compared [145].

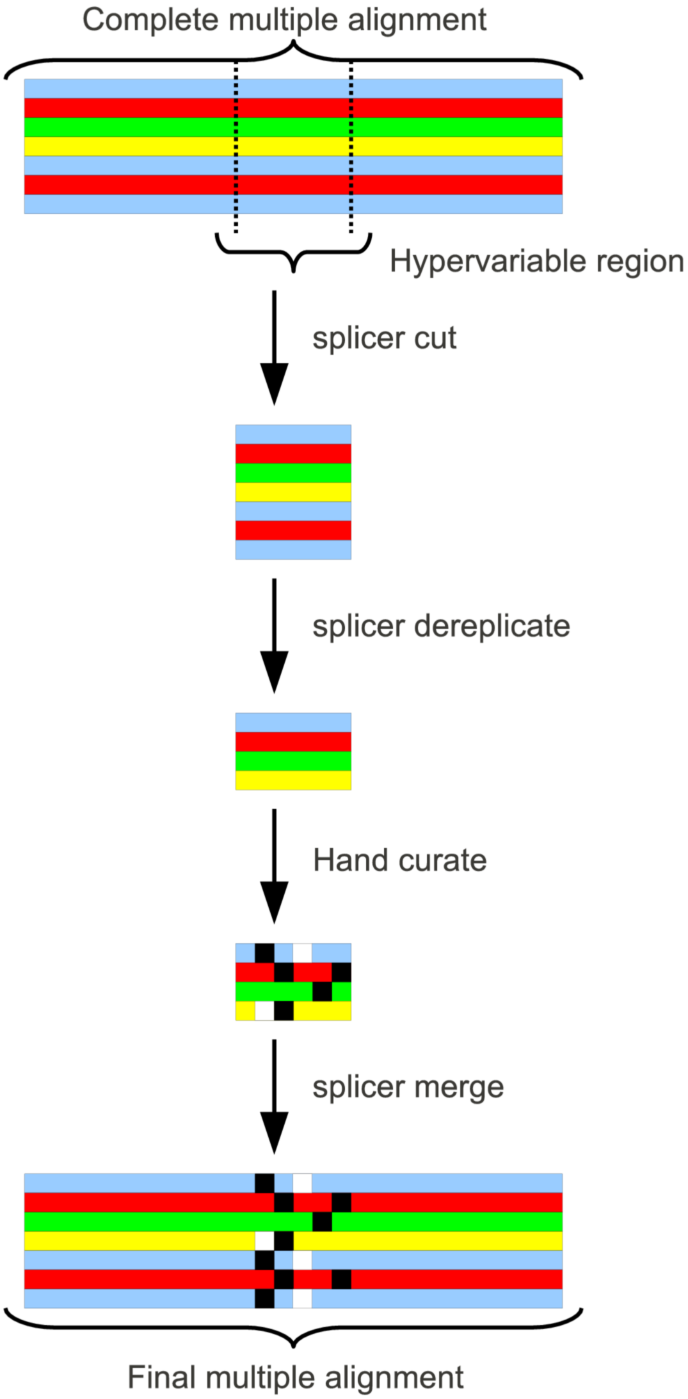


Figure 6.8: Using `splicer`, a part of the TORNADO pipeline, to perform hand curation. Dereplicating the hypervariable region significantly reduces the effective number of snippets of sequences one needs to hand curate (4 instead of 6 in this example). In our dataset of around 20,000 sequences, there were only around 200 unique sequence snippets in the hypervariable region varying in length between 1 and 30 bp.

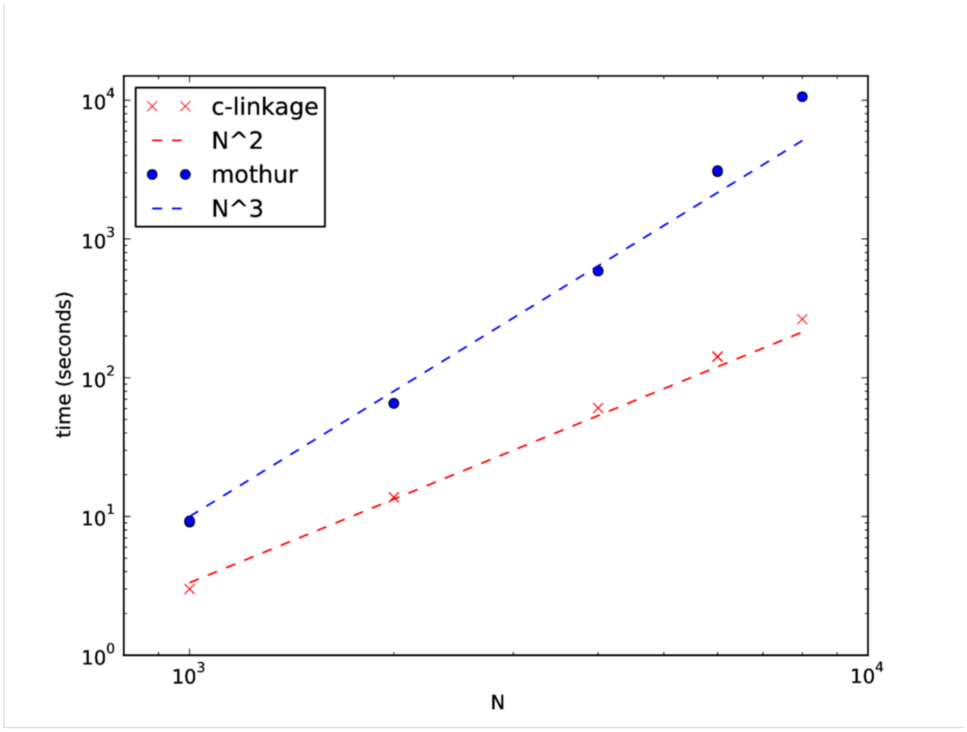


Figure 6.9: Comparison of running times of `c-linkage` with the running times of `Mothur`. The two programs were benchmarked on artificial datasets of 1000, 2000, 4000, 6000 and 8000 elements. The sripts used to generate these datasets and run the benchmarks are available at `http://tornado.igb.uiuc.edu`.

(3) Tree distance calculated from the phylogenetic tree calculated by FastTree in the previous section. The tree distances were acquired by calculating tree branch lengths from the Newick formatted tree using the `tree-distance` program we developed, available at `http://tornado.igb.uiuc.edu/`.

6.5.5 Clustering algorithms

Three different clustering algorithms were compared, all on the hand-curated dataset.

1. Complete linkage clustering with furthest neighbors, as implemented in `c-linkage`, a program we developed that is available at `http://tornado.igb.uiuc.edu/`. We tested that the program produces the same results as Mothur, but much faster and with less memory usage since it works in $O(N^2 \log N)$. For a comparison of running times of `c-linkage` and Mothur version 1.12.3, when clustering up to a clustering cutoff of 10% see Figure 6.9.

2. FastGroup[159] with no trimming, PSI difference of 97% with gaps.

3. ESPRIT[160], for which the dataset was first degapped.

Chapter 7

Fast clustering algorithms for sequence data

In Chapter 6, we described some ways in which the analysis of 16S rRNA datasets can be improved (in the sense of minimizing processing artifacts). In this Chapter we focus our attention on the clustering of sequence data. As shown in Chapter 6, a clustering is optimal when the complete-linkage algorithm is used. However, the big disadvantage of complete-linkage clustering is that it takes a long time. In Chapter 6 we referred to a $O(n^2 \log n)$ algorithm, where n is the number of sequences. The technical question that we answer in this chapter is: Can we find a heuristic algorithm that scales better with time, but keeps clustering artifacts at a minimum? As metagenomic datasets become bigger, this question will become very important and it has already attracted research interest [176]. For our own research in understanding the phase transitions in complex biological systems, clustering is a key technical component of the analysis that needs to be done correctly and efficiently before one can move on to the more interesting physical questions. But one also needs to keep in mind that cluster analysis is a first-order approach: it is possible that the structure of a dataset is such that it is not amenable to clustering (i.e. dataset does not have clear clusters).

The approach presented here has several key steps. First we use a greedy algorithm to find a local solution for the clustering. This is described in Section 7.1. Then we follow up by an optimization step that attempts to optimize the clustering, by finding a better local optimum in the neighborhood of the greedy solution. The optimization step is described in Section 7.2. We conduct some tests on how well our solution performs in Section 7.3 using one artificial and one representative data set. Then we discuss an alternative, heuristic way to cluster sequences, one that scales with fast $O(n)$ time. This algorithm is described in Section 7.4. Finally, we describe other approaches through which clustering of sequence data can be improved, and we list some open problems in Section 7.5.

7.1 Greedy modal clustering

Metagenomic amplicon reads such as 16S are generally short and degenerate. The degeneracy here means that many reads may share the same sequence. The number of degenerate reads is generally used as a proxy for sequence abundance though this assumption can introduce artifacts [177, 178]. In this chapter, we use this degeneracy to improve the greedy cluster assignment. The reads with a large degeneracy are used as priority seeds for the clusters. The clusters then have a uniform radius of size r around them.

The concrete algorithm is the following:

1. Sort the reads into an input list $\mathbb{I}$ according to their degeneracy with the most-degerate (modal) reads first.
2. Initialize an empty list $\mathbb{C}$ of clusters.
3. For each subsequent read R in the list $\mathbb{I}$:
    4. For each subsequent cluster C of $\mathbb{C}$ with seed S:
        5. Calculate sequence distance d of R to S: $d(R, S)$
        6. If $d(R, S) \leq r$ then add R to C and go back to step 3.
        7. Otherwise, go back to step 4.
    8. If no suitable cluster has been found in step 6, then create a new cluster in $\mathbb{C}$ with R as its seed.

The runtime of the above algorithm is $O(nm)$ where n is the number of reads in $\mathbb{I}$ and m is the final number of clusters. Evidently, $m \leq n$, and hence the worst-case runtime of the algorithm is $O(n^2)$. In practice, however, $m \ll n$ and the algorithm performs much faster than complete linkage (see Section 7.3). The major problem though, is that artifacts can occur due to the greediness of the algorithm, namely, the clustering solution is stuck in a local minimum due to the greediness of the algorithm. We deal with this problem in the following optimization step.

7.2 Voronoi diagram optimization

The following situation happens while running the greedy clustering algorithm. Consider a cluster C_1 with seed S_1, and a read R_2 from the input list $\mathbb{I}$. Suppose that in step 6 of the greedy algorithm, it is found that $d(R_2, S_1) < r$ and hence R_2 is added to C_1. At a later time, read R_3 is read from the input list $\mathbb{I}$. Suppose that $d(R_3, S_1) > r$ and hence R_3 nucleates a new cluster, C_3 with seed $S_3 = R_3$. In this case, an artifact is possible, since, whereas $d(R_2, S_1) < r$ and $d(R_3, S_1) > r$, and $R_2 \in C_2$ it is quite possible that $d(R_2, S_3) < d(R_2, S_1)$.

To resolve the above artifact we run an optimization step. For each non-seed read R in each cluster C, we compute the distance of R to seeds of all other clusters. If R is found to be closer to another cluster's seed, it is moved to that cluster. Evidently, this step runs in $O(nm)$ just like the greedy step in Section 7.1 and hence it doesn't add any additional computational complexity. At the end of the this optimization step, every read belongs to the cluster of the nearest seed. Hence, the reads' cluster assignment resembles a Voronoi diagram in high-dimensional space [179].

7.3 Test results

In this section, we test the modal clustering algorithm by comparing it against other common clustering algorithms, Mothur FN (furthest neighbor, i.e. complete linkage), Mothur AN (average neighbor, i.e. average linkage) and NN (nearest neighbor, i.e. single linkage) [133] as well as uclust [180]. We have previously introduced the linkage algorithms in Section 6.2.2. Uclust is an algorithm equivalent to our greedy algorithm from Section 7.1, except for the lack of step 1. Uclust does not sort the input data according to degeneracy of sample reads. Instead, it processes the sequences in whatever order they are in, and performs no optimization.

7.3.1 Artificial data

For our first test, we generated an artificial dataset of 57 different sequences. The first sequence (X) was the initial 1500bp of the Escherichia coli 16S rRNA sequence (Genbank accession J01859.1). Two lineages of variation (A and B) were then gradually evolved from X, such that each (1-28) varied by 1% from the previous sequence. Sequences from linage A (A1 A28) each switched a further 15nt (1%) from A → T, T → A, G → C and C→ G. In contrast, sequences from lineage B (B1 B28) each switched 15nt (1%) from A → C, T→ G, C→ A and G→ T. So for example, A1 and B1 were each 1% different from X but 2% different from each other and A2 and B2 were each 2% different from X, 4% different from each other and 1% different from A1 and B1 respectively. The two lineages were formed as a control, such that the clustering patterns in set A should be the same as those in set B. The abundance of each sequence in the dataset was varied such that there were 100 copies of X and every 4th sequence in each lineage had 12 fewer copies (i.e. A4 and B4 had 88 copies, A8 and B8 had 76 copies etc.). All other sequences had a randomly generated number of copies between 1 and 20.

Each tool we tested clustered identical sequences together. Results are shown in Figure 7.1. Our algorithm performed as expected, with clusters in the A series reflecting those in the B series. All OTU clusters centered on the abundant sequences (X, A4, B4, A8, B8 etc.) and did not extend beyond a 3% margin from the seed. The other tools did not perform as well. Uclust formed OTUs with sequences as divergent as 9% (up to 5% from the seed; X, A1-A5, B1-B4) despite being given a 97% threshold (command: `--id 0.97`). The NN

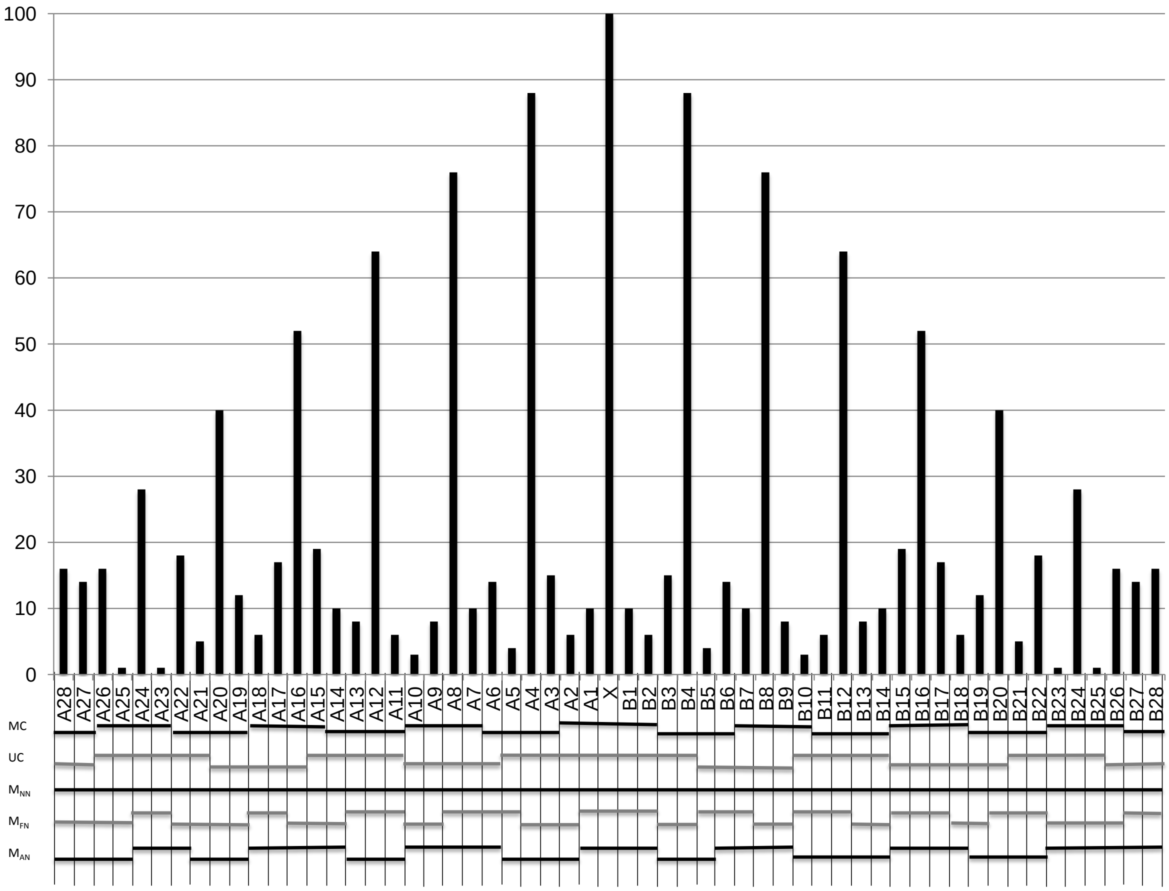


Figure 7.1: Clustering results for each of the 5 methods considered. Vertical axis indicates sequence abundances for the input set. Lines at the bottom indicate cluster assignments for each tool.

(nearest neighbor, i.e. single linkage) implementation of Mothur grouped all sequences into a single OTU as expected, but clustering patterns in the FN and AN mothur implementations were inconsistent. Complete linkage dimensions ranged from 2% to 5% in AN, and 1% - 3% in FN. Further, the clustering patterns of Uclust and Mothur's NN and AN implementations differed for series A and B.

To assess the programs robustness we randomly shuffled the sequences in the continuous dataset 100 times and tested for variability in the resulting OTU clusters as well as variation in the number of OTUs formed. These tests were repeated with the other clustering algorithms and the results were compared. Our algorithm and the NN implementation of mothur formed the same fifteen and one OTU cluster(s), respectively, in all 100 shuffles (Fig. 7.2 A and C). The clustering patterns for UClust, FN and AN all showed considerable variation with sequence order (Fig. 7.2 B, D and E).

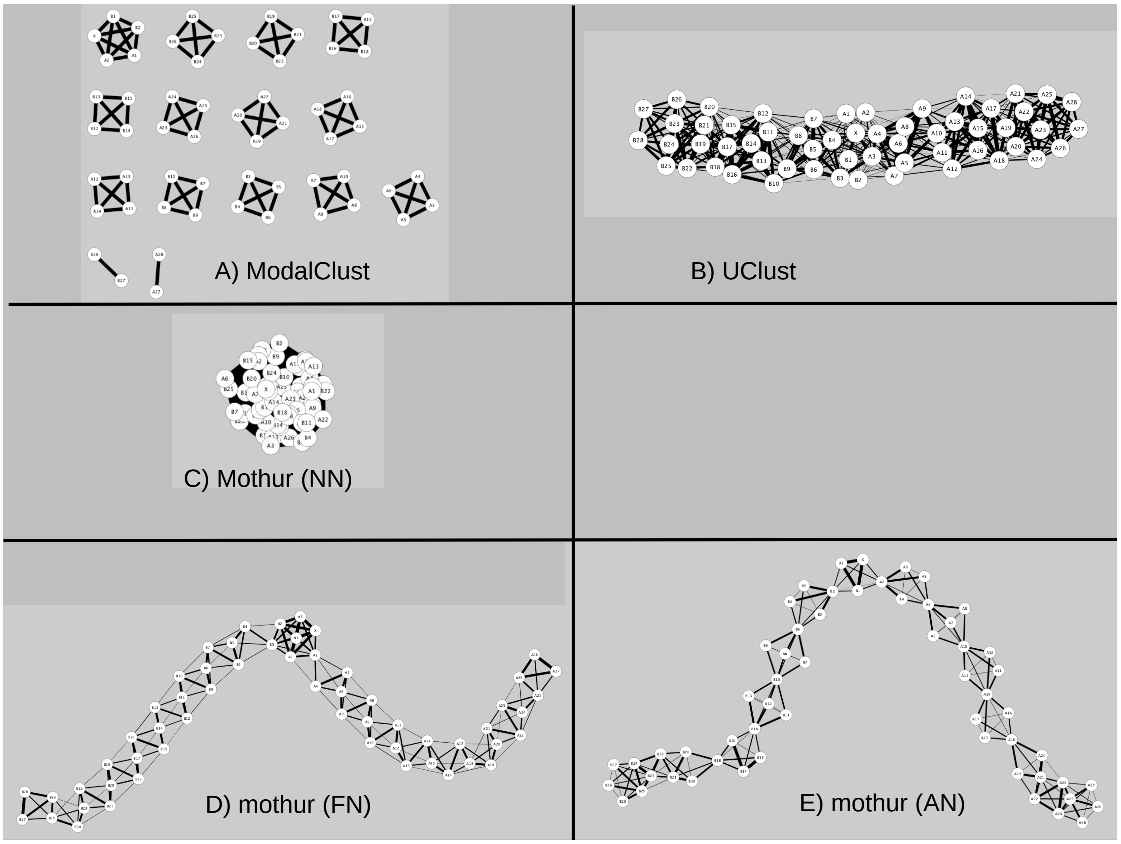


Figure 7.2: Clustering results for different methods after 100 shuffled data sets. Thickness of the line indicates the number of shuffles in which the two elements were members of the same cluster.

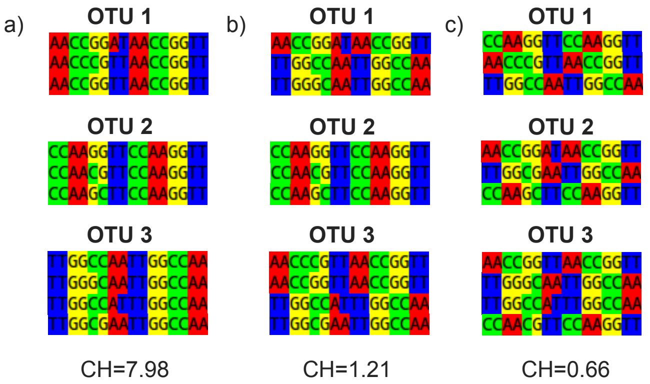


Figure 7.3: Illustration of an artificial data set with 3 possible cluster assignments and the CH value. (a) Optimal clustering. (b) Worse clustering. (c) Worst clustering. As expected, the quality of the cluster assignments correlates with CH value.

7.3.2 Calinski-Harabasz metric

To evaluate the quality of the cluster results, we use the Calinski-Harabasz (CH) metric introduced in Section 6.2.2. To test the Calinski-Harabasz metric on sequence data, we generate an artificial dataset and we permute the OTU assignments. The results are shown in Figure 7.3. As expected, the worse the OTU assignment is, the lower the CH score is. However, there is a subtle issue with the CH score assignment. If a clustering assignment has a different number of clusters k from another clustering assignment, then the comparison of CH indices is not necessarily indicative of clustering quality. To be able to compare clustering quality for different number of clusters, we define separate measures A and B for the sums in the CH score, correctly normalized for comparison. We define:

$$A = \frac{\sum_{K,L}(\bar{x}_K - \bar{x}_L)}{k} \tag{7.1}$$

and

$$B = \sum_K \frac{\sum_{i,j\in K}(x_i - x_j)}{|K|(|K|-1)}. \tag{7.2}$$

where the sum for B only runs over clusters K with at least 2 elements. Here A is the average distance between two clusters, and B is the average width of each cluster, and both quantities are normalized so that they can be compared for different k. Ideally, a good cluster assignment will have a large score for A and as small score as possible for B. Hence, we propose A/B as a good metric for comparison.

7.3.3 Real metagenomic data

Method	**Radius**	**Num. OTUs**	**CH**	**Runtime (s)**	**A**	**B**	**A/B**
Complete Linkage (Mothur FN)	0.03	2868	1479.1	1708.1	0.107	0.003	30.8
Single Linkage (Mothur NN)	0.03	671	2.6	1724.3	0.114	0.012	9.4
Uclust	0.03	1137	42.1	1.6	0.110	0.778	0.14
Modal Clustering	0.03	1039	11.7	40.5	0.113	0.007	15.5
Modal Clustering (no optimization)	0.03	1039	7.4	4.1	0.114	0.008	13.5
Modal Clustering	0.015	4077	375.1	182.1	0.108	0.004	24.9
Modal Clustering (no optimization)	0.015	4077	126.4	25.4	0.109	0.005	22.8

Table 7.1: Summary of results of clustering the Cattle rumen 1 dataset with $n = 31,201$ sequences.

To further evaluate the quality and speed of the algorithm we turn our attention to realistic sequence data. Here we use the cattle rumen 1 dataset from Chapter 8. This is an aligned, quality-filtered dataset with

$31,201$ sequences with alignment width of 471. The results of the algorithm and the quality of clustering are presented in Table 7.1. The results show that complete linkage clustering still outperforms heuristic methods, but the modal clustering performs well at comparable radius (which is half of the complete linkage cluster diameter). Optimization step described in 7.2 significantly improves the CH index value at a cost of 8-10 longer computation time, but it still outperforms complete linkage in speed by a factor of 10. Looking at the A/B column, we see that the modal clustering performs the best from all the heuristic methods.

In this test, complete and single linkage were performed using Mothur 1.28 from October 2012, and the measurements of time include the time spent computing the distance matrix. Uclust 1.2.15 was used, and sequences were sorted by length prior to clustering as advised by the uclust manual.

7.4 Simhashing applied to biological sequences

Recent datasets are breaking the barrier of 100,000 sequences per sample. Complete linkage clustering quickly becomes inviable beyond 40,000 sequences or so, simply because the distance matrix becomes too large to fit into the available memory a typical server. One can apply some patchwork solutions (such as in-memory compression of matrix rows), but complete linkage of more than 100,000 sequences is impossible. Therefore, alternative clustering methods will still be necessary. One method, called similarity hashing (simhashing), introduced in 2007 by Sadowski and Levin [181], is capable of clustering N elements in $O(N)$ time and space. A requirement for the algorithm is that the elements have to be *sets*. In the case of sequences, substrings of sequences can be used as elements of each set (each read is a set). One or more pseudo-random hash functions are applied to each substring, and the minimum (or maximum) value of a hash function is used as the label for the cluster. Since the hash functions are pseudo-random, the probability that the value of the minimum is same for two sets A and B is proportional (in the statistical limit of many elements in a set) to the Jaccard index J of similarity between two sets:

$$J = \frac{|A \cap B|}{|A \cup B|}. \tag{7.3}$$

Evidently, this algorithm is extremely heuristic – in $O(N)$ time, one cannot consider all N^2 distances between the sets. We do not expect that this algorithm should perform as well as complete linkage, but it could nevertheless be an important addition to our present bioinformatics pipeline and widely used in contemporary and future datasets. However, the major challenge at this time is to find a way to associate the parameters of simhashing algorithm (choice of hash function and number of substrings) to the percent sequence identity radius. Until then, the clusters formed by the uncontrolled choice of hash functions are not biologically relevant.

7.5 Conclusion

In this chapter, we introduced a novel, fast, clustering algorithm and we showed that it runs faster than complete linkage while not having as many clustering artifacts as the greedy solution. While complete linkage clustering still outperforms other heuristics, it is important to point to consider the scaling of the algorithms used and the trade offs between accuracy, efficiency, and computational complexity. As datasets become bigger, the $O(nm)$ scaling of our algorithm will significantly outperform complete linkage scaling of $O(n^2 \log n)$. But there are other interesting questions to answer regarding clustering.

One should note that the complete linkage clustering methodology is not ideal. It is a method that builds clusters from bottom up. It merges small clusters into larger ones, until the desired clustering distance r is reached. Hence, it is a "quenching" algorithm because the clustering can be trapped in a local minimum by one of the early steps of the hierarchical algorithm. For an example of such a quenched clustering solution, see Figure 6.3. The Figure shows how a clustering assignment is made as the hierarchical process of complete linkage happens. The quenching happens because once two elements are in a cluster together, they stay in a cluster together throughout the hierarchical clustering process. The question of whether this quenching is detrimental to quality of clustering has not been addressed in the bioinformatics community yet according to our knowledge. One easy solution to the problem would be to implement a Simulated Annealing or Quantum Annealing method to minimize the Calinski-Harabasz function. One could then measure and compare how much better this works in comparison to existing methodologies.

A further interesting question in the literature is the question of the validity of taxonomic assignments at the microbial level [182]. A taxonomy is an organization of the tree of life according to evolutionary relationships inferred from the observed phenotypes and fossil evidence as opposed to phylogeny, which is inferred from the genomic evidence (though that is not the only possible taxonomy, see [183] for a statistical physics perspective). It is known that the scaling of the topology of the tree of taxonomy is significantly different from that of phylogeny [184]. Does a clustering analysis show the same result? One can use a clustering quality index such as Calinski-Harabasz, in conjunction with reference full-length ribosomal RNA sequences, to evaluate the quality of the taxonomic assignment. One can then compare this quality with that generated from a phylogenetic tree and in that way contribute to the debate of phylogeny versus taxonomy.

Finally, existing high-throughput sequencing technology can only sequence partial reads of genetic sequences. Depending on which subregion of 16S gene is sequenced, it is known that tree phylogenies can vary widely [185]. The dependence of OTU clustering assignment on short reads is not known. Typically, the same clustering radius of 3% is used for all short reads, but different subregions of 16S rRNA have different proportions of conserved columns. Having a map of 16S rRNA for effective clustering radii would facilitate easier comparisons of datasets in metagenomics. To calculate this map one can use a reference database, artificially simulate the primers commonly used in the literature and evaluate how much do the clustering assignments change when short reads are used (as opposed to reference full-length sequences).

Chapter 8

Quantification of the relative roles of niche and neutral processes in microbiomes

The theoretical description of the forces that shape ecological communities focus around two classes of models. In niche theory, deterministic interactions between species, individuals and the environment are considered the dominant factor, whereas in neutral theory, stochastic forces, such as demographic noise, speciation and immigration are dominant. Species abundance distributions predicted by the two classes of theory are difficult to distinguish empirically, making it problematic to deduce ecological dynamics from typical measures of diversity and community structure. Here we show that the fusion of species abundance data with genome-derived measures of evolutionary distance can provide a clear indication of ecological dynamics, capable of quantifying the relative roles played by niche and neutral forces. We apply this technique to six gastrointestinal microbiomes drawn from three different domesticated vertebrates, using high resolution surveys of microbial species abundance obtained from carefully curated deep 16S rRNA hypervariable tag sequencing data. Although the species abundance patterns are seemingly well fit by the neutral theory of metacommunity assembly, we show that this theory cannot account for the evolutionary patterns in the genomic data; moreover our analyses strongly suggest that these microbiomes have in fact been assembled through processes that involve a significant non-neutral (niche) contribution. Our results demonstrate that high-resolution genomics can remove the ambiguities of process inference inherent in classical ecological measures, and permits quantification of the forces shaping complex microbial communities.

8.1 Introduction

Ecological species distributions are determined by the interplay between environmental factors and evolutionary processes. In classical ecological theory, niches, characterized, for example, by nutrients and other environmental factors, determine species abundance distributions and populations primarily through deterministic partitioning of resources amongst species [186]. Species populations are limited by niche carrying capacity, rather than interspecies competition, thus tending to promote coexistence [187]. In niche theory, diversity is determined primarily by the number of available niches, raising the issue of how to account quantitatively for the apparent observed diversity [188, 189, 190, 191] from well-documented instances of niche differences [192].

An alternative perspective is the class of neutral theories, in which species are functionally equivalent, and stochastic factors such as immigration, birth-death processes and speciation are the primary drivers of ecological diversity and community structure [193, 194, 118, 195, 196, 197]. This class of models has been reported to be capable of accurate predictions for the species abundance distributions in (e.g.) riverine fish populations [198] or microbial populations [199], in addition to the early successes in forest ecosystems, a planktonic copepod community, and a bat community in Barro Colorado Island (BCI) [118]. However, the methodology used in such comparisons is contentious when examined carefully [200, 201], with sampling issues, parameter estimation, and model definition being some of the key factors that require careful attention. The assumptions of neutral theory, in particular functional equivalence, are not transparently biological [119], and additionally have been criticized on a variety of empirical grounds [202, 203], including the predictions for species lifetimes, speciation rates and the incidence of rare species [204]. Other technical assumptions, for example that the number of individuals competing for a resource is a constant (the "zero-sum" assumption), may be unrealistic, but can be extended or relaxed [120, 124, 197]. Perhaps a more useful insight into the applicability of neutral theory comes from considering the interplay between niche stabilization mechanisms and fitness [205]. A recent study of a sagebrush steppe community, where strong niche stabilization mechanisms were identified even in the presence of apparently small fitness differences [206], underscores the fact that weak functional inequivalence need not necessarily mean that niche dynamics are negligible. On the other hand, a study that attempted to infer pairwise interaction strengths among the most abundant species in the BCI site found that interspecies interactions were much weaker than intraspecies one, in apparent agreement with neutral assumptions [207]. One study on plant, bacteria and ammonia-oxidizing archaea in steppe ecosystems found that strength of neutral and niche effects could be tuned by controlling the artificial Nitrogen deposition rate [208].

Despite their fundamental differences, and the plethora of studies nominally supporting each side of the niche-neutral dichotomy, these theories predict species abundance distributions that are difficult to distinguish empirically [190, 209], with similar mathematical properties for asymptotically large diversity [210]. The inverse problem of inferring ecological dynamics from measures of diversity does not appear to have a unique solution, either theoretically or empirically. Accordingly, a more nuanced perspective has arisen [187, 202, 211], in which elements of both types of theory may contribute to a proper description of the ecological dynamics and a variety of mathematical frameworks for accomplishing this type of synthesis have recently appeared [212, 213, 214, 215, 216, 207, 217]. For example, such theory may apply in tropical arthropod populations, where there is evidence that communities assemble neutrally within each niche defined by habitat [218]. Nevertheless, it remains an open question as to how to properly characterize community dynamics, and how to usefully quantify the relative roles of niche and neutral processes in the evolutionary dynamics of ecosystems.

These questions are of particular relevance to microbial communities, which play functionally important roles in ecosystems, but are typically rich in diversity, suggesting the presence of sub-populations shaped primarily by stochastic forces. Such communities would not be expected to represent end members of the niche-neutral continuum, and quantification of their structuring process represents a complex problem that has recently attracted attention. Most studies find evidence for a mixture of neutral and niche processes in microbial community assembly [219, 220, 221, 222, 223]. These seem to arise for different physical reasons. One indication is that the neutrally-assembling taxa are generalist microbes, that can exist in a wide variety of environments [221], whereas the niche portion of the microbiota are adapted to the media conditions [224]. There are also indications that that microorganismal coocurence patterns are shaped by the same processes and interactions that shape macroorganismal coocurence patterns [225].

In this chapter we propose a methodology for addressing the problem of quantifying the relative role of niche and neutral processes in structuring microbial communities, by fusing measures of abundance with phylogenetic information. The merging of classical ecological measures with phylogenetic analysis is growing in importance, but is still in its infancy [226, 227, 228, 229, 230]. The method presented here is particularly applicable to uncultured microbial communities that are characterized by a high level of diversity, and are amenable to modern metagenomic tools, such as pyrosequencing.

In order to explain the basic idea of how we quantify an ecosystem on the niche-neutral continuum, it is necessary to recall how microbiomes can be probed by genomic methods. The first step in an ecological study of a microbiome, following sequencing, cleanup and alignment, is the assignment of sampled sequences into Operational Taxonomic Units (OTUs) through a clustering process [7]. The OTUs are then used as a proxy for estimating microbial species abundance [231]. The OTU data are two-fold. On the one hand, the OTUs have relative abundances that are estimations of the species' abundances in the environment. On the other hand, OTUs also have representative sequences associated with them. Typically a representative sequence of an OTU is the most abundant of the identical clones within the OTU, and also it is more than 97% similar to every other sequence within that OTU. This genomic data associated with the representative sequence allows us to think of OTUs as points in a sequence space as illustrated in Fig. 8.1. We can think of distances between points in this space as corresponding to the phylogenetic or sequence distances between the sequences in these OTUs.

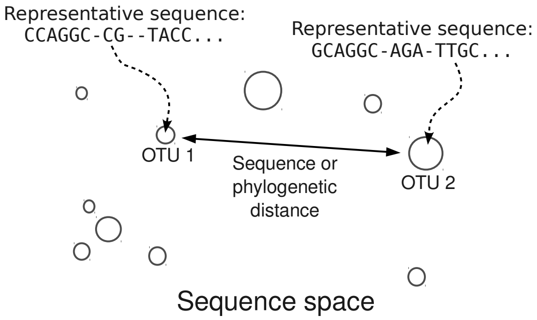


Figure 8.1: Sketch of the starting point for a metagenomic analysis of an environment. Circles indicate OTUs, and abundance (number of sequences within the OTU) is labeled by the size of the circle. A representative sequence is associated with each OTU. The OTUs are embedded in a sequence space such that the distance between the circles in the sequence space corresponds to e.g. sequence or phylogenetic distance between the representatives.

This cloud of points in high-dimensional sequence space can also be labeled by OTU abundance. In our work, this is determined by sequence abundance (after every effort has been made to account for artifacts), but in principle OTU abundance labels could be obtained from any other source, such as Q-PCR. In this space, we can categorize the OTUs into two sorts: the most abundant OTUs (which we term "modal" OTUs, and define this precisely below) and the other, less abundant, OTUs (which we term "rare" OTUs, and define this precisely below). The correlations between the modal and rare OTUs will depend upon the evolutionary dynamics, and in fact exhibit sharp mathematical differences that can be used to discriminate different putative dynamics. To see the essential idea, we will now explain how this would work in two caricatures of ecosystem dynamics: a simplified neutral model and a simplified niche model. A significantly more elaborate analysis is carried out below, in the main body of this chapter, but the key concepts are captured by these simplified models.

First, suppose that the evolutionary dynamics is itself neutral, so that the rare and modal OTUs are distributed at random in the high-dimensional sequence space. We are going to be interested in measuring the distances between sequences corresponding to different OTUs, and comparing their similarity. Let us assume that the sequences being analyzed are all of the same length, containing L nucleotide bases from the usual 4-letter alphabet (ACGT); here we are ignoring real life complications such as insertions, deletions and gaps. We label the sequences by S^i_α, where $\alpha = 1 \ldots L$ labels position along the sequence and i labels the OTU; S^i_α can take the values 1,2,3,4 corresponding to the alphabet of bases ACGT. We define the normalized Hamming distance H_{ij} between two sequences i and j as the fraction of bases in i that are different from the base in the corresponding position in j:

$$H_{ij} \equiv \frac{1}{L}\sum_{\alpha=1}^{L}(1-\delta(S_\alpha^i - S_\alpha^j)) \tag{8.1}$$

where δ denotes the Kronecker delta. The mean $\langle H\rangle$ of H_{ij} averaged over a large sample of random sequences would be 3/4, because there is a 1/4 chance that two bases at the same position are identical. Thus, the probability distribution of H would be expected to be a roughly bell-shaped curve, peaked around $H = 3/4$, with a width dependent on the number of sequences. In practice, there are complications due to insertions, deletions and gaps, but most importantly, conserved positions. Bases that are highly conserved cannot be appropriately modeled as being chosen randomly from the alphabet. This can be taken into account by simply restricting the above analysis to bases that are strongly non-conserved: let us call the number of highly conserved bases $M < L$, so that the expected value of H will now be reduced by the fraction of conserved bases: $\langle H\rangle = 3(L-M)/4L$. Thus, taking into account conservation, the bell-shaped curve will shift its peak to a smaller value of H. In the data presented below, we found that $L \sim 200$ and $M \sim 160$, so that the distribution of H should be peaked at about 0.15, in the case of a neutral system. Now consider a subset $\{E_k\}$ of distances$\{H_{ij}\}$. For each "rare" OTU k, we rank all of the distances between OTU k and each "modal" OTU l. Then, we select the shortest such distance and label it E_k. In this way, the set $\{E_k\}$ is the set of distances of "rare" OTUs to their nearest niche neighbor. For the above case where the evolutionary dynamics is neutral-like, the distribution of E is also a bell-shaped curve like the distribution of H. However, its mean is slightly shifted to the smaller values, and its standard deviation is smaller (because $\{E\}$ is the subset of shortest distances from the set of $\{H\}$). In other words, $\langle E\rangle < \langle H\rangle$.

Second, let us consider a caricature of a system that is dominated by niche dynamics. In the extreme (and unrealistic) case where there is only one niche, occupied by one particular modal OTU, the probability distribution of E will be a delta distribution peaked at $E = 0$. In a more realistic model, where there is a cloud of rare OTUs surrounding the modal OTU, having evolved from it by a few point mutations, one would expect the probability distribution of E to be peaked at $E = 0$, and then to monotonically decrease for $E > 0$. In the case of a system with several niches, the probability distribution for E will be somewhat more complicated, because one needs to calculate the normalized Hamming distance from each rare OTU to the nearest modal OTU, and this requires making a Voronoi polyhedron construction in sequence space. Nevertheless, for small values of E, the probability distribution will be dominated by the single niche argument given above, and the functional form will be unchanged: peaked at the origin and monotonically decreasing for $E > 0$. These two caricatures for simplified models of ecosystem structure are sketched in Fig. 8.2, and show that there are clear and distinct signatures arising from the nature of the processes that have structured the community.

In the remainder of this chapter, we numerically evaluate the metric for model systems in order to quantitatively and concretely confirm the above heuristic description. We then describe how we have implemented these ideas in a proof-of-principle study of vertebrate gastrointestinal microbiomes. These experimental systems were chosen, not only because of the growing recognition of the importance of microbiomes as a determinant of host health [232], but also because these are systems that have high diversity, and are likely to be shaped both by stochastic and niche processes. Indeed, as we will see, they can be well-described naively by neutral theory, although in fact niche processes play a fundamental role in structuring these communities.

8.2 Model calculations

In this section we evaluate our metric on model systems parametrized by a single parameter, α, the proportion of the system undergoing a niche dynamic. We perform 5000 Monte Carlo simulations of the following process. We simulate N OTUs (here $N = 1000$) each with representative sequences of length $L = 200$. A subset αN ($0 \leq \alpha \leq 1$) of the OTUs undergo a niche dynamic in the following way. A single random OTU is chosen to be the center of the niche. The remainder of the $\alpha N - 1$ OTUs (niche OTUs) are are generated by performing random mutations of the genome of the OTU representing the niche center. The placement and number of the mutations were chosen randomly in the following way. Placements of mutations were sampled uniformly (without replacement) across the entire genome. The number of mutations for each of the niche OTUs was sampled from an exponential distribution thereby modeling the evolution of OTUs under multiplicative fitness pressure (larger number of mutations corresponds to smaller fitness, and hence smaller abundance of OTU). The remaining $(1 - \alpha)N$ OTUs (neutral OTUs) are randomly distributed throughout the sequence space, and they represent the sequences undergoing dynamics under no evolutionary pressure (neutral dynamics).

Each OTU in the model system is associated with an abundance. The abundances of neutral OTUs are randomly sampled from an exponential distribution. (In the Hubbell Neutral Model, the OTU rank abundances are exponentially distributed.) On the other hand, the abundance of niche OTUs exponentially scales with their closeness to the niche:

$$N_i = A \exp(-d_i) \tag{8.2}$$

where N_i is the abundance of OTU i and d_i is the distance from the OTU to the center of the niche (in sequence space). The results of our metric, the distributions of $\{E_k\}$ are shown in Fig. 8.3 for 3 model

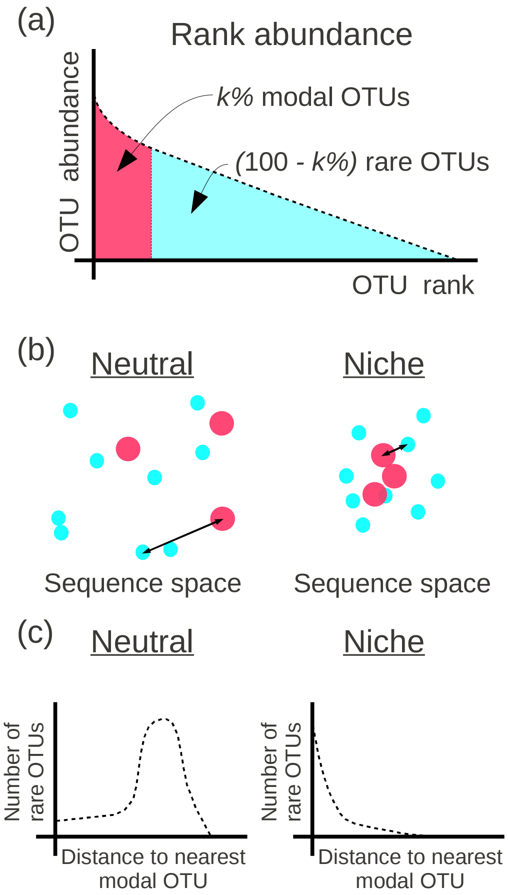


Figure 8.2: (a) Classification of the OTUs into two groups based on the rank abundance. The top $k\%$ of OTUs are labeled modal, whereas the remainder of the OTUs are labeled rare. (b) Sketch of the neutral and niche evolution processes in sequence space. Light blue OTUs are rare, whereas red OTUs are modal. For the neutral process, the average distance of a rare OTU to its closest modal OTU is large (indicated by the arrow). For the niche process, this distance is much smaller since rare OTUs cluster about the modal OTUs which define the niches. (c) Sketch of the expected distributions of distance to the closest modal OTU. For the neutral process, this distribution is peaked around some non-zero distance, which is close to the average distance between the OTUs in the dataset. In the niche process, the distribution monotonically decays with distance since the rare OTUs are attracted to the niches.

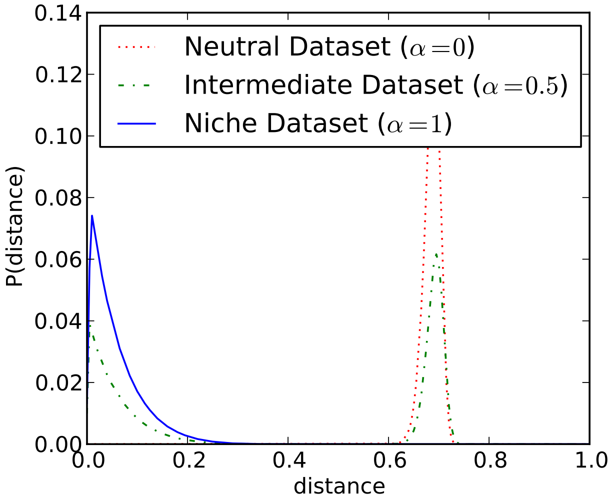


Figure 8.3: The results of our metric, the distributions of E shown for a fully Niche-like model dataset ($\alpha = 1$), a fully Neutral-like model dataset ($\alpha = 0$) and an intermediate dataset ($\alpha = 0.5$). The results shown are the average of 5000 Monte Carlo simulations for each dataset.

systems characterized by values of $\alpha = 0, 0.5$ and 1. We see that the heuristic arguments we described in the previous section and sketched out in Fig. 8.1(c) are consistent with these model numerical calculations.

It is instructive to demonstrate the effects of two factors on our metric, in order to highlight some of the mathematical considerations that went into the design of the metric, in particular our use of an extremal measure (the shortest distance aspect of our metric) and the influence of sampled abundance distributions. First we demonstrate the role of extremality introduced by choosing the subset $\{E\}$. Instead, if we choose to plot the distribution of $\{H\}$ we obtain qualitatively the same results for neutral-like models (compare models 1 and 2 in Fig. 8.4). However, for niche-like models, the peak at zero moves to a nonzero peak which corresponds to the average size of the niche (compare models 5 and 6 in Fig. 8.4). Thus, the choice of an extremal measure is important in making sure that the end member distributions (pure niche, pure neutral) are clearly distinct.

Second, we demonstrate what might appear at first to be a rather counter-intuitive fact: the distribution of distances is only weakly dependent on the abundance distribution of the OTUs. If the abundance of an OTU k is N_k then we could imagine modifying our procedure by weighting the contribution of E_k in the distribution $\{E\}$ by a factor of N_k. Such a weighting introduces no change whatsoever to the neutral dataset (compare models 2 and 4 in Fig. 8.4), and no qualitative change in the niche dataset (models 6 and 8 in Fig. 8.4). Finally, we can also weigh the distribution of $\{H\}$ in such a way that each distance H_{ij} between OTUs i and j gets weighted by a factor of $N_i N_j$. The results are exactly the same as with no weighing for the neutral dataset (compare models 1 and 3 in Fig. 8.4) and qualitatively the same for the niche dataset

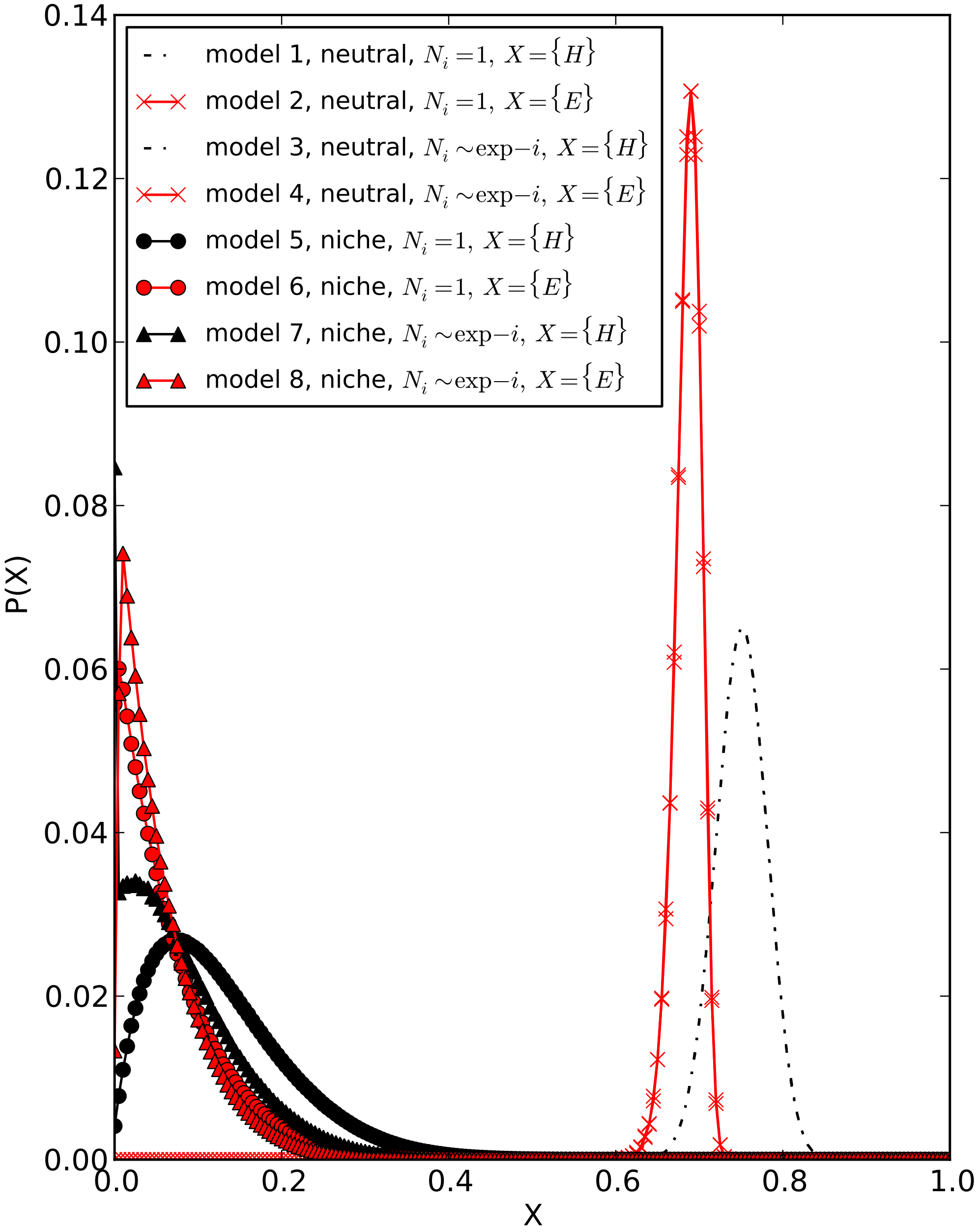


Figure 8.4: Explicit numerical calculations of our metric on 8 model systems. In these systems, we study the difference between the effects of the metric on neutral (models 1-4) and niche model systems (models 5-8). We also study the effect of choosing the closest distance (even-numbered models) compared to considering all distances (odd-numbered models). Finally, we consider the weighted models (3-4 and 7-8) versus the unweighted ones (1-2 and 5-6).

(compare models 5 and 7 in Fig. 8.4).

8.3 Results

We performed a pyrosequencing study of the gastrointestinal (GI) microbiomes of 3 pairs of domesticated vertebrates: 2 swine, 2 cattle and 2 chickens. These pairs of organisms were chosen as pilots for probing specific microbiome issues of relevance to animal science. In particular, we attempted a comparative study looking at the effects of diet on identically cloned swine, and the effects of a microbial challenge on two identically-raised chickens. For the purposes of this chapter, these comparisons and the outcomes of the experiments are not of interest: full details of the comparisons and other studies will be published elsewhere. In this study, two genetically identical cloned swine were fed different diets and then their fecal samples were collected for sequencing. Cattle rumen 1 and cattle rumen 2 were rumen fistula sampled at 0 and 8 hours after feeding, respectively [233]. Chicken caecum 94 was inoculated with *Campylobacter jejuni* one week prior to caecal sampling. Chicken caecum 1 was kept under the same conditions but without oral gavage of *C. jejuni* [170]. See the Methods for details regarding the laboratory protocols. The GI Samples were subjected to deep hypervariable 16S rRNA tag sequencing using a 454 Life Science Genome Sequencer GS FLX[231]. Table 8.1 shows the average read length and number of reads obtained for each sample.

Following their acquisition, we aligned the pyrosequenced reads using NAST [151] to the SILVA [154] database. We also aligned the reads using RDP's front end [152] to the Infernal [155] structural aligner. For each dataset, the NAST+SILVA and RDP+Infernal multiple alignments were merged and hand curated using the methodology and tools described in Sipos *et al.* [7]. Short reads and sequences with unknown nucleotides were removed. Spurious "tails" in the multiple alignment, sequences that extend beyond the region of 16S common to all the sequences in the dataset, were also removed. Distance matrices were generated from the multiple alignments, and were then fed to a complete linkage clustering algorithm to generate the OTUs. The careful multiple alignment procedure led to a vast reduction in the number of resulting OTUs in the datasets as previously reported in Sipos *et al.* [7]. See Table 8.1 for multiple alignment, species diversity and richness metrics for each of the 6 GI microbiome samples. Rarefaction curves show how the number of sampled OTUs varies as a function of the number of organisms sampled. Our rarefaction curves are shown in Fig. 8.5 for each of the 6 datasets.

We plotted the abundances of the OTUs for each of the 6 datasets in our study, and we find a very good agreement with the Neutral Model. These are displayed in rank-abundance form in Fig. 8.6, and in alternative forms in Fig. 8.7 and 8.8. The early ranks (high abundance OTUs) show some systematic deviation from the

	# reads	Unique reads	Avg length	Aligned width	OTUs at 3%	Simpson diversity	Shannon diversity	Jack-knife	ACE	Chao1
Swine 1	33283	14122	165.0	420	1509	0.0070 ±0.0003	5.8 ±0.02	2000 ±260	1472 ±55	1540 ±150
Swine 2	36254	16198	175.3	418	1856	0.0068 ±0.0003	5.9 ±0.02	2300 ±300	1633 ±53	1720 ±150
Cattle 1	31201	18264	180.7	471	2580	0.0044 ±0.0002	6.3 ±0.02	3300 ±260	3070 ±88	2640 ±190
Cattle 2	19642	10074	183.6	385	1509	0.0110 ±0.0006	5.9 ±0.03	2070 ±110	1818 ±62	1830 ±130
Chicken 1	17585	2151	136.5	310	396	0.084 ±0.003	4 ±0.03	770 ±120	655 ±75	620 ±150
Chicken 94	21646	2223	138.9	332	354	0.046 ±0.001	3.9 ±0.02	560 ±90	426 ±57	460 ±100

Table 8.1: Summary statistics of our six datasets.

abundances expected from neutral theory but at face value, these results are consistent with the majority portion (thousands) of the OTUs evolving in the absence of any apparent selection acting on the individual OTUs. Given all the factors that influence the gastrointestinal microbiome [234, 235, 102, 236, 237, 238], and the reproducible, thereby seemingly host-selected, microbial abundances [239], it seems counterintuitive that there should be no apparent selection for the vast majority of OTUs in the exponential tail of the rank abundance. However, if we compare taxonomic assignments of microbes across each pair of animals in our study (Fig. 8.9), we find that there is a correlation between the relative abundances of taxa in members of each animal pair. Namely, we observe that the most abundant taxonomic orders are the same for each animal pair (Clostridiales for swine and chickens, and Pseudomonadales for cattle). This correlation also extends to other taxonomic orders. Hence, our dataset indicates that certain taxa are favored more than others within the GI tract of these 6 vertebrates.

We now attempt to resolve this apparent contradiction, namely that the Neutral Theory fits the rank abundance patterns well, with only 2 fitting parameters, even though the taxonomic data suggests Niche selection. In order to do this, we must turn our attention to other information contained within the pyrosequenced reads. As shown in Fig. 8.1 the OTUs with their characteristic sequences and associated abundances form patterns within a high-dimensional space. Each read constitutes a point in this space, defined by its nucleotide sequence. One way in which we can attempt to comprehend the structure of this space is through dimensional reduction. We use Principal Component Analysis (PCA) in order to place the OTUs into a 2 dimensional space spanned by the two principal components. We perform a weighted version of PCA [240] where we assign a weight to the OTUs proportional to their abundance. The resulting patterns in the space of two principal components are shown in Fig. 8.10. Each circle in the figure is an OTU and the circles' size and color indicates the logarithm of the OTU abundance.

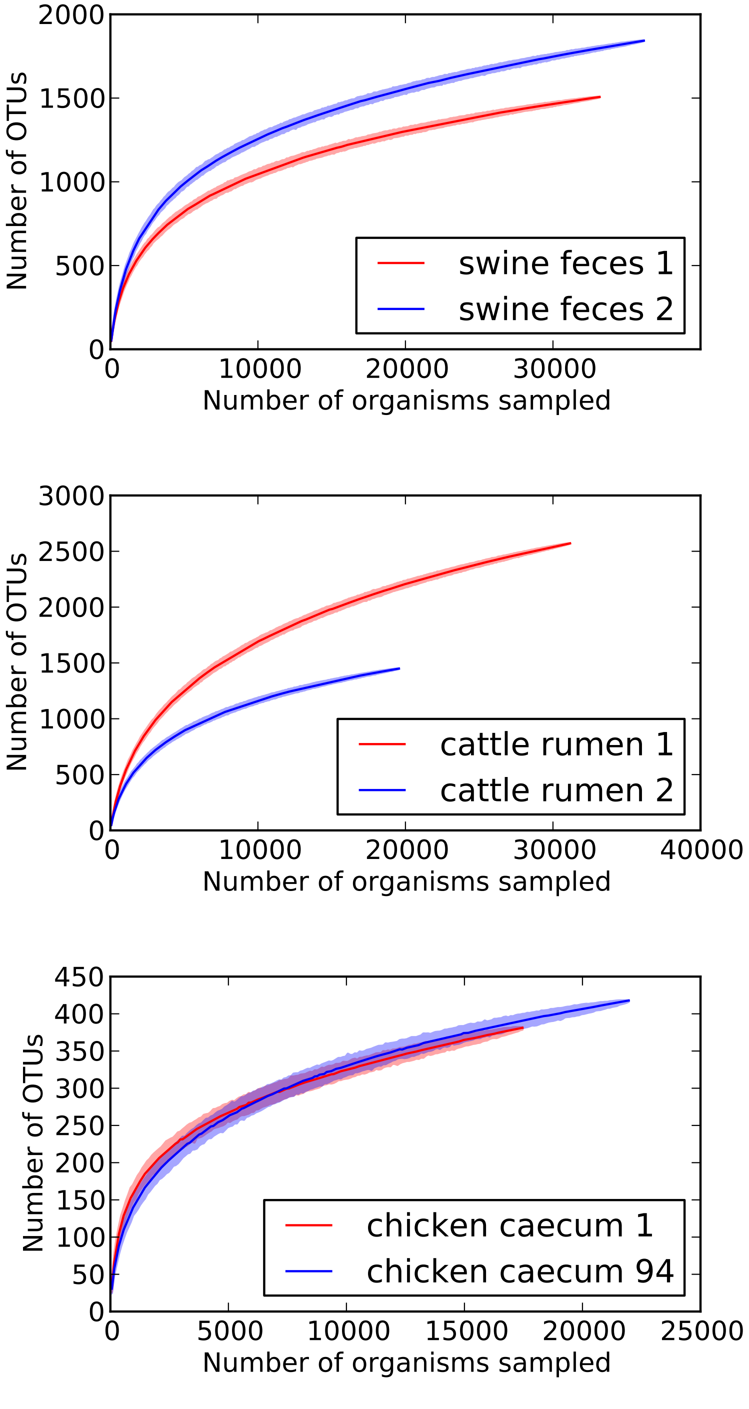


Figure 8.5: Rarefaction curves for the 6 vertebrate GI microbiomes. Solid line represents the median number of OTUs (100 resamplings) whereas the shaded area represents the 95% confidence interval.

As a control, we generate datasets of artificially generated sequences (hereafter referred to as neutral datasets). We generate a neutral dataset for each of the 6 experimental datasets to facilitate a 1-to-1

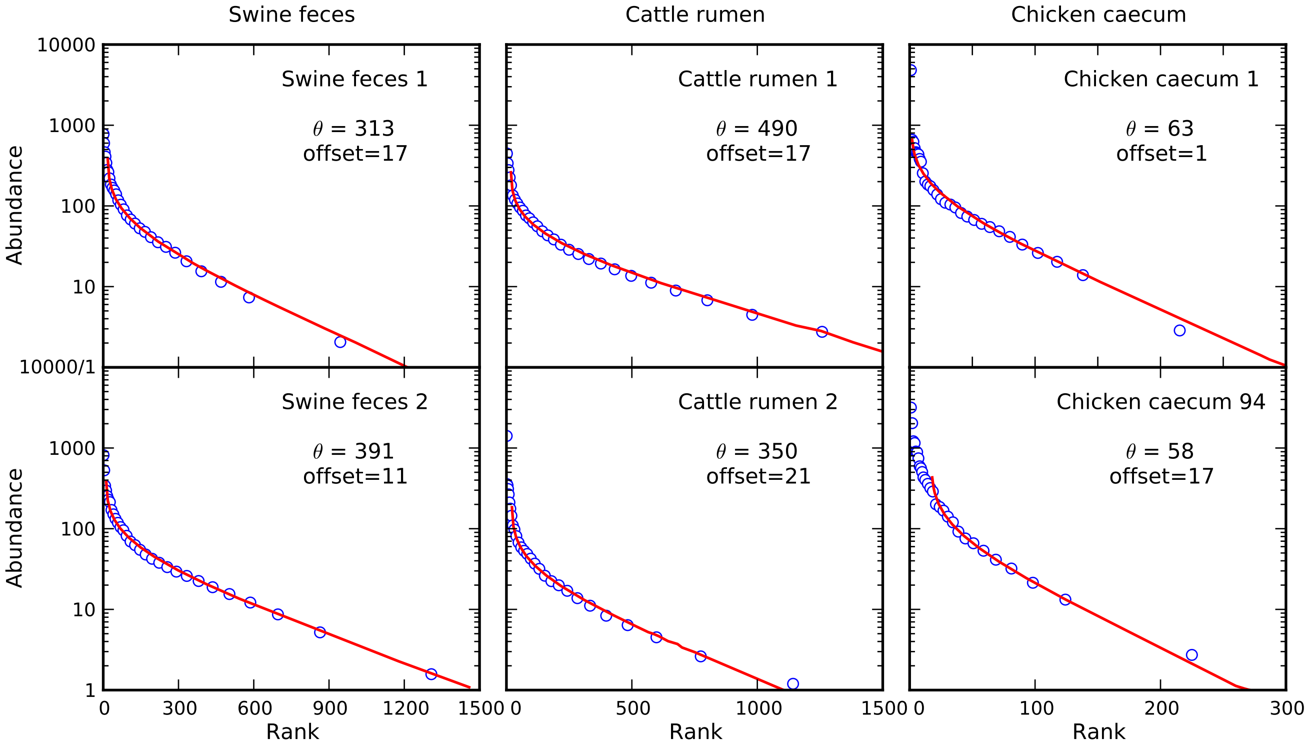


Figure 8.6: Comparison of rank abundance curves and neutral model fits for the six animal GI microbiomes. Lines indicate fits to the Hubbell's neutral metacommunity model. Parameter θ of the model is fit to correspond to the exponential tail in rank abundance. Offset represents the number of high-abundance OTUs that do not fit the neutral model.

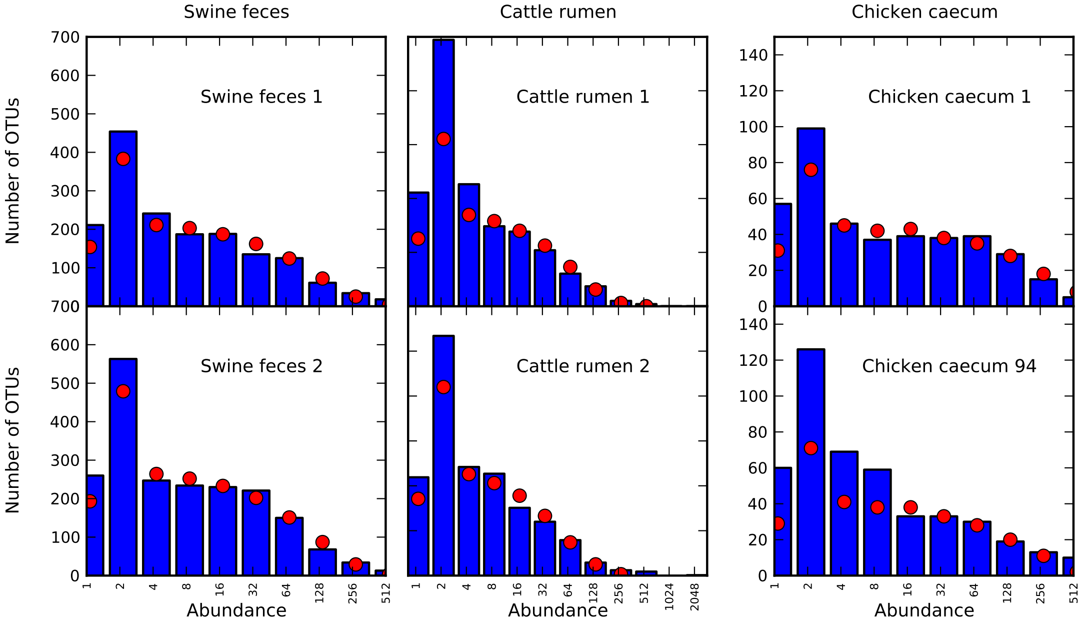


Figure 8.7: Preston plot for swine feces, cattle rumen and chicken caeca samples. In a Preston plot, the height of the bar indicates the number of species observed with abundance 1, 1-2, 2-4, 4-8, etc. Note that in all 6 datasets most OTUs are singletons. In this plot, 1-2 bars are highest because of an artifact. Traditionally, in a Preston plot, the OTUs with borderline abundances split evenly between two neighboring bins.

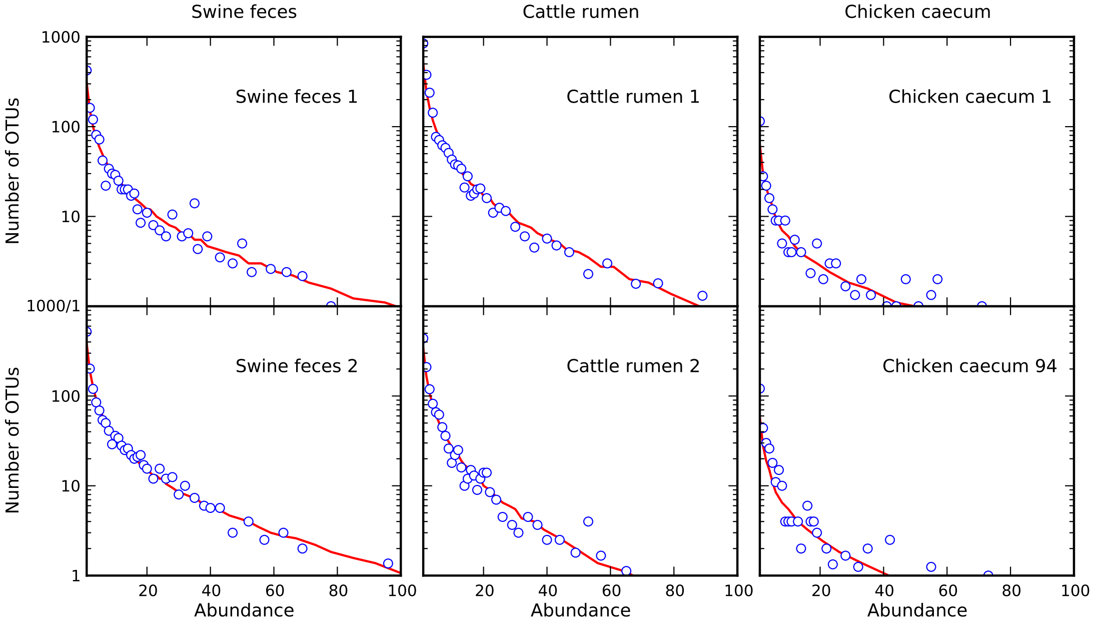


Figure 8.8: Species abundance distribution for swine feces, cattle rumen and chicken caeca. The species abundance distribution indicates the number of OTUs collected for each abundance.

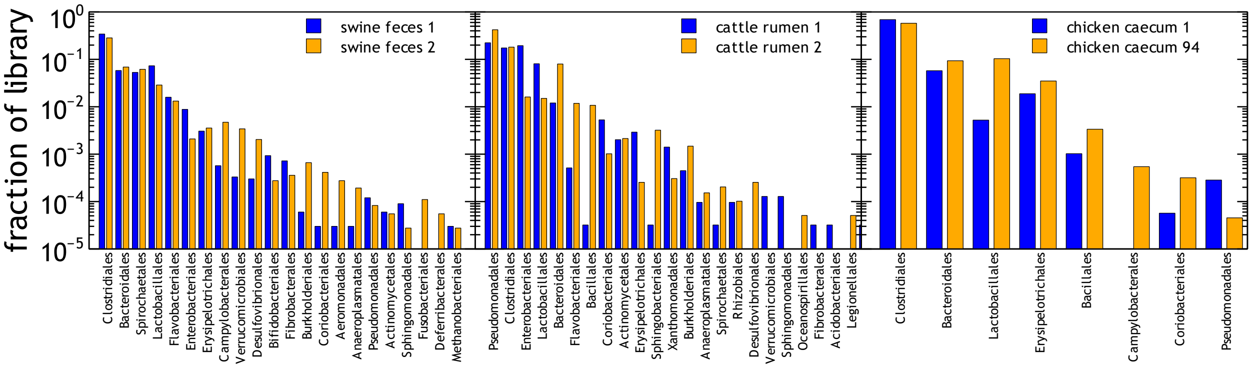


Figure 8.9: Taxa Comparisons. Taxonomic assignments at order level for all libraries, at 80% confidence threshold, sorted by combined abundance. Though there appear to be no differences in the form of the rank-abundance curves, we see differences in the taxonomic distributions here as the result of changes in diet or challenges to the microbial ecosystem.

comparison. Each neutral dataset is constructed in a way such that it has the same number of OTUs and the same OTU abundance distribution as the associated experimental dataset. However, the representative sequence for each OTU is artificially generated and has a randomized sequence, with constraint such that it has the same sequence statistics as the original dataset (probability of observing a nucleotide at a position in the multiple alignment) and column conservation. This ensures that the sequences are randomly distributed along a realistic sub-manifold of sequence space (the subset of 16S sequences that are allowed by secondary structure). We then run the PCA on the neutral datasets (Fig. 8.11). Comparing Fig. 8.10 and 8.11, we notice the following pattern in the experimental GI data: the low-abundance OTUs cluster around the high-abundance OTUs in the dimensionally reduced space. In the neutral datasets, this is not observed, instead the PCA distributes the OTUs approximately uniformly in the dimensionally-reduced space.

The fact that, in neutral datasets, OTUs distribute uniformly in dimensionally-reduced space, may seem to be at odds with the process underlying Hubbel's model. In Hubbel's model, "daughter" species appear, initially rare, in the genetic neighborhood of their abundant "parent" species. Thus, one may be led to believe that the PCA of a neutrally simulated dataset should appear the same as that of niche selected one, a set of abundant OTUs surrounded by a cloud of low-abundance OTUs. In fact, this argument will not hold at long time scales since the rare lineages will have the time to diffuse away from the modal peaks. Nevertheless, one needs to keep in mind that in the local community neutral model (which is at more realistic time scales for the case of ecological community assembly), the appearance of rare species is through migration and has no connection to genomic space. In principle, every event of appearance of a new rare species is an act of migration of a random organism from a diverse metacommunity (this difference is illustrated in Figure 8.12). There is a larger question: the metacommunity itself is unlikely to be strictly neutral, being shaped by diet

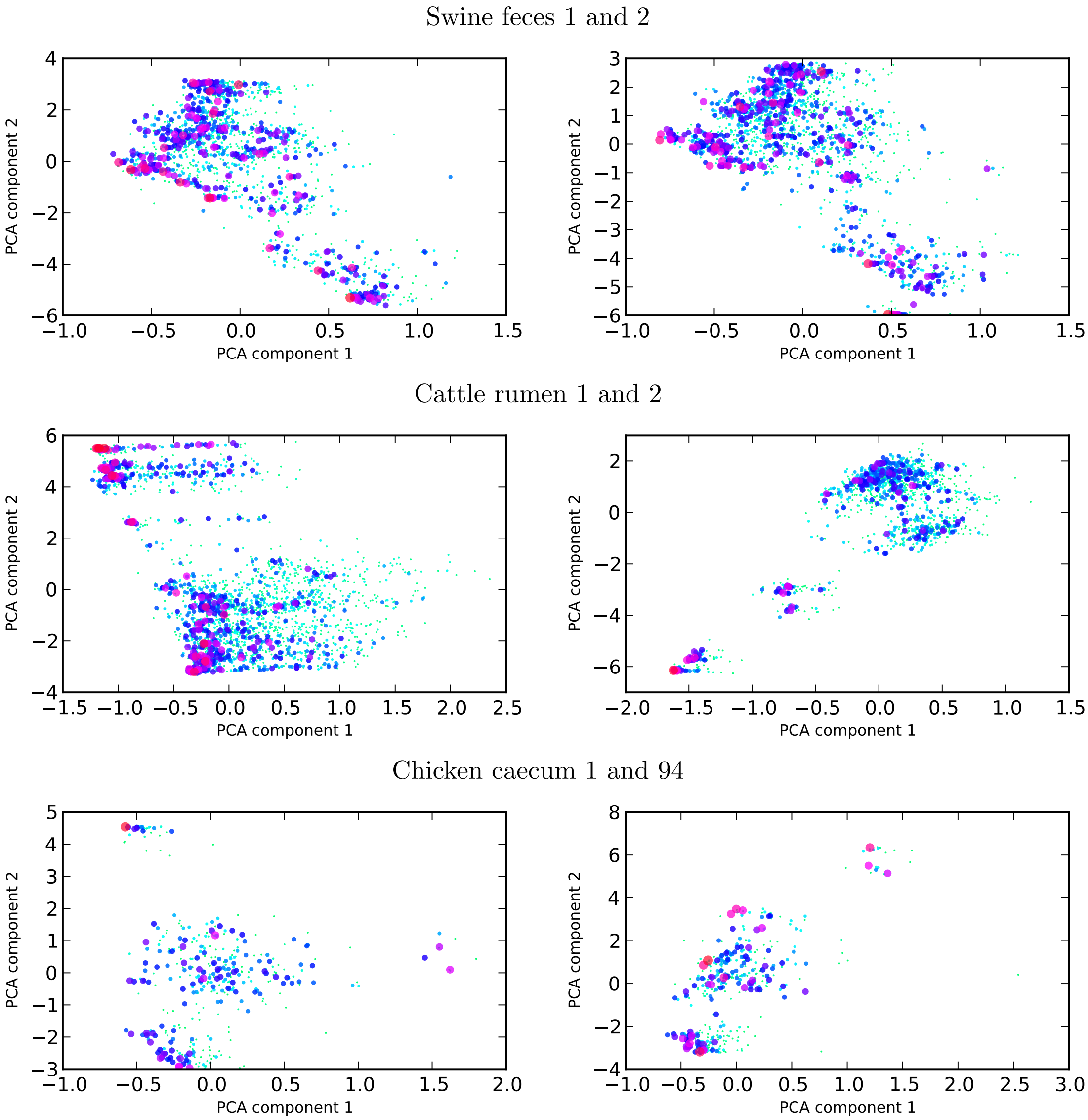


Figure 8.10: Weighted PCA ordination applied to the 6 experimental datasets. See the main text for details on how weighted PCA was performed. Each circle in this Figure represents an OTU and its size and color indicates the logarithm of OTU abundance.

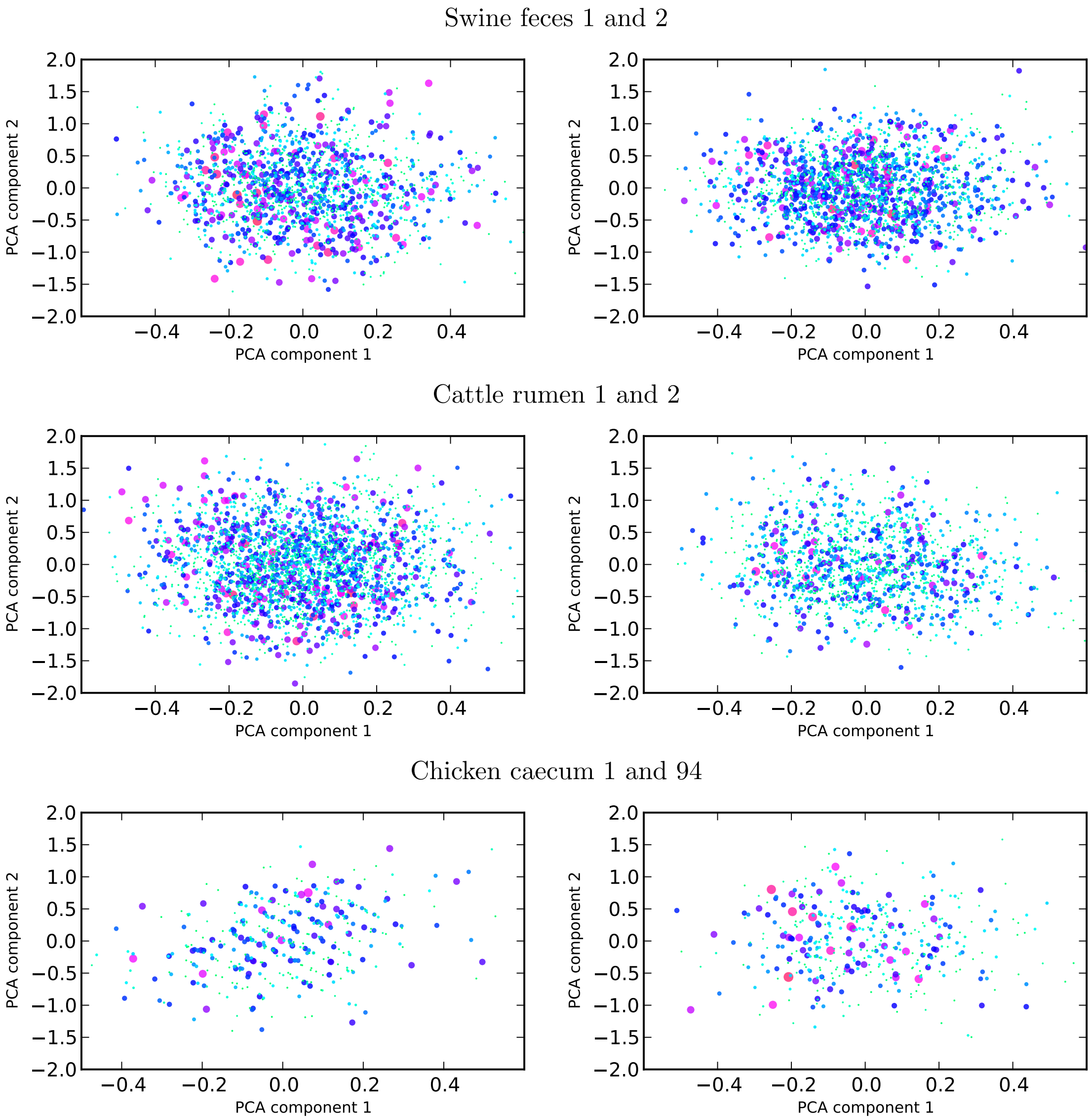


Figure 8.11: Weighted PCA ordination applied to the randomized datasets. Compare with Fig. S5. See the main text for details on how the randomized datasets were generated, and how weighted PCA was performed. Each circle in this Figure represents an OTU and its size and color indicates the logarithm of OTU abundance.

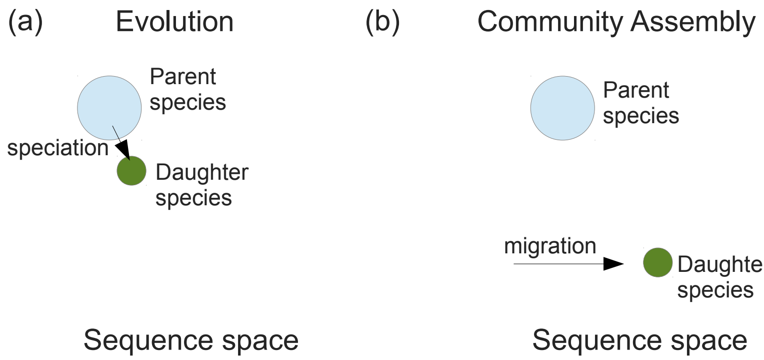


Figure 8.12: A qualitative difference between (a) evolution and (b) community assembly. In the case of evolution, the daughter species is close to the parent species in sequence space. But for the case of community assembly, there is no such requirement.

and environment of the host organism. Thus the OTUs migrating into the microbiome will be sampled from a biased distribution and this could in effect act as a niche, or as a constraint on the allowed sub-manifold in sequence space.

We now formulate a heuristic to clearly discriminate between the randomly assembled model sequences and those assembled from a niche-driven process. On a rank-abundance curve, we label the $k\%$ of the most abundant OTUs as modal OTUs. We label the remaining OTUs as rare OTUs (Fig. 8.2(a)). Instead of using the whole-dataset rank-abundance curve, one can also use per-order rank-abundance curves if additional resolution is necessary. Once modal and rare OTUs have been assigned, for each rare OTU we compute the distance to the modal OTU that is closest to it. The motivation behind this heuristic is the following. The spread pattern of sequence abundances gives us an indication of whether organisms are evolving neutrally or toward defined niches. In long time behavior, Neutral evolution leads to the expectation that organisms have an equal chance of being anywhere in this space. Niche selection, however, suggests a very biased distribution of organisms. In particular, organisms would be densely clustered about the local optimum for each niche (Fig. 8.2(b)). These two scenarios lead to very different distributions of distance to nearest niche. If the OTUs are undergoing a niche-driven dynamic, then the rare OTUs will tend to drop off exponentially in abundance around the modal OTUs. If on the other hand, the OTUs have been sampled from a community shaped by neutral evolutionary dynamics, then the rare OTUs' distance to closest modal OTU will be peaked around some non-zero distance that is the average distance between any two OTUs in the dataset (Fig. 8.2(c)).

We apply the above analysis to the case of gastrointestinal microbiome datasets of the 6 vertebrates. The results are summarized in Fig. 8.13. In this figure, the blue bars indicate the results of our metric

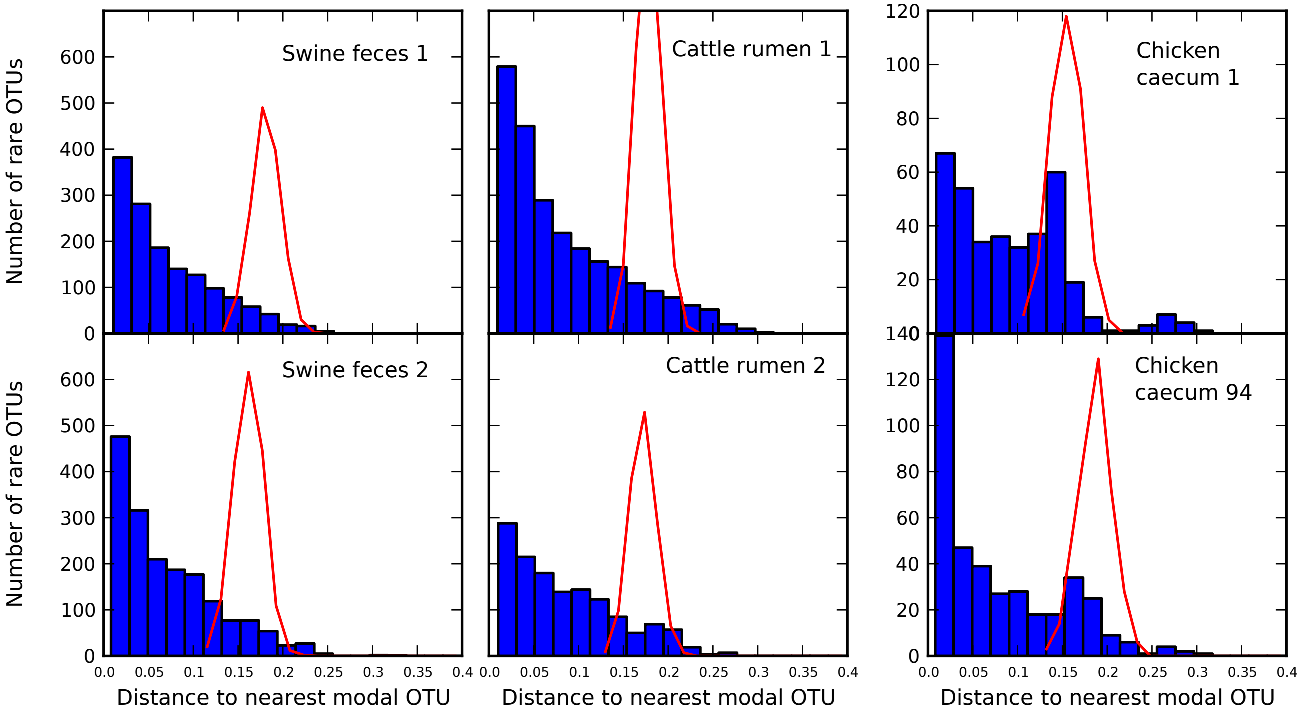


Figure 8.13: Histogram of distances of rare OTUs to the nearest modal OTU for each of the 6 gastrointestinal microbiomes with cutoff $k = 5\%$ (blue bars indicate experimental data). Red dashed lines indicate the results of the metric applied to sequences that were randomized while preserving rank abundance and sequence statistics (see text). Cattle and swine datasets share the same y-axis.

applied to experimental data. The dashed red lines indicate the results of the metric applied to a dataset of sequences that were randomized in the way described above. The results indicate that the GI tracts of the 6 vertebrates largely undergo niche dynamics, with the possible exception of a subpopulation of the chicken GI tracts. The chicken datasets have a small non-zero peak corresponding to the average distance between sequences chosen at random. Our study indicates that the sequences within this peak may be undergoing neutral dynamics. The results that we obtain are robust in that they do not qualitatively depend upon the choice of the cutoff k. In Fig. 8.14 we show the metric for $k = 5\%$ and $k = 7\%$. Similarly, the results of the metric on model systems are virtually unchanged when k is changed between 2% and 10% (Fig. 8.15) indicating robustness. Whereas our metric is robust in this way, the reader is reminded that phylogenetic resolution is nevertheless important: some niches may appear as a single OTUs at 97% percent sequence identity.

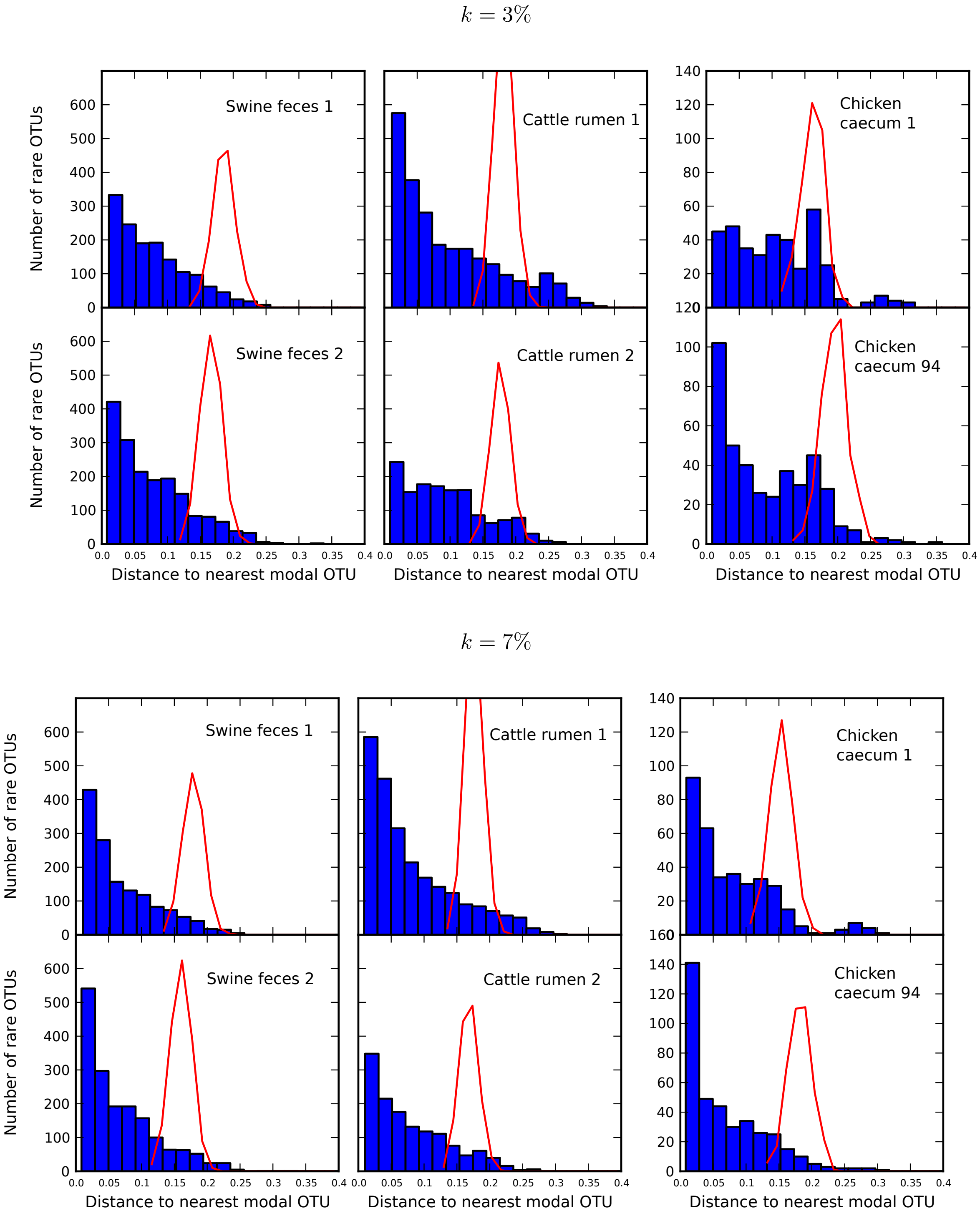
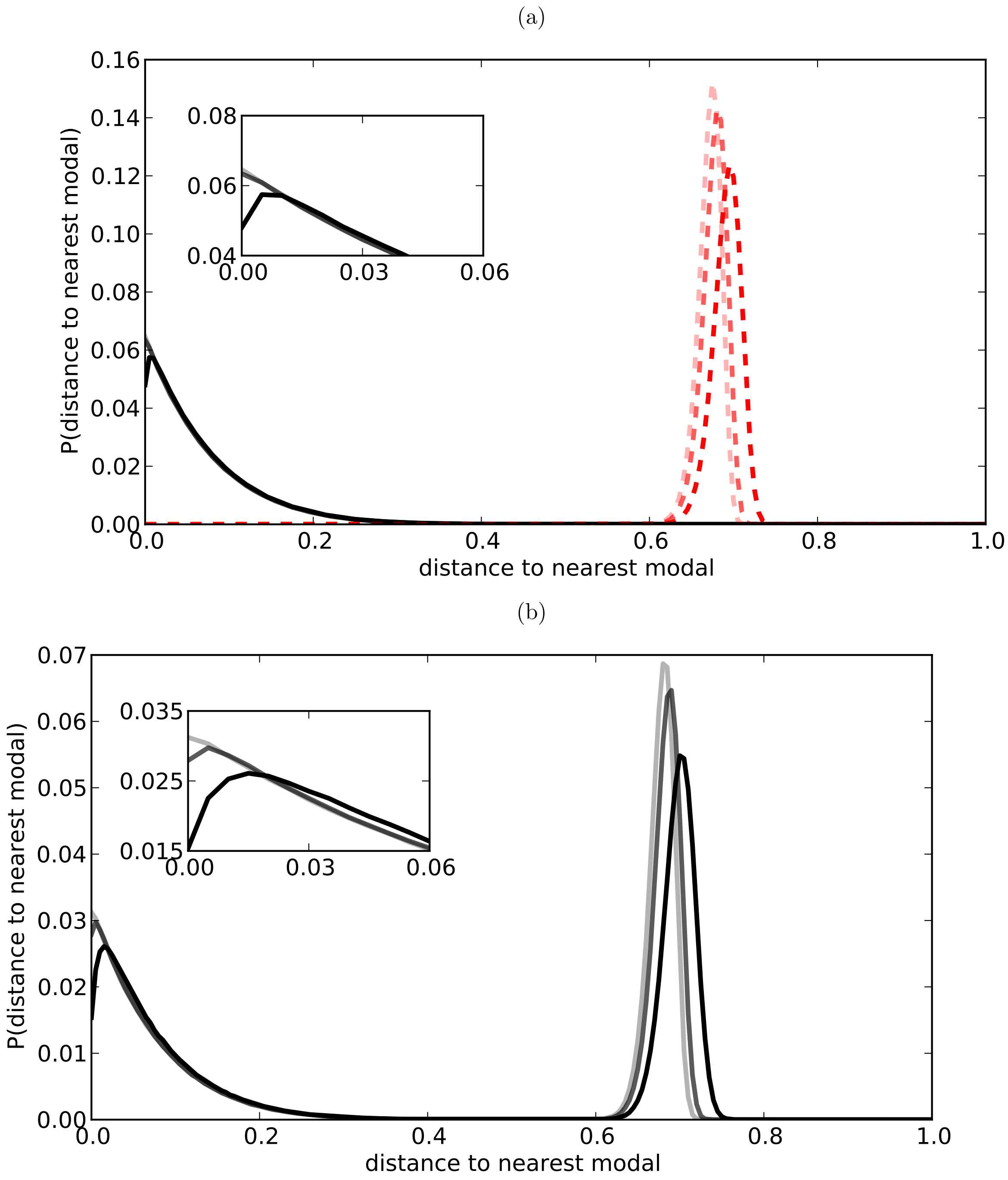


Figure 8.14: Histogram of distances of rare OTUs to the nearest modal OTU for each of the 6 gastrointestinal microbiomes with cutoffs $k = 3\%$ and $k = 7\%$. Red dashed lines indicate the results of the metric applied to sequences that were randomized while preserving rank abundance and sequence statistics (see main text)).


(a)

(b)

Figure 8.15: Measuring the effect of the choice of k on our metric. Darkest lines indicate $k = 2\%$, medium lines indicate $k = 6\%$ and lightest lines indicate $k = 10\%$. (a) $\alpha = 0.0$ model (red dashed lines) and $\alpha = 1.0$ model (black solid lines). (b) $\alpha = 0.5$ model (black solid lines).

8.4 Discussion

In this work, we set out to construct genomic-based measures of ecosystem diversity and abundance that can provide evidence for process. We focused on understanding the processes that structure microbial communities, because these play functionally important roles in many ecosystems, yet are rich in diversity. Thus, such systems would *a priori* be expected to contain at least sub-populations shaped primarily by stochastic forces. The dual features of high diversity and foundational role functionally in their host ecosystem suggest that microbial communities would not be simple to characterize as either niche or neutral. At the same time theoretical arguments suggest that such high-diversity communities might appear, for fundamental statistical reasons, as neutral.

We succeeded in creating a quantitative metric that fuses abundance and genomic data in order to determine whether an ecological system is dominated by neutral evolution or by niche selection. The key concept was to explore the correlations and associated probability distributions between the most abundant members of the community and the long, low abundance tail members. We showed that the signature of the probability distribution describing the distance in genomic sequence space from each rare OTU to the nearest modal OTU provided a signature of the strength of niche dynamics. We tested this construct on large datasets from 6 animal gastrointestinal tract microbiomes, finding in all cases that the results are inconsistent with neutral assembly. We conclude that niche selection largely dominates within the GI microbiome, despite the fact that the rank abundance patterns are apparently well-modeled by Neutral Theory.

Our results provide firm evidence from an empirical dataset that apparently neutral patterns of diversity and abundance can arise from niche-dominated dynamics, in agreement with earlier theoretical expectations [190, 209, 210, 187, 202, 211]. Our results establish definitively that simple ecological measures need to be, and can be, augmented by genomic data in order to provide insight into the processes that structure communities.

8.5 Methods

8.5.1 Sample Preparation

All procedures involving animals were approved by the Institutional Animal Care and Use Committee of the University of Illinois. For each animal, we used two different samples for our test that vary in some aspect such as diet or sampling times. The Duroc sow (2-14; TJ Tabasco) was used as the genomic template for producing cloned animals using somatic cell nuclear transfer. Tabasco was used to produce the CHORI 242

BAC library which was used to generate the full pig genome sequence [241]. The clones were born by vaginal delivery and allowed to suckle. They were weaned at 4 weeks of age and continuously housed together. They were not vaccinated or ever in contact with other pigs after weaning. Pigs were fed once daily in the morning and had free access to water. Fecal samples were collected on day 14 (the last day of that feed rotation) of each diet for a total of 4 samples for each animal. Samples were collected from the rectum into a sterile tube and frozen at -80 °C until time of analysis. Bovine rumen samples were collected as previously reported in ref. [233]. Chicken caeca were collected as previously reported in ref. [170].

8.5.2 Sequencing

Swine and cattle samples were sequenced using PCR product from PCR specific primers flanking the V1-V3 region of bacterial 16S rDNA [242]. The forward fusion primers for pyrosequencing included 454 Life Science's A adapter, and barcode A fused to the 5' end of the V1 primer 27F. In chicken the V3 primer 341F was used. In all samples, the reverse fusion primer included 454 Life Science's B adapter (lowercase) fused to 5' end of V3 primer 534R. The fragments in the amplicon libraries were subjected to a single pyrosequence run from the V3 primer end using a 454 Life Science Genome Sequencer GS FLX (Roy J. Carver Biotechnology Center, University of Illinois). The reads for chicken 1, cattle 1 and 2 and swine 1 and 2 have been deposited to the NCBI Sequence Reads Archive with the accession number SRA052136.3. Chicken 94 reads have been previously deposited as reported in ref. [7].

8.5.3 Rank-abundance, Species-abundance, Preston Plots and Taxa Distributions

The reads from cattle and swine microbiomes were cleaned up using the method recommended in ref. [139]. For the chicken caecum microbiome we removed all sequences shorter than 100 bp. The ends of all reads were trimmed so that the sequences start and end in the same place in the 16S rRNA consensus structure. All remaining sequences were then aligned using the method described in ref. [7]. The OTUs were clustered using complete linkage [133] 3% sequence identity with the denominator 4 from [145] (counting indels as differences). The OTU abundance data for rank-abundance was then binned into a histogram using the method in Adami and Chu [243]. Species-abundance and Preston plots were generated following ref. [244]. Neutral model curves were generated using the algorithm for the sampling organisms from a neutral metacommunity [118]. Hubbell's θ parameter was fixed to match the exponentially decaying tail of the rank abundance. Offset was chosen by a least-squares method. Taxonomy assignments and comparison of libraries was made with the Library Compare tool [245] at RDP [152].

8.5.4 PCA Ordination

In Fig. 8.10 we show the results of Principal Component Analysis on our OTU data. In performing this calculation, each OTU was associated with a vector of real numbers of dimension $4L$ where L is the length of the multiple alignment. The elements of the vectors were calculated in the following way. Each nucleotide within the multiple alignment is represented by a sub-vector of 4 numbers, A is $(1, 0, 0, 0)$, C is $(0, 1, 0, 0)$, G is $(0, 0, 1, 0)$, T is $(0, 0, 0, 1)$, whereas the gap is represented as $(0, 0, 0, 0)$. The vector associated with the OTU is then the arithmetic average of the vectors associated with each sequence within the OTU. We then perform the weighted PCA procedure [240] where we weigh each OTU by its abundance.

8.5.5 Closest-distance metric

We used the percent sequence distance metric in Fig. 8.13. The randomized dataset (red line) was generated in the following way. Each OTU (with its associated abundance) was replaced by a representative randomized sequence. This sequence was generated by selecting each nucleotide from a distribution of probabilities generated from the sequence reads. In this way, the base pair distribution for each position in the multiple alignment of the model dataset is the same as that of the experimental dataset. Furthermore, since the abundances of OTUs are kept, the rank abundance of the model dataset is exactly the same as that of the experimental dataset.

Chapter 9

Theoretical models of ecosystems with energy flow

In the previous chapter, we saw that complex gastrointestinal microbial ecosystems in several vertebrates are clearly shaped by non-neutral forces. One of the most widely-used ways to model this state of affairs is by using models of population ecology. These are typically ordinary differential equations describing the interactions between species and resulting in functional forms for their spatially-averaged time-dependent populations. These models usually involve abstract terms for species competition and species fitness. Our goal in this chapter is to make such models somewhat more concrete and applicable to microbial ecosystems in particular. To that end, we will acquire an energy-driven perspective of the ecosystem in the tradition of Odum [246].

In this chapter we consider marine plankton, organisms responsible for more than half of world's primary productivity, and important factors in global climate [247]. We will consider plankton living in ocean surface waters, a low-nutrient environment [248], characterized by a genomically unique microbiome [249]. The questions we answer are: What physical variables matter for the evolution of organisms in these specific conditions? Can we map the results of abstract models to the metagenomic data? In our model, we introduce concrete terms for energy balance and the cost for maintenance of the genome. Through our model we intend to characterize the phase transition between the observed modes of survival in this environment: cooperative specialization and opportunistic generalism.

We describe a niche structuring model in Section 9.1 to introduce our energy-centric approach. We will not turn our back to experimental data completely; in Section 9.2 we will introduce the many interesting genomic features of the world of ocean surface plankton. These features will motivate our model in Section 9.3. This model has some qualitative aspects of plankton microbe systems: we describe these in Section 9.4. The model shows genome streamlining (Section 9.4.1) and spontaneous symmetry breaking (Section 9.4.2) consistent with experimental observations. In Section 9.4.3, we consider the model in fluctuating non-equilibrium conditions. Finally, we conclude in Section 9.5 by describing the future of energy-driven ecosystem modeling.

9.1 Niche structuring through emergent community assembly

One of the models we developed, while working on the niche-neutral dichotomy, is the “energy–cascade” model. This model is interesting because it is a direct example of a system that produces neutral-like abundance patterns, even though it has explicit interactions. Here, we are interested to model the cascading energy flow in an ecosystem. Different trophic levels are created when the organisms symbiotically adjust to consuming side-products of other organisms in the ecosystem (see Figure 9.1).

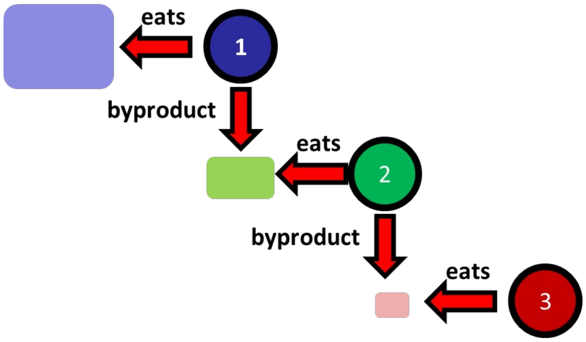


Figure 9.1: Schematic diagram of the niche structuring model.

Specifically, in our artificial life model, both the organisms and the available “units of energy” are given a “genome” g. There is explicit competition between the organisms for the consumption of energy, hence this model is non-neutral. If organisms are unable to find energy to consume, they die at a constant rate. The probability for an organism with genome g_o to consume a unit of energy with genome g_e is given by $F(\epsilon(g_o - g_e))$ where $g_o - g_e$ denotes the distance of the two genomes in some metric, ϵ denotes the strength of the interaction, and F denotes the fitness function. Every time a unit of energy is consumed by an organism, another unit of energy with genome g_p may be created with probability p, as a side product. The genome of the product is $g_p = G(g_o, g_e)$, where G is a deterministic function of the organism and input energy genomes. For example G could be a hash function of g_o and g_e. In this way, the organisms can create chains of producers and consumers, and possibly even cycles. They can self-organize into niches that can further influence how they evolve. In that sense, this model is in the “niche construction” class [250].

Interestingly, simulations of abundance patterns with this model show curves that are of the form given by neutral theory. Namely, we can calculate steady-state abundance distribution for various ϵ (the strength of the competition), with F an exponential function. The genome is a binary string of 100 bits, and G is the bitwise inverse function of g_e. The resulting abundance patterns in steady-state are displayed in Figure 9.2. Whereas the time to reach the steady-state depends on ϵ, the final abundance distribution does not, and it

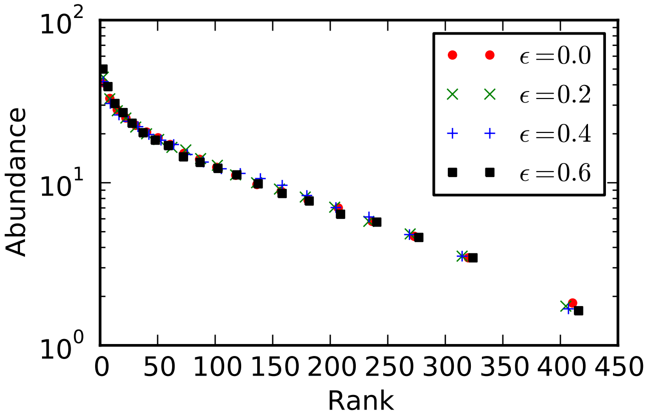


Figure 9.2: Rank abundance curves obtained from a model interacting community. The shape of the curves is described by neutral theory.

has the same form as abundance distributions we've seen in vertebrate gastrointestinal systems in Chapter 8 (see Figure 8.6).

9.2 Microbiome of the Ocean Surface Plankton

The model in Section 9.1 is interesting because it produces neutral-like curves that are observed in ecological measurements of rank abundances. But, we have shown that the rank abundance measures are known not to be good discriminators unless supplemented with genomic data (see Chapter 8). What other questions about existing metagenomic systems can we answer using the energy-driven approach?

In recent years, metagenomic approaches have given us an understanding of many specific single environments, but also of spatially extended systems such as soil [251]. Perhaps the most extensive measurement of a spatially extended system is the Global Ocean Survey (GOS) [252], a metagenomic survey of 52 sites in the Atlantic, Pacific and Indian ocean taken on board of the ship *Sorcerer II*.

The authors of GOS collected samples of marine picoplankton (microbes in the $0.1\mu\text{m} - 3.0\mu\text{m}$ range) and generated 10.97 million metagenomic reads, assembled into 6.12 million protein-coding genes [253] and more than 45,000 16S sequences, The data consisted of whole genome assemblies of 197 marine genomes (of which 137 were previously sequenced). To understand the abundance distribution of the picoplankton, Yooseph *et al.* [249] devised a "recruitment" procedure of the metagenomic reads to whole marine genomes. For each metagenomic read a BLAST [254] is performed against each genome. The read is then recruited to the highest scoring hit if BLAST matches at least 50% of the nucleotides. Less than 25% of the reads are recruited in this fashion to the genomes, indicating that the marine microbiome is still far from fully understood. Yooseph *et al.* defined the depth of coverage as the average number of reads covering a base pair in the reference genomes. Of the 197 genomes, 34 are covered by the recruited reads with a depth of coverage > 1. These are termed the HRG (high recruiting genomes) and the rest are the LRG (low recruiting genomes).

The 197 genomes of Yooseph *et al.* are plotted in Figure 9.3(a) in terms of their recruitment coverage and number of protein-coding genes in the genome. Yooseph *et al.* realized that the highly recruiting genomes tend to be the short ones (with small number of protein-coding genes). This indicates that the abundant organisms living in the picoplankton community have *streamlined* genomes, i.e. they seem to shed unnecessary genes and specialize. Yooseph *et al.* also provide evidence that these organisms may be physically small as well [249]. Since HRG organisms have been found at all GOS sampling sites, they have been termed CAPs (cosmopolitan abundant populations) by Yooseph *et al.*.

To gain some understanding into the functions present in the marine genomes, Yooseph *et al.* provided a mapping of the protein-coding genes present in the genomes to the COG (clusters of orthologous groups) database [255] and the KEGG (Kyoto Encyclopedia of Genes and Genomes) module database [256]. Two genes that map to the same COG indicates that those two genes are orthologous based on phylogenetic data, whereas a mapping of a gene to a particular KEGG module indicates that the gene is a part of a particular metabolic pathway. However, not all genes map to COG and KEGG with sufficient (50%) identity, but we find that of those that do map, their number is still inversely proportional to recruitment coverage: see Figure 9.3(b) and (c).

We can use the above COG and KEGG mappings as a proxy for answering the question of what functions are available to the microbes: we will assume that each COG/KEGG module provides a different function to the organism. For the elements of the HRG dataset, we plot a presence/absence map of each COG/KEGG module in Figure 9.5. If at least one gene is found to match a COG or a KEGG module, then that function is considered present (and colored red in the figure), otherwise it is unavailable to the organism (and colored blue in the figure). Notice that in this map the order of the rows does not matter (we do not sort the genomes by any specific measure). Similarly, the order of the columns of the map has no meaning. This means that we can use a procedure by which we can make the visualization of the presence/absence map easier to interpret. In this chapter, we will sort the map's rows by similarity. The procedure is the following. First, a random row is chosen to be at the top. Then, a similarity (Hamming distance) of this row to each other row is computed. The most similar row is placed to be second in the map and it is then compared to all other remaining rows, etc. After sorting the map by rows, we proceed to do the same for columns. For the case of a random map, the sorting procedure has no effect–the map still has random appearance (See

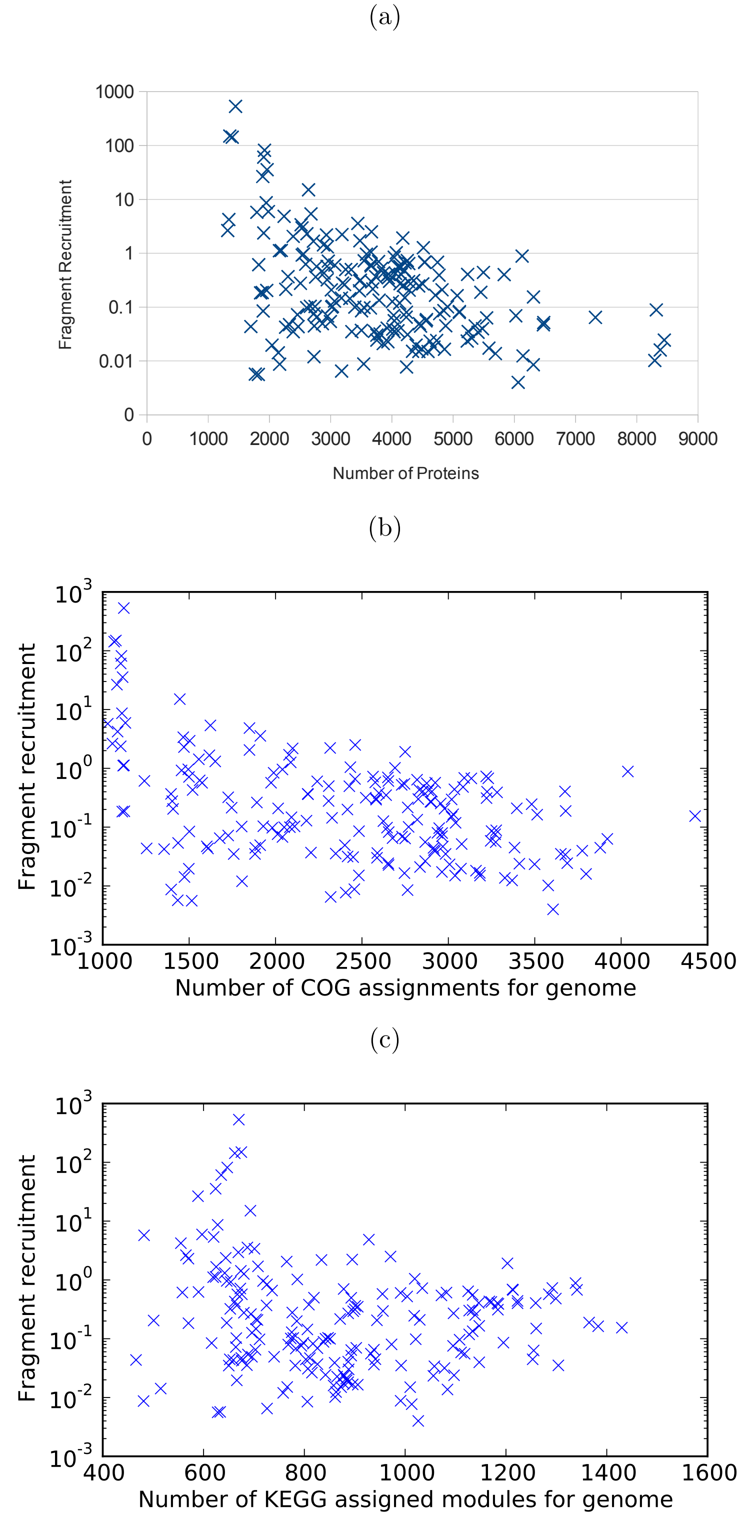


Figure 9.3: Fragment recruitment coverage as function of (a) number of protein-coding genes in the genome, (b) number of protein-coding genes with a match to COG, (c) number of protein-coding genes with a match to KEGG. This figure has been plotted from data by Yooseph *et al.* [249].

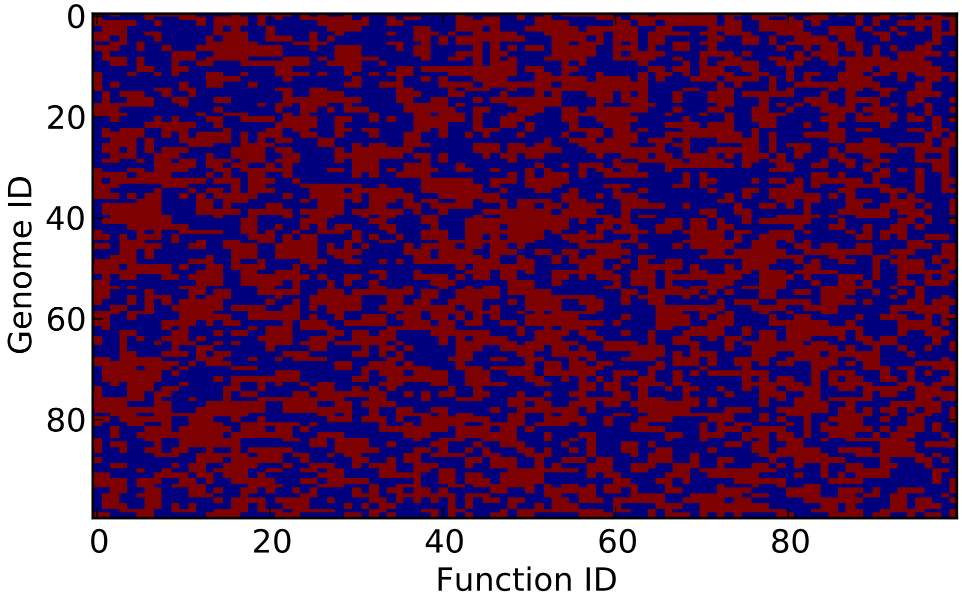
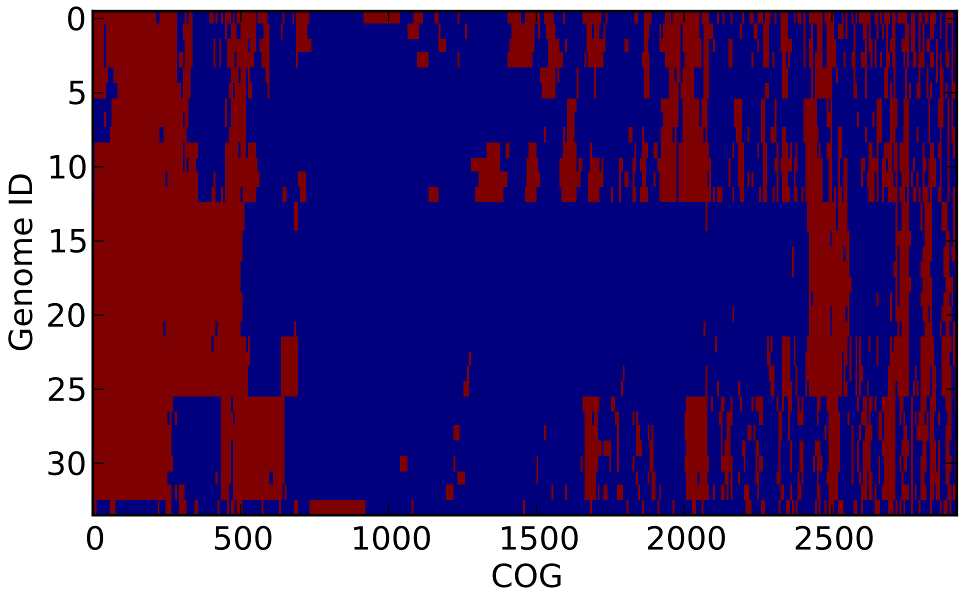


Figure 9.4: Plot of presence (red) or absence (blue) of functions for a random distribution. Rows and columns have been sorted by similarity but this doesn't destroy the random appearance.

Figure 9.4). However, suppose that there are lineages in the dataset that have a certain set of functions, and lineages that don't. In this case, we should expect to see a block of presences and block of absences in the map for the columns in question. In fact, this is precisely the what we observe in Figure 9.5. This indicates that the CAPs specialize, in agreement with the fact that they have streamlined genomes.

The correlated relationship of number of genes that match to COGs or KEGG modules to the genome coverage from Figure 9.3 is not as pronounced when the number of unique COGs and KEGG modules is considered (see Figure 9.6). In Figure 9.7 we show the presence/absence map for COGs and KEGG modules for all genomes (HRG and LRG). In the Figure, the HRGs are represented by teal color. No significant discernible difference can be observed between LRG and HRG from the figure, but the LRGs do tend to have more functions than the HRGs. This idea has led Yooseph *et al.* to denote the LRGs as SAPs (stochastically abundant populations). The motivation is that, unlike the streamlined and specialized CAPs, the SAPs preserve large genomes with wide functionality which allow them to make use of changing conditions and favorable fluctuating circumstances should they arise. We now turn our attention to seeing if this sort of empirical result is predicted to be a generic outcome of an energy-based population model.

(a)



(b)

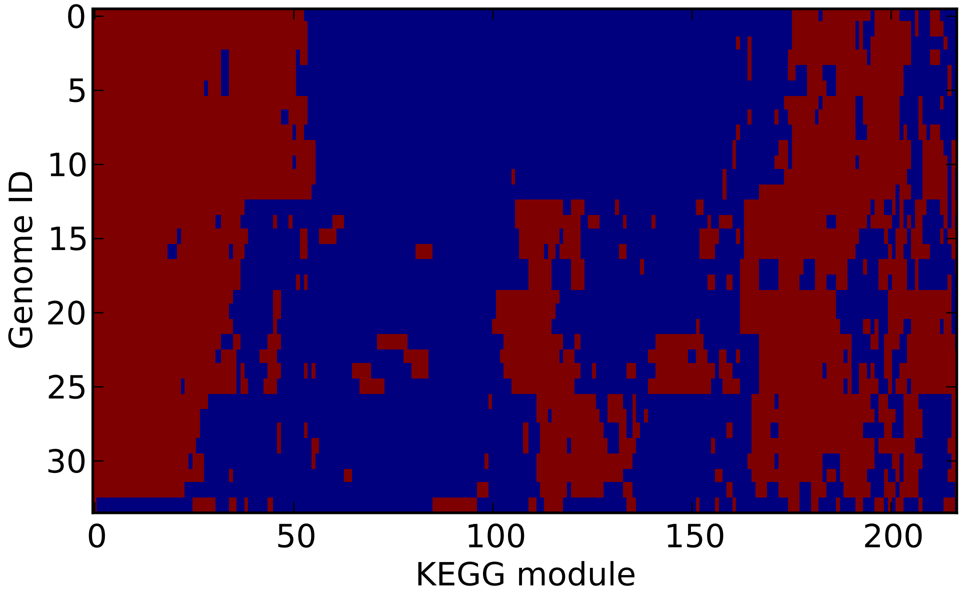
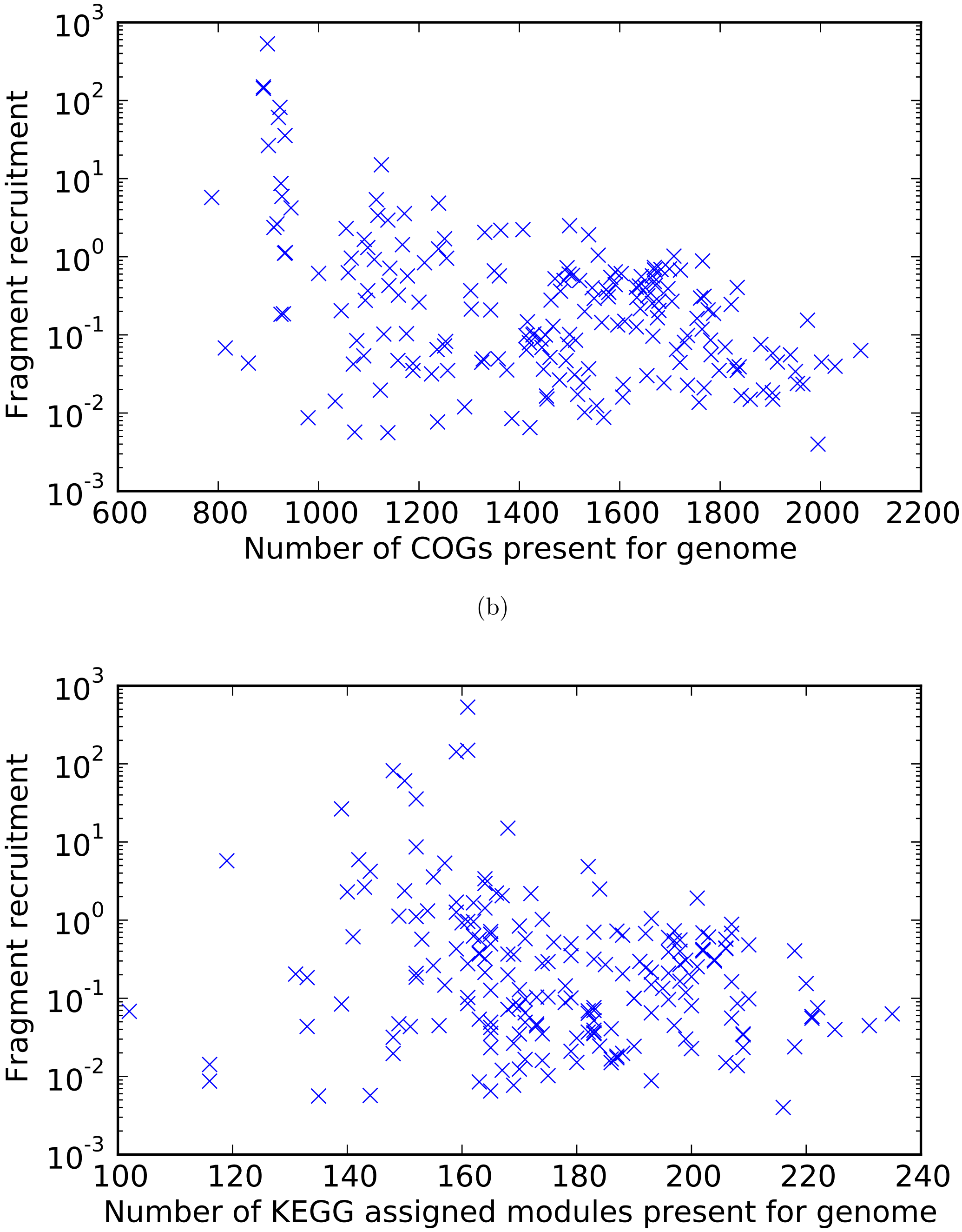


Figure 9.5: Presence (red) and absence (blue) of protein-coding genes that match (a) COGs and (b) KEGG modules in Yooseph *et al.* [249] HRG data.


(a)

(b)

Figure 9.6: Fragment recruitment coverage as function of (a) number of COGs matched by at least one gene, (b) number of KEGG modules matched by at least one gene. Figure plotted from Yooseph *et al.*'s [249] data.

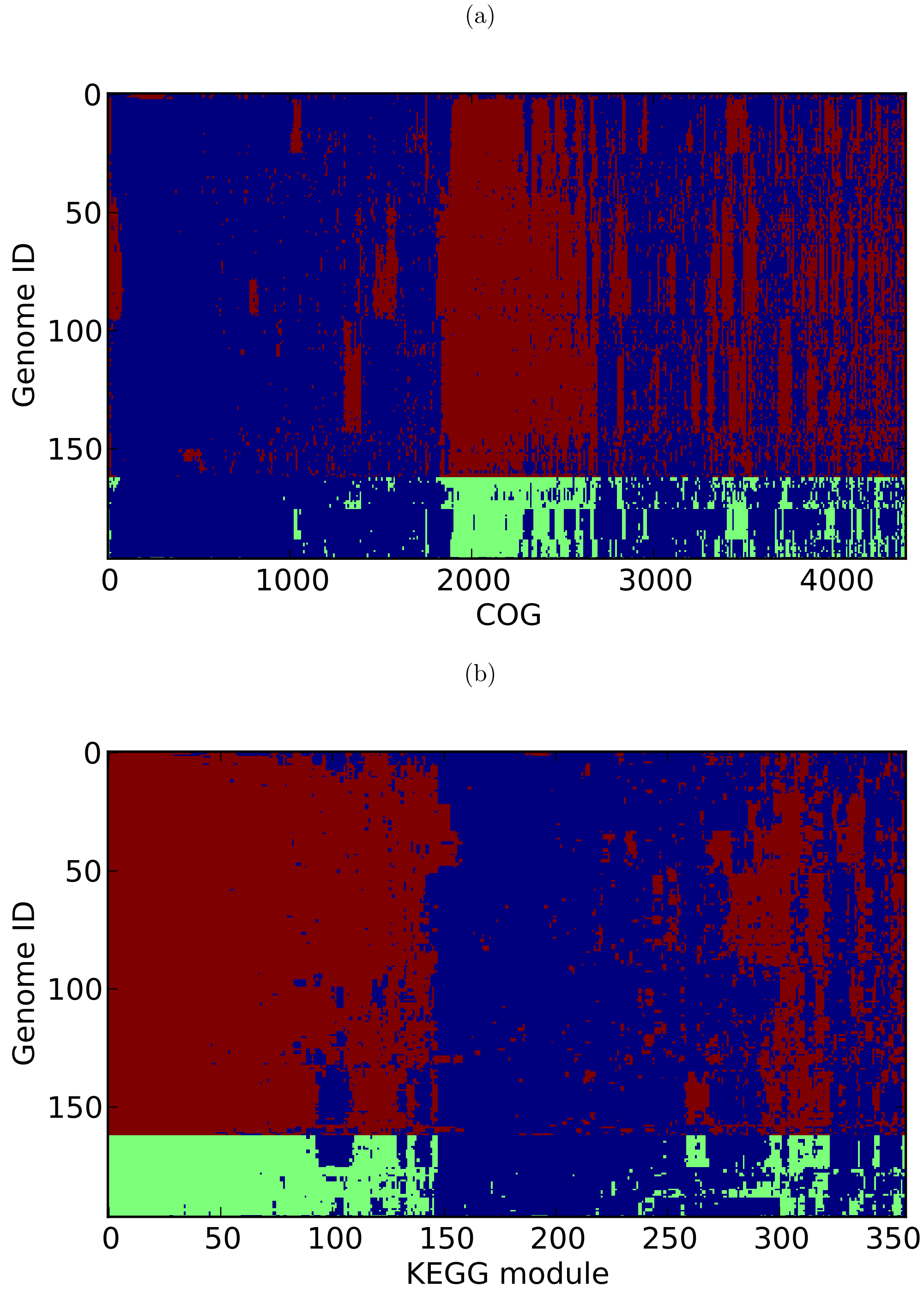


Figure 9.7: Presence (red or teal) and absence (blue) of protein-coding genes that match (a) COGs and (b) KEGG modules in Yooseph *et al.* [249] data. Here, teal represent HRG (genomes with recruitment coverage > 1).

9.3 Model of an Interacting Community

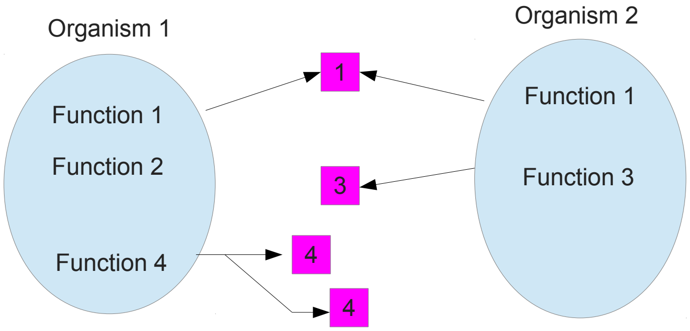


Figure 9.8: Schematic diagram of the function presence/absence model. Here a presence of a function in the genome of an organism allows the organism (displayed as a blue oval) to access nutrients (purple squares).

Our model is schematically illustrated in Figure 9.8. In this model, each organism has a genome that consists of up to N possible functions denoted by f_i^n (in this chapter, $N = 100$ in all simulations). Here n is a label of the organism. The $f_1^n, \ldots, f_N^n$ can each be 0 (function absent) or 1 (function present). There are also N different possible types of nutrients (units of energy). If an organism n has a function f_i^n, then it competes with other organisms with that function for the consumption of the nutrient i. Through fluctuations, the organisms can either compete for the same resources in which case they are divided equally (such is the case for nutrient 1 in the figure), or the organisms can consume resources unchallenged (nutrients 3 and 4 in the figure). Of course, a presence of a function $f_i^n = 1$ incurs a cost c for organism n ($c = 2$ in this chapter). The organisms die at unit rate, and their birth rate b^n is proportional to their net energy consumption:

$$b^n = I^n - c \sum_i^N f_i^n - C \tag{9.1}$$

where C is a fixed cost ($C = 0$ in this chapter) and I^n is the energy input for organism n:

$$I^n = \sum_i^N F_i(t) \left(f_i^n / \left(\sum_j^{M(t)} f_i^j \right) \right) \tag{9.2}$$

where $F_i(t)$ is the flux of nutrient i and $M(t)$ is the total number of organisms at time t.

When a birth occurs, the genome of the daughter organism is evolved according to a mutation rate μ. We consider two mutation strategies:

1. Any function f_i can switch from 0 to 1 and vice versa at uniform rate μ.

2. Any function with $f_i = 1$ can switch to $f_i = 0$ with uniform rate μ. However, a function f_i can only

be reacquired ($0 \to 1$) with probability μ if either of functions $f_{i\pm1} = 1$. Here, boundary conditions for the f_i are periodic.

We refer to mutation strategy 2 as the DP (directed percolation) rule. The motivation is that an organism shouldn't be able to reenable a function i unless it has a phylogenetically close function $i \pm 1$. For example, in a real biological system, a cell can acquire the ability to resist an antibiotic by acquiring an SNP mutation in cell membrane or ribosome proteins but this is usually precluded by a sequence of other SNP mutations that need to happen first. In the opposite sense, we term the mutation strategy 1 as HGT (horizontal gene transfer) rule. Rather than acquiring a function through piecewise steps, a function may arrive through HGT from another organism in the environment. In that case, any function is available to any organism at any time. In the simulations in this chapter, we set $\mu = 0.01$ per base per generation (except in Section 9.4.3). This value is not too big as to entropically wash out the results of the simulation, nor too small to make the simulations slow.

With the notation in this section, $f_i = 0$ (absent) or 1 (present), we can plot the maps of f_i for all organisms in the simulation and sort them in the same way as described in Section 9.2. We can then compare these simulation results directly with the presence and absence maps for COGs and KEGG modules in data from Yooseph *et al.*'s experiment shown in Figures 9.5 and 9.7 and we proceed to do that in the following section.

9.4 Results

In this Section, we introduce 3 results of running our model. The genomes undergo streamlining and they specialize (or spontaneously break symmetry) in order to optimize their energy use. However, non-streamlined genomes can still provide a benefit over the streamlined genomes in conditions where the energy flux is fluctuating.

9.4.1 Genome Streamlining

In this section we simulate our model with a constant energy flux $F_i(t) = 100 + 2\xi$ for $i = 1, \ldots, 100$ and ξ is a normally distributed random variable with unit standard deviation. We consider the average genome length (or number of present functions per genome):

$$\langle f(t) \rangle = \frac{1}{M(t)} \sum_{j}^{M} \sum_{i}^{N} f_i^j(t). \tag{9.3}$$

We run four simulations of the model, two starting from single lineage with all functions present and two starting from 100 lineages, each with 50 functions chosen at random. We also run the simulations for each of the two mutation strategies: HGT and DP. The results for $\langle f(t)\rangle$ are shown in Figure 9.9. We observe that the DP mutation strategy enhances genome streamlining. With DP, the genomes tend to streamline to about 1/6 of their full length. On the other hand, with HGT, the genomes spontaneously only streamline only up to 2/3 of their full length. When started from half-length, the genomes stay at that length and do not streamline further.

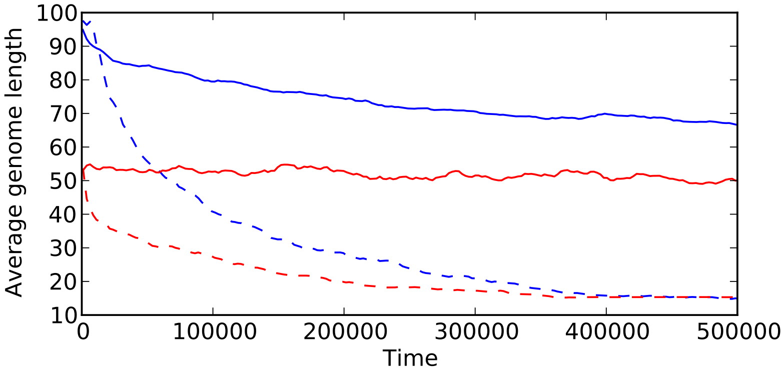
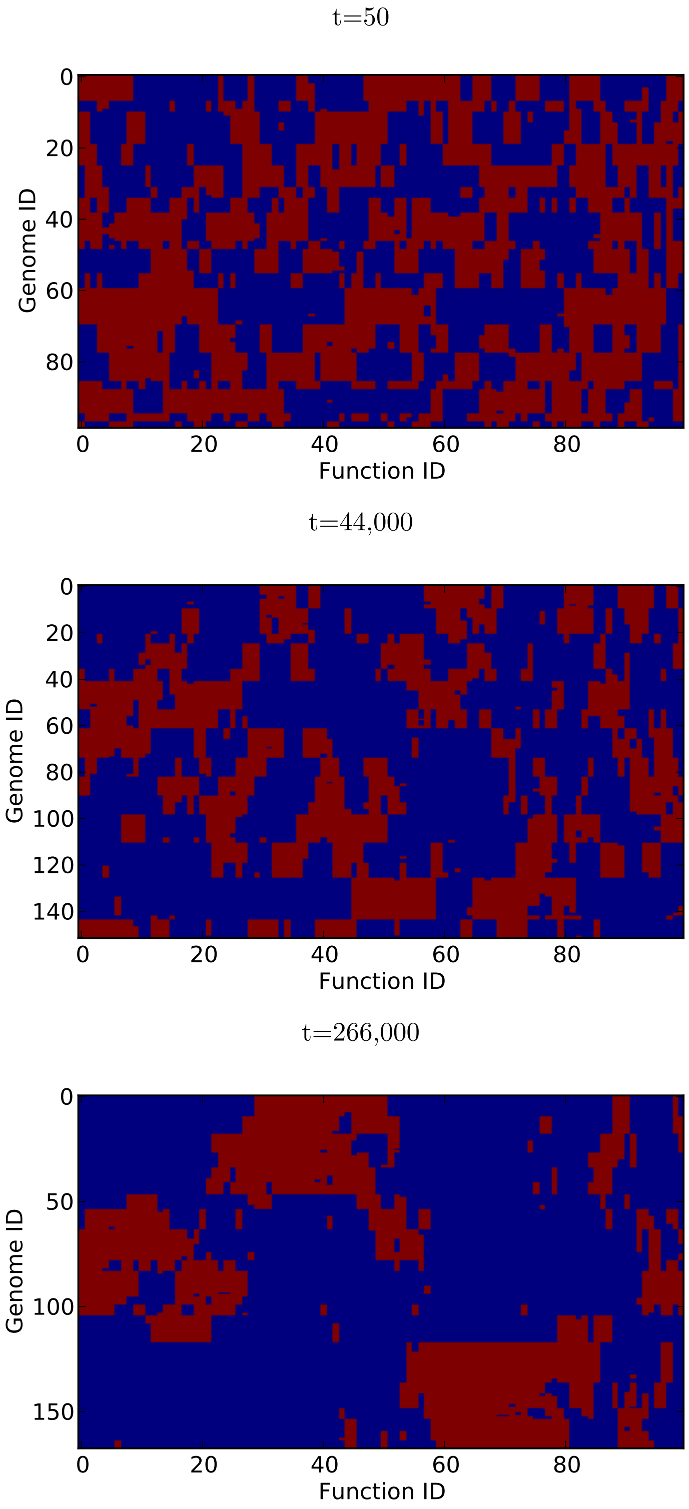


Figure 9.9: Time-trace of the average genome length in 4 simulations of process in Section 9.3. Two simulations are with all mutations allowed (solid lines) and two simulations are with DP-like mutations only (dashed). Two simulations begin with a single lineage with all functions available (blue), and two simulations begin with 100 lineages, each with 50 functions chosen at random (red).

9.4.2 Spontaneous Symmetry Breaking

For the case of DP-like mutations, and same constant energy flux, we observe that the genomes of the organisms in the simulations undergo a spontaneous symmetry breaking. The state of $f_i^n(t)$ is shown in Figure 9.10 for the case of 3 times. We see that, as t becomes large, the lineages specialize and stop competing with each other for most of the nutrients i. This transition is much like in ferromagnetic Ising systems below T_c where single-spin domains form. Here, however, f_i^n prefers to be different for different n and this leads to the emergent block structure observed in Figure 9.10. Notice also that, as time passes, and genomes become more streamlined and specialized, the carrying capacity (total number of organisms in the simulation) grows (from 50 organisms at $t = 0$ to more than 150 at $t = 266,000$).



t=50

t=44,000

t=266,000

Figure 9.10: Snapshots of presence (red) or absence (blue) of functions in digital genomes undergoing a simulation described in Section 9.3.

9.4.3 Energy strategies

The evolution of energy strategies of organisms (how resources are allocated towards growth and reserves), and the resulting population dynamics have been studied before [257]. In this section we study this problem in context of our model. We turn off the mutation, setting $\mu = 0$ and we consider the behavior of 5 lineages, each fixed in terms of their genome. First lineage has all 100 functions, whereas the other 4 lineages each have 25 non-overlapping lineages. Evidently, in the steady state case of previous 2 sections, the first lineage will be driven to extinction. The first lineage competes with all 4 other lineages, whereas the 4 lineages only compete with the first. Is there an environment in which this setup of lineages has benefits for the first lineage?

In this section we show that there is. We consider a sinusoidally fluctuating environment where $F_i(t) = 200 + 180\sin(2\pi t/P) + 2\xi$ for $i = 1, \ldots, 100$ where P is the period of the fluctuation and ξ is a normally distributed random variable with unit standard deviation. We find that for $P < 60$, the first lineage is able to outcompete the other lineages even though it is outcompeted in the steady-state condition. This is because, when the energy is plentiful (in the rising phase) the first lineage is able to get a bigger share of the surplus energy.

The results are shown in Figure 9.11 for three different P. The time trace of the first lineage is shown in blue, the other lineages are shown in red. We observe that for $P = 55$, the first lineage drives the other lineages to extinction. Similarly, for $P = 65$, the other lineages drive the first lineage to extinction. A metastable coexistence can be observed for $P = 60$ but we find that, given sufficiently long time of simulation, it will result in either of the lineages winning. These results, in the context of marine microbiome, indicate that whereas CAPs are more adapted to steady-state conditions, SAPs can outperform CAPs in fluctuating conditions.

9.5 Conclusion

Ocean picoplankton communities are presumed to be divided into 2 subcommunities: CAPs and SAPs. The CAPs are abundant, cosmopolitan specialists, with streamlined genomes and small cell size. The SAPs are rare, and they have large genomes. Yooseph *et al.* proposed that CAPs are undergoing a strategy of “cryptic escape”, whereby they avoid predators by keeping a small biomass. Our model has some properties of CAPs, namely it provides an emergent mechanism for genomes to streamline (Section 9.4.1) provided that the effects of HGT are not too strong. Our model also shows spontaneous symmetry breaking. In the past, symmetry breaking has been proposed as a method for allopatric speciation [258]. Our model, in

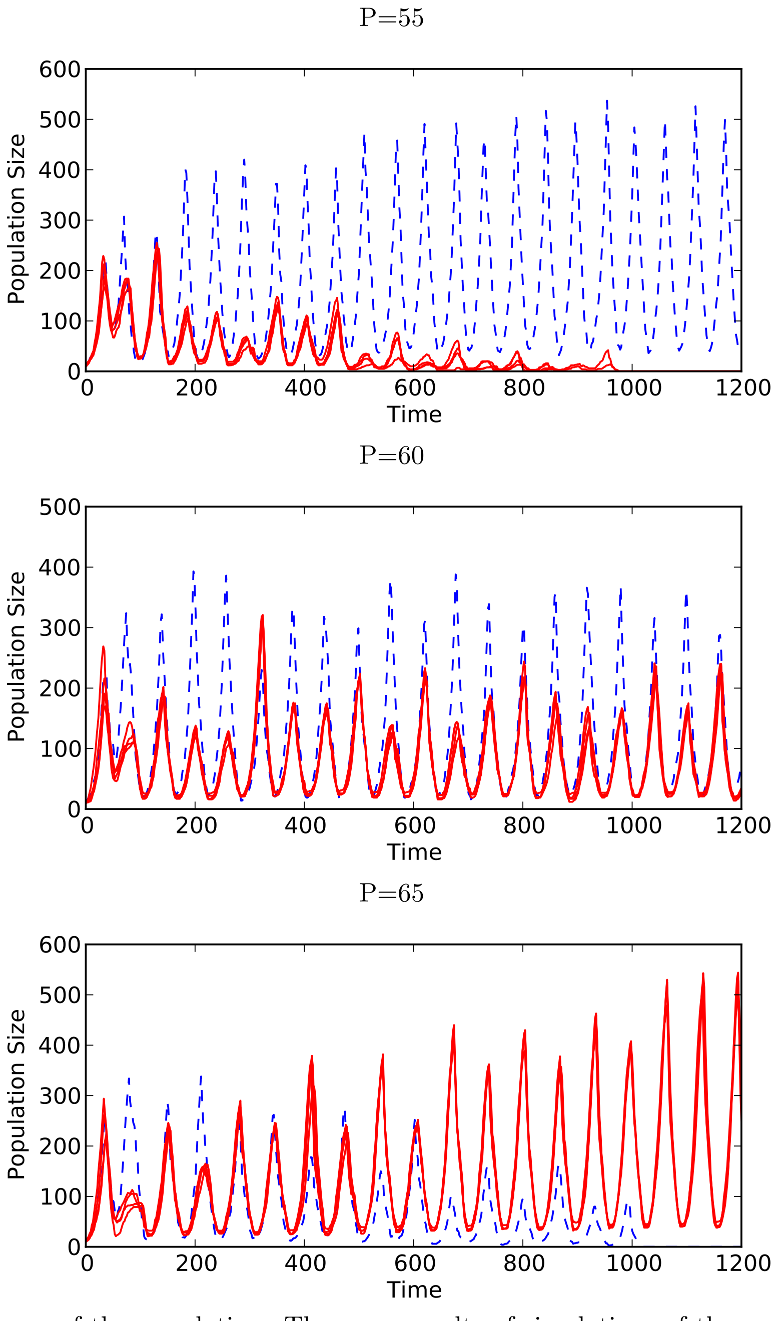


Figure 9.11: Time trace of the population. These are results of simulations of the model from Section 9.3 for the case of a fluctuating nutrient flux (with period P). There are 5 lineages, 1 with all possible functions (indicated with blue dashed line), and 4, each with a quarter of the available functions (labeled with the red lines).

which new lineages can arise through spontaneous symmetry breaking, is a concrete example of that. The specialization of CAPs has been observed in metagenomic data. Compare our results in Figure 9.10 with the data in Figure 9.5. Our model also shows that specialization is not always beneficial: SAPs can outperform the CAPs in case of fluctuating conditions.

At about the same time that we performed our analysis described in this chapter, Rogers, McKane and Rossberg published a mathematical model for a competing population, where the competition is proportional to Hamming distance between organisms [259]. They are able to analytically find that a pattern-forming instability forms, and they propose that this instability leads to genetic clustering, i.e. speciation. In addition, Rogers, McKane and Rossberg find that the pattern-forming instability is enhanced by the effects of demographic noise. We believe that our observations of symmetry breaking in this chapter are the same as those of Rogers, McKane and Rossberg.

The specialization and streamlining observed in CAPs may be indicative that they are syntrophic organisms [260] (they are extremely symbiotic, i.e. they share electrons or metabolites in order to process free energy) but there is yet no concrete evidence for syntrophy in microbial ocean systems. It is clear that the surface ocean waters are a nutrient-poor and energy-starved environment for the picoplankton and this may be the key driving force for the CAP/SAP phase transition. However, this phase transition remains not fully understood: more experimental data is necessary and a correct model may require a more elaborate, syntrophic description.

Chapter 10

Conclusions and reflections

10.1 Conclusion

In Part I, I gave evidence that the transition to turbulence may be in the directed percolation universality class. While the abstract phenomenological model of directed percolation cannot predict many of the structures observed in shear flows, the scaling of the lifetimes and growth rates of turbulent regions seem to be fully described by the model (Chapter 3). I also showed that the Burgers equation, a simplified model of the Navier-Stokes equations, may also undergo a transition in the directed percolation class (Chapter 4). These results imply that statistical, emergent features of fluid dynamics are sufficient to describe many of the important critical properties of the laminar-to-turbulent transition.

In Part II, I showed how to accurately and efficiently process and analyze metagenomic datasets (Chapters 6 and 7). I also wrote down simple models for ecological assembly (Chapter 8) and energy balance (Chapter 9) and showed how one can relate the conclusions of these models to real metagenomic data in vertebrate gastrointestinal and ocean surface microbiomes.

What about the future work? Throughout this thesis I have given a number of open problems and possible future avenues of research. Generally, in the field of transition to turbulence, many of the features of the transition in pipe flows are now considered fully understood. Of course, in turbulence, there is still much work to do. A full statistical theory of turbulence still doesn't exist, and the Millenium prize for the uniqueness and existence of solutions to Navier-Stokes equations is still not claimed (its value for physics is in any case small). In comparison, though, the field of quantitative biology is exploding. The amount of data generated and the number of problems still unresolved is staggering in comparison. The coming biomedical and technological applications will profoundly change our society and it will be exciting to witness these changes.

10.2 Perspective on Computational Biology

In the the first nine chapters of this thesis, I have given the results of my scientific endeavors for the past four and a half years. This thesis was to provide a starting point for further research and numerous references to relevant theory and experiments. In that way, another graduate student can pursue the conjectures given in chapters 4 and 9 for his or her own thesis work. In this section I hope that the reader will permit me a slight foray away from scientific rigor into a realm of opinions and speculation about the future as I give my admittedly young and audacious perspective on the future of computational biology and interdisciplinary research.

10.2.1 Better computational biology tools

Moore's law has followed the development of computing power for the past five decades [261]. At the same time, in the past two decades, the sequencing ability has outpaced Moore's law [262]. As such, it is plausible that $O(N)$ algorithms, such as the ones presented in Chapter 7 will be just as important in the future as they are now.

However, algorithmic challenge is only one aspect of computationally processing future data. Another big challenge is the human understanding and analysis. It is doubtful that the number of people trained to analyze this data will grow with Moore's law. Hence, changes have to be made in terms of how the analysis is performed. Typical approach in computational biology has been to develop pipelines and so-called "black boxes". These are automated tools that can process data, filter it and perform analyses. However, these typically have a slew of parameters associated with them and require a decent training of the operator to use correctly. In reality, in biological research it is common that the users of these tools do not have the time to learn how to correctly use them and set correct parameters. As such, processing artifacts such as the ones described in Chapter 6 are introduced into the analyses.

This problem is also evident in the peer review process of biological papers where the referees are not able to evaluate bioinformatic methods, and they do not insist that authors provide sufficient bioinformatic detail. This makes drawing comparisons with other studies very difficult. A set of standards for reporting bioinformatic analysis in papers should be established. I believe that each bioinformatic study should contain all the computer code including all the invocations of the tools and parameters used to perform the analysis.

There is another approach that could improve the current state of affairs. I believe that the developers of future bioinformatic tools must take the issues raised above into consideration as they design their tools. Instead of developing a pipeline, or a "black box", one should instead develop an expert system. An expert system is an adaptive system that verifies input data, assumptions, selects parameters, and gives a detailed report of precise reasons for the parameters and algorithms used [263] (much like a human expert performing the analysis would do). The added effort of developing such programs would be offset by their wider utility and better automation. This is a perspective that I hope more bioinformatic developers will consider in the future.

10.2.2 Abstractions in Computational Biology

Scientific development occurs through a sequence of abstractions. A research idea in one paper is an analysis method in the follow-up paper. Abstractions are important ideas in computer science as well, since they quickly enable development of large and complex computer systems. For instance, with today's cloud computing platforms, it is possible for a single person to run a web application serving millions of people. Similarly, a single researcher can have very many simulations running on a remote computing cluster, without having to worry about scheduling, or particularities of the hardware running the simulation.

Abstractions are particularly important in programming languages. They enable writing programs that would otherwise not be possible (due to complexity and time constraints). One cannot imagine writing a modern program without using arrays, yet those didn't exist until the high-level programming languages of the 1950's and 1960's such as FORTRAN, COBOL and ALGOL. In scientific programming, I believe the success of Mathworks' Matlab$^{\mathrm{TM}}$ and Wolfram Mathematica$^{\mathrm{TM}}$ are in part due to the suitable choice of abstractions. In Matlab$^{\mathrm{TM}}$, everything is a matrix, which makes many scientific and engineering problems easily representable in its syntax. In Mathematica$^{\mathrm{TM}}$, everything is a mathematical expression, making it easy to represent mathematical relations and mathematical operations such as factoring and integration.

In computational biology, there are no obvious data structures or abstractions. As such, people use general-purpose programming languages, with a library of code for common operations (such as loading data, doing a BLAST or an alignment). In practice, computational biological research is an endless game of converting data from one format to another, piping data through filtering steps, matching relations in one database to another one and plugging data into black boxes, all the while having various parameters and half-processed data floating around.

I don't have a good solution to this problem and I am not sure that it exists. It seems to me that a network should be the most fundamental data structure in computational biology, but there would also have to be some sort of an ontology associated with each network. In practice, building a good programming language for computational biology based on this idea is hard. Successful programming languages and frameworks are only successful if many people use them, and advantages would be gained when all databases and data are available in this "ontological network" framework. But of course, getting to this point is hard. My opinion is that this would make it possible to perform systems biology research much more efficiently. It would allow data mining of relationships that are not readily evident at this time. But, it is also possible that the gains would not be worth the effort. In any case, it is encouraging to know that there is at least one product, NextBio$^{\text{TM}}$, that seems to attempt something similar to this approach in biomedicine.

10.2.3 Scientific method in interdisciplinary research

Another problem is the understanding of scientific research. There were an estimated 1.3 million papers published in 2006 alone [264]. It is impossible for any one person to read all papers in their field, let alone comprehend them. Staying in touch with literature is especially difficult in biological research where many major results are made by making connections across different scientific disciplines. Whereas internet has helped in tracing citations and locating papers, more can be done along these lines. One powerful approach is the semantic web [265]. I hope for an era where scientific papers will have semantic metadata attached to them, allowing data mining programs and simulations to make meaningful connections across disciplines quickly.

Any improvements to the process of performing research can lead to a faster progress of science as a whole. As such, it is surprising that the semantic web development is running so slow. Nevertheless, I remain hopeful that the future will bring many improvements for which I have too little perspective to predict, except to anticipate that among them will be the ideas presented in this chapter. I dare to hope that the future scientist will be like Hari Seldon from Isaac Asimov's Foundation [266]. Using a virtual reality-like device, he or she will be manipulating and tweaking relationships, equations and data at a high level, and considering problems outside of our reach at the present.

References

[1] Y. Pomeau. Front motion, metastability and subcritical bifurcations in hydrodynamics. *Physica*, 23D:3–11, 1986.

[2] Björn Hof, Alberto de Lozar, Dirk Jan Kuik, and Jerry Westerweel. Repeller or attractor? selecting the dynamical model for the onset of turbulence in pipe flow. *Phys. Rev. Lett.*, 101(21):214501, Nov 2008.

[3] KR Sreenivasan and R. Ramshankar. Transition intermittency in open flows, and intermittency routes to chaos. *Physica D: Nonlinear Phenomena*, 23(1-3):246–258, 1986.

[4] K.G. Wilson. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Physical Review B*, 4(9):3174, 1971.

[5] N I Lebedev and Y C Zhang. On the connection between directed percolation and directed polymers. *Journal of Physics A: Mathematical and General*, 28(1):L1–L6, 1995.

[6] M. Sipos and N. Goldenfeld. Directed percolation describes lifetime and growth of turbulent puffs and slugs. *Physical Review E*, 84(3):035304, 2011.

[7] M. Sipos, P. Jeraldo, N. Chia, A. Qu, A.S. Dhillon, M.E. Konkel, K.E. Nelson, B.A. White, and N. Goldenfeld. Robust computational analysis of rrna hypervariable tag datasets. *PLOS ONE*, 5(12):e15220, 2010.

[8] P. Jeraldo, M. Sipos, N. Chia, J.M. Brulc, A.S. Dhillon, M.E. Konkel, C.L. Larson, K.E. Nelson, A. Qu, L.B. Schook, et al. Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc. Natl. Acad. Sci. USA*, 109(25):9692–9698, 2012.

[9] J. Pfitzner. Poiseuille and his law. *Anaesthesia*, 31(2):273–275, 1976.

[10] S.P. Sutera and R. Skalak. The history of poiseuille's law. *Annual Review of Fluid Mechanics*, 25(1):1–20, 1993.

[11] B. Hof, A. Juel, and T. Mullin. Scaling of the turbulence transition threshold in a pipe. *Phys. Rev. Lett.*, 91:244502, Dec 2003.

[12] O. Reynolds. An experimental investigation of the circumstances which determine whether the motion of water shall be direct or sinuous, and of the law of resistance in parallel channels. *Phil. Trans. Roy. Soc. A*, 174:935–982, 1883.

[13] D. Jackson and B. Launder. Osborne reynolds and the publication of his papers on turbulent flow. *Annu. Rev. Fluid Mech.*, 39:19–35, 2007.

[14] I. Wygnanski and F. H. Champagne. On transition in a pipe. part 1. the origin of puffs and slugs and the flow in a turbulent slug. *J. Fluid Mech.*, 59:281–335, 1973.

[15] I. Wygnanski, M. Sokolov, and D. Friedman. On transition in a pipe. part 2. the equilibrium puff. *J. Fluid Mech.*, 59:283–304, 1975.

[16] H. Salwen, F.W. Cotton, and C.E. Grosch. Linear stability of Poiseuille flow in a circular pipe. *J. Fluid Mech.*, 98(02):273–284, 1980.

[17] A. Meseguer and L.N. Trefethen. Linearized pipe flow to Reynolds number 10^7. *J. Comput. Phys.*, 186(1):178 – 197, 2003.

[18] Steven H. Strogatz. *Nonlinear Dynamics and Chaos with Applications to Physics, Biology, Chemistry and Engineering.* Perseus Books, 1994.

[19] H. Faisst and B. Eckhardt. Traveling waves in pipe flow. *Phys. Rev. Lett.*, 91(22):224502, 2003.

[20] H. Wedin and RR Kerswell. Exact coherent structures in pipe flow: travelling wave solutions. *J. Fluid Mech.*, 508(333-371):2–5, 2004.

[21] B. Eckhardt, H. Faisst, A. Schmiegel, and T.M. Schneider. Dynamical systems and the transition to turbulence in linearly stable shear flows. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1868):1297–1315, 2008.

[22] B. Hof, C.W.H. van Doorne, J. Westerweel, F.T.M. Nieuwstadt, H. Faisst, B. Eckhardt, H. Wedin, R.R. Kerswell, and F. Waleffe. Experimental observation of nonlinear traveling waves in turbulent pipe flow. *Science*, 305(5690):1594–1598, 2004.

[23] A. de Lozar, F. Mellibovsky, M. Avila, and B. Hof. Edge state in pipe flow experiments. *Phys. Rev. Lett.*, 108(21):214502, 2012.

[24] RR Kerswell. Recent progress in understanding the transition to turbulence in a pipe. *Nonlinearity*, 18(6):R17–R44, 2005.

[25] B. Eckhardt, T.M. Schneider, B. Hof, and J. Westerweel. Turbulence transition in pipe flow. *Annu. Rev. Fluid Mech.*, 39:447–468, 2007.

[26] Bruno Eckhardt. Introduction. Turbulence transition in pipe flow: 125th anniversary of the publication of Reynolds' paper. *Phil. Trans. Roy. Soc. A*, 367(1888):449–455, 2009.

[27] Kerstin Avila, David Moxey, Alberto de Lozar, Marc Avila, Dwight Barkley, and Bjrn Hof. The onset of turbulence in pipe flow. *Science*, 333(6039):192–196, 2011.

[28] David Moxey and Dwight Barkley. Distinct large-scale turbulent-laminar states in transitional pipe flow. *Proc. Natl. Acad. Sci. USA*, 107(18):8091–8096, 2010.

[29] B. Hof and M. Avila and A. de Lozar and K. Avila. Onset of sustained turbulence in pipe flow. Talk given at Turbulence conference in Eilat, Israel, 2010.

[30] D. Samanta, A. de Lozar, and B. Hof. Experimental investigation of laminar turbulent intermittency in pipe flow. *Arxiv preprint arXiv:1008.2294*, 2010.

[31] Bruno Eckhardt. A critical point for turbulence. *Science*, 333(6039):165–166, 2011.

[32] A. de Lozar and B. Hof. Universality at the onset of turbulence in shear flows. *arXiv preprint arXiv:1001.2481*, January 2010.

[33] RG Deissler. Derivation of the navier-stokes equation. *American Journal of Physics*, 44:1128–1130, 1976.

[34] S.B. Pope. *Turbulent flows.* Cambridge university press, 2000.

[35] LF Richardson. *Weather prediction by numerical processes.* Cambridge University Press, 1922.

[36] A.N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 434(1890):9–13, 1991.

[37] HL Grant, RW Stewart, and A. Moilliet. Turbulence spectra from a tidal channel. *J. Fluid Mech.*, 12(02):241–268, 1962.

[38] R.H. Kraichnan. Inertial ranges in two-dimensional turbulence. *Physics of Fluids*, 10(7):1417–1423, 1967.

[39] G. Batchelor. *The Theory of Homogeneous Turbulence.* Cambridge University Press, Cambridge, England, 1982.

[40] T. Tran, P. Chakraborty, N. Guttenberg, A. Prescott, H. Kellay, W. Goldburg, N. Goldenfeld, and G. Gioia. Macroscopic effects of the spectral structure in turbulent flows. *Nature Physics*, 6(6):438–441, 2010.

[41] Malte Henkel, Haye Hinrichsen, and Sven Lubeck. *Non-Equilibrium Phase Transitions, Vol. 1.* Theoretical and Mathematical Physics. Springer, 2008.

[42] S.R. Broadbent and J.M. Hammersley. Percolation processes i. crystals and mazes. In *Proceedings of the Cambridge philosophical society*, volume 53, pages 629–641, 1957.

[43] E.V. Albano. Spreading analysis and finite-size scaling study of the critical behavior of a forest fire model with immune trees. *Physica A: Statistical Mechanics and its Applications*, 216(3):213–226, 1995.

[44] S. Davis, P. Trapman, H. Leirs, M. Begon, and JAP Heesterbeek. The abundance threshold for plague as a critical percolation phenomenon. *Nature*, 454(7204):634–637, 2008.

[45] JD Cowan, J. Neuman, and W. van Drongelen. Self-organized criticality in a network of interacting neurons. *arXiv preprint arXiv:1209.3829*, 2012.

[46] H. Hinrichsen, A. Jiménez-Dalmaroni, Y. Rozov, and E. Domany. Flowing sand: A physical realization of directed percolation. *Phys. Rev. Lett.*, 83(24):4999–5002, 1999.

[47] H. Hinrichsen. On possible experimental realizations of directed percolation. *Brazilian Journal of Physics*, 30(1):69–82, 2000.

[48] P. Grassberger. On phase transitions in Schlögl's second model. *Zeitschrift für Physik B Condensed Matter*, 47(4):365–374, 1982.

[49] H.K. Janssen. On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. *Zeitschrift für Physik B Condensed Matter*, 42(2):151–154, 1981.

[50] Iwan Jensen. Low-density series expansions for directed percolation: I. a new efficient algorithm with applications to the square lattice. *Journal of Physics A: Mathematical and General*, 32(28):5233, 1999.

[51] Kazumasa A. Takeuchi, Masafumi Kuroda, Hugues Chate, and Masaki Sano. Directed percolation criticality in turbulent liquid crystals. *Phys. Rev. Lett.*, 99:234503, 2007.

[52] Kazumasa A. Takeuchi, Masafumi Kuroda, Hugues Chate, and Masaki Sano. Experimental realization of directed percolation criticality in turbulent liquid crystals. *Phys. Rev. E*, 80:051116, 2009.

[53] P Manneville. Spatiotemporal perspective on the decay of turbulence in wall-bounded flows. *Phys. Rev. E Rapid Communications*, 79:025301(R), 2009.

[54] Ashley P. Willis and Rich R. Kerswell. Critical behavior in the relaminarization of localized turbulence in pipe flow. *Phys. Rev. Lett.*, 98(1):014501, Jan 2007.

[55] Holger Faisst and Bruno Eckhardt. Sensitive dependence on initial conditions in transition to turbulence in pipe flow. *J. Fluid Mech.*, 504(1):343–352, 2004.

[56] Haye H. Hinrichsen. Non-equilibrium critical phenomena and phase transitions into absorbing states. *Advances in Physics*, 49(7):815–958, 2000.

[57] Mina Nishi, Bülent Ünsal, Franz Durst, and Gautam Biswas. Laminar-to-turbulent transition of pipe flows through puffs and slugs. *J. Fluid Mech.*, 614:425–446, 2008.

[58] Eytan Domany and Wolfgang Kinzel. Equivalence of cellular automata to ising models and directed percolation. *Phys. Rev. Lett.*, 53(4):311–314, Jul 1984.

[59] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Cambridge Phil. Soc.*, 24:180–190, 1928.

[60] Nigel Goldenfeld, Nicholas Guttenberg, and Gustavo Gioia. Extreme fluctuations and the finite lifetime of the turbulent state. *Phys. Rev. E*, 81(3):035304, Mar 2010.

[61] William Feller. *An Introduction to Probability Theory and Its Applications.* John Wiley and Sons, 1968.

[62] Martin Z. Bazant. Largest cluster in subcritical percolation. *Phys. Rev. E*, 62(2):1660–1669, Aug 2000.

[63] A. Hansen and EL Hinrichsen. Some remarks on percolation. *Physica Scripta*, T44:55–61, 1992.

[64] K.T. Allhoff and B. Eckhardt. Directed percolation model for turbulence transition in shear flows. *Fluid Dynamics Research*, 44(3):031201, 2012.

[65] D. Barkley. Simplifying the complexity of pipe flow. *Physical Review E*, 84(1):016309, 2011.

[66] H. Chaté and P. Manneville. Transition to turbulence via spatio-temporal intermittency. *Phys. Rev. Lett.*, 58(2):112–115, 1987.

[67] JM Burgers. On the resistance experienced by a fluid in turbulent motion. In *Proc. KNAW*, volume 26, pages 582–604, 1923.

[68] J. Bec and K. Khanin. Burgers turbulence. *Physics Reports*, 447(1):1–66, 2007.

[69] J.M. Burgers. *The nonlinear diffusion equation: asymptotic solutions and statistical problems.* D. Reidel Publishing Company, 1974.

[70] J.P. Bouchaud, M. Mézard, and G. Parisi. Scaling and intermittency in burgers turbulence. *Physical Review E*, 52(4):3656, 1995.

[71] D. Forster, D.R. Nelson, and M.J. Stephen. Large-distance and long-time properties of a randomly stirred fluid. *Physical Review A*, 16(2):732, 1977.

[72] E. Medina, T. Hwa, M. Kardar, and Y.C. Zhang. Burgers equation with correlated noise: Renormalization-group analysis and applications to directed polymers and interface growth. *Physical Review A*, 39(6):3053, 1989.

[73] M. Kardar, G. Parisi, and Y.C. Zhang. Dynamic scaling of growing interfaces. *Phys. Rev. Lett.*, 56(9):889–892, 1986.

[74] D.A. Huse and C.L. Henley. Pinning and roughening of domain walls in ising systems due to random impurities. *Phys. Rev. Lett.*, 54(25):2708–2711, 1985.

[75] D. Chowdhury, L. Santen, and A. Schadschneider. Statistical physics of vehicular traffic and some related systems. *Physics Reports*, 329(4):199–329, 2000.

[76] T. Gotoh. Inertial range statistics of burgers turbulence. *Physics of Fluids*, 6:3985, 1994.

[77] A. Chekhlov and V. Yakhot. Kolmogorov turbulence in a random-force-driven burgers equation. *Physical Review E*, 51(4):2739–2742, 1995.

[78] V. Yakhot and S.A. Orszag. Renormalization group analysis of turbulence. i. basic theory. *Journal of scientific computing*, 1(1):3–51, 1986.

[79] J.D. Cole. On a linear quasilinear parabolic equation in aerodynamics. *Q. Appl. Math*, 9:225–36, 1951.

[80] E. Hopf. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Communications on Pure and Applied mathematics*, 3(3):201–230, 1950.

[81] D.A. Huse, C.L. Henley, and D.S. Fisher. Huse, henley, and fisher respond. *Phys. Rev. Lett.*, 55(26):2924–2924, 1985.

[82] Mehran Kardar. Roughening by impurities at finite temperatures. *Phys. Rev. Lett.*, 55:2923–2923, Dec 1985.

[83] Daniel S. Fisher and David A. Huse. Directed paths in a random potential. *Phys. Rev. B*, 43:10728–10742, May 1991.

[84] R. Bellman. Bottleneck problems and dynamic programming. *Proc. Natl. Acad. Sci. U. S. A.*, 39(9):947, 1953.

[85] B. Derrida and J. Vannimenus. Interface energy in random systems. *Physical Review B*, 27(7):4401, 1983.

[86] JM Kim, MA Moore, and AJ Bray. Zero-temperature directed polymers in a random potential. *Physical Review A*, 44(4):2345, 1991.

[87] L. Balents and M. Kardar. Directed paths on percolation clusters. *Journal of statistical physics*, 67(1):1–11, 1992.

[88] E. Perlsman and S. Havlin. The directed-polymerdirected-percolation transition. *EPL (Europhysics Letters)*, 46(1):13–17, 1999.

[89] Ehud Perlsman and Moshe Schwartz. Ultrametric tree structure in the directed polymer problem. *EPL (Europhysics Letters)*, 17(1):11, 1992.

[90] Yi-Cheng Zhang. Ground state instability of a random system. *Phys. Rev. Lett.*, 59:2125–2128, Nov 1987.

[91] M. Kardar and Y.C. Zhang. Scaling of directed polymers in random media. *Phys. Rev. Lett.*, 58(20):2087–2090, 1987.

[92] T. Halpin-Healy. Directed polymers versus directed percolation. *Physical Review E*, 58(4):4096–4099, 1998.

[93] R. Hidema, Z. Yatabe, M. Shoji, C. Hashimoto, R. Pansu, G. Sagarzazu, and H. Ushiki. Image analysis of thickness in flowing soap films. I: effects of polymer. *Experiments in Fluids*, pages 1–8, 2010.

[94] A. Fortin, M. Jardak, JJ Gervais, and R. Pierre. Old and new results on the two-dimensional Poiseuille flow. *J. Comput. Phys.*, 115(2):455–469, 1994.

[95] H. Chaté and P. Manneville. Continuous and discontinuous transition to spatio-temporal intermittency in two-dimensional coupled map lattices. *EPL (Europhysics Letters)*, 6:591, 1988.

[96] T. Tél and Y.C. Lai. Chaotic transients in spatially extended systems. *Physics Reports*, 460(6):245–275, 2008.

[97] KR Elder, JD Gunton, and N. Goldenfeld. Transition to spatiotemporal chaos in the damped Kuramoto-Sivashinsky equation. *Physical Review E*, 56(2):1631–1634, 1997.

[98] K. Orihashi and Y. Aizawa. Turbulence in diffusion replicator equation. *Physica D: Nonlinear Phenomena*, 237(23):3053–3060, 2008.

[99] W.B. Whitman, D.C. Coleman, and W.J. Wiebe. Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. USA*, 95(12):6578–6583, 1998.

[100] P.G. Falkowski. Tracing oxygen's imprint on earth's metabolic evolution. *Science*, 311(5768):1724–1725, 2006.

[101] R.D. Bardgett, C. Freeman, and N.J. Ostle. Microbial contributions to climate change through carbon cycle feedbacks. *The ISME Journal*, 2(8):805–814, 2008.

[102] P.J. Turnbaugh, R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, and J.I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[103] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J.A. Schloss, V. Bonazzi, J.E. McEwen, K.A. Wetterstrand, C. Deal, et al. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

[104] R.I. Amann, W. Ludwig, and K.H. Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59(1):143–169, 1995.

[105] Christian S. Riesenfeld, Patrick D. Schloss, and Jo Handelsman. Metagenomics: Genomic Analysis of Microbial Communities. *Annu. Rev. Genet.*, 38:525–552, 2004.

[106] J.F. Petrosino, S. Highlander, R.A. Luna, R.A. Gibbs, and J. Versalovic. Metagenomic pyrosequencing and microbial identification. *Clinical chemistry*, 55(5):856–866, 2009.

[107] C.R. Woese and G.E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA*, 74(11):5088–5090, 1977.

[108] J. Shendure and H. Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[109] J.R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315, 2010.

[110] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443 – 453, 1970.

[111] M.E. Sardiu, G. Alves, and Y.K. Yu. Score statistics of global sequence alignment from the energy distribution of a modified directed polymer and directed percolation problem. *Physical Review E*, 72(6):061917, 2005.

[112] J. Raes and P. Bork. Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews Microbiology*, 6(9):693–699, 2008.

[113] J.R. Bray and J.T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological monographs*, 27(4):325–349, 1957.

[114] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. Von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome research*, 20(7):947–959, 2010.

[115] R. A. Fisher, A. Steven Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, 12(1):42–58, 1943.

[116] M. Tokeshi. Niche apportionment or random assortment: species abundance patterns revisited. *The Journal of Animal Ecology*, pages 1129–1146, 1990.

[117] Johan Chu and Cristoph Adami. A simple explanation for taxon abundance patterns. *Proc. Natl. Acad. Sci. USA*, 96(26):15017–15019, 1999.

[118] Stephen P. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography.* Monographs in Population Biology. Princeton University Press, 2001.

[119] S.P. Hubbell. Neutral theory in community ecology and the hypothesis of functional equivalence. *Funct. Ecol.*, 19(1):166–172, 2005.

[120] R.S. Etienne, D. Alonso, and A.J. McKane. The zero-sum assumption in neutral biodiversity theory. *J. Theor. Biol.*, 248(3):522–536, 2007.

[121] M. Vallade and B. Houchmandzadeh. Analytical solution of a neutral model of biodiversity. *Physical Review E*, 68(6):61902, 2003.

[122] D. Alonso and A.J. McKane. Sampling Hubbell's neutral theory of biodiversity. *Ecol. Lett.*, 7(10):901–910, 2004.

[123] R.S. Etienne. A new sampling formula for neutral biodiversity. *Ecol. Lett.*, 8(3):253–260, 2005.

[124] Omri Allouche and Ronen Kadmon. A general framework for neutral models of community dynamics. *Ecol. Lett.*, 12(12):1287–97, 2009.

[125] G J Olsen, D J Lane, S J Giovannoni, N R Pace, and D A Stahl. Microbial ecology and evolution: A ribosomal RNA approach. *Annual Review of Microbiology*, 40(1):337, 1986.

[126] Stephen J. Giovannoni, Theresa B. Britschgi, Craig L. Moyer, and Katharine G. Field. Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345:60–63, 1990.

[127] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nat. Methods*, 5(1):16–8, 2008.

[128] Luiz F W Roesch, Roberta R Fulthorpe, Alberto Riva, George Casella, Alison K M Hadwin, Angela D Kent, Samira H Daroub, Flavio A O Camargo, William G Farmerie, and Eric W Triplett. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*, 1(4):283–90, 2007.

[129] Florent E Angly, Ben Felts, Mya Breitbart, Peter Salamon, Robert A Edwards, Craig Carlson, Amy M Chan, Matthew Haynes, Scott Kelley, Hong Liu, Joseph M Mahaffy, Jennifer E Mueller, Jim Nulton, Robert Olson, Rachel Parsons, Steve Rayhawk, Curtis A Suttle, and Forest Rohwer. The marine viromes of four oceanic regions. *PLOS Biology*, 4(11):e368, 2006.

[130] Robert A Edwards, Beltran Rodriguez-Brito, Linda Wegley, Matthew Haynes, Mya Breitbart, Dean M Peterson, Martin O Saar, Scott Alexander, E Calvin Alexander, and Forest Rohwer. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, 7:57, 2006.

[131] Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA*, 103(32):12115–20, 2006.

[132] Alain Leprêtre and D. Mouillot. A comparison of species diversity estimators. *Population Ecology*, 41(2):203, 1999.

[133] PD Schloss, SL Westcott, T Ryabin, JR Hall, M Hartmann, EB Hollister, RA Lesniewski, BB Oakley, DH Parks, CJ Robinson, JW Sahl, B Stres, GG Thallinger, DJ Van Horn, and CF Weber. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75(23):7537–7541, 2009.

[134] Z. Liu, T. Z. DeSantis, G. L. Andersen, and R. Knight. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.*, 36(18):e120, 2008.

[135] N. Youssef, C. S. Sheik, L. R. Krumholz, F. Z. Najar, B. A. Roe, and M. S. Elshahed. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. *Applied and Environmental Microbiology*, 75(16):5227, 2009.

[136] Vicente Gomez-Alvarez, Tracy K Teal, and Thomas M Schmidt. Systematic artifacts in metagenomes from complex microbial communities. *ISME J*, 3(11):1314–7, 2009.

[137] Hui-Hsien Chou and Michael H. Holmes. Dna sequence quality trimming and vector removal. *Bioinformatics*, 17(12):1093–1104, 2001.

[138] Christopher Quince, Anders Lanzén, Thomas P Curtis, Russell J Davenport, Neil Hall, Ian M Head, L Fiona Read, and William T Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods*, 6(9):639–41, 2009.

[139] V Kunin, A Engelbrektson, H Ochman, and P Hugenholtz. Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, 12(1):118–123, 2009.

[140] Susan M Huse, David Mark Welch, Hilary G Morrison, and Mitchell L Sogin. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, 2010.

[141] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Simulation and Computation*, 3(1):1, 1974.

[142] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):e9490, 2010.

[143] Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–90, 2006.

[144] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228–35, 2005.

[145] Alex C W May. Percent sequence identity; the need to be explicit. *Structure*, 12(5):737–8, 2004.

[146] Patrick D Schloss and Jo Handelsman. Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, 71(3):1501–6, 2005.

[147] J. Felsenstein. Phylip - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[148] David J. Lane, Bernadette Pace, Gary J. Olsen, David A. Stahl, Mitchell L. Sogin, and Norman R. Pace. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. USA*, 82:6955–6959, 1985.

[149] Carl R. Woese. Bacterial evolution. *Microbiol. Rev.*, 51(2):221–271, 1987.

[150] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480, 2009.

[151] T. Z. DeSantis, P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.*, 34(W):394–399, 2006.

[152] J R Cole, Q Wang, E Cardenas, J Fish, B Chai, R J Farris, A S Kulam-Syed-Mohideen, D M McGarrell, T Marsh, G M Garrity, and J M Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, 37(Database issue):D141–5, 2009.

[153] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72:5069–72, 2006.

[154] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, 35(21):7188, 2007.

[155] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–7, 2009.

[156] David A. Morrison and John T. Ellis. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. *Mol. Biol. Evol.*, 14(4):428–441, 1997.

[157] Kevin Liu, Sindhu Raghavan, Serita Nelesen, C Randal Linder, and Tandy Warnow. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–4, 2009.

[158] D. Krznaric and C. Levcopoulos. Fast Algorithms for Complete Linkage Clustering. *Discrete and Computational Geometry*, 19(1):131–145, 1998.

[159] Yanan Yu, Mya Breitbart, Pat McNairnie, and Forest Rohwer. FastGroupII: a web-based bioinformatics platform for analyses of large 16S rDNA libraries. *BMC Bioinformatics*, 7:57, 2006.

[160] Yijun Sun, Yunpeng Cai, Li Liu, Fahong Yu, Michael L Farrell, William McKendree, and William Farmerie. Esprit: estimating species richness using large collections of 16s rrna pyrosequences. *Nucleic Acids Res.*, 37(10):e76, 2009.

[161] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, 1990.

[162] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, 2002.

[163] Glenn W. Milligan and Martha C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[164] Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, 33(8):2433–2439, 2005.

[165] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[166] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797, 2004.

[167] Robert C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(113), 2004.

[168] Anna Engelbrektson, Victor Kunin, Kelly C Wrighton, Natasha Zvenigorodsky, Feng Chen, Howard Ochman, and Philip Hugenholtz. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME journal*, 4:642–647, 2010.

[169] Patrick D. Schloss. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16s rrna gene-based studies. *PLoS computational biology*, 6(7):e1000844+, July 2010.

[170] Ani Qu, Jennifer M Brulc, Melissa K Wilson, Bibiana F Law, James R Theoret, Lynn A Joens, Michael E Konkel, Florent Angly, Elizabeth A Dinsdale, Robert A Edwards, Karen E Nelson, and Bryan A White. Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLOS ONE*, 3(8):e2945, 2008.

[171] Patrick D. Schloss. A high-throughput dna sequence aligner for microbial ecology studies. *PLOS ONE*, 4(12):e8230, 12 2009.

[172] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, 1994.

[173] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31(13):3429–3431, 2003.

[174] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

[175] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–148, 1981.

[176] Weizhong Li, Limin Fu, Beifang Niu, Sitao Wu, and John Wooley. Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, 2012.

[177] Ameet J. Pinto and Lutgarde Raskin. Pcr biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLOS ONE*, 7(8):e43093, 08 2012.

[178] C.K. Lee, C.W. Herbold, S.W. Polson, K.E. Wommack, S.J. Williamson, I.R. McDonald, and S.C. Cary. Groundtruthing next-gen sequencing for microbial ecology–biases and errors in community structure estimates from pcr amplicon pyrosequencing. *PLOS ONE*, 7(9):e44224, 2012.

[179] F. Aurenhammer. Voronoi diagramsa survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[180] R.C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461, 2010.

[181] C. Sadowski and G. Levin. Simhash: Hash-based similarity detection, 2007.

[182] C.R. Woese. Default taxonomy: Ernst Mayrs view of the microbial world. *Proc. Natl. Acad. Sci. USA*, 95(19):11043, 1998.

[183] Y. Oono. Renormalization and taxonomy. *Journal of statistical physics*, 110(3):1369–1374, 2003.

[184] Patricio Jeraldo Maldonado. *Computational approaches to stochastic systems in physics and biology.* Dissertation, University of Illinois, 2012.

[185] P. Jeraldo, N. Chia, and N. Goldenfeld. On the suitability of short reads of 16s rrna for phylogeny-based analyses in environmental surveys. *Environ. Microbiol.*, 13(11):3000–3009, 2011.

[186] M. Tokeshi. *Species coexistence: ecological and evolutionary perspectives.* Wiley-Blackwell, New York, 1999.

[187] P Chesson. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.*, 31:343–366, 2000.

[188] G E Hutchinson. Homage to Santa Rosalia, or why are there so many kinds of animals? *Am. Nat.*, 93:145–159, 1959.

[189] G E Hutchinson. THe paradox of the plankton. *Am. Nat.*, 95:137–145, 1961.

[190] S A Levin J Chaveé, H C Muller-Landau. Comparing classical community models: theoretical consequences for patterns of diversity. *Am. Nat.*, 159:1–23, 2002.

[191] J Silvertown. Plant coexistence and the niche. *Trends Ecol. Evol.*, 19:605–611, 2004.

[192] S.J. Wright. Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia*, 130:1–14, 2002.

[193] H Caswell. Community structure: a neutral model analysis. *Ecol. Monogr.*, 46:327–354, 1976.

[194] G. Bell. The distribution of abundance in neutral communities. *Am. Nat.*, 155:606–617, 2000.

[195] G. Bell. Neutral macroecology. *Science*, 293(5539):2413–2418, 2001.

[196] J Chave. Neutral theory and community ecology. *Ecol. Lett.*, 7:241–253, 2004.

[197] J. Rosindell, S.P. Hubbell, and R.S. Etienne. The Unified Neutral Theory of Biodiversity and Biogeography at Age Ten. *Trends Ecol. Evol.*, 26:340–348, 2011.

[198] R. Muneepeerakul, E. Bertuzzo, H.J. Lynch, W.F. Fagan, A. Rinaldo, and I. Rodriguez-Iturbe. Neutral metacommunity models predict fish diversity patterns in Mississippi–Missouri basin. *Nature*, 453(7192):220–222, 2008.

[199] S. Woodcock, C.J. van der Gast, T. Bell, M. Lunn, T.P. Curtis, I.M. Head, and W.T. Sloan. Neutral assembly of bacterial communities. *FEMS Microbiol. Ecol.*, 62(2):171–180, 2007.

[200] B.J. McGill. A test of the unified neutral theory of biodiversity. *Nature*, 422(6934):881–885, 2003.

[201] B.J. McGill, B.A. Maurer, and M.D. Weiser. Empirical evaluation of neutral theory. *Ecology*, 87(6):1411–1423, 2006.

[202] M.A. Leibold and M.A. McPeek. Coexistence of the niche and neutral perspectives in community ecology. *Ecology*, 87(6):1399–1410, 2006.

[203] D.W. Purves and L.A. Turnbull. Different but equal: the implausible assumption at the heart of neutral theory. *J. Anim. Ecol.*, 79(6):1215–1225, 2010.

[204] R.E. Ricklefs. The unified neutral theory of biodiversity: Do the numbers add up? *Ecology*, 87(6):1424–1431, 2006.

[205] Peter B. Adler, Janneke Hille Rislambers, and Jonathan M. Levine. A niche for neutrality. *Ecol. Lett.*, 10:95–104, 2007.

[206] P.B. Adler, S.P. Ellner, and J.M. Levine. Coexistence of perennial plants: an embarrassment of niches. *Ecol. Lett.*, 13(8):1019–1029, 2010.

[207] I. Volkov, J.R. Banavar, S.P. Hubbell, and A. Maritan. Inferring species interactions in tropical forests. *Proc. Natl. Acad. Sci. USA*, 106(33):13854–13859, 2009.

[208] X. Zhang, Wei Liu, Y. Bai, G. Zhang, and X. Han. Nitrogen deposition mediates the effects and importance of chance in changing biodiversity. *Molecular ecology*, 20(2):429–438, 2011.

[209] D.W. Purves, S.W. Pacala, D. Burslem, M.A. Pinard, and S.E. Hartley. Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. In D. F. Burslem, M. A. Pinard, and S. E. Hartley, editors, *Biotic interactions in the tropics: their role in the maintenance of species diversity*, pages 107–138. Cambridge University Press, Cambridge, 2005.

[210] R.A. Chisholm and S.W. Pacala. Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proc. Natl. Acad. Sci. USA*, 107(36):15821–15825, 2010.

[211] D. Gravel, C.D. Canham, M. Beaudet, and C. Messier. Reconciling niche and neutrality: the continuum hypothesis. *Ecol. Lett.*, 9(4):399–409, 2006.

[212] D. Tilman. Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly. *Proc. Natl. Acad. Sci. USA*, 101(30):10854–10861, 2004.

[213] M.W. Cadotte. Concurrent niche and neutral processes in the competition–colonization model of species coexistence. *Proc. R. Soc. B*, 274:2739–2744, 2007.

[214] T. Zillio and R. Condit. The impact of neutrality, niche differentiation and species input on diversity and abundance distributions. *Oikos*, 116(6):931–940, 2007.

[215] M. Loreau and C. de Mazancourt. Species Synchrony and Its Drivers: Neutral and Nonneutral Community Dynamics in Fluctuating Environments. *Am. Nat.*, 172(2):48–66, 2008.

[216] C.P. Doncaster and S.J. Cornell. Ecological Equivalence: A Realistic Assumption for Niche Theory as a Testable Alternative to Neutral Theory. *PLOS ONE*, 4:e7460, 2009.

[217] B. Haegeman and M. Loreau. A mathematical synthesis of niche and neutral theories in community ecology. *J. Theor. Biol.*, 269:150–165, 2011.

[218] F. Ellwood, A. Manica, and W.A. Foster. Stochastic and deterministic processes jointly structure tropical arthropod communities. *Ecol. Lett.*, 12(4):277–284, 2009.

[219] A.J. Dumbrell, M. Nelson, T. Helgason, C. Dytham, and A.H. Fitter. Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J.*, 4(3):337–345, 2009.

[220] Q.G. Zhang, A. Buckling, and H.C.J. Godfray. Quantifying the relative importance of niches and neutrality for coexistence in a model microbial system. *Funct. Ecol.*, 23(6):1139–1147, 2009.

[221] Silke Langenheder and Anna J Székely. Species sorting and neutral processes are both important during the initial assembly of bacterial communities. *ISME J.*, 5(7):1086–1094, 2011.

[222] Joaquín M Ayarza and Leonardo Erijman. Balance of neutral and deterministic components in the dynamics of activated sludge floc assembly. *Microb. Ecol.*, 61(3):486–495, 2011.

[223] Irina Dana Ofiteru, Mary Lunn, Thomas P Curtis, George F Wells, Craig S Criddle, Christopher a Francis, and William T Sloan. Combined niche and neutral effects in a microbial wastewater treatment community. *Proc. Natl. Acad. Sci. USA*, 107(35):15345–15350, 2010.

[224] Catherine Burke, Peter Steinberg, Doug Rusch, Staffan Kjelleberg, and Torsten Thomas. Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. USA*, 108(34):14288–14293, 2011.

[225] M Claire Horner-Devine, Jessica M Silver, Mathew a Leibold, Brendan J M Bohannan, Robert K Colwell, Jed a Fuhrman, Jessica L Green, Cheryl R Kuske, Jennifer B H Martiny, Gerard Muyzer, Lise Ovreå s, Anna-Louise Reysenbach, and Val H Smith. A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology*, 88(6):1345–1353, 2007.

[226] B.C. Emerson and R.G. Gillespie. Phylogenetic analysis of community assembly and structure over space and time. *Trends Ecol. Evol.*, 23(11):619–630, 2008.

[227] C.K. Kelly, M.G. Bowler, O. Pybus, and P.H. Harvey. Phylogeny, niches, and relative abundance in natural communities. *Ecology*, 89(4):962–970, 2008.

[228] J. Cavender-Bares, K.H. Kozak, P.V.A. Fine, and S.W. Kembel. The merging of community ecology and phylogenetic biology. *Ecol. Lett.*, 12(7):693–715, 2009.

[229] S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–1464, 2010.

[230] M.W. Cadotte, T. Jonathan Davies, J. Regetz, S.W. Kembel, E. Cleland, and T.H. Oakley. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol. Lett.*, 13(1):96–105, 2010.

[231] Susan M. Huse, Les Dethlefsen, Julie A. Huber, David Mark Welch, David A. Relman, and Mitchell L. Sogin. Exploring microbial diversity and taxonomy using ssu rrna hypervariable tag sequencing. *PLOS Genet.*, 4(11):e1000255, 2008.

[232] Jonathan H. Badger, Pauline C. Ng, and J. Craig Venter. The Human Genome, Microbiomes, and Disease. In Karen E. Nelson, editor, *Metagenomics of the Human Body*, pages 1–14. Springer, New York, 2011.

[233] J. M. Brulc, D. A. Antonopoulos, M. E. Berg Miller, M. K. Wilson, A. C. Yannarell, E. A. Dinsdale, R. E. Edwards, E. D. Frank, J.B. Emerson, P. Wacklin, Pedro M Coutinho, Bernard Henrissat, Karen E Nelson, and Bryan A White. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc. Natl. Acad. Sci. USA*, 106(6):1948–1953, 2009.

[234] F. Bäckhed, R.E. Ley, J.L. Sonnenburg, D.A. Peterson, and J.I. Gordon. Host-bacterial mutualism in the human intestine. *Science*, 307(5717):1915–1920, 2005.

[235] Les Dethlefsen, Margaret McFall-Ngai, and D.A. Relman. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature*, 449(7164):811–818, 2007.

[236] P.J. Turnbaugh and J.I. Gordon. The core gut microbiome, energy balance and obesity. *J. Physiol.*, 587(17):4153–4158, 2009.

[237] Min Li, Baohong Wang, Menghui Zhang, Mattias Rantalainen, Shengyue Wang, Haokui Zhou, Yan Zhang, Jian Shen, Xiaoyan Pang, Meiling Zhang, Hua Wei, Yu Chen, Haifeng Lu, Jian Zuo, Mingming Su, Yunping Qiu, Wei Jia, Chaoni Xiao, Leon M. Smith, Shengli Yang, Elaine Holmes, Huiru Tang, Guoping Zhao, Jeremy K. Nicholson, Lanjuan Li, and Liping Zhao. Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl. Acad. Sci. USA*, 105(6):2117–2122, 2008.

[238] Emma Slack, Siegfried Hapfelmeier, Brbel Stecher, Yuliya Velykoredko, Maaike Stoel, Melissa A. E. Lawson, Markus B. Geuking, Bruce Beutler, Thomas F. Tedder, Wolf-Dietrich Hardt, Premysl Bercik, Elena F. Verdu, Kathy D. McCoy, and Andrew J. Macpherson. Innate and Adaptive Immunity Cooperate Flexibly to Maintain Host-Microbiota Mutualism. *Science*, 325(5940):617–620, 2009.

[239] D.A. Antonopoulos, S.M. Huse, H.G. Morrison, T.M. Schmidt, M.L. Sogin, and V.B. Young. Reproducible community dynamics of the gastrointestinal microbiota following antibiotic perturbation. *Infect. Immun.*, 77(6):2367–2375, 2009.

[240] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms. In Bertram Ludäscher and Nikos Mamoulis, editors, *Scientific and Statistical Database Management*, volume 5069 of *Lecture Notes in Computer Science*, pages 418–435. Springer, Heidelberg, 2008.

[241] Sean Humphray, Carol Scott, Richard Clark, Brandy Marron, Clare Bender, Nick Camm, Jayne Davis, Andrew Jenks, Angela Noon, Manish Patel, Harminder Sehra, Fengtang Yang, Margarita Rogatcheva, Denis Milan, Patrick Chardon, Gary Rohrer, Dan Nonneman, Pieter de Jong, Stacey Meyers, Alan Archibald, Jonathan Beever, Lawrence Schook, and Jane Rogers. A high utility integrated map of the pig genome. *Genome Biol.*, 8(7):R139, 2007.

[242] G. Muyzer, EC Dewaal, and AG Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.*, 59:695–700, 1993.

[243] Cristoph Adami and Johan Chu. Critical and near-critical branching processes. *Physical Review E*, 66(011907):8, 2002.

[244] J.S. Gray, A. Bjorgesaeter, and K.I. Ugland. On plotting species abundance distributions. *J. Anim. Ecol.*, 75(3):752–756, 2006.

[245] Q. Wang, G.M. Garrity, J.M. Tiedje, and J.R. Cole. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73(16):5261–5267, 2007.

[246] H.T. Odum. *Ecological and general systems: an introduction to systems ecology.* University Press of Colorado Niwot, CO, 1994.

[247] M.K. Thomas, C.T. Kremer, C.A. Klausmeier, and E. Litchman. A global pattern of thermal adaptation in marine phytoplankton. *Science*, 338(6110):1085–1088, 2012.

[248] B.M. Uz, J.A. Yoder, and V. Osychny. Pumping of nutrients to ocean surface waters by the action of propagating planetary waves. *Nature*, 409(6820):597–600, 2001.

[249] S. Yooseph, K.H. Nealson, D.B. Rusch, J.P. McCrow, C.L. Dupont, M. Kim, J. Johnson, R. Montgomery, S. Ferriera, K. Beeson, et al. Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468(7320):60–66, 2010.

[250] K.N. Laland, J. Odling-Smee, and M.W. Feldman. Causing a Commotion. *Nature*, 429(6992):609–610, 2004.

[251] K. Vetsigian, R. Jajoo, and R. Kishony. Structure and evolution of streptomyces interaction networks in soil and in silico. *PLOS biology*, 9(10):e1001184, 2011.

[252] D.B. Rusch, A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman, K. Remington, et al. The sorcerer ii global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PLOS biology*, 5(3):e77, 2007.

[253] S. Yooseph, G. Sutton, D.B. Rusch, A.L. Halpern, S.J. Williamson, K. Remington, J.A. Eisen, K.B. Heidelberg, G. Manning, W. Li, et al. The sorcerer ii global ocean sampling expedition: expanding the universe of protein families. *PLOS biology*, 5(3):e16, 2007.

[254] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[255] R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, and E.V. Koonin. The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, 29(1):22–28, 2001.

[256] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(D1):D109–D114, 2012.

[257] J.R. Clark, S.J. Daines, T.M. Lenton, A.J. Watson, and H.T.P. Williams. Individual-based modelling of adaptation in marine microbial populations using genetically defined physiological parameters. *Ecological Modelling*, 222(23):3823–3837, 2011.

[258] I. Stewart, T. Elmhirst, and J. Cohen. Symmetry-breaking as an origin of species. *Bifurcation, Symmetry and Patterns*, pages 3–54, 2003.

[259] T. Rogers, A. J. McKane, and A. G. Rossberg. Spontaneous genetic clustering in populations of competing organisms. *Physical Biology*, 9(6):066002, December 2012.

[260] J.R. Sieber, M.J. McInerney, and R.P. Gunsalus. Genomic insights into syntrophy: The paradigm for anaerobic metabolic cooperation. *Annual Review of Microbiology*, 66(1), 2012.

[261] C.A. Mack. Fifty years of moore's law. *Semiconductor Manufacturing, IEEE Transactions on*, 24(2):202–207, 2011.

[262] P.G. Higgs and T.K. Attwood. *Bioinformatics and molecular evolution.* Wiley-Blackwell, 2009.

[263] S.H. Liao. Expert system methodologies and applicationsa decade review from 1995 to 2004. *Expert systems with applications*, 28(1):93–103, 2005.

[264] B.C. Björk, A. Roos, and M. Lauri. Scientific journal publishing–yearly volume and open access availability. *Information Research*, 14(1):391, 2009.

[265] J. Hendler. Science and the semantic web. *Science*, 299(5606):520, 2003.

[266] I. Asimov. Foundation. 1951. *New York: Avon*, 1966.