COLLECTIVE EVOLUTION OF BIOLOGICAL AND PHYSICAL SYSTEMS

BY

KALIN HOREN VETSIGIAN

B.S., Massachusetts Institute of Technology, 2000
B.S., Massachusetts Institute of Technology, 2000

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, 2005

Urbana, Illinois

COLLECTIVE EVOLUTION OF BIOLOGICAL AND PHYSICAL SYSTEMS

Kalin Vetsigian, Ph.D.
Department of Physics
University of Illinois at Urbana-Champaign, 2005
Prof. Nigel Goldenfeld, Advisor

In this dissertation, I study the evolution of solidification fronts propagating in undercooled liquids, the evolution of microbial communities through diversification fronts propagating along microbial genomes, the evolution of the universality and optimality of the genetic code, and the emergence of genome biases.

I present a new phase-field model of solidification which allows efficient computations in the regime when interface kinetic effects dominate over capillary effects. The asymptotic analysis required to relate the parameters in the phase-field with those of the original sharp interface model is straightforward, and the resultant phase-field model can be used for a wide range of material parameters.

I model the competition between homologous recombination and point mutation in microbial genomes, and present evidence for two distinct phases, one uniform, the other genetically diverse. Depending on the specifics of homologous recombination, I find that global sequence divergence can be mediated by fronts propagating along the genome, whose characteristic signature on genome structure is elucidated, and apparently observed in closely related genomes from the *Bacillus cereus* group. Front propagation provides an emergent, generic mechanism for microbial "speciation," and suggests a classification of microorganisms on the basis of their propensity to support propagating fronts.

I propose that selection on the speed, accuracy and energy efficiency of template-directed synthesis processes such as translation, transcription and replication can lead to the *spontaneous* emergence of genome biases. Selection on translation leads to codon usage bias; selection on transcription or replication leads to nucleotide composition biases such as the GC content. These biases result from the generic tradeoffs inherent to template-directed synthesis and occur even in the absence of biased mutation or direct selection on the nucleotide composition coming from, say, DNA or mRNA stability. In the case of translation, it is the coevolution between codon usage and tRNA expression levels that creates a fitness landscape that enforces quasi-stable patterns of codon usage.

Occasional transitions between patterns are expected, due to genetic drift or hitchhiking of slightly deleterious adjustments of the translational system on other beneficial traits.

Then, I show that the above coevolutionary dynamics provides an efficient mechanism for optimization of genetic codes, even if, as the frozen accident theory assumes, every amino acid substitution is lethal at least at some genome sites. This research shows that it is possible to account for the optimality of the code within the framework of translation as a standardized competition between tRNA adaptors.

Finally, I investigate the proposition that genetic exchange dominating the early evolution of life naturally leads to a common genetic code for all organisms, while promoting their incredible diversity in all other aspects. I present three possible mechanisms through which HGT brings universality - communal advantage of popular codes, HGT of translational components and HGT of protein coding regions. A possible consequence of the interplay of these mechanisms is the concerted evolution towards optimality of a community of organisms sharing the same genetic code and having compatible translational machineries.

To my parents

Acknowledgments

The first class on the very first day of my PhD was taught by Nigel Goldenfeld. Right there, before I had seen anything else I knew I would like to become his graduate student. Perhaps, it helped that, coming from MIT, I had an absolute scale for what exceptional is. Six months later, just before I joined his group, he drew me a graph on the black board. A curve steeply went down at first, then stayed at the bottom for a long time, and then turned into a dashed one and went up again. Then he said, "This is the time evolution of your self confidence/satisfaction level during your PhD. Are you ready for this? There is a chance that, at the end, your happiness will bounce back to its present value." What a funny guy, I thought. Obviously, he doesn't know *me*! He turned out to be right about this curve and a zillion of other things, and I guess this is what makes him a mature advisor. He caringly let me break my head and explore on my own most of the time (because this is the road that turns a student into a researcher), but provided the crucial scientific, organizational and psychological support that guaranteed that it won't be all a disaster. This, combined with a continuous exposure to his daring, adventurous, knowing no boundaries curiosity and a profoundly deep and unusual scientific outlook made this PhD what it is. It is not my job to evaluate its scientific significance, if any, but here is how I feel at the end of it: happy, enthusiastic, daring and adventurous; looking forward to future explorations. I have no doubt that I made the best for me choice of an advisor, and I feel incredibly lucky to have had the option. Nigel, many thanks for four incredible and unforgettable years!

The other two professors that influenced me tremendously are Yoshi Oono and Carl Woese. Whenever I dropped to Yoshi's office he spared at least three-four hours to talk with me, which is rather exceptional for a busy professor. These talks were a treat for me, and constituted a singular perturbation of my understanding of biology. I can't pinpoint it exactly, especially at the end of this draining thesis marathon, but it is the combination of his knowing and keeping up with an

incredible amount of biology, and at the same time thinking deeply and provocatively in a most unique way. This combination is possible, he proves, and is the *honest* and meaningful way to go. I have never heard him speak superficially like a macho physicist that would save biology. I also enjoy his style. He would interrupt me and fire at me, "Do you *really* believe that elephants can evolve?" And I would go home with an important message. He was particularly helpful in the months before my prelim when I was frantically looking for direction. It was reading *Dear Mr. Darwin* on the beaches of the Black Sea, a book that he lent me, that I started thinking about the interplay of homologous recombination and point mutations. If I was not feeling at the bottom of the curve, mentioned above, most of the time, I would have interacted with him much more.

As is obvious from the content of this thesis, Carl Woese's ideas have been most influential for my research. I was also incredibly fortunate to attend one of his unorthodox mind-rewiring courses, and eventually work with him! It has been an honor. As great a scientist as he is, he is a most kind and encouraging person. His encouragement, I think, was crucial for this thesis, because it gave me confidence and strength to explore. He also supported me as a research assistant during the last two summers. Without the safe nest that he and Nigel provided, I would have never been so daring, at this stage of my carrier, to explore subjects such as the evolution of the genetic code. The genetic code is a subject that cannot be addressed with conventional scientific tools, and I was simply forced to go berserk. This process reshaped my thinking and my understanding of evolution. Apart from the code, my discussions with Carl led to some extraordinary ideas to be pursued in the future.

During my first year of research I had the pleasure to interact with Jon Dantzig. I would like to thank him for his invaluable support while working on the phase field model of solidification. I am indebted to Paul Goldbart for letting me stay in his comfortable home for a large part of the summer of 2005. His tranquil back porch inspired chapter 4, and after my laptop died, a significant part of this thesis was written on his home computer.

Now I would like to shift gears and briefly acknowledge love, friendship and meaning. To do justice to their significance for the last five years of my life, I should have written a 150 page essay accompanied by an abstract of my research, instead of a 150 pages of research, accompanied by a meager acknowledgment section. To make things worse, most of what I would like to convey to the

people below falls outside the public domain.

Boriana, without your love, I wouldn't have known peace and harmony, and the passion of the mind, useful for research and hence acknowledged, wouldn't have been awakened. In addition, you influenced my scientific outlook and sensitivities in a way complementary to that of the people mentioned above. You are a living proof that deeply analytical does not equal mathematical. Our endless discussions on the philosophy of science, the conceptual differences between the questions asked in Political Science and Physics, limitations of modeling, etc., are all partly reflected in the introductory chapter and throughout the thesis. All in all, the more interesting and relevant problems I tried to attack, the more useful I found the approaches you taught me. Having a private brilliant political scientist makes a difference.

Several people made Urbana-Champaign home. Above all I would like to thank Smitha Vishveshwara, Swagatam Mukhopadhyay and Abhishek Singh for showing me boundless friendship, love, beauty, good times, and helping me to discover and unleash layers upon layers of madness, creativity and, at the same time, humility. I wouldn't have been the same person without you, and the change was for good, mind you. Sunayana Saha and Namitha Vishveshwara – more of the above, though I spent a notch less time with you. This gang has been too good to be true.

I would also like to thank all present and former members of Nigel's group that I was fortunate to meet: Hector Garcia Martin, John Veysey, Patrick Chan, Nick Gutenberg, Tae Kim, Vivek Aji, Michelle Nahas. Guys, you were always kind, considerate, generous, funny, brilliant and silly when appropriate! Please forgive me for not having the energy, in my sleep deprived state, to give everyone of you his/her due individually.

It never hurts to have an extra parent. Dimitar Mekerov was, and is, an extraordinary father figure, *Life is Beautiful* type of person, and a guiding light in my life. Many thanks to Ivan Sirakov - the first brilliant scientist that I've got to know.

Thanks to my brother Vladimir Vetsigian, well, for being the older brother. Here is an example of what it implied: When I was seventh grade, I took a great pride in going to physics competitions. One day he challenged me with the problem of computing the time it takes to empty a pool through a pipe at the bottom. He laughed his guts out, as I was insanely writing equations never figuring out what is going on. Embarrassed, I learned some calculus in the following weeks. He was doing

such horrible things to me all the time.

To my parents: This thesis is dedicated to you! In difficult times you gave me everything, trusted me, and let me figure out who I want to be.

I gratefully acknowledge support by the Department of Physics at the University of Illinois at Urbana-Champaign, the National Science Foundation under grant numbers NSF-EAR-0221743, NSFDMR-99-70690 and NSF-DMR-01-21695, and the National Aeronautics and Space Administration under grant number NASA-NAG-8-1760.

Table of Contents

List of Figures

List of Abbreviations

HGT Horizontal Gene Transfer

DLMEM Distribution of Lengths of Maximal Exact Matches

Chapter 1

Introduction

Of the many topics I explored several made it to the thesis: pattern formation during solidification, the role of genetic exchange for microbial evolution, spontaneous emergence of genome biases, the optimality and universality of the genetic code. The first is concerned with quantitatively relating different levels of description and the rather unusual idea that a macroscopic phenomenology can be simulated accurately and efficiently by a suitably chosen microscopics. The last three projects are concerned with intrinsically *communal* aspects of biological evolution, the respective communities being closely related microbes, different tRNA species within a cell, all primitive organisms.

The first project lives in the realm of well formalized problems. It starts from the known equations of solidification, but uses a different level of description to make quantitative predictions tractable. The other three projects are quite different. They proceeded by accumulating knowledge about poorly-characterized and messy biological worlds and, then, isolating *systems* within them - certain levels of description which can be understood by minimal modeling. The motivation comes from conceptual questions such as: What is the way to classify microorganisms given that they exchange genetic material? How did the genetic code evolve if every change of the code leads to many amino acid substitutions? If it evolved, why is it universal and optimal in a specific sense to be discussed later?

1.1 Why is modeling not very popular in biology?

For good reasons, most researchers do not even consider modeling an option for answering the above questions. How do you model something that has an enormous number of different kinds

of degrees of freedom, most of which you don't know anyway? In addition, unlike well-studied condensed matter systems, you cannot perform controlled experiments on the system of interest as a whole. For example, it is conceivable that you can learn enough to keep most microbes alive in isolation from their environment and study them, but you cannot perform evolutionary experiments on (realistic) ecosystems. If you ask about early life it is even more hopeless - it is *history*. You can't even play with the system components - they might not exist anymore. History is a scary word to a physicist; history - not hysteresis. History is something that happened once to a single idiosyncratic complex system. Where is the universality that models must capture, if they are to be useful and predictive?

In summary, because a biocomplex system cannot usefully be broken down into component parts, it seems that modeling is an all or nothing proposition. Is there an alternative to giving up on theory, or to massive, fully-realistic molecular dynamics simulations of every atom?

1.2 From simplicity to complexity and back

Understanding something means finding its simplest explanation, or better still - a hierarchy of explanations that branch out from this simplest explanation to its different disguised instantiations. Therefore minimal modeling is a natural way to proceed, if you believe that, at least, the simplest insightful explanation is simple. But why would the explanation of a complex behavior be simple?

Well, to start with, we now know that some simple models lead to a spectacularly complex behavior. For example, the solidification equations lead to snowflakes. This gives hope: "Perhaps, the complexity of *my* system is generated by a simple model!" If you can't solve the inverse problem of complexity with incomplete experimental data (I can't), you can compare the *qualitative* behavior of the system against that of a virtual database of patterns generated from simple models. In this thesis, limited by the messiness of the biological context, and the fact that I spent most of the time looking for the right systems and questions, I never went into using simple models that can generate *baroque* complexity. I stuck with plain vanilla types of emergence. And the amazing thing is that once you learn how to think in terms of basic emergence motifs, instead of basic interactions, you start noticing different things. For example, if I see opposing tendencies, I think of phase transitions. If there is dynamics on top of this, perhaps there is pattern formation or

characteristic relaxation. If A changes B, and B changes A - this is a feedback. If it is positive, then it tries to blow up and pushes something called C. If it is negative - there is bi-stability. It is simpler than I would like to admit.

But notice what thinking in terms of motifs means. Just as a *group representation* can consist of many different types of irreducible representations, a given complex system can be a hierarchy, or, at least, a collection of many different simple models. So the question is not: is my system equivalent to a simple model? But, what is its decomposition into simple motifs? Therefore, it is useful to think in terms of easily identifiable motifs. It shifts you into another dimension. It is the difference between looking at a chip as a collection of logical gates and a collection of transistors. I emphasize this, because in physics we are used to focusing on the irreducible models where the most beautiful examples of emergence are. In biology, I believe, the reducible descriptions are the norm.

1.3 The renormalization group

Above we assumed that some complex systems are *equivalent* to simple models. Now we replace the word "equivalent" with the phrase *asymptotically equivalent when coarse-grained*. This is the realm of *the renormalization group*. Its philosophy has been so deeply ingrained in my thinking, for such a long time, that it is hard to convince myself that it is not obvious to all human beings, and is therefore worth writing down.

Simple models of coarse-grained systems are called *effective theories*. The equations of fluid dynamics are the effective theory for describing a system of many particles under a wide range of temperatures and pressures. The sharp interface model from the next chapter is the effective theory of a collection of atoms during solidification, and at the same time, it is the effective theory of the phase field model. There is a hierarchy - many different microscopic models have the same effective theory. What differentiates between them are *irrelevant degrees of freedom*. In condensed matter systems, out of an astronomical number of degrees of freedom at the microscopic level, only a few remain in the effective theory - *the relevant degrees*. The only thing that depends on the irrelevant degrees are the few parameters of the effective theory. A biological example of a parameter that is sensitive to the details is N_e - the effective population size. Now we are ready to do biology

as physicists. If we know *a bit* about a biological system we can suppose that this particular bit contains all the relevant degrees of freedom. Well, there is a danger that wishful thinking can push us in such a direction.

The fact that the parameters, of the otherwise universal theories, depend on the details, means, among other things, that it is non-trivial to relate quantitatively experimental measurements on the basic system components and their interaction strengths with the parameters of the *effective model*. This is exactly the problem treated in the next chapter. Then in chapter 3, I deduce that the ratio of mutation and recombination rates is a relevant parameter for determining the mode of bacterial evolution. But the ratio one measures in the lab, or deduces from looking at very closely related organisms, is different from the ratio that enters into the effective theory. Similarly, in chapter 4, I say that the degree of spontaneous codon usage bias in a genome depends on the mutations per genome per generation. But what the effective generation is, depends on the details of the actual life style of bacteria. It makes a difference whether most of the organisms leave one descendent, or one in a million leaves one million descendants.

Constructing a minimal model (an effective theory) for a class of systems is extremely useful even if you don't know how the system details affect the parameters of your minimal model. First, you can correlate different measurements performed on the same system, and in particular, make predictions about its future behavior, or response to various perturbations. Second, since the minimal model has only a few parameters, you can vary them and explore (analytically or numerically) the possible qualitatively different regimes of the dynamics. Finally, you can try to understand what different complex systems with the same minimal model have in common. Or, how apparently similar systems have different minimal models because of subtle *relevant differences*. These approaches are harnessed in chapter 3, and lead to a classification of microorganisms based on properties relevant to their communal evolution.

The renormalization group has provided us with numerous examples of non-trivial questions about complex systems that have simple answers, independent of most of the details. To give but one example: consider the question of determining the dependence of the size of a polymer in very dilute solution on the number of monomers. Common sense dictates that the answer is different for different polymers, and depends in a complicated way on their chemistry; to answer the question

theoretically you have to perform full blown molecular dynamics simulations for each specific case. It turns out, however, that the only fact that matters is that the polymer chain cannot cross itself. The universal, i.e. independent of any details, answer is

$$R \propto N^{0.588}, \tag{1.1}$$

where R is the polymer size, N is the number of monomers, and the result is valid in the asymptotic limit of large N. The corresponding minimal model of a polymer in very dilute solution is a *self-avoiding random walk* [1, 2, 3].

1.4 The role of context

Well, I am still a physicist. I almost missed this one. Time to wrap it up, I thought... If I were a mouse, would I have forgotten that it is a cat world?

Above we talked about systems as if there were not even bigger systems in which they are embedded. It is not just the low level that determines the effective theory or complexity of the upper level. The level above constrains the level below; it determines the *context*. An important example is that the ecosystem is the context of the cellular evolution.

As I am writing, I am refining my thinking. (This introduction evolves in the context of my thinking, and my thinking evolves in the context of writing an introduction). I want to deemphasize the notion of the context as the "bigger" level, since it is misleading. For example, (part of) the context of ecosystem evolution is the universal genetic code, which enables the continuous regrouping of biosynthetic pathways by horizontal gene transfer. The genetic code is neither bigger nor smaller. In a sense it is smaller, since it is determined at the molecular level. This is, by far, not the most trivial example but it is relevant to chapter 6. The fundamental notion is that of a coevolution between different levels.

In most of the condensed matter systems considered today, the rules at the microscopic level are fixed - the single electron properties do not adapt themselves to the shape and regime of use of a superconductor. So, the microscopic context is considered fixed. How about biology? Classical genetics assumes that the level below the unit of selection has an almost trivial structure - it ignores

the turbulent molecular context. Conversely, most of molecular biology focuses on the details of the level below, ignoring the context in which organisms live. As useful these approaches have been, to understand the evolution of complexity one has to understand the coevolution of levels above and below the level of selection. Welcome to (co)evolution!

This discussion is relevant to chapter 4 which deals with explaining the ubiquity of genome biases. Coevolutionary considerations [4] have been marginalized, in favor of the context independent selection-mutation-drift framework [5] that dominates current thinking. The good reason for this is that it enables analysis of experimental data in a model independent way. But it also creates a context that limits the conceptual progress in the field, at least in my opinion.

In chapter 5 we discuss the evolution of the genetic code towards optimality in terms of two coupled coevolutionary feedbacks. One is between the genetic code and the encoding of proteins, and the other - between codon usage and tRNA expression levels.

After saying all of the above, it is important to emphasize that the biological community is increasingly recognizing the importance of context and coevolution. This leads to the merging of ecology with evolution, especially through metagenomic [6] and phylogenetic approaches [7, 8], and the merging of evolution with development. Other students in Nigel Goldenfeld's group study the coevolution between microbial communities and the geochemical and environmental context in which they are embedded.

Trying to contrast the context-dependent phenomena with the pure universality of fundamental physics I run into some difficulties. Consider the law of inertia as an example of a universal truth. What Galileo Galilei said is effectively: "If you ignore the tendency of things to fall down, and if there were no air, then if you throw something, it will move in a straight line forever." So what? Who cares? In this world if you throw something it falls on the ground, moving with a constant speed is tiring - you stop to push and you stop to move. Even sitting in a chair for a long time is tiring[1].

This is science at its best. At the same time, it is amazing how irrelevant the universal truth seems at first. It is marginalized by the context. You have to look up at the stars, not around, to believe it is useful. The universal truth of Bose-Einstein condensation is even more puzzling.

[1] I have collected extensive experimental data.

The cubic micrometer where the universal truth holds, for perhaps a few seconds, is surrounded by many cubic meters of equipment, lasers, mirrors and cables; people spend their careers making it work. It is universal in the sense that, if there is intelligent life in another galaxy, they can make it work too.

The universal truth is manifested only in a very special context. This context can be an experimental setup, or an engineered device, such as this computer, which manifests the universality of computation, oblivious to the fluctuations of the physical world. Computers invaded and changed the context we live in. Even if the universality is marginalized by the context, harnessing it brings amazing technology that can radically change the very context that "humiliated" it.

As messy as this world is, with its zillions of degrees of freedom, it is full of highly precise *systems* that work according to *universal principles*. "These are artifacts", William Paley would say. Well, the context I live in is different... So I would say, "It is time for one more try on why universality and minimal models are useful in biology".

1.5 Evolution is the proud mother of simple models; physicists around the globe help yourselves

The weather has a mind of its own. Well, what is it thinking? What is the weather going to be on Sep 1st? I don't know, but I know that despite the complexity, randomness and messiness of their context, four very complex systems, namely, Yoshi Oono, Robert Clegg, Taekjip Ha and Nigel Goldenfeld will be in the same room at 3pm on Sep 1st, 2005. This is despite the fact that they have minds of their own. In fact, it is precisely because they have minds that they will be able to plan, overcome all the contingencies, and come.

There is this special type of complexity, called *life*, that not only wrecks havoc, but creates remarkably simple rules and order at all levels. Compare the complexity of the Earth's atmosphere with that of a microbial ecosystem. With both systems we can't make control experiments; both consist of many degrees of freedom. Are they equally amenable to minimal modeling?

The atmosphere is what it is[2]. It also coevolves with microbes, but let's forget about that. The microbial community has emerged through evolution. The microbes fine-tuned themselves to the

[2]Still on the same page with Paley in that regard

environment, they remodeled it, they coordinated their interactions. What were they doing?

They tried to position themselves in a context that *makes sense to them.* They like comfort too, you see. There is *homeostasis*, for example. Selves go at great lengths to ensure the relevance of their universal mechanisms by shielding them. This involves talking with other selves, since they are the most relevant part of the environment. This leads to communication protocols.

It might not come naturally to humans to think as bacteria, selfish genes, viruses, and who knows what else, but there is little doubt that humans would understand their games, if they were explained to them. And since there is no one to guide us through, we have to crack the codes ourselves. Many simple models - abstract games played at all levels, will be revealed. Some of them have been revealed.

Evolution has been actively generating abstract games - zillions of condensed matter systems to be explored.

1.6 Overview of the thesis

The thesis consists of five research chapters plus conclusions.

Chapter 2 describes my work on the phase field model of solidification. The physical process of interest here is the freezing of an undercooled liquid from an initial seed. This is a far from equilibrium process that leads to pattern formation[3]. Mathematically, the dynamics is formulated as a moving boundary problem - a partial differential equation for which non-trivial boundary conditions are imposed at a moving interface, and the motion of the interface is, in turn, determined by the field near it. Moving boundary problems are conceptually difficult, and the phase field model was proposed as a means to make numerical solutions tractable. The moving boundary formulation is the effective theory of the phase field model. My work was concerned with understanding quantitatively the connection between these two levels of description. It culminated with the construction of a new class of phase field models that enable efficient simulations in the experimentally relevant regime where attachment kinetics dominates surface tension effects. Apart from its scientific results, it introduced me, in a very explicit way, to the notion and practice of thinking about systems at two levels of description simultaneously, to fronts as relaxational modes of unstable states in

[3]This process is very similar to the one generating snowflakes.

some extended systems, to the renormalization group, asymptotics and singular perturbations.

These concepts proved invaluable once I started exploring the evolutionary role of the interplay between *homologous recombination* and *point mutations*, presented in chapter 3. Homologous recombination is a form of genetic exchange in microbes, in which a DNA fragment from a donor cell replaces a very similar portion of the genome of a recipient cell. This process is suppressed when the donor and recipient sequences are different. Mutations create sequence differences in microbial populations, opposing and suppressing recombination. The surprising result is that a microbial population in which homologous recombination is strong, resembles a one-dimensional undercooled liquid - diversification fronts, propagating along the genomes on the scale of many generations, can be triggered by seeds, such as the edges of genome rearrangements, or non-homologous horizontal gene transfer islands. In certain scenarios, these diversification fronts lead to speciations. The existence of diversification fronts is contingent on sufficiently strong homologous recombination, but also on certain details of the cellular mechanisms suppressing homologous recombination. Correspondingly, we propose a *classification of microorganisms* based on the *relevant* details of homologous recombination mechanisms. Organisms belonging to different classes have different collective modes of their evolution. In addition, we elucidated the expected signature of front propagation events and scanned the completely sequenced genomes, and looked for the signature in closely related pairs of genomes. A plausible candidate for a front propagation event was identified.

In chapter 4, I give a condensed matter physics perspective to the problem of genome biases, such as codon usage bias and GC-content bias. As I argued above in a light-hearted way, evolution leads to beautiful examples of universal mechanisms shielded from the environment. Here, I focus on some of the universal aspects of *template-directed synthesis*. A template-directed synthesis is a process in which a sequence of letters in the template is converted to a sequence of letters/monomers[4] in the synthesized product, according to some rule. The synthesis machinery moves along the template, waiting at each step for the binding of the correct product monomer. A pool of product monomers is generated independently. Examples of template-directed processes are translation, replication and transcription. The main insight is that, in the presence of some functional/language degeneracy, the alphabet composition of the template *coevolves* with the allocation of resources for

[4]The alphabets of the template and the product are different in general.

the maintenance of the pool of monomers. Selection on the speed, accuracy and energy efficiency of template-directed synthesis leads to *spontaneous symmetry breaking* which favors some letters over others. This process is opposed by mutations which tend to randomize the alphabet usage. In the long template limit, there is a continuous phase transition, with the mutation rate as a control parameter, and the alphabet usage bias as the order parameter.

The above idea is coopted in chapter 5 to explain how the genetic code can change and evolve towards optimality. There is evidence, coming from comparing the standard genetic code with ensembles of random codes, that the genetic code is optimized to minimize the phenotypic effect of mistranslations and mutations. Similar amino acids are coded by codons that differ by a single letter. This evidence strongly suggests that the code evolved. At the same time, every change of the code seems to lead to many simultaneous amino acid substitutions, and is therefore lethal. How could the code have evolved if it could not change? I propose that the coevolution between tRNA expression levels and codon usage catalyzes changes to the code, and present a closed model of the genetic code evolution that confirms this expectation.

After we exposed evolutionary mechanisms that enable code change, the *universality* of the genetic code seems more puzzling than ever. Given that there is an astronomical number of possible genetic codes, and that the codes can change, why isn't there a great diversity of codes today? In chapter 6, a communal explanation of universality is presented and contrasted with more conventional explanations. Following a suggestion by Carl Woese [9] that early life was dominated by horizontal gene transfer (HGT), we explored the mechanisms through which HGT can bring universality. Both exchange of protein coding regions and translational components is important. HGT of protein coding regions between organisms with similar codes provides an effective attractive force between the codes. At a larger scale, it partitions the organisms into communities of incompatible codes. Driven by innovation sharing, these communities compete for niches. Because larger communities have larger pools of innovations, more popular codes are favored. This is a "winner takes all" dynamics which generically leads to universality. In addition, the universality of the code went hand in hand with a universal translational standard that was maintained by and enabled exchange of *code specificators*[5]. Exchange of code specificators facilitated the coordinated

[5]tRNAs for example.

optimization of the genetic codes of organism occupying diverse niches, i.e. the universality of the code might not have come after optimality. I extended the model of chapter 5 to include the effects of HGT.

In the conclusion this section is rewritten emphasizing the results and their significance. A synthesis of what we learned is provided and mapped on the tree of life.

1.7 Accomplishments

The work presented in chapters 2 and 3 is already published, as listed below. Chapters 4, 5 and 6 will appear as three separate articles.

Here is a list of the main results of this thesis:

- Proposed a new class of phase field models that allows efficient quantitative simulations of solidification in the regime where interface kinetics dominates capillary effects.

- Suggested the possibility that the interplay between homologous recombination and point mutations leads to diversification fronts and speciations in microbial communities. Proposed a classification of the modes of communal microbial evolution based on properties of the cellular mechanisms of homologous recombination. Examined the genome data and singled out potential signatures of diversification fronts in the *Bacillus cereus* group.

- Proposed that genome biases are instances of spontaneous symmetry breaking, and are a generic result of selection on the speed, accuracy and efficiency of template-directed synthesis. Constructed an exactly solvable model of selection on speed, which exhibits a continuous phase transition, and singled out the *mutation rate per genome per generation* as the only relevant parameter. Discussed the role of "tunnelling" between different quasi-stable genome bias patterns. The extensive genome data enables tests of corollaries of this picture.

- Offered a conceptual explanation of how the genetic code can evolve despite apparent barriers to its change. Constructed a closed model of the evolution of the genetic code, and showed that the coevolution between tRNA expression levels and codon usage "lubricates" the evolution of the code towards optimality. Pointed out the limitations of the popular notion of an overall *amino acid similarity matrix* for studies on the evolution of the genetic code.

- Presented a conceptual shift in thinking about the universality of the genetic code - the code is universal as a result of extensive exchange of genetic material during the early cellular evolution. Elucidated the different channels through which horizontal gene transfer leads to universality. Introduced the notion of a universal translational machinery protocol that makes code specificators compatible between cells.

1.8 List of Publications

1. "Global divergence of microbial genome sequences mediated by propagating fronts", Kalin Vetsigian and Nigel Goldenfeld, PNAS **102**, 7332 (2005).

2. "Computationally efficient phase-field models with interface kinetics", Kalin Vetsigian and Nigel Goldenfeld, Phys. Rev. E **68**, 060601(R) (2003).

Chapter 2

Phase Field Model

2.1 Introduction

Solidification processes have attracted a lot of attention in the last two decades as a rich source of pattern formation physics and due to their technological relevance [10, 11, 12, 13]. A major and challenging goal has been the accurate and efficient prediction of solidification patterns starting from a set of equations describing the evolution of the thermal field around the solidification front. Typically one is interested in the growth of a crystal seed into undercooled liquid. Depending on the initial conditions and the material properties the growing crystal can develop into many different morphologies such as dendrites and sea weeds.

The difficulty in numerically integrating the solidification equations comes from the fact that they constitute a moving free boundary (sharp interface) problem. The position of the interface has to be determined self consistently and nontrivial boundary conditions have to be imposed on it. Furthermore, numerical artifacts tend to be amplified by an inherent instability of the interface, called the Mullins-Sekerka instability [10], that is cutoff only on a scale many orders of magnitude smaller than the patterns it helps to create.

To cope with the above difficulties the phase field approach has been proposed [10] and after the work of [14] it became the method of choice for studying solidification. Provatas et al. [12] further popularized the approach by developing an efficient adaptive mesh refinement algorithm for it. The phase field model avoids the problem of tracking and imposing boundary conditions on a complicatedly moving interface by the introduction of a new field that distinguishes the two phases and smears the interface over a finite length scale W, effectively regularizing the problem.

A major obstacle to the application of the phase field model for quantitative studies is the need to relate the parameters of the phase field model to that of the sharp interface model. The theoretical question here is that of relating two different levels of description: a macroscopic one - corresponding to the sharp interface model, and a microscopic one - corresponding to the phase field. The Renormalization Group way of thinking about the problem suggests that there are many ways to construct phase field models that have the same macroscopic behavior. The difficult part is to design phase field models that are both computationally efficient and easily analyzable, i.e. ones for which a simple recipe can be found for choosing the microscopic parameters that solve a problem with given macroscopic parameters. The difficulty stems from the fact that the asymptotic analysis required to relate the two levels of description cannot be performed exactly, and typically, singular perturbation techniques are used.

Caginalp and Chen [15] were the first to perform asymptotic matching analysis of the phase field equations giving a connection between the parameters. However their analysis was valid only in the limit of interface thickness much smaller than the capillary length, which is prohibitively expensive computationally even if adaptive mesh refinement is used. This limit is too severe because it requires that the new length and time scales introduced by the phase field are much smaller than the smallest scales present in the formulation of the sharp interface model. Karma and Rappel improved the asymptotic analysis for the same class of phase field models by demonstrating that even for interface thickness of the order of the capillary length the desired sharp interface limit is achieved, though with a different relation between the parameters of the sharp interface and the phase field models. Their work was essentially a second order asymptotic matching analysis - an order higher than that in [15].

In light of the above, a question that naturally arose early in my research was whether it is possible to improve Karma and Rappel's result even further by using more sophisticated techniques for performing asymptotic analysis such as the RG for differential equations proposed in [16]. The answer to this question was unfortunately negative - higher order asymptotic analysis would only reveal that the phase field model used with large W generates interface terms not present in the sharp interface model. There is no known tractable procedure for nullifying or controlling these terms by fine-tuning the parameters and functional forms that enter into the phase field model.

Even if such procedure were found it would most likely introduce delicate fine structure in the phase field profile that would offset the computational benefits because of the need to resolve length scales even smaller than W. Therefore we focused our attention on designing new classes of phase field models for which larger values of W can be used without loss of accuracy.

Here I present a modification of the phase-field model of solidification so as to enable efficient computations in the regime where interface kinetics is the dominant factor. This regime is of current theoretical interest stimulated by the experimental observation of the puzzling morphological transition of the solidification front of Ni at high undercooling [17, 18, 19, 20]. The modification allows one to use an interface thickness many times larger than the capillary length. The methodology for performing the asymptotic analysis is different and much easier to perform systematically to high order than the asymptotic matching employed in all previous analyses, and is capable of being used in other free boundary problems.

2.2 Two levels of description

2.2.1 Sharp interface model

The symmetric model for the solidification of a pure melt from the liquid (L) phase to the solid phase (S) is defined by the equations:

$$\partial_t u = D\nabla^2 u \tag{2.1}$$

$$\partial_n u|_S - \partial_n u|_L = V/D \tag{2.2}$$

$$u_i + d_0 k = -\mathcal{B}(V) \ . \tag{2.3}$$

Here $u = (T - T_M)c/L$ is the dimensionless temperature and u_i is its value at the solidification front, with T being the temperature in the liquid or solid, T_M being the melting temperature of a planar interface, c being the specific heat and L being the latent heat of fusion per unit volume. The curvature of the solidification front is given by k and the capillary length is d_0. D is thermal diffusivity (assumed here to be the same in both phases), and V is the normal velocity of the front. Equation (2.2) expresses heat conservation at the interface, and equation (2.3) is a modified Gibbs-

Thomson condition, which is a statement of local equilibrium at the interface with the attachment kinetics included through the term $\mathcal{B}(V)$. Traditionally, a linear kinetic undercooling $\mathcal{B}(V) = \beta V$ is used. It should be stressed, however, that the linearity of u_i with respect to velocity is a just a simplifying assumption, valid at small enough undercoolings. The only constraint that follows from fundamental considerations is that u_i is linear in k at zero velocity. Molecular dynamics simulations [21, 22] suggest that there are substantial deviations from linearity of $\mathcal{B}(V)$ at large undercooling. In this chapter we are interested in materials with large dimensionless parameter $\tilde{\beta} \equiv \beta D/d_0$. This constant is a measure of the importance of interface kinetics for a given material. It takes very different values for different materials, and for Ni it is estimated from molecular dynamics simulations to be as high as 90 [23].

2.2.2 Phase-field model

The phase-field equations can generally be written in the form:

$$\tau \partial_t \psi = W^2 \nabla^2 \psi - f_\psi(\psi) - \lambda u g_\psi(\psi) \tag{2.4}$$

$$\partial_t u = D \nabla^2 u + \frac{1}{2} \partial_t h(\psi). \tag{2.5}$$

Here ψ represents the phase-field, $f(\psi)$ is a double well potential, $g(\psi)$ shifts the relative height of the two minima making one of the phases metastable for $u \neq 0$. Note that we have used the subscript notation to denote differentiation: hence $g_\psi(\psi)$ means $\partial g/\partial \psi$. The sharp interface boundary is recovered as the locus of points where $\psi = 0$, and we are interested in the behavior of the phase-field equations as the phase-field interface width W and relaxation time τ tend towards 0. In order to solve the desired sharp interface model, we need to ascertain what phase-field parameters $\{\tau, W, \lambda, f(\psi), g(\psi)\}$ should be used given the values of $\{D, d_0, \beta$ or $\mathcal{B}(V)\}$.

2.3 Asymptotic analysis

In the regime when u and ψ profiles near the interface adiabatically adjust to changes in u away from the interface and for $kW \ll 1$ the problem can be reduced to a one dimensional nonlinear

eigenvalue problem:

$$\begin{aligned} \partial_\xi^2\psi - f_\psi(\psi) + (v+q)\partial_\xi\psi - \lambda u g_\psi(\psi) &= 0 \\ \partial_\xi^2 u + (p+q)\partial_\xi u - \frac{1}{2}p\partial_\xi h(\psi) &= 0 \\ \psi(\pm\infty) = \mp 1,\ |u(\pm\infty)| &< \infty \end{aligned}$$

where

$$q = kW,\ v = \frac{V\tau}{W},\ p = \frac{VW}{D}\ ,$$

$\xi = r/W$ and r is the local coordinate perpendicular to the interface. For each $\{v, p, q\}$ unique solutions for $u(\xi)$ and $\psi(\xi)$ exist. Given $u(\xi)$ one can extract the outer limit $u_{out}(r) = \lim_{W\to 0} u(r/W)$, and the value at the interface of this outer limit to compute the effective Gibbs-Thomson condition $u_i = u_i(v, q, p)$.

The limit of small W is a singular one because W multiplies a highest order derivative in the original variables. All previous works use asymptotic matching to deal with the singularity and assume $v, q, p \ll 1$. Here we will demonstrate a simpler approach based on the fact that the equation for u is linear so that it can be trivially solved in terms of ψ. Requiring that u is finite at $r \to \pm\infty$ and requiring that $\psi(r) \to \mp 1$ sufficiently fast one obtains

$$u(\xi) = u_0 + \frac{1}{2}pe^{-(p+q)\xi}\int_{-m\infty}^{\xi} e^{(p+q)\eta}h(\psi(\eta))d\eta \tag{2.6}$$

$$= u_0 + \frac{1}{2}p\left(\hat{u} + \frac{1}{p+q}\right)\ , \tag{2.7}$$

where $m \equiv \mathrm{sgn}(p+q)$ and the last line defines $\hat{u}$. The dependence on $p+q$ expresses the singular nature of the problem with respect to this parameter. From $u(\xi)$ one can compute the outer limit $u_{out}(r)$ and thus obtain

$$u_i = u_{out}(0) = u_0 + \frac{1}{2}\frac{p}{p+q}\ . \tag{2.8}$$

Despite the singularity, if we are interested only in the profile of u near the interface one can

expand in powers of p and $p+q$ and to first order obtain

$$u(\xi) = u_i + \frac{1}{2}p\int_{-\infty}^{\xi}\left(h(\psi(\eta)) - 1\right)d\eta + O\left(p(p+q)\right) \ . \tag{2.9}$$

Using this expression and substituting it back in the equation for ψ we get an equation for ψ which can be solved using regular perturbation theory. In this way we recover the standard asymptotic result of Karma and Rappel:

$$W(\lambda) = \lambda\frac{d_o}{a_1}, \qquad \tau(\lambda) = \lambda^2\left(\frac{\beta D}{d_0} + a_2\lambda\right)\frac{d_0^2}{a_1^2 D}, \tag{2.10}$$

where a_1 and a_2 are constants depending on $f(\psi)$ and $g(\psi)$. Notice also that since ψ is monotonic one can compute the distance from the interface ξ from ψ and in this way write to order p

$$u(\xi) = u_i + \frac{1}{2}pF_1\left(\psi; \{\psi(\xi)\}\right), \tag{2.11}$$

where F_1 is defined as the integral in equation (2.9).

From (2.10) follows that with the functions $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ fixed, λ is the only free parameter.

2.4 Computational complexity

We now examine how the computation time t_c for the phase-field model scales with the free parameter λ in a discretized calculation with adaptive mesh refinement and uniform grid elements in d dimensions. Clearly, t_c depends on the width of the phase-field boundary layer $\tilde{W}$, the space resolution Δx, and the time step Δt. The inverse computation time scales as:

$$t_c^{-1} \propto \frac{\Delta t}{t}\frac{\Delta x}{\tilde{W}}\left(\frac{\Delta x}{L}\right)^{d-1}, \tag{2.12}$$

where L^{d-1} is the order of the surface area and t is the maximum time one wishes to evolve the system. For a spatially explicit numerical scheme $\Delta t \leq \frac{1}{2}\tau\left(\Delta x/W\right)^2/\lambda$. The factor of λ is included to guarantee accuracy in the presence of the term $\lambda u g_\psi$. Collecting terms, and using $\tilde{W} \approx W \propto \lambda d_0$

and $\tau \propto \lambda^2$ we obtain

$$t_c^{-1} \propto \lambda^d \left(\frac{\Delta x}{W}\right)^d \tag{2.13}$$

showing that the computation time is highly sensitive to λ and also depends on the spatial resolution required by the shape of the interface profile of ψ: smoother profiles are better!

While it would be computationally efficient to work with large λ, doing so would introduce higher order terms in the curvature k and velocity V into the Gibbs-Thomson condition (2.3). How can we do better?

2.5 A phase field model for large undercooling

Using equation (2.10) we see that we require $v = (\tilde{\beta} + a_2\lambda)p \ll 1$ which puts a very severe constraint on p if $\tilde{\beta}$ is large. Computationally this would be more important at large undercooling when a thin temperature boundary layer forms around the solidification front, and correspondingly $q < p$ so that the smallness of p is the limiting factor. (For example at undercooling $\Delta = 0.8$ the theory predicts $q/p = 1/7$ at a steady state dendrite tip.) Therefore, a good objective is to modify the phase-field in a way that relaxes the constraint on v.

A step in that direction was made by Bragard *et al.* [23], who replaced λu by $H(\lambda u)$ in equation (2.4). $H(\cdot)$ is computed numerically by solving the following non-linear eigenvalue problem with appropriate boundary conditions on ψ:

$$\frac{d^2\psi}{dx^2} - f_\psi(\psi) + v\frac{d\psi}{dx} - H(v)g_\psi(\psi) = 0 \ . \tag{2.14}$$

With $H(\cdot)$ chosen in this way the non-linearities appearing in the standard phase-field model at large v are cancelled by the non-linearities in $H(\cdot)$. To relate the parameters they use

$$d_0 = \frac{W}{\lambda}, \ \beta = \frac{\tau}{\lambda W}. \tag{2.15}$$

However, this relationship is valid only in the limit of vanishing p, i.e. when u is approximately constant across the diffuse interface - a result analogous to that of Caginalp for the standard phase-field. Correspondingly, in [23] a value of p close to 0.01 is used in computations.

Since $v+q$ is no longer a small parameter it is analytically more involved to derive corrections for finite p. To compute the linear part it is enough to consider small v and the result is

$$\tau = \lambda W(\beta + a_2 \frac{W}{D}) \tag{2.16}$$

which is the analog of Karma and Rappel's formula.

Replacing (2.15) by (2.16) allows much larger values of p to be used and is a straightforward way to make better use of the model proposed in [23]. Numerically we observed that for $g_\psi = (1-\psi^2)^2$, the form used in [23], the non-linearities in k and V are weak even for values of v of the order of 20. However, the Bragard *et al.* phase-field model, even with the improved asymptotics (2.16) that we derived, still does not provide the desired degree of computational improvement because the phase-field profile develops a new length scale of order $W/H(v)$ which needs to be resolved numerically in order to avoid artifacts. In addition, $H(v)$ increases very rapidly with v.

2.6 A new class of models

We propose to replace τ by $\tau_R(\psi)$ in such a way so that the effective equation for ψ becomes

$$\tau \partial_t \psi = W^2 \nabla^2 \psi - f_\psi(\psi) - \lambda u_i W |\nabla \psi| \ . \tag{2.17}$$

What are the advantages of doing this? The asymptotic analysis is greatly simplified because the equation for ψ can be analyzed separately from that for u. The solution for ψ is simply

$$\psi(\xi) = \psi_0(\xi) \ , \tag{2.18}$$

where $\psi_0(\xi)$ is the solution of $\partial_\xi^2 \psi_0 - f_\psi(\psi_0) = 0$, and the relation between the parameters is simply given by (2.15).

Now we can rewrite (2.17) as

$$\tau \partial_t \psi = W^2 \nabla^2 \psi - f_\psi(\psi) - \lambda u W |\nabla \psi| - \lambda \frac{1}{2} p F_1(\psi) W |\nabla \psi| \tag{2.19}$$

with the only problem being the presence of p in the evolution equation. The final trick is to express p in terms of $\partial_t\psi$:

$$p = \frac{W}{D}V = -\frac{W}{D}\frac{\partial_t\psi}{\partial_x\psi} = \frac{W}{D}\frac{\partial_t\psi}{|\nabla\psi|}. \tag{2.20}$$

The equation for ψ becomes

$$\tau_R(\psi)\partial_t\psi = W^2\nabla^2\psi - f_\psi(\psi) - \lambda u W|\nabla\psi|, \tag{2.21}$$

where

$$\tau_R = \tau - \frac{1}{2}\lambda\frac{W^2}{D}F_1(\psi)\ . \tag{2.22}$$

For $f(\psi) = \frac{1}{4}\left(1-\psi^2\right)^2$ and $h(\psi) = \psi$ we have $F_1(\psi) = \sqrt{2}\ln\left((\psi+1)/2\right)$. It follows that $\tau_R \geq \tau$ which means that the model is well behaved. The expression for τ_R can be compared with Karma and Rappel's formula which can be rewritten in the form $\tau' = \tau + a_2\lambda\frac{W^2}{D}$. a_2 is approximately the value of $-1/2F_1(0)$. As it stands $\tau_R(-1) = \infty$ and points with $\psi = -1$ cannot evolve, so some cutoff near $\psi = -1$ should be introduced. Experiments show that results are insensitive to the exact form of the cutoff.

The restriction of order p accuracy comes from expanding $u(\xi)$ near the interface. It is possible to go to higher orders in p or simply use the full expression (2.6). This would result in an *implicit* equation for $\partial_t\psi$

$$\tau\partial_t\psi = W^2\nabla^2\psi - f_\psi(\psi) - \lambda u W|\nabla\psi| - \frac{1}{2}\lambda\frac{W^2}{D}\partial_t\psi\hat{u}(\psi, p+q). \tag{2.23}$$

The function $\hat{u}$ can be tabulated in advance and equation (2.23) can be solved iteratively at each time step.

Different approximations to eqn. (2.23) lead to different schemes. For example if we consider the next order term in (2.9) $\frac{1}{2}p(p+q)F_2\left(\psi;\{\psi(\xi)\}\right)$ we end up with a quadratic equation for $\partial_t\psi$.

The model including p^2 corrections is:

$$\begin{aligned}
\tau_R &= \tau - \tfrac{\lambda W^2}{2D}\left(F_1(\psi) + qF_2(\psi)\right) \\
\tau_R(\partial_t\psi)_0 &= W^2\nabla^2\psi - f_\psi(\psi) - \lambda W u\,|\nabla\psi| \\
\alpha &= (\partial_t\psi)_0 \tfrac{1}{\tau_R}\tfrac{\lambda W^4}{2D^2}\tfrac{F_2(\psi)}{g_\psi(\psi)} \\
\partial_t\psi &= (\partial_t\psi)_0 \tfrac{1-\sqrt{1-4\alpha}}{2\alpha} = (\partial_t\psi)_0(1+\alpha+2\alpha^2+\ldots).
\end{aligned} \tag{2.24}$$

In the evolution equation $q = Wk = W\nabla\cdot\mathbf{n}$ with $\mathbf{n} = \nabla\psi/|\nabla\psi|$.

Another application of the technique is to correct for terms of order pq at small undercooling when $p \ll q$. In this regime terms of order pq are not negligible compared to terms of order p. To achieve this it is enough to use $(\partial_t\psi)_0$ from above without correcting it.

The phase-field model can be generalized to handle arbitrary interface kinetics $u_i = -d_0k - \mathcal{B}(V)$ at order p and arbitrary v. The resultant phase-field model equation is

$$\begin{aligned}
\tau_R(\psi,u)\partial_t\psi &= W^2\left(\nabla^2\psi - k(\nabla\psi\cdot\mathbf{n})\right) - f_\psi(\psi) \\
&- \mathcal{B}^{-1}\left(\lambda(u+d_0k)\right) W\,|\nabla\psi| \qquad (2.25) \\
\tau_R(\psi,u) &= \tau - \frac{\lambda W^2}{2D}F_1(\psi)\mathcal{B}^{-1\prime}\left(\lambda\left(u+d_0k\right)\right). \qquad (2.26)
\end{aligned}$$

The above recipe for improving phase-field models can be used also in cases when the ψ profile changes with V and k. For example it can be applied to the model of Bragard *et al.* by effectively replacing u with u_i yielding

$$\tau_R(\psi,u)\partial_t\psi = W^2\nabla^2\psi - f_\psi(\psi) - H(-\lambda u)g_\psi(\psi), \tag{2.27}$$

with

$$\tau_R(\psi,u) = \tau + \frac{\lambda W}{2D}H'(-\lambda u)\tilde{F}_1\left(\psi, H(-\lambda u)\right)\frac{g_\psi(\psi)}{|\nabla\psi|}\ . \tag{2.28}$$

In this case the expression for τ_R is more complicated because ψ changes with v. The functions $H(v)$ and $\psi_v(\xi)$ which solve equation (2.14) and F_1 can be pre-computed numerically leading to a very efficient numerical scheme.

2.7 Numerical experiments

We now compare the performance of different phase-field models in one-dimensional simulations. The benchmark problem solved is $u(t=0,x) = -\Delta$ for $x \in (-\infty, \infty)$ with the solid-liquid interface initially at $x = 0$. The interface velocity, $V(t)$, is compared to that for the sharp interface model obtained via direct numerical integration.

The models compared are identified as follows. *ST*: standard phase-field model (2.4); *BR*: Bragard *et al.* model with asymptotic relation (2.15); *BR+*: the above model with the improved asymptotic relation (2.16); τ_R: the new model (2.21); $\tau_R + p^2$: the new model with p^2 corrections (2.24); τ_R *BR*: the improved version of *BR* given in (2.27). We used $h(\psi) = \psi$ and $g_\psi = (1-\psi^2)^2$ everywhere.

As an example, we computed the velocity of the front after $t = 3.5 \times 10^4 d_0^2/D$ for $\tilde{\beta} = 10$, $\Delta = 1.2$ and $\lambda = 15$. The exact result is $Vd_0/D = 0.021$. The results for models *ST*, *BR*, τ_R *BR* were 0.012, 0.044, 0.020 respectively, showing that the previously existing models are inadequate in this regime.

Figure 2.1 compares the systematic deviations of various phase-field models from the sharp interface solution as a function of λ. One clearly sees that *BR* model leads to errors linear in λ. For *BR+* unintended nonlinearities in the Gibbs-Thomson condition quickly increase the error with λ. In contrast τ_R and $\tau_R + p^2$ models yield approximately the same velocity for the entire range of λ considered. Using the values for Ni cited in [23], $D = 10^{-5}\mathrm{m}^2/\mathrm{sec}$ and $d_0 = 5.56 \times 10^{-10}$, the choice $\Delta = 1.2$ corresponds to a steady state velocity $V = (\Delta - 1)D/(\tilde{\beta} d_0) = 40\mathrm{m/sec}$ which is approximately where the experimentally observed morphological transition occurs.

To match the accuracy of τ_R model with $\lambda = 30$ we need to take about $\lambda = 5$ in *BR* (measuring deviations from the limiting phase field value). In 3D, this leads to about $(30/5)^3 \approx 200$ times increase in computational speed as compared to the simulations in [23]. If we use the Bragard *et al.* model with our improved asymptotics, the new $\tau_R(\psi)$ models will be about $3^3 = 27$ times faster. The above figures are just for illustration, the precise computational gains will depend on the desired accuracy and the regime of interest.

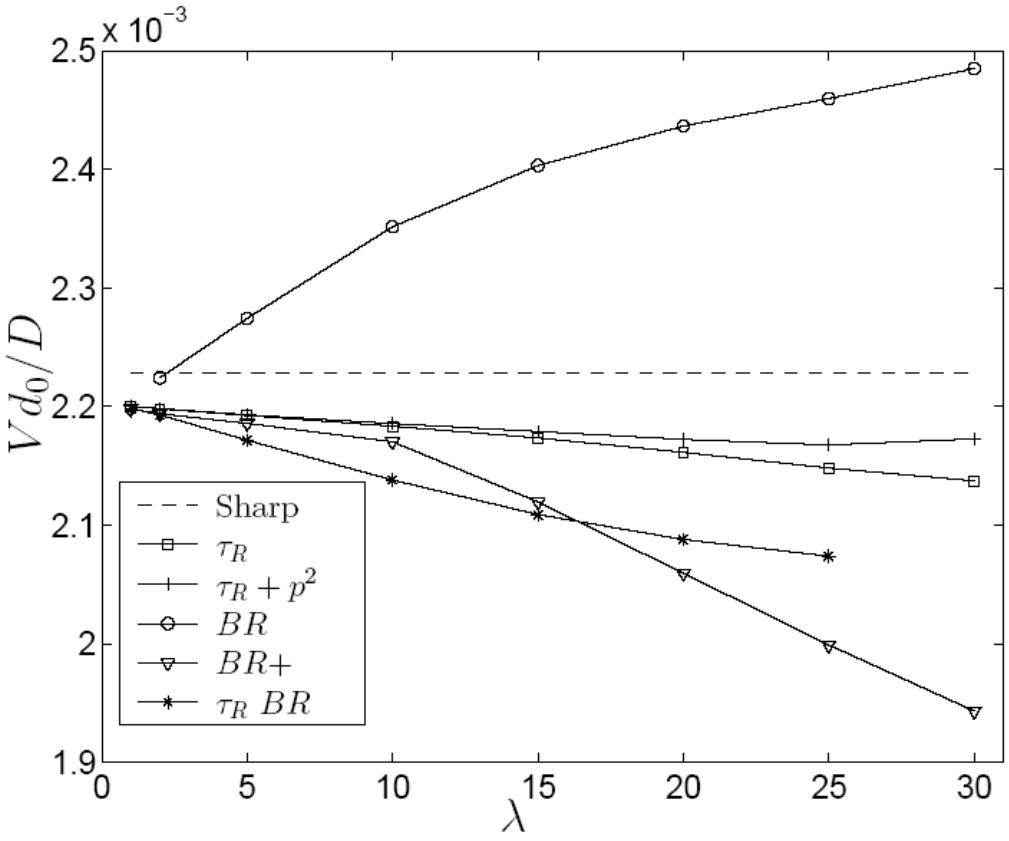

Figure 2.1: Interface velocity from different phase-field models as a function of λ. $\tilde{\beta} = 90$, $\Delta = 1.2$. At $t = 8 \times 10^6 d_0^2/D$. $\Delta x = 0.5W = 0.5\lambda d_0$.

In conclusion, the models described here are the first that can systematically handle interface kinetics dominated growth in and beyond the thin-interface limit enabling huge gains in computational efficiency.

Chapter 3

Global divergence of microbial genome sequences mediated by propagating fronts

3.1 Introduction

The transfer of genetic material between microbial cells plays a crucial role in their evolution, and poses fundamental questions to microbiology. Is there a tree of life for microbes [9, 24, 25]? Are there bacterial species [26, 27]? What are the mechanisms driving their diversification [28, 25, 29, 30]? These questions arise because genetic transfer couples the evolution of different genomes in a way that not only complicates their dynamics but obscures their very identity over time: the evolution is communal. While in sexual organisms the communality of genome evolution is restricted to species, the major elements of microbial evolution—genetic transfer followed by illegitimate or homologous recombination, point mutations, genome rearrangements—do not *a priori* imply sharp genetic isolation boundaries. If there are none, notions such as species and speciation, despite being widely used heuristically, are misleading. Also, it is not clear whether there are classes of microbes with qualitatively different modes of communal evolution and what are the cellular properties that distinguish between them.

Gene transfer results when foreign DNA is taken up from the environment (transformation), delivered by a virus (transduction) or acquired via a direct cell to cell exchange (conjugation), and then permanently incorporated in the recipient genome by homologous or illegitimate recombination. Homologous recombination, mediated by dedicated cellular machinery, plays a vital error

correction role in genome replication [31] but also allows a foreign DNA fragment to replace a sufficiently similar portion of the recipient genome. The probability of successful replacement in homologous recombination is proportional to the exponential of the number of sequence mismatches [32], the mechanism being organism-specific [33, 34, 35]. Illegitimate recombination can be mediated by bacteriophage integrases, selfish genetic elements, or occur by chance DNA breakage and repair, and allows the acquisition of entirely novel traits from evolutionary distant organisms. Illegitimate genetic transfer, also known as horizontal gene transfer (HGT), can be inferred from the genome data through its atypical sequence composition [28] and the phylogenetic incongruences it causes [36]. While the extent of HGT is under heated debate [24], it is clear that it is much less frequent than homologous recombination. Relative rates of homologous recombination and point mutations in natural populations have been estimated by sequence diversity studies using multi-locus sequence typing data in recently-formed bacterial strains [37, 38]. The probability that a gene changes as a result of homologous recombination can be many times higher than that for point mutations. Another manifestation of the pervasiveness of homologous recombination is that the evolution of strains within many named species cannot be represented by a phylogenetic tree [39, 40, 41]. While the importance of genetic transfer, and homologous recombination in particular, is firmly established [42], there are only a few sharp predictions about the resulting modes of microbial evolution. Relevant to our work is the observation of Lawrence [26] that HGT islands locally inhibit recombination. He concludes that global genetic isolation can be achieved through the gradual accumulation of hundreds of HGTs.

The purpose of this chapter is to explore the emergent properties of the collective evolution of closely related bacterial genomes. We model the interplay of homologous recombination and point mutations in bacterial populations, and show that elementary genome changes such as HGT, genome rearrangements, insertions or deletions can trigger diversification fronts that in evolutionary short time propagate along the bacterial genomes and eventually lead to global sequence divergence of sub-populations. The diversification fronts can occur even in the absence of natural selection and demonstrate that fast neutral evolution can have non-trivial long-term evolutionary consequences. The robustness of this mechanism is sensitive to some of the details of homologous recombination, and suggests a way to classify the spectrum of evolutionary modes in bacteria based on specific

details of their homologous recombination mechanisms. We establish a methodology for analyzing closely related genomes and give evidence for a large-scale step-like variation of homologous recombination rates in the *Bacillus cereus* group, which might be a signature of a diversification front. Finally, we discuss the biological implications of the propagation of diversification fronts, as a mechanism for speciation, a force favoring the formation of sharp genetic isolation boundaries, and a dynamical barrier for HGT and genome rearrangements.

3.2 Homologous recombination

The details of homologous recombination are by now reasonably well understood [32, 33]. There are at least two common obstacles to successful integration of a DNA fragment. First, the end of the fragment must find a short region (≈ 20 bp) of sequence identity with the target genome in order to initiate the process. Second, the cell's mismatch repair system can abort the recombination process if it encounters mismatches between the fragment and the portion of the genome being replaced. Both of these obstacles lead to an exponential decrease of recombination with sequence divergence. There are also potentially important variations in the mechanism. While in *E. coli* sequence identity at only one end is required, in *Bacillus* very high sequence similarity at both ends is needed [33, 34] and mismatch repair seems less important. In *Streptococcus* the effect of mismatch repair is intermediate in strength [35] but the overall dependance of sexual isolation on sequence divergence is very close to that in *Bacillus*. In addition, the underlying bases for distinguishing between donor and recipient DNA can differ. Do these differences in the details translate into qualitatively different evolutionary behavior? If so, then the details of the homologous recombination mechanism could be an important criterion for classifying bacteria. The computational studies described here clarify which details are the relevant determinants of the long-term evolutionary dynamics.

3.3 Models

Based on the above considerations, we construct sets of model rules that describe the interplay between homologous recombination and point mutations.

1. There are N circular strings of length L written in an alphabet of n symbols.

2. Each position in each genome is subject to point mutations with rate m. A point mutation changes a symbol to any other symbol with equal probability.

3. Each genome receives fragments at an average rate r. Each fragment is of size F, is derived from an arbitrary position from an arbitrary donor genome and attempts to recombine at the same genome position in the recipient.

4. To be considered for incorporation the fragment must find an identical segment of length M at an arbitrary chosen end (Model I) or at both ends (Model II).

5. The probability of incorporation is $\exp(-\alpha d)$, where α is a coefficient expressing the strength of the mismatch repair system and d is the pointwise sequence difference, i.e. d counts the number of mismatches between the fragment and the genome sequence it is about to replace. We will also consider Model III, where rule 4 is absent.

The genome strings can be thought of as representatives of different strains possessing at least partial ecological distinctiveness, so that random genetic drift is much stronger within strains than between strains. With this interpretation we do not include random genetic drift but it can be straightforwardly added.

3.4 Propagation of diversification fronts

In these models, mutation and recombination play opposing roles: point mutations generate sequence diversity in the population, whereas recombination tends to make sequences more similar. At high recombination rates an initially uniform population will remain close to uniform; at high mutation rates all sequences will diverge from each other. An important property of homologous recombination is that the probability that a recombination event is successful decreases with sequence divergence and becomes negligible even for small levels of divergence [32].

These considerations suggest that the uniform phase is *metastable*: even when recombination is strong enough to maintain a state of near uniformity it will not succeed in bringing together sufficiently diverged sequences. The diverged phase on the other hand is stable. If there is a boundary between a stable and a metastable phase the generic expectation is that the stable phase

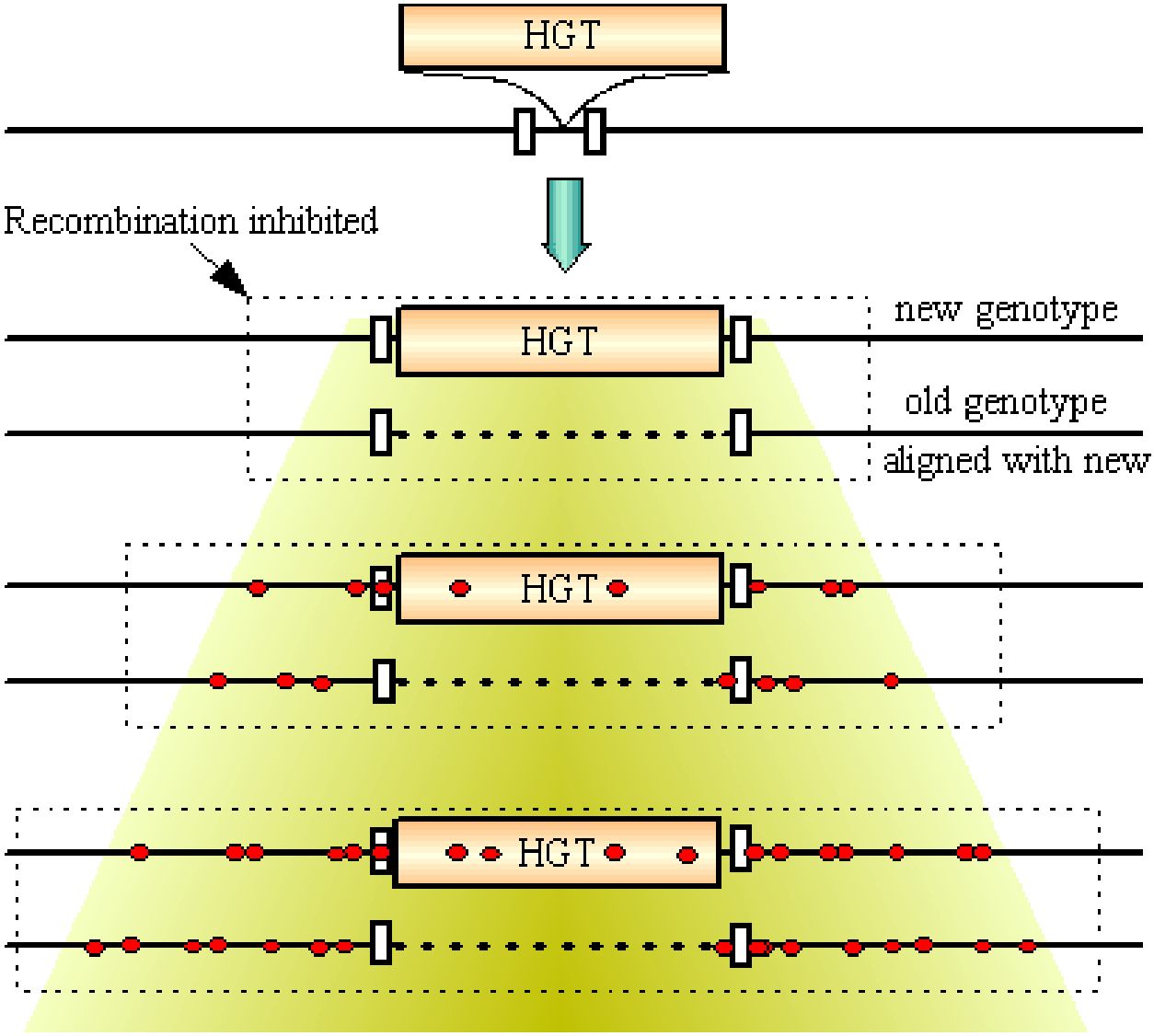

Figure 3.1: Schematic illustrating the process by which a diversification front propagates along a genome in a selection neutral situation. In the vicinity of the HGT island, recombination is suppressed relative to point mutations, allowing point differences to build up in the region flanking the HGT island. The newly accumulated sequence differences lead to the extension of the region where recombination is inhibited and, in turn, an accumulation of point differences further away from the HGT island. The process repeats itself.

will grow at the expense of the metastable one, as shown in Figure 3.1. This will happen because homologous recombination is inhibited not only in the diverged phase but also in a finite region flanking it within the uniform phase. Mutations will accumulate in the flanking region, and as a result the diverged phase will grow. We will refer to the boundary between the uniform and diverged phases as a *diversification front*. Therefore, the system has the potential to sustain the propagation of diversification fronts. Such diversification fronts can be nucleated by processes that create regions of sequence difference between genomes in the population, such as HGT, genome rearrangements, deletions or insertions and have important biological consequences for the evolution and diversification of microbes, as will be discussed later.

3.5 Simulations and results

To clarify this intuition we performed a series of simulations of a population of interacting genomes, starting from two different initial conditions: 1) all sequences are the same, and 2) all sequences are the same except for a strip, long compared with the typical size of recombining fragments, in which the sequences are random. We used three different models for the rules governing the dynamical behavior of homologous recombination: Model I, requiring sequence identity at one end of the recombining fragment; Model II, requiring sequence identity at both ends; and Model III, with no requirement of sequence identity. The central questions addressed are: Under what circumstances is there a well defined front propagation region; is it readily observable or is fine tuning of the parameters required? Do the three models differ qualitatively? To address these questions in a quantitative manner, we define an order parameter

$$\psi(x) = \frac{n}{n-1}\frac{1}{N(N-1)}\sum_{i,j}(1-\delta_{A_{xi},A_{xj}}) \tag{3.1}$$

where A_{xi} denotes the letter at position x of genome i. The order parameter ψ measures the average difference in the population between the sequences at genome position x normalized so that $\psi = 1$ when the genomes are uncorrelated. This corresponds to the *diverged phase* of the system. In the opposite limit, $\psi = 0$, the genomes in the system are highly correlated, giving rise to the *uniform phase* of the system.

For each model, we studied the time evolution of the order parameter for different values of m/r and α. Typical values used for the other parameters are $F = 500$, $M = 10$, $L = 10000$, $N = 20$, $n = 2$. For each separate run we measured ψ as a function of position within the genome and time. By varying α, we control the strength of the mismatch repair mechanism, and hence the success rate of recombination. The most important trend probed by our simulations is the behavior of the order parameter as a function of the ratio $\mu \equiv m/r$, the relative strength of point mutations versus recombination.

3.5.1 Results for Models I and III

For sufficiently low values of α, the equilibrium value of the order parameter varies gradually with $\mu = m/r$, as shown in Figure 3.2. The uniform and random strip initial conditions always relax to the same final state. The random strip simply dissolves and no front propagation is observed. This situation arises when recombination is allowed almost regardless of the degree of sequence divergence.

Above a threshold value of α, the uniform and diverged phases become distinct: for small values of μ, the order parameter is 0, and the system is genetically uniform. However, for large values of μ, the order parameter is close to unity, indicating that the system is genetically diverged. This transition appears to be sharp, as shown in Figure 3.3. There is further interesting dynamical behavior as a function of μ. For $\mu > \mu_u$ the uniform phase becomes unstable and the sequences diverge everywhere simultaneously. For $\mu < \mu_s$, the uniform phase is stable, and a finite region of diverged phase shrinks as a function of time, i.e. the uniform phase invades the diverged one. For $\mu_s < \mu < \mu_u$, diversification proceeds through nucleation and growth of the diverged phase; in this parameter range, front propagation occurs.

From this behavior, we deduce the qualitative phase diagram presented in Figure 3.4a. Model III, with no sequence identity requirement, shows qualitatively similar results (data not shown).

3.5.2 Results for Model II

For Model II, with sequence identity requirement at both ends, we observe front propagation even for $\alpha = 0$. Moreover, the width $w \equiv \mu_u/\mu_s$ of the interval $\mu_s < \mu < \mu_u$, where front propagation

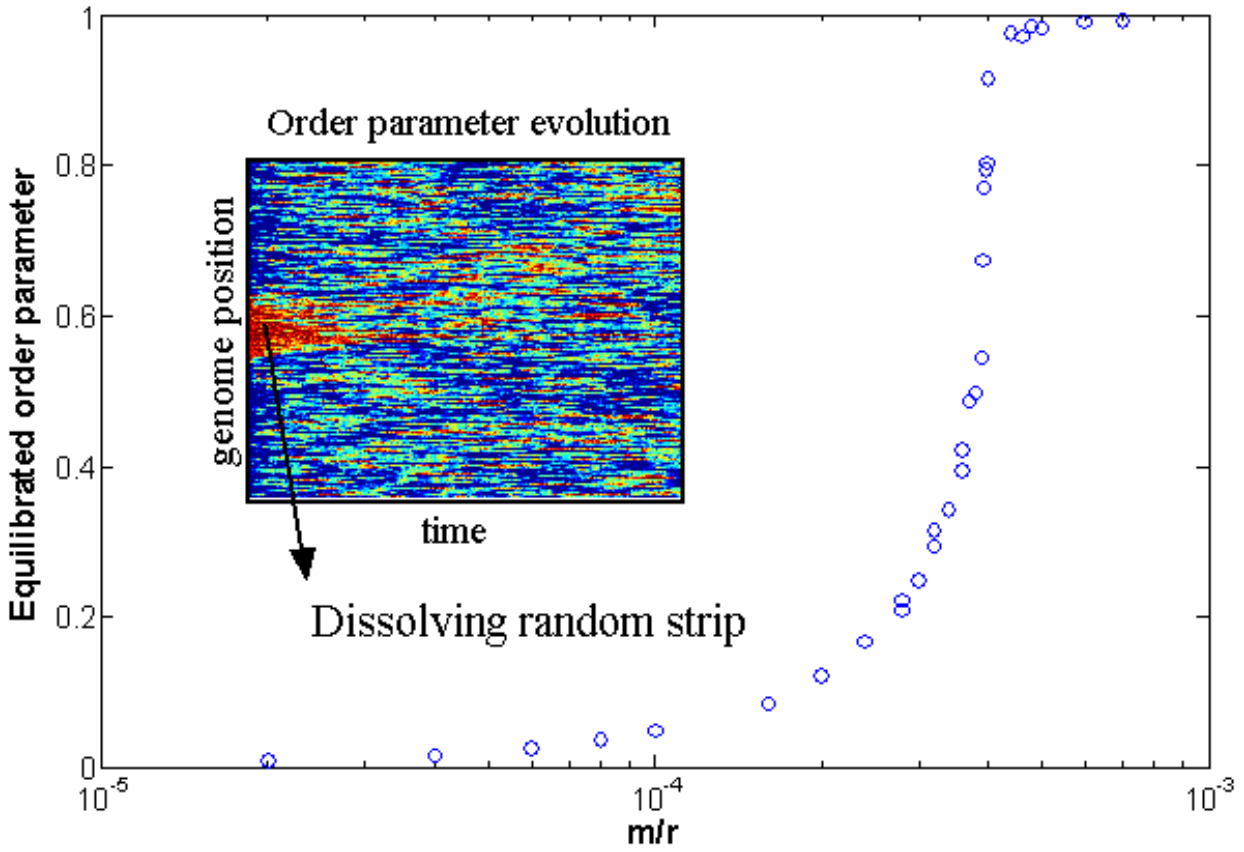

Figure 3.2: The equilibrium value of the order parameter changes gradually with m/r in Model I with $\alpha = 0$, $F = 500$, $M = 10$, $L = 10000$, $N = 20$ and $n = 2$. The inset figure depicts a typical time evolution of the genome population. The vertical axis represents position along the genome, the colorscale indicating the value of the order parameter (blue denoting uniform phase, red denoting diverged phase), while the horizontal axis is simulation time. A random strip dissolves without triggering a diversification front.

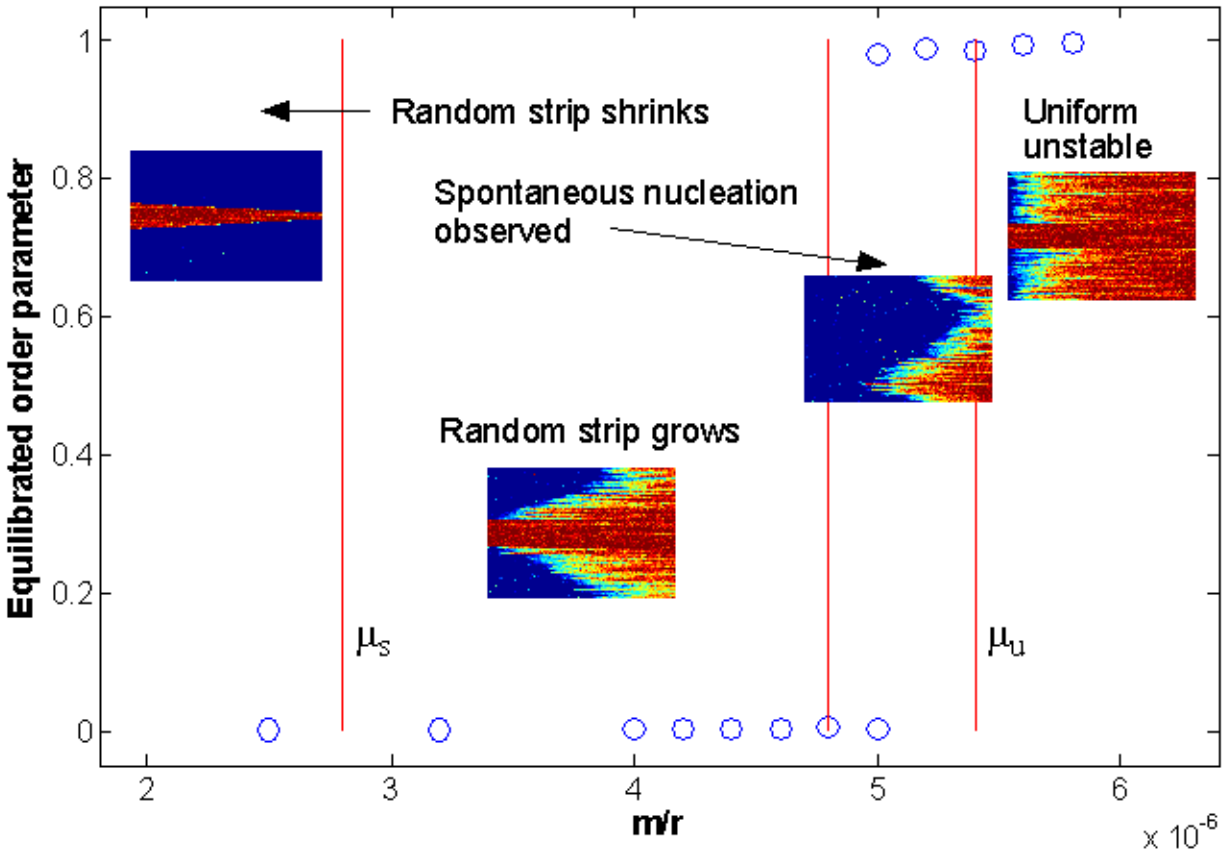

Figure 3.3: Starting from a uniform state, the order parameter equilibrates to values close to 0 or 1 in Model I with $\alpha = 0.4$, $F = 500$, $M = 10$, $L = 10000$, $N = 20$ and $n = 2$, indicating the existence of distinct uniform and diverged phases. The inset figures depict the genome population for the indicated value of m/r, as a function of time. The vertical axis represents position along the genome, the colorscale indicating the value of the order parameter (blue denoting uniform phase, red denoting diverged phase), while the horizontal axis is simulation time. For $\mu_s < \mu < \mu_u$ the random strip triggers a diversification front. For μ close to μ_u spontaneous nucleation is possible.

occurs, is very wide. While for Models I and III we always observed $w \leq 2$, for Model II we could not even observe the point μ_u, and $w > 100$. This results in the phase diagram qualitatively represented on Figure 3.4b. The front speed can be as high as several times the fragment size per average point mutation time near the transition, and is a rapidly decreasing function of the recombination rate.

To summarize, there is a qualitative difference between the situation with no sequence identity requirement (Model III) or sequence identity requirement at only one end (Model I) and Model II with sequence identity requirement at both ends. The difference is manifested in the phase diagram and the width of the front propagation region.

3.6 Microbe classification

These theoretical predictions imply that we can classify microbial genomes according to the details of the recombination dynamics: class I, consisting of models I and III, and class II, consisting of model II. The distinguishing feature of the classes is whether or not the recombination dynamics requires sequence identity at both ends of the incorporated segment. For Class II, as long as the uniformity of a population is maintained by homologous recombination, it will support propagating diversification fronts. For Class I, diversification fronts are possible only within a narrow interval of the ratio of mutation to recombination rates and are therefore unlikely.

The existence of class I and class II indicates that the details of homologous recombination are important beyond the fact that the probability of recombination exponentially decreases with sequence divergence. Therefore it is necessary to elucidate further the differences between homologous recombination mechanisms in different bacteria and work out their consequences for front propagation. For example, if mismatch repair is nick-directed and not methyl-directed [35] then more mismatches will be detected near the ends of the recombining fragments. This, in turn, will make front propagation more robust, because a greater fraction of the average homogenizing capability of recombination will be inhibited by a phase boundary. Also, if non-homologous DNA loops formed during the recombination process are not corrected efficiently, then small deletions, insertions, slippage and inversions would not trigger diversification fronts. Since micro rearrangements are presumably frequent, the efficiency of loop repair will be an important factor in determining

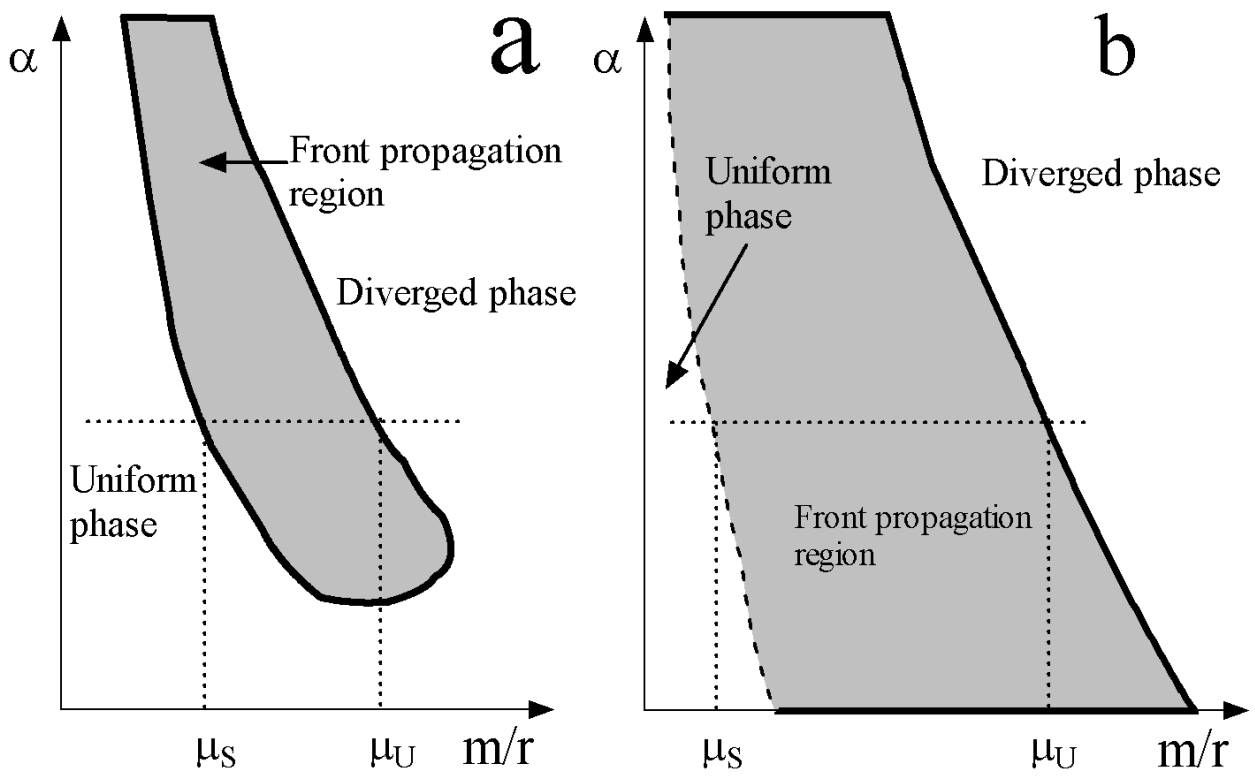

Figure 3.4: **a.** The phase diagram of Models I and III. Distinct phases exist only above a threshold value of α and the width of the front propagation region, μ_u/μ_s, is less than 2. **b.** The phase diagram of Model II. Distinct phases exist for all values of α and the front propagation region is very wide: $\mu_u/\mu_s > 100$.

the rate of nucleation of fronts. Finally, it is important to know whether or not and how the length of the incorporated fragments is dynamically dependant on the differences between the donor and recipient.

In order to seek evidence for the front propagation mechanism, we now compare available completely sequenced genomes of closely-related microbes. The most direct evidence for front propagation from genome data alone would be an extended step-like pattern in the sequence divergence of closely-related well-aligned genomes, with the diverged region centered around a region of HGT, deletion or genome rearrangement. The front profile reflects the different times after genetic isolation of different parts of the chromosome. Under conventional uniform molecular clock assumptions, it will be approximately linear, with a slope determined by the distance the front travels during the time it takes the sequences to fully diverge once recombination is inhibited.

Slowly changing components of the sequence divergence, such as non-synonymous substitutions, leads to more extended profiles.

3.7 Analysis of genome data

We consider the sequenced genomes in the genus *Bacillus*. It is in *Bacillus* that Majewski and Cohan [34] discovered the requirement for sequence identity at both ends, and our simulations indicate that front propagation is more likely to occur in such systems.

We obtained the complete genome sequences from the NCBI database, together with the positions and orientations of the known or predicted protein coding regions, tRNAs and rRNAs. We globally aligned all pairs using the `nucmer` script of the MUMMER package [43] (`nucmer -b 50 -g 300 -c 65 -mum`), obtaining a list of well aligned regions for each pair. Three *Bacillus cereus* strains - ATCC 10987, ATCC 14579 and ZK [44, 45], three *Bacillus anthracis* strains - Ames, Ames Ancestor and Sterne, and *Bacillus thuringiensis serovar konkukian str. 97-27* genomes were close, highly co-linear and analyzed further. The three anthracis strains were practically identical and only Ames was used in the analysis.

For each pair, we mapped the well-aligned regions on one of the genomes, and constructed a series of coarse-grained profiles by sliding a window of width W along the genome while excluding non-aligned regions (resulting from insertions and deletions) from the averaging, as depicted graphically in Figure 3.5. The profiles have gaps where the window covers less than a threshold fraction f of fW unambiguously aligned nucleotides. We used W in the range of 40k to 120k and f between 0.5 and 0.8. We looked at the coarse-grained profiles for the DNA point differences, as well as intergene, intragene, 3rd codon, 1st and 2nd codon, synonymous and non-synonymous (as defined in [46]) differences.

Cereus ATCC 10987 exhibits a distinct step-like pattern of sequence difference when compared to *cereus ZK*, *antracis ames* and *thuringiensis serovar konkukian str. 97-27*. The pattern is also present in each of the other difference components - synonymous, non-synonymous, gene, intergene. What is the explanation for this pattern? Does it involve homologous recombination or not? Is it a result of a front propagation during the separation of *cereus ATCC 10987* with the common ancestor of *cereus ZK*, *antracis ames* and *thuringiensis serovar konkukian str. 97-27*?

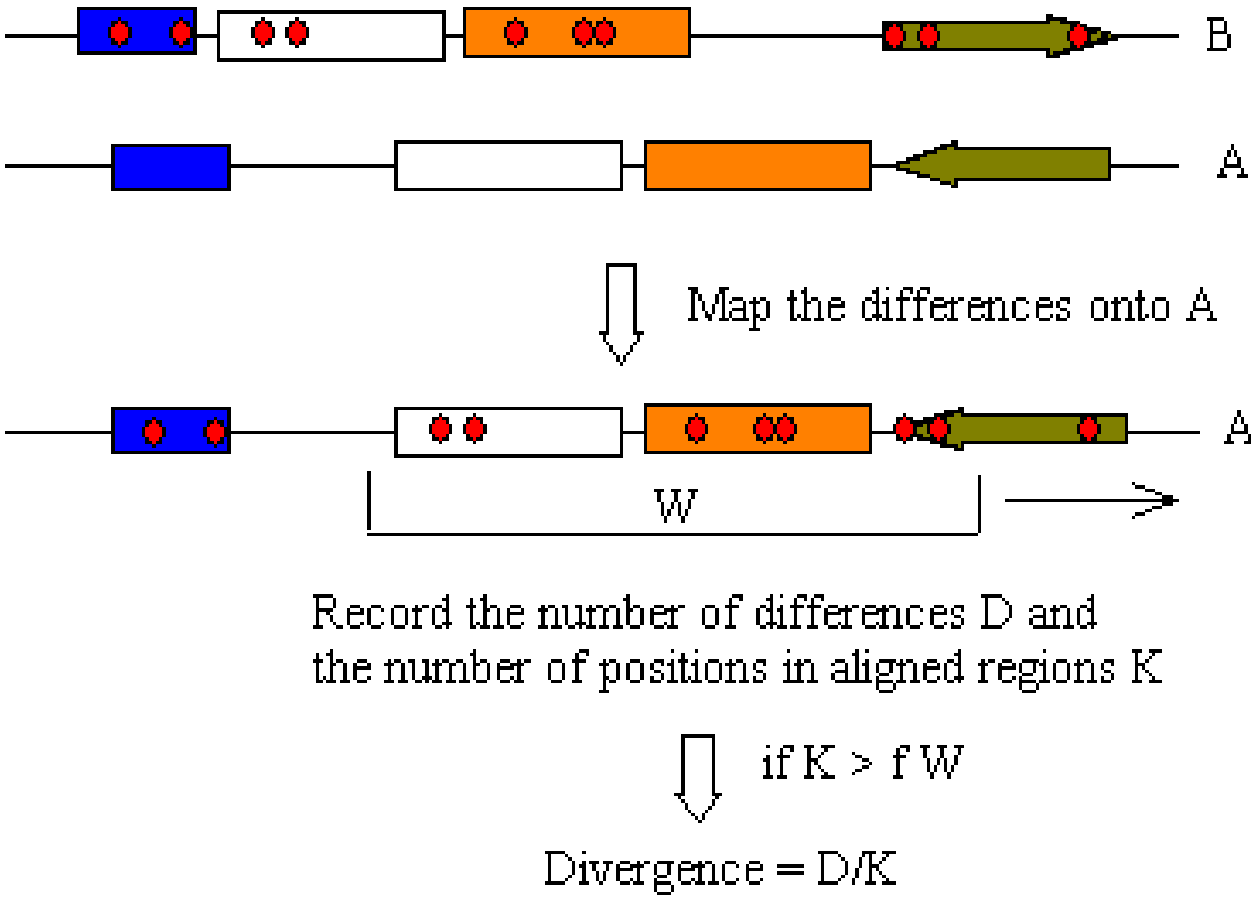

Figure 3.5: To construct the divergence profiles we first identify the well aligned regions (represented by color bars and arrows) using MUMMER, then map the differences (represented by red circles) onto the reference genome and slide a window of width W along the genome.

To answer these questions, we first examined the variation of the nucleotide composition along the genome. Based on the GC and AT skews the replication terminus is located at around 2.6Mb – away from the position of the difference profile step. The GC content varies smoothly along the genome and does not exhibit a step pattern. It has a minimum near the replication terminus.

The step pattern is partially correlated with the density of protein coding regions in the above genomes, the sequence differences being larger where the density is lower. However, since all difference components exhibit the pattern, it cannot be simply an artifact due to different proportions of gene and intergene regions with different mutation rates. Moreover, within the well aligned regions, the intergene regions are, on average, only about 15% more divergent than protein coding regions and the gene density varies only in the 75-90% percent range. Therefore, the small differences in the proportions of sites with different mutation rates would have to have been somehow amplified if varying coding density were the underlying cause of the pattern. The non-aligned regions have a higher intergene fraction than aligned ones suggesting a possible mechanism by which the density of protein coding regions can indirectly affect sequence divergence by a preferential accumulation of inter-strain alignment gaps in intergene regions and a corresponding reduction of recombination rates.

Could it be that not just the proportion of site types, but the point mutation rates themselves vary gradually along the genome, leading to the above pattern? To answer this question, we turn to the distribution of lengths of maximal exact matches (DLMEM) between pairs of aligned sequences. If differences had accumulated by a Poisson mutational process, then we would expect an exponential distribution. Recombination, on the other hand, will lead to a broader distribution and, for example, a deviation from the Poisson statistics value (unity) for the ratio of the standard deviation and the mean [47].

Whether these deviations are statistically significant can be determined by comparing with the distribution of this ratio for the case without recombination.

We gathered DLMEM statistics for different well-aligned regions. The ratio of the standard deviation and mean is significantly above 1, as shown in Figure (3.7a). Moreover, there is a positive correlation between this ratio and the length of the uninterrupted well-aligned regions, a trend which agrees with the notion that non-aligned parts inhibit recombination within the adjacent

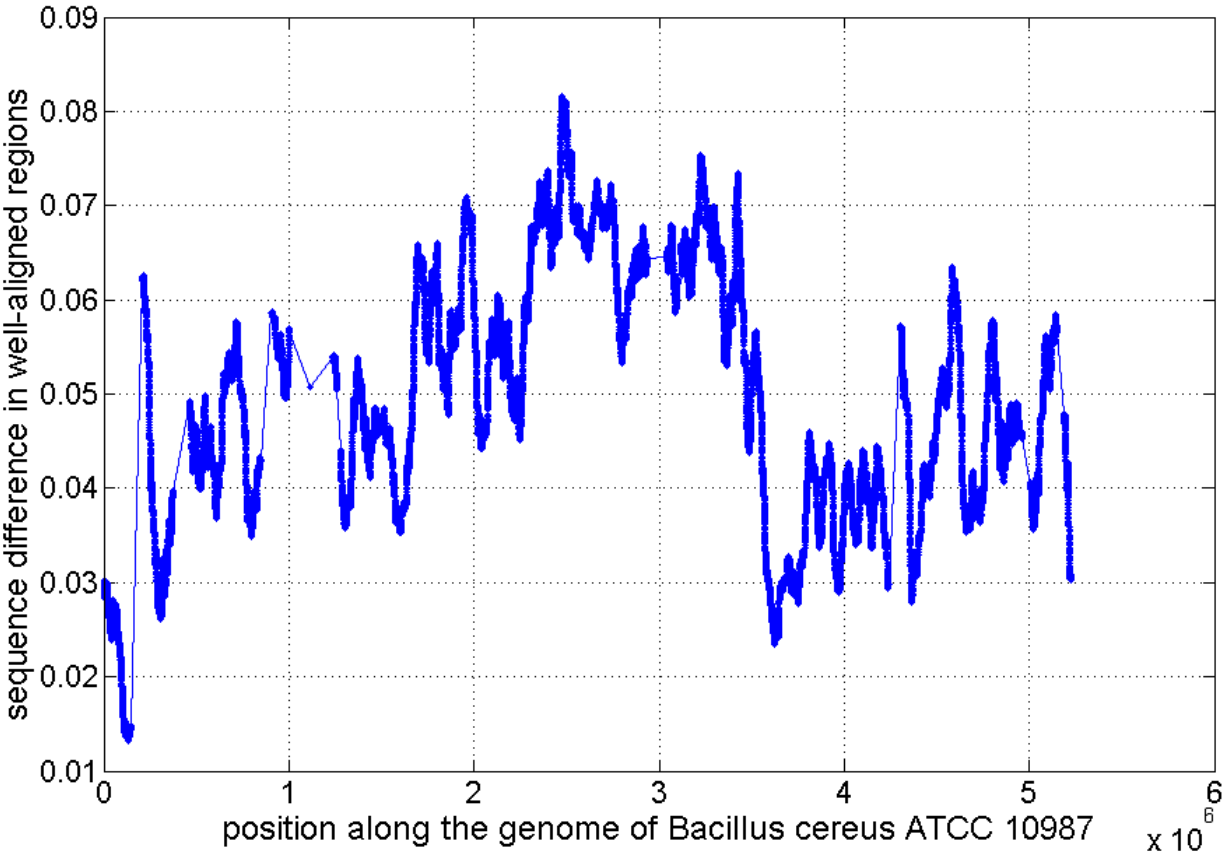

Figure 3.6: The step-like profile of the sequence difference between *Bacillus cereus ATCC 10987* and *Bacillus cereus ZK* obtained by sliding a 60k window with $f = 2/3$ along the genome.

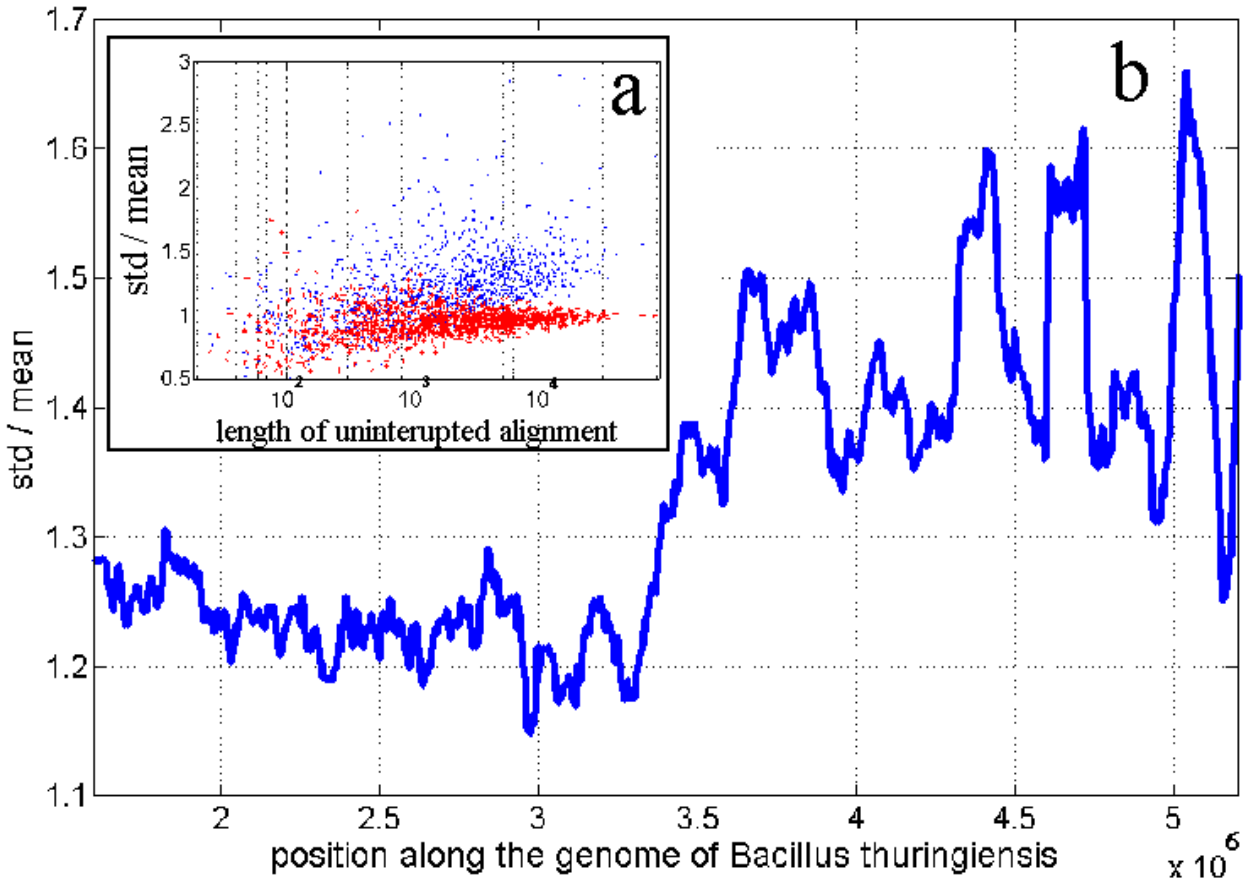

Figure 3.7: DLMEM statistics resulting from the comparison of *Bacillus thuringiensis* and *Bacillus cereus ATCC 10987*. **a.** The std/mean for the distribution of lengths of maximal exact matches within a well-aligned region is positively correlated with the length of the region. The actual data (blue dots) is contrasted with a null hypothesis with matched sequence difference for each region (red *) **b.** The std/mean DLMEM profile obtained using a 120k window with $f = 0.5$ along *Bacillus thuringiensis* exhibits a step-like pattern.

aligned regions.

We then looked for evidence of different rates of homologous recombination along the chromosome by studying the changes in the DLMEM statistics in a sliding window. There is again a step-like pattern for the ratio of the standard deviation and the mean, as shown in Figure (3.7b).

Deviation of the ratio of the standard deviation and the mean of a DLMEM is a sign of clustering of the differences along the chromosome. Are there reasons for clustering which do not involve homologous recombination? If different genes have very different evolution rates, then this can lead to apparent clustering. For example, different gene expression levels can lead to different synonymous mutation rates and an apparent clustering of differences within the weakly expressed

genes. To control for this, we compare the DLMEM for neutral mutations with a null model with matched neutral divergence of each protein coding region separately. The pattern is present in the real data but almost completely disappears in the control. The residue is due to correlations of the divergences of adjacent proteins which are expected in the presence of homologous recombination. Since, presumably, there is no reason apart for recombination for clustering of synonymous substitutions within each gene separately, this test not only rules out genes with different evolutionary rates as an explanation but also gives confidence that the standard deviation over mean deviations from unity are predominantly due to homologous recombination.

Further evidence supporting the homologous recombination interpretation of the ratio of the standard deviation and the mean of DLMEM comes from contrasting the above observations with the results of the comparison between the completely sequenced *Buchnera aphidicola* strains *APS, BP and SG*. Because, these are intracellular parasites lacking the RecA gene we expect no homologous recombination. Indeed, we find that there is no statistically significant deviation from unity of the standard deviation over mean and a highly uniform difference profile.

In summary, the above data indicate that there are large-scale step-like variations of the rates of homologous recombination along the analyzed microbial genomes, apparently consistent with the hypothesis that diversification proceeded by front propagation.

3.8 Discussion

In this section, we discuss the consequences of the front propagation mechanism for the fate of bacteria that have acquired useful skills through HGT or have undergone a large-scale genome rearrangement. We argue that the front propagation mechanism facilitates global genetic isolation between strains, and, as such, is a mechanism for what may be loosely termed "speciation". On the other hand, the front propagation mechanism reduces the chances that chromosomal changes, such as incorporation of HGTs or rearrangements, will be evolutionary successful, thus creating a dynamical barrier to the accumulation of such mutations in evolutionary time.

A bacterium can acquire a new skill by means of HGT. This can lead to the extinction of those bacteria which do not possess the beneficial (under appropriate selection pressure) HGT fragment. Alternatively, HGT can allow the invasion or foundation of a new biochemical niche,

while being disadvantageous in the former one, or lead to specialization within the old niche. (Indeed, ecological distinctiveness without spatial isolation is not unusual for microbes. Even in the simplest of environments - mono culture lab experiments - coexisting strains emerge spontaneously [48]. However, the creation of coexisting genotypes by HGT cannot properly be termed speciation, because the genotypes are not genetically isolated with respect to homologous recombination, except for a small region surrounding the HGT.)

The front propagation mechanism makes local isolation unstable, because the HGT event nucleates a diversification front leading eventually to a global isolation of the carriers of the HGT event from the rest of the population. Therefore, ecological distinctiveness accompanied by local isolation is enough to generate speciation, even when homologous recombination is not reduced by the ecological distinctiveness. Note that this outcome is different from the one proposed by Lawrence [26], who suggested that global isolation is only achieved through the accumulation of hundreds of HGTs. Our work has demonstrated that even a single HGT or genome rearrangement can lead to global sequence divergence.

It is difficult to apply the biological species concept to groups of strains that are isolated at some loci and not at others [49]. Because of diversification front propagation, a community of bacteria in which pairs of bacteria are genetically isolated at some loci, but not others, is unstable and tends to partition itself into groups which are globally isolated from each other with respect to homologous recombination. This is because genetically isolated regions will suppress recombination and trigger fronts into neighboring non-isolated regions. This instability will be even stronger if the different genomes are not colinear or do not have the same set of genes. Therefore, well defined genetic isolation boundaries emerge spontaneously through the front propagation mechanism even if there is no functional barrier to gene transfer.

What happens when a HGT or a rearrangement brings some advantage, but without enabling the recipient to adopt an entirely distinct ecological role? Achieving complete ecological distinctiveness might be a gradual process. In this case the new genotype will be successful initially but not necessarily in the long run because it will be competing with other beneficial mutations at other loci that emerge throughout the population. Beneficial mutations trigger selective sweeps that can be either global, purging the diversity throughout some ecological niche or, because of homologous

recombination, local, purging the diversity only around the locus of the beneficial mutation. In a population in which relative sequence uniformity is maintained by homologous recombination, local selective sweeps will be the norm. However, front propagation, nucleated in the carriers of a HGT or a rearrangement will propagate by accumulation of neutral mutations, and potentially lead to global genetic isolation of the carriers long before they have a chance to achieve a full ecological distinctiveness.

New strains are easily formed by readily absorbing foreign genetic material, rearranging the genomes, etc. But they are typically short-lived entities, because following front propagation they are excluded from the communal evolution. Front propagation implies that the evolutionary rate of HGT accumulation is less than the rate suggested by looking at strains. This can be, in principle, tested against the data. This mechanism can also explain why gene order is highly conserved in some bacterial groups: there exists a dynamical barrier to the survival of rearranged genomes.

These considerations also have implications for the applicability of molecular phylogenetics, and the ongoing debate about the nature of the impact of HGT on the tree of life. Front propagation limits the impact of HGT, reinforcing in a complementary way Woese's concept of a complexity barrier to HGT [9]. Our argument is complementary, because it does not rely on the nature of the interactions between the genes: there is a barrier to HGT arising from the population dynamics alone.

Our work leaves open a number of interesting issues related to the effect of highly conserved regions on front propagation. A large immutable region can present an impassable obstacle to front propagation. Candidates for such obstacles are rRNA operons, tRNA genes and overlapping genes. Such regions lack the flexibility arising from the degeneracy of the genetic code. HGTs islands inserted near front obstacles will lead to the diversification of a smaller fraction of the recipient genome, and have a greater chance to avoid extinction. Is there a correlation between evolutionary persistent HGTs and RNA gene positions? If a genome region is already diversified there is no penalty for the incorporation of another useful HGT island. Is there clustering of HGT islands? How is front propagation modified for clonal bacteria [41]? Finally, is front propagation beneficial? If front propagation obstacles are allowed to evolve or at least reposition themselves, what configuration of obstacles would result?

3.9 Conclusions

On the basis of computer simulations, we have suggested that the interplay between homologous recombination and point mutations can lead to propagating fronts, in whose wake a population of microbes becomes genetically diverse in evolutionary short time. Thus, even in the absence of selection pressure and ecological barriers to genetic exchange, gene-exchange boundaries can emerge as a statistical consequence of the detailed dynamics of recombination. We have presented a preliminary analysis of available genome data for the *Bacillus cereus* group, which is consistent with the presence of front propagation. These findings prompt speculations about the implications for the evolution and the classification of microbes.

Our model can be extended in a number of directions, including explicit accounting for the role of space, the existence of a non-trivial network of gene exchange connectivity and the effects of sharing of beneficial mutations.

A promising approach to looking for diversification fronts is metagenomics data. Such data can give us a consensus genome for an ensemble of closely related organisms, inhabiting the same environment, and an estimate for the sequence diversity along the consensus genome [6]. This diversity can be directly related to the order parameter $\psi(x)$. A step like variation in $\psi(x)$ might be an indication of a diversification front.

Chapter 4

Spontaneous emergence of genome biases due to selection on the speed, accuracy and energy efficiency of template-directed synthesis

4.1 Introduction

4.1.1 Motivation and history

Statistical studies of genome composition have been one of the driving forces behind evolutionary biology. They have been at the heart of the debate between selectionist and neutralist points of view, i.e. whether evolution is dominated by functional or neutral changes, and helped establish the importance of horizontal gene transfer. The origin of genome biases is still a topic of controversy and intense research, fueled by the vast amount of sequence data which is increasingly being put in the context of high resolution measurements of cellular properties. Better understanding of genome biases would suggest new ways to extract evolutionary information from sequence data, and, as we will see in the next chapter, would even allow us to answer conceptually difficult questions about early evolution. In this chapter, I suggest that different types of genome biases are different manifestations of the same universal mechanism. Therefore, "genome bias" is not just a common name for different statistically significant features of genome composition, but is a meaningful evolutionary category.

Here is a brief summary of the study of genome biases. In 1951, the observation by Chargaff [50]

that the amount of adenine equaled the amount of thymine, while the amount of guanine equaled the amount of cytosine helped bring about the discovery of the DNA structure. By the early sixties, small within species and large between species variations of GC content (especially in microbes) was established. Sueoka [51] proposed that this variation results from species-specific mutational biases. To this day, mutational biases remain one of the primary candidates for explaining genome biases [52, 53]. The deciphering of the genetic code exposed its redundant structure and lead to the neutral theory [54]. A dichotomy, bitter at times [55], developed between neutralist and selectionist points of view, and studies of codon usage bias became the battlefield after it was discovered that, just like the GC content, it is highly variable and species-specific [56, 57]. The neutralist perspective is that synonymous substitutions have no fitness consequences and the preference of one synonymous codon over another reflects mutational biases. Selectionists examine the effects of synonymous substitutions on the translational speed, accuracy, energy efficiency, mRNA and DNA stability, etc. [58, 59, 60, 61, 62]. A strong positive correlation between codon usage and cognate tRNA abundance [63, 64, 65, 66] supported the role of selection and was initially interpreted as a stabilizing selection on codon usage to match the relative tRNA abundance [67, 54]. Reversing the direction of causality, Kurland and Ehrenberg [68] argued that if the translational system is optimized to provide maximal growth rate, the tRNA levels should scale as the square root of the cognate codon usage, in agreement with later experimental data [69, 66]. It was also suggested that codon usage and tRNA expression levels coevolve with each other [4] leading to the same square root scaling. The case for selection on translational speed is strengthened by the fact that more expressed proteins have stronger bias [63, 70]. Selection on translational accuracy was supported by studies in *Drosophila* that showed stronger bias at evolutionary conserved sites [71].

In the late 1980's the selection-mutation-drift framework emerged [5, 64]. Independent of the precise nature of selection, this framework emphasizes that there is a unique optimal (or major) codon for each amino acid and the occurrence of other codons is due to the combined effects of mutation and genetic drift. The observable outcomes depend on the fitness difference between major and minor codons, mutation rates (including mutational biases) and the effective population size. This approach proved extremely useful in analyzing data and dominates current thinking about codon usage [72, 60, 58, 73]. With the framework seemingly clarified, the research efforts

has been focused on clever statistical studies and experiments to delineate the selection pressures operating at different levels and the relative importance of selection versus mutational bias. In this chapter, we put the selection-mutation-drift framework in the *context* of universal properties of template-directed synthesis.

While the origin of biases remains controversial, their very presence is a window to microbial evolution. In the late nineties after the first genome sequences appeared it was noticed that there are distinct genome islands with atypical codon usage and GC content. In view of the already established species-specific nature of the biases, these islands were interpreted as evidence of HGT between different species [74, 28]. This had a profound effect on our understanding of modern day microbial evolution and started to revolutionize the way we think about early life [9].

4.1.2 Overview

The purpose of this chapter is to demonstrate that selection on the speed, accuracy and energy efficiency of template-directed synthesis processes such as translation, transcription and replication can lead to the *spontaneous* emergence of genome biases. Selection on translation leads to codon usage bias; selection on transcription or replication leads to nucleotide composition biases such as the GC content and GC skew, and the different biases influence one another. These biases result from the generic tradeoffs inherent to template-directed synthesis and occur even in the absence of biased mutation or direct selection on the nucleotide composition coming from, say, DNA or mRNA stability. In other words, they are a manifestation of a spontaneous symmetry breaking that cannot be understood within a simple cause and effect framework. In the case of translation, it is the bidirectional interaction between codon usage and tRNA expression levels that creates a fitness landscape that enforces quasi-stable patterns of codon usage. Occasional transitions between patterns are expected, due to genetic drift or hitchhiking of slightly deleterious adjustments of the translational system on other beneficial traits.

My interest in the problem of codon usage was an offshoot of my efforts, presented in the next chapter, to understand the mechanisms by which the genetic code can change and hence evolve towards optimality. It seemed to me that the coevolution between tRNA expression levels and codon usage leads spontaneously to highly uneven codon usage, which then catalyzes the changes

to the code. First came the results described in the next chapter and from there, upon working my way backwards towards clarification and simplification emerged this one. As it turned out during the writing, the idea of codon usage tRNA expression coevolution resulting from selection on translational speed was already proposed by Bulmer in 1987 [4]. The new results and *future directions* that distinguish the framework presented in this thesis from earlier work are:

- The same coevolutionary framework is applicable to all template directed synthesis processes. In particular, GC-AT bias can result from selection on speed of replication or transcription.

- Because of its generic nature the spontaneous emergence of genome biases is potentially important not only for explaining modern day variations of genome biases, but also for understanding the evolutionary constraints during the early evolution of life. In fact, selection on the speed and energy efficiency of the basic information processes was perhaps far more important for primitive organisms; selection on accuracy was harsher before the discovery of proofreading. In the next chapter I look at the implications of the spontaneous codon and tRNA biases for the evolution of the code.

- The relevant parameter that controls the strength of the codon/nucleotide bias due to selection on the translational/replicational speed, in a population with a large effective size, is the *number of mutations per genome per generation*. The codon bias exhibits a continuous phase transition as a function of the above parameter.

- GC and AT skews can result from selection on speed during replication. The two DNA strands replicate by different mechanisms, and correspondingly have different speeds for the same nucleotide composition. The overall replication speed is limited by the slower strand, and therefore an evolutionary optimum is achieved when the compositional asymmetry of the strands and the investment in free nucleotides is such that the two strands have the same replication speed. There is a spontaneous symmetry breaking leading to uneven production of G and C nucleotides and bias of the slow strand towards more optimal nucleotides. The fast strand is preferentially using the less optimized nucleotides. In case there is an intrinsic mutational bias, this mechanism would amplify it.

- Selection on translational *accuracy* can lead to spontaneous symmetry breaking within the

same framework. In particular, spontaneous codon disappearance is expected at sites at which amino acid substitutions are lethal. This has implications for the evolution of the genetic code.

- Selection on translational speed would lead to a stronger synonymous bias for weakly used amino acids (as compared to other amino acids with the same degeneracy), provided that different synonymous codons are encoded by different tRNAs. More generally, if we know the tRNA species present and their affinities we can predict the relationship between the different synonymous biases.

- Non-uniform gene expression leads to stronger symmetry breaking when compared with a uniform one. This suggests that, everything else being equal, organisms with more uneven distributions of gene expression levels have higher bias.

- Intermediate, non-beneficial adjustments of the translational system (translational noise) can induce transitions from one stable state to another. This has implications for the evolution of the code and for understanding the phylogenetic distribution of codon usage and GC-content patterns.

- For the translational system, the coevolution is, perhaps, not just between tRNA expression levels and codon usage, but between the *tRNA pool* as a whole and codon usage. This includes "speciations of tRNAs", i.e. one can allow the number of tRNA species to vary. The coevolutionary models of this and the next chapters should, in future work, be generalized to account for this. For a minimal model, each tRNA species can be characterized by an anticodon (which can bind to several codons according to fixed approximately known rules) and an expression level. This will potentially bring insights not only to the genetic code evolution but also to the modern day evolution and phylogenetic distribution of tRNA pools. Perhaps, a classification of the different stable sets of tRNAs is possible.

- Different stable patterns of codon usage, that result from the coevolution, have different amino acid usage in general. This is expected, since at many genome sites several amino acids are functionally acceptable. The juxtaposition of the codon usage and amino acid usage patterns can yield new amino acid similarity measures.

I will start by clarifying the notion of template-directed synthesis. Then I will give a picture of how the coevolution between the letter usage of the template and the distribution of resources for production of monomers or adaptors leads to spontaneous emergence of genome biases. A coevolutionary modeling framework will be set, and useful relations - independent of the nature of the selection pressures - derived. Then, I will discuss selection on speed, and arrive at a model that is explored both numerically and analytically. The results support the points listed above. I will show through simulations that selection on accuracy can also lead to genome biases. Finally, I simulate transitions between different stable codon usage patterns in the presence of "translational fitness noise".

4.2 Tradeoffs of template-directed synthesis lead to genome biases

In this section I introduce the logic behind the spontaneous emergence of genom biases due to selection on the efficiency of template directed synthesis.

4.2.1 Template-directed synthesis

A template-directed synthesis is a process in which a sequence of letters in the template guides the synthesis of a product according to some rule. The product is also a sequence of letters, and the alphabets of the template and the product are different, in general. The correspondence between the two alphabets is made by adaptor molecules. A pool of free adaptor molecules is maintained. A synthesis machinery (synthetase) moves along the template, at each step waiting for an adaptor to diffuse to its active site. It then discriminates between *correct* and *incorrect* ones with respect to the encoding rule.

On the most fundamental level, template-directed synthesis proceeds through the competition of different adaptors for the next available slot on the template. The competition aspect is elaborated upon in the next chapter where I propose that the collection of different tRNA species within a cell should be viewed as an ecosystem.

Examples of template-directed processes are translation, transcription and replication. In the

case of contemporary translation, the template alphabet consists of 61 letters - the different codons coding for amino acids, and the product alphabet consists of 20 letters - the different amino acids. The correspondence rule is the genetic code. The adaptor molecules are the tRNAs charged with the corresponding amino acids (ternary complexes). The number of different tRNA species is, in general, different from the number of amino acids. In the case of transcription, the template alphabet consists of the four standard deoxyribonucleotides (A, G, T and C), and the product alphabet consists of the four standard ribonucleotides (A, G, T and C). The correspondence rule is - Watson-Crick complementarity. Since the correspondence is simple, the adaptors are also simple - the adaptors are the ribonucleotides themselves. The case of replication is similar to that of transcription, but the template and the product alphabets are identical - the four deoxyribonucleotides. The replication adaptors are the deoxyribonucleotides.

4.2.2 Intrinsic tradeoffs of template-directed synthesis

The time a synthetase waits for an adaptor to bind depends on the concentration of the adaptors. The higher the adaptor concentration is, the less is the waiting time. In addition, given the fact that the synthetase discriminates between correct and incorrect adaptors imperfectly, the accuracy of synthesis depends on the relative concentrations of the different adaptors. Therefore, for a given template letter, both the speed and the accuracy of synthesis increase with the increase of the concentration of its cognate adaptor(s). On the other hand, the production of adaptors uses up material resources and energy. As we will see, these generic opposing tendencies lead to an active feedback loop between resource allocation for adaptor production and letter usage in the template that, in certain regimes, drives highly uneven letter usage and adaptor concentrations.

There are many other resource allocation decisions relevant to the synthesis: investment in proofreading, production level of synthetase machineries, etc. But for a given energy investment - higher speed and accuracy is favored in general. Even if higher accuracy per se is not beneficial, higher "intrinsic" accuracy allows to decrease the investment in proofreading, and is therefore favored.

4.2.3 Picture of the coevolution

To illustrate the logic of the feedback consider two synonymous codons that have different cognate tRNAs. If one of the codons is favored at some point because it is more optimized or because its cognate tRNA level is higher (or both) then there will be an advantage in increasing its expression level at the expense of the other tRNA. As a result the favored codon becomes even more favored. The cycle continues until the tendency of the disfavored codon to disappear is balanced by mutational pressure. The system is bistable - there is selection on the bias but not on its direction. While the system is bistable, in the presence of translational fitness noise coming from all other fitness components or a codon usage perturbation as a result of HGT, a lineage can occasionally switch form one stable pattern of codon usage to another.

For the case of replication or transcription, the cost of tRNA (ternary complex) production above is replaced by the cost of nucleotide synthesis (or extraction from the environment) which includes the expression cost of the entire molecular machinery associated with it. The relative usage of the synonymous codons can be replaced by, say, the relative usage of A and G or T and C at the third codon positions or in non-coding regions.

The key to the tradeoff, is that while every letter in a template puts a tiny selection pressure to increase the production of its cognate adaptor in order to improve the speed and accuracy of the synthesis at its position, there is a cost of production of those adaptors. If, due to a genome bias, these tiny selection pressures are not perfectly balanced, the corresponding beneficial adjustment in the monomer production pattern will only increase the bias. Therefore, a bias is self reinforcing. It is limited by the mutational pressure and the functional constraints; some redundancy is essential.

4.2.4 The role of redundancy

The emergence of genome biases through the above mechanism requires some functional redundancy. One source of redundancy is the redundancy of the genetic code. This redundancy allows spontaneous emergence of genome biases due to selection not only on translation but also transcription and replication. Non-coding DNA is another form of redundancy that influences the spontaneous emergence of biases due to selection on replication. Introns generate redundancy that facilitates the emergence of biases due to selection on transcription (and replication).

4.2.5 The role of genetic regulation of resource distribution

The above coevolutionary picture also required that there are genetic mechanisms that allow the regulation of the distribution of resources for adaptor production. tRNA expression and maturation, all the way to the formation of the ternary complexes, is a complex multistep process in contemporary organisms allowing ample opportunities for regulation [75]. Therefore, we have every reason to believe that the effective concentrations of different adaptors ready for translation are easy to change and fine tune to the needs of an organism. There is also evidence that the tRNA abundance is fine-tuned to maximize the translational speed (at given resource investment) [69, 66].

4.2.6 Which comes first: codon usage or tRNA abundance?

As many people have argued, codon usage adjusts to existing tRNA expression levels. But once the codon usage equilibrates, is there an incentive for the tRNA expression levels to change? The answer seems to be no, since any change will be a change away from the perfectly coadapted state. Here we argue, following [4], that, on the contrary, if we impose an arbitrary tRNA expression pattern, and let the codon usage adapt to it, there will still exist favorable adjustments of the tRNA expression pattern. A change in the tRNA expression pattern will then trigger an adjustment in the codon usage pattern, and the process will continue until one of many stable states is reached. These states are characterized by codon usage bias, uneven tRNA expression levels and a strong correlation between them. Which of the locally stable states a lineage is stuck in depends on history and chance.

It has also been suggested that codon usage is determined by biased mutation or direct selection on the nucleotide composition at the DNA or mRNA level, and then the tRNA levels adjust to it. What we argue here is that, even in the absence of such *exogenous pressures*, the evolutionary stable states have an uneven codon usage which is a manifestation of a symmetry breaking instability inherent to the fundamental aspects of template-directed synthesis. If there are mutational biases, the coevolutionary instability would *amplify* those biases. Moreover, at a given exogenous pressure we will have, in general, more than one stable state.

4.3 Modeling framework

Here I describe the coevolutionary framework that will later be augmented by specific assumptions about the selection pressures acting on translation, replication and transcription. The goal is to create a model, or models, that will help us quantify and explore the consequences of the coevolution outlined above. Here and throughout I will predominantly use the language of translation but the considerations will apply to the other types of template-directed synthesis.

4.3.1 Mutation selection equilibrium

While the selection on the speed and accuracy of translation might be an overall significant evolutionary force, the selection on a single genome position is most likely weak enough to allow frequent occurrence of non-optimal codons due to mutations.

Different codons at a given genome position would have different fitness effects, in general. Therefore, a genome position is characterized by a vector specifying the fitness effects of all possible codons. Genome positions that have the same vectors are said to belong to the same *site type*. For example, all genome positions that are *required* to code for a particular amino acid have the same site type, and all completely neutral positions belong to a different site type. We can characterize selection through the matrix F_{si} specifying the fitness of codon i at a *site of type* s.

Mutational pressure can be characterized by the matrix M_{ij} specifying the probability that codon i will mutate into codon j in one generation. It is assumed independent of the site type and genome position. Any mutational biases can easily be incorporated in M_{ij}. I will focus on equally probable single nucleotide changes. In this case, M is specified by a single parameter μ which is the probability for a change at a given site in one generation.

Let u_{si} be the frequency of codon i among sites of type s . Let x enumerate the genome positions, $s(x)$ be the site type at position x, and $i(x)$ be the codon (template letter) at position x. Following Sella and Ardell [76], if the fitness contributions of the different sites are independent, i.e.

$$\text{fitness} \propto \prod_x F_{s(x),i(x)} \ , \tag{4.1}$$

then the codon usage at a site of type s at a mutation selection equilibrium is given by the eigen-

vector corresponding to the largest eigenvalue of the matrix

$$Q_{ij}^{(s)} = \sum_k M_{ik} F_{sk} \delta_{kj} \,. \tag{4.2}$$

The matrix Q reflects the application of selection followed by mutation. Note that the genome structure $\{s(x)\}$ is assumed fixed. This is because we are interested here only in the evolution of the codon choice $\{i(x)\}$.

In particular, if at a given site type only two codons are viable (or strongly favored over the rest), and have an overall (translational+functional) fitness f_1 and f_2 correspondingly, the equilibrium codon usage (u_1, u_2) is the eigenvector corresponding to the largest eigenvalue of the matrix

$$\begin{pmatrix} 1-\mu & \mu \\ \mu & 1-\mu \end{pmatrix} \begin{pmatrix} f_1 & 0 \\ 0 & f_2 \end{pmatrix} . \tag{4.3}$$

For $f_1 > f_2$ and $\mu^2 \ll \big((f_1 - f_2)/f_1\big)^2$ the solution simplifies to

$$\frac{u_2}{u_1} = \frac{f_1}{f_1 - f_2}(\mu + \mu^2) + O(\mu^3) \,. \tag{4.4}$$

Alternatively, since $f_1 > 0$ and $f_2 > 0$ we can write

$$\frac{f_2}{f_1} = \exp\left(-\frac{\tilde{f}}{L}\right) , \tag{4.5}$$

for some $\tilde{f}$. Then in the limit $\mu \to 0$, $L \to \infty$, we have

$$\frac{u_2}{u_1} = \sqrt{1 + \left(\frac{f}{2\mu L}\right)^2} - \frac{f}{2\mu L} \,. \tag{4.6}$$

for any finite $\tilde{f} \geq 0$. Solving for μL this can be rewritten as

$$\mu L = \tilde{f} \frac{u_2/u_1}{1 - u_2/u_1} \,. \tag{4.7}$$

This treatment is valid in the infinite population limit. In the opposite limit, where the popula-

tion effective population size, N_e, is so small that there are almost no polymorphisms, i.e. $\mu N_e \ll 1$, we end up with [5]:

$$\frac{u_2}{u_1} = e^{-2N_e(1-f_2/f_1)} \,. \tag{4.8}$$

4.3.2 Invasion-equilibration cycle

We will be concerned with the evolutionary fate of changes in the patterns of tRNA expression levels (or nucleotide production). This fate depends on the existing codon usage (or nucleotide composition). For example, a sharp decrease in the expression level of a tRNAs cognate to a popular codons is not favored. If, however, a change is beneficial, it can invade the population. It is biologically reasonable and also mathematically convenient to assume that the time it takes a beneficial change to take over the population is short compared to the time it takes the codon usage to equilibrate to the new expression levels.

Based on the above, the simulations proceed as follows: we equilibrate the codon usage given the existing tRNA expression levels, then change the existing expression levels and let them invade the population if beneficial, under the current codon usage. Now we complete the cycle by equilibrating the codon usage.

More generally, the probability that an elementary change is accepted is a function of the ratio of the fitness values, calculated using the current codon usage. A step function at unity corresponds to the case without noise and infinite population size.

This procedure assumes that changes in the expression levels are rare compared with the mean codon equilibration time. An alternative procedure was also tried in which the expression levels were adjusted many times at fixed codon usage, and only after no further changes were likely, the codon usage was equilibrated. This suggests that the equilibrium properties of the coevolutionary dynamics are not sensitive to the relative ratio of codon usage and expression level changes.

4.3.3 Selection pressures

To complete the framework, we need to model the selection pressures. Let t enumerate the different adaptor species, and c_t be their concentrations. The generic form of the fitness function that we

would use is

$$f = G\left(\sum_t c_t\right) \cdot f_{speed}(\{c\},\{u\}) \cdot \prod_x \tilde{F}_{s(x),i(x)}(\{c\}) \,, \tag{4.9}$$

where $G()$ is a decreasing function of the total adaptor expression level, accounting for the fitness cost of adaptor production, and $f_{speed}()$ is a function to be specified later. The last factor accounts for the fact that the fitness depends on the sequence of the product of the template-directed synthesis. For example, by choosing a particular $\tilde{F}_{si}$ to be zero, we can enforce that codon i cannot be used at a site of type s. $\tilde{F}()$ depends on the adaptor concentrations, if we account for the occasional translations via incorrect adaptors by the synthetase machinery. This effects will be considered in section 4.5. If $f_{speed}()$ is non-trivial, the above fitness form is not manifestly a product of independent sites as required by equation 4.2. This will be also handled below.

Selection on either speed or accuracy can drive highly biased codon usage. To show this, we ignore the effects of mistranslation in the translational speed section (all $\tilde{F}_{si}$ values are either zero or one), and ignore the translational speed selection in the mistranslation section ($f_{speed} = 1$).

4.4 Selection on speed

A large amount of evidence has accumulated that selection on translational speed is important [58] and several models have been proposed [68, 4, 5]. Before we go into details, let's pause and ponder why faster is better. Perhaps, because we are often in a hurry we consider it obvious. The more in a hurry we are the narrower is the scope of the really important things that deserve consideration (it follows from the theory of relativity of meaning), and throw in some competition in doing those few things and you end up in the hurry trap. To survive in the hurry trap you have to be biased. But why are we in a hurry to start with? There is an enormous range of organism generation times - from 20 min to hundreds of years. Is there a scale and what sets it? One possible reason to hurry is competition for resources. Certain microbes are subjected to a very irregular supply of resources. When food is present, it may be present in surplus quantities, and then there are prolonged periods of starvation. In such an environment there are two basic strategies - to be fast so that you prepare more copies of yourself by the time the feast is over and to be good at surviving between feasts. For the first strategy faster is better in a scale-free way. In addition, energy efficiency does not matter,

at least in the limit where you and your backup copies end up eating only a small fraction of the pie. For the second strategy, the scales are set by the characteristics of the environment. Once you meet up the standards of the environment what is left is to try to do it in the most efficient way possible. Faster translation increases efficiency [68], for example by reducing the minimum number of ribosomes that you need to meet the minimum speed standards imposed by the environment (this includes not just external scales but a multitude of biochemical decay constants; for example mRNA degradation rates impose requirements on translation). Of course, this only applies to mechanisms of speed increase that are energetically passive; codon biases are in this category. If, the survival probability is low due to lack of supplies, the effective selective pressure on the energy efficiency of the minimum metabolism is also scale-free, say in the number of ribosomes that you need.

Selection on translational speed has been emphasized in view of the correlations between expression levels of proteins and codon bias. What about DNA replication and transcription? Bacterial duplication times in the exponential phase are up to several times shorter than the genome replication times. This is at least consistent with the hypothesis that the replication speed is a limiting factor. The replication fork speed in bacteria is on the order of 20 times faster than that for eukaryotes indicating different selection pressures. While translation is a large part of the metabolism it is also a highly parallel process. At least in bacteria this is not so for replication (two replication forks plus the occasions when the next genome duplication starts before the end of the first.) The fact that translation is a larger part of the metabolism argues, above all, for an effect of translational speed on energy efficiency. Transcription is also less parallel than translation since there is one genome but many mRNA copies of a given gene. So, the speed of translation might be limited by the rate of transcription, similarly to the way it is limited by ribosome production. Transcription in bacteria is about 20 times slower than replication. Collisions between the DNA and RNA polymerases are costly enough to bias the direction of the majority of genes on the chromosome. Does this put a pressure on the transcription rates? Speed, accuracy, efficiency are budget constrained [59]. It seems plausible that the optimal distribution of resource to independent systems acts to equalize the selection pressures on them, so that many different processes are simultaneously limiting.

Now, we will show how coevolution works through specific models. We start by illustrating that coevolution leads to an extreme symmetry breaking in the simplest possible context and then solve exactly a biologically motivated model that can be applied to translation, replication and transcription. We emphasize that while the overall strength of the selection on speed is variable, the scaling of the selection pressures on the individual genome positions with the genome size, L is not, and results in the dependance of the strength of the genome biases on μL - the number of mutations per genome per generation.

4.4.1 Simple model leading to extreme codon bias

Let t enumerate the different tRNA species and c_t - their concentration. Assume that each codon is recognized by only one tRNA species. Let $t(x)$ be the tRNA cognate to the codon at genome position x. If we ignore mistranslation, the translational fitness f_x of a given codon at a given site type depends only on the expression level of its cognate tRNA. Assuming that codons have independent effects on translational fitness, and that the fitness cost of expression is exponential, we have

$$f = \exp(-\sum_t c_t) \prod_x f_x(c_{t(x)}), \tag{4.10}$$

where the product is over all genome positions, which in general have different fitness dependence, $f_x()$, on concentration. Restricting ourselves to one site type at which only two codons are possible, each with its own cognate tRNA, we end up with

$$f = e^{-(c_1+c_2)} f_1(c_1)^{Lu_1} f_2(c_2)^{Lu_2} \tag{4.11}$$

or

$$\log f = -(c_1 + c_2) + Lu_1 \log f_1(c_1) + Lu_2 \log f_2(c_2) \ , \tag{4.12}$$

where L is the number of copies of this site type, u_1 and u_2 are the fractions of codons 1 and 2 at mutation selection equilibrium and c_1 and c_2 the expressions of their cognate tRNAs.

While at fixed c_1 and c_2 the equilibrium u_1 and u_2 are given by equation 4.4, changes of c_1 and c_2 can invade the population at fixed u_1 and u_2 if this leads to a fitness increase. Restrict ourselves

to gradual changes of c_1 and c_2, the condition for equilibrium is

$$\left.\frac{\partial f}{\partial c_i}\right|_{u_1,u_2} = 0, \tag{4.13}$$

and the equilibrium is stable (in the absence of fitness noise) if

$$\left.\frac{\partial f^2}{\partial c_i^2}\right|_{u_1,u_2} < 0 \tag{4.14}$$

which translates into $(\log f_i)'' < 0$.

The symmetry breaking mechanism requires some tradeoff between the cost of expression of tRNAs and the benefit for a specific codon of a higher expression of its cognate tRNA. Therefore, to understand the mechanism in its purest form we have to look at the regime where the typical tRNA concentrations are below the saturation points, if any, of the translational speed functions f_1 and f_2. Mathematically, we guarantee this by using scale-free functions.

Let $f_i(c) = a_i c^\alpha$, with $\alpha > 0$. Then $(\log f_i)'' < 0$ is satisfied everywhere so that the solutions we find using equation 4.13 are stable. Combining equation 4.13 and equation 4.4 we arrive at

$$\frac{u_2}{u_1} = \mu + O(\mu^{1+\alpha})\,. \tag{4.15}$$

For the linear case $f_i(c_i) = \gamma_i + \beta_i c_i$ we end up with

$$\frac{u_2}{u_1} = \mu + \left(1 + \frac{\beta_2}{\beta_1}\right)\mu^2 + O\left(\mu^3\right)\ . \tag{4.16}$$

Since c_1 and c_2 are positive, there won't be a solution for sufficiently large γ_i's which will reflect that the system will be best off getting rid of both tRNAs (biologically irrelevant case).

The system is bistable since we assumed $f_2 < f_1$ when using equation 4.4. Which state the system ends up with is determined by the initial conditions. The symmetry breaking is extreme in nature leading to the virtual disappearance of one of the two codons.

4.4.2 Selection on the overall speed of synthesis

The model above left open the question of how to choose the functions $f_x()$. Now we put in the expectation that what selection "sees" is the overall synthesis time τ. The three things we need to choose now are: how the translation time τ_i of a codon i depends on the expression $c_{t(i)}$ of its cognate tRNA (in general, several adaptor species contribute to the translation of a given codon; this is handled in section 4.5), how τ depends on $\{\tau_i\}$, and how τ contributes to the total fitness of the organism. Most of the basic elements of the model can be found in [68, 4, 5].

There is evidence [77] that the rate limiting step during the elongation of the protein chain is the diffusion of tRNA ternary complexes to (site A of) the ribosome. Under these conditions the waiting time τ_i for a cognate codon i is inversely proportional to the concentration of its cognate codon $c_{t(i)}$. Under the opposite assumption - that the decision time is limiting, we would have approximately $\tau_i \propto \sum_k c_k/c_{t(i)}$. In either case, $\tau_i = 1/c_{t(i)}$ seems a reasonable choice in some appropriate time and concentration units. (I will not worry about mechanisms changing the adaptor diffusion times, since presumably they don't coevolve with the codon usage.)

Next, I set

$$\tau = \sum_x m_x \tau_x, \tag{4.17}$$

where $\tau_x \equiv \tau_{i(x)}$ is the waiting time for the codon at genome position x, and m_x is an "expression level" of the codon, i.e. the number of times it is used as a template. In the case of genome replication, where every letter is replicated only once, we would have $m_x = 1$ for all x.

Now, I will assume that the cell duplication time is $T = \Theta + \tau$, and the fitness is proportional to T^{-1}. In fact, in exponentially growing cultures, growth rate has been reported proportional to translational elongation speed [78]. This would argue for $\Theta = 0$ under those conditions, but in general it will be nonzero, and controls the extent to which the elongation speed is the rate-limiting process.

Putting everything together we arrive at

$$f = \frac{G\left(\sum_k c_k\right)}{\Theta + \sum_x m_x/c_{t(x)}}\,, \tag{4.18}$$

where $c_{t(x)}$ is the concentration of the adaptor cognate to the codon at position x. In addition, we

can specify which codons are allowed at which sites by using the matrix $\tilde{F}_{si}$ as specified in equation 4.9. Above we used an exponential form of $G()$, and we will continue to use it due to its algebraic convenience, but all the different functional forms do is to renormalize Θ, and in particular have no effect if $\Theta = 0$.

Interpretation

The exact interpretation of the above equation depends on detailed assumptions, but its functional form is rather generic. For example, as discussed by Bulmer [5], in present day organisms, most likely the initiation rate, rather than the elongation rate, is limiting. The elongation rate controls the number of ribosomes on a mRNA in a steady state, and, from there, only indirectly affects the initiation rate and the speed of translation. If the initiation rate is inversely proportional to the concentration of free ribosomes, then the protein production rate of mRNA of type q is

$$r_q = k_q R_f = \frac{R k_q}{1 + \sum_p m_p T_p k_p}, \tag{4.19}$$

where k_q is its ribosome binding rate per ribosome number, R is the total number of ribosomes and R_f is the number of free ribosomes, and m_p is the number of mRNAs at the moment considered, and T_p is the typical time it takes to translate its mRNA once a ribosome binds to it. The rate of protein production of a given mRNA is affected not by its own elongation speed but by the current load of the translational system. Since $\sum_p m_p T_p k_p = R/R_f - 1$, the parameter Θ expresses the fraction of ribosomes that are idle. In the extreme where the template directed synthesis is not parallel, m_p will be the total number of times a part of the template is used. The assumption that the ribosomes wanders off after it finishes is not crucial either so the formulation applies to circular polysomes present in eukaryotes.

In the case of genome replication, we can imagine that cells accumulate resources for time Θ and then replicate for time τ, the total life cycle time being $T = \Theta + \tau$.

Adjustments are needed if the fitness is not dominated by the maximum growth rates but the ability to quickly produce only particular proteins - for example, toxins or antitoxins or quickly switch from one set of proteins to another. This can come from selection on survival rates under specific scenarios, and thus will result in independent contributions to fitness. If 1 to k are functional

groups of proteins subjected to specific selection we might expect factors of the form

$$f \propto r_1 \cdots r_k = \frac{1}{\left\{\Theta + \sum_p m_p T_p\right\}^k} \,. \tag{4.20}$$

It turns out that the larger the exponent k is, the stronger is the symmetry breaking, and the effective genome size L changes to L/k. Therefore, we will restrict ourselves to $k = 1$ - the least favorable case for the spontaneous emergence of biases.

Approximate independence of genome sites

The immediate problem of combining this model with the described mutation-selection equilibrium framework is that the different genome positions do not have independent effects on fitness. Now, we demonstrate the approximate independence of sites for large genomes. We prepare for mean field treatment by rearranging the terms in the denominator:

$$T = \Theta + \sum_x \tau_x = \Theta + \sum_s L_s \sum_i u_{si}\tau_i = T_0 \left\{1 + \sum_s L_s \sum_i u_{si}\left(\frac{\tau_i}{T_0} - \frac{1 - \Theta/T_0}{L}\right)\right\} \,. \tag{4.21}$$

For T sufficiently close to T_0 this can be rewritten as

$$\frac{1}{T} = \text{const} \prod_s \prod_i \left\{e^{-\tau_i/T_0}\right\}^{L_s\, u_{si}} \tag{4.22}$$

From here it follows that $F_{si} = \exp(-\tau_i/T_0)$. This treatment is self consistent only as long we can choose a constant T_0 which is close to T at all times despite the random fluctuations of T. This is indeed possible for large L since $\text{var}(T)/\langle T\rangle^2 < O(1/L)$.

With this we can use equation 4.2 to calculate the equilibrium codon usage with the added complication that we have to solve self consistently for T. This is how the codon equilibration step is performed in the simulations.

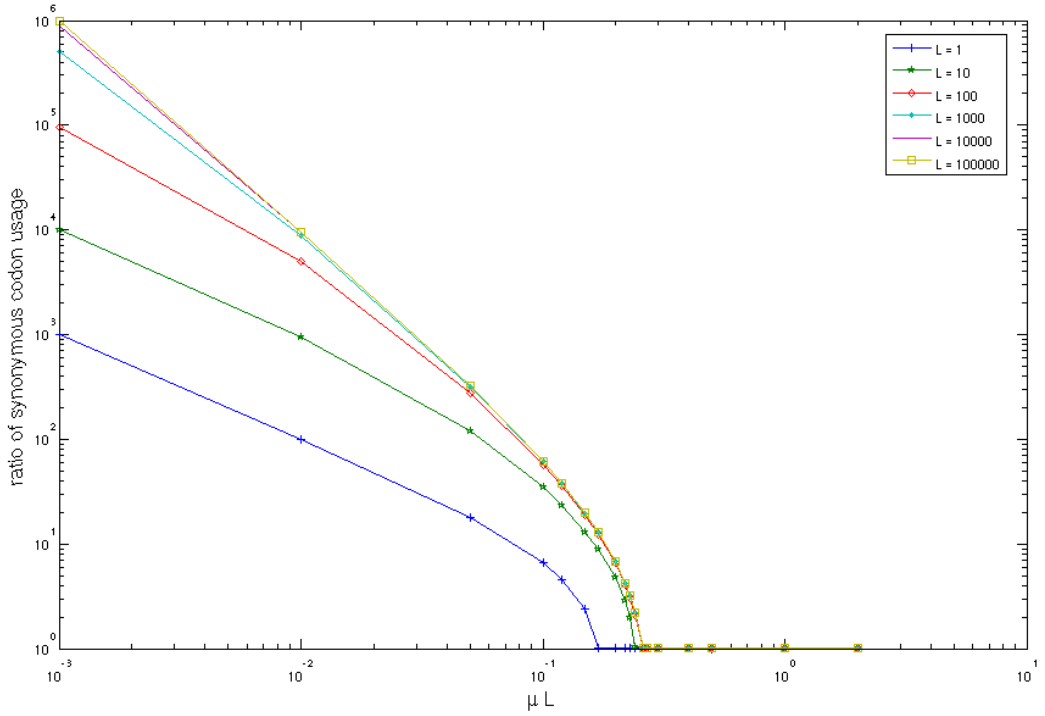

Figure 4.1: Spontaneous emergence of biased codon usage solely due to selection on translational speed in the one site type two codon model. For $L \gg 1$ the solution converges to a universal curve.

4.4.3 Numerical Results from the model

Model with two synonymous template letters

For the one site type two codon case and $\Theta = 0$ the fitness is given by

$$f = \frac{e^{-(c_1+c_2)}}{L\left(\frac{u_1}{c_1} + \frac{u_2}{c_2}\right)} \tag{4.23}$$

Notice that the only parameters of the model are μ and L. A simulation result[1] is presented on Figure 4.1 . The most prominent features are the universal dependence on μL in the large L limit, and the continuous phase transition leading to the spontaneous emergence of codon usage bias for $\mu L < 0.25$. The control parameter for the transition is μL, and the order parameter is the degree of codon bias. The fitness advantage of biased over unbiased genomes is presented on Figure 4.2.

[1]These simulations are performed in the continuous limit, and do not handle the effects of discreteness. This is irrelevant for large L, and the conclusion that the results depend only on μL in the large L limit is valid.

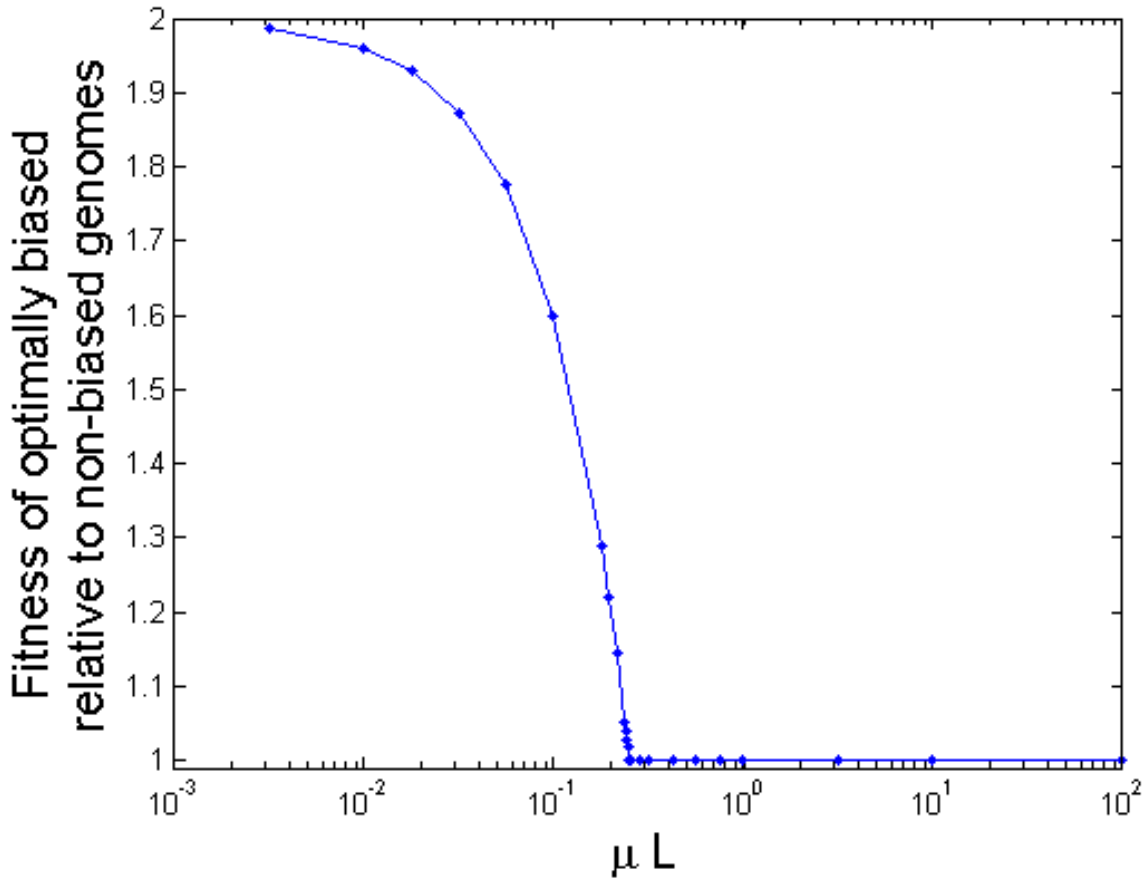

Figure 4.2: Fitness advantage of optimally biased over unbiased genomes.

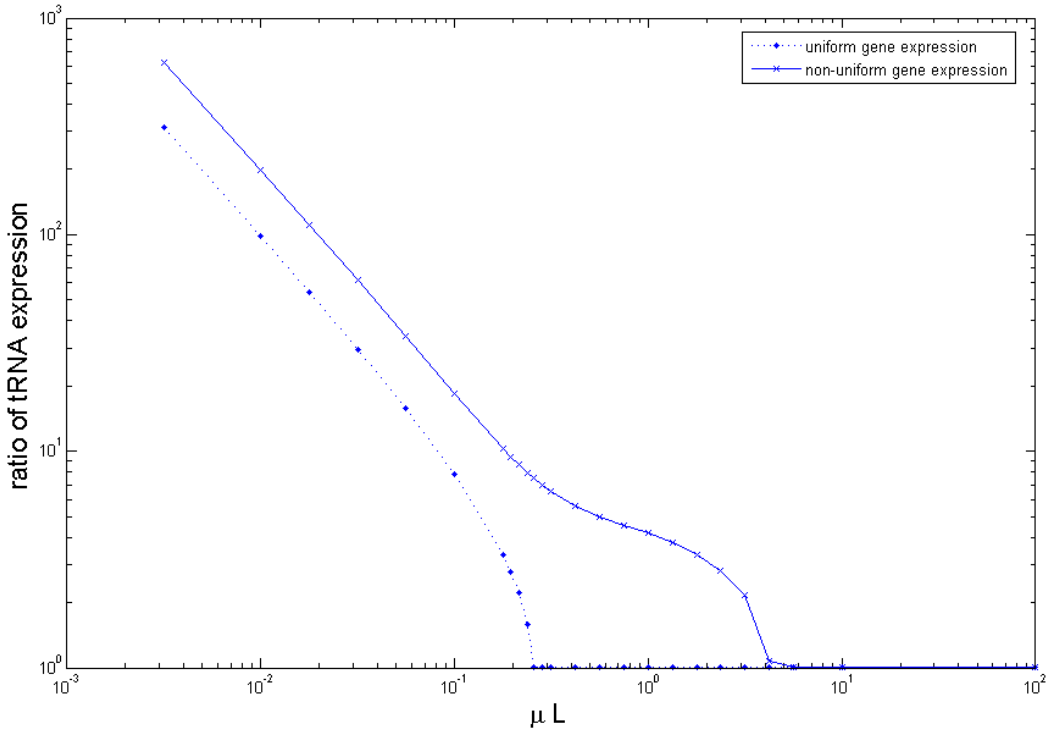

Figure 4.3: The spontaneous emergence of uneven tRNA expression levels is enhanced if some genome sites are more expressed than others. The onset of asymmetry is shifted towards higher μL. The dotted curve shows the result for a uniform expression level of all sites. The solid line curve presents the ratio of tRNA expression levels for a genome in which 10% of the sites have a hundred times higher expression.

Next we study the case in which part of the genome is more highly expressed than the rest, Figure 4.3. Non-uniform expression of different parts of the genome facilitates the emergence of codon bias which can be intuitively understood as reducing the *effective* genome size. In addition, the codon usage bias is stronger at the highly expressed sites, consistent with expectations and experiments. Figure 4.4

If there is a mutational bias, the symmetry breaking amplifies the codon usage bias expected from a mutational bias alone. For sufficiently low μL symmetry breaking in the direction opposing the mutational bias is possible, Figure 4.5.

Certain studies report that tRNA expression levels are highly correlated with gene copy numbers [70]. This motivates a discrete model of tRNA expression. The overall picture is unaffected, Figure 4.6, but if the size of the c step exceeds the the equilibrium values of c, the symmetry breaking is

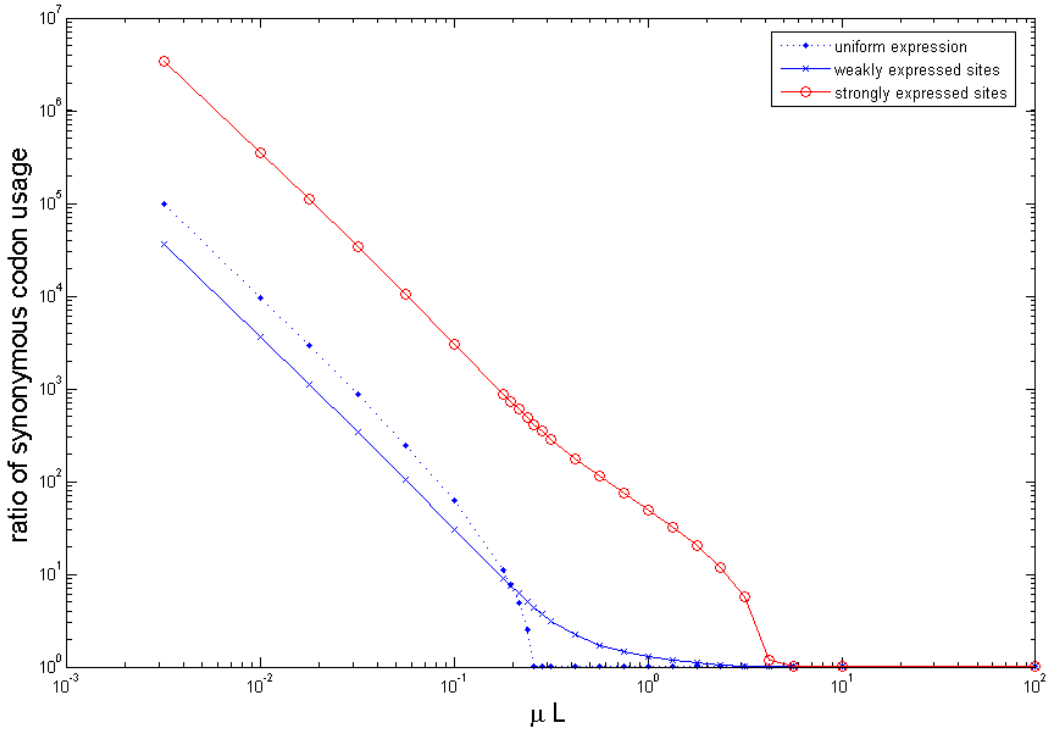

Figure 4.4: The spontaneous emergence of codon bias is enhanced if some genome sites are more expressed than others. The dotted curve shows the bias for a uniform expression level of all sites. The other two curves present the codon bias for a genome in which 10% of the sites have a hundred times higher expression. The codon bias is much stronger at the highly expressed sites. The onset of spontaneous emergence of codon bias is shifted towards higher μL for both the weakly and highly expressed sites.

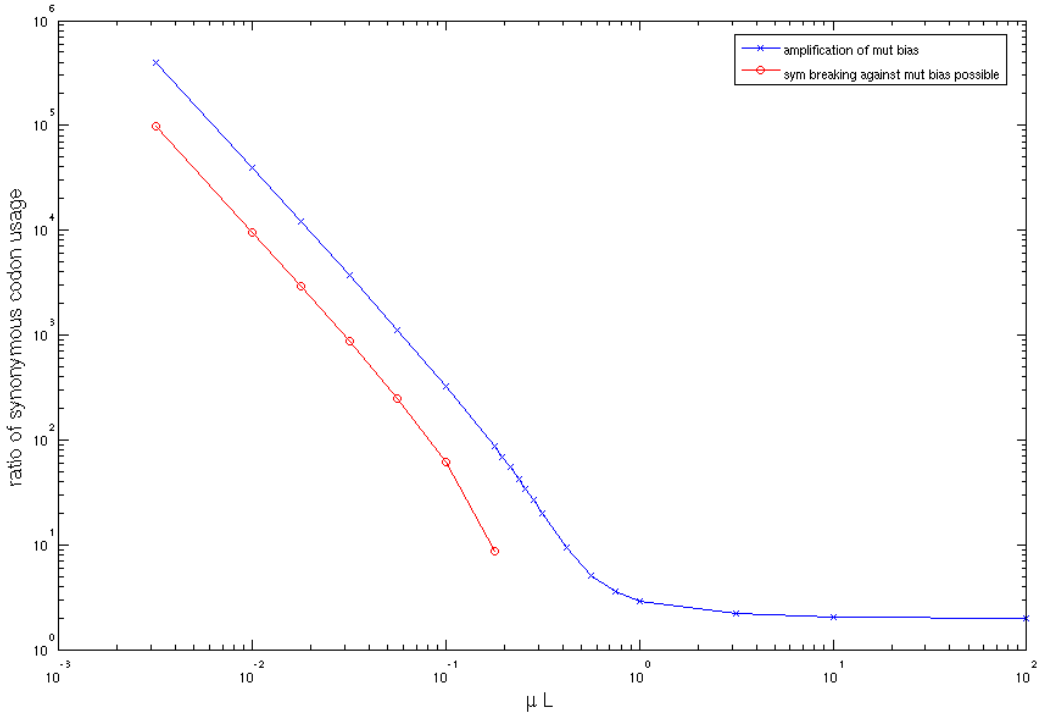

Figure 4.5: Codon usage in the presence of mutational bias. $\mu = \mu_{21} = 2\mu_{12}$, $L = 10^5$. At high mutation rates the equilibrium codon usage is determined by the mutational bias. At low mutation rates the mutation bias is spontaneously amplified. There is a transition point below which codon usage bias opposing the mutational pressure is also stable.

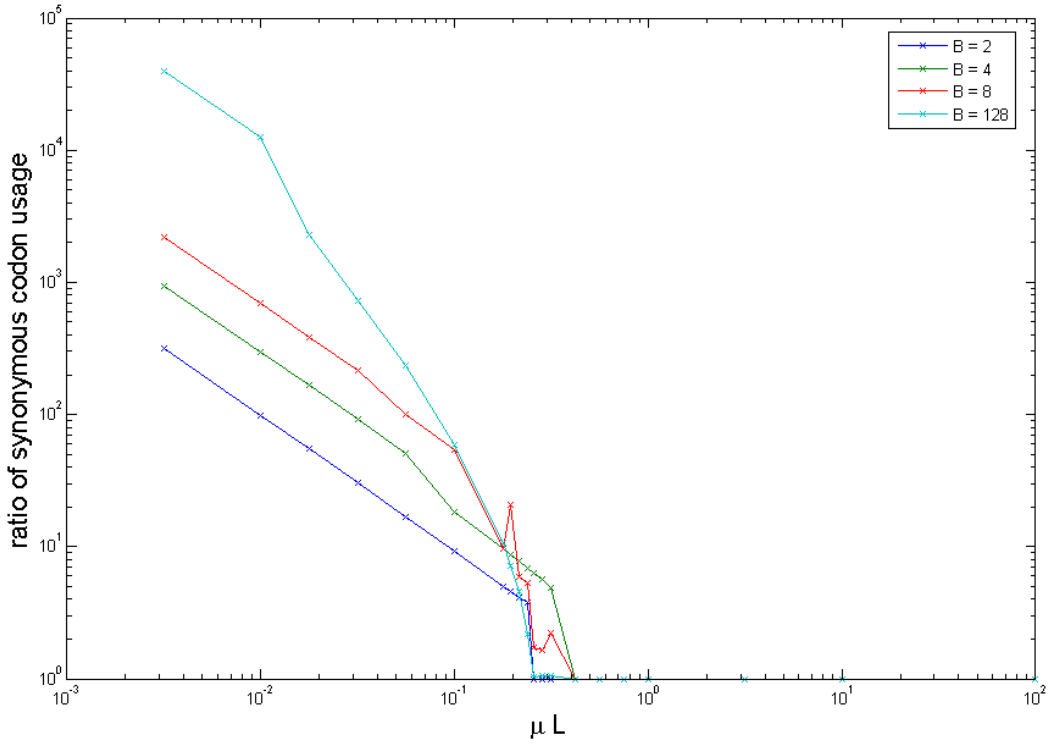

Figure 4.6: Codon usage symmetry breaking with discrete tRNA expression levels. Expression cost factor $\exp\{-\sum c/B\}$ is used . $L = 10^5$. The allowed tRNA levels are $c = k + \epsilon$, for $k = 0, 1, 2, \cdots$, and $\epsilon = 10^{-6}$. For $B < 2$ symmetry breaking is not observed.

suppressed. Figure 4.7 shows the corresponding tRNA levels[2].

Two sets of two synonymous codons

What happens when we have more than one site type? Let's look at the case of two sets of two synonymous codons. Codons 1 and 2 code for one amino acid, and codons 3 and 4 code for another. This can be put in the model by setting

$$\tilde{F} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \tag{4.24}$$

[2]The offset ϵ (see figure captions) is used to resolve the numerical issues when a codon with fitness zero is used zero times.

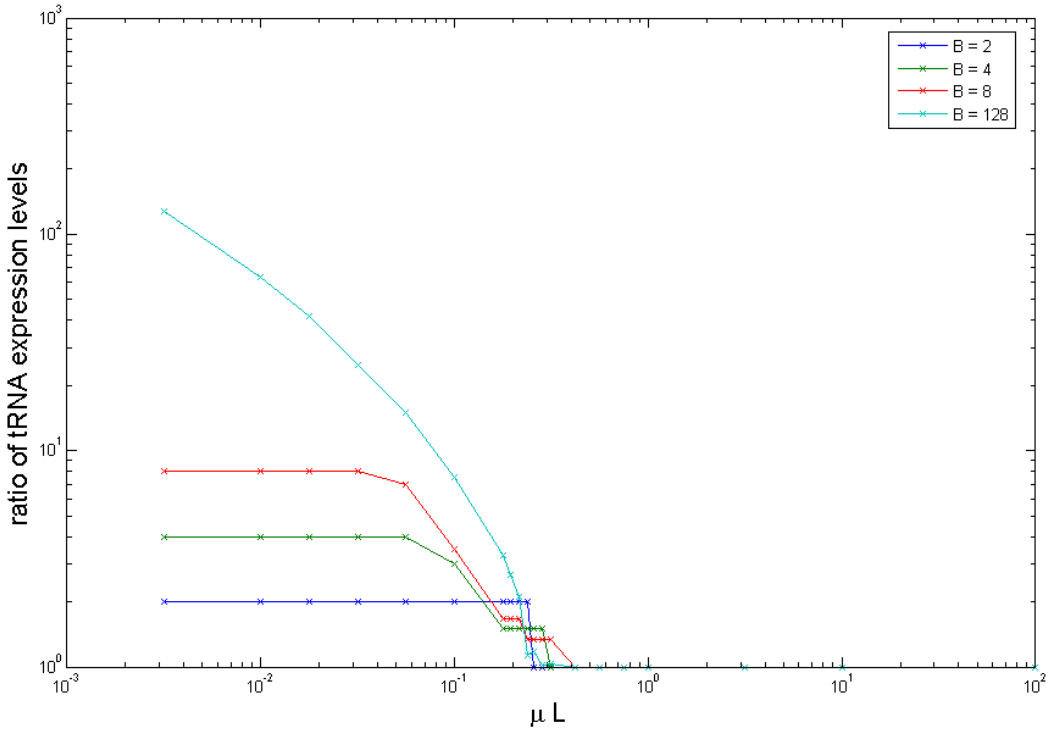

Figure 4.7: tRNA usage bias for a model with discrete tRNA expression levels. Expression cost factor $\exp\{-\sum c/B\}$ is used. $L = 10^5$. The allowed tRNA levels are $c = k + \epsilon$, for $k = 0, 1, 2, \cdots$, and $\epsilon = 10^{-6}$. For $B < 2$ symmetry breaking is not observed.

in equation 4.9, where the row specifies the site type, and the column - the codon. Now, assume that codons 1 and 2 and codons 3 and 4 are mutational neighbors, but 1 and 2 are away from 3 and 4, and 3 and 4 are away from 1 and 2, i.e.

$$M = \begin{pmatrix} 1-\mu & \mu & 0 & 0 \\ \mu & 1-\mu & 0 & 0 \\ 0 & 0 & 1-\mu & \mu \\ 0 & 0 & \mu & 1-\mu \end{pmatrix}. \tag{4.25}$$

Finally, assume that each of the codons has its own cognate adaptor.

Here are the results. If the two site types are used with equal frequency in the genome, then the symmetry breaking curve for both sets of synonymous codons coincides with the curve, obtained above, for the model with only one set of synonymous codons. More interestingly, if the two site types are present in different proportions in the genome, the symmetry breaking is stronger for the synonymous codons of the rare site type, Figure 4.8 and Figure 4.9. This by itself would suggest that the synonymous codon bias is stronger for amino acids that are weakly used, provided that the different codons have different cognate tRNAs.

Different modes of abruptly and randomly changing the tRNA levels were implemented both for the discrete and the continuous levels with highly similar results. The general conclusions are not altered when we generalize to more amino acids and tRNAs. Using the standard genetic code, assuming one tRNA species per codon and equal frequency of all amino acids we see that the tRNAs quickly split into popular and unpopular ones with almost zero expression for $\mu L = 0.1$.

4.4.4 Analytics

Here we solve the model exactly. The analytical solution given below fits the simulation results perfectly as shown on Figure 4.10 for $\Theta = 0$ and Figure 4.11 for $\Theta > 0$. Despite the availability of analytical results, simulations are still useful, because different model assumptions are easily tractable numerically, but not analytically.

Focusing, again, on the two codon (or two nucleotide) case with uniform expression we apply

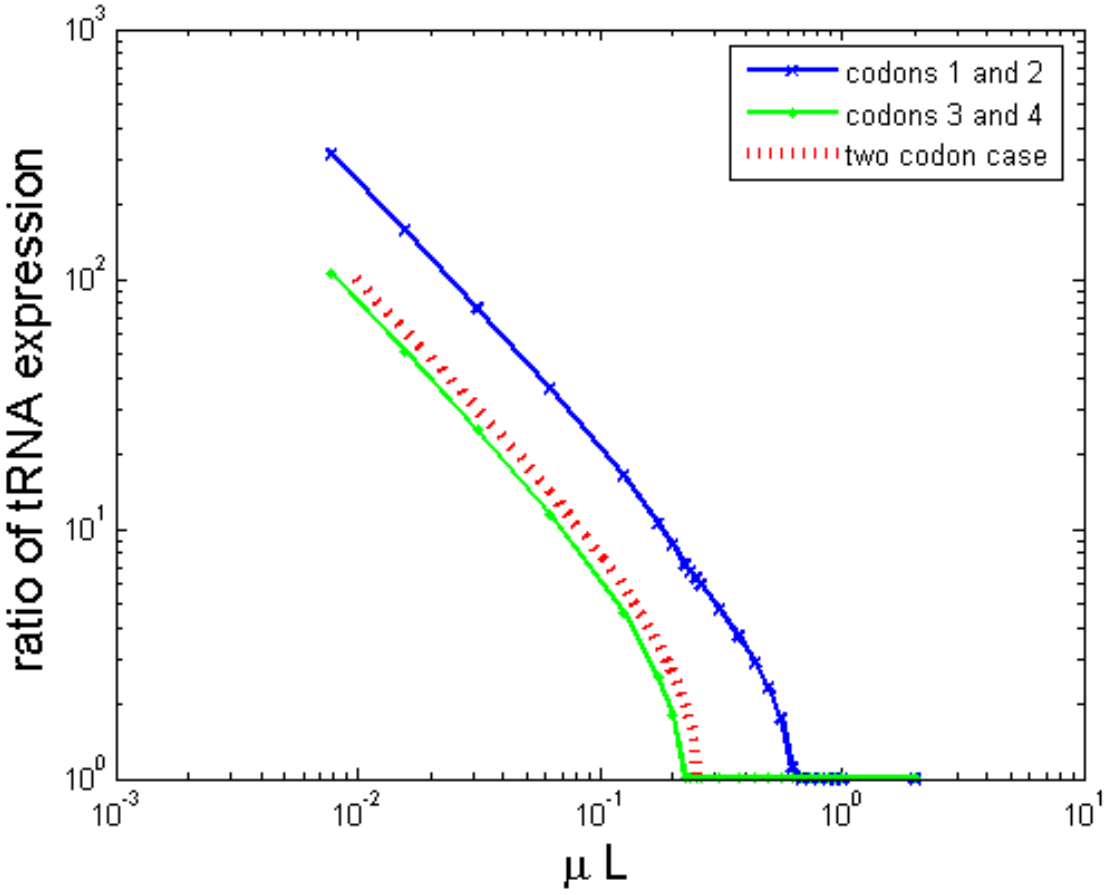

Figure 4.8: Spontaneous emergence of bias for two sets of two synonymous codons. Codons 1 and 2 are synonymous, and so are 3 and 4. The amino acid encoded by codons 1 and 2 is used at 10% of the genome sites. The amino acid encoded by codons 3 and 4 is used at 90% of the genome sites. The symmetry breaking of tRNA expression levels is stronger for the weakly used amino acid. $L = 10^5$.

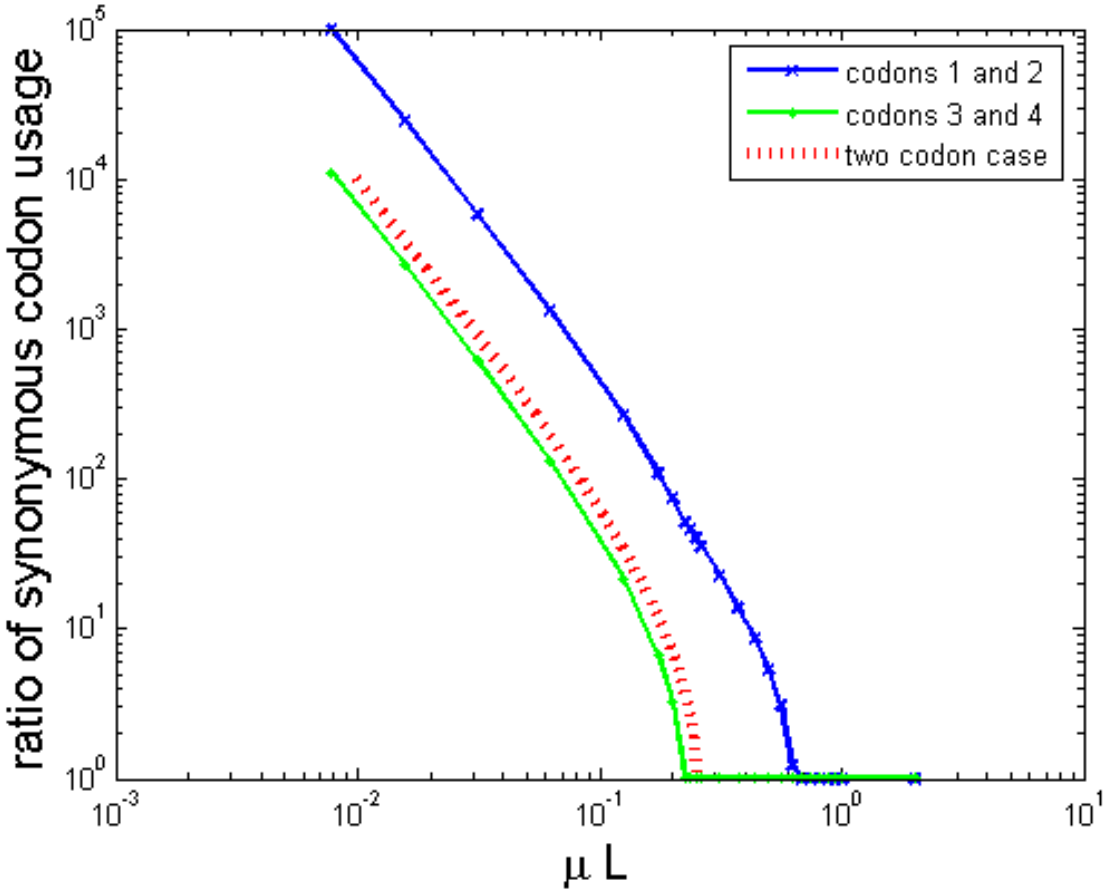

Figure 4.9: Spontaneous emergence of bias for two sets of two synonymous codons. Codons 1 and 2 are synonymous, and so are 3 and 4. The amino acid encoded by codons 1 and 2 is used at 10% of the genome sites. The amino acid encoded by codons 3 and 4 is used at 90% of the genome sites. The codon bias is stronger for the weakly used amino acid. $L = 10^5$.

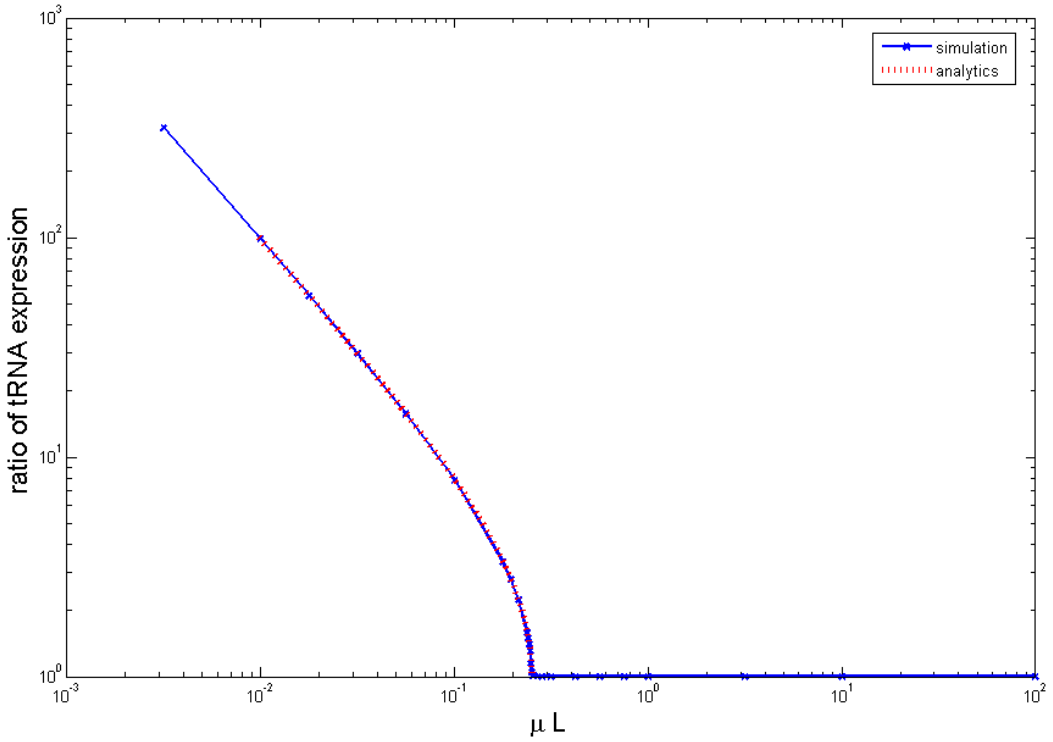

Figure 4.10: Comparison between analytics and simulations for the two synonymous codon case with $\Theta = 0$. $L = 10^5$. There is a perfect agreement.

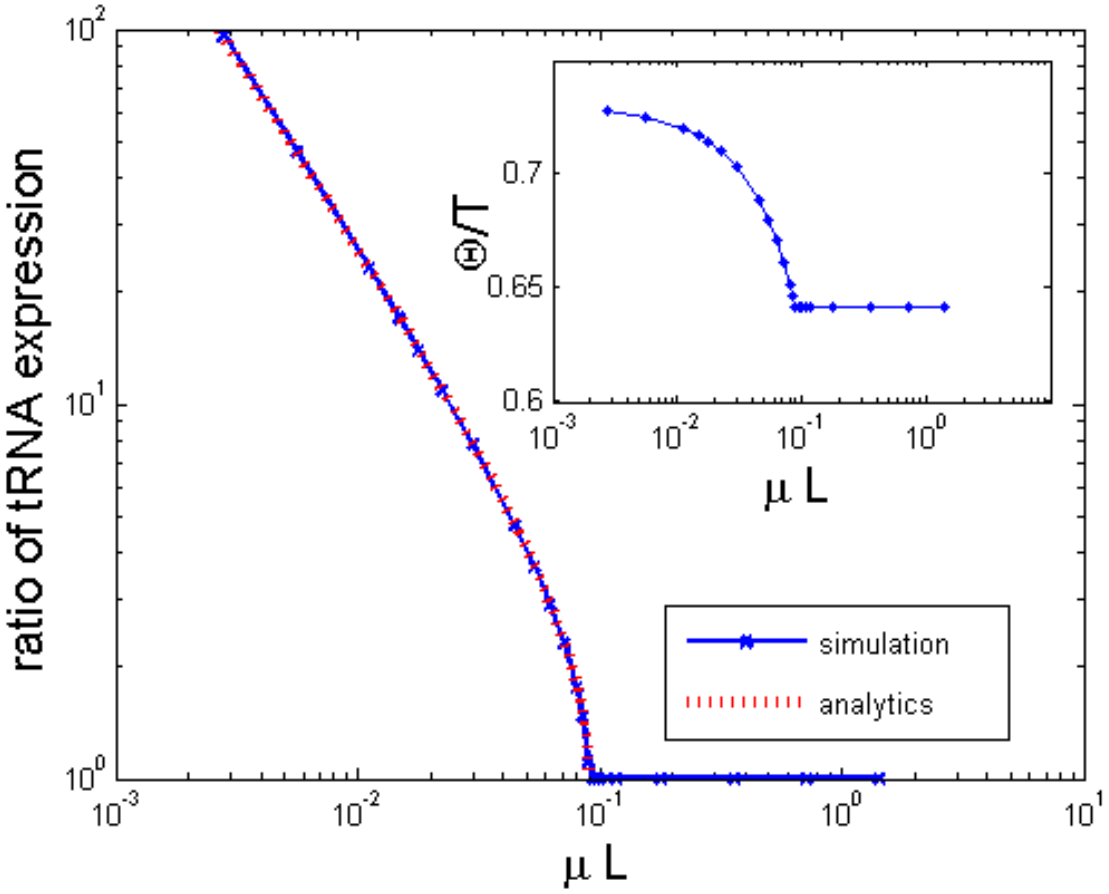

Figure 4.11: Comparison between analytics and simulations for the two synonymous codon case with $\Theta = 10 * L$. $L = 10^5$. There is a perfect agreement. **Inset.** The corresponding Θ/T. The synthesis time is less than half the total generation time, i.e. the symmetry breaking works well even if the synthesis time τ is only a minority fraction of the total time T.

the optimality condition in equation 4.13:

$$\frac{u_i}{c_i^2} = \frac{T}{L}\left(-\frac{G'(\sum c)}{G(\sum c)}\right), \tag{4.26}$$

which reveals a scaling relationship between codon usage and expression levels:

$$\frac{u_2}{u_1} = \left(\frac{c_2}{c_1}\right)^2 \quad . \tag{4.27}$$

Using the normalization $u_1 + u_2 = 1$ it also follows that

$$c_1^2 + c_2^2 = \frac{L}{T}\left(-\frac{G(\sum c)}{G'(\sum c)}\right) \quad . \tag{4.28}$$

Denoting $x = c_2/c_1$, and assuming $x \leq 1$, without loss of generality, it follows from equation 4.7 that

$$\mu L = \frac{x^2 \tilde{f}(x)}{1 - x^4}, \tag{4.29}$$

where $\tilde{f}(x)$, defined as before by

$$\frac{f_2}{f_1} = \exp\left(-\tilde{f}(x)/L\right), \tag{4.30}$$

is to be determined and dependent on $G()$ and Θ. f_2/f_1 is the relative fitness of the two codons.

From equation 4.22 it follows that

$$\tilde{f}(x; c_1 + c_2) = \left(\frac{1}{c_2} - \frac{1}{c_1}\right)\frac{L}{T} = \left(\frac{1}{c_2} - \frac{1}{c_1}\right)(c_1^2 + c_2^2)\left(-\frac{G'(\sum c)}{G(\sum c)}\right) \quad . \tag{4.31}$$

To express $\tilde{f}$ as a function of x alone we need another equation involving only c_1 and c_2. (In a sense, we need to determine the expression scale $c_1 + c_2$.) This can be done by excluding T from equation 4.28 and

$$T = \Theta + L\left(\frac{u_1}{c_1} + \frac{u_2}{c_2}\right) = \Theta + T(c_1 + c_2)\left(-\frac{G'(\sum c)}{G(\sum c)}\right) \quad . \tag{4.32}$$

For $\Theta = 0$ and any decreasing $G()$, we end up with

$$f(x) = \frac{x^2+1}{x+1}\frac{1-x}{x}, \tag{4.33}$$

and therefore we obtain the final result

$$\mu L = \frac{x}{(1+x)^2}, \text{ for } x \leq 1. \tag{4.34}$$

The spontaneous symmetry breaking occurs for $\mu L < \frac{1}{4}$.

For $\Theta \neq 0$, let's denote $\tau_0 = \Theta/L$. Assuming $G = \exp(-\sum c)$ for algebraic convenience, the shape of the curve is

$$\mu L = \frac{x}{1+x^2}\left\{(2\tau_0)^{-1}\left[\sqrt{1+\tau_0\frac{4(1+x^2)}{(1+x)^2}}-1\right]\right\}, \tag{4.35}$$

and the corresponding transition point is

$$(\mu L)^* = \frac{1}{4}\left\{\frac{\sqrt{1+2\tau_0}-1}{\tau_0}\right\}. \tag{4.36}$$

Therefore the smaller τ_0 is, i.e. the stronger the selection on speed is, the higher is the number of mutations per genome per generation at which spontaneous emergence of genome biases is possible. For large τ_0 we have only a square root dependance of the transition point on τ_0.

A typical number for the mutations per genome per generation in bacteria is 10^{-3}, for example [79]. This suggests that bacteria are typically in the symmetry broken phase. Care must be taken in interpreting this number since the population dynamics of bacteria does not follow exactly the population dynamics we use.

4.5 Modeling selection on (translational) accuracy

Here I briefly introduce the modeling of translational accuracy. A more detailed discussion follows in the next chapter.

4.5.1 Mistranslation

Mistranslation is used here as a translation of a codon via a non-cognate tRNA (also known as missense). Let the index t enumerate the different tRNA species within an organism. The relative probability that an adaptor (in contemporary setting, the ternary complex) attempts translation is proportional to its relative concentration c_t. Therefore, *irrespective of the details of translation*, the probability that a codon i will be translated via tRNA of type t can be modeled as

$$T_{it} = \frac{\epsilon_{it}\, c_t}{\sum_k \epsilon_{ik}\, c_k}, \tag{4.37}$$

where ϵ_{it} is the probability that codon i will be translated via tRNA species t if all adaptor concentrations are equal.

In the current study, we impose that every codon has a unique cognate tRNA and that $\epsilon_{ij} = \nu/9$ if i and j are nearest neighbors and $\epsilon_{ii} = 1 - \nu$, with ν being the *mistranslation rate*. It is easy to introduce 1st, 2nd, 3rd codon position asymmetry.

4.5.2 The fitness effect of mistranslation

Each gene is translated many times, possibly to a different amino acid sequence. We model the fitness at a given site s as the average over many translations, i.e:

$$F_{is} = \sum_t T_{it} W_{aa(t),s}, \tag{4.38}$$

where $aa(t)$ is the amino acid charged to tRNA species t and we ignore mischarging. $W_{\alpha,s}$ is the fitness of amino acid α at a site of type s.

Averaging over translations is not the only reasonable choice. If, for example, the proteome is such that many substitutions lead not only to non-functional proteins but to toxic ones, averaging would be unappropriate. The presence of toxic substitutions can greatly amplify the role of mistranslation for the evolution of the code. In addition, if only a small number of protein copies are translated, the width of the distribution might be relevant. Thus, averaging might be a somewhat conservative way of modeling the fitness cost of mistranslation.

Binding of a non-cognate tRNA can have additional fitness effects such as an increase of the

probability of drop off and frameshift errors and an increase of GTP cost per peptide bond due to proofreading. Such effects will only enhance the spontaneous emergence of codon bias and uneven tRNA expression.

4.5.3 Simulation results

To illustrate the fact that symmetry breaking can result purely from selection on translational accuracy we consider a situation in which there are four codons all of which differ from each other by a letter in the same position. Codons 1 and 2 are synonymous, code for an amino acid α, and recognized by different cognate tRNAs. Codons 3 and 4 are also synonymous but code for β. The tRNAs cognate to 3 and 4 have fixed expression levels. First we imagine that substitution of α with β and vice versa is lethal. On Figure 4.12 we focus on the stable expression and codon usage patterns for codons 1 and 2. We see that for a sufficiently low μ the coevolution leads to the virtual disappearance of one of the two codons. The higher the probability that a near cognate tRNA will bind, the higher the effect is at a given μ. Relaxing the lethality assumption the symmetry breaking decreases - Figure 4.13. Finally, I consider the effects on symmetry breaking of a distribution of sites with different sensitivities (functional constraints). At 90% of the sites substitutions of α with β lead to a 1% fitness cost ($s = 0.99$) and at 10% of the sites substitutions are lethal. Figure 4.14 shows that the presence of sensitive sites increases the codon bias at the non-sensitive ones. Codon bias is much stronger at the sensitive sites making codon disappearance at them more likely.

4.6 Selection on replication speed

Simulations were performed with both speed and accuracy components. We show that GC content can emerge from selection on speed of replication even at the expense of reduced precision (or higher energy expenditure due to proofreading) at the non-redundant first and second codon positions. To this end we assumed A, G, C, T, $AGCT$ site types. There is a fitness cost $1 - s$ if G, C or T is present at an A site type, etc. $AGCT$ site type is neutral, i.e. every letter has the same fitness.

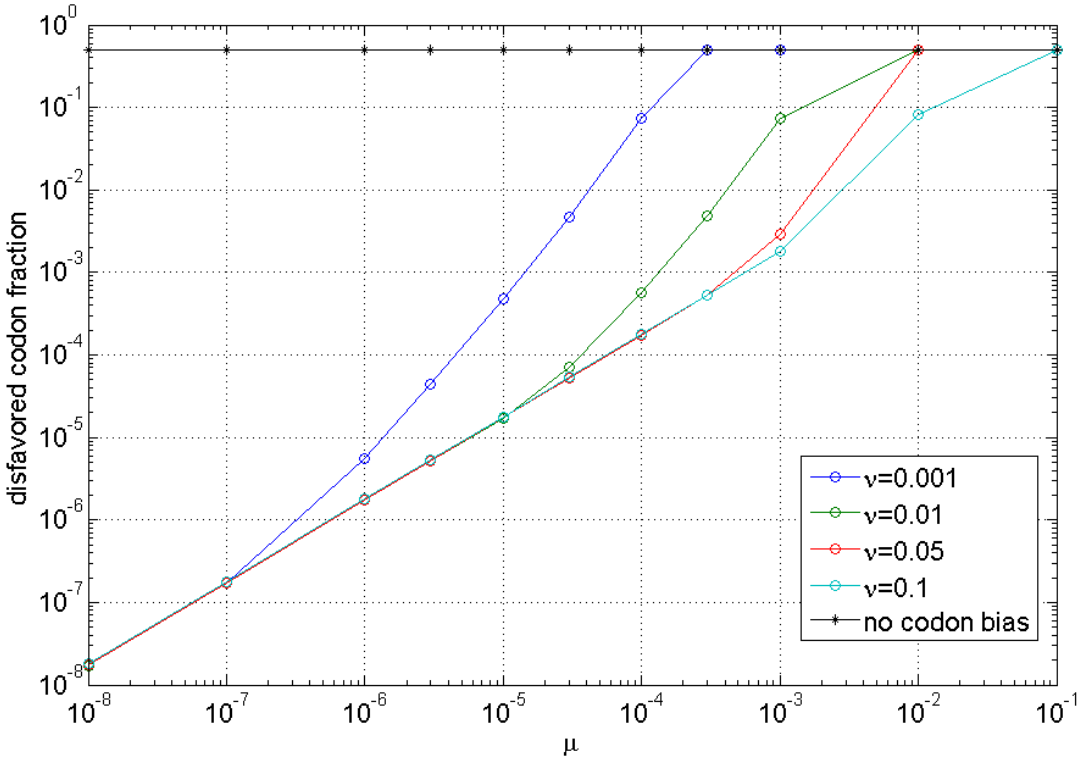

Figure 4.12: Symmetry breaking between two synonymous codons with exactly the same neighborhood at different mutation (μ) and mistranslation (ν) rates. Without tRNA - codon usage coevolution there is no symmetry breaking (black *). The symmetry breaking is solely due to selection on accuracy - there is no translational speed component.

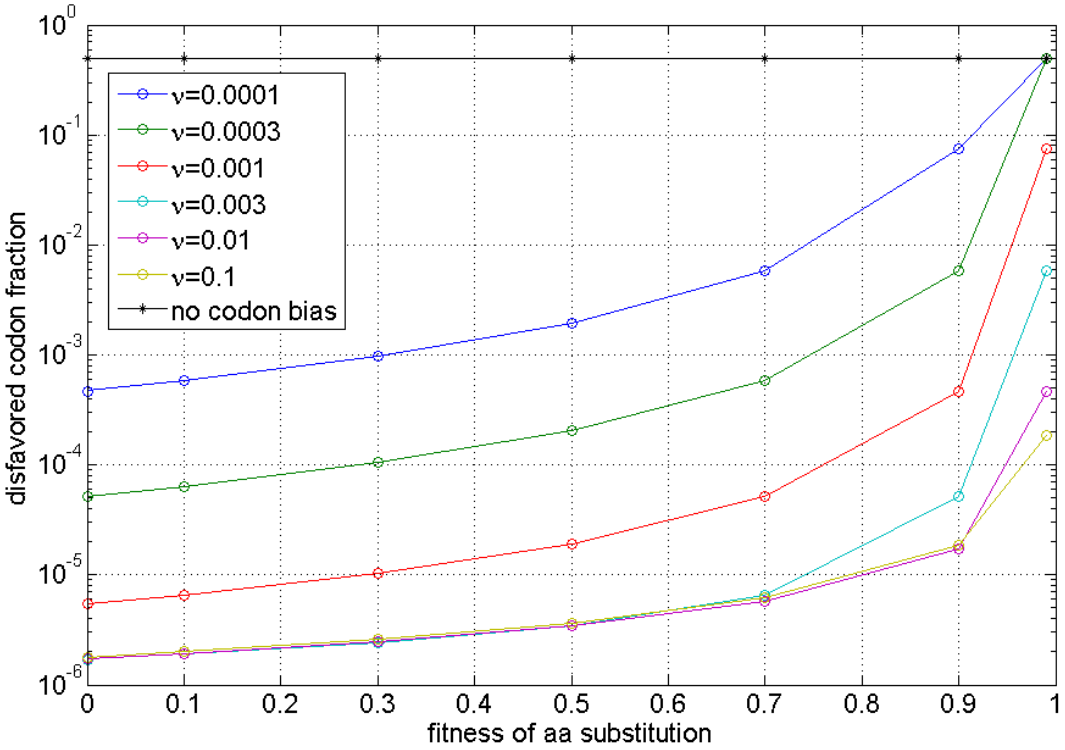

Figure 4.13: Symmetry breaking between two synonymous codons with exactly the same neighborhood solely due to selection on translational accuracy. The asymmetry increases with the mistranslation rate and with the fitness cost of a mistranslation. $\mu = 10^{-6}$. The cognate tRNAs of codons 3 and 4 are kept at a constant expression level. Without tRNA - codon usage coevolution there is no symmetry breaking (black *).

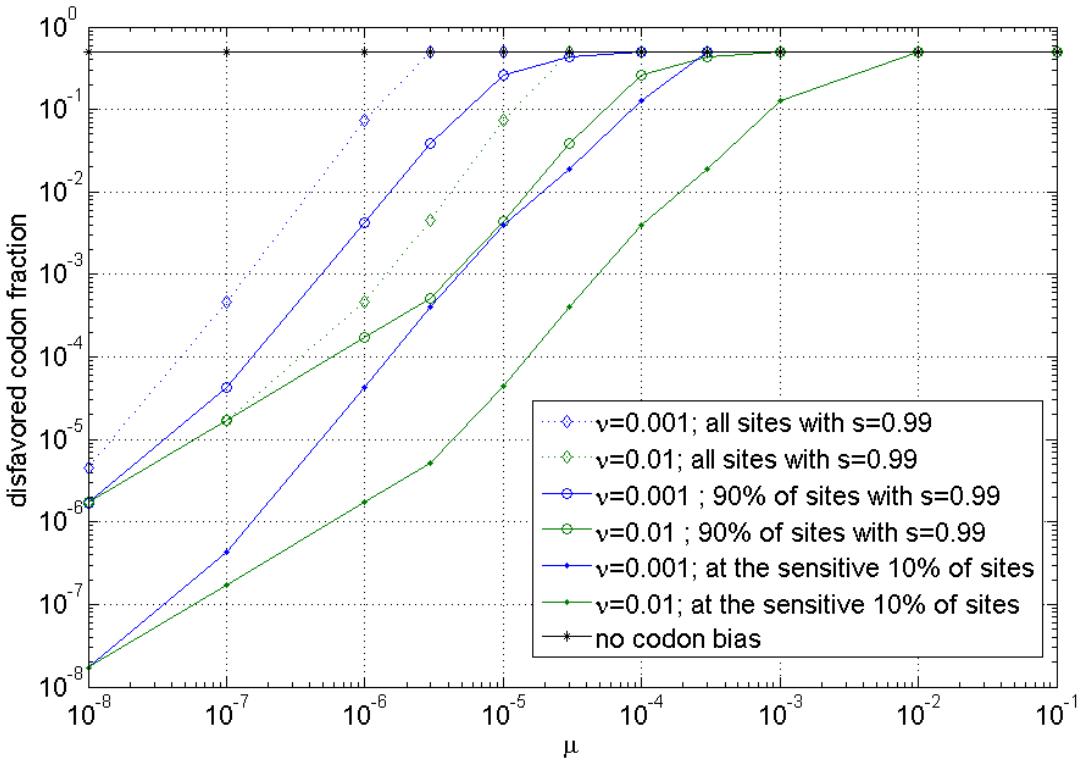

Figure 4.14: Symmetry breaking between two synonymous codons at site types of different sensitivity to amino acid substitutions. The codon bias is stronger at the more sensitive sites. The codon bias at the nonsensitive sites increases when sensitive sites are present. Green lines correspond to $\nu = 0.01$, and blue ones - to $\nu = 0.001$. Without tRNA - codon usage coevolution there is no symmetry breaking (black *). $\mu = 10^{-6}$. The cognate tRNAs of codons 3 and 4 are kept at a constant expression level.

This is captured by the matrix

$$\tilde{F} = \begin{pmatrix} 1 & s & s & s \\ s & 1 & s & s \\ s & s & 1 & s \\ s & s & s & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \tag{4.39}$$

Every third position is assumed to be of $AGCT$ type. All the other site types are present in one sixth of the positions. The formalism was modified to account for the fact that the genome is double stranded, and we assume symmetry between the strands: the replication time of a G position was assumed to be $\tau_G = (1/c_G + 1/c_C)/2$, etc. We used $s = 0.95$, $s = 0.999$ and $s = 1$ (everything is neutral). The results are presented on Figure 4.15. We see a continuous transition in μL with the GC/AT content being the order parameter. If we instead assume A, G, C, T, AG and CT site types, where AG means that there is a penalty if C or T is present, we arrive at a very similar graph - Figure 4.16. All site types are used in one sixth of the positions.

4.7 Transitions between stable states

4.7.1 Fitness noise

Codes and translational machineries do not just evolve towards greater optimality. Mutations decreasing the translational efficiency can hitchhike on beneficial changes in any other of the cellular components. For small populations drift is also important. Temporary mutator phenotypes can also facilitate transitions. This provides a rational for putting stochastic noise in the simulations. For the purposes of demonstration, noise is implemented here by considering not just beneficial mutations of tRNA levels but ones that are above a certain fraction of the current fitness.

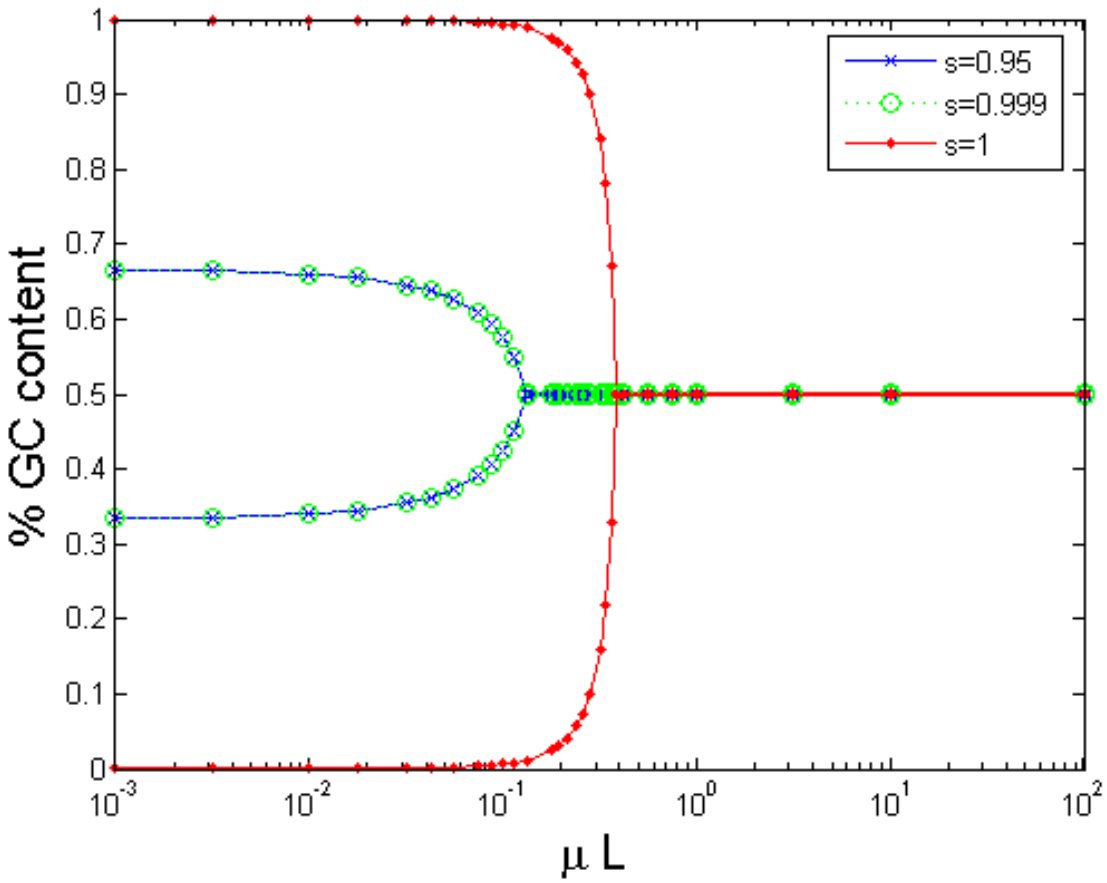

Figure 4.15: Spontaneous emergence of GC bias due to selection on replication speed. Every third position is assumed neutral. There are three curves corresponding to different fitness effects of single letter substitutions at non-neutral sites. $L = 10^5$. The rest of the parameters are given in the text.

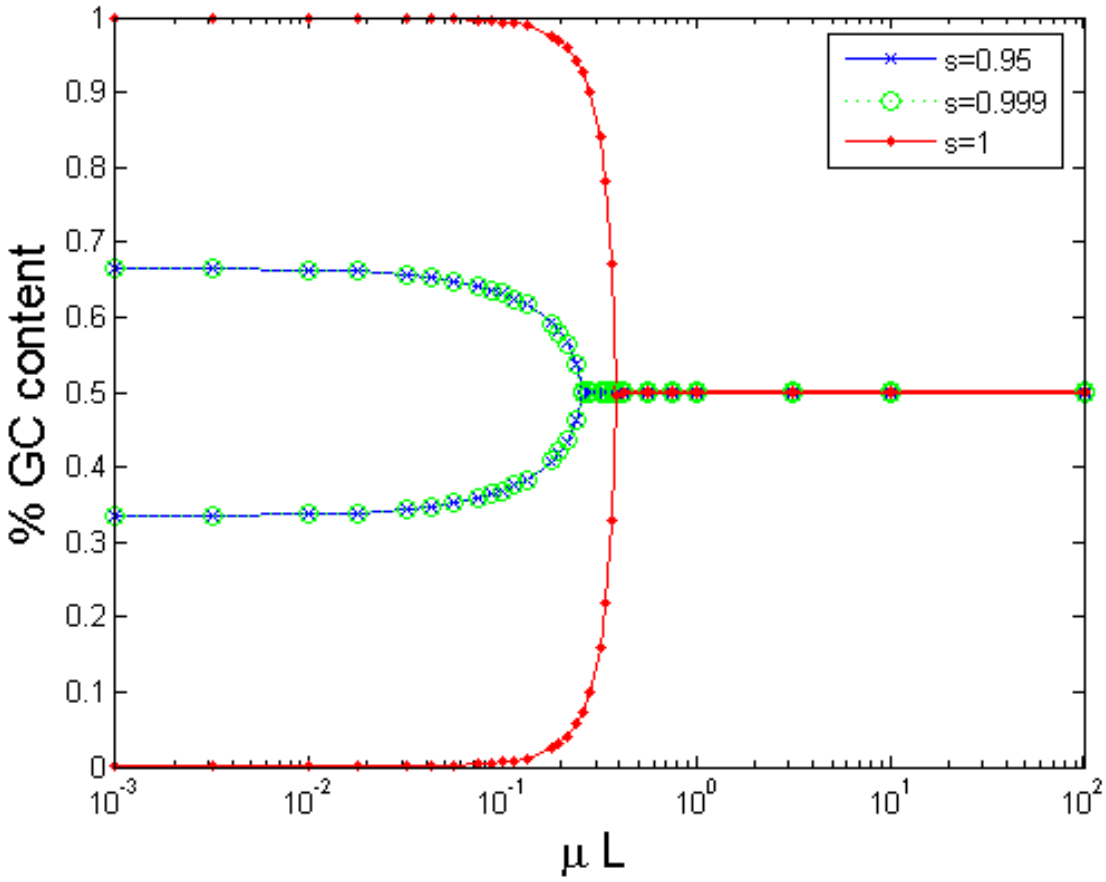

Figure 4.16: Spontaneous emergence of GC bias due to selection on replication speed. There are six equally likely site types named A, G, C, T, AG and CT. $L = 10^5$. There are three curves corresponding to different fitness effects of non-neutral single letter substitutions. The rest of the parameters are given in the text.

4.7.2 Simulation results

Similarly to above, I construct a mini code with four codons. 1 and 2 code for amino acid α and are read by different tRNAs, 3 and 4 code for β and γ. The geometry is

$$\begin{pmatrix} \alpha_1 & \beta \\ \alpha_2 & \gamma \end{pmatrix}. \tag{4.40}$$

α_1 and γ are not nearest neighbors; α_2 and β are not nearest neighbors. Figure 4.17 and Figure 4.18 show the evolution of the codon usage and the corresponding cognate tRNA expression levels of codons 1 and 2. The amino acid similarity matrix used in the figures is

$$W = \begin{pmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.01 \\ 0.9 & 0.01 & 1 \end{pmatrix}. \tag{4.41}$$

There are three site types. The genome consists of equal fractions of the three site types. The fitness noise is implemented by allowing any change with fitness greater than 0.99 times the current fitness to invade the population. The total expression level of the four tRNAs was kept constant.

Figure 4.19 shows the continuation of the above run. In view of the closeness of amino acid α to both β and γ the selection on translational accuracy, perhaps paradoxically, leads to the temporary exclusion of amino acids. A codon coding for one amino acid can be frequently used at site types favoring different amino acids. The codon usage does not vary only within the constraints of synonymous substitutions. In fact, the amino acid usage is very elastic, and covaries with GC content, in modern day organisms [80].

4.8 Epilogue

One of the big questions has been whether the origin of biases is neutral or caused by natural selection. I will end up with a short meditation based on what I've learned in writing this chapter.

While the mechanism of coevolution relies on selection, it is also neutral in the sense that for the same environmental constraints different codon usage patterns are possible and which one is

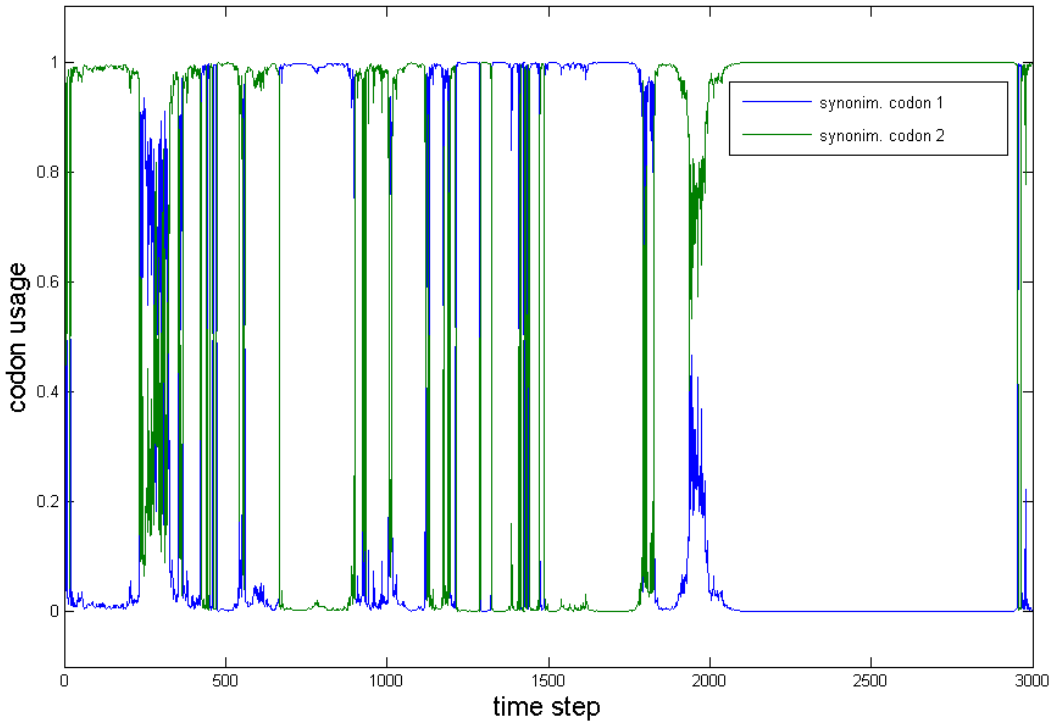

Figure 4.17: Time evolution of the codon usage of two synonymous codons in the presence of translational fitness noise. Most of the time the system is in one of two stable states with strong symmetry breaking. There are occasional transitions between the stable states. Parameters: $\mu = 10^{-5}$, $\nu = 0.01$. The amino acid similarity matrix used is given in the main text. Noise is implemented by allowing any change with fitness greater than 0.99 times the current fitness to invade the population.

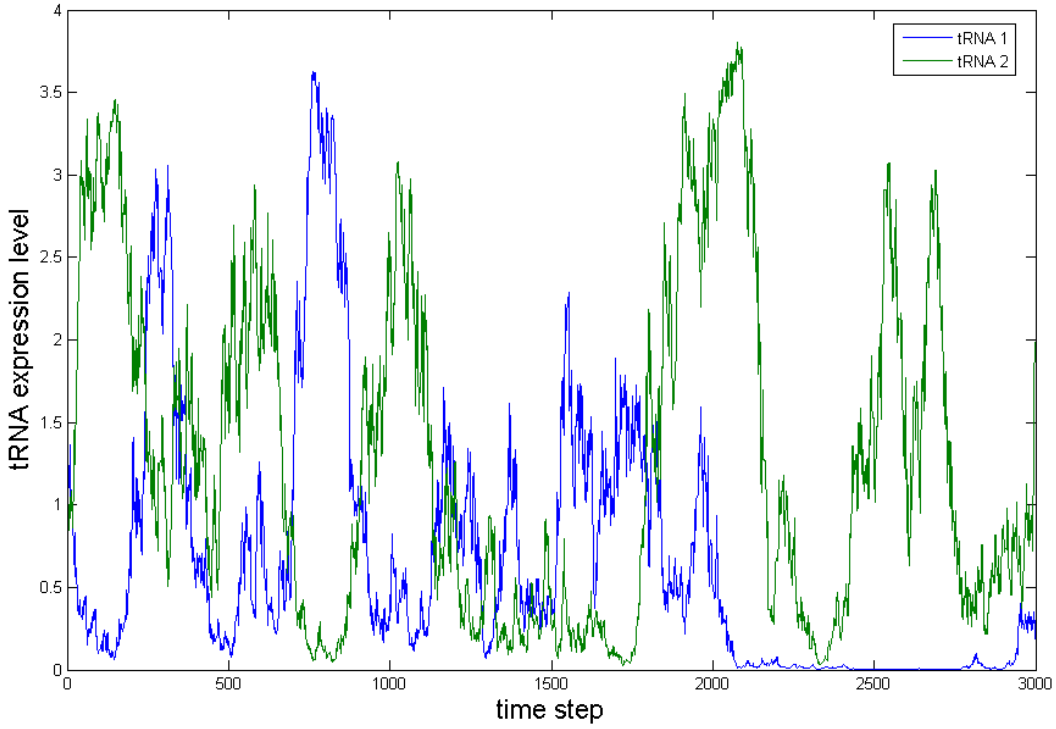

Figure 4.18: Time evolution of the tRNA expression levels corresponding to the previous figure. Parameters: $\mu = 10^{-5}$, $\nu = 0.01$. The amino acid similarity matrix used is given in the main text. Noise is implemented by allowing any change with fitness greater than 0.99 times the current fitness to invade the population.

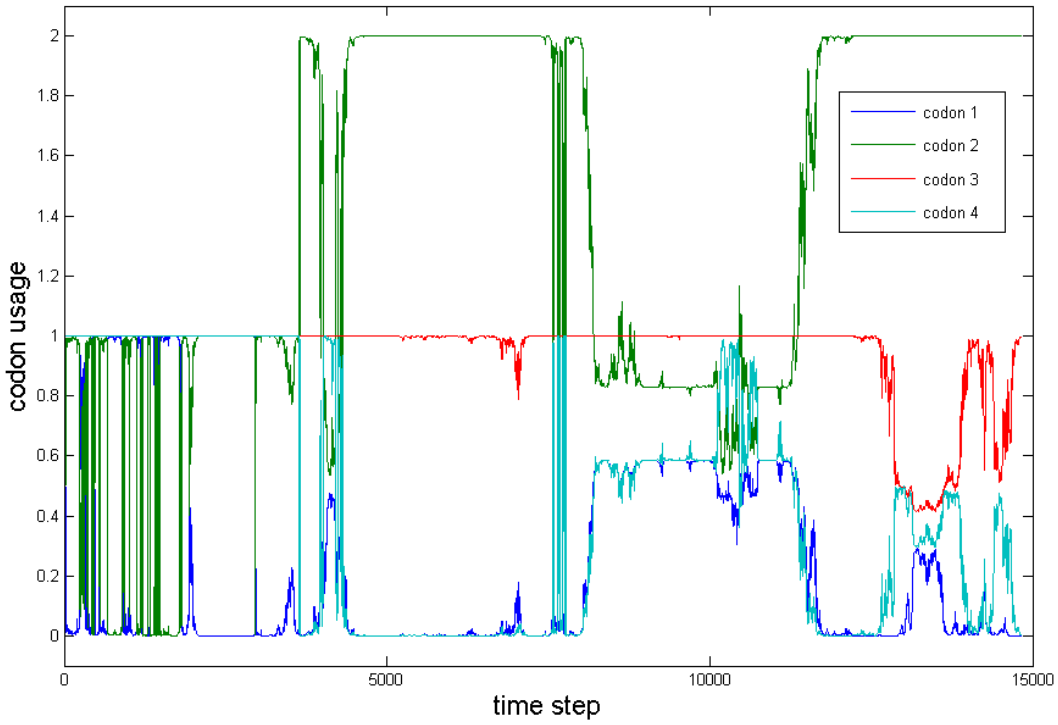

Figure 4.19: Evolution of codon usage in the presence of noise. Codon usage is normalized so that the sum is 3. Codon usage greater than one indicates that a codon coding for one amino acid is used at site types favoring a different amino acid. The four codons translate according to the mini code in equation 4.40. Parameters: $\mu = 10^{-5}$, $\nu = 0.01$. The amino acid similarity matrix used is given in the main text. Noise is implemented by allowing any change with fitness greater than 0.99 times the current fitness to invade the population.

used is a matter of chance. A way to look at this is that selection constrains the space of neutral possibilities. In this way the relation of one codon usage pattern to another is the same as that between two synonymous codons - two discrete (almost) neutral alternatives surrounded by many more non-neutral ones. A synonymous substitution is analogous to a transition from one pattern to another. The only thing that is different is the scale and the fact that codon usage patterns are composite objects.

The relevant distinction seems not to be neutral versus non neutral but universal versus environmentally specific pressures. This is along the lines of a distinction that Charles Darwin made between *natural selection* and *sexual selection*. This distinction was not appreciated and preserved and the canonical view now is that sexual selection is just one more type of natural selection. Of course this is true on the scale of the individual but it misses the point that the feedback between female preference and male advertisement or male preference and female advertisement is of universal nature, it is intrinsic to the sexual species. It is useful to think of sexual dynamics as a system with many spontaneous internal states over which the external pressures act.

So to summarize, in this chapter I argued that every template-directed synthesis process should be looked at as a non-trivial system capable of spontaneous self-organization (mediated by universal selection pressures) and multiple internal states, which is influenced by niche specific external pressures such as temperature, resource availability, life style, etc.

Chapter 5

Coevolution between tRNA expression levels and codon usage lubricates the evolution of the genetic code

5.1 Prologue

The genetic code is believed to be highly optimized and nearly universal. While optimization requires variability, variability opposes universality. Hence we are faced with the challenge of not only explaining the mechanisms through which the code can change but understand how, despite of these changes, the code is universal. In this chapter I focus on explaining how the code can change and optimize and in the next I deal with the question of universality.

The modular nature of the code specification - tRNAs and charging enzymes - seems conducive to change. Why is, then, the evolution of the code a conceptual problem? As Crick expressed it [81], every change to the code is lethal because it leads to many simultaneous amino acid substitutions, and the majority of such substitutions are deleterious. This statement, put in the context of explaining the universality of the code, constitutes the *frozen accident theory*. Historically, this argument has discouraged studies on the evolution of the code. For me, it has been a guiding light because it poses the problem clearly: to understand the evolution of the code we have to understand where Crick's intuitive argument breaks down.

The barrier to change is mitigated in a world in which the protein structures are error tolerant, the coding regions are not too many and are not responsible for the most crucial functions. Is

it, then, that the code was allowed to change only while the proteome was small and translation rudimentary? In this case, understanding the evolution and the structure of the code would require knowledge of the ancient biological context in which translation emerged. Or, did the evolution of the code continue well into the era in which template-generated proteins were carrying the majority of cellular functions and translation proceeded in an orderly fashion through a standardized competition of tRNAs? In this case, we should be able to explain aspects of the code's structure without reference to the complex origin of translation.

After Crick's paper, many non-standard codes were discovered [82, 83]. All variants are close to the standard genetic code and are derived relatively recently [84]. Therefore, even for a modern translation apparatus and in a protein dominated world, changes are possible albeit not very frequent. Thus, it is indeed plausible that the large scale evolution of the code can be understood within the translational framework we observe today, perhaps after allowance for some *quantitative* differences between contemporary and ancient translational mechanisms. In any case, this approach has to be tried since much will be learnt even if it fails. The alternative at this point is wild speculation about the context of early life.

5.2 Introduction

Immediately after the deciphering of the genetic code, regularities were noticed, and it was suggested that the code has undergone natural selection to minimize the lethal effects of mutation [85] and mistranslation [86]. Soon after, Woese [87] defined experimentally an amino acid property, related to hydrophobicity, which he named the *polar requirement*, and showed that similar codons code for similar amino acids. Crick [81] emphasized the barriers to the code evolution, and put forward the frozen accident theory to explain the universality of the code. In 1975, Wong proposed that the number of amino acids expanded by precursor amino acids seceding part of their codons to their biosynthetic products [88, 89]. The evidence that the genetic code is optimized with respect to the polar requirement was statistically quantified by Haig and Hurst in 1991 [90], and further strengthened by taking into account biased mistranslation and mutation [91]. Other amino acid measures were defined, based on free energy changes of protein structures caused by single amino acid substitutions [92] and by looking at substitution frequencies in evolutionary distant protein

homologs [93]. Knight [94] remeasured the polar requirement, confirming Woese's results, and argued that the fact that an amino acid similarity measure based on a single amino acid property leads to highest optimality suggests an early optimization of the genetic code.

A number of authors have proposed mechanisms or elements of mechanisms for code change [95]. In 1963, Crick [96] suggested that biased mutation could lead to the disappearance of codons, enabling their subsequent reassignment. This hypothesis was examined in details and elaborated by Osawa and Jukes [84, 97, 98]. Schultz and Yarus proposed that codon reassignments can proceed through intermediate states in which tRNA modifications lead to ambiguous translation [99, 100]. It was also speculated that pressure to minimize the translational apparatus can lead to disappearance of tRNAs and subsequent code change [101, 102]. Sella and Ardell [76, 103, 104] modeled the coevolution between the genetic code and the codon usage at different functional sites in the genome. They showed that the fitness cost of amino acid substitutions not only does not block the evolution of the code but guides its optimization.

In this chapter we demonstrate the relevance of the variability of tRNA expression levels for the evolution of the genetic code. Considering the finite cost of tRNA expression and the fact that both the speed and the accuracy of translation of a codon increase with the increase of the expression level of its cognate tRNA, we arrive at coevolutionary dynamics between tRNA expression levels and codon usage that provides an efficient mechanism for code optimization, even if, as the frozen accident theory assumes, every amino acid substitution is lethal at least some genome sites. This research shows that it is possible to account for the optimality of the code within the framework of translation as a standardized competition between tRNA adaptors.

5.3 What is like to be an adaptor?

People talk about *tRNA species*. How weird that is! There are protein *families* yet tRNA species. Let's take up this term *literally*.

As I argued in the previous chapter, each cell has a translational budget and translational goals which together with its codon usage constitute the environment that the tRNA species share. From a cell's point of view the *function* of the tRNAs is to decode efficiently and accurately its codons. From the tRNA's point of view the codons are the *environmental resources* that guarantee

their survival. The tRNA species actively shape their environment, which in turn affects their abundance and evolutionary prospects. As is typical for many ecosystems, the coevolution leads to a hierarchical distribution of species abundances which in turn facilitates a turnover of the species.

Especially severe is the competition between tRNA species encoding for synonymous codons. The coevolutionary process leads to the establishment of a dominance of one of the species and the marginalization of the others. The weak species can go extinct and replaced by others that exploit their codons in novel ways. Such events constitute code changes. An extinction can occur either before or after the appearance of a new species. In the first case there will be a period during which a codon is unassigned (codon capture), and in the second a period of ambiguous translation. A new species can remain marginalized as its predecessor or cause a large scale readjustment of the entire ecosystem.

Various speciation scenarios are possible. The number of species is not constant. Particularly important might be the evolution of generalists - tRNAs that decode more than one codon through the mechanism of wobble pairing. Unfortunately for them though, there is simply no biochemical way to translate all synonymous codons with the same accuracy and efficiency. Generalists create uneven distribution of codons, inviting in turn invasions from tRNAs specializing in decoding the popular codon better than the generalists. And unfortunately for this thesis, the study of this interesting dynamics is left out of it. We assume a one to one mapping between codons and species.

What seems worrisome in the discussion above is that we forgot about the cells. In fact, there is no inconsistency. The upper level - the cellular species has just one knob for controlling the level below - differential survival of cells. Having one knob on a complex system does not invalidate a description focused on the complex system itself rather than the owner of the knob. For example, the fact that the Fed can control the money supply to achieve goals of its own, does not necessitate a description of the economy in which the Fed takes the most central role.

The knob that the cells have, determines some of the rules in the ecosystem, and in particular provides an overall directionality to its long term evolution. Apart from chance, this directionality is the factor that prevents the quick equilibration of the ecosystem. Directionality comes from the pressure to reduce the fitness effects of mistranslations, and perhaps mutational load. While the top level provides the incentives for change, the evolutionary barriers are ultimately overcome by

coevolution and competition at the bottom level. The end result is that the optimization of the code is not only possible but inevitable.

The *ecosystem* representation constitutes a mood change from the previous chapter where the emphasis was on broken symmetry and the implicit metaphor was that of a *ferromagnet*. These metaphors focus our attention on different time scales and evolutionary motifs of the complex dynamics. The adventures of the adaptor will continue in the next chapter. Adaptors are not citizens of the cell - they are citizens of the world. They like to travel. Now, time to get serious for a few very serious sections.

5.4 Modeling and discussion of assumptions

In a pioneering work, Sella and Ardell [76, 103, 104] introduced a model that aimed to capture the coevolution between the genetic code and the encoding of a proteome. They introduced the notion of a codon usage at a site type and described a procedure for calculating it at mutation selection equilibrium in a quasi-species framework. The barrier to code change is modeled as the requirement that a change in the genetic code invades the population only if it leads to a higher overall fitness given the equilibrium codon usage of the old code. We extend this framework in several biologically important directions: changing tRNA expression levels, many site types per amino acid, lethal substitutions, and in the next section - HGT. All simulations presented assume one to one correspondence between tRNA species and codons. Mutations, mistranslations, fitness effects of mistranslation and codon usage equilibration are modeled as described in the previous chapter.

5.4.1 Site types, proteome structure and amino acid similarity

Amino acid substitutions at different genome positions are assumed to contribute multiplicatively to fitness. In reality, the fitness effects of individual substitutions or mistanslations, which ultimately shape the codon usage, might be very different from the combined effect of many simultaneous substitutions triggered by a code change (because of net positive or negative epistasis). But since we are interested only on the average genome wide effects of mistranslation, we believe that such correlations are qualitatively irrelevant. Moreover, it turns out that with the tRNA expression

dynamics most reassignments do not actually lead to many simultaneous substitutions.

We model the proteome as a collection of different *site types*. Each site type s is characterized by the fitness $W_{\alpha\, s}$ of different amino acids at that site. The matrix W together with the frequencies $\{L_s\}$ of the different site types in the genome constitutes the *structure of the proteome* in this model.

From this structure we can construct an *amino acid similarity matrix* $S_{\alpha\beta}$ that we use only in the analysis of the results of the simulations, and, in particular, to test the optimality of the code. There is no unique or right way to construct S. Using models, such as this one, one can actually verify that the qualitative presence or absence of optimality with respect to ensembles of random codes is typically not sensitive to the precise construction of S, and thus justify the use of somewhat arbitrary amino acid similarity matrices as a staring point for statistical studies of the genetic code. The particular formula for calculating the amino acid similarity matrix is

$$S_{\alpha\beta} = \sum_s |W_{\alpha,s} - W_{\beta,s}| \ , \tag{5.1}$$

and for the the optimality score:

$$\sum_i \sum_j N_{ij} S_{aa(i),aa(j)} \ , \tag{5.2}$$

where N_{ij} is 1 if i and j differ by a single letter and zero otherwise. $aa(i)$ is the amino acid of codon i, i.e $aa()$ is the genetic code.

The notion of a site type does not imply that there is a target proteome sequence. It implies that the protein structures do not evolve on the time scale on which code or tRNA expression changes invade the population. Moreover, as long as the overall proportion of different site types is preserved, the model dynamics is insensitive to, and thus allowing, evolution of the protein sequences. It is conceivable, though, that the proteome structure systematically adjusts to changes in translation over long time scales, by introducing new site types or changing their relative proportions. Whether or not such coevolution between the genetic code and the protein structures brings qualitatively new aspects is not addressed in the current study.

In this chapter we emphasize the importance of considering *more than one site type per amino acid*. This expression means that for each amino acid there are several site types for which it is the

best fit but with different effects of the rest of the amino acids. For example, at a position in the protein where hydrophobicity matters the effect of the non-perfect amino acids depends on how different the hydrophobicity is from the optimal one. At other site types size might matter. There the effect of the non perfect amino acids depends on how different their size is from the optimal one.

5.4.2 Translational speed fitness and cost of expression

The translational speed increases with the concentration of the adaptors, and saturates at a certain level. Correspondingly, we assumed that the translational speed factor of the fitness of one instance of codon i is given by $f_i^{speed} = 1 - \exp(-cbr_i/\Gamma)$, where $cbr_i = \sum_k \epsilon_{ik}\, c_k$ is the codon binding rate of codon i and Γ is a constant.

We assume that each expression unit brings the same multiplicative fitness burden and correspondingly we arrive at an exponential fitness cost factor $\exp\left(-B \sum_t c_t\right)$, where B is a parameter controlling the magnitude of the burden.

5.4.3 A closed model of the evolution of the genetic code

Putting everything together we end up with the following expression for the fitness of a typical representative of a quasi species.

$$f\left(code, \{c\}, \{u\}\right) = e^{-B\sum_t c_t} \prod_i \prod_s \left\{ f_i^{speed} \left(\sum_k \epsilon_{ik} c_k \right) \sum_j T_{ij} W_{aa(j),s} \right\}^{L_s\, u_{si}} . \qquad (5.3)$$

At each step we attempt to modify the genetic code with probability κ and the tRNA expression levels with probability $1 - \kappa$. In each case we examine in random order the possible changes until we find one that is acceptable. We accept a candidate change if it increases or at least preserves the fitness value above calculated using the existing codon usage $\{u_{si}\}$. This is in contrast with Sella and Ardell's assumption that the change with *maximum* fitness invades the population. Our choice is the consistent one if code changes are rare compared with the codon equilibration time. The simulation ends when no neutral or beneficial changes exist of either the code or the expression levels. After a change is been made we equilibrate the codon usage.

In the future, the acceptance criteria can be generalized to be a probabilistic process that depends on the ratio of the fitness values before and after the change.

5.5 Results

We are concerned with the spontaneous optimization of genetic codes to apparently minimize the average difference of amino acid properties between codons that differ at only one position. Correspondingly, we follow the time evolution of ensembles of initially random codes, and in particular observe the evolution of the distribution of optimization scores.

We compared the optimality score of a code with the distribution of optimality scores for **two different ensembles of random codes**. In the first, *type A*, every codon is equally likely assigned to every amino acid. In the second, *type B*, we randomly permute the amino acids of the code of interest while preserving its redundancy structure. These two tests give different information. The first captures the increase in optimality coming from both improving the redundancy structure (including reduction of the number of amino acids) and improving the arrangement of amino acids. The second factors out the optimality coming solely from improving the redundancy structure. For each set of initial conditions we perform **two runs with different dynamics**: one with variable tRNA expression levels and one with fixed tRNA expression levels.

We compared the two models under **different assumptions about the proteome structure** - different numbers of site types and different distribution of highly harmful amino acid substitutions. All results presented are without noise and employed discrete equally spaced tRNA expression levels of size one (which provides us with a natural criteria for freezing). We used 20 amino acids and 64 codons.

First we used the proteome structure employed by Sella and Ardell [104]. There is one site type per amino acid and a one-dimensional amino acid space. Each amino acid is assigned a number (property value) $A(\alpha)$ randomly chosen from the interval $(0, 1)$ and the fitness of α at site s is taken to be

$$W_{s\alpha} = \phi^{|A(s)-A(\alpha)|}, \tag{5.4}$$

where $\phi \in (0, 1)$ is a parameter characterizing the maximum severity of amino acid substitutions.

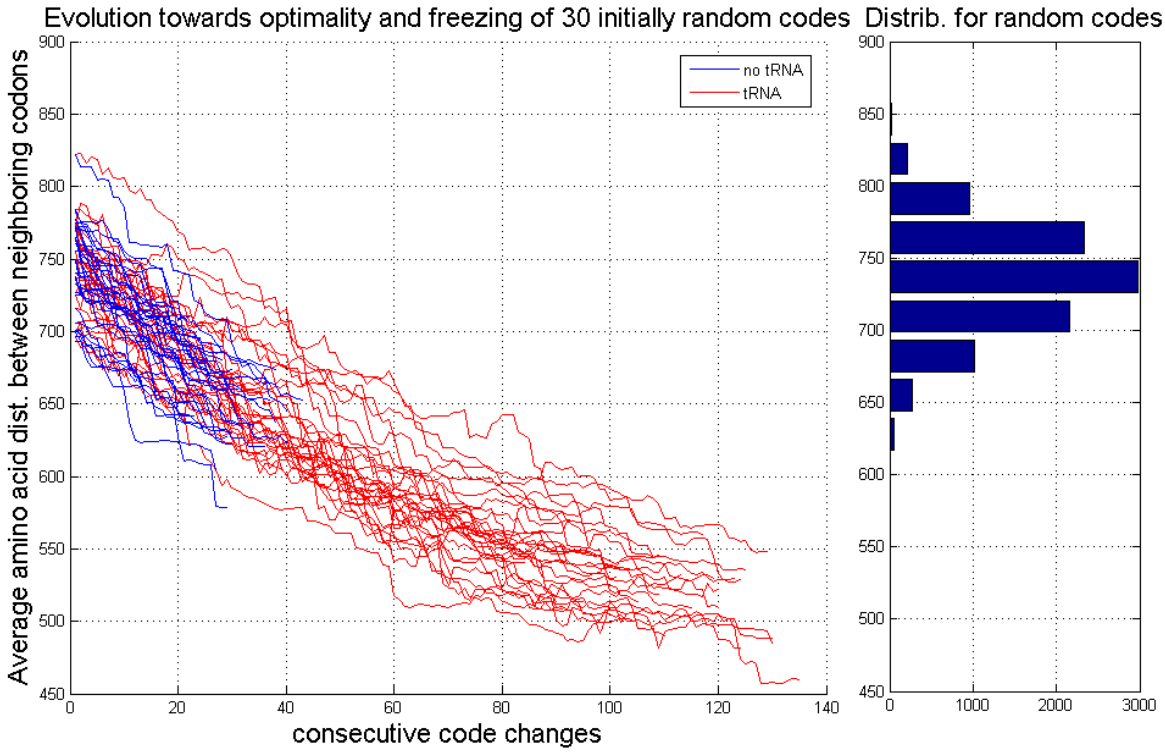

Figure 5.1: **Left panel.** Evolution towards optimality of 30 initially random codes with fixed (blue) and coevolving (red) tRNA expression levels. Only strictly beneficial and neutral changes were considered and the evolution was followed until the codes froze. Parameters: $\mu = 10^{-4}$, $\nu = 10^{-3}$, proteome structure of type I with $\phi = 0.99$, translational speed fitness with $B = 1/60$ and $\Gamma = 1$. **Right panel.** The distribution of optimality scores for random codes drawn from an ensemble of type A.

We label this *proteome structure type I*. This set of assumptions is inspired by the polar requirement of Carl Woese under which the current code has been shown to be highly optimized.

A typical run is presented on Figure 5.1. We see that while the evolution with fixed tRNA levels leads to some optimization, the optimization is greatly enhanced when the tRNA levels are allowed to adjust.

The model with fixed tRNA levels optimizes well at small ϕ and large mistranslation rates as shown on Figures 5.2 and 5.3. Correspondingly, the advantage of models with varying tRNA levels are more pronounced at high ϕ and low mistranslation rates.

One of the points of this chapter is that the proteome structure used by Sella and Ardell is very different from the one implicit in the intuition that the code cannot evolve because every

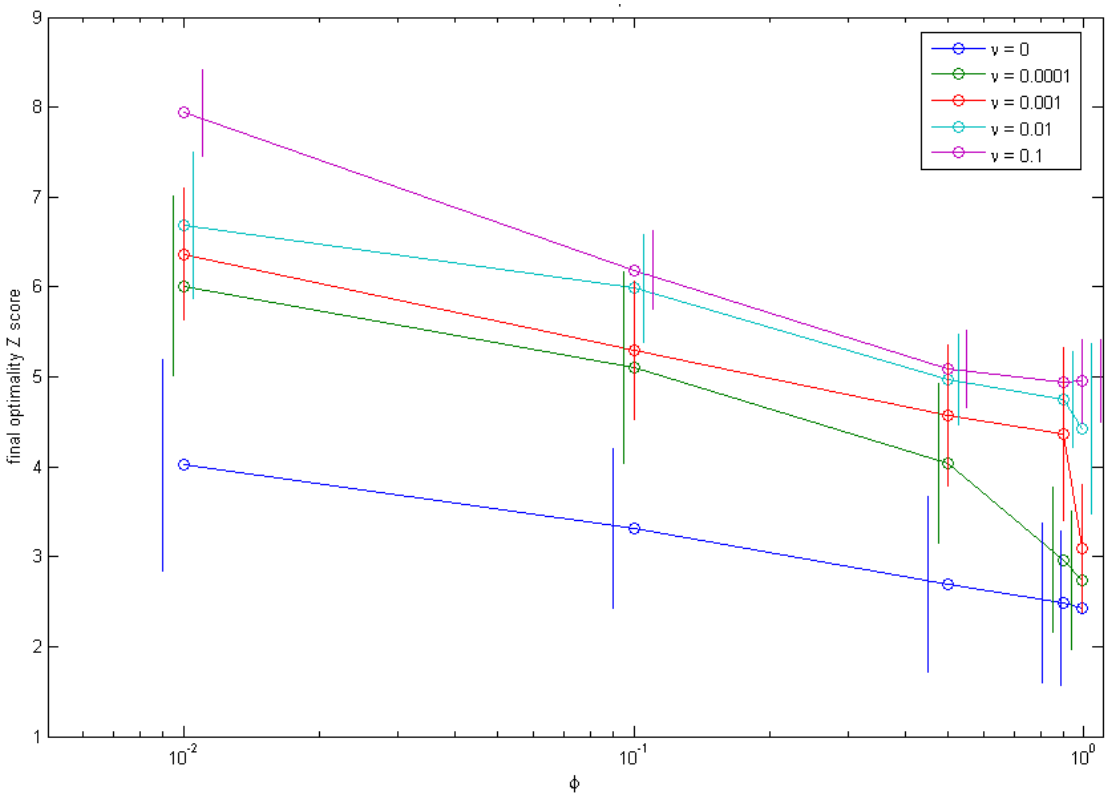

Figure 5.2: Z scores (standard deviations away from the mean) of the final optimization relative to random code ensemble type A. $\mu = 10^{-5}$. The different colors denote different mistranslation rates. The points denote the mean values for an ensemble of 30 initially random codes. Slightly shifted vertical lines denote the standard deviations of the Z scores.

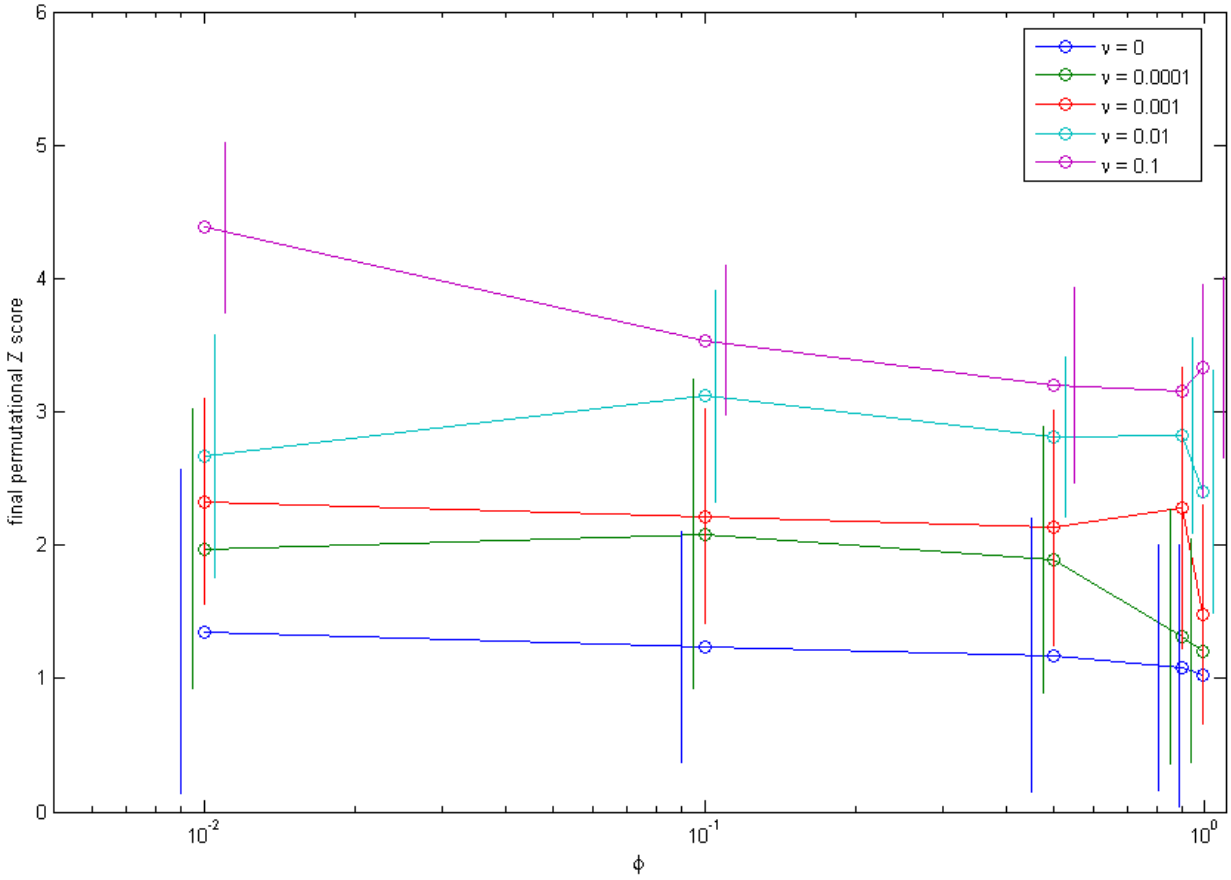

Figure 5.3: Z scores (standard deviations away from the mean) of the final optimization relative to random code ensemble type B. $\mu = 10^{-5}$. The different colors denote different mistranslation rates. The points denote the mean values for an ensemble of 30 initially random codes. Slightly shifted vertical lines denote the standard deviations of the Z scores.

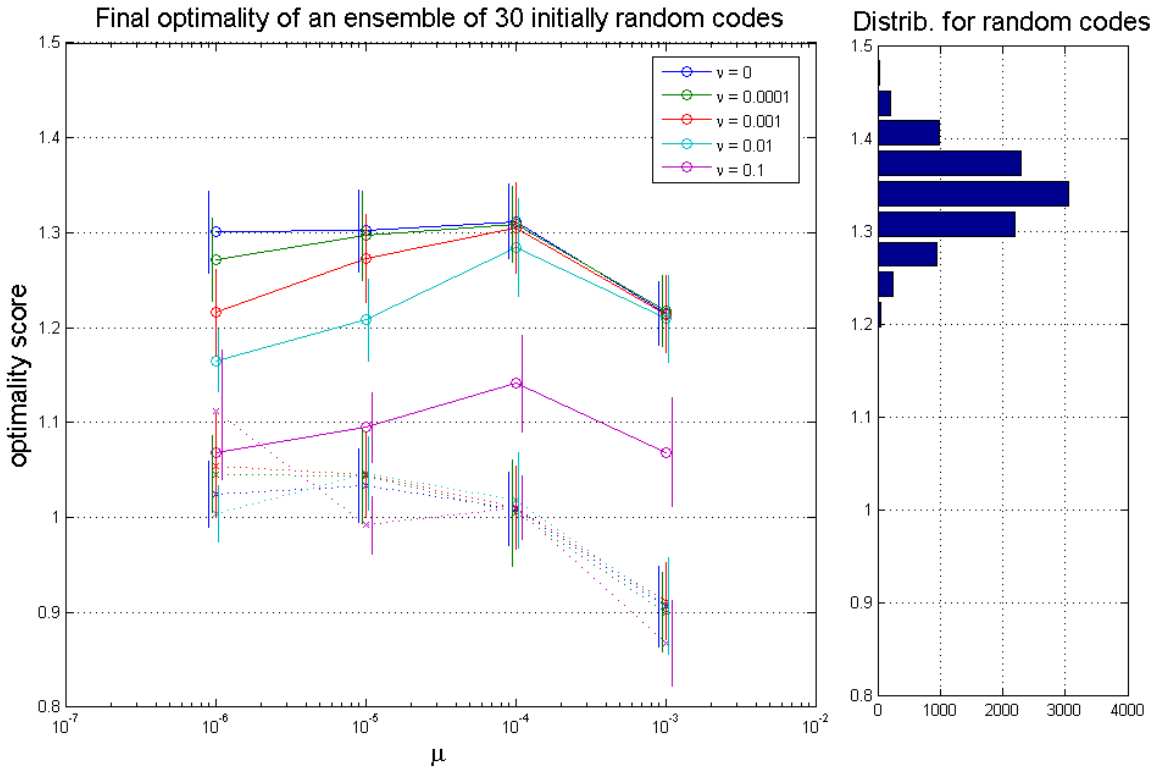

Figure 5.4: **Left panel.** Final optimality of an ensemble of 30 initially random codes with fixed (solid lines) and coevolving (dotted lines) tRNA expression levels. Different colors denote different values of ν. Only strictly beneficial and neutral changes were considered and the evolution was followed until the codes froze. Parameters: proteome structure of type II with $n = 3$ and $\phi = 0.99$, translational speed fitness with $B = 1/60$ and $\Gamma = 1$. **Right panel.** The distribution of optimality scores for random codes drawn from an ensemble of type A.

code change introduces at least some lethal substitutions. To observe freezing we need to consider multiple site types per amino acid and postulate that at some sites some amino acids have severe fitness consequences. To this end we extended the proteome structure to include three, and then five, site types per amino acid. The way we constructed a n site types per amino acid proteome is to generate n sets of $\{A_k(\alpha), \phi_k\}$ with A_k's generated as above. We label this *proteome structure type II*. The results for $n = 3$ are presented on Figure 5.4 and for $n = 5$ on Figure 5.5. In general, the more site types we have, the less efficient is the optimization obtained with the model with constant tRNA levels. Finally, we endowed the proteome not only with many site types but site types for which some of the amino acid substitutions have mild fitness cost and at the same time, others have very high fitness cost. High fitness cost was used to approximate lethality since true lethality

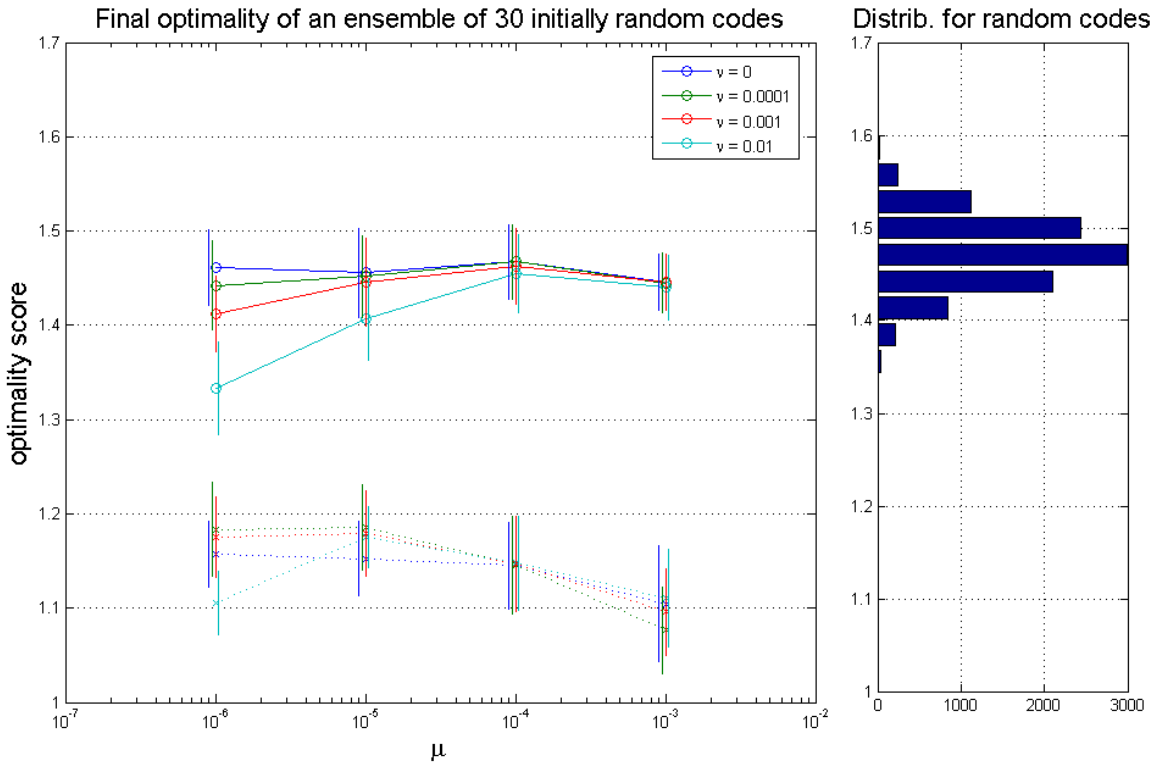

Figure 5.5: **Left panel.** Final optimality of an ensemble of 30 initially random codes with fixed (solid lines) and coevolving (dotted lines) tRNA expression levels. Different colors denote different values of ν. Only strictly beneficial and neutral changes were considered and the evolution was followed until the codes froze. Parameters: proteome structure of type II with $n = 5$ and $\phi = 0.99$, translational speed fitness with $B = 1/60$ and $\Gamma = 1$. **Right panel.** The distribution of optimality scores for random codes drawn from an ensemble of type A.

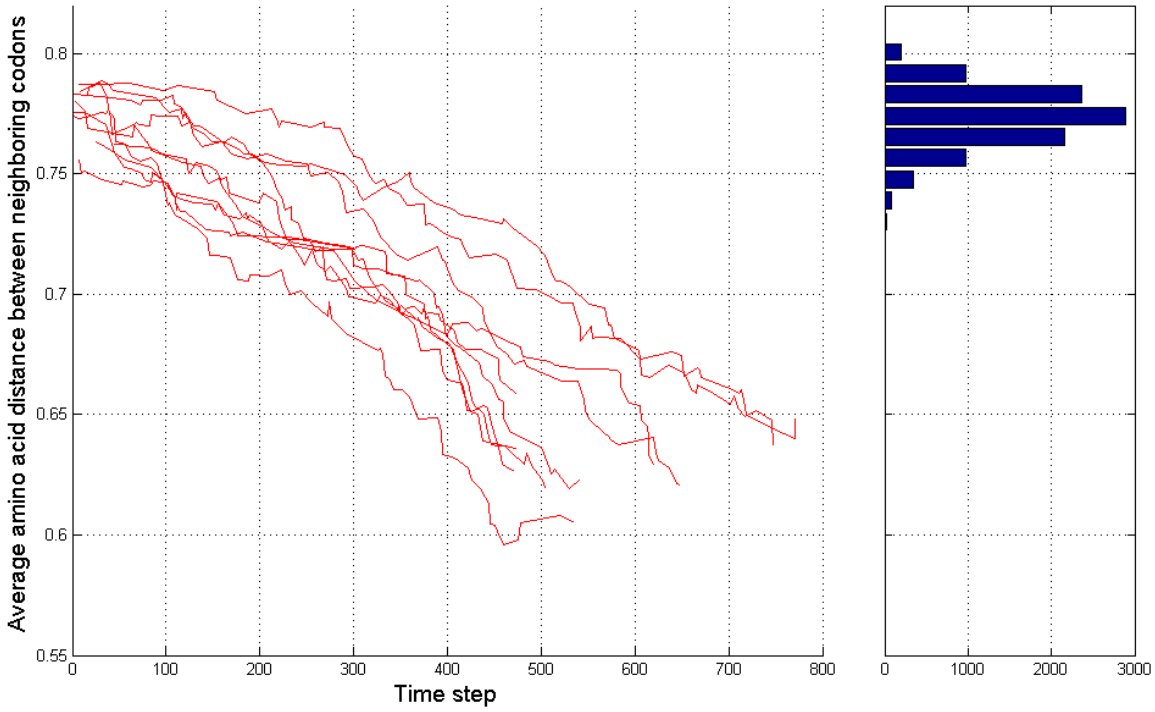

Figure 5.6: Evolution towards optimality of 10 initially random codes. Only results for the model with variable tRNA expression levels are shown, because with fixed tRNA levels all codes are frozen. Parameters: $\mu = 10^{-4}$, $\nu = 10^{-3}$, $B = 1/60$, $\Gamma = 1$. The proteome structure is as described in the text.

(fitness equal to zero) is singular within the continuous description we are using. The particular proteome structure used in Figure 5.6 was generated as follows: Two matrices are generated using equation 5.4 with $\phi = 0.5$ and $\phi = 0.99$. There are 8 times as many site types as amino acids, and each of the above matrices is used at half of the sites. At each site 5 amino acids are considered nonlethal, and the values of $W_{s\alpha}$ are taken from the corresponding matrix. We set $W_{s\alpha} = 10^{-3}$ for the rest of the amino acids at the site. As shown on Figure 5.6, under those conditions the model with fixed tRNA expression levels is frozen as we would expect from the reasoning behind the frozen accident hypothesis but the model with variable expression levels evolves to achieve high levels of optimality.

5.6 Understanding code's evolution.

In this section I will interpret the numerical results and the role of variable tRNA expression levels in the context of previously suggested mechanisms.

A codon reassignment can be beneficial if the net disadvantage of the amino acid substitutions it causes is overcompensated by the improved translational accuracy of all neighboring codons. How is this possible, though, if the improvement of translational accuracy is weaker than the disadvantage of the amino acid substitutions by a factor of the order of the probability of mistranslation? The answer is - highly uneven codon usage - the codon to be reassigned should be weakly used. The required degree of unevenness depends on the nature of the proteome. In the extreme, if most amino acid substitutions are lethal, codon disappearance is required.

Therefore, the code is frozen only as long there are no strong enough mechanisms for generating uneven codon usage. The suggested mechanisms are: mutational bias, selection on the nucleotide composition at DNA and mRNA level, Sella and Ardell's mechanism (which I interpret as symmetry breaking of the fitness of synonymous codons due to their different neighbors), spontaneous emergence of hierarchical tRNA abundance and codon usage distributions in the translational ecosystem.

The barriers to the code optimization can be surmounted from within - *self-catalyzed optimization*, or through external perturbations - *code shaking*. This is my attempt for synthesis of the previous advances. Self-catalyzed optimization was discovered by Sella and Ardell - the genetic code determines the codon usage and codon usage guides the evolution of the code. Code shaking is exemplified by Szathmary's work [105], which, in turn, builds upon the work of Osawa and Jukes [84, 97, 98]. Below we discuss how the tRNA pool codon usage coevolution *amplifies* both of these mechanisms. The numerical results above demonstrate that, after accounting for the variability of tRNA levels, self-catalyzed optimization is feasible, even if at many sites of the proteome the amino acids substitutions are lethal. Investigation of code shaking requires consideration of noise and, perhaps, mutators as discussed below. The discussion on code shaking provides a direction for extending the current research.

5.6.1 Self-catalyzed optimization

Different synonymous codons have different neighbors. Different neighbors result in different fitness effects of mistranslation and different mutational loads. Correspondingly, codons with neighbors that code for similar amino acids, i.e. codons that are more optimized, are overrepresented in the genome. This facilitates the reassignment of the less optimized ones. The beauty of the mechanism proposed by Sella and Ardell is that the codon that is disfavored and reassigned is precisely the less optimized one. In this way, the code guides its own optimization. The strength of the effect naturally, increases with the mistranslation rate.

So how does the Sella and Ardell mechanism overcome the lethality barrier expressed by Crick? It doesn't. The logic outlined above ignores two reasonable biological assumptions about the proteome structure. First, there are many more site types than amino acids, i.e. the same amino acid can be used in different contexts. At some sites polarity matters, at others - size. Some sites are constrained by the local structure that they support, others by the substrate specificity or catalytic activity that they provide. Second, some sites are incompatible with certain amino acids.

Why is the combination of the two assumptions above frustrating for the code optimization with Sella and Ardell's dynamics, i.e. dynamics with fixed tRNA levels? Consider the reassignment of one of two synonymous codons that code for an amino acid α. This reassignment is favored on average but there are some sites which specifically require α. These might be active sites of proteins, for example. At such sites, the only aspect of the codon neighborhood that matters is the number of neighboring α's. Since there are two synonymous codons in our example, the neighborhoods are identical and each codon will be used at approximately half of the sensitive sites. Since every substitution of α at those sites is lethal, the change of the code is frustrated despite the fact that it is favored at many other sites.

What difference does the variability of the tRNA expression levels make? Figure 4.12 in the previous chapter describes exactly the situation discussed. The very sensitivity of these sites induces dramatic symmetry breaking. Even if the sensitive sites are a small fraction, the symmetry breaking is strong as we saw on Figure 4.14. Moreover, the fewer they are - a smaller degree of symmetry breaking will lead to virtual disappearance of one of the synonymous codons at those sites.

Apart from giving some intuition about why the dynamics with variable tRNA expression levels

leads to better optimization, the argument just given illustrates the general point that there are properties of the proteome structure, relevant to the code's evolution, that cannot be encapsulated in an amino acid similarity matrix. This is because for each amino acid pair the fitness cost of a substitution is different at the different site types, and while the optimality of the code is governed by the average values of this distribution, the barriers to the code's evolution are influenced also by its extreme ones (lethality, for example). This point should be addressed in the next generation of statistical studies of the genetic code.

5.6.2 Mutational asymmetry and *code shaking*

Here, I discuss mechanisms through which fluctuating external perturbations can help the optimization of the code. It has been suggested that mutational biases lead to codon capture and subsequence changes of the code [84]. To turn this into a self-contained mechanism for the macroevolution of the genetic code, rather than an explanation of a few recently derived variants, we need to assume further that the mutational biases have been shifting from one extreme to another Different extremes would enable different beneficial code changes to take place. What might be the reason for the extreme nature of the shifts?

Let's look at how exactly the mutational bias translates into uneven codon usage. Consider two synonymous codons 1 and 2 with codon usage u_1 and u_2. Let μ_{12} be the probability that codon 1 mutates into codon 2, and μ_{21} - the reverse. In equilibrium we have

$$\frac{u_2}{u_1} = \frac{\mu_{12}}{\mu_{21}} \ . \tag{5.5}$$

Note that mutational asymmetry is different from mutational bias, as used in *transition-transversion bias*. Two codons different by a transition would be no more unevenly used than two codons different by a transversion (even though the first system would equilibrate faster after a perturbation). We see that extreme codon bias leading to codon capture would require extreme mutational bias and it is not clear what are the environmental forces that would drive it. Biochemical asymmetries between nucleotides are not expected to shift their direction.

One possibility, mentioned in the previous chapter is that selection on the speed and accuracy of replication or transcription leads to bi-stability in the AT-GC composition and occasional tran-

sitions between the stable states. In passing, I introduce the possibility that the *mutational biases themselves coevolve* with the code, the codon usage and the tRNA levels. This might be because mutations during replication are actively suppressed, fist through proofreading and then through mismatch repair. If different units are responsible for detection and repair of different types of mismatches then there will be intrinsic tradeoffs just as in template-directed synthesis. In this case we would have *template-directed repair*. Finally, the main player of this chapter - the coevolution between tRNA levels and codon usage might be responsible.

Small shifts in the mutational biases will be amplified by the tRNA codon usage coevolutionary dynamics. A physics analogy is that *in equilibrium* small external magnetic fields determine the direction of the spontaneous magnetization of a ferromagnet. So the picture is that small external fluctuations of mutational biases are amplified by the tRNA coevolution. On the flip side, hysteresis will prevent flipping. To resolve this problem, I suggest that the flipping is mediated by high temperature, i.e. high mutation rate intermediates. Temporary *mutator phenotypes* might be important for the evolution of the code. Actually, even without external biases, each normal-mutator-normal phenotype cycle can induce flipping of the biases.

5.7 Future directions

I showed that the evolution of the code is possible even in a world dominated by proteins and cells with mature translational systems. This invites the question: has the evolution of the code stopped and if yes why? The answer implicit in the above simulations is that the evolution stopped because it reached a local optimum. Another possibility is that there was a gradual or sudden quantitative change in some of the parameters affecting the process. In particular, as confirmed by the above numerical simulations, the stronger the mistranslation rate, the greater is the incentive to optimize. Perhaps the invention of translational proofreading marked the transition from frequent to unfrequent changes of the code. Perhaps it was the statistical structure of the proteome, to which the dynamics is sensitive, that evolved to inhibit changes. Genome sizes and mutational rates also change. As we saw this affects the symmetry breaking coming from selection on speed. Perhaps it is the use of wobble pairing and tRNA base pair modifications that provides the stability of the code. Investigating this would require relaxing the constraint of one to one correspondence between

tRNAs and codons in the simulations. In the next chapter we will discuss the role of HGT. Perhaps the end of the large scale code's evolution is correlated with the Darwinian transition proposed by Carl Woese [9].

Another interesting question is: Are traces of the initial conditions preserved during the optimization process? My suspicion is that there are dynamical regimes in which this is so.

5.8 Conclusions

Codes can evolve easily despite barriers to their evolution. The driving force is the pressure to optimize the code to reduce fitness effects of mistranslations. tRNA expression level degrees of freedom are the lubricant.

The tRNA expression framework, presented here, incorporates the most attractive features of the previous suggestions concerning the code's evolution. Codon capture emphasizes uneven codon usage, the ambiguous intermediate emphasizes that the code is a continuous probabilistic map, genome streamlining emphasizes that the tRNAs cost something. Finally, code codon usage coevolution emphasizes the ability of the dynamics to self-catalyze its own optimization, despite of some apparent barriers.

On a more philosophical level, the perceived barrier to the evolution of the code stems from its assumed discreteness. Considering tRNA expression levels and mistranslation errors coming from the adaptor competition is one way to explicitly recognize that the code (defined as a probabilistic mapping between codons and amino acids) is a continuous object after all. Small enough changes are, then, non-lethal, and, after each change, the genome, i.e. the codon usage, can equilibrate to the new code.

Chapter 6

Horizontal gene transfer and the evolution of the code

6.1 Introduction

The universal genetic code has become one of the symbols of common descent. Yet it is somewhat at odds with the other symbol - the tree of life. Organismal features are not expected to be universal, especially complex ones; they are supposed to diversify along the branches of the tree of life and display homologies. The universality of the genetic code stands out above the sea of sequence and structural homologies, and demands an explanation.

Crick [81] suggested that what makes the code special is that every change of it is lethal since it leads to many simultaneous amino acid substitutions. At the same time, it is clear that the code did not emerge all at once in its final form, especially in light of evidence that it is optimized. Therefore, it must have evolved before it froze. The question of universality then translates into: Why did the last universal common ancestor (LUCA) have an optimized, (almost) frozen code ?

It is an extraordinary problem that requires an extraordinary solution. Crick himself together with Orgel [106] introduced the possibility that the organisms on Earth were deliberately transmitted to it (among many other targeted planets) by an advanced civilization. Here is a passage from their paper:

> Several orthodox explanations of the universality of the genetic code can be suggested, but none is generally accepted to be completely convincing. It is a little surprising that organisms with somewhat different codes do not coexist. The universality of the code

> follows naturally from an "infective" theory of the origins of life. Life on earth would represent a clone derived from a single extraterrestrial organism. Even if many codes were represented at the primary site where life began, only a single one might have operated in the organisms used to infect the earth.

Keeping *directed panspermia* in mind[1], in this chapter we will entertain the alternative that the genetic code evolved here on Earth.

The question above, "Why did the common ancestor have a frozen code?" is not only extraordinary challenging but also misleading. What is missed is the possibility that all codes are the same for reasons other than common ancestry. It has been missed because, until very recently, the *literal* understanding of common ancestry was implicitly extrapolated to early life. Carl Woese started a conceptual shift [9] by suggesting that what constitutes the root of the tree of life is not an individual cell - the last universal common ancestor, but a transition from predominantly horizontal to predominantly vertical evolution. Early evolution was communal, relying on the genetic exchange of simple modular components. Gradually, coevolving networks of molecules were losing their modularity and freezing in different lineages.

The purpose of this chapter is to investigate the proposition that genetic exchange dominating the early evolution of life naturally leads to a common genetic code for all organisms, while promoting their incredible diversity in all other aspects. I present three possible channels through which HGT brings universality - communal advantage of popular codes, HGT of translational components and HGT of protein coding regions. I will start by systematizing different possible explanations of universality that do not involve HGT. Then, I will discuss the three channels of HGT and the ways in which they reinforce each other. I describe an evolutionary scenario in which organisms evolve towards optimality in concert, maintaining not only similar genetic codes but compatible translational machineries allowing exchange of translational components. Finally, I model aspects of the different mechanisms in order to confirm their plausibility and expose their assumptions, hopefully stimulating further research.

[1]It is an interesting idea, who knows what humankind will be up to in several thousand years. Perhaps a group of rogue individuals will start shipping cocktails of microorganisms to foreign planets from some dark corner of the solar system despite a bill prohibiting the contamination of the galaxy (which in turn was passed after pressure from "Green Peace")

6.2 Scenarios for frozen LUCA without an active role for HGT

6.2.1 Freezing before diversification

One possible explanation of frozen LUCA, hinted by Crick [81] and Wong [107], is that translation emerged, the code optimized itself and froze in a single niche and only then some other evolutionary transition triggered the spread and diversification of its organisms. The code is universal by virtue of competition and genetic drift between organisms occupying this single niche. Or, perhaps, in a localized ecosystem, there was more than one frozen code, but it was a single lineage that diversified, conquering a world that was either unoccupied or inhabited by inferior organisms lacking translation. The original ecosystem was marginalized and the rest of the codes were stochastically lost during the expansion. So, the universality of the code is a result of a large scale "founder" effect.

The problem with this *freezing before diversification* scenario is that it does not explain what was stopping the expansion of organisms endowed with some form of translation well before the genetic code froze. This is especially puzzling, since the evolution of translation and the refinement of the genetic code was most likely a multi stage process that took an extended period of time. A possibility is that the code(s) froze much more quickly than it took for the protein universe to evolve sufficiently so as to allow members of the localized/specialized community to invade the world. Or perhaps the trigger to expansion was a change in the climate, geology, atmosphere composition, etc. of the earth.

The alternative to *freezing before diversification* is *freezing after diversification*, and in the rest we will be concerned with this seemingly more plausible scenario. Even without understanding how codes evolve, it is reasonable to assume that they eventually freeze - perhaps because they reached a local optimum, perhaps because the translational machinery or the proteome became highly complex and coevolved. Whatever the reason, let's look at the first moment of time when all organisms have frozen codes. The codes diversified along with all other organism properties, so we have many somewhat different codes and perhaps even some organisms that do not have translation at all. How do we get from this situation to a universal code? We now outline a variety of logically allowed possibilities, explaining their merits and drawbacks.

6.2.2 Accidental code universality

The universal genetic code is a simple consequence of the fact that after an extended period of time, all organisms will be descendants of a single organism and will inherit its genetic code. This happens for stochastic reasons: the descendants of an organism can invade a neighboring niche, and by mere chance outcompete the organisms that were already there. Since the entire phenotype space is connected, the repeated stochastic takeover of neighboring niches will result eventually in a universe inhabited by organisms with a common ancestor. The problem here is that because of the stochastic nature of the process it will be extremely slow. In addition the phenotype space is constantly growing and it is not clear whether this slow stochastic takeover process can ever saturate the phenotype space.

6.2.3 Evolutionary transition following translation

A universally beneficial cellular property that emerged following the maturation of translation caused a giant selective sweep overcoming the preexisting adaptation of different organisms to different environments. A candidate for such an event is the replacement of RNA by DNA as a carrier of the genetic information. This is an attractive possibility. One of the problems is that by the time of maturation of translation we might already have different cell designs with different ecological roles and it is not immediately clear how even such a great innovation can overcome this. In contrast, if the code was already universal due to the communal evolution, the proteins involved in DNA synthesis could have been easily distributed.

6.2.4 Selection on optimality

Once proteins were established as a major determinant of the phenotype, the quality of the genetic code was the single most important contributor to fitness. The competition between different organisms, coarse-grained over a large period of time, was basically a competition between different genetic codes. The more optimized genetic codes were out competing less optimized ones. Eventually, a very highly optimized one won. Related is the suggestion that optimality is linked to evolvability [108], and therefore, organisms with more optimal codes are evolutionary more successful. The problem here is that it is not a priori clear that we can ignore the diversity and competition

along all the other phenotypic dimensions. Moreover, the more optimized the codes are the less will the differences matter. So we would expect after all a diversification of the genetic codes.

6.3 The three roles of HGT

Our interest here is the role of HGT for the evolution of the genetic code. The natural hypothesis adopted in the previous chapter is that the genetic code was optimized after translation emerged. The extreme alternative to this is that the correspondence between trinucleotides and amino acids was firmly established before the emergence of template-directed synthesis of proteins. In particular, it was suggested that tRNAs were *handles* for *decorating* ribozymes with amino acids [109, 110]. As today, tRNAs were specifically charged, and there existed a protocol for binding which included the genetic code, plus complementarity and compatibility between the anticodon regions of the tRNAs and the local structures of the target sites on the ribozymes.

Little of what I say in this section depends on the nature of template-directed synthesis per se. In the tRNA handles world there is an incentive to optimize the code because attaching the right amino acids to the right places is beneficial. There is a barrier to code change because every tRNA species decorates many different sites in many different ribozymes. Ribozymes that are decorated to become functional cannot be easily transferred between organisms having different codes, and are the analog of protein coding regions.

Therefore, the arguments presented here are rather generic and insensitive to many unknown details. What is assumed, of course, is that whatever was the phase in which the code emerged, there were vectors, or alternatively, lack of evolved barriers to genetic exchange. This said, I will use the language of translation throughout.

6.3.1 Popularity contest

One of the advantages of communal evolution is that universally good traits can spread through HGT to organisms occupying different niches, preserving their diversity. In a protein dominated world most of the innovations will involve proteins, and correspondingly HGT will be most effective between organisms having the same genetic code. In this way, the organisms are partitioned into communities having different/incompatible genetic codes. A single code community can span cells

adapted to different niches and with different organization. Now we look at the evolutionary dynamics of these communities.

The larger the community and diversity of organisms sharing the same code is, the larger the pool of protein innovations accessible to everyone is, the faster they evolve, and therefore a greater potential they have to invade niches occupied by organisms with different incompatible genetic codes. With this dynamics larger communities will tend to become larger at the expense of smaller ones. The only stable solution is a universal genetic code. Thus, it is not better genetic codes that give an advantage but more common ones.

The elementary step in this process is the overtaking of an occupied niche by the descendants of an organism with a different genetic code. If two groups of organisms compete with each other, the one that has access to more innovations (the one belonging to the larger community of common/compatible genetic codes) will on average out compete the other. Unlike the arguments above, there is an active feedback loop driven by innovations sharing through HGT, which not only singles out the genetic code from all other properties a cell, but also provides a mechanism that drives the preferential extinction. The emphasis here is that universality is the single stable solution of the community dynamics.

A human metaphor for this process is the dynamics of languages. There were many different tribes speaking thousands of different languages. For reasons unrelated to the quality of the languages some become more popular. The less popular ones disappear at the expense of the more popular ones. In these days, most of the people speak only a few languages. A few hundred years from now a universal language might emerge. The universality is already achieved for numbers - positional base ten encoding won. Moreover, the ten symbols required for this encoding are universal.

6.3.2 HGT of translational components

The genetic code is a family of modules - it is specified by tRNAs and charging enzymes (aminoacyl-tRNA synthetases in a modern day setting). Moreover, its modularity is universal - all organisms have the same types of modules. The task of improving translation and the code is also universal, i.e. largely insensitive to the niches organisms are occupying. The evolutionary instantaneous

spread of antibiotic resistance is an example of the power of HGT to distribute simple modular components in response to universal challenges. So is it possible then that *HGT of translational components* played an important role in the codes evolution and what might that role be?

A most relevant fact is that the translational mechanism is far more conserved than the transcriptional and replication mechanisms among the three domains of life [9, 111] . This has been interpreted as evidence that translation maturated first [9, 111]. An alternative, consistent with the scenario of this section is that the translational machinery was kept universal by the communal evolution, due to its special role in determining the universal communication protocol, while the other cellular mechanisms naturally diversified.

I will separate the argument into three parts. First, I will discuss mechanisms for maintaining compatibility, then I will outline the effect of this compatibility for the universality of the code, and finally, I will argue that the universality of the code couples back as a cause for maintenance of the compatibility for the code specificators[2].

To do justice to Thomas Kuhn [112], all the arguments here are transfers of common sense observations about today's fast-pace technological world in which we are embedded to the subject of the genetic code. Similarly to today, the genetic code evolved in an era of rapid innovations following an evolutionary transition. Because of HGT, the innovations were distributed to cells in different niches limited only by the compatibilities of codes and functions. Let's just look around; universal protocols are everywhere. I am typing on a standard qwerty keyboard, and by pressing the universal (in the Windows world) Ctrl+S I am saving my files to a standard 3.5" hard drive via a Universal Serial Bus using an almost universal file protocol; the soda can is universal up to decorations, the tea bag (tBag) has a tread and a paper label at the other end - like all the other tBag species. The list can be continued for a long time even if I constrain myself to things that are in front of my eyes. Incidently, none of these standards has emerged through common descent, even if the presence of similar components in different systems *mimics* common descent of the systems. In particular, the presence of similar tRNAs in all cells is consistent with, but does not argue for, common descent of the cells, in the way it argues for a common descent of the tRNAs. To conclude, the evolution of technology in the human society is, as a whole, very different from the evolution

[2]The structure of this section is a historical accident reflecting the evolution of my thinking as I was writing it.

of cells, but certain aspects can be mapped to one another, and I am using this as a source of metaphors.

Generic mechanisms enforcing compatibility

A central premise of this subsection is that the *code specificators* coming from different organisms have been compatible and interchangeable during the evolution of the code. Compatibility is unhampered by differences in the genetic codes as long as the function of the modular *code specificators* and other ribosomal RNA components is encoded in the universal language of RNA biochemistry. Compatibility is opposed by the coevolution between tRNAs and the ribosome, between different tRNA species in the same organism, between charging enzymes and their cognate tRNAs. Lineage specific coevolution eventually won over compatibility. But there are mechanisms that actively opposed coevolution and the balance might have tipped away from compatibility only at a later stage. Below we discuss a few suggestions that are independent of the special role of the translational machinery as a determinant of the universal communication protocol. Later in this section we argue that this special role makes an even stronger case for maintenance of compatibility.

The type of coevolution with which we are concerned is the change of the interface between different modules within a cell. We assume that all cells have a given set of modules and that modules can be efficiently transferred between cells if they are compatible and beneficial. On Figure 6.1.I we have two modules A and B that interact. The ribosome is a special module that interacts serially with an entire family of other modules - the tRNAs, Figure 6.1.II. Sometimes, one of the modules changes in a direction compatible with the old protocol but improving it, Figure 6.2. Such changes can be efficiently distributed through HGT to all organisms and as a result the universality of the protocol is preserved. This is an example of a module upgrade. Figure 6.2.II assumes that the relevant part of the ribosome can be transferred. The ancient ribosome perhaps consisted of only a few RNA molecules, which or parts of which could be transferred and recombined.

Coevolution often requires passing through an intermediate stage of lower fitness as shown on Figure 6.3. In a large population, the effective route to overcoming the barrier is to hitchhike on another beneficial trait. However, HGT can effectively and beneficially decouple the slightly dele-

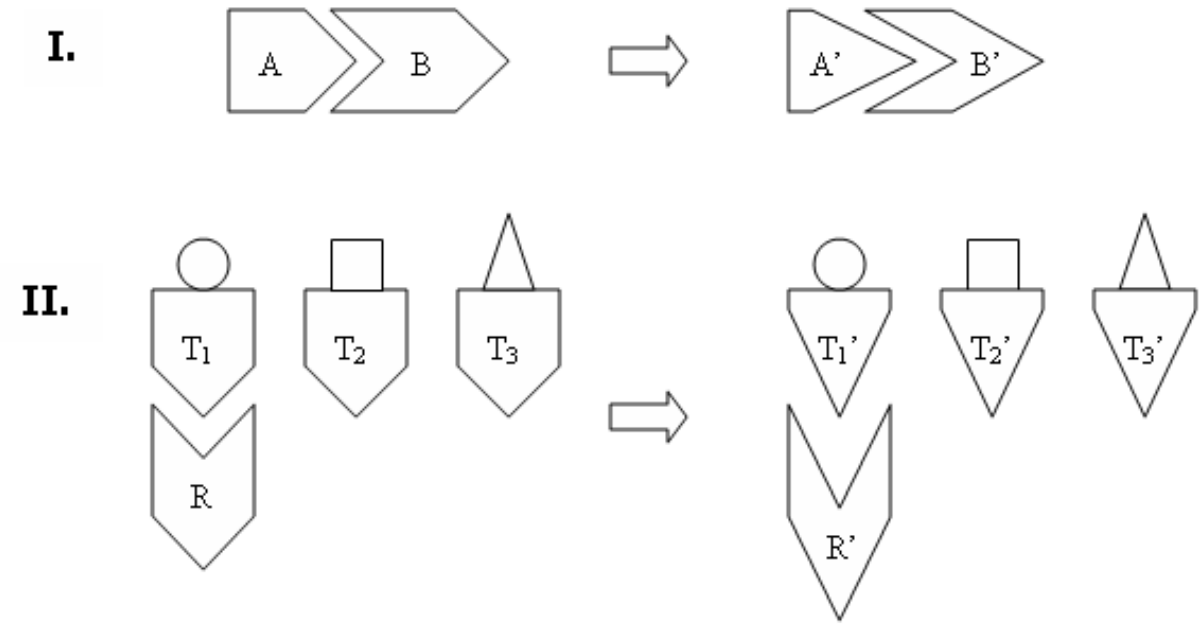

Figure 6.1: Schematic representation of module coevolution. **I.** The simplest case of two modules. **II.** Coevolution when a hub module, R, interfaces with a family of modules, $\{T_1, T_2, T_3\}$. R can be thought of as a ribosome, and T_1, T_2 and T_3 as tRNAs.

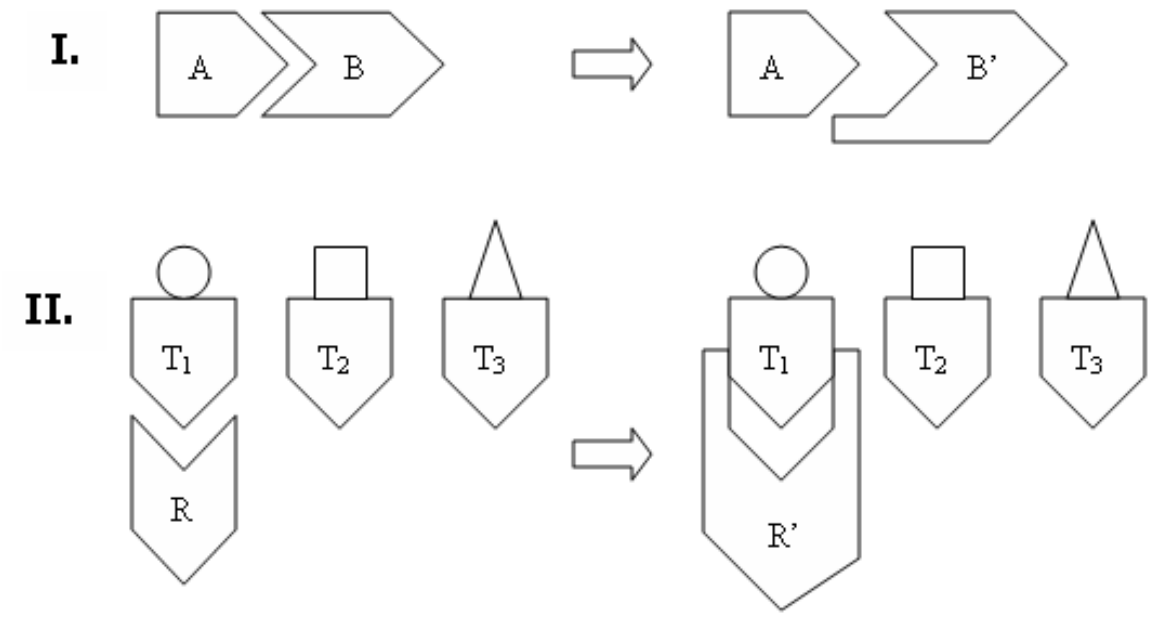

Figure 6.2: Compatible protocol improvements can spread through HGT. **I.** Two modules improve their interface. **II.** An improvement coming from a change of a hub module.

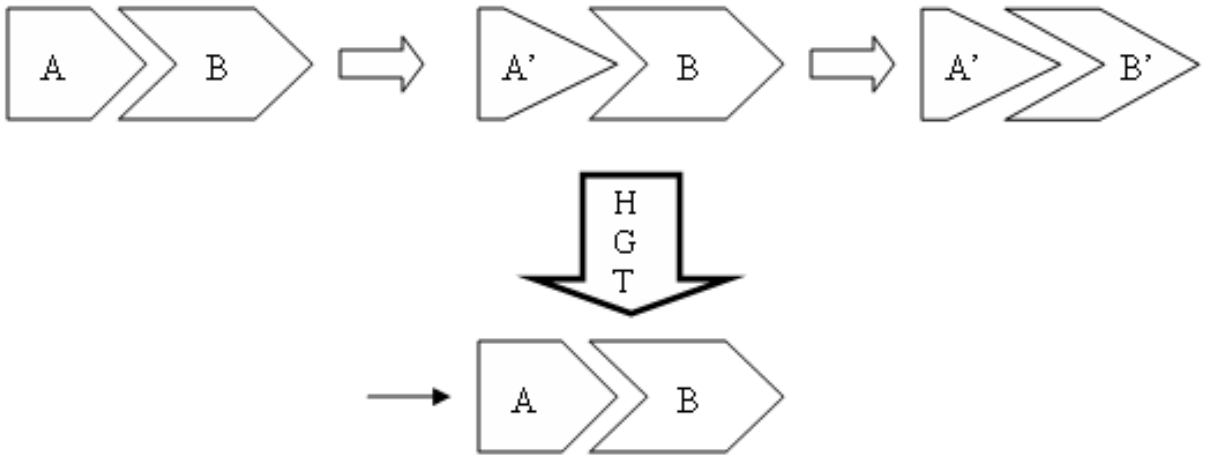

Figure 6.3: A kinetic barrier to coevolution coming from HGT. The slightly deleterious intermediate state is most likely reversed through HGT.

terious intermediate state from the beneficial trait. Moreover, because of the universal modularity, even if the intermediate state invades the population, it can be reversed via a beneficial HGT of the changed module. If the rate of attempts for module exchange is large compared to the rate of spontaneous beneficial module changes, we have an efficient *kinetic barrier to coevolution*.

Compatibility is reinforced by the distribution of module upgrades. There is certainly a benefit in , and therefore a natural tendency for, coevolving the existing components. But this benefit might be small compared to the benefit of modernizing the components themselves. Here is an example of modernization outweighing coevolution. I enter into all kinds of coevolutionary interactions with my laptop by customizing bookmarks, installing the software I need, uninstalling programs that came for free but never use, indexing the harddrive for faster search, adjusting the options of many programs, etc. But these benefits are outweighed by the benefit of getting a faster computer with

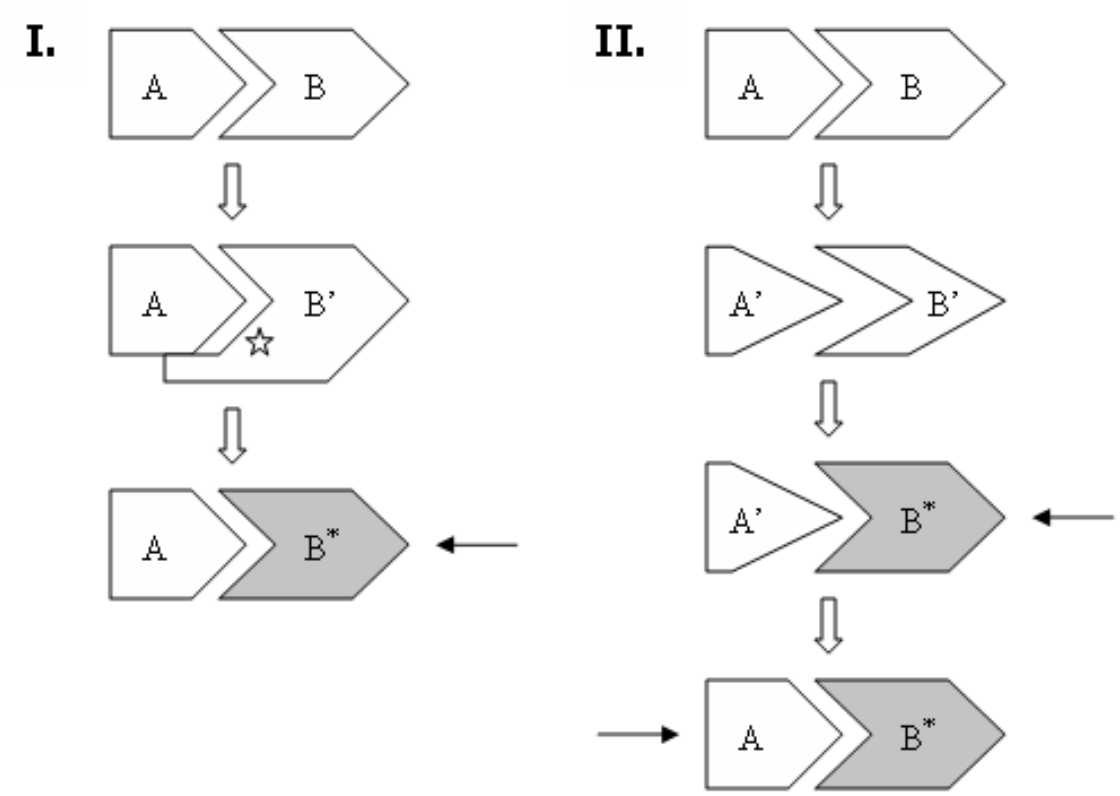

Figure 6.4: Reversal of coevolution through a module upgrade. **I.** Reversal of a non-universal improvement through a model upgrade. **II.** Reversal of a coevolved state through a module upgrade. The disadvantage of the intermediate state is overcome by the beneficial effect of the imported module.

more memory, a longer battery life, wireless and USB 2.0 (instead of 1.0), even if my doing so will lead to a loss of these customizations. Note that there is no need of second order selection here - getting a new laptop is of *immediate* benefit to me. On Figure 6.4.I, there is a protocol improvement idiosyncratic (as denoted by the big star) to the cell in which it is made. This improvement is sacrificed for the sake of obtaining an upgrade of module B. The improvement reinforces the standard protocol. On Figure 6.4.II an isolated coevolutionary event is shown that is reversed via an import of an upgraded component that emerged elsewhere. The upgrade is assumed beneficial even if it does not fit perfectly. Similarly to Figure 6.3, the intermediate state is then resolved in favor of the standard protocol.

Devices which function requires to interface with many different modules at different times are conservative with respect to protocol changes, Figure 6.5. The job of the ribosome is to interface reliably with all tRNA species present in the cell. Coevolutionary deviation from the protocol between the ribosome and one tRNA species will have negative consequences on the interactions of the ribosome with the other tRNAs. The barriers to coevolution increase with the number of molecules involved.

The interface is reinforced, not only by upgrades of existing modules but by import of new modules with new functionality. Imports of tRNAs that change the code reinforce the protocol between tRNAs and the ribosome. Stretching this idea, the loss of global modularity of the translational machinery might be intrinsically correlated with the global freezing of the genetic code. Below we argue for this in a much more convincing way.

Different aspects from above reinforce each other. For example, if coevolution between tRNAs and the ribosome is slow because of the many molecules involved, the likelihood of a significant deviation from the protocol before some upgrade is distributed is low. Similarly, the kinetic barrier is enhanced.

Effects of code specificator compatibility: the universal core

Assuming compatibility, what are the effects on the evolution of the genetic code? In the previous chapter we described the coevolution between the genetic code, the codon usage and the tRNA pool. Now we have a pool of tRNA pools, genetic codes and codon usages. By virtue of the weak

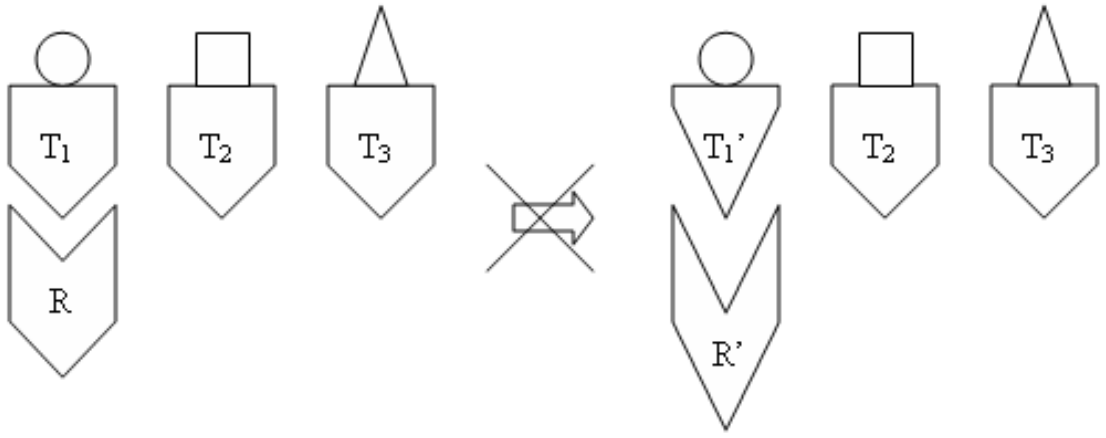

Figure 6.5: Coevolution is difficult when many modules use the same protocol. The coevolution between R and T_1 is disadvantageous for the interactions of R with T_2 and T_3.

dependence of this system on the specifics of different niches, we can study (important aspects of) the communal evolution in isolation.

The tRNA pool of an organism can change as a result of a tRNA speciation within it or by import of a foreign tRNA. There will be different modes of the collective dynamics depending on the relative rates of discovery of tRNA innovations and spread through HGT. One reason for this is that the rates ratio determines the extent to which the exchange of code specificators reduces the phase space of readily accessible code changes. If the probability to invent a code change is small, the readily accessible code changes are those that can be imported. A reduced menu of code change options facilitates the maintenance and achievement of code universality.

Imagine for simplicity the situation where organisms occupying diverse niches have the same non-frozen genetic code. The discovery of a tRNA modification that changes the code and increases its optimality (and therefore the efficiency of translation) in one organism will also be beneficial for organisms in the other niches, due to the universal benefit of optimality. Therefore, a spread of the discovery is beneficial to all recipients. As we mentioned above such a spread can be very rapid. It does not need not be passive either. Think, for example, of the benefit a lysogenic virus (such as lambda phage) would have if it carries with itself the beneficial code specificator. By incorporating itself in the genome it increases the fitness of its vessel and therefore its own replication, while still reserving the option to multiply and flee if the cell starts to die. Therefore, if the spread through HGT is rapid compared to innovations, the code will stay nearly universal while optimizing itself.

So we have the possibility that a core of organisms having the same genetic code maintains its integrity while evolving towards optimality. In the absence of an attractive force that pulls back deviant codes, the core will be decaying. If the decay is slow enough, the deviants will be at a communal disadvantage and disappear as described in section 6.3.1. The core of universality will be like a comet leaving behind garbage that disappears from the evolutionary radar. Better still, the decay will be overcompensated by an expansion of the core due to the benefit of popularity. Notice that what I say below does not assume that the core encompasses the vast majority of the organisms, or that there is only one core. If there are several cores, they evolve independently, and compete through the popularity contest mechanism. In the simulations below we will see even "speciations" of cores.

The universal core enforces compatibility

Lineages that leave the translational compatibility protocol of the core of universal codes, will eventually also leave the core itself, given that the core evolution towards optimality is coordinated via HGT of core compatible code specificators. As in the previous paragraph, leaving the core of genetic codes, leads to evolutional marginalization. Therefore, the universality of the genetic code is interlocked with the universality of the translational protocol. The core of universal codes significantly overlaps (perhaps coincides) with the core of compatible code specificators , and the intersection constitutes the real core. All organisms outside it are *living dead* (the tail of the comet).

Rephrasing the above, given that there is a core of organisms having the same genetic code and compatible translational machineries, and that the common genetic code still undergoes concerted evolution (towards optimality), the compatibility of the code specificators is enforced by the special role of the genetic code as an exchange protocol for innovations. This is because the concerted optimization proceeds through exchange of code specificators. If you cannot receive the distributed upgrade of the code, you leave the cluster of popular codes, and get marginalized through the popularity contest mechanism.

An important corollary is that while the core is being optimized, the compatibility of code specificators is enforced. Once the optimization of the genetic code is complete, there is no pressure to maintain compatibility. Therefore the "freezing" of the universal genetic code might coincide with the radiation of the translational machineries. So, even if translation emerged earlier than the other basic cellular systems, but the optimization of the code took an extended time, it diversified less. As mentioned above, this is consistent with the observation that the translation mechanism is more conserved evolutionary than the replication and transcription ones.

The universal core is an abstract concept in the joint space of genetic codes, code specificators, translational protocols and codon usage. This space is largely independent of the niche space. It is this independence that predisposes for universality. The emergence of the universal core is predicated on the ease of distribution of novelty compared with the spontaneous emergence of novelty.

6.3.3 Code attraction due to HGT of protein coding regions

A universal genetic code enables horizontal gene transfer and communal evolution of the proteomes. Is it also the other way around, i.e. does horizontal gene transfer of protein coding regions between similar but non-identical codes provide an effective attractive force? If this is so, then there is a self-enforcing feedback loop between HGT and universality.

Horizontally transferred genes can be useful for the recipient even if the donor has a (slightly) different code. The codon usage of the transferred gene is adapted to the donor code and therefore different from that of the recipient. Correspondingly, there will be pressure for the recipient code to readjust itself to make a better use of the new gene. The code change can be a change of the anticodon of a tRNA, change in tRNA expression levels, introduction of ambiguous translation, import of a tRNA from another organism, etc. The change can be favored even if the modified genetic code is less optimized with respect to the rest of the genome - the loss of optimization is compensated by the beneficial effects of the new gene. In fact, a change in the code might be a prerequisite for the utilization of a received foreign fragment. Eventually the codon and amino acid usage of the newly transferred segment will equilibrate with the rest of the genome and the indirect pressure of the donor code on the recipient code will disappear but leaving behind its accumulated effects.

More generally, because of HGT, the genomes are a mosaic of different codon usage. This is true today [28] but was perhaps even more pronounced in the era when the template generated proteins were taking over many universal housekeeping functions, and therefore innovation sharing through HGT was most important. Since codon usage guides the code evolution, each code evolves under the influence of others.

HGT requires that the genetic codes of the host and the recipient are sufficiently similar - how similar is sufficient depends on the nature of the proteins and the accuracy of the codes. There are strong reasons to believe that the more primitive the code of the donor is the greater is the genetic code distance over which HGT is possible. This is because the tolerance of the proteins to errors in their primary structure is coadapted to the error rates of the translational machinery. A cell with a non-optimal code cannot afford very capricious and therefore highly fine-tuned proteins because of the cost of discarding defective proteins. A robust to translational errors protein is also

more tolerant to a translation with a different code. On the opposite end, the less optimized the recipient code is, the more error tolerant its proteins are, and therefore the less harmful will be the effect on the old genes of a code change in the direction of the donor code. This has the important consequence that in the initial stages of the genetic code evolution when the diversification tendency of codes was strongest, HGT was possible and extensive despite the presence of many different codes.

Below we model the HGT of coding regions and discover that it not only provides an attraction force between the different codes, but also greatly enhances the communal evolution towards optimality. This means that HGT of coding regions is an effective means to overcome the barriers inherent to the code's evolution. From the point of view of the overall speed and accuracy of translation, HGT is typically non-beneficial, and represents an unstable "excited" state which can be resolved by a code change. Therefore, just as with the tRNA-codon usage interplay from the last chapter, HGT of coding regions can catalyze code changes. HGT is different from the hitchhiking of excited states of the translational system on other beneficial traits because the beneficial role of HGT cannot be decoupled from its perturbing effect.

6.3.4 Interactions

The different mechanisms enabled by HGT synergistically interact with each other. We already saw that the evolutionary expansion of the most popular cluster of codes provides the necessary support for the maintenance of the otherwise weakly decaying universal core.

The opposite is also true. The popularity contest mechanism is ineffective if there are no clusters of sufficient size on which it can operate. The establishment of such clusters is greatly facilitated by the HGT of code specificators and protein coding regions. Distribution of modules enforces modularity that in turn enforces the distribution of modules. Similarly exchange of protein coding regions enforces universality, thus making it easier to exchange genes. Therefore, there is a positive feedback loop that provides at least *local stability* to the protocols and turns them into effective degrees of freedom at a longer time scale. The *global stability* and universality is then guaranteed by the "winner" takes all nature of the popularity contest.

If only the popularity mechanism is at work we can expect that diversification wins at first in the era when codes are far from optimal and evolving fast. Only after the codes become sufficiently

optimal, and thus slowly evolving, the trend is reversed - initially slowly and then at an increasing pace as large clusters emerge. In this case the local stability of the protocols is passive, arising from the saturation of their evolution. Perhaps, this is the story of human languages - extreme diversification followed at a later stage by the evolution towards universality.

The popularity contest mechanism ultimately relies on the benefits of distribution of novelty throughout a community. The community support protects its members against invasions from members of other communities, and provides the technological edge that helps them to invade niches occupied by members of other communities. Therefore it relies, above all, on HGT of protein coding regions (or ribozymes to be decorated by the genetic code protocol if the code emerged before translation).

If you deviate from the core protocol there is an *incentive* to come back since otherwise you are a living dead. Similarly, if you receive a foreign gene from a donor with a different code, there is an incentive to change your code in the direction of the donor. However, these incentives are rendered ineffective by innovation and entropy barriers. It is unlikely that you will invent a change and that it will be of the right type. And here comes the important role of exchange of code specificators. They provide an efficient *means* to respond to your incentives by providing ready for import code changes, and in addition restrict the menu of options in an otherwise huge space of possibilities to the type of changes you need most. If you are outside the giant core the chances are that you will import from the core and therefore become more similar to the core[3]. Similarly, if you obtain a gene from someone, this means that your place in the ecosystem is such that you have access to the genetic material of the donor. Therefore, you, as a "species" have a better than random chance to obtain the right code specificator from the donor before the codon usage of a recently acquired trait ameliorates. It is the interaction between means and incentive that makes the emergence and maintenance of universality robust.

[3]Secondly derived code variants, observed today, emerge in a very different context from the code variants occurring in the early evolution, because the code specificators are not compatible.

6.4 Model

We employ the model from the last chapter with fixed tRNA expression levels. There we evolved ensembles of independent codes. Now we couple their evolution.

The popularity contest mechanism which clearly facilitates universality (and given enough time generically leads to universality) is factored out from the simulations in order to concentrate on the potential weaknesses of the other two mechanisms. Each evolving entity in the ensemble can be thought of as a different "species" (or ecotype). While within each species the evolution proceeds through invasions of code variants with higher fitness, the different species are stable and their number is fixed, thus blocking the popularity contest mechanism.

Here are the features added to the model from the previous chapter.

1. There are N entities. At each time step an entity receives a foreign genome fragment with probability $rLGT$ from a random donor.

2. The fragment is characterized by the *fraction of the recipient genome*, gw, that it represents, and the site type specific codon usage distribution of the donor $\left\{u_{s,i}^{\text{donor}}\right\}$.

3. The fragment is accepted with a probability $Pa = Pa(u^{\text{recipient}}, u^{\text{donor}}; \text{code}^{\text{recipient}}, \text{code}^{\text{donor}})$. We studied the case with *no barrier to HGT* of coding regions, i.e. $Pa = 1$, and a case with a conservative barrier defines as follows:

$$Pa = \begin{cases} 1 & \text{if } \overline{\log f_r} \geq \overline{\log f_d}, \\ \exp\left[-\left(\overline{\log f_r} - \overline{\log f_d}\right)^2 / \left(2 \text{ var}(\log f_d)\right)\right] & \text{, otherwise.} \end{cases} \tag{6.1}$$

   Here $\log f_d$ is the *distribution*, due to mistranslation, of the logarithm of the fitness scores of a gene within the donor, and $\log f_r$ is the distribution for the same gene computed with the recipient code. The distributions are approximately normal because the fitness of a protein is assumed to be the product of the scores of its sites. For the purpose of computing Pa the gene length is assumed to be 100.

4. If the HGT event is accepted, the codon usage of the recipient is updated according to the

rule

$$(1 - gw)\, u_{si} + gw\, u_{si}^{\text{donor}} \rightarrow u_{si}\ . \tag{6.2}$$

5. With probability q an entity looks for a beneficial code change by import of a code specificator from another entity. An imported code specificator *replaces* the existing code specificator for the same codon. With probability $1-q$ the entity tries to find a spontaneous change. In both cases the search continues until a beneficial change is found, or until all the possibilities are exhausted.

6.5 Results

All result presented are with ensembles of $N = 40$ entities. Initially all the codes are identical, and the initial code is generated through randomly assigning amino acids to the codons. As in the previous chapter, there are 20 amino acids and 64 codons. *Proteome structure type I* from the previous chapter is employed with different values of ϕ.

To analyze the results of the simulations we look at the *average code distance* for the ensemble of codes. The average is obtained by considering all pairs of entities with equal weight. The code distance between two entities is the Hamming distance, which counts the numbers of codons that code for different amino acids. The optimality score is calculated as before, equation 5.2.

First, we demonstrate that HGT of coding regions provides an attractive force between codes. To do so we ignore the barrier to horizontal transfers between different codes - $Pa = 1$, and consider only spontaneous code changes — $q = 0$. Figure 6.6 presents the evolution of the average distance between the codes for different *intensities* of HGT. We set $rLGT = 0.99$ and control gw.

The codes quickly diversify at first because there are many different routes towards optimality. Eventually, however, the attractive force provided by the exchange of coding regions manages to make the codes identical again, if the intensity of exchange is strong enough.

HGT of coding regions not only brings universality but greatly enhances the optimality as shown on Figures 6.7 and 6.8. This effect is pronounced for $\phi = 0.99$ and not significant for $\phi = 0.1$. Since ϕ is the scale of the fitness effect of a single amino acid substitution on the entire genome, and the genome is large, we expect ϕ to be close to 1 at a typical site. At the same time, there exist

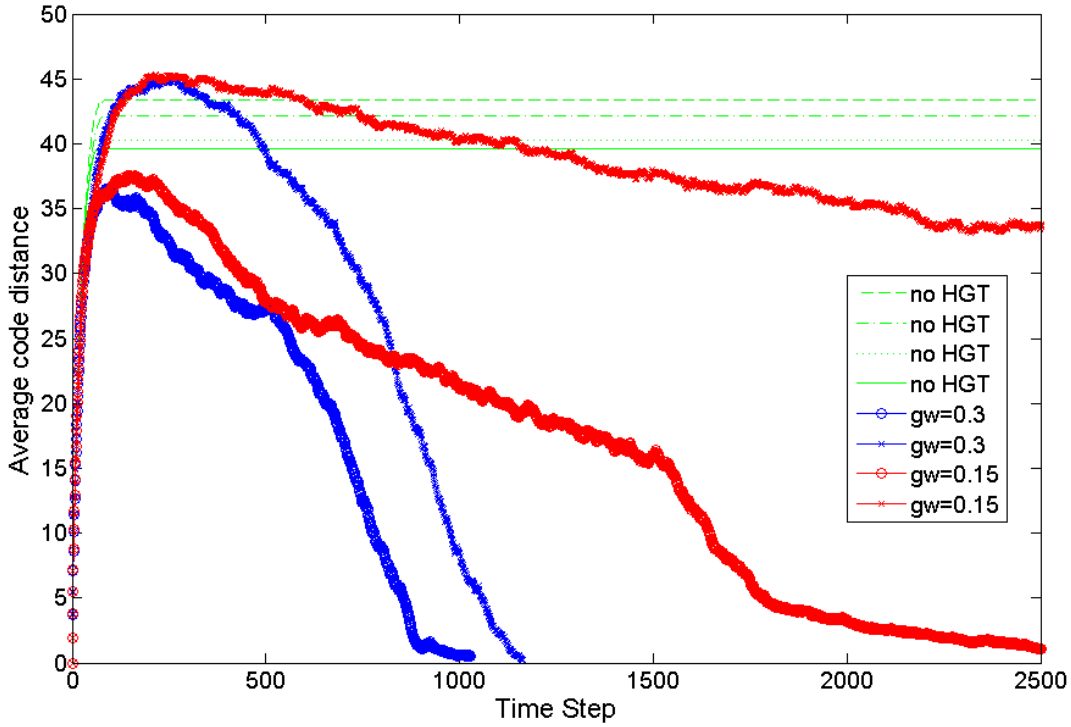

Figure 6.6: Evolution of the average Hamming distance between 40 initially identical codes at different strengths of HGT of protein coding regions. The barrier for HGT between different codes is ignored. The initial conditions are the same for all runs. Parameters: $gw = \{0, 0.15, 0.3\}$, $rLGT = 0.99$, $\phi = 0.1$, $\mu = 10^{-4}$, $\nu = 0.01$. All codes initially optimize in different directions. If HGT is present, the codes eventually converge towards universality.

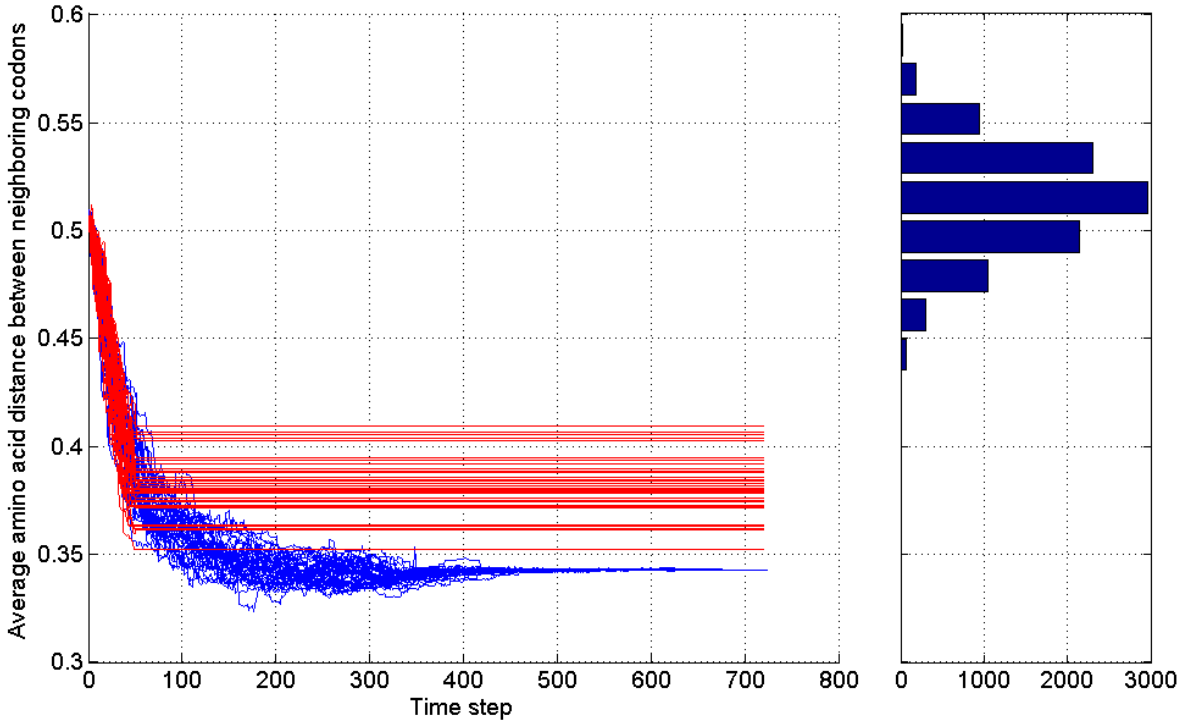

Figure 6.7: Communal evolution towards optimality of 40 codes with (blue) and without (red) HGT of coding regions. There is no barrier to HGT between different codes. The initial conditions are the same for both runs. Parameters: $q = 0$, $gw = 0.3$, $rLGT = 0.99$, $\phi = 0.99$, $\mu = 10^{-5}$, $\nu = 0.01$.

some very sensitive sites in the genomes, and as we discussed in the previous chapter *proteome structure type I* is inadequate in some respects. Therefore, a further clarification of this point will be helpful. The two figures differ in μ, the random initial code, and the randomly generated amino acid similarity matrix.

With the particular choice of initial conditions, the initial diversification tendency of the codes is too strong to be prevented by HGT. After this diversification phase, HGT between different codes is impossible if we account for the barrier of equation 6.1. Still, HGT of coding regions with the barrier synergistically interacts with the HGT of code specificators as we will see below.

Now we turn on HGT of code specificators and study the effect on the dynamics as a function of the ratio of innovations of code changes versus rate of their HGT. In addition, we contrast the cases with and without HGT of coding regions, including the barrier of equation 6.1.

Figures 6.9 and 6.10 demonstrate that for a sufficiently large q a stable core forms that evolves

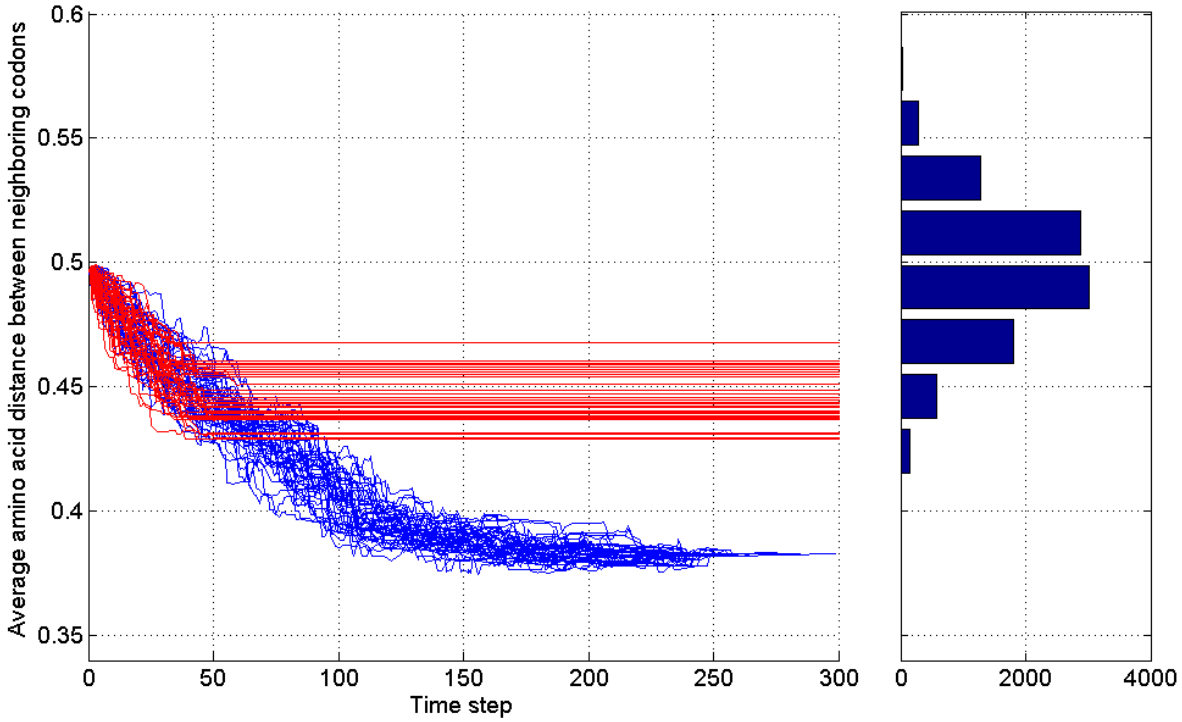

Figure 6.8: Communal evolution towards optimality of 40 codes with (blue) and without (red) HGT of coding regions. There is no barrier to HGT between different codes. The initial conditions are the same for both runs. Parameters: $q = 0$, $gw = 0.3$, $rLGT = 0.99$, $\phi = 0.99$, $\mu = 10^{-4}$, $\nu = 0.01$.

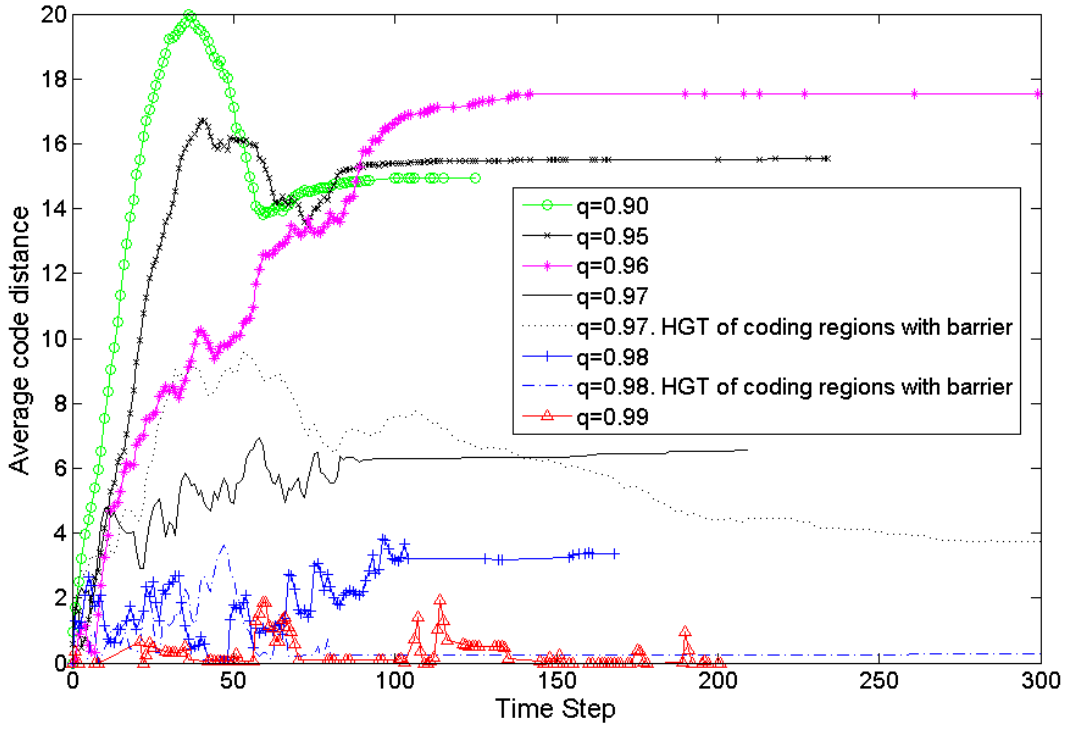

Figure 6.9: Evolution of the average Hamming distance between 40 initially identical codes at different strengths of HGT of code specificators. For two of the runs, HGT of protein coding regions with exchange barrier is turned on. The rest do not have HGT of coding regions. The initial conditions are the same for all runs. Parameters: $\phi = 0.1$, $\mu = 10^{-4}$, $\nu = 0.01$, $q = \{0.9, 0.95, 0.96, 0.97, 0.98, 0.99\}$. For the HGT of coding regions: $gw = 0.3$, $rLGT = 0.99$. At small q the codes diversify. At large q the initial uniformity is maintained, while the codes evolve in concert towards optimality. HGT of coding regions enables maintenance of universality for smaller values of q.

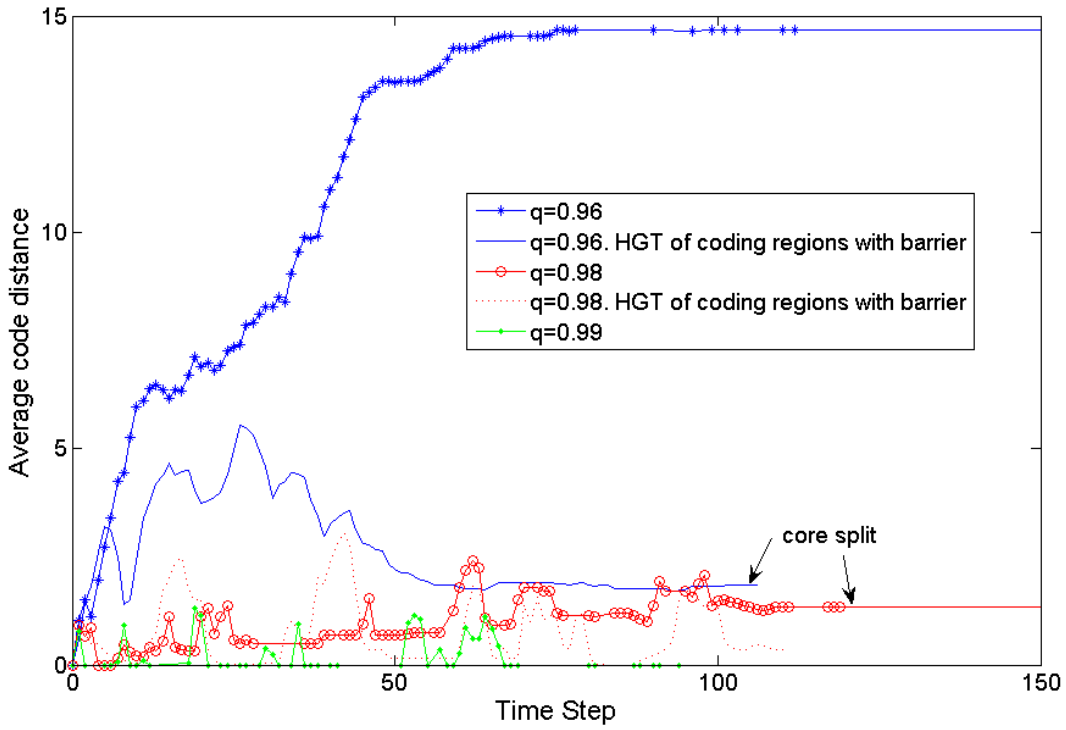

Figure 6.10: Evolution of the average Hamming distance between 40 initially identical codes at different strengths of HGT of code specificators. For two of the runs, HGT of protein coding regions with exchange barrier is turned on. The rest do not have HGT of coding regions. The initial conditions are the same for all runs. Parameters: $\phi = 0.99$, $\mu = 10^{-4}$, $\nu = 0.01$, $q = \{0.96, 0.98, 0.99\}$. For the HGT of coding regions: $gw = 0.3$, $rLGT = 0.99$. At small q the codes diversify. At large q the initial uniformity is maintained, while the codes evolve in concert towards optimality. HGT of coding regions enables maintenance of universality for smaller values of q.

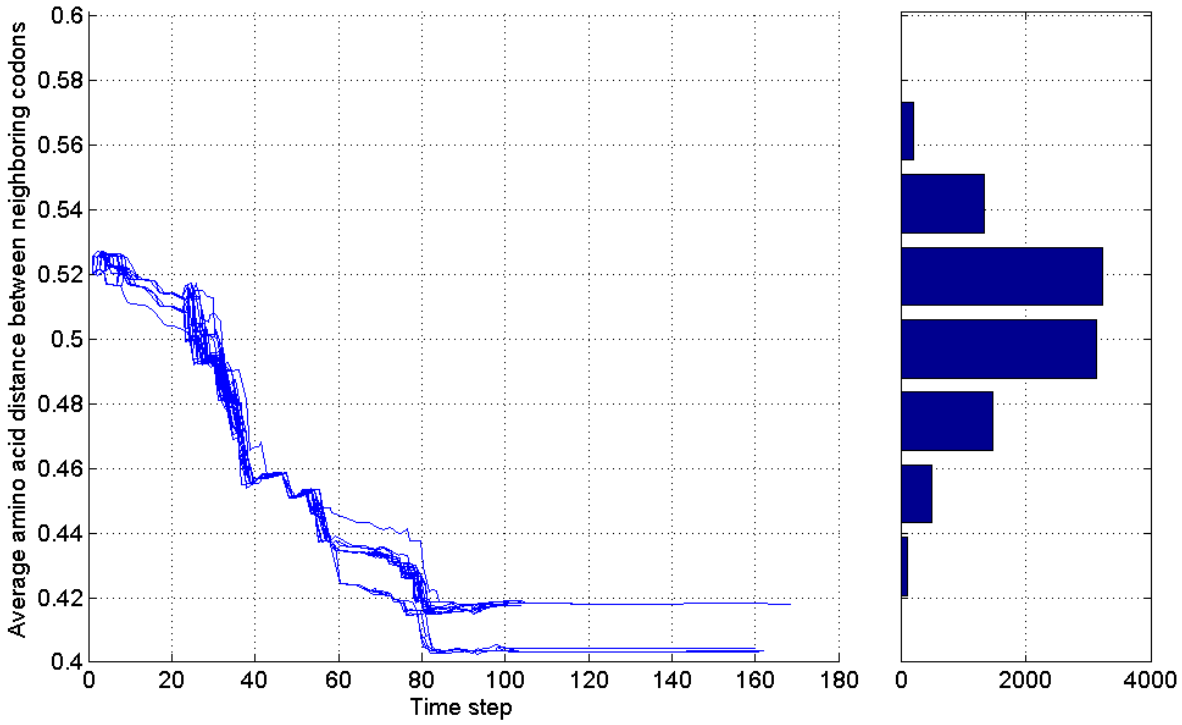

Figure 6.11: Communal evolution towards optimality of 40 codes through HGT of code specificators. No HGT of protein coding regions. Parameters. $q = 0.98$, $\phi = 0.1$, $\mu = 10^{-4}$, $\nu = 0.01$. The community evolves in concert towards optimality except for the split of the core.

towards optimality as a unit. There are values of q for which the core is unstable without HGT of coding regions, and stable with it. Interestingly, for sufficiently large values of q the core does not decay but exhibits "speciations". The effect of a core split on optimality is presented on Figure 6.11.

The data presented above is consistent with the idea that HGT of coding regions renormalizes q in a direction favoring universality, and that by increasing q we pass through three different dynamical regimes: core decay, core break up, apparently stable core.

6.6 Conclusions

In the previous chapter we argued that the code is easy to evolve towards optimality despite apparent barriers to its evolution. This made universality of the code even more puzzling. Previous arguments about universality rely on the existence of a universal common ancestor with a frozen

code. Here we presented the alternative that the universality of the genetic code is a generic consequence of the communal evolution of early life. HGT of protein coding regions and HGT of translational components ensures the emergence of clusters of similar codes and compatible translational machineries. Different clusters compete for niches, and due to the benefits of the communal evolution, the only stable solution of the cluster dynamics is universality. These mechanisms are consistent with two macroevolutionary scenarios: 1) the code stayed nearly universal at all times, 2) The codes proliferated at first but then gradually became universal.

We argued that the code specificators remained interchangeable during the code evolution, because of the special role of the genetic code as a communication protocol for innovations. In turn, exchange of code specificators enabled the concerted optimization of a core of identical genetic codes.

The suggestion that the code specificators remained interchangeable during the code evolution is consistent with the fact that translation is the most conserved cellular mechanism, and must have other signatures in the patterns of structural phylogeny of translational components. For example, one can juxtapose the structural phylogeny of the tRNA species with the phylogeny of other cellular components - protein folds, for example.

Chapter 7

Conclusions

The work I presented spans three broad areas: pattern formation in far from equilibrium systems, microbiology and early evolution. The goal has been to learn some graduate level condensed matter physics, learn even more biology, and identify biological questions where my scientific background can make a difference. It wasn't a one way process - I adapted my physicist thinking to the biological context.

I focused on the following conceptual evolutionary problems: What is the meaningful framework for describing microbial evolution, given that genetic exchange between microbes is frequent? Why is there such a diversity of genome biases among microorganisms? How can the genetic code change, if every change is seemingly lethal due to the simultaneous introduction of many amino acid substitutions? If the code evolved then why is it both universal and optimal?

Here are my contributions to answering these questions:

7.1 Are there bacterial species?

The most frequent type of genetic exchange for modern day microbes is homologous recombination. Homologous recombination is effective only between organisms with very similar sequences. This suggests that the microbes are partitioned into communities of similar sequences, with homologous recombination being frequent within communities and rare between them. Different communities can still influence each other's evolutionary histories by *non-homologous* horizontal gene transfer, which occurs at, perhaps, slower effective evolutionary rate. These communities are related to the notion of microbial species, but there is the problem of *local* versus *global* genetic isolation. Two

organisms can be very similar at some parts of their genomes but different at others. Correspondingly, they will be genetically isolated at some loci but not others. What makes things even worse, different parts of a genome can be similar to the corresponding genome sections of different sets of organisms. It seems that, unlike for sexual organisms, genetic isolation is not a between organism property for asexual microbes.

I show that the interplay between homologous recombination and point mutations makes local isolation unstable. In a community of similar genomes, an introduction of some local difference between the genomes, as a result of a non-homologous gene transfer or a genome rearrangement, triggers a diversification front that propagates along the genomes. Regions of local isolation with respect to homologous recombination grow. Partial genetic isolation is a temporary and unstable state, and, typically, genomes are either globally similar or globally different. Diversification fronts predispose for the existence of well-defined species, and the diversification fronts constitute a *mechanism for speciation*.

Since diversification fronts make a difference to the microbial evolution, it is natural to *classify* microbes based on their ability to support such fronts. This in turn depends on some of the details of the cellular machinery for homologous recombination and its inhibition. If the cellular machinery requires sequence similarity at both ends, diversification fronts are expected, as long as the rate of homologous recombination is high enough to maintain the sequence similarity within the community in this absence of diversification seeds. If the homologous machinery requires sequence similarity at only one end, diversification fronts won't be an expected outcome. This is a new type of classification, based on cellular properties that are not random, or simply experimentally convenient, but ones that are relevant for the communal evolutionary dynamics.

The diversification fronts leave a signature on the genomes of closely related organisms. I examined the completely sequenced genomes and identified plausible candidates for fronts within the *Bacillus cereus* group.

7.2 Why are there genome biases?

Different microorganisms have different codon usage and nucleotide composition biases. This offers a convenient handle on microbial evolution because one can extract the genome biases directly from

the sequences, and there is an ever increasing amount of sequence data. To extract evolutionary information, however, one has to understand the factors that cause and change the genome bias patterns.

My contribution is the realization that different types of genome biases are different manifestations of the same underlying process - spontaneous symmetry breaking due to the selection on the speed, accuracy and energy efficiency of template-directed synthesis. Translation, replication and transcription are the primary examples of template-directed synthesis. Selection on translation leads to codon usage biases, selection on replication and transcription leads to nucleotide usage biases, such as the GC-content. Since all templates are derived from one master template - the genome, different biases are interdependent.

The reason for the symmetry breaking is the *coevolution* of the letter composition of the template with the allocation of resources for the maintenance of a pool of free monomers that is used during the synthesis. In particular, in the case of translation, this is a coevolution between the codon usage and the tRNA expression levels. Having fewer types of tRNAs makes translation faster, more accurate, and more energy efficient. It is faster because, with a given tRNA budget, you can produce more tRNAs of each species. It is more accurate because there are fewer species that the ribosome has to discriminate between. It is more efficient, because you can spend less time and energy in proofreading. If some tRNA species are more popular than others - this is a bias. This is fine, but the number and expression levels of the tRNA species is constrained by the codon usage of the template. Similarly, which codons are preferred to others is determined by the context of the existing tRNA species abundance pattern. It is the coevolution of the two that determines the outcome.

The spontaneous order created by the coevolution is opposed by the randomizing tendency of point mutations. All in all, if the genome has many letters, there is a continuous phase transition controlled by the number of mutations per genome per generation.

One can interpret the existing sequence and tRNA usage data in the context of the framework above. This is a subject of coming work. One of the benefits of realizing the generic nature of the biases is that you can go beyond trying to understand modern day codon usage and GC-content and ask: What were the consequences of this coevolutionary mechanism for early life? After all,

selection on the efficiency of template-directed synthesis was, perhaps, much more pronounced then. If you are barely making a living, you are very careful about how you allocate your resources.

7.3 How can the code evolve towards optimality?

The conceptual problem is that every change of the code seemingly leads to many simultaneous amino acid substitutions, and is therefore lethal. Provoked by the presence of secondary derived non-universal codes, people proposed scenarios for code change. For example, extreme bias leads to the disappearance of some codons, enabling their reassignment in a neutral way [84]. Even if this is the correct explanation for the emergence of particular variant codes, it is not enough to address the question of why the code is highly optimized. Optimization requires many changes, and direction. Optimization is not a neutral process.

What is needed is a *closed model* of the evolution of the genetic code towards optimality. The problem is that we don't know nearly enough about the context of early life. The way we approach the problem consists of two parts. The first is the recognition of universal mechanisms, and the second is the *assumption* that the context of translation is qualitatively similar to the one today. It will be an advance to demonstrate that optimization is possible in a qualitatively modern translational context. The alternative is to say: "Things were somehow different in the dark ages of early evolution, and the problem is not currently tractable".

One universal mechanism was already identified by Sella and Ardell [103] - the codon usage and the genetic code coevolve. The other is the spontaneous emergence of genome biases identified above.

The closed model that Sella and Ardell constructed was abstract. Because of that it misses, I claim in this thesis, two relevant aspects of the context. First, the proteome contains a wide distribution of sites, ranging from neutral to ones specifically requiring a given amino acid. Second, the genetic code is not an abstract map, but an ecosystem of tRNAs. These two claims are based on observations of the contemporary biological context.

I merged together the model of Sella and Ardell with the model of the coevolution between tRNA expression levels and codon usage from chapter 4. Then, I put in by hand the proteome context just discussed. The result is that a code evolving through Sella and Ardell's mechanism is frozen,

as one would intuitively expect, and the model presented here optimized itself. The conclusion is that the proteome context leads to freezing, and the template-directed synthesis context leads to catalysis of code changes. The driving force is the benefit of having a more optimal genetic code, say to reduce the fitness effects of mistranslation, or alternatively, to minimize the resources invested in proofreading.

A level of detail that I have not put in yet is the fact that a tRNA can recognize more than one codon, and that many tRNA species can recognize the same codon. If we add additional relevant features of the context, we can, perhaps, study whether other genetic code features, such as position asymmetry, can be statistically reproduced. If several qualitatively different statistical features of the code can be reproduced by a single closed model, that would be a success.

7.4 How can the genetic code be both optimal and universal?

There is a conceptual problem behind the simultaneous optimality and universality of the code. In order for the code to evolve towards optimality, there needs to be a diversity of codes on which natural selection can operate. However, there are many different solutions to the same problem, and there are many other cellular properties that selection optimizes. The typical outcome is that the cellular properties diversify along the branches of the tree of life. Is there a deep reason that the genetic code is an exception? What does this tell us about early evolution?

I argued in this thesis that the deep reason is that the genetic code is not just one more cellular property, albeit a very central one, but also a communication protocol that enables transfer of protein innovations between organisms. Once you realize this, you can ask: What was the mechanism that led to the establishment of this universal protocol? The answer necessarily involves communication, i.e. HGT. A related question is: Is it only the abstract protocol that is the same, or also the molecular parts that specify it?

In chapter 6 I attempted to decompose the role of HGT for the universality of the code into different possible channels, and, then, offer a synthesis that shows how these different channel synergistically interact.

Both exchange of protein coding regions and translational components is important. HGT of protein coding regions between organisms with similar codes provides an effective attractive force

between the codes. At a larger scale, it partitions the organisms into communities of incompatible codes. Driven by innovation sharing, these communities compete for niches. Because larger communities have larger pools of innovations, more popular codes are favored. This is a "winner takes all" dynamics which generically leads to universality.

In addition, I hypothesize that the universality of the code went hand in hand with a universal translational standard. This universal standard enabled exchange of code specificators, such as tRNAs between different organisms. The exchange of code specificators enabled concerted evolution of a universal code towards optimality, which in turn enforced the compatibility of the translational apparatus through the advantage of the popular codes. I suggest that different translational machineries remained compatible as long as the concerted optimization of the codes lasted.

The suggestion that the code specificators remained interchangeable during the code evolution is consistent with the fact that translation is the most conserved cellular mechanism, and must have other signatures in the patterns of structural phylogeny of translational components. For example, one can juxtapose the structural phylogeny of the tRNA species with the phylogeny of other cellular components - protein folds, for example.

Apart from experimental verification, we offer a natural solution to the conceptual problem of the universality of the code. Perhaps, these ideas will be modified and coopted in future studies of early evolution.

7.5 Communal evolution from the roots to the leaves of the tree of life

In the history of my thesis research, the idea of a transition from predominantly horizontal to predominantly vertical evolution [9] (Darwinian transition) played an important role. At the same time, it is a convenient way to put the different projects together.

Thinking about the Darwinian transition led me to the belief that the transition is, actually, a transition form predominantly non-homologous to predominantly homologous gene exchange. So, I was happy when I learnt that there is a mechanism for homologous exchange in microbes - homologous recombination. Since sex is also a type of homologous exchange - analogous pieces are

swapped, all organisms today have mechanisms for homologous recombination.

Suddenly, this led me to the project in chapter 3, which models the communal evolution near the leaves of the tree of life. On the opposite end of the spectrum is the work on the role of communal evolution before the root of the tree of life. Thinking about the role of HGT for the universality of the genetic code, naturally made me think about how the code can change. This led me to the "discovery" of the communality of the translational system. The genetic code is a community of tRNAs that coevolve with their context. Presumably this coevolution was important before the root of the tree of life, leading to the optimization of the code, and is also relevant today for the codon usage and GC-bias problem.

References

[1] J. Cotton, "Polymer excluded volume exponent nu . an experimental verification of the n vector model for n=0," *J. Physique Lett.(Paris)*, vol. 41, pp. L231–L234, 1980.

[2] J. L. Guillou and J. Zinn-Justin, "Accurate critical exponents from field theory," *J. Physique Lett. (Paris)*, vol. 50, pp. 1365–1370, 1989.

[3] Y. Oono, "Statistical physics of polymer-solutions - conformation-space renormalization-group approach," *Adv. Chem. Phys.*, vol. 61, pp. 301–437, 1985.

[4] M. Bulmer, "Coevolution of codon usage and transfer rna abundance," *Nature*, vol. 325, pp. 728–730, 1987.

[5] M. Bulmer, "The selection-mutation-drift theory of synonymous codon usage," *Genetics*, vol. 149, pp. 897–907, 1991.

[6] G. Tyson, J. Chapman, P. Hugenholtz, E. Allen, R. Ram, P. Richardson, V. Solovyev, E. Rubin, D. Rokhsar, and J. Banfield, "Community structure and metabolism through reconstruction of microbial genomes from the environment," *Nature*, vol. 428, pp. 37–43, 2004.

[7] N. R. Pace, "A molecular view of microbial diversity and the biosphere," *Science*, vol. 276, pp. 734–740, 1997.

[8] C. Woese, "Bacterial evolution," *Microbiol. Rev.*, vol. 51, pp. 221–271, 1987.

[9] C. Woese, "On the evolution of cells," *PNAS*, vol. 99, pp. 8742–8747, 2002.

[10] J. S. Langer in *Directions in condensed matter physics* (G. Grinstein and G. Mazenko, eds.), p. 165, World Scientific Press, 1986.

[11] A. Karma and W. Rappel, "Phase-field method for computationally efficient modeling of solidification with arbitrary interface kinetics," *Phys. Rev. E*, vol. 53, pp. R3017–R3020, 1996.

[12] N. Provatas, N. Goldenfeld, and J. Dantzig, "Efficient computation of dendritic microstructures using adaptive mesh refinement," *Phys. Rev. Lett.*, vol. 80, pp. 3308–3311, 1998.

[13] W. Wheeler, G. McFadden, and W. Boettinger, "Phase-field model for solidification of a eutectic alloy," *Proc. R. Soc. London A*, vol. 452, p. 495, 1996.

[14] A.Karma and W. Rappel, "Phase-field method for computationally efficient modeling of solidification with arbitrary interface kinetics," *Phys. Rev. Lett.*, vol. 77, p. 4050, 1996.

[15] G. Caginalp and X. Chen in *On the evolution of phase boundaries* (M. E. Gurtin and G. B. McFadden, eds.), vol. 1, p. 1, Springer-Verlag, 1992.

[16] L. Chen, N. Goldenfeld, and Y. Oono, "Renormalization group and singular perturbations: Multiple scales, boundary layers, and reductive perturbation theory," *Phys. Rev. E*, vol. 54, p. 376, 1996.

[17] R. Willnecker, "Evidence of nonequilibrium processes in rapid solidification of undercooled metals," *Phys. Rev. Lett.*, vol. 62, p. 2707, 1989.

[18] J. Lum, D. Matson, and M. Flemings, "High-speed imaging and analysis of the solidification of undercooled nickel melts," *Metall. Mater. Trans. B*, vol. 27, pp. 865–870, 1996.

[19] D. M. Matson in *Solidification 1998* (S.P.Marsh, J. Dantzig, R. Trivedi, W. Hofmeister, M. Chu, E. Lavernia, and J. H. Chun, eds.), vol. 1, p. 233, The Mineral, Metals and Materials Society, 1998.

[20] W. Hofmeister, R. Bayuzick, and M. Robinson, "Dual purpose pyrometer for temperature and solidification velocity-measurement," *Rev. Sci. Instrum.*, vol. 61, p. 2220, 1990.

[21] J. Hoyt, B. Sadigh, M. Asta, and S. Foiles, "Kinetic phase field parameters for the cu-ni system derived from atomistic computations," *Acta mater.*, vol. 47, p. 3181, 1999.

[22] J. Hoyt, M. Asta, and A. Karma, "Method for computing the anisotropy of the solid-liquid interfacial free energy," *Phys. Rev. Lett.*, vol. 86, p. 5530, 2001.

[23] J. Bragard, A. Karma, Y. Lee, and M. Plapp, "Linking phase-field and atomistic simulations to model dendritic solidification in highly undercooled melts," *Interface Science*, vol. 10, p. 121, 2002.

[24] C. Kurland, B. Canback, and O. Berg, "Horizontal gene transfer: a critical view," *PNAS*, vol. 100, pp. 9658–9662, 2003.

[25] J. Gogarten, W. Doolittle, and J. Lawrence, "Prokaryotic evolution in light of gene transfer," *Mol. Biol. Evol.*, vol. 19, pp. 2226–2238, 2002.

[26] J. Lawrence, "Gene transfer in bacteria: Speciation without species?," *Theor. Pop. Biol.*, vol. 61, pp. 449–460, 2002.

[27] F. Cohan, "Bacterial species and speciation," *Syst. Biol.*, vol. 50, pp. 513–524, 2001.

[28] H. Ochman, J. Lawrence, and E. Groisman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, pp. 299–304, 2000.

[29] O. Berg and C. Kurland, "Evolution of microbial genomes: Sequence acquisition and loss," *Mol. Biol. Evol.*, vol. 19, pp. 2265–2276, 2002.

[30] E. Joyce, K. Chan, N. Salama, and S. Falkow, "Redefining bacterial populations: a post-genomic reformation," *Nat. Rev. Genet.*, vol. 3, pp. 462–473, 2002.

[31] C. Radding, "Links between recombination and replication: Vital roles of recombination," *PNAS*, vol. 98, p. 8172, 2001.

[32] M. Vulic, F. Dionisio, F. Taddei, and M. Radman, "Molecular keys to speciation: Dna polymorphism and the control of genetic exchange in enterobacteria," *PNAS*, vol. 94, pp. 9763–9767, 1997.

[33] J. Majewski and F. Cohan, "The effect of mismatch repair and heteroduplex formation on sexual isolation in bacillus," *Genetics*, vol. 148, pp. 13–18, 1998.

[34] J. Majewski and F. Cohan, "Dna sequence similarity requirements for interspecific recombination in bacillus," *Genetics*, vol. 153, pp. 1525–1533, 1998.

[35] J. Majewski, P. Zawadzki, F. Cohan, and C. Dowson, "Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation," *J. of Bacteriol.*, vol. 182, pp. 1016–1023, 2000.

[36] W. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, pp. 2124–2128, 1999.

[37] S. Guttman and D. Dykhuizen, "Clonal divergence in *escherichia coli* as a result of recombination, not mutation," *Science*, vol. 266, pp. 1380–1383, 1994.

[38] E. Feil and B. Spratt, "Recombination and the population structures of bacterial pathogens," *Annu. Rev. Microbiol.*, vol. 55, pp. 561–590, 2001.

[39] J. M. Smith, N. H. Smith, M. O'Rourke, and B. Spratt, "How clonal are bacteria," *PNAS*, vol. 90, pp. 4384–4388, 1993.

[40] E. Feil, E. Holmes, D. Bessen, M.-S. Chan, N. Dayi, M. Enright, R. Goldstein, D. Hood, A. Kaliai, C. Moore, J. Zhou, and B. Spratt, "Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences," *PNAS*, vol. 98, pp. 182–187, 2001.

[41] R. Milkman, "Recombination and population structure in *escheria coli*," *Genetics*, vol. 146, pp. 745–750, 1997.

[42] E. Feil, "Small change: Keeping pace with microevolution," *Nat. Rev. Microbiol.*, vol. 2, pp. 483–495, 2004.

[43] S. Kurtz, A. Phillippy, A. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. Salzberg, "Versatile and open software for comparing large genomes," *Genome Biology*, vol. 5, p. R12, 2004.

[44] N. Ivanova, A. Sorokin, I. Anderson, N. Galleron, B. Candelon, V. Kapatral, A. Bhattacharyya, G. Reznik, N. Mikhailova, A. Lapidus, *et al.*, "Genome sequence of bacillus cereus and comparative analysis with bacillus anthracis," *Nature*, vol. 423, pp. 87–91, 2003.

[45] D. Rasko, J. Ravel, O. A. Okstad, E. Helgason, R. Z. Cer, L. Jiang, K. Shores, D. Fouts, N. Tourasse, S. Angiuoli, *et al.*, "The genome sequence of bacillus cereus atcc 10987 reveals metabolic adaptations and a large plasmid related to bacillus anthracis pxo1," *Nucleic Acids Res.*, vol. 32, pp. 977–988, 2004.

[46] W.-H. Li, "Unbiased estimation of the rates of synonymous and nonsynonymous substitution," *J. Mol. Evol.*, vol. 36, pp. 96–99, 1993.

[47] S. Sawyer, "Statistical tests for detecting gene conversion," *Mol. Biol. Evol.*, vol. 6, pp. 526–538, 1989.

[48] D. Treves, S. Manning, and J. Adams, "Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of escherichia coli," *Mol. Biol. Evol.*, vol. 15, pp. 789–797, 1998.

[49] J. Lawrence and H. Hendickson, "Lateral gene transfer: When will adolescence end?," *Molecular Microbiology*, vol. 50, pp. 739–749, 2003.

[50] E. Chargaff, "Structure and function of nucleic acids as cell constituents," *Fed. Proc.*, vol. 10, pp. 654–659, 1951.

[51] N. Sueoka, "On the genetic basis of variation and heterogeneity of dna base composition," *PNAS*, vol. 48, pp. 582–592, 1962.

[52] S. Chen, W. Lee, A. Hottes, L. Shapiro, and H. McAdams, "Codon usage between genomes is constrained by genome-wide mutational processes," *PNAS*, vol. 101, pp. 3480–3485, 2004.

[53] J. Wernegreen and D. Funk, "Mutation exposed: a neutral explanation for extreme bias composition of an endosymbiont genome," *J. Mol. Evol.*, vol. 59, pp. 849–858, 2004.

[54] M. Kimura, *The neutral theory of molecular evolution.* Cambridge: Cambridge University Press, 1983.

[55] E. Mayr, *Populations, species, and evolution: an bbridgment of animal species and evolution.* Belknap Press, 1970.

[56] R. Grantham, "Workings of the genetic-code," *Trends Biochem. Sci.*, vol. 5, pp. 327–331, 1980.

[57] R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave, "Codon catalog usage and the genome hypothesis," *Nucleic Acids Res.*, vol. 8, pp. r49–r62, 1980.

[58] H. Akashi and A. Eyre-Walker, "Translational selection and molecular evolution," *Curr. Opin. Genet. Dev.*, vol. 8, pp. 688–693, 1998.

[59] C. Kurland, "Translational accuracy and the fitness of bacteria," *Annu. Rev. Genet.*, vol. 26, pp. 29–50, 1992.

[60] M. Reis, R. Savva, and L. Wernisch, "Solving the riddle of codon usage preferences: a test for translational selection," *Nucleic Acids Res.*, vol. 32, pp. 5036–5044, 2004.

[61] G. Singer and D. Hickey, "Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content," *Gene*, vol. 317, pp. 39–47, 2003.

[62] E. Willie and J. Majewski, "Evidence for codon bias selection at the pre-mrna level in eukaryotes," *Trends in Genetics*, vol. 20, pp. 534–538, 2004.

[63] T. Ikemura, "Codon usage and transfer-rna content in unicellular and multicellular organisms," *Mol. Biol. Evol.*, vol. 2, pp. 13–34, 1985.

[64] P. Sharp and W. H. Li, "The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias," *Mol. Biol. Evol.*, vol. 4, pp. 222–230, 1987.

[65] P. Sharp, M. Stenico, J. Peden, and A. Lloyd, "Codon usage - mutational bias, translational selection, or both," *Biochem. Soc. Trans.*, vol. 21, pp. 835–841, 1993.

[66] H. Dong, L. Nilsson, and C. Kurland, "Co-variation of trna abundance and codon usage in escherichia coli at different growth rates," *J. Mol. Biol.*, vol. 260, pp. 649–663, 1996.

[67] M. Kimura, "Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons," *PNAS*, vol. 78, pp. 5773–5777, 1981.

[68] C. Kurland and M. Ehrenberg, "Growth-optimizing accuracy of gene expression," *Annu. Rev. Biophys. Biophys. Chem.*, vol. 16, pp. 291–317, 1987.

[69] O. Berg and C. Kurland, "Growth rate-optimised trna abundance and codon usage," *J. Mol. Biol.*, vol. 270, pp. 544–550, 1997.

[70] S. Kanaya, Y. Yamada, Y. Kudo, and T. Ikemura, "Studies of codon usage and trna genes of 18 unicellular organisms and quantification of *bacillus subtilis* trnas: gene expression level and species-specific diversity of codon usage based on multivariate analysis," *Gene*, vol. 238, pp. 143–155, 1999.

[71] H. Akashi, "Synonymous codon usage in *drosophila melanogaster*: natural selection and translational accuracy," *Genetics*, vol. 136, pp. 927–935, 1994.

[72] P. Sharp, E.Bailes, R. Grocock, J. Peden, and R. Sockett, "Variation in the strength of selected codon usage bias among bacteria," *Nucleic Acids Res.*, vol. 33, pp. 1141–1153, 2005.

[73] H. Akashi, "Gene expression and molecular evolution," *Curr. Opin. Genet. Dev.*, vol. 11, pp. 660–666, 2001.

[74] W. Doolittle, "You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear," *Trends Genet.*, vol. 14, pp. 307–311, 1998.

[75] K. Dittmar, E. Mobley, A. Radek, and T. Pan, "Exploring the regulation of trna distribution on the genomic scale," *J. Mol. Biol.*, vol. 337, pp. 31–47, 2004.

[76] G. Sella and D. Ardell, "The impact of message mutation on the fitness of a genetic code," *J. Mol. Evol.*, vol. 54, pp. 638–651, 2002.

[77] E. Rocha, "Codon usage bias from trnas point of view: Redundancy, specialization, and efficient decoding for translation optimization," *Genome Res.*, vol. 14, pp. 2279–2286, 2004.

[78] I. Tubulekas and D. Hughes, "Growth and translation elongation rates are sensitive to the concentration of ef-tu," *Mol. Microbiol.*, vol. 8, pp. 761–770, 1993.

[79] J. Drake, "The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes," *Ann. N. Y. Acad. Sci.*, vol. 870, pp. 100–107, 1999.

[80] R. Knight, S. Freeland, and L. Landweber, "A simple model based on mutation and selection explains trends in codon and amino-acid usage and gc composition within and across genomes," *Genome Biology*, vol. 2, pp. research0010.1–0010.13, 2001.

[81] F. Crick, "The origin of the genetic code," *J. Mol. Biol.*, vol. 38, pp. 367–379, 1968.

[82] B. Barrell, A. Bankier, and J. Drouin, "A different genetic code in human mitochondria," *Nature*, vol. 282, pp. 189–194, 1979.

[83] P. Keeling and W. Doolittle, "A non-canonical genetic code in an early diverging eukaryotic lineage," *EMBO J.*, vol. 15, pp. 2285–2290, 1996.

[84] S. Osawa, *Evolution of the genetic code*. Oxford: Oxford Univ. Press, 1995.

[85] T. Sonneborn in *Evolving Genes and Proteins* (V.Bryson and H. J. Vogel, eds.), (New York), Academic Press, 1965.

[86] C. Woese, "On the evolution of the genetic code," *PNAS*, vol. 54, pp. 1546–1552, 1965.

[87] C. Woese, D. Dugre, S. Dugre, M. Kondo, and W. Saxinger, "On the fundamental nature and evolution of the genetic code," *Cold Spring Harbour Symp. Quant. Biol.*, vol. 31, pp. 723–736, 1966.

[88] J. T. Wong, "A coevolution theory of the genetic code," *PNAS*, vol. 72, pp. 1909–1912, 1975.

[89] J. T. Wong, "Coevolution theory of the genetic code at age thirty," *BioEssays*, vol. 27, pp. 416–425, 2005.

[90] D. Haig and L. Hurst, "A quantitative measure of error minimization in the genetic code," *J. Mol. Evol.*, vol. 33, pp. 412–417, 1991.

[91] S. Freeland and L. Hurst, "The genetic code is one in a million," *J. Mol. Evol.*, vol. 47, pp. 238–248, 1998.

[92] D. Gilis, S. Massar, N. Cerf, and M. Rooman, "Optimality of the genetic code with respect to protein stability and amino-acid frequencies," *Genome Biology*, vol. 2, pp. research0049.1–0049.12, 2001.

[93] S. Freeland, R. Knight, L. Landweber, and L. Hurst, "Early fixation of an optimal genetic code," *Mol. Biol. Evol.*, vol. 17, pp. 511–518, 2000.

[94] R. Knight, *The origin and evolution of the genetic code: statistical and experimental investigations*. PhD thesis, Princeton University, 2001.

[95] R. Knight, S. Freeland, and L. Landweber, "Rewiring the keyboard: evolvability of the genetic code," *Nat. Rev. Genet.*, vol. 2, pp. 49–58, 2001.

[96] F. Crick, "The recent excitement in the coding problem," *F. Crick*, vol. 1, pp. 163–217, 1963.

[97] S. Osawa, T. Jukes, K. Watanabe, and A. Muto, "Recent evidence for evolution of the genetic code," *Microbiol. Rev.*, vol. 56, pp. 229–264, 1992.

[98] S. Osawa and T. Jukes, "Codon reassignment (codon capture) in evolution," *J. Mol. Evol.*, vol. 28, pp. 271–278, 1989.

[99] D. Schultz and M. Yarus, "Transfer rna mutation and the malleability of the genetic code," *J. Mol. Biol.*, vol. 235, pp. 1377–1380, 1994.

[100] D. Schultz and M. Yarus, "On malleability in the genetic code," *J. Mol. Evol.*, vol. 42, pp. 597–601, 1996.

[101] S. Andersson and C. Kurland, "Reductive evolution of resident genomes," *Trends Microbiol.*, vol. 6, pp. 263–268, 1998.

[102] S. Andersson and C. Kurland, "Genomic evolution drives the evolution of the translation system," *Biochem. Cell Biol.*, vol. 73, pp. 775–787, 1995.

[103] D. Ardell and G. Sella, "No accident: genetic codes freeze in error-correcting patterns of the standard genetic code," *Phil. Trans. R. Soc. Lond. B*, vol. 357, pp. 1625–1642, 2002.

[104] D. Ardell and G. Sella, "On the evolution of redundancy in genetic codes," *J. Mol. Evol.*, vol. 53, pp. 269–281, 2001.

[105] E. Szathmary, "Codon swapping as a possible evolutionary mechanism," *J. Mol. Evol.*, vol. 32, pp. 178–182, 1991.

[106] F. Crick and L. Orgel, "Directed panspermia," *Icarus*, vol. 19, pp. 341–346, 1973.

[107] J. T. Wong, "The evolution of a universal genetic code," *PNAS*, vol. 73, pp. 2336–2340, 1976.

[108] S. Freeland, "The darwinian genetic code: an adaptation for adapting?," *Genetic Programming and Evolvable Machines*, vol. 3, pp. 113–127, 2002.

[109] E. Szathmary, "Coding coenzyme handles: a hypothesis for the origin of the genetic code," *PNAS*, vol. 90, pp. 9916–9920, 1993.

[110] E. Szathmary, "The origin of the genetic code: amino acids as cofactors in an rna world," *Trends Genet.*, vol. 15, pp. 223–228, 1999.

[111] G. Olsen and C. Woese, "Lessons from an archaeal genome: what are we learning from *methanococcus jannaschii*?," *Trends Genet.*, vol. 12, pp. 377–379, 1996.

[112] T. Kuhn, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1970.

List of Publications

1. "Global divergence of microbial genome sequences mediated by propagating fronts", Kalin Vetsigian and Nigel Goldenfeld, PNAS **102**, 7332 (2005).

2. "Computationally efficient phase-field models with interface kinetics", Kalin Vetsigian and Nigel Goldenfeld, Phys. Rev. E **68**, 060601(R) (2003).

Author's Biography

Kalin Vetsigian was born in Plovdiv, Bulgaria on 16th of February, 1977. He graduated from the Mathematical High School in Plovdiv in 1995. As a high school student he participated in two International Physics Olympiads - Beijing, China in 1994 and Canberra, Australia in 1995. In Canberra he won a gold medal. He continued his studies it the Physics Department of the University of Plovdiv. In 1996, he also joined the department of Mathematcs and Computer Science. In the summer of 1997 he came to the United States as a transfer student at the Massachusetts Institute of Technology (MIT). He graduated MIT in June, 2000 with a Bachelor's of Science in Physics and a Bachelor's of Science in Mathematics. In August 2000, he started his graduate studies in the Physics Department of the University of Illinois at Urbana-Champaign. The next year, he joined Nigel Goldenfeld's research group. In the summer of 2002 he spent 10 weeks as an intern in the Derivatives Research Department of JP Morgan Chase in New York. In 2005, he obtained the "Renato Bobone Award to the Outstanding European Graduate Student in Physics".