From Molecular Systems to Simple Cells:

a Study of the Genotype-Phenotype Map

DISSERTATION

zur Erlangung des akademischen Grades
Doktor rerum naturalium

Vorgelegt der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

von

Mag. Camille Stephan-Otto Attolini

am Institut für Theoretische Chemie und Molekulare
Strukturbiologie

im November 2005

A Christian, Amalia y Erwin.

A la memoria de mi abuela.

A todos los que estuvieron conmigo, en la distancia y no.

Gracias.

Aknowledgements

I would like to thank my supervisor Peter Stadler for all his support and immense scientific help during my studies. To Christoph Flamm, who became an excellent teacher and colleague and thanks to whom my work is much better. To all the people in the TBI. To Kurt, Roman, Ulli, Luky, Andrea, Rainer, Caro and Xtina for very good talks, especially non-academic ones.

This work was supported by the program of international schollarships from Consejo Nacional de Ciencia y Tecnologia, CONACyT, Mexico.

Abstract

The study of the genotype-phenotype map has been approached in many ways and from many different disciplines. A large amount of knowledge has been produced in the field of molecular evolution as well as in development and differentiation. Nevertheless, a better understanding of the relation between the molecular level and the complexity of organisms is still far from being successful.

This thesis reviews some of the results on the topic and addresses questions about neutrality, plasticity and evolution in models of interacting RNA molecules as well as regulatory networks in simple cells.

We found some resemblances between lower level systems and more complex layers. Neutrality may well be a result of the sequence to structure map in molecules, while plasticity is more probable achieved only when more than one layer lies between genotype and phenotype.

The instability of molecular systems against parasites and in general deleterious mutations is amazingly replaced by a robustness and self regulation in higher order maps that forces to focus on modular organizations and not stoichiometric regulated systems.

Zusammenfassung

Die Beziehungen zwischen Genotyp und Phänotyp sind in unterschiedlichen Fachrichtungen auf viele verschiedene Arten untersucht und charakterisiert worden. Besonders in den Bereichen der molekularen Evolution, der Entwicklungsbiologie und der zellulären Differenzierung hat es in den letzten Jahrzehnten viele erfolgversprechende Ansätze und grosse Fortschritte gegeben. Dennoch ist man auch in diesen Feldern noch weit davon entfernt, die komplizierten Beziehungen zwischen den molekularen Vorgängen und den daraus entstehenden komplexen Organismen genauer zu verstehen.

Die vorliegende Arbeit beschäftigt sich mit einigen Ergebnissen in diesen Gebieten und verfolgt Fragestellungen bezüglich Neutralität, Plastizität und Evolution in Modellen interagierender RNA Moleküle sowie auch genregulatorischer Netzwerke in einfachen Zellen.

Wir fanden einige grundlegende Ähnlichkeiten zwischen einfachen und komplexer organisierten Systemen. Während Neutralität eine Folge der Beziehungen zwischen Molekülsequenz und -struktur zu sein scheint, erfordert Plastizität meist mehr als eine Ebene zwischen Genotyp und Phänotyp.

Überraschender Weise wird die Anfälligkeit molekularer Systeme für Parasiten und für schädliche Mutationen im Allgemeinen bei Systemen mit mehreren Ebenen durch eine immanente Stabilität und Selbstregulation ersetzt,weshalb man sie eher als modular organisiert, denn als stoichiometrisch reguliert betrachten kann.

Contents

1 Introduction

There is without any doubt, a relation between genotype and phenotype in living organisms. The carrier of information and its expression into a living being are linked by an extense chain of events which involve molecular reactions and spacial, temporal and hierarchical organizations. From the level of stoichiometric related molecules to the cells of even the simplest organisms, there are several levels of increasing complexity each of them linked with the others by feedback loops and/or spacial vecinity.

How these reactions emerged in the first place? How could the information be stored and inherited within simple molecular systems? When are the different levels created and how was it possible to control their interactions?

Many of this questions have been partly answered from several disciplines ranging from dynamical systems applied to molecular evolution, to the detailed study of development and cell differentiation. In this dissertation we review some of the main aspects of the genotype-phenotype map developed so far. We focus on three levels of complexity: the molecular level, without any spacial information; a spatially organized molecular system with the hypercycle at its core; and, a simplified cell with a basic genome and a gene regulation network.

We stress the importance of neutrality in any gentype-phenotype map capable of evolution. The search for fitter phenotypes is only possible, in a non-trivial fitness landscape, when mutations with no effect in the phenotype exist. This requirement is even more important when there are interactions among species, or in other words, coevolving species.

Studying isolated populations is a first approach to this complicated task, nevertheless, the environment and the rest of the coevolving species drastically change the way evolution occurs. We use different ways of defining interactions among molecules and with the environment. In the molecular level, for example, we use the cofolding of a pair of RNA sequences or a fixed topology in a structures target set.

The last approach we take in this dissertation on studying genotype-phenotype maps is centered in the formation and evolution of gene regulatory networks. In recent times, the importance and influence of RNA as an extra regulatory layer in the developmental process has been strongly emphasized. We simulate

a population of cells with a simple genome decoded into proteins which regulate each other's transcription. The main task of this model is to understand the effect of a changing environment in the evolution of a genome regulated by a very simple network.

In the first chapter we introduce general aspects of the genotype-phenotype map. Topics as neutrality, plasticity, evolvability, gene transcription and regulation are presented. The second chapter addresses some aspects of molecular and population dynamics, as well as two models of molecular evolution and their implications. The third chapter presents a spatial model of the hypercycle with fold and cofold as different genotype to phenotype maps. In the fourth chapter we introduce `CelloS`, a model of simple cells with a genome and regulatory network which is intended to investigate on the emergence and evolution of gene regulatory networks. Work from sections 3, 4 and 5 has been published in (Stephan-Otto Attolini and Stadler, 2005), (Stephan-Otto Attolini *et al.*, 2005) and (Stephan-Otto Attolini and Stadler, 2004) respectively. Finally we present conclusions and further work in these directions.

2 A review on Genotype-Phenotype maps

2.1 Genotype-Phenotype-fitness maps

In any living organism, phenotype refers to the physical, organizational and behavioral expression during its lifetime. Genotype refers to a heritable repository of information that instructs the production of molecules whose interactions, in conjunction with the environment, generate and maintain the phenotype (Fontana and Schuster, 1998). The process by which the phenotype is decoded from its genotype is poorly understood due of the complexity of interactions and control mechanisms in several distinct levels. To investigate general features of this map is one of the most demanding tasks in theoretical biology. One approach widely developed (e.g. (Schuster, 2002)) is the analysis of mathematical models directed both to the study of evolutionary processes and to the reproduction of existing living systems.

In order to attain this, any comprehensive theory of evolution must handle the phenotype as an integral part of the model. Genotype-phenotype maps should be introduced in a formal mathematical way in response to this requirement. Mathematical functions will assign one phenotype to each genotype in the genotype space; the inverse may not be true, having many-to-one maps where one phenotype may be produced by many genotypes.

This unique assignment of phenotypes to genotypes, however, is an approximation of real biological systems. Phenotypes are not exclusively determined by genotypes, since environmental factors and epigenetic effects are also relevant.

In order to complete evolutionary dynamics, the fitness relevant properties must be extracted from the phenotype. There are many different ways to assign a fitness value to a given genotype. One class of models use direct random or nonrandom model assignments of fitness values to genotypes (Tarazona, 1992), while a second and more realistic way of doing this is to use a two-step relation with the phenotypes as intermediate states.

Both of these steps can be expressed in mathematical terms. The first one as a function f, from the genotype G to the phenotype P, $G \xrightarrow{f} P$, which if the environmental effects E are taken into account, or two genomes combined to create a single phenotype, takes the form: $G \times E \xrightarrow{f} P$ and $G_1 \times G_2 \xrightarrow{f} P$. And

the second one $P \xrightarrow{g} F$ as a function g from the phenotype to the fitness value F (usually the set of real numbers but where many variations are possible) which may consist of the aptitude of a single individual in a given environment, or of the complex set of interactions between several species coevolving in the same scenario each one with a given fitness which is modified by the interactions. It is also possible to take into account environmental changes by means of time dependent mappings.

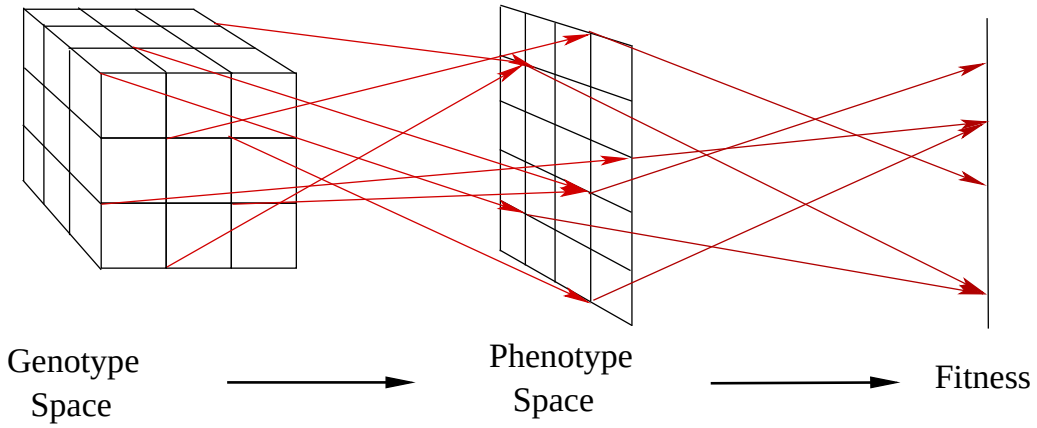


Figure 1: The genotype-phenotype map followed by the assignment of fitness values to phenotypes. Both maps may result to be many-to-one functions.

Regardless of the large variety of genotype-phenotype maps and the corresponding fitness functions, some regularities might be encountered in all kinds. If there are generic features in this mappings, the main goal of the study depends on the success to find a suitable model and an experimental system to which it applies and where further generalizations are possible.

It has been proposed (Kauffman, 1993) that genotype-phenotype-fitness maps must be highly non-linear. One reason to believe this is that linear systems would stop evolving when local peaks are found since small perturbations in linear systems are reflected also as small changes in phenotype. It is also clear that linear systems are less able to react to changing conditions, therefore fluctuating environments would lead in most cases to the extinction of entire populations (Bak, 1996).

A common way of studying this characteristics is through the modeling of evolving populations. The importance of this method resides in the fact that selection does not act on genes but on organisms. “Organisms are the ones that make the

struggling out there. If organisms could be described as the additive accumulation of what their genes are, then you could say that organisms are representing their genes, but they're not." (Brockman, 1995). Interactions among gene products result in emergent characteristics which are only possible when all the elements are present and the control mechanisms are acting on them. The result is high non-linearity from the individual genes to the final outcome. Only through this interactions is the organism realized. There is no decomposition of the living system in independent gene products, therefore, reduction of the process to the understanding of low-level entities alone is not possible (Gould, 2002). These features are invisible until the process reaches the next level of organization and emergent properties appear.

The high non-linearity of these maps is at the same time cause of the interesting behavior we observe and a reason for the difficulty to understand the processes behind them. As Bak says in (Bak, 1996):

> ...we may be dealing with highly non linear systems in which there is no simple way (or no way at all) to predict emergent behavior.

Even when there are cases, usually in the small scales, that are better understood, we cannot extrapolate from the microscopic scale (which could work under the laws of Darwinian evolution) to the macroscopic scale (which present extinctions and punctuated evolution that is impossible to predict from the microevolutionary theory). In (Simpson, 1944), Simpson argues in this direction that

> Geneticists can explain what happens to a population in controlled conditions and short time scales but not over large periods and fluctuating environments.

When regarding the problem as a system of entire populations, isolated subpopulations may present not-representative fractions of the genes in the entire populations. Therefore, evolutionary dynamics may lead to the displaying of all alternative alleles for a gene by chance or random drift.

Another way of approaching the problem is through the study of precise mappings from a given genotype to the corresponding phenotype. Development is the process through which the phenotype is created. From the genetic information to the actual organism, many regulatory steps, influenced by the environment,

are realized to give rise to the final shape of the phenotype. Nevertheless, plasticity may bring further changes in this phenotype by reacting to changes in the environment and activating translation and transcription of the genome. This is one of the most important characteristics of living cells: its capacity to receive impulses from the environment and react to them. Simple systems, as those at the molecular level, can fairly express this kind of behavior, while higher organizations, like those found in living organisms are not only capable but forced to develop this skill in order to survive. With the present knowledge and tools it is naive to try to follow all the steps in the development of the information carrier to the end product given the large amount of components involved and the even larger number of interactions among them. Various attempts of simplified models are at present being studied and computational tools have gained an important place in the pursue of this task ((Stephan-Otto Attolini *et al.*, 2005), (Marée and Hogeweg, 2002), (Reil, 2000), (Stanley and Miikkulainen, 2003)).

2.2 Characteristics of the genotype-phenotype map

2.2.1 Neutrality

Neutrality is the property of a map to allow mutations in the genotype without changing the correspondent phenotype. Mutations of this type give rise to neutral networks, that is, sets of all possible genotypes which have the same phenotype as image under the map (Schuster, 1997). These networks can be connected through neutral neighbors, i.e. genotypes which differ by applying one time the mutational operator. The number of connected components and qualities of the net depend on the properties of the map (Reidys *et al.*, 1997).

According to the neutral theory of evolution developed by Kimura, a large portion of all mutations is neutral and only a small fraction is actually beneficial (Kimura, 1983). This results in redundant maps where many genotypes code for identical phenotypes, i.e. many-to-one maps. “Selectively neutral genetic variations” may be responsible for most part of the evolution by random drift.

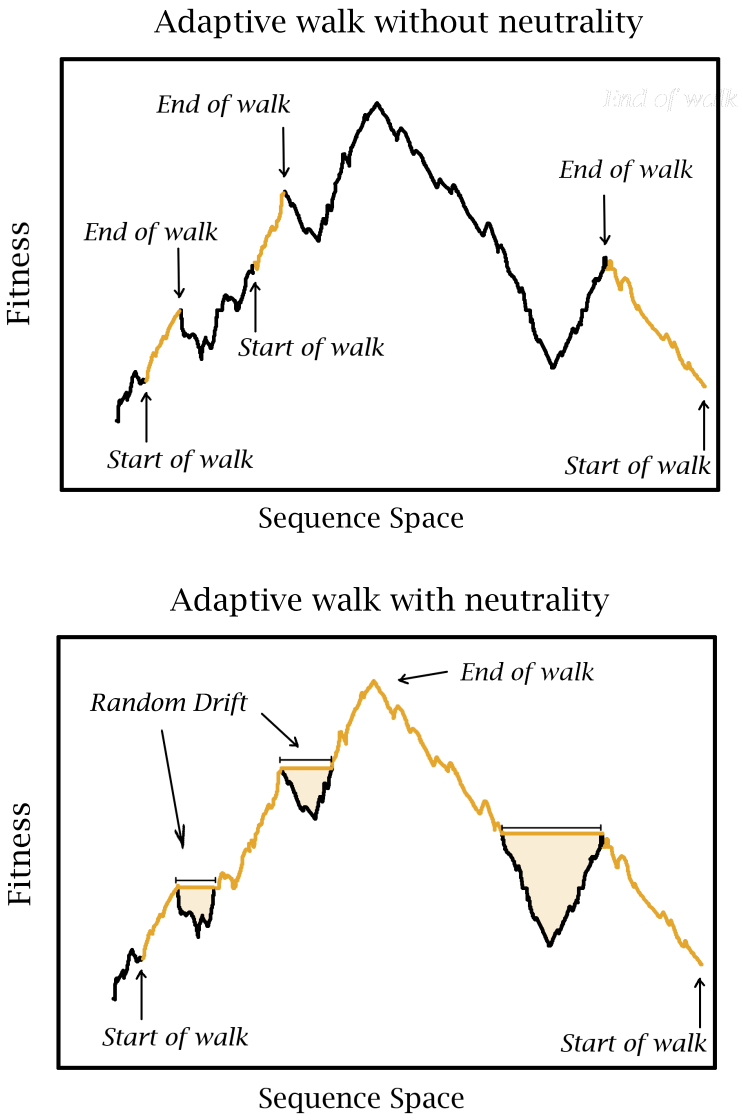


Figure 2: Optimization in sequence stape through adaptive walks of populations. Adaptive walks allow to choose the next step arbitrarily from all directions of where fitness is (locally) non decreasing. Populations can bridge over narrow valleys with widths of a few point mutations. In absence of neutrality (upper part), they are unable to span larger distances and thus will approach only the next major fitness peak. Populations on rugged landscapes with extended neutral networks evolve by a combination of adaptive walks and random drift at constant fitness along the network (lower part). Eventually populations reach the global maximum of the fitness landscape (if it exists at all).

Making use of a combination of graph theory and exhaustive folding in the case of evolution of RNA molecules, Schuster shows in (Schuster, 1997) that the sequence space is highly connected and neutral networks are randomly distributed along this space. It is of crucial importance for optimizing fitness in molecular evolution to have random drift in zones of equal fitness values (Schuster, 1986). This can be seen in Fig. 2, where random walks are able to find the global maximum only in the presence of neutrality (Schuster, 1986). Random walks which which found a local peak, will end there because no fitter genotype is in the mutational neiborhood. With the existence of neutral plateaus, it is possible to cross valeys and find better phenotypes around the neutral network of the corresponding genotype. A detailed description of neutrality in the mapping from RNA to secondary structures is presented in Section 3.2.

In the evolution of ribozymes *in vitro*, many mutations are allowed which do not have an impact in fitness (Wright and Joyce, 1997). This indicates that in adaptive evolution, the majority of point mutations are neutral (van Nimwegen *et al.*, 1999). Nimwegen showed also that the population moves in a neutral network towards highly connected regions, where neutrality is high and phenotypes are more robust against mutations. This implies that selection acts not only on phenotypes, but also on the evolvability of the genotype (van Nimwegen *et al.*, 1999).

The benefit of neutrality is the increase of possibilities for a search algorithm to find a superior phenotype without getting trapped in a local optimum (Ebner *et al.*, 2001b). There are two basic properties which help in the search for improvement: since the population is allowed to move inside the neutral networks without changing the phenotype, the individuals are spread along the network whenever a local optimum is found. In case the fitness landscape changes generating better optima, the population has more possibilities to find these genotypes in small radios around any of the particular individuals. It is also important to notice that even when neutral mutations leave the phenotype of an isolated individual unchanged, it is possible that interactions among species will be modified because of a neutral mutation. This could be the case of interacting molecules, when the structure of the molecule is thought as the phenotype (Stephan-Otto Attolini and Stadler, 2005), or of prey-predator relations or mating activities among the same species.

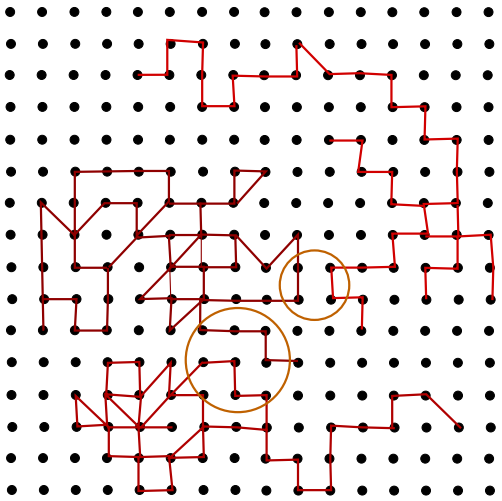


Figure 3: Schematic representation of neutral networks in sequences space. Colors code for individual networks, the circles show some regions where mutations will take sequences from one network to a different one.

Another characteristic of these maps is the possibility of random drift in sequence space. This allows the population to explore larger areas of the genotype space. However, if this map would be completely random, there would be no forces acting on the population to push it towards zones of larger evolvability. Ebner et al (Ebner *et al.*, 2001a) have shown that highly redundant mappings increase evolvability, defined as the ability of random variations to sometimes produce improvement in a population of coevolving species. It is clear then, that characteristics such as evolvability depend strongly in the neutrality of a map. Other research approaches exploring the effects of redundancy include the work of Barnett (Barnett, 1997), who introduced redundancy into kauffman's NK fitness landscapes and analyzed population dynamics on these static fitness landscapes.

It is crucial to note that not all mappings have the right type of redundancy. Redundancy without extensive and highly interwined neutral networks simply slow the rate of finding adaptive mutations because of the random drift and the difficult accessibility from one neutral net to the other (Ebner *et al.*, 2001b).

The most important global characterization of neutral networks is its average fraction of neutral neighbors, usually called the degree of neutrality. Neglecting the influence of the distribution of neutral sequences over sequence space, the degree of neutrality will increase with size of the pre-image. Generic properties of neutral networks (Reidys *et al.*, 1997) are derived by means of a random graph model.

2.2.2 Phenotypic plasticity

Phenotypic plasticity is any change in an organism's characteristics in response to an environmental signal. These responses are stimulated by signals from the environment, having as a result the change in protein production, physiological activity, growth or behavior. Whatever the type of impulse and response, signals must be internally processed at the level of cells. As an hypothesis of how this works, Schlichting and Smith in (Schlichting and Smith, 2002) propose that these changes are produced by different regimes of gene expression, no matter in which level the response occurs.

Metamorphosis is a differentiation event which is usually triggered by an external event but mediated and realized by a change in the internal behavior. Many traits may be involved in a single event, thus coordination among these is necessary in order to obtain the right response (Ballare *et al.*, 1997). It may also occur that metamorphosis is activated by a threshold in a single impulse from the environment. Once the crucial level is surpassed, the rest of the activity is mediated via concomitant changes in gene expression and the interaction between their products.

It is clear that phenotypic plasticity is selected in response to a variable environment (Schlichting and Smith, 2002). A phenotype with the possibility to transform in the fittest option every time a change in the conditions occur, will be favored among those who stay with the same shape or behavior.

Survival of a whole population can be assured by a high plasticity of their individuals or by a high variation among the genotypes of the population (Lloyd, 1984). If many individuals exist which are not optimal in all conditions, but can adapt to any change in the environment that may occur, this increases the probability of the survival of at least a part of the whole. This of course brings selection to a different level, one acting in the totality of genotypes in the population, and the other in the capacity of a single individual to adapt in a life time.

If an individual presents high plasticity, genotypic variation is kept hidden from selection, since external conditions wont lead to the organism's death. Thanks to this, further improvements can be achieved by accumulating mutations which are at first not under the pressure of selection.

Plasticity is a characteristic of the genotype-phenotype map which is selected

from other variants as a fitter option among the rest (Waddington, 1953). Nevertheless, development events responding to external or internal conditions must anyway be controlled by the genetic network and the correspondent interactions among products. This means that, the result of any change in phenotype resulting from impulses in the organisms reflects the intrinsic relation between the genotype-phenotype map and the environment. It is meaningless to study the relationship between genotype and phenotype without the environmental context (Schlichting and Pigliucci, 1998).

The mechanisms which promote plasticity are very abundant at the molecular-genetic level, due to the large plurality that exists in molecular reactions. Thousands of genes are interconnected between each other through their products and the reactions amongst them. Environmental signals also affect the way genes are expressed, having information traveling through several layers in both directions. The complexity of these systems is possible only because of the flexibility and multiplicity of the processes involved. Plasticity allows the organisms to explore the possible control mechanisms and find the best without reducing the already attained fitness.

As a direct consequence of all this, any theory looking for evolutionary algorithms should provide the system with enough plurality and a way to control and drive the process of selection (e.g. via canalization).

2.2.3 Evolvability and variability

The evolvability of a system (or organism) depends crucially on the genotype-phenotype map. Evolvability is understood as the capacity of the system to vary. The existence of possible adaptive mutations is originated by the mutational operator as well as the relation between genome and phenotype (Wagner and Altenberg, 1996).

It is important to point out that not every genetic system with its correspondent genotype-phenotype map is able to produce beneficial mutations. Therefore, it is valid to ask how an evolvable genome appeared in the first place and how does a map like this evolves (Wagner and Altenberg, 1996). The hypothesis of Wagner and Altenberg is that the map itself is under genetic control as well as the outcome and the resulting phenotype.

Two concepts must be defined in this framework. Variation of the phenotype is the actual difference between phenotypes, usually reflected among populations or species. Variation in the genome is produced primary by errors in the copying process. Depending on the genotype-phenotype map, these changes may not be expressed as variation in the phenotype. Neutrality is partly responsible for this as well as interactions between organisms and environmental factors shaping the phenotype. Variability, on the other hand, is the intrinsic capacity of the phenotype to change, belonging to the group of "dispositional" characteristics (Goodman, 1955). It is more difficult to measure variability than variation since the first implies a process while the second is a fixed characteristic for a given population.

The genotype-phenotype map constraints the possible outcomes from the variation in the genome. If fitness is understood as reproduction rate, then the fittest individuals are those more variated. This implies that a perturbation wont bring down the whole construct achieved by evolution (Kingman, 1978). In mathematical terms, this can be rephrased as a dynamical system having a stable attractor. The question here is how is the process of mutation and selection capable of escaping these stable peaks. *Strong causality* is the characteristic of a system by which small changes in the parameters are reflected as small changes in the system performance (Rechenberg, 1994).

The rate by which fitter adaptations are produced depends on the genetic mutation rate and the correlation with their possibility of generating fitter offspring.

Bentley and Kumar present in (Bentley and Kumar, 1999) a clasification of the main types of evolutionary algorithms used to model *embriogenies* (i.e. the process by which the body is growth from the genotype): the genetic algorithm, evolutionary programming, evolutionary strategies and genetic programming. According to them, an embriogeny must have an "indirect correspondance between alleles and phenotypic effects" and "polygeny", i.e. phenotypic traits being produced by multiple genes acting in combination (Bentley and Kumar, 1999), in order to be evolvable and able to find fitter phenotypes. Such a system would provide with a reduction of the search space, more complex solutions in solution space and adaptation among other benefits (Bentley, 1996). The problem is to design such embriogenies, which in nature are defined by the interactions of genes, their products and the environment. Embriogenies can be classified

as *external*, *explicit* and *implicit*, depending on how is the phenotype treated with respect of the genotype and whether this interaction can evolve or is fixed thoughout the process. All this clases have benefits and drawbacks, from the impossibility of evolve complex phenotypes to the capacity of doing it fast and with short genomes. A review of this study can be found in (Bentley and Kumar, 1999). One of the main conclusions in that contribution is the much better performance of embriogenies which are implicit, with non-fixed and many-to-one genotype-phenotype maps.

2.3 Fitness landscapes

Fitness landscapes were first introduced by Wright in 1931 (Wright, 1931), with the idea of assigning fitness values to every possible genome in a population and studying the characteristics of the resulting configuration. A landscape can be thought of as a kind of "potential function" underlying the dynamics of evolutionary optimization. Implicit in this idea is both a fitness function that assigns a fitness value to every possible genotype (or organism), and the arrangement of the set of genotypes in some kind of abstract space that describes how easily or frequently one genotype is reached from another one (Stadler, 2002b).

The space of RNA or DNA sequences is a metric space, i.e. a distance can be uniquely defined for every pair of sequences and this assignment fulfills all requirements of a metric (Hocking and Young, 1988). On the other hand, the space of phenotypes is usually much more complex and in nature never metric. Relation among phenotypes is important because of the different "accesibilities" between them.

A good case study is the RNA sequence to secondary structure map. If the secondary structure of a molecule is taken as its phenotype, then this space lacks a metric and exhibits an unfamiliar topological structure responsible for classic evolutionary patterns such as punctuation and irreversibility (Stadler *et al.*, 2001a). Studies in RNA shape space forces to abandon the notion of a vector space and to replace it by less intuitive structures based in neiborhoods and not in distances. The appropiate topological formal structure for the set of RNA secondary structures is a pretopology (Stadler *et al.*, 2001a), this being a result of requiring the genotype-phenotype map to be continues everywhere. Once this

pretopologies are assigned to the phenotype space, evolutionary trajectories can be tested for continuity. Punctuation is one example of discontinuos transitions between two different phenotypes.

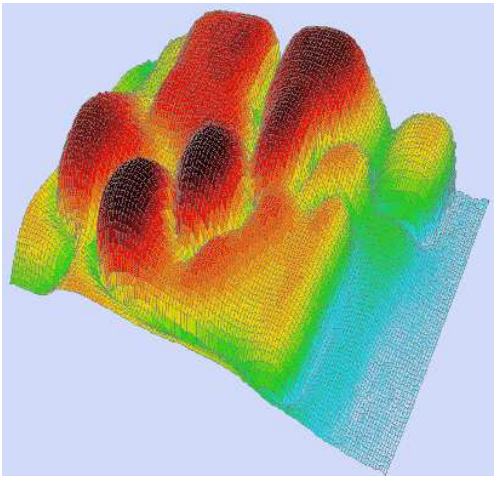


Figure 4: Fitness landscape of a two dimensional space into the real numbers. Level curves are the equivalent of neutral networks: movements inside them do not produce changes in phenotype.

In combinatorial optimization the fitness function is usually referred to as the cost function, and a move-set allows to inter-convert the elements of the search space. The application of evolutionary models to combinatorial optimization problems has lead to the design of so-called evolutionary algorithms such as Genetic Algorithms, Evolution Strategies, and Genetic Programming (Koza, 1994).

In this dissertation, the genotype space is always the set of RNA sequences of same length; the move-set consists of point mutations which are introduced in the copying process and several genotype-phenotype maps are studied as well as the way fitness is assigned to individuals. A detailed description of the characteristics of these maps is presented in the next chapter.

The intuitive notion of ruggedness is closely related to the difficulty of optimizing (or adapting) on a given landscape. It depends obviously on both the fitness function and the geometry of the search space, which is induced by the search process (Stadler, 2002c). Understanding the geometric features of landscapes is of crucial importance when studying evolutionary processes and the capacity of a given population under a certain phenotype-genotype map to attain fitter phenotypes. Characteristics as mountain massifs, valleys, basins, peaks, plains and ridges in multidimensional combinatorial objects may look quite different from our 3D experience and oftentimes require a mathematical description in

terms of algebraic combinatorics rather than calculus (Stadler, 2002b).

Landscapes can also be studied from a dynamical point of view, focusing on the features of a dynamical system, for instance an evolving population, that uses the landscape as its substrate. The challenge for a theory of landscapes is therefore to link these two points of views, for instance by determining how geometric properties influence the dynamical behavior.

It is worth mentioning that even when evolution is often view as a climbing process in fitness landscapes, evolution is not about an increase in complexity or a race towards a predefined goal (Brockman, 1995). The pathways followed by evolving populations may not be the most advantageous but the most easily reachable.

2.4 Regulatory networks

The picture long time accepted of one gene - one trait in genotype-phenotype maps has been dramatically changed thanks in part to the discovery of the very complex and interconnected networks between gene products. The regulation of gene expression is one of the most complex and fascinating problems in biology (Reil, 2000). The importance of regulatory networks in gene expression was noticed when the Human Genome Project released the results on the number of genes found in organisms considered "complex". The number of genes on *D. melanogaster* is not significantly smaller than those in human or *C. elegans*. It is then easy to imply that the complexity of higher eukaryote is not due to an increase on the number of genes but rather to the complexification of the regulating networks (Geard and Wiles, 2003).

However, complexity in networks can be easily pushed towards chaos if the connectivity is increased (Kauffman, 1993), and the number of necessary regulatory genes required scales quadratically with the number of genes regulated (Croft *et al.*, 2003). Regarding this explosion, Mattick proposes that an extra layer of controling RNA molecules in complex organisms, is the one that makes possible to escape this problem (Mattick, 2005). The issue of non-coding RNAs is disscused in the following paragraphs.

With the technological advances providing information about the decoding of the genome, the challenge lies in integrating this vast amount of information in

order to extract principles and paradigms that might help us understand gene regulation on a level above the molecular.

Traditionally, a gene has been defined, depending on the field of study, as: an inheritable phenotype, despite of how the information and variation is encoded; or from the biochemical perspective, as a protein-coding region with some associated regulatory regions. Nevertheless, the evidence of ncRNAs being important for developmental and regulatory processes without encoding proteins has reshaped the concept of gene as a "transcription unit" (Okazaki *et al.*, 2002) or a "complete chromosomal segment responsible for making a functional product" (Snyder and Gerstein, 2003). Even this definition could be erroneous if alternate promoters are taken into account with their corresponding splicing and polyadenylation signals.

In order to simplify the study of gene regulation, we stay with a more conservative definition of gene with the characteristics described bellow. Gene expression is the process of reading and interpreting a given stretch of DNA to make a functioning protein. In principle, control of gene expression can take place at any of the intermediate stages:

- 1) transcription
- 2) RNA processing
- 3) mRNA transport
- 4) mRNA degradation, and
- 5) protein activity.

In practice, however, transcriptional control constitutes the most important level of gene regulation. A typical eukaryotic gene is structured as depicted in Fig. 5.

Its most important components are the coding region (coding for the protein), the promoter (at which RNA polymerase docks to read the coding sequence - a process called transcription), and various regulatory sequences. The regulatory sequences serve as binding sites for gene regulatory proteins (transcription factors), whose presence on the DNA affects the rate of transcription initiation. These sequences can be located not only adjacent to the promoter, but also far upstream - and even downstream - of it. As proteins, transcription factors are themselves subject to the gene regulatory processes outlined above. It follows

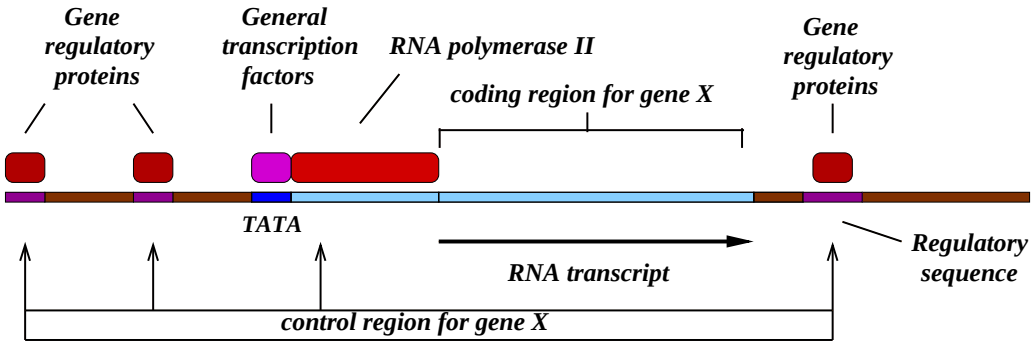


Figure 5: Simplified eukaryotic gene. Regulatory regions can be found far up or downstream from the coding region. Regulatory proteins may travel to other zones of the genome to interact with different genes.

that gene expression control systems typically take the form of networks of transcription factors that regulate each other Fig. 6.

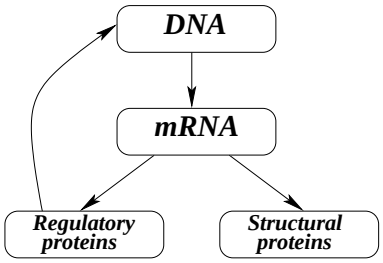


Figure 6: Architecure of decoding of regulatory proteins and structural proteins acording to a simplified model of gene regulation.

The first model of gene regulation was proposed by Jacob and Monod in 1961 (Jacob and Monod, 1961). The operon model consists of two classes of genes, the "structural" class, which encode the proteins that play some function in the metabolism of the cell; and, the "regulatory" class encoding for proteins which regulate the rate of transcription of other genes as transcription factors (TFs).

In bacteria, operons have been recognized and their interactions and function studied (Jacob and Monod, 1961). One example of an operon in *Escherichia colli* is the *lac operon* (Fig. 7), which controls the metabolism of lactose in the cell. It consists of three structural genes which are transcribed as long as the repressor is not bound to the *operator*. The repressor is encoded by a gene called I and it is located just upstream of the promoter. In the presence of lactose, the repressor binds to it leaving the operator and causing the transcription of the structural genes. This mechanism is often complicated by corepressors and

other interactions, but the main idea is captured by the structural/regulatory characterization of the operon model.

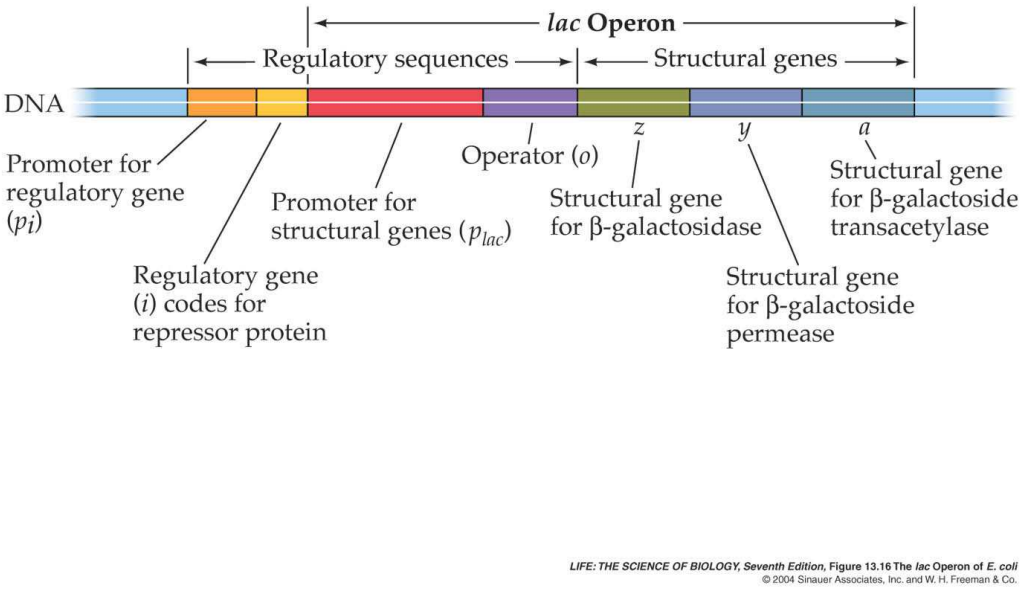


Figure 7: The *lac operon* in *Escherichia colli*. Regulatory and structural proteins encoded in the genome.

Davidson's studies on the see urchin development process propose a different way of approaching the relation between genes and their regulation. According to (Davidson *et al.*, 2003), the nodes of a regulatory network should be seen as a gene encoding a transcription factor or signaling component together with a regulatory module controlling the expression of the same gene.

Large gene networks are increasingly thought of as being built from smaller sub-network modules. It is thus important to understand the structure and dynamics of small functional building blocks (François and Hakim, 2004).

Networks can be induced to switch between different states, oscillate or simply rest in a fixed state. This behavior depends on the quantitative and qualitative interactions between its components. Important examples in biology are found to switch between two stable states triggered by external influences (Widder, 2003).

To attain this behavior, repression described by simple Michaelis-Menten kinetics is not sufficient to produce a working switch. The high-order Hill functions are required with, for instance, protein dimers or higher multimers interacting with DNA. When considering an existing gene regulatory network or the design of a new one, it would be useful to know whether a bistable switch can be made only out of two mutually repressing transcription factors or whether other interaction networks, less easily conceived, could serve the same purpose, perhaps even in a better and more robust way (Cherry and Adler, 2000).

In (François and Hakim, 2004), François and Hakim present an algorithm capable of creating a variety of small networks which behave in a prescribed manner. The design of the different networks is found via an evolutionary algorithm, while the only imposition is the search for bistable switches and oscillating networks. An important result from this study is the crucial use of posttranscriptional interactions, in fact, the behavior of the networks could not be understood from the transcriptional interactions alone (François and Hakim, 2004). It is also of great importance the diversity of topologies found by the algorithm. Simple architectures were already proposed for this kind of behavior, nevertheless, in some cases the result of the simulations reflected more stability and robustness than in the simple predefined networks.

Another remark on regulation regards the importance of RNA as a principal player in this process. Large amount of DNA which was before thought as "junk" DNA has been recently found to contain noncoding regions transcribed into RNAs, same that if damaged produces problems in development. This suggests that the specific content of this "junk" DNA is significant (Mattick, 1994).

One important example of RNA regulation are the *Riboswitches*. Riboswitches are structure that form in *mRNA* and regulate gene expresion in bacteria (Vitreschak *et al.*, 2004). Riboswitches has been shown to regulate several metabolic pathways involved in the biosynthesis of vitamins, amino acids and purines (Vitreschak *et al.*, 2004). The basic mechanism of regulation consists on two alternative secondary structures for a given sequence. One of these configurations is stabilized by a ligand, and a hairpin is formed which terminates transcription or binds directly to the ribosom-binding site thus inhibiting translation. In the derepressing condition, the riboswitch is not bound by the ligand, and an alternative srtucture folds which does not repress transcription. Riboswitches are of great

importance because of the diversity of organisms in which can be found. The most diverse distribution is that of *THI*-elements which are foudn in eubacteria, archaea and eukaryotes. This feature and the diversity of molecular mechanisms of regulation suggest that riboswitches are one of the oldest regulatory systems (Vitreschak *et al.*, 2004).

This is specially notorious in procaryote genomes, around 97-98 percent of the transcribed human genome is non-protein-coding RNA (ncRNA) (Mattick, 2001). Many of the microRNAs, a special sort of short ncRNAs, have been found in "intergenic" regions that were not considered as being transcribed (Lau *et al.*, 2001). The ncRNA molecules although not translated, play an important role in regulation (Ambros, 2000).One of the characteristics that make ncRNAs important is the evidence of being developmentally regulated, i.e. they are expressed differently according to the gender, tissue or cell on which they are found. For example, four of the seven major transcripts found in the *bithoraxabdominalAB* region do not code for proteins but are regulated and the alteration of the DNA that encodes them results in phenotypic consequences (Mattick, 2003).

A large quantity of RNA is transcribed from the genome, and only a small fraction of this mRNA is translated into proteins. Since noncoding regions and introns are characteristic from higher order organisms, it has been proposed that regulatory RNA may be at the base of the regulation processes and a necessary characteristic for complexity and diversity as found in eukaryote (Mattick, 1994). Some of the phenomena involving RNA include: co-suppression, transcriptional and post-trasncriptional gene silencing, and dsRNA-targeted mRNA destruction, in plants, or transcriptional silencing via DNA methilation (Sharp, 2001). RNA has also the property of binding to some proteins thought to be transcription factors, which have high affinity for RNA and other structures that contain RNA (Mattick, 2003). Direct functions of ncRNA have only recently been studied, nevertheless, some of them have been predicted as for example chromatin modification and epigenetic memory, transcription initiation or alternative splicing. Of course one of the main tasks attributed to ncRNAs is the constitution of an extensive but (yet) unrecognized regulatory network within higher organisms (Mattick, 2003).

It is worth noticing that these networks' functionallity relies not only on the connections made via the different transcription factors and regulatory sites, but

also by a very well defined spatial and chronological succession of regulatory states which give rise to shape and cell specification during the organism's development ((Davidson, 2001), (Davidson *et al.*, 2003)).

In order to study the evolution, characteristics and behavior of regulatory networks two approaches can be distinguished: a generative approach, which tries to infer principles and rules from theoretical and computational models that are constructed with no particular model system in mind; and, an analytic approach, in which a particular, known, gene regulation system is modeled. It is clear that most of the studies lay somewhere between these two categories (Reil, 2000).

2.5 Looking for answers in the origins

Complexity of the living organisms we know nowadays makes it too difficult to understand the main features of genotype-phenotype maps. Therefore, answers to this question may be found in the studies of origin of life and the theories around them.

It is clear from what is known from molecular evolution, genetics, gene regulation and development, that the step from molecular systems to living beings is one of the most striking and intellectually demanding questions in biology. Linking these two levels and defining the moment where life can be called that, has been approached from many points of view, ranging from the origins of life to the search of a simplified cell which is fully understood (Rasmussen *et al.*, 2004).

The studies of molecular evolution set the scenario for one of the most accepted theories on the origins of life. The RNA World hypothesis (Gilbert, 1986; Gesteland and Atkins, 1993) proposes a self-contained biochemical system preceding the origin of modern cellular life-forms, in which RNA molecules act both as genetic material and as enzymes (Orgel, 1998). The possibility of an RNA World depends on the capability of the RNA molecules to catalyze the chemical reactions necessary to replicate RNAs (Bartel and Unrau, 1999). This scenario is supported both by the wide range of catalytic activities that can be realized by relatively small ribozymes (Illangasekare and Yarus, 1999; Johnston *et al.*, 2001; Joyce, 2002; Lee *et al.*, 2000; Unrau and Bartel, 1998), and by the usage of RNA catalysis at crucial points in modern cells (Jeffares *et al.*, 1998; Doudna and Cech, 2002; Moore and Steitz, 2002). Plausible ribozyme catalyzed pathways for a late-stage

ribo-organism are discussed in (Joyce, 2002), the role and evolution of co-enzymes in a putative RNA world is explored in (Jadhav and Yarus, 2002). While the template-induced synthesis of oligonucleotides from smaller oligonucleotide precursors was successfully demonstrated in the laboratory (von Kiedrowski, 1986; von Kiedrowski *et al.*, 1989; Sievers and von Kiedrowski, 1994; Wlotzka and McCaskill, 1997), it seems impossible to replicate longer sequences without an enzyme (Orgel, 1998). Approaches to engineering a ribozyme-replicase have been very promising (James and Ellington, 1999; Johnston *et al.*, 2001; Paul and Joyce, 2003). These experiments show that self-replication is most likely within the catalytic repertoire of nucleic acids (McGinness and Joyce, 2003). So far, however, they have not resulted in an RNA ribozyme that can catalyze its own replication with an efficiency that could have sustained a genetic system on the early Earth.

A central issue in models of prebiotic evolution is the integration of information that is necessary to bridge the gap between a simple system of replicating molecules and the complexity of a modern cell (Eigen and Schuster, 1979; Kauffman, 1993). The template length is limited by the accuracy of the replication mechanism, which is necessarily error-prone due to mutations (Eigen, 1971). In principle the error threshold can be circumvented by evolving more accurate replicases that could be encoded by longer sequences (Scheuring *et al.*, 2003; Poole *et al.*, 1999; Szabó *et al.*, 2002). Such a bootstrapping mechanism, however, requires a functional replicase-ribozyme to start with. By comparison with known ribozymes such a molecule would probably be about 100nt long, while the current limit for non-catalyzed replication is less than 20nt.

An alternative mechanism that allows the accumulation of heritable information is the cooperation of self-replicators, introduced in the *Hypercycle model* (Eigen and Schuster, 1979). It was soon noticed, however, that hypercycles and similar models are vulnerable to various kinds of parasites in homogeneous solution (Maynard Smith, 1979; Bresch *et al.*, 1980). Not surprisingly, the number of coupled replicators increases only very slowly in models of self-replicators with mutation (May and Nowak, 1994; Happel and Stadler, 1998).

The shape of the fitness function, and more generally the accessibility of mutants from a given population, crucially influences the dynamics of evolution (Schuster *et al.*, 1994; Fontana and Schuster, 1998; Stadler *et al.*, 2001a). In the case of RNA it has been demonstrated that the genotype-phenotype map is dominated

by so-called neutral networks that percolate through sequence space, thereby allowing efficient exploration by means of neutral drift confined to the neutral networks (Schuster *et al.*, 1994; Huynen *et al.*, 1996; Huynen, 1996). More of the characteristics of this specific system are presented in the next chapter. Recently, it was shown that a similar mechanism allows population of autocatalytic self-replicators to explore sequence space in a diffusion-like manner (Stadler, 2002a).

Simple finite population models of hypercycles have been considered e.g. in (Andrade *et al.*, 1991). For larger, not necessarily hypercyclic, networks destabilization in homogeneous solution has been observed as a consequence of stochastic fluctuations (Nuño and Tarazona, 1994). The only study of sequence evolution of a hypercycle based on an explicit genotype-phenotype map can be found in (Forst, 2000), which concentrates in short cycles in a homogeneous medium.

2.6 On coevolving species

General characteristics of genotype-phenotype maps can be expressed by means of the behavior of populations under mutation and selection. Even when there is not a complete theory of fitness landscapes, a few remarks can be done by observing those already studied. One main difference among these studies is whether interactions among coevolving species are taken into account.

In almost any system with a non trivial landscape, species will only aspire to find local maxima. Whenever a part of a population sits on a maximum, the only possibility to move with smooth searching algorithms is to places of smaller fitness values. An equilibrium between many local maxima in a cohabited habitat would be the best option (Bak, 1996). At the same time, changing fitness landscapes due to interactions between species may allow changes in phenotype that would otherwise imply a reduction in fitness and therefore impossible to find in classic Darwinian evolution. Once a fitness peak is found in a certain configuration of the interactions network, a change in these conditions may reduce the fitness value at this peak, creating new ones which will be reachable via mutations and natural selection. This way, evolution is further impulsed by a changing environment consisting of both climate or geographical characteristics and interacting species.

It is clear then, that a more realistic and broad concept of the genotype-phenotype-fitness map would be one taking into account cooperativity and strug-

gle between species, as well as influence from the environment in the population. The fitness value assigned to an individual becomes then relative to those properties of other organisms that may influence the behavior of a single one.

Wills points out in (Wills, 2001) that if the process of autocatalysis involves more than one component, it is difficult to delineate where is the information carried. Moreover, in this case fitness cannot be assigned to an individual but to a set of interacting components (Wills, 2001).

The importance of interactions among species is remarked by Bak in (Bak, 1996) saying that: "In the absence of interaction between species, evolution would come to an abrupt halt, or never get started in the first place". Kauffman also makes the point with the concept of "interacting dancing fitness landscapes" (Kauffman, 1993) referring to interacting/coevolving species. Further studies of Kaufman in this direction showed that a highly interconnected system falls easily into a chaotic state, where species do not have the time to reach fitness peaks before the landscape changes again. There is no real evolution in this case since any improvement is lost before optimization is permitted (Kauffman, 1993).

2.7 Self-organization

Living systems are the paramount examples of organized complexity. From the genetic expression and regulation, to the neural networks of the nervous system, self-organization is at the core of these systems (Kauffman, 1993). Many questions arise about the origin of such systems and the evolution which took place in order to generate them. It is important to ask whether all systems are suitable to accumulate beneficial mutations, and if selection is capable of bringing systems to this regime.

In order to answer these questions, Kauffman writes in (Kauffman, 1993):

> The task must be to include self-organizing properties in a broadened framework, asking what the effects of selection and drift will be when operating on systems which have their own rich and robust self-ordered properties. [...] It seems preeminently likely that what we observe reflects the interactions of selection processes *and* the underlying properties of the system acted upon.

Dynamical systems are the main tool for studying self organization. Different kinds of attractors constitute the possible alternatives for the long time behavior of the variables. Between oscillating and random behaviors, strange attractors may arise, which are sensitive to initial conditions, meaning any two arbitrarily close points will, after a sufficient period, become as far apart as desired in the attractor. Attractors are usually low dimensional even when exhibit in high dimensional spaces and are found in chaotic systems.

When studying the interaction between species, one well known model is that of boolean networks: sets of species are represented as graphs and joined by edges whenever two species interact. In general terms, three states are visited by boolean networks: the ordered, the complex and the chaotic regimes. Ordered networks exhibit percolation of frozen regions to the whole space, while chaotic ones are the opposite, presenting only islands of frozen behavior. The complex regime, the most interesting for living systems, lays just in between these two, presenting percolation of the frozen region together with isolated unfrozen regions. "Adaptive evolution achieves the kind of complex systems which are able to adapt [...]. These are those which live at the edge of chaos." (Kauffman, 1993). Thus, ecology must be situated exactly in the critical state separating both cases, the frozen one and the chaotic one. That is, at the phase transition.

Although this theory could give important information about the evolution and creation of life, it is far away from being a comprehensive and complete theory of the origin of life. Vast research must still be done in order to define and understand what the "edge of chaos" is and what are the characteristics that make this systems so complex and interesting.

In the case of gene regulatory networks, it is important that the right degree of connectedness is reached in order to have a system which is stable and flexible at the same time, allowing the organisms to adapt to and resist the changes in the environment.

3 Molecular Evolution

3.1 Population dynamics

Population genetics became the mathematical basis of the synthetic theory and is still seen by many biologists as the current frame for understanding evolution. It is based on the study of gene frequencies and their change over time due to natural selection ((Crow and Kimura, 1970), (Wright, 1931), (Lewontin, 1974)).

Population genetics saw a major extension by Motoo Kimura (Kimura, 1955) who introduced the idea of neutrality. This theory was further impulsed by results from comparative sequence analysis (King and Jukes, 1969) which showed that within epochs of phenotypic stasis, the changes in genotypes occur at rates which are as high as, if not higher than, those recorded during adaptive periods.

There are two main problems with this view of population genetics. The first is the fact that mutation is considered as some external event, which is not part of the regularly considered dynamics. The second has to do with the phenotype represented only by its fitness values and sometimes mutation rates which are assigned as parameters to the corresponding genotype.

As a response to this problems, Eigen published his work on self-organization of macromolecules in 1971 (Eigen, 1971), where replication and mutation are seen as parallel chemical reactions and evolution is visualized as a process in an abstract space of genotypes, called sequence space. In his studies, every RNA or DNA sequence is a point in sequence space and the Hamming distance induces a metric in this space. The temporal development of the distribution of genotypes in populations is described by the selection-mutation equation:

$$\frac{d\xi}{dt} = \dot{\xi} = \xi(Q_{ii}a_i - \Phi(t)) + \sum_{j=1, j\neq i} Q_{ij}a_j\xi_j \; ; \qquad i = 1, ...n \tag{1}$$

Where $\xi_i(t)$ are the frequencies of individual genotypes I_i at time t, $\Phi(t)$ takes care of the normalization of the frequencies and the square matrix $Q = Q_{ij};\ i, j = 1, ...n$ contains replication accuracies in the diagonal terms and mutation probabilities from species i to species j in Q_{ij}.

At sufficiently accurate replication, that means low enough mutation rates, populations modeled by eq. 1 approach stationary mutant distributions, called quasi-

species, which are centered around a most frequent genotype, the master sequence.

At rates above the threshold value, populations do not approach stationary states but drift randomly through sequence space and genetic information is lost. Evolution is confined to mutation rates between a lower and an upper limit: The lower limit is given by the maximal accuracy of the replication machinery and the upper limit is set by the maximal sustainable fraction of error copies determined by the error threshold (Eigen, 1971).

Fitness relevant properties of phenotypes in this model appear only as parameters of genotypes in the differential equations, for example a_i and Q_{ij} in eq. 1.

3.2 About RNA Secondary Structures

3.2.1 RNA Secondary Structures and Their Prediction

As with all biomolecules, the function of RNAs is intimately connected to their structure. While successful predictions of RNA tertiary structure remain exceptional feats, RNA secondary structures can be predicted with reasonable accuracy, and have proved to be a biologically useful coarsegrained representation of the tertiary structure.

A secondary structure of a given RNA sequence is the list of (Watson-Crick and wobble) base pairs satisfying two constraints:

(i) each nucleotide takes part in at most one base pair, and

(ii) base pairs do not cross, i.e., there are no knots or pseudo-knots.

The restriction to knot-free structures is necessary for efficient computation by means of dynamic programming algorithms ((Hofacker *et al.*, 1994), (Hofacker *et al.*, 2002), (Wuchty *et al.*, 1999)) . The memory and CPU requirements of these algorithms scale with sequence length n as $O(n^2)$ and $O(n^3)$, respectively, making structure prediction feasible even for large RNAs of about 10000 nucleotides, such as the genomes of RNA viruses (Witwer *et al.*, 2001). There are two implementations of variants of these dynamic programming algorithms: the `mfold` package by Michal Zuker, and the the `Vienna RNA Package`. The latter is freely available from http://www.tbi.univie.ac.at/ and is used throughout this

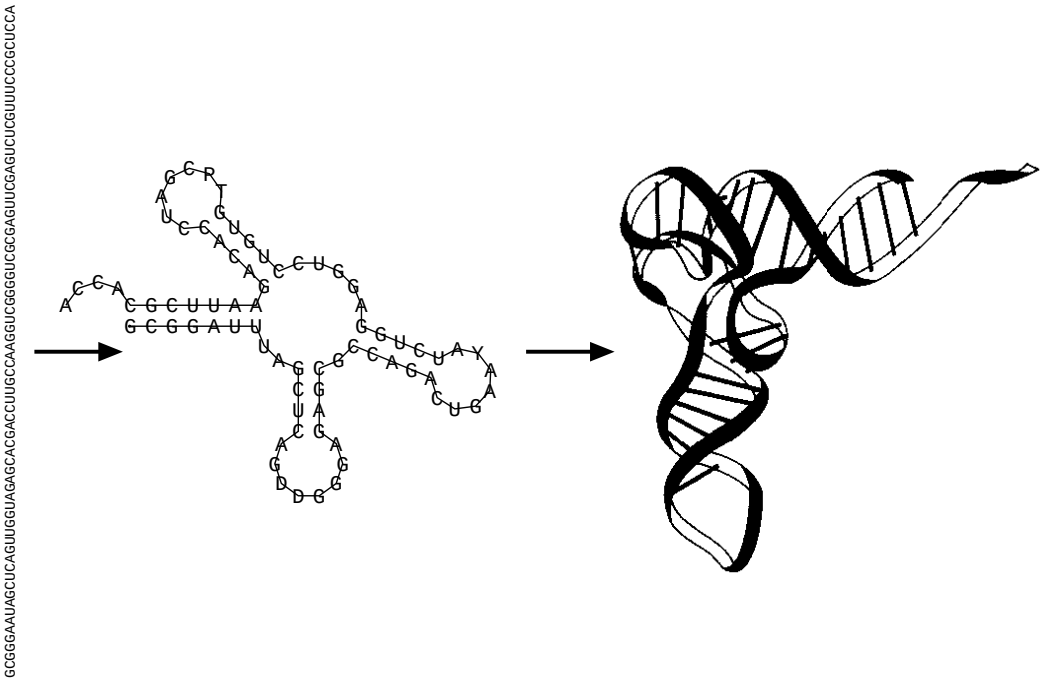


Figure 8: Schematic representation of the stages of RNA folding into the tertiary (three dimensional) structure.

work.

These thermodynamic folding algorithms are based on an energy model that considers additive contributions from stacked base pairs and various types of loops, see e.g. (Mathews *et al.*, 1999). Two widely used methods for determining such nucleic acid thermodynamic parameters are absorbency melting curves and microcalorimetry, see (SantaLucia Jr. and Turner, 1997) for a review.

3.2.2 Neutral Networks in Sequence Space

A more detailed analysis of functional classes of RNAs shows that their structures are very well conserved while at the same time there may be little similarity at the sequence level, indicating that the structure has actual importance for the function of the molecule.

In the RNA case, the genotype-phenotype map can be approximated by the folding process of the molecule, where the sequence of nucleotides is interpreted as "genotype" and the minimum free energy structure (mfe) as "phenotype", see e.g. (Schuster, 2001) for a review. There is ample evidence for redundancy in

genotype-phenotype maps in the sense that many genotypes cannot be distinguished by an evolutionarily relevant coarse grained notion of phenotypes which, in turn, give rise to fitness values that cannot be faithfully separated through selection.

Regarding the folding algorithms as a map f that assigns a structure $s = f(x)$ to each sequence x we can phrase our question more precisely: We need to know how the set of sequences $f^{-1}(s)$ that folds into a given structure s is embedded in the sequence space (where the genotypes are interpreted as nodes and all Hamming distance one neighbors are connected by an edge). The subgraphs of the sequence space that are defined by the sets $f^{-1}(s)$ are called neutral networks.

Theory predicts a phase transition like change in the appearance of neutral networks with increasing degree of neutrality at a critical value depending on the size of the genetic alphabet. If the fraction of neutral neighbors is less than this threshold, the network consists of many isolated parts with one dominating giant component. On the other case, the network is generically connected. The critical value is the connectivity threshold. This property of neutral networks reminds of percolation phenomena known from different areas of physics, although the high symmetry of sequence space, with all points being equivalent, introduces a difference in the two concepts. A series of computational studies ((Fontana *et al.*, 1993), (Schuster *et al.*, 1994)) has in the last decade drawn a rather detailed picture of the genotype-phenotype map of RNA.

- (i) **More sequences than structures**. For sequence spaces of chain lengths $n > 10$ there are orders of magnitude more sequences than structures and hence, the map is many-to-one.
- (ii) **Few common and many rare structures**. Relatively few common structures are opposed by a relatively large number of rare structures, some of which are formed by a single sequence only ("relatively" means here that the number of both common and rare structures increases exponentially with n, but the exponent for the common structures is smaller than that for the rare ones).
- (iii) **Shape space covering**. The distribution of neutral genotypes is approximately random in sequence space. As a result it is possible to define a spherical ball, with a diameter being much smaller than the diameter n

of sequence space, which on the average contains at least one sequence that folds into every common structure.

- (iv) **Existence and connectivity of neutral networks**. Neutral networks of common structures are connected except in cases with specific non-random distributions of the alphabet A, C, G, U. Neutral networks for the RNA-folding map show a percolation-like behavior.

Shape space covering, item (iii) above, is a consequence of the high susceptibility of RNA secondary structures towards randomly placed point mutations. Computer simulations ((Fontana *et al.*, 1993), (Schuster *et al.*, 1994)) showed that a small number of point mutations is very likely to cause large changes in the secondary structures: mutations in 10 percent of the sequence positions already lead almost surely to unrelated structures if the mutated positions are chosen randomly.

The set of nodes of the neutral network is embedded in a compatible set $C(s)$ which includes all sequences that can form the structure s as suboptimal or minimum free energy conformation. Sequences at the intersection between the compatible sets of two neutral networks in the same sequence space, $C(s_0)$ and $C(s_1)$, are of actual interest because these sequences can simultaneously carry properties of the different RNA folds.

It is known ((Reidys *et al.*, 1997)) that each two sets of compatible sequences with respect to the pair of secondary structures have a nonempty intersection. It turns out that the intersection is of particular relevance for transitions of finite populations optimizing fitness (Forst *et al.*, 1995a) since individual sequences folding into the intersection may take the population to structures of increased fitness.

A very good example for this, show sequences which have been evolved to exhibit catalytic activities of two different ribozymes at the same time (Schultes and Bartel, 2000). As was mentioned before, the intersection theorem ((Reidys *et al.*, 1997)) states that for all pairs of structures s_0 and s_1 the intersection $C(s_0) \cap C(s_1)$ is always non-empty. In other words, for each arbitrarily chosen pair of structures there will be at least one sequence that can form both as minimum free energy or suboptimal configuration. If s_0 and s_1 are both common structures, bistable molecules that have equal preference for both structures are easy to

design ((Flamm *et al.*, 2000), (Höbartner and Micura, 2003)).

A particularly interesting experimental case is described in (Schultes and Bartel, 2000). At least, the features (i), (ii), and (iv) of the neutral networks of RNA seem to hold for the more complicated protein spaces as well, see e.g. (Keefe and Szostak, 2001) for experimental data. The impact of these features on evolutionary dynamics is reflected in the fact that a population explores sequence space in a diffusion-like manner along the neutral network of a viable structure. Fast diffusion together with perpetual innovation makes these landscapes ideal for evolutionary adaptation (Fontana and Schuster, 1998) and sets the stage for the evolutionary biotechnology of RNA.

3.3 Properties of the gen-phen map in the RNA model

Due to the fast computation of the secondary structure of an RNA sequence, this map has been largely studied and some of its properties have been formulated in general evolutionary terms.

The energy landscape of a sequence is the RNA analogue of Waddington's developmental or epigenetic landscape. (Waddington, 1957). Sequences folding into the same mfe shape can differ profoundly in their energy landscapes. In this limited sense, the RNA model is capable of mimicking an "evolution of development" (Fontana, 2002). The analogy breaks down when the mechanisms of development themselves evolve. Evolvability and variability, for example, are characteristics also encoded in the genome and regulate the process of development. These features can change along time and evolve depending on external and internal conditions.

Plasticity in the framework of RNA folding may be understood in two different ways: the so called *norm of reaction* (Scheiner, 1993) refers to persistent phenotypic transformation due to changes in the environment; on the other hand, *intrinsic* plasticity is induced by molecular energy fluctuations at non-zero temperature. The first definition is reflected as transitions between mfe shapes as the free energy landscape is deformed by temperature, while plasticity understood as intrinsic phenotypic variance refers to transitions between different shapes on a fixed free energy landscape (Fontana, 2002).

As already mentioned, a neutral mutation is a nucleotide substitution that pre-

serves the mfe shape (but it may affect everything else, such as free energy, plastic repertoire and kinetic folding landscape). The neutrality of a sequence is the fraction of neutral (one-error) neighbors. Neutrality is here defined with respect to mfe shape, not fitness. Fitness is a function from phenotypes to numbers and if phenotype is defined as mfe shape, then neutrality extends to fitness as well. If phenotype and fitness are defined in terms of the plastic repertoire of a sequence, sequences that share the same mfe shape are taken as neutral, even when their plastic repertoires (and fitness) differ.

Epistasis means that the phenotypic consequences of a mutation at gene i depend on the genetic background provided by the remaining genotype. This dependency is mediated by networks of interactions among gene products. The same concept applies to RNA, when substituting "gene" with "sequence position" (Fontana, 2002). The transparency (but also the limitation) of the RNA genotype phenotype model derives from the identity of epigenetic and epistatic interactions, since phenotype is defined directly in terms of interactions among sequence positions. A mutation changes the base pairing possibilities of a sequence and hence the network of epistatic interactions. The mfe shape shown at the top left of Fig. 9 remains the same if C is substituted by G at the position labeled x. Yet, whether x is C or G determines which mfe shape is obtained as a result of mutating position y from U to C. More subtly, the neutral substitution from C to G at x alters the number and identity of neutral positions.

The tendency of a sequence to adopt a different shape upon mutation (variability) is a prerequisite for its capacity to evolve in response to selective pressures (evolvability). In this sense, variability underlies evolvability. Variability (quantified as the number of nonneutral neighbors) is sequence dependent. Variability can therefore evolve. Canalization (Waddington, 1942) is a biological concept related to robustness in physics and engineering aimed at quantifying a system's resilience to perturbation. Biologists distinguish between environmental and genetic canalization, depending on the nature of the perturbation. In our highly simplified RNA context, genetic canalization is phenotypic robustness to mutation and environmental canalization is phenotypic robustness to environmental change or noise. Neutrality, as defined here, is basically a measure of genetic canalization, while plasticity is the converse of environmental canalization.

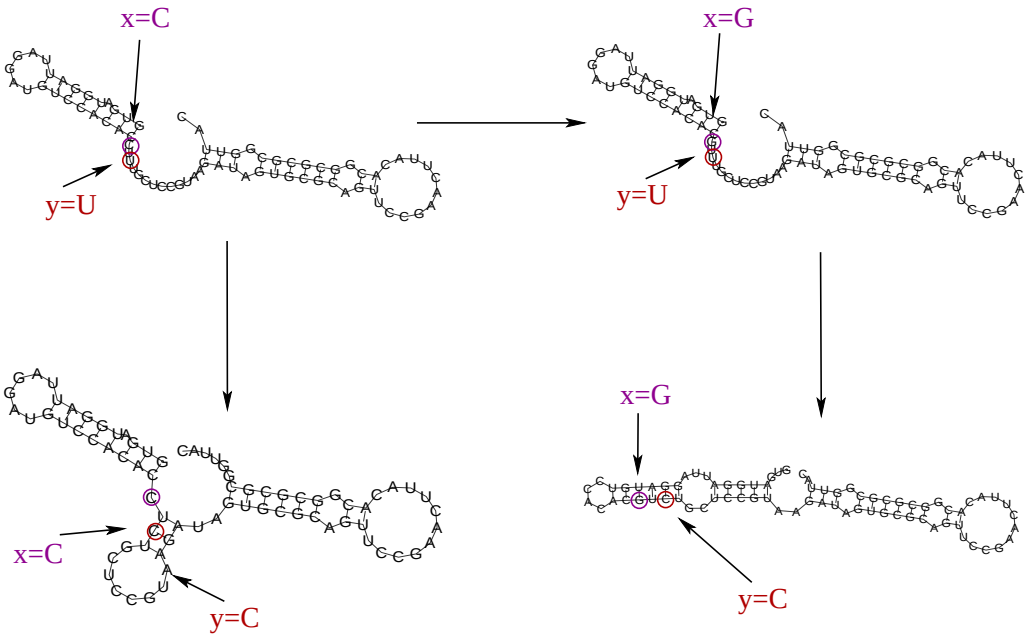


Figure 9: Epistasis in RNA folding. A first mutation in the position labeled by x has no impact on the secondary structure. Nevertheless, once this mutation is introduced, a second mutation in position y changes dramatically the shape of the structure.

3.3.1 Plastogenetic congruence and neutral confinement in RNA

Plastogenetic congruence is the direct relation of plastic accessibility and genetic accessibility. Ancel and Fontana [(Ancel and Fontana, 2000)], showed for the RNA sequence to secondary structure map that the set of shapes compatible with a given sequence is strongly correlated with the mfe structures of the one mutation neighbors of the sequence.

There are at least three reasons to confirm this fact. First, the frecuency of a structure in the one-mutation neighbors of a sequence is much higher if the same structure is present in the plastic repertoire than if it is not. Also, the minimum free energy structure of a sequence is present at high frequency in its one-mutation neighbors. And, structures in the plastic repertoire of a sequence with energy close to that of the mfe, are present as mfe in a neighborhood of on average 5 point mutations from the original sequence [(Ancel and Fontana,

2000)].

From this assumption it is clear that any suboptimal structure with non-neglectible probability will be realized in the one-mutation neighborhood of the sequence. This could represent an evolutionary advantage, as long as the new mfe has a better fitness. Nevertheless, by linking the reduction of plasticity to the reduction of genetic variability, plastogenetic congruence strongly affects evolution. If neutrality is reflected as robustness against mutations and platicity as robustness against thermodinamical perturbations, then genetic canalization is a direct consequence of environmental canalyzation. Plasticity is costly because many subotpimal structures mean less time in each of them, no matter which one has higher fitness. Natural selection then reduces plasticity and therefore evolvability, in the sense that the set of shapes attainable in the close veicinity of a sequence will be decreased. If the probability of finding new advatageous shapes depend on the neutrality of a sequence and this in turn on its platicity, then evolution will be eventually halted and the population "neutrally confined" [(Ancel and Fontana, 2000)].

3.4 About cofolding and its properties

In recent studies, simple models of strongly interacting RNA molecules have been studied in which selection for a common resource is replaced by frequency-dependent fitness terms. In these models, each RNA species depends on the presence of specific catalysts. A prime example of this class of models is the hypercycle model of interacting replicators (Eigen and Schuster, 1979). While such a system has not (yet?) been realized experimentally, there has been substantial progress in constructing RNA replicase ribozymes. We refer to (McGinness and Joyce, 2003) for a description of the state of the art. It is thus worthwhile to study the evolutionary properties of such models.

In (Stadler, 2002a), the diffusion (in sequence space) of a population of interacting replicators is studied, where the replication rates depend only on the sequence similarity of the parent molecules. A model of hypercycles with interactions depending on the secondary structures of the individual RNAs is described in (Forst, 2000) and late in more detail in (Stephan-Otto Attolini and Stadler, 2004). In the latter contribution we emphasize the importance of the neutrality of the

genotype-phenotype map in order to maintain the hypercycle and at the same time have diffusion in sequence space.

In previous work, the basic assumption was that the action of each RNA molecule is determined by its own secondary structure. For example, the replication rate of sequence x under the influence of sequence y as catalyst is $a_{xy} = a(f(x), f(y))$, i.e, a function of the (ground state) secondary structure of both molecules.

The common secondary structure $f(x \circ y)$ of two RNA molecules can be computed using a simple extension of the usual dynamic programming algorithms for computing RNA secondary structures, see e.g. (Hofacker *et al.*, 1994; Dimitrov and Zuker, 2004). The basic idea is to compute the secondary structure of the concatenated RNA sequences $x + +y$ (or $y + +x$), where the "loop" that contains the split between x and y does not contribute to the folding energy. We use the program `RNAcofold` implemented in the `Vienna RNA Package` (Hofacker *et al.*, 1994; Hofacker, 2003). Figure 10 shows the secondary structures of two molecules and the combined structure with intermolecular base-pairs computed with `RNAcofold`. If the ground structure is unique then $f(x \circ y) = f(y \circ x)$, otherwise the structures will in general be different since the backtracking routine implemented in `RNAcofold` yield one of the group state structure in a deterministic way.

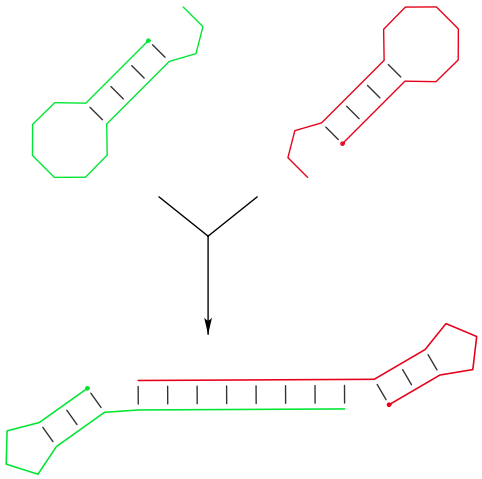


Figure 10: Individual folding of two sequences followed by the combined structure found by *RNAcofold*.

The concept of two sequences interacting to form a single secondary structure could be generalized to many sequences. In Fig. 11, we show the difference between folding one, two or three sequences together. Starting with a closed sequence, single fold is obtained by making one cut. The graph represents the

bases as dots and the pairing with edges between them. If one more cut is made to the original sequence, the case of cofold with two sequences is obtained. It is clear that the order on which the sequences are ordered makes no difference as to the structure found. The requirement of avoiding pseudo-knots in secondary structures is transformed into non-crossing edges in this representation. As the third cut is made, the ordering of the sequences takes importance, as different structures could be found depending on it. In our example, one ordering will result in the crossing of edges (pseudo-knot) and thus in a forbiden configuration. The combinatorial possibilities explode as the number of sequences to fold is augmented.

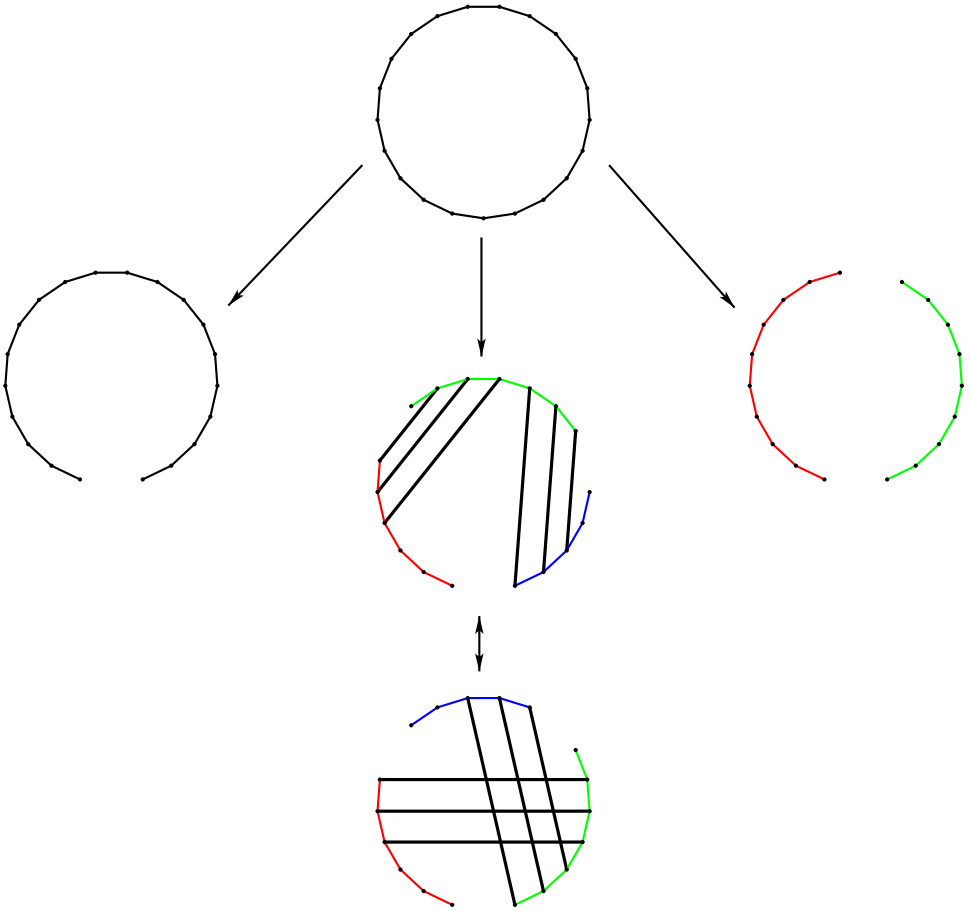


Figure 11: Relation between the folding of one, two or three sequences. The ordering of the sequences is important if more than two are cofolded, the possible structures grow as the number of sequences is increased.

3.4.1 Measuring neutrality in Cofold

In the next section we present a model where we explore the situation where the rate of replication of a molecule's replication catalyzed by another molecule is a function of the structure of the interaction complex of the two secondary structures, i.e. $a_{xy} = a(f(x \circ y))$. To this end, we study in detail the statistical properties of the *RNA co-folding map* $f : (x, y) \mapsto f(x \circ y)$ which assigns to each pair of RNA sequences the secondary structure of their thermodynamically most stable co-folding.

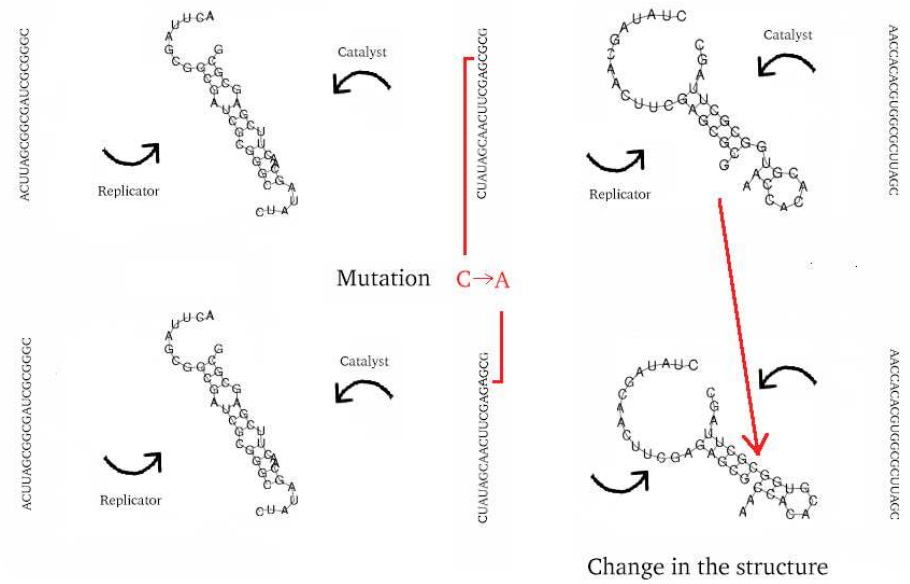


Figure 12: One sequence folding with two more. If one sequence is required to interact with two more, neutrality is defined by those mutations which preserve both structures after the mutation. In this case one of the structures is modified and so the mutation is not considered as neutral.

In the following we will study two different versions of defining neutrality in a cofold map:

1. We say that a mutant x' of x is neutral when $f(x' \circ y) = f(x \circ y)$ for a given partner sequence y. This scenario corresponds to RNA switches or RNAs that bind to target molecules in a specific way, e.g. microRNAs (Rehmsmeier *et al.*, 2004).
2. We say that a mutant x' of x is neutral when $f(y \circ x') = f(y \circ x)$ and

$f(x' \circ z) = f(x \circ z)$. This scenario corresponds e.g. to an RNA hypercycle: the mutant x' simultaneously must be a template (and hence retain the structure of its complex with the catalyst z), and a catalyst (and hence be able to replicate the template y) (See Fig. 12).

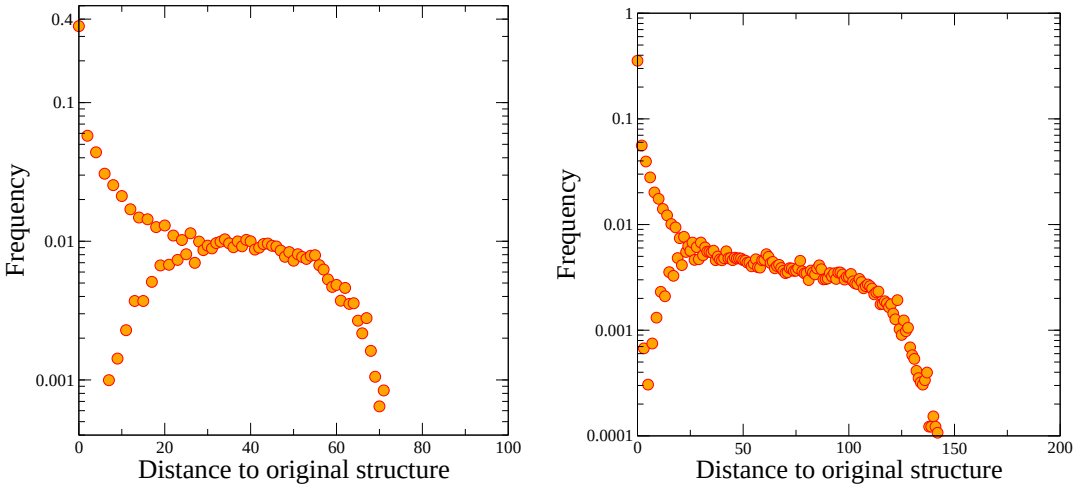


Figure 13: Distribution of distances from the original structures to the new complexes after point mutations. (a) Point mutations. Data extracted from 600000 sequences of length 100. Fraction of neutral mutations: $\bar{\lambda} = 0.32$. (b) Compensatory mutations: 1000 sequences of length $n = 100$ cofolded with a fixed one of length also $n = 100$. On average there are 66 possible compensatory mutations, out of which $\lambda^c = 0.35$ are neutral.

Two mutation operators are used in this model. First, we introduce point mutations, i.e. the change of a single base in the RNA sequence. The second, called "compensatory mutation", consists in the replacement of a base pair by any other of the Watson-Crick or wobble base pairs. The two bases involved are supposed to change simultaneously.

In the first case studied here, i.e., cofolding of the mutating sequence with a fixed partner, we consider both point and compensatory mutations. In order to obtain accurate statistics we compute all point mutations and all compensatory mutations using samples of 600000 and 1000 sequences, respectively. We use the symmetric difference of the set of base pairs as a measure for the structural distance of two RNA secondary structures.

This first case is similar to folding the concatenated sequence $f(x++y)$ instead

of the co-folding complex $f(x \circ y)$, the only difference being the energy contribution from the "exterior loop" that contains the split between the two sequences. Indeed, we observe neutral mutation rates $\bar{\lambda}$ similar to those reported in (Gruener *et al.*, 1996) for an individual RNA sequence.

The second case, where a sequence is mutated and cofolded with two different partners is more important e.g. in the context of models of prebiotic evolution, where a single sequence has to satisfy at least two different constraints: it have to be a recognizable template and it has to perform its catalytic function in two different contexts. In this case we sample in the following way. We randomly generate three different RNA sequences of the same length, x, y, and z. and compute $f(x \circ y)$ and $f(x \circ z)$. We then mutate x and recompute both cofolding structures and determine the distance form the original structures. In this case a compensatory mutation must be compensatory with respect to *both* $f(x \circ y)$ and $f(x \circ z)$, i.e., only base pairs shared by both cofolding structures are candidates for compensatory mutations.

We sampled approximately 300,000 point mutations for chain length $n = 50$, about 570,000 of length $n = 100$, and 450,000 of length $n = 200$. Furthermore 3000 triplets were constructed and compensatory mutations introduced for each of the three chain lengths.

In addition to estimating the fraction of neutral mutations, we also estimated the length of neutral paths (Schuster *et al.*, 1994). A neutral path $\mathcal{L}$ is defined as follows. Starting from a sequence x_0 a sequence of RNA sequences $\{x_i | i = 1, \dots\}$ is constructed such that (i) $f(x_i) = f(x_0)$, i.e., the structures do not change along the path, (ii) x_i is a point mutant or compensatory mutant of x_{i-1} and (iii) the Hamming distance from the starting point x_0 strictly increases with each step. The path terminates after at most n steps when no mutant can be found. The Hamming distance between x_0 and the last point in the path is the length L of the neutral path. Here we constructed 1200 neutral path for sequences of length $n = 100$. In the case of one sequence cofolding with two other sequences, the algorithm is basically the same except that compensatory mutations must be possible in both structures and only neutral mutations for both are accepted into the path.

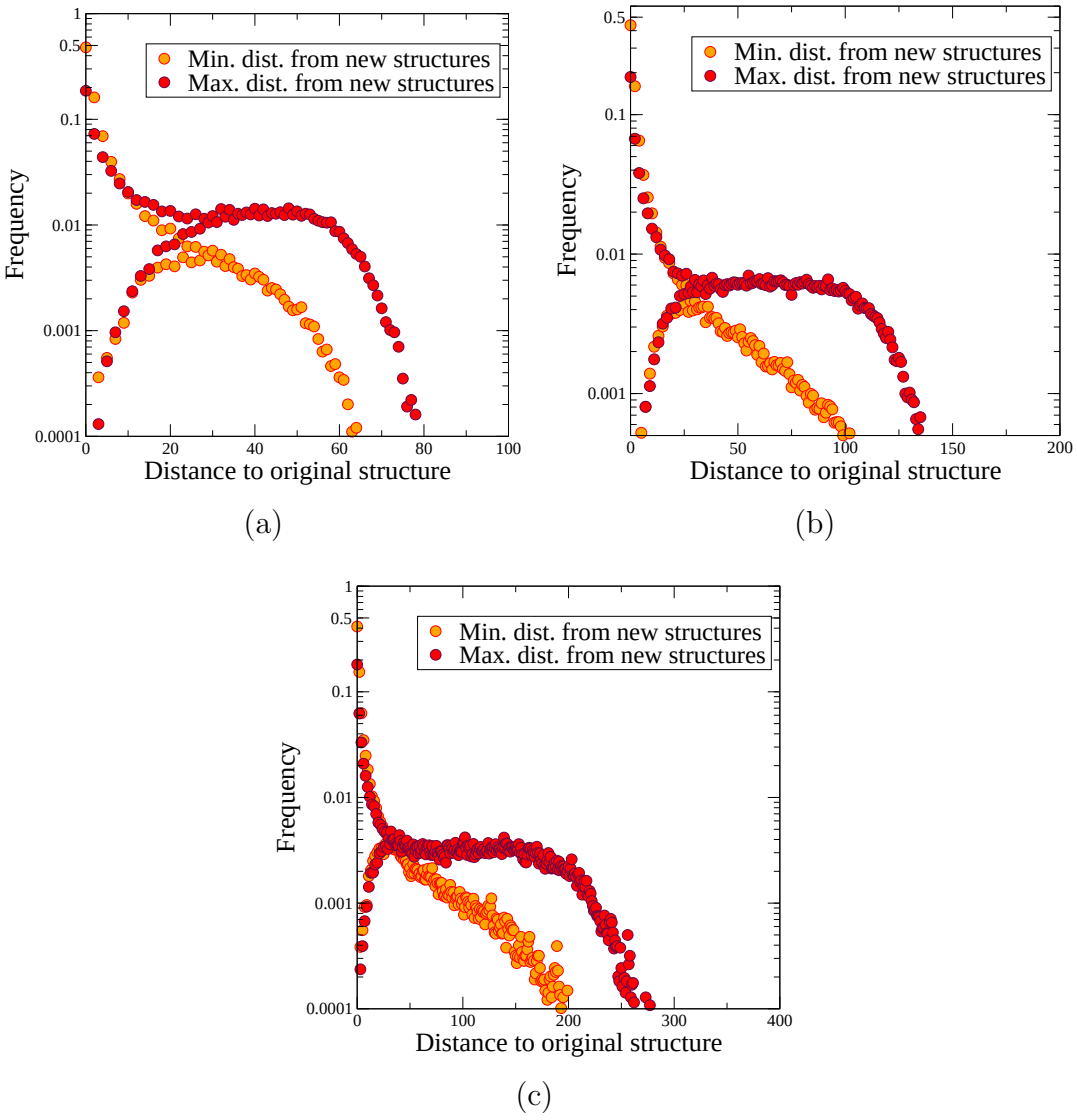


Figure 14: Distribution of distance to the original structure after point mutations for different sequence lengths. a)300,000 sequences of length 50. For point mutations, fraction of neutral mutations: 0.185. b)568,000 sequences of length 100. For point mutations, fraction of neutral mutations: 0.186. c)Length 200, 445,000 sequences tested. Fraction of neutral mutations: 0.18.

3.4.2 Results

The behavior of `RNAcofold` when taking into account only two sequences is very similar to that of `RNAfold` for a single RNA sequence of the same length (Hofacker *et al.*, 1994). The fraction of neutral point mutations is almost a third of the total. One difference from single fold is that almost no point mutations change all base pairs of the structure.

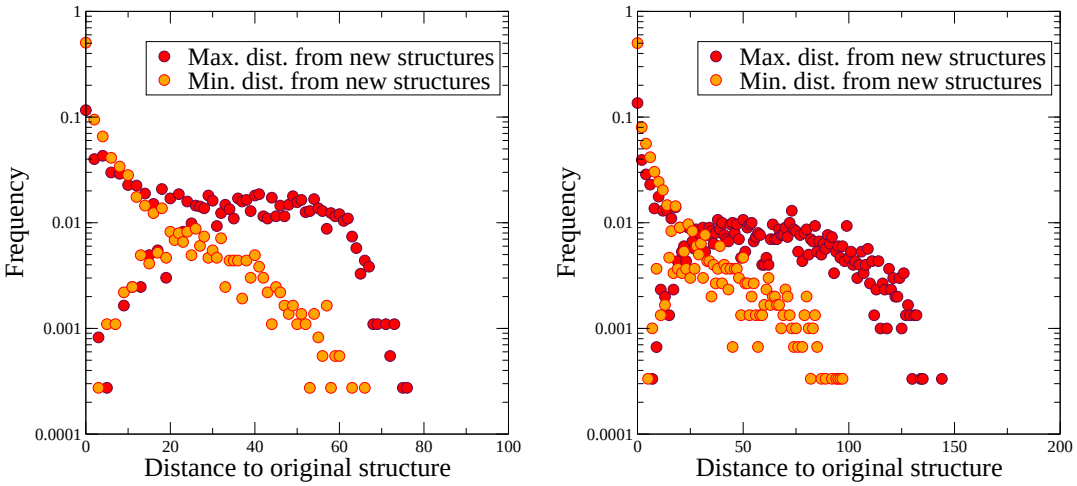
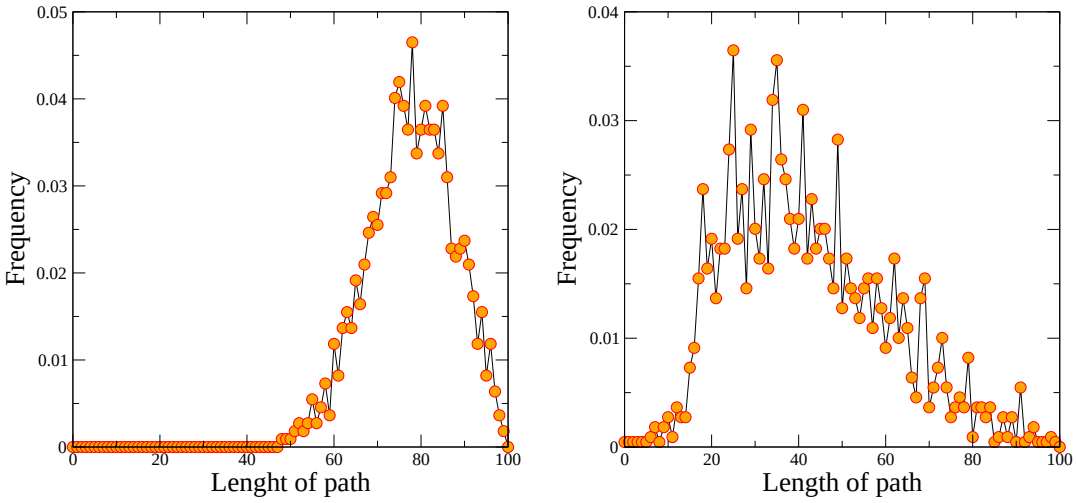


Figure 15: Distribution of distance to the original structure after compensatory mutations for different sequence lengths. a)3500 sequences of length 50 were used. b) 3000 sequences of length 100 were used. On average, we found only 15 possible compensatory mutations for both structures at the same time. Of these, only 0.15 resulted to be neutral.

In the case of compensatory mutations, the situation is different, since we allow mutations only in one of the two sequences, so that inter-molecular base pairs can only change from **GU** to **AU** or **CG** to **UG**. Therefore, two thirds of the possible compensatory mutations are not allowed anymore and neutrality is hardly affected by compensatory mutations: Only 35 percent of the remaining mutations are neutral. From (Schuster, 2001) we know that in order to change from one connected component to some other inside the neutral network, compensatory mutations may be needed. This is important from the evolutionary point of view since a fitter structure may be accessible from a particular connected component of the neutral network.

In the case of more than a single structural constraint, however, the situation becomes difficult. As shown in Fig 15b, the degree of neutrality is drastically decreased both for point mutations and for compensatory mutations. This fact is of crucial importance for models where cofold defines the interactions between RNA molecules.

Fig. 14 shows that neutral mutations occurring simultaneously for both cofolding structures are only about 18 percent of all possible mutations, i.e., less than two thirds of those that are present in single fold.



(a) (b)

Figure 16: Length of neutral paths for different sequence lenghts. a) Length of neutral paths for 1200 sequences of length 100 cofolded with only one fixed sequence. b) Length of neutral paths for 1200 sequences of length 100 when cofolding with two fixed sequences.

	Neutral mutations	Length of path
Single fold	0.33	100
Cofold with one sequence	0.32	75
Cofold with two sequences	0.18	40

Table 1: Summary of results. Fraction of neutral mutations for sequences of length 100 and typical path length for each case.

It is known that for single folding sequences, it is possible to exchange almost all nucleotides without leaving the neutral network (Hofacker *et al.*, 1994; Gruener *et al.*, 1996). In the case we study, the length of neutral paths when cofolding one sequence with one that remains fixed, is shorter than in single RNA fold. Since there are intramolecular base pairs, for some of these it would be impossible to find neutral mutations and therefore some bases will never change without leaving the neutral network. In Fig. 16a we show the results for 1200 sequences of length $n = 100$ cofolding with fixed sequences of the same length.

The length of the path when cofolding one sequence with two different interacting RNAs is much shorter than in the previous case and, of course, than in the case of folding an isolated RNA. Indeed, there are no paths along which all nucleotides of x could be replaced, Fig. 16b.

An overview of the results obtained in this study can be seen in Table 1. It is clear that whenever three sequences are involved, neutrality is largely lowered. For the purposes of modeling interactions of molecules via cofold, we can conclude that diffusion in such a model, will be much more difficult than in the case where single folding sequences are used.

3.5 Two models of molecular evolution

The replicator equation mentioned at the beginning of this chapter deals with the production of one molecule by self-replication or mutation of a different kind into the first one. The probability of mutation is usually modeled in pure mathematical terms, which means there is no relation to the actual chemical rates. In this general formulation, the equations do not take into account the possibility of one molecule's replication being catalyzed by another one. It is of crucial importance for the study of living systems to address this question since the mechanism of catalysis is the main process of molecular interaction in living organisms.

In this direction, the *second order replication equation*,

$$\frac{dx_k}{dt} = \dot{x_k} = x_k \left(\sum_{j=1}^{n} a_{kj} x_j - \sum_{i,j} a_{ij} x_i x_j \right) \; ; \qquad i = 1, ...n \tag{2}$$

proposes a closed system, both in the number of species involved and the constant total concentration. This equation has been target of many studies, ranging from population genetics, mathematical ecology and economics to some applications in physics. The fact that no new species are included in the system, decreases the power of this equations when addressing questions of evolution and emergence of new features.

Happel and Stadler proposed a modified model in (Happel and Stadler, 1998) where random generated rates where used for the catalysis between species. The interaction matrix was filled with random numbers and the equations integrated for a certain period. Mutants were introduced as a modified species, changing its interactions with the rest of the system by small perturbations. It was found that overall fitness increases in a strong but non-monotonous way.

In a similar way, we present a model which considers catalyzed replication of RNA

molecules and its evolution via mutation and selection. Since most of the models so far take into account fitness values measured only in terms of single molecules' properties but do not take care of the interaction between molecules, we make use of the fast prediction algorithm for concatenated sequences, RNACofold, in the `Vienna RNA Package` to simulate interacting molecules.

3.5.1 Model One: Fold, many targets

The model works basically with an evolving population whose dynamics are simulated with equation (2).

In our implementation, the replicator equation is used to model a system of interacting species where the individual replicators are implemented as RNA sequences. The equation used for all the simulations is

$$\dot{x}_k = x_k \sum_i a_{ki} x_i - \sum_i \sum_j a_{ij} x_i x_j \tag{3}$$

Where x_k denotes concentration of species k and $M = (a_{ij})$ is the interaction matrix representing catalysis in the off-diagonal terms and self-replication rates in the diagonal.

As a variation from the model from Happel and Stadler, the rates here depend on the RNA sequence, both for auto-replication and catalysis.

We define interactions between species depending on a set of fixed targets. Together with the structures, the interactions among them are also fixed. The folded structure of single molecules are intended to evolve towards this configuration.

Since cycles are at the core of the molecular interactions in all living beings, we define a very simple 3-members cycle as target set as seen in figure 17.

During the simulation, each sequence is folded into its secondary structure using the RNAFold algorithm from the Vienna RNA Package, and then compared to all targets via different distances: the *hamming*, *base-pair* or *structure distance* all of them computed using the structure's dot-bracket representation. The hamming distance is the comparison of each entry of the strings. The base-pair distance counts how many base-pairs should be open or closed to convert one structure into another; and, the structure distance cyclically searches for a good match

between the structures comparing the hamming distances.

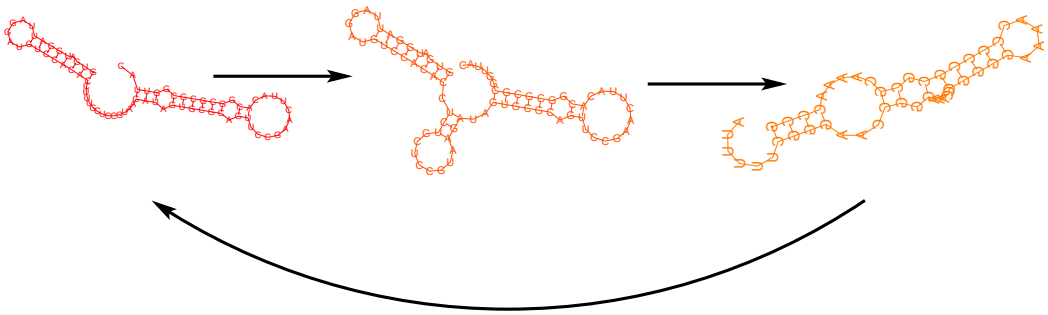


Figure 17: Target structures and topology of the target set. Sequences folding close to one of these targets will behave following the cyclic topology outlined in the figure.

The given molecule is then assigned to the "group" of the closest target, this class defines the interactions of species belonging to this group and those of other groups. Groups can be thought as "phenotypes" which are assigned to the sequences and only modulated by the distance to the correspondent target. This distance defines how good a sequence self-replicates and also how high its catalytic activity is.

The replication rates a_{ij} are calculated as a function of the distance between the folded structures and the predefined target structures and the topology defined by the target cycle. The weights for the interaction between species i and j are calculated as $a_{ij} = \exp(\gamma \cdot \frac{1}{d(S_i,T_i)+d(S_j,T_j)}) \cdot \phi$, where S_i is the secondary structure of species i, $d(S_i, T_i)$ the distance between S_i and the target structure T_i given that species i belongs to the group of T_i and species j to that of T_j. γ and δ are tunable parameters.

In order to generate evolution in the system, point mutations are allowed with certain rate and introduced into the system creating new structures. Depending on the mutation rate μ, a random sequence is chosen from all species in the system and mutated in a single base. A small percentage of the original species' concentration is given to the new one. Each generation, the interaction matrix is filled with the rates of the new species, and the equation integrated. Species are removed from the system whenever their concentration drops below a threshold level.

3.5.2 Results of model One

Targets are approached in a step-wise manner, as seen in other models (Schuster *et al.*, 1994) (Fig. 22).

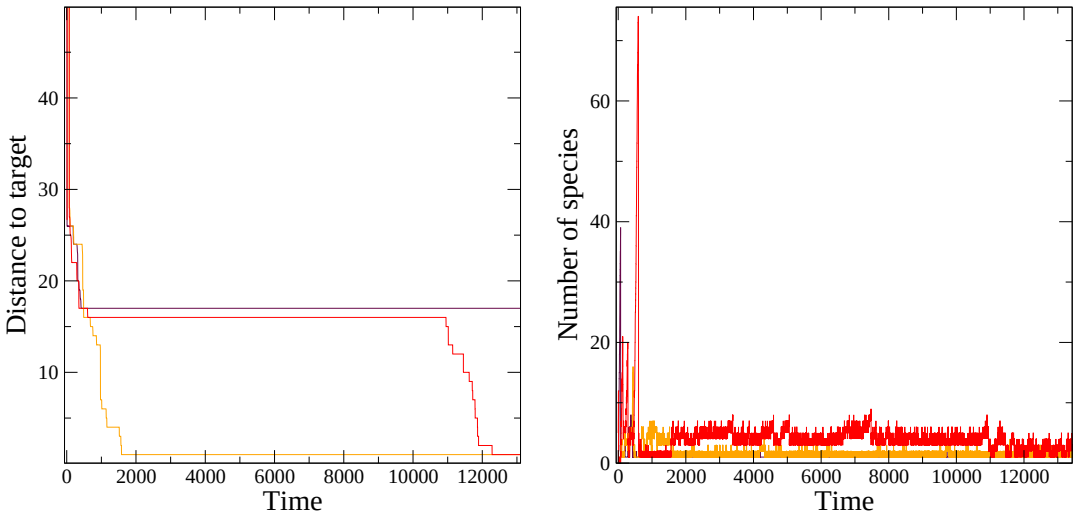


Figure 18: a) Distance to targets for each group. Colors code for different groups, the target is approached in a step-wise manner. b) Number of species for each group. After a period of selection, only few molecular species are left in the system.

When the hamming distance is used, jumping from one group to the other is easy because this metric takes into account only geometric properties and no evolutionary characteristics. Structures that are far away from each other from an evolutionary point of view, may be very close according to the hamming distance. On the other hand, base-pair distance encapsulates the structures in a region were it is difficult to move from one structure to the other, since the opening or closing of a single base-pair implies two point mutations in many cases. The structure distance is somewhere between these two, allowing the sequences to move more freely from one target to the other.

As already mentioned, the folding map generates neutral nets within the sequence space. This is reflected in the fact that once inside a neutral network, more sequences are found which maintain the same structure, and therefore the same interactions with other molecules. This implies an explosion in the population, due to the incorporation of individual species belonging to the same group. Neutrality comes also from the way phenotype is defined. Different structures may

belong to the same group, even if they are not identical. As long as they are closer to the same target, their function in the system will be the same. Therefore, not only mutations leaving the structure unchanged are neutral, but also those leading to a structure which belongs to the same group. This can be seen in Fig. 19, where the increase of species in a group is accompanied by the improvement of the group's fitness.

In this sense, we could say that the system shows a kind of "canalization", meaning that once a good phenotype is found, molecules will try to keep it and no more changes in phenotype will occur. In order to change a sequence from one group to another, a mutation must occur such that the sequence is shifted to the neutral network of a structure closer to an other group's target shape. This means that only sequences in the border of the neutral network which are one point mutation away from the other neutral network can jump from one group to the other. This way, the sequence space is divided in the interior or "canalization" region of each group, and the border, where real phenotypic changes in the sense defined here are possible.

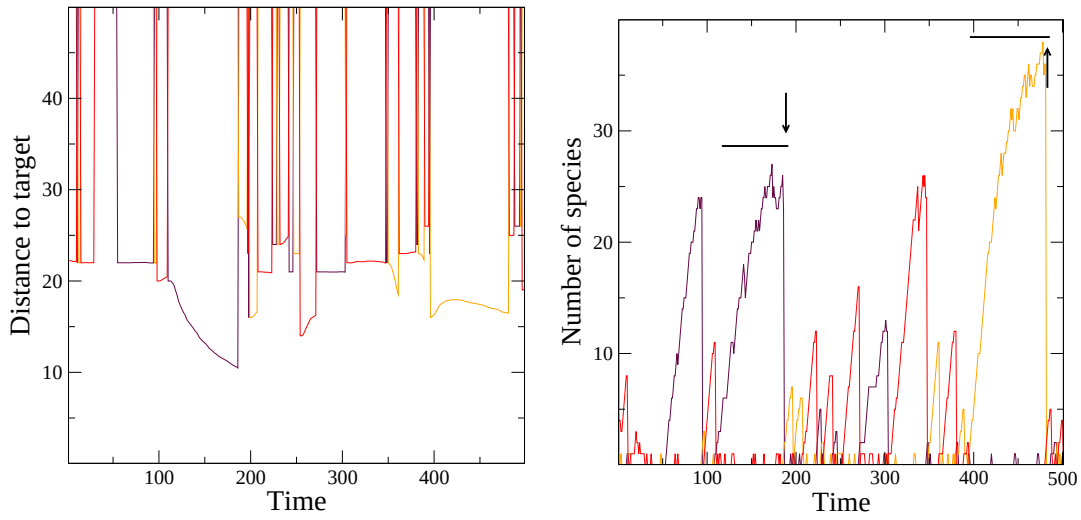


Figure 19: a) Distance to the targets. At each time step only one group dominates the system. Better structures are found until a new species from the next group appears and replicates faster due to better catalysis. b) Number of species of each group. After increasing one of the groups, the population falls due to catalysis of the next species in the cycle. Arrows point to the moment when a species from other group is found killing all the existent species.

In some simulation runs, one or two of the targets were found, while the others

would stay in a close but not perfect structure. Since catalytic rates depend on the distances to the targets, once a group has found the goal, the corresponding catalyzed species will profit from that, not being forced anymore to become themselves perfect self-replicators.

3.5.3 Model Two: Cofold

There are cases where interactions among molecules depend on already folded structures, nevertheless, it is also possible that molecules start to interact before the mfe structure is completely folded. The second model we present, takes this possibility into account and computes the interaction among molecules as a combined folding process which in most of the cases differs from the independent secondary structures of each molecule thanks to the creation of intermolecular base pairs.

To define the values of the matrix M in eq. 3, we use the structure of the co-folding complex of two species based on the thermodynamic rules of RNA folding.

For each pair of molecular species i and j, their sequences are concatenated and the secondary structure of the resulting sequence is calculated. The replication rates a_{ij} are then calculated as a function of the distance between the co-folding complex and a prescribed target structure: $a_{ij} = \exp(\gamma \cdot \frac{1}{d(S_{ij},T)}) \cdot \phi$, where S_{ij} is the cofolded secondary structure of species i and j, $d(S_{ij}, T)$ the distance between S_{ij} and the target structure T and γ and δ tunable parameters.

When concatenating the sequences, the first one is always taken as the replicator while the second acts as catalyst. Since usually the order of the concatenation is not important because of the circularity of the cofold map, this is not an arbitrary setting.

The rest of the implementation follows the same algorithm as in the previous model.

Simulations of this model give information about the number of species in each generation and their concentration, distance of each pair's structure to the target as well as the average distance and fitness of the system. Weighted graphs defined by the interactions between species are studied in order to get an overall view of the system's behavior and self-organization.

3.5.4 Results of model Two

The dynamical behavior depends strongly on the parameters of the system and ranges from the survival of only one single dominating species in each generation to the creation of intricate networks. In the latter case the fitness increases in a stepwise manner as the system approaches the target transition state and maintains, in almost every generation, a number of species greater than a certain lower bound.

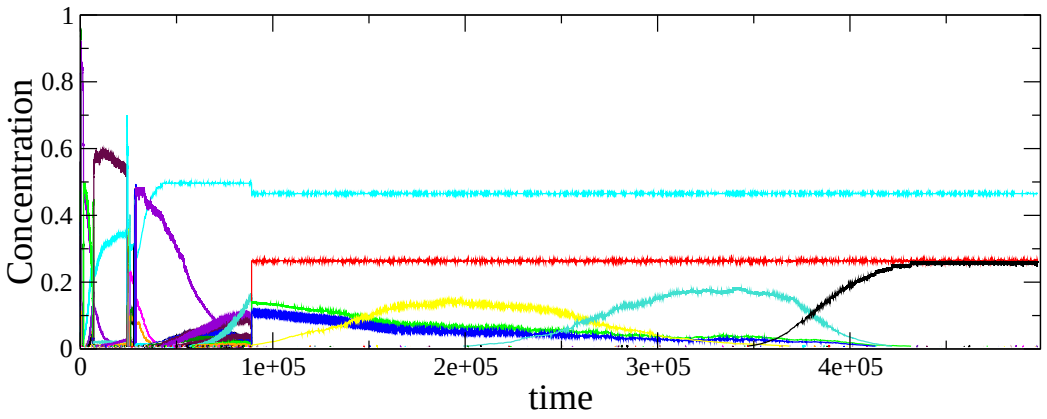


Figure 20: Concentration of molecular species. Few species dominate the system after the first period of selection. The system then reaches a stable state and no further improvement is possible.

Few simulations actually approach the target, and none actually finds it. In fig. 20 the concentrations of all species are plotted. It is clear that most of the time only one or two species take all the space available (total concentration of the system is normalized), and mutated species survive only when they represent a large improvement in auto-replication or are well catalyzed by others. The number of species is depicted in fig. 21 (a), while fig. 21 (b) shows the evolution of the overall fitness of the system. Fig. 21 (c) shows the survival of a fitter variant until another species is found, possibly catalyzed by the old one, killing almost all the rest of the population.

In many cases the model falls in a fixed configuration: existing species are trapped in a fixed point, their concentrations become stable and new species are accepted in the network only if their interactions are stronger than those already existent. Moreover, a new mutation will be accepted only if the rates by which it is catalyzed are good, no matter if it is a good catalyst for the rest of the species.

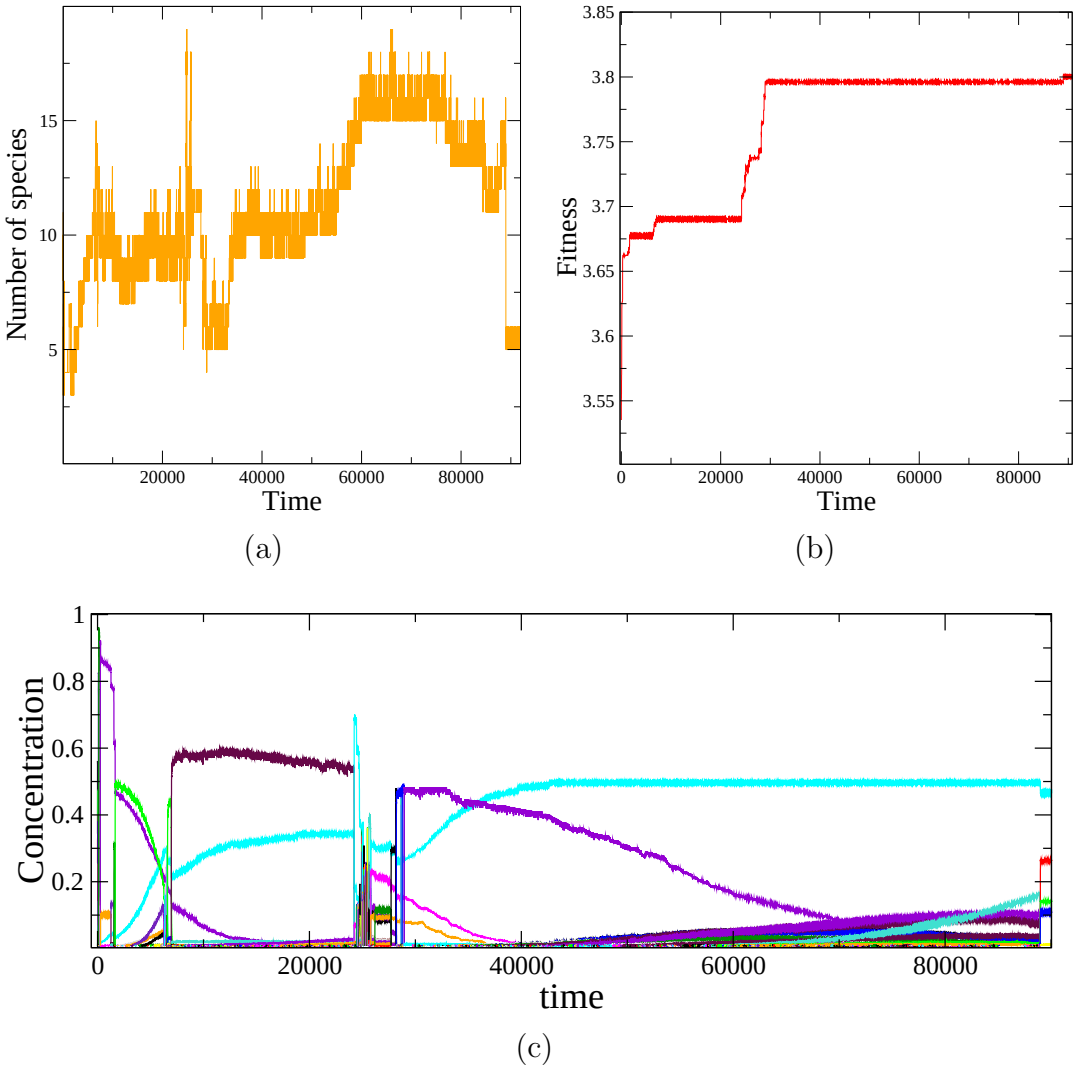


Figure 21: a) Number of species. Variants from the fittest species are added to the system as long as their catalysis is not larger than the self-replication of the principal species. b) After improving the fitness of the system in the first steps, a period of stasis is found. c) Coexistence of more than one species is impossible due to the good self-replication rate of the fittest variant. It is replaced at the end by a better self-replicator.

The system is easily trapped in local minimum because of the high number of interactions between species as well as the low neutrality of the co-folding map. Species are unable to search sequence space due to the strong interactions they must maintain in order to survive and to keep the network working. The fact that only one target structure is approached, reduces strongly the possibilities for different sequences, since they must act both as catalysts and replicators at the same time. In most of the cases, sequences folding with themselves to a

structure close to the target will be the ones accepted, since mutations may leave the structure unchanged thus creating catalytic interaction between the old and new species.

Once a network is created, which consists most of the time of old species catalyzing the replication of new ones and being good self-replicators, the only possibility for the system to increase the overall fitness is to find a sequence which improves the catalytic rates with most of the species in both directions, meaning that it is not only catalyzed but returns to the system some of the help it is taking from other species.

It was found that the minimum distance to the target reached by the system, depends extremely on the sequences taken as initial conditions. To show this, structures formed after a large number of generations are used in a subsequent simulation as target structures. Even when changing other initial conditions or the random numbers used in the program, the system approaches the target much faster than in the first case.

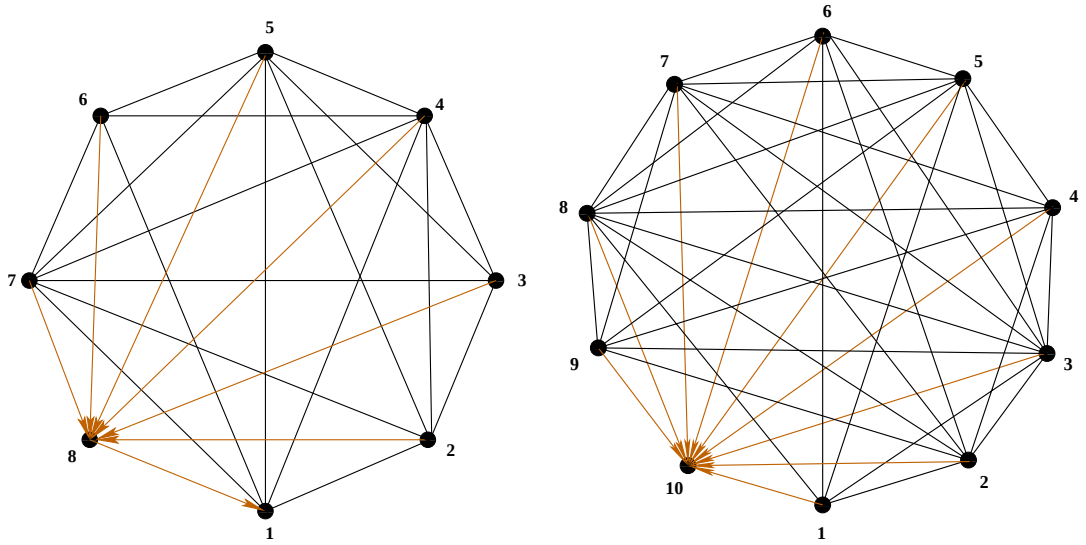


Figure 22: a) Network graph. Black lines code for catalysis in both directions while red lines are interactions in the direction of the arrow. The first figure shows one species being catalyzed by all others, and catalizing the first one of the graph. b) Parasite graph. The new species added to the system (No. 10), acts as a parasite reducing the population dramatically in the next time step.

Comparison between this model and the one with single folded species, makes

clear that interaction between species changes completely the way how evolution to a fixed target occurs. Survival of one species depends not only on its self-replication rate, but in the way it is catalyzed by the others. It may happen too that one species is catalyzed by all the rest (Fig. 22), making its concentration grow very fast. This species will then take all the available resources and kill the rest of the species. This kind of parasites can destroy the whole net loosing the possible improvements attained so far (Fig. 23).

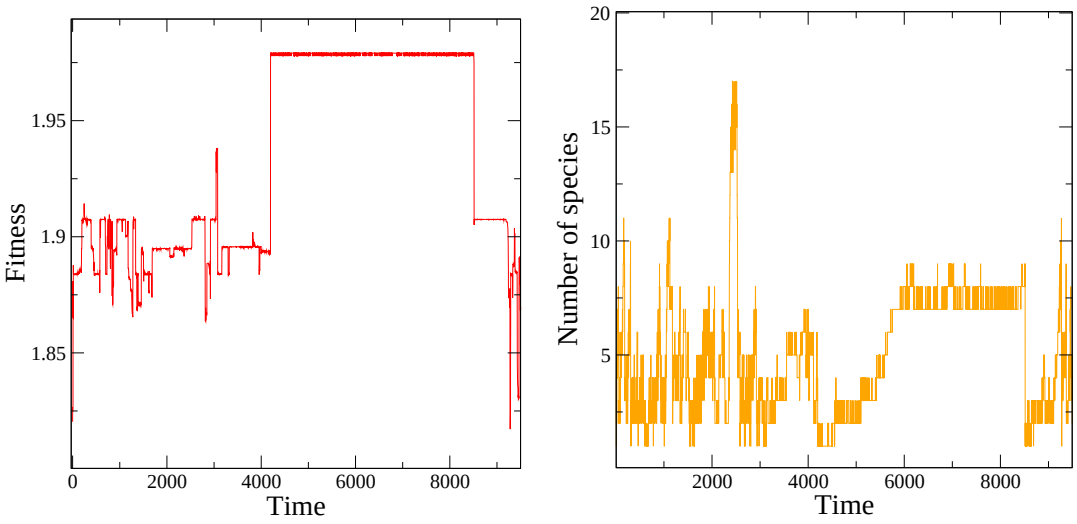


Figure 23: a) Parasite invasion. A parasite invades the system around timestep 8100 and kills almost all species leading to the decrease of overall fitness. b) Population decline. Population decreases due to parasite invasion.

4 Hypercyles

4.1 An answer to the hypercycle's parasites

Boerlijst and Hogeweg (Boerlijst and Hogeweg, 1991) and, later, Streissler (Streissler, 1992) (in a PDE setting) and Cronhjort and Blomberg (Cronhjort and Blomberg, 1994) showed that the problem of parasite invasion can be alleviated by considering spatially organized systems. Most theoretical studies have demonstrated that some kind of spatial structure is indispensable for the persistence and/or the parasite resistance of any feasible replicator system, see e.g. (Tereshko, 1999; Altmeyer and McCaskill, 2001; Zintzaras *et al.*, 2002), although a chemical kinetics with product inhibition can have a similar effect in some parameter ranges (Stadler *et al.*, 2000, 2001b).

In our model we use a two dimensional lattice where molecules diffuse, replicate and catalyze. Making use of the genotype-phenotype map given by the folding of an RNA sequence, we study the evolution of the system towards a fixed target; the diffusion and diversity of the population; and, the resistance to parasites derived from the spacial organization and the different fitness values given by the molecules' structure.

In thi section, we combine the macroscopic modeling of the spatio-temporal population dynamics of self-replicators with the microscopic modeling of the motion of populations of replicators in sequence space. To this end, replicating polymers are explicitly represented by their sequence in a CA-like universe. All reaction rates are derived from the (secondary) structures of the molecules which can be computed directly from their sequences. The parameters of the population dynamics are therefore not external ingredients of the simulation but intrinsic in the model itself (Schuster, 1998). In addition to demonstrating that we recover the typical dynamical features of simpler models of hypercyclic systems, we focus here on the dynamics in sequence space and show that Kimura's model of neutral evolution is applicable at least when time-scales are considered that are much larger than the oscillations of species in the population dynamics of a hypercycle.

4.2 The hypercycle and RNA fold

We consider a stochastic version of a second order replicator equation (Schuster and Sigmund, 1983) with mutation, i.e., a replication mechanism of the form

$$x + y \longrightarrow x + y + z \tag{4}$$

The symbol x represents the sequence of a template RNA molecule that, with the aid of the replicase ribozyme y, is copied to produce an RNA sequence z, which can be the same as the template, $x = x$, in the case of correct copying, or a *mutant* $z \neq x$. In addition we consider a slow uncatalyzed replication mechanism of the form $x \longrightarrow x + z$.

Each RNA sequence is interpreted as a self-replicator that also has the ability to catalyze the replication of other RNAs. Catalytic activities and replication rates are dependent on the molecules' secondary structure[1]. Secondary structures of RNA molecules can be computed efficiently by means of a dynamic programming approach (Zuker and Sankoff, 1984) based on empirical parameters (Mathews *et al.*, 1999). We use the `Vienna RNA Package` (Hofacker *et al.*, 1994; Hofacker, 2003) for this purpose. The optimal reaction rates are realized by the "perfect" target-hypercycle in Fig. 24. It is known that self-organization providing resistance to parasites is possible only in cycles of 6 or more members, while cycles of 3-5 members are quickly destroyed (Hogeweg and Takeuchi, 2003). Therefore we choose an 8 members hypercycle for our model.

The interaction topology of our target set is a hypercycle with 8 members $\mathcal{T}_1$ through $\mathcal{T}_8$. The target structures $\mathcal{T}_k$ were picked at random. In order to investigate resistance against parasites, we consider selfish parasites and short-cut parasites besides the ordinary members of the hypercycle. To do this, one more target-structure is chosen randomly and the corresponding reactions are defined depending on the nature of the parasite (Fig. 25). All rates for parasite sequences are computed in the same way as for the target-set members. Indeed, technically, the parasites are treated as additional target-structures.

For each sequence x in a population $\mathbb{P}$ we compute its secondary structure $\mathcal{S}(x)$ using the `Vienna RNA Package` (Hofacker *et al.*, 1994). Then we determine its

[1] A *secondary structure* $\mathcal{S}$ is a special type of contact structure, represented by a list of base pairs $[i, j]$ with $i < j$ on a sequence x, such that for any two base pairs $[i, j]$ and $[k, l]$ with $i \leq k$ holds: (i) $i = k$ if and only if $j = l$, and (ii) $k < j$ implies $i < k < l < j$.

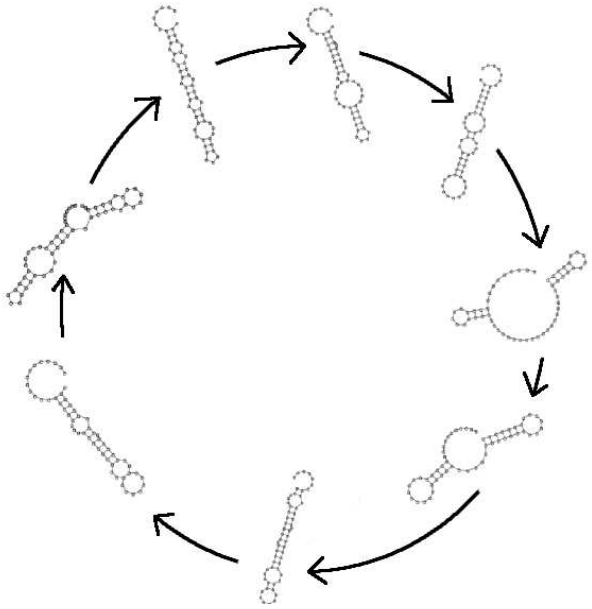


Figure 24: The target set is a hypercycle with 8 members. All sequences have length $n = 56$.

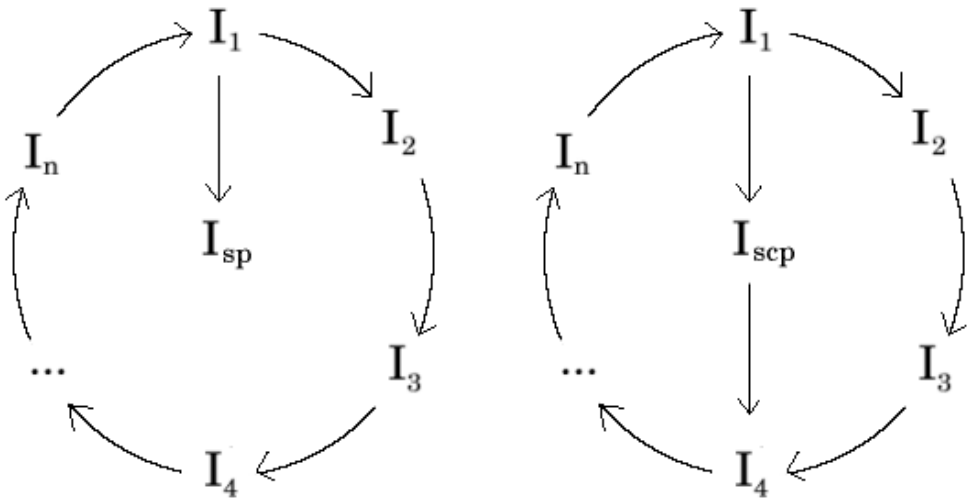


Figure 25: Topology of the target set for selfish and short-cut parasites.

structure distance $D(\mathcal{T}_k, \mathcal{S}(x))$ to the target shapes $\mathcal{T}_k$. For simplicity we define $D(\mathcal{X}, \mathcal{Y})$ as the number of base-pairs that $\mathcal{X}$ and $\mathcal{Y}$ do not share. Finally, we assign $\mathcal{S}(x)$ to the hypercycle-member h that minimizes the distance $D(\mathcal{T}_k, \mathcal{S}(x))$. We write $\mathbb{P}^h$ for this sub-population of sequences whose structure is closest to the

target shape $\mathcal{T}_h$.

Once the group h has been determined for every sequence the replication-decay-catalysis process in simulated as outlined in (Boerlijst and Hogeweg, 1991), Fig. 26:

Decay: Sequence x has a decay probability that depends linearly on the distance to the target structure:
$\delta_x = 1 + D(\mathcal{T}_k, \mathcal{S}(x))$

Replication: Sequence x has a probability to self-replicate without the help of a catalyst that depends inversely on the distance to the target structure:
$\alpha_x \sim \frac{1}{1+D(\mathcal{T}_k,\mathcal{S}(x))}$

Catalyzed Replication: When a self-replicator has neighbors that correspond to their catalysts in the direction of the reaction, the probability (rate) of catalysis is largely improved. As well as self-replication rates depends on fitness, also the performance of catalysts is defined by their distance to the target. The similar a phenotype is to the corresponding target, the better its rate as catalyst will be. The total replication rate is therefore

$$\rho_x = \frac{1}{1 + D(\mathcal{T}_k, \mathcal{S}(x))} + \sum_{y \text{ catalyzes x}} \frac{C}{1 + D(\mathcal{T}_y, \mathcal{S}(y))} \tag{5}$$

where $C = 8000$ is the relative rate of catalyzed versus uncatalyzed replication.

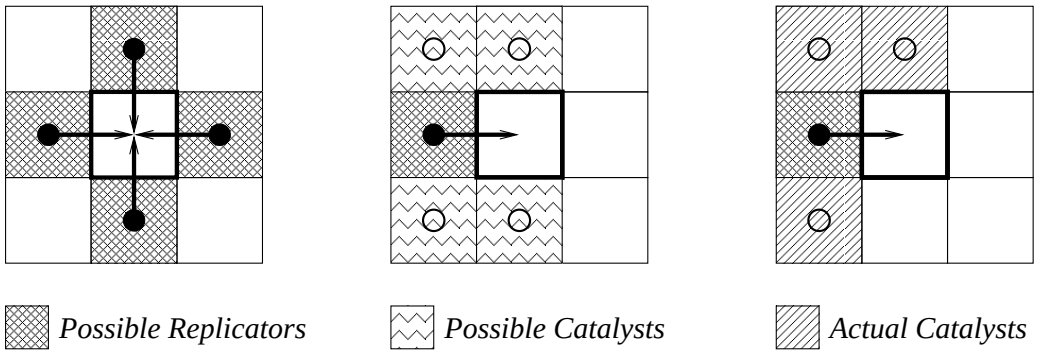


Figure 26: Rules of replication. For each of the neighbors (•) of the empty cell (marked by a bold outline) the replication rate ρ_z is computed taking into account their neighbors in the direction of the replication (∘) as potential catalysts. The neighbor with the largest values of ρ_z invades the empty position. In this example, for the chosen replicator, only three of its neighbors are catalysts according to the hypercycle topology.

From the way rates are computed follows that parasites and members of the

hypercycle may have equal replication and catalysis rates, depending on their sequence and the distance of the folded structure to the targets.

Mutations occur as errors during replication. As in Eigen's quasispecies model (Eigen, 1971) we assume a uniform per-nucleotide rate p of incorporating an erroneous letter. These point mutations of the parental sequence x have a high probability of changing the secondary structure. Since these structural changes may be large (Schuster *et al.*, 1994) we have significant probability that a mutant sequence will belong to either a different class of hypercycle members or to one of the parasite classes.

The population $\mathbb{P}$ of replicators is spread out on a 2-dimensional grid with periodic boundary conditions, typically consisting of 200×200 cells. In this respect our simulation resembles those described in (Boerlijst and Hogeweg, 1991; Cronhjort and Blomberg, 1994). Each cell can be empty or occupied by a RNA single sequence. Diffusion is modeled using the Toffoli-Margolus scheme (Toffoli and Margolus, 1987). The number of diffusion steps within each simulation time unit ranges from 0.01 (meaning that we wait 100 simulation steps between each diffusion step) and 20.

Simulations are initialized by randomly placing 200 to 1000 initial sequences on the grid. The sequence in an occupied cell dies with a rate proportional to δ_x. For every empty cell we compute the replication rates ρ_z for all its neighboring cells, assuming that the replication of z is catalyzed only by those neighbors that correspond to the preceding class in the hypercycle topology, Fig. 26. According to the model presented in (Boerlijst and Hogeweg, 1991), we consider possible catalysts only in the direction of the replication. The sequence with the largest values of ρ_z invades the empty cell. Cells are chosen for update in random order until every occupied cell has been updated.

Several variables are measured throughout the simulations: the number $N_k = |\mathbb{P}^k|$ of individuals per group, the average distance $\bar{D}$ to target over the whole system and over each group, the diversity θ_k between individuals in a class of replicators and number Y_k of *different* sequences belonging to target class k. The diversity of a group is computed as proposed in (Stadler, 2002a)

$$\theta_k = \frac{1}{N_k(N_k-1)} \sum_{x \neq y \in \mathbb{P}^k} d_H(x,y) \tag{6}$$

where $d_H(x,y)$ is the Hamming distance of the sequences x and y. In (Stadler,

2002a) it is shown that replicators with interactions tend to minimize diversity until they end in a quasispecies-like distribution.

4.3 Results

4.3.1 Spatial Pattern Formation

We first consider a universe without parasites, i.e., all sequences are assigned to one of the structures of the hypercycle-members. As in (Boerlijst and Hogeweg, 1991) we observe spiral waves when every member of the cycle has a minimum concentration, Fig 27. In almost every run with two diffusion steps for simulation time unit, a first period of disorder is followed by the birth of a spiral which contains sequences of every group, ordered depending on the topology of the targets. It is important to notice that without a minimum fitness, individuals of that group would die before they could get any help to replicate, so that evolved enough sequences of every group must be present in order to the spatial patterns to emerge. Once the spiral is formed, the sequences continue approaching the target but in a much slower pace, in fact, in some simulations we observed an oscillatory behavior of the fitness average depending on the number of sequences present in the system at any given moment. If the ratio of replication to diffusion steps is increased, we observe multiple smaller spirals. For very small spatial diffusion constants, however, a significant part of the lattice remains empty, and the system usually dies out due to fluctuations.

Some groups could reach the target, while others may stay away without breaking the dynamics. In the case where one group will get to the target while the others had a poor fitness, however, the system sometimes collapses to the survival of only the single fittest species. This is only possible when the group which is catalyzed by this "master species" is not present in the system: since the rate of catalysis depends also on the fitness, so that the "follower" will increase its concentration at the expense of the "master species".

When all members of the cycle are present with a minimum concentration and fitness, a change of behavior occurs and oscillations in the number N_k of sequences per group is observed, see Fig. 28. The amplitude of this waves depends on the ratio between self-replication and catalyzed replication rates and the spatial diffusion parameter. If this ratio is too large, the abundance of sequences of one

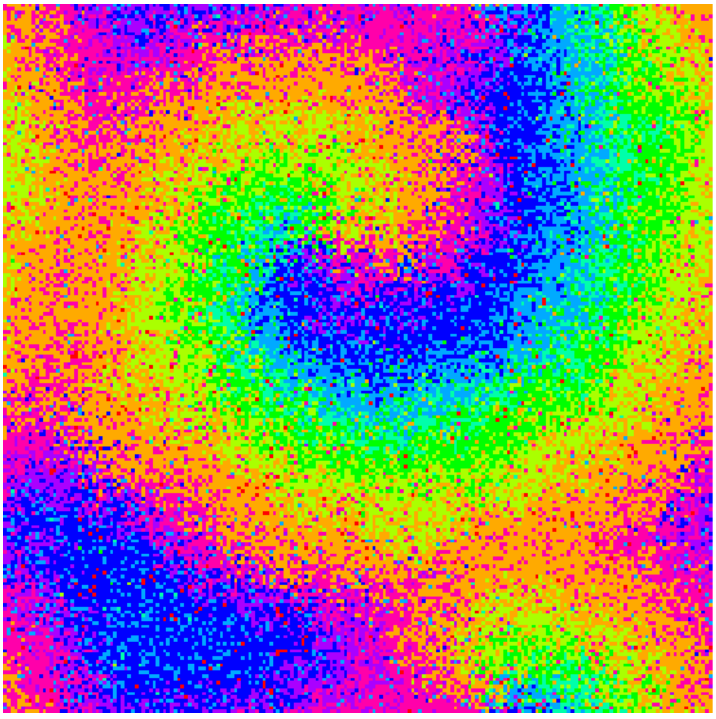


Figure 27: Spirals formed after 3000 generations in an evolution experiment started with 300 random sequences in the absence of parasites.
Simulation parameters: grid size $L \times L = 200 \times 200$, sequence length $n = 56$, mutation rate $p = 3.5 \times 10^{-4}$, 2 diffusion steps between replication steps. Simulation parameters are the same in all figure unless explicitly stated otherwise.

group will lead to a very fast growth of the next group in the hypercycle, giving almost no space for other members to replicate. Only one or two groups fill the entire lattice at any given point in time, making it more difficult or impossible to create the spirals. Also, when spatial diffusion is low, many smalls spirals appear making the amplitude of the oscillations lower.

When a selfish parasite is introduced, a first period is observed where members of the hypercycle, as well as the parasite, appear and disappear from the system without much order. For some time both parasite and hypercycle can coexist, but it ends in the parasite being expelled from the system and the spirals arise. Of course, mutations from regular sequences may jump to the parasite group, implying that the parasite has members almost all the time without being harmful for the system. These parasitic sequences typically are eradicated before they can evolve towards high replication rates. The spirals in this case are not as regular

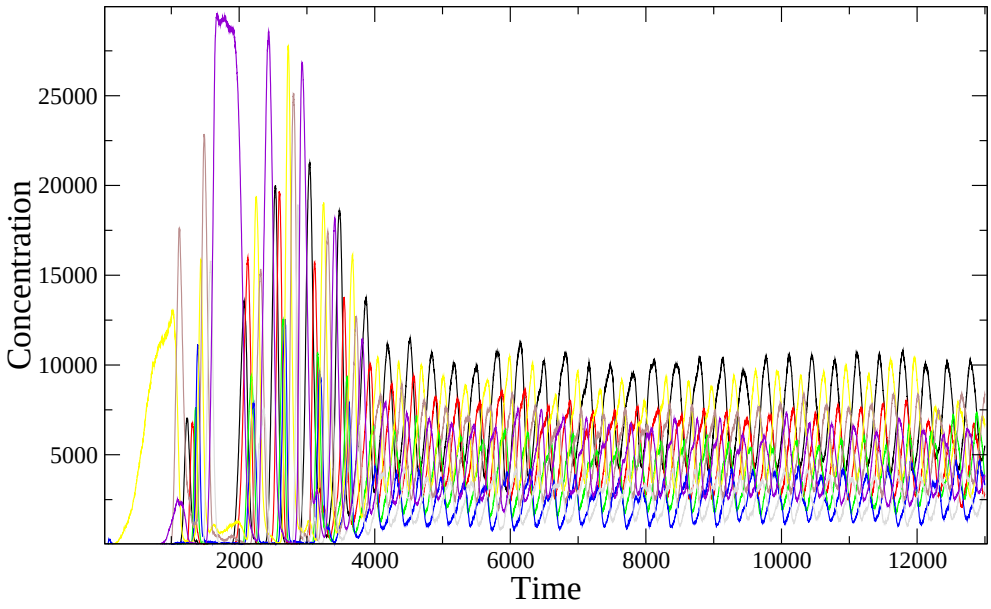


Figure 28: Evolution towards the target hypercycle. After a transient period of disorder, the concentrations Y_k of the individual member-classes of the hypercycle (different colors) exhibit regular oscillations.

as those without parasite, nevertheless they are stable and can coexist with an invading parasite.

The case of the short-cut parasite is quite similar. The system is stable against this kind of parasite and only a few times the runs ended with the shorter cycle formed in this topology. In the majority of the simulations, however, the parasite was expelled from system after some time (Fig. 29 (a)). The reason for this increased resistance appears to lie in the the genotype-phenotype map derived from the RNA folding algorithm. The fact that fitness depends on the secondary structure allows the hypercycle to evolve towards a stronger configuration while the parasite is left behind: from the fitness plot (Fig. 29 (b)) one can see how for some period the parasite evolves more or less the same way as the other members of the hypercycle. Nevertheless, every time the parasite is expelled from the system, it looses the fitness it could have won before, becoming a much weaker species. It is clear that stability of the hypercycle is due not only to the spatial configuration but also to the advantage of its members in an evolutionary way.

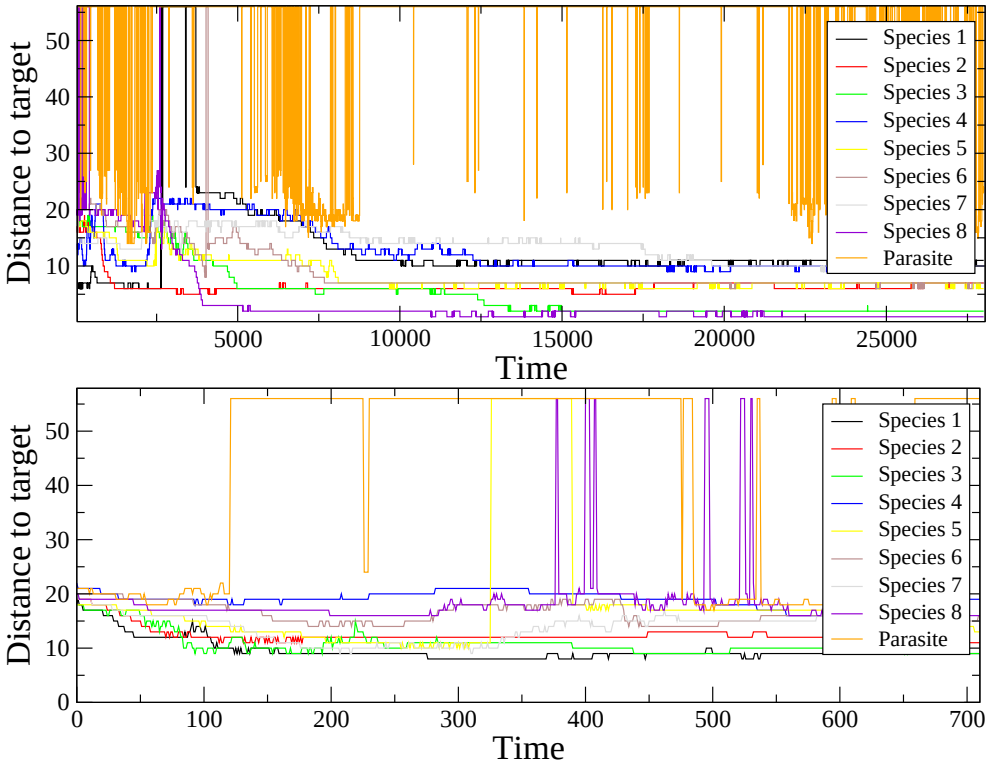


Figure 29: Evolution of the mean fitness of the individual classes (different colors). After a period of disorder, the parasite (orange curve in the upper part of the plot) is unable to re-invade the system. It dies out before it can reach sequences that are near-optimal parasites (distance 0 to target).

4.3.2 Population Structure

Diversity depends strongly on the initial conditions, in particular on the number of sequences first introduced to the system and the spatial diffusion rate. To make replicators evolve towards the targets, it is important to keep a high selectivity among them, this in turn can make it harder for the system to reach the desired organization. Starting with less than 100 sequences leads, almost in almost every case, to the death of all species or the survival of only one. If selection is lowered it is possible to start with one sequence but evolution towards the targets will be slower. In most cases, if the spatial diffusion is kept fixed, starting with 200 or 300 sequences allows the system to survive even with higher selection rates. In this case, diversity falls very quickly to almost zero and is maintained very low until the end of the simulation, see Fig. 30(a). This is due to the fact that only

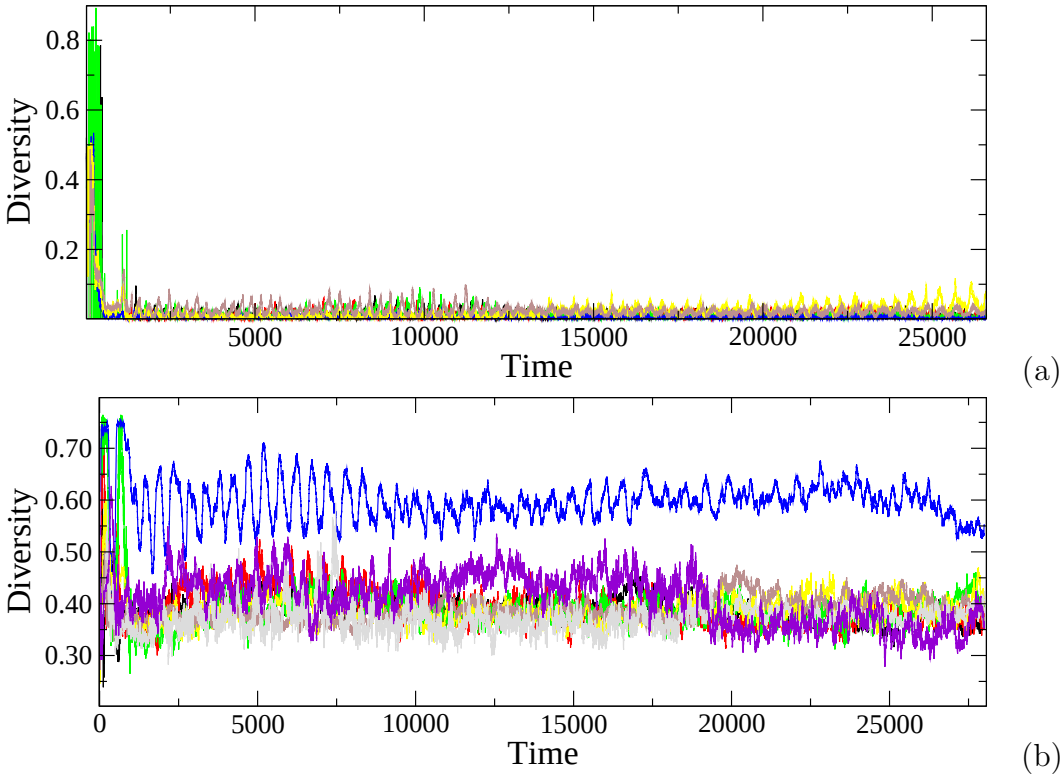
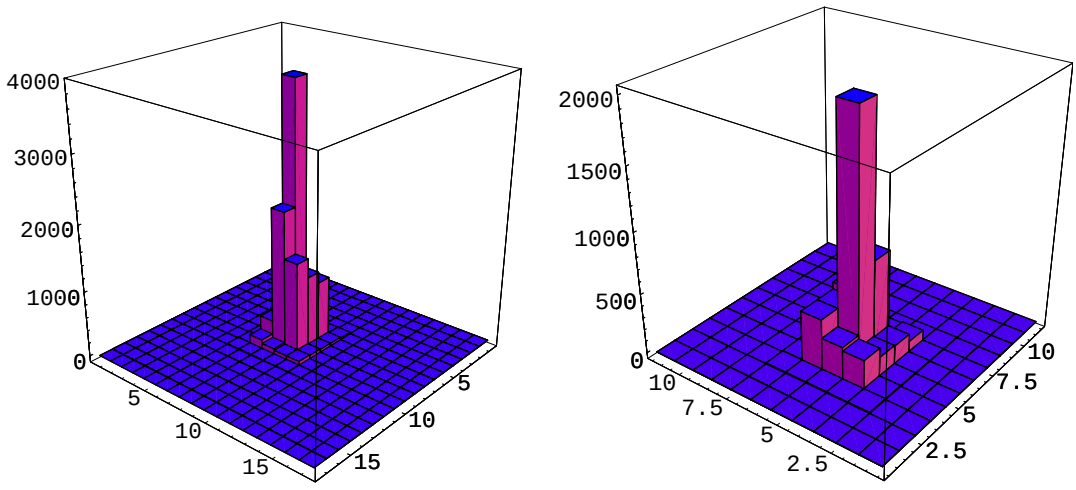


Figure 30: Diversity of the system with different initial conditions. Simulation parameters for both plots are the same except for the number of initial sequences in the lattice. For figure (a) 300 sequences were used while in figure (b) 800 individuals started the simulation. It can be seen that when the number of species at the beginning is high enough, diversity is kept high until the end.

a few sequences will be fit enough to survive at the beginning. After this first selection step, only mutations from the surviving species will produce variation.

When the number of initial sequences is increased, there is a higher probability that good structures will be found, even with totally different sequences. Therefore, diversity is high at the beginning and is maintained by the system, oscillating depending on the number of species per group, Fig. 30(b). We believe that this is due to the interactions and catalyzed replications: selection on a single member becomes less important when its replication is improved by the others. Even a fast dying sequence can stay in the lattice because of its even faster catalyzed replication.



(a) (b)

Figure 31: Distribution of sequences in each class of hypercycle members. Only a few sequences are present with almost all the individuals of the group while the rest of sequences are represented by only one individual. Figures/Hyper (a) and (b) show the distribution of two different members in the same simulation.

A quasispecies-like behavior is observed if diversity is low. The distribution of the number of individuals with the same sequence is centered around a "master sequence" in each class of hypercycle-members; a large fraction of the populations consists of individuals that occur only in a single copy. These "explorers" of the sequence space are lost and replaced by others within a few generations, Fig. 31.

4.3.3 Drift and Diffusion in Sequence Space

The *profile* of the class k of the hypercycle at time t is defined as the $4 \times n$ vector $\mathbf{p}^k(t)$ whose components are the frequencies of the 4 types of nucleotides at each of the n sequence positions (Stadler, 2002a). The overall movement of the population in sequence space can be quantified in terms of the correlation function

$$g(\tau) = \frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} \|\mathbf{p}(t+\tau) - \mathbf{p}(t)\|^2 \qquad (7)$$

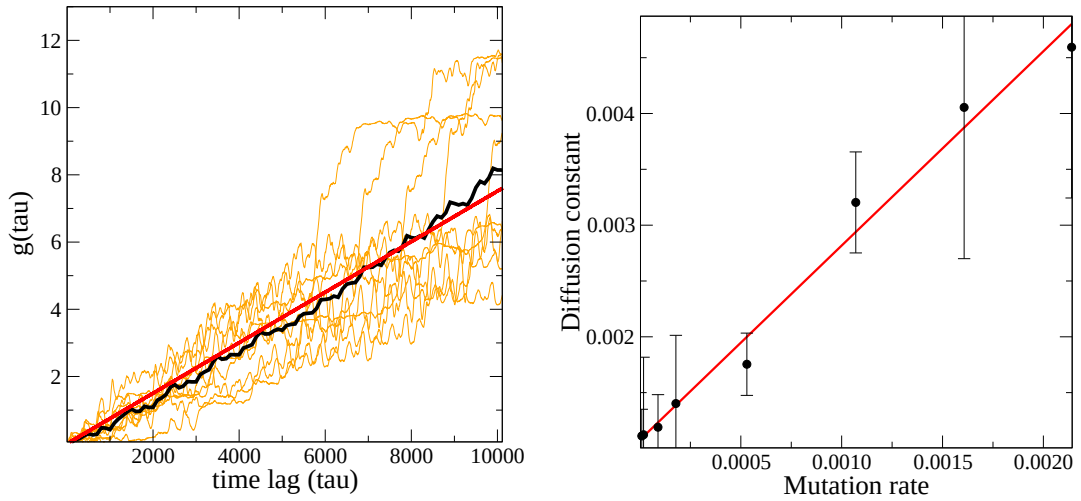
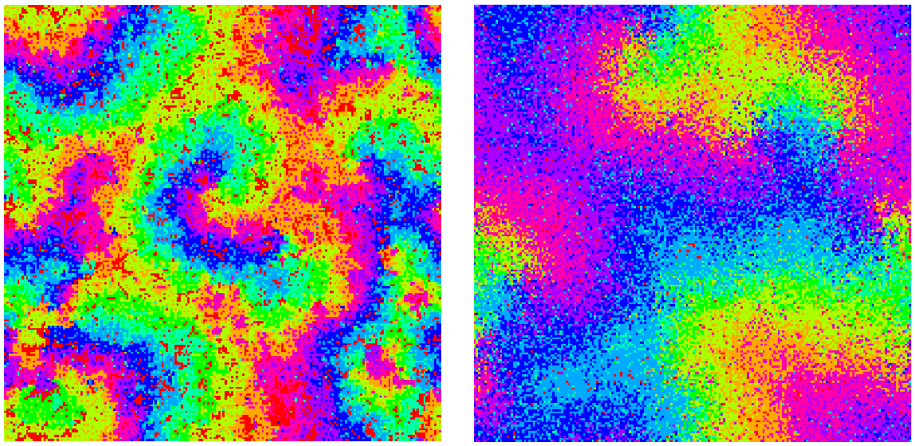


Figure 32: Displacement of the profile and diffusion constant D vs mutation rate. (a) The Displacement of the profile with time, $g(\tau)$, is shown in gray for individual sub-populations $\mathbb{P}^k$. The average of $g(\tau)$ over all sub-populations is displayed as bold black line. (b) The slope of $g(\tau)$ defines the diffusion constant in sequence space. As expected, there is a linear dependence between diffusion constant D and single digit mutation frequency d. The data are averaged over 16 different runs and all species.

computed for suitable intervals of measurement $[T_1, T_2]$. The mobility of the population in sequence space is conveniently quantified in terms of the *diffusion constant* D which is defined as the slope of $g(\tau)$, i.e. as the slope of the linear approximation of $g(\tau)$.

As expected from simulations both of RNA based quasispecies (Huynen *et al.*, 1996) and from a simple model of interacting molecular replicators (Stadler, 2002a) we observe a linear dependence of the diffusion constant on the per-site mutation rate p, see Fig. 32b. We should expect that small differences in the diffusion constants of different sub-populations should exist since the diffusion constant should depend on the fraction ν of mutations that do not change the secondary structure. It is known that ν depends on the secondary structure in question (Huynen *et al.*, 1996). We have not been able to detect significant difference in the diffusion constants of individual sub-populations (data not shown) since the effects are small and would require much more extensive simulations in order to obtain sufficiently accurate estimates of D for each species separately.



(a) (b)

Figure 33: Different size and number of spirals due to the change of spatial diffusion. Same parameters as previous simulations except for: a) Spatial diffusion: 0.1 (1 diffusion step every ten replication events). b) Spatial diffusion: 20 (diffusion steps between two replications)

Diversity and diffusion in sequence space depend on the relative strength of spatial diffusion. As can be seen in Fig. 34(a), reducing the number of diffusion steps between two replication events leads to an increase of diversity. Very small values for this parameter, however, kills the system because molecules take too long to find the correspondent catalysts. A phase transition was found between the regimes of slow and fast spatial diffusion. For small numbers, the population breaks into several spirals, so that individual evolution of these subgroups is possible. Instead of having a single nucleus from where all the molecules arise, many replication basins are created in the center of each spiral. Figure 33 shows the pattern formation for two different values of the spacial diffusion. In figure 33 (a), many more spirals are formed than in fig 27 and figure 33 is an intermediate case. As a consequence of the independent evolution of subpopulations, the diversity of the entire population increases approximately linearly with time. The rate of this increase is used to distinguish the regimes of low and fast spatial diffusion. For slow enough rates, this slope vanishes. After a large number of generations, diversity finally saturates. A similar behavior is observed in diffusion. Since the population splits in many subpopulations when spatial diffusion rates are low, exploration of sequence space is increased. The face transition mentioned above can be seen also in Fig. 34(b).

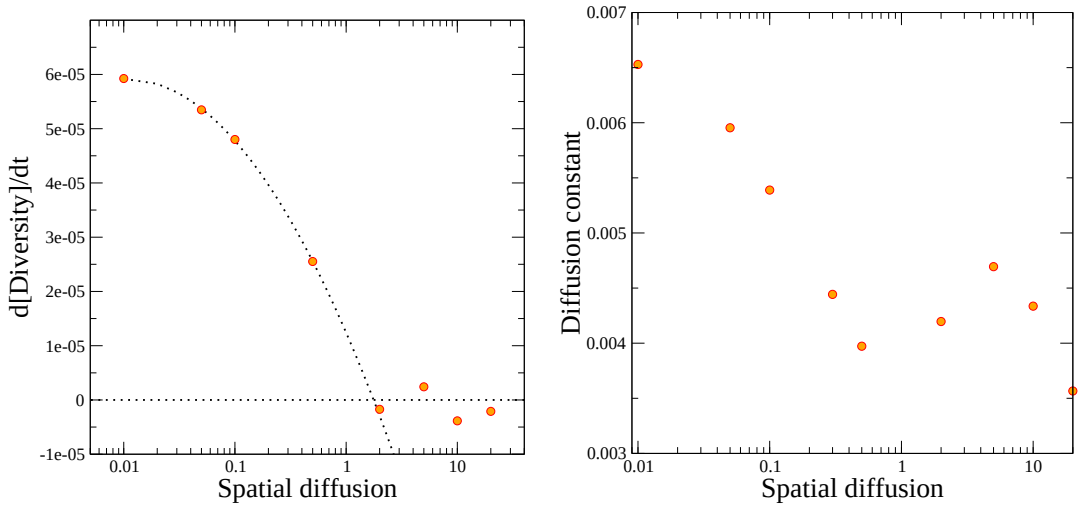


Figure 34: Change of diversity with time and diffusion in sequence space with changing spatial diffusion. Spatial diffusion is measured as the ratio of diffusion steps to replication events in the simulation. "D[Diversity]/dt" is the slope of the linear approximation of individual diversity curves averaged over several species and simulations. For both Figures 600 sequences were introduced in a lattice of size $L \times L = 200 \times 200$ and the mutation rate was fixed to $p = 3.5 \times 10^{-4}$.

4.4 Hypercycle with RNACofold and the implications of low neutrality

In this model we use the same equations and procedures as in the previous one, being the only difference the way we define the interaction rates among molecules. For the exact form of the equations and reactions, please refer to the last section, eq. 4 and the corresponding text.

Coevolution at the molecular level is a predecessor regulating networks in living systems. In this approach we use the secondary structure of a molecule to define its self-replication rate, we combine two molecules to define the catalyzed replication rate, making the genotype-phenotype map depend on the interaction and not only on the single molecule.

It is known that environmental changes greatly influence the development of organisms from the genetic information into the phenotypic expression. In

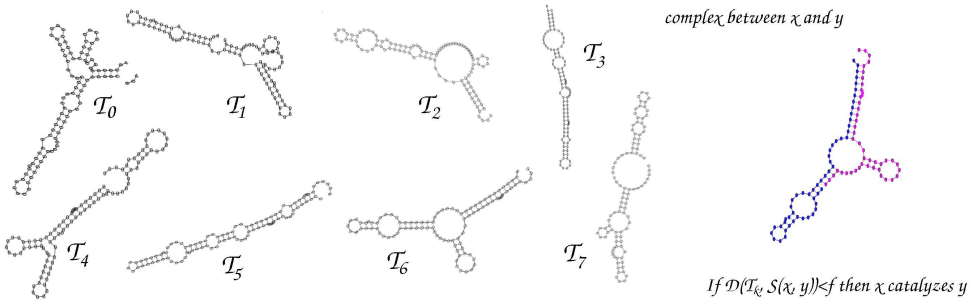


Figure 35: List of targets. If the distance between the complex formed by sequences x and y and some of the targets $\mathcal{T}_1$ is less than a threshold f, then x will catalyze the replication of y.

our model we translate this characteristic into the interaction with neighboring molecules. Mutations in any of them may extremely change the behavior of the system by affecting the resulting phenotype of the pair, i.e. the cofolded structure of the concatenated sequence.

Each RNA sequence is interpreted as a self-replicator that also has the ability to catalyze the replication of other RNAs. Catalytic activities and replication rates are dependent on the inter-molecular complex formed by each pair of sequences. This complex is represented by the secondary structure formed by both sequences. Secondary structures of pairs of RNA molecules can be computed efficiently by means of a dynamic programming approach (Zuker and Sankoff, 1984) based on empirical parameters (Mathews *et al.*, 1999). We use the RNAcofold program from the `Vienna RNA Package` (Hofacker *et al.*, 1994; Hofacker, 2003) for this purpose. The optimal reaction rates are realized by the "perfect" target set in Fig. 35. Our target set is a list of 8 secondary structures $\mathcal{T}_1$ through $\mathcal{T}_8$. The target structures $\mathcal{T}_k$ were picked at random and were chosen in a way that it is possible to close a cycle if the right sequences are cofolded. Nevertheless, the targets should be regarded as a list of possible reactions, and not as a predefined hypercyclic system.

Again, we look for emergence of spatial patterns and the evolution of the population in sequence space.

For each sequence x in the population $\mathbb{P}$ we compute the complex formed with every other sequence y, $\mathcal{S}(x, y)$, using the `Vienna RNA Package` (Hofacker *et al.*, 1994). Then we determine its structure distance $D(\mathcal{T}_k, \mathcal{S}(x, y))$ to the target shapes $\mathcal{T}_k$. For simplicity we define $D(\mathcal{X}, \mathcal{Y})$ as the number of base-pairs that $\mathcal{X}$

and $\mathcal{Y}$ do not share. Finally, we assign $\mathcal{S}(x, y)$ to the target set member h that minimizes the distance $D(\mathcal{T}_k, \mathcal{S}(x, y))$ given that this distance is below a certain threshold f. This way, the catalytic activity between sequences x and y depends on the cofolded structure they form and so any mutation in any of the structures may change the relation they keep. Since the cofold algorithm used to predict these structures may be sensitive to the order of the sequences, we determine that the first sequence will act as catalyst and the second as the replicating molecule.

The replication-decay-catalysis process is simulated in the same way as the case described in the last section.

Mutations occur as errors during replication. Given the low neutrality of the cofolded map, as seen in the second chapter of this dissertation, structural changes due to mutations may be large (Schuster *et al.*, 1994) and thus there is a significant probability that a mutant sequence will change its old relation to the other sequences by changing the cofolded structure.

The sequence in an occupied cell dies with a rate proportional to the distance of the cofolded structure with itself and, as in the previous model, replication rates are computed according with the set of reactions in Fig. 35.

Several variables are measured throughout the simulations: the number of interactions belonging to the class of the different targets, the diversity θ_k between individuals and number Y_k of *different* sequences belonging to target class k.

In order to study the behavior of the hypercycle under these conditions, we start with an equal number of sequences of each kind necessary to close the cycle defined by the target set. The color assigned to each sequence will remain the same after mutations, meaning that parasites will be part of the same population of the original sequence.

4.5 Results

4.5.1 Spatial Pattern Formation

As we said before we start the simulations with sequences folding exactly to the targets in the list. It is then clear that patterns will form in the lattice (Fig. 36) as those first shown in (Boerlijst and Hogeweg, 1991). After the spirals are established, mutation is allowed. We study two main cases depending on the

threshold f: the case when $f = 0$, which means that only exact hits to the targets are taken as good complexes, and when $f = 20$ percent of the total length of the targets, giving the opportunity of some mismatches in the cofolded structures. In both cases the system is unstable and the spirals disappear very soon in the simulations.

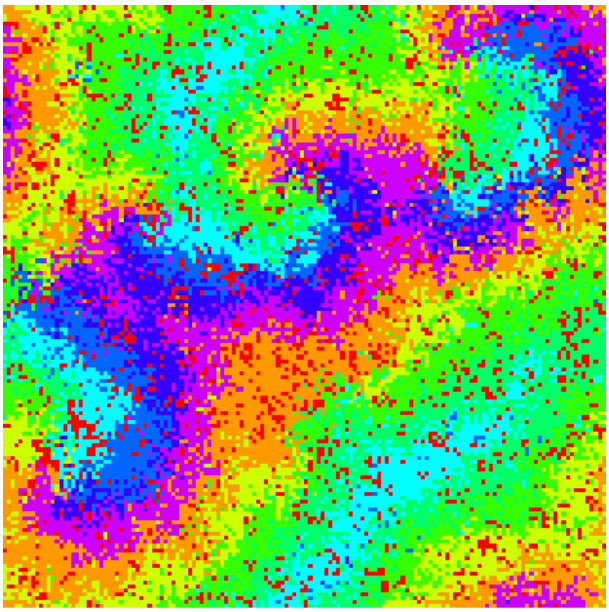


Figure 36: Spirals form in the lattice after few generations. Simulation parameters are: lattice size $L \times L = 200 \times 200$ and mutation rate $p = 3.5 \times 10^{-4}$ and $n = 56$.

4.5.2 Instability of the Hypercycle

The reason for the instability in the first case, is the low neutrality of the genotype-phenotype map derived from the cofolding of two sequences as pointed out in the second chapter of this work. Only a small fraction of mutants will conserve their old reactions (secondary structures) leading to a diminution of catalyzed reactions. As can be seen in figure 37 (a), the number of reactions corresponding to all targets drops to zero in a few hundred generations. Since the sequences in the model are supposed to self replicate, after catalyzed reactions are over, the species conserve their concentrations for some time, Fig. 37. It may happen that new mutants react again by forming a complex in the target's list or that old reacting species find each other in the lattice, Fig. 37 (a).

In the second case, the system is invaded by short-cut parasites. The relaxed

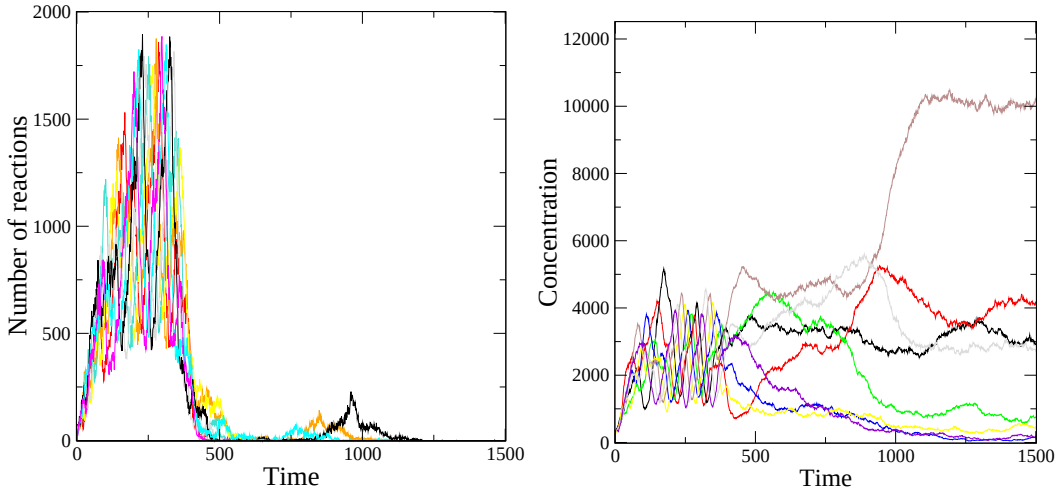


Figure 37: a) Drop of catalyzed reactions after the start of mutations at time 300. Color codes for reactions between different species. b) Concentrations stay constant after the end of catalyzed reactions.

requirement to interact with other sequences, favors the catalysis not only of the usual species but also of others inside the hypercycle. Even when the rates for these new interactions are usually lower than those existing before, the combination between the loss of the reactions belonging to the hypercycle and the appearance of the new ones, results in shorter cycles after some generations and finally to the survival of only one or two species.

In Fig. 38 we show the emergence of short-cut parasites 900 generations after mutation starts. The interaction matrix showed in the $x - y$ plane is built by counting the total number of interactions per generation. Each entry in the matrix represents the interactions for the corresponding column and row groups. In Fig. 38 (a) and (b) a 6-cycle can be seen. The x-axis codes for replicators catalyzed by the y-axis. A cycle can be followed by taking a non-zero entry and moving parallel to the x-axis until the diagonal is found. Then moving in the y direction until a species is found and then again to the diagonal. By repeating this process, it is clear how a 6-cycle emerges when short-cut parasites invade the system. It is also clear that rows and columns for both species 1 and 2 are empty. After generation 1200 the cycle is established and reflected in the oscillatory behavior of the concentrations (Fig. 38(a)).

Since the cycles are rapidly destroyed in the simulations, is it impossible to follow the evolution of the species in sequences space. Unfortunately, the instability of

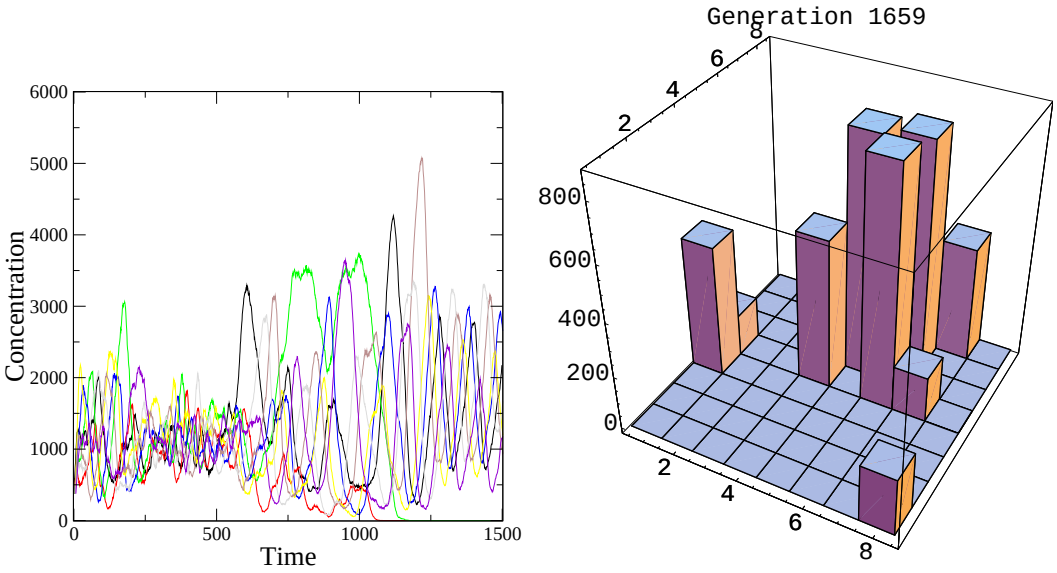


Figure 38: a) Concentration of species against time. Mutations start at timestep 300. By timestep 1200 a 6-cycle is established. b) Number of reactions in generation 1659. The interaction matrix is represented in the x-y plane.

the hypercycle against parasites in this case is not solved by placing the molecules in a two dimensional space.

5 CelloS

5.1 From molecules to simple cells

In our non-trivial task to devise genotype-phenotype maps, many of the difficulties stem from the complexity of even the simplest cells, which precludes a representation of an entire cell at the molecular level. Spatial and temporal organization take a principal role in higher order organisms, thus making it impossible to reproduce all the ascpects of the developemental process. On the other hand, at present there are no established "intermediate-level" theories that would provide consistent but simplified representations of cellular processes (energy metabolism, biomass production, cell division, sensory responses, intracellular transport, gene expression, etc.) which would help to understand common and basic characteristics of these processes. One therefore has to resort either to simulations based on a large number of *ad hoc* assumptions, or to the construction of minimal models based on biophysical and biochemical principles.

As we have seen in previous sections, the process of RNA folding can be viewed as a minimal model of a genotype-phenotype map. Here, the sequence of the RNA molecule acts as the genotype (the sequence information is actually heritable in *in vitro* selection (SELEX) experiments (Klug and Famulok, 1994)), while the (secondary) structure of the molecule is interpreted as the phenotype (SELEX experiments indeed often demonstrate a strong structure dependence of the selected nucleic acids). As we showed before, detailed investigations of the RNA model lead to the development of important concepts, such as neutral networks percolating sequence space, the phenomenon of shape space covering, and the importance of accessibility for phenotypic evolution (Schuster *et al.*, 1994; Fontana and Schuster, 1998). The structure of the genotype-phenotype map determines the structure of the fitness landscape (Stadler, 1999) which in turn determines the dynamics of an evolving population. The high degree of neutrality of the RNA folding map, for example, explains punctuated equilibria in the absence of external events (Forst *et al.*, 1995b; Huynen *et al.*, 1996), leads to a selection for robustness against mutations (van Nimwegen *et al.*, 1999) and influences evolvability (Ebner *et al.*, 2001b).

Concepts such as epistasis and phenotypic plasticity easily translates into this RNA folding metaphor (Fontana, 2002), however, important characteristics of

the genotype-phenotype maps of biological organisms, do not have a counterpart in this framework: While genotype and phenotype are embodied in the same physical entity in the RNA model, there is a rather strict separation between genomic information and functional molecules in all biological organisms. This allows an organism to exist in different internal states (that depend on its *individual* history) which may cope with environmental conditions in different ways. Regulatory networks are at the core of the mechanism by which cells individually adapt to changing conditions, see e.g. (François and Hakim, 2004; Deckard and Sauro, 2004). The majority of the artificial gene regulation models used today (Banzhaf, 2003; Eggenberg, 1997; Geard and Wiles, 2003; Reil, 1999) are based on the well established "operon model" of gene expression (Jacob and Monod, 1961), which divides the genes into two classes: (i) the transcription factors capable of binding to the DNA thereby modulating the expression of downstream located genes; and, (ii) structural proteins which perform some functions different from the regulation of the gene expression. In the simplest case, regulatory networks arise when transcription factors also enhance or inhibit the expression of other transcription factors. (Note that such models still ignore crucial regulation mechanisms of real cells such as signal transduction networks and post-transcriptional gene silencing.)

The `CelloS` model described in the following pages combines a simple computational cell model, the extended Potts model (see (Merks and Glazier, 2005) and references therein), with an artificial genome and a minimal model of gene expression (Reil, 1999). This combination allows us to study the coupling of the environmental dynamics to the cell internal dynamics of gene expression within the framework of an evolving cell population. It also aims to be a step further in the direction of separating the genotype from the phenotype, contrary to the rest of the models in this dissertation which regarded the phenotype as another face of the same molecule or molecules representing the genotype.

The simlulations presented here are motivated by the cell differentiation of the amoeba *Naegleria gruberi*, which is capable of changing cell shape, from a crawling amoeba to an asymmetric elongated cell, and of growing flagella when nutrients are scarce in order to move following a concentration gradient. It has been shown (Fulton and Walsh, 1980) that all proteins necessary for the differentiation are synthesized *de novo*, i.e., due to transcriptional regulation. The initiation of morphological changes require the synthesis of sufficient amounts of proteins,

i.e., a significant investment. The transformation is temporal and the organism returns back to the amoeba state when nutrients are again available. *N. gruberi* divides in the amoeba state only, while the flagellate state is much more mobile and hence better suited to explore novel nutrient sources.

5.2 The model

The basic tool for our simulations is the Potts model with some extensions (Marée and Hogeweg, 2002) on a two dimensional lattice. A *cell* C is a maximal connected subset of the lattice such that all lattice points in C have the same type or "color" u. Lattice points belonging to a single cell are only distinguished between border and interior sites, i.e. cells are homogeneous. Cells interact with each other with strength J_{uv} at neighboring lattice points depending on their types u and v. This interaction is defined as the energy increase provoked by a neighboring cell. A special type 0 denotes empty lattice sites. Each cell is characterized by its energy

$$E_C = \sum_{i \in \partial C} \sum_{j \in N(i) \setminus C} J_{u_i,u_j} + \lambda(\mathrm{vol}(C) - V)^2 \tag{8}$$

where $\mathrm{vol}(C)$ is the volume of the cell, i.e., its number of lattices points, ∂C its boundary, V is a user-defined target volume, $N(i)$ is the set of neighbors of i and λ is a compressibility parameter. The double sum runs over all lattice edges that point from the boundary (surface) of the cell C to other cells or into the environment. The environment contains nutrient spots distributed randomly along the surface. These sources produce a concentration gradient described by c_i at lattice point i. Diameter of the sources and the amount of nutrient they contain is variable, and only inside them cells are allowed to profit from the nutrients.

Cell motion is implemented by a simple Metropolis Monte Carlo step in which a cell attempts to modify its boundary at lattice point $i \in \partial C$ by changing the type of an adjacent site i' to its own type, or by changing one of its boundary sites to 0. The transition probability is

$$\begin{array}{rll} 1 & if & \Delta E_C < H_\partial \\ \exp -(\frac{\Delta E_C + H_\partial}{T}) & if & \Delta E_C \geq H_\partial \end{array} \tag{9}$$

where H_∂ is the energy cost of deforming the cell's boundary and T a temperature-like parameter. In order for the cells to feel the gradient in the nutrient, the energy

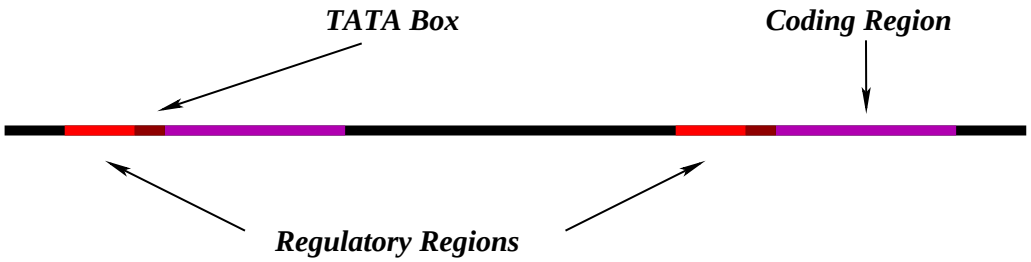


Figure 39: Genome of the `Cellos` model. Markers are localized along the genome to define the start of a gene. The following 40 bases are taken as the "coding" region. The previous section is the regulatory region for the corresponding gene.

change is reevaluated as

$$\Delta E_C^* = \Delta E_C - \mu_0(c_{i'} - c_i) \tag{10}$$

where μ_0 describes the reactivity of the cell to changes in the nutrient concentration.

Note that cell motions are internally driven and hence consume energy rather than the result of molecular Brownian motion. Our cells have a finite life expectancy and require energy to stay alive. This is modeled by a "battery" that is used up when enzymes are synthesized. When the "battery" is empty, the cell dies and the corresponding lattice sites are reset to 0.

Each cell on the lattice contains an RNA sequence of length 200 which represents its genome and contains the information necessary to decode the cell's behavior. This genome can encode two types of effector molecules (corresponding of course to proteins in *N. gruberi*, but modeled as RNAs here for computational convenience) and a simple regulation mechanism.

A short signal sequence (corresponding e.g. to the TATA box in real cells) marks the beginning of a "coding region" on the genomic sequence. We use the signal GC and define a gene to be the following 40 nucleotides (Fig. 39).

This subsequence is folded into its secondary structure using the `RNAfold` program of the `Vienna RNA Package` (Hofacker, 2003). The structure is then compared with two target shapes for the "motion effectors" and the "nutrient importers", which are kept fixed throughout the simulation. The closer target shape determines the function of the gene, while the number of base pairing differences measures the gene's efficiency (Fig. 40).

In the current implementation we keep the gene regulation network fixed. In order

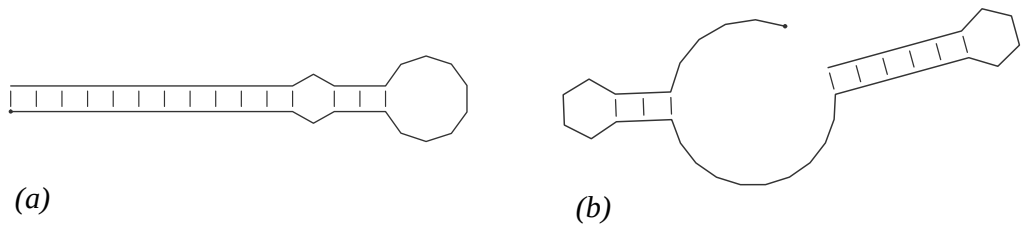


Figure 40: Target Shapes. Two classes of functionally different RNAs are distinguished by archetypic shapes: (a) motion effectors and (b) metabolic effectors that act as nutrient importers.

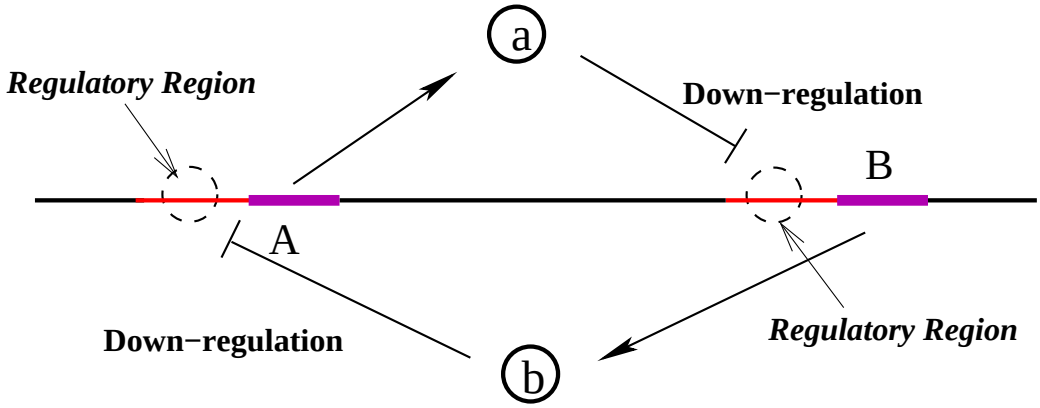


Figure 41: Gene Regulatory Network. The simple mutually repressing network. Products of each type of gene repress the translation of the other type.

to implement the switching between the motion effectors and nutrient importers we use the simple negative feedback system shown in Fig. 41. The differential equations for this scheme are:

$$\begin{aligned} \frac{dG_A}{dt} &= \gamma_A \cdot k \frac{1}{1+G_B^3} - d \cdot G_A \\ \frac{dG_B}{dt} &= \gamma_B \cdot k \frac{1}{1+G_A^3} - d \cdot G_B \end{aligned} \tag{11}$$

where G_A and G_B are the concentrations of the two types of gene products, γ_A and γ_B are their efficiencies, and k and d fixed constants. A 4th order Runge-Kutta method is used to numerically integrate these differential equations.

Once the genome is decoded, the concentrations of the gene products are computed. The cell is then able to feed depending on the available nutrient in the environment provided it expresses nutrient importers, and to move if motion efforts are expressed. The battery level B is decreased depending on the amount of gene products that are produced and it is recharged if the cell is in a food source:

$$B' = B - c_0(G_A + G_B) + \phi_0 G_B \tag{12}$$

The parameters c_0 and ϕ_0 describe the ratio of nutrients obtained from the environment against the cost of producing the importers and motion effectors, respectively. The mobility of the cell depends on the concentration of expressed motion effectors which is reflected in a modified transition probability for changing the cells boundary by replacing the constant μ_0 with $\mu_0 \cdot G_B$. Therefore, equation (10) reads as

$$\Delta E_C^* = \Delta E_C - \mu_0 \cdot G_B(c_{i'} - c_i) \tag{13}$$

In order to link the internal state of the cell to the environment, we give an impulse to the concentration of nutrient importers every time the cell is touching a food spot. This is done by increasing the importers concentration by a fixed amount and then integrating the equations again. If the gene effectiveness are in the correct range, the equations will react to this impulse and concentrations will flip to the desired values, i.e. concentration of importers will surpass that of movement effectors and stabilize in that state. On the other hand, an impulse is given to the movement effectors whenever they are not touching a food source. The amplitude of this impulse regulates how different the efficiencies can be in order to obtain the switching of the effectors' concentrations.

The products of metabolic genes play two different roles: first, they recharge the battery of the cell; and second, they increase the cell's target volume. Both the battery and the target volume are increased proporcionally to the metabolic effectors concentration.

Once a cell has doubled its normal size, it divides by fission copying its genome to the new cell. This process is usually inaccurate, producing mutations in the new RNA string. In this model every replication implies one random point mutation in the genome. Even when this is the simplest way of mutating the genome (among others as deletion, gene duplication or insertion, for example), non-linearity and complexity arises from the characteristics of the genotype-phenotype map used. Genes may be destroyed or created whenever a marker (TATA box) is deleted or

formed. At the same time, if the "coding" region of a gene is touched, the non-linearity of the folding map between sequence and secondary structure is reflected in the overall phenotype of the cell. Total efficiency of the genes is obtained by adding up individual gene efficiencies, therefore, efficiency may increase either by optimizing the structure of existing genes or by creating new ones.

Food sources are depleted when cells feed from them. Once a source is empty, it is replaced by a new one in a randomly chosen spot of the lattice. This way, cells are forced to switch between the metabolic and movement states, reinforcing the selection of only those capable of doing so.

Individual cells with very similar genomes belong to the same *species*. The definition of species in our model is similar to that proposed by Kenneth and Risto in (Kenneth and Risto, 2002). Each gene in the population has a unique historical number. Every time a mutation creates a new gene or changes the type of an old one, this global variable is increased and assigned to the new gene. In order to compare two genomes, we use a linear combination of the number of excess (T) and disjoint (D) genes, and the average efficiency difference between common genes (W). If the result of

$$\delta = \frac{c_1 T}{N} + \frac{c_2 D}{N} + c_3 \cdot W \tag{14}$$

is below a threshold value, the new cell is assigned to the same species as the old one. Whenever a new species is created, a genome is set to represent the whole species. Every time a new cell is born, its genes are compared to all species' genomes and included in the first one for which the distance is below the threshold.

5.3 Results

5.3.1 Population size

Throughout all of our simulations, some parameters are kept fixed: we use a lattice of 200×200 sites with periodic boundary conditions, $J_{x,0} = 11$ for the contact with an empty site, $J_{ab} = 37.5$ for the contact between different cell types, and $J_{aa} = 35$ for the contact with a cell of the same species. Furthermore $T = 3$, $H_\partial = 0.8$, $\mu_0 = 5000$, $c_0 = 0.4$, $V = 30$, $\lambda = 5$.

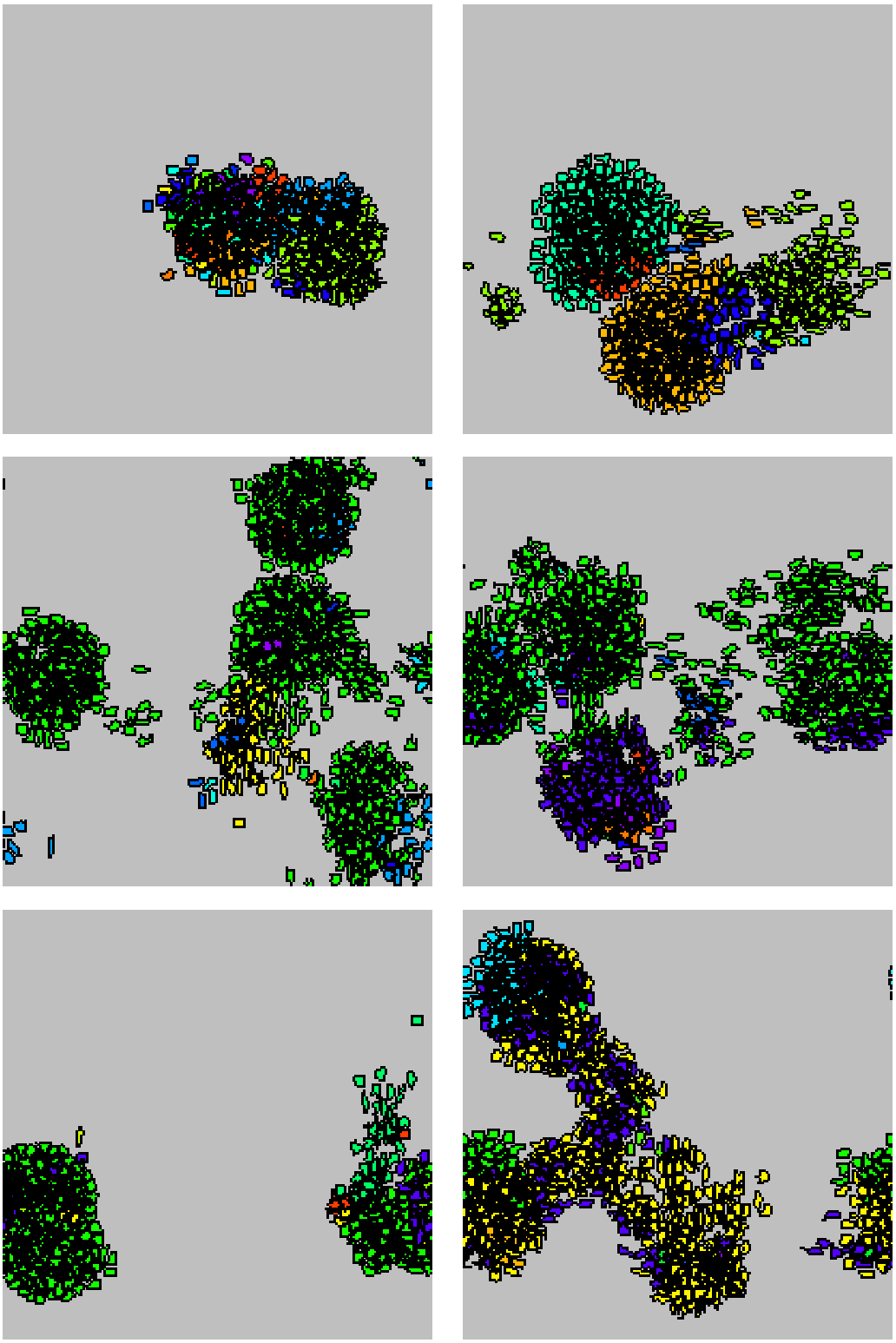


Figure 42: Snapshots of a run with three food sources. The evolution of the system is shown at timesteps: 135, 495, 6000, 12225, 15030 and 19800.

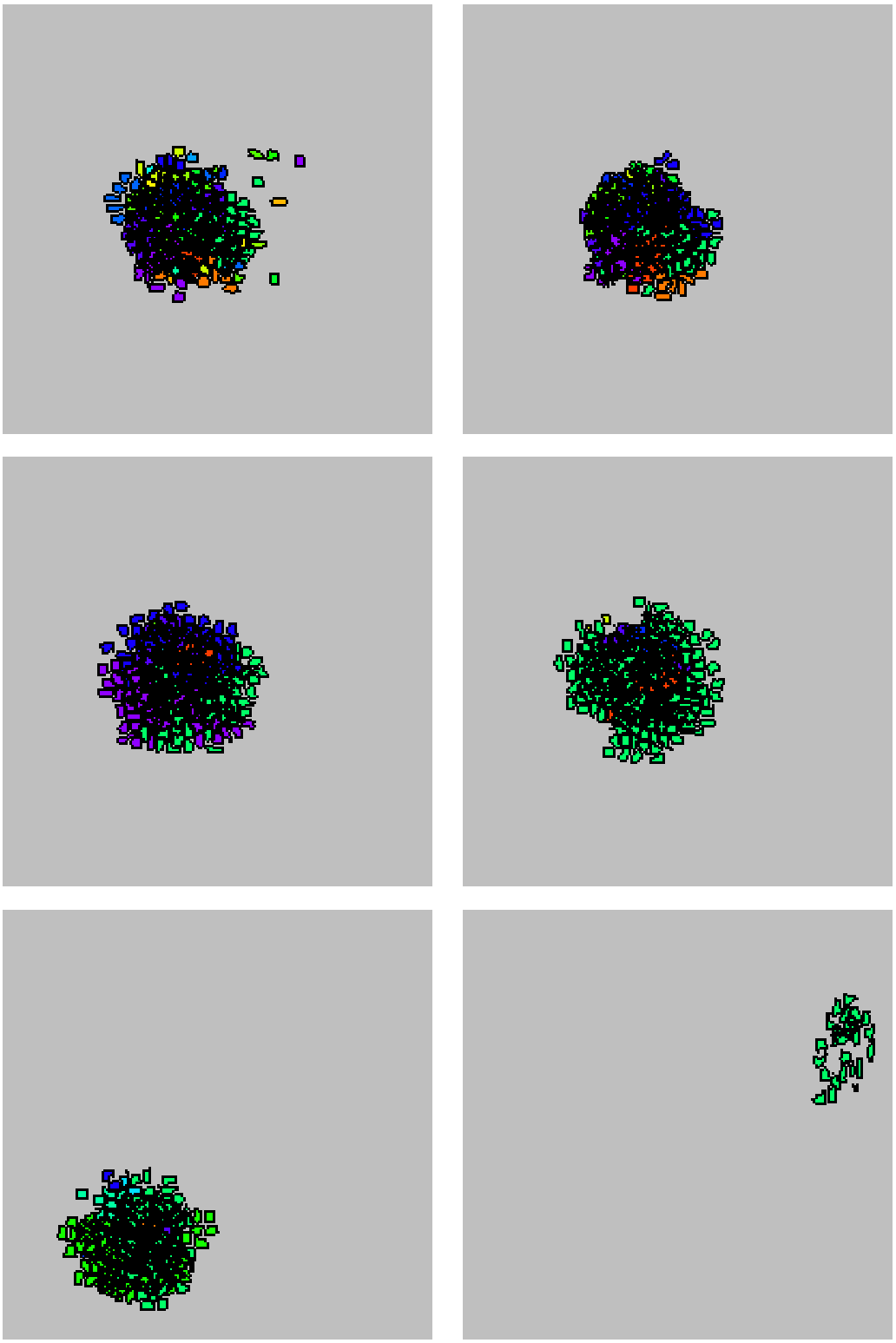


Figure 43: Only one source in the lattice. One source with finite energy is placed in the lattice. Cells are able to feed only from this spot, thus making the population much smaller than in the previous case. Snaps are from generations: 120, 300, 525, 1050, 2250 and 7500.

Figure 42 shows the evolution of the system for a typical simulation run. This images were created with three food sources available for the cells to eat. The location of these is not visible for clarity purposes, nevertheless, it can be implied from the accumulation of cells in certain places of the lattice.

Population size changes depending on the conditions. As cells feed from available nutrients, their volume increases and more duplication events occur. Since cells have a finite life span, population size cannot increase arbitrarily inside certain range of parameters. The relation between the life time and the increase of volume per generation is the factor regulating population size, together, of course, with number of food spots.

Figure 43 shows a run with only one food spot. In this case a single food source is moving around the lattice, maintaining the population size small and relatively constant. In cases with more than one source, cells may or may not feed from all of them. On the contrary, in order for the system to survive with this setting, cells must be feeding from this single spot all the time. This is why variation in the population size is smaller than in the previous case.

The population grows depending on the availability of nutrients. Every time a food source is depleted, cells must migrate to the next one. This periods are usually reflected in a diminution of the population and increase in the average number of movement genes in it. The second panel in Fig. 44 shows the energy of the sources and the change in the number of cells. Source energy staying at its maximum means that there are no cells feeding from it. This is clearly related with a decrease in the population size (Fig. 44).

5.3.2 Genome structure

We measure the impact of the external conditions in the genome by looking at the number of metabolic and movement genes, their efficiencies and the effectors expression inside and outside a food source.

The regulatory network we are using, imposes a well defined range in which gene efficiency must lay in order to obtain the necessary switch between states. In our simulations it is clear how these numbers are controlled by natural selection when the genome is mutating randomly. In Fig. 45 it can be seen how after a period of adjustment, the population falls in a regime where both efficiencies are

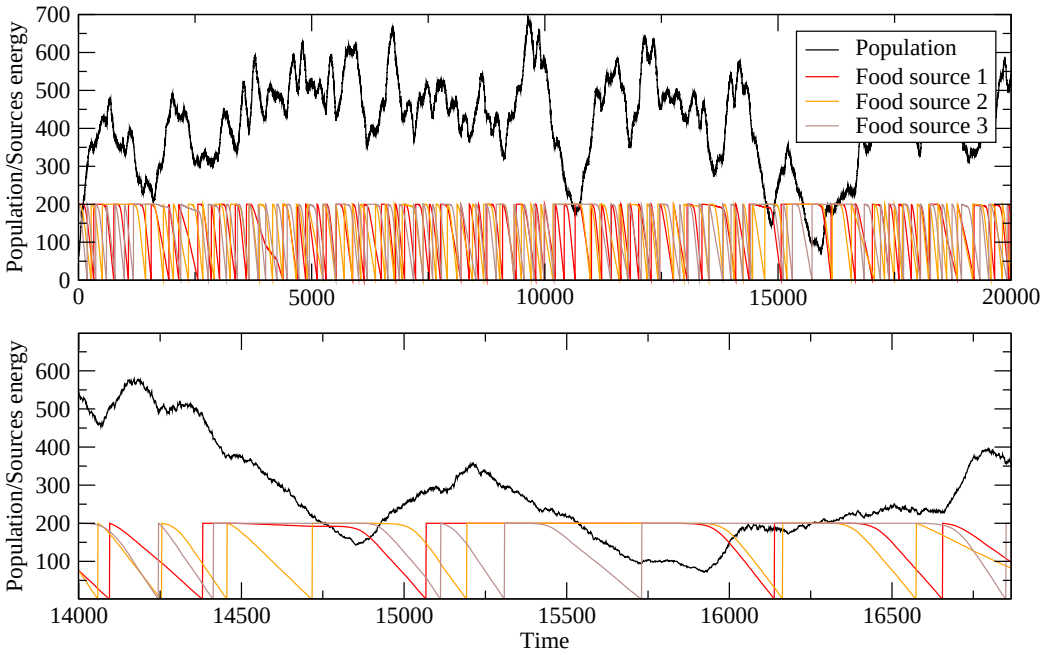


Figure 44: Population and energy of the sources. The zoom in the bottom shows how population grows only when cells are feeding from the sources.

inside a small interval which allows switching. Selection is also reflected in the larger number of metabolic than movement genes. Even when the structure for movement genes is more common than that of metabolic, fact that is reflected in the start of all simulations with more genes of the first than the second, this is soon reverted and stays like that througout most of the simulation (Fig. 45, 46).

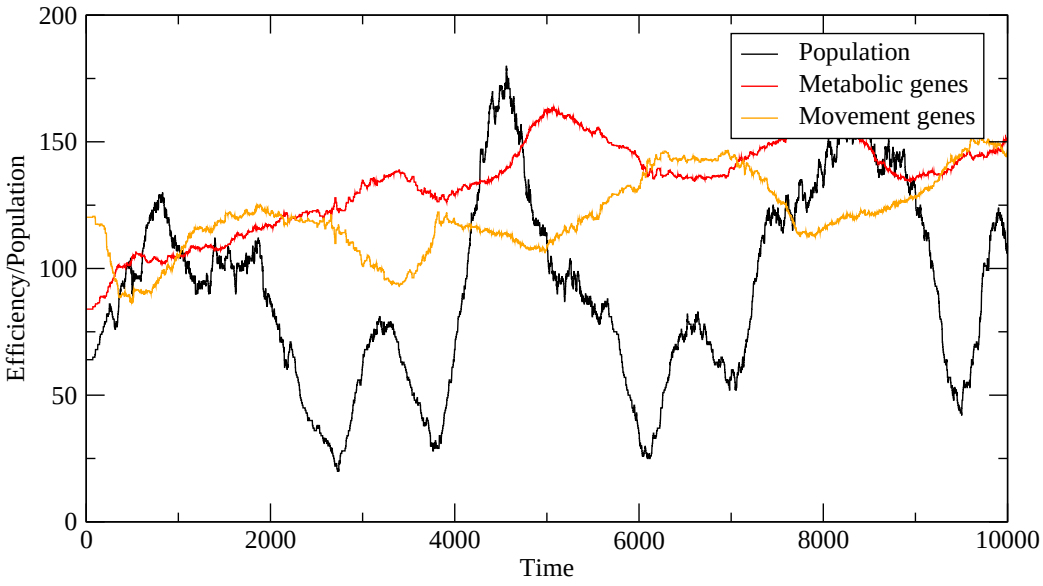


Figure 45: Efficiency of genes and population in a simulation with parameters: number of food spots 3, mean life 500, volume increment per generation 0.07

Since genome size is fixed all along the simulation, the total number of genes starts around the expected number of genes . Whenever a mutation occurs, it is easier to hit the "coding" region of a gene (length 40 nb), than the start-marker (2 nb), or than create a new starter. Therefore, mutations produce the change between gene types in most of the cases. This is the reason why graphs show very symmetric curves for gene number and gene efficiency (Fig. 45 or 49). Nevertheless, environmental pressure and selection increases the number of total genes in around 50 percent. The expected number of genes G is computed as

$$G = \frac{L_{genome} - (L_{tata} - L_{gene})}{(L_{tata})^A} \tag{15}$$

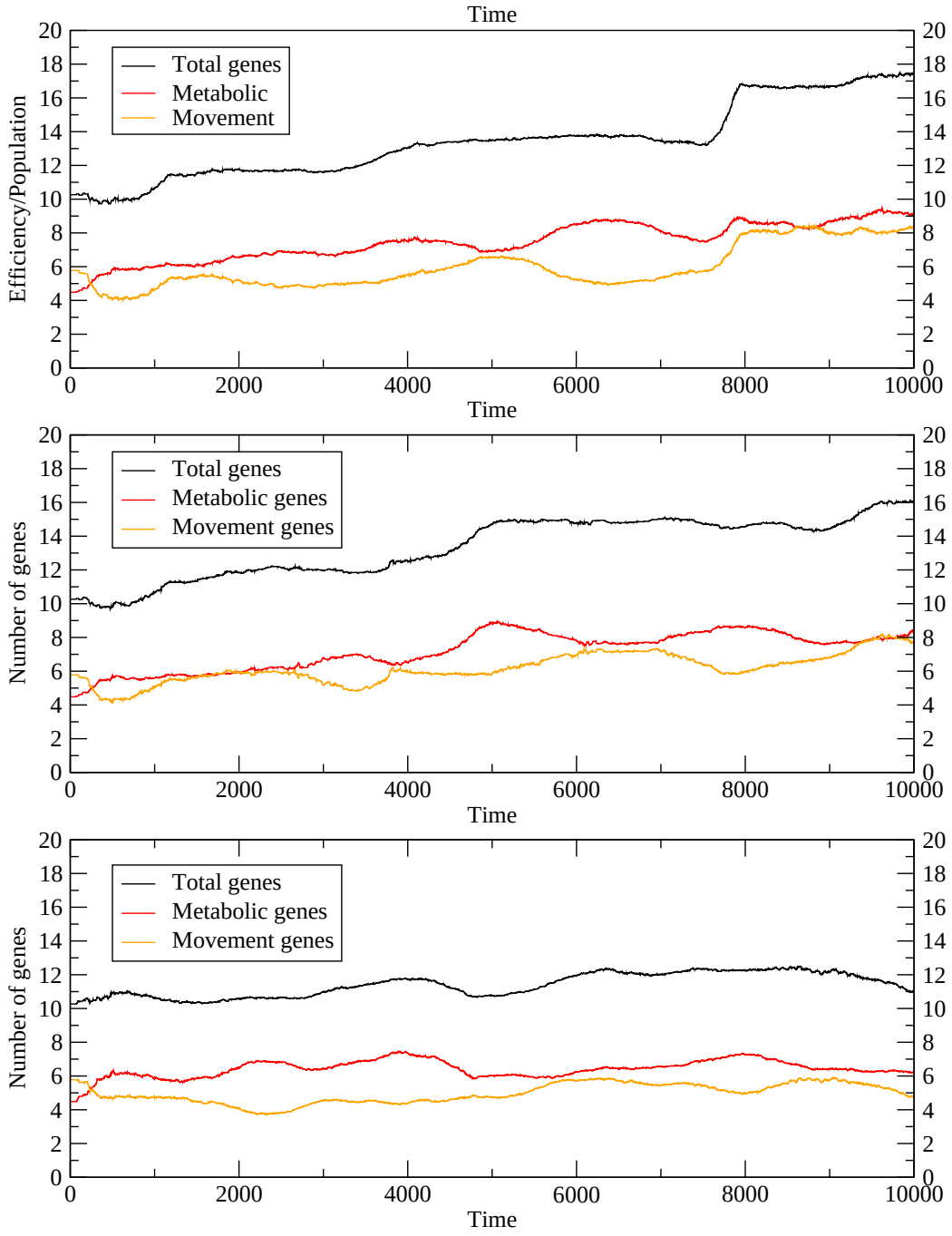


Figure 46: Above, gene number increases due to selection and environmental pressure. The expected number of genes 9.875 (begining of the simulation), is increased up to 18 at the end of the run. The difference among efficiencies is larger than in the second graph because of the change of the impulse from the environment from 10 to 13. The third graph shows how gene number is kept almost constant due to very low selection pressure. The 7 food spots in the lattice give enough food to the cells to survive without increasing number or efficiency of their genes. Same paremeters as in Fig. 45 except for impulse stength and spot number.

Where L_genome, L_gene and L_tata are the length of genome, gene and "TATA box" repectively and A is the size of the alphabet. With the parameters in these simulations, the expected number of genes is 9.875.

The right combination of gene efficiencies allows a switching in their products expression depending only in the presence or absence of food from the environment. The numerical integration of the equations show how this switching is possible when giving a strong enough impulse. This can be seen in Fig. 47. The difference between the efficiencies of both kind of genes is controled by the strength of the impulse. In figures 46(up and middle), the impulse was variated from 10 in the first one to 13 in the second. This also gives the population more resistance to variation in the environment.

If the amount of nutrients increases in the environment (i.e. by increasing the number of food spots), the population size grows and the gene number stays very close to the expected one. This is because the pressure to increase the efficiency of genes is lower given that there is enough nutrients to survive(Fig. 46(up and down)

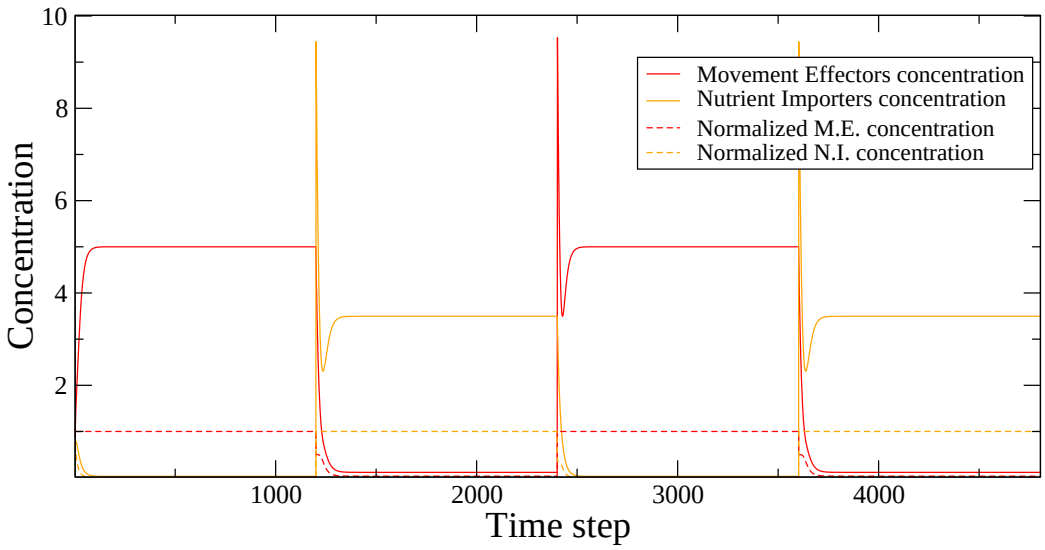


Figure 47: Concentrations of gene products. Numerical integration of the regulation network equations. The efficiency of metabolic genes is 70 and of movement genes 50. An impulse (of 10 arbitrary units) is given to the concentrations every 1200 timesteps

Figure 48 shows the behavior for a single cell with the right number of genes. It can be seen how whenever there is food present, nutrient importers are expressed and movement effectors repressed.

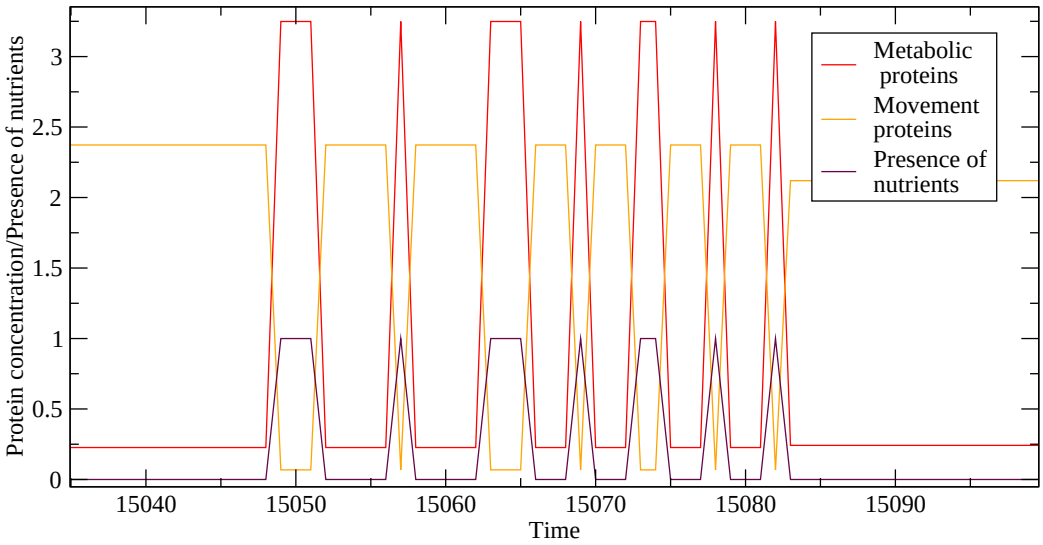


Figure 48: Switching of gene products expression depending on the presence/absence of nutrients. Same parameters as in the previous figure.

In the special case when there is only one food spot of infinite life in the lattice, cells that are in the spot are thrown out of it by the newborns. Even when there is no need of traveling long distances, the fact that cells have to be constantly coming back into the source makes the presence of movement genes indispensable.

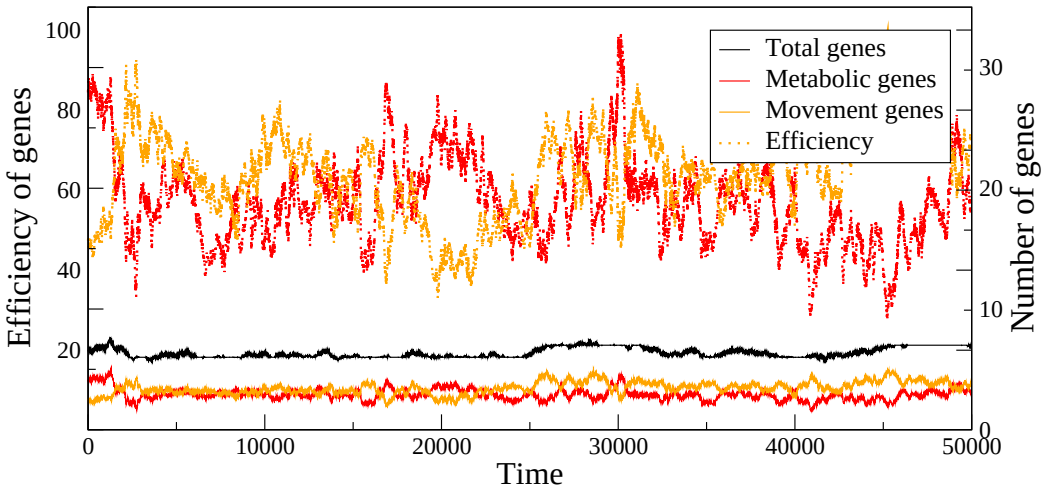


Figure 49: Gene number and efficiency for one food spot.

At the same time, since food is easily available, there is no need to increase the efficiency of metabolic genes. Battery may be refilled slowly without killing the cell since the time it spends outside the food source is usually very short. This

can be seen in Fig. 49, where efficiencies and number of genes are still close one to each other keeping a minumum number of movement genes.

5.3.3 Phylogenetics

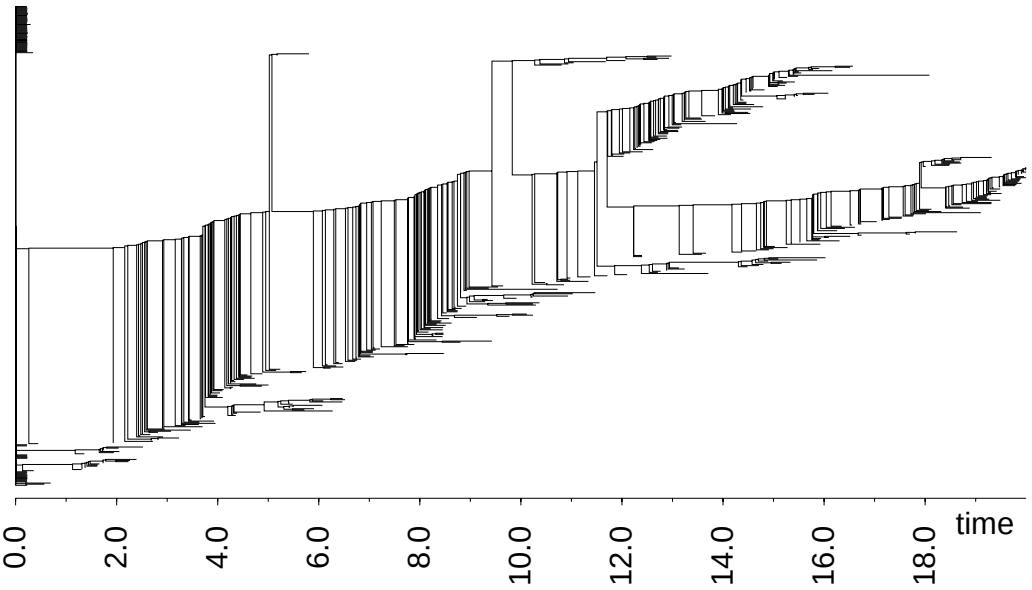


Figure 50: Phylogenetic tree for a run with three food spots. Nodes in the tree represent the disappearance of a species, while saddles stand for the split of two of them. Time unit is 1000 simulation steps.

With our simple definition of species, the number of species depends directly on the volume increase per generation. Phylogenetic trees can be recorded based on the speciation events, see Fig. 50 for a characteristic example. The Darwinian evolution is dominated by one or a few species at any given point in time. The coexistence of distinct lineages over longer times is comparably rare. In some runs one of the initial species survives until the end of the run, failing to find any important improvement in phenotype via mutations. In the case where only one food spot is present, the coexistence of more than one species is very rare.

When more than one spot are available, the population may split for a while, until food sources disappear and put the species in direct competition for the same nutrient source. The survival of one species usually depends on the efficiency of the nutrient importers. Once species are capable of moving at enough speed, those who feed faster replicate better and therefore oversize the other species.

6 Conclusion and Outlook

6.1 Three approaches, one goal

The study presented in this dissertation on evolution of interacting molecules differs from previous studies on molecular evolution ((Eigen, 1971), (Schuster and Sigmund, 1983)) in the way fitness is computed. Traditionally, fitness was calculated from properties of individual species alone, here we take interactions between one or more species into account.

As it was expected, we found that populations in well stirred environments are prone to the invasion of parasites (Eigen and Schuster, 1979). This parasites are produced by mutations of already existing species and are not introduced from the exterior. This is important because in principle membranes or spacial separation of the population would not be enough to protect the system against invasions. From the results on the hypercycle in a two dimensional space, we know that some spatial organizations can result in the expulsion of the parasites, but this is not always the case. We will come to this point again in the next paragraphs.

In all models presented in this dissertation, there was always the opposition of fold-cofold as a first step in the genotype-phenotype map. On the molecular level, the usage of the cofold map resulted in the impossibility of a network to evolve towards a predefined target, given that the neutrality of the map assigning fitness values is much lower than in the single molecule folding map. It was also shown that with too few interactions, the population falls into a frozen configuration when local maxima are found. Therefore the question arises, how a stable and robust molecular reaction network can be constructed which admits enough variation to change and evolve.

The fitness landscape in these models is continually changing, given that fitness values for single species depend on the interactions with the rest. When perfect catalysts exist in the system, the fitness landscape of its catalyzed species changes, taking them to zones where their own phenotype is not crucial anymore but the interaction with the catalyst becomes the only important factor. At the same time, reducing the pressure for these species, may reduce also their catalytic activity, thus forcing the next species in the cycle to improve their self-replication

rates. This is a clear example of how changing fitness landscapes influence the behavior of co-evolving species. The fitness of individual species loses importance, and the system must be explored as a combined set of species and their interactions and evaluated as a whole.

In presence of epistasis, the variability of one trait depends on more than one gene, thus making it more difficult to change the expression of the given trait (Wagner and Altenberg, 1996). When using the cofold map, a "combined" genome is created, which produces only one phenotype. We showed that the neutrality for such a map is nearly as high as the one presented by the single molecule folding map. Nevertheless, if a molecule is required to interact with more than one molecule, the neutrality is radically lowered. This result has crucial implications for systems where molecules must carry out more than one function. The hypercycle is a good example of such a system.

In chapter 3 we presented a model of an hypercyclic network that incorporates strong interactions between species and hence a complicated population dynamics, spatial organization, and an explicit representation in sequence space. Our first main conclusion is that the behavior of such an integrated computer simulation is consistent with earlier findings on both the population dynamics (such as the existence of limit cycles) of hypercycles and on the effects of considering a spatially extended system (such as the formation of spiral waves and resistance against various types of parasites). The resistance of the system against short-cut parasites in addition to "dead-end" parasites is a very important result since it shows that spatially extended hypercycles are indeed evolutionarily very stable systems as long as the fitness (i.e. replication rates) depends only on a single molecule and not on the interactions among more than one molecule. This is in sharp contrast to hypercycles in homogeneous solution (Eigen and Schuster, 1979; Bresch *et al.*, 1980; Stadler and Happel, 1993; Stadler and Schuster, 1996).

Furthermore, we demonstrate here a mode of sequence evolution that is dominated by drift and hence can be described in terms of Kimura's Neutral theory (Kimura, 1955, 1983). This does not mean, of course, that selection does not play a role: the exclusion of parasites, the internal dynamics of the population, as well as the sequence-evolution in the initial phase of the simulation are clearly dominated by selection. It is important to point out that this kind of sequence evolution is achieved thanks to the neutrality of RNAfold used as the genotype-

phenotype map.

The main characteristics of the system are robust and differ only in small details from one set of conditions to the other. We therefore conclude that the ability of such an RNA based system to evolve towards a robust spatially extended organization with diffusion in sequence space is intrinsic to autocatalytic self-replicating molecules as soon as the sequence-structure relationship is dominated by extensive neutral networks, as is the case for RNA.

In contrast to the case described above, the hypercycle with cofold as genotype-phenotype map in not stable against any kind of parasites. In this implementation, interactions among molecules depend on the base pairing of the concatenated sequences, this being responsible for the main difference with the folding model: with cofold, the interaction between RNA strings occur before the molecules fold into their secondary structures, while in the single molecule folding, the interactions are decided according to the already found minimum free energy configuration of the individual sequences. This means that a mutation in a single sequence is more probable to change its interactions with the rest, given that the neutrality of the cofolding map is much lower than the neutrality of the single molecule folding map. Research about kinetic properties of the folding map is being pursued at the moment (Wolfinger *et al.*, 2004), which is of great importance to decide which of the maps is closer to the actual interaction among RNA molecules.

This model results to be unstable against parasites because of the very low neutrality of the cofold map and the freedom given to the molecules to interact with any other molecule in the system. In the first implementation, the topology of the interactions was fixed to a cycle and parasites where explicitly introduced into the program. The possibility of having more interactions created by the cofolding of all sequences, changed dramatically the topology of the network. Short-cut parasites where immediately produced and reactions belonging to the original cycle were lost because of the low neutrality of the map. The connectivity of the network is the most important parameter to be controlled in order to obtain interesting behaviors (Kauffman, 1993), in our model this value depends on the cofolded structures and so it is difficult to control externally.

Another important characteristic for the study of the genotype-phenotype map is the relation between the evolving species and its environment. In our first (and

very simple) implementation of `CelloS`, we observe the response of the genome to variable environmental conditions. After an initial phase of selection the number of genes stays approximately constant. The cells can then use their gene regulatory network to cope with environmental changes. Population dynamics also reflect the presence or absence of nutrients, together with an increase of the number and/or the efficiency of movement genes. We found that, at least in our simple environment, it is not important to have a large number of genes, but to have the right amount of them depending on the environmental inputs and the regulatory network modifying their products' expression.

The genotype-phenotype map in this model can be clearly separated into three levels: the first one is the folding of gene sequences into their secondary structures, assigning function and efficiency to each of them. The second part is the interaction between the gene products which modify gene expression and thus the resulting concentrations of functional proteins. The last level defines the behavior of the cell given the changes on protein expression and the influence of the environment upon them. The interaction between these levels is still hard coded in the model and is far away from being a realistic representation, nevertheless, we consider this model a good tool to study different aspects of the genotype-phenotype map. Neutrality exists in the map thanks to the use of RNAfold in the characterization of genes; epistasis is a direct consequence of the regulatory network used; plasticity is the main attribute of switching expression levels and a changing environment forces the population to adapt and evolve towards a well defined task.

6.2 Different levels of genotype-phenotype maps

The relation kept between the molecules conforming a cell and the cell itself is one of belonging and bringing to being at the same time. It is an autopoietic system (Maturana and Varela, 1980) in the sense that molecules need and are needed by the rest of the construct in order to maintain and reproduce themselves. This organization can be understood only when studied as a whole. Breaking the machinery down to its primary components may give us some hint on how this parts work, but will hardly tell us how or why they are organized the way they are.

One possible way of approaching this question is by defining and studying the different levels forming a living organism. Much has been done on the level of molecular interactions and on the much larger scale of tissue differentiation and morphogenesis. In present time, big efforts are being done to reveal the intricate networks of gene regulation. This level was unknown until recently, and will for sure throw much information about the relation between the two levels mentioned before.

These networks are in part responsible for giving a system more plasticity than it is possible to achieve without. The first model presented here, in which there is only the molecular level to represent genotype and phenotype, is constrained to a reduced number of outcomes depending on the concentrations and structures of the individual molecules. Moreover, once this structures are defined, there is no change allowed until mutations are introduced. Although plasticity has been compared to the repertoire of suboptimal structures in a molecule, this is far from being a coordinated response to environmental changes, and more like a forced effect. Neither the system nor the molecules have any active possibilities to influence the repertoire of suboptimal structures.

There is one more level which can be defined in this hierarchy. Before a clear separation between structures in an organism could be reached, an intermediate state where the different components are stoichiometrically coupled is foreseeable. The tight relation between molecular structure and its function must be at some point relaxed in order to increase the possible outcomes of such a system. Increasing the number of molecules and the interactions among them is one way of doing this, nevertheless, this is clearly not the way evolution shaped higher order organisms. Stoichiometrically coupled systems, like minimal cells which combine metabolism, container and genetic information ((Rasmussen *et al.*, 2004)), must have at some point branched into modular organizations which could specialize in one side, and cooperate in the other, linking these different substructures through feedback with molecules acting both as transmitters and receptors.

In this sense, the study of coevolving species occupies a principal role. Modular or multi-species systems should be able to search the genotype space without reducing the overall fitness of the organization. That is, not only individual fitness must be evaluated, but most important, the behavior of the system as a whole.

6.3 Evolution on different levels

We know from Kauffman's boolean networks that a system in a chaotic state cannot remember the past ((Bak, 1996), (Kauffman, 1993)) because trajectories are doomed to separate no matter how close the initial conditions are. This makes evolution impossible and thus has a strong impact when studying molecular networks. To simulate interactions there is always the question about how to decide which molecules are to interact with others. Models of artificial chemistry ((Benkö *et al.*, 2003)) have tried to solve this matter, nevertheless they are still far from being a tool for evolutionary simulations. In the case we present here, the cofold map resembles somehow the possible interactions between RNA molecules. The result shows that if molecules are permitted to interact with many other molecules, the connectivity of the network is too high and the networks falls into a chaotic state. On the other side, if the interactions are too strict, most mutations will result in the lost of links between molecular species, thus breaking cycles that could have been beneficial to the network.

According to Kauffman (Kauffman, 1993), the connectivity of molecular networks in living organisms must lay in the edge between the chaotic and frozen states. From our models we could see that the parameters responsible for this connectivity value must be artificially tuned in order to get some interesting behavior. One straightforward question to ask to these models is how could a system self-tune this parameters in order to keep itself in the range needed for life to exist.

Going up in the level of organization, we proposed a system where regulation of gene expression is studied. One important aspect to remark of this model is the relation between the molecular activity of the gene products and the resulting behavior of the cell as a whole. Fluctuations in molecular concentrations are controlled by the regulatory network is acting upon them. At the same time, impulses from the environment are being collected by the cell and transformed into molecular signals which the network is able to detect. Once a network is defined, all the reactions between molecules are controlled and therefore small fluctuations in their concentrations are not important anymore for the overall behavior. This interplay between the two levels may be responsible for controlling and tuning the parameters defining the behavior at the molecular level. It is the goal of many attempts nowadays ((Rasmussen *et al.*, 2003), (Goldbeter, 2002), among many others) to understand the coupling of these reactions with the overall

function of the cell. Hopefully, the combination of new techniques together with the knowledge accumulated so far will bring some answers to the striking question of the origin and evolution of life.

6.4 Outlook

Many improvements to the `CelloS` model are currently being implemented. Since the mechanism of the regulation of gene expression in the current implementation can itself not be a target of evolution, we plan to add transcription factors as a third class of gene products to the artificial genome. This will allow the cells to find innovative regulatory strategies based on post transcriptional interaction. A fruitful route will then be to study the mixing of regulatory strategies under sexual reproduction of the cells.

The basic mechanism to decode the genome is based on the one used in the current implementation. Genes are defined through a marker (TATAbox) (see Fig. 51). In mediately before and after this marker, subsequences for regulatory and coding regions are defined, respectively. In order to set the function of a given gene, the coding sequence is folded into its secondary structure and compared to a set of targets. Genes can belong to the movement, metabolic or regulation classes. Reactions among monomers are allowed to create dimers. The probability for this is computed with the energy of the cofolded structure of the monomers' sequences. Monomers and dimers are then able to bind to regulatory regions in order to up or down regulate gene expression. This probability is computed again via the cofolded structure of monomer sequence and regulatory sequence. We expect to see different network topologies which exhibit a switching in protein expression as in the case of the simple regulatory network used in the current implementation.

Currently `Cellos` implements only point mutation to evolve sequences. Adding more sophisticated operations like gene duplication or horizontal gene transfer, turns `Cellos` into a tool for generating test data for phylogenetic reconstruction methods. Comparing the simulated evolutionary scenario with the reconstructed one will allow to evaluate the performance of such methods. `Cellos` could also be a good tool to study the way the genetic coding evolves in replicating systems that have more than one level or organization (Wills, 2001).

Extending the set of mutation operators from point mutation to gene duplication

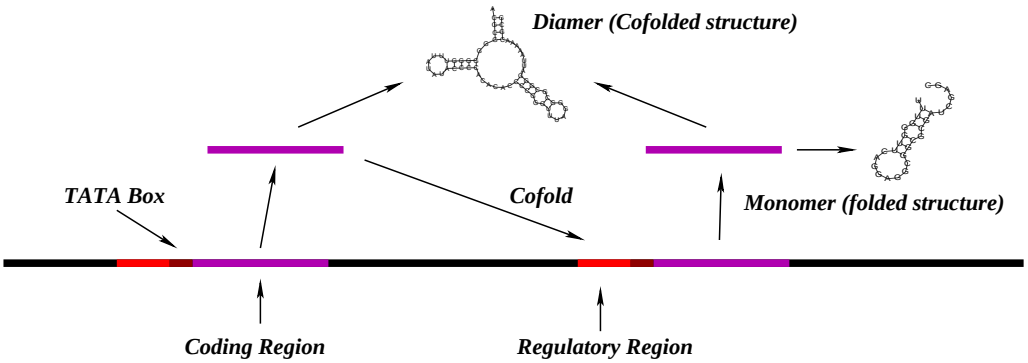
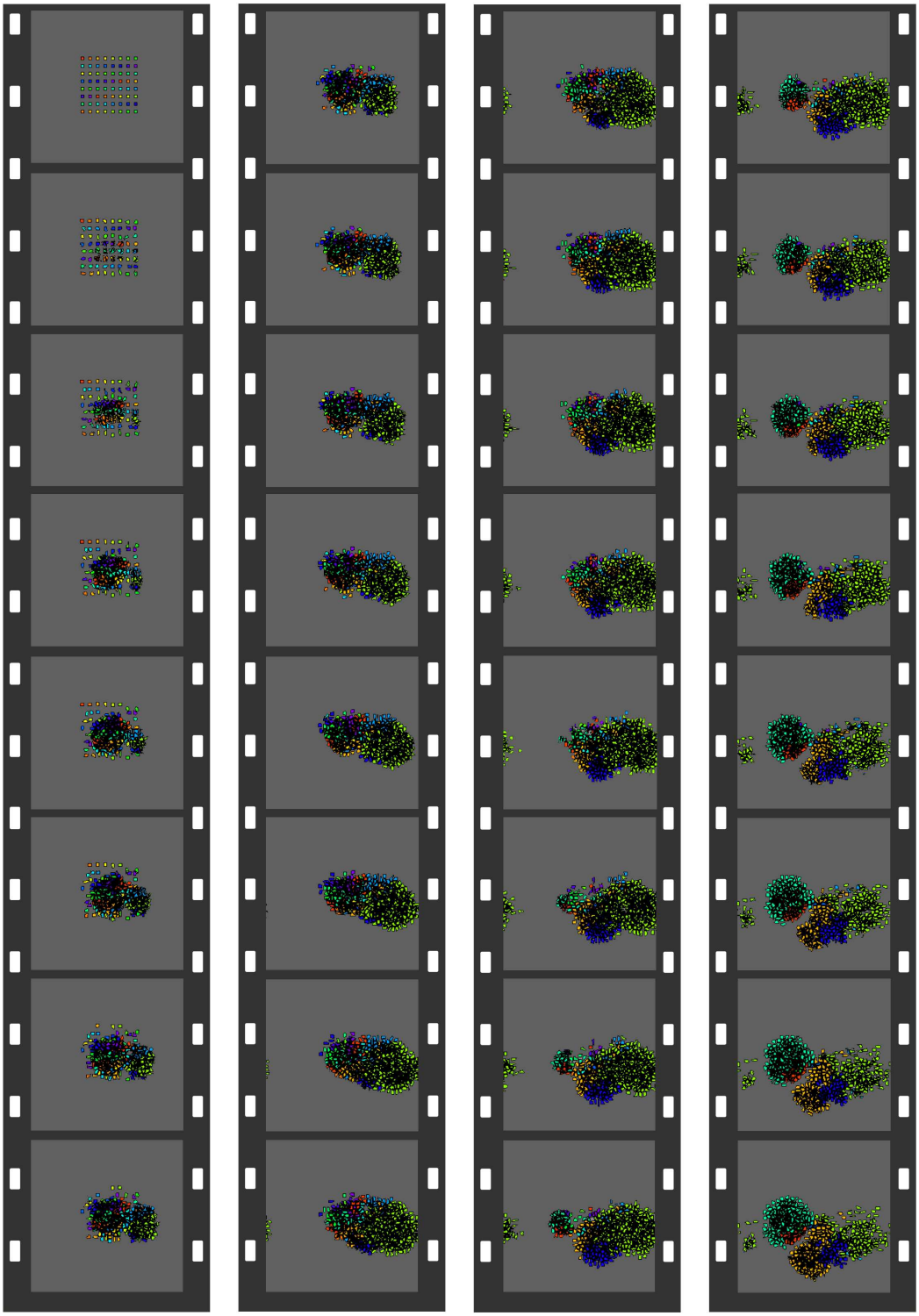


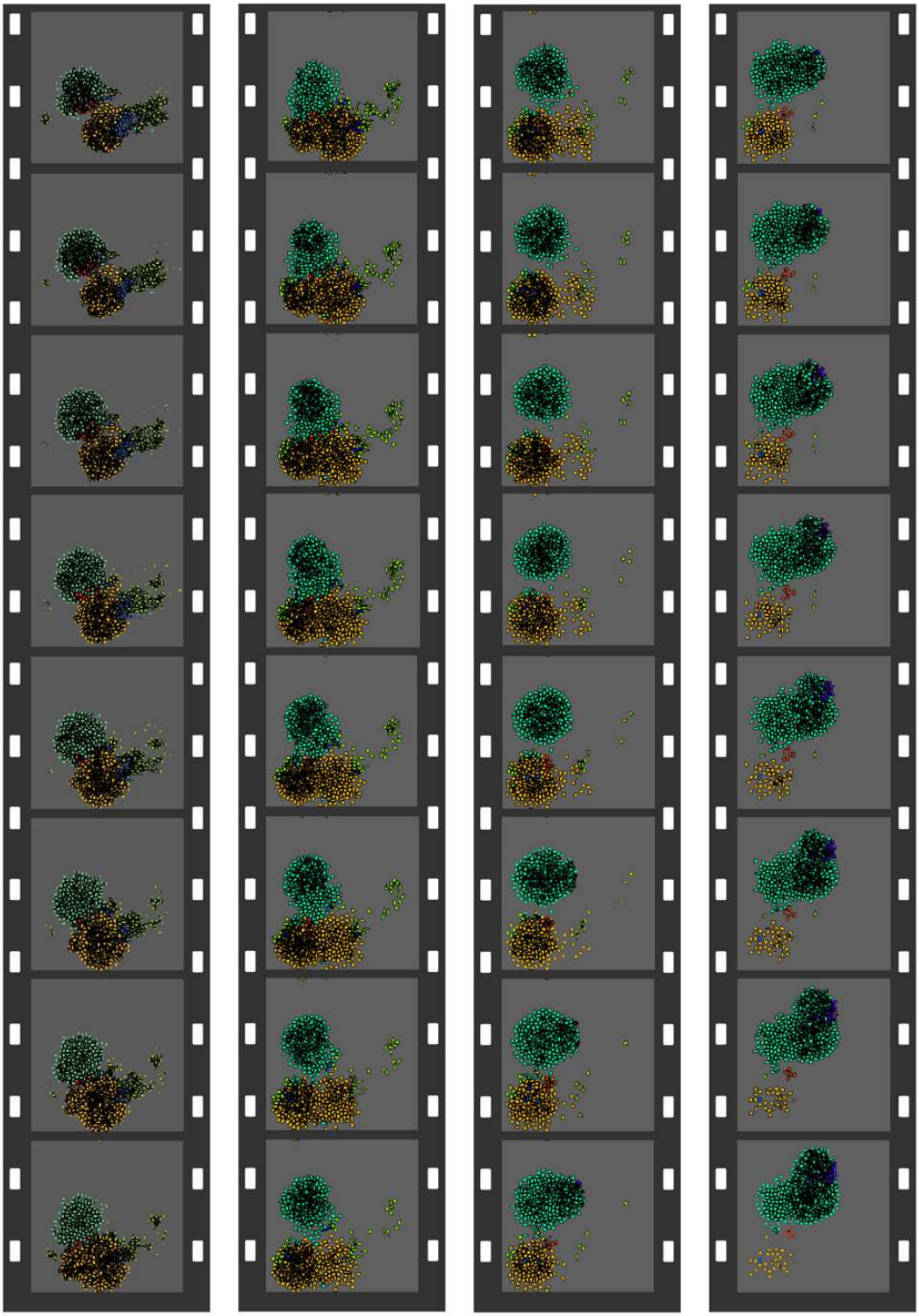
Figure 51: Decoding of the regulatory network from the genome. Genes are located via a marker or TATAbox. Coding and regulatory regions corresponding to this gene are taken immediately after and before this marker, respectively. Monomers function is decided depending on the secondary structure and the probability of dimerization according to the energy of the cofolded structure. Binding probabilities are computed from monomers or dimers according to the cofolded structure with the regulatory regions.

and horizontal gene transfer, turns `Cellos` into a tool for generating test data for phylogenetic reconstruction methods. Comparing the simulated evolutionary scenario with the reconstructed one will allow to evaluate the performance of such methods.

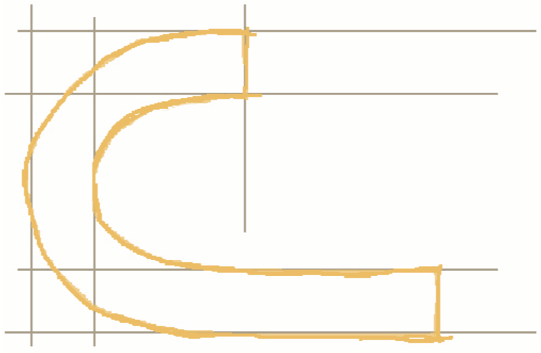
The environmental dynamics can also be improved by switching to an artificial chemistry like the Toy Chemistry Model (Benkö *et al.*, 2003). This forces for an additional decoding layer in the internal structure of the cells, which links our representation of the nutrient importers to organic molecules in the environment. Improvements of the `CelloS` model along these lines are under way.

7 Apendix: CelloS Movies





8 Curriculum



Camille Stephan-Otto Attolini 2005

Curriculum Vitae

address:
Waehringer str. 17.
A-1090
Vienna, Austria

General Information

Place and date of birth
October 18th,1978
México City

Nationality
Mexican

Contact
yocamille@tbi.univie.ac.at
www.tbi.univie.ac.at/~camille

Language skills
Spanish: mother tongue
English: fluent both speaking and writing
French: fluent speaking and some writing
Italian: fluent speaking and some writing
German: well spoken

Computer skills
C, Perl, html, Flash, Photoshop

Academics

Studies

2002-2005	Phd. Program in Biomathematics in the TBI, Institute for Theoretical Chemistry. University of Vienna, Austria. 2002-2005
1997-2001	MSc in Mathematics. Faculty of Sciences, National Autonomous University of Mexico (UNAM). Average Grade: 9.9/10.0

Academics

Congresses and Workshops

2005	Sep	Talk. European Conference on Artificial Life. Canterbury, England.
	Mar	Individual talk. KLI, Altenberg, Austria.
	Feb	Talk. Winter Seminar of the TBI, Bled, Slovenia
		Individual talk. Science Faculty. UNAM, México City, México
2004	Nov-Dec	STSM (Short Term Scientific Mission) from the Cost Action D27, EU. Work with Prof. Peter Stadler. ITZBI, University of Leipzig, Germany
		Talk. Fall Seminar of the TBI, Chribska, Check Republic
	Oct	Attended the COST Action D27 Workshop invited as Young Scientist in Herklion, Crete
	June	Poster. Congress MATH/CHEM/COMP'04. Dubrovnik, Croacia

Academics

2004	Mar	Individual talk. Science Faculty. UNAM, México City, México
	Feb	Talk. Winter Seminar of the TBI, Bled, Slovenia
2003	Nov-Dec	Research visit to Prof. Peter Stadler, ITZBI, University of Leipzig, Germany
	Nov	Oberwolfach Seminar: Some Mathematical Challenges from Life Sciences. Imparted by Prof. Peter Schuster
	Oct	Talk. Fall Seminar of the TBI, Chribska, Check Republic
	June	Poster. Congress MATH/CHEM/COMP'03. Dubrovnik, Croacia
	Feb	Talk. Winter Seminar of the TBI, Bled, Slovenia
2000	Sep	Biomathematics Fall School, CIMAT, Guanajuato, México
1999	Sep	Congress of the Mexican Mathematics Society, Guadalajara, México
1998	Oct	School of Geometric Algebra, CIMAT, Guanajuato, México

Publications

Neutral Networks of Interacting RNA Secondary Structures. Stephan-Otto Attolini, C. and Stadler, P. Submitted to Advances in Complex Systems, April 2005

Cellos: A Multi-level Approach to Evolutionary Dynamics. Stephan-Otto Attolini, C., Flamm, C. and Stadler, P. Published in Proc. of ECAL05, March 2005

Evolving Towards the Hypercycle: A Spatial Model of Molecular Evolution. Stephan-Otto Attolini, C. and Stadler, P. Submitted to Physica D, December 2004

Awards & Scholarships

Awards

Honorific Mention for the thesis: "*Caos en redes neuronales*" (Chaos in Neural Networks), Science Faculty, UNAM. 2002

Award for10/10 average during the degree in Mathematics in the period 1999-2000. UNAM.

Scholarships

Scholarship to attend the CSSS'05 from the Santa Fe Institute, Beijing, China. July-August 2005.

Scholarship. PhD studies in the University of Vienna. CONACyT (National Board for Science and Technology, México). 2002-in progress

Telmex Scholarship due to academic excellence during the MSc degree. 1997-2001

Scholarship S.E.P. (National Board for Education) for primary, Secondary and high school due to academic excellence. 1981-1996

Work

Research visit to Los Alamos National Laboratory. Participation in the project for Protocell Assembly (Pas) in EES-6 under the supervision of Steen Rasmussen. New Mexico, USA. August-October 2005.

Cooperation in Project No. P-14898-MAT with research on RNA molecules' interactions financed by Fonds zur Förderung der Wissenschaftlichen Forschung. February-July 2004

Revision of Mathematics textbooks for use in secondary schools, SEP, Mexico, 2001-2003

References

Altmeyer S, McCaskill JS, 2001. Error threshold for spatially resolved evolution in the quasispecies model. Phys Rev Lett 86:5819–5822.

Ambros V, 2000. Control of developmental timing in *Caenorhabditis elegans*. Current Opinions in Genetics and Development 10:428–433.

Ancel LW, Fontana W, 2000. Plasticity, evolvability, and modularity in RNA. URL `citeseer.ist.psu.edu/article/ancel00plasticity.html`.

Andrade MA, Garcia-Tejedor AJ, Montero F, 1991. Study of an error-prone hypercycle formed from two kinetically distinguishable species. Biophys Chem 40:43–57.

Bak P, 1996. How nature works: the science of self-organized criticality. New York: Springer-Verlag.

Ballare C, Scopel A, Sanchez R, 1997. Foraging for light: photosensory ecology and agricultural implications. Plant Cell and Environment 20:820–825.

Banzhaf W, 2003. On the dynamics of an artificial regulatory network. In: Banzhaf W, Christaller T, Dittrich P, Kim JT, Ziegler J, editors, Advances in Artificial Life, vol. 2801 of *LNCS*, (pp. 217–227). Heidelberg, Germany: Springer-Verlag. Proc. ECAL03.

Barnett L, 1997. Tangled Webs: Evolutionary Dynamics on Fitness Landscapes with Neutrality. Master's thesis, Brighton.

Bartel DP, Unrau PJ, 1999. Constructing an rna world. Trends Cell Biol 9:M9–M13.

Benkö G, Flamm C, Stadler PF, 2003. A graph-based toy model of chemistry. J Chem Inf Comput Sci 43:1085–1093.

Bentley PJ, 1996. Evolutionary Design by Computers. Morgan Kaufman Pub.

Bentley PJ, Kumar S, 1999. Three ways to grow designs: A comparison of embryogenies for an evolutionary design problem. Genetic and Evolutionary Computation Conference (pp. 35–43).

Boerlijst MC, Hogeweg P, 1991. Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. Physica D 48:17–28.

Bresch C, Niesert U, Harnasch D, 1980. Hypercycles, parasites, and packages. J Theor Biol 85:399–405.

Brockman J, 1995. The third culture: beyond the scientific revolution. New York: Touchstone.

Cherry J, Adler F, 2000. How to make a biological switch. J Theor Biol 203:117–133.

Croft LJ, Lechner MJ, Gagen MJ, Mattick JS, 2003. Is prokaryotic complexity limited by accelerated growth in regulatory overhead? Genome Biology 5:P2.

Cronhjort MB, Blomberg C, 1994. Hypercycles versus parasites in a two dimensional partial differential equations model. J Theor Biol 169:31–49.

Crow JF, Kimura M, 1970. An Introduction to Population Genetics Theory. New York: Harper and Row.

Davidson EH, 2001. Genomic Regulatory Systems: Development and Evolution. San Diego, USA: Academic.

Davidson EH, McClay DR, Hood L, 2003. Regulatory gene networks and the properties of the developmental process. Proc Natl Acad Soc 100:1475–1480.

Deckard A, Sauro HM, 2004. Preliminary studies on the in silico evolution of biochemical networks. ChemBioChem 5:1423–1431.

Dimitrov RA, Zuker M, 2004. Prediction of hybridization and melting for doublest randed nucleic acids. Biophys J 87:215–226.

Doudna JA, Cech TR, 2002. The chemical repertoire of natural ribozymes. Nature 418:222–228.

Ebner M, Langguth P, Albert J, Shackleton M, Shipman R, 2001a. On neutral networks and evolvability. In: Proceedings of the 2001 Congress on Evolutionary Computation CEC2001, (pp. 1–8). IEEE Press.

Ebner M, Shackleton M, Shipman R, 2001b. How neutral networks influence evolvability. Complex 7:19–33.

Eggenberg P, 1997. Evolving morphologies of simulated 3D organisms based on differential gene expression. In: Husbands P, Harvey I, editors, Proc. ECAL97, (pp. 205–213). The MIT Press/Bradford Books.

Eigen M, 1971. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523.

Eigen M, Schuster P, 1979. The Hypercycle. New York, Berlin: Springer-Verlag.

Flamm C, Fontana W, Hofacker I, Schuster P, 2000. RNA folding kinetics at elementary step resolution. RNA 6:325–338.

Fontana W, 2002. Modelling 'evo-devo'with RNA. BioEssays 24:1164–1177.

Fontana W, Schuster P, 1998. Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. J Theor Biol 194:491–515.

Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P, 1993. RNA folding landscapes and combinatory landscapes. Phys Rev E 47:2083–2099.

Forst CV, 2000. Molecular evolution of catalysis. J Theor Biol 205:409–431.

Forst CV, Reidys C, Weber J, 1995a. Evolutionary dynamics and optimization: Neutral networks as model-landscapes for rna secondary-structure folding-landscapes. (pp. 128–147). Spain.

Forst CV, Reidys CM, Weber J, 1995b. Evolutionary dynamics and optimization: Neutral Networks as model-landscape for RNA secondary-structure folding-landscapes. In: Morán F, Moreno A, Merelo J, Chacón P, editors, Advances in Artificial Life, vol. 929 of *LNAI*, (pp. 128–147). ECAL95, Springer.

François P, Hakim V, 2004. Design of genetic networks with specified functions by evolution *in silico*. Proc Natl Acad Sci USA 101:580–585.

Fulton C, Walsh C, 1980. Cell differentiation and flagellar elongation in *Naegleria gruberi*. J Cell Biol 85:346–360.

Geard N, Wiles J, 2003. Structure and dynamics of a gene network model. In: Sarker R, Reynolds R, Abbass H, Tan KC, McKay B, Essam D, Gedeon T, editors, Proc. CEC2003, (pp. 199–206). IEEE Press.

Gesteland RF, Atkins JF, editors, 1993. The RNA World. Plainview, NY: Cold Spring Harbor Laboratory Press.

Gilbert W, 1986. The RNA world. Nature 319:618.

Goldbeter A, 2002. Computational approaches to cellular rhythms. Nature 420:238–245.

Goodman N, 1955. Fact, fiction, and forecast. Cambridge, MA: Harvard University Press.

Gould SJ, 2002. The Structure of Evolutionary Theory. Boston: Harvard University Press.

Gruener W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P, 1996. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. Monath Chem 127:355–374.

Happel R, Stadler PF, 1998. The evolution of diversity in replicator networks. J Theor Biol 195:329–338.

Höbartner C, Micura R, 2003. Bistable secondary structures of small mas and their structural probing by comparative imino proton nmr spectroscopy. J Mol Biol 325:421–431.

Hocking JG, Young GS, 1988. Topology. New York: Dover Publications.

Hofacker IL, 2003. Vienna RNA secondary structure server. Nucl Acids Res 31:3429–3431.

Hofacker IL, Fekete M, Stadler PF, 2002. Secondary structure prediction for aligned RNA sequences. J Mol Biol 319:1059–1066.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P, 1994. Fast folding and comparison of RNA secondary structures. Monatsh Chemie 125:167–188.

Hogeweg P, Takeuchi N, 2003. Multilevel selection in models of prebiotic evolution: Compartments and spatial self-organization. Origins of Life and Evolution of the Biosphere 33:375–403.

Huynen MA, 1996. Exploring phenotype space through neutral evolution. J Mol Evol 43:165–169.

Huynen MA, Stadler PF, Fontana W, 1996. Smoothness within ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401.

Illangasekare M, Yarus M, 1999. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. RNA 5:1482–1489.

Jacob F, Monod J, 1961. On the regulation of gene activity. Cold Spring Harbor Symp Quant Biol 26:193–211.

Jadhav VR, Yarus M, 2002. Coenzymes as coribozymes. Biochimie 84:877–888.

James KD, Ellington AD, 1999. The fidelity of template-directed oligonucleotide ligation and the inevitability of polymerase function. Orig Life Evol Biosph 29:375–390.

Jeffares DC, Poole AM, Penny D, 1998. Relics from the RNA world. J Mol Evol 46:18–36.

Johnston WK, Unrau PJ, Lawrence MJ, Glasner ME, Bartel DP, 2001. RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. Science 292:1319–1325.

Joyce GF, 2002. The antiquity of RNA-based evolution. Nature 418:214–221.

Kauffman SA, 1993. The Origin of Order. New York, Oxford: Oxford University Press.

Keefe AD, Szostak JW, 2001. Functional proteins from a random-sequence library. Nature 410:715–718.

Kenneth SO, Risto M, 2002. Efficient Reinforcement Learning through Evolving Neural Network Topologies. Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2002.

Kimura M, 1955. Solution of a process of random genetic drift with a continuous model. Proc Natl Acad Sci USA 41:144–150.

Kimura M, 1983. The Neutral Theory of Molecular Evolution. Cambridge, UK: Cambridge University Press.

King JL, Jukes TH, 1969. Non-darwinian evolution: Random fixation of selectively netrual variants. Science 164:788–798.

Kingman JFC, 1978. A simple model for the balance between selection and mutation. J Applied Probability 15:1–12.

Klug S, Famulok M, 1994. All you wanted to know about SELEX. Mol Biol Reports 20:97–107.

Koza JR, 1994. Genetic programming: On the programming of computers by means of natural selection. Statistics and Computing 4.

Lau NC, Lim LP, Weinstein EG, Bartel DP, 2001. An abundant class of tiny rnas with probable regulatory roles in c*aenorhabditis elegans*. Science 294:858–862.

Lee N, Bessho Y, Wei K, Szostak JW, Suga H, 2000. Ribozyme-catalyzed tRNA aminoacylation. Nat Struct Biol 7:28–33.

Lewontin RC, 1974. The Genetic Basis of Evolutionary Change. New York: Columbia University Press.

Lloyd D, 1984. Variation strategies of plants in heterogeneous environments. BIOL J LINN SOC 21:357–385.

Marée AF, Hogeweg P, 2002. Modelling *Dictyostelium discoideum* Morphogenesis: the Culmination. Bull Math Biol 64:327–353.

Mathews D, Sabina J, Zuker M, Turner H, 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J Mol Biol 288:911–940.

Mattick JS, 1994. Introns: evolution and function. Current Opinions in Genetics and Development 4:823–831.

Mattick JS, 2001. Non-coding rnas: the architects of eukatyotic complexity. EMBO Reports 2:986–991.

Mattick JS, 2003. Challenging the dogma: the hidden layer of non-protein-coding rnas in complex organisms. BioEssays 25:930–939.

Mattick JS, 2005. The functional genomics of noncoding rna. Science 309:1527–1528.

Maturana H, Varela F, 1980. Autopoiesis and Cognition: The realization of the living. Boston: D. Reidel.

May RM, Nowak MA, 1994. Superinfection, metapopulation dynamics, and the evolution of diversity. J Theor Biol 170:95–114.

Maynard Smith J, 1979. Hypercycles and the origin of life. Nature 280:445–446.

McGinness KE, Joyce GF, 2003. In search of an RNA replicase ribozyme. Chem Biol 10:5–14.

Merks RMH, Glazier JA, 2005. A cell-centered approach to developmental biology. Physica A In press.

Moore PB, Steitz TA, 2002. The involvement of RNA in ribosome function. Nature 418:229–235.

Nuño JC, Tarazona P, 1994. Lifetimes of small catalytic networks. Bull Math Biol 56:875–898.

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al, 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cdnas. Nature 420:563–573.

Orgel LE, 1998. The origin of life - a review of facts and speculations. Trends Biochem Sci 23:491–495.

Paul N, Joyce GF, 2003. A self-replicating ligase ribozyme. Proc Natl Acad Sci USA 99:12733–12740.

Poole A, Jeffares D, Penny D, 1999. Early evolution: prokaryotes, the new kids on the block. Bioessays 21:880–889.

Rasmussen S, Chen L, Deamer D, Krakauer D, Packard N, Stadler P, Bedau M, 2004. Transitions from nonliving to living matter. Science 303:963–965.

Rasmussen S, Chen L, Nilsson N, Abe S, 2003. Bridging nonliving and living matter. Artificial Life 9 9:269–316.

Rechenberg I, 1994. Evolutionsstrategie'94. Fromman-Holzboog.

Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R, 2004. Fast and effective prediction of microRNA/target duplexes. RNA 10:1507–1517.

Reidys C, Stadler PF, Schuster P, 1997. Generic properties of combinatory maps: Neutral networks of RNA secondary structures. Bull Math Biol 59:339–397.

Reil T, 1999. Dynamics of gene expression in an artificial genome – implications for biological and artificial ontogeny. In: Floreano D, Nicoud JD, Mondada F, editors, Proc. ECAL99, vol. 1674 of *Lecture Notes in Computer Science*, (pp. 457–466). Berlin: Springer-Verlag.

Reil T, 2000. Models of gene regulation – a review. In: Maley C, Boudreau E, editors, Proc. ECAL00, (pp. 107–113). MIT Press.

SantaLucia Jr. J, Turner D, 1997. Measuring the thermodynamics of RNA secondary structure formation. Biopolymers 44:309–319.

Scheiner SM, 1993. Genetics and evolution of phenotypic plasticity. Annu Rev Ecol Syst 24:35–38.

Scheuring I, Czaran T, Szabo P, Karyoli G, Toroczkai Z, 2003. Spatial models of prebiotic evolution: Soup before pizza? Origins of life 33:329–355.

Schlichting C, Smith H, 2002. Phenotypic plasticity: linking molecular mechanisms with evolutionary outcomes. Evol Ecology 16:189–211.

Schlichting CD, Pigliucci M, 1998. Phenotypic evolution: A reaction norm perspective. Sunderland, Massachusetts: Sinauer Associates, Inc. 387 pp.

Schultes EA, Bartel DP, 2000. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. Science 289:448–452.

Schuster, 1997. Landscapes and molecular evolution. Physica D 107:351–365.

Schuster P, 1986. Selforganization, chap. Mechanisms of Molecular Evolution, (pp. 57–91). Adenine Press.

Schuster P, 1998. Evolution at molecular resolution. In: Matsson L, editor, Nonlinear Cooperative Phenomena in Biological Systems, (pp. 86–112). World Scientific.

Schuster P, 2001. Evolution in Silico and in Vitro: The rna model. Biol Chem 382:1301–1314.

Schuster P, 2002. Biological Evolution and Statistical Physics, chap. A testable genotype-phenotype map: modeling evolution of RNA molecules, (pp. 55–81). Springer Berlin.

Schuster P, Fontana W, Stadler PF, Hofacker IL, 1994. From sequences to shapes and back: A case study in RNA secondary structures. Proc Roy Soc Lond B225:279–284.

Schuster P, Sigmund K, 1983. Replicator dynamics. J Theor Biol 100:533–538.

Sharp PA, 2001. Rna interference-2001. Genes Dev 15:485–490.

Sievers D, von Kiedrowski G, 1994. Self-replication of complementary nucleotide-based oligomers. Nature 369:221–224.

Simpson GG, 1944. Tempo and Mode in Evolution. New York: Columbia University Press.

Snyder M, Gerstein M, 2003. Defining genes in the genomics era. Science 300:258–260.

Stadler BMR, 2002a. Diffusion of a population of interacting replicators in sequence space. Adv Complex Systems 5:457–461.

Stadler BMR, Stadler PF, Schuster P, 2000. Dynamics of autocatalytic replicator networks based on higher order ligation reactions. Bull Math Biol 62:1061–1086.

Stadler BMR, Stadler PF, Wagner G, Fontana W, 2001a. The topology of the possible: Formal spaces underlying patterns of evolutionary change. J Theor Biol 213:241–274.

Stadler BMR, Stadler PF, Wills PR, 2001b. Evolution in systems of ligation-based replicators. Z Phys Chem 21-33:216.

Stadler P, 2002b. Biological Evolution and Statistical Physics, chap. Fitness Landscapes, (pp. 187–207). Springer Berlin.

Stadler PF, 1999. Fitness landscapes arising from the sequence-structure maps of biopolymers. J Mol Struct THEOCHEM 463:7–19. Santa Fe Institute Preprint 97-11-082.

Stadler PF, 2002c. The genotype-phenotype map. URL `www.tbi.univie.ac.at/papers/Abstracts/02-pfs-002abs.html`.

Stadler PF, Happel R, 1993. The probability for permanence. Math Biosc 113:25–50.

Stadler PF, Schuster P, 1996. Permanence of sparse autocatalytic networks. Math Biosc 131:111–134.

Stanley KO, Miikkulainen R, 2003. A taxonomy for artificial embryogeny. Artificial Life 9:93–130.

Stephan-Otto Attolini C, Stadler P, 2004. Evolving towards the hypercycle: A spatial model of molecular evolution. Physica D Submitted.

Stephan-Otto Attolini C, Stadler P, 2005. Neutral networks of interacting rna secondary structures. Adv Complex Syst in press.

Stephan-Otto Attolini C, Stadler P, Flamm C, 2005. Cellos: A multi-level approach to evolutionary dynamics. Proceedings of ECAL05 (pp. 500–509).

Streissler C, 1992. Autocatalytic Networks Under Diffusion. Ph.D. thesis, University of Vienna.

Szabó P, Scheuring I, Czaran T, Szathmáry E, 2002. In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. Nature 420:278–279.

Tarazona P, 1992. Error threshold for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models. Phys Rev A 45:6038–6050.

Tereshko V, 1999. Selection and coexistence by reaction-diffusion dynamics in fitness landscapes. Phys Let A 260:522–527.

Toffoli T, Margolus N, 1987. Cellular Automata Machines: A New Environment for Modeling. MIT Press.

Unrau PJ, Bartel DP, 1998. RNA-catalysed nucleotide synthesis. Nature 395:260–263.

van Nimwegen E, Crutchfield JP, Huynen MA, 1999. Neutral evolution of mutational robustness. Proc Natl Acad Sci USA 96:9716–9720.

Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS, 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? TRENDS in Genetics 20:44–50.

von Kiedrowski G, 1986. A self-replicating hexadeoxynucleotide. Angew Chem Int Ed Engl 25:932–935.

von Kiedrowski G, Wlotzka B, Helbing J, 1989. Sequence dependence of template-directed syntheses of hexadeoxynucleotide derivatives with 3'-5' pyrophosphate linkage. J Angew Chem Int Edn engl 28:1235–1237.

Waddington C, 1942. Canaliztion of development and the evolution of evolvability. Nature 3811:564–565.

Waddington C, 1957. The Strategy of the Genes. Allen & Unwin, London.

Waddington CH, 1953. Genetic assimilation of an acquired character. Evolution 4:277–282.

Wagner GP, Altenberg L, 1996. Complex adaptations and the evolution of evolvability. Evolution (p. In press). URL `citeseer.ist.psu.edu/148444.html`.

Widder S, 2003. Self-Regulating Gene Switches and Molecular Evolution. Ph.D. thesis, University of Vienna.

Wills PR, 2001. Autocatalysis, information and coding. BioSystems 60:49–57.

Witwer C, Rauscher S, Hofacker IL, Stadler P, 2001. Conserved RNA secondary structures in picornaviridae genomes. Nucl Acids Res 29:5079–5089.

Wlotzka B, McCaskill JS, 1997. A molecular predator and its prey: Coupled isothermal amplification of nucleic acids. Chemistry Biology 4:25–33.

Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF, 2004. Exact folding dynamics of rna secondary structures. JPhysA MathGen 37:4731–4741.

Wright MC, Joyce GF, 1997. Continuous in vitro evolution of catalytic function. Science 276:614–617.

Wright S, 1931. Evolution in mendelian populations. Genetics 16:97–159.

Wuchty S, Fontana W, Hofacker IL, Schuster P, 1999. Complete suboptimal folding of rnaand the stability of secondary structures. Biopolymers 49:145–165.

Zintzaras E, Santos M, Szathmáry E, 2002. "Living" under the challenge of information decay: The stochastic corrector model vs. hypercycles. J Theor Biol 217:167–181.

Zuker M, Sankoff D, 1984. RNA secondary structures and their prediction. Bull Math Biol 46:591–621.