Evolving the structure of genomes, networks and ecosystems

OMSLAG Anton Crombach
ISBN 978 90 393 5029 4

Evolving the structure of genomes, networks and ecosystems

Evolutionaire structurering van genomen, netwerken en ecosystemen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus prof. dr. J.C. Stoof ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 22 april 2009 des middags te 2.30 uur

door

Antonius Benedictus Maria Crombach

geboren op 18 juni 1980
te Veldhoven

Promotor: Prof. dr. P. Hogeweg

The studies described in this thesis were financially supported by the Netherlands Organisation for Scientific Research (NWO) through grant number 635.100.001 of the Computational Life Sciences program.

Contents

1 Introduction

Arguably the most important concept in biology is evolution. In 1859 it was Charles Darwin who laid down the theory of evolution by means of natural selection in his book *On the Origin of Species* (Darwin, 1859). Today, evolution is defined as the change in biological systems arising from random mutation and natural selection. Mutations generate variation in a population, while natural selection weeds out the individuals that are ill-adapted to their environment.

We use the method of computer simulations to study how evolution integrates information on changing environments into living systems. In this first chapter we give a short overview of the biological and evolutionary concepts central to the work described in this thesis. Next, an outline of our studies is given.

1.1 Genes and genomes

Classically the gene was defined as the unit of heredity. With the discovery of DNA as the bearer of heritable traits, it was viewed as a stretch of DNA nucleotides that was transcribed to messenger RNA (mRNA) and subsequently translated into protein. In eukaryotes the gene was found to be composed of introns and exons, where introns were spliced out and the resulting mRNA contained the actual protein code. Nowadays, the definition has been refined to allow also for genes that do not code for proteins at all, but whose RNA transcripts perform a certain function (named noncoding genes or RNA coding genes). To complicate matters further, recent insights have also shown genes with both coding and noncoding functions (Mattick, 2003).

The collection of protein coding genes forms an important part of the genome. In bacteria roughly 90% of the genome consists of such coding genes. As we move from bacteria via (relatively) simple unicellular eukaryotes to more

complex multicellular organisms, the amount of protein coding genes plateaus. However the genomes continue to grow: single genes have been split by multiple introns, introns have increased in size and copies of transposable elements have accumulated. These DNA sequences have been called junk DNA due to their apparent lack of function. This points to a long-standing question: to what extent is the growth and architecture of many eukaryotic genomes a neutral process (Lynch & Conery, 2003; Lynch, 2006; Knibbe *et al.*, 2007; Koonin, 2009) and, if partly under selection, what function do these junk DNA elements have?

It has been theorized that junk DNA is necessary as a skeletal frame for the nucleus (Cavalier-Smith, 2005). A certain cell volume appears to require a minimal nucleus size, which in turn relates to genome size. Such physical properties and constraints certainly play a role, yet are probably only part of the story. For instance, the last years it has been found that a large part of this junk DNA is transcribed to RNA (ENCODE Consortium, 2007). Among these transcribed regions are repeat elements, transposable elements and RNA coding genes. This could indicate a certain functionality, and it is now known that many of these junk elements are recruited in various cellular tasks. Their functions range from transcriptional and translational control (Rubin & Spradling, 1982; Smalheiser & Torvik, 2005; Häsler & Strub, 2006) to guidance of alternate splicing (Sorek *et al.*, 2002). Moreover, they may cause mutational effects, for instance the frequent and reversible addition and deletion of microsatellites (Kashi & King, 2006).

In many eukaryotes, genomic variation between individuals points at a relatively neutral process of genome growth and shrinkage (Han *et al.*, 2008), yet we conclude that the above findings suggest the junk may not be as useless as it was once deemed to be.

1.1.1 Genome organization

In bacteria, the genome is arranged in so-called operons (Jacob *et al.*, 1960). Often functionally related genes, for instance occurring in the same metabolic pathway, are located in sequence and, as they share a regulatory upstream region, are transcribed together. This clear structuring of the circular bacterial genome is usually not found in eukaryotes, where co-expressed genes not necessarily co-localize on the genome. A notable exception is the operon-like structure found in nematodes (Qian & Zhang, 2008). Nevertheless gene order and organization is not random in eukaryotes either. A classical example is the clustering of Homeobox genes in vertebrates. An example from the model organism baker's yeast (*S. cerevisiae*) is that genes that function in the same pathway or that are recruited to the same protein complex, are significantly closer together on the genome than expected by chance (Teichmann & Veitia, 2004).

In eukaryotes, due to histones and a large repertoire of modifications of such proteins, epigenetic regulation plays an important role. Epigenetic phenomena are thought to determine which groups of genes are made accessible for transcription (Meaburn & Misteli, 2007), how chromatin folds and loops (van Driel *et al.*, 2003; Batada *et al.*, 2007), and how chromosomes occupy specific areas of

the nucleus (Cremer & Cremer, 2006a,b) (but also see Rosa & Everaers (2008)). These extra layers of complexity allow eukaryotic genomes to compartmentalize the nucleus and to have subtle regulatory interactions between different chromosomes (Meaburn & Misteli, 2007). In contrast to bacterial operons, the genome structure of eukaryotes is distributed over several 'hierarchical' levels and it is unclear how much can be distilled from the location of genes and other elements on individual chromosomes alone.

From comparative studies of genome organization the picture emerges that eukaryotes evolve by macro and micro rearrangements of chromosomes (Britten *et al.*, 2003; Dujon *et al.*, 2004; Lynch & Wagner, 2009). The examination of various mammalian genomes shows a mosaic of 'reused' chromosome segments (Murphy *et al.*, 2005). Large segments have been rearranged, telomeres have fused, (neo)centromeres have been broken and remodeled as telomeres, and within segments many small inversions, duplications and deletions are found as well. In addition, human cancer cells show recurrent chromosomal breakpoints that to certain extent overlap the evolutionary fragile breakpoints (Darai-Ramqvist *et al.*, 2008). Again the question arises if mostly neutral mutational processes are shaping the eukaryotic genome and the observed organization is simply a signature of the specific mutations. On the other hand, if not, to what degree has a specific genome organization been selected for (Poyatos & Hurst, 2006; Knibbe *et al.*, 2008; Koonin, 2009)?

Large chromosome rearrangements have been observed on a much shorter time scale as well. A well-known example is the experimental evolution of yeast in a low-glucose environment with eight replicate experiments (Dunham *et al.*, 2002). After $\sim$300 generations some strains showed small mutations (Brown *et al.*, 1998), while others had large duplications and deletions of chromosome segments, called gross chromosomal rearrangements (GCRs). In all cases these mutations lead to similar changes in gene expression and there was a beneficial effect on the fitness of these strains compared to the parental strain (Ferea *et al.*, 1999). Moreover, it was observed that in this short evolutionary episode several evolved strains had undergone GCRs at the same genomic locations, and at these fragile breakpoints repeat elements derived from transposons were found. Also, repeat elements have been indicated at or near other documented, advantageous GCRs (Schacherer *et al.*, 2005).

Has the genome been shaped such that these beneficial mutations were likely? It is tantalizing to hypothesize a specific organization for rapid evolutionary adaptations. In this thesis, we provide a proof-of-concept of such evolution of genome structuring in which transposable elements and repeat elements play a pivotal role. On the other hand, there may be other processes at work in the above mentioned yeast populations. Perhaps GCRs give only a minor fitness improvement or there are few advantageous mutations and selection for these produces the observed patterns.

In summary, despite the lack of an 'obvious' operon structure as we find in prokaryotes, also eukaryotic chromosomes certainly have a nonrandom gene order.

common name	scientific name	fraction of genome
Baker's yeast	*S. cerevisiae*	0.03
Fruit fly	*D. melanogaster*	0.15
Human	*H. sapiens*	0.45
Maize	*Z. mays*	0.60
Mouse	*M. musculus*	0.38
Thale cress	*A. thaliana*	0.14

Table 1.1 – The approximate fraction of transposable elements and other repeat sequences in the genome of several model organisms (Goffeau *et al.*, 1997; Arabidopsis Genome Initiative, 2000; International Human Genome Sequencing Consortium, 2001; Meyers *et al.*, 2001; Kaminker *et al.*, 2002; Hoskins *et al.*, 2002; Dennis, 2002). Different families of transposable elements dominate in different species. A few abundant class I families are Ty1-copia, Ty3-gypsy and LINE; class II families are Mariner/Pogo, P element and Helitron.

1.2 Transposable elements

In the 1940s Barbara McClintock discovered a special class of genomic elements in maize (*Zea mays*) (McClintock, 1950, 1953), for which she was awarded the Nobel Prize in 1983. She described a process that is now known as transposition. Certain stretches of DNA are capable of copying or moving their own sequence to other positions in the genome. As a result several types of mutations were observed in maize, such as insertions, deletions and translocations. These mobile elements are now known to pervade the eukaryotic kingdoms (see Table 1.1), while prokaryotes hold their own, similarly functioning, insertion sequence elements.

Two main categories of mobile genetic elements have been distinguished in eukaryotes. Firstly, class I elements are called retrotransposons. They use an RNA intermediate and the enzyme reverse transcriptase to insert copies of themselves across a genome. They come in two flavors: with and without long terminal repeats (LTRs). As an example, baker's yeast has several families of LTR-retrotransposons, and the human genome contains over a million copies of a non-LTR element called LINE (Long Interspersed Nuclear Element).

The life cycle of LTR retrotransposons resembles the cycle of retroviruses, excluding the release phase (see Figure 1.1). Several stages are strikingly similar, such as the fact that a dimer mRNA enters a virus-like particle, which suggests that these mobile elements ultimately share a common origin with retroviruses (Beauregard *et al.*, 2008). Non-LTR retrotransposons copy-and-paste in a qualitatively different manner. It is likely that they reverse transcribe RNA copies in the nucleus at the site of integration into the host DNA (Luan *et al.*, 1993; Deininger *et al.*, 2003).

Secondly, class II elements directly integrate a copy of their DNA sequence into the host genome. Hence they are simply known as transposons, and three subcategories are observed: cut-and-paste elements that encode transposase to

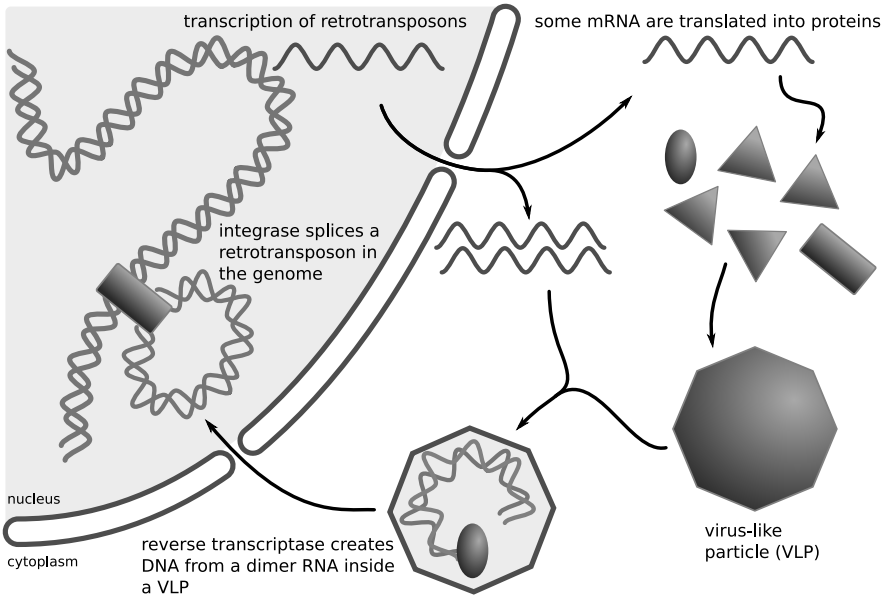

Figure 1.1 – Suggested life cycle of LTR retrotransposons (Sabot & Schulman, 2006). The LTR of a retrotransposon is a weak promoter (Forbes *et al.*, 2007), giving rise to mRNA transcripts that are exported to the cytoplasm. Exported mRNA is either translated into proteins such as GAG-like coating proteins, reverse transcriptase and integrase, or recruited as a dimer into virus-like particles (VLP) (Feng *et al.*, 2000). In the VLP the dimer-RNA is reverse transcribed to a double strand DNA sequence (Cheng & Menees, 2004), which is subsequently imported to the nucleus and integrated into the host genome by integrase.

move around, elements with a rolling-circle replication, f.i. Helitrons (Kapitonov & Jurka, 2001) and so-called Mavericks, that have multiple, virus-like proteins and probably encode a DNA polymerase to copy themselves (Pritham *et al.*, 2007). In mammals only few active class II transposons have been found. For instance, primates have a diverse set of transposon families ($\sim$ 40), yet none have been active in the last 40-50 My (Feschotte & Pritham, 2007).

The relative occurrence of class I and II elements in a genome depends on the species; each species has its own unique composition and even related species may differ substantially (Feschotte & Pritham, 2007).

1.2.1 Mutagenic agents

Even though McClintock already described the potentially beneficial effects of transposable elements (TEs) on regulation, they have long been, and still often are, regarded as parasites of the genome (Feschotte, 2008). By copying and moving through the host genome TEs disrupt genes, change regulatory regions and cause chromosome rearrangements. As such TEs have simply been viewed as

powerful mutagens and therefore a common explanation of their presence in genomes has been an equilibrium between transposition and selection (Rouzic & Capy, 2005).

However, the last decade their positive role has been rediscovered: it has become increasingly clear that TEs have been co-opted for various tasks (Kazazian, 2004). The hosts, it appears, have turned junk DNA into treasure. A few examples are: Wang *et al.* (2007) found that in primates an endogenous retrovirus deposited binding sites for the *p53* transcription factor throughout the genome, enriching *p53* with target genes involved in cell-cell adhesion. Alu elements are extremely abundant in primates, and have been recruited in transcriptional regulation, perform essential RNA editing tasks and alter proteins by incorporating themselves in exons (Häsler & Strub, 2006). The TEs HeT-A and TART have replaced the enzyme telomerase in order to maintain chromosome integrity in *Drosophila* (Pardue & DeBaryshe, 1999). Furthermore during the course of evolution various enzymes have been recruited from TEs: RAG1 is most likely derived from transposase and functions in the adaptive immune system where it interacts with other proteins to catalyze V(D)J recombination (Gellert, 2002). In fission yeast (*S. pombe*) transposase-derived CENP-B not only binds to centromeres, but also silences the transcription of retrotransposon family Tf2 by aggregating and packaging the various copies throughout the genome in so-called Tf-bodies (Cam *et al.*, 2008). This creates a rather direct effect of transposons on genome organization in the nucleus, complementing the previously described indirect, mutational interaction between TEs and GCRs.

Instead of parasitic, it might be better to describe the relation between transposon and host as mutualistic.

1.2.2 Regulation by the host

Despite the positive contributions of TEs to the evolution of genome organization and regulation, a too successful copying of TEs is negatively correlated with host fitness. Hence the host controls TE activity, for instance by transposon silencing. We focus on two major silencing mechanisms based on RNA interference (RNAi).

First of all, transposons are post-transcriptionally silenced (PTGS) by small interfering RNAs (siRNAs). These short 21-25 nucleotide RNAs are derived from double-stranded RNA (dsRNA) by the protein Dicer, and incorporated into a RNA-degrading protein complex (RISC) by binding the Argonaute protein. Next, RISC cuts mRNA with a complementary match to the siRNA. In this manner the host inhibits the translation of transposon RNA transcripts (Chung *et al.*, 2008; Golden *et al.*, 2008; Siomi *et al.*, 2008). It is still unknown how post-transcriptional RNAi is triggered to silence transposons. A main question is how dsRNA is formed in the context of TEs. In the standard RNAi pathway, RNA dependent RNA polymerase (RdRP) creates dsRNA from RNA transcripts. It has also been observed that class II TEs have inverted repeats that on read-through may lead the mRNA to fold into a hairpin, thus resulting in dsRNA (Sijen & Plasterk, 2003). Furthermore, bidirectionally transcribed regions, or simply anti-parallel

TE transcripts, allow for pairing of the complementary mRNAs, again resulting in dsRNA. For instance, the LINE1 element in humans is transcribed from both DNA strands, inducing siRNA production (Yang & Kazazian, 2006).

Second, epigenetic silencing, or transcriptional gene silencing (TGS), is used to control TEs. Both DNA and the accompanying histones may be modified to signal transcription repression. In plants and mammals, DNA methylation is such a signal, which can even be inherited for a few generations. In addition, in eukaryotes nucleosomes are often methylated ((di)methylation of histone 3 at lysine 9; H3K9me) to establish heterochromatin that is transcriptionally repressed. Both DNA and histone methylation are used to silence transposons (Lavrov & Kibanov, 2007); here we focus on the latter.

The paradox is that the chromatin modifications are guided mainly by small RNAs produced from the same TEs that are silenced. Briefly, the following is known: as a transposon is transcribed, its mRNA is captured by RdRP, which creates a dsRNA. Immediately this dsRNA is sliced Dicer. The resulting small RNAs are loaded onto an Argonaute that is part of the RNA-induced transcriptional silencing (RITS) complex. In turn the RITS complex engages with several other proteins, which leads to histone (di)methylation and recruitment of heterochromatization proteins such as SWI6/HP1 (Grewal & Jia, 2007; Slotkin & Martienssen, 2007). In other words, in order to silence TEs, they need to be transcribed (Grewal & Elgin, 2007). These findings come from research on repeat elements in *S. pombe* centromeric regions, which function as a model for epigenetic silencing. And there is additional evidence that the results obtained from these centromeres generalize to TEs (Girard & Hannon, 2008).

Recent research on *S. pombe* has resulted in some, partial, explanations: first of all, antisense transcription is associated with siRNA-mediated heterochromatin formation in *trans* (Iida *et al.*, 2008). Second, during the cell cycle histone modifications are altered, allowing for a build-up of siRNAs in S phase (Kloc & Martienssen, 2008). And it is likely that the trafficking of siRNAs from cytoplasm to nucleus is also involved (Guang *et al.*, 2008). Still there remain many open questions: how well does heterochromatin silence transcription, what are the essential components of the silencing mechanisms, do cytoplasmic siRNAs trigger heterochromatization in the nucleus, how effective is recruitment of SWI6/HP1 by RITS?

A deeper understanding of PTGS has been obtained through modeling the pathway of mRNA degradation by RNAi Groenenboom *et al.* (2005). We adopt their model and extend it with transposon dynamics and TGS. In this manner we investigate minimal requirements for transposon silencing on the nuclear and cytoplasmic level: which proteins are necessary, how do PTGS and TGS interact?

1.3 Gene regulation

Up to now we have viewed genes and genomes as entities that are subject to various mutational processes over many generations. Naturally, such adapta-

tions interact with the regulation of genes during the lifetime of an individual.

Organisms have evolved to deal with many of the changes in the outer environment that occur during their lifetime by sensing these and responding with the activation and/or inhibition of gene expression. In similar fashion signals from the cellular state (or organismal inner environment) are integrated, and lead to specific gene expression patterns. Furthermore, the development of multicellular organisms is based on many interactions between cells and the coordination of their gene expression.

The first evidence of genes being regulated in their protein production originates from work by Jacob & Monod (1961). They discovered that the bacterium *E. coli* regulates its production of lactose digesting proteins. Quickly it became apparent that they had found a general principle of gene regulation that holds across the three kingdoms of life. In the following years this was expanded to the notion of a network of genes that regulate each others protein production. It was found that a special type of proteins, named transcription factors, are responsible for adjusting transcription rates of genes.

As mentioned previously, the operons found in bacteria have a single upstream region of *cis*-regulatory elements, usually called binding sites, to which transcription factors bind. In this straightforward fashion they recruit protein-machinery, such as RNA polymerase, to transcribe the downstream genes (Figure 1.2A). Again, the case of eukaryotes is more complex. In general, binding sites are found in the direct upstream region of a gene. However, in addition distal enhancers and silencers have been found that can be located thousands of basepairs upstream, downstream or simply within the gene coding region (Figure 1.2B). This shows how genome architecture, like chromatin loops, interacts with the regulatory mechanisms of eukaryotes. Interestingly, RNAi has been implicated in the formation of specific euchromatin configurations, in addition to the previously discussed heterochromatin formation (Grimaud *et al.*, 2006).

Recently, with the advent of high-throughput techniques, the regulatory interactions of the transcription networks of both *E. coli* and *S. cerevisiae* have been largely unraveled (Lee *et al.*, 2002; Gama-Castro *et al.*, 2008). These show a rather sparse network in which most transcription factors regulate few others, while a small group of genes, named hubs, influences many target genes. This results in a power-law distribution of gene outdegree and thus a small-world, scale-free network. As a consequence of this topology, networks are robust against random 'failure', that is random deletion of genes. However, scale-free networks do suffer from so-called attack vulnerability: if by mutation a hub gene is deleted, the network is split into several small components. Furthermore, the transcription networks show clustering on many levels: from motifs (i.e. subgraphs of three to five genes) (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002) to a hierarchical layering of modules (Lagomarsino *et al.*, 2007).

Interestingly, simple mutational growth processes without selection for specific network architectures already produce many of the above named characteristics (Barabási & Oltvai, 2004; Teichmann & Babu, 2004; Kuo *et al.*, 2006). For instance, some network motifs appear overrepresented in regulatory networks

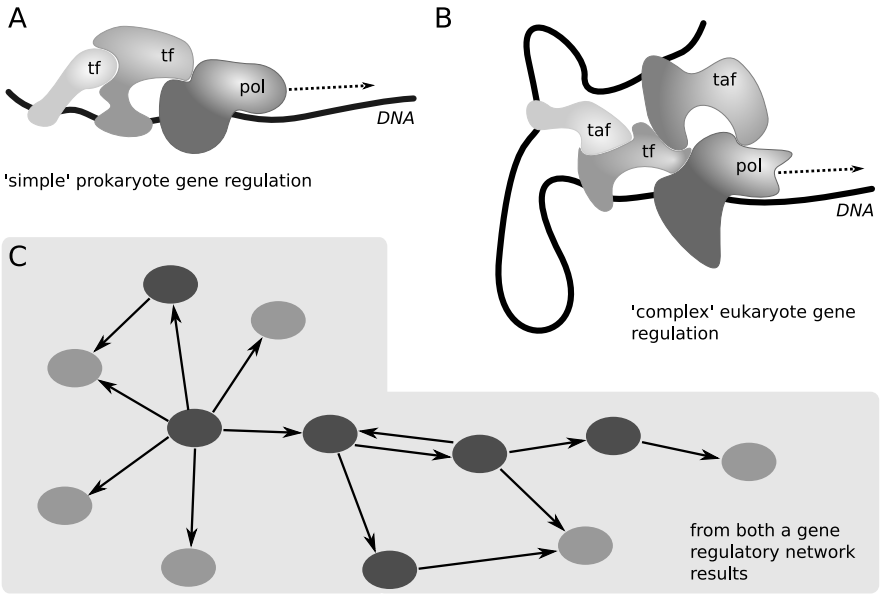

Figure 1.2 – Schematic view on gene regulation. Legend: *tf*, transcription factor; *taf*, trans-acting factor; *pol*, polymerase. A. In prokaryotes transcription factor proteins rather directly activate or repress transcriptional activity. B. In eukaryotes many more proteins are involved than in prokaryotes. Both local transcription factors and trans-acting factors that are in close proximity due to chromatin folding, interact with protein complexes and polymerase. C. In both cases the result is a complex network of genes regulating each other. Here dark-gray nodes symbolize genes that regulate other genes, and light-gray nodes are target genes.

(Milo *et al.*, 2002). These motifs have been ascribed an adaptive value as they allow for a more robust signal processing (Mangan & Alon, 2003), yet they can already be explained as a neutral signature of the mutational processes at work (Cordero & Hogeweg, 2006; Solé & Valverde, 2006, 2008). Thus, again, a pressing question is: how does it evolve? Which parts of the gene regulatory network structure are side-effects of mutational processes, and where does selection play a major role? In order to look at the latter – what are the capabilities of natural selection and how does it shape the network – we have investigated the evolution of gene networks under changing environmental conditions.

1.4 Concepts in evolution

So far we have introduced several topics on genome organization and gene regulation. We now present a selection of concepts from contemporary theory on evolution relevant to this thesis.

Closely connected to evolution is the notion of fitness. Darwinian fitness is

a measure of the ability of one individual or genotype to leave a number of offspring relative to other genotypes, with a genotype consisting of the heritable information of an individual. A visualization of all possible genotypes and their corresponding fitness results in a fitness landscape, where the altitude of the landscape corresponds to the fitness of a genotype (Wright, 1932). As individuals with high fitness have an increased probability of survival and reproduction, in a smooth landscape evolution would resemble the hill climbing of a population towards a peak. Note that we assume here that the environment, i.e. the landscape, does not change.

Fitness is of course defined as an immediate effect on reproduction. However, it has been shown that thus implemented fitness leads over many generations to many indirect fitness effects through an integration of encountered resources, pathogens, environments and so on (Savill *et al.*, 1997; Rauch *et al.*, 2002; Hogeweg, 2007).

1.4.1 Genotype-Phenotype mapping

The relation between genotype and fitness has as an important intermediate the phenotype. The phenotype of an organism describes such physical properties as its morphology (its size and shape), behavior (for instance movement, feeding and mating) and metabolic activities. Natural selection does not directly select for certain genotypes, instead it operates on the level of phenotypes.

Often a simple one-to-one mapping from genotype to phenotype is assumed. However, research on realistic mappings – first RNA folding (Huynen, 1993; Fontana *et al.*, 1993a,b) and recently also gene regulatory networks (GRNs) (Ciliberti *et al.*, 2007a; Aldana *et al.*, 2007; Munteanu & Solé, 2008) – has revealed a qualitatively different picture. In both paradigm systems there are many interdependent elements, leading to epistatic interactions and pleiotropy. In addition, very similar genotypes may code for rather different phenotypes, yet dissimilar genotypes can still result in the same phenotype. The result is a high-dimensional, non-linear, many-to-one relationship between genotypes and phenotypes.

A crucial insight gained from these models is the concept of neutral networks. It has been found that there are networks of equally-fit genotypes connected via single mutations percolating the entire genotype space (Schuster *et al.*, 1994). These genotypes result in the same phenotype, which implies that a phenotype may be reached from many starting points in genotype space. Furthermore, a population of individuals thus resides mostly on such a neutral network. This is rather different from the classical landscape metaphor we introduced previously. It may be better envisioned as a high-dimensional intertwining of neutral networks. Schematically speaking, an evolutionary process now is a random walk on a neutral network of phenotypes as new individuals are born and unfit ones are selected against. The 'hill climbing' is a shift of the population from one neutral network of genotypes to another that results in phenotypes with a higher fitness. Such a shift is rapid, after which the random walk continues (Fontana & Schuster, 1998; van Nimwegen & Crutchfield, 2000).

1.4.2 Evolvability and robustness

Evolvability, the ability to evolve, is a term with multiple definitions. In its most basic incarnation, evolvability is the idea that in a system with heritable genetic variation, mutations may be beneficial, neutral or deleterious and that natural selection is able to remove maladapted individuals. If this type of evolvability is not present, Darwinian evolution is not possible (Pigliucci, 2008).

A second definition is the capacity to discover beneficial, heritable adaptations. Here, not the current variation in the population is important, but the variation that may be generated by this population (Wagner & Altenberg, 1996). If this type of evolvability can evolve, evolution can shape the mutational neighborhood of individuals. At first sight, it contradicts the common knowledge that mutation is blind and natural selection can only select for the variation that is present in a population. Neither process can cater for future needs. However, adaptive evolution is a process of integrating information on the environment in a population. As a second order process also recurring changes in environment may be integrated into the population. In this manner evolution of evolvability may occur if past environments predict future ones well (Draghi & Wagner, 2008).

Considering the evolution of a genotype-phenotype map, evolvability may occur at the genotype or phenotype level. That is to say, we can discern two levels: mutations may be biased (so-called mutational priming (Hogeweg, 2005)), and the phenotypic effect of mutations may be biased. In this thesis we show a proof of principle for both cases of evolvability: we apply transposons as modifier genes that influence mutation rates and their locations on the genome, and we show that the evolution of gene regulatory networks may channel the phenotypic effect of mutations such that most mutations remain neutral, while few specific ones allow for adaptations.

Sometimes a third usage of the term evolvability is used to generalize the idea that individuals become well-adapted at adapting. Instead of evolvability being a question on reliable environmental changes, it is now defined as the ability of an individual to acquire novel functions through genetic change (Wagner, 2005). Importantly, such functions help the organism survive and reproduce. An example would be the recent novel enzymatic function of degrading nylon-related chemicals by bacteria in factory waste water.

Robustness is usually defined as robustness against mutations: the more neutral mutations, the higher robustness. Paradoxically, an increase in robustness actually hinders the first version of evolvability. As neutrality increases, there is less for natural selection to act upon. However the paradox is resolved if we take a neutrality as a more transient property: neutral mutations occur, but at some point may become visible to natural selection. Then an increase in neutrality allows for a better exploration of the neutral networks of the genotype-phenotype mapping. Due to the extra opportunities for innovation, the other types of evolvability may actually benefit from robustness (Huynen, 1996; Wagner, 2005).

1.4.3 Environment

So far we have presented the genotype-phenotype map as a rather isolated, static process, only hinting at the impact of the environment in which the evolutionary processes takes place. It should be clear from the fact that the genotype contains information on how to sustain and thrive in given circumstances, that the environment is an important component in evolutionary processes. As mentioned previously, competition for resources, predation, pathogens and many other types of interactions heavily influence the probability of survival and hence of successful reproduction. Also, the development from genotype to phenotype is often influenced by the environment, called phenotypic plasticity. A classical example is the growth of cloned plants on a mountain slope. The phenotypes of these clones showed a highly non-linear relationship with their environments (Clausen *et al.*, 1958; Lewontin, Fall 2008). A practical example is the usage of temperature sensitive mutants of yeast, fruit fly and other model organisms in the laboratory.

Although it has been common practice to separate ecological and evolutionary modeling, the last decade a string of studies have shown that integrating the ecological and evolutionary time scales allows for the evolution of phenomena such as niche creation, resulting in an increased ecological diversity (Lindgren, 1991; Ray, 1991; van der Laan & Hogeweg, 1995; Kaneko & Yomo, 2000; Takeuchi & Hogeweg, 2008), species stability (Savill & Hogeweg, 1998), but also ecosystem instability (Solé *et al.*, 2002) and individual 'smartness' (Hillis, 1990; Pagie & Hogeweg, 1997; de Boer & Hogeweg, 2009). Two factors the referred studies have in common are (co-)evolving interactions among individuals and the spatial structure of a population. The latter introduces a degree of locality: individuals only interact with closeby ones. Mesoscale patterns, such as patches, spirals and chaotic waves may arise and introduce additional levels of selection that strongly impact the evolutionary process.

Organisms are not only influenced by their environment, they in turn structure their surroundings as well. This occurs both through direct interactions with other individuals (e.g. predation, parasitism, cooperation), and simply by taking resources and excreting waste products. For instance, the distribution of food in the environment determines the foraging opportunities of a group of individuals, which may affect the food learning opportunities of the entire group and shapes their diet preferences (van der Post & Hogeweg, 2006, 2008).

We extend our model of gene regulatory networks by allowing individuals to sense and alter their environment. In this manner we study an eco-evolutionary model where individuals accommodate their environment to their needs.

1.5 Outline

In this thesis we use computational and mathematical models to study the evolution of genomes, gene regulatory networks and ecosystems. The studies we

perform revolve around several overlapping themes. Using individual-based models we investigate the evolution of populations in a dynamic environment. In **chapter 2** we focus on the evolution of genome organization as a side-effect of adapting to recurrently changing environments. Inspired by the experimental evolution on yeast strains (Ferea *et al.*, 1999; Dunham *et al.*, 2002), we find that the combined mutational processes of transposable elements and gene duplications/deletions may lead to organized genomes. Genes group by function and in this manner allow for fast adaptations via large chromosome rearrangements.

In **chapter 3** we apply a similar protocol of evolution in a dynamic environment, yet here we focus on the gene regulatory network. We let a population of individuals evolve in a randomly changing environment. That is to say, the timing of the change is random, the different evolutionary targets are predefined. Most interestingly, we find that evolution organizes the network architecture of the individuals such that they adapt to recurring changes in the environment with few mutations. Moreover, a single hub-gene occupies a special position in the network: its duplication and/or deletion causes the network to switch from one evolutionary target to another. This is observed repeatedly as the environment changes, and analogous to a normal sensor we named it an evolutionary sensor.

In chapter 2 and 3 we imposed environmental changes on the population. Another class of dynamic environments is generated by the evolving population itself. Each individual takes up resources and excretes waste products, and thus influences its own surroundings. In **chapter 4** we study a basic ecosystem in which individuals may shape their neighborhood. One of the main issues we investigate is the effect of a spatially extended system versus a well-mixed system. We show that being able to create a local environment, and "co-evolve" with it, results in smart individuals. Furthermore, well-mixing the system such that there is only a global feedback between the population and its environment enhances specialization and the evolution of cooperative communities.

Finally, in our last study, in **chapter 5**, we return to the transposable elements of chapter 2. Since the discovery of RNAi many connections have been found between RNAi and the regulation of transposon activity. However, given the lack of an obvious pattern in the presence and absence of various proteins related to RNAi (Shabalina & Koonin, 2008), it is unclear which are required for regulating transposon activity. One such protein is RdRP, which performs important steps such as dsRNA formation and amplification of siRNA. To shed light on the presence and absence of RdRP, we investigate two mathematical models of the transcriptional and post-transcriptional silencing RNAi machinery. We show that for the standard pathway – with RdRP– transposon control is based on a positive feedback loop in the cytoplasm, while in the alternative pathway – without RdRP– control is established through a positive feedback in heterochromatin formation. Thus we provide an alternative mechanism that may operate in human, mouse, fly and other organisms that lack RdRP.

2

Chromosome Rearrangements and the Evolution of Genome Structuring and Adaptability

Abstract

Eukaryotes appear to evolve by micro and macro rearrangements. This is observed not only for long-term evolutionary adaptation, but also in short-term experimental evolution of yeast, *S. cerevisiae*. Moreover, based on these and other experiments it has been postulated that repeat elements, retrotransposons for example, mediate such events.

We study an evolutionary model in which genomes with retrotransposons and a breaking/repair mechanism are subjected to a changing environment. We show that retrotransposon-mediated rearrangements can be a beneficial mutational operator for short-term adaptations to a new environment. But simply having the ability of rearranging chromosomes does not imply an advantage over genomes in which only single gene insertions and deletions occur. Instead, a structuring of the genome is needed: genes that need to be amplified (or deleted) in a new environment have to cluster. We show that genomes hosting retrotransposons, starting with a random order of genes, will in the long run become organized, which enables (fast) rearrangement-based adaptations to the environment.

In other words, our model provides a "proof of principle" that genomes can structure themselves in order to increase the beneficial effect of chromosome rearrangements.

2.1 Introduction

The sequencing of several eukaryotes and the research that followed in its slipstream lead to important insights. Transposable elements are found to be a source of genetic innovation and to have regulatory functions in many organisms (Biemont & Vieira, 2006), gene order is not random (Hurst *et al.*, 2004) and genomes evolve by micro and macro rearrangements (Britten *et al.*, 2003; Fischer *et al.*, 2001; Seoighe *et al.*, 2000). Micro rearrangements include inversions of a couple of genes, single gene duplications and deletions (indels). Comparative sequence analysis shows that macro rearrangements, are localized at the telomeres and centromeres (Eichler & Sankoff, 2003; Murphy *et al.*, 2005). It appears that sites on the genome are being reused in the movement and copying of large segments.

In short-term evolution these processes appear to play a role as well. A striking example is given by Dunham *et al.* (2002). Several repeated experiments were performed where baker's yeast, *S. cerevisiae*, was placed in a glucose-limited environment for about 300 generations. Looking at the genomic changes of the resulting strains, they made several observations. Firstly, large chromosome segments are copied and deleted in the majority of strains. Such events are called gross chromosomal rearrangements (GCRs). Secondly, many strains have such huge mutations at the same location on the chromosomes. Dunham *et al.* (2002) suggest the locations to be fragile sites. Thirdly, although GCRs are abundant, there are also non-GCR ways of adapting to a low-glucose environment (Brown *et al.*, 1998). Adaptation via GCRs has also been observed by Hughes *et al.* (2000); Infante *et al.* (2003); Schacherer *et al.* (2005), who evolved various deletion mutants. Regaining the function was in half of the cases accompanied by GCRs.

Dunham *et al.* (2002) found repetitive DNA originating from retrotransposons at the flanking regions of GCRs. In addition we know that the transcription of the yeast Ty1 family of retrotransposons is activated under various stress conditions (Lesage & Todeschini, 2005). Both suggest that retrotransposons are a means of evolutionary adaptation to environmental changes. In a broader perspective retrotransposons have been implicated in chromosome evolution of yeast (Fischer *et al.*, 2000; Hughes & Friedman, 2004; Umezu *et al.*, 2002). In other words retrotransposons are linked to genome evolution on different time scales.

Though details need to be filled in, the mechanisms behind GCRs are established. It is known that replication forks tend to stall in regions of repeat elements (e.g. tRNA genes, transposons, telomeres) and subsequently cause a double-strand break (DSB) (Cha & Kleckner, 2002). From fission yeast we know this promotes ectopic recombination (Lambert *et al.*, 2005). Alternatively, Koszul *et al.* (2004) explains segmental duplications by hypothesizing a role for break-induced replication after DSBs. These modes of homologous recombination result in translocations, deletions and sister-chromatid exchanges (Mieczkowski *et al.*, 2006).

From the above discussed results on short-term adaptation of yeast by sev-

eral, yet reoccurring GCRs, it is tempting to hypothesize that the genome is structured to increase the probability of favorable mutations in alternate environments and hence have an increased adaptability. Such a hypothesis is almost impossible to substantiate experimentally, as one cannot rule out that there exists only a tiny set of beneficial mutations and as a consequence selection produces the observed pattern. We therefore apply a computational approach to investigate whether well-established mutational mechanisms could lead to such an outcome.

We define a simple evolutionary model with random mutations in the form of single gene indels, retrotransposons which are the source of repeat elements and DSBs on retrotransposons that possibly lead to chromosome rearrangements. Given this set of mutational events we study the evolutionary dynamics in a changing environment. We assume different environments require more or less of certain gene products. For simplicity we ignore in the present model gene regulation and instead assume adaptation to protein requirements is only through the gene copy number.

In this paradigm system we show that long-term evolution leads to a structuring of the genome, which in turn leads to faster short-term adaptation to the environment. The results suggest that retrotransposons despite their deleterious effects, may have a beneficial effect in the long term.

2.2 Methods

2.2.1 Model structure

The model (Figure 2.1A) consists of an asexually reproducing population of individuals on a grid (spatial structure) adapting to an environment that is homogeneous in space and changing in time. The grid is updated as a standard cellular automaton, i.e. all grid cells are updated synchronously. The environment provides the evolutionary goal of the individuals. In its simplest form it switches between two target genotypes according to a Poisson process (unless mentioned otherwise $\lambda = 1.5 \cdot 10^{-4}$). The general idea is that the environment defines the copy number of a subset of genes in the target genotypes.

2.2.2 Individual

An individual performs two actions: reproduce and die. Death happens with a specified probability (fixed at 0.1). Reproduction r_i requires an empty grid cell to place the offspring. Given such an empty location, the eight neighboring individuals, called *nbh*, compete on basis of their fitness score f_i

$$r_i = \frac{f_i^p}{\max\left(\sum_{j \in \text{nbh}} f_j^p, \theta\right)}$$

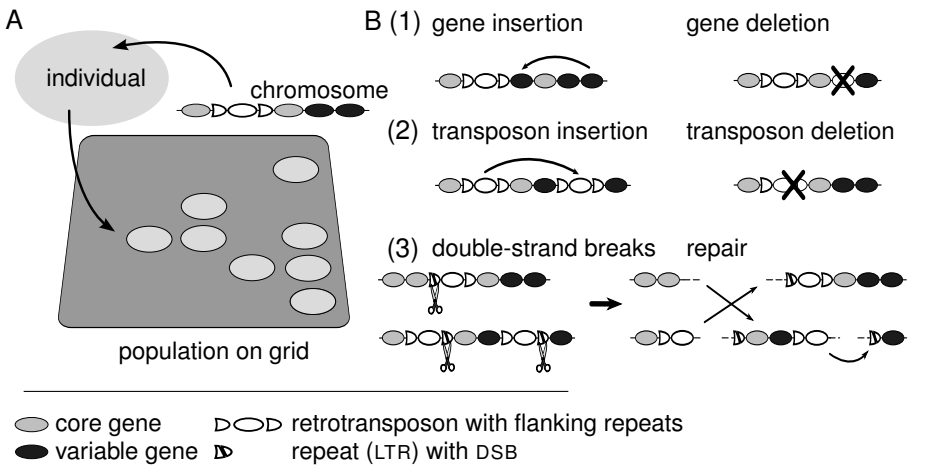

Figure 2.1 – Individual-oriented model of retrotransposon dynamics. (A) The model structure. (B) Three types of mutations: (1) single gene indels; (2) retrotransposon copying and removal, removing single LTRs is not shown; (3) DSBs followed by repair, with rearrangements possibly occurring.

The threshold θ (fixed at $1 \cdot 10^{-4}$) ensures that nothing may happen if there are very few individuals in *nbh* or all individuals are very unfit. Given the relative fitness r_i of each individual in the neighborhood *nbh*, one is selected according to the fitness proportional selection scheme. Reproduction itself encompasses copying the genome, mutating and dividing into two daughters. One of the two replaces the parent, the other is placed in the empty grid cell. For simplicity we only consider asexual reproduction.

In most runs selection pressure is increased by raising f_i to a power p (fixed at 10). It increases the chance that a beneficial mutant spreads in the population, which in turn allows for faster simulations while the results remain qualitatively equivalent. In our results we discuss the effects of selection pressure in more detail.

2.2.3 Fitness

An individual holds a genome, which is a single linear "beads-on-a-string" chromosome. In the majority of our simulations it is initialized with two types of genes, 20 'core' and 20 'variable' ones, and 10 retrotransposons with long terminal repeats (LTRs).

The fitness of an individual f_i is determined by the environment. The environment switches between two states, each associated with an optimal genotype. In one state one copy of each of the variable genes is optimal, and in the other two copies is optimal. In contrast, one copy of each of the core genes is required in both environmental states. Missing any of the genes is lethal, while having extra

copies results in lower fitness. A biological interpretation is that core genes correspond to genes responsible for essential functions (e.g. cell cycle) and variable genes relate to the ones that process resources (metabolites) from the surroundings.

Fitness is a value in the interval $[0, 1]$ and defined in terms of a raw score s_i. Maximizing the fitness amounts to minimizing the raw score.

$$f_i = \begin{cases} 1 - s_i/30 & \text{if } s_i \leq 30 \\ 0 & \text{if } s_i > 30 \end{cases}$$

The raw score is quantified as follows

$$s_i = D_i + \max(0, t_i - 25)$$

D_i is the distance between the copy number of each gene (both core and variable) and the current optimal genotype. As the retrotransposon dynamics lack any control mechanism, a penalty is added to the gene distance if i has more than 25 retrotransposons (t_i is the number of retrotransposons in the genome). A penalty on the number of single LTRs is included in a genome size penalty (threshold size 250). However it is generally not applied during runs and therefore left out of the formulas.

We also perform simulations with an extra group of 20 variable genes that follow an additional, independent environmental cue. This creates a setting with four different environmental states that the individuals adjust their two subsets of variable genes to.

2.2.4 Mutational events

At reproduction the chromosome is duplicated, after which three types of mutational events may occur on the diploid genome (Figure 2.1B). The first type is *gene indels*: gene insertion and deletion. The former is the act of copying a gene and placing it at a random position in the genome, though it is never inserted in between a retrotransposon and its LTRs. The latter is deleting a gene. The second type is *retrotransposon dynamics*: retrotransposon insertion and deletion, and LTR deletion. Insertion is copying a retrotransposon (including the flanking LTRs) and inserting it at a random position in the genome. Deletion is always done via single-strand annealing, which leaves a single LTR. Such a single element (i.e. one that is not next to a retrotransposon) can be removed as well. The third type of mutation is GCRs which happen through DSBs at LTRs. These DSBs are repaired by randomly reattaching chromosome segments to each other, with the constraint that the beginning and end of each chromosome are kept as such. In other words, the first and last segment of a DSB-damaged chromosome are the first and last of a recombined chromosome. At least one DSB per chromosome is needed for a rearrangement to occur ("swapping tails"). During this process no chromosome segments are lost, however the resulting two chromosomes may be of unequal length and/or content. Thus as each daughter cell receives a chromosome, they may have deletions and/or translocations of chromosome segments.

2.2.5 Ancestor tracing

During a simulation each individual has its own unique identification and knows its parent's. This enables us to reconstruct genealogies: one of the best individuals at the end of the simulation is selected and all its ancestors are traced back to the start. By recording genomes and which mutations occur along this lineage, we are able to look at the mutational mechanics in detail.

2.3 Results

2.3.1 Two typical runs

In the evolutionary runs the starting point is that retrotransposons have successfully invaded the population. We discuss two typical runs: a run of $1 \cdot 10^6$ time steps with one set of 20 variable genes per individual and one of $2 \cdot 10^6$ time steps with two sets of 20 variable genes. In both runs a homogeneous population is initialized on a 100x100 grid and subjected to a changing environment.

One group of variable genes

The run is shown in Figure 2.2A. Each time the environment switches, the average gene distance jumps to 20 (maximal distance) and the population adapts to the new environment.

In Figure 2.2B, the close-up shows a small GCR with a nett distance gain of 4, followed by single gene indels. The variable genes make a little jump due to the GCR and then slowly increase to a double copy number (shaded area). We see a genotype with a few extra core genes spread through the population immediately after the environmental switch. This hitchhiking of core genes is caused by the small GCR. It is interesting to note that GCRs are readily applied, even though genes are randomly ordered on the genome and the extra core copies need to be removed again. The close-up at the end of the run (Figure 2.2C) shows a rather different style of adapting. Mutations by GCRs spanning a full set of variable genes cause the population to adapt extremely fast. The doubling of the variable genes still shows a few hitchhiking core genes, which are subsequently deleted.

These observations indicate a long-term process of genome restructuring. To quantify the ordering of genes on the chromosomes we look at all adjacent pairs of genes, while ignoring retrotransposons and LTRs. We take Cramer's phi coefficient[1] of these pairs, which gives a value in the range $[0..1]$. The higher the phi coefficient, the higher the degree of clustering of genes by type (core or variable).

In Figure 2.3A the average organization of genomes in the population is shown. The first third of the run the genomes do not show any clustering, after which the population evolves toward a high degree of organization (phi $= 0.76$).

[1] Cramer's phi is derived from the chi-square statistic, in formula: $\phi = \sqrt{\chi^2 / Nk}$ with, in our case, N the total number of genes and k the number of variable groups.

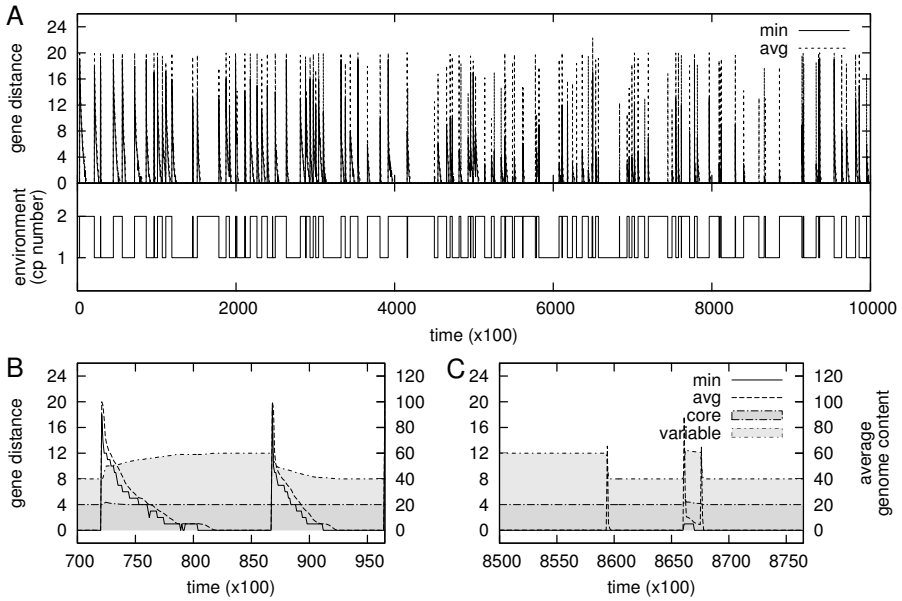

Figure 2.2 – Typical run. Parameters (per gene, retrotransposon, LTR): single gene copy and removal $= 0.5 \cdot 10^{-5}$, retrotransposon copy, removal and LTR removal $= 1 \cdot 10^{-5}$ and DSB repair $= 6 \cdot 10^{-4}$. Note that at least two breaks are needed for a chromosome rearrangement. The environment switches at $\lambda = 1.5 \cdot 10^{-4}$. (A) The top part shows the average (avg) and minimum (min) gene distance of the population. Every environment change, shown at the bottom, is accompanied by a peak in the gene distance. (B) and (C) are close-up graphs of (A). Both show on the left y-axis the average and minimum gene distance and on the right axis (shaded) average genome content (i.e. the number of core and variable genes). (B) $2.5 \cdot 10^4$ time steps at the start, showing a gradual decrease of gene distance by small steps (indels). (C) $2.5 \cdot 10^4$ time steps at the end, with large and immediate decreases of gene distance. The fast mutation is also observed in the shaded area of variable genes, while the area of core genes remains constant.

This level is kept, although it sometimes drops slightly (e.g. at $t = 6500$ and $t = 9800$). The temporary declines are explained by the fact that there are only very few sequences in gene-order space with a high phi coefficient and apparently the indirect selection is not strong enough to maintain them. The main cause is retrotransposons copying themselves through the genome, creating alternative breakpoints and hence via rearrangements the gene order is randomized to a small degree.

The clustering of genes, separated by strings of LTRs is clearly visible if we look at the genomes in Figure 2.4. Figure 2.4B shows from left to right the core genes and two clusters of variable genes. Apparently the individuals are increasing the probability of DSBs at certain locations in their genome and therewith enable fast adaptations to the environment. For instance, a scenario of a rearrange-

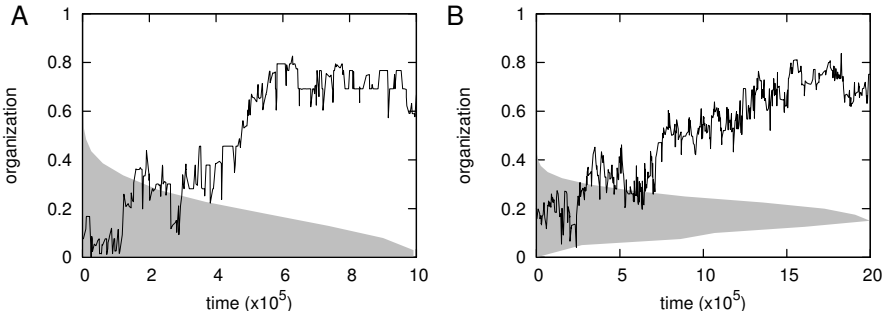

Figure 2.3 – Average genome organization in the population. (A) One set of variable genes. The run of Figure 2.2 is shown. (B) Two sets of variable genes. In both figures simulation time is set against average organization of the population. The latter is expressed as Cramer's phi coefficient. To demonstrate the significance we show shaded areas that give, in the horizontal direction, the frequency distribution of organization from a sample of one million randomly generated genomes. Note that the second simulation is twice as long.

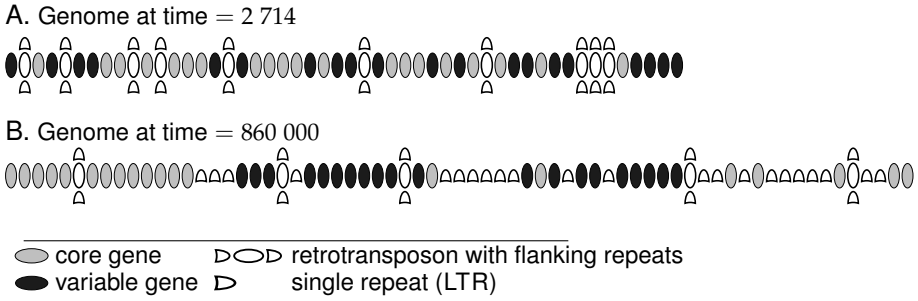

Figure 2.4 – Gene order in two randomly picked individuals, early and late in the simulation. The genomes are taken from the run of Figure 2.2. Genome A shows no organization, in contrast to genome B which displays a clear clustering of the two types of genes and an increase in single LTRs.

ment would involve a DSB in the middle of one chromosome and at the right hand side of the other chromosome. In this manner swapping the right-hand tails, would result in copying half of the variable genes with only one or two core genes hitchhiking.

Two groups of variable genes

The evolution to an organized genome in which only two types of genes are present, seems a rather simple task. We show this behavior can be extended to multiple groups of variable genes. This creates a more complicated task where in one environment more of one group of genes is needed, in another more of the other group of variable genes and in yet other environments both or none are

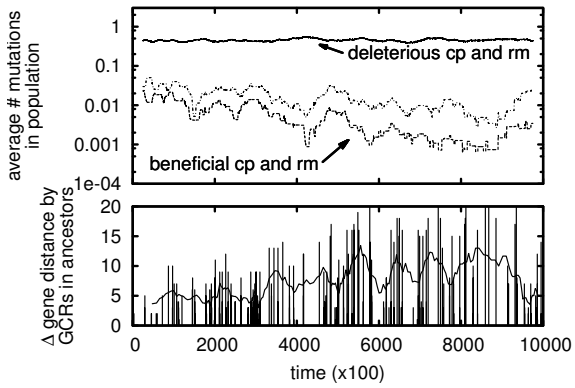

Figure 2.5 – Mutations in the run of Figure 2.2. In the top graph the population running-average (window size 100001) of beneficial and deleterious single gene indels, i.e. copy (cp) and remove (rm), is shown. The bottom figure shows fitness gains expressed as Δ gene distance made by GCRs in the ancestor trace: the decrease in distance to the target genotype. The vertical lines show the actual gains made by each GCR event, while the running-average curve (window size 9) shows the trend of increasing gains.

needed in more than one copy. For $2 \cdot 10^6$ time steps we have run a population with individuals having 20 core genes and two groups of 20 variable genes. The environment has an independent signal for each variable gene group, i.e. a total of four environmental states, both with a probability of change $\lambda = 1 \cdot 10^{-4}$.

We observe qualitatively similar behavior in the adaptations to a new environment as the one-group case (data not shown). For instance, core genes hitchhike with both variable groups, and the variable groups show hitchhiking among each other as well. Again, indirect selection causes genes to cluster by type. In Figure 2.3B Cramer's phi is plotted as the measure of average organization in the population. If we compare the curve to the 'one-group' case, it is remarkable that similar levels of organization are reached in this more complex case and that it takes no more than twice the amount of time.

The majority of the results that we discuss next, is based on the behavior of the 'one-group' case, as the simulations are rather computationally intensive. The 'two-group' case is considered to be a strong indication of the generality of these results within our framework.

2.3.2 Mutational dynamics

We study the 'one-group' run in more detail. In the top graph of Figure 2.5 the running average of single gene insertions and deletions in the entire population is shown. Deleterious mutations make up the bulk and have a fairly constant rate throughout the run, while the number of beneficial indels decreases. The

rearrangements show a different behavior. If we look at the ancestor trace, it provides us with a detailed view on the dynamics of the *beneficial* mutational mechanics. The bottom graph (Figure 2.5) shows the usage of GCRs in the ancestor trace. In contrast to the change in frequency such as the indels show, the rearrangements keep a rather constant rate. However GCRs become more effective as the simulation progresses.

The above observations are reinforced by Table 2.1. If we take the first $3 \cdot 10^5$ time steps (Table 2.1, first row) we know genomes are not yet organized. Hence we observe many single gene indels in the ancestors and they are mostly beneficial ones. Obviously few mutations are deleterious due to the strong selection. Even though we do not have organization, GCRs are applied, as the number of beneficial ones more or less equals the overall occurrence (Table 2.1, third row). This is in accordance with the previous observation that rearrangements occur at about the same rate during the run. The difference with the population average shows that virtually all beneficial GCRs occur in the ancestor lineage.

After $6 \cdot 10^5$ time steps the population consists of organized individuals (Table 1, second row). Compared to the beginning of the run single gene insertions have nearly disappeared. Gene removals have decreased drastically too, but still occur. The explanation is that, as mentioned in the previous section, core genes tend to hitchhike on rearrangements and subsequently one of the two genes is removed via a single gene deletion. GCRs are applied less often than on average in the simulation, but this is compensated by having GCRs that span more variable genes. The slight rise of deleterious GCRs seems paradoxical, but is caused by mutations that by accident anticipate an environmental switch. These mutations are categorized as deleterious, but change to beneficial within the individual's lifetime. Such rearrangements of anticipatory nature strengthen the idea that there are locations on the genome with a higher probability of a DSB. Correcting for these anticipatory GCRs, the rate of deleterious ones in the ancestor lineage is $0.49 \cdot 10^{-4}$ per reproduction, which is lower than the population average.

Interval (x100)	# Individuals	Mutations (x10^{-4}) ben : del		
		Copy	Remove	GCR
0 - 3000	64386	16.93 : 0.15	34.17 : 0.77	10.56 : 0.62
6000 - 10000	82466	0.12 : 0	9.34 : 0	9.82 : 1.21 (0.49)
0 - 10000	209720	6.77 : 0.05	18.60 : 0.29	10.11 : 0.95
0 - 10000	pop $\approx 9 \cdot 10^9$	0.0064 : 0.50	0.0184 : 0.49	0.0026 : 0.71

Table 2.1 – Mutations in the ancestor lineage and population. Average beneficial (ben) and deleterious (del) mutations per individual, with the individuals being from the ancestor trace or the population (fourth row). In the first interval individuals have not organized their genomes yet, in the second interval they have. The third interval is the entire simulation (see Figure 2.2). In the second row the deleterious GCR value in parentheses is corrected for anticipatory rearrangements. As a reference to the ancestor trace, the bottom row shows average mutations for the population (pop) in an entire run.

We conclude that the population is forced to deal with retrotransposons and the frequent rearrangements they cause. One would expect to see the retrotransposons removed by selection, because of the higher death rate they cause. Indeed, chromosomal rearrangements are almost always bad as seen from the three orders of magnitude difference between advantageous and deleterious GCRs in the population (Table 2.1, bottom row). But by reordering their genomes individuals utilize the effects of GCRs and gain a novel (fast) way of adapting, which overrules the negative effects of retrotransposons and their flanking LTRs. The mechanism of restructuring the genome depends on GCRs, the selective loss of surplus genes and the reordering due to single gene indels.

2.3.3 Invasion of retrotransposons

Invasion and maintenance of retrotransposons in a sexually reproducing population has been explained in terms of transposition and recombination (Rouzic & Capy, 2005). It is argued that in a clonal population selection is necessary to explain the presence of retrotransposons (Edwards & Brookfield, 2003). In our model we have seen maintenance of retrotransposons through the indirect selection for evolvability. However we expect such selection to be minimal during an invasion, as the genomes are still unorganized.

We study the paradox of fixing retrotransposons in a host genome by means of invasion simulations. A GCR-enabled population is introduced in an 'optimal' indel-only population for different initial population sizes and retrotransposon numbers. The settings of a typical run are taken with one alteration. The indel-only individual mutates fast if it needs to adapt, yet it hardly mutates if it is perfectly fit. Such behavior is accomplished by evolving the mutation rates as well: during reproduction an individual may mutate (probability $5 \cdot 10^{-4}$) its single gene mutation rates. The new value is drawn from a uniform distribution in the range $[0, 2.5 \cdot 10^{-4}]$. The upper bound is above the maximal mutation rate ($2.0 \cdot 10^{-4}$) for which adaptation to the environment can be maintained. Thus we assume a worst case scenario to study the invasion dynamics.

If we let a fraction of the population of 0.1 ($\approx$ 1000 individuals) have 5 retrotransposons per genome, 32/400 runs result in a successful takeover. Increasing the number of individuals to 2000 gives success in 52/400. Naturally, the more GCR-enabled individuals, the higher the probability of taking over the population, though even a majority does not guarantee a successful invasion (data not shown). The interesting observation is that for smaller numbers of invading individuals we still have successful invasions. If we reduce the fraction of individuals with retrotransposons to 0.05 ($\approx$ 500 individuals), 8/400 invasions succeed. Taking a smaller patch of 100 individuals gives success in 6/400.

We conclude that although retrotransposons do not give an inherent selective advantage in an unordered genome, on an evolutionary timescale it seems fairly easy for them to invade a population based on the positive effect they have on evolvability if they happened to be at the appropriate location in the genome.

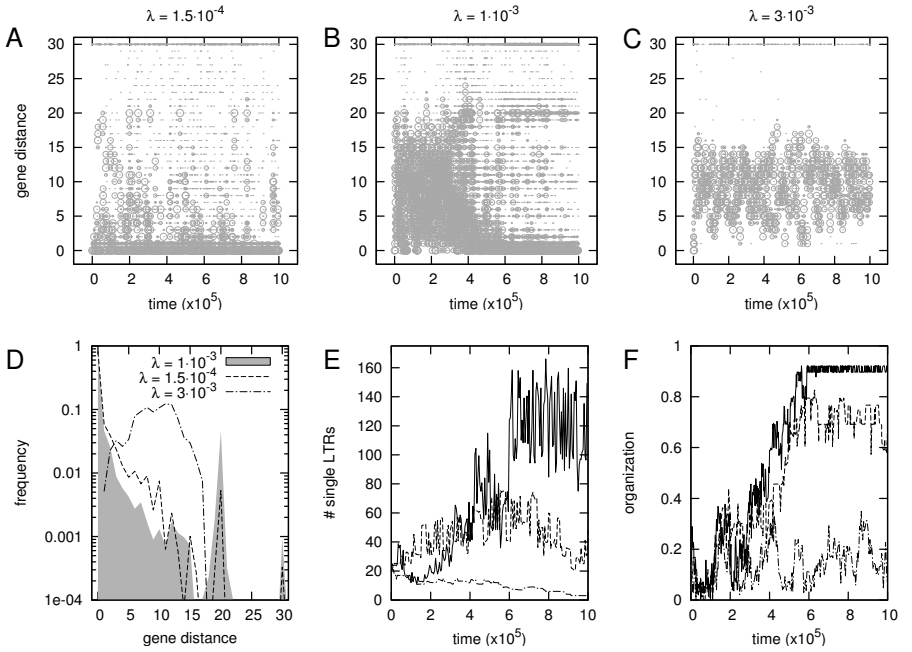

Figure 2.6 – Behavior for different rates of environmental change (λ). (A), (B) and (C) show the distribution of distances to the optimum (0) during the run. Populations are sampled at regular intervals (period of 10000 in (A), 1000 in (B) and 4000 in (C)) and also at environmental switching in (A). The size of each circle represents (logarithmically) how many individuals have such a distance to the optimal genotype. (D) The frequency distribution of the gene distances during the second half of the runs. The intermediate rate of change is shaded. (E) The number of single LTRs for each of the three runs. Please note that the upper curve depicts the intermediate value of environmental change, while the bottom curve is the highest rate of change. (F) The average organization in the populations expressed as Cramer's phi. Upper, middle and lower curve have the same λ values as in (E).

Thus we have a convenient mechanism to introduce repeat elements into a genome.

2.3.4 Rate of environmental change

An important parameter in our model is the rate of switching from one state to another. In our typical runs we used $\lambda = 1.5 \cdot 10^{-4}$ for the one variable-group simulation and $\lambda = 1.0 \cdot 10^{-4}$ (per environmental signal, thus effectively $\lambda = 2.0 \cdot 10^{-4}$) for the more challenging two group case. As we see clearly in Figure 2.6A, in these ranges of λ it is usually possible for the population to fully adapt to the current environment.

When we set $\lambda = 1 \cdot 10^{-3}$, the global behavior changes. In the first half

(Figure 2.6B) often the population does not reach the optimal genotype. Instead the fast fluctuations of the environment cause the population to average over the two environmental states. However there is still a bias toward low distances. The bias and as a consequence the small adaptations to each environmental switch are sufficient to trigger the rise of gene ordering by group. In the second half of the run the population consists of organized genomes. Even more interesting is that on a population level we always have two genotypes, i.e. we have population-based diversity where the opposite genotype, which is present at almost 5%, is constantly being generated from the current optimal genotype.

We increased λ to $3 \cdot 10^{-3}$ (Figure 2.6C). The population basically experiences an average environment and settles at an average gene distance of 10, half of a variable-group size. Yet if we start the run with organized individuals, the gene ordering is kept and the population switches from one to the other environment, behaving like Figure 2.6B. There also appears to be a trend to increase the average number of LTRs to cope with such a fast switching rate (data not shown).

The second half of these three runs is summarized in Figure 2.6D. For populations with organized genomes the peaks at distance 0 and 20 are clearly visible. The latter is more prominent for the run from Figure 2.6B, which emphasizes the idea of population-based diversity. The 'averaging' run (Figure 2.6C) shows a characteristic curve with a mean distance around 10.

The role of LTRs

The interesting behavior we see in Figure 2.6B is that despite the near absence of adaptation in the first half, organization, and therewith adaptation, may still arise and even reach higher levels than in our two typical runs, as discussed above. In most runs this is accomplished by both retrotransposon dynamics and LTRs. Figure 2.6B is a special case as due to a stochastic fluke retrotransposons are purged from the population before any truly organized genomes dominate the dynamics. Only single LTRs are left and as the only mutation LTRs can have is deletion from the genome, we would expect them to disappear. However due to positive indirect selection of small beneficial GCRs the removal of LTRs is extremely slow. Such GCRs occur already in non-organized genomes and are often fixed in the population (i.e. the mutant takes over the population).

As organization starts to develop, the number of single LTRs per individual increases (see Figure 2.6E). The growth is accomplished by rearrangements. LTRs tend to cluster between core and variable genes, thus increasing the probability of creating a fit daughter for the opposite environment. GCRs resembling unequal crossover then may enlarge or shorten these sections of single LTRs. Eventually the organization stabilizes (Figure 2.6F), while the number of single LTRs fluctuates rapidly around a mean of approximately 120.

Apparently the reordering is the result of gene indels and the maintenance of single LTRs by indirect selection. This leads to the conclusion that retrotransposons are not necessary for genome structuring, only repetitive elements like the long-terminal repeats are. This is in accordance with simulations started with

only single LTRs (data not shown).

The role of selection

The actual value of λ for which we observe these modes of behavior depends on the selection power p (see Methods). We relax selection pressure if we choose low values of p ($p = 2$ for example). Consequently there is only a small beneficial effect of having one distance less, resulting in a slow adaptation of the population to the optimal genotype. Thus to observe the evolution of adaptability (reordered genomes) a low λ is needed and due to a tighter error threshold, it is necessary to lower mutation rates as well. On the other hand, as p increases the error threshold is relaxed, hence higher values of λ can be chosen. This enables us to speed up simulations. Except for the timescale there is no qualitative difference in the obtained results.

2.3.5 Parameter sensitivity

The organization as described in the typical runs is a robust phenomena. However for identical parameter values the onset of organization ranges from soon after the start to halfway through the run.

The robustness is further explored by starting at different mutation rates, by removing any spatial patterns that may influence the results and by introducing extra penalties per retrotransposon. Besides initializing at various rates, we let the individuals evolve their mutation rates. The rates may change during a simulation, and each changes with small steps. We have three different step sizes, one for each type of mutation rate (indel step $1.5 \cdot 10^{-7}$, retrotransposon $1 \cdot 10^{-7}$ and DSB repair $1 \cdot 10^{-3}$). By allowing mutating mutation rates we broaden our view on parameter space with a limited set of runs. For instance, if we start with the rates of a typical run and these rates would evolve to very different values and behavior, we could say the behavior of a typical run is an artifact of 'forced' rates, not natural behavior.

Changing mutation rates

We vary each type of mutation events by decreasing or increasing the rate, while the rest of the parameters is initialized as for the typical run (see Figure 2.2). For each setting we have run three simulations, which we discuss shortly. For three times lower DSB rates ($2 \cdot 10^{-4}$) organization develops more slowly, simply because less breaks occur. An almost twice as high rate of DSB repair ($10 \cdot 10^{-4}$) does not show any qualitative difference. Ten times lower retrotransposon rates ($0.1 \cdot 10^{-5}$) make the mobile genetic elements more vulnerable to stochastic fluctuations. In two of the three runs, retrotransposons are removed from the population, but single LTRs are kept. Individuals manage to gain and keep organization in these runs, albeit less stable. For single gene rates we observe that more than twice higher rates ($1.2 \cdot 10^{-5}$) result in better organization, even above

0.80. Because the organization scale is not linear, but represents half a normal distribution, such levels of organization are extremely rare. They are due to the increased reordering effect of single gene mutations. Indeed, a five times lower indel rate ($0.1 \cdot 10^{-5}$) than in a typical run results in a lower level of organization, around 0.6, within an equal amount of simulation time.

Retrotransposon penalty

Throughout a run, retrotransposons cause many lethal mutants by generating erroneous chromosome rearrangements. In other words, retrotransposons increase the death rate of individuals. We examine if the addition of a fitness penalty per retrotransposon (0.04 extra gene distance per retrotransposon) instead of a penalty if a threshold of retrotransposon copies is exceeded, is a viable alternative. In this scenario retrotransposons are rather quickly removed from the population. Yet, individuals still reshuffle their genomes. They apply the trick of using single LTRs, as described above.

Structural stability

In our typical run, variable genes comprise a large part of the genome; either 20 or 40 genes, compared to 20 core genes. The question arises if the ratio core-variable genes influences the evolution of organization. We have performed simulations with more core than variable genes, such that the number of core genes (50 genes) is always more than the variable ones. The runs show genomes still develop organization.

Another test of the structural stability of our model is the introduction of a special element: the centromere. We add the constraint that each chromosome has to have one centromere, while both having more than one (dicentric) and missing the centromere (acentric) is lethal. The simulations resulted in organized genomes, with the centromere being pushed in the direction of the beginning or end of the chromosome.

2.4 Discussion

The concept that "some mutations are more equal than others" is rather controversial. Generally mutations are considered random events, with selection acting on them. However, our results show that some classes of mutations may occur preferentially and, furthermore, these mutations are likely to be advantageous. In our paradigm model we take random, well-established mutational events and show in our results that some (phenotypic) mutations are being favored. We observe that genomes restructure their gene order such that chromosome rearrangements are very likely to produce the target genotype of an alternate environment. However it is not the only possible behavior. Depending on the rate of change of the environment and the previously achieved degree of ordering of

the genomes, we may see the evolution of an average genotype with low fitness that 'integrates' over the environmental states.

There are three assumptions of our model we would like to discuss. Firstly, we make a worst case assumption by using a very simple genotype-phenotype relation, in which each alteration in gene copy numbers affects an individual's fitness. In contrast, experimental findings show that the detrimental effects of having an extra gene copy are not so large (Wilke & Adams, 1992). Often it is compensated for on the gene regulation or metabolic level.

Secondly, our method for resolving DSBs only approximates the *in vivo* mechanism. We ignore explicit homologous repair as an option for resolving DSBs and we assume randomness in concatenation of chromosome parts. In vivo DSB repair is not completely arbitrary, but there are many blanks in our knowledge. Therefore we apply our random repair protocol. For the parameter ranges we investigated, if a GCR occurs, on average two breaks occur in the genome. This allows for simple relocation of a chromosome segment; an entire randomization of the chromosomes is not observed. We also assume that retrotransposons insert themselves at random positions, although Ty families in baker's yeast have a preference for certain genome locations (Lesage & Todeschini, 2005).

Thirdly, if selection is reduced, drifting by gene indels takes place. This is a source of randomizing the gene order. Yet we mainly use a strong selection pressure, hence mutational drift does not occur. It means that keeping the organization in a constant environment becomes rather straightforward. Therefore we may not extrapolate into environmental switching at low rates, unless we lessen selection pressure.

Another model feature worth mentioning is that the individuals evolve in an explicit spatial setting with local competition. By performing simulations in which the population is mixed at each time step, we established that spatial pattern formation does not play a role in our results (data not shown). This is in agreement with previous results that show well-mixed populations employ evolutionary adaptation and mutational priming (Hogeweg, 2005) as their main strategy.

Pepper (2003) studied gene linkage given unequal crossing over and inversions. While ignoring retrotransposons or repeat elements he finds a clustering of genes too. In his discussion he proposes an evolutionary scenario with, on a longer time scale, a positive feedback between rearrangements and gene linkage which resembles our results. The long term result of adaptability is only indirectly selected for in our simulations, as the fitness criterion, the short term advantage, does not contain any reference to the gene linkage we observe at the end. There is selection on lineage level leading to higher levels of gene ordering and a higher degree of adaptability. What actually happens is that the genomes restructure their mutational landscape with the retrotransposons. They develop a 'coding scheme' that allows them to change swiftly between two phenotypes in a genotypic manner: also known as mutational priming. In principle repetitive elements such as LTRs are sufficient for generating and sustaining the organization. Retrotransposons may be regarded as a vector for introducing such

elements into a genome. At first the mobile elements only hinder their host by causing chromosomal rearrangements. Such arrangements are in almost all cases deleterious. However retrotransposons can establish themselves by providing individuals with the opportunity of using any bit of clustering of genes via GCRs to adapt to a new environment. Single gene indels, together with retrotransposon insertions and deletions then amplify the grouping of genes by their type.

In the introduction we already stated that eukaryotes have a nonrandom gene order. This organization of the genome is observed in terms of clusters of co-expressed genes (Singer *et al.*, 2005), clustering of genes encoding subunits of protein complexes (Teichmann & Veitia, 2004) and functionally related genes (Hurst *et al.*, 2004). All are clustering seen in the light of gene expression. In yeast bidirectional promoters also give a direct relation for co-expression between gene pairs (Cohen *et al.*, 2000). How the eukaryotic genome ordered itself is an open question. Naturally the null hypothesis is that it is not under selection. As gene expression is a noisy process, it could just be an effect of expression leaking (Hurst *et al.*, 2004). Or it could be a side-effect of mutational dynamics. For instance highly expressed genes are within open chromatin, which in turn facilitates invasion of new genes (Hurst *et al.*, 2004).

With our model we add a hypothesis to the 'gene expression' and 'open chromatin' ones, which may be labeled as the 'evolvability' hypothesis, i.e. gene ordering evolves as a consequence of chromosome rearrangements and increases adaptability. Interestingly in yeast a lot of remnants of retrotransposon activity are observed, which we could now hypothesize to still have a functional role in evolution.

In the evolutionary experiments with yeast (Brown *et al.*, 1998; Ferea *et al.*, 1999; Dunham *et al.*, 2002; Schacherer *et al.*, 2004) it is clear that both evolutionary and regulatory adaptation play a role. First, in most experiments GCRs occur and, second, different GCRs and cases without rearrangements appear to lead to similar changes in gene expression. In addition the majority of genes that are over or under expressed are not located on the duplicated or deleted chromosome segments. Thus regulation creates a far more complicated mapping from genome to phenotype than we have considered. We should note that the amount of organization we observe is much larger compared to, for instance, yeast. At present we cannot rule out that the GCRs observed in the evolutionary experiments only cause minor improvements and resemble the ones we frequently observe for randomly ordered genomes. In future work we aim to investigate the interplay between evolutionary adaptation as studied here and regulatory adaptation by extending our model with gene regulation.

Our model provides a "proof of principle" that genomes can structure themselves so as to utilize the beneficial effects of chromosome rearrangements. In short we provide a simple but sufficient model that shows evolution of evolutionary adaptability.

3

Evolution of Evolvability in Gene Regulatory Networks

Abstract

Gene regulatory networks are perhaps the most important organizational level in the cell where signals from the cell state and the outside environment are integrated in terms of activation and inhibition of genes. The last decade the study of such networks has been fueled by large-scale experiments and renewed attention from the theoretical field. Different models have been proposed to, for instance, investigate expression dynamics, explain the network topology we observe in bacteria and yeast, and for the analysis of evolvability and robustness of such networks. Yet how these gene regulatory networks evolve and become evolvable remains an open question.

An individual-oriented evolutionary model is used to shed light on this matter. Each individual has a genome from which its gene regulatory network is derived. Mutations, such as gene duplications and deletions, alter the genome, while the resulting network determines the gene expression pattern and hence fitness. With this protocol we let a population of individuals evolve under Darwinian selection in an environment that changes through time.

Our work demonstrates that long-term evolution of complex gene regulatory networks in a changing environment can lead to a striking increase in the efficiency of generating beneficial mutations. We show that the population evolves towards genotype-phenotype mappings that allow for an orchestrated network-wide change in the gene expression pattern, requiring only a few specific gene indels. The genes involved are hubs of the networks, or directly influencing the hubs. Moreover, throughout the evolutionary trajectory the networks maintain their mutational robustness. In other words, evolution in an alternating environment leads to a network that is sensitive to a small class of beneficial mutations, while the majority of mutations remain neutral: an example of evolution of evolvability.

Author summary

A cell receives signals both from its internal and external environment and responds by changing the expression of genes. In this manner the cell adjusts to heat, osmotic pressures and other circumstances during its lifetime. Over long timescales, the network of interacting genes and its regulatory actions also undergo evolutionary adaptation. Yet how do such networks evolve and become adapted?
In this paper we describe the study of a simple model of gene regulatory networks, focusing solely on evolutionary adaptation. We let a population of individuals evolve, while the external environment changes through time. To ensure evolution is the only source of adaptation, we do not provide the individuals with a sensor to the environment. We show that the interplay between the long-term process of evolution and short-term gene regulation dynamics leads to a striking increase in the efficiency of creating well-adapted offspring. Beneficial mutations become more frequent, nevertheless robustness to the majority of mutations is maintained. Thus we demonstrate a clear example of the evolution of evolvability.

3.1 Introduction

Gene regulatory networks (GRNs) have become a successful tool for understanding the organization within cells and their dynamics. In GRNs information from the cell state and the outside environment is translated into a correctly timed expression of genes. As such, one may argue that GRNs are the nexus of physiological adaptations. However, as soon as the time scale of environmental change exceeds an individual's lifespan evolutionary adaptations will also play a role. In this work we concentrate exclusively on this evolutionary side of the equation.

GRNs have been studied extensively. Randomly generated networks have been investigated, for instance in deriving various characteristics of homogeneous random networks (Kauffman, 1969), assessing attractor landscapes (Aldana *et al.*, 2007) and evolutionary potential (Aldana *et al.*, 2007; Wagner, 2005; Ciliberti *et al.*, 2007a,b; Fernández & Solé, 2007). With the recent insights from the regulatory networks of model organisms, experimentally inspired networks have been investigated as well (Milo *et al.*, 2002; Lee *et al.*, 2002; Li *et al.*, 2004; Teichmann & Babu, 2004; Buchler *et al.*, 2005).

Mutational dynamics (i.e. neutral evolution) have been applied to GRNs in order to explain the global and local topology of GRNs (Milo *et al.*, 2002; van Noort *et al.*, 2003; Barabási & Oltvai, 2004; Kuo *et al.*, 2006; Cordero & Hogeweg, 2006). Evolution with Darwinian selection and gene expression dynamics has been used to generate small biochemical networks realizing specific mathematical functions (François & Hakim, 2004; Paladugu *et al.*, 2006) and to assess the requirements for evolving specific expression patterns (Quayle & Bullock, 2006).

Evolution has also been applied in the closely related areas of signal transduction pathways and metabolic regulation (Pfeiffer *et al.*, 2005; Soyer & Bonhoeffer, 2006; Soyer *et al.*, 2006; van Hoek & Hogeweg, 2006, 2007). Predominantly these networks evolved to a fixed target. This has been successfully extended by evolving towards changing fitness regimes both in a genetic programming context (Pagie & Hogeweg, 1997) and for evolving electronic circuits (Kashtan & Alon, 2005). The latter also demonstrated that alternating the evolutionary targets can decrease the total time needed to reach every target at least once (Kashtan *et al.*, 2007).

In this work we alternate evolutionary targets and focus on the long-term evolution of adapting toward these targets. Reaching an evolutionary target is therefore only the first step: we study the effect of repeatedly evolving towards it. That is to say, we investigate the evolution of the genotype-phenotype mapping, from genome to network, on a longer time scale. To achieve this we do not directly operate on the network level. Instead we explicitly model a genome where mutations occur, and a network derived from this genome. We do not provide the individuals with direct input from the environment and consequently they are absolutely blind to environmental changes. Hence our observations are not influenced by physiological adaptations.

We concentrate our analysis on the evolution of evolvability. The concept of evolvability has been formalized in various ways (Wagner & Altenberg, 1996; Pigliucci, 2008) and we define it as the efficiency of an organism in discovering beneficial mutants. Hence our question is whether evolution can modulate the mutational efficiency of 'generating' well-adapted offspring via the genotype-phenotype mapping. In other words, through the encoding of the network in the genome.

Evolutionary experiments with yeast, *S. cerevisiae*, resulted in compelling evidence for such evolvability (Ferea *et al.*, 1999). Only a small number of mutations were needed to change the expression levels of many genes as well as causing an increase in fitness. In addition, while almost all strains showed gross chromosomal rearrangements, equivalent restructuring of the transcriptome and similar fitness gains were observed in strains with only minor mutations (Dunham *et al.*, 2002). The observation of multiple, short mutational paths suggests the genetic system of yeast is capable of efficiently discovering advantageous adaptations. Similarly, in several independently evolved *E. coli* strains beneficial mutations on the same genes were found to influence large parts of the gene regulatory network (Philippe *et al.*, 2007). These empirical studies strongly suggest the genotype-phenotype mapping itself is a product of evolution and may have become optimized to increase evolvability.

We show that in a changing environment individuals evolve their genotype-phenotype mapping such that they become more efficient at generating adaptive mutants. The evolvability manifests itself as a sensitivity to gene duplication and deletion mutations of one particular gene, an "evolutionary sensor". The duplication or deletion of this gene results in the network switching its state toward the evolutionary target set by the environment. We show that these evolutionary

sensor genes are either hubs of the regulatory network or directly provide input to a hub gene. In addition, our mapping from genome to network introduced a large degree of mutational neutrality. During the long-term evolutionary process the vast majority of mutations remained neutral, in other words, the population was constantly on a mutationally neutral network and evolvability hardly impacted the mutational robustness.

Summarizing, we show that in a dynamically changing environment long-term evolutionary processes and short-term gene regulation dynamics interact such that our gene regulatory networks become extremely efficient at generating advantageous mutations, while they remain mutationally robust.

3.2 Results

To study the evolution of the genotype-phenotype mapping we employed an individual-oriented model (Figure 3.1). At the start it was initialized with a homogeneous population of genomes, from which gene regulatory networks were built (Figure 3.1A). On the genomes mutations occurred, such as gene and binding site duplications and deletions (indels), which influenced the network topologies of these individuals (Figure 3.1B). The individuals were selected for reproduction on the basis of their gene expression pattern, i.e. scoring if genes were correctly *on* or *off*.

As shown in Figure 3.1C, the environment determined the evolutionary target of the population. The goal was always to minimize the Hamming distance of the network state to a predefined expression state, yet which genes were to be turned *on* or *off* changed through time. For the simulations we selected by hand two network attractors from the initial network as the evolutionary targets. In other words, the population adapted to an attractor state and the environment alternated the attractor over time. As mentioned in the Introduction, the individuals could not sense these changes in the environment.

Due to the computationally intensive nature of our model the main results presented are based on a set of 15 replicate runs. In all cases the population evolved evolvability, and 11 (73%) runs showed so-called evolutionary sensors (ES). The latter ones are our focus in this work[1]. For an in-depth analysis we randomly selected a single run with an evolutionary sensor, which we refer to as a typical run.

3.2.1 Evolving to the targets

During a run the population repeatedly adapted as the environment changed. The individuals in the initial population had the desired network states as attractors in their attractor space, yet we observed that in the beginning they were unable to reach these gene expression patterns (Figure 3.2A, distance 1 is reached

[1]We have been unable to pinpoint a specific strategy in the four runs in which the population failed to reach the solution of an evolutionary sensor.

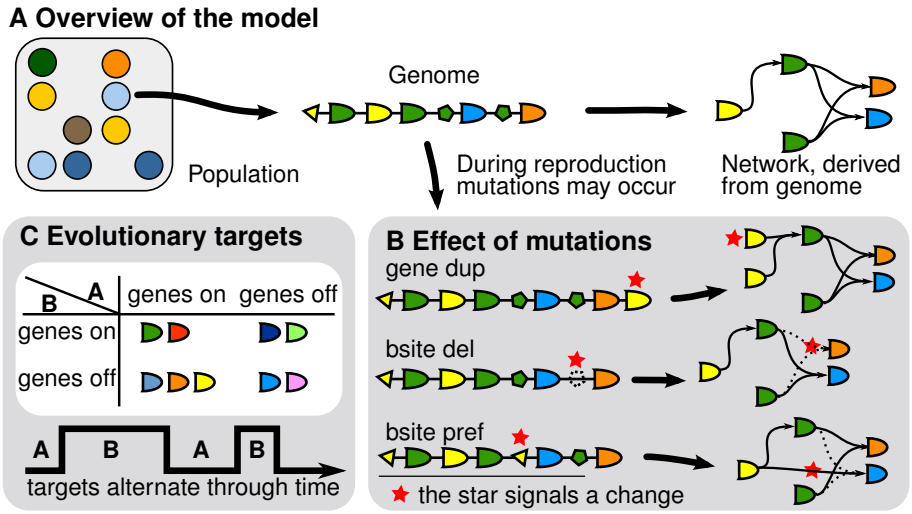

Figure 3.1 – Overview of the model. A. Simulations are run on a 150x50 lattice for $6 \cdot 10^5$ time steps. The lattice harbors a population of genomes, where a genome is a linear chromosome of genes with binding sites. From a genome a Boolean threshold network is built. During each time step the network may update the expression level of the genes for 11 propagation steps. B. The impact of several gene and binding site mutations is shown. The change in the genome and network topology is signaled by a red star. In a typical simulation the parameters are (per gene, binding site): gene duplication (dup) $2 \cdot 10^{-4}$, deletion $3 \cdot 10^{-4}$, threshold $5 \cdot 10^{-6}$, binding site (bsite) duplication $2 \cdot 10^{-5}$, innovation $1 \cdot 10^{-5}$, deletion (del) $3 \cdot 10^{-5}$, preference (pref) $2 \cdot 10^{-5}$ and weight $2 \cdot 10^{-5}$. See Methods for an explanation on each type of mutation. C. Typically the environment changes over time with a probability of $\lambda = 3 \cdot 10^{-4}$. The two evolutionary targets A and B determine which genes should be expressed (*on*) or inhibited (*off*). The result is four categories of genes; some should be always on, some should toggle their expression state and some should never be expressed. In a typical simulation, the target expression states are, from gene 0 to 19, A: `00011 11000 00000 11111` and B: `11010 01001 01100 01011`.

by the population). That is to say at the start adaptation was slow and unsuccessful, though eventually the population evolved a swift mode of switching correctly between attractors. As can be observed in Figure 3.2B, when the environment switched, the population had to change the expression state of nine genes, which caused the mean Hamming distance to jump to 9, while within the resolution of 10 time steps the minimum already jumped back to 1. Hence the best individuals had virtually immediately activated and/or inhibited eight genes via a mutational adaptation: a clear sign of the evolution of evolvability.

To assess the improvement we calculated for all runs the time differences between consecutive Hamming distances to the evolutionary target (Figure 3.3A). Sustained gains in the speed of adaptation were observed until $t \approx 1.2 \cdot 10^5$ for reaching at least a distance ≤ 4, and until $t \approx 2 \cdot 10^5$ for a distance ≤ 1. Ad-

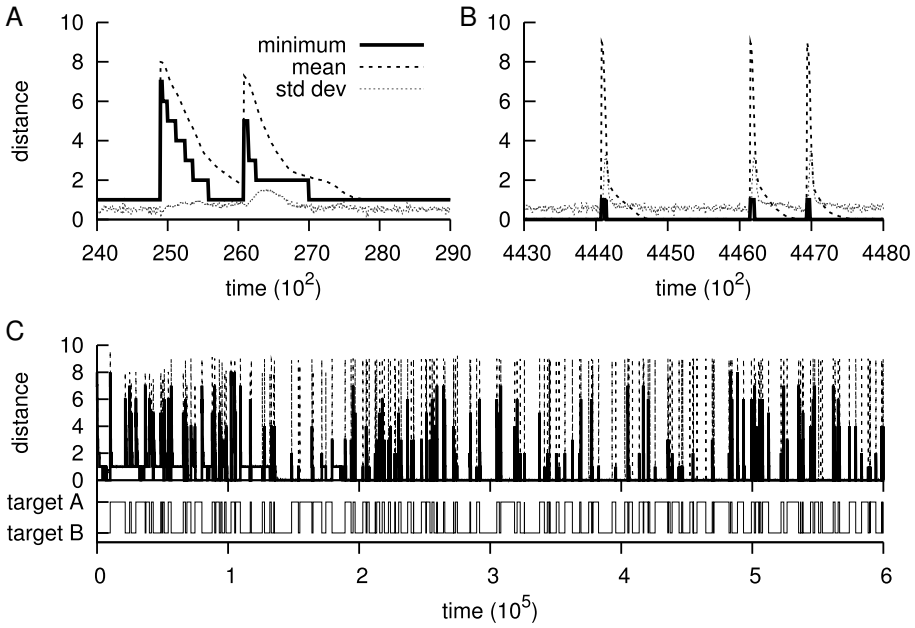

Figure 3.2 – A typical run. A, B. Close-up of the population dynamics. The population is minimizing the Hamming distance to the evolutionary target. For two intervals (one at the start of the run, the other to the end), the minimum and mean population distance with the standard deviation are plotted at a resolution of 10 time steps. C. An overview of the entire run. The top panel shows the population minimum and mean distances as in figure A and B, while the bottom panel shows the random timing of alternations between the evolutionary targets. There were 191 switches between the two targets.

ditionally, as a signature of the global evolutionary dynamics, we have taken the median of the population median distances (Figure 3.3B). In agreement with the time differences shown in Figure 3.3A, until $t \approx 2 \cdot 10^5$ the populations improved their ability of simply reaching the evolutionary targets, followed by a long transient of slowly decreasing population median distances.

Concluding, after an initial phase, the population had evolved a ten-fold improvement in its mutational speed of alternating network states. In other words, the individuals had arrived at a genotype-phenotype mapping that allowed for rapid and accurate switching. We now turn our focus to the long-term evolutionary dynamics.

3.2.2 An evolutionary sensor

As a proxy for the genome content of individuals we measured the average copy number of genes in the population (Figure 3.4). Besides the drift of gene 18, two remarkable periods were observed in this run. In *Period I* the copy number of

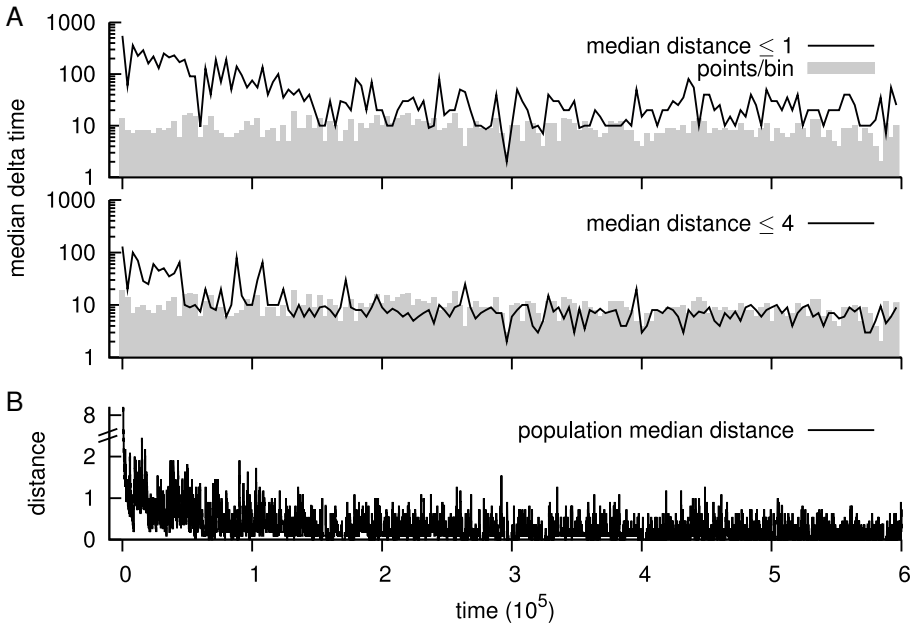

Figure 3.3 – Evolving to the targets. A. The median time differences (delta time) of the 11 runs are plotted. After an environmental change we recorded the time to reach at least the mentioned Hamming distance (1 and 4) from a previous higher one at a resolution of 10 time steps. The top panel shows the median time to almost reach the evolutionary target, while the bottom panel shows the median time to get at least halfway between the two evolutionary targets. Due to the random timing of environmental changes in the runs, we binned the time differences and show the number of points per bin in the background of both figures (bin size $= 4 \cdot 10^3$ time steps). Note the logarithmic scale of the ordinate. B The median of the population median Hamming distances for the 11 runs. Due to the random environmental changes of all the different runs, the populations appear not to reach the evolutionary targets. However, as shown in Figure 3.2, in a single run it is clearly visible that the populations do so.

gene 3 alternated between 1 and 2 as the environment switched 44 times back and forth between the two attractors. The behavior was lost around $t = 4.7 \cdot 10^5$, and gene 6 quickly took over the same behavior of switching copy numbers. The remaining part of the simulation was marked by *Period II* and contained 38 alternations of the evolutionary target. Immediately the hypothesis arose that these genes were responsible for the multiple events of extremely rapid evolutionary adaptation.

In order to verify the validity of the hypothesis a detailed picture of the evolutionary dynamics was needed. Therefore we closely examined the evolutionary process by performing an ancestor trace (see Methods). Due to adaptive mutants sweeping the population after each environmental change, the entire population had a recent single common ancestor, and hence by looking at a single ancestor

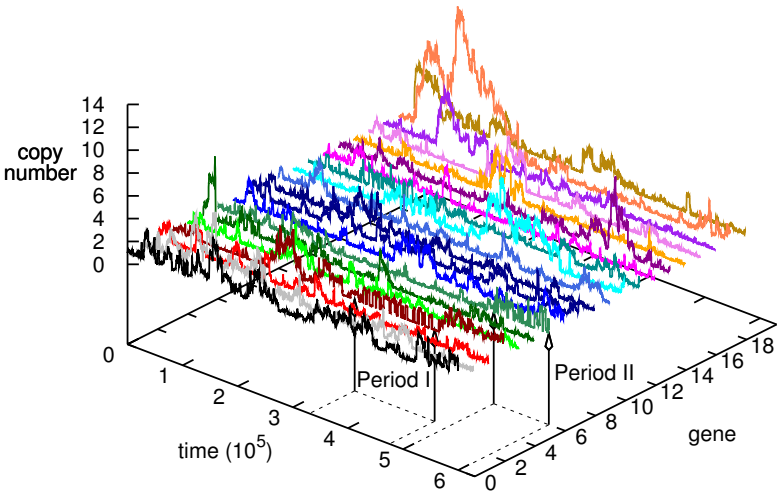

Figure 3.4 – The change in copy number for each gene. For each of the 20 different gene types the average copy number in the population is plotted through time (see Methods for an elaboration on the concept of gene types). Two intervals are highlighted: *Period I* stretches from $t = 3.4 \cdot 10^5$ to $4.7 \cdot 10^5$, where gene 3 shows switching behavior and *Period II* from $t = 5 \cdot 10^5$ to $6 \cdot 10^5$ where gene 6 takes over as evolutionary sensor. See also Figure S3.1 and Figure S3.2 for other runs with ESs.

trace, we essentially looked at the one lineage that has survived from the start. This complementary analysis allowed us to characterize genes 3 and 6 in much more detail: What their impact was on the adaptation, how networks utilized them to switch the gene expression state and how they altered the local mutational landscape.

Adapting with an evolutionary sensor

Figure 3.5 shows the cumulative fitness gain over time (A) and final fitness gain (B) of mutational events. We could see that the involvement of genes 3 and 6 in adapting to the environment was unmistakable. Initially the networks in the ancestor trace appeared to avoid gene mutations and from t $\approx 1 \cdot 10^5$ to $3.4 \cdot 10^5$ a variety of genes was used. Clearly in both *Period I* and *Period II* the evolutionary sensor genes accounted for the large majority of adaptive mutations (Figure 3.5A).

In Figure 3.5B we see that a few genes played pivotal roles in this run, while binding site mutations had only a minor effect on the dynamics. The inset of Figure 3.5B shows that both gene 3 and 6 also had a high effect per mutation. Interestingly, from Figure 3.5A we found gene 3 to have been a rather 'dominant' gene throughout the run, while gene 6 was active solely during the last part. This

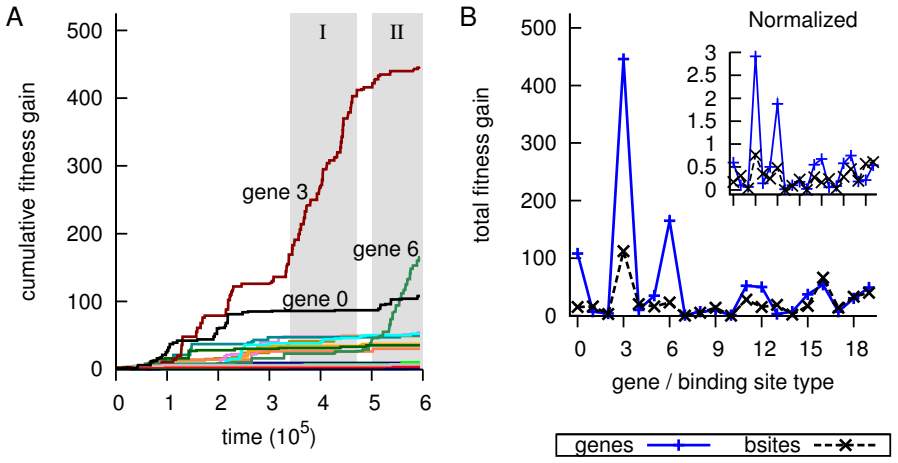

Figure 3.5 – Fitness gained per gene or binding site type. From an ancestor trace the adaptive mutations were categorized by gene or binding site identification tag. A. The cumulative fitness gain of genes is shown through time. Highlighted are *Period I* and *Period II*. B. The total fitness gain of gene and binding site mutations is shown. The *inset* figure shows the contributions normalized by the number of mutations of each type. Note that the peak at binding site 3 is a fluke; during the transition from *Period I* to *Period II* binding sites with a preference for gene 3 were involved in a few very effective mutations (data not shown).

implies that before gene 6 became a sensor its mutations were mostly deleterious and hence were hardly encountered in the ancestor trace.

With respect to our initial hypothesis: indeed gene 3 and 6 played a central role in the long-term evolutionary adaptation through their copy number alterations and accompanying fitness gains. However, the observations in the above paragraph indicate that the question why specifically these genes became sensors is not trivial.

Case study at time = 457755

As an example of how a single mutation pushed a network from one evolutionary target to the other, we selected two consecutive individuals from the ancestor trace. As shown in Figure 3.6 they differed by a deletion of gene 3.

Both genes 3 had been active in the parent network, thus the deletion caused a reduced sum of inputs at the target genes $1, 5, 9, 11, 12, 14$. Of the 31 genes 13 were targets of gene 3, yet most had compensatory input from other genes and as a result only genes of type 11 were affected. As gene 5 was still active, they were no longer sufficiently repressed and therefore the dosage effect of lacking one gene 3 activated 11.

It followed that genes 11 activated gene 0 and turned off gene 4, 15 and 17, which in turn silenced gene 5. Furthermore, once gene 17 was off, 11 turned

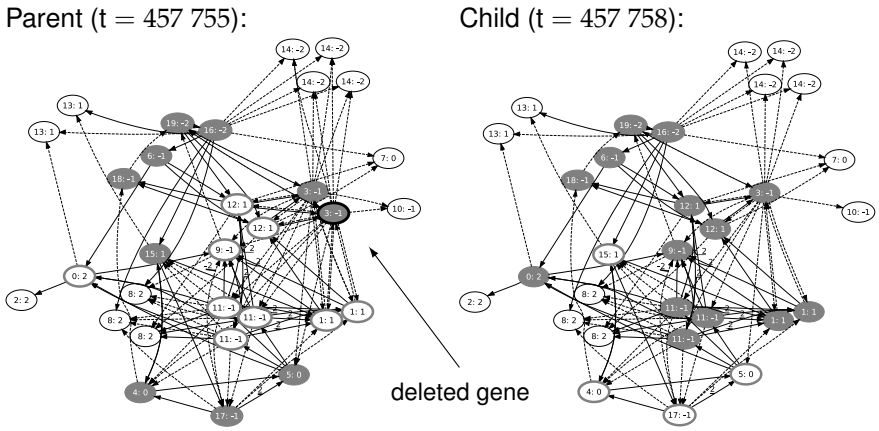

Figure 3.6 – A network switching attractor. Genes are represented by nodes, labeled as (identification tag: expression threshold), colored grey if active, or a thick grey outline if they are active in the opposite attractor. Activating interactions are solid edges, inhibiting ones are dashed. The deletion of gene 3 silences genes 4, 5, 15 and 17 and activates 0, 1, 9, 11 and 12, which is exactly a switch between the two evolutionary targets. As we study a long-term process, the networks are representative of the 'final solution' that the population evolved to (a switch guided by gene 6 is shown in Figure S3.4). For visibility, both networks were pruned for interactions originating from genes that are always silent and for parallel interactions that cancel out.

on 1 and 9. Interestingly the two copies of gene 12 were activated via different pathways. One is brought to expression via genes 11, the other via genes 1. Thus within four propagation steps of the network, the individual changed the expression of nine genes and ended in a different attractor.

3.2.3 The local mutational landscape

We have shown that evolvability can evolve in a changing environment and by what kind of mutations this process takes place. Now we study the changes in the local mutational landscape of the individuals as a direct measure of the evolved genotype-phenotype mapping.

All one-point mutants were generated for each individual in the ancestor lineage and, after allowing the mutant networks to settle in a new (stable) gene expression state, we computed the Hamming distance of the mutant network to the opposite evolutionary target. Next, we subtracted the mutant's distance from the individual's distance to the opposite target, which gave us the improvement of the mutant in Hamming-distance units. By opposite target is meant the network state that the individual did *not* evolve to at the time of birth. Since the population was for a large part of the simulation close to either target, we looked for signs of evolvability in the local mutational neighborhood of these individuals; how well they were able to switch to the opposite target.

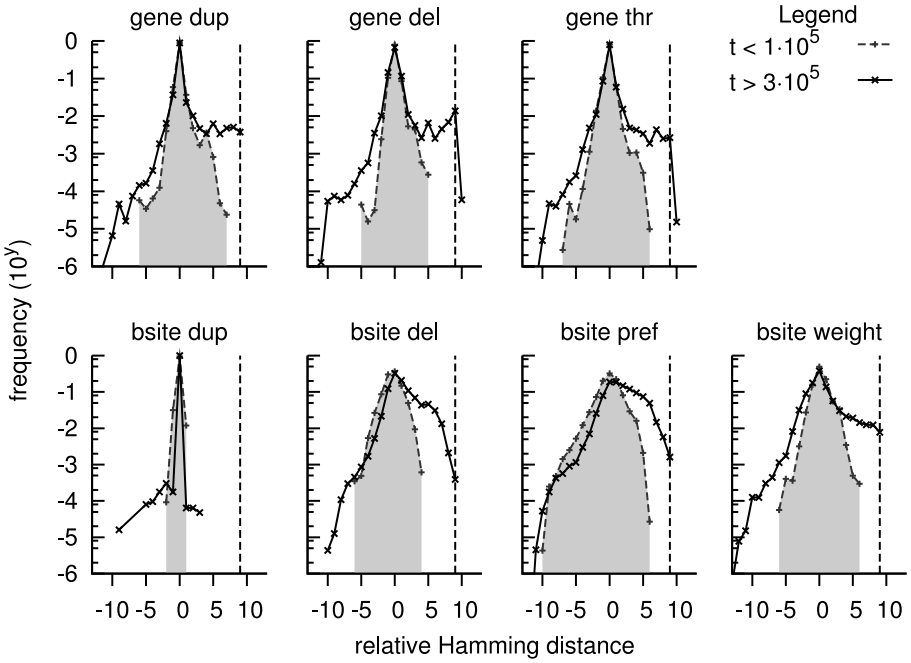

Figure 3.7 – Hamming distance improvement to the opposite evolutionary target. Each sub-figure gives for a mutational operator the frequency plot of the Hamming distance of a mutant compared to its ancestor with respect to the opposite evolutionary target. Positive distances signal that mutants are closer to the target, a distance of 0 is a neutral mutant and negative distances indicate mutants are farther from the target. The evolutionary targets have a distance 9 from each other, indicated by the vertical dotted line. The grey dotted line shows a frequency plot integrated over all ancestor trace individuals until $t = 1 \cdot 10^5$, the black solid line integrates from $t = 3 \cdot 10^5$ to $6 \cdot 10^5$. Note that the ordinate is in log-scale and that binding site innovations have been grouped with duplications. (*dup* duplication, *del* deletion, *thr* threshold, *bsite* binding site) and *pref* binding preference.

In Figure 3.7 we observe that for every type of mutation the mutants initially ($t < 1 \cdot 10^5$) peaked at distance 0 from the ancestor and they were approximately symmetrically distributed around this peak. Thus the vast majority of the one-point mutations was neutral, and few mutations allowed the individuals to change their gene expression either towards or from the opposite target. In the second half of the simulation ($t > 3 \cdot 10^5$), where the sensors dominated the evolutionary adaptation, the distribution of mutants was strikingly different. Still the majority of mutant networks ended in the ancestor's state, indicating a maintained mutational robustness against gene expression changes. But in contrast to the initial variety of mutants, the ones that were adaptive for the opposite target were now overrepresented. Except for binding site duplications, all types of mutations showed a difference of several magnitudes for mutating

towards the opposite target. Especially the gene mutations (duplication, deletion and threshold changes) and binding site weight changes were capable of generating adaptive offspring close to the opposite target. The evolutionary change in the effect of binding site deletions and binding preference was less focused. They became less likely to mutate away from the environmental target and more likely to mutate (a bit) towards the target. In other words they influenced only a few genes, but with a high probability of improving in the direction of the target. These observations suggest that genes performed the large mutations, while binding site mutations resulted mainly in small adaptations (which is nicely in concordance with Figure 3.5B).

Thus it appears that the evolution of the genotype-phenotype mapping maintained mutational neutrality that is inherently present in the mapping, while it increased the number of one-point mutants near the opposite evolutionary target.

3.2.4 Mutations and neutrality

The first trial simulations we ran on a grid of 100x50. These rather small populations reached both evolutionary targets and showed a tendency to develop evolvability, but were never able to keep it for more than a few environmental switches. Neutral mutations were accumulating in the population (data not shown), and the (secondary) evolutionary process of creating an ES was faced with its own Muller's ratchet.

Subsequently we enlarged the lattice, which lead to the presented results. We recorded the mutations both on the level of the population and the ancestor trace and categorized them by the direct fitness effect. Naturally, the individuals in the ancestor trace (in short 'ancestors') received a magnitude more beneficial mutations than the average individual in the population as shown in Figure 3.8. Nevertheless the ancestors also appeared to have had rather many deleterious mutations. The majority of these mutations had their effect altered during the lifetime of an individual: an environmental switch of the evolutionary target turned the mutation into a beneficial one. We found that 70/106 deleterious mutations were in fact advantageous.

Most interestingly the majority of mutations was still neutral, as we also observed in Figure 3.7. By comparing in Figure 3.8 the top panels with the bottom ones, we observed that both the ancestors and the average individual from the population had fixed a similar number of neutral mutations in their genomes. As shown in Figure 3.9 there was a constant rate, almost clock-like, of acceptance of neutral mutations. From this we may draw two conclusions: (a) the population was drifting on a neutral network of network topologies (Fontana *et al.*, 1993b; Ciliberti *et al.*, 2007a,b) and (b) even though the networks achieved greater evolvability, the neutrality of their mutational neighborhood was largely maintained.

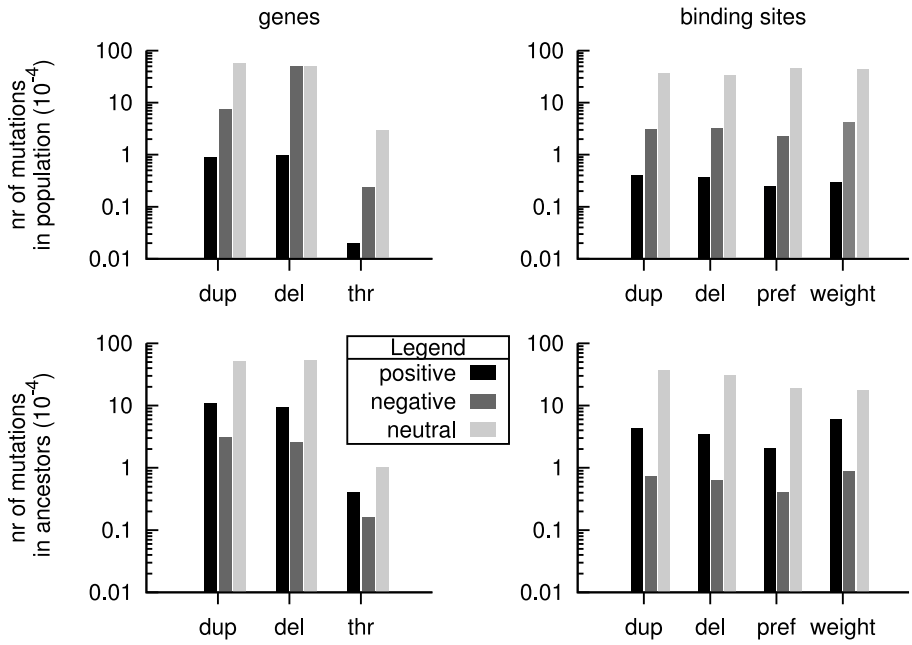

Figure 3.8 – Mutations categorized by their immediate fitness effect. The mutations are: duplication (dup), deletion (del), threshold (thr), binding preference (pref) and weight. Per type three categories are distinguished: positive effect (gaining fitness), negative effect and neutral. Each bar is an average over the entire run per individual in the population (top panels) and per individual from the ancestor trace (bottom panels). In the population $4.10 \cdot 10^9$ individuals were born and of them $1.26 \cdot 10^5$ belonged to the ancestor lineage. The total number of events is comparable for genes and binding sites, in the ancestor it is respectively 1660 and 1545. Note that the ordinate is in log-scale, that binding site innovations have been grouped with duplications and that the large number of deleterious gene deletions in the population (compared to the binding sites as well as the ancestor's gene mutations) is explained by the lethality of missing a gene type.

3.2.5 Properties of an evolutionary sensor

One way to rephrase the evolution of the genotype-phenotype mapping is to say it is the evolution of the network topology. The identification of the ESs was a dynamic characterization of how the network was shaped and altered. Thus it is interesting to study the properties of these sensor genes with respect to the network topology. In order to provide a more general and cohesive picture, we characterized several properties of the ESs in all the 11 runs in which these evolved.

First of all, in the typical run gene 3 and 6 needed to be expressed in both environmental targets, which enabled the networks to create a dosage effect that changed gene expressions. As shown in the list of runs in Figure 3.10E such con-

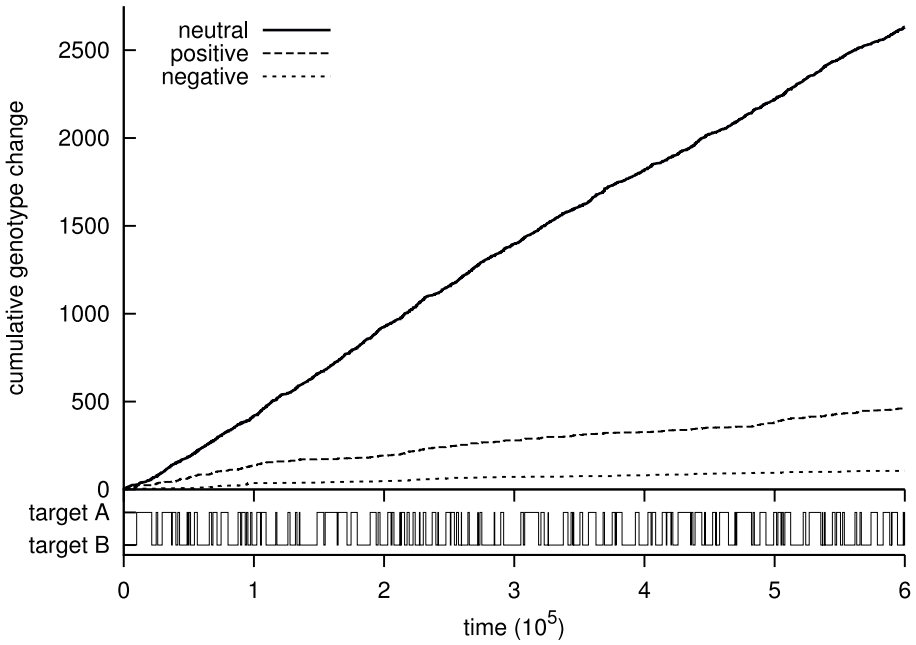

Figure 3.9 – Traveling on the neutral network. The top panel shows the accumulated genotype changes categorized by their effect on fitness (positive, negative and neutral). The bottom panel is a reminder of the switching of the environment between the two targets.

stitutively expressed genes (genes 3, 6, 16, 18 and 19) are clearly overrepresented among the sensor genes. So genes that should be always *on* were favored by evolution.

Next, if we consider that the evolutionary process must structure the raw material of binding sites and genes in order to produce the correct expression state of the genes, the layer in-between the genotype and phenotype, i.e. the network topology, should provide key insights into the evolvability we observe. Therefore we related two important characteristics of nodes in a network, the outdegree and indegree, to the already known 'dynamic' property of the gene copy number change. The latter we looked into indirectly in the section "Adapting with an Evolutionary Sensor", where we found a positive association of ESs with gene copy number change. That is to say, the more a gene alters its copy number, the more likely it is to be an evolutionary sensor.

First of all, for each run we visually identified ESs by their copy number changes in graphs like Figure 3.4. Then we selected from the simulations two intervals. For the long term evolutionary dynamics we took the second half of the runs ($t > 3 \cdot 10^5$), and as a reference we picked the initial period until $t = 1 \cdot 10^5$. For a gene to influence the state of the network, it needs an outdegree. Thus we

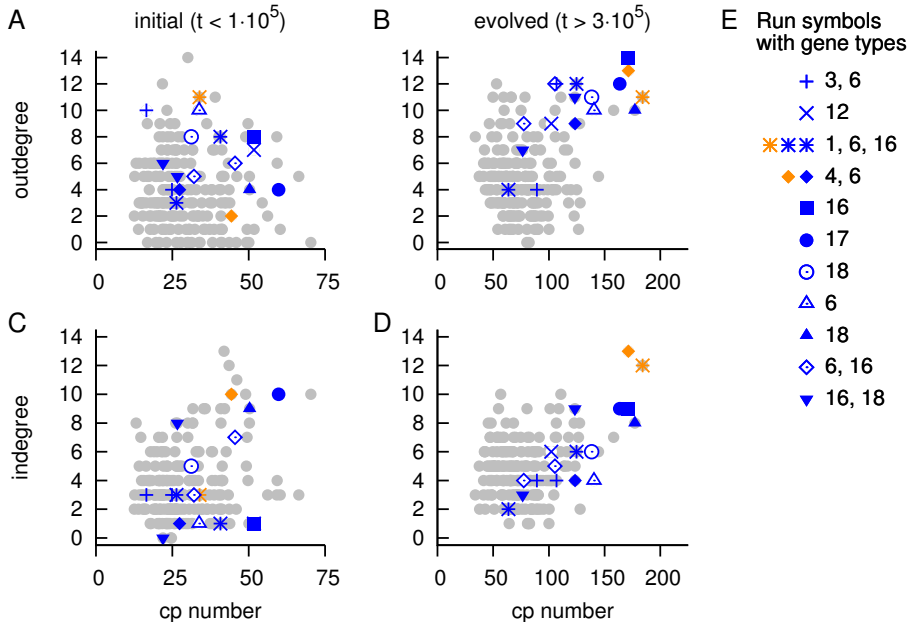

Figure 3.10 – Scatter plots of gene properties. All data points are population averages per gene. See Methods for details. Plotted are the initial (A, C) and evolved distributions (B, D) of accumulated copy number change against outdegrees (A, B) and indegrees (C, D) of each gene. The accumulated copy number change (cp number) is a measure for how often a gene is duplicated or deleted in the entire population, thus showing fixation of such mutations in the population (i.e. indicating it may have been adaptive). The outdegree and indegree are topological properties of genes in a gene regulatory network indicating respectively how many genes they influence and by how many they are influenced. In each subfigure genes that we identified as ESs are shown in *blue*, while to indicate hub genes receiving input from ESs *orange* is used. E. A list of the runs. For each run a different symbol is used, with the gene types of the involved sensors/hubs. Constitutively expressed genes are 3, 6, 16, 18, 19.

first studied the change in outdegree of the ESs. As shown in Figure 3.10A initially their copy number change and outdegree were uniformly distributed and one could not distinguish ESs from other genes. This strongly contrasts to Figure 3.10B, where the majority of ESs had evolved to a high copy number change and large outdegree. We still observed four ESs with a combined low copy number change and outdegree. This is explained by the fact that their ES behavior was observed only for a short period of time. For instance, gene 6, which we discussed previously, is among these genes. The secondary evolutionary process of creating ESs had been acting for too short a time to distinguish these genes from the rest.

Two runs showed a different strategy. By sampling networks through time in the third run (third from the list in Figure 3.10E) we established that gene 6 and

16 had been providing input to gene 1, which is a hub gene. The network state change involved either copying an ES or a state-switching hub gene. Thus the network was controlled by two hub genes, one of which was an ES. A similar scenario holds for the other case (fourth run from top in Figure 3.10E) (data not shown).

Secondly we studied the indegree of genes (Figure 3.10C and D). In the beginning copy number and indegree were uniformly distributed, as was the case for the outdegree. However, unlike the outdegree, we did not observe any clear long-term evolutionary effects on the indegree. In other words, there was only a selective increase in outdegree. Moreover the fact that such a signature of the network topology is still visible after averaging over populations and over the chosen time intervals is astonishing and shows that the result is robust.

Of the 15 runs, we already mentioned four did not show any signature of ESs. These runs also had no genes which evolved towards high outdegree (data not shown). Thus the topological characterization provides a general procedure for discriminating among runs with and without ESs, and for identifying genes as evolutionary sensors.

3.2.6 A general strategy

We expanded the scope of the problem by introducing a third evolutionary target. Instead of a "simple" toggling between two attractors, evolution needed to generalize the process of duplicating and deleting genes in order to change the network state. Although observing a clear duplication and deletion pattern for an evolutionary sensor as in Figure 3.4 was hard, the scatter plot of outdegree and indegree against copy number change showed evolutionary sensors (Figure S3.3). That is to say, the genes most likely to be an evolutionary sensor, i.e. genes which should always be expressed, showed ES behavior with respect to outdegree and indegree. From this we conclude that an 'evolutionary sensor' strategy has been applied in this extended case as well.

Finally, the results we have presented in this work were cross-checked against a variety of 8 initial networks and different evolutionary targets with qualitatively equivalent outcomes. The ancestor trace analysis was checked against ancestor traces of two other randomly selected runs with a sensor, again with qualitatively the same results. Additionally, a broad range of mutation rates and environmental change rates resulted in networks with evolutionary sensors (Supplementary Text).

3.3 Discussion

The accepted framework in which evolution operates is that mutations are random events and selection acts on the generated variation. Nonetheless, even if we assume mutations are random, their phenotypic effect may be strongly biased. In our simple model of GRN evolution we recognized that only a specific

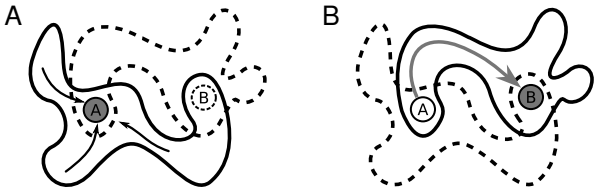

Figure 3.11 – The local attractor landscape around target *A* and *B*. A. The network is in attractor *A*, and its basin of attraction is shown by the black arrows and the solid-outlined 'cloud' around them. B. Attractor *B* has come into existence, forcing the network state of attractor *A* to propagate through the basin of attraction into *B*, as shown by the thick grey arrow.

subset of mutations was selected for. The networks had become sensitive to the indel mutations of a particular gene, the evolutionary sensor. That is to say, the genotype-phenotype mapping from genome to network had evolved such that a small class of mutations was adaptive and therefore repeatedly observed. This demonstrates a clear example of mutational priming and hence of evolution of evolvability (Hogeweg, 2005).

Previously we have studied a similar process at the level of the genome, where a genome composed of genes and transposons structured itself in a manner that favored mutations at specific locations on the chromosome (see chapter 2). In our current model we lack the transposons and consequently mutations are not biased to a location on the genome. Instead the networks have been shaped by evolution to allow for swift adaptation to different environments. Over time the evolutionary sensor genes became hubs of the regulatory networks. In contrast to the genome being shaped by evolution, there has been a structuring of the evolutionary substrate on a higher hierarchical level, the network topology.

3.3.1 Attractor landscapes

The case of evolvability that we presented is elegantly explained in terms of the attractor landscape and its basins of attraction. A conceptual representation is shown in Figure 3.11, where in the left panel the network is in target attractor A. Attractor B need not exist at this moment, but due to a gene duplication or deletion of the evolutionary sensor, it will be created in tandem with the destruction of attractor A. In this manner the network state is suddenly in the basin of attraction of target B and the network ends in the correct state, target B.

A priori we did not anticipate such a dynamic attractor landscape. Recent work on the evolvability of networks had focused mainly on conserving old attractors, while gaining new ones Aldana *et al.* (2007), or keeping gene expression patterns, while altering the interactions Ciliberti *et al.* (2007a,b). Our work complements these, as we show networks are also capable of (re)generating 'known' attractors that are not necessarily present in the current attractor landscape. Remarkably, the networks have evolved toward a topology that allows them to

establish and move to a new attractor and to do so in a reversible fashion.

3.3.2 Discussing the model

With respect to our modeling formalism we would like to highlight two assumptions. First of all, we let binding sites determine via their weight whether the effect of a transcription factor (TF) is activating or inhibiting. One could argue this should be a shared decision or perhaps rather that binding sites should be passive, as TFs are generally regarded as being either only activating or inhibitory and not a mixture of the two. On the other hand the yeast cell cycle regulation does show that TFs can have opposite effects on different genes (Li *et al.*, 2004). Because we only model the TFs of a gene regulatory network, it is not unrealistic to allow for a more delicate tuning among them.

As mentioned above, the genomes and networks constitute the transcriptional core of a complete gene regulatory network. Hence an evolutionary target defines which combination of transcription factors is required to activate the correct (but not explicitly present) target genes. A few exploratory simulations with explicit target genes suggest the more 'realistic' case creates an easier task for evolution. One plausible explanation is that not predefining the exact wiring of which TF controls which target genes provides extra flexibility to the mutation-selection process. In that case, we have presented here a worse-case scenario of GRN evolution.

3.3.3 Generalizing the model

In gene expression dynamics noise is an important player (especially if one looks into the details). It is known that evolution may even exploit these random fluctuations, for instance to switch between distinct gene expression patterns (Kashiwagi *et al.*, 2006). We explored how evolvability was affected by noisy gene expression in our model. With a certain probability a gene toggled its state during the lifetime of an individual. It is important to realize that completely activating or silencing a gene is a strong type of noise, however evolution of evolvability was still observed. In three runs with high amounts of noise ($p = 0.04$, which translates to more than one gene affected per network propagation step) we identified evolvability in all cases and an evolutionary sensor in one of them. Thus the model appears to be able to cope with expression noise.

Secondly, instead of copying the expression state of a gene when it is duplicated, we initially silenced the new copy both as a biologically more sound setting and to test the resilience of the evolutionary process in discovering the evolutionary sensor. The latter relates to the observation that copying a gene basically results in a dosage effect of the sensor gene that then percolates the network. We still observed evolution of evolvability and the accompanying sensors. The straightforward solution was to have a negative gene expression threshold for the evolutionary sensor. As long as the sensor gene received more activating

than inhibitory inputs, the copy would start expressing with a delay of one time step compared to the original setting.

The last alteration we performed involved the copying of the gene expression pattern at reproduction. If one imagines a cell splitting into two daughter cells, it sounds reasonable to copy the state of the genes. However we ignored the cell cycle and therefore we introduced a birth state for each gene that mimics the starting point of the cell cycle. Again we found evolvability and the presence of a sensor gene.

3.3.4 From *in silico* to *in vivo*

In order to focus on evolution, we have imposed the restriction of no environmental sensor. This implies that the gene expression state cannot be pushed out of an attractor; if the network has to switch its state, the only option is to turn the current attractor into a transient state. Naturally, the question arises whether organisms actually perform gene expression changes via mutations. In bacteria it is known that specialized DNA recombination events, such as DNA inversions (Dworkin & Blaser, 1997), replication slippage in combination with methylation patterning (Srikhanta *et al.*, 2005) and other rearrangement events (Lysnyansky *et al.*, 1996) underlie phenotypic switching. Most importantly, these occur without prior environmental signals. This allows for a heterogeneous population that is resistant to sudden environmental changes (Dybvig, 1993). Whether such mutational mechanisms were involved in the evolutionary adaptation of the *S. cerevisiae* strains from which we have drawn our inspiration remains an open question. Still the more general idea of using, in a reversible manner, mutations to alter gene expression and consequently the phenotype seems to be a widespread mechanism.

3.3.5 Summary

By combining the time scales of evolution and gene expression with a dynamic environment, we have shown that networks become evolvable while their robustness to mutations is maintained. Mutational mechanisms to stochastically switch phenotype are applied by bacteria and single-celled eukaryotes, and we have demonstrated a scenario for the evolution of such survival mechanisms. In addition, our work provides a new search-image with respect to the effect of mutations on short-term evolutionary adaptation, which may be of importance in an upcoming field like synthetic biology.

3.4 Model

We study an individual-oriented model with a population on a lattice subjected to an environment that changes over time (see Figure 3.1). The simulation is initialized with a homogeneous population and usually run for $6 \cdot 10^5$ time steps,

during which the environment alternates between two evolutionary targets, that is two gene expression patterns, according to a Poisson process (usually $\lambda = 3 \cdot 10^{-4}$).

A lattice has been used for two reasons. Firstly, it enables a computationally efficient method for competition among individuals and, secondly, it is biologically sound, as organisms virtually always live in a spatial system with a certain degree of locality. Given a default lattice size of 150x50 and a fixed death rate for each individual of 0.1, the population size averages around 6750.

3.4.1 Individual with a genome

Each individual starts with a linear chromosome containing n different genes ($n = 20$) and on average 2 binding sites per gene. The network is derived from this chromosome with genes as nodes and interactions between genes defined by which gene binds to which binding site. The fitness of an individual is defined on the level of network states, i.e. which genes are activated or inhibited. Reproduction of an individual is based on this fitness. Note that we only model the transcription factors explicitly, and hence assume a certain combination of activated transcription factors would result in the correct activation and inhibition of 'phantom' target genes.

3.4.2 Network

At the start of the simulation all individuals have the same network. The network has been selected from a pool of randomly generated networks according to the following criteria: (1) the network is connected, (2) there are no parallel edges in the network, (3) the average Hamming distance between the attractors with a basin size > 10 is ≥ 5. The evolutionary targets have been chosen at random from the available attractors in the network, with a Hamming distance between them > 6.

Genes and binding sites The network consists of genes with interactions among them. A gene has a state of expression s (on $= 1$, off $= 0$), a threshold $\theta \in \{-2, -1, 0, 1, 2\}$ and an identification tag $t \in \{0, 1, 2, \ldots, n\}$. Binding sites specify which gene may bind to them via their own identification tag (i.e. if tags are equal), which is called the binding preference. They also determine the type of interaction w: *activation* ($w = 1$) or *inhibition* ($w = -1$). If there are multiple copies of a binding site present in the upstream region of a gene, there will be parallel edges in the resulting network. Symmetrically, if there are multiple copies of a gene, they all bind to a binding site.

Gene types The duplicates of a gene all have the same identification tag. As mentioned above they behave equivalently in terms of binding and, as described below, we map the expression states of all genes with the same tag to one state.

Therefore we introduce the concept of a gene type: a group of genes which all have the same identification tag.

We do not allow for new types, nor do we allow genes to change their identification tag. It creates a closed system of gene types that simplifies the definition of the evolutionary targets, i.e. the network states the population has to evolve to. Hence the copies of a gene may be viewed as constituting a family of transcription factors.

Throughout the text we use both gene and gene type for the collection of genes with the same identification tag, unless this would result in ambiguities. In similar fashion we group binding sites by their identification tag.

Updating the network On the network the gene expression dynamics are defined. Similar to classical Boolean networks, the genes in the network are updated in parallel. However, as it is a threshold network, for a gene i its state of expression s_i at time $t+1$ is defined as:

$$s_i^{t+1} = \begin{cases} 0 & \text{if } \sum_j w_{ij} s_j^t < \theta_i \\ s_i^t & \text{if } \sum_j w_{ij} s_j^t = \theta_i \\ 1 & \text{if } \sum_j w_{ij} s_j^t > \theta_i \end{cases}$$

This network approach has been successfully applied to the yeast cell-cycle network (Li *et al.*, 2004).

3.4.3 Fitness and reproduction

The fitness f of an individual is based on the Hamming distance D between the current state of its genes and the target network state as defined by the environment. If gene indel mutations have resulted in multiple copies of a gene (all of them have the same tag), that type of gene is regarded as *on* if at least one of the copies is on, and *off* if all copies are off. This is based on the fact that duplicated genes are usually capable of substituting for each other. Both missing a gene and not having any gene in the network activated are lethal. The Hamming distance is normalized and rewritten as a similarity measure. In formula the fitness is defined as:

$$f = \left(1 - \frac{D}{D_{max}}\right)^p$$

Selection pressure is increased by raising f to a power p, which increases the chance that a beneficial mutant spreads in the population. We fix $p = 10$, which closely resembles an exponential function in the range $[0, 1]$.

The fitness determines the probability of producing offspring r, if there is an empty grid cell in the neighborhood to place the offspring. Given such an empty

location, the eight neighboring individuals, called *nbh*, compete on basis of their fitness score f

$$r_i = \frac{f_i}{\max\left(\sum_{j \in \text{nbh}} f_j, \Theta\right)}$$

The threshold Θ (fixed at 0.4^p) creates the probability that if there are only a few individuals in *nbh* or these individuals are very unfit, nothing may happen. Given the relative fitness r_i of each individual in the neighborhood *nbh*, one is selected according to the fitness proportional selection scheme. Reproduction itself encompasses copying the chromosome, mutating and dividing into two daughters. The state of the genes is copied as well, in other words there is inheritance of the network state. Subsequently one of the two daughters replaces the parent, the other is placed in the empty grid cell.

3.4.4 Mutational events

While selection acts on the network, mutations act on the chromosomes. During reproduction the genome is duplicated, creating a diploid individual, after which mutations may occur on both chromosomes. We have defined the following events on genes:

- *duplication*: a gene with its binding sites is copied to a random location on one of the chromosomes. The expression state of the gene is copied as well,
- *deletion*: a gene with its binding sites is removed from the chromosome,
- *threshold change*: the current gene expression threshold is changed to a randomly chosen, valid, other value,

Binding sites have several types of mutations as well:

- *duplication*: a binding site is copied to the upstream region of a random gene in the genome. This introduces a new connection in the network,
- *deletion*: a binding site is deleted. If one or more genes bind to the deleted binding site, multiple connections are deleted in the resulting network,
- *innovation*: a new binding site is inserted in the upstream region of a random gene, with a random weight and a random binding preference,
- *weight change*: a binding site toggles from being activating to inhibiting or vice versa,
- *preference change*: a binding site changes its binding preference, in other words the gene type that binds to the binding site is changed. This may involve multiple connections being deleted and created in the network.

In our simulations we assume deletions should occur more often than duplications for two reasons. Firstly, the growth of a genome is bounded in this manner and, secondly, deleting a gene or binding site is regarded as an inherently 'easier' task than duplicating it.

3.4.5 Ancestor tracing

We trace lineages of individuals by attaching to them a unique identification and recording the parent-child relationships. The result is a "perfect fossil record". A single trace from one of the fittest individuals in the final population back to the initial population allows us to dissect the exact mutational dynamics, to calculate attractor state spaces of the networks and see their evolution. It also enables us to perform mutational experiments on each individual and to visualize the resulting mutational landscapes. In order to perform such ancestor tracings we only consider asexual reproduction in our model.

3.4.6 Analyzing indegree and outdegree

During a simulation, population averages of the gene copy number, indegree and outdegree are saved to disk categorized by gene type. At a resolution of 1000 time steps we get a good view of the general evolutionary trends.

As a measure for the amount of change in gene copy number, we sum, for each gene type, the absolute differences between adjacent sampling points. The more a gene fluctuated in copy number, the higher the sum. Hence evolutionary sensors have a tendency for high sums, as do genes that drift a lot. Both indegree and outdegree averages were binned in histograms, and from the resulting distributions the medians were taken as a representative number.

By using the above described procedure for the interval $[0, 1 \cdot 10^5)$ and $[3 \cdot 10^5, 6 \cdot 10^5)$ we constructed Figure 3.10. Evolutionary sensors were identified by hand using graphs as Figure 3.4 and then marked in the graphs of copy number against indegree and outdegree.

3.5 Supporting information

Parameter dependencies

In order to assess the dependencies of our results on the model parameters, we varied several key parameters such as the mutation rates and the rate of alternating the evolutionary targets.

Mutation rates

We increased and decreased all mutation rates by a factor of 2.5 with qualitatively equal results. If we raised the rates by a factor of 10, we observed evolution of evolvability, but no evolutionary sensors. Decreasing the rates further than 2.5 resulted in a population that adapted less to environmental changes. Due to the fact that the large majority of mutations is neutral, the proper advantageous mutation often just did not occur. Reducing the rates a factor of ten gave an averaging solution. The individuals tended to integrate over the two environments.

Next we investigated the ratio of gene to binding site mutation rates. In our set of 15 runs the number of events on genes was similar to these on binding sites. We increased binding sites rates 5 times and found that the influence of the evolutionary sensor diminished. In other words, the events on binding sites began to outweigh those on genes. In an extreme case we set binding site rates equal to gene mutation rates. The networks showed no topological adaptation and adapted via binding site mutations (data not shown).

Environmental rate of change

In analogy to the mutation rate survey we increased the evolutionary target switching rate. For simplicity we performed these simulations in a periodically changing environment rather than a Poisson one. If we started a run with an evolved population and a fast environment (period $= 1000$ time steps, compared to the typical run: $1/\lambda = 3333$), the evolvability was maintained. In fact, due to the fast switching, the indirect selection pressure on keeping evolvability was higher and the population showed less neutral drift. Interestingly it was possible to evolve evolvability in this fast setting. However often the process ended in a suboptimal solution. A gene that needed to switch expression from one target to the other was recruited as the evolutionary sensor.

Summary

If we summarize the collection of all performed simulations, we have $31/65$ simulations showing faster adaptations and an evolutionary sensor gene, $24/65$ show only faster adaptations, but not a clear signature of an evolutionary sensor and in $10/65$ runs the population failed to evolve evolvability. Note that we have included all parameter settings in these counts.

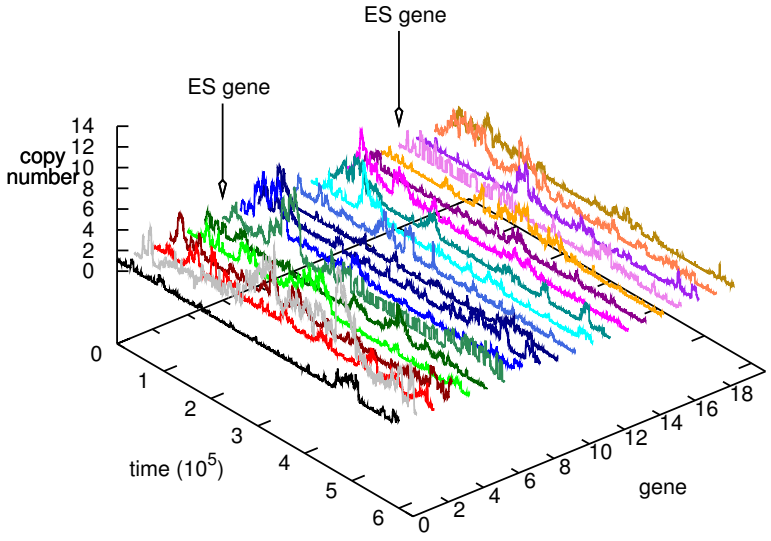

Figure S3.1 – For each of the 20 different types the average copy number in the population is plotted through time. There are two clear ESs visible in this run. From $t \approx 0.5 \cdot 10^5$ to $2.5 \cdot 10^5$ gene 16 is the ES, and from $t = 3 \cdot 10^5$ to $6 \cdot 10^5$ gene 6 is the sensor. With the exception of gene 1 most of the other genes do not show large fluctuations in the long term. This run is number ten in Figure 3.10E (first before last).

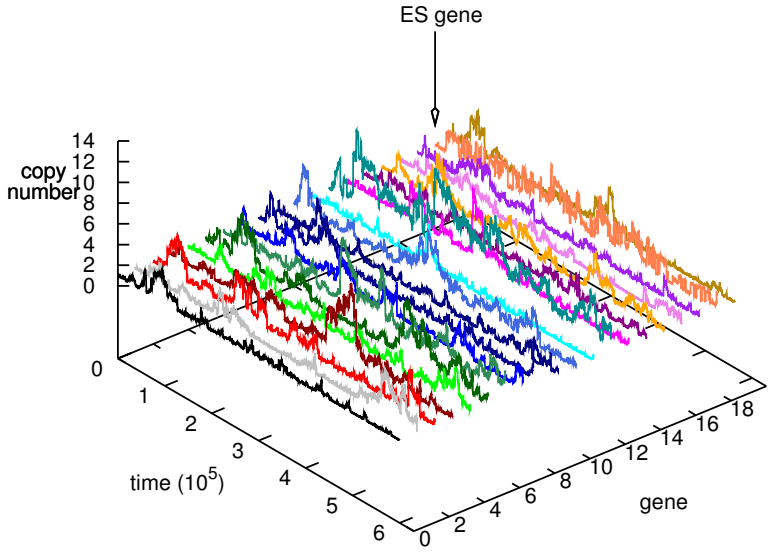

Figure S3.2 – For each of the 20 different types the average copy number in the population is plotted through time. In contrast to Figure S3.1, we observe more fluctuations in copy numbers. Gene 18 is the ES, and for most of the run its copy number alternates between 1 and 3, or 2 and 4, which results in a fuzzier signal. Still the gene is responsible for the adaptation. After $t \approx 5 \cdot 10^5$ gene 18 shows the clear-cut behavior of an ES as we have seen in Figure 3.4 and Figure S3.1. This run is number nine in Figure 3.10E (second before last).

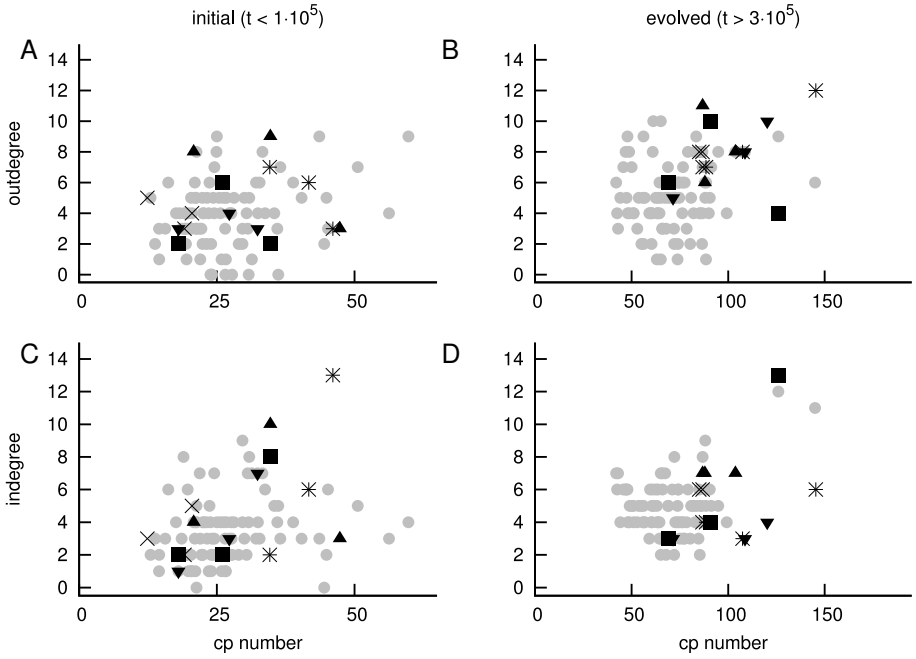

Figure S3.3 – Scatter plots of gene properties for the case of three evolutionary targets. All data points are population averages per gene type. See Methods for details. A–D. Plotted are the initial (A, C) and evolved distributions (B, D) of accumulated copy number change against outdegrees (A, B) and indegrees (C, D) of each gene. The accumulated copy number change (cp number) is a measure for how often a gene is duplicated or deleted in the entire population, thus showing fixation of such mutations in the population (i.e. indicating it may have been adaptive). The outdegree and indegree are topological properties of genes in a gene regulatory network indicating respectively how many genes they influence and by how many they are influenced. In each subfigure genes that are expressed in all three evolutionary targets are shown as *black* symbols and for each run different symbols are used. These genes are most likely to become evolutionary sensors, and indeed show such behavior.

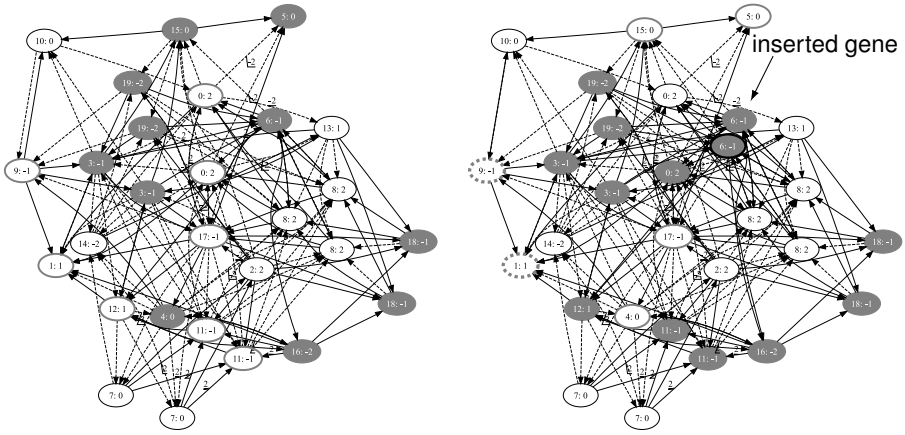

Figure S3.4 – A network switching attractor. Genes are represented by nodes, labeled as (identification tag: expression threshold), colored grey if active, or a thick grey outline if they are active in the opposite attractor. Activating interactions are solid edges, inhibiting ones are dashed. An insertion of gene 6 changed the expression state of 7 genes. The genes 1 and 9, depicted by a dashed grey ellipse, have not changed their expression yet, though they should be activated in order to obtain maximal fitness. Compared to Figure 3.6, the networks have gained interactions. For visibility, both networks were pruned for interactions originating from genes that are always silent and for parallel interactions that cancel out.

4

Evolution of Resource Cycling in Ecosystems and Individuals

Abstract

Resource cycling is a defining process in the maintenance of the biosphere. Microbial communities, ranging from simple to highly diverse, play a crucial role in this process. Yet the evolutionary adaptation and speciation of micro-organisms have rarely been studied in the context of resource cycling. In this study, our basic questions are how does a community evolve its resource usage and how are resource cycles partitioned?

We design a computational model in which a population of individuals evolves to take up nutrients and excrete waste. The waste of one individual is another's resource. Given a fixed amount of resources, this leads to resource cycles. We find that the shortest cycle dominates the ecological dynamics, and over evolutionary time its length is minimized. Initially a single lineage processes a long cycle of resources, later crossfeeding lineages arise. The evolutionary dynamics that follow are determined by the strength of indirect selection for resource cycling. We study indirect selection by changing the spatial setting and the strength of direct selection. If individuals are fixed at lattice sites or direct selection is low, indirect selection result in lineages that structure their local environment, leading to 'smart' individuals and stable patterns of resource dynamics. The individuals are good at cycling resources themselves and do this with a short cycle. On the other hand, if individuals randomly change position each time step, or direct selection is high, individuals are more prone to crossfeeding: an ecosystem based solution with turbulent resource dynamics, and individuals that are less capable of cycling resources themselves.

Concluding, in a baseline model of ecosystem evolution we demonstrate different eco-evolutionary trajectories of resource cycling. Varying the strength of indirect selection through the spatial setting and direct selection, the integration of information by the evolutionary process leads to qualitatively different results from individual smartness to cooperative community structures.

4.1 Introduction

Organisms influence their surroundings by taking up nutrients from the environment and excreting waste products in it. As Earth is a closed system with respect to its chemical components, this leads to resource cycles. Moreover, in doing so organisms may create a specific local environment for their offspring and competitors. From an ecological point of view these are rather basic observations, yet the overall consequences of such feedback between organisms and their abiotic environment on the evolution of a population, community and ecosystem are not well-studied.

One possible outcome of such organism-environment interaction is metabolic crossfeeding. Crossfeeding is an indirect interaction between two or more species, usually microorganisms: it is often observed as the dependence between bacterial strains on each others metabolites. Especially determining the ecological preconditions for such cooperative communities has received much attention. Both experimentally (Helling *et al.*, 1987; Rosenzweig *et al.*, 1994; Treves *et al.*, 1998) and theoretically (Porcher *et al.*, 2001; Pfeiffer & Bonhoeffer, 2004; Johnson & Wilke, 2004; Gudelj *et al.*, 2007) it has been shown that crossfeeding may evolve due to trade-offs in resource uptake and processing, but also simply through the excretion of secondary metabolites.

While crossfeeding and related experimental evolution studies (Rainey & Travisano, 1998; Kassen & Rainey, 2004; Tyerman *et al.*, 2005; Maharjan *et al.*, 2006; Philippe *et al.*, 2007; Spencer *et al.*, 2007, 2008) have been done mostly in 'artificial' well-mixed environments, the last few years metagenomics has been shedding light on the interplay between microbial communities and their 'natural' local environment (Raes & Bork, 2008). Such whole-ecosystem views show various eco-evolutionary solutions on different spatial and temporal scales to nutrient processing (Torsvik *et al.*, 2002; Frias-Lopez *et al.*, 2008; Strom, 2008): shallow phylogenetic divergence, yet large ecological divergence in the human intestine (Ley *et al.*, 2006), generalist bacterial lineages performing carbon processing (Mou *et al.*, 2008), but also temporal and spatial specialization through resource partitioning among *Vibrionaceae* strains in coastal waters (Hunt *et al.*, 2008). Closely related are analyses in evolutionary functional genomics. A striking example is the finding that many bacterial species contain only parts of the citric acid cycle, suggesting extensive metabolic cooperation among bacterial lineages (Huynen *et al.*, 1999).

In this work we use a computational modeling approach to gain a more general, qualitative insight in the spatial and temporal dynamics and mechanisms of evolving organism-environment interactions. Previous studies have shown the importance of an interplay between ecological and evolutionary processes. It plays a crucial role in the generation and turnover of ecological diversity (van der Laan & Hogeweg, 1995; Savill & Hogeweg, 1998). In addition, the spatial locality of ecological and evolutionary processes has been shown to strongly influence the outcome and dynamics of evolutionary processes (Savill *et al.*, 1997; Savill & Hogeweg, 1998; Pagie & Hogeweg, 2000b). Furthermore, evolving inter-

actions via resources has been shown to facilitate niche creation and selection on an ecosystem level. Stable genotypic and phenotypic diversity through resource partitioning was shown by Chow *et al.* (2004), while evolution of ecosystems as an example of multilevel selection was investigated by Williams & Lenton (2007, 2008). With respect to evolving interactions among individuals, it has been shown that this may lead to niche creation and ecological diversification (Lindgren, 1991; Ray, 1991; Takeuchi & Hogeweg, 2008).

Hence we include interlocking ecological and evolutionary processes and a spatial embedding of these processes. As we concentrate on the dual feedback between organisms and their (local) environment, we restrict interactions between individuals to competition for reproduction via nutrients. Furthermore, motivated by metagenomic studies showing the dominant role of microbes in the process of nutrient, or resource, (re)cycling (Falkowski *et al.*, 2008; Strom, 2008), resources can be altered according to a simple artificial chemistry that allows for cycling. Schematically speaking, individuals reproduce by taking up a resource, processing it with their gene regulatory network and excreting the resulting resource as waste. An environmental feedback is established and as a consequence of eating the environment is changed and future feeding opportunities in the neighborhood are affected. Importantly, a frustration arises as direct selection for resource processing can be antagonistic to the indirect selection for cycling resources. In this manner we have a simple evolving ecosystem, with the important feature that individuals determine how much fitness they derive from a resource and which waste product, that is new resource, they produce.

Note that in contrast to studies on crossfeeding (Pfeiffer & Bonhoeffer, 2004; Gudelj *et al.*, 2007), we abstract from reduction/oxidation and energy constraints in order to focus the analysis on the qualitative effects that organism-environment interactions have on the eco-evolutionary outcome. Also, with respect to ecological studies on food webs (Drossel *et al.*, 2001; Loeuille & Loreau, 2005; Guill & Drossel, 2008), our model leaves out any predatory or parasitic relationships between individuals (see also Discussion).

In this model, we study the effect of indirect selection. We do this by varying two parameters, namely the spatial setting and the strength of direct selection. Firstly, we compare local feedback to a null model that lacks the local feedback due to the mixing of individuals, yet still has recycling on a lattice-wide scale. Secondly, we study the relative balance between direct selection for processing resources against indirect selection for cycling.

We show that local feedback enhances indirect selection, as it allows individuals to shape their local surroundings. This results in evolutionary stagnation: resource distributions are more in equilibrium over long periods of time and resource cycling is slower than in the null model. Furthermore, local feedback shows a long-term trend for independent, 'smart' individuals. Individuals are adapted at cycling resources themselves, and do so with shorter cycles than individuals from the mixed model. As such, especially for relatively high indirect selection (i.e. low direct selection) we find single generalists dominating the population eventually. In contrast, the mixed model – with only a global

cycling of resources – displays more turbulent resource dynamics both qualitatively and quantatively, and a preferred evolution for cooperating, crossfeeding lineages. By shifting the balance of selection pressures by adjusting direct selection, we find similar changes in evolutionary behavior as for the two spatial settings. In both the local feedback and mixed model, low direct selection results in 'quiet' resource dynamics and an evolutionary trend for self-sufficient individuals. Also, if direct selection increases, resource dynamics become more turbulent and crossfeeding lineages become a favored behavior.

Thus strong indirect selection for resource cycling, both via local feedback and low direct selection, favors the evolution of self-sufficient individuals, while weak indirect selection, accomplished through only a global feedback and strong direct selection, lead to an ecosystem based resource cycling via crossfeeding lineages.

4.2 Methods

We describe our model from a high-level perspective first, followed by several sections covering the details. Central to the local and null model is the processing of resources: organisms have to evolve their regulatory network such that they gain energy from nutrients in the environment. As depicted in Figure 4.1B, a resource is a bit string, and as an abstraction of metabolic activity an individual has to reproduce the bit string as a temporal output pattern of its gene regulatory network. We named this a "bite". The example in Figure 4.1B shows a bite of 13 bits. Next, this bite determines both the fitness of an individual and what waste product is left in the environment: the bite is cut from the left of the resource and re-attached at the right side, effectively rotating the resource bit string.

Thus we combine a model of genes and binding sites on a genome, that are translated into a regulatory network, with a very abstract approach to metabolic processing: processing a resource through a gene expression pattern in time. Various chains of such resource processing steps reflect the different paths of nutrient conversions, such as occurring in the microbial nitrogen cycle. For example nitrogen can be cycled through few, large steps: $NO_3^- \rightarrow NH_4^+ \rightarrow NO_2^- \rightarrow NO_3^-$, or with intermediate metabolites: $NO_3^- \rightarrow NO_2^- \rightarrow N_2 \rightarrow NH_4^+ \rightarrow NO_2^- \rightarrow NO_3^-$.

The above described interaction between organism and environment is embedded on a two-dimensional grid (Figure 4.1A), where each grid site contains a single resource and at most one individual. Given a grid site without an individual, the organisms in the 8 neighboring sites compete for the opportunity to reproduce and place a daughter in the empty site. This competition is based on how well each individual can process the resource at the empty site: the longer an individual's bite, the larger is its probability to reproduce. Naturally, during reproduction mutations may occur (Figure 4.1C).

As a consequence, in the local model a lineage of fit individuals shapes the resource distribution in its vicinity. This causes an effect over several generations

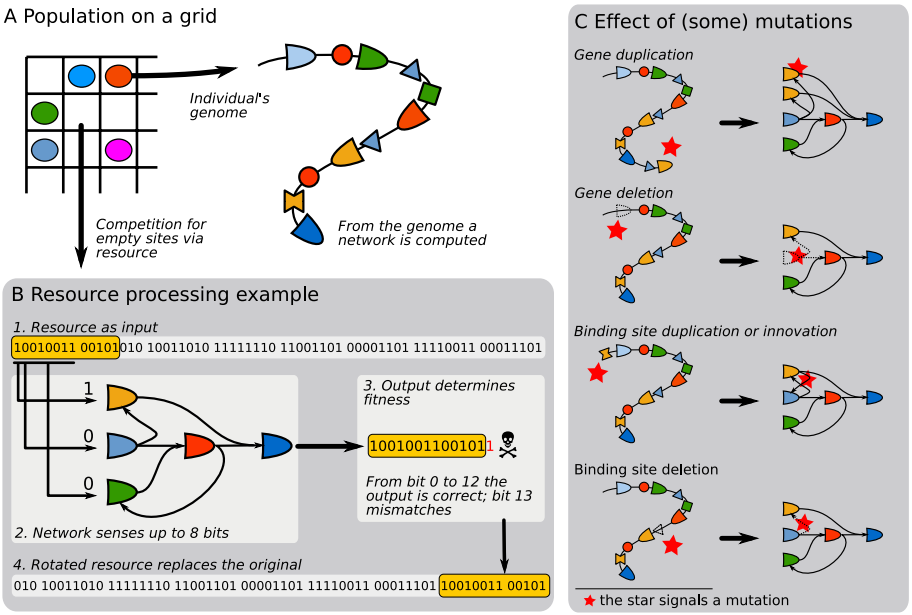

Figure 4.1 – Overview of the model. A. Individuals and resources are placed on a grid (size 100×100). Individuals consist of a genome, from which a network is computed. They compete for reproduction into empty grid sites by processing resources. B. The resource is a bit string of length 64. Maximally the first 8 bits can be sensed by a network, which then produces a sequence of bits at its output. The output is matched to the original bit string, and the length of the correct sequence (matching the bit string from the leftmost bit) is the raw score of the individual, which in this example is 13. If the individual reproduces, the resource is rotated from right to left for 13 bits and placed back in the grid site. C. The effect of a few types of mutation on the genome (left) and the topology of the network (right). By default the parameter values for each type of mutation are: gene duplication $16 \cdot 10^{-4}$, deletion $24 \cdot 10^{-4}$, binding specificity $4 \cdot 10^{-4}$, gene expression threshold $4 \cdot 10^{-4}$, binding site duplication $4 \cdot 10^{-4}$, deletion $10 \cdot 10^{-4}$, innovation $1 \cdot 10^{-4}$, binding specificity $4 \cdot 10^{-4}$, weight $4 \cdot 10^{-4}$. In order to balance the growth of the network, we apply a small penalty per gene and binding site of $pen = 2.5 \cdot 10^{-5}$.

with respect to what resources are available in the local environment and thus impacts the evolution of the local population or community. We compare such local change to a null model in which we remove this opportunity of shaping the local environment. In the null model we randomly relocate all individuals on the grid at the end of each time step. It leaves the process of biting and excreting waste in place (i.e. there is still a global feedback), but destroys any phylogenetic relationship between neighboring individuals – in effect individuals are not located close to their parent anymore – and removes the multi-generational effect of a lineage establishing itself in a patch and shaping its surroundings.

4.2.1 Model description

Grid By default the grid is initialized with an identical resource at each site and a homogeneous population of size 8000 (individuals are randomly placed, at most one per grid site). Both the initial resource and individual have been randomly generated in advance and are reused in replicate runs. In addition to the argumentation in the Introduction on the use of space, the grid allows for a computationally efficient method of competition among individuals and is biologically sound, as organisms virtually always live in a spatial system with a certain degree of locality.

With a standard grid size of 100×100 and a fixed death rate per individual of 0.1, the population size is on average 9000 and approximately 1000 new individuals are born each time step. Note that in the null model, the individuals are mixed each time step, which means each individual is moved to a random location on the grid.

Resources A resource is a bit string of length n (default $n = 64$) and, as explained above, can be altered into another resource by rotation from right to left. A minimal rotation of zero bits returns the original bit string, as does the maximal rotation of n bits. Thus the 'chemical universe' consists of n different bit strings, each a rotated version of another, with a total of $n^2 = 4096$ transitions or reactions between them.

We identify each resource by the number of rotation steps performed on the original bit string. Thus resource 5 is equal to the original resource rotated for five steps. As a reference, the input bit string in Figure 4.1B is the original resource used in the main results. Furthermore, resources diffuse by Margolus diffusion (Toffoli & Margolus, 1987), by default one diffusion step per time step.

Genotype and network The genotype of an individual is a single chromosome, on which genes with their binding sites are placed. From this chromosome a Boolean threshold network is computed that can process resources. To this end each gene and binding site have a binding specificity, or integer number. The network is constructed by connecting genes with binding sites if their binding specificity is equal. Thus all genes with specificity 3 bind to all binding sites with a specificity 3, possibly connecting multiple source genes with various target genes. We allow for self-loops and parallel connections.

There are three categories of genes: input genes (with a binding specificity in the range $[0, 7]$), processing genes (specificity in $[8, 14]$) and output genes (binding specificity 15, which is for identification purposes only, i.e. there are no binding sites with specificity 15). Not having any output genes is lethal, the other genes are all optional.

Resource processing by the network Resource processing is performed in two steps. First, the first 8 bits of the input string are assigned as starting states to the input

genes by binding specificity. The rest of the genes are set to zero. As an example: all genes 4 are assigned the state of the fourth bit of the bit string. Note that input genes may be deleted by mutations and the sensing of bits may thus be impaired.

Second, after a few calculation steps (default 2) the task of the network is to reproduce the bit string as a gene expression pattern through time on its output genes. The state of the output gene is read and matched to the bit string. If there are multiple copies of the output gene, the output is considered 'on' if at least one of the output genes is 'on', else it is 'off'. As soon as a mismatch is detected, the network updating is stopped. We refer to such a stretch of matching bits as a "bite". Both the calculation steps and the expression output steps are done by updating the Boolean network in parallel (Li *et al.*, 2004), see also chapter 3. For a gene this is defined as:

$$s_i^{t+1} = \begin{cases} 0 & \text{if } \sum_j w_{ij} s_j^t < \theta_i \\ s_i^t & \text{if } \sum_j w_{ij} s_j^t = \theta_i \\ 1 & \text{if } \sum_j w_{ij} s_j^t > \theta_i \end{cases}$$

with s_i^t the expression state of gene i at time t (s=0 is 'off' and 1 is 'on'), w_{ij} the weight (1: activating and -1: inhibiting) of binding site j, and θ_i the threshold of expression of gene i ($\theta \in \{-2, -1, 0, 1, 2\}$).

The network dynamics are performed in one simulation time step and the number of updates may vary from 2 to $n + 2$ depending on when a mismatch occurs.

Reproduction If no individual is present at a site, the neighbors compete for the empty site in order to reproduce and place a daughter at the empty site. The neighborhood *nbh* consists of the 8 adjacent sites, also known as a Moore neighborhood. The competition consists of each neighbor trial processing the resource of the empty site as described previously. The fitness of an individual, f, is defined as

$$f = e^{\sigma * \max(0.0,\, l - pen * g)} - 1.0$$

with l the length of bite, g the genome size (number of binding sites and genes), *pen* the genome size penalty coefficient, and σ the selection coefficient. From the fitness we calculate a relative fitness r_i of each individual i in the neighborhood: $r_i = f_i / \sum_{j \in \text{nbh}} f_j$. Next, one is selected according to the fitness proportional selection scheme. Reproduction consists of duplicating the chromosome, applying mutations and dividing into two daughters. Subsequently one of the two daughters replaces the parent, the other is placed in the empty grid cell. And, the rotated resource (waste product) of the winner replaces the resource at the empty site.

Mutations On the genes and binding sites mutational events have been defined as follows. Genes may (a) duplicate, which includes the accompanying binding sites, (b) be deleted, which also includes the binding sites of the gene, (c) have their expression threshold mutated to a random other threshold value, and (d) have their binding specificity changed to a random other value. Thus the effect of a change of binding specificity may be that a lost gene is rediscovered.

Binding sites may (a) duplicate, where a copy of a binding site is inserted in front of a random gene, (b) be deleted, (c) have their weight w changed from activating to inhibiting and vice versa, (d) have their binding specificity changed to a random other value. Default rates of gene and binding site mutations are mentioned in Figure 4.1.

4.2.2 Analysis of individuals

Phenotype For an individual we can calculate the bites it produces for each resource. This results in 64 bites, each in the range $[0, 64)$. Next, we store the bite lengths in a vector, and we define this vector as the phenotype of an individual. A vector index corresponds to the number of rotation steps from the original resource. The phenotype is a measure of overall performance, as an individual is evaluated on only few resource in its lifetime.

Network dynamics If networks are allowed to continue generating output bits after the first mismatch with the input resource, repeating patterns of zeros and/or ones are observed. In other words, if we view a network as a dynamical system, it initially moves through a transient and then settles in a fixed point or cyclic attractor. Per individual, we calculate the transients and attractors for all resources. Next we group individuals by these network dynamics as a measure of diversity in the population. In addition, such dynamics allow us to examine how the different bites are generated. Even if the phenotypes of individuals are equal, the network dynamics may be different.

Phylogeny tracing Every individual has a unique identification tag and its parents' identification tag. We record these relationships, and in addition, we periodically log a sample of the population (every $1 \cdot 10^4$ time steps) and the entire population (every $2.5 \cdot 10^4$ time steps). This allows us to compute the 'true' ancestry of individuals in a phylogenetic tree, to correlate branching depth to phenotypes, and to display the evolution of various properties.

4.2.3 Analysis of ecosystems

In order to visualize the ecosystem at a specific time point, we depict it as a network with resources as nodes and bites that transform one resource in another as edges (see also Figure 4.3). In the resulting figures the following ecological concepts are easily observed.

Shortest cycle The rotations of bit string resources leads to cycles. As we have $n = 64$ resources, a cycle of bites has a maximum length of 64 (i.e. each bite is one rotation step). It follows that shorter cycles have longer bites, which means higher fitness for the individuals involved. This brings us to the shortest cycle: the cycle that contains the smallest set of bites that cycles a set of resources. We require that a bite can be performed by at least 5 individuals at a time step.

Crossfeeding A special case of division of labor is crossfeeding: the shortest cycle is formed by the cooperation of multiple lineages instead of a single one. To be more precise: we compute the shortest cycle as described previously and take a 10% sample from the population ($\approx$ 900 individuals). Crossfeeding is present if the shortest cycle cannot be performed by a single group of phenotypically identical individuals, with the group having at least 5 individuals.

Ecological simulations The stability of crossfeeding is studied with so-called ecological runs. Such runs are initialized with a population and accompanying environment – that is the resources on the grid – from a specific time point of an evolutionary run. In these runs we omit the mutational process and therefore eliminate much of the 'noise' generated by mutants and their aberrant bites.

4.3 Results

Due to the nature of resource processing (rotating a bit string) the ecosystem evolved to contain one or more, overlapping, cycles in which the bites neatly map onto each other. This optimization of resource cycling was selected for indirectly, and importantly, implied a frustration in the evolutionary process. This frustration is a crucial feature of the dynamics of our model: in the long run it can pay off to not increase the direct fitness, but to produce resources which the offspring or local community can process well. We refer to this secondary level of selection for a cycle of bites as indirect selection.

At the core of our studies is this long-term effect on resource changes and how it affects the eco-evolutionary solution. We investigate this by comparing the outcome of the local and null model and adjusting the relative strength of direct to indirect selection. We have taken 3 levels of direct selection ($\sigma = 0.2$, 1.0 and 5.0) and for each level we have performed 25 runs of both models (150 runs in total). Note that it is the relative contribution of direct to indirect selection that matters. Therefore strong direct selection implies weak indirect selection and vice versa.

We first introduce an overview of the dynamics of both models. Next we give a dynamical view of how individuals process resources, what kind of individuals evolve and what the resulting ecosystem looks like. This is followed by the ecological property of population diversity and how it evolves over time: how does the ecosystem "solve the problem" of cycling resources? We focus on the following: do local interactions among individuals promote cooperativity, such

as crossfeeding? Or do individuals evolve to cycle resources 'all by themselves'? If so, how 'smart' are the individuals? And how does this compare to individuals and their community structure if we vary direct selection?

Additional simulations have been run to test for the robustness of our main body of results against various parameter changes, such as mutation rates, starting networks, different resources and resource size (Supporting information).

4.3.1 Overview of the local and null model

In Figure 4.2 we show for both models and each selection regime a representative run. Clearly, both models show that from low to high selection (and thus from relatively high to low indirect selection) the resource dynamics in the environment become more turbulent. Importantly, there is also a clear difference between the models as the local model is more 'quiet', or stable in its resource dynamics.

With respect to fitness (Figure 4.2D – F), we find that the local model often results in a lower maximum bite length than the null model. In both models, the difference between maximum bite and median bite is large, and median bite length increases hardly. Furthermore, it is important to realize both models have many resources present, that is to say there are many evolutionary targets. Thus if a maximum bite length is lost (as we observe multiple times in Figure 4.2E and F), this is due to a new lineage invading at resources that were poorly processed by the dominant lineage.

The most striking property of the diversity (Figure 4.2G – I) in these runs is that the sudden burst of diversity corresponds with the end of the so-called initial phase. It is also interesting to note that a diversity of $\sim$600, given a sample size of 1000 individuals, implies that there is a lot of diversity in the populations of both models at $\sigma = 1.0$ and 5.0, and in the null model at $\sigma = 0.2$.

From the runs in Figure 4.2A, C, J and L we have taken a single time point and visualized the ecological interactions between individuals and resources (Figure 4.3). Again we observe two 'gradients' from rather simple dynamics to complex interwoven cycles of resource modifications: both from local to null model, and from low to high direct selection. Moreover, under low direct selection, we find a single lineage performing the resource cycling, while for high selection multiple cooperating lineages are shown.

4.3.2 The first resource cycles

We consider the example run with local feedback shown in Figure 4.2A. It can be divided into two periods: an initial phase characterized by an equal abundance of a subset of resources, and the long term evolution of the run in which resource numbers varied from equilibrium dynamics to turbulent patterns in time. The initial phase as we observed it in the example run was found across all runs, though its duration varied and was shortened with higher values of σ.

We now present this period in terms of the sequences of zeros and ones that individuals output, allowing us to show the mechanics of our model. This nicely complements the higher-level ecosystem view we take in the subsequent study of long-term evolution.

In the initial phase the population was characterized by individuals that produced simple output bit strings in response to resources. We found that in the example run the networks generated first a transient of zeros, followed by the point attractor of value one. In Table 4.1 all bites are enumerated that lead to reproduction at time step $5 \cdot 10^4$. Many bites were still small (2 or 3 bits) and thus of low fitness. All bites were of low complexity: for instance, both the most abundant (`01`) and most fit (`01111111`) would be generated by the same transient and attractor. Only the most abundant output would have a mismatch much earlier. Also, a single output 'strategy' was often applied to multiple resources (see third column in Table 4.1). To study which bit strings individuals could produce, we analyzed the underlying network dynamics. We found that the strings were generated by a single lineage with the following 5 typical outputs: `00001*`, `0001*`, `001*`, `01*` and `1*`, where the star signals the continued emission of the last bit. Thus the individuals were capable of modulating the transient length

output	count	bites
`01.......`	291	11→13, 13→15, 15→17, 21→23, 38→40
`0011.....`	208	4→8, 17→21, 34→38, 52→56
`011......`	132	31→34, 62→1
`001......`	94	1→4, 8→11
`011111...`	54	46→52
`000011...`	46	40→46
`01111111.`	44	23→31
`000111...`	41	56→62
`11.......`	24	6→8, 19→21, 32→34, 36→38, 44→46
`00.......`	21	1→3, 4→6, 17→19, 34→36, 52→54, 56→58
`1........`	12	3→4, 12→13, 22→23, 39→40
`000......`	8	40→43, 56→59
`0........`	6	11→12, 13→14, 21→22, 38→39, 46→47, 62→63
`0000.....`	5	40→44
`111......`	5	59→62
`11111....`	1	47→52

Table 4.1 – Output bit strings at time $= 5 \cdot 10^4$ for the example run of Figure 4.2A. In the first column output bit strings are shown, with mismatched bits as dots (.). As the length of a correct output is the main determinant of an individual's fitness, one can easily differentiate between high and low fitness bites. The second column contains the number of individuals which generated the output. Matching resource rotations (bites) are given in the third column, corresponding to the ecological snapshot in Figure 4.3A. The upper half of the table contains all bites from the shortest cycle, the lower half contains 'mutant' bites.

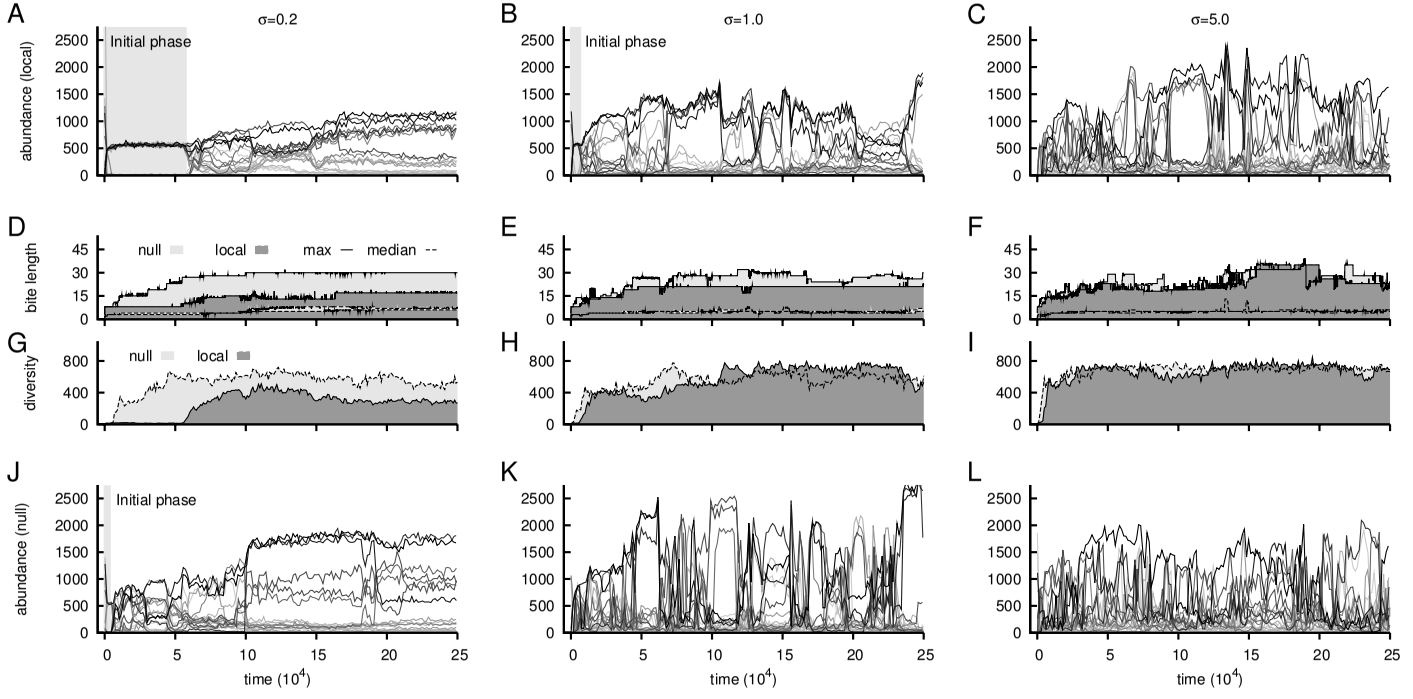

Figure 4.2 – Resource abundance, population fitness and population diversity in a set of typical runs. A – C and J – L. The 20 most abundant resources are plotted through time. The darker a curve, the higher the abundance of this resource throughout the run. The top and bottom row contain, respectively, runs from the local and null model (for each selection regime a run). If visible at this scale, the initial phase is indicated by a gray background. D – F. Maximum and median fitness of the population through time. Dark shaded areas indicate the local model, light shaded areas the null model. G – I. Diversity in the population measured as the number of different network dynamics through time. Each 1000 time steps 1000 individuals were sampled and grouped by their network dynamics. The number of different groups is plotted. Grouping by phenotype gave qualitatively similar results.

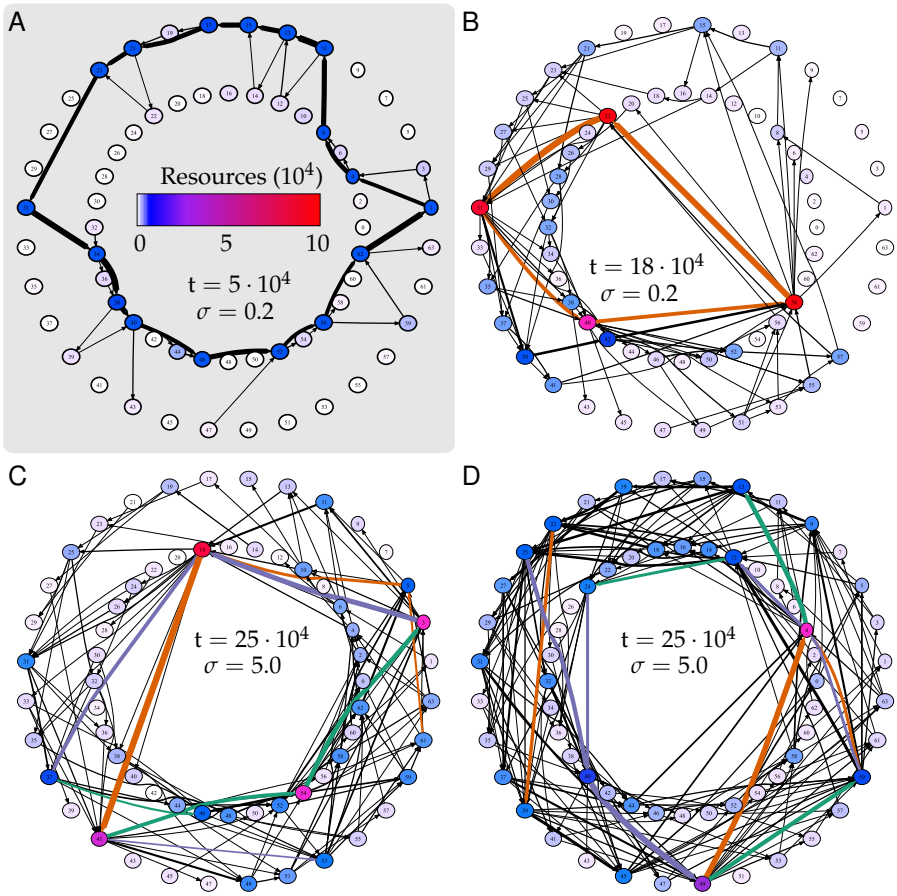

Figure 4.3 – Ecological views of several runs. Each 'wheel' shows the 64 resources (nodes) and the bites (edges) that rotate one resource into another. For visualization purposes even and odd numbered resources are plot on the inner and outer circle, respectively. Resources are colored by abundance, see the legend in panel A. The edge-width is logarithmically scaled according to popularity: the more a bite occurred the thicker the edge. In B, C and D the shortest cycles are colored to distinguish up to 3 different phenotypic groups (orange, purple and green). This coloring does not indicate any relationship between the phenotypes in different 'wheels'. A and C are taken from runs with local feedback (Figure 4.2A and C), B and D are from runs of the null model (Figure 4.2J and L). A. Run with low direct selection ($\sigma = 0.2$) at time $5 \cdot 10^4$, which is still in the initial phase of evolution. B. Run with selection $\sigma = 0.2$ at time $18 \cdot 10^4$. The shortest cycle (22, 31, 40, 58) is performed by a single lineage. C. Local model run with $\sigma = 5.0$ at time $25 \cdot 10^4$. The shortest cycle (3, 18, 41, 54) is composed of three different phenotypic lineages. D. Null model run with $\sigma = 5.0$ at time $25 \cdot 10^4$. There several shortest cycles, composed of multiple lineages. One of these cycles is: 4, 12, 24, 40, 49.

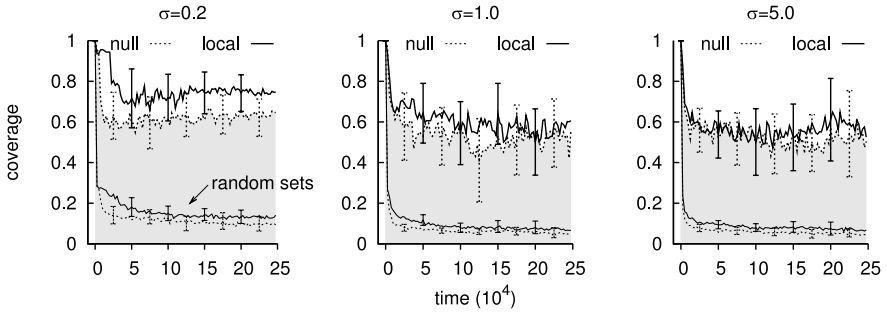

Figure 4.4 – Resource coverage of shortest cycles. Mean coverage of the grid by resources on the shortest cycle, and random sets of resources. Coverage is expressed as a fraction of the total grid size. The coverage of the null model is shaded gray for clarity. The random sets are composed of resources selected at random from the grid such that the size of the random set equals the number of resources on the shortest cycle.

depending on the input before ending in a point attractor. This behavior was typical for the initial phase of the runs.

In the ecosystem snapshot of time step $5 \cdot 10^4$ (Figure 4.3A) we observed a dominant cycle that was composed of popular bites and abundant resources. As expected, this was in concordance with Table 4.1: row 1 to 8 contain all bites from this shortest cycle. The positive correlation between shortest cycle and resource abundance was a general property that held in all runs. The resources on the shortest cycle occupied over half of the grid, while a randomized ensemble of resources covered 3 to 4 times less space (Figure 4.4). We can understand this as an equilibrium between the dominant lineage and the mutants that were present. The dominant lineage maintained resources on the shortest cycle, and possibly channeled other resources onto the shortest cycle. On the other hand the various mutants took bites such that the resulting resources were likely not on the shortest cycle anymore. This explained the equilibrium we observed in Figure 4.2A.

Furthermore, as the shortest cycle was a concatenation of relatively large bites (compared to the bite-composition of other cycles, which were by definition longer), the dominant lineage would often win the competition for reproduction. Thus while this lineage established itself in the population, it out competed the other individuals and there would be an increase in shortest-cycle resources, until an equilibrium was reached. The result was a shortest cycle that was composed of abundant resources and the bites were performed by the dominant lineage.

Summarizing, our ecosystem was initially composed of a population of simple individuals that shaped their environment with a low variety of processing steps. Resource cycles quickly emerged, of which the shortest cycle of bites is an im-

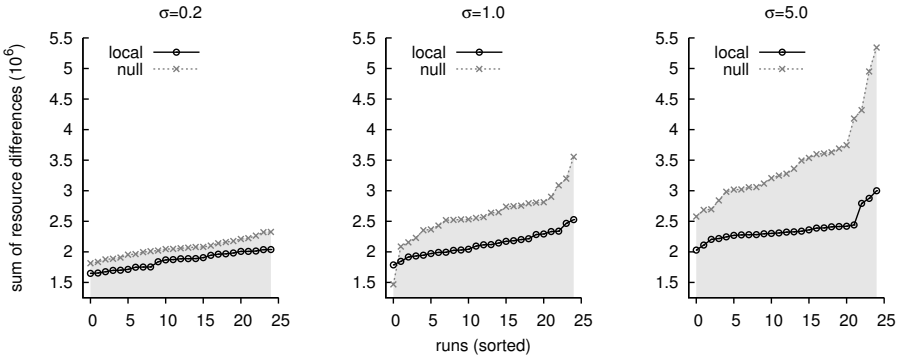

Figure 4.5 – Resource turbulence. For each run, we calculate for all resources the sum of absolute changes in resource abundance through time, and then sum over the resources. Thus we arrive at a single number indicating how turbulent resource dynamics have been through a run. For each selection regime the difference between the local and null model is statistically significant. For $\sigma = 0.2$, 1.0 and 5.0 the Wilcoxon rank-sum test (alternative hypothesis: local less than null) results in, respectively, $p < 2.73 \cdot 10^{-6}$, $p < 5.97 \cdot 10^{-8}$ and $p < 1.38 \cdot 10^{-12}$.

portant property of the system. In all runs, this first phase of evolution abruptly ended with the innovation of oscillatory output patterns (Supplementary Text). It was accompanied by a sudden dramatic increase in phenotypic diversity (Figure 4.2D – F). And while the initial phase had been rather similar for all runs, the long term outcomes were different, as we show next.

4.3.3 Evolutionary stagnation

From a detailed account on the initial phase, we now move to an ecosystem point of view. As mentioned previously, we focus on the effect of indirect selection. We do this by studying the differences in behavior of the local - and null model, and by examining the effect of different levels of direct selection. We show that relatively high indirect selection, especially by local feedback between individual and its environment, leads to evolutionary stagnation in three different ways.

Resource dynamics

In the comparison between the local and null model, resource dynamics in runs with local feedback were less turbulent and more often in equilibrium (Figure 4.2 and Figure 4.5). In addition, the local model showed only a minor increase in turbulence from low to high σ, while for the null model this increase was much larger (Figure 4.5). Thus local feedback enhanced a stable cycling of the same resources (i.e. indirect selection for a cycle of resources), which implies fewer mutants established in the populations and therefore a slow down of the evolutionary process.

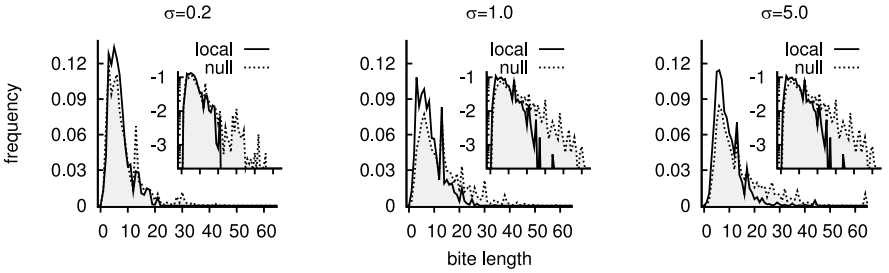

Figure 4.6 – Frequency distributions of bite lengths that lead to reproduction. The mean distributions are shown, summed over the period $t = 12.5 \cdot 10^4$ to $25 \cdot 10^4$, and based on 25 runs of both the local and null model. The inset in the left panels show the frequency plot in log scale, such that the differences in the tail are highlighted. Note that null model curves are shaded gray for clarity.

Bites

Next, we considered the bites that occurred in the population over time. As argued previously, in case an individual influences its local environment, there is an incentive to have resource modifications that nicely follow up on each other: if an organism can increase the probability of its descendants having an "easy bite", this will lead to more successful offspring in the long run. On the other hand, in the null model this process is not present as local kin relationship is destroyed by the mixing of individuals. How does this difference reflect in the distributions of bite lengths over time?

The average bite that lead to reproduction in all selection regimes was around 4-6, for both the local and null model. However, as we observed clearly for low direct selection ($\sigma = 0.2$), the local feedback resulted in less long bites (Figure 4.6, left panel). There were no bite lengths > 21, while the null model showed bite lengths up to 48. Also, for $\sigma = 1.0$ and 5.0 bite lengths in the interval $[20, 40]$ were less present in case of local feedback (Figure 4.6).

Considering that mutation rates were equal and genome lengths comparable, mutants arose at a similar rate in both models. Yet in case of local feedback mutants with larger bites failed to establish. The only difference between the two models was that in a spatial setting mutants were polluting their local surroundings and they were 'confronted' with this. That is to say, while the residing individuals were well-adapted to local resources and due to indirect selection also to the resources at the next time step, mutants found themselves among resources they could not process well. Thus most mutants could not establish a lineage locally and were out competed by the expanding lineages that cycled resources more efficiently. In the null model, however, by mixing the connection between individual and resource composition was much weaker. Mutants were not 'forced' to process their locally produced resources, hence it was easier for them to establish in the population. Thus, in the long run, mutants were subjec-

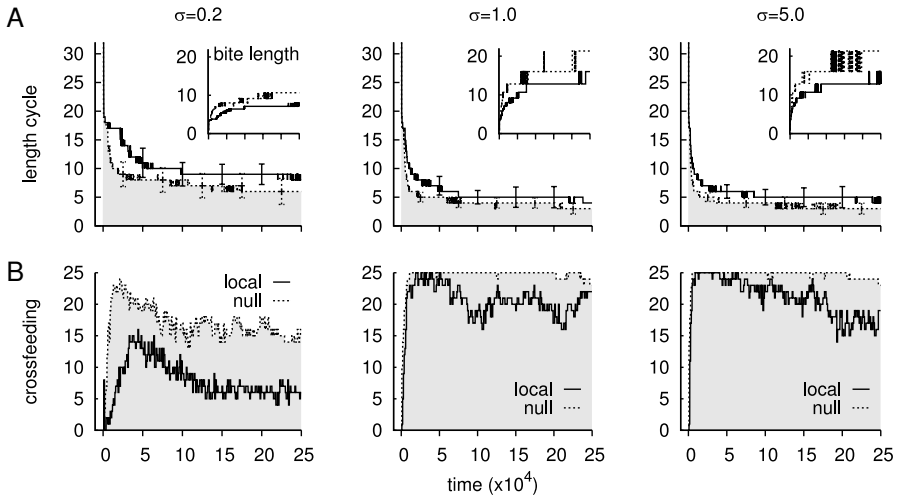

Figure 4.7 – Shortest cycles, their length and crossfeeding. A. For the three selection regimes the mean shortest cycle with standard deviations is plotted through time. Inset panels show the corresponding average bite length on the shortest cycle. In all panels the mean and standard deviation are computed from 25 replicate runs. Using permutation tests, we established all three curves are significantly different (all $p < 0.004$). B. Number of runs with crossfeeding through time. In short, crossfeeding is present if the shortest cycle cannot be performed by a single group of phenotypically identical individuals, with the group having at least 5 individuals (see Methods). Note that in all panels the null model curves are shaded gray for clarity.

ted to stronger competition in the local model. And the higher indirect selection, the stronger this competition was.

From this also follows that for higher values of σ direct selection overruled indirect selection and mutants were able to invade. In concordance with this argumentation is the coverage of the grid by the resources of the shortest cycle. In Figure 4.4 we observed that while low selection had a high coverage (> 0.7), thus indicating few established mutants, $\sigma = 1.0$ and 5.0 showed a distinctly lower coverage. This points at alternative resource modifications taking place.

Shortest cycle

As we have shown the core of the ecosystem dynamics was formed by the shortest cycle, we compared these cycles in the different settings. For $\sigma = 0.2$ not only bite length, but also cycle length had clearly stagnated (Figure 4.7A). Translated into bite lengths, a local feedback cycle of length 9 resulted in an average bite of length 7.11, while in the null model average bites were $64/6 = 10.7$ bits long. Thus on average approximately three bits were processed less if local feedback was present.

Moreover, these differences in cycle length extended to $\sigma = 1.0$ and 5.0. At first sight a difference of one step may not seem significant, but as cycles become shorter the bites become progressively larger. For instance, for $\sigma = 5.0$, if local feedback was present cycle length was 5, which equaled a bite length of 12.8, while the null model had an average bite of $64/3 = 21.3$. Thus there was a difference of more than 8 bits, which indicates quite some change in the network dynamics of the individuals.

Concluding, locality of the feedback between individual and environment enhanced the effect of indirect selection for cycling resources. Naturally, this was most obvious for low direct selection ($\sigma = 0.2$). However in comparison to the null model for all selection regimes we observed less change in the resource dynamics, a stagnated evolution towards long bites and a stagnated evolution of short cycles.

4.3.4 Population structure

Given the evolutionary stagnation, we turn to the underlying population and community structure. We focus on long-term dynamics of the community for the different selection regimes in both the local and null model.

Crossfeeding

As described before, the innovation of oscillatory outputs marked a sudden increase in phenotypic diversity. From this large variety of different individuals cooperative cycling of resources emerged. Subsequently the lineages participating in shorter cycles, thus performing larger bites, took over and this was mirrored in a rapidly decreasing length of the shortest cycle (Figure 4.7A). Eventually a few groups of individuals became dependent on each other for resources on the shortest cycle. We labeled such structuring of the population as crossfeeding, and examined it in detail.

First of all, we looked at the evolutionary dynamics of crossfeeding. It is clear that for all three selection regimes the local model resulted in less crossfeeding than the null model. However, there was a stronger dependence on selection. As shown in Figure 4.7B, for low selection crossfeeding was a transient phenomenon. In both the local and null model, there was a peak before $5 \cdot 10^4$ and in $\sim$ 55% of the runs a single lineage eventually took over, performing the shortest cycle by itself. Such was also the case in the example run we discussed previously, and its null model counterpart at $\sigma = 0.2$ (Figure 4.2A, J). The ecological network of the null model run is shown in Figure 4.3B. A single lineage performed the shortest cycle of length 4, while its close mutants created detours such as $31 \rightarrow 39 \rightarrow 42 \rightarrow 58$ and $58 \rightarrow 11 \rightarrow 15 \rightarrow 21 \rightarrow 31$.

Contrastingly, runs with average and strong selection showed prolonged periods of crossfeeding (Figure 4.7B): it was the main long term evolutionary outcome. This impacted the length of the shortest cycle. Even though there was a

5-fold difference in σ between average and high selection, the lengths are comparable and substantially lower than for $\sigma = 0.2$ (Figure 4.7A). In Figure 4.3C and D snapshots show the partitioning of the resources among different phenotypes. Large bites were performed by specific phenotypes, while shorter ones were more prone to be shared (data not shown).

Ecological stability

Next, we studied the population dynamics of these crossfeeding runs. A population structuring while mutations occur does not imply ecological stability (van der Laan & Hogeweg, 1995). We performed ecological runs and tested if crossfeeding was maintained. The results stressed that for low selection the structuring is not stable: only few runs preserved crossfeeding (Table 4.2). However, for average and high selection all runs showed maintenance of crossfeeding. Thus in these cases we have a stable phenotypic partitioning of the population.

Phylogenetic basis

Still the question remained if multiple lineages were present if there was crossfeeding? We verified that the phenotypic structure of the population had indeed a phylogenetic basis, that is to say there was a genotypic grouping as well. In Figure 4.8 we show for the local model the phylogenetic distance between pairs of individuals against their phenotypic distance. It is important to realize that for short phylogenetic distances we observed mostly quasispecies variation among the phenotypes, while for large phylogenetic distances we have the actual population or community structuring.

One of the first observations was that all 3 figures have a red-colored peak close to the origin $(0, 0)$, which implied that a large fraction of the pairs was both phenotypically and phylogenetically closely related. Second, low selection

σ	runs	time (10^4)		
		5	15	20
0.2	10	3/7	0/3	0/2
1.0	10	10/10	9/9	9/9
5.0	10	9/9	9/9	6/6

Table 4.2 – Ecological stability of crossfeeding. For each of the selection regimes, we have tested the ecological stability of crossfeeding in a random set of 10 runs, for populations taken from 3 time points. The total number of runs, including the ones that lack crossfeeding, is shown in the second column. The first time point is chosen early in the simulations, after the innovation event. The other two time points are indicative of the long term evolutionary dynamics. Each entry gives the fraction of runs with crossfeeding that show ecologically stability. A run is labeled as ecologically stable if crossfeeding is maintained in short ecological simulations (25 000 time steps, no mutations).

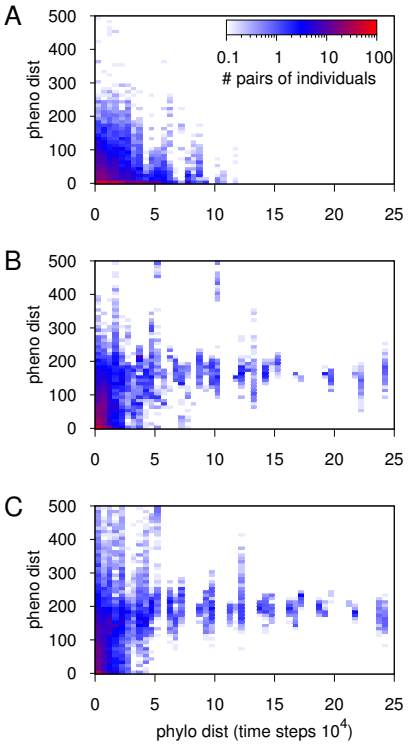

Figure 4.8 – Phylogenetic distance (*phylo dist*) against phenotypic distance (*pheno dist*). We computed phylogenetic trees of all runs with local feedback (for an example tree see Figure S4.3), and for each run we sampled a 1000 random pairs of individuals with a time of birth difference < 20 time steps and traced their last common ancestor. The phylogenetic distance is the difference in time of birth between the pair and their ancestor. Phenotypic distance is expressed as the Manhattan distance between two phenotypes. The colors, as given in the legend, give the number of pairs averaged over 25 runs. Note that the regular spacing in the data of B and C is an artifact of the periodicity of logging populations. Low, average and high selection are shown in A, B and C respectively ($\sigma = 0.2$, 1.0 and 5.0).

showed only limited phylogenetic diversity, which nicely corresponded to the evolutionary solution of a single dominant lineage. Third, there was an obvious difference between low selection and the other two regimes. While for $\sigma = 0.2$ the population remained similar as branch depth increases, for the other two average phenotypic distance increased. Furthermore, the latter two showed deep branches of dissimilar pairs of individuals. In fact, considering some pairs of individuals had a last common ancestor almost $25 \cdot 10^4$ time steps ago, in a few runs the population must have split into separate lineages only a relatively small period after the start.

Overall, the null model agreed well with the observations above. There was, however, one clear difference. We found that the last common ancestor was quickly different (Figure S4.4), contrasting the rather long stretch of similar ancestors found in the local model. This indicates that lineages diverged, and thus specialized, quickly. Such faster divergence of the population is facilitated by mixing: the spread of a new adaptation is not limited by space, hence a mutant may establish faster.

In summary, both the local and null model show a clear dependency of cross-feeding on the strength of indirect selection, and this is most obvious along the different degrees of direct selection. High indirect selection – that is low direct selection or local feedback – often resulted in a single lineage performing a relat-

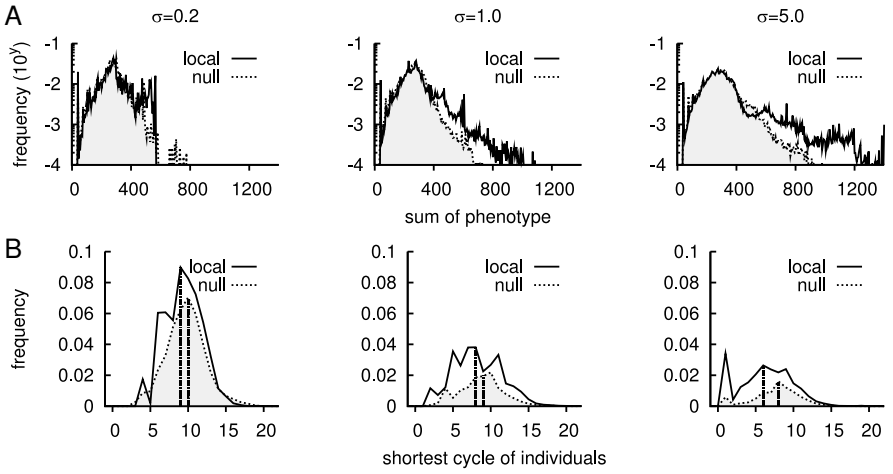

Figure 4.9 – Distributions of individual 'smartness'. In both series of frequency plots we have taken per 1000 time steps a sample of 100 individuals, over the interval $[12.5 \cdot 10^4, 25 \cdot 10^4]$. With 6×25 runs this results in 6 data sets of 325000 individuals. A. Frequency plot of smartness defined as the sum of an individual's phenotype. Note the logarithmic y-axis. B. Frequency plot of smartness as an individual's shortest cycle. We calculated the shortest cycle given the presence of the most abundant resources at each sampling point. A minimal number of resources was selected such that the grid was covered by a 0.95 fraction. The three panels show the distribution of shortest cycle lengths of individuals that could actually perform a cycle. For individuals that were incapable of doing so see Figure S4.5. Note that in all panels the null model curves are shaded gray.

ively long shortest cycle. With decreased indirect selection, via average and high direct selection or mixing of the population, prolonged periods of crossfeeding were observed. This crossfeeding was based on a phenotypic and phylogenetic differentiation in the population, and was ecologically stable.

For high direct selection ($\sigma = 5.0$), indirect selection was relatively low. Hence most of the decline in crossfeeding was not caused by local feedback, but by an impressive result of the evolutionary process. In 4 out of 25 runs with local feedback and 1 null model run a lineage evolved that was capable of outputting a correct sequence of 64 bits for a specific resource, as one can also observe in Figure 4.6. This specialist lineage out competed everyone and thus dominated the population.

4.3.5 Individuals

From the population and community structure we concluded that strong indirect selection favored the evolution of generalist lineages. If we look at the individuals of the runs, an interesting question is to assess their 'smartness'. How well would they process all the resources by themselves? A straightforward defini-

tion of smartness is the sum of an individual's phenotype. Thus per individual we established for 64 resources its corresponding bite length and summed these lengths.

For each setting (3 selection regimes, local and null model) we sampled $3.25 \cdot 10^4$ individuals in the second half of the runs, from $12.5 \cdot 10^4$ to $25 \cdot 10^4$. In Figure 4.9A the resulting distributions of smartness are shown. In all cases there was a peak of 'stupid' individuals close to zero. These were simply mutants that had an extremely deleterious mutation. Secondly, all showed another peak around 300. Thus independent of the selection pressure or spatial setting most individuals were equally smart. Thirdly, it was in the right tail of the distributions that we found distinct behavior for each σ. With increasing σ, individuals became smarter in both models, and even more so in the local model.

However, the following alternative definition of smartness resulted (partly) in the opposite observation. Previously we argued that the evolutionary stagnation of resources and their shortest cycles was caused by the indirect selection for cycling resources. To study how this affected individuals we wondered what shortest cycles would be generated by the individuals themselves (that is without crossfeeding). Thus we introduced a second definition of smartness: to have a short cycle of resources. We allowed the individuals to use only resources that were abundant during their lifetime, and we found a clear difference between individuals along the two 'gradients' of indirect selection. More individuals from the local model were capable of performing a cycle, and they also had a shorter cycle than individuals from the null model (for each selection regime: Wilcoxon rank-sum test, alternative hypothesis: local less than null, $p < 2.2 \cdot 10^{-16}$), see also Figure 4.9B and S4.5. Also, an obvious decline in the number of individuals that can perform a full cycle, was observed from low to high direct selection. Note that in the latter case the individuals that are capable of performing a cycle, actually do this in fewer steps for higher σ.

At first sight it was perhaps paradoxical that different definitions of smartness lead to contrasting results. Considering the first definition – the sum of bite lengths – the explanation is that under strong direct selection, individuals with large bites are strongly selected for and reproduce. We examined runs with local feedback by hand and in a subset of them ($\sim$ 25%) we found that the evolution for large bites for many resources had been very successfull. Local feedback enhanced the capability to integrate information and as a result phenotypic smartness ensued (Pagie, 1999). Hence the tail of smartness in Figure 4.9A. With respect to the second definition, this smartness is simply a consequence of indirect selection: in both the local and null model resources had to be cycled. Thus strong indirect selection, either via locality or low direct selection, enabled the evolution of individuals that cycle resources.

In addition the following argumentation highlights the role of local interactions, as compared to the mixed case. We established that indirect selection for resource cycling via a local feedback resulted in a trend for single lineages performing the cycling across all three selection regimes (Figure 4.7B). Due to the locality of the interactions, crossfeeding was slightly obstructed, and individuals

would (and could) rely on their own lineage (i.e. neighbors on a lattice were most likely closely related). Contrastingly, the mixing of individuals in the null model facilitated crossfeeding and due to the higher long-term evolutionary stability of crossfeeding (Figure 4.7B) individuals may have specialized.

4.4 Discussion

In this study we examine the influence of an individual on its environment in an evolutionary setting. We let individuals evolve an abstract metabolic process of consuming resources and excreting waste products in a closed ecosystem. This introduced a new, indirect selection for resource cycling. As a result a short cycle of resource modifications dominated the ecological dynamics, and depending on the strength of indirect selection eco-evolutionary phenomena such as stagnation, crossfeeding, generalists and specialists were observed.

4.4.1 Evolutionary stagnation

In case of local feedback a lineage "co-evolves" with its surroundings, while mixing of the population (i.e. only global cycling) removes the relation between individuals and their local environment. We observe that local resource cycling results in stagnated evolution. Especially for low direct selection ($\sigma = 0.2$) evolution slows down drastically by the interaction between organism and environment. Looking at all selection regimes, we observe a lower throughput in the resource cycling (that is longer shortest cycles) compared to the well-mixed null model.

Previously, evolutionary stagnation has been observed in the context of predator-prey co-evolution on a lattice (Savill & Hogeweg, 1998). It was found that spatial patterns in the form of patches hindered the invasion of mutants. Recently, in a model on marine microplankton a different mechanism with similar outcome was found: an increasing number of sibling species competing for resources slowed down the evolution of the entire ensemble (Alizon *et al.*, 2008). It is tantalizing to associate such evolutionary slowdowns with the well-known observation of morphological stability in many fossil species (Eldredge *et al.*, 2005). Could it be that interactions with the environment increase robustness of species on an ecosystem level? However, at the moment indirect, resource mediated ecological interactions among individuals have hardly been acknowledged as a potential mechanism that is contributing to the observed evolutionary stasis in the fossil record (Alizon *et al.*, 2008).

4.4.2 Crossfeeding and self-sufficiency

In case of strong direct selection ($\sigma = 1.0$ and 5.0), crossfeeding lineages evolve and maintain with ease in both models. Two, occasionally three, genotypically

and phenotypically distinct lineages form a cooperative community that is ecologically stable. The specialization on specific resources by different lineages could also be interpreted as a partitioning of the resources. Still, as the different lineages clearly depend on each others 'waste', crossfeeding is a more appropriate term for the ecological dynamics.

The common hypothesis is that crossfeeding in (microbial) populations originates from rate-yield trade-offs (Helling *et al.*, 1987; Pfeiffer & Bonhoeffer, 2004; MacLean & Gudelj, 2006; Gudelj *et al.*, 2007). As we omit such thermodynamic constraints, an energy related trade-off is not present. However, due to the fact that a network has a maximum of 16 different genes, there is an 'information storage' trade-off. Nonetheless, we consider it unlikely that the maximum network capacity is reached, given the distribution tails in Figure 4.9A. Instead a more likely cause of generalization and specialization is the difficulty of evolving yet another recognition of a resource bit pattern or another long bite. In other words, the metabolism is practically not constrained as is the case for the rate-yield trade-off, but simply difficult (MacLean, 2008).

If we look at the evolutionary trajectories of the various runs, an interesting succession of phases is found. In the initial phase there is a single lineage that performs the cycling: an individual-based ecosystem. The era of this lineage ends when oscillatory outputs are discovered. This innovation leads to a phase of cooperating lineages: a community-based ecosystem. The long-term behavior that follows depends strongly on the indirect selection for bites that neatly map onto each other. Local feedback and low direct selection both favor an individual-based ecosystem. In the long run the crossfeeding lineages tend to be replaced by a 'smart' generalist lineage that performs the cycling of resources by itself. On the other hand, both mixing the population and a high direct selection override the selection for resource cycling and foster a community-based solution of cooperating lineages. Thus the evolution of our models appears to show a gradient from smart populations with 'stupid', cooperating individuals to 'smart', self-sufficient individuals. Both the distribution of environmental knowledge over different lineages that compose a population, and the embedding of this knowledge in single individuals has been named "information integration" (Hogeweg, 2007).

Such different modes of evolution have been reported previously in an evolutionary model on bacterial restriction-modification (RM) systems under pressure by a continuous stream of phages (Pagie & Hogeweg, 2000a). Either individuals were 'smart' in the sense that they contained many RM systems to guard against many phages, or the population was composed of mutually uninfectable bacteria with few RM systems. Similar results have been found in a study on individual-based versus ecosystem-based function approximation (Hogeweg, 2007), where predators and scavengers subdivide a mathematical function in order to approximate it.

4.4.3 Evolutionary innovation

We reported the innovation of oscillatory output patterns for both the local and null model, followed by the rapid takeover of the population by the lineage that 'discovered' the oscillations (see also Supplementary Text). Though it is tied to the modeling approach we apply, we find it extremely intriguing that a fitness gain may not be reducible to a single gene, but is found on a higher level, namely that of the network discovering a novel behavior. From a classical point of view one expects that as a mutant takes over the population, the phenotypic diversity decreases. In our model this is not the case, even if we decrease the mutation rate by two orders of magnitude. Instead, as the oscillatory behavior establishes in the population, there is a striking increase in phenotypic diversity, genome length and fitness (Figure S4.1 and S4.2). Our conjecture is that the innovation of a new mode of generating output makes a large variety of phenotypes accessible, and this is subsequently exploited by the evolutionary process. In other words, innovation leads to species radiation. Further research is necessary to establish the generality of this phenomenon.

4.4.4 Modeling choices

From an ecosystem modeling perspective, we made choices that deserve some attention. First of all, we wondered if the fact that our ecosystem is closed with respect to the resources affects our results. Therefore we performed various runs with a slightly modified system. Instead of rotating the resources, we let the individuals bite chunks from one side of the bit strings. Thus the bit strings become smaller and at some point are finished. A finished resource is then replenished with a new resource, identical to the original. As a result we now have an open ecosystem with an influx of nutrients. In this model we observed both stagnation and crossfeeding. Also the abundance of the different resources (leftovers of different lengths) changed over time in a qualitatively equal manner as in our default model. Thus our results are not directly dependent on the precise topology of the 'chemical universe'.

Second, questions regarding ecosystem evolution are often approached from an energy point of view, yet we have omitted any explicit constraints on the anabolic and catabolic activities of micro-organisms. This has allowed us to replace the thermodynamic constraints and complex nutrient pathways between bacteria, archaea and other micro-organisms with a much simpler chemical universe, and focus on the concept of having a feedback mechanism. As such our approach is a baseline study for the evolution of ecosystems.

Finally, ecosystems usually consist of (and are modeled as) food webs in which resources not only exist as 'edible' items in the environment, but also are immobilized in individuals and made available via predation and parasitism. In this study we have focused solely on individuals and their abiotic environment, and explicitly left out any direct interactions between individuals, except for competition. Thus in order to study the evolution of more complete ecosys-

tems, we could in future work extend our model with the ability of individuals to evolve interactions among each other.

4.5 Conclusion

The dynamics of our model amounts to how information on the environment is stored in the population and in single individuals. We used selection and the spatial setting to vary environmental structuring, population structuring and individual ecological roles. Most importantly we show the effect of different degrees of indirect selection on the eco-evolutionary solution via the contrast of local against global feedback and the different levels of direct selection.

In short, locality enhances the integration of information from the environment into single individuals. However, as a consequence resources are cycled more slowly and in that sense the ecosystem is less efficient. Though locality also plays a role in the evolution of crossfeeding, the latter is more dependent on the strength of direct selection. Crossfeeding is always observed after the initial phase of evolution, yet it is likely to be only a transient phenomenon if there is (relatively) weak direct selection. In contrast, strong direct selection leads to sustained crossfeeding, and therewith more efficient resource cycling and faster environmental change.

Despite our simplified *in silico* approach, it is suggestive to associate our results with the diverse range of ecosystems formed by microbial communities: from the individual-based, single-species ecosystem in a South African mine (Chivian *et al.*, 2008), simple endolithic ones (Walker & Pace, 2007) to complex soil communities (Dunbar *et al.*, 2002).

4.6 Supporting information

Oscillatory output

As explained in the Results, the initial phase of evolution consisted of individuals that produced simple output bit sequences. Such a sequence consisted of an initial transient of zeros (or ones) that switched to a perpetual stream of ones (or zeros) as the gene network ended in an attractor state.

In each simulation this initial phase ended as individuals evolved networks that ended not in a fixed point attractor, but in a cyclic attractor. The individuals apply these cyclic attractors to oscillate the sequence of bits at the output gene(s). We found simple oscillatory patterns at first, for instance `01*` or `110*`, with the star signaling the continued repetition of the pattern. As simulations progressed, evolution produced individuals that are capable of oscillating differently depending on the input patterns they observe, while the output patterns became more intricate: as an example `01101011110*`.

Such oscillatory output patterns enabled the individuals to process much larger parts of the resource bit string and hence gain a high fitness. Moreover, the oscillatory behavior of the network seems not reducible to a single gene, but is found on a higher level, namely that of the network topology. As described in the Results and Discussion, the discovery of this new class of behavior lead to a rapid radiation of new phenotypes in the population.

Parameter Dependencies

In order to assess the dependencies of our results on the model parameters, we varied several key parameters such as the mutation rates, resource diffusion rate and bit string length.

Mutation rates

We know from previous work that the network modeling formalism we apply has a large fraction of neutral mutations in its mutational neighborhood (Chapter 3). Therefore we have chosen the default mutation rates relatively high, as this increases the probability beneficial mutations are discovered. Hence we focus here on the effect of mutation rates that are lower than the default rates.

We ran simulations with mutation rates one order of magnitude smaller (e.g. gene insertion now is $16 \cdot 10^{-5}$). Except for slowing the speed of evolution, for instance under weak selection the innovation of oscillatory output may take almost $20 \cdot 10^4$ time steps (of a total $25 \cdot 10^4$ steps), we do not observe qualitatively different outcomes. An initial period is visible, followed by the invention of oscillatory output patterns. Also, as expected, only a small period of crossfeeding is found for weak selection, while sustained resource interdependence is observed for average and strong selection.

Different bit strings as resource

Our main body of results is based on a default resource. The resource was randomly generated and we do not expect any specific effects from our particular choice of bit string. Nevertheless we checked the validity of our results against ten other (randomly generated) bit strings of length 64.

In addition we performed simulations ($\sigma = 1.0$) with a shorter bit string (length 32), and a twice as long bit string (length 128). For short bit strings, in 2 out of 3 runs there evolved a specialist that performed a 32 bit output and thus out competed the rest and dominated the population. In case of long bit strings, we performed 3 runs and observed rather quiet resource dynamics (though within the range of $\sigma = 1.0$), long shortest cycles ($> 10 \approx > 5$ for 64 bits) and extensive crossfeeding. The length of the shortest cycles is most likely affected by to the fact that a 128 bits generate many more input patterns to classify and process. The 'problem' of cycling resources became more difficult.

In case of short bit strings the environment is relatively simple and a single individual can specialize on a few resources, thus giving rise to extreme specialists that output 32 bits. With longer bit strings there is more information to be stored and input patterns to be recognized, resulting in extensive crossfeeding.

Different starting networks

Initially each individual has an identical genome and network. We replaced the default starting network with five different, randomly generated, networks, and for each we ran 3 simulations at $\sigma = 1.0$. We observe qualitatively equivalent dynamics as described in the main text.

Next, we ran 10 simulations with a smaller network (8 input genes, no processing genes, 1 output gene) at average selection ($\sigma = 1.0$). The resource dynamics were often rather 'quiet', yet crossfeeding was in 8 of 10 runs the evolutionary stable outcome. In addition, the distribution of bite lengths that lead to reproduction appeared to be an intermediate between low and average selection. Thus by making the network smaller, evolving large bites became more difficult.

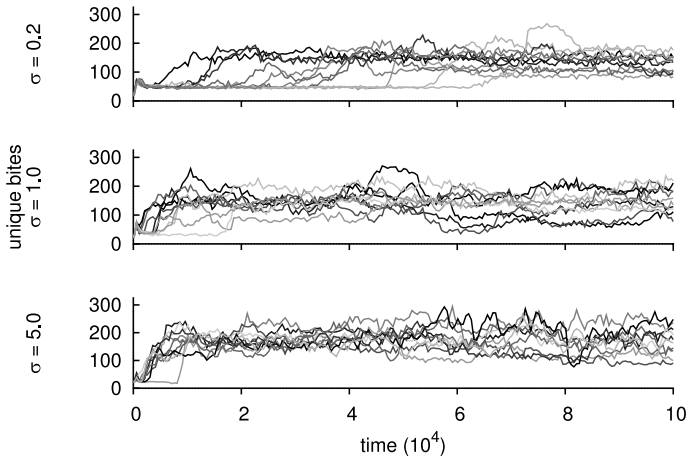

Figure S4.1 – Sudden increase in diversity. The number of unique bites is shown for the first $10 \cdot 10^4$ time steps. The number of unique bites in an ecosystem is directly related to the phenotypic diversity. If there is more variety in the ways individuals process resources, the number of unique bites increases. In the same manner a phenotypically uniform population will have a low number of unique bites, as one can observe in the initial phase of evolution (clearly visible for $\sigma = 0.2$). Furthermore, though we plot only the runs of the local model, the null model simulations result in qualitatively equivalent plots.

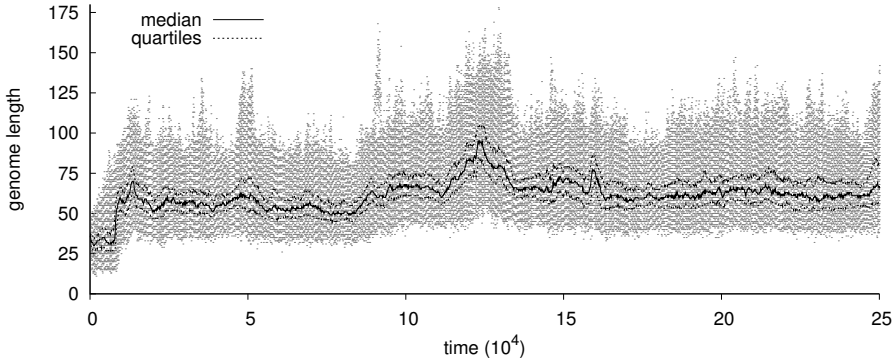

Figure S4.2 – Genome length through time for a run with local feedback and $\sigma = 1.0$. Genome length is defined as the sum of genes and binding sites. The total spread of different genome lengths at each time step is given by the gray dots, with the median and the 1st and 3rd quartile given by the solid and dashed black lines. We observe that the genome length penalty *pen* does not inhibit the evolution of large genomes. See also Figure S4.3 for the phylogenetic tree of this run.

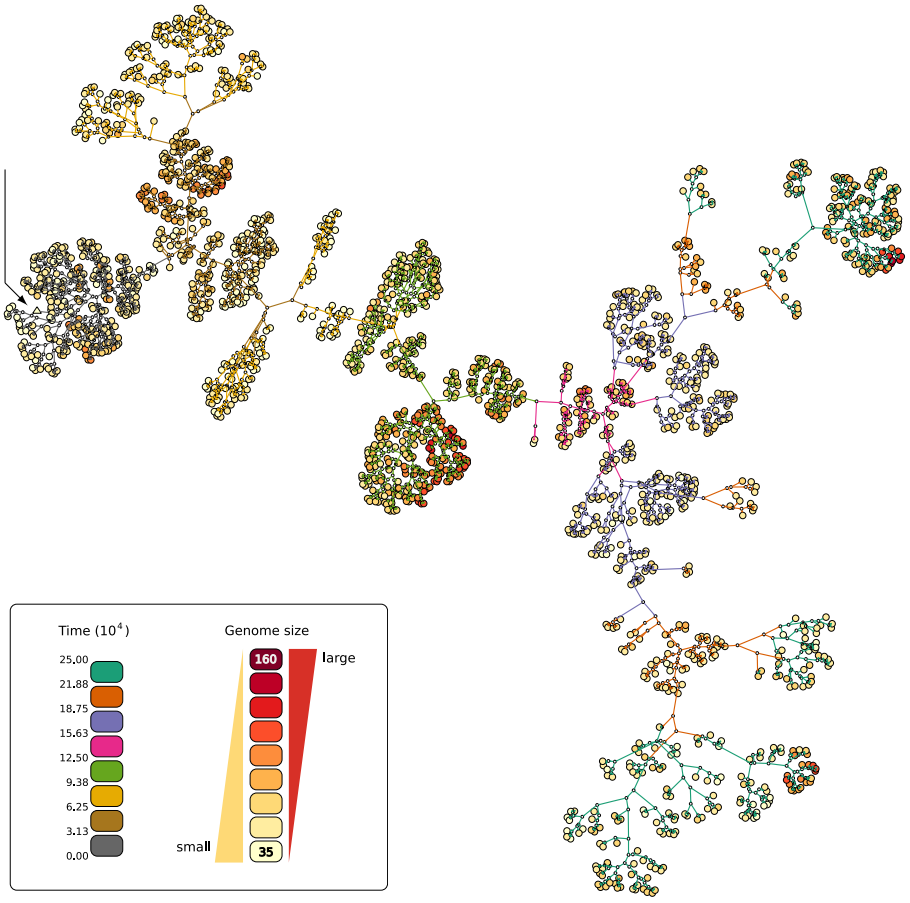

Figure S4.3 – Phylogenetic tree with leafs colored by genome size. We plot the phylogenetic tree of a run with local feedback and $\sigma = 1.0$. Each $2.5 \cdot 10^4$ time steps a population is logged to disk and used in combination with ancestor tracing (see Methods) to build the tree. Nodes are individuals from the logged populations and ancestors at lineage-splitting events. In other words, we prune the tree for intermediate ancestors. The edges thus represent branches from last common ancestors, and are scaled and colored by time interval. For the coloring of the leaves the genome length, genes plus binding sites, is mapped to a color from yellow to red. The arrow in the top-left corner points to the ancestor in the initial population (triangle node). We observe an overall modest genome size, with occasional branches evolving toward long genomes. See also Figure S4.2.

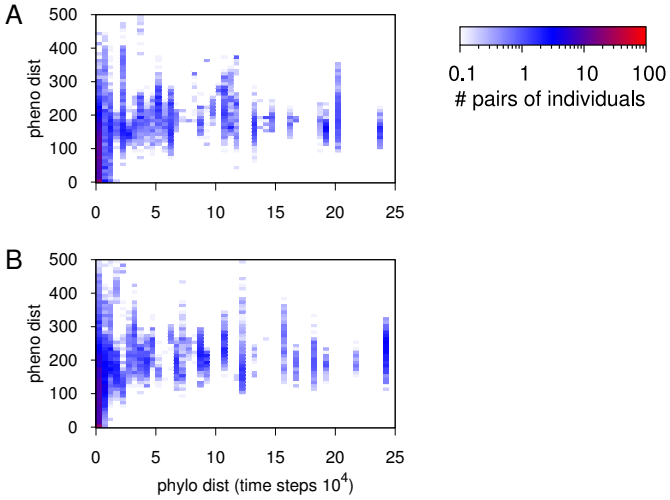

Figure S4.4 – Phylogenetic distance (*phylo dist*) against phenotypic distance (*pheno dist*) in the null model. We computed phylogenetic trees of all null model runs with $\sigma = 1.0$ and 5.0 (due to technical reasons data for $\sigma = 0.2$ was not available). For each run we sampled a 1000 random pairs of individuals with a time of birth difference $<$ 20 time steps and traced their last common ancestor. The phylogenetic distance is the difference in time of birth between the pair and their ancestor. Phenotypic distance is expressed as the Manhattan distance between two phenotypes. The colors, as given in the legend, give the number of pairs averaged over 25 runs. Note that the regular spacing in the data is an artifact of the periodicity of logging populations. Average and high selection are shown in A, B respectively ($\sigma = 1.0$ and 5.0).

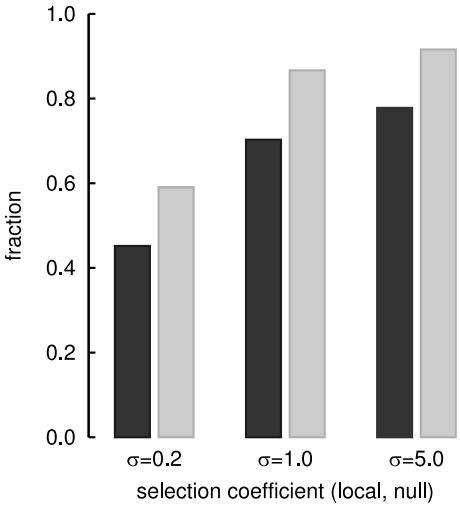

Figure S4.5 – Fraction of individuals incapable of cycling resources by themselves. We have taken per 1000 time steps a sample of 100 individuals, over the interval $[12.5 \cdot 10^4, 25 \cdot 10^4]$. With 6×25 runs this results in 6 data sets of 325000 individuals. For each selection coefficient we find that local feedback (dark gray bars) results in a smaller fraction of individuals that cannot cycle resources on their own, compared to global feedback (light gray bars). This plot complements Figure 4.9B.

5

Modeling RNAi Transposon Control: Interactions between Transcriptional and Post-Transcriptional Gene silencing

Abstract

Eukaryotes use RNA interference and epigenetic regulation to control transposons. The standard pathways of this post-transcriptional and transcriptional silencing, in which RdRP plays an essential role, are found in many organisms. However, animals such as fly, mouse and human lack RdRP, while they still silence the many transposon copies in their genomes. An important question thus is what alternative RNAi regulatory strategies are used by organisms. To this end we study the dynamic interplay between host RNAi pathways and transposon copying.

Using mathematical models, we model three variants of small RNA mediated transposon regulation. We show that the standard RdRP-based model controls transposons well, mainly via the positive feedback of cytoplasmic small RNA amplification. Next, we consider two alternative non-RdRP models that use antisense transcription of transposons to form dsRNA for nuclear silencing. In the first alternative model we assume antisense RNA (asRNA) is exported from the nucleus and used in cytoplasmic dsRNA formation. We observe that, due to having a positive feedback now only in the nucleus, nuclear silencing dominates the dynamics, though cytoplasmic silencing is important too. In the second alternative model there is no transport of asRNA to the cytoplasm. Instead, cytoplasmic dsRNA forms via mRNA hairpins. In this model nuclear silencing is also dominant, yet transposons are not tightly controlled. Finally, we study the transport of small RNAs between nucleus and cytoplasm in all three models. In the standard model this results in a larger fraction of silent transposons at the cost of a larger total number of transposons. The dynamics in both alternative models are not affected by the transport of small RNAs.

We show how three RNAi-based pathways can be used for controlling transposons. Depending on the pathway, particularly its positive feedback loops, silencing of TEs is biased towards either nuclear or cytoplasmic silencing. Moreover, we find that the role played by siRNA transport between cytoplasm and nucleus is ambiguous.

5.1 Introduction

Transposons, or mobile genetic elements, are found in virtually all eukaryotes. They are DNA sequences that have the ability to create copies of themselves in the genome. In order to do so, transposons have an intimate relation with their hosts: they only encode few proteins themselves and are entirely dependent on cellular machinery such as host DNA replication and repair mechanisms and RNA processing enzymes (Beauregard *et al.*, 2008). This copying activity has been linked to deleterious mutations such as chromosomal aberrations, and faulty activation and inhibition of genes (Kano *et al.*, 2007). Hence transposable elements have long been classified as parasites of the genome.

In recent years it has become increasingly clear that transposable elements (TE) have also been recruited by various essential cellular processes, such as alternative splicing (Goodier & Kazazian, 2008), telomere maintenance (Feschotte & Pritham, 2007), and even transposon control (Cam *et al.*, 2008) (but also see (Urrutia *et al.*, 2008)). Furthermore, their tendency to integrate new copies into specific spots on the genome, such as regulatory and coding regions of genes (Wu *et al.*, 2007; Feschotte, 2008; Conley *et al.*, 2008), has been recognized as a source of variation on which Darwinian selection may act. Also, during periods of stress TEs are more active, possibly increasing host adaptability (Slotkin & Martienssen, 2007; Muotri *et al.*, 2007). Thus we may conclude that transposons play a major role in a large range of adaptations (Kazazian, 2004).

As TEs are a powerful mutagenic agent, hosts need to regulate their self-copying activity. The large diversity of TEs, their different sequences and copy strategies have resulted in a large variety of host responses (Lippman *et al.*, 2003). In this study we focus on two main components that are employed in many eukaryotes: transcriptional and post-transcriptional gene silencing (TGS and PTGS) via RNA interference (RNAi), or template matching small RNA molecules. Note that in this study we are not considering the much-related Piwi-based defense against transposons in the germline of multicellular organisms.

PTGS takes place in the cytoplasm. A double-stranded RNA molecule (dsRNA) is cleaved into small interfering RNAs (siRNAs), 21–25 nucleotides long, by a protein of the Dicer family. Next, single siRNA are loaded onto Argonaute proteins, which are part of RNA Induced Silencing Complexes (RISC). RISC identifies complementary RNA transcripts and subsequently degrades them by cleavage. In this manner TE mRNAs cannot be translated and thus TE activity is inhibited. In many eukaryotes the silencing response is enhanced by amplifying the number of siRNAs by means of creating secondary siRNAs (Baulcombe, 2007). Complementary basepair matching of siRNA with mRNA recruits RNA directed RNA polymerase (RdRP, in this case primed RdRP) that synthesizes a complementary strand resulting in a new dsRNA (Sijen *et al.*, 2001).

In the nucleus transposons are silenced by TGS. Both DNA methylation and histone modification patterns lead to the inhibition of expression of the underlying DNA sequence. Here we focus on (di)methylation of histone 3 at lysine 9 (H3K9me) as the signal resulting in heterochromatization. Such modifications

are initiated and maintained by small RNAs (Girard & Hannon, 2008). Schematically, the following process takes place: DNA is transcribed into RNA, which is used by RdRP to form dsRNA. This dsRNA is subsequently sliced by a Dicer protein that physically and functionally interacts with RdRP (Colmenares *et al.*, 2007). Small RNA is then loaded on an Argonaute protein in the RNA induced transcription silencing complex (RITS) (Irvine *et al.*, 2006). Next, this complex recruits methyltransferase, CLR4, that methylates the nearby histones (Lavrov & Kibanov, 2007). Additionally the complex appears to recruit other chromatin modifying proteins such as SWI6/HP1, which bind to H3K9 methylated nucleosomes, compacting them and thus inhibiting transcription of the DNA (Grewal & Jia, 2007). Though the activity of RITS is mostly shown to operate in *cis* (Grewal & Jia, 2007), we assume there are *trans*-effects as well (Iida *et al.*, 2008).

In both above described silencing processes RdRP seems to be a crucial component. However, in fly (*D. melanogaster*) and mammals this protein is not present (Shabalina & Koonin, 2008), yet these organisms also employ RNAi to control TEs (Ghildiyal *et al.*, 2008; Chung *et al.*, 2008). Thus, an important question is whether these organisms have an RdRP that has evaded detection despite extensive searches, or whether RdRP is not as essential to silencing as perceived. In this study we aim to shed light on this matter by modeling RNAi-based silencing of transposons in the nucleus and cytoplasm, i.e. TGS and PTGS, both with and without RdRP. We study an RdRP-based system first, after which we investigate two alternative scenarios for silencing without RdRP. In addition we compare the mechanisms in terms of active and silenced transposons, and abundance of TE products in the cytoplasm.

Our model consists of a set of ordinary differential equations based on the PTGS model by Groenenboom *et al.* (2005) and extended with transposon dynamics and TGS. We take the life cycle of a 'generalized' retrotransposon to represent the transposons, and the key feature of heterochromatization in our model is that its rate is modulated by small RNAs. We show that in this model RNAi with RdRP is a robust method of transposon silencing. Both TGS in the nucleus and PTGS in the cytoplasm are capable of controlling a TE invasion. As we combine the two mechanisms, because of siRNA amplification by primed RdRP the emphasis is on cytoplasmic silencing. Furthermore, transport of small RNA to the nucleus decreases the number of active transposons at the cost of an overall higher number of TEs.

If we assume RdRP is absent, an alternative way of generating dsRNA must be present. Also, no siRNA amplification will occur. Several dsRNA formation strategies have been hypothesized and/or observed: convergent and divergent transcription from both strands (Watanabe *et al.*, 2008), *trans*-acting natural antisense transcripts (NAT) (Werner *et al.*, 2009) and hairpins due to inverted repeats (Watanabe *et al.*, 2008; Okamura *et al.*, 2008). In addition, many dsRNAs derived from TEs have been found in Drosophila and mouse (*Mus musculus*), in which no RdRP has been observed, but how these have been formed is unknown (Okamura & Lai, 2008). Considering the increasing evidence that antisense transcription is both widespread and associated with epigenetic silencing (Katayama *et al.*, 2005;

Iida *et al.*, 2008), we assume in both alternative scenarios dsRNA production via sense-antisense duplex formation in the nucleus. We find that such systems, that lack RdRP, may silence TEs well.

The difference between the two alternative scenarios is in the cytoplasmic dsRNA formation. If we include antisense RNA (asRNA) transport to the cytoplasm and sense-antisense dsRNA formation, TE are silenced strongly, while if we do not have such extra transport and dsRNA originates from hairpin formation of cytoplasmic mRNA, TE are moderately silenced. Moreover, siRNA trafficking between the cytoplasm and nucleus does not affect these systems. It follows that in our alternative silencing scenarios the emphasis of TE control is on the nuclear TGS mechanism, in contrast with the RdRP containing model where cytoplasmic PTGS dominates.

In summary, we show three RNAi-based pathways that the host may utilize in its attempts to control the copying activity of transposable elements. Depending on the constituents and the positive feedback loops that are present, the silencing of TEs is biased towards nuclear silencing or cytoplasmic silencing. Furthermore, the role of siRNA transport between cytoplasm and nucleus remains ambiguous.

5.2 Methods

The model consists of three components: transposon copying with intermediate stages in the cytoplasm, transcriptional silencing (TGS) via small RNA guided heterochromatin formation and translational silencing (PTGS) via small RNA mediated mRNA degradation (Figure 5.1).

Inspired by the cycle of retrotransposons (Sabot & Schulman, 2006), we model a transposon's life as a 4 stage cycle. A functional, active transposable element produces mRNA that is exported from the nucleus. In the cytoplasm the mRNA is translated into various proteins needed for virus-like particles. In such VLPs two mRNAs are reverse transcribed to a single, double strand DNA molecule (Feng *et al.*, 2000). This new transposon copy is transported to the nucleus and integrated into the host genome.

Next, we model RNA-based regulation on two levels. Firstly, transposable elements may be transcriptionally silenced. In the Background we schematically explained the general mechanism. We described the initiation of heterochromatin, which is explained as being triggered by RNA transcription. In contrast, established heterochromatin is defined as silent DNA, and then a paradoxical property of its maintenance is that the necessary siRNAs do not bind to the silent DNA directly, but to an intermediate RNA transcript. Thus in order to silence DNA continuously, it must be transcribed (Grewal & Elgin, 2007). Recently it has been found in *S. pombe* that the transcription is a cell-cycle dependent process, where during S phase DNA is transcribed (Kloc & Martienssen, 2008; Kloc *et al.*, 2008; Chen *et al.*, 2008). Though the advances in understanding the links between RNAi and heterochromatin are rapid, we abstract from most molecular details. As a substitute for the cell cycle dependent transcription, we take a con-

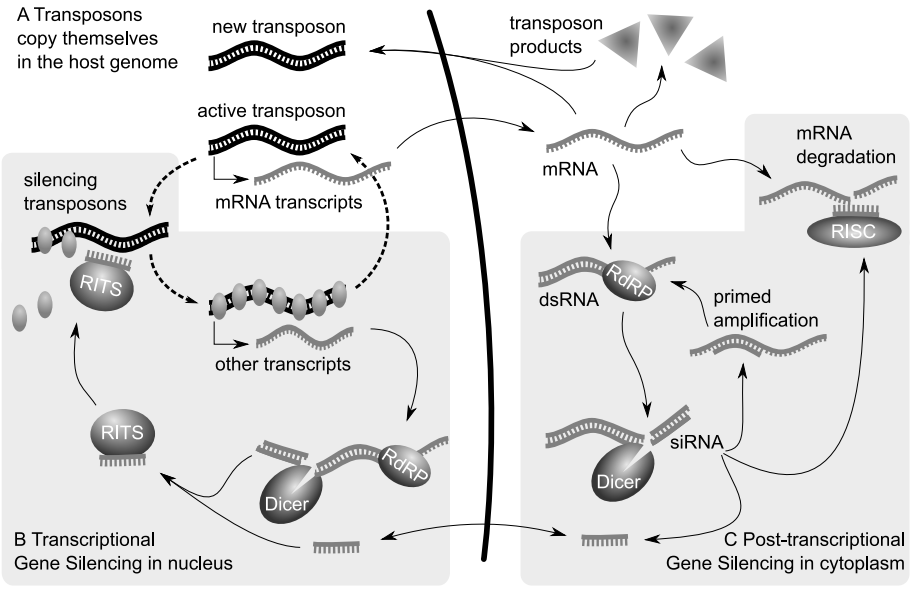

Figure 5.1 – RdRP-based model of transposon silencing. A. The general transposon life cycle is shown in counter-clockwise fashion: active transposons are transcribed, mRNA is transported to the cytoplasm where it is used to produce various proteins, and it acts in duplex formation of mRNA for the generation of DNA. B. Transposon silencing in the nucleus, or TGS. Active transposons are silenced both by a background process, and by influence of nuclear siRNA. As they are silenced, still some transcripts are produced (see Methods). These RNA transcripts are processed by RdRP and Dicer, and the resulting siRNA are loaded onto RITS. C. Cytoplasmic transposon control, or PTGS. RdRP creates dsRNA from mRNA at a small rate, which is then cleaved by Dicer to generate siRNAs. These siRNA prime complementary mRNA for dsRNA formation by RdRP, are loaded onto RISC for mRNA degradation and may be transported to the nucleus. Note that we also allow for siRNA transport from the nucleus to cytoplasm.

stant low level of transcription of heterochromatic regions. Moreover, we model heterochromatin formation as a siRNA modulated process, while maintenance is captured by the "unsilencing" or activation process and possible subsequent re-formation.

Secondly, transposon mRNA is degraded by the RNAi machinery in the cytoplasm. We adopt the model – and accompanying parameter values – developed by Groenenboom *et al.* (2005). They have shown that in order to explain observations on the core RNAi pathway, it is necessary to extend it with either a siRNA degrading RNase, RdRP-based amplification of aberrant RNA fragments or cooperative aberrant RNA that triggers amplification by RdRP. As we also investigate the control of TEs in systems without RdRP, we decided to employ the RNase extension. Furthermore, transport of siRNA from the cytoplasm to nucleus has been reported in *C. elegans* (Guang *et al.*, 2008). Though not as well-understood,

transport in the other direction, from nucleus to cytoplasm, is included in our model as well (Berezhna *et al.*, 2006).

Below the RdRP-based model is shown as a system of coupled ordinary differential equations, with the default parameter values listed in Table 5.1 (see Supporting information 1 for the corresponding stochastic model). Note that we do not consider the decay of TEs, and as a consequence we study a system that is in a transient, not an equilibrium.

5.2.1 Basic model

Transposons

$$T'_{act} = jfV + uT_{sil} - (h_b + h_s S_n)T_{act} \tag{5.1}$$

A fraction f of the virus-like particles manages to integrate, j, a new transposon into the DNA of the host. Some transposons are activated u from the silenced state, while active transposons are silenced with a basal rate h_b and by small RNA h_s. We do not consider the decay of TE.

Silenced transposons

$$T'_{sil} = (h_b + h_s S_n)T_{act} - uT_{sil} \tag{5.2}$$

Silenced transposons originate from active ones, h_b and h_s, and may be activated, u, again. Note that silent TE do not decay.

Nuclear mRNA

$$M'_n = v_{ta}T_{act} - t_m M_n - d_m M_n \tag{5.3}$$

Transcription of active transposons occurs with a rate v_{ta}. The resulting mRNA is either transported to the cytoplasm, t_m, or decays in the nucleus, d_m.

Nuclear other RNA

$$R'_n = v_{ts}T_{sil} - p_n R_n - d_r R_n \tag{5.4}$$

Silenced transposons produce RNA (untranslated transcripts) that is immediately processed by RdRP to dsRNA with rate p_n, or degraded d_r.

Nuclear dsRNA

$$D'_n = p_n R_n - g_s D_n \tag{5.5}$$

Nuclear dsRNA is produced from the transcripts of silenced transposons, p_n, and cleaved by Dicer, g_s. Note that in contrast to cytoplasmic RNA-based silencing we do not consider primed amplification.

Nuclear siRNA

$$S_n' = g_s n D_n - \frac{v_s S_n}{k_s + S_n} - d_s S_n - h_s S_n T_{act} - t_{sn} S_n + t_{sc} S_c \tag{5.6}$$

Nuclear small RNA is produced from dsRNA, g_s. The small RNAs are degraded enzymatically, v_s, and according to mass action d_s (see also (Groenenboom *et al.*, 2005)). Next, nuclear siRNA are used by RITS in the heterochromatization of transposons, h_s. In addition there is export from the nucleus t_{sn}, and import from the cytoplasm, t_{sc}.

Cytoplasmic mRNA

$$M_c' = t_m M_n - q M_c^2 - d_m M_c - p_c M_c - g S_c M_c - b S_c M_c \tag{5.7}$$

From the nucleus mRNA enters the cytoplasm with rate t_m. Here it is used in the production of VLP, q, or simply decays, d_m. With respect to cytoplasmic silencing mRNA is made into a dsRNA by RdRP, p_c. Also, RdRP-based amplification by priming with siRNA occurs, g. Furthermore, mRNA is subjected to RISC degradation, b.

Cytoplasmic dsRNA

$$D_c' = p_c M_c - g_s D_c + g S_c M_c \tag{5.8}$$

Cytoplasmic dsRNA is created from cytoplasmic mRNA, p_c and primed RNA g, and thereafter cleaved by Dicer, g_s.

Cytoplasmic siRNA

$$S_c' = g_s n D_c - \frac{v_s S_c}{k_s + S_c} - d_s S_c - b S_c M_c - g S_c M_c + t_{sn} S_n - t_{sc} S_c \tag{5.9}$$

Small RNAs are produced by cleavage of dsRNA by Dicer, and subsequently degraded by an RNase, v_s, k_s and 'overflow' decay d_s. With respect to Groenenboom *et al.* (2005), this is one of their extensions to the basic pathway of RNAi. Small RNAs are integrated into RISC and facilitate degradation of mRNA, b. Note that we assume that small RNA are degraded as well in the process, which may be a more stringent constraint than biologically needed. Small RNAs also prime mRNA for synthesis of a second complementary strand by RdRP, g. Finally, transport to and from the nucleus is present.

Virus-like particle

$$V' = q M_c^2 - j V - d_v V \tag{5.10}$$

Though we model transposon activity as if it is a retrotransposon that codes for a virus-like particle (VLP), it holds for all TEs that various proteins and intermediate steps are required to create a new DNA copy that can be integrated into the

host genome. Thus the terms may be viewed as the general process of protein production etc. that is required for transposon activity. Throughout this report we refer to V as VLP, or VLP levels. We model the production of a new transposon, q, as a process with some cooperativity among mRNAs. This is most obvious in retrotransposons. In yeast, Ty families require two mRNA to form a dimer in order to produce a single DNA copy. The resulting VLP (or its load, a DNA) moves to the nucleus for integration, j. In addition we have decay of VLPs, d_v.

5.2.2 Alternative models

Below we list the equations that changed due to the absence of RdRP, with extra or changed terms in bold. We now assume silenced transposons produce antisense transcripts at a low level, and hence dsRNA may be formed from duplex formation of antisense RNA (asRNA) with mRNA. In the antisense model we allow this to occur both in the nucleus and cytoplasm, and in this model we also include transport of asRNA across the nuclear envelope. In the hairpin model there is no such asRNA transport present, and we assume cytoplasmic dsRNA is produced from hairpin formation of mRNA (see Eq. 5.15 and Eq. 5.18).

Nuclear mRNA (from Eq. 5.3)

$$M_n' = v_{ta}T_{act} - t_m M_n - d_m M_n - \mathbf{p_{nx}R_nM_n} \tag{5.11}$$

Nuclear mRNA is recruited in the formation of dsRNA, p_{nx}.

Nuclear asRNA (from Eq. 5.4)

$$R_n' = v_{ts}T_{sil} - \mathbf{p_{nx}R_nM_n} - d_r R_n - \mathbf{t_{an}R_n} \tag{5.12}$$

Silenced transposons produce asRNA that may combine with mRNA to dsRNA with rate p_{nx}. Also, asRNA is transported to cytoplasm, t_{an}.

Nuclear dsRNA (from Eq. 5.5)

$$D_n' = \mathbf{p_{nx}R_nM_n} - g_s D_n \tag{5.13}$$

Nuclear dsRNA is produced from asRNA and mRNA at rate p_{nx}, and cleaved by Dicer, g_s.

Cytoplasmic mRNA (a) (from Eq. 5.7)

$$M_c' = t_m M_n - qM_c^2 - d_m M_c - \mathbf{p_{cx}R_cM_c} - bS_c M_c \tag{5.14}$$

The generation of dsRNA is based on mRNA and asRNA, p_{cx}. Also, there is no amplification of siRNAs via primed RdRP.

Cytoplasmic mRNA (b) (from Eq. 5.7)

$$M_c' = t_m M_n - q M_c^2 - d_m M_c - \mathbf{p_{cxx} M_c} - b S_c M_c \tag{5.15}$$

The generation of dsRNA is based on hairpin mRNA, p_{cxx}. There is no amplification of siRNAs via primed RdRP.

Cytoplasmic asRNA (only in antisense model)

$$\mathbf{R_c'} = \mathbf{t_{an} R_n} - \mathbf{p_{cx} R_c M_c} - \mathbf{d_r R_c} \tag{5.16}$$

Cytoplasmic antisense RNA is imported from the nucleus, t_{an}, can be recruited for dsRNA, p_{cx}, and may decay d_r.

Cytoplasmic dsRNA (a) (from Eq. 5.8)

$$D_c' = \mathbf{p_{cx} R_c M_c} - g_s D_c \tag{5.17}$$

Cytoplasmic dsRNA is created from cytoplasmic asRNA and mRNA, p_{cx}, and thereafter cleaved by Dicer, g_s.

Cytoplasmic dsRNA (b) (from Eq. 5.8)

$$D_c' = \mathbf{p_{cxx} M_c} - g_s D_c \tag{5.18}$$

Cytoplasmic dsRNA is created from cytoplasmic mRNA hairpins, p_{cxx}, and thereafter cleaved by Dicer, g_s.

Cytoplasmic siRNA (from Eq. 5.9)

$$S_c' = g_s n D_c - \frac{v_s S_c}{k_s + S_c} - d_s S_c - b S_c M_c + t_{sn} S_n - t_{sc} S_c \tag{5.19}$$

As RdRP is absent, small RNAs in the cytoplasm are not subjected to primed amplification.

Par.	Description	Value	Units
j	Integration of new transposon	0.1	hr^{-1}
f	Fraction of successful integration	0.1	-
v_{ta}	Transcription of active transposons	16	hr^{-1}
v_{ts}	Transcription of silenced transposons	1.6	hr^{-1}
t_m	Export of mRNA from nucleus	0.45	hr^{-1}
t_{sc}	Import of siRNA from cytoplasm	0.45	hr^{-1}
t_{sn}	Export of siRNA from nucleus	0.45	hr^{-1}
t_{an}	Export of asRNA from nucleus*	0.45	hr^{-1}
q	VLP production (proteins etc)	$1 \cdot 10^{-5}$	$\#\text{mol}^{-1}\ \text{hr}^{-1}$
d_v	Decay of VLP	2.0	hr^{-1}
u	Activation of silenced transposon	0.02	hr^{-1}
h_b	Basal heterochromatin formation	0.01	hr^{-1}
h_s	siRNA induced heterochromatin formation	0.001	$\#\text{mol}^{-1}\ \text{hr}^{-1}$
p_n	Rate of dsRNA synthesis from nuclear RNA	0.002	hr^{-1}
p_{nx}	dsRNA synthesis from mRNA and asRNA*	$2 \cdot 10^{-4}$	$\#\text{mol}^{-1}\ \text{hr}^{-1}$
d_r	Decay rate nuclear RNA	0.28	hr^{-1}
p_c	Rate of dsRNA synthesis from mRNA	0.002	hr^{-1}
p_{cx}	dsRNA synthesis from mRNA and asRNA*	$2 \cdot 10^{-4}$	$\#\text{mol}^{-1}\ \text{hr}^{-1}$
p_{cxx}	dsRNA synthesis from hairpin mRNA*	0.002	hr^{-1}
g	Primed amplification rate	0.002	$\#\text{mol}^{-1}\ \text{hr}^{-1}$
b	RISC activity	0.008	$\#\text{mol}^{-1}\ \text{hr}^{-1}$
d_m	Decay rate nuclear/cytoplasmic mRNA	0.14	hr^{-1}
g_s	Rate of dsRNA cleavage by Dicer	2.0	hr^{-1}
n	Number of siRNAs cleaved from single dsRNA	10	-
d_s	Decay siRNA	2.8	hr^{-1}
v_s	Degradation rate by RNase	800	$\#\text{mol}\ \text{hr}^{-1}$
k_s	Saturation constant	5.0	#mol

Table 5.1 – Overview of parameters, their description and default value. Parameters marked with a star ($*$) in their description are used only in the alternative model. Units are number of molecules (#mol), and per hour (hr^{-1}). See Supporting information 2 for references and estimations of parameter values.

5.3 Results

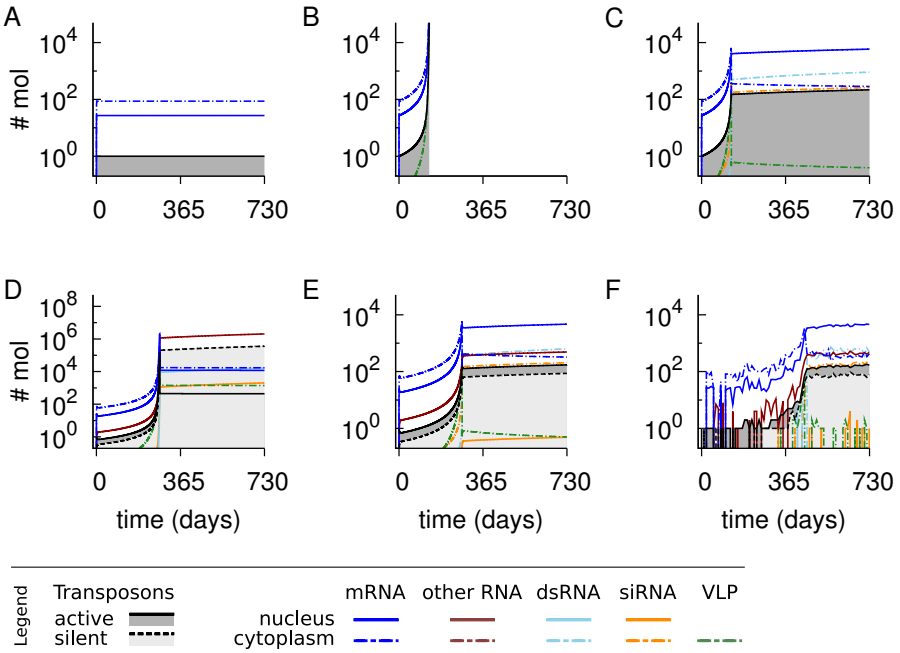

Figure 5.2 – Time plots of the basic RdRP model. Time runs from 0 to 2 years (yr) (17520 hr or 730 days), # *mol* is number of molecules. Active and silenced TEs are shaded in dark and light gray respectively. A. Transposon activity without silencing mechanisms and integration ($f = 0$). B. Copying of transposons in a silencing-defective host. C. Post-transcriptional gene silencing (PTGS). Cytoplasmic silencing is established around $t = 129$ days. D. Transcriptional gene silencing (TGS). The host controls the TE invasion at $t = 275$ days, but many TEs have invaded. Note the differently scaled y axis. E. Both PTGS and TGS are active, but PTGS is clearly dominant (compare to C and D). Around $t = 274$ RNAi regulation starts limiting TE growth. There is no transport of siRNA between the cytoplasm and nucleus. F. Typical stochastic simulation with PTGS and TGS, but no siRNA transport. The number of TEs is slightly lower than in E.

5.3.1 Transposon invasion in basic silencing model

We first study the basic model with a nuclear and cytoplasmic RNAi response to TE activity based on RdRP. The model and default parameter choices are reported in the Methods. In our simulations, we implicitly follow a lineage of individuals. First of all, this is a consequence of the timescale at which transposons invade. Second, there is mounting evidence that transposons are especially active in dividing cells (Kloc & Martienssen, 2008). Furthermore, we define TE silencing as

successful if either VLP levels are already < 1 or if VLP levels approximate 1 and are still decreasing. Thus we regard TEs as well-controlled if their cytoplasmic products are hardly present.

To establish a baseline behavior of a TE invasion, we place a single active transposon in a silencing defective host. First, the fraction of successful integrations f is set to zero (Figure 5.2A). The transposon now functions as a normal gene with an approximate ratio between nuclear and cytoplasmic mRNA of 1:3 (Jarmolowski *et al.*, 1994). If we now change the fraction of successful integrations to the default value of $f = 0.1$, transposon numbers grow exponentially (Figure 5.2B) due to the lack of control from the host.

PTGS and TGS considered separately

Next, we study the effect of PTGS in the absence of TGS. The degradation of mRNA by the cytoplasmic RNAi pathway is clearly able to control the invasion (Figure 5.2C). We observe a threshold effect: initially transposons copy themselves without the interference of the host, yet at ~ 150 active transposons RNAi starts to massively degrade mRNA. The lack of mRNA then causes VLP levels to drop, which stops the invasion of TEs. The nature of this threshold and how the underlying pathways generate it has been discussed in-depth by Groenenboom *et al.* (2005). Interestingly, in an experimental invasion of a Tobacco retrotransposon Tnt1 into Arabidopsis (*A. thaliana*), a similar threshold has been observed (Pérez-Hormaeche *et al.*, 2008). Introducing an active element in hosts having few Tnt1 copies, would not silence the new element. However if the host already contained over 20 copies, no activity of the new element was observed. We note that there is 7.5-fold difference between the TE copy number in our model and that observed in Arabidopsis. It could easily be explained by additional, not in our model incorporated, pathways that are involved in transposon control. Moreover, in the current model it may be overcome, for instance, by increasing the primed amplification to $g = 0.006$.

Nevertheless, there are limitations to the parameter range in which successful cytoplasmic silencing occurs in our model. Most parameters can be varied over roughly 2 orders of magnitude, 10-fold up and down, with only quantitative effects. That is to say, simply fewer or more TEs become fixed before the silencing threshold is reached and the invasion is stopped. However, two parameters have a relatively narrow range for successful silencing. If we take cleavage by RISC 2-fold higher, $b > 0.017$, or the number of siRNA from a single dsRNA 2-fold lower, $n \leq 5$, there is uncontrolled exponential growth of TEs. The latter – the number of siRNA produced from a single dsRNA– is rather straightforward: it appears we need a certain multiplication of the signal in order to degrade enough mRNA. However, the former – the cleavage rate of RISC– requires some further inspection. We realized this is essentially the result of a modeling artifact. By not modeling RISC as a separate entity (in which siRNAs have a certain lifespan), but using simple mass action kinetics, we implicitly assume that for each cleavage a siRNA is used and degraded. In fact, it is much more likely that small RNAs are

reused by RISC for a number of cleavages.

If we now investigate only epigenetic silencing, or TGS, we find that the host again is able to control the invasion (Figure 5.2D). That is to say, the number of transposons does not grow indefinitely but stops at a certain level. However, in comparison to PTGS, the number of transposons in heterochromatin is enormous ($3.57 \cdot 10^5$ against 217.3). This difference is caused by the different architectures of the positive feedback loops present in PTGS and TGS, and the much lower transcription of silent elements. In PTGS the feedback loop consists of primed mRNA on which RdRP acts, whereas in TGS this loop consists of silenced TEs leading to RNA production, dsRNA and siRNA formation and hence more silenced transposons. In other words, in PTGS amplification of silencing requires amplification of siRNAs, which is much less dependent on the number of transposons, whereas in TGS amplification of silencing specifically requires amplification of the number of silenced transposons. If we also take into account the low transcription rate of silenced TEs, it is clear that silencing requires a large number of silenced transposons in TGS but not in PTGS.

Again, there are some limits to the parameter range within which "successful" control of a TE invasion occurs in TGS. No TE silencing occurs if transcription of silent elements is 2-fold lower ($v_{ts} < 0.62$), if decay of nuclear other RNA (not mRNA) is 4-fold higher ($d_r > 0.78$) and if the number of siRNA cleaved from a single dsRNA is too low ($n < 4$). Furthermore, the question remains if a host can handle the sudden increase from 1 to 10^5 transposon copies. Most likely, such an increase severely disrupts essential cellular processes such as transcription, histone production, and nuclear traffic.

In summary, both pathways of small RNA-based transposon silencing may independently control a TE invasion. In case of PTGS we have biologically rather realistic TE copy numbers, however, if we have only heterochromatin-based silencing, from a biological point of view the host is probably not capable of controlling the invasion.

PTGS and TGS studied together

As a next step we study the combination of PTGS and TGS transposon control. As expected, the two mechanisms together are very well able to control a transposon invasion (Figure 5.2E). We observe that for default parameter values – but without siRNA transport between nucleus and cytoplasm – PTGS is the dominant silencing mechanism. As discussed above, the cytoplasmic silencing mechanism is considerably more efficient at producing small RNAs than the nuclear mechanism. Recently, it has been found that specific Argonaute proteins transport small RNAs from the cytoplasm to the nucleus in *C. elegans* (Guang *et al.*, 2008). We hypothesize that such transport may contribute to TE control and therefore incorporate it into our model.

If we include siRNA transport in our model, PTGS remains the dominant pathway (data not shown). Yet, the heterochromatization has clearly been improved by the transport. For a large range of TE abundance that we may start with,

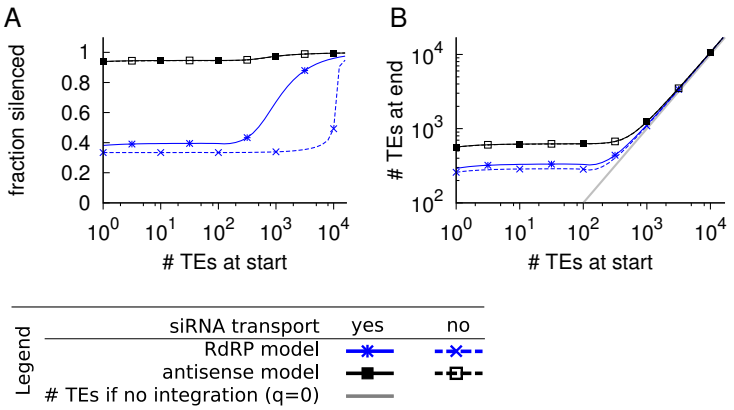

Figure 5.3 – Transport of siRNA from cytoplasm to nucleus. In both subfigures, the number of active and silent TEs is recorded at $t = 2$ yr for a range of starting numbers of active TEs. Also, each subfigure shows the RdRP and antisense model, with and without siRNA transport (see Legend). The antisense model is not affected by siRNA transport; black lines are on top of each other. A. The fraction of silenced TEs. B. Total number of TEs, i.e. active plus silent. The gray line is a reference line that indicates the number of TEs present if there is no integration of new TE copies; this is equal to the number of TEs at the start.

we find that siRNA trafficking from cytoplasm to nucleus shifts the equilibrium of active and silent TEs towards the silent ones (Figure 5.3A). The ratio active-to-silent may change more than 2-fold for larger numbers of TEs. However, the total number of TEs that invade is larger in case of siRNA transport (Figure 5.3B, low starting numbers of TE). We can understand this as follows. Transport of siRNAs from cytoplasm to nucleus leads to a decrease in the cytoplasmic siRNA levels. In turn, a slightly weaker cytoplasmic silencing leads to elevated VLP abundance, which increases the rate of TE integration in the host genome. Thus what is gained by packing more TEs in heterochromatin, is lost in a lower number of mRNA being degraded by RISC.

Finally, we simulated the dynamics of TE invasions in a stochastic model (see Supporting information 1). The results of a single typical run are shown in Figure 5.2F. Due to the stochastic nature of the dynamics the time period of early exponential growth may now differ substantially between runs. In addition, we observe that the number of transposons after RNAi has been activated is lower in the stochastic than in the deterministic setting. In the long run the deterministic, continuous model and its stochastic counterpart converge on the numbers of silenced and active transposons, though at $t = 2$ yr an average over 100 runs shows that the stochastic model still contains fewer TEs and a lower VLP count (Table 5.2).

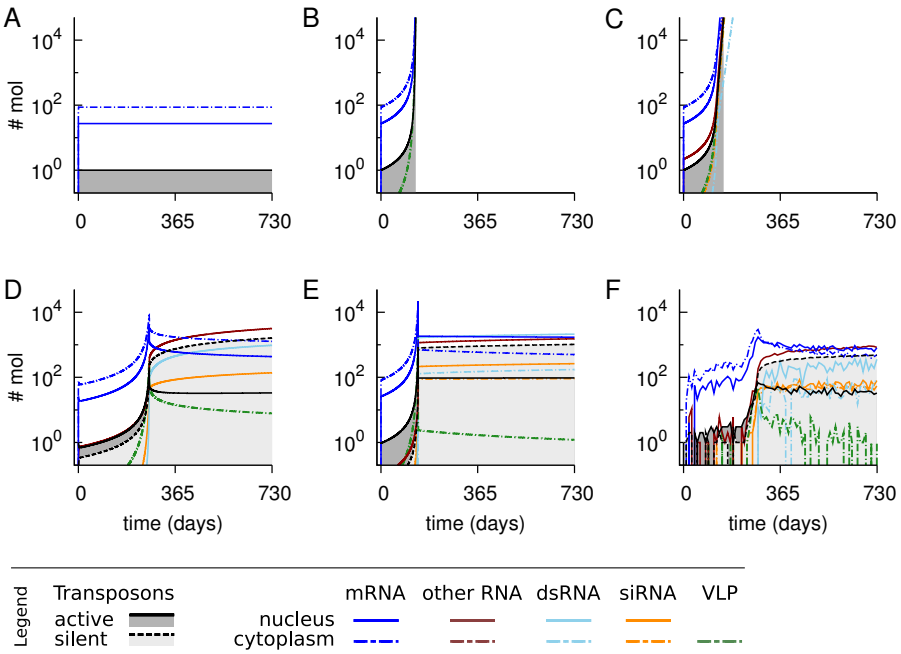

Figure 5.4 – Time plots of the antisense model. Time runs from 0 to 2 years, # *mol* is number of molecules. A. Transposon activity without silencing mechanisms and integration ($f = 0$). B. Exponential growth of transposons in a silencing-defective host. C. Post-transcriptional gene silencing (PTGS) in case active TEs also produce asRNA (which we only assume for this particular case). D. Transcriptional gene silencing (TGS), activated at $t \approx 100$. E. Both PTGS and TGS are active, and TGS is dominant (compare to C and D). There is no transport of siRNA between the cytoplasm and nucleus. F. Typical stochastic simulation with PTGS and TGS. Compared to E, the number of silenced TEs is higher and the abundance of VLPs lower.

5.3.2 Transposon invasion in RdRP-absent system

In the RdRP model silencing in both the nucleus and cytoplasm relies on the presence of RNA-dependent RNA Polymerase. RdRP is important for dsRNA production, amplification of siRNA and initiation of heterochromatin. However, in many animals RdRP appears not to be present, yet RNAi based silencing of transposons occurs. This implies that alternative pathways for dsRNA formation must be present. Naturally, the question arises how these animals silence TEs.

From observations on germline cells of mouse, fly and other animals an alternative silencing pathway is known: a Piwi-based strategy (Aravin *et al.*, 2007). It is based on RNA-coding genes – called Piwi genes – and a ping-pong model of amplification (Girard & Hannon, 2008). This allows for sustained transposon silencing without RdRP. However, the Piwi-based mechanism has only been found to be active in the germline and not in somatic tissues. Here we do not consider

Piwi-related silencing, instead we focus on (somatic) siRNA-based silencing if RdRP is not present.

As described in the Background there are several known pathways for dsRNA formation. Firstly, many transposons are transcribed in both sense and antisense direction (Okamura & Lai, 2008). Secondly, class II transposons often have inverted repeats, and on read-through may form hairpins (Okamura *et al.*, 2008). Considering that antisense transcription is both widespread and associated with heterochromatin, in both alternative models we now incorporate dsRNA production via sense-antisense duplex formation in the nucleus. This implies we now interpret 'other RNA' that is transcribed from TEs in heterochromatin as antisense RNA (asRNA). In the cytoplasm we differentiate between our alternative models: the antisense model has dsRNA production via mRNA and asRNA, both being exported from the nucleus, whereas the hairpin model has dsRNA formation via mRNA hairpins. Transport of asRNA has not been directly observed, but asRNA are reported as being fully capped and polyadenylated (Werner *et al.*, 2009). Such RNAs are capable of being transported to various locations in the nucleus and cytoplasm. The hairpin alternative does not change the modeling of dsRNA formation in our model, only the interpretation from parameter p_c reflecting RdRP activity to p_{cxx} reflecting hairpin formation rate (Eq. 5.7 and Eq. 5.15). Note that in both models we do not assume an amplification pathway in the cytoplasm. In addition, we do not alter the siRNA-based modulation of heterochromatization. The resulting models are given in the Methods.

For the antisense model we follow a similar approach as for the RdRP-based silencing model. We investigate cytoplasmic and nuclear silencing separately, followed by a study of their combined efforts. Since the hairpin model is a variant on the antisense model, we do not discuss it in such detail. For both alternative models we do verify our results with a stochastic model and compare them to the RdRP-based model.

PTGS, TGS and the combination in the 'antisense' model

If we use the default parameter values, post-transcriptional silencing of TEs does not occur in the alternative model. Instead exponential growth is observed (data not shown), due to our assumption that asRNA is only produced by TEs in heterochromatin and the fact that we do not allow for such nuclear silencing in this setting. However, even if active TEs also produce asRNA, there is still no cytoplasmic silencing (Figure 5.4C). Also allowing for a basal rate of heterochromatin formation and the reactivation of TEs ($h_b = 0.01, u = 0.02$) still results in exponential growth. In contrast to the RdRP-based model, the current model lacks the positive feedback loop of primed amplification of siRNAs, which makes TE control harder. It is important to note that this failure of silencing is not an intrinsic property of the alternative PTGS: there are parameter settings that allow for control. As an example, if we take sense and antisense RNA transcription rates equal, the host can control the transposons even though there is no amplification of siRNA (data not shown). Similar behavior was also found in the context of

	model	T_{act}	T_{sil}	V
transport	RdRP	179.90	110.85	0.71
	antisense	36.91	518.13	1.08
	hairpin	25.74	792.27	2.50
no transport	RdRP	171.94	86.41	0.50
	antisense	34.99	525.59	1.13
	hairpin	23.71	788.67	3.38
stochastic (no transport)	RdRP	155.16	77.47	0.33
	antisense	37.50	407.09	2.61
	hairpin	23.37	629.97	3.61

Table 5.2 – Transposons and their protein products at $t = 2$ yr. The columns are active and silent TEs (T_{act} and T_{sil}), and the number of VLPs (V). We distinguish three settings: the deterministic models with and without siRNA transport between nucleus and cytoplasm and the stochastic models without siRNA transport. For the stochastic models the values are averaged over 100 runs.

RNAi defense against viruses (Groenenboom & Hogeweg, 2008).

In contrast to the RdRP-based model, in which TGS allows for the integration of $\sim 10^5$ TEs, transcriptional gene silencing (TGS) is now capable of stopping the exponential growth of TEs at a much smaller number of TEs. In Figure 5.4D we observe that the magnitude of TE copies is in line with the number of copies we see for the full RdRP-based model (Figure 5.2E). This improvement of TGS is based on the fact that both asRNA and mRNA are consumed in the formation of dsRNA. This mRNA thus can no longer contribute to a VLP and hence a potential new incorporation of a transposon in the genome. Despite this improvement, nuclear silencing in the alternative model still does worse than the cytoplasmic silencing of the RdRP-based model, as the level of VLP remains relatively high (≈ 10 VLPs present at $t = 2$ yr).

Next we study the combination of PTGS and TGS. As shown in Figure 5.4E, TE activity is controlled well by the host. Also, VLP levels are decreased significantly by the cytoplasmic silencing. Considering the stability of this silencing under parameter changes, the model allows for increases in b without the negative effects we observed in the RdRP-based model. Instead a 10-fold increase of b strengthens the silencing. We find also that a 10-fold decrease of the basal heterochromatin formation hardly affects active TEs, but does result in a larger number of silenced TEs (around 1.5-fold increase). And because of the large number of silent TEs, there is a large pool of asRNAs, resulting in a strong cytoplasmic silencing and thus low levels of VLP (from 1.13 to 0.36).

Furthermore, in similar fashion as the RdRP-based model, we introduce siRNA transport. For our default parameters, we compare the fraction of silenced TEs and the total number of TEs that have invaded at $t = 2$ yr (Figure 5.3). In contrast to the RdRP model, however, we do not see any difference between the system

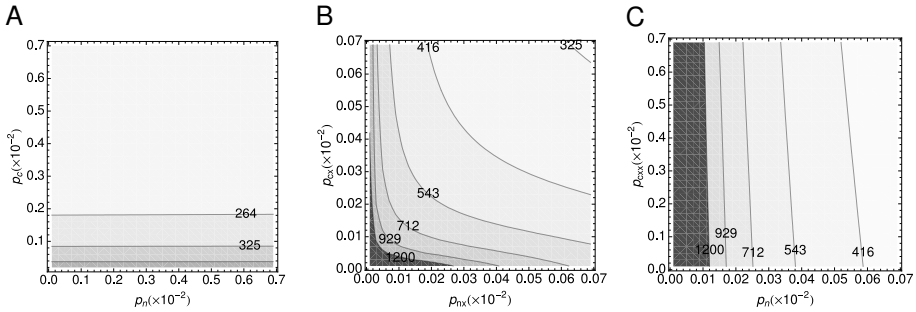

Figure 5.5 – Recruitment for RNAi silencing. Both figures display a contour map of the total number of TEs at $t = 2$ yr, starting with one transposon. Nuclear dsRNA formation is on the x axis, while cytoplasmic dsRNA formation is on the y axis. Note the different scale of the axis between the figures. A RdRP model. There is a distinct gradient of total number of TEs along the y axis. B. Antisense model, with a gradient of total number of TE along both axes. C. Hairpin model, with a clear gradient of total number of TEs along the x axis.

with and without transport. Instead, the remarkable differences are between the RdRP and antisense model. Firstly, especially for small starting numbers of TEs the antisense model allows for the invasion of many more transposon copies. These silent TEs are 'needed' for the production of asRNA in order to silence both in the nucleus and cytoplasm. Secondly, where the RdRP model shows a dependence on the number of active TEs that a simulation starts with, we find that in the alternative model over 90% of the transposons is silenced nearly invariant of this starting number. Only for large numbers of TEs do the two models converge on this behavior.

As a final item we simulated a stochastic version of the alternative model (Figure 5.4F). We observe that silencing is established earlier in terms of the number of active TEs. From this results a lower total number of TEs in the stochastic setting, even in runs that are twice as long (stop at $t = 4$ yr). By averaging over 100 runs, we verified that this is a robust result. We find that while in the deterministic model there are 34.99 active and 525, 59 silent TEs, the stochastic version has an average of 37.50 active and 407.09 silent TEs (see Table 5.2). Thus the deterministic model overestimates the number of transposons by > 100. Since on the longer term other processes, such as the beneficial and deleterious effects of transposons, certainly play an important role as well, an open question is if this difference has implications for a host's future.

Additional comparison between the models

We have reported that two silencing pathways, with RdRP-based dsRNA production and antisense-sense dsRNA formation, are capable of controlling TE activity. We find that the hairpin model silences transposons not as well as the other two.

If we look at Table 5.2, in all cases VLP levels are well above 2.00 for the hairpin model, which is a clear signal TEs are not tightly controlled and the integration of a new transposon copy will occur occasionally. One could hypothesize that if hosts indeed up-regulate TE activity under environmental stress (Slotkin & Martienssen, 2007), shutting down asRNA nuclear export would be a possible mechanism. That would effectively transform the antisense model, with decent TE silencing, into the hairpin model.

In our analysis above we find that RdRP-based transposon control has a bias towards cytoplasmic silencing, while the alternative models are biased towards nuclear silencing. To illustrate these observations, we computed the number of TEs that have invaded at $t = 2$ yr for a range of $p_{n(x)}$, $p_{c(x)}$ and p_{cxx} (i.e. varying the recruitment for both silencing mechanisms). Shown in Figure 5.5, there is a clear distinction between our models. In the RdRP model TE numbers show a clear gradient along p_c, the recruitment for cytoplasmic silencing. On the other hand, in the asRNA-based model the number of invaded transposons changes mostly with nuclear silencing, and cytoplasmic silencing is of secondary importance. In the hairpin model the cytoplasmic silencing is weaker than the asRNA alternative, which results in larger parameter ranges with high transposon levels. In addition we observe a clear gradient along p_{nx}.

5.4 Discussion and conclusion

Transposable elements are regulated in their activity by the host. We here present an initial approach of modeling such RNAi-based regulation. While the silencing of genes and transposable elements in the cytoplasm and nucleus is relatively well understood if the organism has RdRP, in absence of this protein complex it has remained unclear how the cellular silencing machinery functions (Buchon & Vaury, 2006). We report on an RdRP-based and two (hypothetical) asRNA-based model and find that TE activity is well-regulated. If an organism employs RdRP, cytoplasmic silencing dominates the dynamics of transposon expansion, while in the absence of RdRP the nuclear silencing mechanism dictates the behavior. In addition, if RdRP is present we find that siRNA transport between cytoplasm to nucleus can boost the nuclear silencing, though a cost is involved as cytoplasmic siRNA levels are lower and as a consequence PTGS is less effective.

Though RNAi is a powerful mechanism to harness transposons, the more this level of regulation becomes unraveled in different species, the more it becomes apparent that similar functionality, such as gene silencing or chromatin modifications, may be performed by different ensembles of protein complexes and pathways (Golden *et al.*, 2008; Okamura *et al.*, 2008). As an extreme example, while *S. pombe* is capable of RNAi via RdRP, Dicer and Argonaute, these proteins are simply absent in the modeling organism baker's yeast (*S. cerevisiae*). Still this species of yeast is capable of regulating its retrotransposon families. Apparently RNAi is not essential from an evolutionary point of view, and alternatives may come about (perhaps still based on RNA (Berretta *et al.*, 2008)).

With respect to our models, we realize heterochromatin formation has been approached from a rather high level. Since we find that the alternative model is based on a positive feedback loop in the formation of heterochromatin, a more detailed modeling of the nuclear silencing seems imperative. For instance, a feature we have not attended to is the impact of space in heterochromatin silencing. The initiation and spread of histone modifications along chromosomes is a defining factor in the process of establishing a region of heterochromatin. Furthermore, cell cycle dependence of histone modifications appears to play an essential role in the activity of TEs (Chen *et al.*, 2008; Kloc & Martienssen, 2008; Kloc *et al.*, 2008).

We hope that further experimental research will be done to provide additional data to verify our model. We would be interested in the function of transport of antisense transcripts and siRNAs, both between the cytoplasm and nucleus and within the nucleus. As mentioned before, one important assumption in our model is *trans*-acting siRNAs in the nucleus. This process is only starting to be unraveled (Iida *et al.*, 2008). Also, processing of sense-antisense pairs in both the nucleus and cytoplasm is an important factor in silencing without RdRP, but to our knowledge has not been extensively characterized in experiments.

Concluding, this work provides an exploratory modeling approach to transposon dynamics and the subsequent silencing via RNAi in nucleus and cytoplasm. We have shown that in organisms lacking RdRP viable alternative silencing pathways can be based on antisense transcription and asRNA transport, combined with the positive feedback caused by heterochromatin formation.

5.5 Supporting information 1

In this study we have predominantly used a set of ordinary differential equations (ODE) to describe the behavior of transposons and host control. However, it is well known that for small numbers of molecules intrinsic noise may play an important role in the dynamics. In order to verify our results with respect to small numbers of transposons, VLP levels and other molecules we also design and study a stochastic version of our model using the stochastic simulation algorithm, or Gillespie algorithm (Gillespie, 1977). Below we describe the pseudo-reactions of this approach. The reactions, with the corresponding parameters above the reaction arrows, are a straightforward translation from the original models as described in the Methods. Thus parameters are the same as used in the ODE model.

Transposon life cycle

$$T_{act} \xrightarrow{v_{ta}} T_{act} + M_n \qquad M_n \xrightarrow{d_m} \emptyset$$

$$T_{sil} \xrightarrow{v_{ts}} T_{sil} + R_n \qquad V \xrightarrow{j \cdot f} T_{act}$$

$$M_n \xrightarrow{t_m} M_c \qquad V \xrightarrow{j \cdot (1-f)} \emptyset$$

$$2M_c \xrightarrow{q} V \qquad V \xrightarrow{d_v} \emptyset$$

In the transposon life cycle we introduce one extra decay reaction. The decay of VLPs that fail to successfully integrate into the host genome is implicitly present in the original models, and needs to be explicitly modeled here.

Transcriptional gene silencing

$$T_{act} \xrightarrow{h_v} T_{sil} \qquad R_n + M_n \xrightarrow{p^*_{nx}} D_n$$

$$T_{act} + S_n \xrightarrow{h_s} T_{sil} \qquad R_n \xrightarrow{d_r} \emptyset$$

$$T_{sil} \xrightarrow{u} T_{act} \qquad D_n \xrightarrow{g_s} n\ S_n$$

$$R_n \xrightarrow{p_n} D_n \qquad S_n \xrightarrow{d_s,\ (v_s, k_s)} \emptyset$$

Post-transcriptional gene silencing

$$M_c \xrightarrow{p_c,\ p^*_{cxx}} D_c \qquad M_c + S_c \xrightarrow{b} \emptyset$$

$$M_c + R_c \xrightarrow{p^*_{cx}} D_c \qquad M_c \xrightarrow{d_m} \emptyset$$

$$M_c + S_c \xrightarrow{g} D_c \qquad R_c \xrightarrow{d_r} \emptyset$$

$$D_c \xrightarrow{g_s} n\ S_c \qquad S_c \xrightarrow{d_s,\ (v_s, k_s)} \emptyset$$

Both in TGS and PTGS the reactions with a star (*) are the alternative pathways of dsRNA formation, one via duplex formation of mRNA and asRNA and the other as hairpin formation of mRNA. Note that the decay of siRNA both in the nucleus and cytoplasm is composed of two reactions, one according to mass action kinetics (d_s), and one following saturating kinetics (v_s, k_s).

Transport of asRNA and siRNA

$$R_n \xrightarrow{t^*_{an}} R_c$$
$$S_c \xrightarrow{t_{sc}} S_n$$
$$S_n \xrightarrow{t_{sn}} S_c$$

Transport of 'other' RNA, R_n, across the nuclear envelop is marked with a star (*) as we only consider it in the antisense model.

5.6 Supporting information 2

Transposon life cycle The transcription rate of active transposons is estimated to be rather low: we take it a tenth of the mRNA influx of Groenenboom *et al.* (2005), that is $v_{ta} = 16\ \text{hr}^{-1}$. Silent transposons are assumed to be transcribed only occasionally, thus at an even lower rate than active ones: we take their transcription rate a 10-fold lower ($v_{ts} = 1.6\ \text{hr}^{-1}$).

Simple mass action mRNA transport is taken such that a ratio 1:3 between nucleus and cytoplasm is obtained in agreement with experimental observations (Jarmolowski *et al.*, 1994). Decay of mRNA ($d_m = 0.14\ \text{hr}^{-1}$) is taken from Groenenboom *et al.* (2005) and decay of 'other' RNA is assumed to be twice as high ($d_r = 0.28\ \text{hr}^{-1}$).

Finally, the stage of VLP production requires various proteins and processing steps. We assume it is the bottleneck of transposon dynamics, and hence we take $q = 1 \cdot 10^{-5}\ \#\text{mol}^{-1}\ \text{hr}^{-1}$.

Transcriptional gene silencing Parameters that are related to RNAi processes in the nucleus and that mirror RNAi in the cytoplasm (such as p_n and p_c) are taken to be equal to their cytoplasmic counterpart. See below.

To our knowledge estimates for the process of heterochromatin formation are not known (i.e. the time period that is needed for a gene or transposon to become packaged in heterochromatin). In order to arrive at a reasonable magnitude of the parameter values h_b, h_s and u, we perform the following estimation. We decompose heterochromatin formation into two steps: histone methylation and further packaging by proteins such as SWI6/HP1. First, enzyme kinetics of histone 3 lysine 9 (di)methylation, H3K9me, have been measured *in vitro* for the protein Dim5 of *Neurospora crassa*: a maximum turnover of $k_{cat} = 2.3\ \text{min}^{-1} = 138\ \text{hr}^{-1}$

was observed (Gowher *et al.*, 2005). Second, if we take the length of a transposon 6000 basepairs, it contains ~ 30 nucleosomes that may be methylated. Each nucleosome has 2 histone H3 proteins, thus in the order of 60-120 (di)methylations prepare the TE for packaging into heterochromatin. This brings us to a maximum of $138/120 = 1.15$ transposons per hour that become ready for transcriptional silencing. Next, SWI6/HP1 needs to bind the methylated nucleosomes. As a shortcut for determining how quickly a TE is packaged by SWI6/HP1, we take it to be in the same order of magnitude as histone methylation.

Due to other nuclear processes and crowding effects we expect that the maximum rate that we calculated here is usually not reached. Instead, we estimate the full process of heterochromatin formation should be in the range $[0.01, 0.10]$. This brings us to a basal heterochromatin formation rate of $h_b = 0.01\ \text{hr}^{-1}$ and activation – that is removal of SWI6/HP1 such that transcription may take place – by default twice as high ($u = 0.02\ \text{hr}^{-1}$). Thus we assume a rather negative scenario from the view point of silencing: the genome prefers to be open and transcribed. Furthermore, siRNA induced heterochromatization is taken $h_s = 0.001\ \#\text{mol}^{-1}\ \text{hr}^{-1}$, that is in the same order of magnitude as RISC activity.

Post-transcriptional gene silencing The parameters have been taken from literature (Groenenboom *et al.*, 2005; Kim & Yin, 2005). Furthermore, dsRNA formation via mRNA and asRNA ($p_{cx} = 2 \cdot 10^{-4}\ \#\text{mol}^{-1}\ \text{hr}^{-1}$) we estimate as a rare event, in-between RISC cleavage ($b = 8 \cdot 10^{-3}\ \#\text{mol}^{-1}\ \text{hr}^{-1}$) and the assembly of a VLP ($q = 1 \cdot 10^{-5}\ \#\text{mol}^{-1}\ \text{hr}^{-1}$). dsRNA formation via hairpins is taken to be equal to RdRP activity ($p_{cxx} = 0.002\ \text{hr}^{-1}$).

6 Discussion

The common denominators of this thesis are evolution in dynamic environments and mobile genetic elements, though not every chapter contains both of them. In chapters 2, 3 and 4 we study adaptive evolution under fluctuating environmental conditions. In the first two studies the conditions are externally imposed on the population. Conditions one can think of are for example the weather, climate or food abundance. We show that the commonly accepted framework of random mutation and natural selection allows for the at-times-disputed evolution of evolvability. A sequence of repeated short-term events may lead in the long term to the evolution of evolvability, where we define evolvability as the enhanced ability to discover beneficial, heritable adaptations.

In chapter 4 we switch from dictating environmental changes to letting the population shape its own environment. In the setting of resource processing we study the long-term evolutionary outcome of ecosystem evolution. We find that the majority of the simulations can be categorized in few typical evolutionary solutions. The long term ecosystem dynamics may be dominated by a single, 'smart' generalist, multiple specialized crossfeeding lineages and occasionally by an extreme specialist.

The other focal point of this thesis are mobile genetic elements, or transposons, which we study in chapters 2 and 5. In chapter 2 we investigate our hypothesis of an evolutionary beneficial role for transposons in evolutionary adaptation. Our results confirm this hypothesis and lead us to propose that increased evolvability may explain the maintenance of transposons in host genomes despite their large mutagenic power. Genomes may evolve a specific architecture – and employ transposons to do so – to allow for rapid adaptations to changed environments. In chapter 5 we study RNAi and how it is used by the host to control the activity of transposons in its genome. A key protein complex, RNA dependent RNA polymerase (RdRP), is thought to be essential in the standard pathway, yet is absent in many animals. We show that alternative pathways based on an-

tisense transcription are also readily capable of regulating transposon activity. Thus currently known interactions in the nucleus and cytoplasm are sufficient to explain transposon silencing, also if RdRP is not present.

6.1 A few requirements for evolvability

Evolution of evolvability becomes more and more an accepted concept (Pigliucci, 2008), however a question that remains is whether it arises through selection or is an accidental byproduct of the evolutionary process (Lynch, 2007; Pigliucci, 2008). In our models we observe the evolution of evolvability as the result of a second order selection pressure for integrating information on re-occurring environmental changes in the population. Starting with a naive population, mutations create variation, also with respect to being able to adapt swiftly and accurately to an alternate environment. Next, if the environment changes, individuals that are relatively fast at adapting (i.e. that need only few mutations), take over the population. Given that the same, or similar, environmental fluctuations occur repeatedly, over time the population experiences a selection pressure for swift and reliable changes in phenotype. Thus evolvable genotypes are selected for and evolvability may arise.

There are some preconditions to this phenomenon. Evolution of evolvability only occurs within a window of environmental change rates. If the environment varies rarely, once in a thousand generations or more, populations simply adapt to the new conditions and evolve robustness against mutations, rather than evolving evolvability. On the other extreme, if the rate of change of the environment is swift, in the order of ten generations, physiological regulation is expected to evolve. Evolution of evolvability can occur if environmental change occurs on an intermediate time scale (Meyers *et al.*, 2005).

Nevertheless, we observe that once a genome or network has been structured towards larger evolvability, the population survives also at much faster rates of environmental change in which evolvability would not evolve *a priori*. In this way the time scale of change at which evolvability may occur approaches the one of physiological regulation based on a sensor. To cope with the fast dynamics, population diversity is generated by the constant creation of mutants towards alternate environmental conditions. In prokaryotes this is known as phenotype switching (Dybvig, 1993). A different, elegant mechanism for phenotype switching not requiring mutations was investigated by Kashiwagi *et al.* (2006). They observed that if individuals are very unfit, the relatively noisy gene expression – due to low numbers of mRNA– may cause regulatory networks to switch attractor. As noise becomes less with higher expression rates, this creates an adaptive response to the environment.

In addition to the above mentioned time scales, population size and mutation rate are also important parameters for the evolution of evolvability. Similar to the evolution of robustness, which requires a large enough population size and high enough mutation rates (van Nimwegen *et al.*, 1999), the evolution of evolvability

also imposes certain requirements on population size and mutation rates. Lynch (2007) has proposed the concept of constructive genetic drift. The author hypothesizes that the elaborate genome and network architectures (especially in eukaryotes) are a result of drift in small populations: neutral or (slightly) deleterious expansions that could subsequently be used adaptively. Our results of chapter 2 can be interpreted within this framework of constructive drift. Transposons cause GCRs, which are virtually always extremely deleterious in our model. Still, selection is only occasionally able to remove these mobile genetic elements from the population. Thus with respect to transposons the population is drifting, and this is subsequently exploited by evolution, leading to an organized evolvable genome. These results thus suggest that with a rather small population constructive drift may occur. However, we also found that, for a given mutation rate, a too small population experiences too much neutral drift inbetween the alternating environmental conditions, resulting in a loss of evolvability. Therefore, we conclude that for evolvability to evolve, population sizes can be small but not too small, with an intermediate amount of genetic drift.

6.2 Evolvability of the genome and network

We show the evolution of evolvability on two levels, namely the genotype and the genotype-phenotype map. In chapter 2 we focus on the genotype: genome structuring by well-established mutational operators. The basic assumption in Darwinian evolution is that mutations are random events and natural selection acts on the phenotypic variation generated by them. In our study, mutations are, of course, still random events. However, due to transposon activity and gene duplication/deletion, the evolutionary process can shape genome organization such that some types of mutations occur more frequently than others, and some genomic locations are mutated more frequently than others. Interestingly, we see that the mutation types and locations that occur more frequently are those that are likely to be beneficial.

It is interesting to note that evolvability is intimately connected with the mutational operators. Based on experimental findings we defined chromosome rearrangements, taking into account the suggested fragility of LTRs and transposon dynamics. Despite the havoc these GCRs create, the evolutionary process subsequently exploited them in order to cope with the alternating environmental conditions. Thus the available set of mutations has determined the coding structure to a large extent.

In chapter 3 we study evolvability of gene regulatory networks. Again, we apply random, unbiased mutations. In this context, evolution of evolvability implies that the phenotypic effects of random mutations become biased by evolution. We find that only a small set of specific mutations orchestrate changes in a regulatory network such that the correct, alternate expression pattern is established. As a result these mutations are fixed in the population more often than others. So, rather than the preferential occurrence of certain mutations, as was

the case for the genotype evolvability we discussed previously, we get the preferential fixation of certain mutations during the evolution of genotype-phenotype map evolvability. Both the bias in mutation rates and the preferential fixation of a small set of mutations create a similar signature of biased occurrence in the ancestor trace. Of course, in real organisms structuring of both genome and network occurs in tandem. As a consequence, bias in mutation rates and fixation co-occur, rather than acting independently.

Furthermore, we find an unanticipated dynamic interplay between the network attractor landscape and the small set of beneficial mutations. Remarkably, we have been able to pinpoint these mutations: the insertion and deletion of a specific gene would cause the change in expression. We named this gene an evolutionary sensor. If we now represent two alternating environments as two network attractors, the following picture emerges. While a network resides in one of the attractors, the other one need not exist in the attractor landscape. However, as soon as an evolutionary sensor gene is duplicated (or deleted in the mirror case), the current attractor ceases to exist and the alternate attractor is created. Moreover, the network is now in the basin of attraction of the newly created attractor and a few gene expression updates will take it to the correct network state. Thus evolution has lead to networks with a specific topology that allows them to create and destruct specific attractors in a reversible fashion. Not only have these networks a 'known' current attractor landscape, also the mutational neighborhood has been structured in terms of specific attractors with favorably shaped basins of attraction.

Recent literature on the evolvability of gene regulatory networks has concentrated on static analysis of the networks and their potential for innovation and evolvability. Our dynamic view nicely adds to the finding that old attractors are rather easily kept, while new ones are added (Aldana *et al.*, 2007). And the maintained mutational neutrality that we find, is in confirmation with the extensive neutral networks of mutations on gene interactions that exist in regulatory networks (Ciliberti *et al.*, 2007a,b).

6.3 Modularity

One often mentioned feature in relation to evolvability is modularity. Modularity reflects how well a system can be decomposed into relatively independent parts, where the components within these parts are more tightly interconnected than components belonging to different parts (Barabási & Oltvai, 2004). In this discussion we define it as an architectural property. That is to say it is a static feature; we do not consider modularity that is visible in the dynamics of a system. Modularity affects both evolvability and robustness. The latter may be enhanced as a mutation in one module would not affect other modules (lower pleiotropy), restricting the range of effects of the mutation. On a similar note this increases evolvability, as evolution can experiment in one module without strongly influencing the other modules. Furthermore, a rewiring of the interactions between

modules would allow for a quick adaptation of the functionality of the system, further contributing to evolvability (Wagner *et al.*, 2007).

The organized genomes in chapter 2 are modular in the sense that genes are clustered by functional requirement. This can potentially be attributed to both the mutational operators, as mentioned previously, and the fact that the environmental changes can be decomposed into disjunct subsets. In contrast, although we have not explicitly evaluated the evolved networks of chapter 3 for their degree of modularity, our networks appear to be a single densely connected component. In the majority of simulations, similar to the situation in chapter 2, environmental conditions changed in a modular fashion. This suggests that a modular environment may not be sufficient for modularity to arise.

Thus it appears that high modularity is not simply a precondition to or result of all evolvable systems, but does sometimes arise. We started with two hypotheses for the evolution of modularity, "modular" mutational operators and "modular" environmental change. Given our results, modular mutational operators seem to be more strongly linked to modularity than modular environmental change. In this sense we might say it is a side-effect of the mutational operators (Solé & Valverde, 2008). However, in a study by Parter *et al.* (2008) where changes occurred in only a subset of the environmental conditions, thus resulting in modular environmental change, a strong link between evolvability and modularity was observed.

6.4 Direct and indirect selection

Throughout this thesis the concepts of direct and indirect selection play an important role. Direct selection refers to the defined fitness criterion which determines the probability an individual reproduces. In contrast, indirect selection refers to a long term effect that is better characterized as determining the survival of lineages, in our case through a structuring of the genome, network or environment.

In our studies evolvability comes about via indirect selection, as it arises over a time span of many generations by the evolution of populations adapting to fluctuating environmental dynamics. We can also make such a distinction between direct and indirect selection in chapter 4, where we studied an eco-evolutionary model in which individuals process resources.

Fitness is a function of the size of an individual's resource processing steps: bigger bites are better, thus determining direct selection pressure. As the waste of one individual becomes the food for the next individual, resource cycling arises and individuals shape the environment for the next generation. The result is a competition between lineages within the population for how well their offspring can process the resources that have been produced at previous time steps. In turn, this leads to an indirect selection pressure that stretches over several generations: it may actually be beneficial for a lineage that a bite remains at a certain size and does not become bigger, if the outcome is that the offspring can pro-

cess the resulting resource better. Such a limitation or reversal of direct selection by arising indirect selection pressures has been observed in various other eco-evolutionary models (Boerlijst & Hogeweg, 1991; Savill *et al.*, 1997).

We found that the type of ecosystem evolving in this model strongly depends on the interplay between these two levels of selection, one direct and one indirect, and the spatial setting of the environment, locality or global mixing. If over time individual lineages can continuously structure their local environment, lineages form patches in which they repeatedly encounter the same resources. Under these conditions indirect selection for cycling dominates. As a result evolution for long processing steps slows down, but evolution of smart individuals that are able to cycle resources on their own is facilitated. The evolution of such smart generalist individuals, that can handle conditions of which only a subset is encountered during the lifetime of a single individual has also been observed by Hillis (1990); Pagie & Hogeweg (1997); Hogeweg (2005). Key to this generalization is that the locality of spatial interactions ensures that a lineage samples from all conditions, allowing successful integration of information on the environment in the genotype.

If there is only a global feedback between the population and the resource distribution (global mixing), indirect selection for cycling is much weaker and evolution more often leads to crossfeeding, cooperative communities with a high degree of specialization of the various lineages. If lineages cannot structure their local environment, mutants that take larger bites can more easily invade as they do not pollute the environment of their own children with a resource that they cannot process. In practice the situation is a bit more complex, as both in the case of local and global feedback also an increase of direct selection enhances the formation of cooperative communities.

In Boerlijst & Hogeweg (1991); Savill *et al.* (1997) spatial patterns generated the indirect selection pressures. In our eco-evolutionary model the fact that resources need to be recycled shapes the secondary selection pressure. However, as discussed, the strength of its effect does depend on the locality of the processes. As a complementary case, evolvability (chapter 2 and 3) is an indirect selection that does not limit or reverse fitness gain, nor does it depend on locality. Instead direct fitness is augmented by evolvability.

6.5 Modeling transposable elements

As mentioned in the Introduction transposable elements are powerful mutagenic agents. Their origins trace back to the early roots of eukaryotes and prokaryotes, implying a long history of coevolution between the host genome and its "parasitic" elements.

Modeling the evolution of genomes with transposable elements has often been based on population genetics and placed in a setting of sexually reproducing populations (Charlesworth *et al.*, 1994). In such models, the invasion and maintenance of transposons has been explained as a balance between transpos-

ition and recombination (Rouzic & Capy, 2005). In clonally reproducing populations, such as we studied in chapter 2, the occurrence of transposons requires that selection plays a positive role (Edwards & Brookfield, 2003; Dolgin & Charlesworth, 2006).

In contrast, we do not assume that transposons have a beneficial effect on host fitness. We simply initiate the hosts with a small set of TEs and let the population evolve in a dynamically varying environment. The *a priori* expectation would be that transposons are removed from the host, as the GCRs caused by them increase mutational load. In contrast, we find that transposons rather easily invade and are kept over long time periods. The population of hosts actually evolves to employ the mutational effects of transposons to adapt swiftly to the changing environment. Our study thus provides a proof-of-principle of how hosts may evolve to take advantage of transposons.

We would like to stress that our modeling approach is one of few that offer a constructive explanation for the presence of transposons (another example being Quesneville & Anxolabéhère (2001)). By this we mean that we explicitly implement transposons and their dynamics and therewith we study their indirect effects on genome evolvability, rather than having a parameter that gives transposons a certain "beneficial" value – see for instance Rouzic *et al.* (2007) – which in turn would allow for transposon persistence.

6.6 Regulating transposable elements

Transposons and related elements have contributed to many adaptations in the evolution of plants, animals and other eukaryotes (Kazazian, 2004; Biemont & Vieira, 2006; Muotri *et al.*, 2007), though for instance the ascribed functionality of Alu elements may also be explained partly as a transposition bias (Urrutia *et al.*, 2008). In any event, if their copying activity is not under some control by the host, accumulating transposon copies in the genome are bound to be detrimental to the host's fitness. Of course, given the above section one could argue that individual control is not needed as on a population level natural selection may act against individuals with a high transposon load. True as that may be, in the last decade RNAi mechanisms have been discovered that are used by the host to control its TEs. In chapter 5 we developed and investigated a mathematical model of the control of transposons by the host using two RNAi mechanisms, namely transcriptional gene silencing in the nucleus based on heterochromatin formation, and post-transcriptional gene silencing in a cell's cytoplasm by means of mRNA degradation.

Given the scattered presence of various RNAi-related proteins among species, it is likely that there are various alternative pathways for regulating TE activity. Moreover, different transposons require different regulatory strategies (Lippman *et al.*, 2003). We studied three alternative pathways, one based on RdRP, which has a firm basis in experimental data, and two (hypothetical) pathways based on experimentally observed antisense transcription. In these two alternative mod-

els, epigenetic, nuclear silencing is based on sense-antisense double-strand formation (Iida *et al.*, 2008). Cytoplasmic RNAi, however, is different between the alternative models. One assumes antisense RNA transport into the cytoplasm for double-strand formation (Werner *et al.*, 2009), while the other assumes only hairpin formation of cytoplasmic mRNA (Sijen & Plasterk, 2003). We find that in all cases silencing of TEs is based on a positive feedback loop: in the RdRP model cytoplasmic amplification of siRNA determines the dynamics, while in both antisense models nuclear amplification dominates TE silencing.

Hosts use similar RNAi based mechanisms to control transposons and viruses. As transposons have a genomic DNA copy of their code, controlling or clearing a TE may be a more difficult task than clearing an RNA virus. To determine whether this is the case, we need to extend the current study with a more thorough analysis of the transposon copying and siRNA amplification loops, performing for example parameter bifurcation analyses. This would allow for a detailed comparison of our TE regulation model with a recent model on viral infection and the host's RNAi immune response (Groenenboom & Hogeweg, 2008). Another idea is to evolve *in silico* the parameters and interactions of a model TE regulatory network. This would allow us to establish "evolutionary stable strategies" of the regulation of transposons, and to investigate how these strategies depend on the properties of different transposons. This approach has been pioneered by van Hoek & Hogeweg (2006, 2007) in the context of *lac* operon regulation.

6.7 Outlook

We have modeled a series of related phenomena in this thesis ranging from genome organization to network structuring, and to ecosystem evolution. On the intersections of these research areas lie interesting directions for future work.

An interesting direction is, for instance, to combine our studies on evolvability: to study evolution in the presence of both genome order (chapter 2) and a gene regulatory network (chapter 3), and transposons potentially reshaping both of them. After all, baker's yeast has both a specific gene order and a highly structured gene regulatory network. Moreover, yeast senses many environmental changes. Allowing for such physiological regulation provides an opportunity to study various conditions e.g. different space and time scales. This would also allow us to construct a more integrated view on transposons and their effects on the host. We could address questions such as: what mutational side-effects does transposon activity have on genome order and network topology, what minimal requirements must be met for transposons to be of adaptive value (i.e. transposons are detrimental to a host, why does the host keep them?), is there still a clear genome structure if there is a regulatory network in-between genotype and phenotype. And what is the relative importance of genome order and network structure in the evolution of evolvability?

Another, related, topic for future research could be the evolution of transpo-

son control by the host (chapter 5). Given the advances in computing power soon it will be feasible to simulate the evolution of a population of individuals with stochastic internal dynamics, for instance modeling individuals with RNAi-inspired transposon regulation. Perhaps embedding the individuals in a simplified version of the above suggested model, this would allow us to investigate the evolution of transposon activity and host control. The outcome of such *in silico* evolution could then be mapped onto the various RNAi pathways that are found in different species.

In addition, in recent years it has become apparent from transcriptional and post-transcriptional regulation of transposons (chapter 5) and other genes by RNAi that eukaryotes – and most likely prokaryotes as well – have a RNA-based regulatory layer. The idea would be to explore the properties of such RNA regulatory networks by modeling a gene regulatory network on the basis of a genome producing various RNA transcripts, RNA template matching, and cellular machinery based on various proteins that we observe in RNAi (leading to activation and inhibition of transcripts, RNA cleavage, amplification and so on). An interesting question is how the evolution of such a network differs from networks with protein mediated regulation. What evolutionary patterns arise? How do concepts such as evolvability, modularity, and robustness apply to such networks? The modeling approach by Knibbe *et al.* (2007) provides a useful starting point.

Finally, with respect to the evolution of ecosystem complexity, it is attractive to extend our eco-evolutionary study (chapter 4) to a more open-ended evolutionary regime with multiple resource cycles, or in which new resources can arise – for instance through innovations by individuals – or by regarding individuals as resources as well. This would allow for a broader range in the evolution of both direct and indirect interactions among individuals. In other studies it has been shown that this can serve as a rich ground for increases in ecosystem and organismal complexity (i.e. speciation, parasitism, individual smartness) (Ray, 1991; Lindgren, 1991; Pagie, 1999; Takeuchi & Hogeweg, 2008).

6.8 Conclusion

In this thesis we have studied evolution in dynamical environments and mobile genetic elements from several viewpoints. We have shown the evolution of evolvability twice in a setting of imposed environmental changes. Once via genome structure and mutational bias – mutational priming by means of transposons – and once via the structuring of the gene regulatory network leading to a bias in the phenotypic effect of mutations.

If a population can shape its own environment, depending on spatial setting and balance between direct and indirect selection, a range of different eco-evolutionary trajectories can occur: from stable resource dynamics and smart, self-sufficient individuals to turbulent resource patterns and cooperative communities.

Finally, we return to transposons and how the host regulates them. We show that the lack of an apparent key protein in the RNAi control pathways of many organisms can be explained by putting the emphasis of control in the nucleus instead of cytoplasm, an alternative that has a firm basis in available experimental data.

Bibliography

Aldana M, Balleza E, Kauffman SA & Resendiz O. Robustness and evolvability in genetic regulatory networks. *J Theor Biol* 245: 433–448 (2007).

Alizon S, Kucera M & Jansen VAA. Competition between cryptic species explains variations in rates of lineage evolution. *Proc Natl Acad Sci U S A* 105: 12382–12386 (2008).

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796–815 (2000).

Aravin AA, Hannon GJ & Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318: 761–764 (2007).

Barabási AL & Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5: 101–113 (2004).

Batada NN, Urrutia AO & Hurst LD. Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet* 23: 480–484 (2007).

Baulcombe DC. Molecular biology. Amplified silencing. *Science* 315: 199–200 (2007).

Beauregard A, Curcio MJ & Belfort M. The take and give between retrotransposable elements and their hosts. *Annu Rev Genet* 42: 587–617 (2008).

Berezhna SY, Supekova L, Supek F, Schultz PG & Deniz AA. siRNA in human cells selectively localizes to target RNA sites. *Proc Natl Acad Sci U S A* 103: 7682–7687 (2006).

Berretta J, Pinskaya M & Morillon A. A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in S. cerevisiae. *Genes Dev* 22: 615–626 (2008).

Biemont C & Vieira C. Genetics: Junk DNA as an evolutionary force. *Nature* 443: 521–524 (2006).

Boerlijst MC & Hogeweg P. Self-structuring and selection: spiral waves as a substrate for evolution. In *Artificial Life II*, edited by C G Langton, C Taylor, J D Farmer & S Rasmussen, pp. 255–276. Addison-Wesley (1991).

Britten RJ, Rowen L, Williams J & Cameron RA. Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci U S A* 100: 4661–4665 (2003).

Brown CJ, Todd KM & Rosenzweig RF. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol Biol Evol* 15: 931–42 (1998).

Buchler NE, Gerland U & Hwa T. Nonlinear protein degradation and the function of genetic circuits. *Proc Natl Acad Sci U S A* 102: 9559–9564 (2005).

Buchon N & Vaury C. RNAi: a defensive RNA-silencing against viruses and transposable elements. *Heredity* 96: 195–202 (2006).

Cam HP, Noma K, Ebina H, Levin HL & Grewal SIS. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* 451: 431–436 (2008).

Cavalier-Smith T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot (Lond)* 95: 147–175 (2005).

Cha RS & Kleckner N. ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. *Science* 297: 602–606 (2002).

Charlesworth B, Sniegowski P & Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371: 215–220 (1994).

Chen ES, Zhang K, Nicolas E, Cam HP, Zofall M & Grewal SIS. Cell cycle control of centromeric repeat transcription and heterochromatin assembly. *Nature* 451: 734–737 (2008).

Cheng Z & Menees TM. RNA branching and debranching in the yeast retrovirus-like element Ty1. *Science* 303: 240–243 (2004).

Chivian D, Brodie EL, Alm EJ, Culley DE, Dehal PS, Desantis TZ, Gihring TM, Lapidus A, Lin L, Lowry SR, Moser DP, Richardson PM, Southam G, Wanger G, Pratt LM, Andersen GL, Hazen TC, Brockman FJ, Arkin AP & Onstott TC. Environmental genomics reveals a single-species ecosystem deep within Earth. *Science* 322: 275–278 (2008).

Chow SS, Wilke CO, Ofria C, Lenski RE & Adami C. Adaptive radiation from resource competition in digital organisms. *Science* 305: 84–86 (2004).

Chung WJ, Okamura K, Martin Rl & Lai EC. Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Curr Biol* 18: 795–802 (2008).

Ciliberti S, Martin OC & Wagner A. Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A* 104: 13591–13596 (2007a).

Ciliberti S, Martin OC & Wagner A. Robustness Can Evolve Gradually in Complex Regulatory Gene Networks with Varying Topology. *PLoS Comput Biol* 3: e15 (2007b).

Clausen J, Keck D & Hiesey W. Experimental Studies in the Nature of Species, vol. 3: Environmental Responses of Climatic Races of Achillea. *Carnegie Institution of Washington Publication* 581: 1–129 (1958).

Cohen BA, Mitra RD, Hughes JD & Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* 26: 183–186 (2000).

Colmenares SU, Buker SM, Buhler M, Dlakić M & Moazed D. Coupling of double-stranded RNA synthesis and siRNA generation in fission yeast RNAi. *Mol Cell* 27: 449–461 (2007).

Conley AB, Piriyapongsa J & Jordan IK. Retroviral promoters in the human genome. *Bioinformatics* 24: 1563–1567 (2008).

Cordero OX & Hogeweg P. Feed-forward loop circuits as a side effect of genome evolution. *Mol Biol Evol* 23: 1931–1936 (2006).

Cremer T & Cremer C. Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur J Histochem* 50: 161–176 (2006a).

Cremer T & Cremer C. Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present. *Eur J Histochem* 50: 223–272 (2006b).

Darai-Ramqvist E, Sandlund A, Müller S, Klein G, Imreh S & Kost-Alimova M. Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. *Genome Res* 18: 370–379 (2008).

Darwin C. *On the Origin of Species by Means of Natural Selection*. J Murray (1859).

de Boer F & Hogeweg P. Specialization and collective problem solving. *In preparation* (2009).

Deininger PL, Moran JV, Batzer MA & Kazazian HH. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* 13: 651–658 (2003).

Dennis C. Mouse genome: a forage in the junkyard. *Nature* 420: 458–459 (2002).

Dolgin ES & Charlesworth B. The fate of transposable elements in asexual populations. *Genetics* 174: 817–827 (2006).

Draghi J & Wagner GP. Evolution of evolvability in a developmental model. *Evolution* 62: 301–315 (2008).

van Driel R, Fransz PF & Verschure PJ. The eukaryotic genome: a system regulated at different hierarchical levels. *J Cell Sci* 116: 4067–4075 (2003).

Drossel B., Higgs P. G. & McKane A. J. The influence of predator–prey population dynamics on the long-term evolution of food web structure. *J Theor Biol* 208: 91–107 (2001).

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, Montigny J De, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisramé A, Boyer J, Cattolico L, Confanioleri F, Daruvar A De, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wésolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P & Souciet JL. Genome evolution in yeasts. *Nature* 430: 35–44 (2004).

Dunbar J, Barns SM, Ticknor LO & Kuske CR. Empirical and theoretical bacterial diversity in four Arizona soils. *Appl Environ Microbiol* 68: 3035–3045 (2002).

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F & Botstein D. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 99: 16144–9 (2002).

Dworkin J & Blaser MJ. Nested DNA inversion as a paradigm of programmed gene rearrangement. *Proc Natl Acad Sci U S A* 94: 985–990 (1997).

Dybvig K. DNA rearrangements and phenotypic switching in prokaryotes. *Mol Microbiol* 10: 465–471 (1993).

Edwards RJ & Brookfield JFY. Transiently Beneficial Insertions Could Maintain Mobile DNA Sequences in Variable Environments. *Mol. Biol. Evol.* 20: 30–37 (2003).

Eichler EE & Sankoff D. Structural dynamics of eukaryotic chromosome evolution. *Science* 301: 793–797 (2003).

Eldredge N, Thompson JN, Brakefield PM, Gavrilets S, Jablonski D, Jackson JBC, Lenski RE, Lieberman BS, McPeek MA & MillerW III. The dynamics of evolutionary stasis. *Paleobiology* 31: 133–145 (2005).

ENCODE Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816 (2007).

Falkowski PG, Fenchel T & Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science* 320: 1034–1039 (2008).

Feng YX, Moore SP, Garfinkel DJ & Rein A. The genomic RNA in Ty1 virus-like particles is dimeric. *J Virol* 74: 10819–10821 (2000).

Ferea TL, Botstein D, Brown PO & Rosenzweig RF. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci U S A* 96: 9721–6 (1999).

Fernández P & Solé RV. Neutral fitness landscapes in signalling networks. *J R Soc Interface* 4: 41–47 (2007).

Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9: 397–405 (2008).

Feschotte C & Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41: 331–368 (2007).

Fischer G, James SA, Roberts IN, Oliver SG & Louis EJ. Chromosomal evolution in Saccharomyces. *Nature* 405: 451–454 (2000).

Fischer G, Neuvéglise C, Durrens P, Gaillardin C & Dujon B. Evolution of gene order in the genomes of two related yeast species. *Genome Res* 11: 2009–19 (2001).

Fontana W, Konings DA, Stadler PF & Schuster P. Statistics of RNA secondary structures. *Biopolymers* 33: 1389–1404 (1993a).

Fontana W & Schuster P. Continuity in evolution: on the nature of transitions. *Science* 280: 1451–1455 (1998).

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

Forbes EM, Nieduszynska SR, Brunton FK, Gibson J, Glover LA & Stansfield I. Control of gag-pol gene expression in the Candida albicans retrotransposon Tca2. *BMC Mol Biol* 8: 94 (2007).

François P & Hakim V. Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci U S A* 101: 580–585 (2004).

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW & Delong EF. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* 105: 3805–3810 (2008).

Gama-Castro S, Jiménez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Peñaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muñiz-Rascado L, Martínez-Flores I, Salgado H, Bonavides-Martínez C, Abreu-Goodger C, Rodríguez-Penagos C, Miranda-Ríos J, Morett E, Merino E, Huerta AM, Treviño-Quintanilla L & Collado-Vides J. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 36: D120–D124 (2008).

Gellert M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem* 71: 101–132 (2002).

Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z & Zamore PD. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science* 320: 1077–1081 (2008).

Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemisty* 81: 2340–61 (1977).

Girard A & Hannon GJ. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol* 18: 136–148 (2008).

Goffeau A, Aert R, Agostini-Carbone L, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M & D Alexandraki et al. The yeast genome directory. *Nature* 387: 5 (1997).

Golden DE, Gerbasi VR & Sontheimer EJ. An inside job for siRNAs. *Mol Cell* 31: 309–312 (2008).

Goodier JL & Kazazian HH. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* 135: 23–35 (2008).

Gowher H, Zhang X, Cheng X & Jeltsch A. Avidin plate assay system for enzymatic characterization of a histone lysine methyltransferase. *Anal Biochem* 342: 287–291 (2005).

Grewal SIS & Elgin SCR. Transcription and RNA interference in the formation of heterochromatin. *Nature* 447: 399–406 (2007).

Grewal SIS & Jia S. Heterochromatin revisited. *Nat Rev Genet* 8: 35–46 (2007).

Grimaud C, Bantignies F, Pal-Bhadra M, Ghana P, Bhadra U & Cavalli G. RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* 124: 957–971 (2006).

Groenenboom MAC & Hogeweg P. The dynamics and efficacy of antiviral RNA silencing: A model study. *BMC Syst Biol* 2: 28 (2008).

Groenenboom MAC, Marée AFM & Hogeweg P. The RNA silencing pathway: the bits and pieces that matter. *PLoS Comput Biol* 1: 155–165 (2005).

Guang S, Bochner AF, Pavelec DM, Burkhart KB, Harding S, Lachowiec J & Kennedy S. An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* 321: 537–541 (2008).

Gudelj I, Beardmore RE, Arkin SS & MacLean RC. Constraints on microbial metabolism drive evolutionary diversification in homogeneous environments. *J Evol Biol* 20: 1882–1889 (2007).

Guill Christian & Drossel Barbara. Emergence of complexity in evolving niche-model food webs. *J Theor Biol* 251: 108–120 (2008).

Han K, Lee J, Meyer TJ, Remedios P, Goodwin L & Batzer MA. L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A* 105: 19366–19371 (2008).

Häsler J & Strub K. Alu elements as regulators of gene expression. *Nucleic Acids Res* 34: 5491–5497 (2006).

Helling RB, Vargas CN & Adams J. Evolution of Escherichia coli during growth in a constant environment. *Genetics* 116: 349–358 (1987).

Hillis WD. Co-evolving parasites improve simulated evolution as an optimization procedure. *Phys D* 42: 228–234 (1990).

van Hoek MJA & Hogeweg P. In Silico Evolved lac Operons Exhibit Bistability for Artificial Inducers, but Not for Lactose. *Biophys. J.* 91: 2833–2843 (2006).

van Hoek MJA & Hogeweg P. The effect of stochasticity on the lac operon: an evolutionary perspective. *PLoS Comput Biol* 3: e111 (2007).

Hogeweg P. *Self-organisation and Evolution of Social Systems*, chapter Interlocking of self-organisation and evolution, pp. 166–189. Cambridge University Press (2005).

Hogeweg P. From population dynamics to ecoinformatics: Ecosystems as multilevel information processing systems. *Ecological Informatics* 2: 103–111 (2007).

Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, Yasuhara JC, Wakimoto BT, Myers EW, Celniker SE, Rubin GM & Karpen GH. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol* 3: RESEARCH0085 (2002).

Hughes AL & Friedman R. Transposable element distribution in the yeast genome reflects a role in repeated genomic rearrangement events on an evolutionary time scale. *Genetica* 121: 181–185 (2004).

Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, Friend SH & Marton MJ. Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat Genet* 25: 333–337 (2000).

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ & Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320: 1081–1085 (2008).

Hurst LD, Pál C & Lercher MJ. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* 5: 299–310 (2004).

Huynen MA. *Evolutionary Dynamics and Pattern Generation in the Sequence and Secondary Structure of RNA*. Ph.D. thesis, Universiteit Utrecht (1993).

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

Huynen MA, Dandekar T & Bork P. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol* 7: 281–291 (1999).

Iida T, Nakayama J & Moazed D. siRNA-mediated heterochromatin establishment requires HP1 and is associated with antisense transcription. *Mol Cell* 31: 178–189 (2008).

Infante JJ, Dombek KM, Rebordinos L, Cantoral JM & Young ET. Genome-wide amplifications caused by chromosomal rearrangements play a major role in the adaptive evolution of natural yeast. *Genetics* 165: 1745–59 (2003).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921 (2001).

Irvine DV, Zaratiegui M, Tolia NH, Goto DB, Chitwood DH, Vaughn MW, Joshua-Tor L & Martienssen RA. Argonaute slicing is required for heterochromatic silencing and spreading. *Science* 313: 1134–1137 (2006).

Jacob F & Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3: 318–356 (1961).

Jacob F, Perrin D, Sanchez C & Monod J. Operon: a group of genes with the expression coordinated by an operator. *C R Hebd Seances Acad Sci* 250: 1727–1729 (1960).

Jarmolowski A., Boelens W. C., Izaurralde E. & Mattaj I. W. Nuclear export of different classes of RNA is mediated by specific factors. *J Cell Biol* 124: 627–635 (1994).

Johnson TJ & Wilke CO. Evolution of resource competition between mutually dependent digital organisms. *Artif Life* 10: 145–156 (2004).

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, Ashburner M & Celniker SE. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* 3: RESEARCH0084 (2002).

Kaneko K & Yomo T. Sympatric speciation: compliance with phenotype diversification from a single genotype. *Proc Biol Sci* 267: 2367–2373 (2000).

Kano H, Kurahashi H & Toda T. Genetically regulated epigenetic transcriptional activation of retrotransposon insertion confers mouse dactylaplasia phenotype. *Proc Natl Acad Sci U S A* 104: 19034–19039 (2007).

Kapitonov VV & Jurka J. Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A* 98: 8714–8719 (2001).

Kashi Y & King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet* 22: 253–259 (2006).

Kashiwagi A, Urabe I, Kaneko K & Yomo T. Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. *PLoS ONE* 1: e49 (2006).

Kashtan N & Alon U. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A* 102: 13773–13778 (2005).

Kashtan N, Noor E & Alon U. Varying environments can speed up evolution. *Proc Natl Acad Sci U S A* 104: 13711–13716 (2007).

Kassen R & Rainey PB. The ecology and genetics of microbial diversity. *Annu Rev Microbiol* 58: 207–231 (2004).

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C, Group RIKEN Genome Exploration Research, Group) Genome Science Group (Genome Network Project Core & Consortium FANTOM. Antisense transcription in the mammalian transcriptome. *Science* 309: 1564–1566 (2005).

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

Kazazian HH. Mobile elements: drivers of genome evolution. *Science* 303: 1626–1632 (2004).

Kim H & Yin J. Robust growth of human immunodeficiency virus type 1 (HIV-1). *Biophys J* 89: 2210–2221 (2005).

Kloc A & Martienssen R. RNAi, heterochromatin and the cell cycle. *Trends Genet* 24: 511–517 (2008).

Kloc A, Zaratiegui M, Nora E & Martienssen R. RNA interference guides histone modification during the S phase of chromosomal replication. *Curr Biol* 18: 490–495 (2008).

Knibbe C, Coulon A, Mazet O, Fayard JM & Beslon G. A long-term evolutionary pressure on the amount of noncoding DNA. *Mol Biol Evol* 24: 2344–2353 (2007).

Knibbe C, Fayard JM & Beslon G. The topology of the protein network influences the dynamics of gene order: from systems biology to a systemic understanding of evolution. *Artif Life* 14: 149–156 (2008).

Koonin EV. Evolution of genome architecture. *Int J Biochem Cell Biol* 41: 298–306 (2009).

Koszul R, Caburet S, Dujon B & Fischer G. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* 23: 234–243 (2004).

Kuo PD, Banzhaf W & Leier A. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems* 85: 177–200 (2006).

Lagomarsino MC, Jona P, Bassetti B & Isambert H. Hierarchy and feedback in the evolution of the Escherichia coli transcription network. *Proc Natl Acad Sci U S A* 104: 5516–5520 (2007).

Lambert S, Watson A, Sheedy DM, Martin B & Carr AM. Gross chromosomal rearrangements and elevated recombination at an inducible site-specific replication fork barrier. *Cell* 121: 689–702 (2005).

Lavrov SA & Kibanov MV. Noncoding RNAs and chromatin structure. *Biochemistry (Mosc)* 72: 1422–1438 (2007).

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK & Young RA. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298: 799–804 (2002).

Lesage P & Todeschini AL. Happy together: the life and times of Ty retrotransposons and their hosts. *Cytogenet Genome Res* 110: 70–90 (2005).

Lewontin R. The Genotype/Phenotype Distinction. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta (Fall 2008).

Ley RE, Peterson DA & Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124: 837–848 (2006).

Li F, Long T, Lu Y, Ouyang Q & Tang C. The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci U S A* 101: 4781–4786 (2004).

Lindgren K. Evolutionary Phenomena in Simple Dynamics. In *Artificial Life II*, edited by C G Langton, C Taylor, J D Farmer & S Rasmussen, pp. 295–311. Addison-Wesley (1991).

Lippman Z, May B, Yordan C, Singer T & Martienssen R. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. *PLoS Biol* 1: E67 (2003).

Loeuille Nicolas & Loreau Michel. Evolutionary emergence of size-structured food webs. *Proc Natl Acad Sci U S A* 102: 5761–5766 (2005).

Luan DD, Korman MH, Jakubczak JL & Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 72: 595–605 (1993).

Lynch M. The origins of eukaryotic gene structure. *Mol Biol Evol* 23: 450–468 (2006).

Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A* 104 Suppl 1: 8597–8604 (2007).

Lynch M & Conery JS. The origins of genome complexity. *Science* 302: 1401–1404 (2003).

Lynch VJ & Wagner GP. Multiple chromosomal rearrangements structured the ancestral vertebrate Hox-bearing protochromosomes. *PLoS Genet* 5: e1000349 (2009).

Lysnyansky I, Rosengarten R & Yogev D. Phenotypic switching of variable surface lipoproteins in Mycoplasma bovis involves high-frequency chromosomal rearrangements. *J Bacteriol* 178: 5395–5401 (1996).

MacLean RC. The tragedy of the commons in microbial populations: insights from theoretical, comparative and experimental studies. *Heredity* 100: 471–477 (2008).

MacLean RC & Gudelj I. Resource competition and social conflict in experimental populations of yeast. *Nature* 441: 498–501 (2006).

Maharjan R, Seeto S, Notley-McRobb L & Ferenci T. Clonal adaptive radiation in a constant environment. *Science* 313: 514–517 (2006).

Mangan S & Alon U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* 100: 11980–11985 (2003).

Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* 25: 930–939 (2003).

McClintock B. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A* 36: 344–355 (1950).

McClintock B. Induction of instability at selected loci in maize. *Genetics* 38: 579–599 (1953).

Meaburn KJ & Misteli T. Cell biology: chromosome territories. *Nature* 445: 379–781 (2007).

Meyers BC, Tingey SV & Morgante M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res* 11: 1660–1676 (2001).

Meyers LA, Ancel FD & Lachmann M. Evolution of genetic potential. *PLoS Comput Biol* 1: 236–243 (2005).

Mieczkowski PA, Lemoine FJ & Petes TD. Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. *DNA Repair (Amst)* 5: 1010–1020 (2006).

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D & Alon U. Network motifs: simple building blocks of complex networks. *Science* 298: 824–827 (2002).

Mou X, Sun S, Edwards RA, Hodson RE & Moran MA. Bacterial carbon processing by generalist species in the coastal ocean. *Nature* 451: 708–711 (2008).

Munteanu A & Solé RV. Neutrality and robustness in evo-devo: emergence of lateral inhibition. *PLoS Comput Biol* 4: e1000226 (2008).

Muotri AR, Marchetto MCN, Coufal NG & Gage FH. The necessary junk: new functions for transposable elements. *Hum Mol Genet* 16 Spec No. 2: R159–R167 (2007).

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, Hitte C, Meyers SN, Milan D, Ostrander EA, Pape G, Parker HG, Raudsepp T, Rogatcheva MB, Schook LB, Skow LC, Welge M, Womack JE, O'brien SJ, Pevzner PA & Lewin HA. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309: 613–617 (2005).

van Nimwegen E & Crutchfield JP. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull Math Biol* 62: 799–848 (2000).

van Nimwegen E, Crutchfield JP & Huynen M. Neutral evolution of mutational robustness. *Proc Natl Acad Sci U S A* 96: 9716–9720 (1999).

van Noort V, Snel B & Huynen MA. Predicting gene function by conserved co-expression. *Trends Genet* 19: 238–42 (2003).

Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP & Lai EC. The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* 453: 803–806 (2008).

Okamura K & Lai EC. Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* 9: 673–678 (2008).

Pagie LWP. *Information Integration in Evolutionary Processes*. Ph.D. thesis, Utrecht University (1999).

Pagie LWP & Hogeweg P. Evolutionary consequences of coevolving targets. *Evol Comput* 5: 401–418 (1997).

Pagie LWP & Hogeweg P. Individual- and population-based diversity in restriction-modification systems. *Bull Math Biol* 62: 759–774 (2000a).

Pagie LWP & Hogeweg P. Information integration and red queen dynamics in coevolutionary optimization. *Proceedings CEC* pp. 797–806 (2000b).

Paladugu SR, Chickarmane V, Deckard A, Frumkin JP, McCormack M & Sauro HM. In silico evolution of functional modules in biochemical networks. *IEE Proc Syst Biol* 153: 223–235 (2006).

Pardue ML & DeBaryshe PG. Drosophila telomeres: two transposable elements with important roles in chromosomes. *Genetica* 107: 189–196 (1999).

Parter M, Kashtan N & Alon U. Facilitated variation: how evolution learns from past environments to generalize to new environments. *PLoS Comput Biol* 4: e1000206 (2008).

Pepper JW. The evolution of evolvability in genetic linkage patterns. *Biosystems* 69: 115–126 (2003).

Pérez-Hormaeche J, Potet F, Beauclair L, Masson I Le, Courtial B, Bouché N & Lucas H. Invasion of the Arabidopsis genome by the tobacco retrotransposon Tnt1 is controlled by reversible transcriptional gene silencing. *Plant Physiol* 147: 1264–1278 (2008).

Pfeiffer T & Bonhoeffer S. Evolution of cross-feeding in microbial populations. *Am Nat* 163: E126–E135 (2004).

Pfeiffer T, Soyer OS & Bonhoeffer S. The evolution of connectivity in metabolic networks. *PLoS Biol* 3: e228 (2005).

Philippe N, Crozat E, Lenski RE & Schneider D. Evolution of global regulatory networks during a long-term experiment with Escherichia coli. *Bioessays* 29: 846–860 (2007).

Pigliucci M. Is evolvability evolvable? *Nat Rev Genet* 9: 75–82 (2008).

Porcher E, Tenaillon O & Godelle B. From metabolism to polymorphism in bacterial populations: a theoretical study. *Evolution* 55: 2181–2193 (2001).

van der Post DJ & Hogeweg P. Resource distributions and diet development by trial-and-error learning. *Behavioral Ecology And Sociobiology* 61: 65–80 (2006).

van der Post DJ & Hogeweg P. Diet traditions and cumulative cultural processes as side-effects of grouping. *Animal Behaviour* 75: 133–144 (2008).

Poyatos JF & Hurst LD. Is optimal gene order impossible? *Trends Genet* 22: 420–423 (2006).

Pritham EJ, Putliwala T & Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390: 3–17 (2007).

Qian W & Zhang J. Evolutionary dynamics of nematode operons: easy come, slow go. *Genome Res* 18: 412–421 (2008).

Quayle AP & Bullock S. Modelling the evolution of genetic regulatory networks. *J Theor Biol* 238: 737–753 (2006).

Quesneville H & Anxolabéhère D. Genetic algorithm-based model of evolutionary dynamics of class II transposable elements. *J Theor Biol* 213: 21–30 (2001).

Raes J & Bork P. Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol* 6: 693–699 (2008).

Rainey PB & Travisano M. Adaptive radiation in a heterogeneous environment. *Nature* 394: 69–72 (1998).

Rauch EM, Sayama H & Bar-Yam Y. Relationship between Measures of Fitness and Time Scale in Evolution. *Phys. Rev. Lett.* 88: 228101 (2002).

Ray TS. An approach to the synthesis of life. In *Artificial Life II*, edited by C G Langton, C Taylor, J D Farmer & S Rasmussen, pp. 371–408. Addison-Wesley (1991).

Rosa A & Everaers R. Structure and dynamics of interphase chromosomes. *PLoS Comput Biol* 4: e1000153 (2008).

Rosenzweig RF, Sharp RR, Treves DS & Adams J. Microbial evolution in a simple unstructured environment: genetic differentiation in Escherichia coli. *Genetics* 137: 903–917 (1994).

Rouzic A Le, Boutin TS & Capy P. Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A* 104: 19375–19380 (2007).

Rouzic A Le & Capy P. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* 169: 1033–1043 (2005).

Rubin GM & Spradling AC. Genetic transformation of Drosophila with transposable element vectors. *Science* 218: 348–353 (1982).

Sabot F & Schulman AH. Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* 97: 381–388 (2006).

Savill NJ & Hogeweg P. Spatially induced speciation prevents extinction: the evolution of dispersal distance in oscillatory predator-prey models. *Proc Biol Sci* 265: 25–32 (1998).

Savill NJ, Rohani P & Hogeweg P. Self-reinforcing spatial patterns enslave evolution in a host-parasitoid system. *J Theor Biol* 188: 11–20 (1997).

Schacherer J, De Montigny J, Welcker A, Souciet J & Potier S. Duplication processes in Saccharomyces cerevisiae haploid strains. *Nucleic Acids Res* 33: 6319–6326 (2005).

Schacherer J, Tourrette Y, Souciet J, Potier S & De Montigny J. Recovery of a function involving gene duplication by retroposition in Saccharomyces cerevisiae. *Genome Res* 14: 1291–7 (2004).

Schuster P, Fontana W, Stadler PF & Hofacker IL. From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci* 255: 279–284 (1994).

Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Huizar L, Davis RW, Scherer S, Tait E, Shaw DJ, Harris D, Murphy L, Oliver K, Taylor K, Rajandream MA, Barrell BG & Wolfe KH. Prevalence of small inversions in yeast gene order evolution. *Proc Natl Acad Sci U S A* 97: 14433–14437 (2000).

Shabalina SA & Koonin EV. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol* 23: 578–587 (2008).

Shen-Orr SS, Milo R, Mangan S & Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* 31: 64–68 (2002).

Sijen T, Fleenor J, Simmer F, Thijssen KL, Parrish S, Timmons L, Plasterk RH & Fire A. On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* 107: 465–476 (2001).

Sijen T & Plasterk RHA. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature* 426: 310–314 (2003).

Singer GAC, Lloyd AT, Huminiecki LB & Wolfe KH. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol Biol Evol* 22: 767–775 (2005).

Siomi MC, Saito K & Siomi H. How selfish retrotransposons are silenced in Drosophila germline and somatic cells. *FEBS Lett* 582: 2473–2478 (2008).

Slotkin RK & Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8: 272–285 (2007).

Smalheiser NR & Torvik VI. Mammalian microRNAs derived from genomic repeats. *Trends Genet* 21: 322–326 (2005).

Solé RV, Alonso D & McKane A. Self-organized instability in complex ecosystems. *Philos Trans R Soc Lond B Biol Sci* 357: 667–671 (2002).

Solé RV & Valverde S. Are network motifs the spandrels of cellular complexity? *Trends Ecol Evol* 21: 419–422 (2006).

Solé RV & Valverde S. Spontaneous emergence of modularity in cellular networks. *J R Soc Interface* 5: 129–133 (2008).

Sorek R, Ast G & Graur D. Alu-containing exons are alternatively spliced. *Genome Res* 12: 1060–1067 (2002).

Soyer OS & Bonhoeffer S. Evolution of complexity in signaling pathways. *Proc Natl Acad Sci U S A* 103: 16337–16342 (2006).

Soyer OS, Pfeiffer T & Bonhoeffer S. Simulating the evolution of signal transduction pathways. *J Theor Biol* 241: 223–232 (2006).

Spencer CC, Bertrand M, Travisano M & Doebeli M. Adaptive diversification in genes that regulate resource use in Escherichia coli. *PLoS Genet* 3: e15 (2007).

Spencer CC, Tyerman J, Bertrand M & Doebeli M. Adaptation increases the likelihood of diversification in an experimental bacterial lineage. *Proc Natl Acad Sci U S A* 105: 1585–1589 (2008).

Srikhanta YN, Maguire TL, Stacey KJ, Grimmond SM & Jennings MP. The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes. *Proc Natl Acad Sci U S A* 102: 5547–5551 (2005).

Strom LS. Microbial ecology of ocean biogeochemistry: a community perspective. *Science* 320: 1043–1045 (2008).

Takeuchi N & Hogeweg P. Evolution of complexity in RNA-like replicator systems. *Biol Direct* 3: 11 (2008).

Teichmann SA & Babu MM. Gene regulatory network growth by duplication. *Nat Genet* 36: 492–496 (2004).

Teichmann SA & Veitia RA. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics* 167: 2121–2125 (2004).

Toffoli T & Margolus N. *Cellular Automata Machines: A New Environment for Modeling*. MIT Press, Cambridge, Mass (1987).

Torsvik V, Øvreås L & Thingstad TF. Prokaryotic diversity–magnitude, dynamics, and controlling factors. *Science* 296: 1064–1066 (2002).

Treves DS, Manning S & Adams J. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. *Mol Biol Evol* 15: 789–797 (1998).

Tyerman J, Havard N, Saxer G, Travisano M & Doebeli M. Unparallel diversification in bacterial microcosms. *Proc Biol Sci* 272: 1393–1398 (2005).

Umezu K, Hiraoka M, Mori M & Maki H. Structural analysis of aberrant chromosomes that occur spontaneously in diploid Saccharomyces cerevisiae: retrotransposon Ty1 plays a crucial role in chromosomal rearrangements. *Genetics* 160: 97–110 (2002).

Urrutia AO, Ocaña L Balladares & Hurst LD. Do Alu repeats drive the evolution of the primate transcriptome? *Genome Biol* 9: R25 (2008).

van der Laan JD & Hogeweg P. Predator-Prey Coevolution: Interactions across Different Timescales. *Proceedings: Biological Sciences* 259: 35–42 (1995).

Wagner A. Robustness, evolvability, and neutrality. *FEBS Lett* 579: 1772–1778 (2005).

Wagner GP & Altenberg L. Complex Adaptations and the Evolution of Evolvability. *Evolution* 50: 967–976 (1996).

Wagner GP, Pavlicev M & Cheverud JM. The road to modularity. *Nat Rev Genet* 8: 921–931 (2007).

Walker JJ & Pace NR. Phylogenetic composition of Rocky Mountain endolithic microbial ecosystems. *Appl Environ Microbiol* 73: 3497–3504 (2007).

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK & Haussler D. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* 104: 18613–18618 (2007).

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, Surani MA, Sakaki Y & Sasaki H. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453: 539–543 (2008).

Werner A, Carlile M & Swan D. What do natural antisense transcripts regulate? *RNA Biol* 6: 43–48 (2009).

Wilke CM & Adams J. Fitness effects of Ty transposition in Saccharomyces cerevisiae. *Genetics* 131: 31–42 (1992).

Williams HT & Lenton TM. Artificial selection of simulated microbial ecosystems. *Proc Natl Acad Sci U S A* 104: 8918–23 (2007).

Williams HT & Lenton TM. Environmental regulation in a network of simulated microbial ecosystems. *Proc Natl Acad Sci U S A* 105: 10432–7 (2008).

Wright S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the VI International Congress of Genetrics* 1: 356–366 (1932).

Wu M, Li L & Sun Z. Transposable element fragments in protein-coding regions and their contributions to human functional proteins. *Gene* 401: 165–171 (2007).

Yang N & Kazazian HH. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol* 13: 763–771 (2006).

List of Publications

Crombach, A & Hogeweg, P. Chromosome Rearrangements and the Evolution of Genome Structuring and Adaptability. *Mol Biol Evol* 24: 1130-1139 (2007).

Crombach, A & Hogeweg, P. Evolution of Evolvability in Gene Regulatory Networks. *PLoS Computational Biology* (2008).

Crombach, A & Hogeweg, P. Evolution of Resource Cycling in Ecosystems and Individuals. (*Submitted to BMC Evol Biol*).

Crombach, A & Hogeweg, P. Modeling RNAi Transposon Control: Interactions between Transcriptional and Post-Transcriptional Gene Silencing. (*In preparation*).

Samenvatting

Dit jaar, 2009, is het 150 jaar geleden dat Charles Darwin zijn boek *The origin of species*, oftewel *De oorsprong der soorten*, publiceerde. Daarin ontvouwde hij de theorie van evolutie, waarbij kleine verschillen in de erfelijke eigenschappen van individuen leiden tot betere en slechtere overlevings- en voortplantingskansen. Het resultaat zien we dagelijks om ons heen: een enorme diversiteit aan planten en dieren. Sindsdien is evolutie opgeklommen tot de belangrijkste theorie in de biologie en worden de principes van evolutie succesvol toegepast in andere wetenschappen: van de ontwikkeling van nieuwe medicijnen tot het ontwerpen van robots.

De afgelopen vier jaar hebben wij onderzoek gedaan naar de invloed van evolutie op de structurering van het genoom, genregulatie en ecosystemen. Hierbij zijn twee onderliggende thema's kenmerkend geweest voor onze studies, namelijk de invloed van de omgeving en de veranderingen daarin op evolutie – en natuurlijk hoe evolutie vervolgens de omgeving kan beïnvloeden – en mobiele genetische elementen. Wat het tweede thema betreft, zulke genetische elementen zijn genen, ookwel transposons genaamd, die zichzelf kunnen kopiëren en daarmee op allerlei plekken in het genoom terecht komen. Dit leidt vaak tot nadelige mutaties, maar af en toe kan het ook positieve effecten teweegbrengen.

In twee studies over evolutie zijn wij uitgegaan van omgevingsveranderingen die niet beïnvloed kunnen worden door de populatie van individuen. We hebben onderzocht hoe de genoomstructuur (**hoofdstuk 2**) en de genregulatie structuur (**hoofdstuk 3**) aangepast worden door het evolutionaire proces. Onder genoomstructuur verstaat men onder andere de groepering op het genoom van genen die bij dezelfde biologische taken betrokken zijn. Zeer illustratief zijn de Hox-genen, die naast elkaar op het genoom liggen en samen zorgdragen voor het bouwplan van het lichaam. Genregulatie staat een niveau hoger in de hiërarchische organisatie van een cel: het is de bedrading die bepaalt hoe genen andere

genen aan- of uitzetten en hoe deze signalen zich verspreiden.

Wij hebben laten zien dat zich herhalende wisselingen van omgeving op de lange termijn inderdaad kunnen leiden tot gegroepeerde genen en dat het netwerk van elkaar beïnvloedende genen zich zo kan structureren dat slechts specifieke mutaties grote, gunstige effecten teweegbrengen. In beide gevallen is het resultaat dat een kleine set van mutaties herhaaldelijk wordt waargenomen: in het eerste geval omdat de juiste mutaties gewoonweg vaker voorkomen en in het tweede omdat er slechts een beperkt aantal mutaties gunstig zijn. Dus evolutie heeft ervoor gezorgd dat evolutionaire aanpassingen aan de omgeving sneller verlopen. Je zou kunnen stellen dat evolutie zichzelf verbeterde. Dit wordt de evolueerbaarheid van evolutie genoemd, een concept wat door ons werk minder controversieel is geworden.

In ons model over genoomorganisatie zouden de genomen zich niet hebben kunnen structureren als ze geen mobiele genetische elementen hadden bevat. Deze genen zorgen niet alleen voor veel nadelige mutaties, maar ook voor zwakke plekken waar de chromosomen eenvoudig breken. Doordat bij de reparatie van chromosomen soms verkeerde stukken aan elkaar gelast worden, kunnen hiermee tijdens de celdeling groepen genen makkelijk gedupliceerd of weggehaald worden. Op deze manier is de kans groot dat individuen, bij wie de genen die nodig zijn voor dezelfde functie bij elkaar in de buurt liggen op het genoom, aangepaste nakomelingen produceren. Een aanpassing aan een nieuwe omgeving kan dus veel sneller plaatsvinden. Lange tijd heeft men transposons beschouwd als "junk DNA" en parasieten. Tegenwoordig echter worden er steeds meer positieve gevolgen toegekend aan deze elementen, waar wij hier een nieuwe hypothese aan toevoegen: transposons als mechanisme voor evolueerbaarheid.

De structurering van het genregulatie netwerk vond op een andere wijze plaats. Als genen elkaar aansturen en deze besturing kan evolueren, kan er een hiërarchische structuur ontstaan waarbij één gen vele andere aan of uitzet. Dit is precies wat wij ontdekten in ons model van genregulatie. Na een lange tijd van evolutie in een veranderende omgeving, bestond de populatie uit individuen die zeer snel wisselden van genexpressie patroon als de omgeving wijzigde. De individuen waren gevoelig geworden voor mutaties die één specifiek gen kopieerden of verwijderden. Zo'n mutatie versterkte of verzwakte het aan/uit signaal naar vele andere genen waardoor de juiste genen een ander expressieniveau kregen en de individu aangepast was aan de nieuwe omgeving. Dit gekopieerde of verwijderde gen leek te functioneren als een sensor voor een veranderde omgeving, maar deed dit via mutaties. Daarom hebben we het een "evolutionaire sensor" genoemd. Opnieuw is de populatie in staat zich veel sneller aan te passen aan een veranderende omgeving: evolutie van evolueerbaarheid in genregulatie netwerken.

Tot nu toe zijn we uitgegaan van een omgeving die verandert door invloeden van buitenaf. Wat gebeurt er als individuen zelf hun omgeving bepalen? In een model waar digitale organismen voedsel uit de omgeving halen en hun afval er in terugplaatsen (**hoofdstuk 4**), hebben we de koppeling van de korte tijdschaal

van metabolisme en omgevingsveranderingen met evolutie bekeken. De evolutie van dit ecosysteem gaf grofweg twee uitkomsten. Afhankelijk van de spatiale structuur waarin evolutie plaatsvindt – of individuen en hun nakomelingen in de buurt van elkaar blijven, of dat individuen vaak gemengd worden – ontwikkelen 'slimme' individuen of samenwerkende soorten. De evolutionaire strategie van slimme individuen is vooral waarschijnlijk als gerelateerde individuen in elkaars buurt zitten. In dit geval is het van belang je eigen voedingsstoffen te hergebruiken, met als resultaat onafhankelijke, 'slimme' individuen. Als we de populatie constant mengen, zal deze zich makkelijk opsplitsen in twee of drie aparte takken die afzonderlijk niet goed in staat zijn de voedingsstoffen te hergebruiken, maar samen wel kunnen zorgdragen voor deze recycling. De invloed van organismen op hun eigen omgeving kan dus zowel individuen 'slimmer' maken als bijdragen aan de diversiteit die we zien in ecosystemen.

In de laatste studie hebben we in detail bestudeerd hoe een individu zich kan beschermen tegen de activiteit van mobiele genetische elementen (**hoofdstuk 5**). In de afgelopen tien jaar is het duidelijk geworden dat transposons en hun zelfkopiërende activiteiten strikt gereguleerd worden door de cel. Zowel in de kern als in het cytoplasma maakt de cel gebruik van "RNA interference". In het Nederlands zou men dit proces RNA onderbreking of RNA bemoeienis noemen, hier zullen we het aanduiden met de afkorting RNAi, wat ook de gebruikelijke term in de literatuur is. Wij hebben met een wiskundig model onderzocht hoe het kan dat essentiële enzymen lijken te ontbreken in de RNAi machinerie van bijvoorbeeld de mens of fruitvlieg, terwijl beide goed in staat zijn transposons te controleren. Wij vonden dat er verschillende strategiëen mogelijk zijn voor transposoncontrole, en dat elke strategie andere eisen stelt aan welke enzymen of cellulaire processen aanwezig moeten zijn. Waar bij de ene de nadruk ligt op controle in de celkern, zal een andere strategie vooral in het cytoplasma actief zijn. Zo zijn sommige enzymen wellicht toch niet zo essentieel als tot nu toe is aangenomen.

Samenvattend, we hebben met modellen de implicaties van natuurlijke selectie en variatie door mutatie onderzocht. Veranderingen op verschillende tijdschalen, zoals herhaalde omgevingswisselingen en de interacties met de lange termijn van evolutie kunnen leiden tot versnelde evolutionaire aanpassingen aan de omgeving. Dit kan zowel op het genoomniveau, als op het niveau van het regulatoire netwerk. Als de omgeving wordt bepaald door de organismen zelf en we kijken naar het ecosysteem, dan zien we twee strategieën. Namelijk dat onafhankelijke, zichzelf voorzienende, individuen ontstaan, maar ook dat samenwerkende soorten zich kunnen ontwikkelen. Zeer belangrijk voor de evolueerbaarheid van het genoom zijn mobiele genetische elementen. In een aparte studie laten we zien dat controle van deze zelfkopiërende genen op verschillende manieren mogelijk is, waarmee we een eerste verklaring bieden voor de verscheidenheid aan enzymen die worden aangetroffen, of juist ontbreken, bij vele soorten. Dit fundamentele onderzoek heeft bijgedragen aan een dieper inzicht in het proces van evolutie en de belangrijke rol die daarin vervuld wordt door de interactie met een dynamische omgeving en mobiele genetische elementen.

Curriculum Vitae

The author was born in Veldhoven, the Netherlands, on June 18, 1980. From 1992 to 1998 he attended van Maerlantlyceum in Eindhoven, where he obtained his Gymnasium diploma.

In September 1998 he started his studies Computer Science at Eindhoven University of Technology (TU/e) and a year later, in August 1999, he gained his propadeutic diploma. After three years, he chose to specialise on the intersection of computer science and biology. In 2001 he followed the majority of his courses at the Theoretical Biology and Bioinformatics Group of Utrecht University, followed by a traineeship on diploid genetic algorithms at the Biomodelling and Informatics Group at TU/e. In 2002 – 2003 he performed his Master's thesis project on metabolomic fingerprinting at Exeter University. In September 2003 he graduated from Computer Science, receiving his Master's degree from prof. P. Hilbers.

After a half year travel through Australia, he started his postgraduate research under supervision of prof. P. Hogeweg in July 2004, finishing in Januari 2009. The results of this research are described in this thesis.

Dankwoord

Hier wil ik graag enkele mensen bedanken voor vier – bijna vijf – mooie jaren die hebben geleid tot dit proefschrift. In de eerste plaats mijn promotor, Paulien Hogeweg. Je hebt me weten om te vormen van een informaticastudent met een voorliefde voor biologie tot een "computational biologist". Je kennis, ideeën, inspiratie en enthousiasme hebben me door soms toch moeilijke tijden geloodst.

I would like to thank Günter Wagner, Laurence Hurst, Kristian Lindgren and Ricard Solé for being members of my reading committee and finding the time and interest to read this thesis.

De gehele vakgroep wil ik bedanken voor alle plezier, inspiratie, discussies en steun. Mijn kamergenoten, Otto, Nobuto, Boris en Ramiro, wil ik met name noemen. Voor de afleiding in de pauzes en buiten werk dank ik ook Marian, Daniël, Milan, Christian, Laura en Bas. Jan Kees, het is wonderbaarlijk dat je die enorme hoeveelheid computers op de vakgroep zo goed draaiende houdt.

Buiten de vakgroep zijn er ook velen die mijn vier jaar in Utrecht heel aangenaam hebben gemaakt. Natuurlijk wil ik mijn ouders, Anneke en Titus, en zus en broer, Marica en Pieter, bedanken voor alle plezier en steun die ik heb ontvangen. Van vakanties in Milaan tot 'eerste hulp' na fietstochten van Utrecht naar Veldhoven. Jorrit, per toeval hebben we elkaar ontmoet in Australië, en dat heeft geleid tot avonden plaatjes draaien, vele feesten en steun wanneer dat nodig was. Dank je wel dat je mijn paranimf wilde zijn. Sascha, mijn tweede paranimf; vakantie, uiteten, uitgaan en dansen, maar ook een luisterend oor voor de moeilijke momenten, dat is echte vriendschap. Dorus, Jessica, Linda bij deze dank jullie wel voor alle plezier, interessante gesprekken, leuke feestjes en mooie sportmomenten. Christina, we met by accident on an island in Brasil, and even though we do not see each other that often, our friendship means a lot to me. Thank you for your support, entertainment and let us not lose sight of each other.

Club 14 (jullie weten wie ik bedoel), jullie wil ik natuurlijk bedanken voor kerstdiners en de mooie geschiedenis die we samen hebben! Claartje, Tanca, René, Tjalco, Marion en Clement; jullie zijn de tofste huisgenoten die ik me heb kunnen wensen. Maite, dank je voor het lekkere eten, de verfrissende gesprekken en tuinavonturen. Claudia, Matt, Marta and Rocio, what a great thing we still keep in touch.

Kirsten, dank je voor een mooi laatste (onderzoeks)jaar en ik heb enorme zin om met je op vakantie te gaan.