Modelling evolution of genome size in prokaryotes in response to changes in their abiotic environment

A thesis submitted to the School of Environmental Sciences of the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy

By Piotr Bentkowski

January 2014

Abstract

The size of the genomes of known free-living prokaryotes varies from ~ 1.3 Mbp to ~ 13 Mbp. This thesis proposes a possible explanation of this variation due to variability of the physical conditions of the environment. In a stable environment, competition for the resource becomes the main force of selection and smaller (thus cheaper) genomes are favoured. In more variable conditions larger genomes will be preferred, as they have a wider range of response to a less predictable environment. An agent-based model (ABM) of genome evolution in an free-living prokaryotic population has been proposed. Using the classic Hutchinson niche space model, a gene was defined as a Gaussian function over a corresponding niche dimension. The cell can have more than one gene along a given dimension, and the envelope of all the corresponding responses is considered a full description of a cell's phenotype over that dimension. Gene deletion, gene duplication, and modifying mutations are permitted during reproduction, so the number of genes and their phenotypic effect (height and position of the Gaussian envelope) are free to evolve. The surface under the curve is fixed to prevent 'supergenes' from occurring. Change of the environmental conditions is simulated as a bounded random walk with a varying length of the step (a parameter representing variability of the environment). Using this approach, the model is able to reproduce the phenomenon of genome streamlining in more stable environments (analogical to e.g. oligotrophic gyre regions of the ocean) and genome complexification in variable environments. Horizontal gene transfer (HGT) was also introduced, but was found to act in a similar manner as gene duplication and shown no important contribution to the speed of evolution and the adaptive potential of the population.

Acknowledgements

First of all, I wish to express my gratitude to all the members of my advisory committee, i.e. Tim Lenton, Thomas Mock, Hywel Williams and Cock Van Oosterhout, for their incredible patience toward my working style through all the years I have spent at University of East Anglia. This interdisciplinary project would not have been possible without their help and competence in scientific disciplines spanning from mathematics, through molecular biology and ecology, up to Earth system science. But most importantly, I would like to thank them for their enthusiasm in supporting me in my studies. Also, I am very thankful to my colleagues from the Earth System Modelling Group for hours of discussions which often opened my eyes to problems I had previously been fully unaware of. I would also not have been able to reach this point of my scientific career without Joanna Pijanowska and Andrzej Mikulski, with whom I have made my first steps as a researcher at the University of Warsaw.

My life in Norwich would have beeen much less interesting without my friends from grad school. I am grateful to my house mates, Jona Barichivich and Jeppe Graugaard, for their friendly company and for putting up with my ridiculous circadian rhythm. I would also like to thank all my office colleagues, with special thanks to Jan Strauss, Ben Mills, Phil Underwood, Justin Krijnen and Michał Bochenek, for gallons of coffee and beer drank together and also for hours of joint procrastination.

Special thanks are go to my long lasting friends and compatriots from Warsaw, Magda Mantorska and Zbyszek Pietras, with whom I shared the fate of a Slav in East Anglia. Life here would have been harsh had I not had anyone to sympathise with my complains about English weather, local bread and left-hand traffic. It was good to have you around!

But most of all, I am grateful to my parents for years of support and unconditional love.

Symbols and acronyms

Mathematical symbols:

$[a, b]$ – an interval where $x \in \mathbb{R} \mid a \leqslant x \leqslant b$

$]a, b[$ – an interval where $x \in \mathbb{R} \mid a < x < b$

$\langle x \rangle$ – the expected value of variable x

$\in$ – belongs to: e.g. $x \in X$: i.e. element x belongs to set X

$RNG[a, b]$ – a mathematical function which randomly selected one real number from a continuous linear distribution over the range of $[a, b]$

$\mathcal{N}(\mu, \sigma^2)$ – normal distribution function with mean at μ and variation σ^2

Acronyms:

ABM – agent-based model

bp – base pairs, DNA chain length unit

CPU – central processing unit

DNA – deoxyribonucleic acid

fW – femtowatt, 10^{-15} Watt

GTA – gene transfer agent

HGT – horizontal gene transfer

HPC – high performance computing

kbp – kilo base pairs, DNA chain length unit, 10^3 basepairs

Mbp – mega base pairs, DNA chain length unit, 10^6 basepairs

mRNA – messenger RNA, ribonucleic acid molecule encoding a protein sequence

N – nitrogen, an element

NCBI – National Center for Biotechnology Information

NTME – neutral theory of molecular evolution, a theory formulated by M. Kimura

P – phosphorus, an element

Pg – petagram, mass unit, 10^{12} kg

PolI – DNA polymerase, type of enzyme I

PolIII – DNA polymerase, type of enzyme III

RM system – restriction-modification system

RNA – ribonucleic acid

rRNA – ribosomal ribonucleic acid

SD – standard deviation, a statistical measure

TA – toxin/antitoxin complex, a type of genetic regulatory system

TF – transcription factors, type of regulatory proteins

tRNA – transfer ribonucleic acid

W – Watt [$kg \cdot m^2 \cdot s^{-3}$]

Symbols used in the models:

α – surface under gene $u(x, c, \sigma, A)$, constant for all the genes

γ – metabolic cost of one gene, constant

δ – random death factor, constant

$\eta_{0,max}$ – maximum number of genes in a genome at initialisation of the model run

$\eta_{0,min}$ – minimum number of genes in a genome at initialisation of the model run

κ – metabolic cost of living, constant

μ_{del} – mutation, probability of deleting a gene

μ_{dupl} – mutation, probability of duplicating a gene

μ_{mod} – mutation, probability of modification of a single gene

σ – width of gene $u(x, c, \sigma, A)$

τ – maximum amount of resource a cell can gain at one time step, constant

A – height of gene $u(x, c, \sigma, A)$, maximally 1

c – location of the maximum of gene $u(x, c, \sigma, A)$ in x space

$\mathbf{G}_i$ – array containing all the genes of the ith cell

H – Shannon index

h_c – probability of horizontal gene transfer, at cell level

h_g – probability of horizontal gene transfer, at gene level

$K_{i,t}$ – total resource expenses for living of the ith cell at time t

N_i – number of all genes the ith cell has

n_i – number of metabolic genes the ith cell has, $N_i = n_i(1 + n_i)$

$Q_{i,t}$ – amount of resource the ith cell gains at time step t

R – total amount of resource in the model, $R = R_{env} + \sum_i r_{cell,i}$

R_{env} – free resource allocated in the environment

$r_{cell,i}$ – resource allocated in the ith cell

r_{min} – minimum level of resource for a cell to survive, if $r_{cell,i} < r_{min}$ than the cell dies

r_{rep} – level of resource for a cell to reproduce, if $r_{cell,i} \geq r_{min}$ than the cell divides

T – turbulence level

t_{max} – total number of time steps of the model run

t – time, measured in intervals

$U_i(x)$ – genotype of the ith cell, envelope of all the cell's genes

$u(x, c, \sigma, A)$ – the gene, Gaussian function defining resource uptake efficiency

x – environmental conditions value

Contents

List of Tables

List of Figures

Chapter 1

Introduction: how to model genes in their environment?

1.1 Aim of this work

In recent years, with the progress of genomics and metagenomics, we became aware of new facts concerning the molecular foundations of biological evolution. One of them is the huge diversity of genome sizes and spatial arrangements of information stored in DNA among seemingly similar organism. The most prominent question which arises from this new knowledge is that on the reason for such diversity in genome sizes. If one focuses, for the sake of simplicity, only on organisms with the simplest chromosome arrangement and lacking a membrane bound nucleus, namely bacteria and archaea (referred jointly in this study as prokaryotes), the majority of which have a single circular DNA molecule, one can clearly see that there exist both upper and lower limits for their genome size (Koonin and Wolf, 2008). As it has been suggested in previous publications (Koonin and Wolf, 2008; Lynch, 2006a), there are two kinds of forces which affect genomes. One of those causes their expansion, the other – their reduction, with both leading to the emergence of an optimum genome size for a given species.

The size of the species' genome has to have an impact on its population's ability to survive and on the way it evolves. This work aims at investigating the link between how the rate at which the environment changes alters the size of the genome and how the genome size could influence the adaptive potential of a species.

1.2 Environmental variability as a driver in biological evolution

Questions regarding linkages between variability of the environment and biological evolution have been one of the profound problems in the development of theory of evolution and ecology. Charles Darwin formulated his theory of the origin of species by natural selection stating that for further reproduction only the fittest from a large cohort of individuals with varying properties will be selected. Darwin's theory naturally raises the question on how species change if external conditions, which constitute the selection pressure factor, vary in a fairly short time frame?

1.2.1 Niche overlap in variable environments

One the most profound examples of studies of environmental variability and its impact on adaptation is the so called 'paradox of the plankton'. The competitive exclusion principle, known also as Gause's law, named after ecologist Georgii Gause, states that if two species similar in their environmental requirements compete over the same resource one will eventually outnumber the other one and, in the long run, will lead to the extinction of the weaker competitor (Krebs, 1994). In other words, one niche can only have one occupant. Gause experimented with two species of closely related protozoans (a group of heterotrophic unicellular eukaryotes, mostly aquatic) *Paramecium aurelia* and *P. caudatum*, keeping them together in a shared culture. When environmental conditions were kept constant, growth of population of one of the species led the other to extinction via exploitative resource competition (Gause, 1932). Gause's results have been confirmed and widely acknowledged by other ecologists. By extrapolation, we could expect that in the pelagic zones of lakes, which have little spacial variation and are nearly homogeneous on a large surface, only few species of photosynthetic unicellular algae can persist. Meanwhile, there are many species of such algae and many of them have similar ecological requirements (Hutchinson, 1961). The explanation of this phenomenon is the temporal variation of the environment. Before any of the members of the community will manage to outcompete others, the environment changes, suddenly favouring a different species, thus maintaining diversity of the community in the long run (Hutchinson, 1961; Descamps-Julien and Gonzalez, 2005). On the other hand, this line of argument is difficult to accept when attempting to explain the diversity of species of trees in tropical forests (Hubbell, 2001), because the life spans of trees are long and comparable even with the time scale of climatic fluctuations. Some studies have suggested that spatial heterogeneity can sustain species diversity if it is considered on a large enough scale Davies *et al.* (2005), which is probably the case in a large terrestrial biocenosis such as the tropical forest.

Observations such as the 'paradox of the plankton' lead to two questions: 1) how many similar species can coexist in the environment under a certain amount of temporal variability?; and 2) how similar these species be in their ecological requirements for the community to be sustainable? In his classic paper, Hutchinson (1957) proposed to considerer the environment as multidimensional space of all the factors which are important for the biology of a species. All these factors (e.g. temperature, food size, pH, light intensity, etc.) can be represented as orthogonal dimensions and the volume in this multidimensional space, which has suitable factors' values for species to survive, grow and reproduce is the species' fundamental niche. This concept, though difficult to apply to natural species and mathematically challenging, can become, after some simplification, very fruitful and inspiring, especially in numerous modelling works. For the sake of simplicity many modellers take into account only one dimension of the niche and assume Gaussian shape of the species response in terms of biomass growth, fertility, number of offspring or other measurable parameter relating to species' evolutionary success to the whole span of values of the considered environmental factor. The canonical example are the differences of beak shapes of seed-eating birds. Seeds available in the environment as food come in different sizes and a different shape of beak is optimum for a given size of seeds. A beak appropriate for cracking nuts will serve poorly for handling grass seeds and vice versa. The efficiency of a bird species with a particular beak shape to live on a particular seed size can be represented as a 'utilisation function' (Figure 1.1) which is a cumulative measure of how good in terms of biomass growth and/or offspring production a species is with a given value of environmental conditions. And the questions are: 1) how much overlap is possible between utilisation functions of different species?; and 2) how much this overlap is dependent on the variability of the environment? In their work on modelling environmental variability and niche overlap, which was partly inspired by the 'beak size problem' and the Lotka-Volterra competition model (Foryś, 2005), May

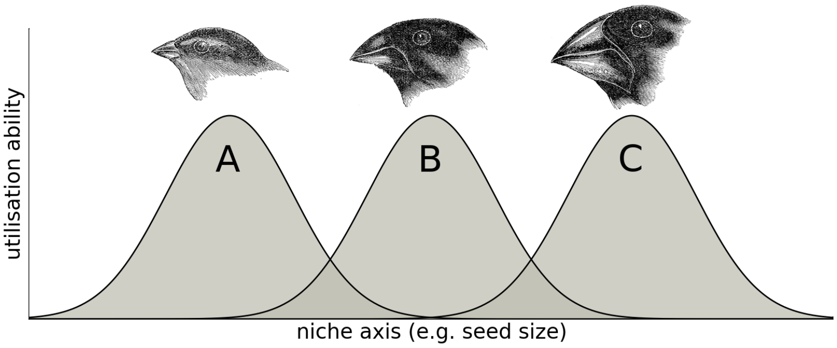

Figure 1.1: Example of niche distribution and overlap along one niche dimension. Different shapes of beaks of species *A*, *B* and *C* are efficient tools for gathering only seeds of a particular size. Or putting it differently, the 'utilisation ability' (also called 'utilisation function') of each species' niche spans and peaks for different seed sizes. The questions is: how tightly can the utilisation functions of different species using the same resource be packed? (figure modified from Nee and Colegrave, 2006).

and Mac Arthur (1972) showed that in a stable unvarying conditions there are no limits to the degree of overlap. At the same time introduction of a stochastically fluctuating variability leads to a difference in the average food size for neighbouring species on the resource spectrum. This difference must be approximately equal to the standard deviation of food size taken by either individual species. This result is robust for environmental fluctuations whose variances relative to their mean span from 0.01% to 30%. This means that there is a limit to the niche overlap and that the level of overlap is only weakly dependent on the environmental variability. a number of field studies, mostly conducted on birds, seem to agree with this model (Storer, 1966; Hespenheide, 1971; Diamond, 1972) showing niche separation of size around one standard deviation of the utilisation function also with regard to environmental factors other than the food size.

A commonly accepted explanation why species have their niches separated is the need to avoid competition. If species are packed uniformly on the niche axis then they suffer from competition 'on both sides' of their utilisation functions and are forced to compromise between selection pressure pushing in different directions. In fact it might happen that species will clump together, as competing with only few but very similar species will give them a greater chance of survival than competing at the same time with quite different species coming in a greater number. Optimising the utilisation function in just one area of the niche space is easier than in two areas. It has been shown in theoretical simulations that when it is allowed for a number species to evolve theirs place along the niche axis they either look a like or are very different. The niche axis is covered by humps of species with similar utilisation functions separated by unoccupied space (Scheffer and van Nes, 2006). This ‘paradox of the clumps’ (Nee and Colegrave, 2006) might be caused by slowness of competitive exclusion as a selection factor. In the model by Scheffer and van Nes (2006) eventually one species in the hump survived, leading to a pattern of equally spaced single species, but this took a long time. In the follow-up study authors have mathematically explained these observations, showing its sole dependence on the eigenvalues and eigenvectors of the community matrix. They also demonstrated the existence of a critical value for the width of the species distribution and for the number of species below which the clustering disappears (Fort *et al.*, 2009). Nonetheless, this long-lasting disequilibriums might be ecologically more significant than the equilibrium states, as it has been argued that transients dynamics play an important and mostly unappreciated role in the ecological processes (Hastings and Higgins, 1994; Hastings, 2004; Olszewski, 2011) and it has been also applied to solving the plankton paradox (Huisman and Weissing, 1999).

1.2.2 Variability of the environment and the genetic evolution

Variability of the environment has been recognised to have an impact not only on the ecological context of species characteristics, but also on low-level molecular organisation of physiology, metabolism and architecture of genomes. One of the elements of the genome architecture is the way how genes interact with each other forming networks. These interactions can have many different characters: genes may depend on common regulatory factors, or regulate each other expression, or they can code different enzymes of the same metabolic pathway. The shape of this networks is a subject of numerous experimental and theoretical research whose overview can be found e.g. in *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Alon, 2007). One of the metrics used in this kind of research is the network modularity defined as the fraction of edges lying within modules rather than between modules relative to a respective value in a network with the same assignment of nodes into modules but with random connections between them (Newman, 2006). It can be further normalised with respect to a randomised network defining a normalised modular network Q_m (Kashtan and Alon, 2005). Using this measure, Parter *et al.* (2007) have taken under consideration 117 species of bacteria which had a good quality of genome annotations and divided them into six environmental groups following the NCBI classification of bacterial lifestyles: obligate symbionts (intra- or extracellular); specialised species (living in specific environments e.g. thermal vents), aquatic species (only fresh water); facultative species (generally free-living but often associated with a host); multiple species (living in a wide range of environments e.g. having different hosts) and terrestrial species (soil bacteria). Each of these types of bacteria live in an environment with a different degree of variability. Authors have shown that the more variable the environment, the more modular is the metabolic network (Figure 1.2, panel C). Also mean fraction of transcription factors (TFs – proteins that bind to specific DNA sequences and by that mean regulate DNA-to-mRNA transcription) among all genes in genotype and mean total number of nodes in the metabolic network are larger for environments with higher variability than the ones with lower (see Figure 1.2, panels A and B, respectively). Authors also found that the fraction of TFs in the total number of genes in a genome is the best predictor of modularity. Genome size is less powerful for that, but better than phylogenetic distance (Parter *et al.*, 2007, supplementary information).

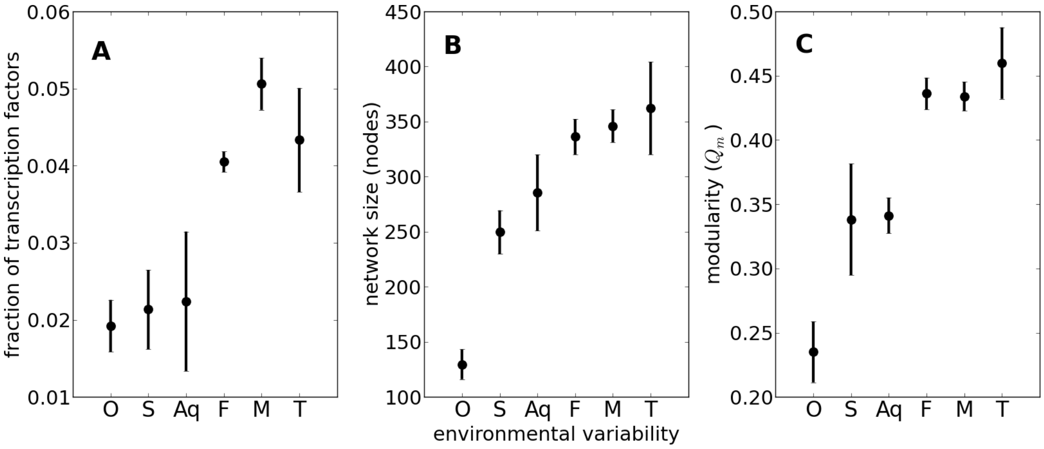

Figure 1.2: Relation between environmental variability and genome architecture. Panel A is the mean fractional number of transcription factors out of the total number of genes in the genome. Panel B is the mean metabolic network size (giant component). Panel C shows normalised modularity measure Q_m (explained in the text) of bacterial metabolic networks versus the environmental class of the organism. Error bars represent standard errors. O – obligate, S – specialised, Aq – aquatic, F – facultative, M – multiple, T – terrestrial (see text for details). Groups are ordered according to their predicted level of variability (figure modified from Parter *et al.*, 2007).

Authors explain these phenomena as an adaptive property of bacteria in variable environments. In a follow-up work they claim that this mechanism allows cells to have the potential to tackle changing condition of the environment and also to evolve faster, as they shown in theoretical study of evolving Boolean logic circuits that forced extinctions in a heterogeneous environment trigger increased modularity of circuits but do not force modular architecture in the case where environment is homogeneous (Kashtan *et al.*, 2009). These discoveries are very tightly linked with problems of evolvability, i.e. the system's capacity to undergo adaptive changes. In evolutionary biology this means a way of organisation of genome which will facilitate genetic mutations beneficial for the population in new environmental conditions, in other words under new directions of selective pressure (Wagner and Altenberg, 1996; Kirschner and Gerhart, 1998; Earl and Deem, 2004; Wagner, 2008). A number of theoretical simulations showed that when certain amount of variation is introduced to the environment, artificial gene regulatory networks will evolve to an architecture which promotes their potential to develop new successful phenotypes after the next change of the environment occurs (Kashtan and Alon, 2005; Crombach and Hogeweg, 2008).

Linked to evolvability is the problem of 'robustness', i.e. the ability of a system to withstand the change (external or internal) without altering its structure. In evolutionary biology this usually means organisation of the genotype in such a way that when a mutation is introduced it will not affect the phenotype. Robustness and evolvability are linked to each other, with robustness pre-conditioning evolvability as it is necessary for a system to be stable against change and not to drift to unfavourable form. But the system has to be robust only to some degree, as a system too robust will not be able to develop new phenotypes (Ciliberti *et al.*, 2007; Pigliucci, 2010). Simulations based on artificial metabolic network selected for biomass production have proven to produce systems robust to gene deletion if evolved under fluctuating environments, e.g. changing in terms of available metabolites (Soyer and Pfeiffer, 2010).

However, results shown by Parter *et al.* (2007) can be explained also by a non-adaptive scenario. There is a number of forces in the evolutionary process, such as genetic drift and genetic draft (discussed further in section 1.5.2), which shape the architecture of prokaryotic genomes in a non-selective manner (see *The Origins of Genome Architecture* by Lynch (2007) and *The Logic of Chance: The Nature and Origin of Biological Evolution* by Koonin (2011) for reviews). Prokaryotes are under constant pressure to perfect their metabolic efficiency and the ultimate reason for higher modularity of metabolic networks in variable environments is that the direction of selective pressure is not focussed on a single goal but on many goals preventing species from achieving maximum efficiency of a cell's metabolism. This view is strengthened by significant evidence that horizontal gene transfer (HGT, discussed in more detail in Chapter 4) plays an important role in the history of prokaryotic evolution (Koonin *et al.*, 2001; Koonin and Wolf, 2008; Koonin, 2009; Isambert and Stein, 2009; Koonin, 2011; Syvanen, 2012) as genes transferred from different species initially form a 'foreign' module in the recipient's genome before evolution will fully integrate them with the recipient's genetic network and regulation systems. Emphasis on non-adaptive forces of evolution may seem more of a change of narrative rather than a change of conclusions, but it is important for gaining the right perspective on the biological evolution as a whole.

1.3 Structure of genomes

Building a successful model of evolving population of bacterial cells needs a good understanding of known data, underlying mechanisms explaining the data and notion of the current theories in the field of evolutionary biology of prokaryotes. Let us move to a brief overview of the up-to-date results in the genomics of prokaryotes.

1.3.1 Prokaryotic genome structure and replication process

To be able to understand the processes behind genome size evolution, we have to have the knowledge about the structure, which has been studied almost since the discovery of the molecular structure of DNA in 1953, and the replication of prokaryotic genomes.

A prokaryotic chromosome is a single circular double strained DNA molecule. There are only a few known bacterial species with linear chromosomes, e.g. the Lyme disease pathogen *Borrelia burgdorferi* (Fraser *et al.*, 1997), plant pathogen *Agrobacterium tumefaciens* (Goodner *et al.*, 2001; Wood *et al.*, 2001) and a soil bacterium *Streptomyces coelicolor*, known for producing natural antibiotics (Bentley *et al.*, 2002). The bacterial chromosome codes for nearly all the proteins produced by a given species. There is also a number of sequences used for regulatory purposes, like the binding sites for e.g. transcription factors (TFs). Prokaryotic genomes are rich in protein-coding information as, with only few exceptions, more than 85% of viral and prokaryotic DNA are coding sequences (Lynch, 2006a).

Most of bacterial and archaean genes are clustered in so-called operons, which are formed by genes located physically close to each other on the DNA strand and are served by one promoter (a DNA sequence recognised by an RNA polymerase and an associated sigma factor) or a single regulatory signal controlling its transcription. All the genes from one operon are transcribed together into a single mRNA strand. Later they can be either translated together, or undergo trans-splicing to form monocistronic mRNA strands that are translated separately into single peptides. An extra layer of regulation mechanisms is provided by enhancers and silencers, i.e. DNA sequences capable of binding the transcription factor proteins (TFs) causing enhancement or suppression of the transcription levels of genes.

The prokaryotic chromosome is replicated in a single continuous process, which is initialised at an origin of replication site (called simply the origin), a DNA sequence recognised by a number of proteins involved in replication initialisation, e.g. DnaA, which takes part in splitting the double helix. Circular bacterial genomes have only one origin site. At the origin, in a multi-stage process, the double strain of DNA is decoupled and two kinds of DNA polymerases get attached, forming, with a number of other proteins, a replication fork. One polymerase is on the 3' → 5' strand (leading strand) and it polymerises the new DNA continuously, while the other one is on the 5' → 3' strand (lagging strand) and it elongates DNA in the opposite direction to the movement of the replication fork (also in the 3' → 5' direction of the template strand). Synthesis of new DNA on the lagging strand is done in short intervals of DNA (1000-2000 nucleotides in bacteria) called the Okazaki fragments. New DNA on the lagging strand is discontinuous and has to be stitched together later on by DNA ligase. The replication fork spreads both ways, as the right-hand side leading strand is at the same time the left-hand side lagging strand and vice versa. After the replication is finished, there are two double helixes, each one having one 'old' strand and one freshly synthesised one. Preservation of half of the template in a newly synthesised DNA means that the replication is a semiconservative process.

The fidelity and integrity of the genome is maintained by three groups of mechanisms. The first and also the least sophisticated one is proofreading by DNA polymerase when elongating the complementary strand. DNA Polymerase III (PolIII), the most important replicating enzyme, introduces an error by mis-pairing the bases with an average frequency of 10^{-6} per base. Those errors are revealed by the failure of the newly incorporated nucleotide to form a hydrogen bound with its complement. The mismatch can be detected and repaired exactly thanks to PolIII, and also to DNA Polymerase I (PolI), which manifest exonuclease activity (it can break the phosphodiester bond within a polynucleotide chain), which allows it to perform post-replication editing. The net rate of spontaneous error occurrence is 10^{-9} per base per replication (Joset *et al.*, 1993).

The second group of mechanisms is formed by a variety of systems preventing foreign DNA from invading the cell, whose expression could bias the existing metabolic trails. Briefly speaking, this restriction-modification system (RM system) marks the cell's DNA with a characteristic nucleotide methylation pattern of methylases ('modification') which is later recognised by endonucleases that cut alien DNA, preventing its expression ('restriction').

The third group are the DNA repair mechanisms. As an effect of interaction with different physical and chemical agents (UV and other high energy electromagnetic radiation, free radicals, toxins and toxic substances, etc.) DNA strands can experience damages to its structure. Most common types of errors include: missing bases, incorrect bases, modified bases, single-strand brakes, double-strand brakes and interstrand cross-links. The catalogue of repair systems is longer, as there is more than one mechanism to repair one type of damage, depending on its scale and chemical arrangement. A broader description of this problem can be found in e.g. *Prokaryotic Genetics: Genome Organization, Transfer and Plasticity. Chapter 8: DNA Repair* (Joset *et al.*, 1993).

The cost of DNA replication is relatively low, as it accounts for just about 2% of the energy budget of microbial cells during growth. For comparison, protein synthesis takes up approximately 75% of a cell's total energy budget (Harold, 1986). An organism can have a different number of chromosomes having identical structure – this feature is called ploidy. Prokaryotes have a single chromosome and they are haploid, but many eukaryotes have diploid genomes, which means having two chromosomes of identical structure, often differing in single alleles (slightly different variants of the same gene). An average size bacterium having a haploid genome of a size of 6 megabase (Mb) and about $5 \cdot 10^3$ genes has approximately $0.49 \cdot 10^{-12}$ W of power per cell. With ploidy, this gives on average $0.12 \cdot 10^{-15}$ W per genome, when an average eukaryote has $1143 \cdot 10^{-15}$ W available per genome (Lane and Martin, 2010). It has been suggested that these low energetic levels, as compared with eukaryotes, might be the reason for a lower genome complexity among prokaryotes (Lane and Martin, 2010).

Data regarding the speed of DNA replication among a wider range of prokaryotes are not available at the moment, however there is detailed information on the speed and efficiency of bacterial DNA replication for specific enzymatic complexes and a wide range of research on the replication of the chromosome of the model bacterium *Escherichia coli*. An interesting observation is that when replication in *E. coli* takes place under optimum conditions, it lasts approximately 40 minutes (not including other phases of the cell). Meanwhile, the generation time can also be shorter than 30 minutes (Cooper and Helmstetter, 1968; Bipatnath *et al.*, 1998). This is possible because *E. coli* is able to run multiple replications at one time by initialisation of replication on a DNA strand which has not been fully completed in the previously started process. It is possible to pass to a daughter cell a full chromosome which is already undergoing next replication, so the daughter cell picks up the process which is under way. In this way, DNA copying can be initialised even two generations ahead (Nordstrom and Dasgupta, 2006). Nevertheless, in *in vivo* conditions replication is rarely a limiting factor for reproduction. Other factors, of more environmental nature, like low nutrient availability, suboptimum temperature, phage activity, etc. are more prone to impact the growth of prokaryotic cells.

As genomics is a well-established discipline, a reliable and detailed overview of the structure and replication of genomes is provided in a number of academic handbooks (e.g. Joset *et al.*, 1993; Brown, 2006). The picture of the prokaryotic DNA replication system which arises from the modern knowledge shows a complicated multilevel structure. Not all of its features can be included in models of evolution of the prokaryotic genome, especially when this model was supposed to form part of a bigger simulation, e.g. an earth system model. Perhaps, it is not even necessary to include some of the features of the system to still obtain a good, predictive model.

1.3.2 Eukaryotic genome structure compared to prokaryotes

This work focuses solely on the evolution of the prokaryotic genome and there are reasons to do so, despite the most fundamental properties of the genetic code and its expression being shared by virtually all the kingdoms of the tree of life: codon usage is fairly universal though not invariant (Osawa *et al.*, 1992), the structure of ribosomes is usually very similar though not always identical (rare exceptions are found e.g. among eukaryotes, where some linages of protists (mostly ciliates), some parasitic bacteria and mitochondria in a number of species differ from the general structure) (Knight *et al.*, 2001a,b; Jenner *et al.*, 2013), and replication mechanisms are also very similar (Chagin *et al.*, 2010). Nevertheless, eukaryotic genomes are characterised by a number of significant differences as compared to prokaryotic ones. Some of them are discussed below.

Eukaryotic chromosomes are often multiple (polyploidy) and always linear, which results a need for some kind of mechanism preventing their shortening and loss of genetic information after each replication as the replication protein complex on the lagging strand needs a number of nucleotides ahead of the replicon (the replicated area of DNA) to attach itself to the DNA strand (Lynch, 2007). Eukaryotic chromosomes often suffer from double-strand brakes, which are usually repaired by so-called non-homologous end joining (Lieber *et al.*, 2003; Daley *et al.*, 2005), which may bring a risk of accidental fusion of the chromosome ends, leading to instability during meiosis as some chromosomes would possess multiple centromeres (regions of chromosome used to separate them during cell division) and while others would have no centromeres at all (Lynch, 2007). These problems are solved by the existence of telomeres: repetitive distinctive sequences of nucleotides at the ends of chromosomes. The telomeric arrays of unicellular eukaryotes are usually short being under a few hundred bp in length, whereas those of plants and animals are about 5 to 150 kbp long (Louis and Vershinin, 2005). The presence of telomeres is usually accompanied by the presence of telomerase enzyme which is part protein and part RNA molecule. The RNA part consists of sequences complementary to the telomeric repeat that gives the telomerase the ability to elongate the telomere area via so called slippage mechanism (Chan and Blackburn, 2004; Lynch, 2007). In mammals telomeres are elongated normally only during early embryogenesis, whereas in somatic cells gradual shorting of the telomeres can be observed, e.g. in humans at a rate of up to 200 bp per cell division (Schaetzlein *et al.*, 2004). Circular prokaryotic chromosomes do not experience the problem of premature replication termination at the chromosome's end thus having no need for telomere presence.

Linearity and larger size of eukaryotic chromosomes call for some modifications of the replication process in comparison to prokaryotes. Prokaryotic single-chromosome genomes fall mostly in the range of 1.0-8.0 Mbp and have a single origin of replication site with only few exceptions like archaea from the *Sulfolobus* genera, which have 3 origins (Lundgren *et al.*, 2004). The rates of the replication forks progression vary from e.g. 50 kb/min in *Escherichia coli*, through 21 kb/min in *Caulobacter crescentus* to just 3 kb/min in slow-growing *Mycobacterium tuberculosis* (Hiriyanna and Ramakrishnan, 1986; Stillman, 1996). This allows for the replication of the whole genome, assuming good growth conditions, in the order of 10 minutes to 4 hours (Lynch, 2007). Meanwhile, eukaryotes have larger chromosomes, with average size between 10-250 Mbp for invertebrates, 40-200 Mbp for vertebrates and 30-2000 Mbp for land plants (Lynch *et al.*, 2006; Lynch, 2007). The rates of replication varies from 0.2 kb/min for angiosperms (Van't Hof and Bjerknes, 1981), through 1.7 kb/min for humans (Jackson and Pombo, 1998) to 2.9 kb/min for the *Saccharomyces cerevisia* yeast (Raghuraman *et al.*, 2001). Assuming a rate of 1.5 kb/min with a single origin on the chromosome it would take around 2 days to replicate a 10 Mb chromosome and over 3 weeks for a 100 Mb chromosome (Lynch, 2007). Eukaryotes prevent this slowdown by possessing multiple origin of replication sites on their chromosomes.

Another important difference is in the fraction of protein-coding DNA in the genomes of prokaryotes and eukaryotes. As mentioned above, eukaryotes have larger chromosomes and they also contain more genes. The 250 fully sequenced procaryotic genomes (data from the year 2006) posses between 350 to 8000 genes in 0.5 to 0.9 Mb long genomes. Meanwhile, fully characterised genomes of animals and land plants have on average $1.3 \cdot 10^4$ genes packed in a genome of at least 100 Mb (Lynch, 2007). Please note that these data are based only on fully annotated genomes. For viruses and prokaryotes the amount of protein-coding DNA scales almost linearly with the size of the genome, covering from 80% to 95% of the organism's DNA. The set of the smallest eukaryotic genomes shows similar scaling, but for genomes exceeding around 10 Mb in size the fraction of coding DNA decreases. For genomes of size around 100 Mb and larger (e.g. vertebrates and land plants), the fraction of coding DNA reaches a plateau occupying only 10% to 2% of the genome. The 10^4-fold range in genome sizes is accompanied by only 10^2-fold range in the amount of protein-coding DNA (Lynch *et al.*, 2006; Lynch, 2007). The question of why large genomes harbour so much non-coding DNA has been widely studied in recent years and will be addressed later in this chapter (see section 1.5.1).

Whereas prokaryotic genes are usually single, undivided sequences of DNA Joset *et al.* (1993), the eukaryotic genes are fragmented consisting of pieces of DNA containing information about the polypeptide chain (exons) and fragments which do not carry it (introns) (Berget *et al.*, 1977; Chow *et al.*, 1977; Gilbert, 1978, 1987). Small self-splicing introns are known to exist also in prokaryotes, but they never exceed 0.2% of the whole genome (Lynch *et al.*, 2006; Lynch, 2007). In the transcription process in eukaryotes a whole fragment of DNA containing both exons and introns is being transcribed into a single messenger RNA molecule (mRNA, serving as a template for protein synthesis), which later has the introns removed during the maturation process by a complex biochemical machinery called the spliceosome. A majority of multicellular eukaryotes have on average at least 5 introns per protein-coding gene, whereas eukaryotic unicells on average have less than one (Lynch, 2007). Existence of the exon-intron system allows to increase protein diversity by producing a number of alternative mature mRNAs from one DNA sequence. Mechanisms involved are: inclusions and exclusions of exons (exon skipping); mixing of various different exons (exon swapping); and modifications of 5' and 3' splice sites (Mayeda *et al.*, 1999; Smith and Valcárcel, 2000; Graveley, 2001; Roberts and Smith, 2002; Xing and Lee, 2006; Lynch, 2007). In humans, at least 75% of genes produces alternatively spliced variants of mRNA (Boue *et al.*, 2003; Johnson *et al.*, 2003). Even if spliceosomal introns arose in eukaryotic genomes as a result of a nonadaptive process, nowadays the ‘genes in pieces’ structure can provide a basis for further adaptive exploration in the search for new proteins (Lynch, 2007).

Yet another difference is the number of copies of chromosomes (ploidy). A large number of eukaryotes, with a majority of multicellular organisms, have their chromosomes in two copies (diploidy) at least at some point of their life. Chromosomes which constitute a pair, called homologous chromosomes, are almost identical in their basic structure (size, location of genes, cetromere location, telomere size) but at corresponding loci they may carry different variants of similar genes (alleles). This doubling of information, known as heterozygosity, can produce different phenotypic outcomes as compared to homozygosity (two identical copies of a gene on homologous chromosomes) regardless of which allele is residing at a given locus. Diploid organism may swap fragments of homologous chromosomes, most commonly during sexual reproduction, in a process called crossing-over (Lynch, 2007). Prokaryotes rarely have more than one chromosome.

There is a large number of other features characteristic for eukaryotic genomes but they are either characteristic for certain groups of eukaryotes or they are of minor importance to the topic of this dissertation. A thorough overview of the current understanding of the genome structure, with special focus on eukaryotes, can be found in e.g. the handbook *The Origins of Genome Architecture* (Lynch, 2007). The differences mentioned above, especially alternative splicing, multiplication of chromosomes (polyploidy) and the overwhelming amount of non-coding DNA, result in extra complication for any model of genotype-to-phenotype mapping in eukaryotes, thus this work will be solely focused on the evolution of prokaryotic genomes.

1.4 Modelling evolution of genomes

With the development of transistor-based computers followed by a fast progress in informatics and computer sciences the art of computer-based modelling flourished, supporting traditional ‘wet‘ biology. Many schools of thought and approaches have been developed, mostly inspired by long experience in building models in fields of mathematics and physics (for review see Morgan and Morrison, 1999b). Let us have a look on the most important developments.

1.4.1 Mathematical models in life sciences

Mathematical models entered the sphere of biology mostly via population studies with *An Essay on the Principle of Population* (1798) by Thomas Robert Malthus being probably the most prominent early example. Originally developed on the ground of economics, Malthus’ studies became an inspiration for the works of Charles Darwin and Alfred Russel Wallace. In these early days mathematical models were build mostly to be able to precisely predict future, similar to the models in physics of those days with Newtonian mechanics being the finest example. Newton’s theory can predict future location of a physical body basing on its properties (mass, size, velocity, current location) and fundamental properties of Nature. The same purpose lies behind the Lotka-Volterra equations (Foryś, 2005) which enables the prediction of changes in population size in time based on the species' simple-to-measure biological properties. These type of models give 'how actually' type of answers. If the Lotka-Volterra equations fit the prey and predator populations' actual sizes perfectly, then we can expect that this model and chosen parameterisation might be the actual explanation of the populations' dynamics, provided that the pray and predator populations really follow this type of principles. But in biology most often models of both types, i.e. mathematical and experimental ones, give a 'how possibly' type of explanations. Most biological phenomena are very complex in their nature, making them impossible to be fully recreated either in the laboratory or in a reasonably computable set of equations. And the researcher is aware of that imperfection from the very beginning of work, thus the positive output of research can only show a possible explanation of the phenomena. Only by gathering a large amount of 'how possibly' type of evidence, pointing in similar directions, can we construct reliable theories (Plutynski, 2001).

But we can build and use models in a different way than just for testing our theories or predicting future events. Having the phenomena well investigated and documented, the modeller can decide on the importance of different factors and use them to form a set of assumptions later used to build a model upon. A mathematical model like this is aimed at stretching beyond current knowledge and existing data and becoming a tool not to test the theory but to build new theories. In this approach models are not just fitted to match existing data, but rather they serve as a way of scouting where the data could lie in the huge space of possibilities given by the imperfect theory. It is not the model's ability to predict the 'right' kind of data that is the essence of the investigation, but the way the model's output changes in response to changes in the parameters or modifications of the underlying assumptions. In that way a model becomes an agent on its own, becoming a sort of technological tool mediating between existing theory and the existing data (Morgan and Morrison, 1999b).

A majority of known biological phenomena are not being explained by complicated mathematical theorems, in fact most of the mathematics in biology is quite basic. It is the overwhelming complexity of biological systems, the huge number of interactions and cross-links between systems' elements which make modelling in biology so challenging. Most of the models we build show a possible, not the actual, answers to the investigated problems. The same complexity also triggers a combinatoric explosion of possibilities predicted by theories, making many problems not only impossible to analyse by one-by-one approach in a laboratory experiment, but also difficult to compute. Fortunately, the rise of powerful computing tools allows us to construct more sophisticated models and analyse them within a reasonable time frame.

1.4.2 Essence of genes

When attempting to build a model of biological evolution the key question is: what will be the agent of its actions? Or, in other words, what will be the carrier of information about the fitness of the organism? First natural choice is a 'gene' which stands in the very centre of the modern theory of biological evolution if not biology as a whole. The link between the content of genes, their arrangement in the organism's genome and this organism's survival and reproduction forms the focus of interests of modern evolutionary biology (Noble, 2008). But this link is very hard to define and recreate in models. First of all, there is the problem of defining what a gene is as biology has a considerable number of definitions. A few of them are worth mentioning:

1. "Special conditions, foundations and determiners which are present [in the gametes] in unique, separate and thereby independent ways [by which] many characteristics of the organism are specified." – definition given by Wilhelm Johannsen who first used the term 'gene' in 1909, long before the discovery of the DNA chain (Gerstein *et al.*, 2007).
2. "A DNA segment that contributes to phenotype/function. In the absence of demonstrated function a gene may be characterised by sequence, transcription or homology." – definition used by the HUGO Gene Nomenclature Committee (Wain *et al.*, 2002). It is used for cataloguing human DNA sequences submitted to the Human Genome Organisation (HUGO).
3. "A locatable region of genomic sequence, corresponding to a unit of inheritance,

which is associated with regulatory regions, transcribed regions and/or other functional sequence regions." – definition adopted by the Sequence Ontology Consortium (Pearson, 2006; Gerstein *et al.*, 2007).

4. "The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products." – definition proposed after summarising the results of the Encyclopedia of DNA Elements (ENCODE) Consortium (Gerstein *et al.*, 2007). This definition tries to overcome the problems with 'classical' gene definition occurring in eukaryotic genomes e.g. the phenomenon of trans-splicing (ligation of two separate messenger RNA (mRNA) molecules, while in the human genome it has been shown that they might originate from two different chromosomes) and other cases when the final product can be rooted in multiple sequences spread on the DNA strand or overlapping on it. The definition focuses on the gene's final product achieved at the price of gene's discontinuity on the DNA strand. According to the authors, in simple cases where the gene is not discontinuous or there are no overlapping products, this definition gives in to the 'classical' one. In this approach a gene can get 'fuzzy' and becomes a "continuum of genes" (Pearson, 2006). It is also difficult to trace alleles with this definition.
5. "Locatable regions of DNA sequences with identifiable beginnings and endings." – minimalistic 'molecular-centric' definition of a gene (Noble, 2008). It is simple and quite generic thus it also covers fragments of chromosomes which do not code peptides but are evolutionarily conserved, e.g. some long non-coding RNA (lncRNA) which can take part in the regulation of gene expression (Kung *et al.*, 2013).

An approach in which a gene is seen purely at a molecular level has been criticised by some, probably most notably by the pioneer of systems biology Denis Noble. In his view, this way of defining genes detaches them from their functions. He also opposes putting the gene in the centre of biology and rejects seeing the genome as a 'program for life' (seeing the genome as an executable computer program which tells a cell what to do), calling it rather a 'database' where the cell stores some, but not all, information about its own structure (Noble, 2006, 2008). The position of genes in the big picture of cell physiology could form the starting point of an interesting and important debate but, unfortunately, it reaches beyond the scope of this thesis. We will focus on liking genes to a phenotype.

It can be seen that the term ‘gene’ began its history by referring to some unspecified agent in the gametes which stores information on the organism’s features, to become defined later as a locatable region (or a collection of regions) of a particular chemical molecule. In the meantime, the gene was losing its linkage with the organism’s characteristics directly related to its fitness, understood as reproductive potential in a given environment. The picture resulting from this short overview suggests the existence of some kind of a ‘genetic-phenotypic uncertainty principle’. The more specific one wants to be about a gene’s location, the less can be said about its impact on the organism’s survival and reproduction. The overall picture is further complicated by the discovery of epigenetics e.g. heritable methylation of DNA which influences patters of gene expression (Jaenisch and Bird, 2003).

This is certainly bad news for someone trying to find a link between genome size and the environmental conditions influencing the organism’s survival and reproduction. Yet let us look at this problem from the perspective of the most important force of evolution, i.e. natural selection. Using an anthropomorphic metaphor we can ask: Does the selection pressure care how and where the organism’s features are stored? In some important aspects yes, it does. A good example is when a particular selection pressure acts on the way this information is stored e.g. if strong UV radiation or high temperature is an important environmental factor, then the manner in which heritable information is kept in the cell does matter. Energies keeping atoms together, absorbed spectra of wavelengths, elemental composition of molecules and other chemical aspects are important for e.g. thermophiles (ecological class of prokaryotes found in hot environments e.g. thermal vents), which on average tend to have a higher percentage of guanine-cytosine pairs (G+C) in their DNA as the G+C Watson-Crick bond is stronger than the adenine-thymine (A+T) one making accidental double helix decoupling under high temperature less likely (Zheng and Wu, 2010). The chemical arrangement of the information carrier also determines the mutation ratio, thus influencing the population’s ability to adapt to the changing environment. But in most cases we can in fact assume that the selection pressure does not care how the heritable information is stored. Using DNA as a master template for protein production and the conservation of the genetic code are the corner stones of the theory of unity of all cellular life on Earth. If an organism has the right proteins, regardless of their origin, it will gain advantage in a given condition. A lot of organisms even ‘outsource’ some of the important characteristics to foreign genomes e.g some corals incorporate photosynthetic algae into their bodies and use them to obtain nutrients (Wooldridge, 2010); poison dart frogs belonging to the *Dendrobates* genus deter predators using a venom they get from the arthropods they eat (Daly *et al.*, 2003); while early eukaryotes acquired oxygen-breathing bacteria as a symbiont (which later became mitochondria), which allowed them to increase their metabolic efficiency (Andersson *et al.*, 2003).

In my opinion any characteristic of an organism can be considered as a subject of selection worth modelling if it is:

1. Durable – it can last through generation in a fairly stable form.
2. Relevant – in given environmental conditions a form of the characteristic has to have an impact on an individual’s survival and/or reproduction.
3. Variable – there are different forms of the characteristic among which the selection pressure can ‘pick the best one’.
4. Traceable – it has to be possible to determine what is the key factor which makes, in given conditions, individuals with one form of the characteristic perform better than individuals bearing a different form. Without traceability, it will not be possible to determine the cause of change in fitness or to track difference in time as species evolve. Note that in natural systems finding a link between genotype and phenotype is very difficult (Pigliucci, 2010) and this fact influences the ability to trace the changes that had an impact on species’ fitness, though to some limit such attempts have been done successful (e.g. Barrick *et al.*, 2009; Blount *et al.*, 2012).

Genes (their particular alleles) meet all these criteria: they last through generations having stable sequences; they stand behind proteins and/or regulatory processes which have an impact on an individual’s survival and/or reproduction; even conserved genes have the potential to alter different forms via mutations and mutations act slow enough not to interfere with the durability criterion; change of DNA sequence can be traced and compared with a corresponding sequence in any other individual in the population. Criteria 2. and 4. are very difficult to attribute in *in vivo* systems which results from problems with defining the genes mentioned earlier in this paragraph.

One can ask why genes, not proteins? Because in fact these criteria are met also by proteins. It is true, but it is also worth noting that protein sequences depend on the DNA structure and the DNA structure does not have any underlying template. Genes are structures found on the lowest level of the cellular organisation hierarchy still containing information necessary to build and maintain subsystems of the cell. Beneath genes there is only fundamental chemistry, but no unique information on how organisms differ from each other. Also, due to degeneracy of the genetic code, even putting post-transcription modifications aside, the same protein sequence can be coded by different DNA sequences (silent mutations), but a single gene defines only one specific protein sequence (with accuracy to an overlap of maximally three possible reading frames).

I would like to postulate that a modeller who wants to simulate a genetic-like system and its evolution in their response to environmental conditions has to build an abstraction of a gene meeting those four criteria. Embedding information about the DNA's internal structure (sequences, structure of the chromatin, genes loci etc.) is useful if investigating phenomena is happening on a sub-cellular level. But when investigating how the evolution affect characteristics not involving details of DNA structure, we can skip how genes are built. We can do so because the selection pressure will not care how we store information about those characteristics.

1.4.3 Building a genome

After discovery of the biological meaning of the double helix DNA, which came on top of the discoveries of Gregor J. Mendel and Thomas Hunt Morgan, genes became seen as sole determiners of the organism's physiology, morphology and anatomy, hence the phenotype. A metaphor of the genome as a 'blueprint for life' has emerged (Noble, 2008). But the development, in 1970's, of DNA and RNA sequencing methods and their rapid improvement in the past two decades gave rise to a large scale genome sequencing effort which started with publishing whole genome sequences of small viruses (Fiers *et al.*, 1976; Sanger *et al.*, 1977) and later led to the revealing of the *Homo sapiens* genome (Lander *et al.*, 2001; Venter *et al.*, 2001) and to recent sequencing of extinct species of humans (Green *et al.*, 2010; Krause *et al.*, 2010). This huge inflow of data gave a boost to a number of sub-disciplines, with bioinformatics (automated analysis of biological data e.g. of

DNA, RNA and protein structures) being the most prominent example. However, the data also revealed a much more complicated picture than the majority of researchers had expected. This encompassed, inter alia, the huge diversity of bacterial genomes, even among related lineages; pivotal role of horizontal gene transfer in the evolution of prokaryotes; importance of non-coding RNAs; alternative splicing in eukaryotes; and many more facts complicating possible explanations. This large body of discoveries (especially gene sequencing and gene knock-outs) have revealed complicated interplay between genes, their products and epigenetic interactions showing that the 'blueprint' metaphor is illusive. It became especially difficult to apply in the field of developmental biology (Alberch, 1991; Pigliucci, 2010).

Discoveries have shown that genes cannot be treated as equivalent to each other, which has been a common simplifying assumption in population genetics, as in the evolutionary process mutations in some genes are more crucial than in others (Stumpf *et al.*, 2007; Chouard, 2008; Pigliucci, 2010). For example, small mutations in genes widely spread among many taxa of animals such as the gene *hox*, *hedgehog* and *shavenbaby* responsible for the development of body plans in *Drosophila* flies can alter the morphology of epidermis. A similar important role is played in *Drosophila* by the *bicoid* gene but, surprisingly, it is not present in the genomes of most groups of insects other than dipterans (Stern and Orgogozo, 2009; Chouard, 2008). Functions played by the *bicoid* are spread between other regulatory genes. It thus becomes apparent that information on the functioning of a living organism is distributed between genes, developmental mechanisms and the environment (Oyama *et al.*, 2003). Denis Noble proposed a metaphor of the genome as a 'database' (Noble, 2006) where information about available routines is stored but which does not determine the whole organisation of an organism's life.

To determine the phenotype when knowing the genotype, one needs to focus more on the interactions between the genes rather than on the individual gene's outcome (Chouard, 2008). This gave rise to two directions of research, i.e. to genotype-phenotype mapping (Alberch, 1991; Pigliucci, 2010) and to network analysis (Strogatz, 2001; Proulx *et al.*, 2005). Genotype-phenotype maps are virtual representations of how genes and their alleles relate to particular phenotypes. In quantitative genetics they are usually based on statistical associations between alleles of genes and phenotypic features of organisms, whereas in systems biology they represent functional associations between genes and the phenotype based on introduced perturbations, e.g. gene knock-out, drug introduction (Landry and Rifkin, 2012). At the moment, due to the complexity of genotype-phenotype maps, RNA folding and protein functions are the best that can be done using this approach (Pigliucci, 2010). Network analysis is also quite a complex method, but it enjoys wide support from other disciplines such as computer sciences, physics and mathematics. Both approaches, i.e. analysis of data obtained from living systems (Isalan *et al.*, 2008; Kashtan *et al.*, 2009) and purely theoretical analysis of simulations (Kashtan and Alon, 2005; Newman, 2006; Crombach and Hogeweg, 2008), have been attempted. Nevertheless, the complexity of live models and insufficient experimental data shifted the scale of conducted research towards theoretical studies, which need to simplify reality to be able to draw conclusions. There has been a large number of genome abstractions proposed, varying vastly and dependent on the problems their authors tried to tackle. The most common generic representations are of the following types (Hindré *et al.*, 2012, Table 2):

- Programs – the genome is represented as a set of algorithmic instructions for certain computation and behaves like a 'domesticated computer virus' (Wilke and Adami, 2002).

- Allelic – the genome can be composed of fixed number of genes which can exist in a finite of infinite number of alleles represented as integers or characters.

- Network – the genome is a graph representing a gene-regulatory network (e.g. neural network, logic circuit) with no DNA abstraction. Mutations are changes in the graph's connectivity and/or number of nodes.

- String-of-pearls – the genome is an ordered string of different types of genes (metabolic, regulatory, binding sites, retrotransposons, etc.) and can evolve by rearrangement of elements, duplication, deletion, change in regulation.

- Sequence-of-nucleotides – most realistic type, some times involving a genetic pseudocode able to simulate point mutations. Also most computationally intensive.

For reviews on *in vitro* and *in silico* evolution experiments with microbes, see Hindré *et al.* (2012) and Mozhayskiy and Tagkopoulos (2013) publications. The choice of a particular approach is the result of a compromise between the need for an accurate representation of the real genetic systems, computational power available to run the simulations and the necessity to have a model which will be easy to analyse and allow to formulate solid conclusions, as complex models are also complicated to analyse.

1.4.4 Brief introduction to agent-based models

Agent-based modelling (ABM), sometimes also referred to as individual-based modelling (IBM), is a technique of constructing simulation models where individual entities (particles, cells, organisms, people, etc.) are represented. The origin of this approach dates back to the late 1940s, when John Von Neumann proposed the first theoretical concept of a self-replicating machine. His co-worker Stanisław Ulam proposed placing it on a grid, and that gave rise to cellular automata, with John Conway's *The Game of Life* being the most notable example. ABM was used through the years along with the rise in the efficiency of computers, but its development accelerated in early 1990's (Grimm, 1999), when computers achieved sufficient power.

Simulation models (a more general class of modelling techniques) are very useful when the researcher has a good understanding of basic rules underlying the phenomenon, but wants to investigate more complex behaviour of the whole system. As, for instance, in the case of Conway's *The Game of Life*, one knows simple rules deciding when a cell dies and when it reproduces, but would like to investigate the behaviour of a whole population on a given lattice. Or when one has data regarding the growth of an individual algae species, but is interested in the growth of a multi-species community.

Three ideas are central to the ABM concept: 'individuality', 'complexity' and 'emergence'. Assumptions made by the modeller are assigned mostly on a level of an individual entity and the number of these entities constitutes the investigated system. From the number of entities and the number of possible interactions between them there rises a huge number of possible states which can be entered into by the system, thus it is impossible to fully analyse it solely with 'pencil and paper' approach. This overwhelming complexity disproves a large body of efforts aimed at deducting systems' properties from just bare assumptions. System properties, on the level above individual entities, emerge as the simulation progresses.

Agent-based modelling, which is a variety of simulation modelling, has gained significant popularity in ecology (Grimm *et al.*, 1999; Grimm, 1999) because it allows to tackle questions in population ecology strictly within a pre-defined set of its paradigms: (1) a population consists of individuals; (2) the individual is the subject of natural selection; (3) properties on a population level emerge from interactions between individuals and between individuals and their environment. On top of that, living communities have a complex network of interactions, making them impossible to track on a sheet of paper using only analytical mathematics. Usefulness of the ABM approach was also noticed and appreciated in a number of studies in evolutionary sciences and ABM accounts for a vast majority of modern theoretical investigations in genome evolution (Hindré *et al.*, 2012; Mozhayskiy and Tagkopoulos, 2013) whether authors state that explicitly or not. In most cases an individual cell is represented as a genome in different forms of abstraction (see section 1.4.3).

1.4.5 Evolution of genetic networks under varying environmental conditions

ABM was found very useful for studying evolvability, which is a high level property of the genome and cannot be reduced to solely the presence or absence of a particular gene or even a set of genes. The questions lie in the way genes interact with each other and how they are regulated to be able to quickly form a new beneficial arrangement under new conditions. Studying regulation in living organisms is extremely difficult due to the overwhelming complexity of regulatory systems and a large body of facts that still remain unknown. A number of works use Boolean logic circuits to simulate regulatory systems in genomes, e.g. when a flock of circuits evolve under varying evolutionary goals, they tend to arrive in architectures which facilitate the emergence of beneficial mutations (Crombach and Hogeweg, 2008). This has been shown by setting the initial flock of circuits (genomes) meeting mathematical criteria of architecture resembling those seen in real gene networks and later allowing them to evolve towards specified properties of circuit architecture (the optimum phenotype is understood here as a desired architecture of the circuit, not the mathematical result that the circuit produces). Genomes consisted of linearly arranged metabolic genes and transcription factors (TFs) which regulated interconnections between genes thus deciding on the shape of the network. If the goal architecture was changing in the long run, it was observed that the dominant phenotype manifested arrangements which allows for network-wide rearrangement by altering only a few specific genes. At the same time the system was still robust to mutations in other sites of the network. In other words, when exposed to variable environments, the system evolved to being sensitive only to a small class of changes in the regulatory sub-system of the genome. Thus a case of developing evolvability with retaining system's robustness was observed.

It has been known that some knock-out mutations introduced in bacteria have little or no effect on their growth. For example, in *Escherichia coli* and the *Saccharomyces cerevisiae* yeast, 37% to 47% of metabolic reactions can be individually removed without blocking the production of any biomass component under any nutritional conditions (Wang and Zhang, 2009). Studies of artificial metabolic networks have shown that circuits selected for maximisation of biomass production from given metabolites which evolved in fluctuating environments had a higher number of multi-purpose enzymes than networks kept in stable conditions. Also, in the case of genomes from stable environments, knock-out of a mono-functional enzyme had a larger effect on the network than deleting a multi-functional enzyme in a genome from fluctuating environments. Variability of the environment was simulated by random fluctuations between three virtual nutrients given to genetic networks, whereas in stable conditions two of those three were selected randomly at the beginning and kept till the end of the simulation (Soyer and Pfeiffer, 2010). It can be observed that fluctuating nutritional conditions can lead to the emergence of robustness in metabolic networks. Also, it is worth noting that after the environmental regime was switched from fluctuating to stable, all robustness was lost, suggesting that the environment plays an absolutely pivotal role in maintaining it. However, analysis of metabolic reactions in *E. coli* and *S. cerevisiae* suggests that not all of this redundancy is necessary and its major part can be attributed to the ability of maximising growth efficiency in varied environmental conditions as different genes have different efficiency maxima in different conditions. Some other redundancy can be attributed, for instance, to recent horizontal gene transfer. But also *E. coli* and *S. cerevisiae* show opposing relationships between the functional importance of a reaction and its redundancy level, which might suggest that some of the redundancy, with its resulting robustness, is not a sign of adaptation to the fluctuating environment but a side effect of the evolutionary dynamics (Soyer and Pfeiffer, 2010) rooted in the stochastic nature of the evolutionary processes.

1.4.6 Artificial life and evolution – *Avida* software platform

An interesting approach, started by Thomas S. Ray's Tierra software, applies the so-called 'digital organism' scheme. Digital organism are self-replicating computer programs which have the ability to mutate and the 'digital environment' they occupy has traits they have to adapt to. Together, this allows for a simple evolution. In a way they are a "form of domesticated computer virus" (Wilke and Adami, 2002). The field of artificial life research developed a number of generic platforms for *in silico* research of the digital evolution e.g. *Avida* (Ofria and Wilke, 2004), *Tierra* (Ray, 1992) (*Avida's* predecessor)). Both systems consist of a flock of small computer programs, designed according to specific rules, which compete for central processing unit's (CPU) time and, in the case of *Tierra*, also for access to the main memory. This sort of approach allows to draw only the interesting properties from the natural system and run a large number of simulations in a very short time. Questions addressed with this system are quite diverse. *Avida* has been particularly successful in generating interesting results and it is subject to ongoing development by its creators.

Avida, used for genome complexity, robustness and gene interaction investigation, showed that complex programs are more robust than the simple ones with respect to the average impact of a single mutation. Also, when mutations are multiple, each next mutation shows a smaller impact on the organism's fitness if it runs a complex program (antagonistic epistasis). The genes were simple mathematical instructions which, stitched together, form a program performing more complex operations. Organisms are awarded, depending on their genetic content, with extra CPU cycles which lead to faster replication (Lenski *et al.*, 1999).

The same system was used to show that, under elevated mutation rate, organisms tend to have a slower replication rate, providing they occupy a flatter peak on the fitness landscape (Wilke *et al.*, 2001). High mutation rate can lead to negative effects like Müller's ratchet and/or mutational meltdown (see section 1.5.2), thus slowing down the replication rate will also slow down the mutation rate. Furthermore, being positioned in the fitness landscape on a flatter peak allows for larger robustness to deviations from the optimum position caused by mutations. Climbing on a higher, and most likely tighter, peaks brings, under a high mutation rate and high reproduction rate, a high risk of 'sliding off' the peak when mutations will temporally shift the population too far from the optimum. This leads to, on the first glimpse, paradoxical situations when a population with slower replication rate can out-compete a faster replicating one, as long as the mutation rate is high enough (Wilke *et al.*, 2001).

A particularly interesting case, from the point of view of this thesis, is using *Avida* for checking whether sexual reproduction will be maintained in a changing environment. This research showed that when values of the environment change rate are mild, sexual reproduction can be sustained, though the parameter space that favours sex is quite narrow. Also, it showed that sexual reproduction maintenance is easier than its origin *de novo* (Misevic *et al.*, 2010).

Avida has proven its usefulness in tackling a number of other problems, e.g. adaptive rise of radiation from resource competition (Chow *et al.*, 2004), selective pressure on genomes (Ofria *et al.*, 2003), or the origin of complex features (Lenski *et al.*, 2003). Yet it cannot be easily extrapolated on a living system, especially with its very abstract representation of a gene as a mathematical operation. Nonetheless, it allows for a handful of fruitful speculations which could result in taking a new direction in *in vitro* research.

1.5 Modern views on genome size constraints

As shown before variability of the environmental conditions can have an impact not only on species ecology but also on its evolution and ultimately on the shape and organisation of its genome. But what are the factors shaping the size of genomes? It has been argued that there are important differences between prokaryotes and eukaryotes in terms of genome size, organisation and expression (see sections 1.3.1 and 1.3.2). Are there differences in these two kingdoms of life also in terms of forces deciding on genome sizes? If yes, the question is what makes prokaryotes' genomes the size we observe?

1.5.1 Why are eukaryotic genomes so big?

The first striking observation made when comparing prokaryotes and eukaryotes is the cell size difference: an average unicellular prokaryote has circa 10^5 to 10^7 smaller biomass than the average single-cell eukaryote. Similar disparity can be observed between unicellular and multicellular eukaryotes (Bonner, 1988; Lynch, 2007). Much smaller cell size is followed by a huge abundance of individuals: it is estimated that the total number of prokaryotic cells on Earth is in the unimaginable range of $4{\cdot}10^{30}$ to $6{\cdot}10^{30}$, constituting up to 350–550 Pg [10^{12} kg] of carbon, and being approximately 60-100% of carbon built into all plants (Whitman *et al.*, 1998). By analysis of ribosomal RNA in ocean water samples, a single clade of marine SAR11 α-proteobacteria is estimated to consist of $2.4 \cdot 10^{28}$ cells worldwide (Morris *et al.*, 2002), being probably the most abundant organism type on the planet. The average prokaryotic species global population consists probably of around 10^{23} cells (Lynch, 2007). Despite this huge number of individuals, there are only 10^4 species of bacteria and archaea known to science (Oren, 2004), as compared to $1.5 \cdot 10^6$ named eukaryotic species, about half of them being insects and circa $3 \cdot 10^4$ being protists (Lynch, 2007). How come eukaryotes, despite being less numerous, have developed larger diversity, larger cells and larger genomes?

A hypothesis which tries to put the above mentioned facts together comes from population genetics and the neutral theory of molecular evolution (NTME), introduced by Motoo Kimura (1983). The NTME states that mutations are rarely beneficial, in fact most of them are insignificant as they neither bring important improvements nor are strongly deleterious. The majority of changes happening in evolutionary processes are neutral modifications not affecting fitness of the individuals and in rare cases, due to random genetic drift, they reach fixation, becoming dominant in the population. Examples of neutral changes are silent mutations: substitutions of nucleotides in the DNA which, due to degeneracy of the genetic code, do not alter the amino acid composition or the resulting polypeptide. Also if allozymes, i.e. variants of an enzyme coded by different alleles of the same locus, do not show different properties in terms of their enzymatic activity nor in terms of biochemical stability, they can be considered neutral in the light of the NTME (Skibinski *et al.*, 1993). In the classic Wright-Fisher model of population genetics (Wright, 1931; Fisher, 1949), population is assumed to consist of N diploid hermaphroditic individuals mating randomly. This is rarely the case in any natural species, however most of these complications can be overcome by introducing a new measure: the genetic effective population size N_e, where always $N_e < N$. A large body of literature is dedicated to solving technical problems with estimating N_e (e.g. Caballero, 1994; Whitlock and Barton, 1997; Rousset, 2003). Known factors which decrease the genetic effective population size are associated with the difference in the number of gametes produced by individuals, e.g.: spatially structured population; patchy distribution of individuals; a sex ratio different than $1 : 1$; population size fluctuations, e.g. series of bottlenecks. All this will impact the number of genes each individual is contributing to the next generation, decreasing N_e in comparison with N (Lynch, 2007). In some of low-fecundity vertebrate species N_e can be as low as circa 10% of all breeding adults in the population (Frankham, 1995). If we want to find the probability of a mutation, which is reducing an individual's fitness by the selection coefficient s, getting fixed (in an infinite population it would always eventually perish), we have to take into account the effective number of genes residing at a locus at the time of reproduction (N_g), which can be considered, in a randomly mating diploid population, as the equivalent of doubling the effective number of individuals in a haploid population ($2N_g$). In finite populations, the stochastic allele-frequency change is given by $1/N_g$ and natural selection will be effective at removing the deleterious mutation if directional change dictated by the selection coefficient s will be greater than $1/N_g$. If s is significantly smaller than $1/N_g$, then this deleterious allele will behave as it though were neutral. Its fixation, or loss, will be dictated purely by chance. For a neutral mutation, its fixation probability is equal to its initial frequency, but under the NTME probability of fixation of a deleterious allele is:

$$\theta_f \approx \frac{2N_g s}{e^{2N_g s} - 1} \tag{1.1}$$

as given by Kimura (1962) and by Lynch (2006b). For sufficiently small $2N_g s$ the probability of fixation θ_f approaches 1.0, which is the expected value under neutral conditions, for example for $2N_g s = 0.1$, $\theta_f \approx 0.951$. But for large $2N_g s$, θ_f rapidly approaches zero, e.g. $2N_g s = 1$, $\theta_f \approx 0.582$ and for $2N_g s = 10$, $\theta_f \approx 0.00045$. Under the assumptions of the NTME, for the selection coefficient s we can expect that in any population with $2N_g s \gg 1$ a deleterious mutation has a negligible chance of drifting to fixation (Lynch, 2006b). As s is difficult to estimate, it can be replaced with two other factors: the number of nucleotides which have to be intact to maintain the function of a gene n and the rate μ, at which each of n hazardous sites mutates per generation. Then piece of excess DNA will shift the rate of production of defective alleles by nu upwards and that can substitute s. We end up with the following inequality $2N_g u \neq 1/n$. If $2N_g \gg 1/n$ (and this is the case for prokaryotes and other microbes), then it is enough to prevent significant colonisation of introns, mobile elements, and excess intergenic DNA. Whereas in multicellular eukaryotes $2N_g$ is sufficiently low to allow the introduction of non-functional DNA in the genome (Lynch and Conery, 2003; Lynch, 2006a,b, 2007). It is argued that the first non-coding DNA, transposons, variety of mobile elements of DNA including endogenous retroviruses, pseudogenes (genes which no longer produce proteins, but still have a recognisable coding sequence), gene duplicates, etc. arose as truly 'junk DNA', as in relatively small populations selection pressure was too weak to purify the genome from invading elements. Later in the history of the affected lineages some of these elements gained significance allowing for the rise of larger genome complexity and new layers of regulations. Existence of smooth transitions in the amount of coding DNA and size of intrones from viruses, through prokaryotes and unicellular eukaryotes to animals and plants (Lynch, 2006a, Figure 1) supports this line of argument. A more detailed description of this concept can be found in *The Origins of Genome Architecture* by M. Lynch (2007). The simplicity of this concept is its strong side, but estimating N_e and N_g in natural populations is problematic. This generic approach brings an interesting insight into the broad pattern, but faces some difficulties when trying to explain differences in genome size between individual species (Charlesworth and Barton, 2004). The line of argument driven from population genetics also fails to explain an important questions: why have eukaryotes evolved only once (e.g. Stechmann and Cavalier-Smith, 2002; Rivera and Lake, 2004)?

Another hypothesis is limitation via the amount of energy available per gene. Genome replication is relatively cheap, as the cost of DNA replication it self accounts for just 2% of the energy budget of a microbial cell during growth (Harold, 1986). The real cost is the proteins synthesis which constitutes up to approximately 75% of a cell's total energy budget (Harold, 1986). That includes all types of proteins: structural ones, enzymes, signalling and regulatory proteins, toxins, etc. It is worth noticing that studies on the *S. cerevisia* yeast showed that the fitness cost of expressing a protein depends rather on its structure than on its function (Tomala and Korona, 2013). It was shown that proteins with transmembrane regions and those with a high proportion of protein length occupied by sequences predicted to be loosely shaped (intrinsically disordered) are especially damaging to fitness when overexpressed (Tomala and Korona, 2013). Growing proteobacteria have an average metabolic rate of 0.19 ± 0.5 W·g^{-1} and mean mass of $2.4 \cdot 10^{-12}$ g. Actively growing protozoa have a mean metabolic rate of 0.06 ± 0.1 W·g^{-1} and mass of $4.01 \cdot 10^{-16}$ g. The average metabolic rate per cell for proteobacteria is 0.49 pW and for protozoa $2,286$ pW. Taking into account the cell size difference and metabolic rate, differing only by a factor of three, it can be seen that an average protozoan has *circa* $5 \cdot 10^3$ more metabolic power than a single proteobacterium (Lane and Martin, 2010). But the metabolic power per mega base is different by one order of magnitude between bacteria and protozoa, that is about 0.08 pW·Mb^{-1} for an average bacterium (assuming 6 Mb of DNA in a gene) and 0.76 pW·Mb^{-1} for an average protozoan. These fairly similar numbers contradict the approximately 10^4 fold difference in genome sizes between these two groups (Lane and Martin, 2010). When we consider power (Watts) per gene available and additionally take ploidy into account, an average bacterium has 0.03 fW per gene while an average eukaryote has 57.15 fW per gene (note that, as discussed previously, eukaryotic genes are arranged on the DNA strand slightly differently from prokaryotic genes). A more detailed comparison, including cell's size classes, can be found in Lane and Martin (2010), Table 1. Authors attribute this difference to the fact that eukaryotes acquired mitochondria and are able to generate a large surplus of power in comparison to prokaryotes. They argue that it is not the higher complexity and sophisticated regulation of the eukaryotic cell which allows for embedding mitochondria into the cellular machinery, but the possession of mitochondria provides extra power which allows for relaxation of selection pressure on energetic efficiency and, consequently, sustaining more proteins in the cell. This, in turn, drives higher diversity.

Structural arrangement of the chromosomes has been considered as a reason for prokaryotic genome size limitation as dominant circularity of the prokaryotic chromosome can be a cause of certain constrains (Bentley *et al.*, 2002). But the existence of prokaryotes with linear chromosomes (see section 1.3.1) rises the question of why are there so few bacteria or archaea with linear chromosomes and cell complexity similar to eukaryotic cells? Those eukaryotes with small genomes are rather similar to a majority of prokaryotes in terms of their ecology e.g. population size and structure, reproduction strategies, nutrient limitations and not vice versa.

Population-genetic-based line of argument and energetic limitation do not oppose each other. Selection pressure, whose strength depends on population size, effectively removes deleterious mutations in numerous species and enhances the chance of fixation of energetically efficient genetic lineages, but in small populations it leads to accumulation of non-coding DNA, selfish genetic elements which can spread themselves independently from the rest of the genome, mobile elements and other kinds of 'junk DNA'. On the other hand surplus of power provided by the mitochondria allows species affected by accumulation of excessive DNA not only to overcome the negative effects of possessing unnecessary DNA but to use e.g. gene duplication and even intronic insertions (see section 1.3.2) as a source of innovation leading to a larger diversity of genes in the eukaryotic domain of life. Also, in this context, it is worth remembering that regulatory systems, which are more sophisticated in eukaryotes than in prokaryotes, are based mostly on protein activity.

Unity of basic mechanisms of evolution in all domains of life, from viruses to multicellular eukaryotes is the basic paradigm of modern biology, nevertheless differences in genome size, organisation and expression regulation between prokaryotes and eukaryotes, followed by differences in cell physiology and in population dynamics justify treating evolution of prokaryotes in a slightly unorthodox way. As already mentioned, they constitute a large portions of Earth's biomass and play an important role in our planet's circulation of elements. Also, most likely they were the first cellular form of life on Earth which later give rise to all eukaryotic organisms (Koonin, 2011). Thus focusing solely on prokaryotes is justified from the point of view of ecology and history of life and it will be done so in this study.

1.5.2 Processes that influence prokaryotic genome size

In their review on bacteria and archaea genomic Koonin and Wolf suggest existence of five kinds of forces shaping the size of prokaryotic genomes (Koonin and Wolf, 2008;

Koonin, 2011):

- Genome degradation
- Purifying selection
- Gene or whole operon duplication
- Horizontal gene transfer (HGT)
- Replicon fusion

Two more mechanisms, responsible for genome degradation, can be added to this list, namely:

- Müller's ratchet
- Mutational meltdown

The list can be further extended two include two additional factors of more generic nature, based on stochastic processes in finite populations (Lynch, 2007), i.e.:

- Genetic drift
- Genetic draft

Let us now have a look at this catalogue.

Genome degradation leads to the reduction of the number of genes among parasitic and endosymbiotic prokaryotes. Most of the lost genes are involved in metabolic tracks no longer necessary in a parasitic lifestyle e.g. genes underlying biosynthesis of amino acids, but surprisingly also those involved in replication, transcription and translation are being lost, too (Andersson *et al.*, 1998; Moran and Wernegreen, 2000; Moran, 2002). The loss of those genes probably triggers lower G+C content observed in obligate pathogens, as A+T rich DNA is less prone to errors introduced at replication, especially integration of uracil instead of cytosine (Glass *et al.*, 2000; Moran, 2002; Dufresne *et al.*, 2005). Many of those changes can be explained by the genetic drift, as populations of the parasitic and endosymbiotic prokaryotes are often limited in numbers by specific dispersive environment they occupy and are exposed to numerous bottlenecks.

Purifying selection acts on the genomes of the organism which are highly abundant and which renders their populations prone to weak selection (Lynch, 2006a). A good example are the free-living cyanobacteria of the *Prochlorococcus* genus. These oxyphototrophic organisms are widely distributed and, at the same time, highly specialised. The niche of most of the *Prochlorococcus* strains are euphotic, oligotrophic waters of low latitude oceans (between 40 °N and 40 °S) (Partensky *et al.*, 1999). This is an environment with very stable physical conditions and low nutrient supply, and also it is spread on a huge fraction of the planet's surface. Enormous population size and selection pressure acting in a constant direction jointly provide conditions where even a small adaptive change will spread very rapidly. Streamlining of the *Prochlorococcus* genome is reflected not only in the low gene number, but also in its compactness and extremely low number of pseudogenes and integrated selfish elements (Koonin and Wolf, 2008), which enhance their own transmission relative to the rest of an individual's genome, but are neutral or even harmful to the individual Werren (2011).

Duplication of a gene or of a whole operon may lead to two kinds of advantages: in rare cases, when a duplicated gene or an operon play an important role, the cell can benefit from having an elevated number of mRNA in comparison to its ancestor, meanwhile the other possible option is neofunctionalisation when a duplicated gene or an operon gains a different function. Rarely it is an entirely new function and in most cases the functions are related (Zhang, 2003). This types of mutations occur as a result of an error in establishing the replication forks or by exchange of DNA between the chromosome and plasmids.

Horizontal gene transfer (HGT), sometimes also referred to as lateral gene transfer (LGT), and replicon fusion are both similar in their nature, as they result in the acquisition of new genes originating from outside the cell's chromosome. These mutations are not a result of replication faults. HGT is when one species gains a new genes which originate from another species, often not related to the recipient species. The most common ways of delivering a new gene are: transformation (acquisition of exogenous DNA); conjugation ('deliberate' exchange of DNA in cell-to-cell contact); and transduction (moving DNA between different bacteria by a viral vector). HGT plays a prominent role in evolution of the prokaryotes, e.g. it is estimated that between 10% and 16% of the *Escherichia coli* chromosome was built through HGT (Ochman *et al.*, 2000). It has also been a very important evolutionary mechanism in the rise of eukaryotes (Rivera and Lake, 2004) and will be discussed in more detail in Chapter 4 of this thesis.

Replicons are DNA or RNA particles which replicate from a single origin, i.e. replication site. Among prokaryotes most common replicons include chromosomes, plasmids and phage DNA/RNA. Their incorporation in the host chromosome is usually an effect of the RM system fault. Some of them actively force theirs incorporation by containing a toxin/antitoxin module (TA). A common mechanism is that the toxin is a more stable compound than the antitoxin and by that it will kill the cell when it will manage to remove the TA-containing DNA from its genome (Buts *et al.*, 2005; Gerdes *et al.*, 2005). This can enforce evolution of the species towards fusing the invasive replicon into the chromosome for the sake of ensuring that this ‘self destruct’ mechanism will not damage the cell.

Müller’s ratchet is is a term introduced to explain the advantages of sexual reproduction in comparison to the asexual one (Müller, 1964; Felsenstein, 1974). Chromosomes of asexually reproducing organisms are transferred to next generation as single unchangeable blocks and then they are affected by rare changes, majority of which are disadvantageous. This must lead to an accumulation of deleterious mutations e.g. deletions of genes which later cannot be regained. Drop in fitness, in comparison to the situation when all individuals in the population would have unaffected genomes, called the genetic load, will increase, eventually leading to the population’s extinction. Selection forces are counteracting this degradation by promoting the fittest cells for reproduction, but positive selection has its limits. Each step towards genome degradation is unlikely to be reverted, but exchanging parts of the genome with other individuals may restore lost genes in the lineage and the ratchet mechanism can be averted. However, asexual unicellular organisms also have other means of avoiding genome degradation by deleterious mutations. One is horizontal gene transfer (discussed in detail in Chapter 4), allowing to acquire new or previously lost genes even from distant linages (Koonin, 2011). Also division of the cell’s genome into a number of quasi-chromosomes, such as e.g. plasmids and other mobile genetic elements can help overcome the ratchet mechanism. Such small chunks of genetic information are easily exchangeable and when damaged by mutation they can be obtained from the environment and expressed without the risky process of integration into the chromosome. So why don’t genomes consist mostly of flocks of small mobile elements? Large single chromosome might give advantage in maintaining complex regulation and it may pay off to have the majority of genes located on a big central storage molecule with only some of the information stored in mobile elements. Müller’s ratchet may seem to be the same thing as genome degradation, but it is not quite the same process.

Genome degradation can be a factor contributing to the ratchet mechanism, but the ratchet work in a longer time frame and unavoidably leads to extinction of the species.

Müller's ratchet does not consider what impact constant genome degradation affecting fitness will have on the species population size and this is done following the mutation meltdown model (Lynch *et al.*, 1993). The process happens in three stages and affects populations which are small and derived from a common founder individual thus having small genetic variability. In the first phase, a number of deleterious mutations accumulate, increasing overall genetic variability in the population and eventually reaching values expected under drift-selection-mutation equilibrium and moving on to the second phase, when variability is large, making the selection process most efficient. This leads to slowing down the rate of accumulation of deleterious mutations but, at the same time, selection decreases variability up to the point where number of surviving offspring is less than one per every adult. This moment marks the third stage, where population size declines entering the meltdown phase. The small number of individuals leads to weakening the selection process and accumulation of deleterious changes at an increasing rate, reaching the mutation rate and finally the population goes extinct. Diploid organisms can avoid meltdown because they can mask a deleterious gene with a valid second copy and are able to repair DNA using the homologous chromosome as a template (Lynch *et al.*, 1993). Haploid organisms can also evade meltdown if they have sufficiently big populations, preventing the overall population fitness from falling too low.

Genetic drift and draft are high-level mechanisms underlying many of the above mentioned processes. Drift becomes increasingly important in small populations and may lead to fixation of deleterious or neutral mutations (more in section 1.5.1). Draft is less linked with population size and it leads to fixation of neutral of slightly deleterious mutations if they are linked (e.g. they are located close on the DNA strand) to a beneficial allele which increases fitness (Brenner *et al.*, 2002; Lynch, 2007).

Recent observations the suggest existence of an upper limit on the prokaryotic genome size. As it was said before (section 1.3.1), an increase in the number of genes in the genome triggers a rise in the number of regulatory genes with a power-law factor of 2 (van Nimwegen, 2003; Molina and Nimwegen, 2008; Molina and van Nimwegen, 2009; Koonin, 2011, p. 97). With the growth of the genome size, this 'cellular bureaucracy burden' can eventually become too big for the genome to be sustainable. Koonin and Wolf refer to that number as the "Van Nimwegen Limit" estimating it to be approximately 13 Mbp (Koonin and Wolf, 2008), just above the size of the largest known bacterial genome of *Sorangium cellulosum* of the size 12.2 Mbp (Pradella *et al.*, 2002).

1.5.3 Reasons for prokaryotic genome size constrains

Ultimate explanations for the size limitation among prokaryotes are still open to debate. Comparison with eukaryotes shows that prokaryotic genomes seem to be not only smaller but also undergo more severe size constrains. Eukaryotic genomes sizes span from 10^1 to 10^5 Mbp, which is a four-orders-of-magnitude difference, whereas prokaryotic genomes vary only by one order of magnitude: from slightly less than 1 Mbp to approximately 13 Mbp (Casjens, 1998; Gregory and Hebert, 1999; Mira *et al.*, 2001; Koonin and Wolf, 2008). Seeking the reason of this impressive difference is one of the approaches taken in investigating genome size limitations.

It has been hypothesised that prolonged nutrient limitation (mostly P, but also N) acting together with pressure for faster growth rate might lead to genome streamlining in eukaryotes (Hessen *et al.*, 2010). High growth rate is coupled with high RNA concentration in some eukaryotes (Elser *et al.*, 1996, 2003) which calls for high P demand. If an eukaryote faces nutrient limitation, mostly via phosphorus, it is forced to let go of some genes to reduce 'spendings' on RNA and keep a high growth rate (Hessen *et al.*, 2010). Knowing how small the genomes of *Prochlorococcus* from low-latitude nutrient-depleted oceans are, this concept sound also promising for prokaryotes. Nevertheless, it all sounds a bit too simple, even just for eukaryotes. Free-living prokaryotes have an extremely wide range of metabolic pathways as ways of acquiring nutrients and this explanation might be true for some of the taxa, such as α-proteobacteria, but it can barely serve as a generic explanation. Also numerous eukaryotes have strategies of overcoming nutrient shortages, e.g. many animals have some kind of structural body support (skeleton, shells, exoskeleton) which can also serve as a storage organ for essential elements in shortage. Oceanic gyres are a type of environment which would support this kind of limitations, but there is a number of environments which change in time, including their nutrient regime, like coastal upwelling zones.

Another hypothesis reflects on issues of genome regulation. Bacteria, with their fairly simple operon-based gene expression regulation system, are limited in their ability to manage any possible interference between different proteins and metabolic paths when expressing a large number of genes. Eukaryotes can expand their genomes further because they have a number of extra layers of regulation, suh as the intron-exon system, DNA methylation, nucleosomal chromatin and cellular compartmentalisation (Bird, 1995). The optimum size is reached when the bacterial genome reaches maximum metabolic complexity (which generates 'revenue' for the cell) for a minimum number of regulatory genes (which can be treated as the logistic cost) (Ranea *et al.*, 2005). It is tempting to ask if the maximum possible genome size which could be constructed using this hypothesis could be equivalent to the mentioned previously "Van Nimwegen Limit"?

Interestingly, differences in the way prokaryotic and eukaryotic chromosomes are organised are rarely discussed in the light of genome size difference. Circular shape of the prokaryotic chromosome may be expected to be more problematic for expansion, by causing difficulties in storage and handling in a small cell in comparison to the linear eukaryotic chromosomes. It also raises a chicken-or-egg type of question: is the reshaping of the chromosome a cause or a consequence of genome size expansion?

Except hypotheses rooted in population genetics and neutral selection theory, most of the possible explanations focus on the issues of nutrient supply and its allocation. Population genetics has its own well-established mathematical apparatus and the ABM approach is not the best way to handle it. Thus this investigation will focus on possible trade-offs between the cost of having a gene (in terms of expression costs and costs of genome regulation) and the need to allocate sufficient resources to growth and reproduction.

1.5.4 Number of genes and size of prokaryotic gnomes

As the growing number of genomes sequenced in the last two decades makes a vast amount of genomic data available (van Nimwegen, 2003; Koonin and Wolf, 2008), scientists started noticing certain regularities not only in the sequences of genes constituting genotypes but also in their number and in the way genes associate with each other or, in other words, in the structure of the genome.

One of the outstanding facts discovered is the bimodal distribution of bacterial genome size with a larger peak around 2 Mbp and smaller near 5 Mbp. The total span of bacterial genome sizes ranges from circa 180 kbp for *Carsonella rudii* up to around 13 Mbp for *Sorangium cellulosum* (Koonin and Wolf, 2008). The distribution of genome sizes in archaea seems to be simpler. Their range is narrower starting at about 0.5 Mbp for parasitic *Nanoarchaeum equitans* to circa 5.5 Mbp in *Methanosarcina barkeri*, showing only one peak at about 2 Mbp, so around the same size as the larger peak in bacteria (Koonin and Wolf, 2008). More detail is shown in Figure 1.3.

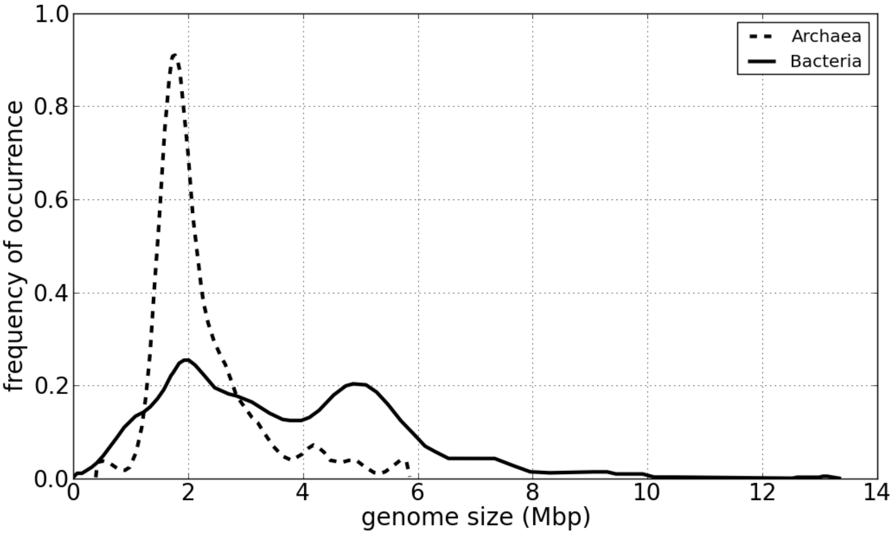

Figure 1.3: Distribution of genome sizes in bacteria and archaea (Koonin and Wolf, 2008). Authors have obtained distribution curves by Gaussian-kernel smoothing of individual data points. Solid line – bacteria, dashed line – archaea.

Another important pattern is the clustering of genome sizes in prokaryotes having different life styles, with obligatory symbionts and obligatory parasites nearer the lower end of the genome size spectrum and free living prokaryotes closer to the higher values (Figure 1.4). Of course, the transition is not sharp and e.g. *Pelagibacter ubique* (a small heterotrophic marine α-proteobacteria from the SAR11 clade) has only 1354 open reading frames (1308759 bp), which makes it the smallest known free-living organism genome, smaller even than a number of the bigger parasitic prokaryotes' genomes (Giovannoni *et al.*, 2005). Yet it still is positioned quite high among parasitic unicellular organism.

A different kind of pattern worth mentioning is the dependency of the number of genes belonging to a different functional category on the size of the genome. A number of researchers have shown a power-law ($y = ax^b$) interdependency between the number of

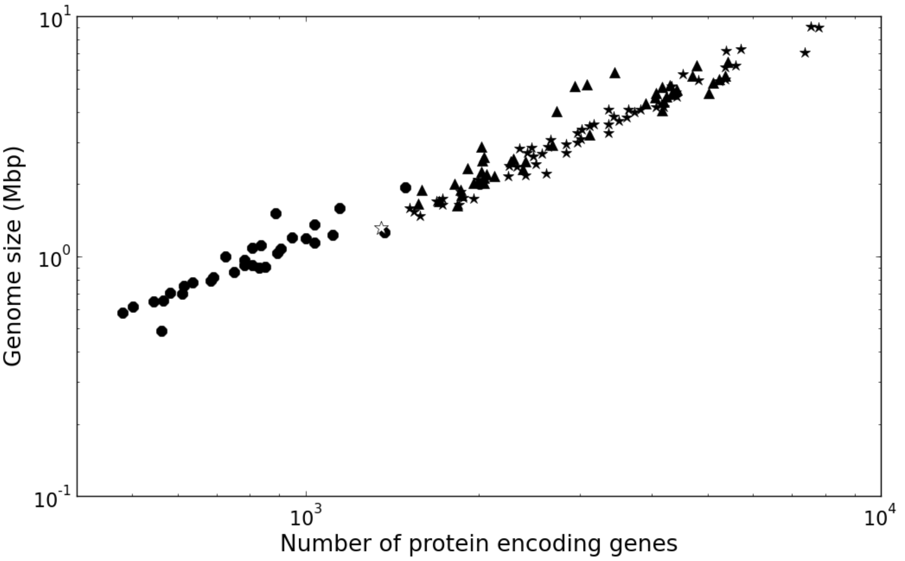

Figure 1.4: Number of predicted protein-encoding genes versus genome size for 243 complete published genomes from bacteria and archaea (Giovannoni *et al.*, 2005). Circle (•) represents obligatory parasites/symbionts, star (⋆) – free-living species, triangle (▲) – host-associated species, white star represents *Pelagibacter ubique* whose genome is the smallest among all the sequenced free-living prokaryotes.

genes in a functional category and the genome's size in eukaryotes, bacteria and archaea (Stover *et al.*, 2000; van Nimwegen, 2003; Konstantinidis and Tiedje, 2004; Galperin, 2005; Ulrich *et al.*, 2005; Koonin and Wolf, 2008; Koonin, 2011, [p. 97). In his fairly detailed research, van Nimwegen (2003) divided genes into three major groups: a large category of metabolic genes, cycle-related genes and transcription regulatory genes. He showed that metabolic genes occupy roughly the same fraction of the genotype as its size grows (exponent is the power-law expression roughly equal to 1), the cycle-related genes fraction shrinks with genome size (exponent smaller than 1) and the transcription factors have their exponent significantly above 1. In bacteria, transcription factors (TFs) have an exponent of almost 2, which means that when the size of the genome doubles, the fraction occupied by TFs quadruplicates (more detailed data are shown in Figure 1.5). In their review, Koonin and Wolf (2008) distinguished also a group of genes involved in cell division which they showed to have an exponent equal nearly to 0 (the number of those genes is similar in all species and is independent from the size of the genome). All these findings suggest that the larger the genome becomes, the larger the effort for its 'managing' and certainly this cannot be without implication for genome structure design (e.g. operon structure, expression regulation, genetic networks) and size in bacteria.

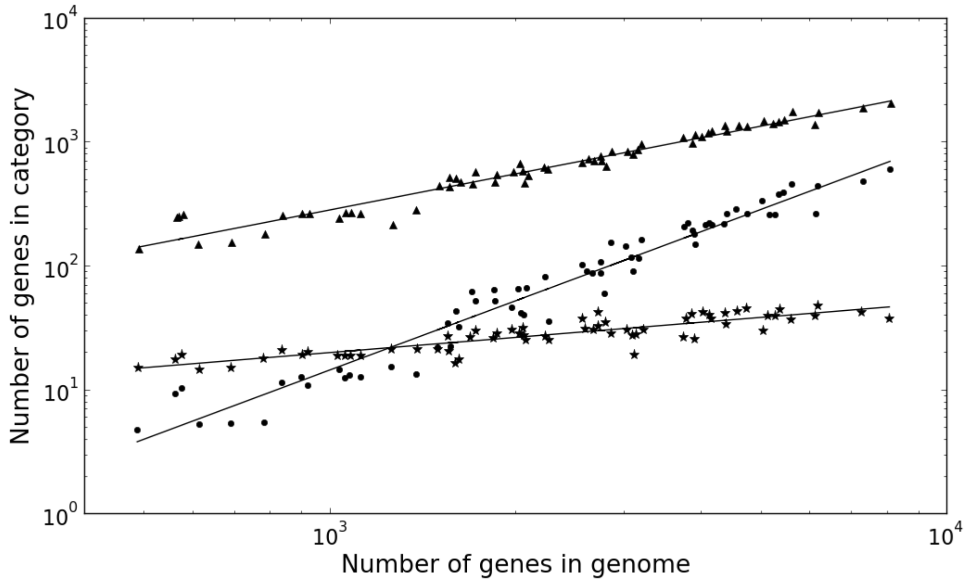

Figure 1.5: Number of genes of different functional types as a function of the total number of genes in the genomes of bacteria. Both axes are shown on logarithmic scales. Each point corresponds to a genome. The straight lines are power-law fits (van Nimwegen, 2003). Circle (•) represents transcription regulatory genes, star (⋆) – cell-cycle-related genes, triangle (▲) – metabolic genes.

All the results mentioned above suggest the existence of forces which on the one hand limit the maximum span of prokaryotic genomes and on the other do not allow them to become smaller than a certain minimum containing the most essential gene set. What kind of evolutionary selective pressures shape the genome size?

Some hopes for an interesting case study have been linked with thermophiles (many of them being archaea) which are a broad ecological class of microorganisms living in hot environments (between $45°C$ and $122°C$). It is also suspected that thermophilic bacteria are one of the oldest known phyla of bacteria (Horiike *et al.*, 2009). It was hoped that the ancient lifestyle and specific environmental niches they occupy (majority of proteins denature above $45°C$) will force them into developing unique features of genome organisation. To some extent, they do differ from mesophiles, known for favouring lower temperatures: as already mentioned previously, they tend to have higher G+C content in many genes (Zheng and Wu, 2010). Also, hyperthermophiles universally possess the reverse gyrase, a protein that is strictly required for DNA replication at extremely high temperatures (Forterre, 2002); and it has been demonstrated that thermophilic proteins tend to have higher charge density and a larger number of disulphide bridges (Beeby *et al.*, 2005), which both help stabilise the molecule structure. But not much more was found, and there are no significant genetic markers allowing to determine ecological preference of thermophiles solely on the genome sequence (Koonin, 2011) thus 'reverse ecology' seems to be not possible. Thermophilic life style puts a number of challenges on the physiology, forcing a fairly narrow set of solutions. Thus horizontal gene transfer (HGT) between thermophilic species has been observed (Koonin, 2011), even between archaea and bacteria. But, apart from temperature preference, there can be quite a wide range of differences between their ecology, e.g. towards optimum nutrients. They can be either chemoautotrophs or heterotrophs and live in deep sea hydrothermal vents, hot springs or decaying organic matter (e.g. compost), depending on the species.

An interesting case study comes from a different extremophile: the *Deinococcus radiodurans* bacterium known for its impressive resistance to irradiation. *D. radiodurans* was discovered during experiments on food sterilisation technologies by using ionising radiation (Anderson *et al.*, 1956) and since then it was shown to be resistant also to UV radiation (from 100 to 295 nm), desiccation and mitomycin C, which induces oxidative damage not only into DNA but also into all cellular macromolecules (Slade and Radman, 2011). *D. radiodurans* is prone to DNA transformation (acquisition of foreign DNA), which makes it a popular laboratory model, and it is typically grown at $32°C$ with aeration where cell doubling time is around 100 minutes (He, 2009). Its genome contains 3.28 Mbp, placing it half way between the two peaks of genome size distribution in bacteria (see Fig. 1.3) and has $3,187$ open reading frames (White *et al.*, 1999) which roughly corresponds to the number of predicted protein coding genes. Although in-depth genome analysis did not reveal the cause of *D. radiodurans*' incredible resistance to stress, it has shown that this bacterium possesses a wide selection of genes involved in stress response, with topoisomerase IB being the most interesting case as this protein was till then known from eukaryotes exclusively. *D. radiodurans* knockout mutant with this gene deleted is more sensitive to UV (254 nm), but not to ionizing radiation, as compared with the wild type (Makarova *et al.*, 2001). As no single genetic agent (gene, operon or network of genes) could be directly linked with this remarkable stress resistance, it has been proposed that the *Deinococcus* genus gained this ability by acquiring a number of independent protective mechanisms and linking them together (Makarova and Daly, 2010; Koonin, 2011).

Is has been presented that *D. radiodurans* has significant redundancy in DNA repair-related enzymes and antioxidant enzymes. Also its genomic redundancy with 2 to 10 genome copies in each cell enables DNA repair based on sequence homology (Slade and Radman, 2011). It seems that *D. radiodurans* tackles various hazards to DNA and protein structure by possessing a variety of different repair systems rather than a single generic one.

This two cases point towards conclusion that it is futile to expect that elevated stress levels can be tackled by single solution. There are no magic genes. Adaptations to extreme conditions tend to be on a level of the whole genome. Studies on genetic networks and Boolean logical circuits seem to stand in line with this observation.

1.6 A new framework

Successful development of artificial life systems like *Avida* shows that ABM simulations are a very good tools for tackling questions in evolutionary biology. But ABMs weakness tends to lie in the way they try to represent a gene. Also, in recent years, with the development of genomics and molecular biology, the way the gene is seen in biology has become tangled, even to the point of becoming confusing. The task of modelling gene-centred evolution has become even more difficult.

Despite the rise in available computing power, researchers still face technical limitations which make realistic approach towards the gene and genome architecture and meeting quantitative requirements of population genetics at the same time rather unreachable. There is a constant need for a good, computationally and intellectually manageable abstraction of a gene.

The definition of species, one of the fundamental concepts in biology since Carl Linnaeus, varies depending on the discipline of biology: it means something slightly different in microbiology, in paleobiology and in ecology. The same has happened to the concept of a gene: it may vary in technical details depending on the subject of research. As a consequence, ways in which the gene is abstracted in a model will vary, depending on the purpose of building that particular model.

1.6.1 Goal of this framework

The goal of this modelling framework is to create a way of investigating genome size evolution in prokaryotes with focus on the issues of how the gene number is impacted by resource limitation and with consideration of the rising regulation burden as the number of metabolic genes gets higher. The framework should make consideration for the basic mechanism of mutations and be ready to later introduce a number of other ways to acquire new genes rather than sticking only to neofunctionalisation through gene duplication (e.g. horizontal gene transfer, viruses, plasmids).

Genes should meet the characteristics mentioned in section 1.4.2 but the natural selection should act on organism level. In other words, not only it has to be important for the selection process that a cell performs for a gene with the particular properties, but it also has to take into consideration what the other genes constituting a genotype are and how they act together.

1.6.2 Assumptions

As living cells are one of the most complex systems found in nature, a number of simplifying assumptions has been introduced, which should capture the essence of the investigated phenomena and serve well for the purpose of investigating genome size constraints.

(I) A single gene has a direct impact on the organism's fitness. As it is hard to distinguish, at the molecular level of real living systems, between the boundaries of contribution of a single gene to wining or losing the selection process, we have introduced this simplifying assumption and took a Dawkinsian perspective on a gene as a standalone unit (Dawkins, 1976, 1982). In real cells, most of the genes are tangled in often complex networks of interactions which make it difficult to estimate the contribution of a single gene to a cell's fitness (Pigliucci, 2010).

(II) Selection pressure does not recognise genes themselves but recognises their phenotypic outcome. Processes leading from defined sequence of purines and pyrimidines, which are the basic chemical blocks building the DNA molecule, to a phenotypic effect are complicated and not fully recognised. For the sake of simplicity we have decided to represent genes directly by their phenotypic outcome as it is the phenotype which becomes the subject of selective pressure.

(III) Cost of maintaining DNA calculated per gene is fixed. Environmental conditions do not alter the costs of maintaining (replicating, repairing, chromatin folding, etc.) of a single gene. We also assume that the amount of energy and nutrients in all the mentioned processes used per base pair is constant because metabolic pathways responsible for various DNA maintenance are evolutionarily conservative and, if they change at all, they change much slower than other metabolic pathways.

(IV) Genes generate genes. None of the genes works on its own, each has to be regulated and regulation acts via other genes. It has been shown that the number of transcription factors grows in scale with genome size (van Nimwegen, 2003) and it has been shown that these scaling laws are universally shared by all prokaryotes (Molina and van Nimwegen, 2009). An overhead of regulatory genes associated with genes responsible for direct response to the environmental conditions (n) proportional to n^2 has been introduced (Koonin, 2011, p. 97-99).

(V) Amount of non-coding DNA is negligible in prokaryotes. With just a few exceptions, the amount of coding DNA in prokaryotes exceeds 85% of the genome (Mira *et al.*, 2001; Lynch, 2006a), thus representation of non-coding DNA is being omitted in our model.

Using these assumptions a model of the evolving population of prokaryotic cells feeding on an abstract resource was constructed. The population occupies a well mixed environment and has to face one abiotic trait which changes in time. Genes are responsible for the efficiency of uptake of the resources under specific abiotic conditions. Properties and the number of genes in the cell can be altered by mutations.

Chapter 2

Construction of the model and testing the parameter space

2.1 Introduction

2.1.1 Aims of the model

From the current experimental and theoretical research arose number of hypotheses about the reasons of the genome size constrains among free-living prokaryotes. The most prominent are: small and fast growing cells need to have a small genome to facilitate rapid growth rates; there is a selection pressure among prokaryotes towards deletion of mutationally hazardous DNA; high demand of DNA material for phosphorus limits the growth rate and thus, by reducing the genome size, it allows to allocate P to metabolic pathways other than DNA maintenance and replication (e.g growth). It might be that all of these hypotheses are correct, depending on the environmental context. Genome size constrains are an effect of many selection pressures which can vary in their intensity, depending on many factors in the environment. This model investigates yet another theory. In this thesis it is hypothesised that the stability of environmental conditions has an influence on the size of genomes. Our aim is to explore plausible mechanisms through which evolution selects optimum genome size and to open the space for further investigations.

There is a strong competition among unicellular organism between individuals of similar species for resources, and thus evolutionary success depends on how effective a clonal lineage is in resource allocation. In environments with stable conditions in time less genes are required to have full responsiveness towards the demands of the outer environment. As an effect, the genes which do not bring adaptive benefits are, together with metabolic and regulatory pathways they code for, nothing more than an unnecessary cost and eventually they are lost. When environment is more turbulent in time, organisms need broader margins of response to its challenges, thus they require more genes which would code for all the necessary metabolic pathways. Eventually more turbulent environments will support genomes which are bigger and by that more expensive, but secure the organisms against a majority of possible changes.

This model investigates possible mechanisms of genome size constraints by the level of stability of the environment, using agent-based simulation of evolving population of virtual cells.

2.2 Methods

2.2.1 Mathematical design of the model

The model represents a resource-limited population of free living prokaryotic cells. Each cell is an independent agent which can uptake resources, has costs of living, can die because of starvation or random causes, reproduces and mutates at reproduction. Their performance depends on the quality of their genes and the size of their genotypes in the context of environmental conditions.

Environment. The state of environmental conditions is represented by one variable x which is a real number assigned from the interval $[-1, +1]$. This interval can be understood as one-dimensional representation of Hutchinson's niche space of abiotic conditions (Hutchinson, 1957), but the novelty is that its accessibility is limited to one value at a time. Examples of this type of environmental variable are pH, temperature, light intensity, moisture, concentration of ions, inorganic substances, toxins or other substances which have an impact on living conditions. These types of abiotic environment variables have just one particular value at a time, but the range and rate of their change in time defines the long time environmental conditions. The boundaries of x are limited as boundaries of most abiotic conditions can also be considered finite. For instance, pH values are limited by definition; environment temperature rarely exceeds certain values; even levels of concentration of different substances can be considered as limited, as the probability of reaching extraordinary high values is close to zero. The state of environmental conditions changes in time: $x(t)$ (the mode of change is discussed later).

Resources R is a real number representing the total amount of a single abstract ‘nutrient’ R in the ecosystem, both free and bound in the cells. Cell i is able to take a fraction of free resources R_{env} from the environment and allocate it as its internal resources $r_{cell,i}$. Total amount of the resources in the environment R is constant in any time step:

$$R = R_{env,t_l} + \sum_{i=0}^{P_{t_l}} r_{cell,i,t_l} = R_{env,t_k} + \sum_{i=0}^{P_{t_k}} r_{cell,i,t_k} = const \tag{2.1}$$

Where t_l and t_k are different time steps and P_{t_l} and P_{t_k} are population sizes in these time steps. Note that the nutrient circulation and x are different factors which do not interact with each other.

Genes and Genotype. Genes, represented directly by their phenotypic outcome, have a form of Gaussian functions over the space of environmental conditions defining a resource uptake rate for the given values of x:

$$u(x, c, \sigma, A) = A\, e^{\frac{-(x-c)^2}{2\sigma^2}} \tag{2.2}$$

Where x is the space of environmental conditions, c is the location of the maximum value of u in the space of environmental conditions ($c \in [-1, +1]$) and A is the maximum value of u ($A \in [0, 1]$), representing the maximum efficiency of resource uptake permitted by a given gene. σ is the dispersion controlling the width of the Gaussian curve, introduced to prevent ‘supergenes’ from emerging (ones which would have a near-to-maximum uptake rate in all of the environmental conditions space by having a large value of dispersion). σ is obtained from the following equation:

$$\sigma = \frac{\alpha}{A\sqrt{2\pi}} \tag{2.3}$$

Where α is a constant factor which scales the surface under the Gaussian curve. Figure 2.1 shows an example of a genotype. Note that setting this surface as s fixed entity adds a constrain for σ being dependent on A. Also, it has to be remembered that the Gaussian function produces real values for all real arguments ($x \in [-\infty, +\infty]; x \in \Re$) and that in this model we are interested only in the finite interval $x \in [-1, +1]$. This means that when a gene has a form of a flat and wide hump, then its left and right tails are outside the $[-1, +1]$ interval and its effective surface, considered within $[-1, +1]$, is smaller than the surface of a gene, which is higher thus narrower.

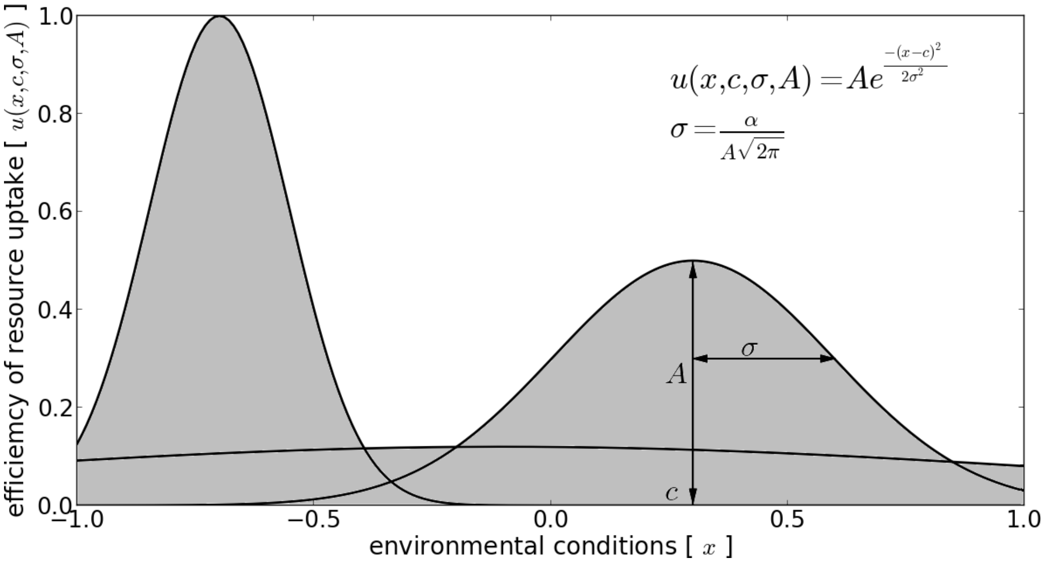

Figure 2.1: A genotype consisting of three genes in the environmental condition space. Surfaces under all the Gaussian curves are equal and scaled by factor α (eq. 2.3). The shaded area represents the fraction of the total environmental space occupied by the genotype which, as discussed later, is one of the focuses of selection pressure.

A single cell has to have at least one gene and up to any integer number of them. A genotype $\mathbf{G}_i$ of the cell i which has n_i genes is an array of the size $3 \times n_i$:

$$\mathbf{G}_i = \begin{bmatrix} A_{1,i} & \sigma_{1,i}(A) & c_{1,i} \\ A_{2,i} & \sigma_{2,i}(A) & c_{2,i} \\ \vdots & \vdots & \vdots \\ A_{n,i} & \sigma_{n,i}(A) & c_{n,i} \end{bmatrix} \tag{2.4}$$

The total size N_i of the genotype of the cell i is calculated as:

$$N_i = n_i + n_i^2 \tag{2.5}$$

Where n_i is the number of genes in the cell i. Factor 2 is based on ‘scaling laws‘ of different functional groups of genes (van Nimwegen, 2003; Koonin, 2011, pp. 97-99).

See also section 1.5.4.

If the cell i has more than one gene, the value of efficiency of resource uptake $U_i(x)$ for a given value of x is taken from a gene which has the highest value for that x:

$$U_i(x) = MAX\left[u(x, c_1, \sigma_1, A_1), u(x, c_2, \sigma_2, A_2), \ldots, u(x, c_n, \sigma_n, A_n)\right] \tag{2.6}$$

The number of genes n a cell has and the parameters A and c of each gene are subjects to mutations and selection with restriction on the value of σ, given by eq. 2.3. On the power of how the Gaussian functions was defined in this model each $u(x, c_i, \sigma_i, A_i)$ and, as a consequence, $U_i(x)$ has the maximum possible value of 1.

In living cells the phenotypic outcome of a gene is a result of the impact of countless factors of physical, chemical and biochemical nature as well as a result of interactions with other genes. Thus the Gaussian-shaped functions to represent the efficiency of resource uptake seems to be a natural choice as on the power of the central limit theorem, the sum of a sufficiently large number of iterates of independent random variables (with a well-defined expected value and well-defined variance) will be approximately normally distributed (Foryś, 2005).

Life cycle of a cell. Life history of a cell is dependent on how good it is in obtaining resources. If the cell i has the internal resource pool $r_{cell,i}$ smaller than the minimum allowed quota for a cell to live:

$$r_{cell,i} < r_{min} \tag{2.7}$$

then cell i is removed from the population and its resources are returned to the general pool R_{env}. Also at any time step, a cell can be killed by a random factor with probability δ.

Maternal cell i divides into two cells when its internal resource pool is larger than the reproduction threshold:

$$r_{cell,i} > r_{rep} \tag{2.8}$$

Which is:

$$r_{rep} > 2r_{min} \tag{2.9}$$

r_{rep} has to be slightly larger than twice the r_{min} to prevent two cells from being on the edge of death after the division, as each gets half of the maternal cell's internal resource pool.

During the division, an offspring cell can become a subject of mutations. The permitted mutations are:

- *deletion* – a gene gets removed from the genome with probability μ_{del}
- *duplication* – an extra copy of an existing gene in the genome gets added with probability μ_{dupl}
- *gene modification* – all the parameters A, σ and c of a given gene get changed by assigning their values once more from uniform distributions: $[0, 1]$ for A and $[-1, +1]$ for c. The values of σ is given by eq. 2.3. Each gene has the probability of being modified of μ_{mod}.

Resource circulation. The model does not consider energy circulation, but represents only the nutrient circulation between the population and the environment. A cell can obtain resources (virtual nutrients) from the environment to reproduce and also can give resources back to the environment as costs of genome maintenance and metabolic costs. The circulation has a perfect 100% efficiency and no resources are lost. Everything that comes back to the environment, either from dying cells or is returned in a form of costs, can be taken up by cells in the next iteration of the model.

The amount of resource taken by ith cell at a given time step t, called here the gain $Q_{i,t}$, depends on two factors. One is the value of the cell's genotype ability to benefit from the state of the environmental conditions at this time point $x(t)$ (eq. 2.6):

$$Q_{i,t} = \tau U_i\left(x(t)\right) \tag{2.10}$$

where τ is a fixed value of the maximum amount of resource units a cell can get in one time step, that is if U_i would be 1, which is an upper bound value of U_i. The other factor is the availability of resources in the environment. A cell is given at a time step t all that it has 'demanded' ($Q_{i,t}$) basing on its uptake efficiency $U_i(x(t))$ for given environmental conditions $x(t)$ and the value of τ set in the simulation. But the gain $Q_{i,t}$ has to be available, so:

$$Q_{i,t} \leqslant R_{env,t} \tag{2.11}$$

If the free resources that are available in the environment at the time t ($R_{env,t}$) are smaller than the gain a cell wants, then the cell gets nothing. The first few cells usually will get what they wanted, but others will be given nothing. To avoid a situation where a dozen cells always get fed and the rest gradually starves to death in each iteration of the model, all cells are aligned in a random queue and 'approach feeding' in a random order. This might seem obscure, but allows to avoid some complications. If the amount of available free resources were shared by all cells in proportion to the population size and respective individual uptake efficiency $U_i(x(t))$, which is a common alternative approach (e.g.: Williams and Lenton, 2007b), then a sharing algorithm would be needed and that is computationally expensive. The function $U_i(x(t))$ would need to be calculated at each time step for each single cell in the population and later available resources would need to be distributed to all the cells. But in the applied random queuing algorithm only a few dozen first cells have to have their $U_i(x(t))$ calculated and there is no need for second stage of computation (distributing the resource). Also, when an even-sharing system is used, then cells which divided at the same time and have similar genomes (in other words their uptake efficiency patterns are similar) will get synchronised in their growth and reproduction producing population-wide artefact in the form of cyclic rises and falls of the population size ('saw teeth' patterns in the population size versus time plots). To avoid that, there is a need for an extra parameter which introduces small differences between the cell's internal resource pool at a division that will break any further synchronisation. Random queuing does not need this parameter.

The chance that a cell will be chosen for feeding is inversely proportional to the population size $\sim 1/P(t)$, where $P(t)$ in the population size (the number of cells) at time step t. Note that due to closed and perfectly efficient resource circulation the amount of free resources $R_{env,t}$ at time t is also inversely proportional to $P(t)$. This property allow to have varying selective pressure on uptake efficiency and on the ability to withstand starvation depending on population size. If the population is small, then cells get fed frequently due to a lesser number of competitors and larger amount of available resources.

But what if we allow for a larger population by increasing the total amount of resources R? Will this sharpen the competition resulting from queuing? If the random death factor δ is fixed, then no, it will not. A larger population size P means also a larger pool of free resources $R_{env,t}$ as δ defines a percentage of dying cells in the total population (whose resources are returned to the free pool). The feeding queue will be longer and chances of getting to the first position will be smaller but, at the same time, number of fed cells will become larger.

Apart of random mortality δ, cells also return their resource to the environment in a form of costs of expressing genes. Living cells produce proteins and other substances (carbohydrates, lipids, wax, etc.) many of which are recycled internally and their elements are used again in the synthesis of new molecules. But some are lost (e.g. are excreted). Also respiration produces, apart from energy, molecules which cannot be reincorporated due to too high energy cost or inefficiency limitations of metabolic pathways, e.g. CO_2, in the case of aerobic respiration. In this model, costs of expressing genes $K_{i,t}$ are based on genotype size N_i (eq. 2.5) of a cell i:

$$K_{i,t} = -(\gamma N_i + \kappa) \tag{2.12}$$

where γ is a constant representing the amount of resource units a single gene costs and κ is a constant metabolic cost used to prevent even small-genome cells' starvation in the long run.

On the one hand, the cell is being charged with the cost $K_{i,t}$ in each time step and, on the other, it can be fed only every few time steps. The sequence of 'feeding' of the cells in the population is being randomly assigned in each time step separately. To maintain stable population, average individual gain of a cell has to be larger than their costs of living. This is fulfilled only when γ, κ and τ are set to be:

$$(\gamma \ll \tau) \wedge (\kappa \ll \tau) \tag{2.13}$$

When this assumption is met, the population is also provided with desynchronisation of development of maternal and offspring cells (no ‘saw teeth’ patterns in the population size). At the same time, it may seems odd that, apart from uptake efficiency $U_i(x(t))$, a second layer of parametrization is introduced. These parameters are useful in setting the model in a way which will secure smooth runs. Also they allow for further investigations of multispecies communities by varying τ and κ values.

Environmental variability. The change in time of the environment conditions $x(t)$ was generated as one-dimensional bounded random walk:

$$\begin{cases} x(t=0)=0 \\ x(t)=x(t-1)+RND[-T,T] \\ |x(t)| \leqslant 1 \end{cases} \tag{2.14}$$

Where t is time and T (called the turbulence level in this study) is a maximum length of a single step in this random walk with ranges $T \in [0, 0.5]$. This type of function has a number of advantages: (1) it is described by one parameter; (2) it has the memory of previous environmental conditions; (3) unlike in many smooth periodic functions which could be used to generate values within limited range, e.g. the sine function, its derivative does not depend on t that would complicate interpretation of the results. It is worth mentioning that T can be also understood as a measure of the amount of ‘memory’ in the system. The bigger the T is, the less memory the system has, as when T has the value of 2, the environmental conditions will change completely randomly within the range of the whole environmental space.

2.2.2 Algorithm

The model’s algorithm has been designed as a single threaded object-oriented program and was compiled and computed on a cluster computer.

First, a population of cells is generated. Each cell has a randomly assigned genotype, which has a number of genes randomly chosen from a given range. Each cell gets a pool of internal resources from the environment until the total sum of resources $\sum r_i$ in the cells reaches environmental limit R. Allocated resources cannot be higher than the r_{min}

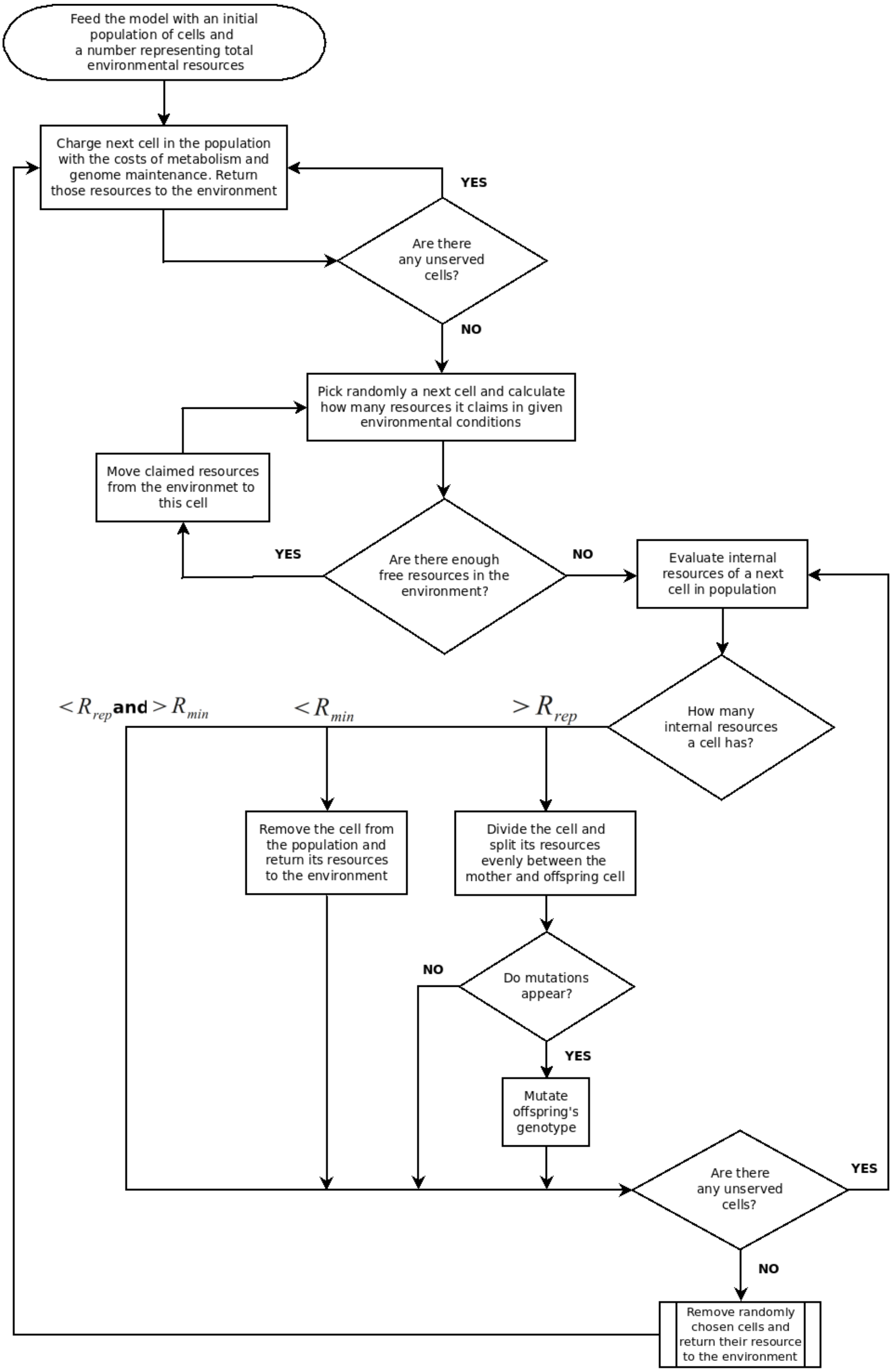
Feed the model with an initial population of cells and a number representing total environmental resources
Charge next cell in the population with the costs of metabolism and genome maintenance. Return those resources to the environment
YES
Are there any unserved cells?
NO
Pick randomly a next cell and calculate how many resources it claims in given environmental conditions
Move claimed resources from the environmet to this cell
YES
Are there enough free resources in the environment?
NO
Evaluate internal resources of a next cell in population
$<R_{rep}$ and $>R_{min}$
$<R_{min}$
$>R_{rep}$
How many internal resources a cell has?
Remove the cell from the population and return its resources to the environment
Divide the cell and split its resources evenly between the mother and offspring cell
NO
Do mutations appear?
YES
Mutate offspring's genotype
Are there any unserved cells?
YES
NO
Remove randomly chosen cells and return their resource to the environment

Figure 2.2: A general algorithm of the simulation run. r_{min} is the survival threshold of the cell's internal resources below which it dies. r_{rep} is the reproduction threshold above which the cell divides.

number or lower than the survival threshold r_{min}.

The population of cells constructed in this manner is ready to evolve. First, all the cells in population (organised in a vector) are being charged with all the cost of living (genotype maintenance and metabolic costs). All the ‘charged’ resources are returned to the environment. Later, a randomly sampled cell is chosen to be fed. If there are enough resources in the environment, a cell gets all the resources it had claimed and a next cell is sampled for ‘feeding’. If not, the ‘feeding’ procedure is terminated and the algorithm moves on to the next step.

In the next step, the program counts how many internal resources each cell has. If the cell has more than the reproduction threshold r_{rep}, then it is divided into two cells: a maternal cell and an offspring cell. Depending on random factors μ_{dupl}, μ_{del} and μ_{mod}, the offspring can mutate or not. If a cell has less resources than the survival threshold r_{min}, then it is killed and its resources are allocated into the environmental pool. If the cell's internal resources are somewhere between those two numbers, then it goes to the next step unaffected.

Next, a number of randomly chosen cells are killed with probability δ and their internal resources are moved back to the environmental pool. Then the loop closes and all the cells are charged with the cost of living again. A schematic representation of the simulation is presented in Figure 2.2.

Assignment of the next environmental condition $x(t)$ value (see eq. 2.15) is implemented as a three-step mechanism: in the first step, a value from the $[-T, T]$ interval (maximum turbulence span) is randomly selected, in the second step, it is added to the previous value $x(t-1)$ and the third step assesses if the new value $x(t)$ is within the $[-1, +1]$ interval. If it is not (when $|x(t)| > 1$), then it is ‘bounced from the edge’, which means:

$$\begin{cases} x\prime(t) = 1 - (x(t) - 1) = 2 - x(t) \text{ if } x(t) > 1 \\ x\prime(t) = -1 - (x(t) + 1) = -2 - x(t) \text{ if } x(t) < -1 \end{cases} \tag{2.15}$$

where $x\prime(t)$ is the value of the environmental condition after applying the boundaries correction (corrected $x(t)$).

2.2.3 Model's statistics

Different model runs were evaluated in regard with certain statistics:

Grand mean number of genes. For the purpose of easy comparison of the distribution of genome sizes in different simulations the grand mean and the grand standard deviation of the number of genes were introduced. They were calculated by averaging the numbers of genes in all the cells and all the time steps after the population's genome size distribution has settled down on a stable level.

Ratio of surface under the genotype to total environmental space. This statistic is introduced for individual cells. It is the proportion of the surface under the envelope of the genome of a cell (shaded area in Figure 2.1) to the total surface of all available environmental space (surface of the whole plot in Figure 2.1). It is an approximation of the trade-off between the need to have low costs of maintaining a genome and the need of having genome 'wide' enough to deal with environmental changes. The ratio of the genotype's surface to total environmental space has the values within the range of $[0, 1]$.

Grand mean ratio of surface under the genotype to total environmental space. It is calculated in similar manner as the grand mean number of genes and can be correlated with it as genes have fixed surface under their curves (note parameter α in Table 2.1), but as genes can overlap, this correlation is not accurate.

Cells' resource intake efficiency. Amount of resources a cell can take in one iteration is given by the combination of its genotype shape, the current value of environmental conditions x and fixed maximum amount of resource a cell can get (τ in Table 2.1). Mean resource intake is the averaged value of the population's resource real intake efficiency (the averaged gain $Q_{i,t}$ over all the cells). A histogram of the cells' intake efficiency should be a good estimation of how well the population it adapted. For a well-adapted population, the cells would have the best possible genes and their individual uptake $U_i(x(t))$ would approach 1.

The rate of evolution. To estimate the speed of evolution, a very simple measure was used. Each single mutation was accounted for (all three kinds were counted separately) and added to the cell's record. The average number of mutations was obtained by calculating the mean number of mutations for all distinctive clonal strains (not all the cells in the population at that time step). In that way, each clonal strain has a track of how many mutation events lead to its current genotype shape. At a selected time step all mutation records were harvested. To allow comparison between model runs of different length, the numbers of mutation were counted per 10^5 time steps. Such a big normalisation number is a result of the fact that mutations are rare events by their definition and normalisation per one time step would be statistically unjustified. Also, considering the accumulative number of mutations in a living clonal strain over a long time interval, instead of in each time step, allows to filter out most of the non-beneficial mutations which lead to the elimination of its bearers thus being only a stochastic noise in the adaptation process.

In most cases the number of mutations was summed up at the end of the model run. The rate of evolution is then the average number of mutations of any of the three kinds separately per clonal strain per 10^5 time steps.

Biodiversity. Biodiversity was measured using the Shannon index H, a very standard measure in ecology derived from the information theory (Cover and Thomas, 2006):

$$H_t = -\sum_{i=1}^{S_t} \frac{n_{i,t}}{N_t} \ln \frac{n_{i,t}}{N_t} \tag{2.16}$$

Where S_t is the total number of clonal strains in the system at the time t, N_t is the total number of cells and at the time t and $n_{i,t}$ is the number of cells belonging to ith clonal strain at the time t. A clonal strain is any group of cells which have identical genotypes and share a common ancestor. Cells which differ even in just one gene will be accounted as belonging to different clonal strains.

The Shannon index was calculated in each time step. For the purpose of comparing different model runs, the mean value and its standard deviation (SD) was computed by averaging the Shannon index after the grand mean number of genes has stabilised.

2.3 Analysis of parameter space

Time t is being measured in the model's iterations, where one iteration is represented as the biggest loop in Figure 2.2. The most difficult challenge was to choose parameter values in a way which would compromise between: (1) computing time of the model; (2) reasonable average life span of the cells as too short life spans would cause unrealistic, fast turn-over of clonal lineages; (3) reasonably large size of the population which is necessary for evolution to work on. Conditions (1) and (3) are in strong oppositions to each other.

2.3.1 Runtime values of parameters.

Parameters values used in the study are presented in Table 2.1. Reasons for this selection and their robustness analysis is discussed below.

Table 2.1: Values of the parameters of the model's runs a described in section 2.2.1. Note that probabilities are dimensionless by definition.

Symbol	Parameter description	Value	Units	Comments
α	Surface under the Gaussian curve ('gene width')	0.08	–	constrain preventing 'super-genes' from emerging
μ_{mod}	Probability of modification of a single gene	0.002	–	considered only during reproduction
μ_{del}	Probability of deleting a gene	0.002	–	considered only during reproduction
μ_{dupl}	Probability of duplicating a gene	0.002	–	considered only during reproduction
γ	Cost of maintenance of one gene	0.005	resource unit	considered in all iterations
κ	Metabolic costs of living	1.0	resource unit	considered in all iterations
τ	Maximum amount of resource units a cell can get	30.0	resource unit	considered when cell is fed, scales the $u(x, c, \sigma, A)$ value
r_{min}	Minimum allowed quota for a cell to live	300.0	resource unit	constant at all times, property of a cell
r_{rep}	Minimum quota for a cell to reproduce	640.0	resource unit	constant at all times, property of a cell
δ	Probability of random death	0.005	–	considered in all iterations
$\eta_{0,min}$	Minimum number of genes in a genome at initialisation	40	No. of genes	considered at start only
$\eta_{0,max}$	Minimum number of genes in a genome at initialisation	60	No. of genes	considered at start only
t_{max}	Number of iterations (time steps)	$2 \cdot 10^5$	time step	
R_{env}	Total amount of resources	$1.5 \cdot 10^6$	resource unit	constant at all times, property of the environment

Turbulence level T is subject to changes during the runtime and it is discussed further in this section.

At initialisation the model is started with a completely random population of cells where the number of genes in each single cell's genome is randomly chosen from discrete uniform distribution between $\eta_{0,min}$ and $\eta_{0,max}$. The values of all the genes are also generated randomly.

2.3.2 Probability of random death

The model is very sensitive to random death factor δ, a parameter which accumulates probabilities of all kinds of death caused by reasons other than starvation (e.g. predation, viral infection, intoxication, radiation, etc.). Too high random death factor will obviously bring the population to extinction. Absolute maximum of the random death factor is limited by the minimum time necessary for a significant number of cells to gather enough resources to reproduce. A large enough fraction of cells has to make it in the time given by the following equation:

$$t_{min} = \frac{r_{rep} - r_{min}}{\tau} \tag{2.17}$$

Which for the parameters given by Table 2.1 is equal approximately to $t_{min} = 11$ time steps. The age structure of the population should be given, assuming no other factors impacting the mortality, by the exponential distribution:

$$f(t; \delta) = \delta e^{-\delta t} \tag{2.18}$$

where t is the time. The expected cell's life span is equal to $1/\delta$ (the expected value of exponential distribution). For expected life span of $t_{min} = 11$ (achievable under maximum possible uptake by all cells) it should be $\delta_{max} \approx 0.09$ per time step, but model runs show that the borderline death factor is approximately 0.03 per time step (when turbulence level is $T = 0.01$). Above this value all the runs are terminated with extinction before reaching steady state. The higher the probability of random death is, the faster extinction occurs (Figure 2.3). The fact that the threshold value of δ is three times lower than the absolute maximum shows there are other factors than the random death factor which also cause mortality. The question is though what they are and what importance they have. Figure 2.4 shows that the age structure fits the exponential model very well when δ is fairly high (0.01 in Figure 2.4), and for the value of $\delta = 0.003$, the exponential curve still explains most of the age distribution and varies slightly from simulation results only for the cohorts of the young cells. But when $\delta = 0.001$ or is lower, then the exponential distribution fails completely to explain cells age distribution, vastly underestimating the number of young individuals in the system (0.001 and 0.0005 in Figure 2.4). There has

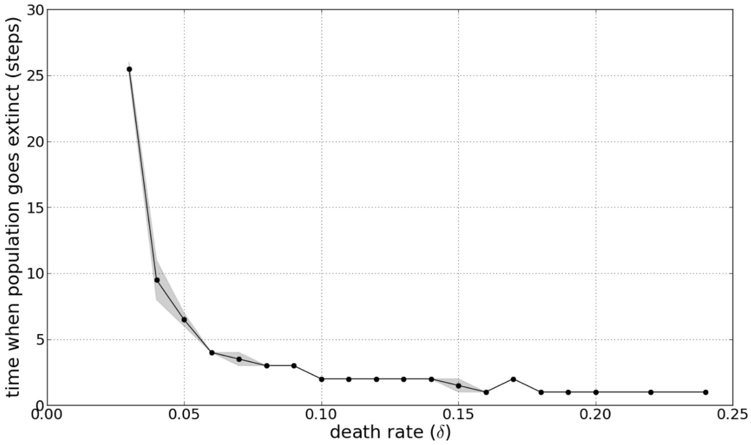

Figure 2.3: High random death rate factor triggers extinction events and the higher is its value, the earlier extinction happens. Circle (●) represent the mean value of 4 runs, the grey area is the SD. Model runs were initialised with parameter values given in Table 2.1 except death rate (δ), which varied between $\delta \in [0, 0.24]$. Turbulence level was set to $T = 0.01$.

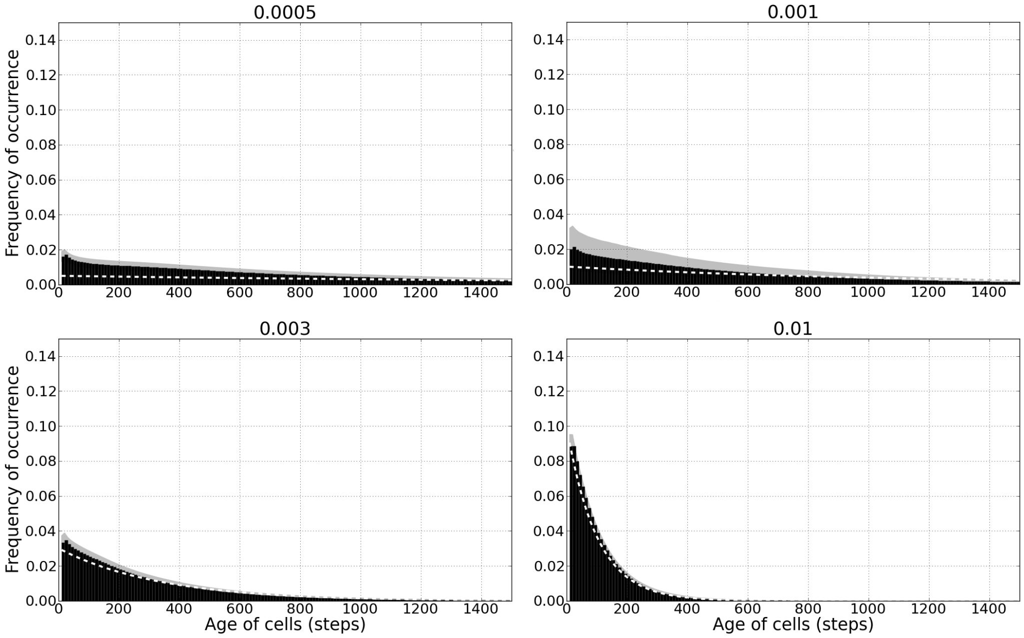

Figure 2.4: Age structure of the population in various random death regimes. Numbers above the panels show the value of the random death factor δ. Black bars indicate fraction of the population of a particular age (they were obtained by averaging all time steps in the run), grey area indicates the SD. White dashed line is given by the equation of the exponential distribution: $f(t;\delta) = 10 \cdot \delta e^{-\delta t}$, where t is the time and 10 is an extra normalising term representing the width of the histogram bins. All four runs have turbulence level $T = 0.25$. All other parameters were set as given in Table 2.1.

to be an another important source of mortality in this population, different from random death alone.

As the value of the random death parameter gets lower, the grand mean number of genes also decreases. Consequently, the ratio of surface under the genotype to total environmental space also reaches lower values as δ decreases. For the high value of the turbulence level $T = 0.25$ the grand mean number of genes, when there is no random death factor ($\delta = 0$), is 11.4 ± 0.9 (mean ± SD) and for a very low value of $\delta = 0.0005$ these are 11.5 ± 1.1. For moderate value of the random death factor $\delta = 0.003$, the grand mean number of genes is 13.9 ± 1.5 (mean ± SD), for quite high value of $\delta = 0.01$ it is 18.8 ± 1.9. For $\delta = 0.02$ the population did not survive till the end of the simulation when turbulence was set to $T = 0.25$ (Figure 2.5, panel B). At the same time, the ratio of the surface under the genotype to total environmental space for $\delta = 0$ is 0.38 ± 0.04 (mean ± SD) and for $\delta = 0.01$ is 0.55 ± 0.04 (Figure 2.5, panel A). Setting no random death factor gives the same mean number of genes and genomes envelope surface as low values of this parameter such as $\delta = 0.0005$ or $\delta = 0.001$ (Figure 2.5). What is noticeable here is that the span of the difference between the ratios of the surface of the genotype's envelope is lower than the span of difference between the grand mean numbers of genes. And that the difference is larger for systems which have evolved fewer genes (Figure 2.5, panel C), which are usually the systems with a very low turbulence level.

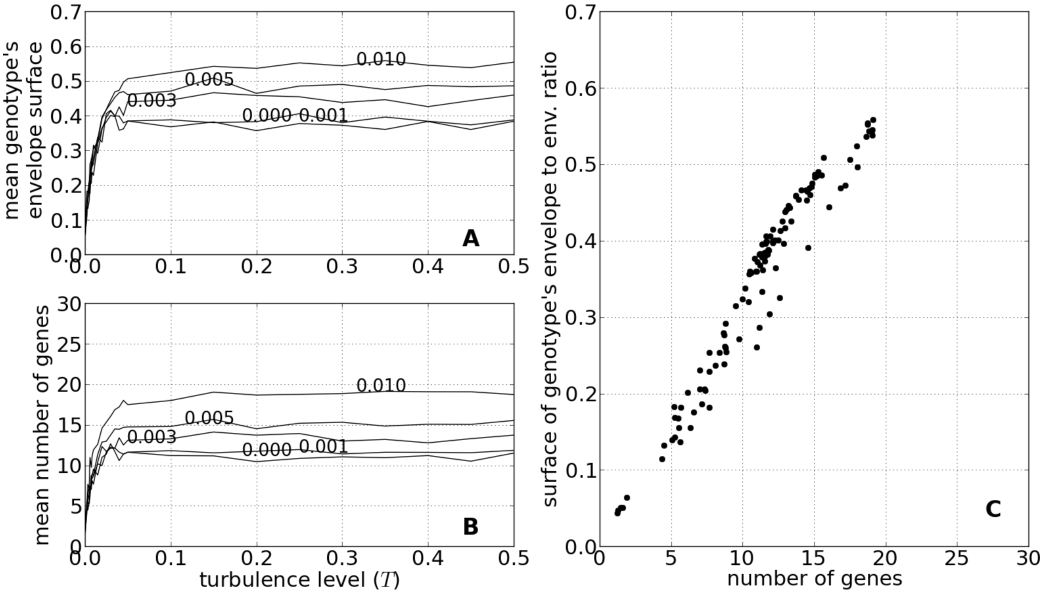

Figure 2.5: Model's sensitivity to changes in the values of the random death factor δ. Numbers above curves in panel A (ratio of surface under the genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are the respective values of the random death factor δ. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1 except the death factor δ.

Also biodiversity measured with the averaged Shannon index $\langle H \rangle$ in systems with lower random death factor decreases. The model is initialised with the population where each cell is different, which means that biodiversity has a maximum possible value of

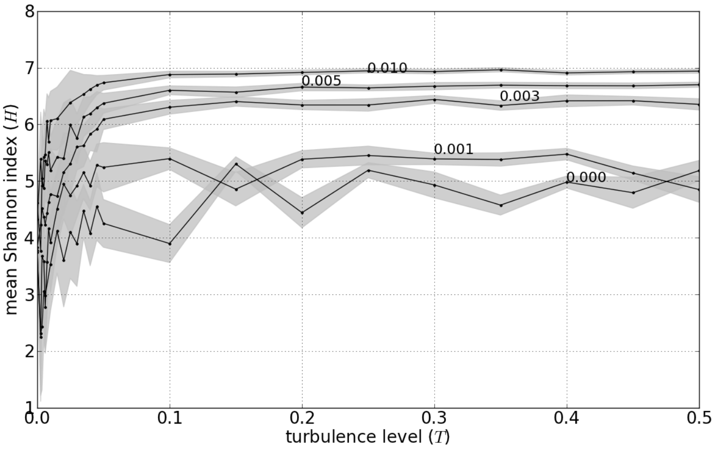

Figure 2.6: Dependence of biodiversity (measured with the Shannon index) on the random death factor. Numbers above curves are the respective values of the random death factor δ. Dots represent mean values of the Shannon index for the given δ and the turbulence level T calculated after the gene numbers stabilised. The grey areas are the SDs. The means and the SDs were calculated over the last 10^5 time steps. All model runs were set to parameter values as given in Table 2.1, except the death factor δ.

$\langle H\rangle = \ln\langle N\rangle$, where $\langle N\rangle$ is the total expected number of individuals in the population (Cover and Thomas, 2006). After the initial phase of biodiversity reduction, the Shannon index is stabilised at a lower,but still relatively elevated level. For $\delta = 0.001$ it is 5.45 ± 0.16 (mean ± SD) when $T = 0.25$. Same parametrisation, only for $\delta = 0.01$, gives $\langle H\rangle = 6.94 \pm 0.04$ (mean ± SD) as a result, meanwhile the maximum Shannon index value for both system is $\langle H\rangle \approx 8$. Shannon index for a system without the random death factor was not much different than for $\delta = 0.001$. Systems with a high random death factor also have less fluctuation of the Shannon index which is represented here by smaller standard deviation. More detailed data are presented in Figure 2.6.

A value of $\delta = 0.005$ was chosen for further analysis.

2.3.3 Surface under Gaussian representation of gene (‘width of gene’)

The parameter α represents the strength of the trade-off between the effectiveness of resource uptake and the ability to cover a wide spectrum of environmental conditions. The system responds to the decrease of α by increasing the number of genes and at the same time decreasing the fraction of the environmental space surface covered by the surface under the genotype. Nevertheless, the system is poorly sensitive to change in this parameter: changing its value from 0.04 to 0.10 decreases, for turbulence level $T = 0.25$, the grand average gene number from 18.9 ± 1.8 genes (mean ± SD) to only 14.7 ± 1.5 genes and increases the fraction of the environment covered by the genotype from 0.16 ± 0.02 (mean ± SD) to 0.27 ± 0.02 (Figure 2.7). This is only a 22% change in the gene number

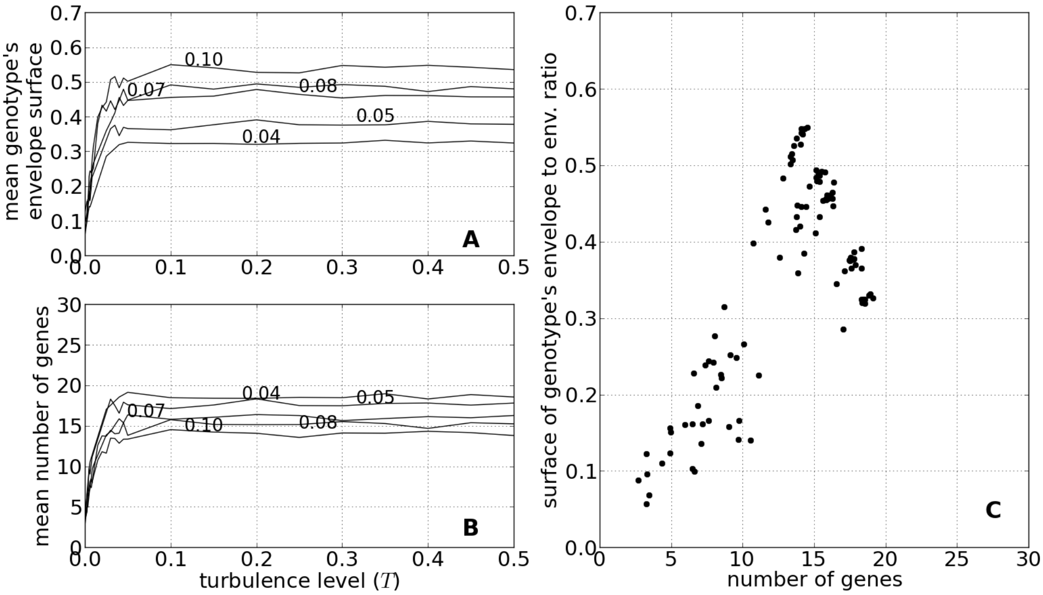

Figure 2.7: Model's sensitivity to change in the surface under Gaussian representation of a gene. Numbers above curves in panel A (ratio of surface under the genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are the respective values of gene width parameter α. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1 except the gene width α.

and 40% in the genome's envelope surface in response to more than doubling the α value. Meanwhile, setting $\alpha = 0.03$ caused all runs with T grater than 0.01 to become extinct.

Biodiversity of the population was affected very weakly. For the investigated parameter α and turbulence level of $T = 0.25$, the Shannon index has the highest value for $\alpha = 0.10$ being 7.18 ± 0.03 (mean ± SD) and the lowest for $\alpha = 0.08$ being 6.34 ± 0.09.

2.3.4 Cost of maintenance of individual gene

The number of genes in the system is sensitive to change in the cost of maintenance of one gene (γ). Increasing the cost of the gene's maintenance from 0.001 to 0.008 decreases the grand mean number of genes from 24.4 ± 2.5 (mean ± SD) to 12.5 ± 1.3, both for $T = 0.25$. That is nearly a 50% decrease in the number of genes. At the same time the mean fraction of the environment covered by the genotype (again considered for $T = 0.25$) changed from 0.58 ± 0.03 (mean ± SD) for $\gamma = 0.001$ to 0.43 ± 0.03 for $\gamma = 0.008$ which is only around a 25% change (Figure 2.8). When the cost of maintenance of one gene was set to 0.09, all the populations with T larger than 0.05 died out before

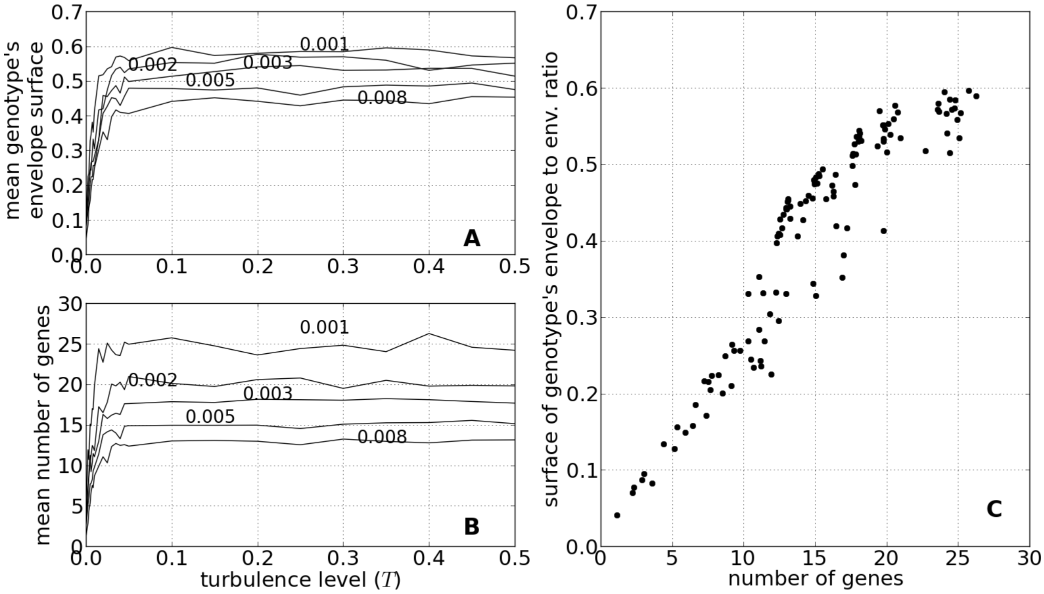

Figure 2.8: Model's sensitivity to change in the cost of maintenance of one gene. Numbers above curves in panel A (ratio of surface under the genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are the respective values of the cost of maintenance of one gene parameter γ. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1 except the cost of maintenance of one gene γ.

reaching the end of simulation.

Biodiversity, for the turbulence level of 0.25, varied from 6.34 ± 0.09 (mean ± SD) for $\gamma = 0.008$ to 7.18 ± 0.03 for $\gamma = 0.001$.

The value of 0.005 was selected for the cost of the maintenance of an individual gene providing strong, but not devastating trade-of between the number of genes and the need to accommodate the population in changing environmental conditions.

2.3.5 Mutation rates

In all the runs all probabilities of all three kinds of mutations, i.e. gene modifications μ_{mod} (change of the shape of the Gaussian curve), duplications μ_{dupl}, and deletions μ_{del} were set to the same value. Changes in the mutation rates parameters did not have a significant influence on the number of genes. For turbulence level $T = 0.25$, it is 14.6 ± 1.8 (mean ± SD) for $\mu = 0.004$ and 14.9 ± 1.3 for $\mu = 0.001$. Only for $\mu = 0.0005$ this number was higher rising to 16.2 ± 1.3 genes on average (Figure 2.9, panel B). The grand mean ratio of surface under the genotype to the total environmental space changed to an even smaller extent: for $T = 0.25$ it was 0.45 ± 0.04 (mean ± SD) when $\mu = 0.004$ and 0.47 ± 0.03

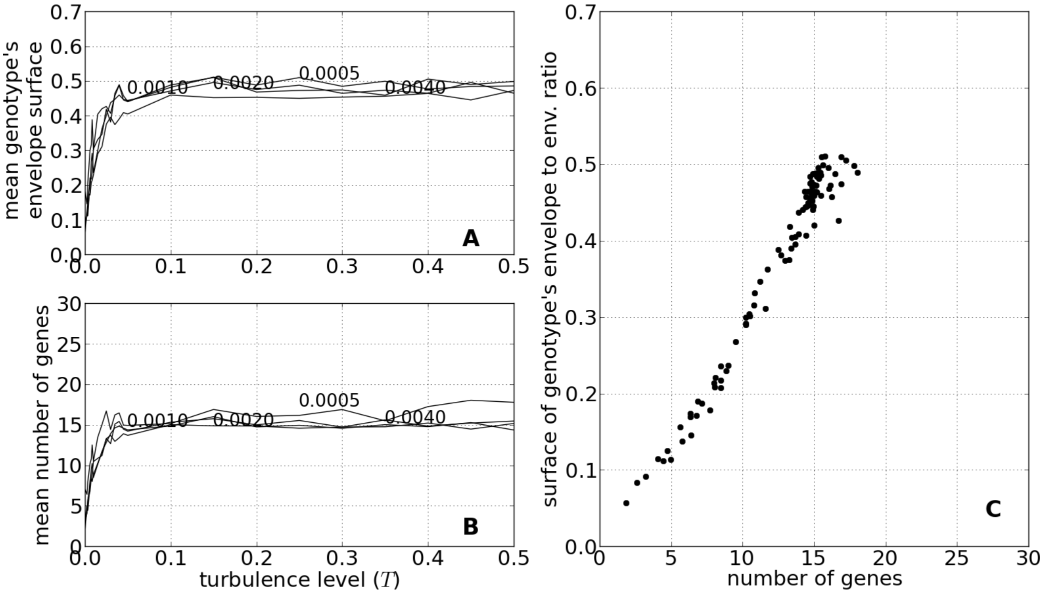

Figure 2.9: Model sensitivity to change in the probabilities of mutations. Panel A is the ratio of surface under the genotype to total environment surface as function of turbulence level T and panel B is the mean number of genes as function of turbulence level T (numbers above curves indicate the value of mutations rates μ_{mod}, μ_{dupl}, μ_{del}). In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1 except the three mutation probabilities μ_{mod}, μ_{dupl}, μ_{del}, which have all been set to the same value.

(mean ± SD) when $\mu = 0.0005$ (Figure 2.9, panel A).

But diversity measured by the Shannon index varied vastly between the runs with different values of the mutation probabilities. The lower the value of μ_{dupl}, μ_{del} and μ_{mod} is, the lower the diversity, e.g. for turbulence level $T = 0.25$ and for $\mu_{mod,dupl,del} = 0.0005$ the H is 4.23 ± 0.39 (mean ± SD calculated over the last 10^5 time steps) and for $\mu_{mod,dupl,del} = 0.004$ it is 7.23 ± 0.03 (Figure 2.10) with approximately 8.0 being the maximum value for a population of this size. It is worth noticing that for higher mutation probabilities the system also has less variable Shannon index values, as shown by the lower SD.

The main difference between runs with different values of probability of mutations is the time necessary for a population to settle down and reach a stable mean gene number for the given turbulence level T. The smaller the values of mutations' probabilities were, the later a steady plateau for the average gene number was achieved.

A value of $\mu_{mod,dupl,del} = 0.002$ was chosen for further analysis.

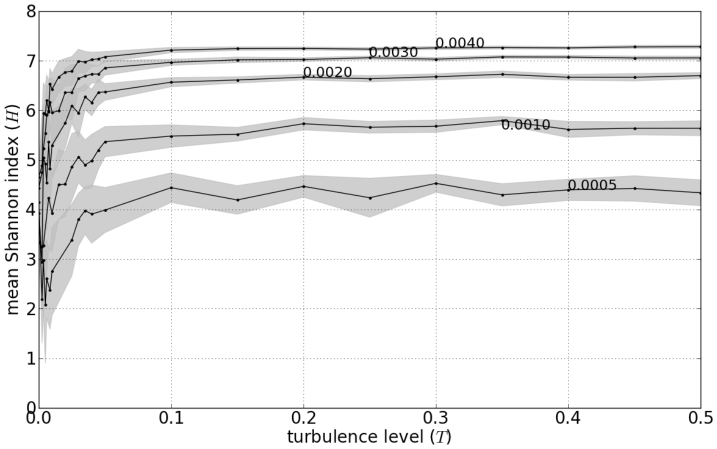

Figure 2.10: Dependence of biodiversity as measured with the Shannon index on the mutation probabilities values. Numbers above curves are the respective values of the mutation probabilities μ_{dupl}, μ_{del} and μ_{mod} (all three are equal). Dots represent mean values of the Shannon index H for given $\mu_{mod,dupl,del}$ and the turbulence level T. The H is calculated after the gene numbers stabilised (after 10^5 time step). The grey areas are the SDs. All model runs were set to parameter values as given in Table 2.1 except the mutation probabilities.

2.3.6 Amount of free resources

The amount of free resources in the system is the key factor limiting the number of cells in the population. The more resources there are in the environment, the larger the population that can be sustained by the ecosystem, e.g. an ecosystem containing $R_{env} = 1.5 \cdot 10^6$ of total resources can sustain approximately $3,200$ cells, when $R_{env} = 10^7$ population has approximately $22,000$ cells. Populations of different sizes do not evolve different number of genes. The grand mean number of genes and the grand mean ratio of surface under the genotype to total environmental space are statistically indistinguishable between systems of different resource levels (Figure 2.11).

2.4 Rationale behind parameter values

No modelling work can fully represent natural phenomena, and this applies even to the mathematical principles of physics. A simple formula for pendulum's period $T = 2\pi\sqrt{I/Mgh}$ can be used to determine the value of a local gravitational acceleration g with accuracy of up to four significant figures (Nelson and Olsson, 1986), with the model being based on rather crude assumptions e.g. that only small angular displacements are applied, that there is no air friction, and that the mass of the pendulum is concentrated in a dimensionless point at the pendulum's tip (Morgan and Morrison, 1999a). In modelling of more complex

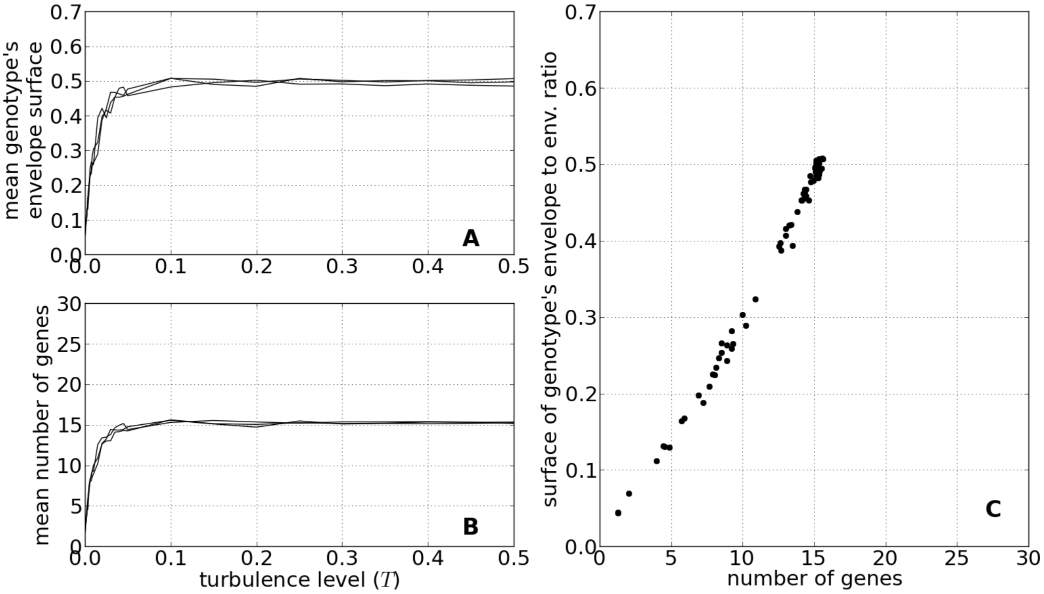

Figure 2.11: Model sensitivity to change in the total amount of resources. The total amount of resources R_{env} were: $1.5 \cdot 10^6$; $4.5 \cdot 10^6$; $7.5 \cdot 10^6$; 10^7. Panel A is the ratio of surface under the genotype to total environment surface as function of turbulence level T and panel B is the mean number of genes as function of turbulence level T. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1 except the total amount of resources.

systems the assumptions place the model even further from the reality and it often happens that the researcher is interested in investigating unknowns of the reality by observing how modelling results deviate from the observations. This is the spirit behind this work.

2.4.1 Parameter selection

This model tackles questions regarding evolution of the genome architecture, which is currently subject to an open scientific debate, with large body of facts still unknown and new discoveries coming practically daily. Similarly to many other theoretical research studies made in this field, the question is of a rather qualitative than quantitative nature thus parameters are selected to observe the desired behaviour of the model, rather than to achieve numerical values comparable with *in vivo* and *in vitro* experiments. Due to the facts mentioned earlier, this is a common approach in this field of research (e.g. Wilke and Adami, 2002; Kashtan and Alon, 2005; Williams and Lenton, 2007a; Crombach and Hogeweg, 2009; Kashtan *et al.*, 2009; Torres-Sosa *et al.*, 2012).

Mutations: Frequencies of mutation have been a subject of interest in molecular studies of evolution for a long time now. Information on how often they occur, if there are any preferred regions of the genome they occur in and what kind of change they introduce is relevant for estimating the speed and significance of evolutionary changes. The concept that mutation rates can change and can be a subject of selection was proposed even before the discovery of the double helix structure of nucleic acids (Sturtevant, 1937). Later Motoo Kimura proposed that there had to be a trade-off between the benefits of reduced mutation rate and the physiological cost of improving fidelity of DNA transcription, and that at some point this cost outweighs the gain in fitness (Kimura, 1967). In the long-term evolution experiment with *E. coli* based on circa $4 \cdot 10^4$ generations (which probably makes it the most accurate experiment to date), the synonymous substitutions rate was estimated to be on average $8.9 \cdot 10^{-11}$ per base-pair per generation, which corresponds to the total genomic rate of $4.1 \cdot 10^{-4}$ per generation, given that the ancestral genome was $4.6 \cdot 10^6$ bp in size (Wielgoss *et al.*, 2011). Synonymous substitutions are neutral mutations as they are replacements of a single nucleotide in protein-coding genes, which do not alter the amino acid sequence of the resulting protein due to redundancy of the genetic code. Thus they do not have an impact on fitness and survival of the mutated cell, whereas non-synonymous mutations were found to have such impact, making them not suitable for estimation of the sheer rate of point mutations. A beneficial mutation will be over-represented in next generations as a result of the selection pressure (Wielgoss *et al.*, 2011). Tracking mutations in the context of the selection process is quite difficult task, even when bacteria are kept in controlled constant laboratory culture. It has been demonstrated that the rise of fitness of population decelerates over time in a stable laboratory environment, whereas the number of mutations grows linearly. Also a constant environment can produce a mutator strain which mutates at a rate 70-fold higher than the ancestral strain (Barrick *et al.*, 2009).

Different organisms have different mutations rates, with RNA viruses being the fastest mutating ones.In the case of DNA-based microbes, mutation rates per base-pair per replication vary over nearly four orders of magnitude: from $7.2 \cdot 10^{-7}$ for Bacteriophage M13 to $7.2 \cdot 10^{-11}$ per base pair per replication for the bread mold *Neurospora crassa* (Drake *et al.*, 1998). But when considering the length of the genome, then mutation rates of DNA microbes are very similar: from $4.6 \cdot 10^{-3}$ for M13, $3.0 \cdot 10^{-3}$ for *N. crassa* and the lowest of $2.5 \cdot 10^{-3}$ for *E. coli* per genome per replication (Drake, 1991; Drake *et al.*, 1998). Note that the estimation for *E. coli* is performed here with a different method than by Wielgoss *et al.* (2011).

Single nucleotide substitutions are the easiest type of mutations to measure and estimate. In this study, we are interested in a higher level of changes in genomes: gene deletions, gene duplications and mutations affecting the gene's performance. Neutral changes like synonymous substitutions are not being represented well in the model. Meanwhile, different types of genome rearrangements which nevertheless have a similar outcome will be considered as the same type of mutation, e.g. deletion in this study is understood as simply a loss of gene expression, thus physical loss of a protein-coding fragment of a chromosome and a single nucleotide substitution in the regulatory region resulting in a loss of function of a gene are considered equivalent. A change of gene performance can be a result of a change in the protein sequence which is coded by the gene, or by altering the regulatory pathways that the gene is a part of. Again, this model does not differentiate between these two situations.

As it is difficult to estimate the mutation rate in this regard, let us try to estimate the chance that a cell will not mutate at the reproduction. In the model there are three types of mutations, each having the probability of occurrence in a single gene at the replication given by μ_{del}, μ_{dupl} and μ_{mod}. When there are N genes in the genome, then the probability for a cell of not mutating p_N is:

$$p_N = (1 - \mu_{del})^N (1 - \mu_{dupl})^N (1 - \mu_{mod})^N \tag{2.19}$$

And as all types of mutations have the same probability of occurrence $\mu = 0.002$ (see Table 2.1) it simplifies to:

$$p_N = (1 - \mu)^{3N} \tag{2.20}$$

For a dividing cell with 15 genes in its genome (see Figures 2.7, 2.8, 2.9) the probability of not mutating is $p_{15} \approx 0.914$, and for 10 genes it is $p_{10} \approx 0.942$. For experimental data for *E. coli* this figure is 0.996 (Wielgoss *et al.*, 2011) and 0.998 (Drake, 1999), respectively. One-fold difference seems to be not that large of a number, considering that the experimental data do not take into account mutations other than single nucleotide polymorphism and are based on studying populations of a size of circa 10^6 cells (Lenski *et al.*, 1991), with genomes consisting of approximately $4.7 \cdot 10^3$ protein-coding genes (Welch *et al.*, 2002).

Number of genes: Genome sizes in the prokaryotic kingdom span from 180 kbp to circa 13 Mbp (see section 1.5.4). The well-investigated model bacterium of *Escherichia coli* has between $4.3 \cdot 10^3$ to $5.1 \cdot 10^3$ protein coding genes, depending on the genetic strain considered (Welch *et al.*, 2002). In the proposed model genome size varies from 1 to the maximum of 50 protein coding genes (see Table 2.1). This two-fold difference is dictated by the limitation of computing performance. As said before, a typical run has around $3.2 \cdot 10^3$ cells and if each has the maximum of 50 genes, then there is a total of $1.6 \cdot 10^5$ genes in the system. If the number from *E. coli* would be used, say $4.7 \cdot 10^3$ genes, then the total number of individual genes rises two-fold, to $1.5 \cdot 10^7$. Assuming all the other parameters being the same, this two-fold difference in the gene number will translate also into two-fold longer runtime. If on a given machine the model runs for one hour, then with a realistic number of genes, based on *E. coli* it would run for slightly over 4 days. Also, considering the introduced simplifications, especially simple single dimension representation of the niche (see Figure 2.1), larger genomes do not guarantee any new emergent properties other than those already discussed in this thesis. Increasing genome size 100 times to reach the values observed in *in vitro* research, is an overkill.

Cell's resource circulation restrictions: There are five parameters influencing resource circulation within the cell: cost of maintenance of one gene γ; metabolic costs of living κ; maximum amount of resource units a cell can acquire in one iteration τ; minimum allowed quota for a cell to live r_{min}; minimum quota for a cell to reproduce r_{rep}. The first two (γ, κ) are restrictions on genome size introduced by generating costs.The cost of maintenance of one gene is variable in nature depending on organism type, particular gene, e.g. genes which are more frequently expressed (with more protein molecules in the cytoplasm) are under more rigorous selection pressure regarding protein's synthesis and folding than rarely expressed genes (Koonin, 2011). Nevertheless, some attempts have been made to estimate these costs, averaged on the genome scale for prokaryotes and for eukaryotes (Lane and Martin, 2010): see sections 1.3.1.

Parameter γ is being investigated in detail in section 2.3.4 as it has an impact on the gene number. Metabolic costs of living are dependent in prokaryotes on the species or even the genetic strain of a species and are one of the main traits to undergo optimisation in the evolutionary process. Furthermore, organisms implement elaborate regulatory systems to optimize their metabolic costs, depending on the state of environmental conditions (for review see Alon, 2007). In this model, only one species and one niche is being considered, so for the sake of simplicity parameter κ was set as fixed.

The last three parameters (τ, r_{min}, r_{rep}) are linked with each other. The minimum allowed quota of resources for a cell to live r_{min} is linked to the minimum biomass of a single cell and if it will get too small, then the cell does not have enough resources to maintain its metabolism. It is not set to 0 as in living cells there is always some biomass fixed into the physical structure of the cell (e.g. membranes, cell wall, chromosomes) which cannot be relocated to other metabolic paths without destruction of the cell. Cell size (and thus biomass) varies widely in nature and have been link to growth rate, especially in plants (Hessen *et al.*, 2010) thus affecting the number of genes via the cost of expression. But it was not the effect studied in this thesis and its significance in prokaryotes is debatable in the light of population genetic effects and purifying selection (Lynch, 2006a), so r_{min} was also set fixed to make the system easier to analyse. The minimum quota for a cell to reproduce r_{rep} is slightly over double the value of r_{min}, so when the cell divides, the resulting two offspring cells have a safety margin above the threshold of dying. Genes are represented as Gaussian functions of resource uptake in given environmental conditions x and a Gaussian function has a maximum value by definition. The maximum amount of resource units a cell can get in one iteration τ is a scaling factor deciding, along with the competition between cells, on how fast cells will grow.

The values of the pair r_{min} and r_{rep} plus the value of τ affect the cells' lifespan and duplication time setting the time frames of the model. A cell has to live long enough to experience a certain amount of environmental perturbation as adaptation to perturbation is the key question behind this study but, on the other hand, generation turnover has to be relatively fast to observe the outcome of evolutionary processes on a population scale in a relatively short time. The shorter, the better as that allows to harvest a large body of results to give the conclusions a statistical significance.

Population size: Population size is directly driven by the total amount of resources that are allocated to the whole environment R_{env}. The idea was that when population size decreases, then some of the resources previously bound in the cells are released freely to the environment, increasing the amount of easily accessible resources and thus relaxing the pressure on uptake efficiency to some extent. It was decided that the model will be analysed for $R_{env} = 1.5 \cdot 10^6$ (around $3.2 \cdot 10^3$ cells in the population). On the one hand, the population size of simulated microbes should be made very big. In the long-term evolutionary experiment with *E. coli* each flask contains approximately $2 \cdot 10^6$ cells (Lenski *et al.*, 1991), i.e. three orders of magnitude more than in these simulations. But this modification would extend the time necessary to compute $2.0 \cdot 10^4$ iterations from around one hour to nearly 42 days giving no significant gain in the quality of results except maybe better resistance to total extinction of the population (see section 2.3.6).

The set value allows for a reasonable compromise between the number of cells allowing for a sustainable population with size big enough for robust statistical analysis and the computing time of the model allowing to run a large number of its copies (on a commercial off-the-shelf PC each run takes approximately an hour). For the purpose of analysing the outcome and generating plots in this and the next chapter, the model was run nearly 2000 times (with varying parameters) on a high performance computing cluster.

Random death: The death rate of bacteria in various aquatic environments was estimated experimentally by measuring the degradation rate of DNA liberated by dead cells and came to be in the range of 0.010 to 0.030 per hour, which stands in an agreement with the growth rates measured in the same environments (Servais *et al.*, 1985). Dynamics of microbial populations in laboratory experiments with colonisation of marine snow showed similar time scales. The encounter volume rates were estimated at about 0.01 and 0.1 cm^3 h^{-1} for bacteria and flagellates, respectively. The model has the random death rate set to $\delta = 0.005$ per time step (simulation iteration) and a value around 0.03 turned out to be always driving populations to extinction (see section 2.3.2). These are no time units to compare, so let us compare growth rates. In the mentioned study of bacterial mortality, growth rates were slightly below mortality rates (Servais *et al.*, 1985, Table 1). In this modelling study, the time between divisions, hence doubling time, was 176 time steps (see section 3.3.3). This gives the growth rate of 0.004 per time step. In the marine snow colonisation experiment, bacterial growth rates were between 0.1 and 1.2 day^{-1} (Kiørboe *et al.*, 2003) which translates into 0.004 and 0.05 h^{-1}. The *in vivo* and *in vitro* experiments happen in a continuous time, whereas any computer modelling has to use discrete time frame due to the nature of the computer construction limitations making modelling tricky to compare with the data. Let us use an approach similar to that shown when discussing the comparison of mutation rates and compare the probabilities of a cell to survive till the age of reproduction. We can use transformation of equation for the growth rate of bacteria $\mu = \ln 2/t_d$, where μ is the growth rate and t_d is the doubling time of bacterial species to estimate the time needed between reproductions. Both mortality rate and growth rate obtained experimentally are available in literature (Servais *et al.*, 1985, Table 1) for a number of aquatic environments found in Europe. Probability of survival till reproduction p_s is given by the following equation:

$$p_s = (1 - \delta)^{t_d} \tag{2.21}$$

Where δ is the probability of a cell dying in one time interval and t_d is the number of time intervals (number of time steps in the case of the model and number of hours in the case of experimental data) needed for the average cell to reproduce itself (doubling time). For the model $p_s \approx 0.41$ and for the experimental data this value varies widely from 0.12 to 0.72, with median around 0.52, while the first and third quantile are 0.40 and 0.63, respectively. The value of the random death parameter used in the model $\delta = 0.005$ seems to reproduce the odds of reaching the reproduction similar to the lower values observed in aquatic environments of the temperate climate.

2.4.2 Overall effects of selected parameters values

A number of parameters are linked to time scales of the model, namely all resource circulation restrictions and the random death factor. As said above, their values should be selected rather to reproduce realistic qualitative behaviour of the modelled system than to strictly follow true quantitative properties of *in vivo* systems. It has been proposed that for modelling of evolutionary processes the time step should be smaller than the frequency of mutations and division and the overall number of time steps should be sufficient to allow for natural selection and fixation of mutants (Mozhayskiy and Tagkopoulos, 2013).

In this model, the selection of respective parameters allows for around 90% of cells to reproduce without mutations, thus making fixation of new characteristics probable. Simulation time is also long enough to observe transitions in the evolutionary process (see 3.2.2). Long runtime of the simulation ($2 \cdot 10^5$ iterations) reduces the effect of stochastic noise in the natural selection process and provides the stability of of the achieved equilibria. It has also been advised that in the case of models simulating gene expression and molecular interactions, the time step should be at least of 10^{-3} fraction of the generation time (Mozhayskiy and Tagkopoulos, 2013). This model does include gene expression, but completely omits any molecular properties of the cell, thus generation time of 180 time steps (fraction slightly over 10^{-2}) seems to be a sufficient value.

Running the model with different seeds of the pseudo-random number generator showed similar and predictable simulation results which all fall within each other's variation limits. This is achieved thanks to sufficient population size and a long time set for the runs to settle down.

Chapter 3

Evolution of genome size in free-living cells under variable abiotic environments

3.1 Introduction

Some of the results of testing the parameter space showed interesting properties of the model and they will be discussed further in this chapter along with new simulations. Problems of interest are as follows: how the gene number depends on the turbulence level of the environment; how robust this dependency is; how resource uptake is shaped by the different turbulence levels; and will populations under high turbulence levels also show elevated rates of evolution and higher diversity?

3.2 Results

3.2.1 Existence of optimum gene number for a given turbulence level

The model was initialised with $R_{env} = 1.5 \cdot 10^6$ of total resources in the environment that supports a population with approximately $3,200$ individuals. The model produces populations with the least genes when the environmental conditions are stable (when T is equal to 0) and it is on average just over two genes. This is one gene more than a cell can have in this model (less than one gene responsible for resource uptake is not permitted).

Though expected, reaching one gene might be difficult due the to stochastic nature of the simulation, with finite runtime and finite population size. Rapid growth of the genome size is observed as the turbulence level grows until it reaches the value around $T = 0.05$. Above this value the growth declines and the number of genes reaches a stable value between 15 and 16 genes per genome after the number of genes has stabilised. This value is the same regardless of the value used to initialize the model's pseudo-random number generator (Figure 3.1). Also changing the number of genes per genome at initialisation of the simulation does not alter the size of the genomes at the end of the simulation. Two series of simulations, each of length $t_{max} = 2.0 \cdot 10^5$, were initialised with the number of genes randomly chosen from a continuous uniform distribution ranging from $\eta_{0,min} = 40$ to $\eta_{0,max} = 60$ (simulations A) and from $\eta_{0,min} = 4$ to $\eta_{0,max} = 10$ (simulations B). In both cases the population's genetic structure changes slowly and reaches a stable state with a narrow range of different sizes of genotypes, which are similar regardless of the initial span of genome sizes. An example of two full runs with $T = 0.25$ is presented in Figure 3.2 and final results of runs for all of T in Figure 3.3.

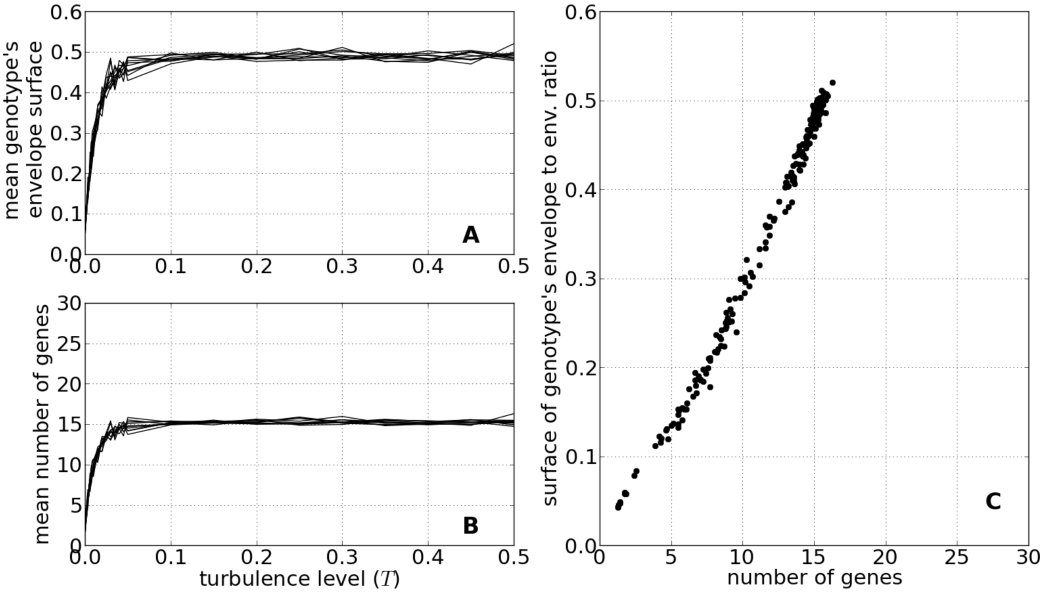

Figure 3.1: Ten simulation runs set to standard parameter values show consistent results regarding the gene number and coverage of environmental space by the genomes. Panel A is the ratio of surface under the genotype to total environment surface as the function of the turbulence level T and panel B is the mean number of genes as function of the turbulence level T. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1. Each series of runs was initialised with a different value of the seed of the pseudo-random number generator.

As expected, in a stable environment ($T = 0$) the population's genome size is the

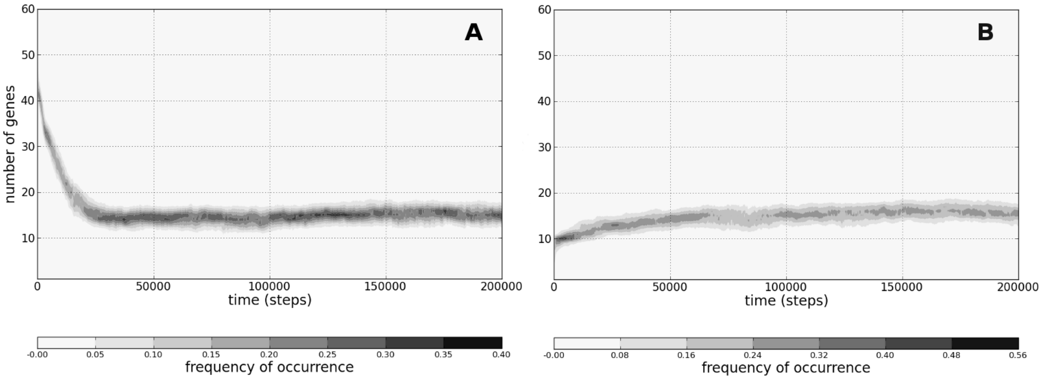

Figure 3.2: Optimisation of the genome sizes does not depend on the genome sizes at initialisation. Two example simulations initialised with different permitted genome sizes. Genome ranges at initialisation were $\eta_0 \in [40, 60]$ (panel A) and $\eta_0 \in [4, 10]$ (panel B). Both simulations had the turbulence level of $T = 0.25$ (for all T values, see Figure 3.3),other parameters were set as in Table 2.1. Both systems evolve a similar grand mean number of genes and similar SD. Note that the colour maps in both panels are scaled differently and that the change of the mean size of the genome is initially so fast that the resolution of the plots is too small to present it.

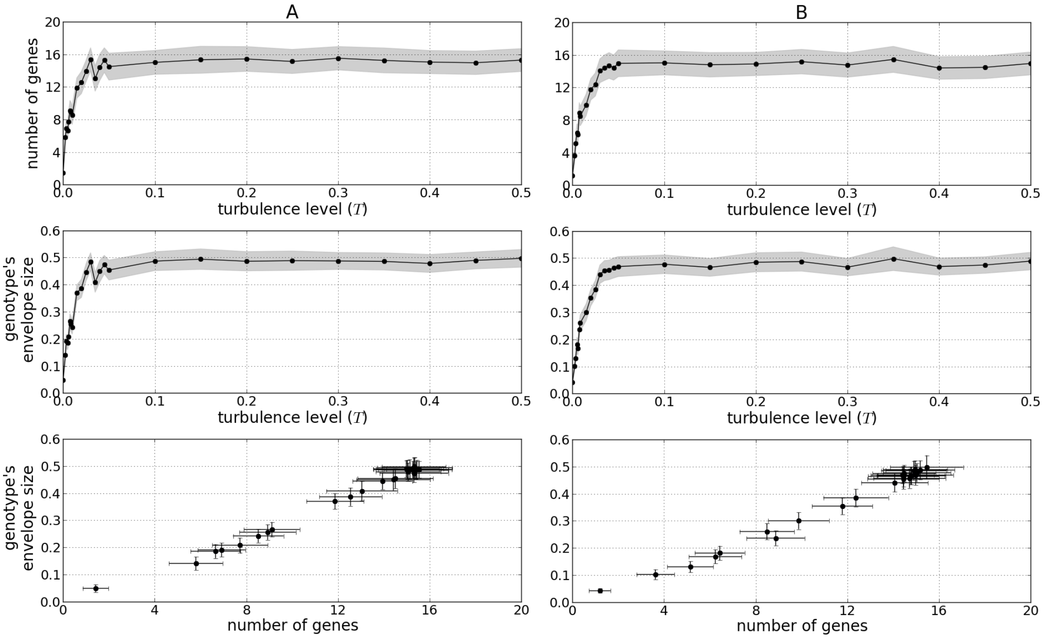

Figure 3.3: Dependence of genome sizes and shapes on the turbulence level of environmental conditions and genome size at initialisation. Genome size ranges at initialisation were $\eta_0 \in [40, 60]$ genes (panels A) and $\eta_0 \in [4, 10]$ genes (panels B). Turbulence level was in the range of $T \in [0, 0.5]$. Black dots indicate the grand mean of gene number (upper panels) or the grand mean of proportion of surface under the genotype envelope to total surface of environmental space (middle panels). All averaging was made for time steps greater than $t \geqslant 10^4$ to ensure that a population in a stable state is being investigated. The grey area is the grand standard deviation calculated in a similar manner. On the lower panels, the bars indicate the same grand standard deviations of the genome size and the envelope surface, respectively. Only the runs which did not suffer from a total extinction were taken into account.

smallest of all simulations (around two genes, when the minimum permitted number is one). As the turbulence level increases, the mean number of genes rises rapidly, reaching a stable number just under 16 genes per genome for $\eta \in [4, 10]$ and around 18 for $\eta \in [40, 60]$ for all systems with turbulence above 0.1. A system with a lower gene number per genome at initialisation reaches equilibrium also at a lower gene number, but both sets of systems are within each other's standard deviation (SD) values. Also the fraction of total environment covered by the genotype's envelope (surface under the function, given by eq. 2.6) reaches a stable average number around 0.25 for $\eta \in [40, 60]$ and around 0.24 for $\eta \in [4, 10]$ with both those values being within each other's standard deviations (Figure 3.3).

The system is also able to adapt to dynamic change of turbulence in the environment. The run was divided into three equal parts: in the first part the turbulence level was $T = 0.005$, during the second $T = 0.2$ and during the third, again $T = 0.005$. The population was able to smoothly evolve between gene numbers optimum for those two regimes (Figure 3.4).

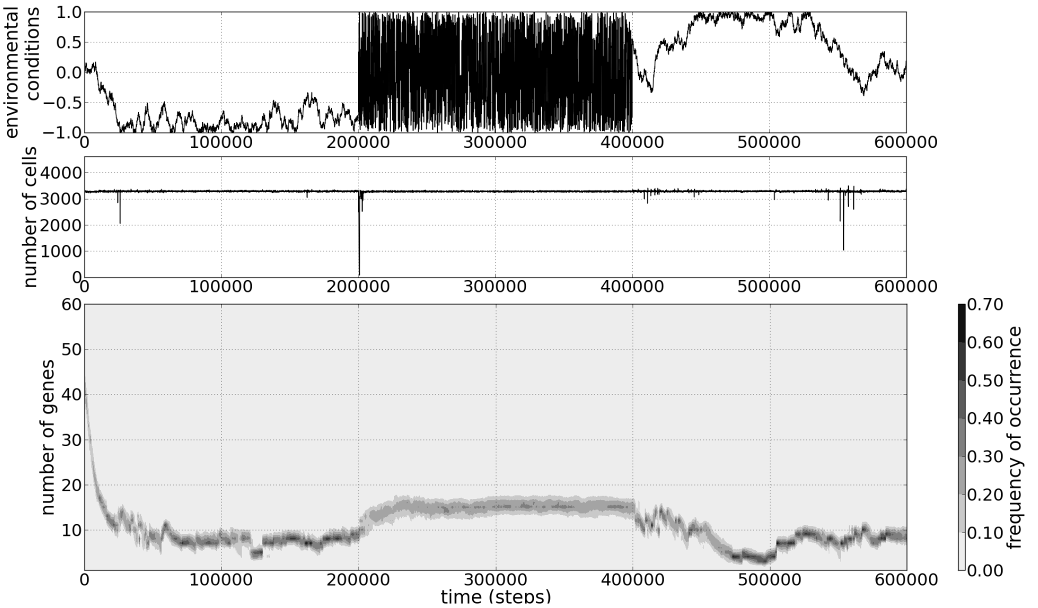

Figure 3.4: Optimisation of genome sizes in an environment with modulated turbulence level. Simulation was initialised with genome ranges of $\eta_0 \in [40, 60]$. Turbulence level was $T = 0.005$ at the beginning, $T = 0.2$ in the middle of simulation and again $T = 0.005$ at the end (see the upper panel). All three panels are shown in the same time scale. All parameters were set as given in Table 2.1

3.2.2 Adaptation of genome size to a given turbulence level

Adaptation to the environment is manifested in the model by the number of genes, their shape and how they are placed in the space defined by the environmental conditions x and their values of the uptake efficiency $U(x)$ (Figure 3.5). If we look at the most frequent genotypes at the end ($t = 2 \cdot 10^5$) of selected runs, then an increase in the gene number with growing turbulence level T can be seen, with the exception of runs within $T = 0.005$ and $T = 0.05$ (Figure 3.5, see description above each panel), which have more genes than the next panel with higher T. Also spatial distribution of genes along environmental conditions x varies, with run $T = 0.005$ having its genes packed very tight in a fraction of x and runs $T = 0.25$ and $T = 0.05$ spaced quite evenly along the whole x axis. In the

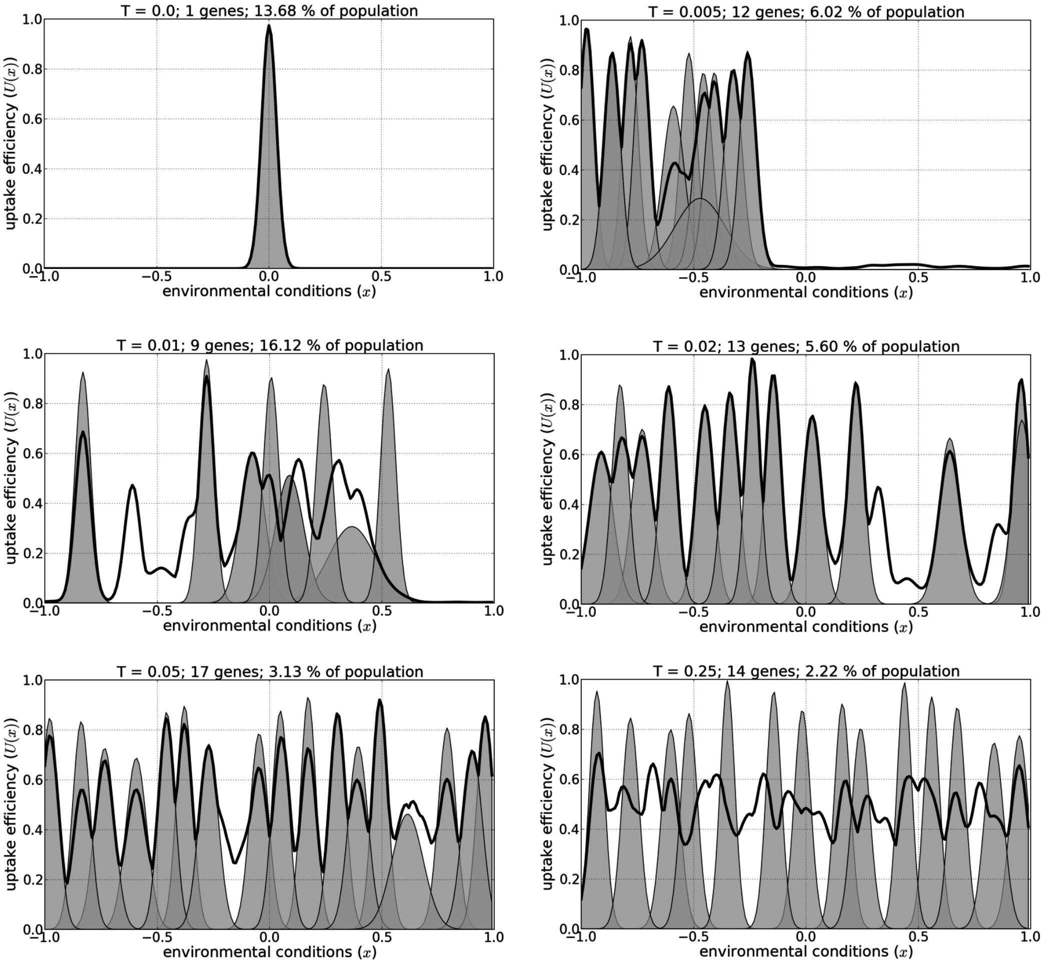

Figure 3.5: Most frequent genomes at the end of a simulation in environments with different turbulence level. Each shaded hump represents one gene in the genotype (note that they can overlap), thick solid line is the averaged uptake efficiency $U(x)$ of the whole population (averaged genotype shape). The following are indicated above each panel: the turbulence level T of the run, number of genes in the most frequent genome and percent of the population that a given genotype constitutes (note that this takes into account only the phylogeny of the genomes, not their apparent similarity). All parameters were set as given in Table 2.1.

run with $T = 0.25$, it can be noticed that each individual gene has its peak higher than the average uptake efficiency of the whole population, but the genotype also has 'wells' in which performs worse than the averaged value. Also runs with $T = 0.01$ and $T = 0.02$ have intervals of x not covered by genotype of the most frequent genome, but on average the whole population has genes suitable to benefit in these values of x. But this intuitive view of individual genomes can be deceiving as it only shows one particular genome in a particular time step.

Let us consider one time step, say half-way through the simulation at $t = 10^5$, and compare it in a number of runs with different turbulence levels T. For stable $T = 0$ the $U(x)$ versus x space is dominated by one single peak placed at $x = 0$, which is the constant value of the environment in the non-turbulent runs (Figure 2.1, $T = 0.000$). These runs evolved on average just over one metabolic gene (see section 3.2.1). As the turbulence level rises, there rises also the number of peaks in the genotype and the level of overlap between genes. But as fewer peaks reach higher value of uptake efficiency (note that the maximum possible value of $U(x)$ is 1.0), the average uptake efficiency of the population also gets lower. For turbulence level of 0.25 and above (not shown), it oscillates around $U(x) = 0.5$ (Figure 3.6, $T = 0.250$).

It can be noticed that for $T = 0.01$ (Figure 2.1, $T = 0.010$) genes are humped in the region of the environmental condition where $x > 0.5$ and the runs with turbulence slightly below $T = 0.005$ and above $T = 0.020$ are not humped. Inspection of environmental conditions x versus time (Figure 3.6, narrow panels) reveals that in the case of $T = 0.01$ environmental conditions have been in the range of values $x \in [0.5, 1.0]$ for at least $2 \cdot 10^4$ time steps and the genome's shape managed to evolve and stabilise when the snapshot was taken, whereas for $T = 0.005$ and $T = 0.020$ the snapshot was taken shortly after the moment of change in the environmental conditions thus the figures present a transition state. Also, it is worth noticing that for all-but-one figure, all runs have a region of environmental conditions where the values of the averaged uptake efficiency is significantly lower than in other regions. Only $T = 0.250$ has the whole space of x covered fairly uniformly by genes.

To gain a better understanding of the time evolution of the genome shapes, let us have the same runs presented as colour maps of values of averaged $U(x)$ versus time (Figure

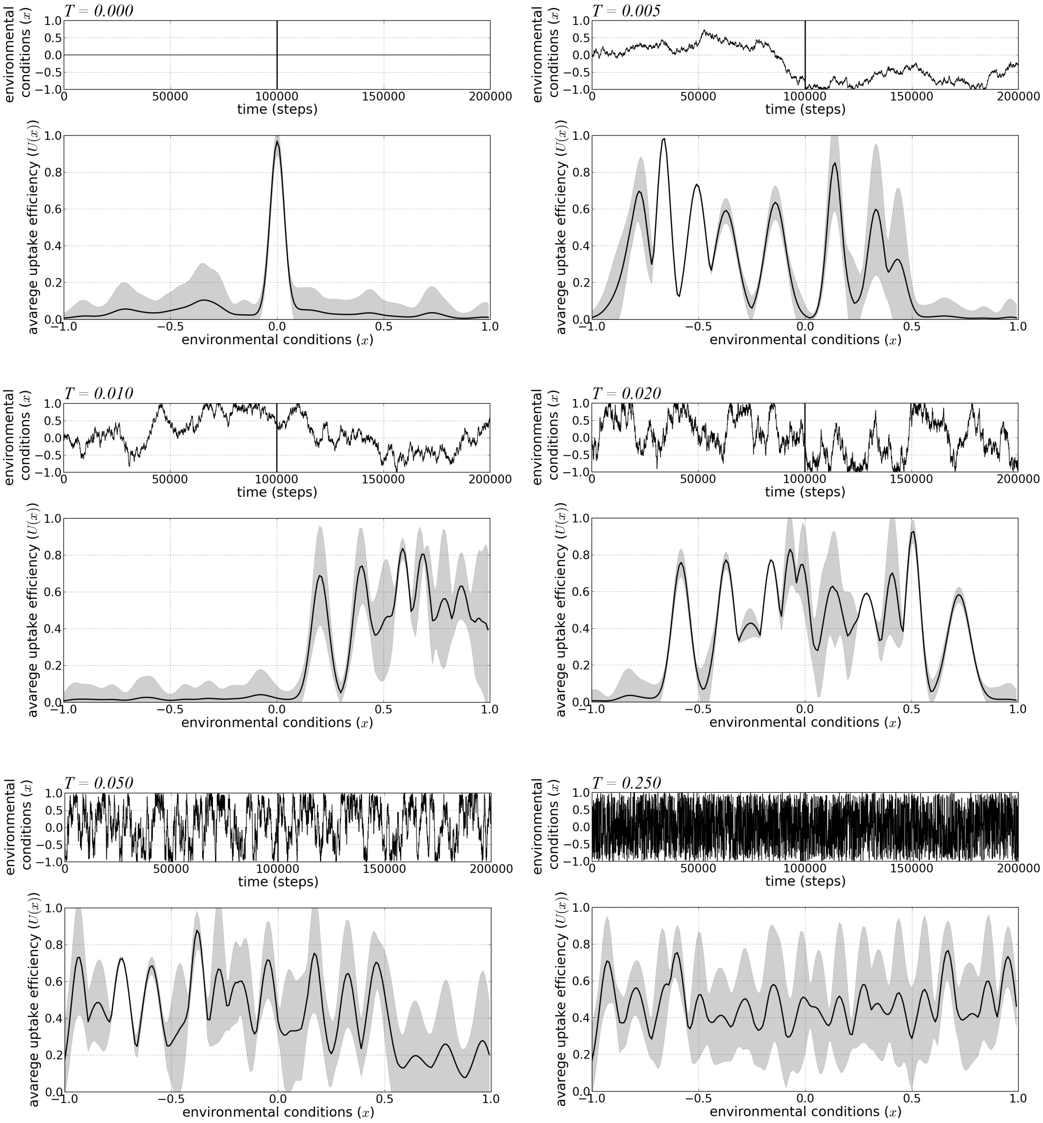

Figure 3.6: Optimisation of genome sizes in environments with different turbulence levels. Each pair of panels (thin and thick) represents one simulation with different turbulence level T values (large numbers between panels). Thin upper panels present the change of environmental conditions x in time during the whole simulation with time point 10^4 marked in the middle with the black solid vertical line. Lower thick panels show a snapshot for time point 10^4 of the mean uptake efficiency $U(x)$ calculated for the whole population (solid line) with a standard deviation (shaded area). All parameters were set as given in Table 2.1.

3.7). They can be understood as averaged genome plots from Figure 3.6, aligned along time axis and flattened. It can be seen that for no turbulence (Figure 3.7, $T = 0.000$) there is an initial phase when the averaged genome has, along the necessary gene, a number of random ones, then there is a phase when one peak dominates but some low noise is still present, eventually, after the time step $t = 1.5 \cdot 10^5$ when the noise is gone, only one peak is left on the stage. When the turbulence level is low (Figure 3.7, $T = 0.005$), the system evolves a number of short-lived peaks with fairly high values of uptake efficiency

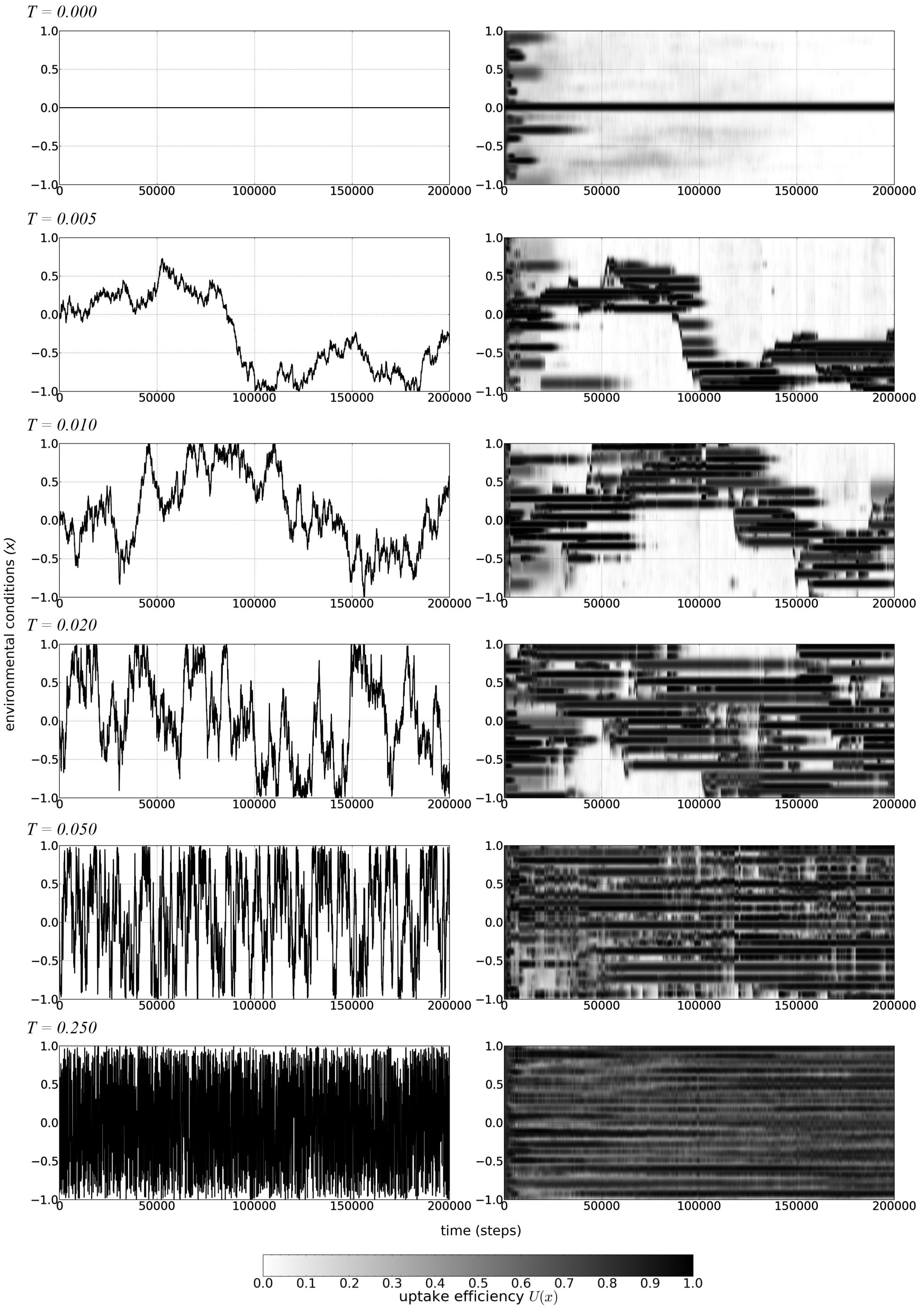

Figure 3.7: Genome shapes evolve in time and depend on the turbulence level. Each pair of panels represents one simulation with different turbulence level T values (number above the panels). Left panels present the change of environmental conditions x in time during the whole simulation. Right panels show colour maps of the average uptake efficiency $U(x)$ calculated for the whole population and evolving in time. Intensity of colours is proportional to the value of average uptake efficiency $U(x)$ for the given environmental conditions x in time step t (see colour bar below panels). Standard deviation of $U(x)$ has been omitted due to clarity of the presentation. All parameters were set as given in Table 2.1.

$U(x) > 0.8$. As the turbulence level gets higher, the time of continuous lasting of a high peak is longer, but also the number of peaks present at a time is larger and the hight of peaks gets lower. Eventually the shape of the genome becomes nearly homogeneous through time, with lower but wider peaks. As T gets higher, the overall coverage of uptake efficiency versus time space also gets tighter by the cost of less effective mean uptake in the population.

A peak of the genome at given x might not be formed by the same gene in all the genomes and that is the case in the runs with $T = 0$. At the end of the run, information about genomes of all the cells in the population was gathered and it was revealed that the peak of the average uptake level $U(x)$ is formed here by a number of genes of different evolutionary origin, but having almost identical shape. A case of convergent evolution. Considering the simplicity of the presented model, this situation might be a regularly occurring phenomenon also in the runs with a higher turbulence level. A look at the two already mentioned runs with (Figure 3.7, $T = 0.005$ and $T = 0.010$) shows that the time step $t = 10 \cdot 10^5$ in run with $T = 0.01$ follows a time of a fairly stable condition of x, whereas in run with $T = 0.005$ it is a moment of a rapid change (Figure 3.7). That would explain why one run has, at $t = 10 \cdot 10^5$, its most recurrent genome in a form of a hump in one region of the x space, when the other has a number of genes spread more evenly. The former seems to undergo a moment of accelerated evolution.

3.2.3 Efficiency of resource uptake

The efficiency of resource uptake is a good indicator of a population's adaptation. We can expect that a well suited population will consist mostly of cells with uptake ratios close to the maximum permitted value τ (Table 2.1). And it is the case when environmental conditions are constant (Figure 3.8, turbulence level $T = 0$): almost all the cells have their uptake efficiency close to the maximum permitted value. When turbulence level has a very low value (Figure 3.8, $T = 0.005$) distribution of uptake efficiencies is much wider and covers all the possible ranges, but still its peak is shifted toward the higher values. For the turbulence level of 0.01 and higher the dominance of cells with high uptake rates is no longer visible and the distribution of uptake efficiency has a shape of a wide hump centred around the median value of uptake efficiency (Figure 3.8, the rest of the panels).

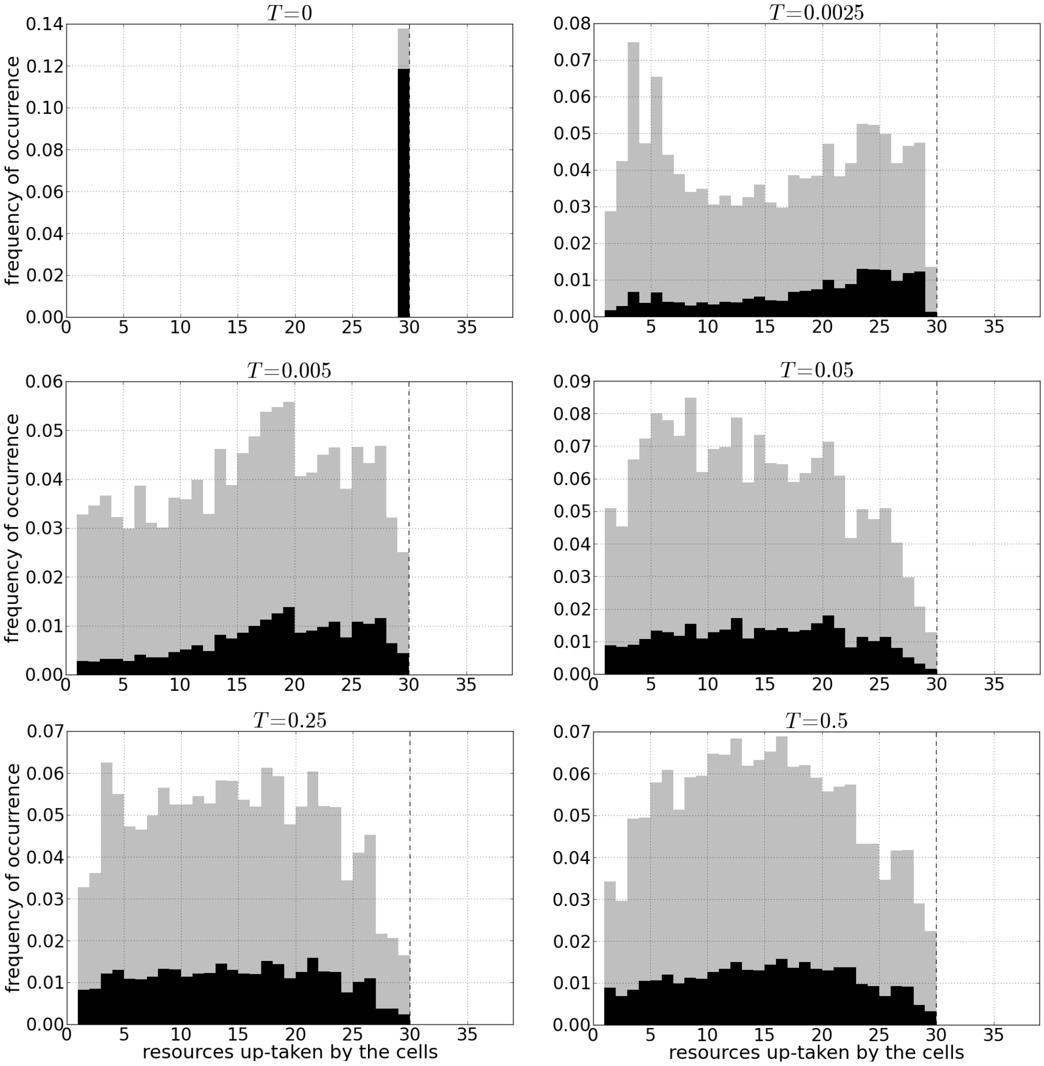

Figure 3.8: Distribution of uptake efficiencies of different cells as functions of the turbulence level. Black bars represent mean frequency of occurrence of different uptake efficiencies averaged over all time steps in the simulation; grey bars represent standard deviation calculated in a similar manner; vertical dashed line is the maximum permitted amount of resources a cell can get in one time step (τ). All runs have parameters as given in Table 2.1. Turbulence level T is shown above the panels. Note that in each case only slightly over 30% of the cells got access to food (data shown in Figure 3.9).

The mean fraction of cells that are fed in each iteration of the model is very small and for the stable environment ($T = 0$) it is between 10% and 15% of all the cells in the population at a time step. This number rises gradually, eventually reaching a plateau for turbulence levels just below $T < 0.05$, stabilising at a level around 35% of the cells in the population. Also SD of the mean fraction of fed cells rises with the rise of the turbulence level, reaching its peak values between 14% to 18% for $T \in [0.05, 0.2]$ and stabilising for $T > 0.1$ at a level of 10% to 11% (Figure 3.9). These quite large variances (note also large error bars in Figure 3.8) show there is a lot of fluctuations among cells regarding

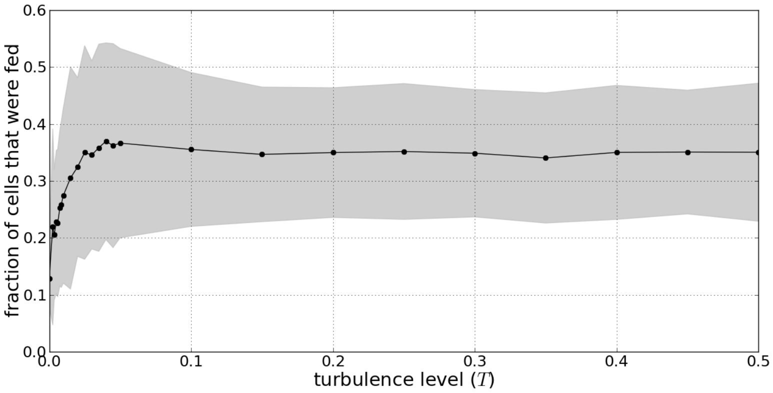

Figure 3.9: Proportion of cells that were fed as a function of the turbulence level. Numbers are calculated in a similar manner as in Figure 3.8. Grey area is standard deviation.

their efficiency in the nutrient uptake ($U(x)_i$).

3.2.4 Population crash in response to turbulence level

All the runs, regardless of their turbulence level, have a very similar average numbers of cells which die and get born in one time step. This number was between 17.9 and 18.9 cells per time step on average. What was different, was the standard deviation which reached the highest values (even up to 16) for populations living under turbulence levels $T \in (0.005, 0.05)$. For other values of T, SDs were around 5.5 (Figure 3.10). The average numbers of dead and born cells are equal, otherwise the population would be at a constant growth or constant decay.

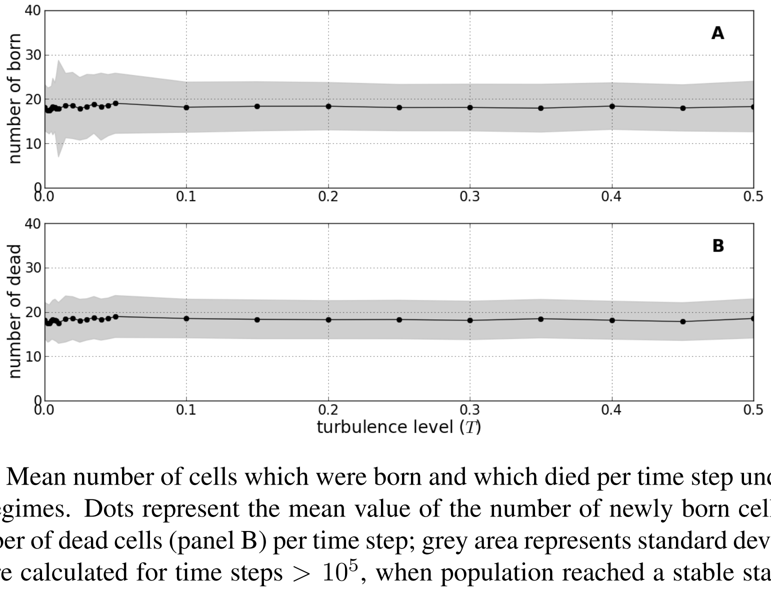

Figure 3.10: Mean number of cells which were born and which died per time step under different turbulence regimes. Dots represent the mean value of the number of newly born cells (panel A) and the number of dead cells (panel B) per time step; grey area represents standard deviation. Both statistics were calculated for time steps $> 10^5$, when population reached a stable state. All runs have parameters as given in Table 2.1.

As the turbulence level increases, we could expect that vast reductions of the number of cells will be happening more often. However, that is not the case. Population crashes are most frequent and are more severe for quite low values of the turbulence level $T \in [0.005, 0.05]$. Figure 3.11 presents a handful of runs under different turbulence regimes. Note that the systems with no population crashes at all are the ones with either no turbulence at all (constant conditions) and the ones with turbulence $T \geqslant 0.1$ (with a system $T = 0.1$ shown in Figure 3.11). The system with the largest number of population crashes has $T = 0.02$. In few rare cases a system with the value of the turbulence level within the abovementioned range got extinct before reaching the end of the simulation.

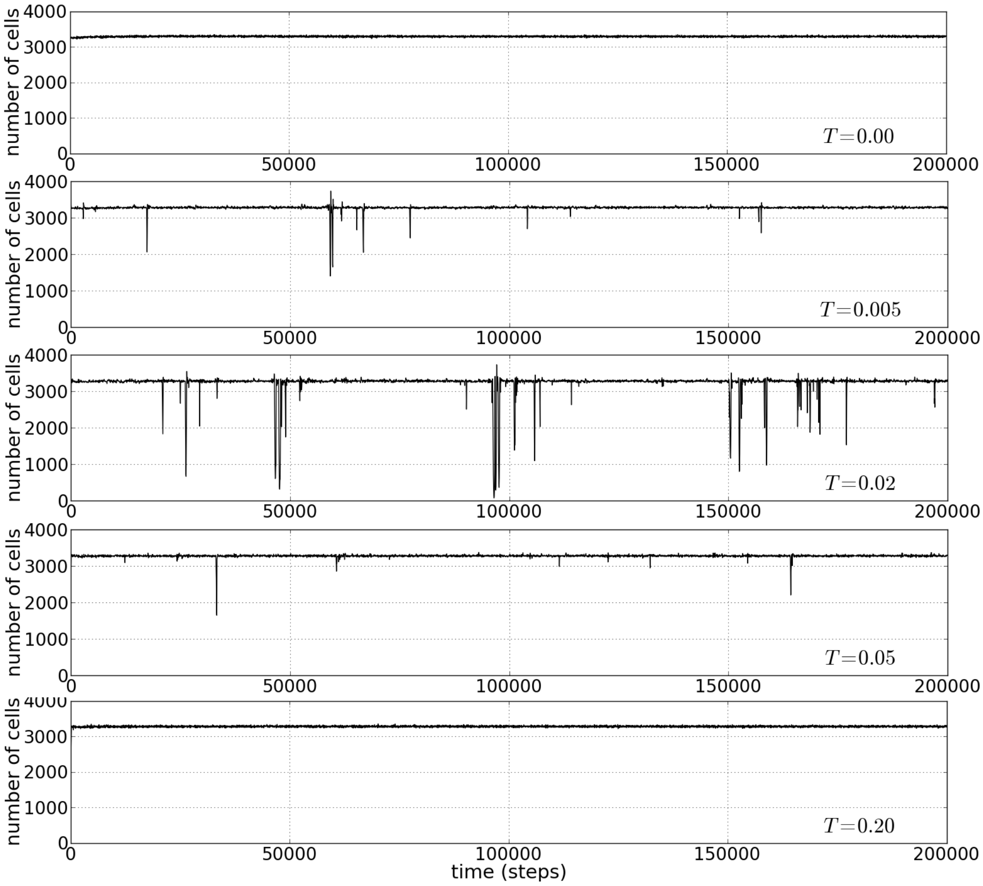

Figure 3.11: Frequency of population crashes depends on the turbulence level. Each panel presents one run with different turbulence level (the number in the lower right-hand corner of each panel). All runs have parameters as given in Table 2.1.

One more thing worth noticing is that when there is a shift in the turbulence level during a run, the model produces a series of vast crashes when changing from a low turbulence regime to a high turbulence one, but there are no crashes when the environment is turning the other way (Figure 3.4). Note that this environmental changes are followed by a change in the numbers of genes in genomes from low to high, when the environment shifts from a low to high turbulence level and vice versa, when the shift is from high to low turbulence.

3.2.5 Mutation rates

As it could have been expected, runs with a constant value of the environmental conditions have the lowest rate of evolution. Yet the highest rates are observed not in the runs with the highest turbulence level. The highest number of mutations per 10^5 time steps (the time after the system has definitely stabilised) is observed for the moderate value of the turbulence level $T \in [0.005, 0.05]$ (Figure 3.12). Gene modifications are the most frequent types of mutations and they are about twice more frequent than duplications or deletions. The number of duplications is almost equal to the number of deletions (note that the numbers of mutations are calculated after the system has stabilised, so any surplus or lack of genes at model initialisation, as compared with the stable state, does not influence this result). The rise in the mutation rate is proportional for all three types of

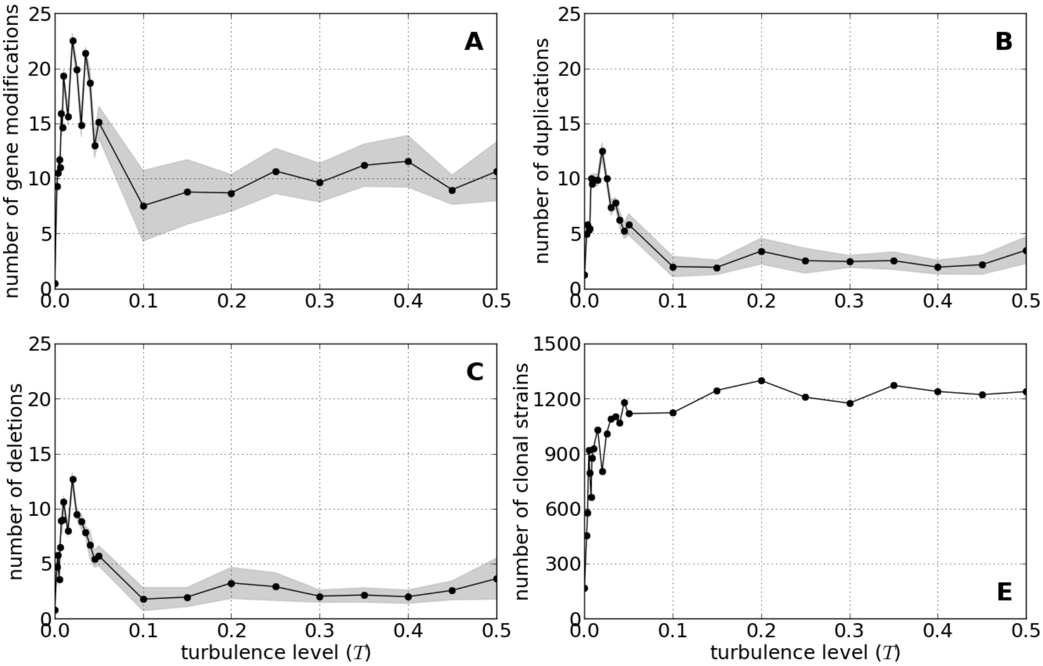

Figure 3.12: Mean number of mutations per clonal strain as a function of the turbulence level. Each dot represents the mean number of mutations of one of the three kinds (panel A – gene modifications, B – duplications, and C - deletions) calculated for the last 10^5 time steps, after the system has stabilised; grey area is SD, calculated in a similar manner. Panel D presents the number of clonal strains at the end of the run (after $2 \cdot 10^5$ time steps). All runs have parameters as given in Table 2.1.

mutations.

There is a strong correlation between the number of mutations of all the three kinds (Figure 3.13). The strongest correlation is between deletions and duplications as the Spearman's rank-order correlation coefficient (which measures if data correlation can be described by any monotonic function) in 10 series of the model runs over the turbulence level space is always $\rho > 0.9$. The correlation between gene modification and the other two types of mutations is weaker but still strong, being $0.7 < \rho < 0.9$ for 10 series of the model runs.

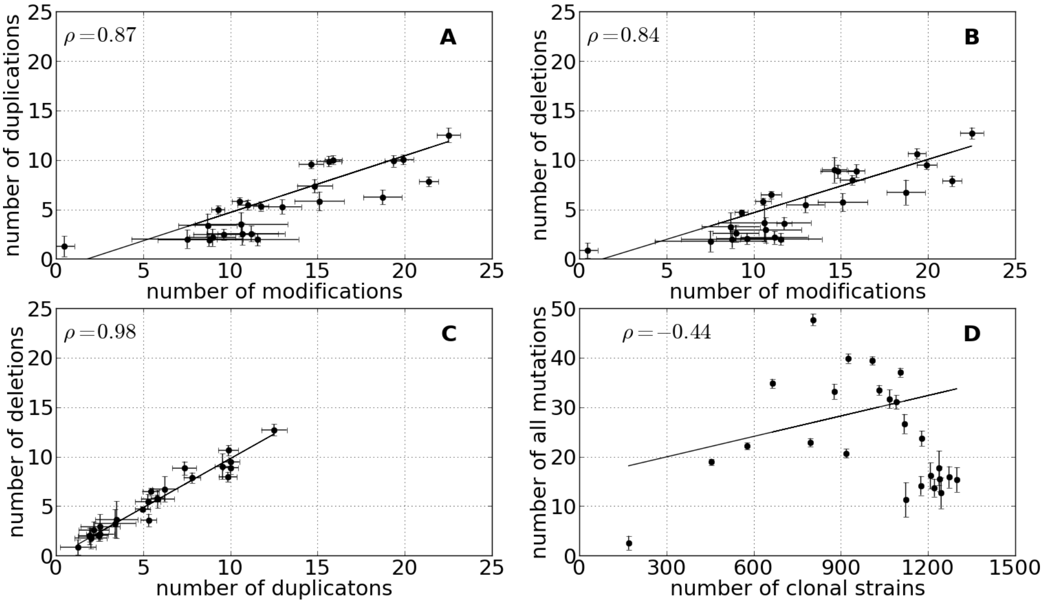

Figure 3.13: Correlation between frequencies of mutations of all kinds and between mutations and the number of clonal strains. In panels A, B and C dots represent the mean number of mutations per clonal strain per 10^5 time steps in model runs with different turbulence levels (same data as in Figure 3.12); bars are standard deviations. In panel D dots represent the averaged sum of all mutations versus the number of clonal strains at the end of the simulation. Straight lines were fitted using the orthogonal distance regression algorithm (fitting a linear model), ρ is the Spearman's rank-order correlation coefficient (checks if any monotonic function applies to explain the data). All runs have parameters as given in Table 2.1.

Another interesting phenomenon is that the sharp rise of the mutation rates is not accompanied by the rise in the number of clonal strains (Figure 3.12, panel D). The total number of all mutations does not correlate well with the number of clonal strains at the end of the simulations ($\rho = -0.44$) for all 10 series of the model runs. It is not the speed of evolution what drives the genetic diversity of the population: in other words, an elevated mutation rate is not the factor behind greater diversity of clones in high-turbulence environment.

It can be concluded that the speed of evolution, understood as the number of mutations which appear and get fixed per 10^5 time steps, depends on the specific turbulence level value T and has nothing to do with the number of the clonal strains a model run produces. A linear correlation between different types of mutations shows that the specific T value influences all three types of mutations in the same manner. Gene modification is the most frequent type of mutational change affecting genomes.

3.3 Conclusions

3.3.1 Turbulence level influences the genotypes' size via variance of the environmental conditions

The turbulence level T, seen as the measure of the environment's instability, has an impact on the number of genes. The gene number initially grows quite fast, but saturates at a relatively small value of the turbulence level of around $T = 0.05$ (Figure 3.1, panels A and B). Let us recall that T is the maximum length of change of the environmental conditions x per one time interval in the bounded random walk of $x(t)$ through time. In other words, in each time step the model picks a random number from the interval $[-T, T]$ and adds this number to the current state of the environmental conditions $x(t)$. Thus T is not a measure of actual environmental variability, as it defines only the maximum possible change per time interval. Variance is more suitable to measure actual variability of x. Variance is the measure of how far a set of values is spread out from the central tendency of the investigated variable (in this case, the mean x). In the discussed model, it can be interpreted as the amount of variation of the environmental conditions x which a cell has to expect to encounter during its life. If $T = 0$, then x does not change at all and variance is also 0. If we compare the variance of $x(t)$ in 25 runs with different T values, we can see that it grows from 0 for $T = 0$ to around 0.33 for $T = 0.05$ and stops growing for any $T > 0.05$. (Figure 3.14). It can be seen that the general pattern of the rise and saturation of variance is similar to the pattern of the rise of the gene number and the rise of the ratio of the surface under the genotype to total environment surface. After the initial increase, it stabilises at around $T = 0.05$, reaching a plateau.

Variance of the environmental conditions x reflects the real instability of the environment. It is the span of values of x that each cell has to be prepared to handle. The

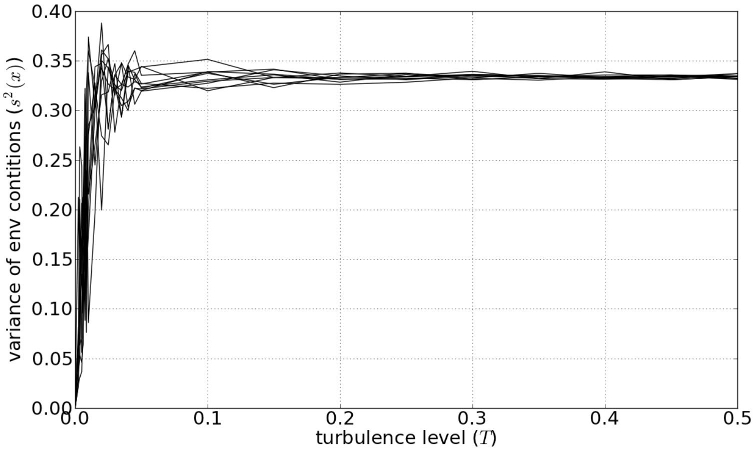

Figure 3.14: Variances of the environmental conditions $x(t)$ of ten series of simulations as a function of the turbulence level. Each line represents a series of 25 runs initialised with the same seed of the pseudo-random number generator. Compare with Figure 3.1, panels A and B.

environmental conditions span from -1 to $+1$. When $T = 0.005$, the maximum change of x value in one time step is 0.25% of the total span of the space of the environment. A change of x from -1 to $+1$ has to take at least 400 time steps, which is extremely unlikely to happen, because it would need 400 changes of the environment in the very same direction at a maximum rate. That is more than the expected life span of a cell (see Figure 2.4). If $T = 0.05$, then the same change of the environment needs a minimum of 40 time steps, which is also more probable than 400 directional steps. That is less than the cell's expected life span. For the maximum considered value of $T = 0.5$, the minimum time for the environmental conditions to shift from one extreme to the other takes just 4 time steps and it is far more likely than 400 directional changes. But from the perspective of the cell, $T = 0.05$ and $T = 0.5$ are not that much different, as in both cases the environment can potentially be in any possible state from the $[-1, 1]$ range during the cell's life. In both cases a cell has to be prepared to handle any of the x values.

3.3.2 The costs of gene maintenance and expression are limiting the size of genomes

The model has two parameters which can be considered to have an impact on the cost of genome expression and thus on its size. The first one is α – the surface under Gaussian representation of a gene; the second is γ – the cost of maintenance of one gene. High α values allow the cells to cover a larger fraction of the environmental space x with less genes. And that is in fact happening: with lower α there are more genes covering a smaller fraction of the environmental space in comparison to systems with higher α values. But despite more than doubling α the mean gene number is kept in a quite tight span of values (see Figure 2.7, panel A). The span of the mean ratios of surface under the genotype to total environment surface is larger (see Figure 2.7, panel B). It seems that there is a force which keeps tight control over the number of genes a cell can have. The cost of gene maintenance γ regulates the 'price' at which one gene comes. The lower γ is, the more genes a cell can have, still spending the same amount of resources. Eight-fold difference in γ values gives approximately two-fold span of the mean gene number values, but produces only one third difference in the mean fraction of the environmental space covered by genomes (see Figure 2.8). Cells tend to have large numbers of genes, even if the benefit obtained from the increase in gene number is not that big.

It can be concluded that cells are driven to have as many genes as possible, and the only force preventing them from expanding their genomes infinitely is the price at which genes come. Each functional gene in our model generates a cost in every time step regardless of the benefit (or lack of it) it brings to its host. Thus, having genes which do not bring any resources to the cell might by dangerously costly. On the other hand, not having a good, beneficial gene at the rare moment of feeding is a loss of an opportunity which might not come again. That can also be lethal. The cost of having genes acts towards streamlining of the genome and the need of being able to draw resources from the environment pushes towards increasing their number. Both of them acting together can generate a tight equilibrium area for the optimum number of genes (see Figure 3.2 and Figure 3.3).

3.3.3 Turbulence level of the environmental conditions impacts the evolution rate

The results have shown that mutation rates are reaching the highest values not for the highest values of the turbulence level but in fact they are the highest for intermediate T. And at the same time they are the values of T in which cells do not get exposed to the maximum level of variance of the environmental conditions x (see Figure 3.14). Also, the same values of the turbulence level are producing population crashes (see Figure 3.11).

The fact that the runs with the elevated mutation rates also suffer from regularly occurring population crashes allows to speculate that we are witnessing a multiple event of an adaptive radiation. Let us remind that the maximum population size, or carrying capacity K as it is referred to in classical ecology (Krebs, 1994; Foryś, 2005), is defined by the total amount of resources R_{env} (see Table 2.1) and that the resource recycling is perfectly efficient. When multiple cells die and their resources are allocated back to the free environmental pool R_{env}, then suddenly there are multiple available resources for the remaining population. More free resources means that cells get fed more often and as a result they absorb more resources on average and grow faster. That allows them to divide more often. Also, more resources means less inter-cell competition. Even a genome normally not appropriate for given conditions will do the job, as the times between feeding episodes are shorter than before the population crash. That is an adaptive radiation event. Later, as free resources deplete, pressure for reasonable genome size increases and feeding becomes less frequent forcing better trimming of the shape of genomes. Many of the newly emerged genomes slowly die out from the population. This hypothesis is in agreement with the observation that when changing from low-turbulence environment (where populations evolve small genomes) to high-turbulence environment (where bigger genomes are required), a sever population crash occurs and when changing the turbulence level the other way around, no crashes are observed (see Figure 3.4).

To understand why systems with rather low T suffer from extinction events and, as a consequence, also show elevated mutation rates, let us ask what are the environmental conditions a cell experiences during its life in comparison to what it should expect if it would live infinitely? As it was said previously, environmental conditions x are randomly generated from a finite interval $x \in [-1, +1]$. The random walk is bound in the way that a value of every next step depends on the values of the previous one and the smaller the turbulence level T is, the stronger this dependence is. The length of every next step is selected randomly from a uniform distribution which means that every sequence of x being near-to-infinity-long will have the mean $\bar{x} = 0$ (a half-cut value of the $[-1, +1]$ interval) and every cell living infinitely long would have to expect that the x will fluctuate around 0. But cells do not live that long. Furthermore, they can starve to death (cells with big genomes will starve faster than the ones with small genomes) and they have to be prepared to accumulate resources at every time step. As a result, cells adapt only to the conditions they are facing in their lifetime – the evolutionary process cannot predict and cannot invest in the future (unless the future is predictably periodic, like e.g. seasons of the year in temperate climates). When T is low then x has a low variance too (see Figure 3.14) which means a cell can expect that the environment will not shift that much from its present state and there is no need to have genes being able to handle remote values of x. We can describe the niche occupied by a clonal strain with two numbers: mean value of the environment within the frame of a cell's expected life span $\bar{x}_f$ and the variance of x within this frame. If, somehow, the environmental conditions will shift from this region, most likely a cell adapted to these conditions will die. To better investigate this hypothesis, lest us consider an average absolute difference between the mean value of x within the stepping frame of length of the life span of an average cell and the grand mean of x over the whole length of the simulation:

$$\bar{D}_f = \overline{|\bar{x}_f - \bar{x}|} \tag{3.1}$$

where $\bar{x}$ is the grand mean of the environmental conditions over all time steps in the simulation and $\bar{x}_f$ is the stepping mean of the environmental conditions over a frame of f time step. The value of f was chosen to be equal to the expected life span of an average cell and it is based on the real death rates measured in 250 simulations and set to 176 time steps. Note that the mean number of deaths does not depend on T (see Figure 3.10). For each step an absolute value of expression $|\bar{x}_f - \bar{x}|$ was calculated. The mean value of this expression is the mean difference of the mean environmental conditions within the expected cell's life span and the grand mean value of x for the whole simulation $\bar{D}_f$. The dependence of $\bar{D}_f$ on the turbulence level is shown in Figure 3.15. On the power of equation 3.1, for $T = 0$ the mean difference $\bar{D}_f$ is also 0. Then it sharply rises to reach its maximum of around 0.5 between $T = 0.005$ and 0.05, after which it slowly and monotonously drops to reach $\bar{D}_f \approx 0.14$ for the maximum considered T of 0.5. If we compare the mutation rates, we can see that the mutation rate is the highest for the same set of T values as the peak of $\bar{D}_f$ (see Figure 3.12). But on the other hand, $\bar{D}_f$ drops slowly, while drop of the mutation rates is rather sudden and reaches a plateau at $T = 0.05$. To understand this, we have to go back to the variances of the environmental conditions

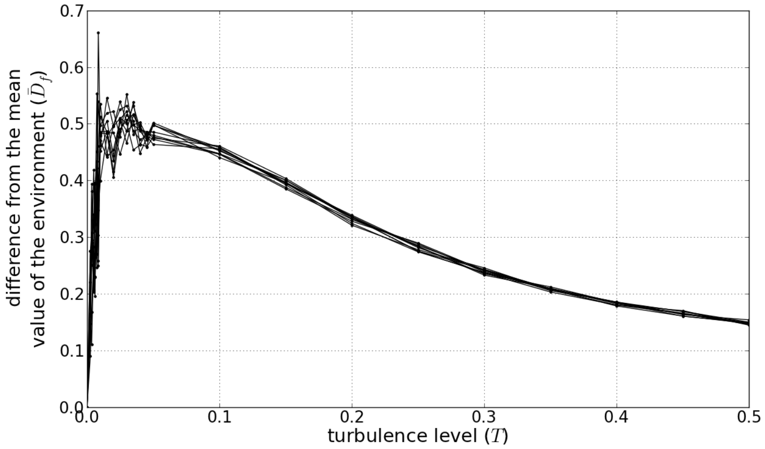

Figure 3.15: Mean difference of the mean environmental conditions x within the expected cell's life span and the grand value of mean of x for the whole simulation. Ten series of 25 runs with different environmental values (the result of each run is represented by a dot) were initialised with a different pseudo-random number sequence. The statistic $\bar{D}_f$ is given by the equation 3.1 and explained in the text. Frame f (see the text for explanation of the symbol) used to generate the plots has 176 time steps.

which also reach their plateau and maximum value at the same time at $T = 0.05$. As we concluded previously, for turbulence level $T > 0.05$ a cell has to be prepared to handle any value of x within its life time. In other words, regardless of the mean value of the environment during the cell's life time, the expected variation of the environment conditions covers the whole space of x defined by the model. In consequence it does not matter for $T > 0.05$ how much the value of x, at which genomes evolved, is different from the grand mean of x as these genomes are adapted to handle any rate of change. Thus, even a rapid shift of the x value will not trigger a population crash, which, as said previously, is followed by an adaptive radiation event, which elevates the mutation rates observed at the end of the simulation. For a similar reason population crash is not triggered by a rapid change from high turbulence level to low turbulence level. Population adapted to a high T is able to handle a low T whatever the new value of x is, because it has the widest possible span of response to change in x. Meanwhile, competition for resources works slowly and reduction of the number of genes under relaxed pressure from the variance of the environment is gradual.

We can see here that we have two types of environments which produce evolutionary stable populations with low mutation rates in their history. One type are environments which do not change at all or change very slowly. Their rate of change is slower than the maximum speed of a gradual evolutionary process. The other type are systems changing fast. Conditions in these environments can change dramatically during an organism's life cycle, forcing its genome to be prepared for a wide span of possibilities. Systems which force a high speed of evolution create a sort of 'illusion' by having a low variation in the short run (and then they force a tight adaptation of genomes), but occasionally they can change quite dramatically. Instead of 'stability' I would suggest to use the term of 'predictability'. The first two types of systems are predictable: the first one has very stable and thus predictable conditions; the second has a high variability all the time, which also makes it predictable. Systems of the third type are unpredictable. They have periods of low variability long enough to force genome size reduction, but at any time they can change very fast to a new state.

Unlike some other systems (Kashtan and Alon, 2005; Crombach and Hogeweg, 2008; Kashtan *et al.*, 2009), this model does not show any traits of evolvability. On of the reasons is the mutation rates are fixed here (see section 2.3.5 and Table 2.1). If the environment is constant, then, after the shape of a genotype settles down fitting the need of environmental conditions, any mutation will be deleterious (will drift the genotype away from the optimum) and cells which mutated will be selected against. As an effect, the observed net mutations numbers in the surviving fraction of the population at the end of the simulation will be low in comparison with populations which went through a period of environmental change. Population originating in high variability environments show greater adaptive potential as there are no population crashes when the system shifts from high turbulence to low turbulence regime (Figure 3.4). This is a result of a pressure to maintain high potential of response to varying conditions in a cell's life span and it is quickly diminished after conditions stabilise. Elevated net mutations numbers in mildly turbulent simulations are not evolvability. They are an outcome of repetitive interplay between phases of selection, acting for trimming down the size of genotype to compete effectively in times of stability and phases of sharp population crashes followed by evolutionary radiation events triggered by a rapid change in the environmental conditions.

3.3.4 Random death factor as an additional dimension of niche

Surprisingly, the random death rate δ has a significant impact on the optimum number of genes in the population. The key factor to understand how δ is altering the optimum number of genes is the cell's life expectancy. If there are no other selection pressures present, the expected length of the cell's life is equal to $1/\delta$. That means a cell has on average $1/\delta$ time steps to gather enough resources to survive and to divide, giving the rise to the new generation. When the death rate is low, this time is quite long. But as the death rate gets higher, there is less and less time to grow and reproduce. Eventually, the time is too short and cells cannot gather enough resources before the random death hits them and the population, step by step, dies out. A comparison of two model runs (both with $T = 0.25$) with the death factor varying by an order of magnitude shows a big difference between how much a cell consumes per one unit of time (Figure 3.16). When the random death rate is $\delta = 0.001$, an average gross gain of a cell which reached reproduction is 2.88 of resource units per one time step and its counterpart exposed to $\delta = 0.010$ (ten times higher) gains on average 7.03 of resource units per time step. In a lower death rate the population evolves less genes which, on average, cost 1.81 of resource units per time step, which leaves 1.06 of resource unit for growth and reproduction. A system with $\delta = 0.010$

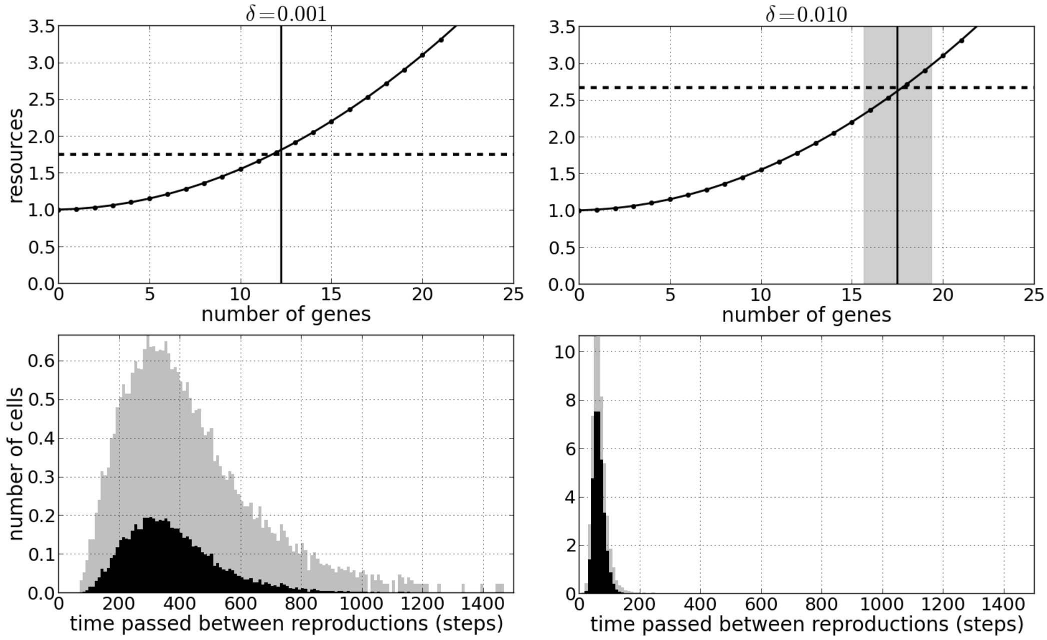

Figure 3.16: Comparison of the reproductive statistics of two model runs with extreme values of the random death factor: left panels – very low random death rate $\delta = 0.001$, right panels – high death rate $\delta = 0.010$. On the upper panels solid-dotted curve is the function showing the cost of maintaining given number of genes averaged per one time step; vertical solid line is the mean number of genes a system evolved with its SD (grey area); horizontal line is the difference between cells gross uptake and the resources it is allocating for the reproduction averaged per one time step (it is 'the cost of living'). Lower panels are the histograms of the number of cells which reproduced after a given time since their last reproduction, black bars are the mean values per time steps, grey bars are the SDs. All values are averaged per one time step. Both runs have parameters as a given in Table 2.1 except the random death δ. Turbulence level was set to $T = 0.25$.

forces larger genomes, which cost 2.61 per time step, but that still leaves 4.41 for growth, which is over four times more comparing to the cells exposed to a lower random mortality. The average time a cell from a $\delta = 0.001$ system needs to grow big enough to reproduce is 290.0 time steps (when the expected life span is $1/\delta = 1000$), comparing to 70.56 (expected life span: $1/\delta = 100$) time steps for a cell from a $\delta = 0.010$ system (see histograms in Figure 3.16). Note two things: that the cell's life span is not the same as the time passed between reproductions and that in each system a cell has to acquire the same amount of resources to reproduce: $\Delta r \approx r_{rep} - r_{min}$. A comparison of the fraction of population which gets fed shows that in the system with a lower random death rate less cells are given access to food in each time step than when random death is more probable (Figure 3.17). That contributes to a lower average gross gain of a cell per time step under $\delta = 0.001$ discussed above. So why do the cells living under more severe probability of death consume more? The answer is in the real death rate. The model has a closed, perfectly efficient resource circulation, which means each dead cell's resources r_i is returned to the free resource pool R_{env} and becomes available for any other living cell to absorb. When more cells die per time step, then more resources are available for

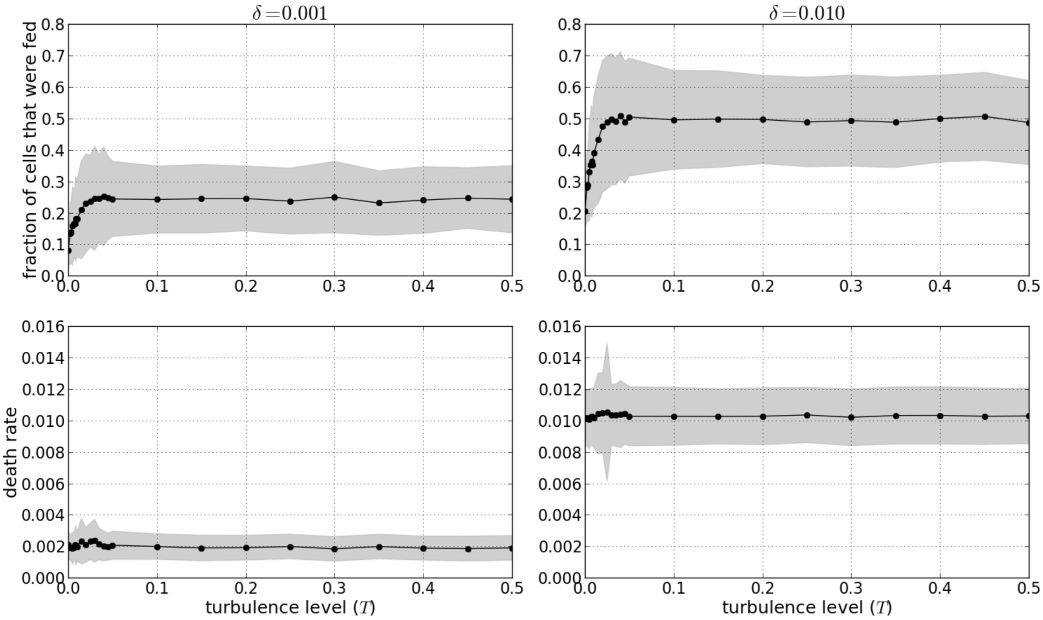

Figure 3.17: Comparison of the fraction of populations of cells which gained access to resources (upper panels) and the real mortality rates (lower panels) between two different random death probabilities under different turbulence levels. Left panels – very low random death rate $\delta = 0.001$, right panels – high death rate $\delta = 0.010$. Dots represent a mean value for each model run with a given turbulence level T calculated after the system has stabilised, the grey area shows the SD calculated in a similar manner. All runs have parameters as given in Table 2.1 except the random death δ.

the remaining living population. The high random death factor increases the resource turnover. Another interesting phenomenon is the actual value of the real death rate. There are just two ways a cell can die: being randomly chosen to die or starving to death. When δ is high then the mean real death rate is just above 0.010, meaning that death from starvation is a rare case in these systems, with δ being responsible for nearly all the dead cells. But when $\delta = 0.001$, then the mean real death rate is approximately 0.002. That is twice as much as the random death probability in these systems, which means that every second dead cell has died because of starvation. That also explains why systems where $\delta = 0.010$ have their cells' age distribution almost perfectly explained by the exponential model and the age distribution of systems with $\delta = 0.001$ cannot be explained in this way (see Figure 2.4). A comparison of how the resources are allocated in the cells shows very similar distributions, with a large number of cells having a small amount of resources and very few having just the amount to reproduce (Figure 3.18, upper panels). But interestingly, systems have different distributions of uptake efficiencies. When the

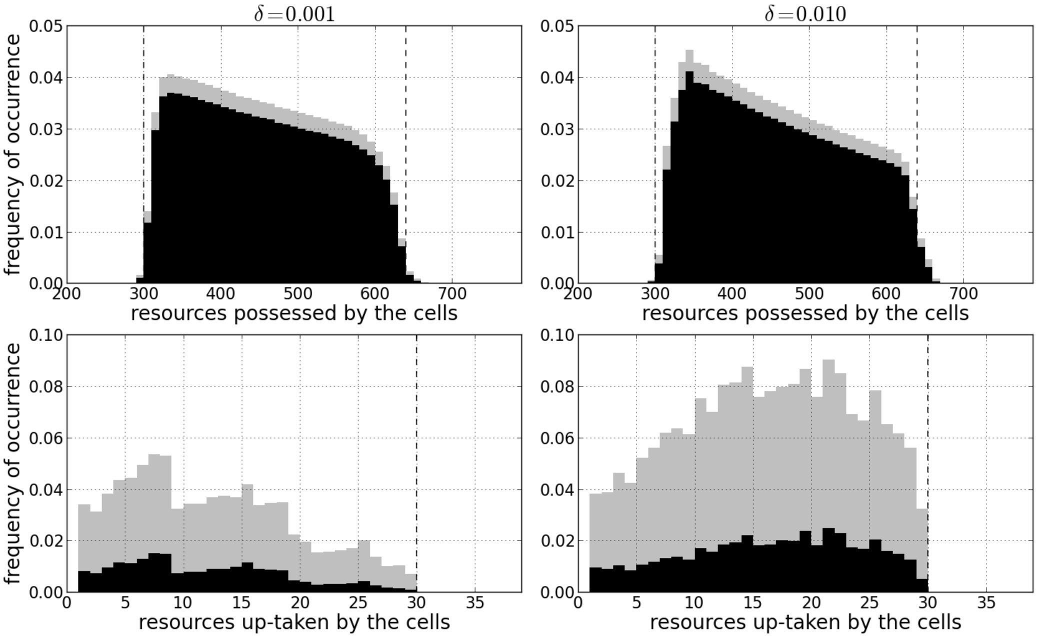

Figure 3.18: Comparison of resources allocated by the cells (upper panels) and distribution of the uptake efficiencies (lower panels) of two runs with different random death rates. Left panels – very low random death rate $\delta = 0.001$, right panels – high death rate $\delta = 0.010$. The turbulence level has been set in both cases at $T = 0.25$. The black bars represent the mean fraction of the population which has the given value, grey bars are SDs. Vertical dashed lines in upper panels are the minimum allowed quota for a cell to live r_{min} and the minimum allowed quota for a cell to reproduce r_{rep}. In the lower panels dashed lines show the maximum amount of resources a cell can get in one time step τ (see Table 2.1). Both runs have parameters as given in Table 2.1 except the random death δ.

$\delta = 0.001$ population has its distribution of uptake efficiency shifted towards low values, that means that if a cell gets fed, it demands little resources on average. When δ is set to 0.010, more cells are located in the upper half of the resource uptake efficiency histogram (Figure 3.18, lower panels). This means that under a low random death rate, cells not only gain access to resources quite rarely, but also if they get access to nutrients, they usually are able to uptake less than their counterparts exposed to greater random mortality.

If the cells' life span is not limited by any factor causing mortality other than starvation (e.g. predation, parasitism, intoxication, sudden worsening of abiotic conditions etc.), then recirculation of resources is very slow because resources get locked in the bodies of living cells. That causes a severe shortage of resources in the environment, which favours the cells being able to survive prolonged periods of starvation. The only way to do this is to have a small genome (see eq. 2.5 and eq. 2.12). If the turbulence T is high, a cell with less genes will have a lower probability of having the right gene for the value of the environmental conditions x at which it was given a chance to feed. In other words, a cell i will rarely have the optimum value of uptake efficiency $U_i(x)$ (see eq. 2.6) for randomly chosen environmental conditions x. This results in low uptake efficiency. But when δ is high, more cells get killed per time step, making the pool of free resources higher. That means that a cell gets access to resources more often and low maintenance costs are not a matter of life and death. Also the ticking clock of random death gives less time to acquire enough resources to reproduce. An optimum strategy is to 'gamble' with the genome size. Cells tend to have bigger genomes allowing to uptake more when given access to the resources because the risk of not being given access to resources long enough to starve to death is smaller than the risk of being randomly killed before reaching the reproduction threshold r_{rep}.

The random death rate δ was designed as one of the model's parameters set to control the resource turnover in the system by preventing resources being locked in the cells for too long. But in fact it turned out to be a trait along which our artificial bacteria had their metabolism and genomes optimised in the evolutionary process. Thus the system initially designed to have just one trait appears to have two. Why are other parameters not seen this way? Because all the other parameters presented in Table 2.1 should be seen as physical (τ), biochemical (α, μ) and physiological (κ, r_{min}, r_{rep}) constraints which have been put on the cell metabolism. Meanwhile, the random death factor δ is dependent on the external environmental factors, just like the environmental conditions x and its turbulence level T.

3.3.5 Turbulence level increases the genetic diversity of the population

For all the parameters chosen the genetic diversity of the population rises in a similar manner as the gene number and the envelope of the gnome to environmental space ratio (see Figure 2.6 and 2.10). Low turbulence environments generate smaller diversity than systems with $T > 0.05$. In low values of turbulence level T, the Shannon index shows a lower mean value, but also sharp peaks of increase occur from time to time (Figure 3.19). Drops and peaks of the Shannon index occur during restructuring of the population.

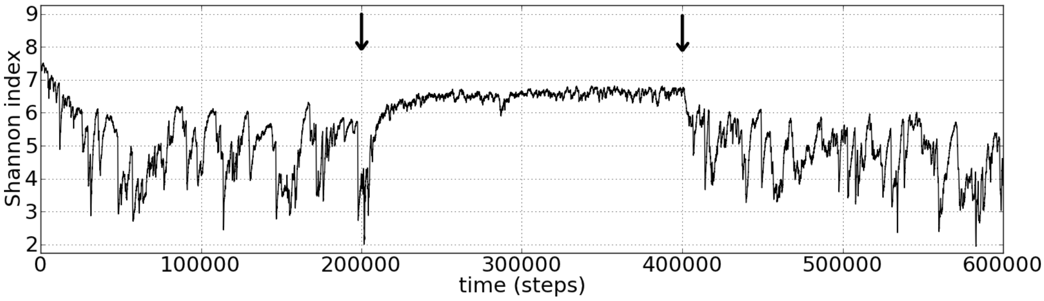

Figure 3.19: Change of the Shannon index in time for a run with three different turbulence regimes: $T = 0.005$ for $t < 2 \cdot 10^5$; $T = 0.2$ for $t \in]2 \cdot 10^5, 4 \cdot 10^5]$ and again $T = 0.005$ for $t > 4 \cdot 10^5$. Black arrows indicate time steps when the turbulence level was changed. That is the same run as the one presented in Figure 3.4. All parameters were set as given in Table 2.1.

A deep drop in the Shannon index around $t = 2 \cdot 10^5$ (a moment of rapid shift from $T = 0.005$ to $T = 0.2$) took place at the same time as a big extinction event (see Figure 3.4).

The lower diversity in stable systems seems to be a bit counter intuitive, as we have gotten used to the paradigm that stable environments produce larger diversity. But this paradigm refers to the diversity (or, more often, species richness) of an ecosystem and in this model we are discussing stability of the niche. The system is more similar to the one in the classic Gause experiments on mixed population of two species of yeast (Gause, 1932). Gause kept two species of yeast in controlled homogeneous conditions. At the end, he always found one species taking over the whole system and the other going extinct. When our models are running in a stable regime ($T = 0.0$), the Shannon index keeps dropping until it reaches a stable value of around 4 (Figure 3.20). This does not

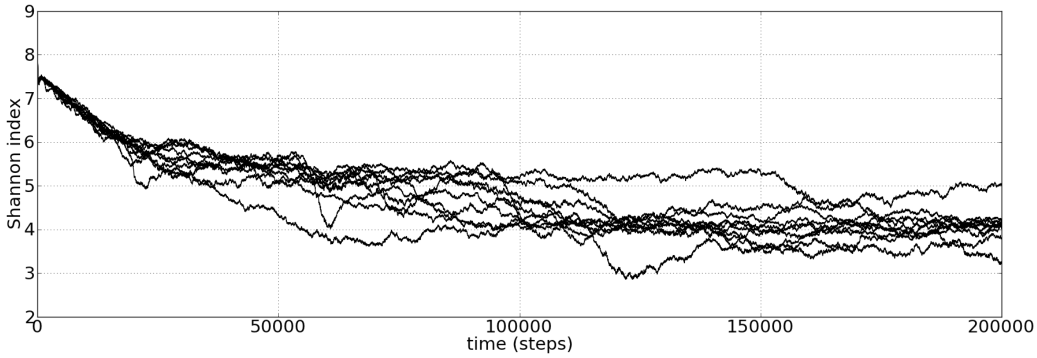

Figure 3.20: Change of the Shannon index in time for ten runs with no turbulence ($T = 0.0$). All parameters were set as given in Table 2.1. Note that the populations at initialisation consist of cells belonging to a different clonal strain each, thus they have at $t = 0$ the maximum possible value of the Shannon index for this size of population: $H \approx 8$. Runs are initialised with different pseudo-random sequences.

happen in high turbulence runs because there are no long periods of stability which would give the best fitted genotype enough time to prove its supremacy in terms of the reproduction rate. The system shifts suddenly, favouring a different genotype and before this one will manage to dominate, the environment changes again and again, maintaining this high genetic diversity. Drops in diversity are due to too deep change in the environmental conditions, causing some of the lineages to go immediately extinct rather than be eliminated by a slower competitive exclusion mechanism. If the environmental change is very severe, we have an extinction event which, as discussed above, is followed by an adaptive radiation, again boosting genetic diversity.

3.4 Discussion

3.4.1 Fluctuating environment has an impact on the size of genomes

The results shown here are in consistency with some of the results obtained from bioinformatics analysis of the number of annotated bacterial genomes: species living in more variable environmental conditions do have more genes (Parter *et al.*, 2007, supplementary information). However, it is not the number of genes that was found to be the strongest indicator of variability of the living conditions in annotated genomes, but the structure of the genetic network. Genomes from more variable environments have more modular organisation of linkages between genes (Parter *et al.*, 2007) (see section 1.2.2). Theoretical research on the impact of environmental variability on genetic networks showed that variability of the environment and forced extinctions will lead to the emergence of a modular network, but when the environment will be switched to homogeneous conditions, the modularity will disappear (Kashtan *et al.*, 2009).

This shows that what is important in tackling turbulent environment and the perspective of change is the potential of the species to undergo fast transformation of genome. This can be ensured by having a larger genome, which gives two kinds of advantage: on the one hand, it might be that genes beneficial in the new conditions are already present in the genome and on the other hand more genes give a greater chance for a beneficial mutation to occur (but it has to be remembered that a majority of mutations are deleterious (Lynch, 2007)). The second important feature is the architecture of the genetic network, which allows for a rearrangement of the metabolic pathways fairly fast and in a way minimising the risk of deleterious changes. As already said before, this property has been described in the literature as ‘evolvability’. But is it really a trait which is a subject of the selection pressure?

The model presented in this thesis, due to its simplicity and assumptions selected, cannot recreate evolvability. It has been shown that the model’s genomes adapt only to the conditions present during a cell’s life time (see sections 3.3.3 and 3.3.4) and that the existence of genes which are not necessary at a particular value of environmental conditions $x(t)$ is a result of inertia and these genes eventually disappear from population (see Figure 3.7, $T = 0.005$ and $T = 0.010$). Nonetheless, some properties of the system resembling evolvability can be observed, e.g. lack of population crashes when the environment shifts from a high-turbulence to low-turbulence regime (see Figure 3.4). But that is not evolvability. As said before, these properties emerged as adaptation to turbulent regime of the environment and the lack of population crash is just a result of the fact that natural selection works here in a gradual manner. It has been demonstrated that the highest turbulence does not in fact produce the highest mutation rates, while mild turbulence does (see section 3.2.5). One might wonder if the high gene number and higher modularity of genetic networks (and their resulting lower integration) in variable environments is not an adaptive ‘evolvability’, but just an artefact of constantly changing goal, towards which the selection pressure aims. Thus we are dealing with a by-product, not a selected trait, and what we observe are constantly ‘half-baked’ genomes always being under construction.

The notion that this is a feature which arose from natural selection might be a bias resulting from under-appreciating transients dynamics in ecology and evolution (Hastings and Higgins, 1994; Hastings, 2004; Olszewski, 2011).

3.4.2 What is stability?

Stability of the environmental conditions is associated with the increase in biodiversity, whereas the presented model showed a decrease of genetic diversity under constant conditions and a rise of diversity when conditions were changing at a high rate. We have to define what is meant by 'stable conditions' in these two cases. Biodiversity and species richness are measures associated with the ecosystem level of biosphere organisation. Meanwhile, stability of the environmental conditions in the model is associated with the niche created *in silico* for a population of virtual prokaryotes. The niche and the ecosystem are very different levels of biosphere organisation.

The system of links between elements of a species-rich biocoenosis is a very complex structure which emerged from a large number of processes, often after a very long time of interactions. An approach which sees at biological systems as networks has been quite successful and has recently gained popularity in ecology and other disciplines of natural sciences (Strogatz, 2001; Newman, 2003; Proulx *et al.*, 2005). Many of the networks observed in biology are large and complex systems. It is imaginable that a big ecological web is easier to build in a stable framework. Predictability of climate (irradiation, precipitation, temperature), geological structure of the underlying continents, low vulnerability to changes in the Earth's orbit – these are the conditions which make a solid and stable framework for the equatorial rain forest, which is one of the world's most diverse biocoenoses. But that does not mean that the nodes in the web cannot change. And the bigger the web, the least vulnerable it becomes to removal or addition of a single node: e.g. it has been shown that in grassland ecosystems stability of the community's biomass is positively correlated with species diversity. When drought-sensitive species decline in their biomass they are compensated by growth of drought-resistant species via the competitive release mechanism (Tilman, 1996). We can expect that once a huge network will emerge, it will be a system fairly robust to perturbation as long it has the support of a solid framework. Macro-scale stability is necessary.

In this model, we asked rather about the stability of a single node in a web. We investigated what are the consequences of the stability of conditions within a niche for the evolution of a free-living prokaryote. Undoubtedly, a complex ecosystem will demonstrate all three types of niches described previously: very stable niches in which conditions have not changed for millions of years; niches with conditions rotating faster than the life span of a generation; and niches which are collapsing after a period of relative stability of conditions. This is a very important level of organisation of the biosphere, i.e. the level where individuals have a direct interaction with their environment and thus it is the level where genetic evolution takes place. In this research, we investigated the influence of micro-scale stability on the evolution of prokaryotes, which can be a significant phenomenon in e.g. evolution of antibiotic-resistant pathogens.

It is also worth noting that the model produces stable and repetitive values as regards the genome size and coverage of the environmental space with genes on a given turbulence level T, but these are not stable equilibria and they can be easily perturbed by small deviation in T. This is not the persistent asymptotic behaviour which we observe in Gause's experiments with *Paramecium* – these are stochastically driven transient dynamics, sensitive to small perturbation. Unstable equilibria can be observed also in laboratory experiments, but transient behaviour was shown to be dependent on the starting conditions (Cushing *et al.*, 1998). This approach gained some attention in ecology (Hastings, 2004) and was used in studies of food webs (Chen and Cohen, 2001). Focus on dynamics might be useful in the analysis of genetic networks as well.

3.4.3 Cost of gene maintenance and expression as limitation of genome size

It has been shown in the model that the decrease of the cost of maintenance of one gene (γ, see equation 2.12) triggers growth of the genome size. This parameter is a cumulative representation of: the costs of replicating the gene, the cost of its reparation, and the costs of its expression (transcription, translation, splicing etc.). In living cells most of the energy is allocated to protein production – up to $\sim 75\%$ of the cells' total budget (Harold, 1986), so the expression of genes is the most significant expense of the cell. It also has been shown that prokaryotes have less power available per one gene than eukaryotes: it is on average 0.03 fW per gene in bacteria and 57.15 fW for an average eukaryotic cell (Lane and Martin, 2010). This difference is also accompanied by disparities in genome size and complexity of the gene expression regulation systems, making eukaryotes far more complicated in those terms than prokaryotes (Lynch and Conery, 2003; Lynch, 2006b, 2007).

The large difference in available power, attributed to the rise of mitochondria in eukaryotes (Lane and Martin, 2010), accompanied by strong purifying selection in huge prokaryotic populations (Lynch, 2006a), might lead to the emergence of small-sized genomes among prokaryotes as compared to those of the eukaryotic domain.

3.4.4 Dormancy as a response to periods of lethal environmental conditions

In the model, intermediate turbulence levels produced population crashes and some of them were very severe, diminishing the population from around individuals to around $3,200$ individuals to around 10. Does this situation happens in nature? Severe population crashes are observed (Hawkins and Holyoak, 1998) and even mass extinctions did happen in the Earth's history (e.g. Raup and Sepkoski, 1982; Stanley and Yang, 1994). Can populations adapt to avert them? In the presented model, it was not possible, mostly due to the way environmental conditions x were generated. The change of x is random and the direction of sudden shifts of x is unpredictable. Also simple representation of the genome does not allow for much adaptation in those terms.

But in the natural environment some species can survive periods of worsening of conditions. If the environmental conditions are predictably periodic or it can be expected that the worsening of the environment will eventually pass, then the ability to undergo dormancy can be a good strategy. We do know a handful of examples among bacteria being harmful to humans e.g. *Clostridium botulinum*, *C. perfringens*, *Bacillus cereus* (Carlin *et al.*, 2000) or an obligatory pathogen *Bacillus anthracis* causing anthrax, infamous for its use as a biological weapon. There are also reports of free-living prokaryotes being able to go dormant, e.g.: in the case of the soil bacterium *Bacillus subtilis*, we know the molecular mechanism which leads to formation of endospores (Errington, 2003); some of the species from the *Clostridium* and *Bacillus* groups found in food products are able to form spores (Postollec *et al.*, 2012; Scheldeman *et al.*, 2005); and a number of other genera later separated from the *Bacillus* group: *Paenibacillus*, *Brevibacillus*, *Aneurinibacillus*, *Geobacillus* and *Virgibacillus* (Reva *et al.*, 2001), many of them free-living. Even in non-spore-forming bacteria it has been shown that some sort of dormant forms do exist (Suzina *et al.*, 2006).

Among pathogenic microbes, especially obligatory pathogens, the existence of a dormant form is quite obvious. It is a non-obligatory developmental stage which allows to survive when being outside the host. Furthermore, because each species has a unique composition of proteins and tends to sustain homeostasis, it is easy for a pathogen to recognise when it is in a host and transform back into an animate form. Free-living microbes have a slightly tougher task. Abiotic signals are more complicated to analyse and often the signal can be misleading, e.g. a rise of the environment's temperature is very short and it quickly goes back to the original 'cold' values. So, going dormant is in fact a gamble in the case of free-living cells. But this sort of risk may pay off, especially when we would look at it from the perspective of a clonal strain instead of a single cell. We should expect slight randomness in the clone's response to signals about the change in quality of the environment. If only a fraction of cells will go dormant, then, if the change in the environmental conditions is temporary, some cells will stay active and the clone will continue to exist. Also, when cells are dormant, we should expect that a positive signal from the environment will not trigger all the cells going back to the animate form. When the signal will appear to be a 'hoax', then there is still a fraction of the population waiting for a better time to come. Diversification of responses of a clonal strain may be one of the answers to changes of a turbulent environment.

Claims that nearly all bacteria have resting stages, called 'microcysts', have been made since the early days of modern microbiology (Bisset and Moore, 1952). Unfortunately, there is a limited number of examples from marine and soil ecosystems of free-living prokaryotes undergoing any form of dormancy. Furthermore, in the light of the difficulties researchers encounter in their attempts to keep and reproduce marine and most soil prokaryotes in laboratory conditions, experimental verification of the ecological and evolutionary relevance of dormant stages in free-living microbes seems to be unreachable at the moment.

In this model, dormancy could be implemented. One way of doing this is to allow cells to be aware if the value of x remains outside of the cell's genotype, in other words, if the uptake function $U_{i,t}$ covers sufficiently the current value of environmental conditions $x(t)$. If the genotype is not covering current environmental conditions for long enough, then the cell goes dormant. The situation becomes more problematic when the timing of bringing a cell back to its animate form is considered. The simplest approach is to revive cells at random with a certain probability. Another problem is how long a cell can be dormant as there has to be a certain cost of being inactive. One, for sure, is being outnumbered and eventually out-competed by lineages which did not go dormant and were able to reproduce normally in given conditions. Another interesting question is how will dormancy impact the size of the genome?

3.4.5 Linkage between traits of niche via time-dependent trade-offs

For a bacterium the main purpose of having a set of specific metabolic pathways is to gather enough resources and energy in its habitat to divide within a reasonable time frame. In most cases bacteria, though potentially immortal, do not have much time to reproduce because of a large number of environmental threats causing mortality, e.g. predation, virus infections, thermodynamic processes of molecule decay, harmful physical factors (e.g. electromagnetic radiation, temperature). These factors define the maximum expected life span of a bacterial cell, constituting a time frame in which a cell has to be able to gather enough resources and energy to live. Of course, it is a stochastic game of numbers, so some cells live shorter, some live longer. Nevertheless, the average time passed from one reproduction to the next one has to be shorter than the time frame of the expected life span. Otherwise the population will not sustain itself in the given environment.

This time constraint draws a trade-off on gathering resources. All the metabolic pathways have to provide the cell with enough resources to reproduce before the time horizon of the maximum expected life span. That puts a tough constraint on efficiency and on costs generated by the metabolic pathways, making the selective pressure on them very strong – poor genotypes are removed immediately, good ones get fixed fast.

The model we proposed does not allow the random death rate δ to change during the simulation and cells cannot alter it in any way. But *in vivo* cells do counteract most of the destructive processes: they have enzymes repairing damaged DNA, they produce toxins to chase away predators and competitors, they have enzymes removing invasive DNA, etc.

All these adaptations do not provide new resources, on the contrary, they cost resources and energy. They can be seen as ways of postponing the expected time of death, ways of 'buying time' necessary to gather enough resource to reproduce. Of course that brings a lot of space for different sorts of optimisations and trade-offs which were not represented in the model. For example, if the abundance of viruses is very low, then the probability of encountering any of those before the time necessary to gather resources will pass is negligible, thus the selective pressure favouring genomes immune to invasive DNA/RNA will not be strong enough to fix the anti-virus genes in the population. It pays off to be vulnerable to infection but at the same time being able to allocate more resources into growth and reproduction.

Dependencies described above provide links between the evolution of the resource-uptake machinery and the evolution of mechanisms preventing the cell from dying prematurely. These links have two underlying mechanisms:

1. Limitation by a finite amount of resources available to a cell which forces resource allocation trade-offs.
2. Limitation by the time frame in which these resources should be collected.

As a result, some of the traits of the species' niche cannot be separated from one another and have to be investigated jointly.

3.4.6 Optimisation of the niche width

A fair body of theoretical work has been put into studying the problem of niche width and niche overlap between competing species, also in response to varying environment (see section 1.2.1)In this study, we are investigating niche width optimisation in response to limitation of physiology, fluctuations of the environmental conditions and competition not with other species, but within the same species. This may seem very similar, but there are certain important difference which are worth pointing out. In the traditional approach (e.g. May and Mac Arthur, 1972; Scheffer and van Nes, 2006; Fort *et al.*, 2009; Yamauchi and Miki, 2009), authors focused on inter-species interactions and this is manifested by the fact that the environmental parameter space is wide and accessible to species at all times at its full width. This in turn means that species can drift far apart in the environmental parameter space as time passes, simulating e.g. divergent evolution of birds feeding on seeds of different sizes. In the model presented in this thesis, the accessible parameter space is time-limited and it is very tight then: just a scalar number. Different genetic lineages observed in the model cannot be called different species as they come from the same ancestral population and occupy the same effective niche, whereas when using the classical Hutchinson's definition of a niche, we can expect at least some difference in the niche occupation pattern between species (Hutchinson, 1957). This model does not give such an opportunity and what we observe is the evolution of a single species with no chance for divergent evolution. Using the metaphor from the work of Scheffer and van Nes (2006), this model does not investigate what is happening between the 'humps' (utilisation functions of species, usually Gaussian-shaped in the literature: see Figure 1.1), it looks at the temporal dynamics inside the hump and processes shaping the hump in response to abiotic forcing. This is important as it was shown that the form of the utilisation function is crucial in shaping the distribution and coexistence of multiple species on the niche axis (Szabó and Meszéna, 2006).

There have been attempts to add evolution and temporal variability to models derived from the classical approach. One way is to introduce time-dependent perturbation to the location of the maximum of the Gaussian-shaped species utilisation function and observe how the multi-species system reacts to such perturbation. It was shown that this type of perturbation elevates the biodiversity of the community (Yamauchi and Miki, 2009).

Furthermore, the difference appears to be more profound than the time-dependent accessibility of the parameter space. The formulation of the niche, though inspired by the Hutchinson approach, turns out to be not exactly a single axis per one environmental factor. The niche is in fact a combination of an abiotic factor (in the model called simply 'environmental conditions x') and time-dependent factors such as turbulence T and random death δ, limiting the possible life span (see section 3.4.5). The emergent properties of the model turned to be more complicated than it was anticipated at the beginning.

Chapter 4

Horizontal gene transfer in single homogeneous population

4.1 Introduction

4.1.1 Evolutionary meaning of horizontal gene transfer

In a classical Darwinian concept of biological evolution, species change through slow gradual steps appearing on transitions between generations. In modern gene-centred evolutionary biology these steps are attributed to various kinds of mutations. But what if a species took a shortcut and 'borrowed' some useful genes from another species? Horizontal gene transfer (HGT), as it is called (also known as lateral gene transfer), has been intensively studied in recent years, being pointed as a significant source of new genes in prokaryotic genomes. The fact that bacteria can exchange their genetic material was known already in the early years of microbiology and that observation led to the formulation of a hypothesis of cross-species gene transfer (Syvanen, 1985). First suggestions of its significance came from studies on *Escherichia coli*. It was shown that some portions of the *E. coli* genome have significant codon frequencies deviations from the pattern observed in the majority of the bacterium's DNA and this affects in about 15% of the bacterium's genome (Médigue *et al.*, 1991). As it was known that a number of genes from this bacterium is similar to bacteriophage genes, a hypothesis was proposed that these parts of DNA came from HGT events (Koonin *et al.*, 2001). Later, comparison of three ecologically distinct stains of *E. coli* showed that only 40% of their genes are the same

(Welch *et al.*, 2002). If it was assumed that the common ancestor had all these genes, which later simply got lost in individual lineages, this would mean that it had an unrealistically big genome. Thus most likely the majority of them are new acquisitions from different species (McInerney *et al.*, 2008). Recent investigations suggest also a great significance of HGT in protein families expansion among prokaryotes: even between 88% to 98% of expansions of protein families are due to HGT (Treangen and Rocha, 2011). Many species from the *Alphaproteobacteria* group possess gene transfer agents (GTAs). GTAs are host-encoded virus-like elements that package random fragments of the host chromosome (Lang and Beatty, 2007). Intraspecific gene transfer based on GTAs has been discovered in marine bacterioplankton (Biers *et al.*, 2008) and it was also shown that GTAs are not uncommon in marine bacteria from different communities (Lang and Beatty, 2007). The estimate is that between 1.6% to 32.6% of each microbial genome comes from HGT, mainly by acquisition of whole operons (Koonin *et al.*, 2001; Price *et al.*, 2008).

A recently proposed hypothesis, at the moment lacking convincing data support, claims that widely spread HGT in microbes is a matter of essential need under genome size constraints. Strong selection pressure forcing development of small genomes will remove ‘unused’ genes very fast and the only way to get new genes when a new need arises is to ‘borrow’ them. Without HGT events microbes would perish during evolution (Isambert and Stein, 2009). Prokaryotic chromosomes are not surrounded by a membrane, which makes HGT easier than in eukaryotes (which have their genome encapsulated inside the nucleus). But that is a ‘chicken or egg’ type of question. Is the chromosome exposed because of the need for HGT, or rather does HGT happen so often because of the exposition of the chromosome?

Voices of criticism regarding the importance of HGT in evolution of prokaryotes are rare and focused on the problem of how often HGT events happen and what fraction of transferred genes settles in the recipients genome. It seems that only between 10% to 15% of acquired genes are retained in a long time perspective (Ochman and Davalos, 2006). It has been argued that introduction of new genes, which most likely will not fit well into the existing regulatory schemes of the cell, especially if they were involved in a complex network (Jain *et al.*, 1999), will cause a drop of the reproduction rate. And even a small reduction in those terms, combined with strong competition in huge prokaryotic populations, will be severely punished by the natural selection. HGT might have been extremely important for turning points in natural history, which happen in special conditions, but frequency of its occurrence on a 'daily basis' might be overestimated due to inaccuracy of mathematical procedures used for genomic sequence analysis (Kurland *et al.*, 2003).

4.1.2 Aim of model including HGT

As the controversy surrounds only the significance of HGT in evolution, not its existence or whether it affects evolution at all, it becomes obvious that horizontal gene transfer has to be considered a force shaping the size of the genome and its investigation becomes unavoidable. On the one hand, it should increase the size of the genome by generating constant inflow of genes to the cell. But, on the other hand, genes circulating in a metagenome of a size of the whole community should allow species to delete genes when they are unnecessary and retrieve them from somewhere, e.g. sister species, when they become beneficial once again.

Furthermore, it might be expected that the evolution will work more 'effectively'. If a gene giving optimum properties emerges somewhere in the population, it will spread across the multi-species community faster than if it had to be invented *de novo* in each independent species (of course, it will not be the very same gene then, but a number of analogous genes possessing similar properties). This might also decrease the frequency and severity of population crashes.

If a beneficial gene is easier to acquire and, as a result, cells can have fewer genes, then it might be expected that the population will be able to survive under tighter constrains on single gene properties. If it is easier to replace a gene, then maybe the gene can cover a tighter surface of the environmental space: parameter α ('gene width', see Table 2.1) can be smaller. And also, if fewer genes are required, then they may be more costly: parameter γ (cost of having a gene, see Table 2.1) can be higher.

To sum up, the first question is whether HGT will decrease or increase the genome size? The second question is how HGT will affect the speed of evolution and, consequently, the dynamics of population crashes? And finally, will the HGT mechanism allow for a tighter constraint on the gene properties?

4.2 Design of the model

All the basic properties of the version of the model with HGT are the same as those described in Chapter 2 (see section 2.2.1) but with two parameters facilitating HGT, instead of just one, as we could expect. The simplest way to implement HGT is to have one parameter which would randomly choose a gene in a population and move it to a randomly selected recipient cell. That means repeatedly browsing though all the genes in the ecosystem in each time step and that is computationally intense. To speed-up computation HGT was split in two steps:

1. h_c – first step is two cells 'meeting' each other. A cell is randomly chosen to become a donor with the probability of h_c (horizontal gene transfer, cell level). When a donor is chosen, then a recipient cell is picked out of the remaining population also at random with the probability of $1/(N_t - 1)$, where N_t is the size of the population at time t.
2. h_g – the second step is genes being transferred from the donor to the recipient cell. Each gene has a probability of (horizontal gene transfer, genome level) to be copied to a recipient cell. Note: copied, not moved. The donor, of course, is keeping it, too.

Two things need to be highlighted here: the first parameter h_c indicates how widespread HGT is across the population or, to put differently, what fraction of the population donates its genes to other cells. If it would be e.g. $h_c = 0.05$ in the population of 1000 cells, then there would be on average 50 HGT events per time step. The second parameter h_g shows what fraction of the donor's genotype is being copied to a recipient cell. If e.g. $h_g = 0.1$, then on average 10% of the donor's genotype gets transferred. Also, the algorithm makes the more abundant genes spread even more, because the more frequent a clonal strain is, the more chances it has to get picked by the random mechanism for donor selection.

In this model, genes are transferred only one way. The model has no spatial dimensions, being one big homogeneously mixed population. As a result, the proposed mechanism can be considered as a representation of any form of an HGT mechanism, e.g. bacterial conjugation, transfer via viruses or different kinds of mobile genetic elements like plasmids. Bacterial conjugation is the most known way in which bacteria exchange

DNA, but it applies to only a limited number of taxa. A given bacterium has to have a plasmid, called the F-plasmid, which codes for the formation of a structure called *pilus* which forms a bridge with another cell (the other cell does not have to have the F-plasmid). This bridge is then used to transfer DNA (Holmes and Jobling, 1996). The GTA mechanism (Lang and Beatty, 2007), mentioned already above, it seems to be the most similar to the scheme proposed in this model.

In the program implementing the model, all the HGT procedures take place after the reproduction procedures and after the random selection of cells for elimination by random death (see Figure 2.2).

4.3 Results

4.3.1 HGT as gene bombardment

As it was already said in Chapter 3, the model appears to impose strong constraints on the number of genes per cell (see section 3.3.2). Adding HGT to the model means a second, after gene duplication & neofunctionalisation, source of new genes in the genomes. This is compensated by a larger number of deletions happening in the simulations (Figure 4.1, panels C and D). HGT parameters were set to $h_c = 0.002$ and $h_g = 0.1$, thus the probabil-

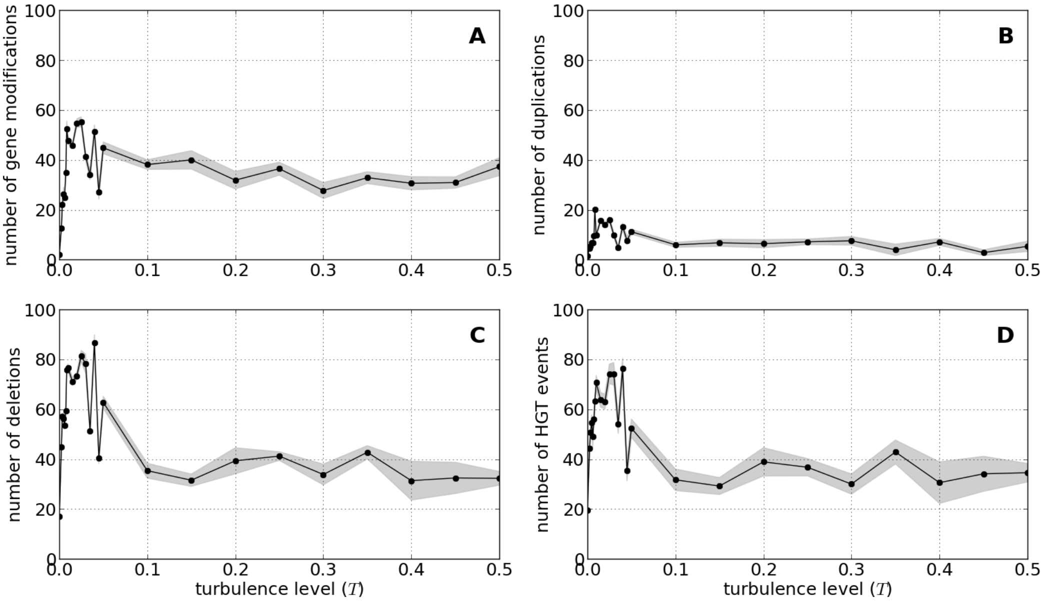

Figure 4.1: Averaged number of mutations of all three types compared to the number of HGT events in the last 10^5 time steps per clonal strain given for different turbulence levels T. HGT parameters were set to: $h_c = 0.002$ and $h_g = 0.1$. Other parameters were as given in Table 2.1.

ity of an HGT event h_c is equal to that of any of the mutations set to $\mu_{mod,dupl,del} = 0.002$. It can be observed that the selection process elevates the deletion rate beyond the basic level only to compensate for an uncontrolled growth of the genome size from two sources: gene duplication and HGT. Gene modification rate is also quite high, but it is consistent with a previously shown pattern that evolution in the model will modify the existing gene rather than create a new one (Figure 3.12). Also, the number of gene modification is proportional to the number of all genes circulating in the system.

After increasing the probability of HGT events h_c (h_g is kept constant at the level of 0.1), it can be noticed that the mean number of genes in an average genotype also increases (Figure 4.2, panel B), but the grand mean ratio of the surface under the genotype to total environmental space does not rise (Figure 4.2, panel A). Nearly all the genotypes cover the same fraction of the environment. That means the extra genes bring no benefit to their new hosts, as they do not expand their potential of harvesting resources towards new values of the environmental space. The new horizontally transferred genes are nothing but cost.

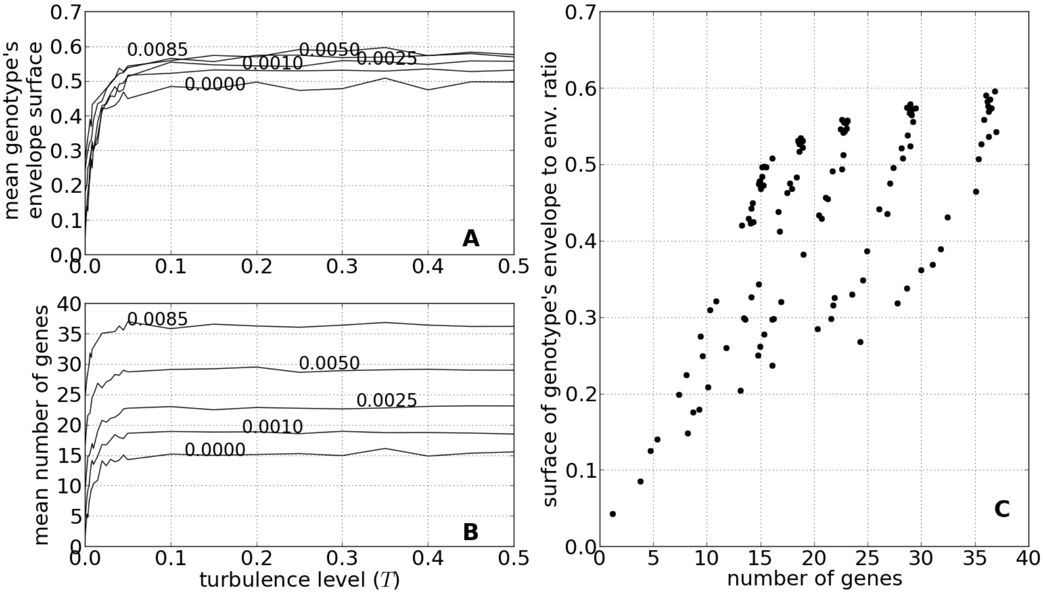

Figure 4.2: Model's sensitivity to change in the probability of HGT event h_c. Numbers above curves in panel A (ratio of the surface under genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are the respective values of h_c. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs.All model runs were set to parameter values as given in Table 2.1. Parameter h_g was set to 0.1.

When HGT is high, then the span of genome sizes is also wider. For example, if the turbulence level is set to $T = 0.25$ for $h_c = 0.001$, the system has on average 17.8 ± 2.0

genes, while for $h_c = 0.0085$ it has $33.8{\pm}4.1$ genes (mean ± SD). In a stable environment ($T = 0$), this dependency is also visible: 2.5 ± 1.2 genes on average for $h_c = 0.001$ and $17.1{\pm}3.3$ for $h_c = 0.0085$ (mean ± SD). Genome size and the span of genome sizes in the population grow proportionally to the rise in h_c (Figure 4.3). It seems that the system has problems with reaching the tight equilibrium discussed in Chapter 3 (see section 3.3.2).

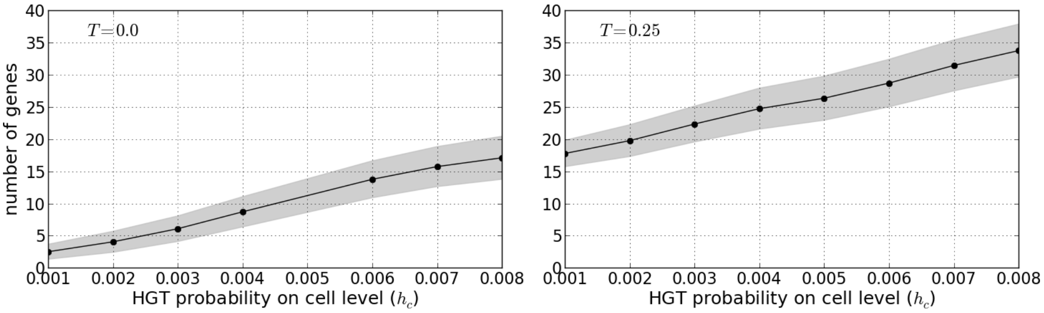

Figure 4.3: Mean genome size and span of genome sizes grows with rising HGT probability h_c. Dots represent the mean genome size, calculated after the gene numbers stabilised, in simulations with turbulence set to $T = 0$ (left panel) and $T = 0.25$ (right panel) with different values of h_c; grey area denotes SD. All model runs were set to parameter values as given in Table 2.1. Parameter h_g was set to 0.1.

4.3.2 Genetic diversity of the population

The Shannon index was used to estimate the genetic diversity of the population. This measure uses the equation identical to eq. 2.16 used previously to estimate population biodiversity (see section 2.2.3), only instead of cells and clonal lineages, the individual genes and their abundance within the population were considered as the key property

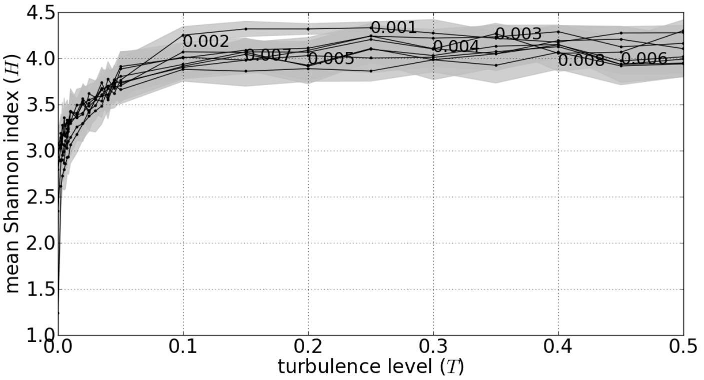

Figure 4.4: Dependence of genetic diversity (as measured with the Shannon index) on the frequency of HGT events h_c. Numbers above curves are the respective values of the probability of HGT event h_c. Dots represent mean values of the Shannon index for a given h_c value and turbulence level T calculated after the gene numbers stabilised. The grey areas are SDs. The mean and SD were calculated over the last 10^5 time steps. All model runs were set to parameter values as given in Table 2.1 except the HGT parameter h_c. Parameter h_g was kept constant $h_g = 0.05$.

of the system (Figure 4.4). It can be firmly concluded that HGT does not change the genetic diversity of the population. This means that there is no linkage between HGT and the selection pressure. Genes just randomly jump between genomes in all directions, generating only evolutionary noise, and not adding anything to the population's adaptive potential.

4.3.3 HGT does not permit the population to have tighter genes

When designing HGT in the model, it was first thought that HGT will replace the mutation & selection process in the process of inventing new genes. It was also hoped that HGT will facilitate the spread of genes already trimmed to the environmental conditions and it was assumed that this mechanism will allow the population to survive with genes which are narrower (have lower α). It was started with a value of $\alpha = 0.08$, which is the standard value used in the simulations in the non-HGT model from Chapter 2, then values were lowered eventually reaching $\alpha = 0.04$, under which in the non-HGT model some of the simulations ended with total population extinction (Figure 4.5). All runs with $\alpha \leq 0.03$ failed to be completed (the population went extinct). Comparison of these results with

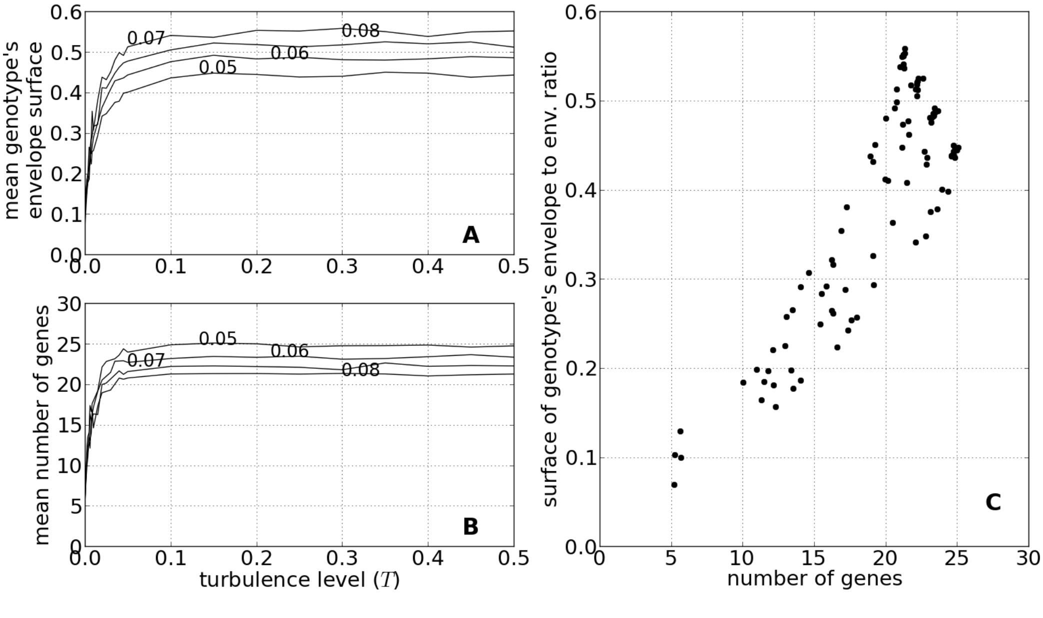

Figure 4.5: Model's sensitivity to change in the surface under the Gaussian representation of a gene with HGT. Numbers above curves in panel A (ratio of surface under genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are the respective values of the gene width parameter α. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1, except the gene width α. HGT parameters were set to: $h_c = 0.0025$ and $h_g = 0.05$.

non-HGT simulations (see Figure 2.7) shows there is no enhancement of tolerance on smaller surfaces under the Gaussian curve.

4.3.4 HGT does not permit for more expensive genes

It has also been tested whether allowing HGT will impact the gene cost parameter. HGT parameters were set to: $h_c = 0.0025$ and $h_g = 0.05$. No significant change was observed (Figure 4.6) in comparison to non-HGT simulations (Figure 2.8). The system was able to sustain itself with the maximum value of the gene cost of $\gamma = 0.008$. For $\gamma = 0.009$, most of high turbulent runs did not make it to the end of the simulation. This means that there is no difference from what the simulation showed when no HGT was involved.

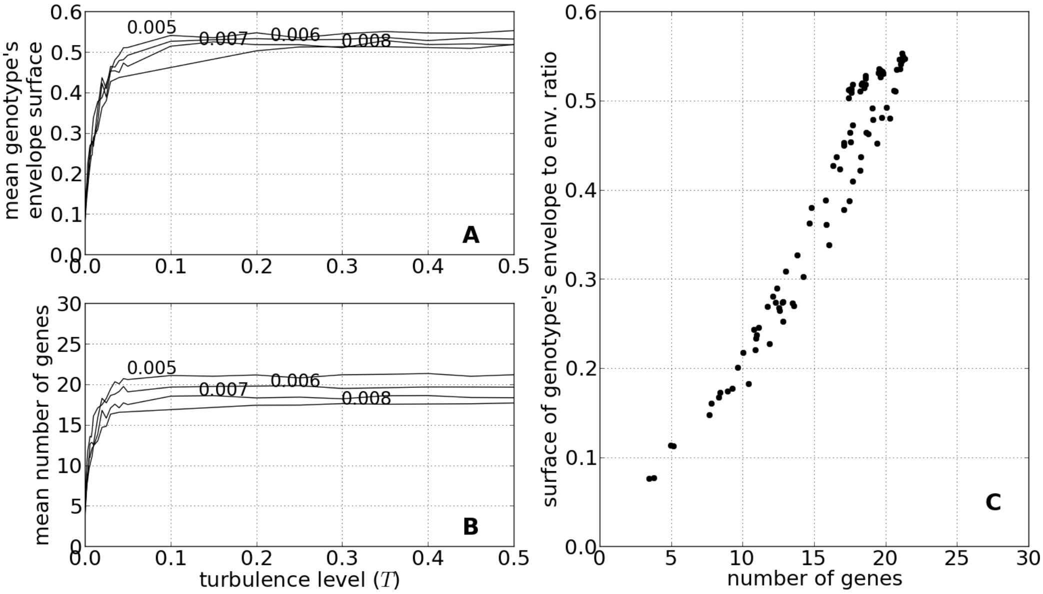

Figure 4.6: Model's sensitivity to change in the cost of maintenance of one gene with HGT allowed. Numbers above curves in panel A (ratio of surface under the genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are the respective values of the cost of maintenance of one gene parameter γ. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs. All model runs were set to parameter values as given in Table 2.1, except the cost of maintenance of one gene γ. HGT parameters were set to: $h_c = 0.0025$ and $h_g = 0.05$.

4.3.5 HGT does not decrease the frequency of population crashes

The presence of horizontal gene transfer does not decrease the probability of a population crash. If only those crashes which lead the population to dropping to 1000 or less individuals were taken into account (an ecosystem can sustain around 3200 with this amount of resource), we can see that the larger the value of h_c, the more frequent and more severe these crashes become (Table 4.1). Figure 4.7 presents population size plots of five model runs with different values of h_c. Turbulence level equal to $T = 0.02$ was chosen because previously this value generated runs with the highest frequency of population crashes (see Figure 3.11). The population with h_c set to 0.01 went extinct before the end of the simulation.

Table 4.1: Number of population crushes as dependent on the probability of horizontal gene transfer, cell level h_c. Turbulence level was set to $T = 0.02$ and HGT on the gene level to $h_g = 0.1$.

Horizontal gene transfer, cell level h_c	**Number of crashes reaching 1000 individuals or less.**
0	5
0.0005	6
0.001	6
0.002	11 (one of them with two peaks)
0.005	9 (two of them with two peaks)
0.0075	17 (some of them with multiple peaks)
0.0085	14 (many of them with multiple peaks)

4.3.6 HGT replaces duplication in delivering new genes for evolution to modify

Two runs were compared: one where HGT was not present and another where HGT was set to $h_c = 0.0025$ and $h_g = 0.05$, but where the gene duplication parameter was switched off ($\mu_{dupl} = 0$). The runs with HGT have more genes, e.g. for $T = 0.25$, the HGT-allowing and no-duplication run has 20.5 ± 2.5 genes on average, when a duplication-allowing and no-HGT simulation only has 15.0 ± 1.5 genes (mean ± SD). The elevated number in the HGT-based system might have accrued due to HGT parametrisation ($h_c = 0.0025$ for HGT-based systems, versus $\mu_{dupl} = 0.002$ for duplication-based systems) and because mutations are allowed when a cell reproduces, whereas the HGT procedure is launched at the end of each iteration. This disproportion produces effective mutation bias by making changes in the genome more probable per time step. In comparison, the no-duplication and no-HGT runs are very unstable with some of them even failing to become completed. For $T = 0.25$ this system showed 12.0 ± 1.0 genes (mean ± SD) and seemed to be struggling with the deletions happening more often then duplications (Figure 4.8). Random loss of genes cannot be compensated for by gene duplication. HGT works very much in the same manner as duplication had worked before introducing HGT into the

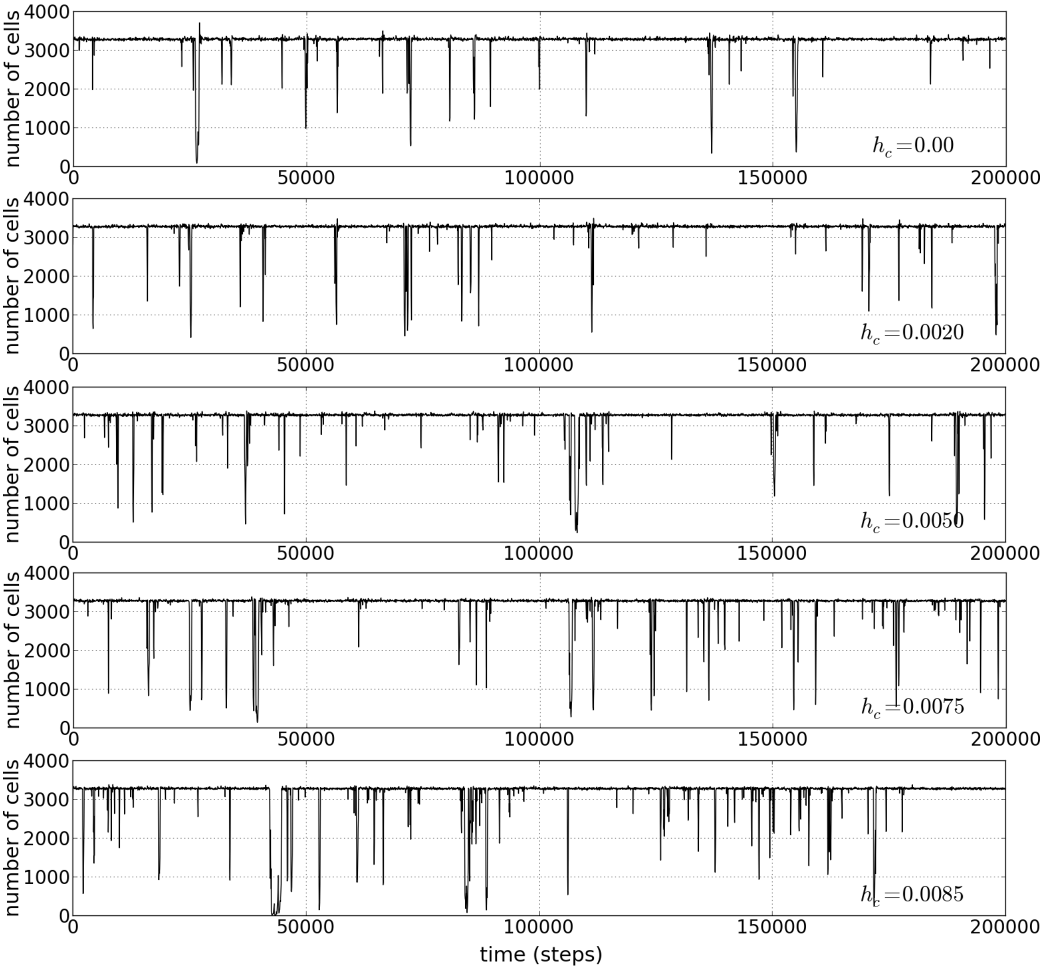

Figure 4.7: The presence of horizontal gene transfer does not decrease the probability of a population crash. Five runs with turbulence set to $T = 0.02$ and with different values of HGT probability (cell level): $h_c = \{0; 0.002; 0.005; 0.0075; 0.0085\}$ show different frequency of population crashes. All model runs were set to parameter values as given in Table 2.1, and HGT probability (gene level) was set to $h_g = 0.1$.

model (see Chapter 3).

4.4 Conclusions

4.4.1 HGT does not allow tolerance for less effective genes

Introduction of horizontal gene transfer does not allow the population to survive under tougher constraints imposed on parameters regarding the gene quality. This goes against the initial expectation that the ‘gap’ in genotypes created by tighter genes (with smaller α) will be filled by more genes per genome. And the system with HGT evolves more genes, but as we could see in Figure 4.2 they do not bring any new quality. They fall in the same regions of the environmental space x as genes already possessed by cells. As a

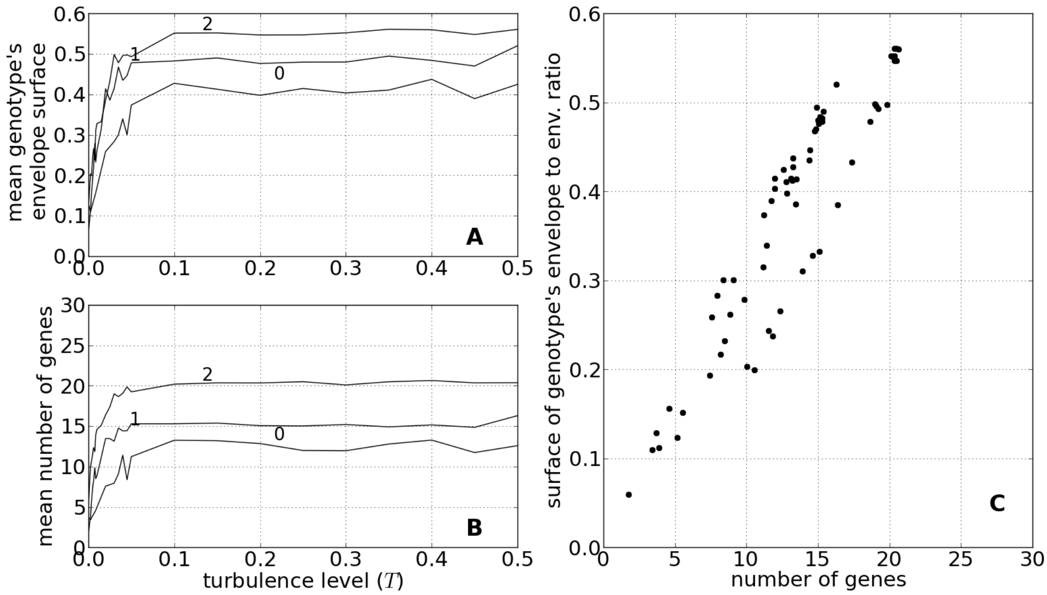

Figure 4.8: Model's sensitivity to distortions in proportions in rates of different mutations. Numbers above curves in panel A (ratio of surface the under genotype to total environment surface as function of turbulence level T) and panel B (mean number of genes as function of turbulence level T) are: 0 – no HGT and no duplication; 1 – no HGT, duplication set to $\mu_{dupl} = 0.002$; 2 - no duplication. In panel C each dot represents one model run. Uncertainty has been omitted for the sake of simplicity of the graphs.ll model runs were set to parameter values as given in Table 2.1. HGT parameters were set to: $h_c = 0.0025$ and $h_g = 0.05$.

result, they cannot compensate for any 'gaps', even if any gaps do occur. Furthermore, a larger number of genes means a higher costs of living. It turns out that the minimum gene width is very similar to the set-up with less genes flowing to the cells during the evolution process.

4.4.2 One-niche environment is too simple for population to benefit from HGT

Introduction of horizontal gene transfer does not speed up evolution, nor does it allow the population to survive under tougher constraints imposed on the parameters regarding the gene cost and shape. Also HGT does not increase the genetic diversity of the population. This is due to the environmental space being too small or not complex enough for the gene pool to get very divergent. This evolutionary system has a very high adaptive potential and also only one optimum at a time to search for, thus all the successful genotypes look alike, even if they do not share a common ancestry. This is similar to flying: no matter if it is a reptile, a bird, a mammal or a very big butterfly, the rules of physics simply force the wings to have a strictly defined geometry. And it is the same here: no matter what the genotype's original shape is, in the given environment eventually everyone has to have a fairly similar shape and distribution of genes.

In other words, because environmental space is so 'flat', horizontal gene transfer does not provide anything new. No genes emerge which a cell could not otherwise acquire by means of mutation. If it is not the same gene, than in most cases a cell has a different gene, but with a very similar profile (similar height, similar localisation of the maximum). And swapping genes around does not bring any benefit.

4.4.3 HGT is equivalent to gene duplication

The gene duplication mutation copies an existing gene and 'pastes' it in the very same place in the genome. Then a cell has two identical copies of the same gene, which generate a double cost but bring no extra benefit. This is a situation causing a slow-down of reproduction, which can be compensated only if in a short time after the duplication a modifying mutation arises, giving the duplicated gene a brand new function and making it beneficial (if the cell was lucky enough in this mutation lottery). As it was already said previously, the model's environmental space is 'flat', with only one optimum at a time, which makes it easy to explore by evolutionary mechanisms. Thus, the peak fitness is reached rather fast and the optimum genes rise quickly. Furthermore, that happens multiple times in different genomes. The population is full of analogous genes not related by a common origin.

Genes transferred by HGT come from the same homogeneous population and the more abundant a gene version is, the more likely it will become a subject of an HGT event. Thus, when transferring a gene between cells, we may expect that the recipient cell will already possess a gene of a similar profile or even the very same gene. Most likely, genes from HGT do not bring novel properties to the cell, but occupy positions in the environmental space already taken by the cell's own genes. And, as in the case of duplication, after an HGT event there are two genes doubling the costs, but with no extra benefit.

HGT, just like duplication, introduces new genes to genomes which later get changed by modifying mutation. Also, when there is no duplication, then HGT balances the deletion, preventing genomes from shrinking. But when existing along with duplication, HGT creates a bias toward expanding genomes. This bias has to be compensated for by an increased deletion rate. Systems which develop more genes also record more gene modification, because there are more genes to work on for modification forces, but the ratio of number of modifications per one gene is kept.

HGT can replace duplication as a way of introducing new genes to the genotype. This is a mechanism of rising new genes competing with the ‘duplicate & modify’ model. Instead, there is the ‘receive & modif’ mechanism.

4.5 Discussion

4.5.1 Are horizontally transferred genes really ”ready to use”?

In the literature, the following slogan can be encountered: “HGT provides genes which are working and are ready to use” (Koonin *et al.*, 2001; Ochman *et al.*, 2000; Treangen and Rocha, 2011). Yet that might be just a part of the truth. New genes entering the genome via HGT have to fit within the existing functional network of the recipient’s genes. In the model presented in this chapter, it could be seen that incoming genes often duplicated existing ones in their function (they cover similar area of the environmental space), thus interfering with the recipient genome by generating extra costs. This results in the reduction of the reproduction rate mentioned above (Kurland *et al.*, 2003). In a real system, the new gene often does not fit in perfectly into the recipient’s physiology, but has to be later tweaked to it by mutations. Those new genes meet just the ‘early pre-requirements’, rather than being immediately ready to work. We can imagine a scenario when a gene arrives in a genome, providing new functions by having an interesting active site performing a very beneficial reaction. But all the other characteristics (stability of the mRNA transcript, membrane binding site, size of the protein etc.) have to be tweaked. The model developed in this chapter cannot simulate that, because the environmental space has just one dimension with just one optimum. The space is too ‘flat’, as there are no secondary dimensions along which the ‘tweaking’ could happen.

4.5.2 Slow mutations and fast HGT

Why does the model treat HGT as a source of ‘junk DNA’ rather than, as in the case of real systems, an opportunity to gain beneficial genes? The reason for that is the speed at which mutations can generate new solutions in real living systems. Biological particles are huge and developing their appropriate composition and shape by means of trial and error, as evolution does, takes a very long time. On top of that, evolution is not able to make investments by temporarily decreasing species fitness for the sake of possible future profits. As a result, many possible intermediate stages of a molecule, which might lead to its optimum form in the future, are ‘unacceptable’ because they reduce fitness here and now. HGT, on the contrary, might offer solutions which might not be perfect, increasing fitness only slightly or even not at all, but which are already there. Unfortunately, the model presented above does not make that distinction and HGT and modifying mutation have the same potential of generating genes in every corner of the environmental space x.

Current mechanism of mutation by modification implemented in the model allows a gene to shift its maximum to any position within the range of the environmental space. A gene simply disappears from one place and pops out somewhere else with a uniform probability for every infinitesimal interval of the environmental space. Genes transferred by HGT appear in a genotype in a similar way: they appear out of nowhere and are pasted into the genotype. The only difference is that genes originating from an HGT event most likely were beneficial for the donor, whereas genes that arose as a result of a modifying mutation are rather useless in most cases. Yet the simplicity of the environment neutralises this difference.

A possible solution to this issue might be putting a limit on the speed of evolution by mutation, i.e. adding another parameter that would impose a constraint on how different a mutated gene can be in comparison with its forefather to the model. In this manner, HGT would become the only way to gain genes completely different from cells in the current genetic composition. HGT would be a shortcut alternative to the slow way mutations change the genome.

4.5.3 HGT and stability of environment

Charles G. Kurland and his colleagues have suggested that HGT becomes important during big evolutionary transition which are triggered by extraordinary shifts of the environment, whereas normally the classical Darwinian evolution dominates the process of forming new lineages (Kurland *et al.*, 2003). The simulation results presented here would comply with this statement.

When responding to changes happening along only one environmental trait, which can be seen as modification of the niche, HGT seems to be unnecessary as mutations do this job better. Yet complete reformulation of the niche was not attempted here. In the previous chapter, the problem of stability on different levels of the biosphere organisation has been lifted (see section 3.4.2). It was argued there that micro-scale stability of the niche impacts the number of genes in a prokaryotic genome. But that changes can be seen as just 'flickering' of a node in a biocenotic network. Then HGT is not necessary. Meanwhile, it may become important when the node is seriously shifted in the network or it has disappeared and the species needs to find a new spot in the biocenosis.

Another problem is where from can a population in need of a big genetic change 'borrow' new genes? The node cannot make a fast shift into the regions of abiotic environment unoccupied by any population because that will trigger extinction. At some point during the evolutionary history of a taxon, its node on the way to the new position in the network has to come close enough to another occupied node for long enough to be able to gain new genes.

Chapter 5

Summary: limitations and possible further developments of the model

Each model is a simplification of a fragment of the reality to which it refers. This simplification comes from our limited cognitive abilities, from limitations of our technology and from the fact that if a model would fully represent reality, then it would be as difficult to analyse as reality itself (Morgan and Morrison, 1999b). The same simplification allows us to trace problems to the very bottom without distraction from side factors which are not interesting for us at the moment. Also, it allows us to draw firm conclusions from what we have observed.

5.1 Model's limitations

Simplifications introduced in this model bring certain limitations to the types of problems which can be investigated with this framework. Some of these limitations were obvious from the moment the model's assumptions were formulated, others were discovered during development and evaluation of the model.

5.1.1 Model has extremely simplified idea of niche

G. Evelyn Hutchinson proposed a model of an ecological niche as volume in a multi-dimension space. Each dimension represents one trait of the environment and somewhere on the trait's axis a species has an optimum interval of values for its survival and reproduction. The sum of all the optimum intervals of all ecologically significant dimensions creates a volume which is the niche. It should not be possible for two species to occupy the same volume (Hutchinson, 1957). The main problem with application of this concept is that it is almost impossible to find out how many and what environmental traits are ecologically significant for a given species. Light intensity is an essential trait in the case of plants, but it is not that significant for fungi. Also some traits are difficult to define in this manner, e.g. nesting preferences of birds.

The model used a modified and simplified Hutchinson's niche concept, i.e. just one dimension was introduced. This one dimension, combined with the idea that the environmental conditions can have only one value at a time and this value can change at different rates ('flickering of the niche', already mentioned before), gave interesting results regarding how many genes a species needs to have in its niche, depending how stable the niche is. But simulating extensive genome rearrangements is rather beyond this framework's abilities. The model can simulate evolution of ecological lineages and genetic strain, but not the emergence of new species.

5.1.2 Niche has just one optimum

The one dimension of the niche is very easy to explore, especially that it has only one optimum at a time and there are no suboptimum peaks a population could explore and potentially get stuck upon. Selection pressure pushes the genetic profile of the population towards only one direction at a time. Heterogeneity of conditions is achieved by shifting the peak in time at divergent rates. The environmental space itself is one-dimensional and this dimension is 'flat'. Exploring this kind of space is fairly easy for a population having three possible types of mutations to modify its genotype. It seems that any fourth mechanism (e.g. horizontal gene transfer) is superfluous.

5.1.3 Model cannot simulate big evolutionary transition

Horizontal gene transfer is proven to be an important force driving the evolution of prokaryotes, but the model has failed to replicate its significance. Some papers underline the role of horizontal gene transfer in spreading protein families across different prokaryotic taxa (Treangen and Rocha, 2011), while other authors attribute the wide spread of the cyanobacteria-like photosystem among higher organisms to HGT as well (Yerrapragada *et al.*, 2009). These are major changes in physiology which gave the recipients a tool to penetrate new environments and triggered further rearrangement of their metabolism. This model is not able to simulate this kind of evolutionary leaps. This framework allows the population to freely adapt along just one trait of the niche and does not allow in any way to 'invent' new traits. Here, evolution can only lead to perfection of adaptation along one dimension of the niche to one specified point, which is a major limitation in the light of the published results. The model cannot investigate the big question of the role of HGT in the evolution of the biosphere as a way to expand it into new environments. Big transitions in evolution need a different approach.

5.1.4 It is not open-ended evolution system

Evidently, one can make an attempt to expand the environment into other dimensions and turn a gene into a multi-dimensional Gaussian surface over this space, with different properties along different axes. The dimensions of the niche could be divergent in their significance for survival and reproduction of the species. This expansion, very difficult to grasp and not much easier to analyse, might allow for investigating more complex problems: for instance, the environmental landscape might be complex enough to distinguish between gene duplication and HGT but it will not be a step towards open-ended evolution. An open-ended system would have to be able to add new traits to the species' environmental landscape and adapt the population along them. In this way, individuals at the end of the simulation would show a complexity and diversity greater than individuals at the beginning (Maley, 1999).

In my opinion, only an open-ended evolution system might be able to catch the essence of species-to-species transition resulting from an evolutionary process.

5.2 Possible further development

Besides the abovementioned constraints, the model also has a number of perspectives which were not investigated in this thesis due to time and resource constraints. Additionally, theoretical reflections suggest certain possible improvements and further developments.

5.2.1 Difference between gene duplication and HGT

It was shown that horizontal gene transfer actually acts in the model in the same manner as gene duplication. It was also argued that the possible main difference between HGT and mutation-based acquisition of genes with new functionalities lies in the average time necessary to gain a new function. It may be worth considering to introduce a new parameter which would slow down the rate of mutation-based evolution in the model, giving HGT a chance to show its potential.

This new parameter could be a constraint on how fast, in the evolutionary time scale, a gene can change. In the current model, when a gene is mutating, all its values can be changed with uniform probability within the permitted value range. In particular, the new value of the maximum of the Gaussian representation of a gene c (see Figure 2.1) could be any of the values in the range $[-1, +1]$ with equal probability. To make horizontal gene transfer a reasonable alternative to mutation by modification, a new value of the mean c' may be selected at random from a normal distribution having the mean at c (at the 'old' value) and variation equal to s_{mod} (the third new parameter):

$$c' \in \mathcal{N}(c, s_{mod}) \tag{5.1}$$

Where c and c' are the values of the mean of the Gaussian representation of a gene before and after mutation, respectively. Probability of occurrence of a mutation modifying a gene is still μ_{mod}.

The larger s_{mod} is, the wider the span of values of c' possible to get when a mutation is modifying a gene becomes. In other words, a gene can move faster to a new optimum value of the environmental conditions x when s_{mod} has a higher value. Of course, the environmental condition space is still limited to the $[-1, +1]$ interval, so if the value of s_{mod} is high enough, then the effective distribution of possible c' values within the permitted range is near to uniform distribution over that range.

5.2.2 Investigation of the gene deletion bias

Is has been suggested that the deletion bias is an important force responsible for streamlining of prokaryotic genomes (Mira *et al.*, 2001; Lynch, 2006a). In the current parametrisation of the model all three types of mutations are set at the same level (see Table 2.1). It has been discussed previously that when a gene mutates, its new values are set at random, whereas HGT moves genes which have probably proven their usefulness in a different genome 4.5.2). Can HGT be a life-savour for the population under elevated gene deletion frequency? It might be possible that deletions will remove beneficial genes at a fast rate while re-inventing similar ones via mutations will take too long. Getting new ones from HGT might be faster.

5.2.3 Two populations occupying different fractions of the environmental space

Horizontal gene transfer has been shown to be redundant in a single homogeneous population. Literature has always underlined that it is a cross-species gene exchange (Syvanen, 1985). Thus, the model would have to somehow diverge the population into two subpopulations.

The current model has one population per simulation, but the framework allows for more. The environmental condition x is randomly selected from an arbitrarily given interval $[-1, +1]$. This interval might be either split into two halves or expanded and then divided, thus creating two niches, each occupied by a different population. If cell exchange between these two populations will be permitted, along with HGT, a certain number of genes will be flowing between ecosystems. The two environmental condition intervals might overlap in some fraction making the exchange of genes possibly more frequent.

The model showed that the genome, when having enough time, will tightly fill this simple environmental space. With two ecosystems with niche parameters changing in time, it might be worth investigating: (1) if the HGT events become more frequent when these niches become similar to each other at some time interval (when the intervals of permitted x values are overlapping); (2) how long a gene from HGT lasts in the population; (3) if a lineage moving to a new environment will acquire necessary genes via HGT or rather via mutations.

5.2.4 Disequilibriums in resource flow

The model is characterised by perfectly efficient resource recycling. Each time a cell dies, its resources are returned to the environmental pool of free resources. But it is not the case in nature, e.g. in marine environments, there is always a fraction of cells which sink to the sediment (Lenton and Klausmeier, 2006) nd this gap in resource supply must be filled from external sources. Also, there are environments with a turbulent flow of resources, e.g. upwelling zones (Ishizaka *et al.*, 1987).

It has been shown in the model that resource supply is an important factor shaping the size of the genome. In the simulations, this was happening via the random death factor γ. Random death influenced how much free resources were available in the environment, thus what the expected length of the starvation period a cell had to face was. That period has a direct impact on the genome size (see section 3.3.4). By changing the sinking rate of cells or by altering the resource supply in different time steps, one can investigate how a turbulent resource supply will change the size of the genomes.

5.2.5 Anabiosis

Formation of the resting stages and other forms of anabiosis of microbes has been discussed as a possible way of overcoming periods of harsh environmental conditions (see section 3.4.4). This option seems to be interesting and not difficult to investigate.

5.3 Summary

Despite certain limitations, the modelling approach shown here proved its usefulness in the study of genome size constraints . Results show a linkage between the reasons for size limitation always mentioned separately in earlier literature, i.e. resource availability (Hessen *et al.*, 2010; Lane and Martin, 2010) and the limitation by power-law growth of the number of regulatory genes (so called ‘bureaucracy burden’) (van Nimwegen, 2003; Koonin and Wolf, 2008; Molina and van Nimwegen, 2009).

An interesting conclusion is that not the most turbulent environment have the highest rate of evolution and the highest frequency of population crashes, but the environments which are only mildly turbulent. They create an ‘illusion’ of stability in a short run (time frame of the individual's life span), trimming genomes' sizes, but later they shift to a new state, triggering harsh times for the species. This result shows that each niche stability has to be measured in the time scale of the species which occupies it.

Horizontal gene transfer has been found to be not only a way of accelerating evolution, as literature shows it, but also a possible threat to the genome's stability. Numerous mechanisms of removing foreign DNA from the chromosome might not only be a way of removing viral or parasitic DNA, but they could also come in useful against any 'unwanted' genes.

References

Alberch, P. (1991), From genes to phenotype: dynamical systems and evolvability, *Genetica*, *84*(1), 5–11.

Alon, U. (2007), *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC, Boca Raton.

Anderson, A. W., H. C. Nordan, R. F. Cain, G. Parrish, and D. Duggan (1956), Studies on a radio-resistant micrococcus. I. Isolation, morphology, cultural characteristics, and resistance to gamma radiation, *Food Technology*, *10*(12), 575–578.

Andersson, G. E., O. Karlberg, B. Canbäck, and C. G. Kurland (2003), On the origin of mitochondria: A genomics perspective, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1429), 165–179.

Andersson, S. G., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Pontén, U. C. Alsmark, R. M. Podowski, A. K. Näslund, A. S. Eriksson, H. H. Winkler, and C. G. Kurland (1998), The genome sequence of Rickettsia prowazekii and the origin of mitochondria, *Nature*, *396*(6707), 133–140.

Barrick, J. E., D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim (2009), Genome evolution and adaptation in a long-term experiment with Escherichia coli, *Nature*, *461*(7268), 1243–1247.

Beeby, M., B. D. O'Connor, C. Ryttersgaard, D. R. Boutz, L. J. Perry, and T. O. Yeates (2005), The Genomics of disulfide bonding and protein stabilization in thermophiles, *PLoS Biology*, *3*(9), e309.

Bentley, S. D., K. F. Chater, A.-M. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.-H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.-A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood (2002), Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2), *Nature*, *417*(6885), 141–147.

Berget, S. M., C. Moore, and P. A. Sharp (1977), Spliced segments at the 5 terminus of adenovirus 2 late mRNA, *Proceedings of the National Academy of Sciences of the United States of America*, *74*(8), 3171–3175.

Biers, E. J., K. Wang, C. Pennington, R. Belas, F. Chen, and M. A. Moran (2008), Occurrence and expression of gene transfer agent genes in marine bacterioplankton, *Applied and Environmental Microbiology*, *74*(10), 2933–2939.

Bipatnath, M., P. P. Dennis, and H. Bremer (1998), Initiation and velocity of chromosome replication in Escherichia coli B/r and K-12, *Journal of Bacteriology*, *180*(2), 265–273.

Bird, A. P. (1995), Gene number, noise reduction and biological complexity, *Trends in Genetics*, *11*(3), 94–100.

Bisset, K. A., and F. W. Moore (1952), *Bacteria*, E. & S. Livingstone Ltd., Edinburgh & London.

Blount, Z. D., J. E. Barrick, C. J. Davidson, and R. E. Lenski (2012), Genomic analysis of a key innovation in an experimental Escherichia coli population, *Nature*, *489*(7417), 513–518.

Bonner, J. T. (1988), *The Evolution of Complexity: By Means of Natural Selection*, Princeton University Press, Princeton.

Boue, S., I. Letunic, and P. Bork (2003), Alternative splicing and evolution, *BioEssays*, *25*(11), 1031–1034.

Brenner, S., J. H. Miller, and W. J. Broughton (2002), *Encyclopedia of Genetics, Vol. 2*, Academic Press, San Diego.

Brown, T. (2006), *Genomes 3*, 3rd edn., Garland Science, New York.

Buts, L., J. Lah, M.-H. Dao-Thi, L. Wyns, and R. Loris (2005), Toxin-antitoxin modules as bacterial metabolic stress managers, *Trends in Biochemical Sciences*, *30*(12), 672–679.

Caballero, A. (1994), Developments in the prediction of effective population size, *Heredity*, *73*(6), 657–679.

Carlin, F., H. Girardin, M. W. Peck, S. C. Stringer, G. C. Barker, A. Martinez, A. Fernandez, P. Fernandez, W. M. Waites, S. Movahedi, F. v. Leusden, M. Nauta, R. Moezelaar, M. D. Torre, and S. Litman (2000), Research on factors allowing a risk assessment of spore-forming pathogenic bacteria in cooked chilled foods containing vegetables: A FAIR collaborative project, *International Journal of Food Microbiology*, *60*(2–3), 117–135.

Casjens, S. (1998), The diverse and dynamic structure of bacterial genomes, *Annual Review of Genetics*, *32*(1), 339–377.

Chagin, V. O., J. H. Stear, and M. C. Cardoso (2010), Organization of DNA replication, *Cold Spring Harbor Perspectives in Biology*, *2*(4).

Chan, S. R. W. L., and E. H. Blackburn (2004), Telomeres and telomerase, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1441), 109–121.

Charlesworth, B., and N. Barton (2004), Genome size: Does bigger mean worse?, *Current Biology*, *14*(6), R233–R235.

Chen, X., and J. E. Cohen (2001), Transient dynamics and food-web complexity in the Lotka-Volterra cascade model., *Proceedings of the Royal Society B: Biological Sciences*, *268*(1469), 869–877.

Chouard, T. (2008), Darwin 200: Beneath the surface, *Nature News*, *456*(7220), 300–303.

Chow, L. T., R. E. Gelinas, T. R. Broker, and R. J. Roberts (1977), An amazing sequence arrangement at the 5 ends of adenovirus 2 messenger RNA, *Cell*, *12*(1), 1–8.

Chow, S. S., C. O. Wilke, C. Ofria, R. E. Lenski, and C. Adami (2004), Adaptive radiation from resource competition in digital organisms, *Science*, *305*(5680), 84 –86.

Ciliberti, S., O. C. Martin, and A. Wagner (2007), Innovation and robustness in complex regulatory gene networks, *Proceedings of the National Academy of Sciences of the United States of America*, *104*(34), 13,591–13,596.

Cooper, S., and C. E. Helmstetter (1968), Chromosome replication and the division cycle of Escherichia coli B/r, *Journal of Molecular Biology*, *31*(3), 519–540.

Cover, T. M., and J. A. Thomas (2006), *Elements of Information Theory. Second Edition*, John Wiley & Sons, Hoboken, New Jersey.

Crombach, A., and P. Hogeweg (2008), Evolution of evolvability in gene regulatory networks, *PLoS Computational Biology*, *4*(7), e1000,112.

Crombach, A., and P. Hogeweg (2009), Evolution of resource cycling in ecosystems and individuals, *BMC Evolutionary Biology*, *9*(1), 122.

Cushing, B. Dennis, Desharnais, and Costantino (1998), Moving toward an unstable equilibrium: saddle nodes in population systems, *Journal of Animal Ecology*, *67*(2), 298–306.

Daley, J. M., P. L. Palmbos, D. Wu, and T. E. Wilson (2005), Nonhomologous end joining in yeast, *Annual Review of Genetics*, *39*(1), 431–451.

Daly, J. W., H. M. Garraffo, T. F. Spande, V. C. Clark, J. Ma, H. Ziffer, and J. F. Cover (2003), Evidence for an enantioselective pumiliotoxin 7-hydroxylase in dendrobatid poison frogs of the genus Dendrobates, *Proceedings of the National Academy of Sciences of the United States of America*, *100*(19), 11,092–11,097.

Davies, K. F., P. Chesson, S. Harrison, B. D. Inouye, B. A. Melbourne, and K. J. Rice (2005), Spatial heterogeneity explains the scale dependence of the native – exotic diversity relationship, *Ecology*, *86*(6), 1602–1610.

Dawkins, R. (1976), *The Selfish Gene*, Oxford University Press, Oxford.

Dawkins, R. (1982), *The Extended Phenotype*, Oxford University Press, Oxford.

Descamps-Julien, B., and A. Gonzalez (2005), Stable coexistence in a fluctuating environment: An experimental demonstration, *Ecology*, *86*(10), 2815–2824.

Diamond, J. (1972), *Avifauna of the Eastern Highlands of New Guinea*, 1St edition edn., Harvard Univ Nuttall Ornithological, Cambridge, Massachusetts, USA.

Drake, J. W. (1991), A constant rate of spontaneous mutation in DNA-based microbes., *Proceedings of the National Academy of Sciences of the United States of America*, *88*(16), 7160–7164.

Drake, J. W. (1999), The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes, *Annals of the New York Academy of Sciences*, *870*(1), 100–107.

Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow (1998), Rates of spontaneous mutation, *Genetics*, *148*(4), 1667.

Dufresne, A., L. Garczarek, and F. Partensky (2005), Accelerated evolution associated with genome reduction in a free-living prokaryote, *Genome Biology*, *6*(2), R14.

Earl, D. J., and M. W. Deem (2004), Evolvability is a selectable trait, *Proceedings of the National Academy of Sciences of the United States of America*, *101*(32), 11,531–11,536.

Elser, J. J., K. Acharya, M. Kyle, J. Cotner, W. Makino, T. Markow, T. Watts, S. Hobbie, W. Fagan, J. Schade, J. Hood, and R. W. Sterner (2003), Growth rate–stoichiometry couplings in diverse biota, *Ecology Letters*, *6*(10), 936–943.

Elser, J. J., D. R. Dobberfuhl, N. A. MacKay, and J. H. Schampel (1996), Organism size, life history, and N:P stoichiometry, *BioScience*, *46*(9), 674.

Errington, J. (2003), Regulation of endospore formation in Bacillus subtilis, *Nature Reviews. Microbiology*, *1*(2), 117–126.

Felsenstein, J. (1974), The evolutionary advantage of recombination, *Genetics*, *78*(2), 737–756.

Fiers, W., R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert (1976), Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene, *Nature*, *260*(5551), 500–507.

Fisher, S. R. A. (1949), *The Theory of Inbreeding*, Oliver and Boyd, Edinburgh.

Fort, H., M. Scheffer, and E. H. v. Nes (2009), The paradox of the clumps mathematically explained, *Theoretical Ecology*, *2*(3), 171–176.

Forterre, P. (2002), A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein, *Trends in Genetics*, *18*(5), 236–237.

Foryś, U. (2005), *Matematyka w biologii*, 1st edn., Wydawnictwo Naukowo-Techniczne, Warszawa.

Frankham, R. (1995), Effective population size/adult population size ratios in wildlife: A review, *Genetics Research*, *66*(02), 95–107.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J.-F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. van Vugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. O. Smith, and J. C. Venter (1997), Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi, *Nature*, *390*(6660), 580–586.

Galperin, M. Y. (2005), A census of membrane-bound and intracellular signal transduction proteins in bacteria: Bacterial IQ, extroverts and introverts, *BMC Microbiology*, *5*, 35–35.

Gause, G. (1932), Experimental studies on the struggle for existence, *Journal of Experimental Biology*, *9*(4), 389–402.

Gerdes, K., S. K. Christensen, and A. Løbner-Olesen (2005), Prokaryotic toxin-antitoxin stress response loci, *Nature Reviews. Microbiology*, *3*(5), 371–382.

Gerstein, M. B., C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder (2007), What is a gene, post-ENCODE? History and updated definition, *Genome Research*, *17*(6), 669–681.

Gilbert, W. (1978), Why genes in pieces?, *Nature*, *271*(5645), 501–501.

Gilbert, W. (1987), The exon theory of genes, *Cold Spring Harbor Symposia on Quantitative Biology*, *52*, 901–905.

Giovannoni, S. J., H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappé, J. M. Short, J. C. Carrington, and E. J. Mathur (2005), Genome streamlining in a cosmopolitan oceanic bacterium, *Science*, *309*(5738), 1242–1245.

Glass, J. I., E. J. Lefkowitz, J. S. Glass, C. R. Heiner, E. Y. Chen, and G. H. Cassell (2000), The complete sequence of the mucosal pathogen Ureaplasma urealyticum, *Nature*, *407*(6805), 757–762.

Goodner, B., G. Hinkle, S. Gattung, N. Miller, M. Blanchard, B. Qurollo, B. S. Goldman, Y. Cao, M. Askenazi, C. Halling, L. Mullin, K. Houmiel, J. Gordon, M. Vaudin, O. Iartchouk, A. Epp, F. Liu, C. Wollam, M. Allinger, D. Doughty, C. Scott, C. Lappas, B. Markelz, C. Flanagan, C. Crowell, J. Gurson, C. Lomo, C. Sear, G. Strub, C. Cielo, and S. Slater (2001), Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58, *Science*, *294*(5550), 2323–2328.

Graveley, B. R. (2001), Alternative splicing: increasing diversity in the proteomic world, *Trends in Genetics*, *17*(2), 100–107.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. d. l. Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Pääbo (2010), A Draft Sequence of the Neandertal Genome, *Science*, *328*(5979), 710–722.

Gregory, T. R., and P. D. N. Hebert (1999), The modulation of DNA content: Proximate causes and ultimate consequences, *Genome Research*, *9*(4), 317–324.

Grimm, V. (1999), Ten years of individual-based modelling in ecology: What have we learned and what could we learn in the future?, *Ecological Modelling*, *115*(2–3), 129–148.

Grimm, V., T. Wyszomirski, D. Aikman, and J. Uchmański (1999), Individual-based modelling and ecological theory: Synthesis of a workshop, *Ecological Modelling*, *115*(2–3), 275–282.

Harold, F. M. (1986), *The Vital Force: A Study of Bioenergetics*, WH Freeman New York.

Hastings, A. (2004), Transients: the key to long-term ecological understanding?, *Trends in Ecology & Evolution*, *19*(1), 39–45.

Hastings, A., and K. Higgins (1994), Persistence of transients in spatially structured ecological models, *Science*, *263*(5150), 1133–1136.

Hawkins, B. A., and M. Holyoak (1998), Transcontinental Crashes of Insect Populations?, *The American Naturalist*, *152*(3), 480–484.

He, Y. (2009), High cell density production of Deinococcus radiodurans under optimized conditions, *Journal of Industrial Microbiology & Biotechnology*, *36*(4), 539–546.

Hespenheide, H. A. (1971), Food preference and the extent of overlap in some insectivorous birds, with special reference to the Tyrannidae, *Ibis*, *113*(1), 59–72.

Hessen, D. O., P. D. Jeyasingh, M. Neiman, and L. J. Weider (2010), Genome streamlining and the elemental costs of growth, *Trends in Ecology & Evolution*, *25*(2), 75–80.

Hindré, T., C. Knibbe, G. Beslon, and D. Schneider (2012), New insights into bacterial adaptation through in vivo and in silico experimental evolution, *Nature Reviews. Microbiology*, *10*(5), 352–365.

Hiriyanna, K. T., and T. Ramakrishnan (1986), Deoxyribonucleic acid replication time in Mycobacterium tuberculosis H37 Rv, *Archives of Microbiology*, *144*(2), 105–109.

Holmes, R. K., and M. G. Jobling (1996), Genetics, in: *Medical Microbiology* (Baron, S., ed.), 4th edn., University of Texas Medical Branch at Galveston, Galveston (TX).

Horiike, T., D. Miyata, K. Hamada, S. Saruhashi, T. Shinozawa, S. Kumar, R. Chakraborty, T. Komiyama, and Y. Tateno (2009), Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs, *Gene*, *429*(1–2), 59–64.

Hubbell, S. P. (2001), *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press.

Huisman, J., and F. J. Weissing (1999), Biodiversity of plankton by species oscillations and chaos, *Nature*, *402*(6760), 407–410.

Hutchinson, G. E. (1957), Concluding remarks, *Cold Spring Harbor Symposia on Quantitative Biology*, *22*(2), 415–427.

Hutchinson, G. E. (1961), The paradox of the plankton, *The American Naturalist*, *95*(882), 137–145.

Isalan, M., C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao, E. Raineri, M. Garriga-Canut, and L. Serrano (2008), Evolvability and hierarchy in rewired bacterial gene networks, *Nature*, *452*(7189), 840–845.

Isambert, H., and R. Stein (2009), On the need for widespread horizontal gene transfers under genome size constraint, *Biology Direct*, *4*(1), 28.

Ishizaka, J., M. Kaichi, and M. Takahashi (1987), Resting spore formation of Leptocylindrus danicus (Bacillariophyceae) during short time-scale upwelling and its significance as predicted by a simple model, *Ecological Research*, *2*(3), 229–242.

Jackson, D. A., and A. Pombo (1998), Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells, *The Journal of Cell Biology*, *140*(6), 1285–1295.

Jaenisch, R., and A. Bird (2003), Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals, *Nature Genetics*, *33*, 245–254.

Jain, R., M. C. Rivera, and J. A. Lake (1999), Horizontal gene transfer among genomes: The complexity hypothesis, *Proceedings of the National Academy of Sciences of the United States of America*, *96*(7), 3801–3806.

Jenner, L. B., A. Ben-Shem, N. Demeshkina, M. Yusupov, and G. Yusupova (2013), X-ray analysis of prokaryotic and eukaryotic ribosomes, in: *Biophysical approaches to translational control of gene expression* (Dinman, J. D., ed.), no. 1 in Biophysics for the Life Sciences, pp. 1–25, Springer New York.

Johnson, J. M., J. Castle, P. Garrett-Engele, Z. Kan, P. M. Loerch, C. D. Armour, R. Santos, E. E. Schadt, R. Stoughton, and D. D. Shoemaker (2003), Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays, *Science*, *302*(5653), 2141–2144.

Joset, F., J. Guespin-Michel, and L. O. Butler (1993), *Prokaryotic Genetics: Genome Organization, Transfer and Plasticity*, Blackwell.

Kashtan, N., and U. Alon (2005), Spontaneous evolution of modularity and network motifs, *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13,773–13,778.

Kashtan, N., M. Parter, E. Dekel, A. E. Mayo, and U. Alon (2009), Extinctions in heterogeneous environments and the evolution of modularity, *Evolution*, *63*(8), 1964–1975.

Kimura, M. (1962), On the probability of fixation of mutant genes in a population, *Genetics*, *47*(6), 713–719.

Kimura, M. (1967), On the evolutionary adjustment of spontaneous mutation rates, *Genetics Research*, *9*(01), 23–34.

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

Kiørboe, T., K. Tang, H.-P. Grossart, and H. Ploug (2003), Dynamics of microbial communities on marine snow aggregates: Colonization, growth, detachment, and grazing mortality of attached bacteria, *Applied and Environmental Microbiology*, *69*(6), 3036–3047.

Kirschner, M., and J. Gerhart (1998), Evolvability, *Proceedings of the National Academy of Sciences of the United States of America*, *95*(15), 8420–8427.

Knight, R. D., S. J. Freeland, and L. F. Landweber (2001a), Rewiring the keyboard: evolvability of the genetic code, *Nature Reviews Genetics*, *2*(1), 49–58.

Knight, R. D., L. F. Landweber, and M. Yarus (2001b), How mitochondria redefine the code, *Journal of Molecular Evolution*, *53*(4-5), 299–313.

Konstantinidis, K. T., and J. M. Tiedje (2004), Trends between gene content and genome size in prokaryotic species with larger genomes, *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 3160–3165.

Koonin, E. V. (2009), Evolution of genome architecture, *The International Journal of Biochemistry. Cell Biology*, *41*(2), 298–306.

Koonin, E. V. (2011), *The Logic of Chance: The Nature and Origin of Biological Evolution*, 1st edn., FT Press.

Koonin, E. V., K. S. Makarova, and L. Aravind (2001), Horizontal gene transfer in prokaryotes: Quantification and classification, *Annual Review of Microbiology*, *55*, 709–742.

Koonin, E. V., and Y. I. Wolf (2008), Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world, *Nucleic Acids Research*, *36*(21), 6688–6719.

Krause, J., Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, A. P. Derevianko, and S. Pääbo (2010), The complete mitochondrial DNA genome of an unknown hominin from southern Siberia, *Nature*, *464*(7290), 894–897.

Krebs, C. J. (1994), *Ecology: The Experimental Analysis of Distribution and Abundance*, 4th edn., HarperCollins College Publishers, New York, NY.

Kung, J. T. Y., D. Colognori, and J. T. Lee (2013), Long noncoding RNAs: Past, present, and future, *Genetics*, *193*(3), 651–669.

Kurland, C. G., B. Canback, and O. G. Berg (2003), Horizontal gene transfer: A critical view, *Proceedings of the National Academy of Sciences of the United States of America*, *100*(17), 9658–9662.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. d. l. Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley,

T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, and M. J. Morgan (2001), Initial sequencing and analysis of the human genome, *Nature*, *409*(6822), 860–921.

Landry, C. R., and S. A. Rifkin (2012), The Genotype–Phenotype Maps of Systems Biology and Quantitative Genetics: Distinct and Complementary, in: *Evolutionary Systems Biology* (Soyer, O. S., ed.), no. 751 in Advances in Experimental Medicine and Biology, pp. 371–398, Springer New York.

Lane, N., and W. Martin (2010), The energetics of genome complexity, *Nature*, *467*(7318), 929–934.

Lang, A. S., and J. T. Beatty (2007), Importance of widespread gene transfer agent genes in Alphaproteobacteria, *Trends in Microbiology*, *15*(2), 54–62.

Lenski, R. E., C. Ofria, T. C. Collier, and C. Adami (1999), Genome complexity, robustness and genetic interactions in digital organisms, *Nature*, *400*(6745), 661–664.

Lenski, R. E., C. Ofria, R. T. Pennock, and C. Adami (2003), The evolutionary origin of complex features, *Nature*, *423*(6936), 139–144.

Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler (1991), Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations, *The American Naturalist*, *138*(6), 1315–1341.

Lenton, T. M., and C. A. Klausmeier (2006), Co-evolution of phytoplankton C: N: P stoichiometry and the deep ocean N: P ratio, *Biogeosciences Discussions*, *3*(4), 1023–1047.

Lieber, M. R., Y. Ma, U. Pannicke, and K. Schwarz (2003), Mechanism and regulation of human non-homologous DNA end-joining, *Nature Reviews. Molecular Cell Biology*, *4*(9), 712–720.

Louis, E. J., and A. V. Vershinin (2005), Chromosome ends: different sequences may provide conserved functions, *BioEssays*, *27*(7), 685–697.

Lundgren, M., A. Andersson, L. Chen, P. Nilsson, and R. Bernander (2004), Three replication origins in Sulfolobus species: Synchronous initiation of chromosome replication and asynchronous termination, *Proceedings of the National Academy of Sciences of the United States of America*, *101*(18), 7046–7051.

Lynch, M. (2006a), Streamlining and simplification of microbial genome architecture, *Annual Review of Microbiology*, *60*, 327–349.

Lynch, M. (2006b), The origins of eukaryotic gene structure, *Molecular Biology and Evolution*, *23*(2), 450–468.

Lynch, M. (2007), *The Origins of Genome Architecture*, 1st edn., Sinauer Associates Inc.

Lynch, M., R. Bürger, D. Butcher, and W. Gabriel (1993), The mutational meltdown in asexual populations, *Journal of Heredity*, *84*(5), 339–344.

Lynch, M., and J. S. Conery (2003), The origins of genome complexity, *Science*, *302*(5649), 1401–1404.

Lynch, M., X. Hong, and D. Scofield (2006), Nonsense-mediated decay and the evolution of eukaryotic gene structure, in: *L. E. Maquat (ed.) Nonsense-mediated mRNA Decay*, pp. 197–211, Landes Bioscience, Georgetown.

Makarova, K. S., L. Aravind, Y. I. Wolf, R. L. Tatusov, K. W. Minton, E. V. Koonin, and M. J. Daly (2001), Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics, *Microbiology and Molecular Biology Reviews*, *65*(1), 44–79.

Makarova, K. S., and M. J. Daly (2010), Comparative genomics of stress response systems in Deinococcus bacteria, *Bacterial Stress Responses, ASM Press, Washington, DC*, pp. 445–457.

Maley, C. C. (1999), Four steps toward open-ended evolution, in: *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-1999)*, vol. 2, pp. 1336–1343.

May, R. M., and R. H. M. Mac Arthur (1972), Niche overlap as a function of environmental variability, *Proceedings of the National Academy of Sciences of the United States of America*, *69*(5), 1109–1113.

Mayeda, A., G. R. Screaton, S. D. Chandler, X.-D. Fu, and A. R. Krainer (1999), Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements, *Molecular and Cellular Biology*, *19*(3), 1853–1863.

McInerney, J. O., J. A. Cotton, and D. Pisani (2008), The prokaryotic tree of life: past, present...and future?, *Trends in Ecology & Evolution*, *23*(5), 276–281.

Médigue, C., T. Rouxel, P. Vigier, A. Hénaut, and A. Danchin (1991), Evidence for horizontal gene transfer in Escherichia coli speciation, *Journal of Molecular Biology*, *222*(4), 851–856.

Mira, A., H. Ochman, and N. A. Moran (2001), Deletional bias and the evolution of bacterial genomes, *Trends in Genetics*, *17*(10), 589–596.

Misevic, D., C. Ofria, and R. E. Lenski (2010), Experiments with digital organisms on the origin and maintenance of sex in changing environments, *Journal of Heredity*, *101*(Supplement 1), S46–S54.

Molina, N., and E. v. Nimwegen (2008), The evolution of domain-content in bacterial genomes, *Biology Direct*, *3*(1), 51.

Molina, N., and E. van Nimwegen (2009), Scaling laws in functional genome content across prokaryotic clades and lifestyles, *Trends in Genetics*, *25*(6), 243–247.

Moran, N. A. (2002), Microbial minimalism: genome reduction in bacterial pathogens, *Cell*, *108*(5), 583–586.

Moran, N. A., and J. J. Wernegreen (2000), Lifestyle evolution in symbiotic bacteria: insights from genomics, *Trends in Ecology & Evolution*, *15*(8), 321–326.

Morgan, M. S., and M. Morrison (1999a), Models as Autonomous Agents, in: *Models as Mediators: Perspectives on Natural and Social Sciences*, Cambridge University Press, Cambridge; New York.

Morgan, M. S., and M. Morrison (1999b), *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge University Press, Cambridge.

Morris, R. M., M. S. Rappé, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, and S. J. Giovannoni (2002), SAR11 clade dominates ocean surface bacterioplankton communities, *Nature*, *420*(6917), 806–810.

Mozhayskiy, V., and I. Tagkopoulos (2013), Microbial evolution in vivo and in silico: Methods and applications, *Integrative Biology*, *5*(2), 262–277.

Müller, H. J. (1964), The relation of recombination to muatational advence, *Mutation Research*, *106*, 2–9.

Nee, S., and N. Colegrave (2006), Ecology: Paradox of the clumps, *Nature*, *441*(7092), 417–418.

Nelson, R. A., and M. G. Olsson (1986), The pendulum—Rich physics from a simple system, *American Journal of Physics*, *54*(2), 112–121.

Newman, M. E. J. (2003), The structure and function of complex networks, *arXiv:cond-mat/0303516*.

Newman, M. E. J. (2006), Modularity and community structure in networks, *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582.

Noble, D. (2006), *The Music of Life*, Oxford University Press, Oxford.

Noble, D. (2008), Genes and causation, *Philosophical Transactions A*, *366*(1878), 3001.

Nordstrom, K., and S. Dasgupta (2006), Copy-number control of the Escherichia coli chromosome: A plasmidologist's view, *EMBO Reports*, *7*(5), 484–489.

Ochman, H., and L. M. Davalos (2006), The nature and dynamics of bacterial genomes, *Science*, *311*(5768), 1730–1733.

Ochman, H., J. G. Lawrence, and E. A. Groisman (2000), Lateral gene transfer and the nature of bacterial innovation, *Nature*, *405*(6784), 299–304.

Ofria, C., C. Adami, and T. C. Collier (2003), Selective pressures on genomes in molecular evolution, *Journal of Theoretical Biology*, *222*(4), 477–83.

Ofria, C., and C. O. Wilke (2004), Avida: A software platform for research in computational evolutionary biology, *Artificial Life*, *10*(2), 191–229.

Olszewski, T. D. (2011), Persistence of high diversity in non-equilibrium ecological communities: implications for modern and fossil ecosystems, *Proceedings of the Royal Society B: Biological Sciences*, p. rspb20110936.

Oren, A. (2004), Prokaryote diversity and taxonomy: Current status and future challenges, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1444), 623–638.

Osawa, S., T. H. Jukes, K. Watanabe, and A. Muto (1992), Recent evidence for evolution of the genetic code, *Microbiological Reviews*, *56*(1), 229–264.

Oyama, S., P. E. Griffiths, and R. D. Gray (2003), *Cycles of Contingency: Developmental Systems and Evolution*, MIT Press.

Partensky, F., W. R. Hess, and D. Vaulot (1999), Prochlorococcus, a marine photosynthetic prokaryote of global significance, *Microbiology and Molecular Biology Reviews: MMBR*, *63*(1), 106–127.

Parter, M., N. Kashtan, and U. Alon (2007), Environmental variability and modularity of bacterial metabolic networks, *BMC Evolutionary Biology*, *7*(1), 1–8.

Pearson, H. (2006), Genetics: What is a gene?, *Nature*, *441*(7092), 398–401.

Pigliucci, M. (2010), Genotype–phenotype mapping and the end of the 'genes as blueprint' metaphor, *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1540), 557–566.

Plutynski, A. (2001), Modeling evolution in theory and practice, *Philosophy of Science*, *68*(3), S225–S236.

Postollec, F., A.-G. Mathot, M. Bernard, M.-L. Divanac'h, S. Pavan, and D. Sohier (2012), Tracking spore-forming bacteria in food: From natural biodiversity to selection by processes, *International Journal of Food Microbiology*, *158*(1), 1–8.

Pradella, S., A. Hans, C. Spröer, H. Reichenbach, K. Gerth, and S. Beyer (2002), Characterisation, genome size and genetic manipulation of the myxobacterium Sorangium cellulosum So ce56, *Archives of Microbiology*, *178*(6), 484–492.

Price, M. N., P. S. Dehal, and A. P. Arkin (2008), Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli, *Genome Biology*, *9*(1), R4.

Proulx, S. R., D. E. L. Promislow, and P. C. Phillips (2005), Network thinking in ecology and evolution, *Trends in Ecology & Evolution*, *20*(6), 345–353.

Raghuraman, M. K., E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer, and W. L. Fangman (2001), Replication dynamics of the yeast genome, *Science*, *294*(5540), 115–121.

Ranea, J. A. G., A. Grant, J. M. Thornton, and C. A. Orengo (2005), Microeconomic principles explain an optimal genome size in bacteria, *Trends in Genetics: TIG*, *21*(1), 21–25.

Raup, D. M., and J. J. Sepkoski (1982), Mass Extinctions in the Marine Fossil Record, *Science*, *215*(4539), 1501–1503.

Ray, T. S. (1992), Evolution, ecology and optimization of digital organisms, *Santa Fe Institute*.

Reva, O. N., I. B. Sorokulova, and V. V. Smirnov (2001), Simplified technique for identification of the aerobic spore-forming bacteria by phenotype., *International Journal of Systematic and Evolutionary Microbiology*, *51*(4), 1361–1371.

Rivera, M. C., and J. A. Lake (2004), The ring of life provides evidence for a genome fusion origin of eukaryotes, *Nature*, *431*(7005), 152–155.

Roberts, G. C., and C. W. Smith (2002), Alternative splicing: combinatorial output from the genome, *Current Opinion in Chemical Biology*, *6*(3), 375–383.

Rousset, F. (2003), Effective size in simple metapopulation models, *Heredity*, *91*(2), 107–111.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith (1977), Nucleotide sequence of bacteriophage ϕX174 DNA, *Nature*, *265*(5596), 687–695.

Schaetzlein, S., A. Lucas-Hahn, E. Lemme, W. A. Kues, M. Dorsch, M. P. Manns, H. Niemann, and K. L. Rudolph (2004), Telomere length is reset during early mammalian embryogenesis, *Proceedings of the National Academy of Sciences of the United States of America*, *101*(21), 8034–8038.

Scheffer, M., and E. H. van Nes (2006), Self-organized similarity, the evolutionary emergence of groups of similar species, *103*(16), 6230–6235.

Scheldeman, P., A. Pil, L. Herman, P. D. Vos, and M. Heyndrickx (2005), Incidence and diversity of potentially highly heat-resistant spores isolated at dairy farms, *Applied and Environmental Microbiology*, *71*(3), 1480–1494.

Servais, P., G. Billen, and J. V. Rego (1985), Rate of Bacterial Mortality in Aquatic Environments, *Applied and Environmental Microbiology*, *49*(6), 1448–1454.

Skibinski, D., M. Woodwark, and R. D. Ward (1993), A quantitative test of the Neutral theory using pooled allozyme data, *Genetics*, *135*(1), 233–248.

Slade, D., and M. Radman (2011), Oxidative Stress Resistance in Deinococcus radiodurans, *Microbiology and Molecular Biology Reviews*, *75*(1), 133–191.

Smith, C. W., and J. Valcárcel (2000), Alternative pre-mRNA splicing: the logic of combinatorial control, *Trends in Biochemical Sciences*, *25*(8), 381–388.

Soyer, O. S., and T. Pfeiffer (2010), Evolution under fluctuating environments explains observed robustness in metabolic networks, *PLoS Computational Biology*, *6*(8), e1000,907.

Stanley, S. M., and X. Yang (1994), A Double Mass Extinction at the End of the Paleozoic Era, *Science*, *266*(5189), 1340–1344.

Stechmann, A., and T. Cavalier-Smith (2002), Rooting the eukaryote tree by using a derived gene fusion, *Science*, *297*(5578), 89–91.

Stern, D. L., and V. Orgogozo (2009), Is genetic evolution predictable?, *Science*, *323*(5915), 746–751.

Stillman, B. (1996), Cell cycle control of DNA replication, *Science*, *274*(5293), 1659–1663.

Storer, R. W. (1966), Sexual dimorphism and food habits in three north american accipiters, *The Auk*, *83*(3), 423–436.

Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson (2000), Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen, *Nature*, *406*(6799), 959–964.

Strogatz, S. H. (2001), Exploring complex networks, *Nature*, *410*(6825), 268–276.

Stumpf, M. P. H., W. P. Kelly, T. Thorne, and C. Wiuf (2007), Evolution at the system level: the natural history of protein interaction networks, *Trends in Ecology & Evolution*, *22*(7), 366–373.

Sturtevant, A. H. (1937), Essays on Evolution. I. On the Effects of Selection on Mutation Rate, *The Quarterly Review of Biology*, *12*(4), 464.

Suzina, N. E., A. L. Mulyukin, V. V. Dmitriev, Y. A. Nikolaev, A. P. Shorokhova, Y. S. Bobkova, E. S. Barinova, V. K. Plakunov, G. I. El-Registan, and V. I. Duda (2006), The structural bases of long-term anabiosis in non-spore-forming bacteria, *Advances in Space Research*, *38*(6), 1209–1219.

Syvanen, M. (1985), Cross-species gene transfer; implications for a new theory of evolution, *Journal of Theoretical Biology*, *112*(2), 333–343.

Syvanen, M. (2012), Evolutionary implications of horizontal gene transfer, *Annual Review of Genetics*, *46*(1), 341–58.

Szabó, P., and G. Meszéna (2006), Limiting similarity revisited, *Oikos*, *112*(3), 612–619.

Tilman, D. (1996), Biodiversity: Population versus ecosystem stability, *Ecology*, *77*(2), 350–363.

Tomala, K., and R. Korona (2013), Evaluating the fitness cost of protein expression in Saccharomyces cerevisiae, *Genome Biology and Evolution*, *5*(11), 2051–2060.

Torres-Sosa, C., S. Huang, and M. Aldana (2012), Criticality is an emergent property of genetic networks that exhibit evolvability, *PLOS Computational Biology*, *8*(9), e1002,669.

Treangen, T. J., and E. P. C. Rocha (2011), Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes, *PLoS Genetics*, *7*(1).

Ulrich, L. E., E. V. Koonin, and I. B. Zhulin (2005), One-component systems dominate signal transduction in prokaryotes, *Trends in Microbiology*, *13*(2), 52–56.

van Nimwegen, E. (2003), Scaling laws in the functional content of genomes, *Trends in Genetics*, *19*(9), 479–484.

Van't Hof, J., and C. Bjerknes (1981), Similar replicon properties of higher plant cells with different S periods and genome sizes, *Experimental Cell Research*, *136*(2), 461–465.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen,

M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu (2001), The sequence of the human genome, *Science*, *291*(5507), 1304–1351.

Wagner, A. (2008), Gene duplications, robustness and evolutionary innovations, *BioEssays*, *30*(4), 367–373.

Wagner, G. P., and L. Altenberg (1996), Perspective: Complex adaptations and the evolution of evolvability, *Evolution*, *50*(3), 967–976.

Wain, H. M., E. A. Bruford, R. C. Lovering, M. J. Lush, M. W. Wright, and S. Povey (2002), Guidelines for human gene nomenclature, *Genomics*, *79*(4), 464–470.

Wang, Z., and J. Zhang (2009), Abundant Indispensable Redundancies in Cellular Metabolic Networks, *Genome Biology and Evolution*, *1*, 23–33.

Welch, R. A., V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S.-R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner (2002), Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli, *Proceedings of the National Academy of Sciences of the United States of America*, *99*(26), 17,020–17,024.

Werren, J. H. (2011), Selfish genetic elements, genetic conflict, and evolutionary innovation, *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 2*, 10,863–10,870.

White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft, M. L. Gwinn, W. C. Nelson, D. L. Richardson, K. S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J. J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K. S. Makarova, L. Aravind, M. J. Daly, K. W. Minton, R. D. Fleischmann, K. A. Ketchum, K. E. Nelson, S. Salzberg, H. O. Smith, J. Craig, Venter, and C. M. Fraser (1999), Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1, *Science*, *286*(5444), 1571–1577.

Whitlock, M. C., and N. H. Barton (1997), The effective size of a subdivided population, *Genetics*, *146*(1), 427–441.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe (1998), Prokaryotes: The unseen majority, *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 6578–6583.

Wielgoss, S., J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R. E. Lenski, and D. Schneider (2011), Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with Escherichia coli, *G3: Genes, Genomes, Genetics*, *1*(3), 183–186.

Wilke, C. O., and C. Adami (2002), The biology of digital organisms, *Trends in Ecology & Evolution*, *17*(11), 528–532.

Wilke, C. O., J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami (2001), Evolution of digital organisms at high mutation rates leads to survival of the flattest, *Nature*, *412*(6844), 331–333.

Williams, H. T. P., and T. M. Lenton (2007a), Artificial selection of simulated microbial ecosystems, *Proceedings of the National Academy of Sciences of the United States of America*, *104*(21), 8918–8923.

Williams, H. T. P., and T. M. Lenton (2007b), The Flask model: emergence of nutrient-recycling microbial ecosystems and their disruption by environment-altering 'rebel' organisms, *Oikos*, *116*(7), 1087–1105.

Wood, D. W., J. C. Setubal, R. Kaul, D. E. Monks, J. P. Kitajima, V. K. Okura, Y. Zhou, L. Chen, G. E. Wood, N. F. Almeida, L. Woo, Y. Chen, I. T. Paulsen, J. A. Eisen, P. D. Karp, D. Bovee, P. Chapman, J. Clendenning, G. Deatherage, W. Gillet, C. Grant, T. Kutyavin, R. Levy, M.-J. Li, E. McClelland, A. Palmieri, C. Raymond, G. Rouse, C. Saenphimmachak, Z. Wu, P. Romero, D. Gordon, S. Zhang, H. Yoo, Y. Tao, P. Biddle, M. Jung, W. Krespan, M. Perry, B. Gordon-Kamm, L. Liao, S. Kim, C. Hendrick, Z.-Y. Zhao, M. Dolan, F. Chumley, S. V. Tingey, J.-F. Tomb, M. P. Gordon, M. V. Olson, and E. W. Nester (2001), The genome of the natural genetic engineer Agrobacterium tumefaciens C58, *Science*, *294*(5550), 2317–2323.

Wooldridge, S. A. (2010), Is the coral-algae symbiosis really 'mutually beneficial' for the partners?, *BioEssays*, *32*(7), 615–625.

Wright, S. (1931), Evolution in mendelian populations, *Genetics*, *16*(2), 97–159.

Xing, Y., and C. Lee (2006), Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes, *Nature Reviews. Genetics*, *7*(7), 499–509.

Yamauchi, A., and T. Miki (2009), Intraspecific niche flexibility facilitates species coexistence in a competitive community with a fluctuating environment, *Oikos*, *118*(1), 55–66.

Yerrapragada, S., J. L. Siefert, and G. E. Fox (2009), Horizontal gene transfer in cyanobacterial signature genes, in: *Horizontal Gene Transfer* (Gogarten, M. B., J. P. Gogarten, L. C. Olendzenski, and J. M. Walker, eds.), *Methods in Molecular Biology*, vol. 532, pp. 339–366, Humana Press.

Zhang, J. (2003), Evolution by gene duplication: An update, *Trends in Ecology & Evolution*, *18*(6), 292–298.

Zheng, H., and H. Wu (2010), Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species, *BMC Bioinformatics*, *11*(Suppl 11), S7.

Appendix A

Documentation of the model

The computer source code and its documentations are attached in an electronic form. Also they can be obtained from the author upon request. The documentation consists of:

00_HGT_CurrentCode – directory containing the source code of the model's implementation

01_PyScripts – directory containing scripts used to generate plots and statistics for the analysis

04_Results – directory containing a number of example outputs of the model

Doc - directory containing documentation of the program in HTML format; use the *index.html* file to launch it in a browser

If parameter No. 42 will be set to zero, then the model should present properties such as the ones analysed in Chapter 3.

All of the computer source code published with this thesis is licensed under a GNU General Public Licence version 3 or later as published by the Free Software Foundation.