DIPLOMARBEIT

Modelling Evolution of Genetically Controlled Reaction Networks

zur Erlangung des akademischen Grades
Magister rerum naturalium

Verfasser	Thomas Taylor
Matrikelnummer	0008247
Studienrichtung	A490 — Molekulare Biologie
Betreuer	Prof. Dr. Peter Schuster
Vorgelegt der	Fakultät für Lebenswissenschaften der Universität Wien

Wien, am 2006-04-05

Acknowledgment

Silent gratitude isn't much use to anyone.

— G.B. Stern

So let me not be silent and thank you as resonantly as possible.

Invaluable help to create this diploma thesis has come from Peter Schuster, Christoph Flamm, Lukas Endler, Ulli Mückstein, Andreas Svrcek-Seiler, Christian Taylor.

And I am eternally indebted to those who kept me sane, Julia Doll, Vazul Litkey, Clemens Wagner, Thomas Walach.

Contents

Life is full of doors that don't open when you knock, equally spaced amid those that open when you don't want them to.

— Roger Zelazny

1 Introduction

But scientists, who ought to know
Assure us that it must be so.
Oh, let us never, never doubt
What nobody is sure about.

— Hilaire Belloc

1.1 A Sample Cell

Cells are the building blocks of organisms. Anything the organism is capable of is a result of functionality provided by cells. In multicellular organisms cells form tissues, a multitude of cells that carry out similar functions, which in turn build organs, a multitude of tissues, each tissue specialised for a different function. Even though cells provide different functionality to the organism and may appear completely different from one another, all have the same cellular identity on the organismic level. Close to every cell has the same information, the same genes encoded by its DNA (desoxyribonucleic acid) in its nucleus and thus the same cellular identity.

The DNA in the nucleus is called the genome, and contains all genes expressed anywhere in the organism. To express a gene, it first has to be transcribed. This process is initiated by a transcription complex forming near or at a promotor on the genome. Then a copy of the information is made, the RNA (ribonucleic acid). From that copy the sequence for the protein is read and the protein synthesised.
Of course this is a simplified example. In any biological cell additional mech-

anisms of control exist, which influence what proteins are made and how much of them are synthesised. Also some RNAs are not meant to be the template for a protein and have entirely different functions and sometimes RNA is the template for DNA. The simplified cellular processes and elements needed are depicted in figure 1.

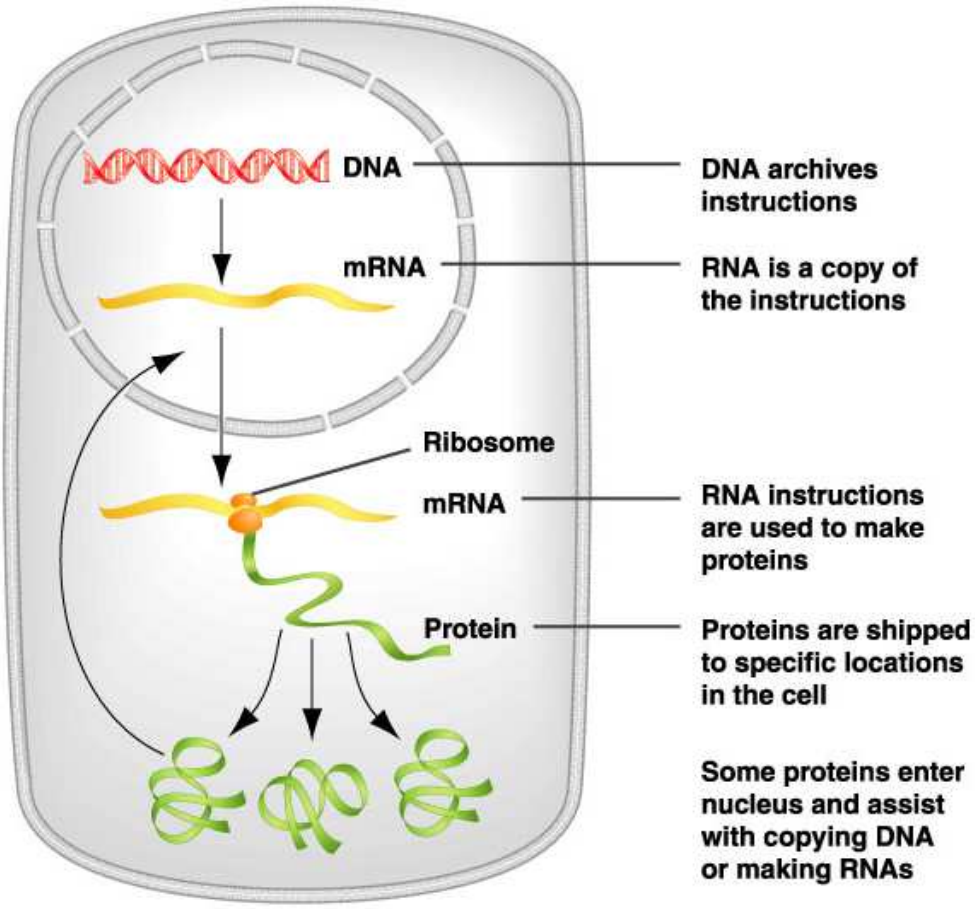

Figure 1: Primary processes of cells. The genome (DNA) is transcribed (RNA) and then translated (proteins).

1.2 Evolution

In 1859 Charles Darwin published his best known book *On the Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life*. This work proposed a mechanism for adaption known as natural selection. Thus gradual change of species, phenotypic variants and to some extent the origin of new species could be explained. However the theory was at odds with the observable rather fast and sometimes quite abrupt creation of a new species, in contrast to a gradual and "unbroken" chain of changes as ought to be expected.

In 1865 Gregor Mendel first made his work available to the public. Through extensive analysis of hereditary properties of certain traits in peas he arrived at three laws that form the basis of genetics. To devise his laws, Mendel identified distinct units of hereditary information now called genes, and realised that they are passed on to offspring in principle without change and without influence of the phenotype of the parents.

So genotype and phenotype become distinct properties of an organism. The genotype refers to what information the genome carries, basically an idea what the organism should be like. On the other hand the phenotype denotes the actual reality of the organism, from shape and size of proteins to wrinkly or smooth peas. To some extent the genotype defines the phenotype but the environment and epigenetics constitute important factors for the unfolding of the phenotype as well. Every mechanism of evolution has to change the genotype, but alterations become meaningful for selection only if the phenotype is different as well.

The synthetic theory of evolution is a synthesis of Mendel's theory on genetics and Darwin's mechanism of natural selection, sometimes also called neo-Darwinism, a version of which has been proposed by Ernst Mayr, Theodosius Dobyhansky and others [11, 34]. When populations of organisms become separated into subpopulations by time or space, any newly developed trait is limited to the subpopulation, in which it originated, and it will not spread to other subpopulations due to lack of interbreeding. As traits and

differences accumulate interbreeding becomes impossible even withouth spatial or temporal barriers, because no fertile living offspring can be concieved any more [33]. But neo-Darwinism is not limited to Mendelian genetics and Darwinian selection, mutation and recombination in populations have been recognised to play an important role and were incorporated into the synthetic theory. The idea of mutation was concieved around 1910 by Thomas Hunt Morgan and his students when they discovered a white eyed mutant of *drosophila melanogaser*. They found that mutations are permanent changes in genes and cause a different phenotype. Furthermore white eyes where only ever present in male flies of *drosophila melanogaster* and thus gave reason to assume genetic linkage, like sex and eye colour. Rigorous studies of coinheritance frequencies of traits lead to the development of the first genetic map and proved that chromosomes are the carriers of genes.

In 1952, a series of experiments done by Alfred Hershey and Martha Chase provided evidence for DNA (desoxyribonucleic acid) to be the molecular substrate of genes and chromosomes, instead of proteins, therefore contradicting common beliefs at that time [5].

In 1953, the structure of DNA was finally unravelled by James Watson and Francis Crick: Two antiparallel chains coiled around the same axis form a double-helix, the interaction between the two chains limited to two different pase pairings [49]. Around that time the term "central dogma" coined by Francis Crick became popular, describing the concept of information stored on the DNA encoding RNA (ribonucleic acid) which in turn encodes proteins, but no information flow occurs in the opposite direction. The central dogma was partially disproven in 1971 by Howard Martin Temin and David Baltimore who independently of each other discovered reverse transcriptase, an enzyme that synthesises DNA from an RNA template [3, 21, 38, 47]. Further the central dogma is based on the assumption that proteins, the last element of the information flow chain, are the catalytically and regulatorily active components of the cell. However recent studies give evidence to catalytical and regulatory activity of RNA [1, 2, 32] and lend credibility to the

RNA world hypothesis [20, 44], a model for the origin of life based on the assumption that RNA fulfilled all needed functions in the early cell or protocell whereas proteins and DNA arose later as specialised elements.
However the "classic" Neo-Darwinian view became recently substituted by modern evolutionary synthesis. The difference lies in the inclusion of several mechanisms for evolution other than natural selection [36]. An abstract of the tenet is given by Futuyma [17]: Genetic variation arises by mutation and recombination. Any population is made up of genetically variant individuals and thus evolution of populations is achieved by change in the frequencies of alleles in the gene pool. Mechanisms of frequency alteration are either random like genetic drift and catastrophic events, or adaptively directed like gene flow and natural selection. The result are gradual phenotypic changes and diversification by speciation, the latter coming into effect by accumulation of alterations.
In contrast to the above interpretation of speciation the theory of "Punctuated Equilibrium" assumes that speciation is a rapid process interspaced with long intervals of stasis. As both models are able to explain key aspects of evolution, a hybrid hypothesis is possible and the relative contribution of each theory to the whole of the process is subject to discussion.

1.3 Neutrality in Evolution

To apply adaptively directed selection mechanisms, phenotypes have to be ordered in some way. The natural ordering is usually refered to as fitness. A fit phenotype will have a high likelihood of survival, an unfit one will more likely become extinct. Now there are alterations in the genotype and often also at basic levels of the phenotype that do not change the fitness. These mutations are considered neutral [27]. At first glance neutral mutations appear to be meaningless but studies have proven the vast impact they have on biological systems. Fit phenotypes are often separated by multiple mutations on the genotypical level. To cross from one such phenotype to another unfit intermediate phenotypes are expected. However, neutral mutations allow for

approach to another fit phenotype, thus enhancing the number of phenotypes that can be switched to by a single mutational event [24]. On the other hand neutrality enables higher mutation rates and therefore faster adaption while minimising the risk of unviable phenotypes.
Already in 1968 Motoo Kimura formulated the neutral theory of molecular evolution [27]. Tenets thereof are neutrality of the majority of single nucleotide exchanges in the genome and genetic drift. As neutral mutations do not alter fitness natural selection does not apply. However the mechanism of genetic drift leads to accumulation or decrease and loss of certain equally fit variants in a population by pure chance. This mechanism has been shown to vastly improve the search capacity in sequence space [14, 15, 25, 43]
Little later, in the 1970s and 1980s, a theory for optimization and maintenance of nucleotide sequences under error prone replication was introduced by Manfred Eigen, Peter Schuster and John McCaskill and described as molecular quasi-species [8–10]. According to the underlying theory the requirements of evolution are only an open system with replication far off the thermodynamic equilibrium and limited resources. For every organism the molecular quasi-species is the sum of all mutants, that can arise from replication and mutation of the original organism. If a sequence specific fitness distribution and fixed mutation rate is assumed, the quasi-species is unambiguously determined by the sequence. Under such circumstances mutation and selection suffice to guarantee maintained diversity of sequences [9].

1.4 RNA as a Model

Beginning around 1970, RNA (ribonucleic acid) became a prime model for molecular evolution. The sequence of RNA is made up of a four letter alphabet. Well defined interactions between distinct monomers lead to formation of structure. Function is largely determined by structure, the three dimensional shape called tertiary structure. Secondary structure has a big impact on possible tertiary structures and can be easily computed by several folding algorithms [23, 31, 50]. As the phenotype is made up of the structure and the

sequence equals the genotype, a substrate for neutrality has been identified: Any structure can be formed by multiple sequences. The importance of neutral mutants for evolutionary optimisation was demonstrated by work done by Peter Schuster and his group [12–15].

By further research from the 1990s it was shown unambiguously that an increase in neutrality results in an increased phenotypic error threshold [16,40, 42], which is a measure for maintainance of phenotypes. Extensive neutral networks were found to exist for RNA. The sum of all genotypes that form the same phenotype constitute the neutral network. Within that network all sequences that differ only at one position are connected and form a graph. For binary alphabets the neutral networks for RNA sequence to secondary structure maps has been exhaustively analysed [18, 19]. Using the natural alphabet statistical evaluation was performed [45, 46].

The above described graph of all sequences with connections between one error neighbours spans the RNA sequence space. Extensive structural neutral networks can be found upon which sequences may differ without losing the optimal structure. For a population this is a sturdy design and increases flexibility as neutral networks intercalate and the more sequences present the more likely an intercalation point is found.

For evolution however the phenotype and its fitness are the predominant criteria. For adequate description thereof the accessibility of genotypes by mutation has to be mapped to phenotype accessibility relations. Hamming distances provide a metric for RNA sequences, but for the structure only statistical neighbourhood relations are available. [15].

Theoretical observation of neutral networks could be proven by experimental evidence gathered by Schultes and Bartel [41]. For two known, phylogenetically unrelated ribozymes with different catalytic activities, an RNA sequence was found that can fold into both secondary structures and exhibits both catalytic activities.

1.5 Genetically Controlled Reaction Networks

RNA as a model for evolution is very well understood and has given countless insights in mechanisms that shape living organisms. Even if the RNA world hypothesis proves true, today's creatures feature RNA as only one of many different levels in transforming genotype into phenotype. To take previous studies further, genetically controlled reaction networks are subject of this work.

Section 2.1 explains the details of the model underlying the approach of this work and also discusses similarities and divergencies to processes in organic cells. Some constraints arise independent of the model and are discussed in section 2.2. The phenotype of RNA is defined as the structure. For genetically controlled reaction networks the definition and measure of a descriptive phenotype is not as straight forward. Section 2.3 details the process of arriving at a phenotype from a given genotype. Additionally every mutation can be typed and section 2.4 deals with all possible types of mutation.

In section 3 the sample sets are described along with all phenotypes and mutations found. Also a little correlation analysis of phenotype and mutation type will be done.

Finally section 4 states possible conclusions based on the data processed for this work and indicates possible further studies.

2 Model and Approach

Reality is that which, when you stop believing in it, doesn't go away.

— Philip K. Dick

Gene — In this chapter the term gene will be used often. However behind this single entity two different concepts are hidden. On the one hand there is the gene in the wider sense that contains everything from the URE (upstream regulatory element) and the promotor down to the last basepair transcribed. On the other hand gene may be used in the stricter sense refering to the sequence being transcribed (URE and promotor are not transcribed). See figure 2.
Although potentially confusing the intended concept behind the term gene should be clear from the context and where not will be pointed out explicitely.

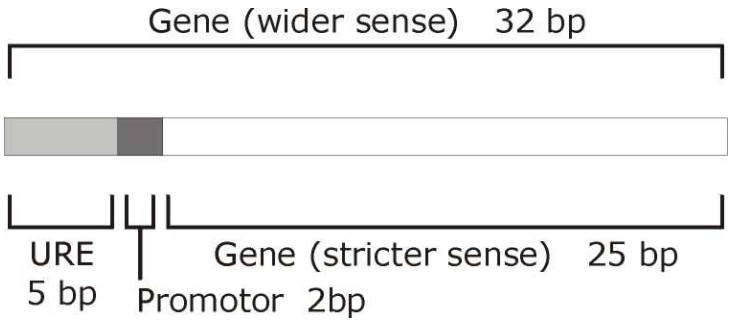

Figure 2: Anatomy of a model gene.

2.1 Model

Please be aware that this model is built on numerous approximations. First the size of genes and genome has been reduced drastically. Also genes may overlap, which means that a single position / basepair on the genome may be part of more than one gene, as is commonly found in viruses only. Further epigenetic mechanisms present in all biological cells are not part of the model. And lastly the complexity of gene regulation and gene product interaction is only a fraction of even the simplest known cell.

2.1.1 A Simulated Cell

This work is based on MiniCellSim-Genome (see appendix B), a set of perl libraries to simulate typical cellular processes. Basic principles of cells found in organisms have been used to design this model, however many things are simplified. The reason for this is that more complex models can hardly be handled by today's hardware and have to sacrifice either precision or variability.
Like in organic cells, cellular identity is maintained by the genome encompassing information on all cellular processes and their mode of execution. The genome is also passed on to offspring causing parental and filial cells to have similar cellular identities, or identical if no mutation is present. Genes located on the genome are templates for RNA, the process of synthesis is called transcription. In turn the RNA gives rise to a protein, by a process known as translation. As not every gene is useful to the cell in the given circumstances of the environment, numerous levels of regulation exist enabling the cell to control gene expression. As described in section 2.3 the regulatory network is part of the evaluated phenotype.

Figure 3 shows the basic cellular processes implemented in MiniCellSim-Genome and the following sections discuss them in more detail. Every cell's identity is determined by the genome (A). On that genes are identified (B). No part of a genes may situated outside of the genome boundaries but genes

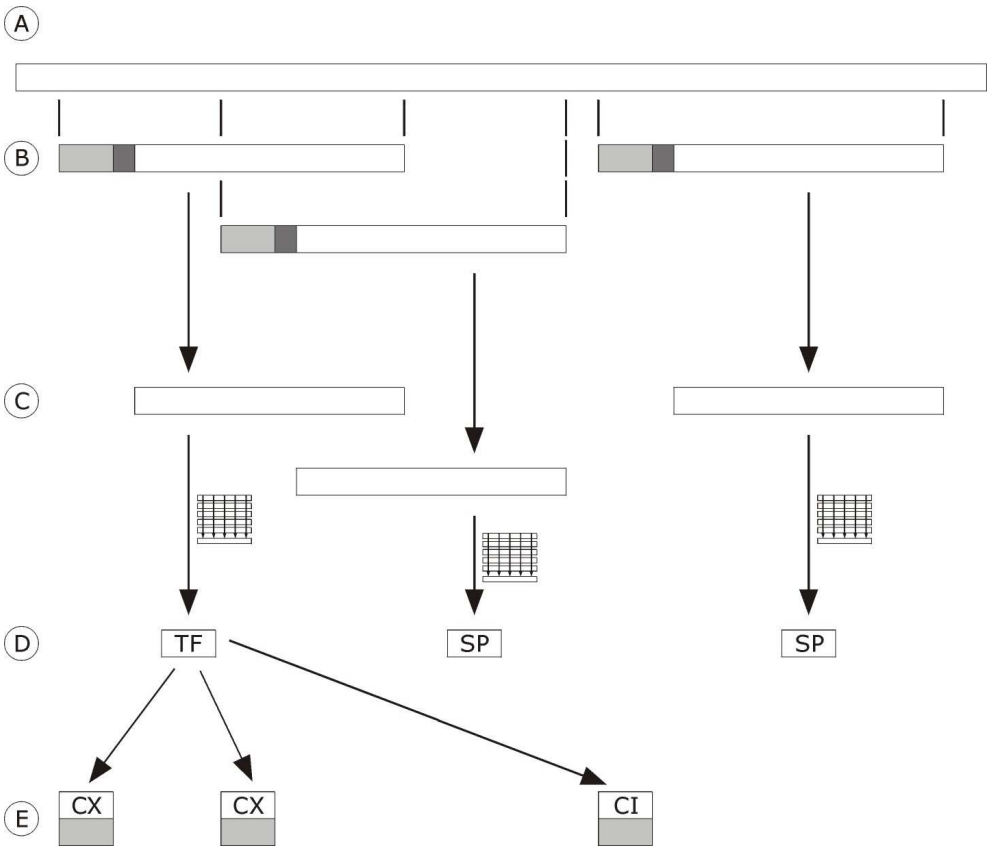

Figure 3: Cellular processes of MiniCellSim-Genome. In (A) the genome can be seen on which (B) genes are located. Genes may overlap. (C) Every gene is transcribed to mRNA, to which a majority rule is applied to translate it into a protein (D). By sequence evaluation half of the proteins are assigned to be structural proteins (SP) and the other half to be transcription factors (TF). The latter is capable of interacting with the UREs of genes on the genome and where two types of interaction are possible: either activating (CX) or repressing (CI). The interaction strength is calculated by cofolding URE and TF. Interaction strength and type then influence the expression rate (transcription rate) of the gene with the corresponding URE.

may overlap each other. Every gene is template for a transcript (C) which represents the RNA of organic cells. As equivalent to translation the transcripts are folded using a majority rule to derive shorter sequences, the gene products (D). The corresponding entity for gene products in living organisms are usually proteins. Two classes of gene products are distinguished: structural proteins (SP) and transcription factors (TF). The latter is capable to influence transcription of certain genes either as activator (CX) or repressor (CI) as shown in (E).

2.1.2 Genome

Every living organism known today has a genome made of DNA or RNA (ribonucleic acid – the latter has as of yet only been found to constitute the genome in viruses). Size and shape may vary considerably as the following examples show:

Bacteriophage phi X174 [48]	ss linear	5,375 bp	5×10^3 bp
Escherichia coli [48]	ds circular	338,534 bp	3×10^5 bp
Homo sapiens [6]	ds linear	3,019,560,019 bp	3×10^{10} bp

Table 1: Sample genomes. All genomes are DNA. ss means single strand; ds double strand. The genome of Homo sapiens is organised in 23 chromosomes and the size displayed is for a female haplotype (humans are of diploid chromosome type, alas every chromosome exists twice with the exception of the sex chromosome where females have XX and males have XY).

DNA and RNA are chemically pretty similar biopolymers made up of monomers called nucleotides. Each nucleotide consists of a base, a sugar and phosphate. The "backbone" is made up of the pentose sugar (ribose for RNA and desoxyribose for DNA) and the phosphate, which can link nucleotide to nucleotide in an infinite chain. The sequence of the biopolymer is determined by the bases, which allow for the interaction of two strands of nucleic acids (either DNA/DNA, RNA/RNA or DNA/RNA). In DNA the purine bases adenosine (A), guanine (G) and the pyrimidine compounds cytosine (C) and

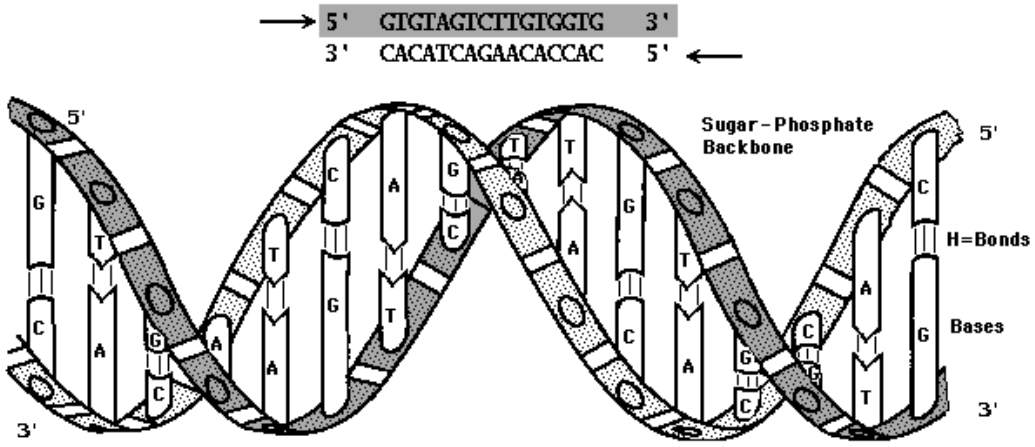

Figure 4: The structure of DNA. On top is the shorthand notation of both strands. The lower part shows the helix those two strand form, which is a more accurate representation of what the DNA actually looks like.
As an example for upstream and downstream look at the A in position 5 of the upper strand (grey background). Any and all of the bases CTGT in position 1 to 4 are considered upstream of A in position 5. The following bases GTC etc. beginning in position 6 are termed downstream. So in the upper strand upstream means left of and downstream right of. Because of anti-parallel alignment, 5' and 3' have switched positions in the lower strand. Upstream and downstream still point towards 5' and 3' respectively, but upstream is to the rigth and downstream to the left.

thymidine (T) are found. Pairing is restricted to the Watson-Crick pairs, adenosine (A) with thymine (T) and guanine (G) with cytosine (C). For RNA thymine (T) is replaced by the pyrimidine base uracil (U), which then pairs with adenosine (A).
All sequence specific amplification processes, like transcription, depend on the the principle of complementarity [49]. For two strands to be complementary each base on one strand has to pair with the corresponding base on the other strand. As only Watson-Crick pairs are allowed the sequence of the complementary strand to any strand is unique and reproducible. Because of the deterministic behaviour cells can reproduce their genome by resynthesising one strand using the other as a template and vice versa.

Both DNA and RNA are directed, their ends are dissimilar. Derived from the atoms of the sugar one end is denoted 5' and the other 3', the direction of the strand being 5' → 3' by convention (see figure 4). All positions nearer to the 5' end of the genome than position X are said to be upstream of X, but if closer to the 3' end are called downstream of X. In biological systems DNA and RNA always pair antiparallel, which means that if position X on one strand pairs with position Y on the other strand, nucleotides upstream of X only pair with nucleotides downstream of Y and vice versa (see figure 4).
Replication, the doubling of the genome needed for passing on one to the filial cell, also is sequence specific. Whereas double stranded (ds) organisms like *Escherichia coli* and *Homo sapiens* simply copy the genome by unwinding the helix and using both strands as templates, single stranded (ss) genomes like in *Bacteriophage phi X174* have to first generate the complementary strand and use it as template to replicate their actual genome. Nevertheless a double stranded form exists in all organisms, therefore both strands are available for serving as a template to RNA and thus both strands may carry genes.
Genomes can be linear or circular. Although linear genomes need more complicated machinery to be replicated without loss and cannot have genes in the outlying regions, the linear form appears to be of advantage for long genomes. Circular genomes are mostly found in bacteria, where small size and ease of replication constitute a considerable selective advantage.

To describe the DNA genome in the model a string containing only the four bases A, C, G, T is sufficient. These four bases may also be termed letters and together they form the alphabet of DNA $\mathcal{A} = \{A, C, G, T\}$.
Genome sizes assumed in this work range from 50 to 150 bp, which is a lot smaller than even the smallest known genome of an organism. To compensate gene size has been reduced greatly as well (see below, section 2.1.3). Because of the short length of the genome, the linearity of the genome in the model has considerable impact (as discussed in section 2.2.1). The genome is single

stranded and in the model replication occurs by direct copying of the strand. So the complementary strand that would be present in biological cells is missing. Such a strand could carry additional genes but they are irrelevant for the model. In regard to reproduction kinetics the simplification has little impact.

2.1.3 Genes

Based on observed genome lengths one might assume that organisms with genomes longer by several orders of magnitude would have proportionally more genes. Of course organisms with longer genomes have more genes, but far less than might be expected. To organisms with short generation time, like bacteria who can doulbe their number in hours or days, a small genome is profitable as it can be replicated faster and the process needs less energy. So a genome that contains all genes in a way that they need little space is advantageous. Organisms with long generation time, like *Homo sapiens*, gain little by optimising the genome for being short. Therefore genes tend to be longer as coding sequences are interspaced with non-coding regions and the amount of non-coding regions in general is higher. To stick with the example of man, presumably 90% of the sequence are non-coding and the remaining 300,000,000 bp code for approximately 30,000 genes, giving a mean length of 1,000 bp or 1 kbp per gene. This calculation is not entirely true as the genome is double stranded and non-coding regions may be located within gene boundaries.
The layout of genes shows a great range. Bacteria have polycistronic genes, where a single promotor induces the transcription of multiple genes that lie adjacent to each other. Regulatory elements usually are in the vicinity of the promotor and genes do not contain non-coding regions. In eukaryotes (plants, animals and others) each gene has its own promotor and contains numerous coding and non-coding regions which need to be correctly arranged after transcription and may yield many different products because of different arrangement. Regulatory elements appear to be located potentially anywhere,

even inside the gene or several thousands of basepairs away.

As genome length is very limited in the model, so is the gene length. All genes have a fixed length of 32 bp and a predefined layout (see figure 2). Also the promotor sequence is the same for all genes which is not the case in organic cells.
Simulating cellular processes starts with finding the genes on the genome. Because of linear genome, fixed gene layout and predetermined gene length, several criteria have to be met for a gene to exist on the genome. First a promotor sequence has to be identified. Second at least the URE has to fit in front of the promotor. In other words a number of basepairs equal to the length of the URE need to be upstream of the promotor. Third the transcribed sequence, the gene in the stricter sense, is required to exist in full length after the promotor. This is only the case when at least a number of basepairs equal to the length of a gene in the stricter sense lie downstream of the promotor sequence. An equivalent way to express the criteria is to require a subsequence of the length of a gene (wider sense) with an exact match to the promotor sequence at the correct nucleotides.

2.1.4 Transcription

In organic cells the mere existence of a gene does not lead to its transcription and expression. Any one gene in a transcriptionally active area will most likely be transcribed whereas the same gene in a transcriptionally inactive area will be not. Correlation of genomic structure and transcription has been identified but the influence of position and genomic structure on gene transcription still is not fully understood [4, 28, 39].
The product of transcription is RNA. Mostly mRNA (messenger RNA) is synthesised which functions as template for translation and thus as template for proteins. Produced in smaller amounts are rRNA (ribosomal RNA is RNA contained in the ribosome, which is the machinery for translating mRNA to proteins), tRNA (transfer RNA which is necessary to transport

the single amino acids to the ribosome where they are then assembled into a protein) and small regulatory RNAs (snRNA - small nuclear RNA; snoRNA - small nucleolar RNA; siRNA - short interfering RNA; miRNA - microRNA). At the level of transcription so-called transcriptional regulation may occur, leading to no, less or more expression of a gene. This is achieved by transcription factors which may block the assembly of the necessary machinery for transcription at the promotor, or make that assembly more likely. After transcription, regulation termed post-transcriptional takes place. Examples are RNA stability, splicing, editing and others. These mechanisms allow for prevention of an mRNA from ever getting translated, reduction of the amount of translated protein or even alteration of the RNA sequence thus creating a new template [4, 35].

MiniCellSim-Genome transcribes all genes found on the genome, or to be more exact the gene in the stricter sense. The transcribed sequence is not complementary to the gene, as it would be in biological systems, but an exact copy of it. Further all genes produce mRNA and therefore proteins.
Transcriptional regulation is realised by transcription factors (see section 2.1.7). All other forms of regulation are not implemented in the model.

2.1.5 Translation

If mRNA exists long enough and is presented properly to the ribosome, it will serve as a template for a protein. A particular sequence is required near the beginning of the mRNA as a place for the ribosome to assemble at and to initialise translation at the start codon. The information on the mRNA is organised in codons, sequences of three nucleotides, that code for a distinctive amino acid. Although there are 64 different codons only 20 amino acids can be encoded by codons, so some amino acids are encoded by multiple codons. Actually a codon is the complementary sequence to the anti-codon on the tRNA, and by a trick, the wobble of the last base, the anti-codon may match more than one codon, thus less distinct tRNAs are needed. Translation is

mRNA					
	A	T	G	T	T
	A	C	A	G	T
	C	G	G	A	T
	C	A	C	G	T
	C	T	C	T	T
occurrences (A,C,G,T)	2,3,0,0	1,1,1,2	1,2,2,0	1,0,2,2	0,0,0,5
Protein					
	C	T	G	A	T

Table 2: Sample majority rule for translating an mRNA into a model protein with mRNA length 25 and protein length 5.

terminated by one of three stop codons, marking the end of the protein [26]. The sequence and shape of a protein determine its function. Currently it is very difficult to predict the shape of a protein from the sequence and even harder to successfully guess its function.
The amount of protein synthesized from a template is influenced by translational regulation. Post-translational regulation for example alters the degradation rate of proteins or their localisation within the cell.

There is no such thing as a ribosome or even amino acids and tRNAs in MiniCellSim-Genome. Instead a majority rule is applied to the mRNA yielding a shorter sequence (still of the genomic alphabet, thus A, C, G, T) which represents the model protein. Also no particular sequence is required near the beginning of the mRNA, nor start or stop codon.
The majority rule works like this: If for example the mRNA has the following sequence ATGTT ACAGT CGGAT CACGT CTCTT (length 25) and the protein length is 5, subsequences of the mRNA of length 5 are written underneath each other. The number of occurrences of each letter per position is counted. So the letter occurring most often in the first position of all

subsequences (C) is the first letter of the protein. If at any position (3) two or more letters occur the same number of times (C and G), the letter being first if sorted alphabetically takes precedence over the others (C). See table 2 for the complete example.
The protein sequence is then evaluated in a deterministic but arbitrary way so that approximatley half are classified as structural proteins and the other half as transcription factors.
Translational and post-translational regulation are not implemented, the degradation rate is constanct and equal for all proteins.

2.1.6 Structural Proteins

Number and diversity of purpose of structural proteins in a living cell seem inexhaustible. Anything from actin, a protein that can form polymers to allow the cell to actively modify its shape and to move, to citrate synthase, an enzyme catalysing a step in an important biochemical pathway, to rhodopsin, which allows to register photons and thus is the molecular basis of vision, may be considered a structural protein.

However in this work the simulated cells can express only one type of structural protein which does not even have a defined function. It may appear desirable to allow for different structural proteins that for example regenerate the nucleotides used for transcription, translation and replication. However the complexity of such a system is beyond the scope of this diploma thesis.
In biological systems monomers have to be activated by an energy consuming process to make them available for polymerisation. Monomers set free by breakdown also need to be reactivated or degraded. In analogy the model dinstinguishes nucleotides (activated form) and used nucleotides (unpolymerisable form). Nucleotides in their activated form need not be replenished by the simulated cell, as a mechanism is present independent of the genome that converts used nucleotides into the active form at a constant rate.

2.1.7 Transcription Factors

After expression most transcription factors exist in an inactive state, usually located in the cytoplasm as opposed to the genome which resides in the nucleus of the cell. To regulate transcription of one or multiple genes a transcription factor needs to be activated and relocated to the nucleus. Activation and relocation are often induced by distinct events as a result of the triggering of the signal transduction cascade. The latter describes a system where a preexisting set of proteins recognise and produce signals. The first protein, a receptor – usually a transmembrane protein on the cell surface, recognises a distinct event or signal outside the cell and produces a signal inside the cell. This signal in turn activates a protein within the cell, which produces its own signal. Many such internal signals are generated and registered, before the latter elements of the cascade produce a less transient change e.g. activate transcription factors. Much like the snowball effect this system gives signal amplification, because the outside signal may activate several receptors, each receptor activates several other proteins where each in turn activates some more proteins, and so on [7, 29, 37].
For a transcription factor to influence a gene it is required to bind to one or more regulatory elements of the gene and/or change the genomic structure in the vicinity of the gene.
The main reason for transcription factors and signal transduction cascades appears to be the comparative ease with which a cell may switch on those parts of its genetic programme needed for survival in given circumstances and to switch off what is not required [7].

As mentioned above, the model features transcription factors, but no equivalent to a signal transduction cascade. MiniCellSim-Genome assumes all transcription factors to be active at all times and in a position to modify transcription of all target genes. Any gene can only be modified thus, if the transcription factor is able to interact with the URE of the same gene. To determine if an interaction is activating (= positive, increasing the rate of

transcription) or repressing (= negative, decreasing the rate of transcription) a majority rule is done for the URE and the protein (without the need of subsequences as both are of the same length). The resulting sequence is then evaluated in an arbitrarily chosen way to decide for activation or repression. The strength of interaction and thus the strength of activation or repression is calculated by cofolding the URE and the protein with RNAcofold (see appendix B).

2.2 Model-Independent Properties

2.2.1 The Ideal Genome

As already described, the model thrives on a great number of approxmiations. Ideally however the model should more or less have the same constraints as appear for real cells (eukaryotic cells to be more precise). If such a very lifelike model were used it would have a genome which would be very large in comparison to gene length.
For purposes of measuring similarity of the genome used in the model and what the genome sould look like for a biological cell, the ideal genome is used. This ideal genome is so large that the regions near the rim of the genome which cannot contain genes have no impact whatsoever on the overall gene number.
In our model the existence (or absence) of a promotor determines if and where genes are located on the genome. Therefore μ_{ng}^{id}, the mean number of genes per kbp for the ideal genome, can be calculated from the probability of occurrence of a promotor P_{Prom}. This is true for all genomes that may fit at least one gene onto their genome and where genes may overlap.
Let us start with a simple case that allows to identify an ideal genome. If $gl > L$, the genome length, the above mentioned situation occurs. Not even a single gene will fit onto the genome so μ_{ng} has to be 0.
However this is not true for circular genomes. Even if the genome consists of only a single basepair this is equivalent to an infinitely long linear genome

of the given nucleotide. Further if any subsequence serves as a promotor a gene can be found. Although a circular genome may be seen as an infinitely long linear one the maximum number of distinct genes is L. If $gl \leq L$ and the genome is circular, it is also ideal. So every circular genome is ideal.
Even though overlapping of genes is very frequently found on circular and linear genomes of viruses, it should be mentioned that this increase in coding density is bought by decreasing coding flexibility. Any change in a position that is part of more than one gene will more likely cause a mutant with lesser fitness. Therefore it is less likely that a variant with a mutation at this position will survive.

For linear genomes there exists a geneless region of length $gl - 1$. Even if a promotor occurs within this region no gene is located there because it would not fit onto the remaining basepairs. The shorter such a geneless region is compared to the genome length L, the smaller is its influence on μ_{ng} and the better the approximation of an ideal genome. Therefore the ideal genome is only theoretically possible, when the geneless region is extremely short and the genome is very large.
In mathematical terms this reads as follows:

$gl \to 0$ and / or

$L \to \infty$

2.2.2 Number of Genes

The number of genes ng in a given genome of random sequence can be calculated from the genome length L, the gene length gl and the probability of a promotor occuring at any one position P_{Prom} by the following formula:

$$ng = (L - gl + 1) \times P_{Prom}$$

Obviously the increase in gene number correlates linearly with the increase in genome length. To measure proximity of the given genome to the ideal

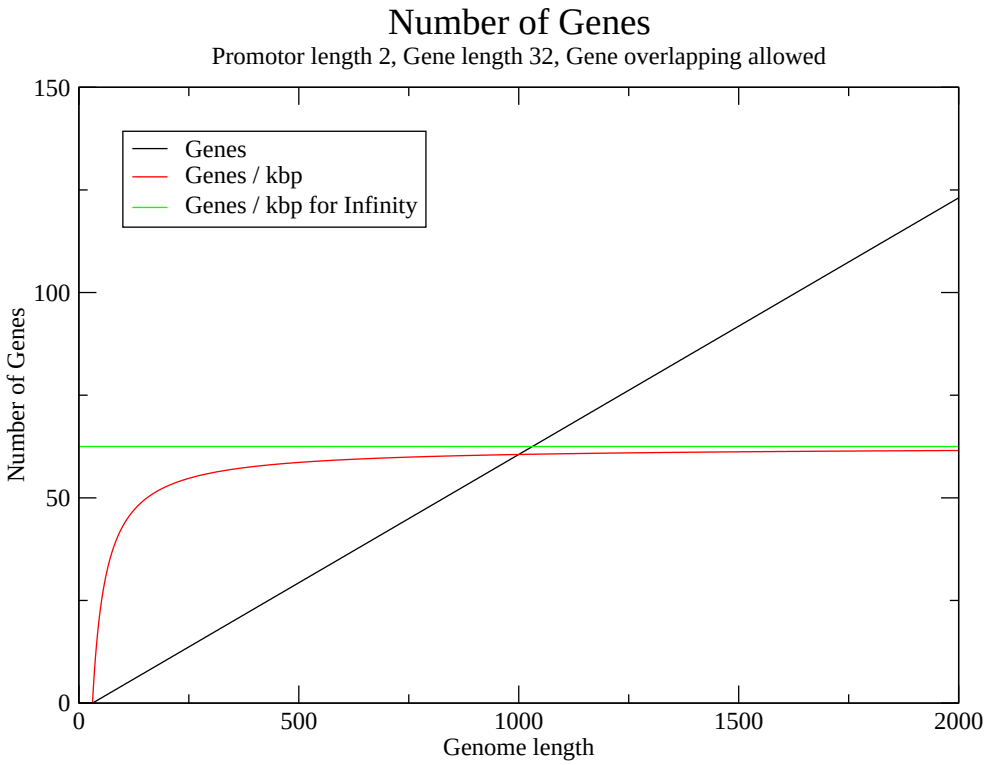

Figure 5: Number of genes in dependency of the genome length.

genome the mean number of genes per kbp μ_{ng} is used, which is easily calculated:

$$\mu_{ng} = \frac{ng}{L} \times 1000$$

With increasing genome length the values of μ_{ng} asymptotically approach the μ_{ng}^{id} for the ideal genome, which is:

$$\mu_{ng}^{id} = P_{Prom} \times 1000$$

Even though the ideal genome is a good approximation of genomes, it differs considerably from linear genomes that have either a small genome length L or a large gene length gl. The sequences used in this work range from 50 to 150 basepairs and therefore the genomes are far from ideal.

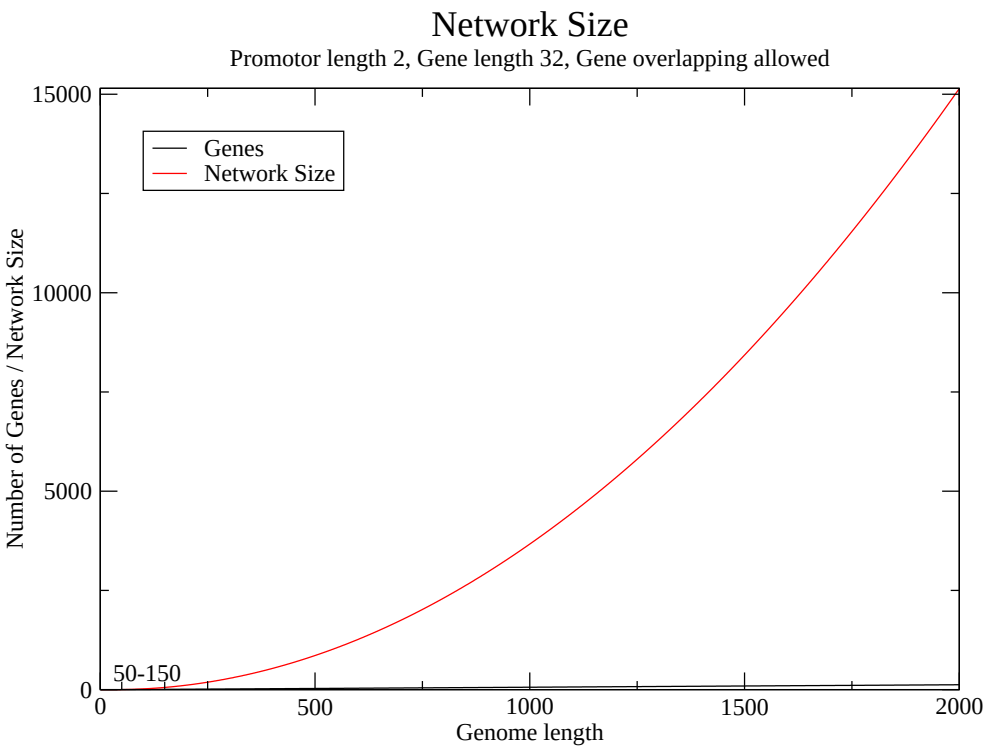

Figure 6: Network size in dependency of the genome length.

2.2.3 Network Size

Every genome has a gene and gene product interaction network, called the reaction network. In the simplest cases with no genes present on the genome the network contains no vertices and no edges. If one or more genes are present there are some vertices and edges representing transcription and translation for every single gene, but on top of that there are vertices and edges representing the formation of activating or repressing complexes that may regulate the transcription. The maximum number of such activating or repressing complexes is present if every single gene product is capable of forming such a complex and every complex regulates all genes. This maximum number calculates as ng^2.

So whereas the number of genes increases linearly with the genome length, the size of the corrsponding gene interaction network increases quadratically with the number of genes, or with the genome length as follows from their linear interrelation.

2.2.4 Shadow

First an arbitrary genome is used as starting point, referred to as parent. Then the shadow of a parent is the sum of all genomes that differ in one position from the parent. Every distinct parent has a unique shadow and there is only one shadow per parent.
Another way to describe the shadow is to call it the sum of all 1-point mutants or all error 1 neighbours. To compare arbitrary strings the Hamming distance can supply a measure of similarity by summing up the number of positions that differ in the given strings. Error 1 neighbours of a string are all strings that differ in exactly one position from the reference string.

As in all known cells the model uses DNA with the Alphabet $\mathcal{A} = \{A, C, G, T\}$. In any one position there may be $|\mathcal{A}| - 1$ alternative letters and multiplied by the number of positions, the genome length L respectively, this yields the size of the shadow:

$$|\mathcal{S}| = (|\mathcal{A}| - 1) \times L$$

With the size of the Alphabet $|\mathcal{A}| = 4$ for DNA the size of the shadow calculates as $|\mathcal{S}| = 3 \times L$.
Ideally the question of stability and neutrality of a network derived from a genome should be concluded from exhaustive study. So every possible sequence for a given length (which would be $|\mathcal{A}|^L$ sequences) should be analysed. However even for small genomes, such as used in this work with a length from 50 to 150 basepairs, handling of all sequences contained in the shadow is a problem.
An approximation can be given by using the shadow. Therefore the question

of how likely the network changes due to even a single mutation is answered. Using the shadow has other consequences too:

- As only one mutation occurs, the original sequence / the parent cannot be the result of the mutation, in contrast to application of two or more mutations where the latter alterations might reverse the effects of the earlier ones.

- All derived sequences are unique, which is not the case with multiple mutations, as many permutations of the mutation order will give the same result.

- Even though only one mutation occurs, more than one gene may be affected in case genes may overlap. As μ_{ng}^{id} was calculated above, we can compute the mean number of genes any single mutation will affect μ_{1p} by multiplying with the gene length gl.

  $$\mu_{1p}^{id} = P_{Prom} \times gl$$

  Throughout this work gene length $gl = 32$ and promotor probability $P_{Prom} = \frac{1}{16}$ are used, thus giving $\mu_{1p} = 2$.

2.3 Phenotype Measure

The MiniCellSim-Genome model allows for easy evaluation of reaction network and kinetic laws. As a consequence those are the characteristics used to define the phenotype. Basically that means the phenotype has a "what" component, the reaction network showing all genes and gene products and all interactions, and a "how much" element, as the quantity of all gene products can be computed by integrating the kinetic laws.
For practical purposes continuous functions like the kinetic laws are compared by solving at various points. The results of the two functions for the same point are then compared.
Two genomes should be considered equivalent, and mutations that transform

one into the other as neutral, if and only if the reaction network and all kinetic laws are identical. Put another way neutral mutations change neither what genes are present nor how much gene product is derived from them. This gives rise to a set of features to define the phenotype, referred to as TFD:

T network

F functional network

D dynamics

Each of the above features is evaluated and denoted 0 if the two compared genomes share this part of the phenotype and 1 otherwise. In this work all genomes of the shadow are compared to the parent yielding a three digit code that gives a rough idea of the differences.

2.3.1 T — Network

T is a description of the reaction network plus identifiers for all genes. The important fact here is that each vertex representing a gene or a gene product can be traced back to its origin on the genome. The origins of the genes are neither important for the reaction network nor for the phenotype. So T actually evaluates data irrelevant for the phenotype. Nevertheless the inclusion of this feature may yield valuable information when analysing the effects of mutation.
Obviously if after mutation the same set of genes is present and all vertices in the reaction network can be traced back to the same origins as before, T will be considered identical, thus denoted 0. Otherwise T will be 1, when the set of genes changes, which is the case if genes / origins are lost or newly created and / or vertices in the network are traced to different origins. This holds true even if the reaction network and kinetic laws are identical and therefore mutation was neutral. In this later case with T = 1 but mutation being neutral, the reason for evaluating T becomes clear. Instead of always

attributing lack of change of the phenotype to lack of change in expression of the genome (as with T = 0), the much more unlikely occurrence of a compensatory mutation can be explicitely detected (T = 1). Such a compensatory mutation is present if all genes lost or changed have been substituted for by other genes created or altered.

2.3.2 F — Functional Network

This feature is the genuine measure of the phenotype. F evaluates the reaction network without caring for the origins of the vertices / genes. The actual comparison of two networks is done by reactions (see appendix C). As the operation is asymmetrical (vertices and edges lost will be noticed as differences, those newly created will not) the two genomes have to be compared both ways to identify all dissimilarities.

2.3.3 D — Dynamics

As pointed out before the same reaction network can lead to very different expression rates in the simulated cell depending on the kinetic laws. By integrating all equations with SOSlib (see appendix D) the quantities of every single entity are calculated for 10 points in time. In this work the requirements for neutrality of the quantities are rather harsh as all quantities of all entities at all calculated points in time have to be identical. Please note that theoretically the functions could fit all points derived by SOSlib without being identical. However this is very unlikely to occur.

2.4 Typing of Mutations

2.4.1 Localisation

The impact of any given mutation on a single gene is largely determined by its position relative to the gene. A gene is defined as a unit on the genome that can give rise to one or more biopolymers (RNA and proteins), which serve a certain function in the organism. This usally includes a continuous

string beginning at the promotor or a site capable of initiating transcription and ending at a terminator. Elements exist that can modify the rate of transcription and expression of a gene, but mostly are not counted as part of a gene, even if these elements are located within the gene. In our simplified approach we have a fixed layout for the gene (as shown in figure 2). First is the upstream regulatory element (URE), adjacent to which lies the promotor and then the gene in the stricter sense. The URE, the promotor and the gene in the stricter sense all have a fixed length. Any mutation outside of the boundaries of the gene (wider sense) can only influence this gene if the mutation lies within the boundaries of a second gene, which acts as a transcription factor on the first. The conclusion is that any mutation located outside the boundaries of all genes cannot affect any gene. Please note that if a new gene is created which may happen inside or outside of boundaries of preexistent genes, the mutation is per definition inside of gene boundaries, namely the boundaries of the gene it created.
This leads to the distinction of 4 cases of localisation:

1. Intergenic region
2. Non-intergenic region
   - 2a. URE region
   - 2b. Promotor region
   - 2c. Gene region

Intergenic region — By definition if a mutation is not in the URE, promotor or gene region of any gene, it has to be located in the intergenic region. As this definition is derived by exclusion a mutation with location intergenic region cannot be located anywhere else in regard to another gene.
To calculate the likelihood of a mutation to be intergenic three cases must be distinguished:

1. $gl > L$
   If the gene length gl is greater than the genome length L all mutations

will be intergenic since there cannot exist even a single gene on the genome.

$$P_{ig}^{M} = 1$$

2. $gl > L_{eff}$
   The effective genome length L_{eff} equals the number of positions where if a promotor occurs a gene can be fitted onto the genome. In other words it is the genome length L minus the geneless region in linear genomes (as discussed in section 2.2.1).

   $$L_{eff} = L - gl + 1$$

   With this equation the above condition $gl > L_{eff}$ can be redrafted to $2 \times gl > L + 1$.
   The first position of the genome can only be within one gene, if a gene starts right at the beginning of the genome. Similarly the last position can only be part of one gene, if a gene reaches right to the end of the genome. For both positions the likelihood of beeing within a gene depends on only one gene, or in other words on one position where a promotor has to occur in order for the first or last position to be within the gene. This is inverse to the likelihood of a promotor not being in that specific place, which is $1 - P_{Prom}$. The second and the second last position are part of a gene if in one or both of two positions a promotor occurred. For both positions the likelihood to be intergenic is $(1 - P_{Prom})^2$. By analogy the same is true for the positions at maximum $L_{eff} - 1$ from an end of the genome. To give an approximate probability the mean of the values is calculated. So far the formula reads

   $$P_{ig}^{M} \approx \frac{\sum_{i=1}^{L_{eff}-1}(1 - P_{Prom})^i \times 2}{(L_{eff} - 1) \times 2}$$

   There still are $L - [(L_{eff} - 1) \times 2]$ positions not evaluated in the above approximation. The number of positions can be redrafted to $2 \times gl - L$

(please note that this formula is only true for $gl > L_{eff}$, which is $2\times gl > L+1$). As the positions are in the middle of the genome they are most likely to be within a gene. To be exact a maximum number of L_{eff} genes may occur which all would encompass these middle positions. The probability for any such middle position not to be part of a gene is $(1 - P_{Prom})^{L_{eff}}$. The full formula is

$$P_{ig}^{M} = \frac{(2 \times gl - L) \times (1 - P_{Prom})^{L_{eff}} + \sum_{i=1}^{L_{eff}-1}(1 - P_{Prom})^{i} \times 2}{L}$$

3. $gl < L_{eff}$
   The above calculations for positions near to one of the ends of the genome also hold true for this case. However the maximum distance to the rim is no longer determined by L_{eff} but by gl, giving

$$P_{ig}^{M} \approx \frac{\sum_{i=1}^{gl-1}(1 - P_{Prom})^{i} \times 2}{(gl - 1) \times 2}$$

   Now $L - 2 \times (gl - 1)$ middle positions exist. This expression may be stated as $L_{eff} - gl + 1$ as well. In this environment the maximum number of genes is gl yielding the probability for these positions not to be within a gene as $(1 - P_{Prom})^{gl}$. Thus the formula reads

$$P_{ig}^{M} = \frac{L_{eff} \times (1 - P_{Prom})^{gl} + \sum_{i=1}^{gl-1}(1 - P_{Prom})^{i} \times 2}{L}$$

Non-intergenic region — The likelihood of a mutation to be non-intergenic is one minus the probability of a mutation to be intergenic. For this work all non-intergenic mutations have been split up in the three subcategories.

$$P_{nig}^{M} = 1 - P_{ig}^{M}$$

URE region — Mutations in the URE are expected to either change the influence of transcription factors on the corresponding gene or alter nothing at all.
As the URE has a fixed length any mutation is positioned there with a chance of the fraction URE length ul divided by gene length gl. Because

of overlapping any non-intergenic mutation may have multiple positions and hence increases the likelihood of every single possible localisation.

$$P^M_{URE} = P^M_{nig} \times \frac{ul}{gl} \times P_{Prom} \times gl$$

$$P^M_{URE} = P^M_{nig} \times ul \times P_{Prom}$$

Promotor region — If the promotor region is affected by mutation the result is either gene gain or gene loss. In systems with more than one promotor sequences a change in transcriptional activity would be possible, but not in this simplified model.
Pricipally the amount of promotor region mutations depends on the same factors as that of URE region but an additional term has to be added to include the gene gain mutations, which have to be located in the promotor.

$$P^M_{Prom} = P^M_{nig} \times pl \times P_{Prom} + P^{GN}_{gain}$$

Gene region — A change in this region can lead to potentially any kind of different behaviour. Classes may switch, and for transcription factors targets may change and interaction strengths may vary.
What was true for the URE is true for the gene. The fraction of gene region is gene length (strict) gnl divided by gene length (wide) gl.

$$P^M_{gene} = P^M_{nig} \times gnl \times P_{Prom}$$

2.4.2 Gene change

Gene creation — Wherever sequences similar to a promotor exist on the genome a single point mutation might be sufficient to transform this sequence into an actual promotor. Even though it is unlikely the possibility exists that such a point mutation may create more than one promotor and thus more than one gene. The maximum number of genes a single mutation can create is given by the promotor length pl. Throughout this work $pl = 2$ has been used.
With given promotor length pl, and thus P_{Prom}, the likelihood of a mutation

creating at least one gene can easily be calculated by calculating the reverse, the probability of a mutation not to create a gene.

$$P_{gain}^{GN} = P_{Prom} \times (1 - P_{Prom}) \times pl$$

Gene loss — Not only may promotors be created by a mutation but destoyed as well. The maximum number of genes affected this way is the same as with promotor creation ergo pl. Again the probability of at least one gene loss by point mutation is easily calculated.

$$P_{loss}^{GN} = P_{gain}^{GN}$$

Class switch — If the mutation is located wihtin the gene in the stricter sense the gene product, the protein, may change class. As mentionend in section 2.1 the model distinguishes between transcription factors and structural proteins. The more targets a transcription factor had before or gains after the switch, the bigger the impact on the reaction network.
Current versions of MiniCellSim-Genome classify a gene by factors which are wholly dependent on the sequence of this single gene. As the classification process is scheduled to be changed in upcoming versions no probability has been computed for this event.

Illegal gene creation — In the given environment and model all mutations that give rise to new genes ought to be located in the promotor. To be more precise the mutation should be situated within the promotor of the recently created gene. If it is not, an illegal gene creation occurred. Even though this should never occur, as a failsafe and quality control a check for illegal gene creation is made.
No instances of illegal gene creation have been documented in the experimental data evaluated for this work.

Illegal gene loss — The same principles hold true for illegal gene loss. If a gene is lost and the mutation is not located within the promotor of the recently lost gene the gene loss is illegal. Checks to identify such occurrences

are made.
The experimental data used contains no instance of illegal gene loss.

2.4.3 Transcription Factors

There are some properties unique to transcription factors in contrast to structural proteins. Transcription factors usually have a bigger impact on the reaction network as they may vary the expression of any gene of the genome. This transcription level control is the reason for the quadratical increase in network size.
Two components determine the effect of a transcription factor on the target gene. The first is the interaction type, which can be either activating (CX) or repressing (CI). Activating complexes always increase the rate of transcription whereas repressing complexes decrease it. The second component is the interaction strength also called activity, which defines the quantity of increase or decrease. Since activity is derived by RNAcofold B predictions of effects by single point mutations are very hard to calculate, which is why probabilities have been omitted.

For every pair of transcription factor and target gene the following events are mutually exclusive.

Target gene loss — Complexes with activators and repressors are only assembled on UREs. In this model, per definition, UREs only occur where there is a gene. The consequence is that gene loss results in loss of all regulatory activities on that gene. Think of the reaction network as a directed graph with every gene as a vertex and every edge between vertices as a regulatory activity of one gene on another. If a gene is lost, a vertex is removed and therefore all edges connected to that vertex (the regulatory activities) are erased as well.
As any gene may be under regulatory control, the loss of a target gene is approximately as likely as the loss of any gene, because even with genomes

of a small size most genes are subject of regulatory control.

Target gene gain — Whenever a new gene is created another URE appears where regulatory complexes may become effective. Depending on the number of transcription factors and the type of the newly created gene new regulatory activies may be exhibited.
Again any newly created gene may be target of expression control by transcription factors, so the likelihood of target gene gain is again approximately equal to the probability of gene gain.

Decrease of activity on target gene — If a transcription factor or an URE is affected by the mutation, the strength of interaction of the regulatory complex and the URE may be altered. If located in an URE all activity on that gene may be affected, if located in a transcription factor all activity by the factor may be altered.
Please note that both an activating complex that cannot activate as strong as before as well as an repressing complex which represses less will be considered a decrease of activity.
As the strength and type of activity of a given transcripiton factor on a gene is calculated by cofolding, it is very hard to estimate the probabilities of events resulting from these calculations.

Loss of activity on target gene — A special case of decrease of activity is the complete loss of it. The target gene then is no longer under regulatory control by the given transcription factor.

Increase of activity on target gene — Again if located in an URE or transcription factor, a mutation may change regulatory activity. Just as the control may be diminished, alternatively it may be amplified.

Gain of activity on target gene — A special case of increase of activity,

gain of activity, occurs when before the mutation the transcription factor had no influence on the target gene.

Switching of activating and repressing complexes — Some mutations may reverse the effects of a transcription factor on a given gene, activating what has been repressed and vice versa. Any such occurrence is termed a switch regardless of the strength of the activating or repressing action as long as both are greater than zero (otherwise it would be loss of activity or gain of activity).

3 Results on Neutrality in the Shadow

...man will occasionally stumble over the truth, but usually manages to pick himself up, walk over or around it, and carry on.

— Winston Churchill

3.1 Data Sets

For calculations three data sets have been used, denoted set A, B and C.

Set A — Consisting of genomes from the length L of 50 to 150 of 100 samples each. This totals to 1100 genomes.

L	50	60	70	80	90	100	110	120	130	140	150
Samples	100	100	100	100	100	100	100	100	100	100	100

Table 3: Samples in set A.

Set B — Consisting of 1000 genomes of the length 100.

Set C — Consisting of 1000 genomes of the length 100. Other than containing different sequences sets B and C are identical.

Genome Length L	100
Sample Size	1000

Table 4: Samples in set B and C.

3.2 Number of Genes

In section 2.2.2 the statistically expected mean number of genes ng for certain genome lengths has been computed. Table 5 compares expected and actual mean ng for sets A, B and C. The data indicates the correctness of the formula for ng.

Set	L	Sample Size	calculated ng	calculated σ_{ng}	calculated $\sigma_{ng}\%$	exp. ng	Δng	cal. μ_{ng}	exp. μ_{ng}
A	50	15100	1.23 ± 1.04		84.80%	1.19	−0.04	24.6	23.8
A	60	18100	1.59 ± 1.12		70.44%	1.81	+0.22	26.5	30.2
A	70	21100	2.31 ± 1.28		55.61%	2.44	+0.13	33.0	34.8
A	80	24100	2.67 ± 1.46		54.56%	3.06	+0.39	33.4	38.3
A	90	27100	3.55 ± 1.77		49.75%	3.69	+0.14	39.4	41.0
A	100	30100	4.57 ± 1.99		43.46%	4.31	−0.26	45.7	43.1
B	100	301000	4.29 ± 2.07		48.19%	4.31	+0.02	42.9	43.1
C	100	301000	4.30 ± 1.94		45.10%	4.31	+0.01	43.0	43.1
A	110	33100	5.32 ± 2.15		40.42%	4.94	−0.38	48.4	44.9
A	120	36100	5.58 ± 2.09		37.54%	5.56	−0.02	46.5	46.3
A	130	39100	6.02 ± 2.17		36.04%	6.19	+0.17	46.3	47.6
A	140	42100	6.61 ± 2.53		38.34%	6.81	+0.20	47.2	48.6
A	150	45100	7.28 ± 2.45		33.71%	7.44	+0.16	48.5	49.6

Table 5: Number of genes for sets A, B and C. Set describes which set the samples are from. L is the genome length. Sample size is the number of genomes used from which values have been calculated. The next three columns calculated ng, σ_{ng} and $\sigma_{ng}\%$ are calculated from the samples, ng being the mean number of genes, σ_{ng} the standard deviation and $\sigma_{ng}\%$ the standard deviation in percent of the mean value ng. Exp. ng lists the expected mean number of genes for the given length. Δ_{ng} shows differences in expected and calculated ng. Cal. μ_{ng} is the calculated mean number of genes per kbp, whereas exp. μ_{ng} describes the expected mean number of genes per kbp.

The sample size from which the mean number of genes is calculated corresponds to the sample size of the sets. As the number of genes is computed for each genome (i.e. the whole shadow and the parent) each set or subset has $3 \times L + 1$ genomes per parent, where L is the genome length.

3.3 Neutrality

Next comes the analysis of neutrality of the mutants in the shadow in respect to the phenotype. In the model the phenotype is measured in TFD, alas network, functional network and dynamics, each assigned either 0 or 1. Through all sets only three of five possible cases of phenotypic difference between mutant and parent have been identified. They occur with the same frequency in all sets. It has to emphasised that only differences in phenotype of mutant and parent are checked. Mutants of a single shadow might be neutral to each other but were not analysed for that possibility.

3.3.1 Cases found

Case 000 — All three features being 0, literally nothing has changed. Although mutant and parent differ in one point mutation the phenotype is exactly the same.

Case 001 — Here only the dynamics have been affected by the mutation. All genes are still present and of the same type, but a change in at least one interaction between a transcription factor and its target gene has changed the quantity of one or more gene products.

Case 111 — One or more genes have been lost, gained or switched type. In addition the reaction network, formed by the genes and gene products, has changed. Because of that the dynamics necessarily were altered as well.

3.3.2 Cases not found

Case 011 — This case would only be seen, if a mutation caused the loss or the gain of an interaction between a preexistent transcription factor and a preexistent target gene. The conclusion is that a single mutation either within the transcription factor or the URE does not suffice to gain or loose interaction.

	Occurrences						
TFD	total	%	mean		σ		σ_{ln}
Set A							
000	188672	57.17%	171.52	±	54.04	×	1.38
001	5129	1.55%	4.66	±	5.34	×	2.85
111	136199	41.27%	123.82	±	68.68	×	2.21
Sum	330000	100.00%					
Set B							
000	171923	57.31%	171.92	±	43.68	×	1.31
001	4885	1.63%	4.88	±	5.11	×	2.76
111	123192	41.06%	123.19	±	41.01	×	1.49
Sum	300000	100.00%					
Set C							
000	171938	57.31%	171.94	±	41.95	×	1.29
001	4857	1.62%	4.86	±	5.16	×	2.78
111	123205	41.07%	123.20	±	39.28	×	1.46
Sum	300000	100.00%					

Table 6: Phenotypes of sets A, B and C. Please note that a both a normal distribution was calculated using σ and the in some cases more useful lognormal distribution, where the mean is multiplied or divided by σ_{ln}.

Case 101 — In this situation at least two genes are affected by mutation, but in respect to the reaction network all changes are compensated for by each other. Because of altered gene names the dynamics are considered different.

3.3.3 Cases impossible

Cases 010 and 110 are actually impossible. The reason for that is based on the fact that an altered reaction network necessarily results in changed dynamics. 010 and 110 however would indicate altered network and unchanged dynamics.
The case 100 cannot occur because the comparison of dynamics is bound to the names of the genes and their products. A change in network necessarily results in a change of name for at least one gene therefore yielding different dynamics.

3.4 Mutation Types

Section 2.4 dealt with the different types of possible mutations. With the three sets of data the occurence of each distinct mutation was quantified. As with neutrality in the above section the frequency for mutation types are very similiar to each other between the sets.

The interpretation of the data from table 7 is somewhat involved. First the seemingly odd sums are evident. Please note that the sums do not have to sum up to any specific value as genes may overlap. Any position that is part of two or more genes will show up two or more times in this statistic.
Gene (strict) and URE are strictly linked in their behaviour, because any gene (wider) contains both. As lengths are fixed the relative amount of gene (strict) to URE is the same for a genome with one gene as a genome with any number of genes. Even though the promotor is part of the gene (wider) as well its behaviour is not linked. This results from an additional impact by the mutations that generate new genes.

Localisation					
	Occurrences				
Location	total	%	mean	σ	σ%
Set A					
Intergenic	94232	16.49%	85.67 ±	50.76	59.26%
Gene (strict)	350475	61.34%	318.61 ±	205.85	64.61%
Promotor	56563	9.90%	51.42 ±	26.19	50.93%
URE	70095	12.27%	63.72 ±	41.17	64.61%
Sum	571365	100.00%			
Set B					
Intergenic	86011	16.40%	86.01 ±	54.20	63.01%
Gene (strict)	322050	61.42%	322.05 ±	155.19	48.19%
Promotor	51850	9.89%	51.85 ±	12.07	23.28%
URE	64410	12.28%	64.41 ±	31.04	48.19%
Sum	524321	100.00%			
Set C					
Intergenic	85804	16.35%	85.80 ±	51.84	60.42%
Gene (strict)	322650	61.49%	322.65 ±	145.52	45.10%
Promotor	51716	9.86%	51.72 ±	11.56	22.35%
URE	64530	12.30%	64.53 ±	29.10	45.10%
Sum	524700	100.00%			

Table 7: Comparison of mutation localisation between the sets A, B and C. Location gives the localisation of the mutation. Total states the number of occurrences in the given location; % the relative amount of mutations with specific location for the set; mean gives the mean number of occurrences of the particular location; σ the standard deviation of the mean; σ% the standard deviation in percent of the mean.

Standard deviation σ appears broader for all non-intergenic locations in set A than in sets B and C. Whereas the latter only contain genomes of length 100, set A also has smaller genomes. As the length diminishes the distribution of number of genes ng is more roughed and causes greater standard deviation.

Again in table 8 standard deviation σ is broader in set A than in the other sets and this is so for the same reason. Of greater interest is the difference between gene gain and gene loss. The two events should be equally frequent

Gene change					
	Occurrences				
Mutation	total	%	mean	σ	σ%
Set A					
Gene loss	28038	18.46%	25.49 ±	16.47	64.61%
Gene gain	28525	18.78%	25.93 ±	12.47	48.08%
Class switch	95315	62.76%	86.65 ±	57.61	66.48%
Sum	151878	100.00%			
Set B					
Gene loss	25764	18.47%	25.76 ±	12.42	48.19%
Gene gain	26086	18.70%	26.09 ±	3.98	15.25%
Class switch	87643	62.83%	87.64 ±	45.82	52.28%
Sum	139493	100.00%			
Set C					
Gene loss	25812	18.44%	25.81 ±	11.64	45.10%
Gene gain	25904	18.51%	25.90 ±	4.03	15.57%
Class switch	88252	63.05%	88.25 ±	43.67	49.49%
Sum	139968	100.00%			

Table 8: Comparison of mutations causing gene change in sets A, B and C. See table 7 for explanantions of column headers

and that they are, however σ is very different for them. Gene gain is basically independent of what the parent looks like, whereas gene loss depends on the number of genes existing in the parent. The variances in ng level out the frequency but singular cases are much more likely to be off the mean for gene loss than the independent gene gain.

For table 9 (next page) differences in standard deviation σ betweens Sets A and B, C as well as those between target gene gain and target gene loss remain present, their cause already explained above. On top of that the very high σ for all mutations is a very obvious result. As gene loss depends on the number of genes ng, thus being more variant, all mutations of the transcripiton factor type depend on the reaction network which in turn depends on ng, causing vast differences in singular cases.
Because the σ is greater than the mean in most cases, the probably more useful distribution is the lognormal one with σ_{ln}. Instead of adding or subtracting σ from the mean, it is multiplied or divided by σ_{ln}.

3.5 Correspondence of Mutation Types and Phenotypes

Analysis of mutation types is only a preliminary step to the search of correlation between mutation type and phenotype. As some correlations are to be expected because of the model underlying the simulation the results may also serve as quality control.

3.5.1 Localisation

According to the TFD model phenotype 000 is characterised by total absence of change. Therefore it is a small surprise, that most mutations simply are outside all gene boundaries and thus intergenic. More important is the result of the reversed point of view: All mutations outside of all gene boundaries

Transcripition factors							
	Occurrences						
Mutation	total	%	mean		σ		σ_{ln}
Set A							
Target gene loss	64673	30.07%	58.79	±	82.74	×	6.77
Target gene gain	66369	30.86%	60.34	±	65.63	×	5.73
Activity switch	25791	11.99%	23.45	±	39.89	×	5.61
Activity loss	17548	8.16%	15.95	±	23.56	×	4.71
Activity decrease	11424	5.31%	10.39	±	17.45	×	4.12
Activity gain	17355	8.07%	15.78	±	21.92	×	4.33
Activity increase	11910	5.54%	10.83	±	17.81	×	4.06
Sum	215070	100.00%					
Set B							
Target gene loss	50396	29.44%	50.40	±	62.74	×	4.76
Target gene gain	50670	29.60%	50.67	±	35.57	×	3.70
Activity switch	22081	12.90%	22.08	±	35.10	×	5.17
Activity loss	14190	8.29%	14.19	±	19.46	×	4.20
Activity decrease	10318	6.03%	10.32	±	16.43	×	3.96
Activity gain	13630	7.96%	13.63	±	16.33	×	3.60
Activity increase	9901	5.78%	9.90	±	14.67	×	3.75
Sum	171186	100.00%					
Set C							
Target gene loss	47720	29.34%	47.72	±	50.86	×	4.86
Target gene gain	49284	30.31%	49.28	±	34.46	×	3.91
Activity switch	20345	12.51%	20.34	±	28.95	×	4.96
Activity loss	13818	8.50%	13.82	±	17.01	×	4.11
Activity decrease	9221	5.67%	9.22	±	13.30	×	3.82
Activity gain	12900	7.93%	12.90	±	14.74	×	3.60
Activity increase	9333	5.74%	9.33	±	12.80	×	3.67
Sum	162621	100.00%					

Table 9: Comparison of mutations affecting transcription factors in sets A, B and C. σ is the standard deviation for a normal distribution, σ_{ln} for the lognormal distribution. See table 7 for explanantions of other column headers.

Phenotype 000 — Localisation							
	Occurrences						
Location	total	%	mean	σ	σ%	T→M	M→T
Set A							
Intergenic	94232	38.62%	85.67±	50.76	59.26%	100.00%	49.94%
Gene	130277	53.40%	118.43±	62.10	52.43%	37.17%	44.75%
URE	19476	7.98%	17.71±	11.86	67.00%	27.79%	9.63%
Sum	243985	100.00%					
Set B							
Intergenic	86011	38.36%	86.01±	54.20	63.01%	100.00%	50.03%
Gene	119598	53.34%	119.60±	47.56	39.76%	37.14%	44.71%
URE	18609	8.30%	18.61±	11.64	62.54%	28.89%	10.01%
Sum	224218	100.00%					
Set C							
Intergenic	85804	38.13%	85.80±	51.84	60.42%	100.00%	49.90%
Gene	120489	53.54%	120.49±	45.38	37.66%	37.34%	44.81%
URE	18742	8.33%	18.74±	11.79	62.91%	29.04%	9.98%
Sum	225035	100.00%					

Table 10: Correlations between phenotype 000 and localisation of the mutation for sets A, B and C. Location gives the localisation. Total gives the number of occurrences of the specific combination; % the relative amount of localisation for the phenotype; mean states the mean number of occurrences of the combination; σ the standard deviation of the mean; σ% the standard deviation in percent of the mean. T→M, M→T: e.g. first line: 100.00% and 49.94% means that 100% of intergenic mutations are found correlated to the phenotype 000, but only 49.94% of phenotype 000 have an intergenic mutation.

lead to no change in the phenotype, as they have to because of the model used.
A considerable number of mutations in genes and UREs also lack any effect on the phenotype. These are truly neutral mutations.

Phenotype 001 means only the dynamics are altered. Obviously transcription factors (genes) and UREs are the locations where mutations are expected that cause this phenotype. Expectations prove true, but with a suprising twist: All 001 phenotypes have a mutation in the URE which might indicate that in all the samples not a single point mutation was capable of changing a transcription factor in a way that altered its function.

Phenotype 001 — Localisation

	Occurrences						
Location	total	%	mean	σ	σ_{ln}	T→M	M→T
Set A							
Gene	4374	43.39%	3.98	± 6.29	× 2.92	1.25%	52.02%
URE	5707	56.61%	5.19	± 5.97	× 2.97	8.14%	100.00%
Sum	10081	100.00%					
Set B							
Gene	4016	42.53%	4.02	± 6.01	× 2.90	1.25%	50.42%
URE	5426	57.47%	5.43	± 5.68	× 2.89	8.42%	100.00%
Sum	9442	100.00%					
Set C							
Gene	3901	42.15%	3.90	± 5.83	× 2.84	1.21%	48.63%
URE	5353	57.85%	5.35	± 5.69	× 2.89	8.30%	100.00%
Sum	9254	100.00%					

Table 11: Correlations between phenotype and location of the mutation for set B. See table 10 and 9 for explanation of column headers.

If the phenotype is 111 network and dynamics have changed. The model demands every mutation in a promotor to change the network and result in phenotype 111, which table 12 shows to be the case. Also changes in genes that switch the class are required to produce this phenotype, which presumably is the case in some of the mutations localised in genes. The mutations in the URE however cannot cause a 111 (but only a 011), so they only appear because the position is also part of a gene or promotor.

Phenotype 111 — Localisation

	Occurrences						
Location	total	%	mean	σ	σ%	T→M	M→T
Set A							
Gene	215824	68.02%	196.20±	156.15	79.58%	61.58%	83.01%
Promotor	56563	17.83%	51.42±	26.19	50.93%	100.00%	39.10%
URE	44912	14.15%	40.83±	36.21	88.70%	64.07%	29.03%
Sum	317299	100.00%					
Set B							
Gene	198436	68.27%	198.44±	126.56	63.78%	61.62%	83.37%
Promotor	51850	17.84%	51.85±	12.07	23.28%	100.00%	39.45%
URE	40375	13.89%	40.38±	29.30	72.57%	62.68%	28.75%
Sum	290661	100.00%					
Set C							
Gene	198260	68.27%	198.26±	119.43	60.24%	61.45%	83.41%
Promotor	51716	17.81%	51.72±	11.56	22.35%	100.00%	39.36%
URE	40435	13.92%	40.44±	27.68	68.45%	62.66%	28.78%
Sum	290411	100.00%					

Table 12: Correlations between phenotype and location of the mutation for set C. For explanation of column headers see table 10.

3.5.2 Gene change

As shown in table 13, all instances of gene loss, gene gain and class switch produce phenotype 111. Considering the model this behaviour is expected. Because the gene (strict) is much longer than the promotor, class switches are more common than either gene loss or gain.

Phenotype 111 — Gene change							
	Occurrences						
Mutation	total	%	mean	σ	σ%	T→M	M→T
Set A							
Gene loss	28038	18.46%	25.49±	16.47	64.61%	100.00%	20.01%
Gene gain	28525	18.78%	25.93±	12.47	48.08%	100.00%	20.35%
Class switch	95315	62.76%	86.65±	57.61	66.48%	100.00%	54.32%
Sum	151878	100.00%					
Set B							
Gene loss	25764	18.47%	25.76±	12.42	48.19%	100.00%	20.25%
Gene gain	26086	18.70%	26.09±	3.98	15.25%	100.00%	20.53%
Class switch	87643	62.83%	87.64±	45.82	52.28%	100.00%	54.65%
Sum	139493	100.00%					
Set C							
Gene loss	25812	18.44%	25.81±	11.64	45.10%	100.00%	20.32%
Gene gain	25904	18.51%	25.90±	4.03	15.57%	100.00%	20.33%
Class switch	88252	63.05%	88.25±	43.67	49.49%	100.00%	54.95%
Sum	139968	100.00%					

Table 13: Correlations between phenotype and location of the mutation for sets A, B and C. Mutation gives the exact type of the mutation. Other headers like in table 10.

3.5.3 Transcription Factors

Table 14 on the next page shows activity decrease and increase are equally likely as expected. Phenotype 001 could also have activity loss, gain and switch. The lack of these mutation types combined with the findings in table 11 seems to indicate that activity loss, gain and switch are very unlikely events, at least if the URE mutates.

Most frequent events are target gene loss and target gene gain as can be seen in table 14 on the page after the next. They occur whenever a gene is lost or gained that was under the influence of a transcription factor. As the number of genes ng increases, more genes that can be transcription factors and more potential targets exist. With the genome lengths around 100 the reaction network appears sufficiently complex to have a lot of genes subjected to transcriptional control, therefore target gene loss and gain happen often. Activity loss, gain and switch are more frequent than activity decrease and increase. Together with results from tables 14 and 15 this lends support to the hypothesis that activity loss, gain and switch can easily be caused by mutations in the transcription factor but hardly in mutations located in the URE.

Phenotype 001 — Transcription factors							
	Occurrences						
Mutation	total	%	mean	σ	σ_{ln}	T→M	M→T
Set A							
Act decrease	3611	48.10%	3.28	± 4.93	× 2.59	31.61%	54.88%
Act increase	3896	51.90%	3.54	± 5.06	× 2.63	32.71%	57.63%
Sum	7507	100.00%					
Set B							
Act decrease	3655	51.73%	3.65	± 5.24	× 2.67	35.42%	57.71%
Act increase	3410	48.27%	3.41	± 4.60	× 2.56	34.44%	54.04%
Sum	7065	100.00%					
Set C							
Act decrease	3303	49.54%	3.30	± 4.59	× 2.58	35.82%	55.24%
Act increase	3365	50.46%	3.37	± 4.21	× 2.49	36.05%	55.92%
Sum	6668	100.00%					

Table 14: Correlations between phenotype and mutations of transcription factor type for sets A, B and C. Mutation states the exact type, where 'Act' is short for activity. Total gives the number of occurrences of the specific combination; % the relative amount of localisation for the phenotype; mean shows the mean number of occurrences of the combination; σ the standard deviation of the mean; σ_{ln} the standard deviation for the lognormal distribution. T→M, M→T: e.g. first line: 31.61% and 54.88% means that 31.61% of activity decreases correlate with the phenotype 001, but full 54.88% of phenotype 001 have suffered activity decrease.

Phenotype 111 — Transcription factor							
		Occurrences					
Mutation	total	%	mean	σ	σ_{ln}	T→M	M→T
Set A							
Tgt gene loss	64673	31.16%	58.79±	82.74×	6.77	100.00%	17.04%
Tgt gene gain	66369	31.98%	60.34±	65.63×	5.73	100.00%	17.41%
Act switch	25791	12.43%	23.45±	39.89×	5.61	100.00%	8.65%
Act loss	17548	8.45%	15.95±	23.56×	4.71	100.00%	8.88%
Act decrease	7813	3.76%	7.10±	14.15×	3.67	68.39%	4.31%
Act gain	17355	8.36%	15.78±	21.92×	4.33	100.00%	8.85%
Act increase	8014	3.86%	7.29±	14.28×	3.64	67.29%	4.34%
Sum	207563	100.00%					
Set B							
Tgt gene loss	50396	30.71%	50.40±	62.74×	4.76	100.00%	17.33%
Tgt gene gain	50670	30.87%	50.67±	35.57×	3.70	100.00%	17.65%
Act switch	22081	13.45%	22.08±	35.10×	5.17	100.00%	8.70%
Act loss	14190	8.65%	14.19±	19.46×	4.20	100.00%	8.13%
Act decrease	6663	4.06%	6.66±	13.06×	3.49	64.58%	4.10%
Act gain	13630	8.30%	13.63±	16.33×	3.60	100.00%	7.96%
Act increase	6491	3.96%	6.49±	11.94×	3.42	65.56%	3.92%
Sum	164121	100.00%					
Set C							
Tgt gene loss	47720	30.60%	47.72±	50.86×	4.86	100.00%	17.14%
Tgt gene gain	49284	31.60%	49.28±	34.46×	3.91	100.00%	17.25%
Act switch	20345	13.05%	20.34±	28.95×	4.96	100.00%	8.41%
Act loss	13818	8.86%	13.82±	17.01×	4.11	100.00%	8.31%
Act decrease	5918	3.79%	5.92±	10.37×	3.37	64.18%	3.79%
Act gain	12900	8.27%	12.90±	14.74×	3.60	100.00%	7.73%
Act increase	5968	3.83%	5.97±	10.19×	3.30	63.95%	3.74%
Sum	155953	100.00%					

Table 15: Correlations between phenotype 111 and mutation of transcription factor type in sets A, B and C. 'Act' is short for activity, 'Tgt' for target. Column headers are explained at table 14.

4 Conclusion and Outlook

Research is what I'm doing when I don't know what I'm doing.

— Wernher von Braun

The aim has been to determine the neutrality of systems simulated by Mini-CellSim-Genome. On the whole after analysing 931100 genomes an average of 57.3% are neutral and therefore have the same phenotype after mutation. A closer look reveals that approximately half of the neutral mutations lie in regions of the genome that do not take any influence at all on the phenotype. The other half may be considered truly neutral mutations as the process of deriving the phenotype from the information on the genome allows for some changes without altering the resulting phenotype. This leaves 25% of all mutations to be ignored or compensated for by cellular mechanisms. So many sequences lead to the same phenotype which in turn makes the phenotype rather stable.
Such stability is bought at a price. This price is loss of information. In the model a majority rule is applied for translating mRNA into proteins and here information is lost. The protein encoded by the mRNA is unambiguous, but there are lots of different mRNAs that encode the same protein. Biological systems are similar at translation and have an additional loss with structure. For a rather large percentage of catalytically active proteins and RNAs the structure is important but not the sequence and all structures can be made up by more than one sequence.

The longer the genomes the less likely becomes the neutral phenotype 000, whereas the others gain in probability. This is not surprising as the number

of genes increases linearly with genome size. Also the fact that all three test sets yield similar data shows that sets with only a single genome length behave like sets with many different genome lengths but the same mean genome length.
Because of similar results derived from the different sets it is a well justified assumption that enough cases have been sampled in order to give representative results.

Further experiments with more than one mutation could lead to the finding of other phenotypes which are possible but did not show up for the analysed one point mutants. The used sets centered on a genome lenght of 100 nucleotides have roughly 50% neutral mutations, but as pointed out above longer genomes have less. If compared to figure 5 genome sizes of 1000 or 2000 nucleotides would give a very good approximation of the number of neutral mutations any genome with the promotor and gene length used in this work. It appears reasonable to assume that this number will be around 25% because that is the amount of neutral mutations in the analysed sets that occur within genes, and very long genomes would have very little intergenic sequences.

References

Appreciation is a wonderful thing: It makes what is excellent in others belong to us as well.

— Francois Marie Arouet Voltaire

[1] I. Alvarez-Garcia and E.A. Miska. MicroRNA Functions in Animal Development and Human Disease. *Development*, 21:4653–4662, 2005.

[2] A. Aravin and T. Tuschl. Identification and Characterization of Small RNAs Involved in RNA Silencing. *FEBS Lett.*, 579(26):5830–5840, 2005.

[3] D. Baltimore. RNA-dependant DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(252):1209–11.

[4] H Boeger, D A Bushnell, R Davis, J Griesenbeck, Y Lorch, J S Strattan, K D Westover, and R D Kornberg. Structural basis of eukaryotic gene transcription. *FEBS Lett.*, 579(4):899–903, 2005.

[5] M. Chase and A.D. Hershey. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, 1:39–56, 1952.

[6] DEOgenomes.org funded by the U.S. Department of Energy Biological and Environmental Research program. Human reference sequence. http://www.ornl.gov/sci/techresources/Human_Genome/-posters/chromosome/faqs.shtml, accessed on 12 Dec 2005.

[7] J E Dueber, B J Yeh, R P Bhattacharyya, and W A Lim. Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry. *Curr Opin Struct Biol.*, 14(6):690–699, 2004.

[8] M. Eigen. Self Organization of Matter and the Evolution of Biological Macro Molecules. *Naturwissenschaften*, 58(10):465–523, 1971.

[9] M. Eigen, J. McCaskill, and P. Schuster. The Molecular Quasispecies. *Adv. Chem. Phys.*, 75:149–263, 1989.

[10] M. Eigen and P. Schuster. The Hypercycle. *Naturwiss.*, 64:541–565, 1977.

[11] R.A. Fisher. *The Genetical Theory of Natural Selection. A Complete Variorum Edition.* Number ISBN 0-1985-0440-3. Oxford University Press., 2000/1930.

[12] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA Scondary Structures. *Biopolymers*, 33:1389–1404, 1993.

[13] W. Fontana, W. Schnabl, and P. Schuster. Physical Aspects of Evolutionary Optimization and Adaptation. *Phys. Rev. A*, 40(3301-3321), 1989.

[14] W. Fontana and P. Schuster. Continuity in Evolution: On the Nature of Transitions. *Science*, 280(5368):1451–5, 1998.

[15] W. Fontana and P. Schuster. Shaping Space: The Possible and the Attainable in RNA Genotype-Phenotype Mapping. *J. Theor. Biol.*, 194(4):491–515, 1998.

[16] C. V. Forst, C. Reidys, and J. Weber. *Advances in Artificial Life*, volume 929, chapter Evolutionary Dynamics and Optimization, pages 128–147. Springer, Berlin, Heidelberg, New York, 1995.

[17] D.J. Futuyma. *Evolutionary Biology*. Sinauer Associates, Massachusetts, 1979.

[18] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, and P. Schuster. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration. I. Neutral Networks. *Monatsh. Chem.*, 127:355–374, 1996.

[19] W. Gruener, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, and P. Schuster. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration. II.Structure of Neutral Networks and Shape Space Covering. *Monatsh. Chem.*, 127:375–389, 1996.

[20] A.J. Hager, J.D. Pollard, and J.W. Szostak. Ribozymes: aiming at RNA replication and protein synthesis. *Chem Biol.*, 3(9):717–725, 1996.

[21] A. Herbert. The Four Rs of RNA-directed Evolution. *Nat. Genet.*, 36(1):19–25, 2004.

[22] I.L. Hofacker and P.F. Stadler. *Vienna-RNL.* TBI.

[23] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[24] M.A. Huynen. Exploring Phenotype Space through Neutral Evolution. *Journal of Molecular Evolution*, 43:165–169, 1996.

[25] M.A. Huynen, P.F. Stadler, and W. Fontana. Smoothness within Ruggedness: The Role of Neutrality in Adaptation. *Proc. Natl. Acad. Sci. USA*, 93(1):397–401, 1996.

[26] L D Kapp and J R Lorsch. The molecular mechanics of eukaryotic translation. *Annu Rev Biochem.*, 73:657–704, 2004.

[27] M. Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(624-626), 1968.

[28] H Kohzaki and Y Murakami. Transcription factors and dna replication origin selection. *Bioessays*, 27(11):1107–1116, 2005.

[29] R D Kornberg. Mediator and the mechanism of transcriptional activation. *Trends Biochem Sci.*, 30(5):235–239, 2005.

[30] R. Machné. Sbml ode solver library (soslib). http://www.tbi.univie.ac.at/~raim/odeSolver/, accessed on 19 Dec 2005.

[31] D.H. Mathews, J. Sabina, and D.H. Zuker, M.and Turner. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol*, (288):911–940, 1999.

[32] J.S. Mattick. Challenging the Dogma: the Hidden Layer of Non-protein-coding RNAs in Complex Organisms. *Bioessays*, 10:930–939, 2003.

[33] E. Mayr. *Systematics and the Origin of Species*. Columbia University Press, New York, 1942).

[34] E. Mayr. *What Evolution is*. Number ISBN 0-465-04425-5. Basics Books, 2001.

[35] S Meyer, C Temme, and E Wahle. Messenger rna turnover in eukaryotes: pathways and enzymes. *Crit Rev Biochem Mol Biol.*, 39(4):197–216, 2004.

[36] L. Moran. What is Evolution? online.

[37] I I Moraru and L M Loew. Intracellular signaling: spatial and temporal control. *Physiology (Bethesda)*, 20:169–179, 2005.

[38] T. Mourier. Reverse Transcription in Genome Evolution. *Cytogenet. Genome. Res.*, 110(56-62), 2005.

[39] M Ptashne. Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci.*, 30(6):275–279, 2005.

[40] C. Reidys, C. V. Forst, and P. Schuster. Replication and Mutation on Neutral Networks. *Bull. Math. Biol.*, (63):57–94, 2001.

[41] E.A. Schultes and D.P. Bartel. One Sequence, two Ribozymes: Implications for the Emergence of new Ribozyme Folds. *Science*, 289(5478)):448–52, 2000.

[42] P. Schuster and W. Fontana. Chance and necessity in evolution: Lessons from rna. *Physica D*, 255:279–284, 1999.

[43] P. Schuster, W. Fontana, P.F. Stadler, and I.L. Hofacker. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc.Roy.Soc.Lond.B*, 255:279–284, 1994.

[44] A.S. Spirin. Omnipotent rna. *FEBS Lett.*, 530(1-3):4–8, 2002.

[45] M. Tacker, W. Fontana, P.F. Stadler, and P. Schuster. Statistics of RNA Melting Kinetics. *Eur. Biophys. J.*, 23(1):29–38, 1994.

[46] Manfred Tacker, Peter F. Stadler, Erich G. Bornberg-Bauer, Ivo L. Hofacker, and Peter Schuster. Algorithm Independent Properties of RNA Structure Prediction. *Eur. Biophy. J.*, 25:115–130, 1996.

[47] H.M. Temin and S. Mizutani. RNA-dependant DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(252):1211–1213.

[48] (TIGR) The Institute for Genomic Research. World's Longest Contiguous DNA Sequence. http://www.tigr.org/tdb/contig_list.shtml, accessed on 12 Dec 2005.

[49] J.D. Watson and F.H.C. Crick. Molecular Structure of Nucleic Acids. *nature*, 171:737–738, 1953.

[50] M. Zuker and D. Sankoff. RNA Secondary Structures and their Prediction. *Bull. Math. Biol.*, 46:591–621, 1984.

List of Figures

Reason shapes the future, but superstition infects the present.

— Iain M. Banks

List of Tables

The beginning of knowledge is the discovery of something we do not understand.

— Frank Herbert

A Used Symbols

Mathematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true.

— Bertrand Russell

symbol	meaning
$\mathcal{A}$	an alphabet e.g. $\mathcal{A} = \{A, C, G, T\}$ for DNA
$\|\mathcal{A}\|$	the size of a given alphabet = the number of letters in the alphabet
gl	gene length (wider)
gnl	gene length (strict)
L	genome length
L_{eff}	effective genome length
μ_{1p}	mean number of genes affected by a single point mutation
μ_{ng}	mean number of genes
μ_{ng}^{id}	mean number of genes of the ideal genome
ng	number of genes
P_{gain}^{GN}	probability of a mutation to lead to gene gain
P_{loss}^{GN}	probability of a mutation to cause gene loss
P_{gene}^{M}	probability of a mutation to be within a gene (strict)
P_{ig}^{M}	probability of a mutation to be intergenic
P_{nig}^{M}	probability of a mutation to be non-intergenic
P_{Prom}^{M}	probability of a mutation to be within a promotor
P_{URE}^{M}	probability of a mutation to be within a URE
P_{Prom}	probability of occurrence of a promotor at any one position
pl	promotor length
$\mathcal{S}$	the shadow of a sequence
$\|\mathcal{S}\|$	the size of a shadow = the number of sequences in the shadow

abbreviation	meaning
A	adenosine (a DNA and RNA nucleotide)
C	cytosine (a DNA and RNA nucleotide)
CI	inhibitory activity / transcription factor
CX	activating activity / transcription factor
D	dynamics (TFD model)
DNA	desoxyribonucleic acid
F	functional network (TFD model)
G	guanine (a DNA and RNA nucleotide)
GN	gene
mRNA	messenger RNA
RNA	ribonucleic acid
rRNA	ribosomal RNA
SP	structural protein
T	network (TFD model)
T	thymidine (a DNA nucleotide)
TF	transcription factor
tRNA	transfer RNA
U	uracil (an RNA nucleotide)
URE	upstream regulatory element

B MiniCellSim-Genome

In the beginning the Universe was created. This has made a lot of people very angry and has been widely regarded as a bad move.

— **Douglas Adams**

MiniCellSim-Genome is a set of Perl libraries and sample Perl scripts that allow to simulate cellular processes for a single cell. This project is under heavy work and will soon encompass more features.
The version used for this work is the checkout of the 16^{th} of December 2005. Cellular processes are depicted by figure 7.

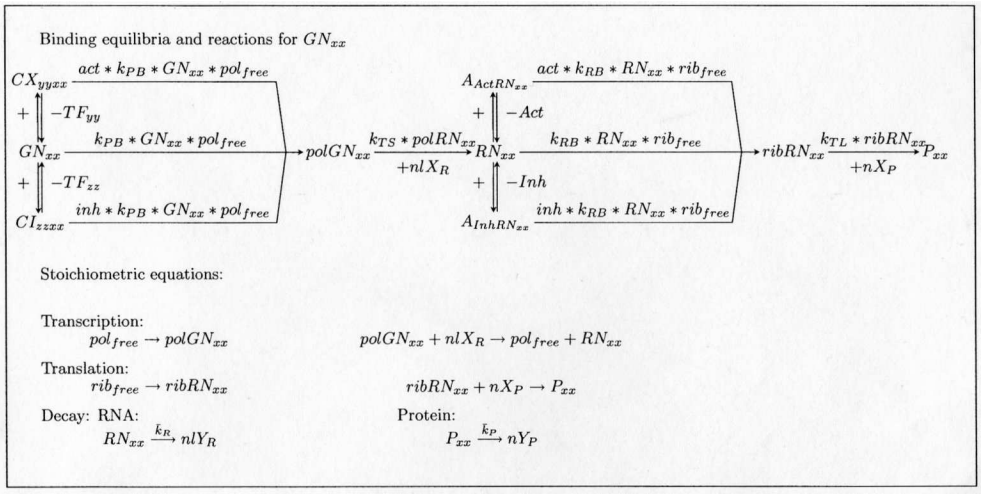

Figure 7: Cellular processes in MiniCellSim-Genome.

C Reactions

It has yet to be proven that intelligence has any survival value.

— Arthur C. Clarke

The Vienna-RNL is a library for chemical reaction networks. The code is written in ANSI C and can be used for C programs or perl scripts. SBML and its own rct files are valid input files and numerous output formats are supported. More information and a download is available at http://www.tbi.univie.ac.at/software/Vienna-RNL/

As pointed out reactions was used to generate a simpler rct file from the SBML output of MiniCellSim-Genome. The rct files then served for direct line to line comparison via the unix program diff to determine T of the phenotype. Further parent and mutant rct files were compared by netcomp which uses the function NetDiff of the Vienna-RNL to compare two reaction networks. This function "computes the difference of the two Networks N1 and N2. The reaction difference Network contains all reactions occurring in N1 but not N2, and all substrates occurring in the remaining reactions." [22] As such an operation is asymmetrical parent and mutant have to compared in both directions to yield the phenotypical feature F.

D SOSlib – odeSolver

You climb to reach the summit, but once there, discover that all roads lead down.

— Stanislaw Lem

"The SBML ODE Solver Library (SOSlib) is both a programming library and a command-line application for construction, symbolic and numerical analysis of a system of ordinary differential equations (ODEs) derived from a chemical reaction network encoded in the Systems Biology Markup Language (SBML). It is written in ISO C and distributed under the terms of the GNU Lesser General Public License (LGPL). The package employs libSBML's AST (Abstract Syntax Tree) for formula representation to construct ODE systems, their Jacobian matrix and other derivatives. SUNDIALS' version of CVODE is incorporated for numerical integration and sensitivity analysis of stiff and non-stiff ODE systems." [30]
Download and more information at http://www.tbi.univie.ac.at/~raim/odeSolver/

The last phenotypical feature D is computed using SOSlib. All kinetic laws computed by MiniCellSim-Genome are integrated to yield quantitative results for the amount of the gene products in the cell. To save calculation time only 10 points in time were used equally spaced until the integration end time, which was 10000.

E About the Author

I have great faith in fools — self confidence my friends call it.

— Edgar Allan Poe

Personal

Thomas Taylor
anubis@tbi.univie.ac.at
1982-07-30 at Wien
Austrian nationality

Education

since 2002	Medizin / Medicine at the Medizinische Universität Wien
since 2000	Molekulare Biologie / Molecular Biology and Ägyptologie / Egyptology at the Universität Wien
June 2000	Matura
1992—2000	humanistic branch, BG/BRG XXII Bernoullistraße

Languages

German	mother tongue
Englisch	CCAE grade A
French	intermediate
Arabic	beginner
Latin	advanced intermediate
Ancient Greek	beginner

EDV

OS	UNIX (Linux), Windows
Programming	C, C++, QT, VisualBasic
Script	Perl, Bash, Tcl/Tk, OpenGL
Databases	PostgreSQL
Applications	LaTeX, Emacs, Microsoft Office Suite, OpenOffice Suite

Interests

RPG, SF/Fantasy, Ancient Egypt, Jiu Jitsu, Dancing, Piano