Computational Models of the Genetic Code Evolution Based on Empirical Potentials

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

Vorgelegt von

Mag. Günther Weberndorfer

im April 2002

Für Sylvie und Johannes.

Acknowledgment

An erster Stelle möchte ich mich bei all jenen bedanken, die zu dieser Arbeit beigetragen haben. allen voran meinen Betreuern Peter Stadler und Ivo Hofacker. Sie haben diese Arbeit ins Leben gerufen und mich durch den wissenschaftlichen Dschungel dieses Gebietes begleitet. Weiters möchte ich mich bei Peter Schuster, meinem Doktor-Grossvater bedanken dafür, daß er mich an seinem Institut freundlich aufgenommen hat. Christoph Flamm hat mir sehr bei der Programmierung und der Auswahl wissenschaftlicher Papers geholfen. Er war ein sehr angenehmer und inspirierender Partner für Diskussionen während der letzten Jahre. Auch gebührt mein Dank all jenen die das Arbeiten am TBI ermöglicht und angenehm gemacht haben, allen voran Judith Jakubetz sowie vielen netten Kollegen, die im Laufe der Jahre kamen und gingen. Besonders meinen Zimmerkollegen, Roman, Michael und Bärbel danke ich dafür, meine Launen all die Zeit ertragen zu haben.

Den allergrößten Dank sowie den Rest dieser Arbeit widme ich jedoch meiner Familie. Meine geliebte Frau Sylvie Marie hat mir in den schweren Zeiten der Frustration und finanziellen Entbehrung immer beigestanden. Gemeinsam mit unserem Sohn Johannes Elias hat sie es verstanden, mich immer wieder aufzumuntern und zu motivieren. Ohne sie würde es diese Arbeit nicht geben. Weiters danke ich meinen Eltern, meiner Mutter für die Unterstützung und meinem Vater, der diesen Tag leider nicht mehr erleben durfte.

Abschließend möchte ich mich bei meinen Freudnen und Geschäftspartnern Andreas Wernitznig, Alexander Renner and Stephan Kopp bedanke, die mich auch tatkräftig unterstützt haben. Mit ihnen beginne ich nun meine Zukunft, "*Insilico*" aufzubauen.

Danke Euch allen!

Abstract

The building plan of a cell lies within its genes. The story of this building plan consists of nucleotide words, that are interpreted using a pivotal table, *the genetic code*. This table is an interface between the linear information stored in nucleic acids and the interpretation of this data in the sequence and folding of proteins. The genetic code is the key switch within the biochemical dogma and shared among almost all living beings, certain variations give hint that evolution takes place at this level as well. Although the genetic code is well known since over thirty years, its origin remains an enigma.

The primordial code is hypothesized to have been a simplified ancestor of the canonical code used in contemporary cells. But how could the language of the building plan change without destroying the information? There is no compelling theory of the mechanism of code development in terms of the early evolution. The motivation of this work was to design and implement a realistic model for the stage of evolution, where a primitive genetic code existed, and proteins took some of the duty from nucleic acid enzymes. Such a model can be used to test the feasibility and mechanism of genetic code evolution.

Our goal was to develop an object oriented computer application framework that provides easy access to artificial model organisms evolving in a tank reactor. The organisms are simplified due to limitations in the knowledge of all necessary

components of a living cell as well as limited computer resources. An artificial organism is built of a set of tRNAs and a native RNA dependent RNA polymerase gene. The tRNAs are loaded using Boolean operators on suppositional *determinant positions*, which are generated via folding the tRNA sequence to its mfe secondary structure and applying a constant mask onto its sequence. These tRNAs are employed in translating the replicase gene and determine the genetic code. The resulting amino acid sequence is threaded onto the native structure of T7 phage's RNA polymerase $3d$ structure using an empirical four point contact potential. The z-score of the sequence (fitness) is used to perform a tournament replication within the tank reactor. The variation of the energy of residues within the core of the protein determine the accuracy (mutation rate) of the replicase.

We were able to observe that if a system like this is evolved using a restricted alphabet set (such as HP amino acids) the organisms tend to optimize the mutation rate and expand the alphabet of known amino acids. Since only very limited number of mutations yield a protein sequence with increased performance, not all findings are fixated within the population. But in a larger timescales, the whole population drifts from two residues to three and more letter alphabets manifested in fixed codons. These simulations, based on pure evolutive and biophysical laws show that it is not necessary to stress any metabolic pathways to extend a coding biological system, and that ambiguous coding is a suitable mechanism to explain codon changes.

Zusammenfassung

Der Bauplan der Zelle ist in den Genen niedergeschrieben. Der Text dieses Plans ist aus Nucleotid Wörtern zusammengesetzt, die durch eine zentrale Tabelle, den *genetischen Code*, übersetzt werden. Diese Tabelle ist die Schnittstelle zwischen der linearen Information der Nucleotid Basen und der Interpretation dieser Daten in der Sequenz und räumlichen Faltung der Proteine. Der sogenannte universelle genetische Code ist die Schlüsselstelle im biologischen Dogma und wird von beinahe allen Organismen eingesetzt. Variationen des genetischen Codes geben einen Hinweis darauf, daß auch dieser den Prinzipien der Evolution unterliegt. Obwohl der Code seit über 30 Jahren entschlüsselt ist, bleibt sein Ursprung ein Mysterium.

Es wird angenommen, daß der primordiale Code eine vereinfachte Form des heute gültigen war. Wie konnte sich jedoch die Sprache des zellulären Bauplans ändern, ohne die gesamte Information unlesbar zu machen? Es gibt keine überzeugende Theorie für den Mechanismus der Entwicklung des genetischen Codes in dieser frühen Phase der Evolution. Die Motivation dieser Arbeit war es, ein möglichst realistisches Modell der Entwicklungsstufe zu entwerfen, wo ein einfacher Ur-Code besteht und Proteine begannen Nucleinsäuren zu ersetzen. Dieses Modell bietet die Möglichkeit, Machbarkeit und Mechanismus der Änderung des genetischen Codes zu untersuchen.

Es ist uns gelungen, ein objektorientiertes Computer Framework zu entwickeln, das es ermöglicht, künstliche Modellorganismen in einem Flussreaktor zu modellieren. Der Organismus ist in vielerlei Hinsicht vereinfacht, da nicht alle molekularen Details bekannt sind, und moderne Computer nicht genügend Rechenleistung für die vollständige Simulation von Leben bieten. Das Genom des Modellorganismus besteht aus RNA und kodiert für wenige tRNAs und eine native RNA Replicase. Durch Anwendung Boolscher Operatoren auf hypothetische *Determinanten Positionen* werden tRNAs mit Aminosäuren beladen. Die Determinanten Positionen werden durch Falten der tRNA Sequenz in die MFE Sekundärstruktur und Überlagerung der Sequenz mit einer konstanten Maske errechnet. Die so durch den genetischen Code beladenen tRNAs werden zur Translation des Replicase Gens verwendet. Das naszive Polypeptid wird mittels eines empirischen vier-Punkt Kontaktpotentials in die native $3D$-Struktur der T7 Phagen RNA Polymerase gefaltet. Der z-score der Sequenz bestimmt die Fitness in Konkurrenzkampf um die Replikation innerhalb des Flussreaktors. Die Variation der Energie gewisser Sequenzpositionen im aktiven Zentrum der Replikase bestimmen deren Genauigkeit (Mutationsrate).

Wenn ein System dieser Art evolviert und zu Beginn nur ein eingeschränktes Aminosäurealphabet (z.B. eine hydrophobe und eine polare Aminosäure) zur Verfügung war, tendieren die Organismen dazu, erst die Mutationsrate zu optimieren und dann mehr Aminosäuren zu verwenden, indem sie den genetischen Code erweitern. Da nur wenige Mutationen zu erhöhter Fitness der Replicase führen, werden nicht alle Alphabet Erweiterungen in der Population fixiert. In längeren Zeitabschnitten ist es jedoch möglich zu beobachten, daß die gesamte Population des Flussreaktors zur Codierung erweiterter Alphabeten driftet. Diese Simulationen beruhen ausschließlich auf den Gesetzen der Evolution und biophysikalischen Erkenntnissen, und zeigen, daß es nicht nötig ist metabolische Pfade zur Erklärung von Erweiterungen des genetischen Codes heranzuziehen, und daß mehrdeutige Codierung ein geeigneter Mechanismus ist, um Änderungen im genetischen Code zu erklären.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

> "Science is made by observation and quantification, and only a mathematical description of a problem enables a deep and undoubtive understanding" *Peter G. Wolynes* 1998 [153].

It is widely accepted that life as we know it has *evolved* during the past 4 billion years and is the result of a developmental process. Men in all epochs wanted to illuminate the history and roots of life, and found that history left traces. First of all life is not homogeneous. Life found millions of ways to express itself in species. In the 18th century scientific evidence was found, that all species on earth are brothers and sisters. Using homologies it was possible to draw trees, that showed relationships of the families of species and the three phyla could be separated. A detailed description of the biological evolution, however is only possible on the basis of biological process, and these processes happen at a molecular level. The discovery of biopolymers (RNA, DNA and proteins), that are responsible for the biological processes revolved the understanding and description of life completely.

A molecular description of life was extremely fruitful during the past decades. Entire genomes of organisms were sequenced, opening the door to a world that

was *terra incognita* only a few years ago. This gave rise to new scientific disciplines such as bioinformatics, a biosciences branch that focuses the analysis of biological data. Beside the academic interest, new industries have been established and biotechnology [79] is now considered to be the cradle for future technologies. Biopolymers are used in technological applications already: gene therapies, biosensores and micro array analysis are impressive applications of the building blocks of life, and more is on the way. Even the principles of evolution itself are applied in the evolutionary design of macromolecules, that in parts replaced rational drug design.

Darwinian evolution is based on the principle of *survival of the fittest* [27]. In the language of computer science this is an optimization problem: One has to find an individual equipped with properties that are optimally suited to solve the entities problems. This optimization becomes very hard in populations of limited size, but nature's strategy of optimizing life as we know it is extremely efficient and simple: Increase variation on the basis of genotypes and select the phenotypes to decrease diversity. Variations of the genotypes arise by mutation in all organisms and by recombination in sexually reproducing organisms. The execution of the genotype results in a phenotype, whose properties determines the individuals fitness and prospects for reproduction.

The genotype of contemporary cells is laid down in a linear sequence of four nucleotides (A,C,T,G) in the biochemical more or less inert DNA. The genetic information is transcribed into an intermediate RNA messenger, that is used as instructional input for protein translation. Translation requires an unambiguous mapping of the four nucleotides to the 20 standard amino acids, that are building blocks for genetic executives, the proteins (figure 1.1 shows a schema of the translation). In the ribosome *triplets* of the four different RNA bases are read sequentially from the mRNA, what obviously results in $4^3 = 64$ possible codons. The degeneracy of triplets is used to encode START and STOP signals and to make the genetic code redundant.

The relatively simple nature of the genetic code (see table 1) was discovered in the 1960ies by Marshal Nirenberg [103, 104] and honored with the Nobel prize in 1968. The origin of this pivotal table remains an enigma, Nevertheless it is

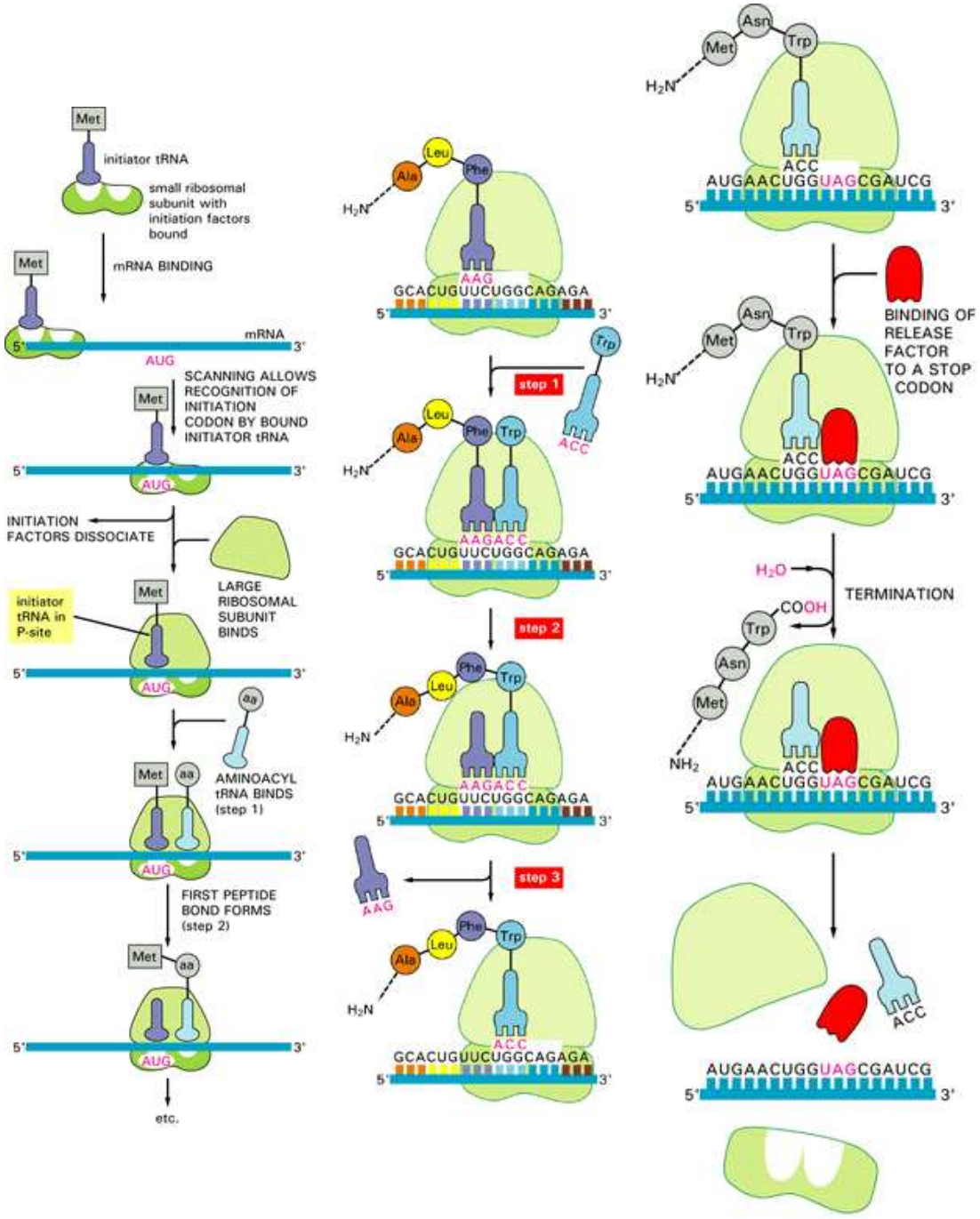


Figure 1.1: Molecular details of the translation (figure copyrighted by Alberts *et. al.* [1]). The three phases, initiation, elongation and termination are presented in the columns of the figure.

expected to contain clues about the origin of life itself.

1. Position	2.Position				3.Position
	U	**C**	**A**	**G**	
U	Phe	Ser	Tyr	Cys	**C**
	Phe	Ser	Tyr	Cys	**A**
	Leu	Ser	STOP	STOP	**G**
	Leu	Ser	STOP	STOP	**U**
C	Leu	Pro	His	Arg	**C**
	Leu	Pro	His	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
	Leu	Pro	Gln	Arg	**U**
A	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Asn	Ser	**A**
	Ile	Thr	Lys	Arg	**G**
	Met	Thr	Lys	Arg	**U**
G	Val	Ala	Asp	Gly	**C**
	Val	Ala	Asp	Gly	**A**
	Val	Ala	Glu	Gly	**G**
	Val	Ala	Glu	Gly	**U**

Table 1.1: The universal genetic code

Surprisingly, the genetic code is almost the same in all organisms [102]. This can be interpreted as result of continuous heredity in evolution. On the other hand there are small differences between the genetic codes of different phyla which imply that the genetic code is also subject to evolution [106]. These code variants can be represented as tree as shown in figure 1.2.

Blindly changing the codon table of an organism would of course correspond to re-wiring a keyboard and thus would be absolutely lethal. The highly complex information that had evolved would be rendered unreadable.

This view is at odds with the observations that codons can change their specificity. UGA is read as STOP signal in most organisms, but mapped to selenocysteine under some circumstances [98]. Since there exist three STOP signals in the *universal* code, this redundancy can be exploited, to enlarge the coded alphabet from 20 to 21 amino acids. Another common mechanism to enlarge the code is

to use codons that have a low frequency of usage. An experimental verification using cystein as miscoding replacement was successful [34]. Cystein is due to its size and structure suitable to replace most other amino acids. It was possible to show that a reassignment *in vivo* in *e. coli* is tolerated. It should in principle be possible to follow such a re-assignment in an artificial life model as well, because known physicochemical and evolutionary constraints influence this experiments.

The importance of a proper understanding of the universal genetic code and its variations also arises from contemporary biology. In the so-called "post-genomic" era where genomic information of entire organisms on single-nucleotide resolution is available and high-throughput methods steadily deliver more sequences, computational analysis became an indispensable tool. It is a common technique to perform *in silico* translation of open reading frames for annotation. Molecular modeling and inverse protein folding make rational drug design attractive. The design of biologically active recombinant proteins requires a decent understanding of the host and expression system. The substitution of single amino acids by non-standard codes may have major impact on the folded protein structure. Hence the usage of the correct codon table is crucial in selecting organisms for gene expression and analysis of DNA sequences.

Insight into evolutionary process is usually acquired by back-extrapolation from currently living, highly developed life forms to simpler precursors. This approach becomes infeasible if intermediate species are missing or if the focus goes beyond the first common ancestor of all living beings. The higher developed the observed metabolism is, the more speculative the theories become. Translation with all its complexity in modern cells so far successfully resists the bombardment by scientific theories. Because of the lack of evidence most of the relevant questions concerning the origin of the genetic code fall into the twilight zone of speculation. One has to accept that the same biochemical laws and conditions were valid under prebiotic conditions to accept evidence from existing metabolisms. It is common consensus that the chemical and physical properties of nucleic acids and amino acids were the same as today and therefore a molecular level can be used to model ancestral scenarios.

In the last two decades computer experiments simulating molecular evolution

were carried out with considerable success. Especially RNA molecules are well investigated because of their well known and simple genotype-phenotype mapping. The interaction of redundancy (there are by far more sequences that structures) and dispersion (sequences folding into the same structure are spread all over sequence space) led to neutral networks and "*shape space covering*". Simulations of RNA molecules were able to explain rare jumps and diffusion in evolutionary dynamics [52].

The computer model becomes increasingly complex if one progresses from simulating single macro molecules to populations of molecules and finally to entire biological processes and phenomena such as translation and the genetic code development. A major obstacle to perform *in silico* studies of genetic code evolution is that a system must be designed that couples the translation apparatus and replication because selection only can act upon many generations. Translation itself is a very complex procedure to which dozens of very elaborated proteins and RNAs contribute. Not every biochemical detail of protein synthesis is understood well at present. The ribosome, for instance, the particle that catalyzes the mRNA directed protein polymerization is one of the most complex biomolecular structures known. It has long resisted to crystallization and the detailed atomic structure of the complete *Thermus thermophilus* 70S ribosome has only partially been solved [157].

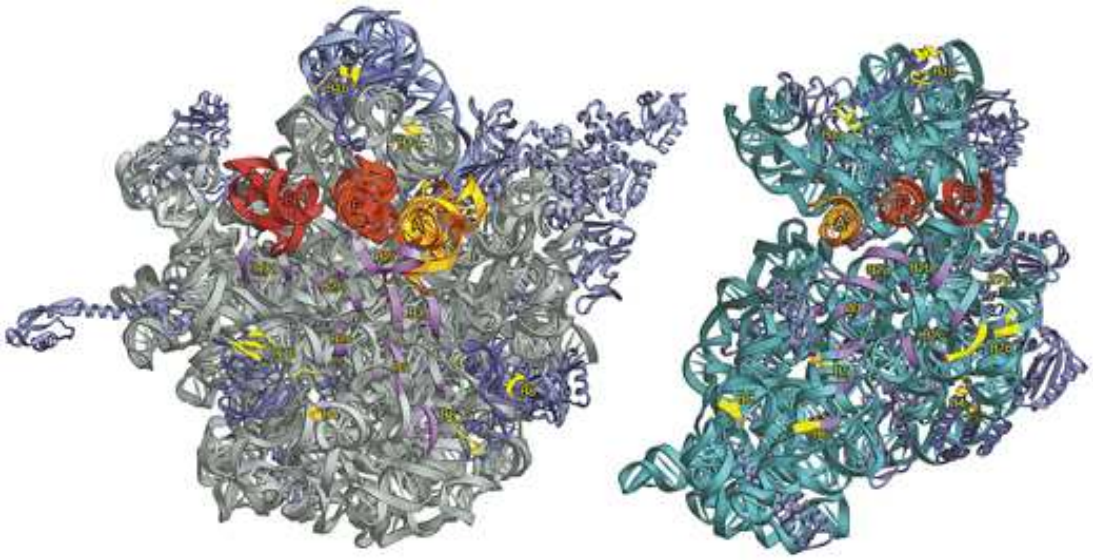


Figure 1.3: The figure shows the ribosomal subunits (50S (left) and 30S (right)) of the 70S *Thermus thermophilus* ribosome . Magenta, RNA-RNA contacts; yellow, protein-protein and protein-RNA contacts; A, P, and E mark tRNAs at left and tRNA anticodon stem loops at right.

The situation is not better for much smaller structures: Transfer RNAs (tRNAs) play a role as genetic adapter. They are only about 76 nucleotides long. Their duty is "reading" the sequence and executing the genetic code. To perform this task the correct loading with amino acids is crucial for replication accuracy. It is well known that a combination of structural and sequential information is identified by the aminoacyl tRNA synthetase, but by far not all factors have been identified. The specific synthesis of the aminoacyl tRNAs is crucial for the maintenance of the genetic code, since no more prove reading is done on the loading.

The replication of a genome is part of the cell cycle that employs hundreds of genes in complex organisms. It requires the coordination of events at the replication fork such that progress on the leading strand matches that of the lagging strand. Beside this complex interaction of biomolecules the $3D$-structure of the proteins, that are responsible for the biological function, is unknown for most participants. The so called "folding-problem", that is to find a set of rules that determine the spatial structure for a linear amino acid sequence, is still unsolved for proteins

in general. And the fold alone is not enough! Function requires movement and simulating atomic detailed molecular dynamics of a protein of several hundreds of amino acids is by orders of magnitude off the limit of computable problems. Simulations of entire cells are at the moment restricted to models of interactions like metabolic networks. At the state of computer hardware development it seems not possible however to simulate larger interacting biological compounds on the basis of structures.

Nevertheless, it is possible to simplify all these process extremely and build a consistent computer model, that is able to explain the extension of the genetic code solely on the basis of biophysical and evolutionary laws. It is the aim of this work to describe the design and implementation of such a model and investigate how this stands respect to existing hypothesis of the origin of the genetic code.

The next section will give a brief description of the state of knowledge about the origin of life. This comprehensive information of the chemical framework of the molecular origins of life will lead to the emergence of the genetic code and its time of occurrence. The high symmetry and importance of the genetic code have provoked numerous hypothesis and theories. In chapter 3 prevalent theories about the origin and evolution of the genetic code are presented and compared. In the methods chapter 4 a model is designed and presented that enables consistent simulations of the genetic code evolution solely based on biophysical and evolutionary constrains. Results from some representative computer experiments are presented in chapter 5 and discussed in the following outlook section.

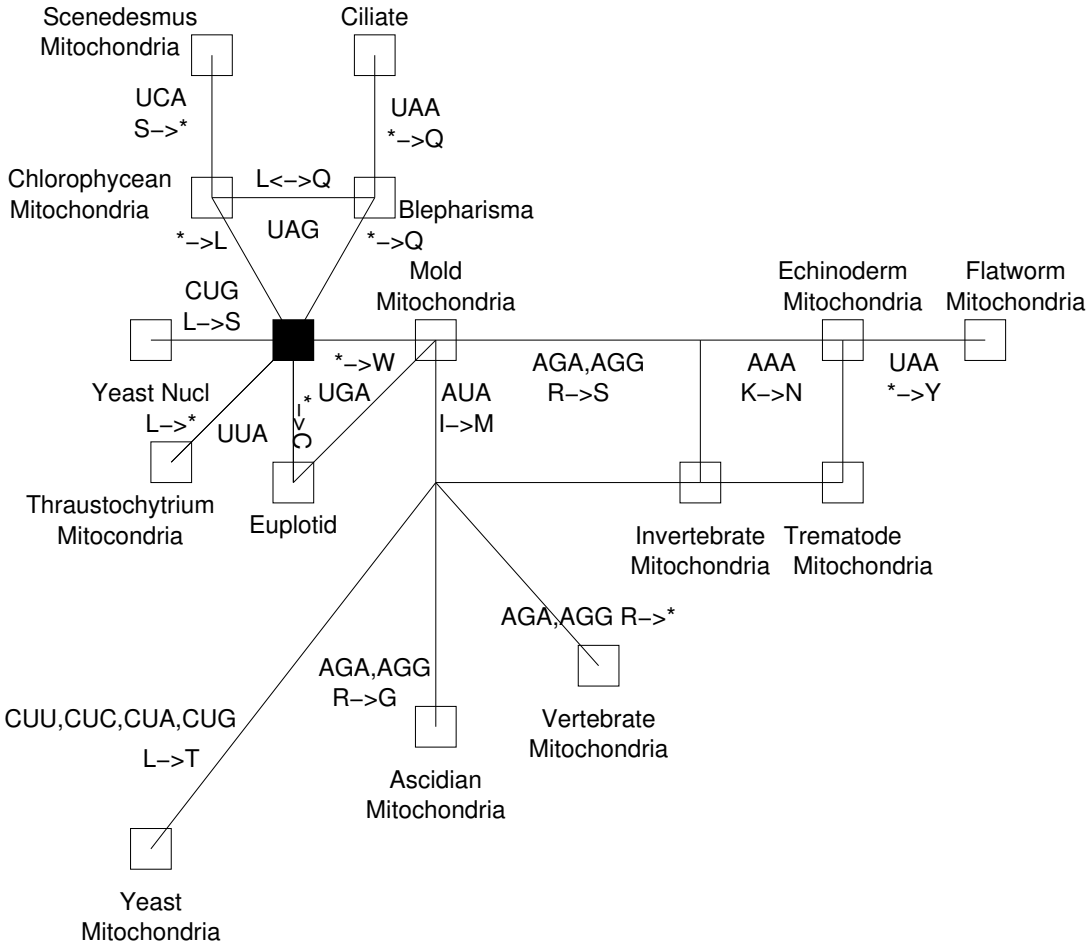


Figure 1.2: The genetic codes shows variations among different species, that can be represented as tree. Each node in the tree represents a species, traveling along the edges increases the number of differing codons. The black square marks the so called *universal* code, that is the most common among all phyla. Variants that are in the same subtree share the same codon differences. Edges denote the least common changes. The codes were taken from the National Center for Biotechnology Information (NCBI) and can be found on the Internet under the URI `http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c`.

CHAPTER 2

The Origin of Life

Apart from the fact that life arose at all on this planet, the speed is most surprising. Earth was created from a cloud of cosmic dust about 4.5×10^9 years ago. But this young planet was not a very hospitable place, there is good evidence that Earth was almost completely molten at that time. After a short period of crust formation and withstand of heavy meteoric impacts organic chemistry could have started about 400 million years later. The oldest microfossils discovered on Earth are bacterial and cyano bacterial structures and were found in Apex cherts of the Warrawoona Group in Western Australia [99, 122]. They were dated to be at least 3,465-million-years old, but the exact composition of organisms is still under dispute [17]. These microfossils show considerable structural complexity pointing toward an earlier, not yet identified root. The oldest, living organisms known cyanobacteria. Thus providing clues of how early life looked like. Therefore, only a few hundred million years must have been enough time to bring life to Earth. The uniformity of the biochemistry in all living cells indicates that the tree of life is rooted in a so called *last universal common ancestor (LUCA).*

The phylogenetic analysis of rRNA has so far been a successive method to separate different phyla and construct the tree of species. Nevertheless rRNA phylogenies are unsuitable to find the root because the concept of linages is doubtful

at this level, since a consensus history of genes is not understood [150]. A primitive translation apparatus and the lack of error correcting make a genetically drifting population reasonable. Therefore, high mutation rates and lateral gene transfer as common feature dominate the evolutionary dynamic of the progenote population, and it becomes likely that the common ancestor was more an entire population than a single cell. The development of highly specific, optimized genes made lateral transfer impossible and the phylogenetic tree started to grow.

2.1 Prebiotic Evolution

Almost all data and evidence for the first steps of life were wiped out through the last billions of years. Nevertheless, the molecular history and the laws that governed them are still the same, and using contemporary molecular biological techniques, it is possible to trace back life. But for a molecular biologist as well as a chemist the most fundamental question before explaining a reaction is to define the reacting compounds and condition during a reaction. Therefore, before it is possible to draw a consistent picture of the origin of life the composition of the prebiotic atmosphere must be clarified.

This was the starting point for the historic Miller-Urey experiments performed in the 1950ies [96]. To model the prebiotic atmosphere a mixture of methane, ammonia and hydrogen was exposed to thunder and lightning: electric discharges supplied energy and the products were diluted in liquid water. The solution contained numerous small organic molecules with several of the standard amino acids among them. Closer investigations revealed that glycine, for example, is formed from formaldehyde, cyanide and ammonia in a Strecker reaction. Despite the inspiring results serious doubts about the reductive character of the early atmosphere came up [78] and led to the *impact theory*. This contemporary view of the ancient atmosphere states that organic carbon infected the early Earth by meteorites. Experimental evidence comes from the investigations of carbonaceous chondrites, such as the much cited Murchison meteorite that impacted on Earth and contain non-racemic mixtures of amino acids [110].

Regardless of the detailed origin of the components, these theories have in com-

mon that they postulate simple organic molecules in aqueous solution build a kind of broth termed *primordial soup* [105]. This soup was the starting point for the polymerization that led to the development of the first genes.

A very different, hydro-thermal view of this evolutionary stage is getting more evidence nowadays [141]. The theory of a pressurized iron-sulfur world suggests a fast origin by an autotrophic metabolism of low-molecular weight constituents, in an environment of iron sulfide and hot magmatic exhalations of deep sea vents. The reaction of FeS and hydrogensulphide yields pyrites that offer strongly reductive surfaces. These sulphur catalysts in combination with heat and high pressure are able to reduce CO_2, thereby enriching environment with a wide palette of small organic molecules [24] such as pyruvate which is an essential intermediate metabolite. Wächtershäuser's theories are based on the synthesis of genetic monomers via a complex cycle of non-enzymatic chemical reactions but the reaction schemes seem to be rather complicated for an ancient system.

The various theories leading to the first organic molecules already disagree, but the next step is even more disputed. If one assumes that nucleotides, similar or equal to those existing in present day cells, were the basis for an ancestor that was a life-like aggregate of self-replicating molecules, activated building blocks must have been available for polymerization. Nucleotides are however from a chemical point of view extremely complicated molecules. Some major problems concern the available building pathway such as the auto-catalytic properties of the formose reaction [111] which irreversibly produces complex mixtures of sugars, of which ribose is only a minor component. Nitrogenous substances that are also needed for prebiotic nucleotide base synthesis would interfere with the formose reaction by reacting with formaldehyde and sugar products in undesirable ways. Nevertheless pathways for model prebiotic nucleotide synthesis have been shown to be achievable [126].

One key step at this level of development was the ability for chiral separation because all the known reactions in the prebiotic environment produced racemic mixtures of D- and L-enantiomeres. On the other hand a typical property of life as we know it is its specificity for distinct optical isomers. But fortunately enantiomeric fossils help to explain the preference of nature for distinct enan-

tiomeres: crystal facets. The absorption of small molecules such as amino acids enriched the prebiotic soup in the concentration of just one rotamere. This was shown to happen using the very common rock-forming mineral calcit ($CaCO_3$) when exposed to a racemic mixture of D- and L enantiomeres [70].

In the face of all the difficulties that a prebiotic nucleotide synthesis had, it was proposed that simpler template molecules preceded RNA. Such a system must have been simple enough to be accessible under prebiotic conditions, but still able to evolve and in turn "learn" to synthesize nucleotides [60]. A first candidate were self-replicating inorganic clays [19], but experimental evidence is missing. Also the question of how information was transfered from a mineral to RNA remains unclear. However, it is possible to imagine a kind of intermediate nucleic acid-like polymer that could serve as template. Various polynucleotide analogues using different sugars have been proposed. For instance, Eschenmoser and his colleagues [47, 48] systematically investigated the base-pairing properties of nucleic acid analogues and in particular the pyranosyl analogue (p-RNA) is appealing. Complementary p-RNA strands interact in a way that is stronger and more selective than either RNA or DNA. Eschenmoser presented an elegant theory for an prebiotical synthesis pathway, but experimental verification is missing. Another proposed backbone variant is "Peptide Nucleic Acid" (PNA), that binds bases via Nitrogen. This results in truly non-racemic polymers, that can base pair and perform template directed oligomerization [14]. In figure 2.1 the structure of p-RNA and PNA are drawn.

From a contemporary point of view, the formation of cell-like compartments is straight forward and the logical consequence of parasitism. It is well known that lipids and other amphiphiles have the capacity to undergo spontaneous self-organization into supra-molecular structures such as micelles and bilayers. This behavior is responsible for the use of this class of molecules to form stable cell compartments. Such compartments would be advantageous to establish specialized reaction conditions and protection to nucleases.

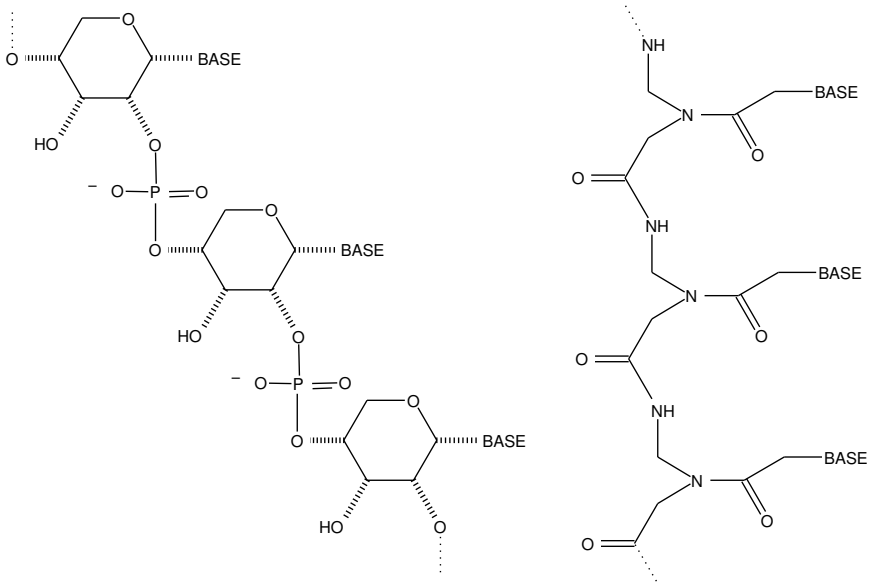


Figure 2.1: The structure of pyranosyl-RNA based on ribose-2,4-diphosphate.(from [47]) is drawn on the left side, PNA is shown on the right side.

2.2 The RNA World

The exact avenue from a primordial soup or submarine vents to simple cellular organism is still unclear, and the gap between the time where simple organic molecules evolved and condensed to form genes, that were able to evolve, still is unbridged. A widely accepted scenario after the establishment of genes is the so called *RNA-world*. This term was introduced by Gilbert in 1986 [62] after the discovery of the self-splicing *Tetrahymena* intron. This hypothesis [60] places RNA into the functional and informational center of primordial life.

The choice of RNA as basis of life has two good reasons: First RNA molecules are excellent templates for self-replication. Therefore they are source of information and target at the same time. Although enzyme-free template-induced synthesis of longer RNA molecules from monomers has not been achieved so far, more basic reaction could be demonstrated. Günther von Kiedrowski [139, 140] successfully demonstrated auto-catalytic template-induced synthesis of oligonucleotides from smaller oligonucleotide precursors.

Another property that makes RNA a good choice for a molecular basis of life

is that RNA molecules showed out to fold into complex $3D$-structures including pockets and binding sites which give them the capability of enzymatic activity. The finding of RNA enzyme activity in 1981 by Thomas Cech [22] was the breakthrough for the RNA-world hypothesis. In analogy to protein catalysts RNA enzymes were named *Ribozymes*. The finding of new reaction for RNA enzymes was accelerated by the development of the SELEX technique. This approach uses a transition state analogue as epitopal target for a large random library of RNA molecules. By this means many different reactions were added to the record of abilities of RNA: RNA-catalyzed RNA polymerization [44], aminoacyl esterase activity [109] or even peptide bond formation [158]. This is just a small excerpt of the variety of reaction that has been exposed over the last two decades, although some important ones, such as the phosphorylation of free ribose are still missing.

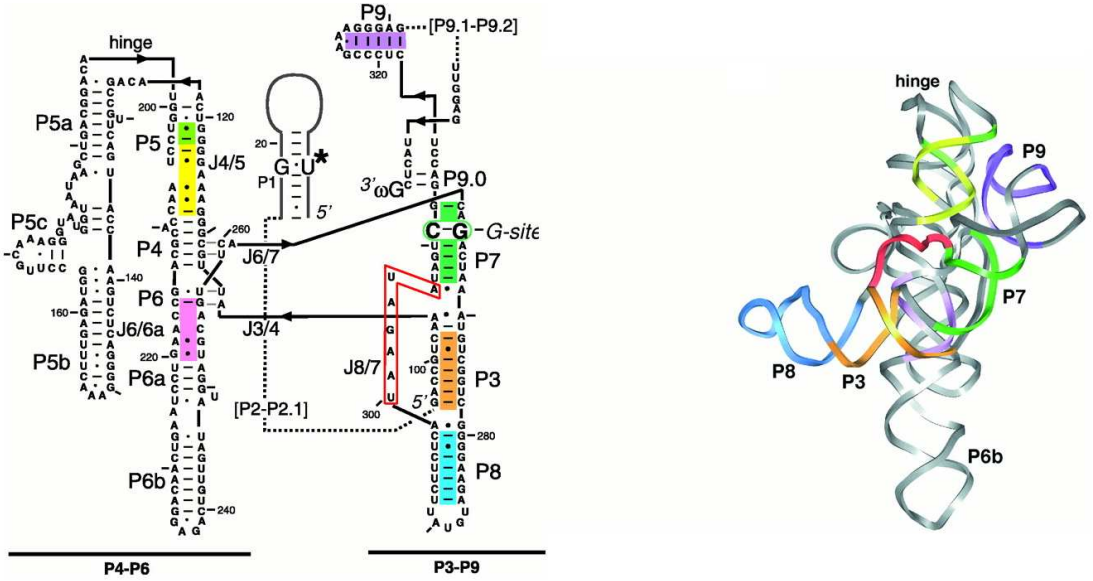


Figure 2.2: Secondary structure and crystal structure of the *Tetrahymena* ribozyme taken from [64]. This group I of *Tetrahymena thermophila* catalyzes self-splicing from a precursor RNA. Conserved helical (paired) elements are designated P1 through P9.2, and joining regions are designated with a "J". This large ribozyme is largely pre-organized for catalysis, much like a globular protein enzyme

RNA catalysis is far less efficient than protein enzymes, as a consequence of its chemical simplicity. Proteins are built from 20 different building blocks, RNA from four. RNA lacks for instance a general acid base with a pKa in the neutral range, as occurs in histidine. Some of these handicaps can be overcome by the

use of modified bases. But since RNA has to work as genetic information carrier as well, the increase of chemical diversity is disadvantageous. Post translatoric modifications could solve this problem, but they need to be catalysts as well and we are facing a chicken-egg problem. Protein enzymes enhance their catalytic abilities by the use of co-factors for otherwise unaccessible chemical reactions (eg. NADH). It is thinkable, that might also be an opportunity for RNA enzymes. An example where ribozymes make use of co-enzymes is Mg^{2+}, that is known to be necessary for the folding for many RNA structures. Roth and Breaker were even able to generate a histidine dependent DNA enzyme that performs RNA cleavage [116].

The $3D$ folding observed for ribozymes effect to the molecules: it increases the resistance to hydrolytic cleavage, which was probably a serious problem to early RNA species. The support for the RNA-world hypothesis is based on the following findings:

- RNA has excellent template properties.
- The discovery of the catalytic RNAs.
- The requirement for RNA in many essential, and presumably ancient, cellular processes such as translation, splicing, and priming of DNA synthesis.
- The presence of ribonucleotides or derived components thereof in most biological co-enzymes.
- The biosynthesis of deoxyribonucleotides by the reduction of ribonucleotides rather than by a *de novo* pathway.

But despite the appealing data some clouds of doubt still cover the sky of the RNA world. It seems likely that an intermediate, pre-RNA world existed, based on a much simpler polymer that was later displaced by RNA. Therefore it is supposed that the origin of genetic information was found in an other template heteropolymer and *transcribed* to RNA subsequently. Later on the genetic information was moved to DNA, and mechanisms for this direction of information

flow still can be seen in the reverse transciptase of retroviruses (such as HIV) and retrotransposons .

The concept of an RNA world and the extensive studies on nucleic acids have led to a rather deep understanding of template chemistry and evolutionary dynamics. The concept of the molecular evolution that hold for the RNA should hold as well for other template polymers.

2.3 Molecular Evolution

The bacteriophage Q_β, which affects *Esterichia coli*, is due to its small and simple genome a well suited model system to study RNA replication. Its 4200 nucleotide genome codes for four different proteins, one of which is a highly specific replicase. The purification of this enzyme opened the door to a series of experiments [8–11, 97] in which the kinetics of RNA replication could be studied. It was possible to demonstrate that Q_β replicase was able to synthesize RNA in absence of a template. In these experiments the *in vitro* evolution of RNA molecules can be followed directly: In a so called serial-transfer procedure (see figure 2.3) the selection of optimal templates could be observed showing that Darwinian evolution directly acts on molecular basis.

In an evolving population of self-replicating RNA molecules competing for nucleotides the faster growing species would sooner or later take over. If limited RNA stability is taken into account the best competitor is the mutant sequence with the most favorable combination of copying fidelity, stability and replication rate. This mutant would together with its "comet tail" of variants compose the so called *master sequence*. The rigorous mathematical description of the model led to the development of the *Quasispecies Model* [38] in the 1980ies. One of the most important conclusions of this theory is that there is a threshold condition for the stable replication of genetic information. Therefore, the accuracy of replication determines the maximum gene length l_{max} of the master sequence calculates from:

$$l_{max} = \frac{\ln \sigma}{1 - \overline{q}_m}$$

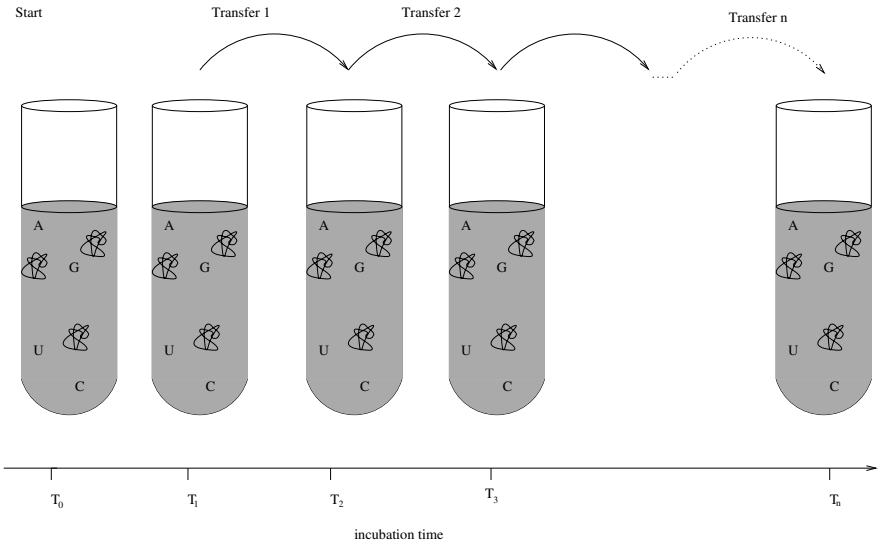


Figure 2.3: An aqueous stock solution highly purified containing Q_β replicase, monomers of A, G, U and C, but no RNA is incubated for a period of time. Then a portion of that solution is transfered to fresh stock solution and incubated again, over and over again. During this procedure a *fitter* template strand of RNA is selected from the randomly generated population of molecules.

Here σ denotes the advantage factor of the master sequence and q_m its copying fidelity. If the error-threshold is violated, the quasi-species is destabilized. The master-sequence is then unable to withstand the accumulation of errors, the population starts to drift and all information is lost. In a prebiotic world the error-threshold would tolerate sequences that must not be longer than 100 nucleotides and rich in GC content (higher GC content lowers the mutation rate of self replicating RNAs) self-replicating . Only the invention of the translation, and the development of enzymes that have improved copying fidelity and succeeded this information crisis. This was the point where phenotype and genotype were separated.

Another approach to increase the copying fidelity is to use an information carrier that can by itself distinguish between right or wrong. A successive approach would be that the daughter strand remains at the parental template, where a wrong base simply would not pair. This called DNA onto the plan of life. DNA forms stable double helices and gains additional bonus by its higher resistance to hydrolytic cleavage because of its missing 2'-OH group in the β-D-2 Deoxyribose.

The improved fidelity led to a second information crisis: A lower mutation rate decreases the variation of the species. This could be overcome by the development of a recombinative process that led to sexual reproduction. So the step from Darwinian driven self reproduction to Mendelian Genetics was taken.

Selection has to take place on the genotype, however the fitness evaluation affects the product of the genotype. This requires that the gene product feeds back information to its gene. Such double-feedback loop were extensively investigated by Manfred Eigen and Peter Schuster, who called this behavior a *Hypercycle* [39–41]. Hypercycles alone allows many quasi-species distributions to coexist within the same soup. In a primordial soup many interacting RNA and protein molecules formed hypercyclic networks. To evaluate the fitness of a single gene, compartimentation had to take place, separating the cell from the environment and this was the basis for evolutionary optimization of genes and their products, thereby being natures solution for the genotype-phenotype dichotomy.

2.4 Toward a Riboprotein World

If one takes the RNA-world hypothesis is taken for granted, the question what came next remains open. RNA delegated its function: information storage was shifted to DNA, catalytic functions were deputed to proteins. As stated above the error threshold requires either of the two to enable longer genomes, but does not predict the order. Desoxyribonucleotides are bio-synthesized by reducing ribonucleotides, and thymine by methylating uracil. The responsible enzyme in extant organisms is ribonucleotide reductase. This protein was shown to be monophyletic and uses an energetically expensive and biochemically unusual radical reaction. It is extremely difficult to design a Ribozyme that performs the ribonucleotide reduction, and it was so far not possible to retrieve it in SELEX experiments [55]. Taking into account that almost no significant amount of desoxyribonucleotides was accessible prebiotically it is unlikely that DNA occurred before amino acid portions enlarged the catalytic possibilities of RNA. Another evidence for this order of occurrence comes from the distribution of catalytic RNA within extant metabolism: In almost all important steps of translation ri-

bonucleotides take key roles. DNA is transcribed to mRNA that uses an RNA adapter (tRNA) to interpret nucleotides in amino acids. The loaded tRNAs are processed in the polymerizing step of peptide synthesis at the ribosome, which consists of two unequally sized subunits.

Stripping a large portion of proteins of the large ribosomal subunit still remains the peptidyl transfer reaction [81]. It seems that only structural constrains of the 23s subunit limit the complete removal of peptides from the subunit. It could be hypothesized that in a much less elaborated interaction positively charged amino acids could stabilize the polyanionic ribozymes. Only in a later stage the sequence specificity gave rise to the complex process of peptide translation.

The invention of proteins by prebiotic molecular species required a collaboration between nucleotides and amino acids. This relationship presumably evolved stepwisly and could have started by the use of amino acid cofactors for ribozymes at a first step [136]. In a time where RNA and proteins were "sharing work" there was an interplay of structure and function. Proteins for instance can provide a protective shield against nucleases, whereby RNA performs catalysis. An example for a present day enzyme, where the proteins serves a scaffold and the RNA acts as catalyst is RNase P and even the ribosome itself.

The more complex the chemical patterns of amino acids were, the higher their number became. In an RNA-world shifting to the employment of proteins RNA must also have carried out amid bond formation. Again the results from *in vitro* selection experiments provided evidence: Several laboratories were able to select RNA molecules that catalyze amid bond formation from a large set of random RNA polymers [92, 147, 158].

As amino acids overtook more and more of the catalytic duties, the genetic information established so far had to be rewritten, a *translation* into the language of amino acids by specific interaction was inevitable. The translation required a common table of nucleotide-to-amino acid equivalence hence this was the time to write down the genetic code.

2.5 The Origin of Translation

Translation raises a typical chicken-egg problem: To perform protein translation an elaborated machinery of specialized enzymes is necessary. This machinery must be produced before translation can take place at all. It seems reasonable to start this process in a simplified form using only a restricted set of amino acids that were of prebiotic origin.

Two plausible scenarios for the invention of a genetic code can be drawn for the RNA world:

In vitro selection experiments were used to evolve aptamers that specifically bind to amino acids [89]. The nucleotide distributions found in these small RNA molecules strongly suggest a role of chemical determinism in shaping the codon assignment for distinct amino acids. Ribozmes might have assembled short peptides that were able perform a feedback. This feedback of proteins and the mechanism of their translation led to a stabilization of the genetic mapping [13].

Another possible scenario arises from the usage of amino acids in RNA catalysis: the more RNA depended on proteins, the higher the peptide content became and proteins started to acquire more and more of the ribozymes abilities.

A broader overview of the common theories about the origin of the genetic code is given in the next section.

CHAPTER 3

The Genetic Code

Linear nucleotide sequence are the major information carrier of all living organisms. The interpretation of the information depends on the application of a code to translate the four letter alphabet to the 20-letter alphabet of proteins. This code must be independent of the specificity and meaning of the genetic message, because it has to enable any kind of "communication" between DNA and proteins. So the problem nature was facing before the invention of the genetic code can be seen with the eyes of information theory. To say it in the words of this discipline's pioneer:

> "The fundamental problem of communication is that of reproducing at one point either exact or approximately a message selected by another point." *Claude E. Shannon 1948* [125] .

3.1 Deciphering the Code

The Big Bang and the genetic code are two scientific ideas that dramatically changed most our view of the world in the twentieth century. The big bang

tries to explain the creation of the universe, while the genetic code manifests the phenotype of an organism from the inherited material. Interestingly both ideas were introduced by the same man: George Gamow [124]. As a response to the historic *Letter to Nature* of Watson and Crick [145], Gamow suggested the existence of 20 amino acids and a direct correlation of DNA as information carrier for proteins.

He was the first to recognize that this was actually an abstract *problem of coding* and in his "diamond code" he suggested that the bases are read from the edges of the DNA grooves in the double helix [59]. This was the first triplet code (because 2 bases in the diamond were assumed to pair and therefore having the information content of just 1), and by eliminating the symmetry underneath the permuted diamond codon words he ended up in 20 codons. The diamond code was thought to be an overlapping code, therefore a sequences of length 4 would be translated to a di-peptide. Arguments for this hypothesis were storage efficiency and the elimination of the frame-shift problem. This was the point where Francis Crick could falsify it: There are 400 (20^2) possible amino acid sequences of length two, but only 256 (4^4) combinations of four nucleotides in a sequence. It was therefore not possible to represent all amino acid sequences as DNA, and this violates the requirement for a code in general, as well as experimental evidence for failure were soon found. Soon after all overlapping codes were ruled out by nearest-neighbor correlations of all known protein sequences. But this brought the frame-shift problem back.

Until that point the amino acids were thought to interact *directly* with the nucleotide, and Crick postulated the existence of adapter molecules. To overcome the frame-shift problem of non-overlapping codes he postulated that since only a limited number of adapters exist, some codons are non-sense. Therefore the right frame was the one with maximum number of sense-codons. This was a so called *comma-free* code, a code that retains its meaning even without the existence of special separators (commas, spaces, ...). In a comma-free code the homogeneous codons (AAA, UUU, CCC, GGG) had to be excluded and many speculations and mazes of coding-schemes followed. This era was ended by Marshal W. Nierenberg, as he published cell-free *in vitro* protein synthesis [104]. In these

experiments poly-U was translated and yielded an oligopeptide of phenylalanine. This proved general comma-free codes to be insufficient to explain the genetic code. Nirenberg and Matthaei's experiments involved incubating RNA samples with a "soup"(cell-free extract) of bacterial ribosomes, enzymes, ATP (an energy source), tRNA, and amino acids tagged with carbon-14 for later detection. By a similar protocol more codons were solved and soon the genetic code as was solved by Nirenberg [102, 103].

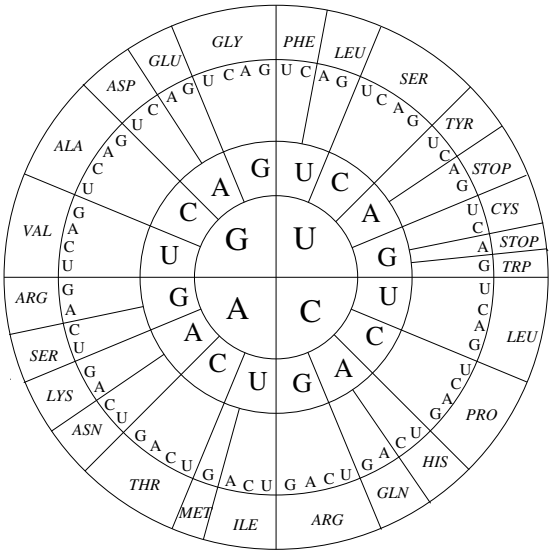


Figure 3.1: The universal genetic code in circular representation.

3.2 The Universal Genetic Code

The genetic code is both, a physico-chemical system and a communication system. Information can be transfered from DNA to RNA and from RNA to protein, but not from protein to protein. This *Central Dogma* of molecular biology was introduced by Francis Crick in 1968. The reason for this theorem roots in coding theory: the four letter alphabet of DNA (2 bits per nucleotide required) is expanded to a 64-codon letter alphabet (3×2 bits). It is necessary to use 6 bits of information since one position in a nucleotide sequence encodes 2 bits, and the maximum number in a 4 bit alphabet is 16. This is insufficient to code for 20 amino acids. The mapping of the code is analogous to a logical ADD-gate that

can only be passed in one direction without loss of information. It could be shown in mathematical generality that a loss of information happens if communication between systems where the information entropy of the source alphabet is larger than that of the receiver is forced. This argumentation also holds to be true for the information flow of the *reverse transcription* of some retro viruses where RNA is transcribed to DNA, because the information entropy remains equal.

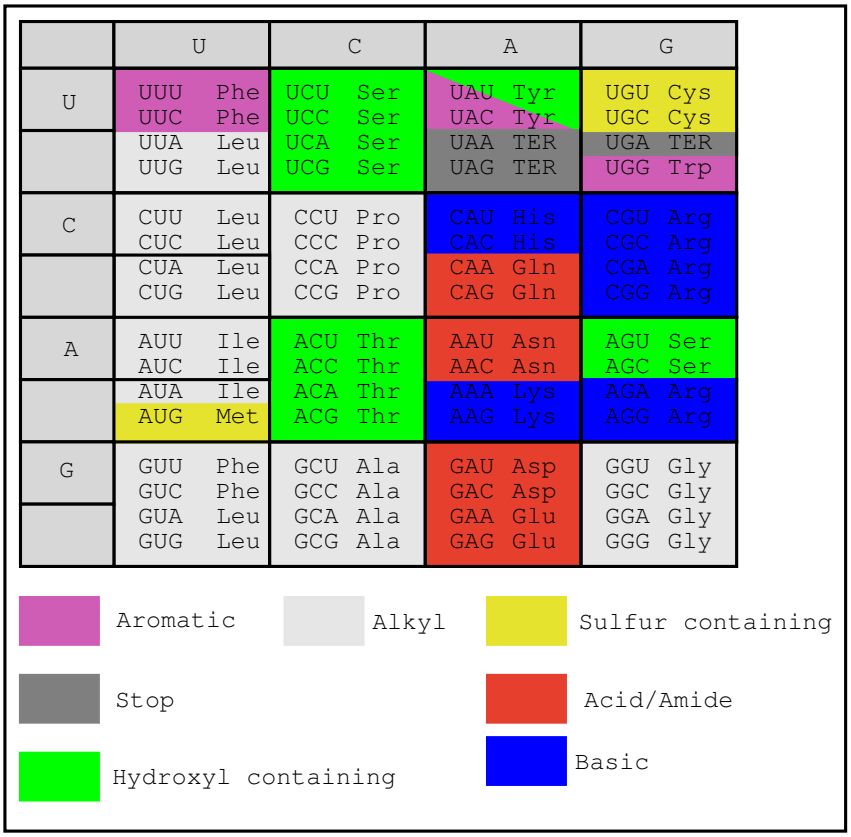


Figure 3.2: The universal genetic code is a block code. Amino acids with similar chemical properties are found within mutational nearness of each other. (Figure adapted from [83])

The universal genetic code is a block code as easily can be seen in figure 3.2. This means that codons that differ in one base are usually assigned to the same or a similar amino acid forming a so called *family box*. There are only seven groups of codons, where 2 amino acids share the first two bases of the codon. These "split" boxes have in common that they have either A or U (or a combination of them) in the first two positions. Since GC pairs are characterized by a significantly higher base pairing, energy the code redundancy can be caused to thermodynamics. The

codons can be grouped by similarity and ordered by their Hamming distance. Physico-chemical similar amino acids are observed to appear in close mutational proximity within the code. This implies interesting symmetries with respect to physico-chemical properties within this code. A certain degree of *fault tolerance* is achieved by the fact that the nucleotide in the third codon position is neutral with respect to amino acid mapping in many cases.

The canonical genetic code is remarkably redundant. Its degree of degeneracy is determined by the fact that some codon-anticodon interactions are indistinguishable. The block-like structure can be drawn as six dimensional Boolean Hypercube [76], each node represents a codon and is separated by a one-bit change from other nodes. The Hamming distance between two nodes is therefore determined by the number of bits differing between two nodes. Within a four dimensional subspace of NXN with X $\in$ {A,C,G,U} changes lead to silent mutations whereas mutations of the XNN class are non-conservative as frequently found in proteins. This illustrates well the interplay of redundancy and innovative opportunity within the structure of the genetic code. Codons that code for similar amino acids typically form clusters in the table structure (see figure 3.2, for example codons that have a U at the second position (NUN) code for hydrophobic amino acids, whereas codons that have an A at that position map to hydrophilic amino acids. Aromatic amino acids (Phe, Trp, Tyr) are encoded by triplets that carry a U at the first position.

Furthermore it is apparent, that neighboring amino acids tend to be related by polarity value [149], biosynthetic relationship[155], or both. A statistical rules, a kind of "code within the codon" [137] predicts that the first and second codon base indicates biosynthetic relationship and amino acid polarity. The relations between amino acids and anticodon nucleotides led to the hypothesis that both effects shaped the code. A correlation formulated by Jungck [77] relates nucleotide hydrophobicity with amino acid polar requirement and bulkiness(the ratio of side chain volume to length), what is consistent with the idea that a stereochemical effect influenced the early evolution of the code.

A more general description of the code deals with the embedding of codes [101,

148]. A generic code $\mathcal{O}$ is defined by

$$\mathcal{O} = \mathcal{C} \times \mathcal{A}$$

whereby $\mathcal{C} = \{c_1, \ldots, c_\lambda\}$ is the set of codons and $\mathcal{A} = \{a_1, \ldots, a_\lambda\}$ denotes the set of amino acids both of finite length λ. A subset of ordered pairs $S_\pi = \{c_i \mapsto a_j, i, j \in \{1, \ldots, \lambda\}$ is called a code, where π denotes a permutation of i, j. The embedding of a code into protein sequence space describes the reflexive relationship between information and function that evolved. This goes back to the concept of auto-catalysis [80], which is required in a feedback system like the translation to evolve and to select a distinct code among its vast number of variants in its primordial origin.

Polynucleotide sequences of simple organisms or self-replicating molecules, having little or no aid by efficient enzymes to reproduce were shown [41] to require a high GC-content to maintain mutational stability. The efficient translation of the first self-organized sequences is crucially dependent on the ability to keep information acquired by self-organization. An emerging coding system must have been able to read-off nucleotide systems uniquely, and since no "separator" appears at the messenger, the code must be able to act without frame-shifts, therefore *comma-free*. A frame-shift error is especially grave since it exterminates all subsequent information coded.

3.3 The Genetic Code is Not Universal

The most stressed evidence for the evolution of the genetic code is the fact that the code is not "universal" as originally proposed. The first derivates were observed in vertebrate mitrochondria, soon many more were identified among different phyla (see figure 1.2). Interestingly, some changes occur independently in related linages implying multiple changes within a short period of time during evolution. Several codons seem to be more easy changeable and were assigned to different amino acids. For instance AGG has been reassigned from Arg to Ser, Gly, and STOP. Especially STOP-codons seem to be an evolutionary degree of freedom. Their neutrality may be achieved due to their rareness (they occur once per

gene) and the fact that transcriptional release factors are easy to change [107]. Another factor that makes reassignment evolutionary feasible is the frequency by that codons occur.

The codon usage among different species is extremely biased. For synonymous codons this means that some organism have distinct preferences while others use redundant codons equally. Table 3.1 reveals, that for instance the two lysine codons (AAA and AAG) are used with opposite affinity in *Lactobacillus acidophilus* and *Streptomyces venezuelae*. The inhomogeneous codon usage among taxa has direct impact on practical applications such as PCR primer design or phylogeny reconstruction. Hypothesis that correlate codon usage with GC content can be shown to match the observed distribution of codons under respect of the codon position and the frequencies of nucleotide exchange [84].

Changes in the genetic code can be introduced by several components of the translation apparatus, eg. mutation of the tRNA (change identity elements), mis-pairing of codon and anticodon or post transcriptional modifications. The possibilities of changes are limited by the impact of change (most changes will be deleterious as proposed by the *frozen accident* hypothesis). There also seems to be a restriction within the recognition ability of the codon-anticodon pairing: no evidence is found that any C can be identified in the third position. This is manly based on the wobble effect of base pairing.

In recent years three mechanisms of codon changes especially in mitrochondria were published and each of them predicts certain codon changes that have not yet been observed.

Codon Capture Hypothesis

The "codon capture" theory states [106] that specific codons disappeared by AT or GC pressure from the code. Hence mutations in tRNAs coding for these codons are neutral and if the pressure relieves the codons reappear and may code for a different amino acid. Support for this theory comes from mitrochondria code, where genes are AT rich and small.

Codon	Amino Acid	freq *S. venezuelae*	freq *L. acidophilus*	Codon	Amino Acid	freq *S. venezuelae*	freq *L. acidophilus*	Codon	Amino Acid	freq *S. venezuelae*	freq *L. acidophilus*	Codon	Amino Acid	freq *S. venezuelae*	freq *L. acidophilus*
UUU	F	–	0.65	UCU	S	0.01	0.19	UAU	Y	0.03	0.50	UGU	C	0.08	0.57
UUC	F	1.00	0.35	UCC	S	0.42	0.05	UAC	Y	0.97	0.50	UGC	C	0.92	0.43
UUA	L	–	0.40	UCA	S	0.01	0.36	UAA	*	–	0.66	UGA	*	0.88	0.11
UUG	L	0.01	0.22	UCG	S	0.29	0.04	UAG	*	0.12	0.23	UGG	W	1.00	1.00
CUU	L	0.02	0.21	CCU	P	0.02	0.33	CAU	H	0.05	0.61	CGU	R	0.08	0.48
CUC	L	0.51	0.06	CCC	P	0.44	0.06	CAC	H	0.95	0.39	CGC	R	0.47	0.09
CUA	L	–	0.07	CCA	P	0.01	0.54	CAA	Q	0.02	0.88	CGA	R	0.02	0.10
CUG	L	0.45	0.04	CCG	P	0.54	0.08	CAG	Q	0.98	0.12	CGG	R	0.39	0.07
AUU	I	0.02	0.68	ACU	T	0.01	0.66	AAU	N	0.02	0.53	AGU	S	0.01	0.20
AUC	I	0.96	0.24	ACC	T	0.65	0.12	AAC	N	0.98	0.47	AGC	S	0.26	0.16
AUA	I	0.02	0.08	ACA	T	0.02	0.15	AAA	K	0.03	0.48	AGA	R	0.01	0.23
AUG	M	1.00	1.00	ACG	T	0.31	0.06	AAG	K	0.97	0.52	AGG	R	0.03	0.03
GUU	V	0.01	0.58	GCU	A	0.02	0.50	GAU	D	0.04	0.68	GGU	G	0.10	0.64
GUC	V	0.64	0.08	GCC	A	0.62	0.14	GAC	D	0.96	0.32	GGC	G	0.67	0.18
GUA	V	0.03	0.26	GCA	A	0.03	0.30	GAA	E	0.14	0.86	GGA	G	0.07	0.14
GUG	V	0.32	0.08	GCG	A	0.33	0.06	GAG	E	0.86	0.14	GGG	G	0.16	0.05

Table 3.1: Codon usage of *Lactobacillus acidophilus* and *Streptomyces venezuelae* (codons that are not in use are marked with a dash '–'). Data has been taken from the *Codon usage database* available via the Internet under URI `http://www.kazusa.or.jp/codon/`. Distribution given as frequency per thousand in species' genes available from GenBank Release 127.0

Ambiguous Intermediate Hypothesis

The "Ambiguous Intermediate Hypothesis" proposes [156], that codons do not disappear while under change, but undergo a period of ambiguity. In this phase single codons are translated to two different amino acids. This takes into account that RNA mis-pairs in some cases ($\mathsf{G} \cdot \mathsf{A}$ and $\mathsf{C} \cdot \mathsf{A}$ pairing at the third and $\mathsf{G} \cdot \mathsf{U}$ airing at the first position). Support also comes from yeast where it has been reported that a mistranslation between Ser and Leu at the CUG codon occurs.

Genome Streamlining Hypothesis

The "genome reduction" theory proposes [3] that simplification of the translation apparatus is the driving force for codon reassignment in mitrochondria. The shortening of the genome brings direct selective advantage, and the size of a single tRNA is significant for very small genomes. This is the driving force for the loss of tRNAs.

3.4 Origin of the Code

Based on symmetry considerations and simple base pairing logic it is possible to construct patterns that are able to produce comma-free codes. GC-stability has to be considered as well as plus-minus symmetry from an evolutionary point of view. From known features of the anti-codon loop codes matching the "RNY" patterns are considered to be particularly interesting. Manfred Eigen proposed [42] that the first codons were GGC, GCC, GAC and GUC today coding for the Gly, Ala, Asp and Val. Interestingly, these are some of the amino acids suspected to be primordially available according the experiments of Stanley Millers experiments [96]. Statistical analysis of tRNAs and genomic sequences in general revealed a periodic re-occurrence of the RNY pattern and showed a high predominance of this structure, reflecting genetic code properties.

3.5 Hypotheses on Genetic Code Evolution

The key role in living beings and the mysterious block structure of the genetic code inspired many scientists to yarn their theory, some based on facts, others not. In the 30 years since the discovery of the chart of amino acid nucleotide mapping key methods such as SELEX, automated DNA sequencing and synthesis were developed, and each technique brought new insights that contributed the puzzle. A major obstacle for models concerning the origin of the genetic code is the fact that for an efficient protein synthesis powerful enzymes are required, what ends in a chicken-egg problem. In the next four sections common theories about the genetic code origin are reviewed.

3.5.1 Frozen Accident

It was Francis Crick himself, who proposed that the genetic code was an evolutionary *accident*. Crick suggested that the sacrosanct, generic code was established in the last common ancestor and *frozen* since then [26]. Therefore the observed pattern requires no further explanation and makes any further analysis unnecessary. The block structure is simply explained by the wooble hypothesis, thereby explaining the base mis-pairing by chemical reasons of base-pairing mechanisms. The necessity for this redundancy comes from the fact that a single *adapter* (tRNA) decodes many codons, but is charged by only one single amino acid. The Frozen accident model explains where the genetic code comes from, but does by no means predict the observed order. This is in contradiction to the code variations that were observed among different taxa (see figure 1.2 on page 9). There are variations in the translation of synonymous, initiation and termination codons, indicating that the genetic code cannot be considered as truly universal.

Crick's major argument was that a change in the genetic code causes changes in *all* proteins of the organism, which are likely to be deleterious or at least very strongly selected against. Therefore successful changes of a code that an organism once relied on, are very unlikely. This fact locks the organism's code and makes it inaccessible to evolution any more.

Mechanisms and linages for the code's isomers and their transmutation are suggested and shown [106]. The patterns that are observed within the codon table are real in a statistical mathematical sense. This could be shown by various analysis, e.g. the comparison codon correlations of randomly generated codes [2]. Nevertheless the *frozen accident* provides a valuable "null-hypothesis" that can be used to test other theories against.

3.5.2 Stereochemical Similarities

Stereochemical theories propose that the specificity of a codon for a particular amino acid is based on a direct interaction of amino acid and nucleotides. Using semi-empirical potentials it was possible to verify a key-lock like fitting of the anticodon-loop plus the discriminator base (the first base upstream the anti-codon) and the cognate amino acid (*C4N model*). Such stereo-chemical correlations explain well the universality of the genetic code, since there should only exist one optimal matching of RNA to amino acid interaction and the nature of this interaction would be a *frozen stereochemical accident*. The selective benefit of such a behavior is obvious: single RNA mutations change the chemical pattern of the RNA trimer only slightly, and the amino acid that fits a mutated pattern best would be chemically similar to its wild type. Suggested sites of this interaction are the anti-codon loop of the tRNA. Alternatively the amino-acid RNA recognition was proposed to take place at the tRNA acceptor stem [75]. This is consistent with some evidence that the acceptor-stem and anticodon-loop might have evolved independently [120].

Stereochemical affinities might have influenced early codon-amino acid pairings, but evidence for many amino acids is still missing, though the repertoire is expanded from originally arginine to leucine and tyrosine. Some amino acids such as tryptophan, glutamin and asparagine may have entered the code relatively late, what is consistent with Wong's co-evolution theory (see section 3.5.3). On the other hand chemical similarity requirement set tight restrictions on the amino acids that were code-able at all. Not surprisingly, some amino acids present under prebiotic conditions were excluded. This was blamed to be the reason for the selection of the standard amino acids used in translation. Therefore elaborated

enzymatic modification mechanisms had to be evolved to enlarge the repertoire of available amino acids and increase catalytic activities and structural robustness of proteins.

An explanation of the primordial translation could under a stereo-chemical point of view of the code have happened by direct interaction of the amino acids with the RNA template. Their close spatial arrangement enabled a ribozyme catalyzed condensation. The binding sites would in this model act more as a structural template, than as a sequence. The assumption of ribozyme mediated condensation is well supported by the fact that in modern ribosomes that a lion's share of work is done by RNA, and peptide bond forming ribozymes could be isolated.

Another possible explanation of the stereochemical correlation of amino acids and RNA statues that the genetic code arose before translation, and was originally used to select amino acid cofactors for ribozymes [136]. This theory, called Coding Coenzyme Handle (CCH), views the anti-codon loop as stereochemical adapter, that was charged and used by primordial ribozymes to compensate missing functional groups. Amino acids were covalently linked to particular oligonucleotides (handles), which could than base-pair with ribozymes, although a direct binding of amino acids to nucleotide triplets is usually not observed in solution. In a later bifurcative step the adapter and the enzymatic moiety were separated and became tRNA and mRNA.

An experimental protocol to test stereo-chemical theories seems reasonable by *in vitro* selection experiments. RNA oligomeres have to be selected that have the highest affinity to bind all possible amino acid. This can be done by repeated selection of large pools of randomly generated RNA molecules over several generations [90]. Such a scenario mimics the RNA world, where short RNA strands and amino acids were available.

In particular arginine has been studied intensively and aptamers reflecting the arginine assigned codons were found more often that expected by chance [85]. Arginine is special in many respects: its guanidino moiety is able to mimic the hydrogen-binding face of guanidine and arginine is positively charged, what makes electrostatic interactions with the poly-anionic RNA molecule probable (structure shown in figure 3.3). This unspecific interaction seems not to be sufficient to

explain the structure of the aptamers because it has been reported that the selection of lysine aptamers failed [49].

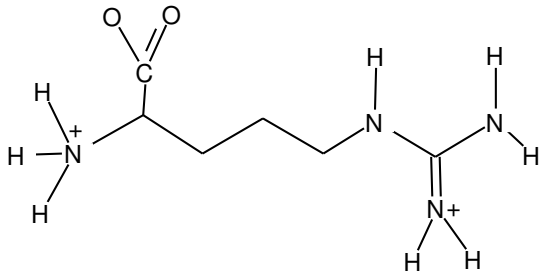


Figure 3.3: Structure of arginine

In their work Knight and co-workers investigated arginine aptamers, that is small RNA oligomeres selected to specifically bind arginine. They correlated the aptamer sequences in structural motifs and sequential properties and tested significance against non-related aptamers to eliminate effects of general aptamer sequences. They were able to demonstrate a high over-representation of purines, which is typical for arginine codons (AGN and AGR) and compositional effects (permutations of the triplet) could be eliminated. It has been concluded that a significant correlation between arginine and its codon exist in contrast to the anticodons. Additional experiments revealed that the over-representation of arginine codons in aptamers is not restricted to RNA, but can also be found for DNA aptamers as well. This independence from the backbone chemistry is important with respect to theories suggesting a pre-RNA world based on simpler molecules (see section 2.2 for examples) because of the difficulties of a prebiotic RNA synthesis.

A problem with this kind of direct adapter-amino acid interaction is that it does not explain the appearance of the tRNA [45]. This class of biomolecules forms the canonical cloverleaf and lots of evidence support a monophyletic origin. Comparative sequence analysis of transfer RNA by the method of statistical geometry in sequence space revealed, that a significant part of present day tRNAs date back to the time where archaebacteria separated from eubacteria [37]. Using an aptamer-codon hypothesis it can be explained that specific codons might have been selected, but does not take into account the great benefit of tRNAs as their

structure greatly facilitates poly peptide chain elongation. There is no evidence for a direct participation of the codon in modern translation, the tRNA keeps it spatially separated (> 70Å) from the transpeptidyl reaction [61].

The adapter hypothesis also lacks an explanation of the "parity problem": it is equally likely that amino acids have originally contacted codons and anti-codons as pointed out by Knight and Landweber [86]. The authors propose a solution of this dilemma by a modified CCH theory. It is suggested that RNA sequences acquired a selective advantage by specifically binding an amino acid (e.g. increased catalytic activity, higher resistance to degradation or energetically favorable charge distribution). Later catalytic activity for *in trans* or *in cis* aminoacylation activity was gained by the RNA. As ribo-organism relied more and more on amino acids an independent carrier for the amino acid would be favorable, because re-using the carrier yielded a higher turn-over. This could have led to the development of tRNAs.

It is a fact that the genetic code is a kind of "optimal" with respect to single base mutations. This could be demonstrated by Freeland and Hurst [54]. But this reveals another problem with the direct amino-acid nucleotide interaction: There is no biophysical reason to assume that particular codons are related to amino acid structure. For instance the arginine aptamers are of built of bases matching the pattern CGN, whereas isoleucine aptamers contain a preponderance of A and U (AUN codons). There is no evidence to assume better interaction of GC to hydrophobic or of AU to hydrophilic residues.

For messenger RNAs that are functionally selected such as rev responsive element (an RNA secondary structure involved in regulation the transport of unspliced RNA to the cytoplasm) additional evolutionary pressure limits the amino acid sequence that can be coded. The hypothesis of compatible coding [88] emerges from the necessity for elaborated secondary structure in an RNA world and predicts an influence of secondary structures on the selection of codons. It is further predicted that that RNY patterns are predominantly found in stem regions of mRNA secondary structures.

Ellington *et.al.* [45] criticize massively the arginine hypothesis because of shortcomings in the aptamer selection and statistical methodology. He suggests al-

ternatively that an RNA aptamer amino acid interaction is responsible for an intermediate step in the evolution as function specificity taken by peptides came before sequence specificity.

3.5.3 Co-evolution Theory

In 1975 Wong postulated the so called co-evolution theory [155]. This theory tries to explain the non-randomness of the code by stating that the code system is an imprint of the prebiotic pathways of amino acid formation. Hence the genetic code and its evolution reflect the precursor-product relationship among amino acids and their bio-synthesis. On this basis it is possible to embed a graph that maps the codons (single base chance per edge) to groups of biosynthetically related amino acids. Two amino acids are defined to be near each other, if their bio-synthetic pathways are related and separated by few enzymatic steps. The product-precursor pairs that were originally compiled by Wong [155] are listed in table 3.2.

Asp → Asn	Glu → Gln	Ser → Trp	Thr → Ile	Aln → His
Asp → Thr	Glu → Pro	Ser → Cys	Thr → Met	Val → Leu
Asp → Lys	Glu → Arg			Phe → Tyr

Table 3.2: Product-precursor relations as used by the co-evolution theory.

The earliest code used only a small subset of prebiotically synthesized amino acids (such as Gly, Ala and Ser) which were coded by an extremely degenerated code. The evolutionary development of biosynthesis made amino acids (such as Arg, His, Trp) available that are present in current day organisms. The degenerated code "learned" to specify more detailed and incorporated the newly available amino acid words. The evolution of the contemporary code can therefore be followed by a detailed analysis of synthesis pathways.

If a codon is reassigned, the newly incorporated amino acid is derived from its metabolic precursor, and hereby similar. This implies that error minimization is generated even without explicit selection for it.

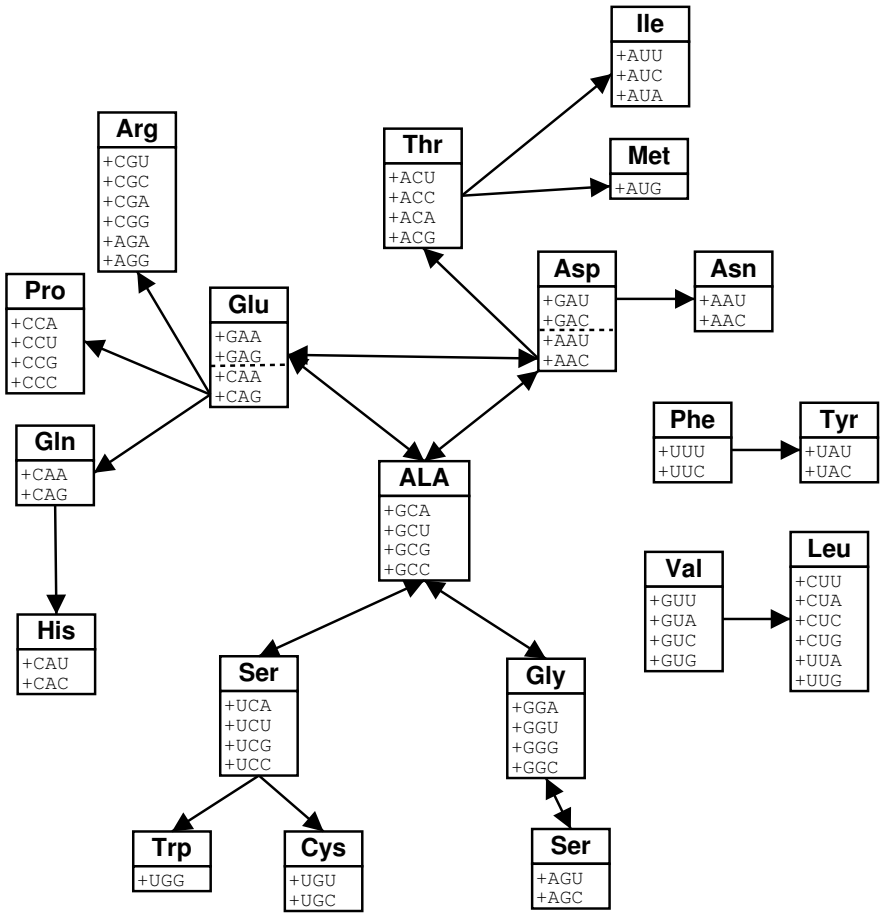


Figure 3.4: Evolutionary map of the genetic code (adapted from [155]): Each amino acid and its contemporary codons are represented as single box, if a dashed line separates codons, those were available at earlier stages. Single headed arrows show precursor-product relations, double headed arrows imply biosynthetic interconversion. Each connection lines corresponds to a hamming distance of 1 in codons.

A possible scenario for the direct interaction of amino acid metabolism and codon selection was proposed by Di Giulio [30]. RNA hairpin structures, reflecting precursor tRNAs were directly used to locate the biosynthetic pathways of amino acids [32]. These hairpins, charged with amino acids gave rise to a primitive protein synthesis, and the organization of the genetic code.

To test the co-evolution theory Amirnovin [2] generated a huge number of random codes to determine if similarities found between codons of related amino acids could be generated by chance. The codes were produced by assigning the 20 amino acids plus STOP to groups of codons that reflect the blocks observed

in the universal code. The codes are analyzed for correlation of the biosynthetic related amino acids using Wong's originally designed list as well as a more recent map. Amirnovin showed clearly that a significant fraction of random codes can even outperform the natural cousin in creating correlations. Therefore the code correlations between related amino acids cannot be taken as proof for the biosynthetic co-evolution theory and that the observed pattern is extremely dependent on the choice of amino acid similarity.

The strong relationship of the selection of the product-precursor pairs provoked a closer investigation by Ronneberg and co-workers [115]. Co-evolution theory defines a precursor amino acid as one in which any portion of the amino acid (side chain or backbone) is metabolically incorporated into the product amino acid. A thorough analysis of the fundamental biochemical relations showed that assumptions of product-precursor pairs are wrong. On the one hand the assignment of product-precursor seems wrong. For instance the Glu $\rightarrow$ Arg are separated by six enzymatic steps, whereas Asn is only twp steps from Arg. On the other hand some product-precursor pairs are rather alternative branches in the metabolic pathways than products and precursors (e.g. Val $\rightarrow$ Leu). Some interconversion proposed by the co-evolution theory are thermodynamically prohibitive since their inversion is mediated by ATP hydrolysis in modern metabolisms. Ronneberg and coworkers recalculated the significance of codon pattern correlation to product/precursor pairs of amino acid and found it to be vanishingly small [115]. Taking the available evidences together it is questionable if the co-evolution theory is an adequate explanation for the structure and origin of the genetic code.

3.5.4 Adaptive Codes

Another attempt to explain the observed patterns within the genetic code and its development are those that postulate optimality to the code. The pattern of the code is hypothesized to result from adaptation that optimizes a function such as minimizing the number of errors arising from mistranslation. This seems reasonable because it is disadvantageous to accumulate lots of errors in a protein and translational or replicational errors always occur. In fact, the genetic code is supposed to be optimized to reduce the effect of mutations and mistranslations. The

degeneracy observed within the genetic code certainly helps to neutralize changes, but can obviously be explained on basis of the Wobble-Hypothesis. There are two fundamental approaches to show that the genetic code is optimal: statistical and engineering approaches. Statistical approaches compare the natural code with many randomly generated ones. Engineering approaches on the other hand compare the natural code with the best possible alternative.

Carl Woese was among the first to think about adaptive codes. In his pioneering work [149, 151] he suggested that the patters within the genetic code reflect physico-chemical properties of amino acids. Woese introduced a measure for the polarity of an amino acid, the so called *polar requirement* [149] that is defined as a partitioning coefficient of an amino acid in a water/pyrimidine system. The distribution of amino acid polar requirement or hydrophobicity are built in a way to minimize the effect of point mutations. The conservation is explained by a greater frequency of translational misreadings in the first and second position as observed in vitro. Hence the genetic code adds another dimension of neutrality to the evolutionary frame of molecules.

Massimo Di Giulio compared [33] random codes with the native one with respect to the do-called polarity distance (a normalized, chemical scale, derived from an ethanol to water interaction parameter [29, 32]). The arbitrary codes were generated by relabeling the amino acids of the canonical code, thus conserving the block structure of the native code. Thereby Di Giulio estimated that the genetic code has achieved 68% minimization of polarity distance. Or to put this in other words: the genetic code is far from optimal. Di Giulio further postulate from these results that the genetic code is not the product of adaptation. However one can not imply that the product of optimization is necessarily optimal.

This approach of an engineered code is derived from a comparison of the distance between the mean and the optimal code. Under the assumption of single base changes let N_{ij} be the number of times the i-th amino acid changes into the j-th amino acid, and X_i be the polarity index (as in [29, 32]) of the i-th amino acid, the percent minimization is defined by:

$$\frac{\Delta_{mean} - \Delta_{code}}{\Delta_{mean} - \Delta_{opt}} \tag{3.1}$$

where

$$\Delta^2 = \frac{\sum_{i,j}(X_i - X_j)^2 N_{ij}}{\sum_{ij} N_{ij}}$$

Δ_{mean} is the average Δ value for many random codes and Δ_{opt} is an approximation of the lowest possible Δ value. Di Giulio, calculated Δ_{opt} analytically using the method of Lagrange multipliers to solve a constrained minimization problem. Other authors used heuristic computer search algorithms (e.g. [65]).

Haig and Hurst attempted to quantify the effect of minimizing the effect of point-mutations [68] by comparing the native code with randomly generated ones in a statistic approach. These codes were designed to have the native block structure as well, partitioning the 64 codons into 21 non-overlapping sets of genotypes that the 21 phenotypes (20 amino acids and the STOP signal) map to. The authors found that of 10000 artificial codes only two perform better error minimization with respect to amino acid polar requirement compared to the canonical code. The polar requirement (introduced by Woese [151]) was identified to be more significant for error minimization than other mutational effects such as hydropathy, molecular volume and isoelectric point. This measure of distance for amino acids is clearly reasonable, because changing a non-polar for a polar amino acid most probably destroys the well folded protein structure and causes lethal changes. Especially mutations in the second position base could be responsible for altering the polar requirement of the coded amino acid. The first and the third codon position seem to be the result of optimization.

The former analysis was performed without considering the biases in errors that are produced by mutation and mistranslation. This does not consider that in native genetic systems transition errors (i.e. C $\leftrightarrow$ T and A $\leftrightarrow$ T) occur by far more often that and transversion errors ($\{\mathsf{C},\mathsf{U}\} \leftrightarrow \{\mathsf{A},\mathsf{G}\}$) (see e.g. [82]). Mistranslations were empirically studied and were shown to vary in a complex manner. The frequency of misread codon positions is $P_2 < P_1 < P_3$, emphasizing that the second position seems most significant. In a later work this effect has been added to the model. In an enlarged sample of one million random codes the above statistics shifts even toward a more conservative code by a factor of two and that the universal genetic code is "*one in a million*" [54] with respect ot

mistranslation mutational bias. The weighting of translation/transversion mutational effect also led to the insight that the relative efficiency of the second base is more pronounced in this model system, depending on the ratio of the bias.

The *a priori* assumption of a block code might bias the effect of similarity calculations because of the special structure and symmetry. Also there is no justification in nature that the block structure is the only possible code. This problem has been taken up by Goldman [65] who considered more general *shuffled codon* codes, which does not require the block structure of the standard genetic code, but still has the same amount of codons per amino acid. Goldman used a simulated annealing technique to generate the sample of artificial codes.

A further generalization has been proposed by Schönauer [121], who tried to model a heuristics that searches *all* possible codes. Because the space of all possible codes is extremely huge (There are more than 10^{65} possible maps for the generalized codes that assign $64 \rightarrow 21$) the computational effort for this large search space is formidable. This limits a detailed search with available computer resources, but first simulations showed that applying a more sensitive amino acid similarity measure (WAC matrix, amino acid micro-environments in 1Å shells) it shows that the canonical code *extremely* fault tolerant. The optimized artificial codes often showed to have three STOP codon as well and did not show the block structure.

The similarity measure of amino acids is a source of bias since the *ad hoc* assumption of a distinct optimization of a certain physical property of amino acids is arbitrary. The standard genetic code is not special with respect to all amino acid properties and fault tolerance is only showed to be granted for the polar requirement [68]. This weakness has been addressed by Freeland and co-workers [53] by employing *point accepted mutation* (PAM) 74-100 data which derives from frequent *observed* substitution patterns of amino acids in naturally occurring pairs of homologous proteins. Thus this matrix provides a direct measure of similarity. To avoid the problem of just reflecting the effect of neutrality in the genetic code this special PAM derivate was built solely from evolutionary diverged proteins.

To compute the optimization of the native code it is necessary to generate the possible codon space. This was performed using a powerful technique known as

"The Great Deluge Technique" [35]. Large scale simulations revealed that the adaptation within the code can be verified by this refined definition of amino acid similarity. The standard genetic code is found to be close to the global optimum of all codes with regard to error rates.

As shown in figure 1.2 the universal genetic code shows variation in a significant portion of taxa. Investigations of these variant codes revealed slightly lower optimality in terms of error minimization. This can be understood under the assumption that in a primordial organism errors were much more severe that in a extant genome. The occurrence of DNA as information carrier and sophisticated protein machines with elaborated error checking mechanisms made error minimizing codes less important.

Comparing the statistic and engineering approach to quantify code optimization it becomes evident that the statistic reflects reality better. In its linear dependency the engineering endeavor neglects the Gaussian distribution (increasing optimal codes are increasingly rare) and therefore the global optimum is unattainable [83]. This led to an almost emotional debate in literature [31, 56]. Nevertheless it has to be noted that theories [33] derived from doubtful statistics have to be treated carefully.

3.6 Summary

The experimental and theoretical findings all point toward an evolving code, it is unsustainable to assume the standard genetic code is *frozen*.

The theories outlined in this section are at least based on *ad hoc* assumptions should therefore be regarded as informed opinions rather than well tested scientific theories. None of these theories is able to explain all aspects of genetic code origin and evolution and beyond the methodological dispute it becomes clear that despite technological advance, no progress is made by the hypothesis.

The major obstacles are the complexity of the modern translation apparatus that is difficult to explain in terms of a primitive prebiotic environment and the proteins by themselves. Since it is impossible to predict the spatial and functional

properties of a poly peptide solely based on its primary structure, it is impossible to exactly determine each amino acids function in the folding network.

Nevertheless the next section will give advice on how to perform shortcuts and simplifications to implement such a model on a personal computer. Regardless the simplifications the model is built consistently on the basis of known biophysical and evolutionary constrains, observed *in vivo*. Reasonably this model focuses the *extension* of an existing coding system rather than inventing one from scratch and enables to observe modifications of the genetic code along an evolutionary trajectory.

CHAPTER 4

Methods

Despite the powerful molecular biological toolkit available to contemporary biologists and despite the plurality of hypothesis discussed in the previous section, not much light has yet been brought into the origin and evolution of the genetic code. The ribosome and translation apparatus in general seem to be the most complex structures researchers focused ever. It is almost unimaginable that a sequence instructed peptide synthesis apparatus operated without the aid of elaborated enzyme catalysts. In favor of an RNA-world precursor world it is assumed that at this stage all catalysis was performed by ribozymes, the RNA analogs of present day enzymes. Confronted with the technical details of translation such as initiation, frame-selection, concerted movement or chain termination it becomes obvious that an adequate molecular machinery is necessary for the duty. The molecular "stone of rosetta" is buried by this biochemical network of molecular machines in modern cells what aggravates the search for a general explanation of the genetic code. Although by far not all relevant components of translation are known in detail, even the known components are to complex to fully model them on a computer.

Building block for the presented model are virtual organisms that are able to perform a cell-cycle, comparable to a modern cell: Translation, replication, mu-

tation and tournament with competitors happen in the memory of the computer. By cloning a common ancestor a homologous population can be produced that is offered a convenient chemical environment for growth: a stirred flow-reactor. The construction and details of flow-reactors is explained in detail in section 4.6. Within this protected environment of the tank reactor the population is monitored while it underlays the laws of variation and selection.

4.1 Model Organisms

A typical eukaryotic cell contains thousands of genes, coding for an impressive number of proteins. Metabolic pathways and regulatory networks keep the unit alive, while the necessary executives, namely the proteins, are synthesized at the ribosome. It is neither possible to build a living cell from scratch in a lab bench at the time of this work, nor is sufficient computation force available to explicitly simulate a single cell in structural and metabolic detail. However a realistic model of a cell can focus distinct molecular feature and simulate the "rest" as abstract framework.

Therefore our setup reflects a compromise from simplification and physical reality. The virtual organisms that are described in detail in the next section are designed to contain all the components that are necessary to observe the behavior of the coding system. With our model we focus to observe that an existing code can be expanded, not that a new one is found. The biophysical background of the mechanisms involved in replication and translation are documented in the next few sections. The success of an organism is determined by the replication rate, that in terms is dependent on the metabolism. The organisms fitness, measured as efficiency of its replication system, can be compared among a population of mutants and determines if the organism gives rise to a progenitor. The translation does not require a ribosome because the ribosome is not engaged in the code interpretation. The ribosome provides an elaborated framework for the mechanics of translation, that is not required in our model.

The organism that is designed here is prokariotic, therefore has a cell membrane to separate its components from surrounding media. The software equivalent of

the cell membrane is a part of memory within the computer. Everything that is included by the memory block that was assigned a certain organism is said to be a component of it. The cell membrane of the virtual cell is permeable to small molecules like nucleic acids and amino acids. These are the only required building blocks, because no primary metabolism is simulated in detail. This restriction is based on the fact that this kind metabolism does not influence genetic coding and requires immense computational effort.

To efficiently simulate the evolution of the genetic code it is necessary to couple replication, translation and tRNA loading. Replication is required to enable genetic changes and to evolve properties. We reduce replication to a simple "copy" process, that does not act specifically but copying is performed by a specifically acting replicase. Efficiency and accuracy of the replication depend on the properties of this RNA depended RNA polymerase.

The translation apparatus executes the genetic code, and therefore needs to be as realistic as possible, though not all components of the contemporary ribosome are necessary. On a molecular level tRNA aminoacyl synthetases encode the genetic code. In a highly specific reaction they attach an amino acid to a tRNA that in turn offers an anticodon to be used in mRNA reading during translation. The correct loading of the tRNA is the most crucial step in the maintenance of the genetic code, but unfortunately not know on a molecular level in general.

The building plan of a cell is written in the letters of nucleic acids in its genome. It is reasonable to assume that progenitor cells used RNA as information carrier as many evidences point to the existence of an RNA world (see 2.2) in early days of life. The RNA genome of our virtual organism is very simple: only one gene coding for an RNA dependent RNA polymerase (replicase) used to create self-copies and a set of tRNAs to perform translation. The replicase sequence is designed to adopt the fold the native replicase of a phage living nowadays (T7). This design is performed via inverse protein folding, a method that is based on the fact that there are extensive neutral networks within protein space [4, 5]. This protein space is spanned for a predefined alphabet of amino acids, which are the building blocks of our sequence. This amino acid sequence is reverse translated using the reversed genetic code of the contributing tRNAs.

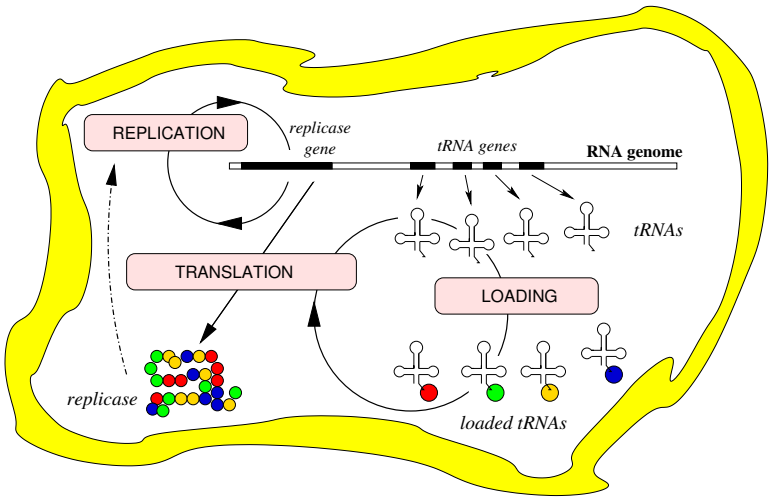


Figure 4.1: The minimal organism model

The tRNA sequences are also designed using inverse folding. Randomly generated sequences are optimized to fold into predefined tRNA structure. The organism obtains only one copy of tRNA coding for a distinct start amino acid. Therefore the starting point of the simulation is an organism that owns exactly those tRNAs that are required to translate the replicase gene, that had been optimized for the accessible amino acids. As the hypothetical ribosome slides along the nucleotide sequence it finds the codons and looks for a tRNA whose anticodon loop can match an exact base pairing with the codon (only canonical base pairs are implied). If no exact matching codon is available, it is attempted to find one, that matches the first two or at least the first position of the anticodon. If non-matching tRNA could be found at all, a random amino acid from the set of available amino acids is used for peptide chain elongation. This is justified by the observation that the modern genetic code has a significant level of neutrality with respect to the third position.

The gene of our model organism was built to have fixed positioned borders at the functional blocks. This relieves the model from an assumption of how a distinct gene is identified, and which gene is translated. One might alternatively postulate that splicing is available in an RNA world already and the transcript is able to spit itself into the functional fragment by itself. For ancient precursor organisms this problem was not to severe since the translation apparatus was

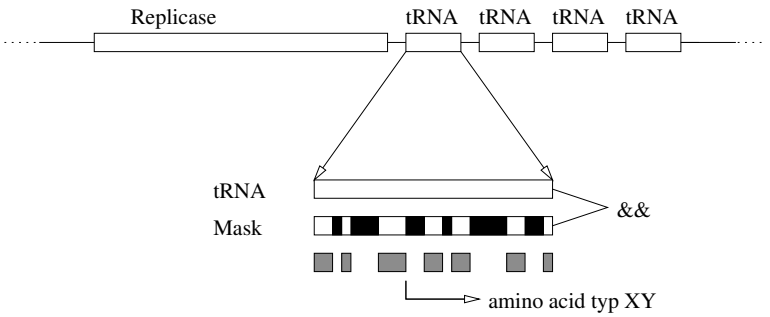


Figure 4.2: The genome of a model organism. Only two types of genes are required: a replicase and tRNA. The tRNAs are mapped to an amino acid via a defined rule rather than an explicitly modeled aminoacyl tRNA synthetase.

simple, and accuracy was low. Building non-functional proteins was not that severe since amino acids were available in high concentrations in the primordial cocktail. Therefore no energetic constraint forces the replication of individuals from strong energetic bookkeeping as is the case for modern cells.

The expression of the genetic information requires the translation of the polynucleotide into a functional protein. This requires a transcription of DNA into messenger RNA by RNA polymerase in a first step. RNA polymerases are large, multi domain proteins that slides randomly along the DNA chain, only initiating transcription if a promoter sequence is identified. There the DNA molecule is unwound and template directed polymerization of RNA starts in the 5'-to-3' direction. This very simplified view of the highly regulated transcription is very rough, but also completely omitted in our model since we assume an RNA genome that can be directly translated.

In a living cell the ribosome takes care of the further processing of the mRNA. This complex machinery consists to a high extend of RNA and is responsible for the coordinated movement along the mRNA chain and the handling of the growing peptide chain. It has been proposed that the main part of the work is done by RNA, supported by tha aid of ribosomal proteins. It could be shown that stepwise peptide extraction retained peptidyl transferase activity of the large ribosomal subunit [81], but protein-free peptide bond formation could not be produced. However, no isolated protein, or mixture of proteins, has ever been shown to catalyze the peptidyl transferase reaction (see [69] for review). Typically

the ribosome (see figure 1.3) consists of two subunits of different size that form a complex of several million Daltons and has 2 binding sites for tRNAs and one for mRNA. The translation machinery moves in the 5'-to-3' direction, using tRNAs charged with amino acids to decipher the genetic information until a STOP codon is reached. The growing peptide chain is released from the P-site (peptidyl-tRNA-binding site) tRNA to the A-site (aminoacyl-tRNA-binding site), where the kinetically verified tRNA molecule offers the activated amino acid and the peptide bond is formed catalyzed by a peptidyl transferase enzyme. In the last step the peptide chain is translocated back to the P-site.

The function of the ribosome is to facilitate codon-anticodon recognition, what improves translation performance, but does not change its mechanism. The mechanism of translation itself and the function of the ribosome do not evolve any more, i.e. translation is established already. Therefore our studies do not require to model the ribosome explicitly. Variation within the translation apparatus arises by mutation of the tRNA codons in codons or identity elements and the tRNA loading.

Based on the following assumptions we do not explicitly simulate the ribosome:

1. The function of the ribosome is performed by RNA. The rRNA catalysis peptide bond formation without the aid of proteins has been shown by experiment [81].

2. Translation is not the time limiting step in the cell cycle.

3. The accuracy of the translation is not limiting the quality of the polymerase.

The replication of the virtual organism is erroneous, showing mutations that lead to genetic variation. Since genes are assumed to have fixed length, no insertions or deletion are assumed to happen. This is necessary since threading amino acid sequences onto a $3D$ structure is computationally extremely costly. The most common kind of mutation is point mutation. In our model all point mutations were treated equally to this end, i.e. transitions and transversion errors happen with the same probability. To relieve parts of the competitive pressure, and add "a little neutrality" we allow duplication of the tRNA genes as mutation event

to happen. The replicase gene is not duplicated, since the optimization of more than one protein sequences would shift the time scale of the whole system, and it is not focus of this study to follow protein evolution.

4.2 The Transfer RNA

The existence of an adapter molecule that would carry an amino acid and interact with messenger RNA was hypothesized by Crick in 1955 [25]. This central role in translation and the universality of the genetic code made tRNAs good candidates for the earliest genes in evolution [42]. The tRNA gene family can be partitioned by the amino acid specificity, several of these groups contain so called *isoacceptors*. Isoacceptors are tRNAs that accept the same amino acid, but have different mRNA codon selectivity. In yeast for example the two $\mathrm{tRNA}^{\mathsf{Phe}}_{\mathsf{GAA}}$ and $\mathrm{tRNA}^{\mathsf{Thr}}_{\mathsf{AGU}}$ are identical except two nucleotides.

Transfer RNAs typically comprise 76 nucleotides that fold into a canonical *cloverleaf*-like structure on the basis of secondary structure. This structure is built from three hairpin region and a variable region, a terminal stack and a single stranded NCAA-end where the amino acid is linked to. The stems are the same length in all species: seven basepairs in the amino acid acceptor stem, five basepairs in TΨ-stems and anticodon stems and three to four basepairs in the D-stems. Within the stem regions non-canonical base pairing is frequently observed, especially GU pairs are common. The anticodon and TΨ-loop are seven nucleotides long in all tRNAs. Two classes of tRNAs can be distinct by the length of the variable region (see figure 4.6). In the early 1970ies several conserved sequence positions were identified that can also be seen from figure 4.4.

The cloverleaf, in turn, is organized into an L-shaped three-dimensional structure composed of two domains: The amino acid acceptor CCA group and the anticodon at both ends of the folding. This structure reveals the importance of conserved and semi-conserved residues of tRNA.

tRNAs contain many modified nucleotides that are produced by a posttranscriptional editing. Some are common to almost all species, such as dihydrouridine

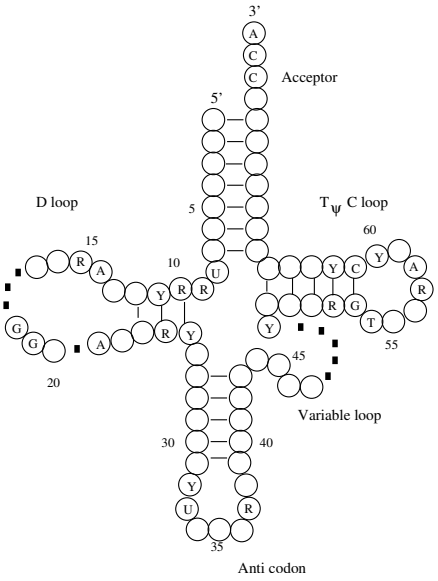


Figure 4.3: The canonical clover leave structure of a tRNA: conserved nucleotides are marked (R=purine, Y=pyrimidine). The black blocks represent extra variable loops. The loops are named after sequences of modified bases that typically occur in that region: The TψC -loop is named after Ribothymin-Pseudouracil-Cytosin, D-loop Dihydrouracil. The anti-codon loop is typically located at the residues 34-36 from the 5$\prime$ terminus.

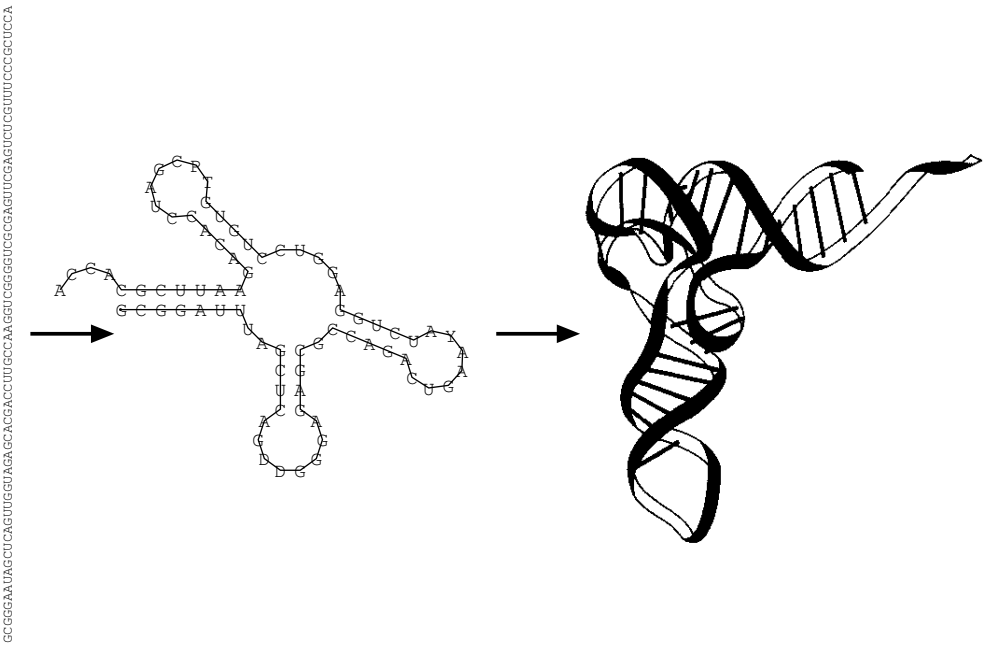


Figure 4.4: Folding of a tRNA sequence via its secondary structure. The canonical cloverleaf folds into the characteristic L-shaped 3D structure.

in D-loops or ribothymidine in T-loops. Other modified bases are characteristic of specific tRNAs. They are located mainly in loop regions. The number and amount increases with evolutionary complexity and can be as high as 20% in higher organisms. Although most of the tRNA transcripts keep their specificity of amino acylation, if deprived from modified bases, some are directly involved in the recognition process.

In our virtual organism tRNAs are modeled explicitly and they are the component that is responsible for the genetic code. *In vivo* their special structure is essential for the function as molecular adapter. Therefore the basic requirement for the identification of a sequence as tRNA in our model is that it adopts the typical structure. Since no efficient *ab-initio* prediction of $3D$-structures is available at the time, the structural requirement has to be reduced to secondary structure. This is reasonable because the canonical cloverleaf structure is the basis of the spatial folding and part of the folding pathway of the tRNA.

Given a sequence x that is assigned a secondary structure s by mfe-folding, and s is represented by the bracket-dot notation, x is said to fold into a cloverleaf like structure if it matches the regular expression presented in table 4.1.

```
(^\({5,9}\.*                        # closing loop
\({3,5}\.+\){3,5}                   # first stem loop
\.*                                 # variable region
\({3,7}\.{2})(\.{3})(\.+\){3,7}     # codon loop
\.{2,7}                             # variable region
\({3,6}\.+\){3,6}                   # third stem loop
\.*                                 #
\){5,9}                             # closing loop
\.+)$                               # trailing base pairs
```

Table 4.1: Regular expression to match tRNAs

This expression set is shown in the `perl` flavor of regular expressions [57] since it is taken from the `GCE` package (see 4.7). The secondary structure shown in figure 4.5 of yeast tRNA^{Phe} would exactly match the above condition. This secondary

structure has been used for inverse folding of tRNA templates that were as start condition in the simulations.

```
(((((((..((((........)))).(((((.......)))))).....(((((.......)))))))))))....
```

Figure 4.5: Secondary structure of $tRNA^{Phe}$ in bracket-dot notation (see section 4.3.1)

4.3 RNA Folding

Nucleic acids as well as proteins form compact, well defined structures in aqueous solution, This structure determines its physical properties and biological function. But in contrast to proteins, where the formation of the hydrophobic core is the main driving force, RNA structures are determined by base pairs, base triplets and other ordered motives. The base pairs tend to form double helices because of the stacking of consecutive doublets. The mapping of a nucleic acid sequence to its secondary structure is simple due to the simple logic of basepairing. For RNA the paired regions will consist almost exclusively of Watson-Crick C·G and A·U pairs as well as G·U wobble pairs. This is the basis for secondary structures and their prediction, but secondary structures are just an intermediate, giving rise to the spatial order of the molecule. RNA $3D$-structures are due to immense effort not computable by energy minimization [135], even nowadays, and crystallographic date is rare [123]. Therefore the focus lies on secondary structures, and for the identification of relevant sequence positions for tRNAs this is sufficient.

4.3.1 RNA Secondary Structures

A secondary structure is defined as a set of base pairs $[i, j]$ on a sequence such that $i < j$ and for any two base pairs $[i, j]$ and $[k, l]$ with $i < k$ it is valid that:

1. $i = k$ if and only if $j = l$
2. $|i - j| \geq 4$

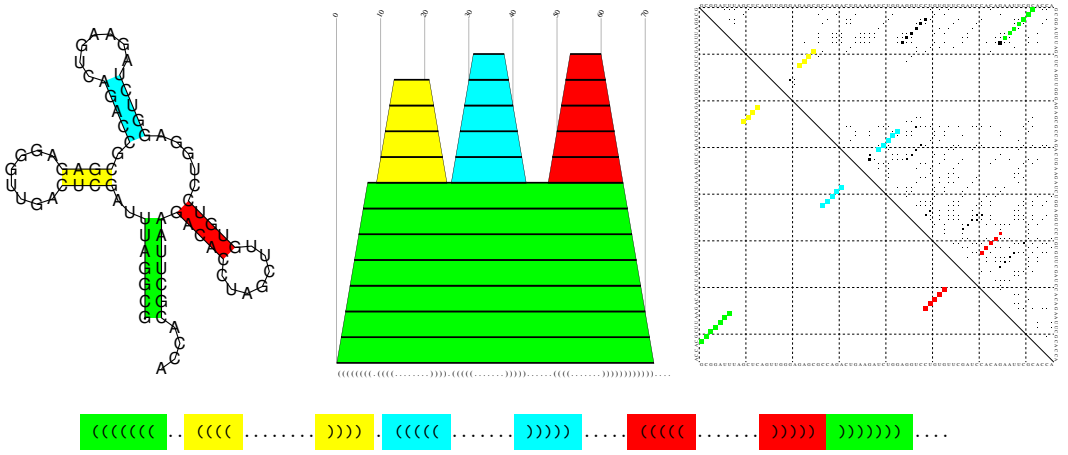


Figure 4.6: The tRNA clover leaf structure as secondary structure graph, mountain plot, dot plot and in bracket notation.

3. $k < l$ implies $i < k < l < j$

The first condition postulates that each base only pairs with one other (i.e. no base triplets are permitted), the second condition ensures that the backbone of the nucleic acid strand is not bent too sharp. The third condition forbids the existence of knots and pseudo-knots. This is important because on the one hand most folding algorithms cannot deal with pseudo-knots and on the other hand most structures violating condition 3 are also sterically very unfavorable.

Most frequently secondary structures are presented as graphs, whereby each node represents a base, a vertex connects consecutive nucleotides as well as basepairs. Figure 4.6 shows the secondary structure for the well known tRNA^{Phe} in alternative representations. For large structures the so called mountain representation is very handy. A secondary structure is plotted in a two dimensional graph, in which the x-coordinate is the position k of a nucleotide in the sequence and the y-coordinate the number $m(k)$ of base pairs that enclose nucleotide k.

For computational uses the string representation of a secondary structure is an efficient storage: A dot "." represents an unpaired position of the sequence, for each pair (i, k), $i < k$ an open bracket "(" is placed at position i, correspondingly bracket ")" at position j closes the pair.

Secondary structures can be uniquely decomposed into loops, stacked base pairs are treated as loops of zero size. The energy of the secondary structure is the sum of the energy contributions of all loops. Due to the additivity of energy contributions, the minimum free energy can be calculated recursively by dynamic programming [144].

The energy parameters were determined experimentally [93], and depend on loop type, loop size and partly on its sequence. For pseudo-knots only the H-type variant was measured [67] and so this is another obstacle for including this pattern in secondary structures of nucleic acids.

Zucker and co-workers were the first to formulate the algorithm for the minimum energy problem [162, 163] using the standard energy model. A modified version of this algorithm also allowed to calculate suboptimal structures within a predefined energy band [161]. The idea to calculate a partition function over all secondary structures $Q = \sum_{\psi} \exp\left(-\Delta G(\psi)/kT\right)$ using dynamic programming was introduced by John McCaskill [94]. An improvement of the secondary structure prediction can be achieved by the reconstruction of folding pathways as aimed by the kinetic folding approach, and elementary step folding trajectories could be computed using kinetic folding [50]. A performance optimized implementation of the mentioned algorithms are part of the `Vienna RNA package` [1)] [73].

4.4 tRNA Aminoacylation

The aminoacylation of the tRNAs is catalyzed by the aminoacyl-tRNA synthetase in a highly specific two step reaction. Each of the 20 amino acids has its distinct synthetase. The aminoacylation is specific to the determinant positions of the tRNA. In many cases an obvious choice for an such a determinant position is the anticodon triplet of the tRNA. Therefore at the biochemical level, the genetic code is established by aminoacyl-tRNA synthetases.

These enzymes are divided into two families on the basis of the architectures

[1)] accessible via the Internet URI `http://rna.tbi.univie.ac.at/`

of their active sites [46]. Each class derives from an ancient distinct single-domain protein. The core feature of this domain is the adenylat synthesis, the condensation of an amino acid with ATP to form aminoacyl-adenylate which is illustrated in figure 4.7. The first step of this reaction is an in-line displacement followed by a nucleophilic attack by the carboxyl group of the amino acid on the α-phosphate of the ATP. After the covalent linkage of the amino acid to the 3'-end of the tRNA, the charged target is available for polypeptide elongation in the ribosome. The major challenge of understanding tRNA charging is how the recognition of amino acid and tRNA is performed. Once loaded, no further check for the accuracy of the genetic code is performed.

This insight was established by an ingenious experiment in which an cystein was chemically converted to alanine after it was covalently bound to its specific tRNA. When such "hybrid" tRNA molecules were used for protein synthesis in a cell-free system, the wrong amino acid was inserted at every point in the protein chain where that tRNA was used. The same mechanism is used by nature itself to enlarge the number of usable amino acids. Selenocystein-inserting tRNAs have been found in many species [91]. These tRNAs are recognized and charged with serine by SerRS and afterwards converted to selenocystein while attached. Selenocystein is incorporated in several proteins where the UGA – stop codon is remapped.

Since all tRNAs have similar structures, the identification must take place on a sequence level in combination with subtle structural variations. Therefore the existing of so called *identity determinants* has been proposed, making a tRNA distinguishable for the synthetase (see [61] for review). It is not surprising that in most known cases (17 out of 20 for *E. coli*) the anticodon bases are part of this set of identity elements. In a minority of cases, notably those of leucine, serine and alanine, the tRNA anticodon is not recognized by the synthetase and other identity elements elsewhere on the tRNA are crucial for tRNA recognition. The recognition elements are located at the same site in all tRNAs, the synthetases discriminate on the basis of a distinct nucleotide at such a site. Beside the anticodon loop the acceptor stem, position 73, the variable loop and the variable pocket. The observation that the active site domains of some synthetases

are able to specifically aminoacylate RNAs comprised of only the acceptor-TΨC stem [18] has led to the proposal that the ancestral tRNA recognition system primarily involved the acceptor stem and position 73. However, at the stage of the establishment of a more complex translation machinery and the availability of the full genetic code the recognition system had to be enlarged, anticodon recognition was added.

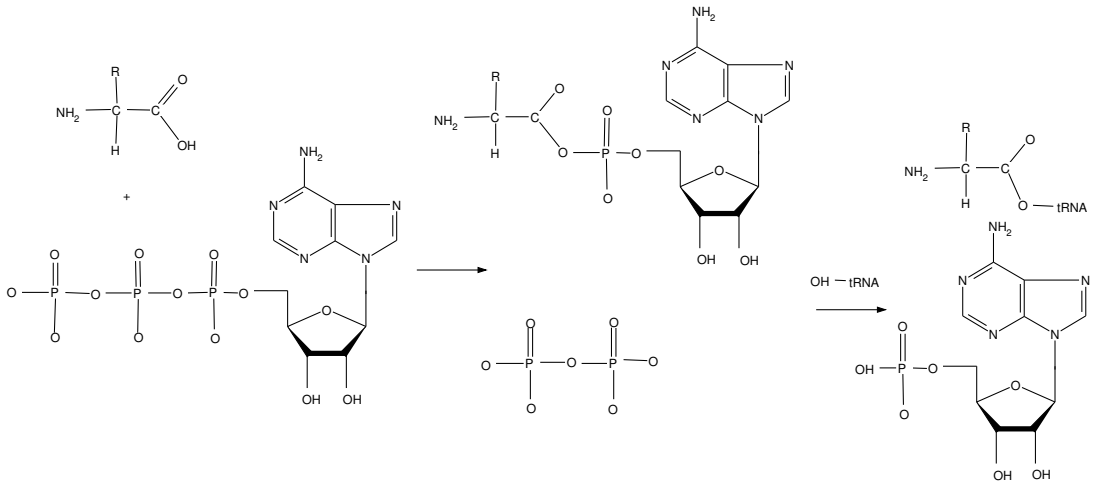


Figure 4.7: The activation reaction of an amino acid to form an adenylate amino acid, which is then

The two families of tRNA synthetases, also known as class I and II, have different active centers: class I synthetases carry a Rossman nucleotide binding fold, composed of alternating β-strands and α-helices [117]. In contrast the active site of class II enzymes are built from a seven-stranded β-sheet with flanking α-helices [118].

The representatives of each class are divided into subgroups, whereby each subgroup identifies chemically related amino acids. For both classes, the subclasses have been denoted 'a', 'b', and 'c'. For example members of class Ia enzymes recognize hydrophobic amino acids (Ile, Leu and Val) and the sulfur containing amino acids Met and Cys. Each subclass is thought to have its own ancestor that arose after the progenitor of the entire class. The members of each class are listed in figure 4.8, which implies a certain symmetry for the subclasses having the same denotation [112]. This is most obvious for the subclasses Ic and IIc

(aromatic amino acids) and Ib and IIb (charged amino acids).

The symmetry is also seen in the side of approaching the tRNA acceptor stem: class I enzymes approach from the minor groove side, class II from the major groove side [118].

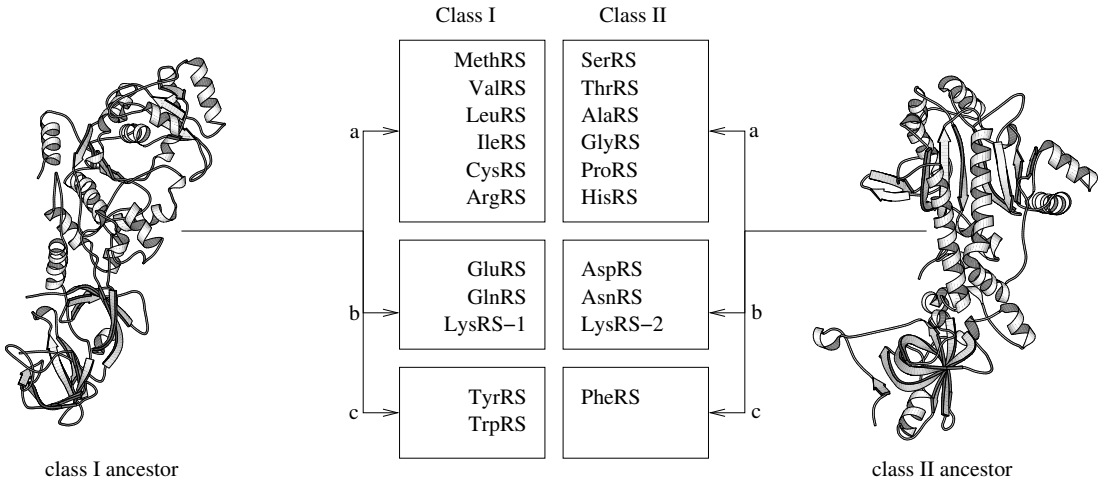


Figure 4.8: The two classes of aminoacyl tRNA synthetases and their subtypes: the representation shows the implied symmetry of the subclasses. On the left side of the figure the class I synthetase is represented by GlnRS (`1qtq`), on the right side AspRS (`1eov`) represents the class II synthetases.

Ribas de Pouplana and coworkers showed that it is possible to find pairs of class I – class II enzymes that can simultaneously bind to the acceptor stem of a tRNA [113] without sterically hindrance. This could be interpreted as evidence that synthetases developed as a protection for the acceptor helix in a hostile (e.g. hot) environment. The amino acid transfer could have been performed by a ribozyme at this early stage of evolution [109].

In our simplified artificial organism tRNA loading is simulated by an heuristic rather than a model enzyme. However to build a realistic model the specific aminoacylation of the tRNAs the `XOR` filter was applied to a combination of structural and sequential parameters derived from the tRNAs and their secondary structure. But this XOR map (called $\oplus$–aminoacyl synthetase) applied to a limited set of positions and nucleic acids narrows the entire space of possible genetic codes of 10^{65} codes to a vanishing small subspace (see table 4.3)that is searched by the genetic heuristics described before. The resulting codes are no

block codes *per se*, but the most plausible scenario why block codes are essential is the optimality in respect to translational errors. However it is possible to construct block codes similar to those found in the standard genetic code, because codons that are not assigned explicit by the mapping are affiliated according their match to defined ones. In other words if for instance GCC was assigned to leucine, and CAA to serine then all codons starting in G are assigned to L, those having C at the first position are translated to S and all others are randomly assigned (to eitherL or S) for each time they are needed. However the simulations performed by Schoenauer [121] revealed that it is not compelling to have block structure for a fault-tolerant coding.

a	b	$a \oplus b$
0	0	0
1	0	1
1	1	0
0	1	1
1	1	1

Table 4.2: Behavior of the function $a \oplus b$.

The complete codes that can be reached by our simulations are easily computable by exhaustive assignment of all permutations of identity nucleotides. It shows that each of the 20 standard amino acids (N,P, Q, A, R, S, C, T, D, E, V, F, W, G, H, Y, I, K, L, M) occurs exactly 32 times in these codes, 384 combinations map to the STOP signal (which means miscoding to our model as described previously). This distribution results from the amino acid assignment mechanism, which only holds one position for each amino acid, if the calculated position is beyond the available positions, it is interpreted as STOP. Table 4.3 shows the code sub-space of the XOR based mapping.

The number and sequences of accessible codons also depends on the tRNA structure since not all mutants of a tRNA fold into the clover leave structure using all permutations of nucleotides in the denominated identity element sequence positions. A primordial organism also must have faced this problem, therefore it might be suspected along with Eigen [43] that a precursor tRNA was simply a stem-loop motive that is easier to maintain.

	N	P	Q	A	R	S	C	T	D	E	V	F	W	G	H	Y	I	K	L	M	STOP
AAA	×			×			×	×	×	×	×		×			×		×	×	×	4
AAC		×	×		×	×						×		×	×		×				8
AAG	×			×			×	×	×	×	×		×			×		×	×	×	4
AAU		×	×		×	×						×		×	×		×				8
ACA	×			×			×	×	×	×	×		×			×		×	×	×	4
ACC		×	×		×	×						×		×	×		×				8
ACG	×			×			×	×	×	×	×		×			×		×	×	×	4
ACU		×	×		×	×						×		×	×		×				8
AGA	×			×			×	×	×	×	×		×			×		×	×	×	4
AGC		×	×		×	×						×		×	×		×				8
AGG	×			×			×	×	×	×	×		×			×		×	×	×	4
AGU		×	×		×	×						×		×	×		×				8
AUA	×			×			×	×	×	×	×		×			×		×	×	×	4
AUC		×	×		×	×						×		×	×		×				8
AUG	×			×			×	×	×	×	×		×			×		×	×	×	4
AUU		×	×		×	×						×		×	×		×				8
CAA		×	×		×	×						×		×	×		×				8
CAC	×			×			×	×	×	×	×		×			×		×	×	×	4
CAG		×	×		×	×						×		×	×		×				8
CAU	×			×			×	×	×	×	×		×			×		×	×	×	4
CCA		×	×		×	×						×		×	×		×				8
CCC	×			×			×	×	×	×	×		×			×		×	×	×	4
CCG		×	×		×	×						×		×	×		×				8
CCU	×			×			×	×	×	×	×		×			×		×	×	×	4
CGA		×	×		×	×						×		×	×		×				8
CGC	×			×			×	×	×	×	×		×			×		×	×	×	4
CGG		×	×		×	×						×		×	×		×				8
CGU	×			×			×	×	×	×	×		×			×		×	×	×	4
CUA		×	×		×	×						×		×	×		×				8
CUC	×			×			×	×	×	×	×		×			×		×	×	×	4
CUG		×	×		×	×						×		×	×		×				8
CUU	×			×			×	×	×	×	×		×			×		×	×	×	4
GAA		×	×		×	×						×		×	×		×				8
GAC	×			×			×	×	×	×	×		×			×		×	×	×	4
GAG		×	×		×	×						×		×	×		×				8
GAU	×			×			×	×	×	×	×		×			×		×	×	×	4
GCA		×	×		×	×						×		×	×		×				8
GCC	×			×			×	×	×	×	×		×			×		×	×	×	4
GCG		×	×		×	×						×		×	×		×				8
GCU	×			×			×	×	×	×	×		×			×		×	×	×	4
GGA		×	×		×	×						×		×	×		×				8
GGC	×			×			×	×	×	×	×		×			×		×	×	×	4
GGG		×	×		×	×						×		×	×		×				8
GGU	×			×			×	×	×	×	×		×			×		×	×	×	4
GUA		×	×		×	×						×		×	×		×				8
GUC	×			×			×	×	×	×	×		×			×		×	×	×	4
GUG		×	×		×	×						×		×	×		×				8
GUU	×			×			×	×	×	×	×		×			×		×	×	×	4
UAA	×			×			×	×	×	×	×		×			×		×	×	×	4
UAC		×	×		×	×						×		×	×		×				8
UAG	×			×			×	×	×	×	×		×			×		×	×	×	4
UAU		×	×		×	×						×		×	×		×				8
UCA	×			×			×	×	×	×	×		×			×		×	×	×	4
UCC		×	×		×	×						×		×	×		×				8
UCG	×			×			×	×	×	×	×		×			×		×	×	×	4
UCU		×	×		×	×						×		×	×		×				8
UGA	×			×			×	×	×	×	×		×			×		×	×	×	4
UGC		×	×		×	×						×		×	×		×				8
UGG	×			×			×	×	×	×	×		×			×		×	×	×	4
UGU		×	×		×	×						×		×	×		×				8
UUA	×			×			×	×	×	×	×		×			×		×	×	×	4
UUC		×	×		×	×						×		×	×		×				8
UUG	×			×			×	×	×	×	×		×			×		×	×	×	4
UUU		×	×		×	×						×		×	×		×				8

Table 4.3: The codon sub space that is accessible by the `XOR` based mapping

Since the principles that govern tRNA identity are not fully understood and a fitting of the $3d$ structure of a tRNA synthetase would significantly increase the complexity of the model, the synthetases were modeled in an abstract set of rules rather that explicit. The tRNA sequence is folded into its MFE structure, and evaluated to fit onto the tRNA schema. It is reasonable to assume that a molecular machinery that is as complex as the ribosome did not evolve in a single step, but the first step must have been the logic of the loading. The bases at a predefined set of positions are translated into a binary string that needs to be mapped onto the set of possible amino acids. To model the non-linear effects of loading the string is folded onto itself applying the XOR (noted as "$\oplus$") operator on the halves of the coded string (table 4.2 shows the behavior of this operator). The result is interpreted to be a binary number that specifies the position of the loaded amino acid within a predefined table.

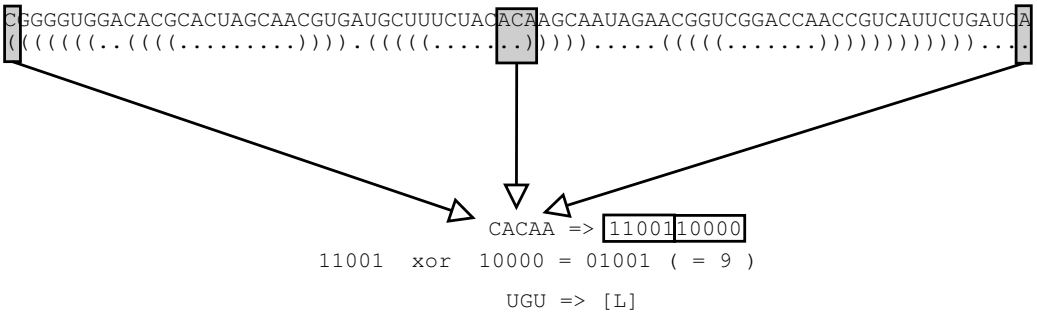


Figure 4.9: Modeling the tRNA aminoacyl synthetase with an "$\oplus$–aminoacyl synthetase"

As easily can be seen, this procedure results in an ambiguous mapping: It is obvious that the coding nucleotides participate the loading, but since the other positions may vary free, a single codon may be mapped to several different amino acids. A translation apparatus using ambiguous codons does not have the opportunity to fixate any codon that are advantageous. This problem has been solved by a simple strategy: for a single codon each tRNA in the model may contribute the chance to code fo an amino acid, but during the translation process only one amino acid id used. A shift of specificity is enabled by introducing an artificial mutation type: codon shuffle. This means within the set of possible mappings for one codon the one that is selected may change. This schema corresponds to an adaptive codon change, where the loading of a tRNA becomes ambiguous and

one phenotype is selected by mutation or deletion.

Since the whole genome is replicated with a certain mutation rate, tRNAs may also be affected by mutations. On the one hand this can make the tRNA unusable (e.g. not folding to the required clover-leaf structure any more), on the other hand the mutation may be neutral with respect to structure. The drift on the neutral network of tRNA sequences may enable the structure to change for instance an identity position. This in combination with gene duplication can enlarge the accessible amino acids for the translation. A tRNA that keeps its amino acid specificity and changes a base in the anticodon loop could shift its family membership to another isoaccepting tRNA. A scenario like this has been investigated by *in vitro* mutation experiments and phylogenetic analysis of *E.coli* tRNAs by Saks et.al. [119]. The author reports that $\text{tRNA}^{\mathsf{Arg}}$ showed coding for Thr after mutating A→U at sequence position 20 and changing the codon from UCU to UGU in *in vitro* amino acylation experiments.

4.5 The Evaluation of Protein Structures

The $3D$ structure ψ of a protein is determined by its sequence x. Although this relation is straight forward, the folding problem is not yet solved for this class of biomolecules. The ,,inverse" folding problem is astonishingly much easier. Inverse folding is, however, *not* just minimization of the energy function in sequence space for a given conformation. This would be the case only if the energy function were normalized such that the native state (ground state) of *each* sequence is equal to 0. This, of course, amounts to solving the protein folding problem for each possible sequence first.

4.5.1 Knowledge Based Potentials

So called knowledge-based or empirical potential functions try in contrast to the molecular mechanic approach not to model the Hamiltonian, but "extract" energetic parameters from given protein structures. Prominent representative of

this energy functions are Sippl's PROSA Potential [128], Lapedes' Neural Network NN Potential [66] or Alexander Tropsha's Four-Point Potential [100, 127, 160].

Given a knowledge-based potential function $W(x, \psi)$ it is possible to evaluate the energy of a sequence x when folded into a structure ψ. The structure (or fold) ψ is defined by the spatial coordinates of a subset of its atoms, usually the C^{α}- and/or C^{β}-atoms. A whole series of studies [15, 21, 66, 71, 129–131] using different empirical potentials $W(x, \psi)$ showed that the z-score

$$z(x, \psi) = \frac{W(x, \psi) - \overline{W}(x)}{\sigma_{W(x)}} \tag{4.1}$$

can be used to decide whether a sequence can adopt a given protein structure ψor not. The symbols $\overline{W}(x)$ and $\sigma_{W(x)}$ denote the mean value and the standard deviation of the distributions of energy values, which can be calculated from ψ for a constant x is compared to a database of native like structures. The relevance of the z-scores recently has been shown by thermodynamic measurements [159]. A sequence x can be assigned to fold into a distinct structure ψ if the z-scores exceeds a sequence length dependent limit. Empirically, native folds have z-scores in a narrow characteristic range [130].

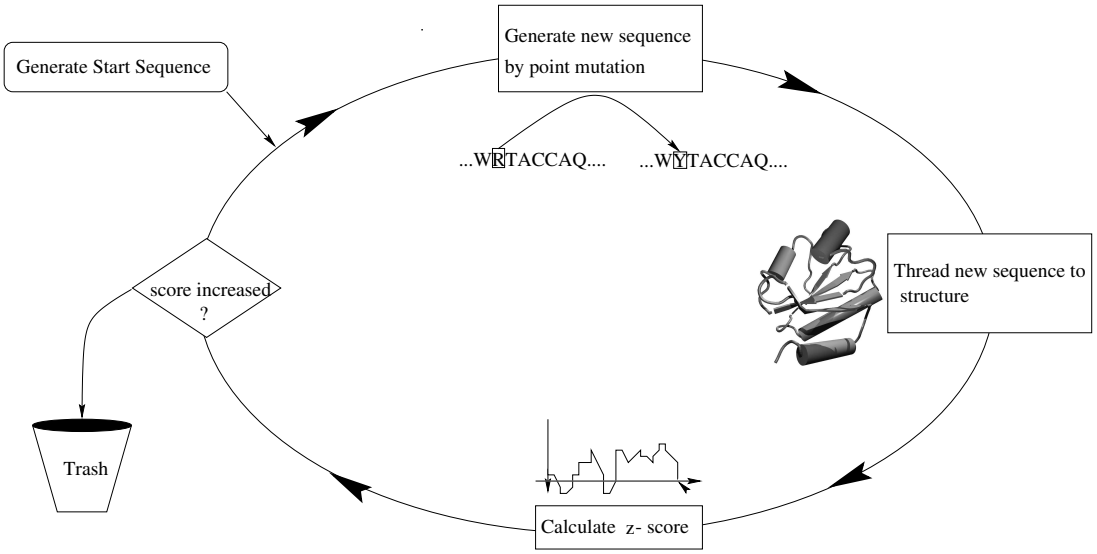


Figure 4.10: Schema showing the inverse folding procedure.

Using inverse folding techniques, large scale statistical surveys of the sequence-structure map of poly-peptides can be conducted without actually solving the

folding problem [4, 5]. These investigations revealed the existence of large neutral networks spanning the protein sequence space. Base on these findings it is possible to find optimal sequences for a structure Ψ even though the fitness-landscape is extremely rough, and local optima are spread.

4.5.2 Delauney Tessellation of Protein Structures

Protein potentials based on Delauney tessellations were originally proposed by A. Tropsha *et al.* [100, 127, 160]. In his work Tropsha derives a parameter free definition of a contact potential using the Delauney decomposition. The protein structure is reduced to a set of points in $3d$ space, to further simplify the systems only the C^{α} and/or C^{β} atoms of an amino acid is selected. The coordinate set is tessellated using the *Delauney triangulation*. The result of this geometric procedure is a partitioning of the space included by the set into irregular tetrahedra with the points as vertices. The quadruple of amino acids represented by these points are considered to be nearest neighbors.

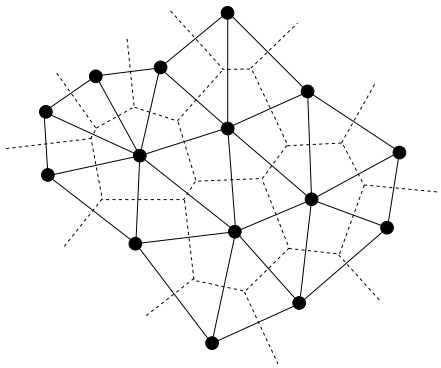


Figure 4.11: Tessellation of a set of points in $2d$: The straight line are the Delauney simplices defining nearest neighbor contacts, the dashed lines show the corresponding Voronoi Diagram of each point.

An efficient tessellation algorithm has been implemented by Barber and coworkers [16], their `qhull` program and library is freely available via the Internet[2)]. A convex hull of a set in $\mathbb{R}^{d+1}$ by lifting the points to a paraboloid and adding the sum of the squares of the coordinates and computing their convex hull. The

[2)] URI: `http://www.geom.umn.edu/software/download/qhull.html`

set of ridges of the lower convex hull is the Delauney triangulation of the original set. The `qhull` algorithm is a derivate of the randomized incremental algorithm that constructs an additional point at the hull that aids to decide which facet belongs to it.

To obtain the parameters for energy evaluation, a calibration step is necessary. By counting the occurrence of all quadruple combinations in a database of native, non-redundant protein structure the energy parameters can be defined by

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}} \,. \tag{4.2}$$

Where f_{ijkl} is the observed frequency of occurrence of amino acids i, j, k, l, which is compared to the *a priori* expected frequency of such a tetrahedron, which is proportional to the product of the single amino acid frequencies.The contact energy $W_C(x, \psi)$ of sequence x when threaded onto structure ψ is then the sum of the energy contributions of all tetrahedra into which the structure is decomposed. This results from the application of the *inverse Boltzmann law* as introduced by M. Sippl [128].

$$W(x, \psi) = \sum_{contacts} q_{ijkl} \tag{4.3}$$

An efficient implementation of an empirical protein potential using Delauney tessellation has been created by P.F. Stadler, I.L. Hofacker and the author [146].

In a recent study Carter *et.al.* [20] were able to show that four-body contact potentials as derived by the Delauney Triangulation for different proteins scale to experimental $\Delta(\Delta G_{unfold})$ values, and could successfully be used, to identify stability changes in mutant proteins.

Unfortunately, the pure mathematical description of protein structure in the presented manner suffers from some shortcomings that are discussed in detail in sections 4.5.3, 4.5.4 and 4.5.5.

4.5.3 Superposition of the Surface

The Delauney tessellation by itself does not distinguish between surface residues and buried amino acids. This neglects the fact that surface exposed residues strongly interact with the surrounding media, what derives the driving force for folding. Following Bowie and Eisenberg, solvent exposure is treatable as superposed term. The parameters for this term are hardly known from experiment, but it is straight forward to identify those triangles (faces of the tetrahedra) that are exposed to the outside of protein from the tessellation. A buried triangle will appear in two tetrahedra, while a surface triangle appears only once.

A surface term based on the log likelihood ratios q^s_{ijk} of triples i, j, k of amino acids in surface triangles can be computed analogous to 4.3 and contributes to the energy simply by:

$$W^{comb} = W^{cont} + \gamma W^{surf}$$

where the combined energy W^{comb} is built from the contact energy W^{cont} and the surface energy W^{surf}, heightened by a factor γ.

4.5.4 Sparse Data Correction

Due to the vast amount of parameters some quadruple combinations are under represented within the limited data-set of non-redundant experimental protein structures. Using Bayesian reasoning it is possible to circumvent this problem using the following strategy [72]:

Given the database of structures provides N contacts, whereby a contacts share a distinct property (eg. Ala-Ala-Ala-Ala contact), a good estimation of the probability of occurrence λ is: $\lambda \approx a/N$. In case of limited database size and a prior expectation of what λ will be, the measured frequency $f = a/N$ will be a good approximation for λ if $N \gg \frac{1}{p}$.

If the exact value of λ was known, the probability $P(D|\lambda)$ could be stated via Bayes' theorem [6]:

$$P(D|\lambda)P(\lambda) = P(\lambda|D)P(D) \tag{4.4}$$

and therefore:

$$P(\lambda|D) = \frac{P(D|\lambda)P(\lambda)}{P(D)}. \tag{4.5}$$

Since $P(D)$ is independent of λ, it can be treated as normalization constant, $P(\lambda)$ is the so called *prior*. Since we are not interested in the entire probability distribution $P(\lambda|D)$, a maximum likelihood estimate is performed. For the prior the following assumptions are made:

1. The maximum of $P(\lambda)$ should be at p, so that we'll estimate $\lambda = p$ for $N = 0$ (i.e. no data).

2. λ should never be 0 or 1, so that the potentials stay finite. Thus we have $P(0) = P(1) = 0$.

A reasonable approximation is a linear function:

$$P(\lambda) \sim \begin{cases} \frac{\lambda}{p}, & \lambda \leq p \\ \frac{1-\lambda}{1-p}, & \lambda > p \end{cases}$$

If we assume independence of our N measurements, $P(D|\lambda)$ is a simple binomial distribution $P(D|\lambda) \sim \lambda^a(1-\lambda)^b$, with $b = N - a$. The ML estimate for λ is value that maximizes $P(\lambda)\lambda^a(1-\lambda)^b$. To find it, we have to distinguish three cases:

The maximum could be at some $\lambda \leq p$, in which case it has to fulfill

$$\frac{d}{d\lambda}\lambda \cdot \lambda^a(1-\lambda)^b = 0$$

$$(a+1)\lambda^a(1-\lambda)^b - b\lambda^{a+1}(1-\lambda)^b = 0$$

$$\lambda = \frac{a+1}{N+1}.$$

If $\lambda \geq p$, we have

$$\frac{d}{d\lambda}(1-\lambda)\lambda^a(1-\lambda)^b = 0,$$

which eventually yields

$$\lambda = \frac{a}{M+1}.$$

Else the maximum might by at $\lambda = p$.

Combining the three cases we have

$$\lambda = \begin{cases} \frac{a+1}{N+1}, & \frac{a+1}{N+1} < p \\ \frac{a}{N+1}, & \frac{a}{N+1} > p \\ p, & \text{else.} \end{cases}$$

4.5.5 Filtering of the Tetrahedra

The `qhull` algorithm generates *per definitionem* the convex hull. This creates a totally smooth surface for protein structure coordinates, thereby loosing important information about pockets and holes. In order to recover the original structure one has to apply a filter rule on the set: Tetrahedra with edges longer than a threshold λ as well as tetrahedra with a circumsphere radius greater that a limit ρ (the smaller the radius, the tighter the packing).

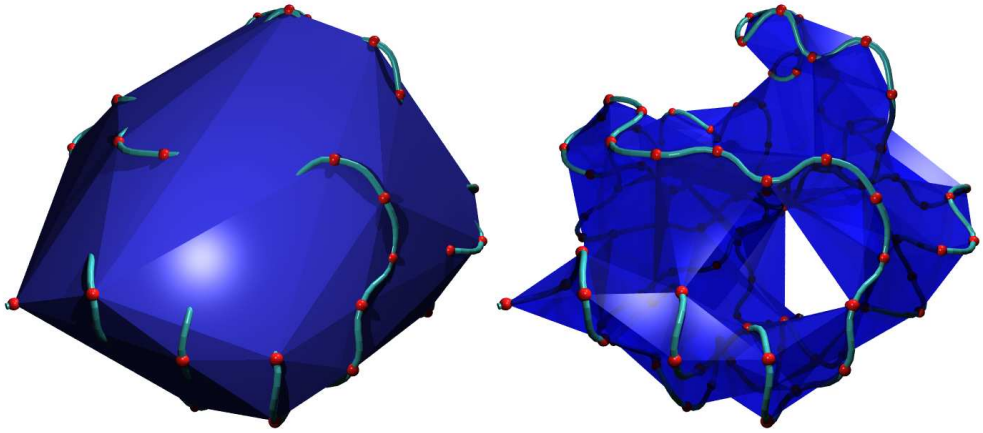


Figure 4.12: The cartoon compares the result of the Delauney tessellation of the crystal structure of bacterial Thioredoxin (pdb-id: `2trx`) with filtering applied (right) and without(left). The blue surfaces are the resulting surface triangles as resulted from pure tessellation. The Backbone is outlined by the green tube, C^{α} atoms are shown as red balls.

4.5.6 RNA Polymerase

If the tasks of an organism are reduced to replication that means no metabolic activity takes place, the only molecular machine that is required must be able

to copy the genome. This is reasonable because most RNA viruses (e.g. Polyo virus) make a living with only one single protein.

In our organism an RNA-dependent RNA polymerase (replicase) is necessary. As model protein bacteriophage T7 RNA polymerase has been chosen because as a viral replicase this is the only extant class of polymerases that recapitulate the replication requirements of the RNA world. So replicases could have played a key role in the switch of an RNA to a DNA genome. A portion of the molecule shows extensive structural homology to the polymerase domain of Klenow fragment [28] indicating that these enzymes share a progenitor polymerase. Biebricher and coworkers could show that T7 RNA polymerase is able to perform RNA polymerization without a template [12]. In these experiments it could also be shown that the enzyme has distinctive specificity for RNA, if offered DNA templates as well, but initiation as well as polymerization can be performed as well using desoxy ribonucleotides. Another argument for choosing this distinct replicase is the availability of a high-resolution crystal structure [132] which is absolutely necessary for the fitness evaluation of the protein.

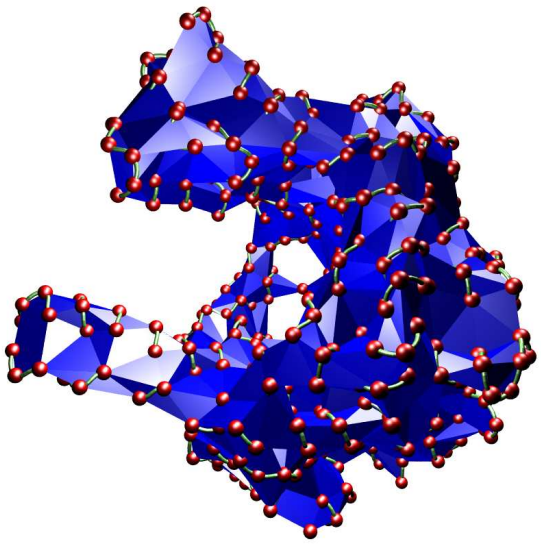


Figure 4.13: Chain A of T7 RNA polymerase (Protein Database access id:`4rnp`), The structure was derived from x-ray diffraction at an resolution of 3.3 Å only C^{α} atoms are given. The blue surface is calculated via tessellation, the red balls mimic the C^{α} atoms.

The properties of the replicase determine the organisms fitness in two ways: First

if the structure of the enzyme is optimal with respect to its task, the reaction rate can be accelerated that is more copies in less time. Second the accuracy of the replication solely depends on the enzymes ability to read the template. Certain position at the active site are responsible for the identification of the template base, and direct the recruitment of a nucleotide for elongation. In our model we relate the overall fitness of the replicase with the z-score of the sequence, if threaded onto the native structure of the T7 RNA polymerase. This is justified by the fact that the z-score gives a good measure of the compatibility of a sequence with a structure in overall. To obtain a sequence dependent measure for the accuracy of the replicase the energy of distinct positions was compared with the energy of the tessellated wild type T7 sequence. The sequence position taken into account were chosen by studying the result of the mechanism found in the RNA polymerase-promoter complex. To achieve a reasonable initial mutation rate the mutation rate μ has been scaled to

$$\mu = \sum_i \left\{ \frac{w_{i(wt)}}{\sum_{j(wt)}} - \frac{w_{i(mut)}}{\sum_{k(mut)}} \right\} \tag{4.6}$$

Where $w_{i(wt)}$ stands for the energy contribution of the wild type amino acid at position i and $w_{i(mut)}$ denotes its mutated counterpart. The energies are normalized using the total putative energy of the tessellated structure. Table 4.4 lists the contributing sequence positions and their native function.

The coupling of translation and replication in our model is achieved via fitness and mutation rate. The fitness of a model organism is computed as the the z-scores of the replicase gene[3)]. Attempts to include the number of tRNAs in the fitness calculation revealed, that this additional pressure delayed innovations (data not shown), but does not alter the results qualitatively. The mutation rate is calculated from the energy map of the tessellated replicase structure using equation 4.6.

[3)] Actually the sum of z-scores of all replicases, but we restricted the current model to the use of one single copy of replicase.

wt amino acid	sequence pos.	explanation
res	93 ..101	flexible loop
leu	136	Melts base pair (G2-C71)
val	237	Formation of a β-hairpin to melt double strand
trp	422	Introduces a sharp bend in the template strand by stacking of the aromatic side chain with position+1
asp	537	Binding of incoming nucleotide
gly	542	discriminates ribo- from desoxyribonucleotides
arg	746	Hydrogen bonds with A-8 and G-7
asn	748	Major groove interaction with bases of non-template strand (DNA-template)
arg	756	interaction with 6-keto and 7-imino groups of G-9
gln	758	Hydrogen bonds with A-8 and G-7
his	784	Discriminates ribonucleotides from desoxyribonucleotides
his	811	Binding of incoming nucleotide
asp	812	Binding of incoming nucleotide
lys	613	Catalytically involved (mutant study)

Table 4.4: Catalytically significant amino acids involved in the polymerase reaction. Investigated in studies of co-crystallized T7 RNA polymerase with blunt-ended promoter DNA[23, 108]. These positions were used to calculate the mutation rate via equation 4.6.

4.6 Flow Reactors

In order to observe a microbial population over a long period of time in vitro, the environment must be kept with a constant supply of energy and a kind of garbage collection must take care of metabolic end products. One approach to provide such an experimental framework is the so called serial transfer experiment where a subsample of the culture is transfered to a fresh solution. This approach has been successfully applied in the historical experiments of Sol Spiegelman in the 1960s [133]. In this pioneer work it was possible to select RNA molecules that are capable of fast replication in a cell free environment.

Another possibility to breed an asexual reproducing population is in the so called

flow-reactor. This system is built as chemo-stat, this means that the required nutritions and medium is supplied in a constant flow. A dilutent flow of solvent removes metabolic toxins as well as organisms (illustrated in fig. 4.14). This kind of system is easily implementable *in silico*, and follows the principle of Darwinian variation and selection.

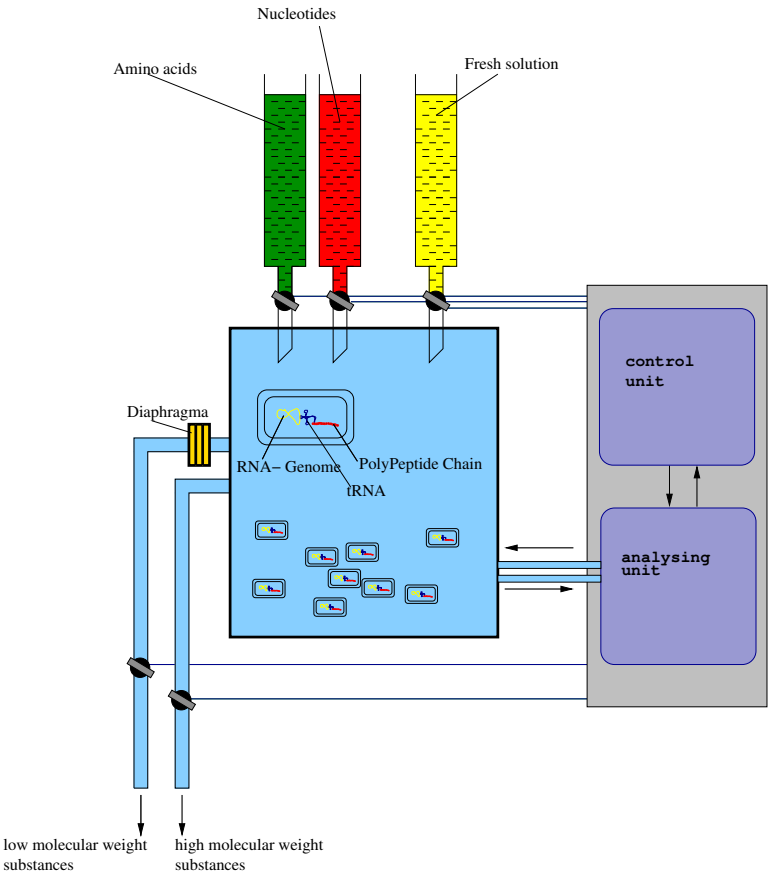
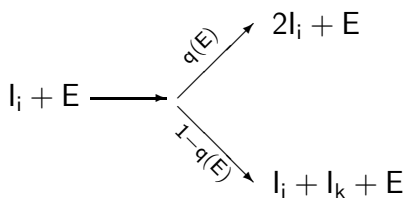


Figure 4.14: The flow reactor is a chemo stat and provides a population with fresh solution as well as energy. The population size is diluted via a constant flow of fresh solution, thereby acting as selective force.

A mathematical rigorous description of flow-reactors is based on the theory of stochastic chemical reaction networks. Possible reactions are: exact replication, erroneous replication (i.e. mutation) and dilution [63]. The species of replicators

$$I_1, \ldots, I_n$$

are competing for resources an the auto-catalytic replication (either correct or erroneous) can be written using the reaction scheme:



The quasispecies equation for this kind of reaction may be written as

$$Q_{ki} = \left(\frac{P_i}{\alpha - 1}\right)^{d(k,i)} (1 - P_i)^{n-d(k,i)}$$

where Q_{ki} is the probability of mutating species i to k, under the assumption of solely point mutations. P_i denotes the mutation rate per digit and replication, α is the size of the alphabet (four in our case). The kinetic of the reaction is therefore given by:

$$\dot{x}_k = \sum_i \{Q_{ki} A_i x_i - Q_{ik} A_k x_k\}$$

For RNA molecules, where the genotype–phenotype mapping is known in detail, there is a well developed theory, namely the theory of the molecular Quasispecies [36]. Manfred Eigen was the first to apply chemical reaction networks to molecular evolution [39–41].

Organisms competing for resources are forced to optimization. This result of the Darwinian theory has led to the design of genetic algorithms and genetic programming: an application of evolutionary models to optimization problems. Holland performed fundamental research on genetic algorithms, which were published in his outstanding book [74]. The *item* targeted by optimization is coded in the gene of an organism, like the building plan (DNA) of biological organisms. This genotype is mapped to a phenotype with a distinct fitness. The fitness of the phenotype determines the chance for the organism to reproduce (replication rate). The survival of the fittest leads to the establishment of a predominant *master species* [38] that is surrounded by a "tail" of mutants. If a mutant becomes fitter than the master, the population drifts toward this species.

Our simulations were started using organisms that had a comparable high fitness and were all clones from one single cell. Each organism inherited beside its genes a unique token (cookie) from its parents, that was equal to the parents token if they were exact copies from the replicant. Otherwise the token was renewed.

The population of test organisms (see section 4.1) were allowed to reproduce in a so called *tournament replication* process. In a replication step two individuals are picked randomly, their fitness is evaluated and compared and the fitter one is permitted to replicate. To limit population size of the reactor a dilution flow is applied. The child organism replaces another randomly picked individual, therefore a finite chance exists for an organism to replace a fitter one and the population size remains constant.

The protocol of a simulation contained a header, consisting of the start parameters and the time. During the run reports were dumped periodically protocoling the average fitness and mutation rate of the population as well as the totally available amino acids and their concentration. A dump of the codon table, protein sequence and magic cookie make the changes traceable for single species.

4.7 Software Implementation

The proposed model has been implemented in an object oriented programming framework in the computer language `perl` [134, 143]. This dynamic typed language permits rapid development, in combination with object oriented techniques the software projects scale well. As `perl` is available in source code [4)] and compiles for almost any hardware platform the problem of portability hardly exists. Also, `perl` has excellent text processing tools in its standard toolkit. This is especially important for the manipulation of RNA secondary structures, where `perl`'s powerful regex engine comes into play. Computationally demanding procedures such as protein threading and RNA folding are performed in optimized `C` libraries that are interfaced using the `SWIG` (Simplified Wrapper and Interface Generator) package[5)] [7]. All required modules have been grouped to a package within the `GCE` (**G**enetic **C**ode **E**volution) name space. A UML diagram of the package is given in figure 4.15

The base class is the simple root of the `GCE::` name-space implementing common interfaces for all classes. It is an abstract class, therefore not to be instantiated

[4)] down-loadable at URI `http://www.cpan.org/src/latest.tar.gz`
[5)] freely available via the Internet URI `http://www.swig.org/`

(it also lacks a constructor). The `GCE::Config` class is a generic configuration module that corresponds to the singleton pattern [58], the constructor can create an instance from an an `XML`-file to create a unified access to configuration variables. `XML` (Extensible Markup Language) becomes increasingly important in bioinformatics.

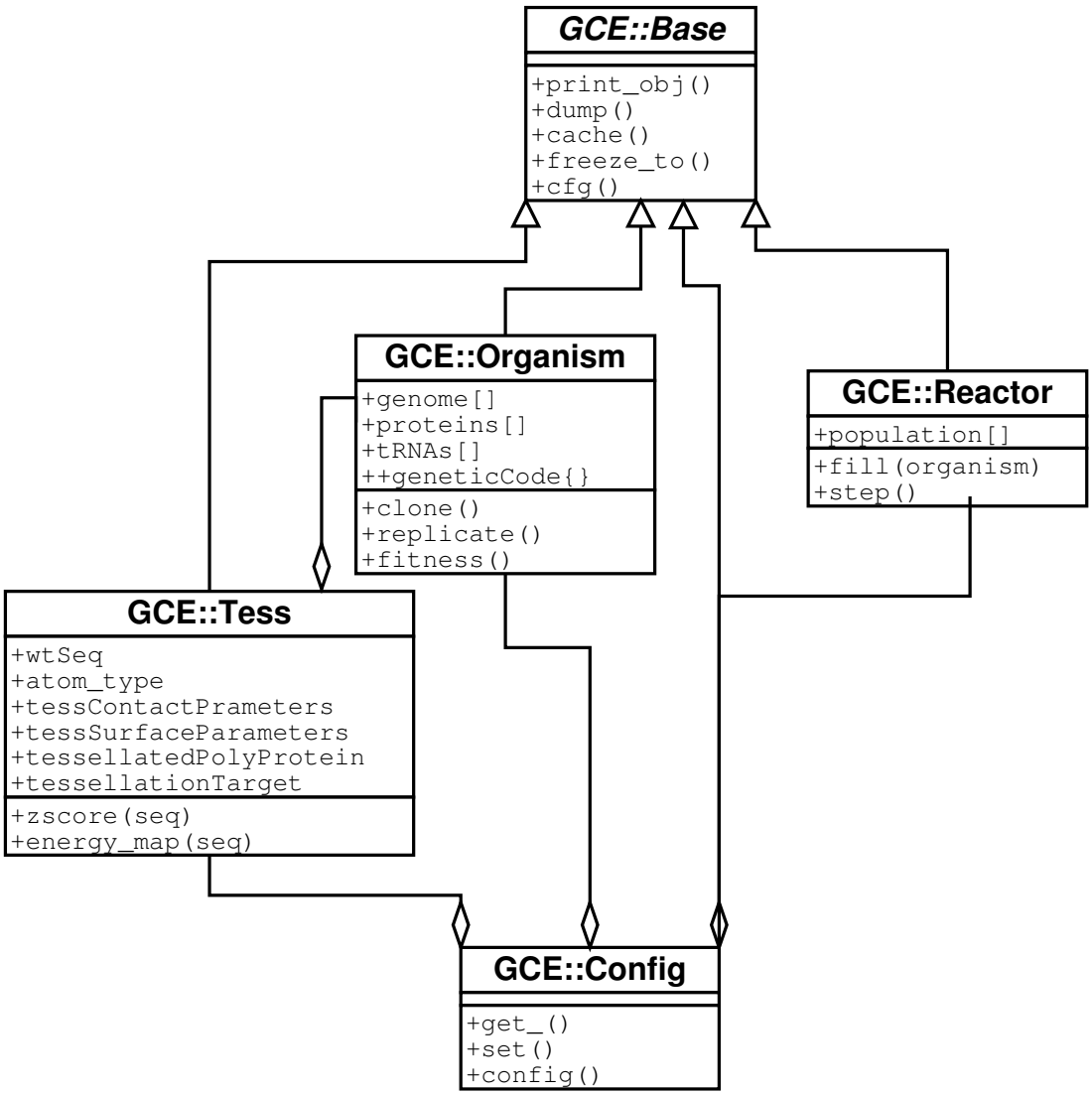


Figure 4.15: A UML diagram of the GCE:: package.

The fundamental module is a reactor class that instantiates container objects from a static configuration class. The filling (initialization) of the reactor requires objects that can be of any desired class, but have to provide at least the following

methods:

cloning:
The item is able to produce exact copies of itself.

replication:
The production of erroneous copies leads to variation within the population.

fitness:
The initial fitness of the template individual must not be lower than a defined threshold.

The most important action to be performed on the reactor is the `step()` routine, where "dead" (fitness $<$ threshold) organisms are eliminated and tournament replication takes place.

The organism class used for the studies of the genetic code held a genome, an evolvable hash table of the genetic code and translation products. The computationally most expensive step is the initialization of the tessellation of the polyprotein for the calculation of the z-scores of the proteins. But since the entire population is generated via cloning a common ancestor, this step has to be performed only once per run. Efficient folding of the tRNA is performed using the `perl` interface of the `Vienna RNA Package` [73] that is a very efficient implementation of Zuckers' dynamic programming algorithm [163].

Reporting functions and special flow control can easily be implemented and modified via driver scripts, where the required objects are instantiated. The memory consumption mainly is caused by the protein evaluation via the tessellation potential. The simulations presented in section 5 were computed on Intel-based

Mutation Rate	0.01
Identity positions:	1, 77, anticodon loop (3 bases)
Population size:	1000
Poly protein for z-score calculation	`poly10k.pdb`
replicase pdb file	`4rnp`
minimum required fitness	0

Table 4.5: Default parameters used for the calculations used in this section.

computers running Linux as operating system. The default parameter setting can be seen from table 4.5, this set was used unless explicitly stated otherwise.

CHAPTER 5

Results

5.1 Overview

In this section we discuss some simulations that were performed using the software described in section 4.7. We focus on the evolution of the amino acid coding table. To this end we track the average fitness of the population and the coding schemes of the individuals inhabiting the tank reactor as a function of time. It does not come as a surprise that modifications of the code are rarely fixated in the population although code variants frequently arise, usually as deleterious mutants.

The procedure of expanding an organisms code table starts by a point mutation of a tRNA. After folding the tRNA sequence and verification of the secondary structure pattern (see table 4.1) our $\oplus$–aminoacyl synthetase acts on its substrate. If the designated identity elements code for an amino acid, the anticodon of the tRNA is said to code for this amino acid. The fixation of such an alphabet extension depends on the fitness of the replicase translated with the new code. In most cases the mutant is deleterious, this means the organisms replicase is less efficient than that of competitors. A new amino acid only results in a more

efficient replicase, if the sequence positions of the translated gene contribute higher energy to the threaded protein structure in the tessellation potential. The fitter competitors start to spread over the reactor, but there is still a chance that mutation vanishes an tRNA before more offspring was generated and the parent falls victim to dilution in the tank. Therefore lots of "birth" and "death" process of new amino acids are observed before a code extension is fixated among the whole population.

The loss of an amino acid – codon pair mostly happens because the tRNA that was coding for that amino acid gets unusable by point mutation. In figure 5.1 examples of tRNA mutations taken from a simulation run are shown, at first successive reassignment ($\mathsf{I} \rightarrow \mathsf{E}$) and the subsequent mutant cannot be loaded with an amino acid by the $\oplus$–aminoacyl synthetase.

```
          ************************************ ****************************************
tRNA_Ile ACUGUUAUCGAGGUCAUCUACACCUUAAUGGGAAUAAUGCUCAUCCGUCAUGCGCCAGUUAUGCAUUAACAGUACGC    77
tRNA_Glu ACUGUUAUCGAGGUCAUCUACACCUUAAUGGGAAUAAAGCUCAUCCGUCAUGCGCCAGUUAUGCAUUAACAGUACGC    77
tRNA_--- ACUGUUAUCGAGGUCAUCUACACCUUAAUGGGAAUAACGCUCAUCCGUCAUGCGCCAGUUAUGCAUUAACAGUACGC    77
```

Figure 5.1: Multiple sequence alignment of wild-type and mutated tRNAs. The wild type is coding for an Ile tRNA, the anticodon identity is changed by a point mutation ($\mathsf{U} \rightarrow \mathsf{A}$) and the $\oplus$–aminoacyl synthetase loads this tRNA with Glu. In a final mutation the tRNAs identity is altered again to a non-coding pattern. The tRNA does not code any more, the codon is lost.

Another example is given for a mutation where the tRNA secondary structure violates our definition of cloverleaf fold. In figure 5.2 this is process is illustrated by an example.

Based on these sequences it is proposed that our model favors Yarus' "ambiguous intermediate model" [156]. In fact our model is built to be able to code ambiguous: a single codon may map to different amino acids because its tRNA keeps the anticodon loop and mutates other identity elements. We did not test the "genome streamlining hypothesis" because additional selective pressure would slow down the simulations significantly. Though it would be easy to perform such a test in longer runs. We were not able to observe a codon change based on the "codon capture hypothesis". If this were the case the organisms should have acquired many non-coding tRNAs as neutral reservoir for codon reassignments, in fact this was not observed.

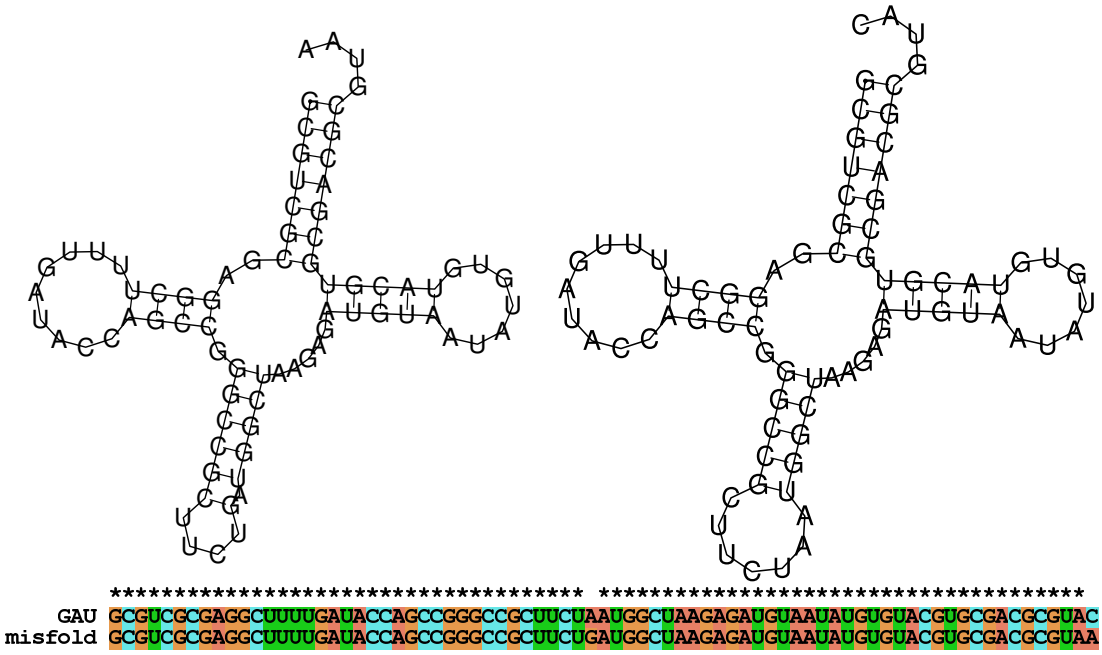


Figure 5.2: The tRNA $\mathsf{GAU}^{\mathsf{G}}$ (right structure) is altered by point mutations, that destroy the canonical cloverleaf. The bulge near the anticodon loop in the mutant (left structure) is not tolerated. The alignment of the two structures shows the mutation at position 39 that was responsible for the different fold. The additional mutation in position 77 ($\mathsf{C} \rightarrow \mathsf{A}$) would have remapped the codon to H.

In computational studies on lattice proteins the most common starting point is HP, one of any hydrophobic amino acid and a polar one. This goes back to the first theories of protein folding that argued that folding simply arise from the hydrophobic effect. The burial of hydrophobic residues is a thermodynamic most favorable state and X-ray and crystal structures confirm this pattern well. We performed our first experiments under the assumption of a HP world migrating to more amino acids.

5.2 HP Computations

Measurements of the folding speed of beads on lattices models revealed, that three different kinds of beads produce more efficient folding than two kinds (as used in HP experiments) [154]. Therefore, it is expected that an organism coding for only one polar and one hydrophobic amino acid will be driven to expand the number of

used amino acids as proposed by lattice simulations. This was the starting point for the following experiments. Several combinations of hydrophobic and polar amino acids were used as initial alphabets for simulations and our organisms were optimized to build proteins solely from these. The HP amino acids were selected to be potentially available under prebiotic conditions.

IG

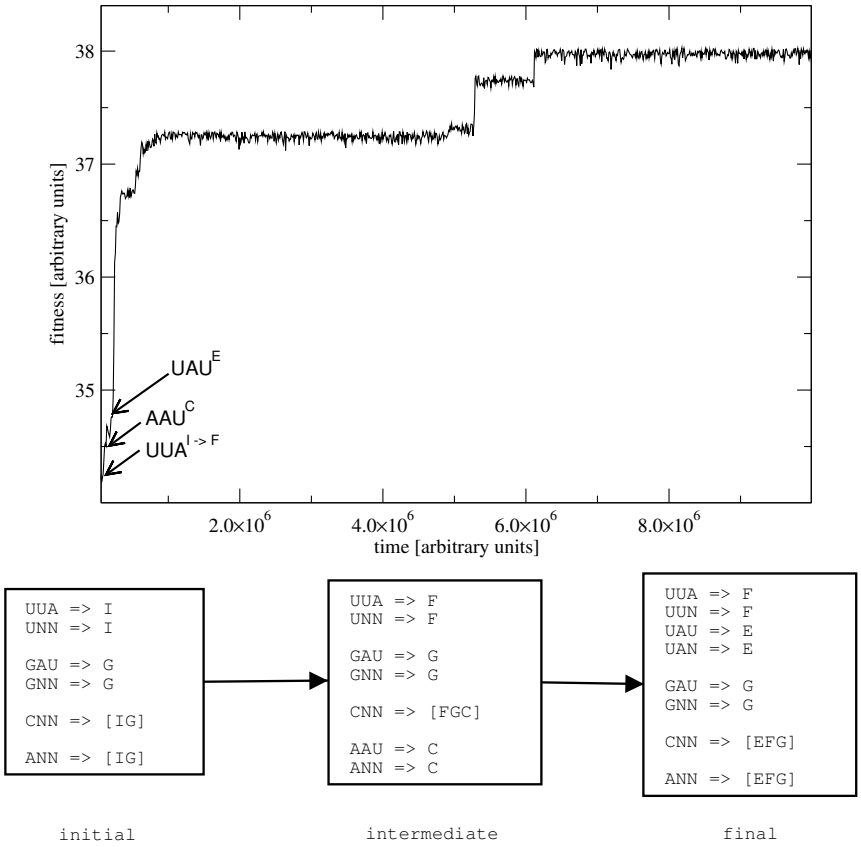


Figure 5.3: Upper graphics: time versus average fitness in the tank reactor with a start alphabet IG. The first transition is caused by the code refinements and extensions, subsequently jumps are due to sequence innovations. The block schema shows the code tables

In the simulation where the *Last Universal Common Ancestor* (LUCA) of a tank reactor population had isoleucine and glycine as start amino acids we were able to observe codon reassignment. A codon (UUA) that was assigned to I at simulation start was remapped to F after a very short period of time (only about 2×10^5

replications). This caused an enormous step in the overall fitness of the individuals as can be seen in the first transition in figure 5.3. Small saddles within this transition are caused by the code shifts outlined in the same figure. Subsequent pronounced transitions are optimizations on sequence level, caused by the inverse folding.

It is further remarkable that the code transition steps toward *diversification* not coverage. The "intermediate" code shown in the block schema of figure 5.3 covers more codons with specific assignments, than the next (marked "stop") code table does. This corresponds with the observation in contemporary codes, where the degenerated codon families are split.

LS

The LUCA for this simulation was an organism that had two tRNAs, one coding for the hydrophobic amino acid Leu and the polar searing. After a total of 4×10^7 replications the tank reactor showed a high concentration of a species that coded for another polar amino acid in addition: Lysine. After this innovation major transitions are observed that optimize the protein discontinuous on sequence level.

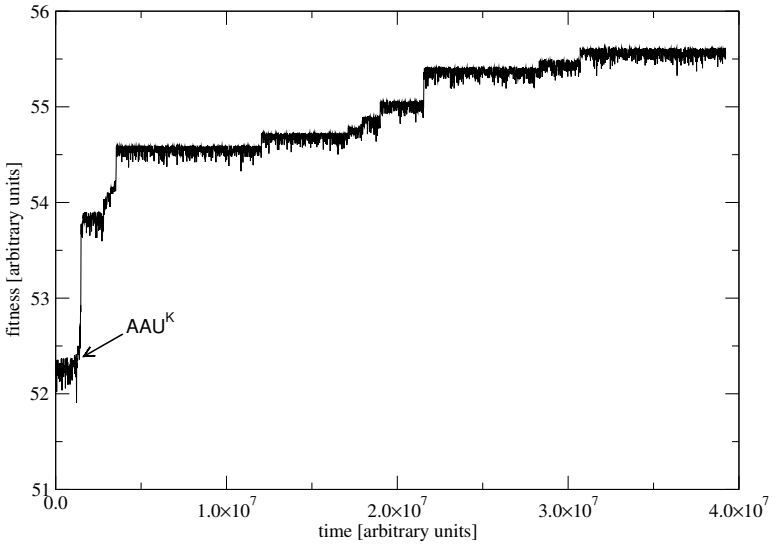


Figure 5.4: Time-Fitness Plot of start alphabet LS. The fitness value is the mean of the entire reactor, a time step is defined every thousand replication cycles (size of population).

The entire table before and after expansion of the genetic code is drawn in figure 5.5. This table shows the block structure of the code that is used very well.

AAA => [LS] ACA => [LS] AGA => [LS] AUA => [LS]	AAC => [LS] ACC => [LS] AGC => [LS] AUC => [LS]	AAG => [LS] ACG => [LS] AGG => [LS] AUG => [LS]	AAU => [LS] ACU => [LS] AGU => [LS] AUU => [LS]
CAA => S CCA => S CGA => S CUA => S	CAC => S CCC => S CGC => S CUC => S	CAG => S CCG => S CGG => S CUG => S	CAU => S CCU => S CGU => S CUU => S
GAA => [LS] GCA => [LS] GGA => [LS] GUA => [LS]	GAC => [LS] GCC => [LS] GGC => [LS] GUC => [LS]	GAG => [LS] GCG => [LS] GGG => [LS] GUG => [LS]	GAU => [LS] GCU => [LS] GGU => [LS] GUU => [LS]
UAA => L UCA => L UGA => L UUA => L	UAC => L UCC => L UGC => L UUC => L	UAG => L UCG => L UGG => L UUG => L	UAU => L UCU => L UGU => L UUU => L

AAA => K ACA => K AGA => K AUA => K	AAC => K ACC => K AGC => K AUC => K	AAG => K ACG => K AGG => K AUG => K	AAU => K ACU => K AGU => K AUU => K
CAA => S CCA => S CGA => S CUA => S	CAC => S CCC => S CGC => S CUC => S	CAG => S CCG => S CGG => S CUG => S	CAU => S CCU => S CGU => S CUU => S
GAA => [LSK] GCA => [LSK] GGA => [LSK] GUA => [LSK]	GAC => [LSK] GCC => [LSK] GGC => [LSK] GUC => [LSK]	GAG => [LSK] GCG => [LSK] GGG => [LSK] GUG => [LSK]	GAU => [LSK] GCU => [LSK] GGU => [LSK] GUU => [LSK]
UAA => L UCA => L UGA => L UUA => L	UAC => L UCC => L UGC => L UUC => L	UAG => L UCG => L UGG => L UUG => L	UAU => L UCU => L UGU => L UUU => L

Figure 5.5: Code evolution: LS expands to KLS. Ambiguous codons are specified to code for the amino acids in brackets.

LD

The basis for this simulation were organisms that had two defined codons, UGU was assigned to lysine, and CGG to aspartic acid. The population of organisms in the flow reactor shows the typical behavior: Few mutants are dominated by one single fitter (master) species that has a "tail" of surrounding competitors. The transitions along the time scale of our evolutionary experiment are discontinues. This behavior is well investigated and a typical pattern of evolutionary dynamics [51, 52]. In figure 5.6 different plateaus are observed, which are a drift on the neutral net of proteins as well as the neutrality within the genetic code.

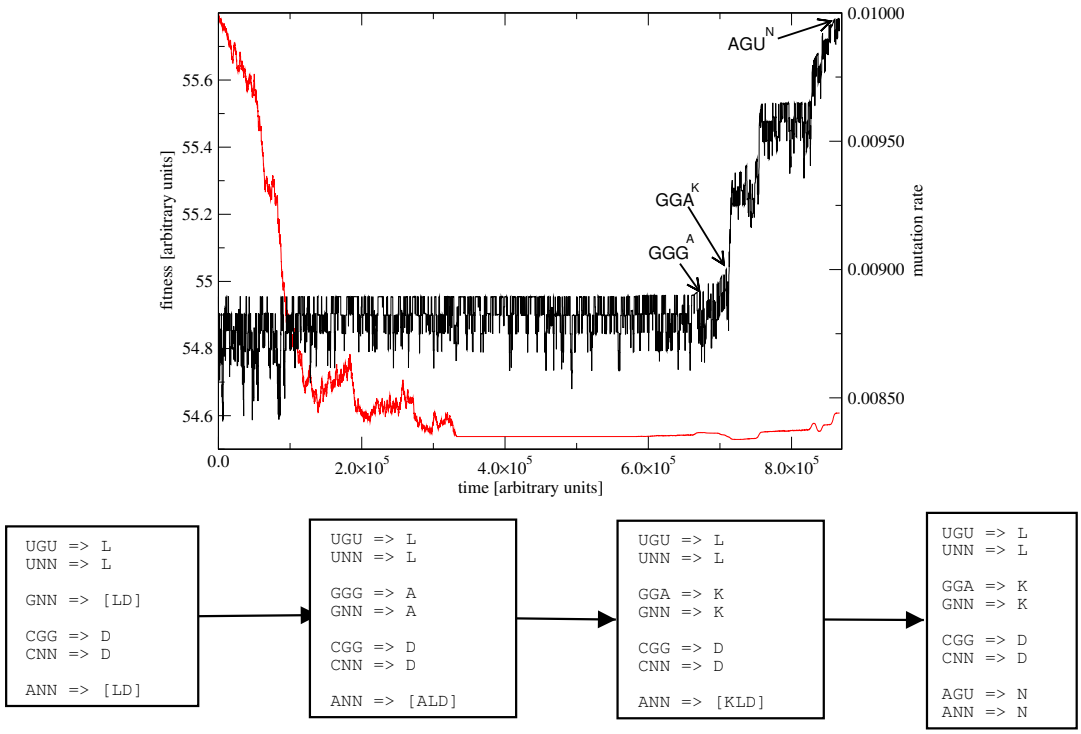


Figure 5.6: Plot time vs. fitness for a start population coding for LD (drawn in black). The red plot is the development of the mutation rate. In the lower block scheme the code development is represented.

The first transition observed at about 7×10^5 can be traced back to the migration from population coding for LD to individuals that code ADL. This extension enabled the specific loading of GNN codons. On following plateau ADL is converted to KDL, what is interestingly a neutral movement and does not alter the

size of the alphabet. The (small) fitness increase can be reasoned to be due to the incorporation of a basic amino acid (lysine) that is able to neutralize aspartic acid. The next major transition is caused by sequential optimization, followed by another code invention: the incorporation of asparagine that offers another fitness booster. This transition completes the set of specific coding triplets: from then on all triplets are ambiguously coding.

The second graph in the plot of figure 5.6 (in red) displays the mutation rate. We were able to observe that a population at first has to stand extreme pressure to optimize the mutation rate. Hence the first plateau phase is characterized by a drift on the neutral net toward a more favorable replication accuracy.

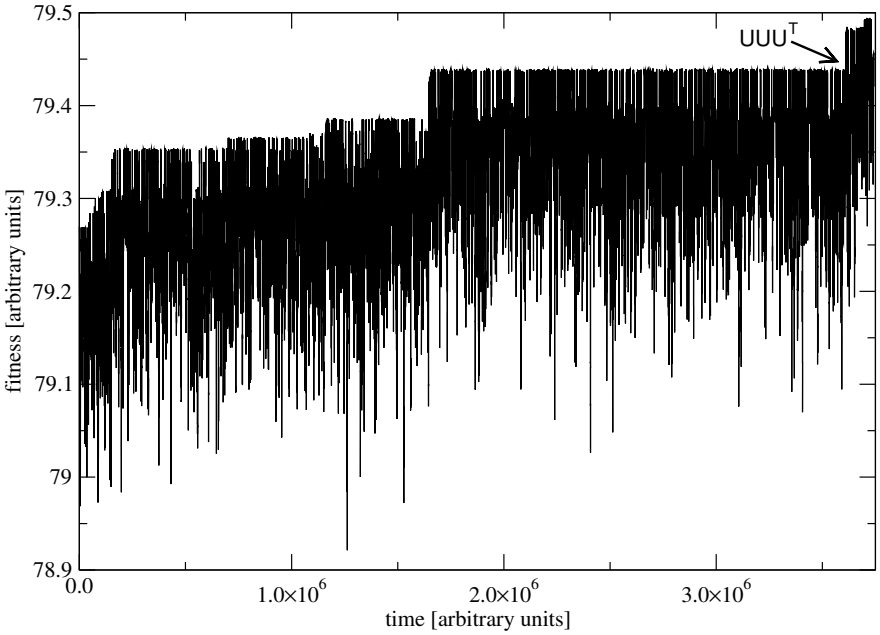
5.3 ADLG

The amino acid sub-set ADLG has been proposed [95] as primordial set to be in place before the genetic code had reached the total of todays 20 amino acids. Extensive studies [4, 5] on neutral networks in protein space revealed that this set forms long neutral paths, therefore the structures built by ADLG sequences are easily accessible by inverse folding. According to adaptive walk experiments that we performed using the tessellation potential ADLG sequences are able to form energetically favorable structures.

Our genetic code evolution studies on the ADLG subset revealed that structures built from this amino acid subset are more stable than those of solely HP. As a consequence it takes longer to surmount the local minimum and find individuals that extend the alphabet *and* have increased fitness. Nevertheless figure 5.7 documents the fitness development of a population started with this restricted alphabet set and extended it.

The organisms found the usage of the UUU codon for Thr to be advantageous. At the first glance it is astonishing that the alphabet extension does not happen for ANN codons. This would offer complete codon coverage to the organism. This observation becomes less mysterious when the locations of ANN codons on the genome sequence are considered. Only at nucleotide positions 1408-1410 the

codon AUA is found as a representative of the ANN class in this reading frame. The corresponding amino acid position (407) is located at the protein surface, and this implies that most of the hydrophilic residues should be contribute an equal energy portion. This codon 407 was acquired by mutation, because the wild type sequence (LUCA) that is product of inverse protein folding and *in silico* translation, did not have any codon of this type.




```
UAU => A
UAN => A
UGU => L
UNN => L

GAU => G
GNN => G

CGG => D
CNN => D

ANN => [ADLG]
```

```
UAU => A
UAN => A
UGU => L
UUU => T
UNN => L

GAU => G
GNN => G

CGG => D
CNN => D

ANN => [ADLGT]
```

Figure 5.7: Top: time vs. fitness plot for the ADLG start set. Every 1000 replications a mean fitness value for the population is plotted. After about 3.6×10^6 replication steps the set is enlarged for threonine. Bottom: The wild type and expanded code table for this ADLG simulation

5.4 IKEAG

If real amino acids are used instead of abstract beads, as used in lattice simulations even more types of residues are required. A phage display experiment [114] resembling the evolution of the SH3 domain (an important part of intracellular signaling) identified two hydrophobic (Ile and Ala) and three hydrophilic (Lys, Gln and Gly) as mainly sufficient to build the binding site protein. Increasing the number of amino acids used to encode a sequence decreases the ruggedness of the energy landscape for the competing structures while keeping the stability gap the same. Protein-like folding under thermodynamic control can become more reliable in the funnel-like landscape [152].

It is expected that a an organism coding for the set of amino acids that has been shown to result in sequences adopting a natural fold (like IKEAG) is very stable. Therefore the organism has little affinity to expand the amino acid alphabet. The experiments performed on organisms containing tRNAs loadable with solely {IKEAG} could not show this behavior. The population was able to find individuals that were fitter using an enlarged alphabet. Also some modifications were found: instead of Gly Asp was used in the loading. The plot shown in figure 5.8 shows the temporal behavior of an example simulation. It is worth noting that the mutation rate is optimized very rapid, and remains low (a negative value means that the population is trapped, and can not modify any more. This is of course an artifact produced by our simulation, nevertheless it shows that the mutation rate is the most significant item of optimization.

The changing of the alphabet again demonstrates that pure availability is not sufficient for the fixation of an alphabet. The thermodynamic differences of protein structures are more important for reliable folding a sequence, than the biosynthetic availability and physicochemical nearness of the residues. Our model does not contain any information about the dynamics of folding, the protein structures (in this case just the replicase) are thought to be in thermodynamical equilibrium. Though this is inadequate for a correct description of protein folding this does not significantly affect the evolution of the genetic code. Therefore the evolution of the protein is driven by thermodynamic stability of the structure, the evolution of the genetic code depends on the loading of the tRNA adapter as well.

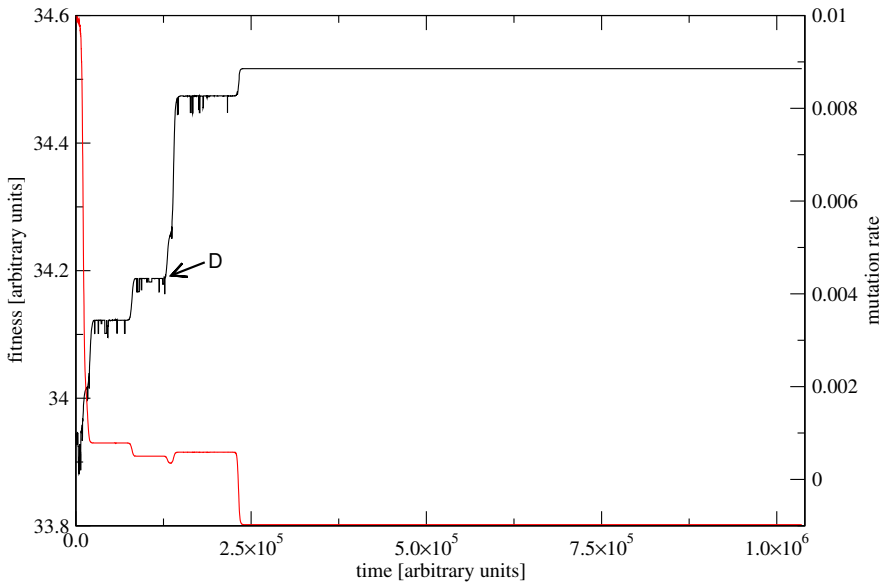


Figure 5.8: Progression of a simulation started with IKEAG, ended at IKEAD. The red curve shows the mutation rate, the black line is the fitness progression. The extension of the alphabet is marked by an arrow.

CHAPTER 6

Conclusion and Outlook

We have presented a simplified model of a primitive cell that shows that and how a primordial genetic code can evolve. The model is based on well-known biophysical principles: We explicitly simulate the evolution of a population of haploid "cells" in a flow reactor in such a way that the fitness (replication rate) is explicitly dependent upon an encoded protein product. Both the gene for this replicase protein and the translation mechanism itself is subject to mutational change.

The cell consists of an RNA genome, that codes for a replicase and for tRNAs as well as all necessary monomers for translation and replication. Replication happens as erroneous copy process of the RNA genome. The tRNAs are folded into their mfe secondary structure using the Vienna RNA package. The tRNA can only be used as adapter in translation if the secondary structure matches our definition of the cloverleaf structure and an $\oplus$–aminoacyl synthetase is able to load the tRNA with an amino acid (by identifying designated identity elements). The RNA genome is directly translated by the tRNAs and the polypeptide is threaded onto the $3D$ structure of a native $T7$ phage replicase via an empirical tessellation potential. The z-score of the sequence on this structure determines the replication rate, the energy variation of position within the active center of the

enzyme of *T7* wild type and translated replicase is responsible for the mutation rate of replication. Therefore the replication rate is directly influenced by the product of translation via the z-score.

Our simulation revealed that the genetic code of an organism modeled after the biophysical reality is able to evolve and expand the number and type of amino acids and codons. The mechanism of the code expansion is comparable to the one proposed in the ambiguous coding mechanism [156], in fact our model is built in that way. We could not observe a codon change using the mechanism proposed by the codon capture hypothesis. In this case it would be expected, that the organisms would acquire numerous unused tRNAs, what is not the case in our experiments. We sis not test for the genome streamlining hypothesis since this would require the usage of multiple genes and variations in genome length, what is computationally to costly at the moment.

The fitness behavior of a flow reactor environment that is framework for minimal organisms is discontinuous. Since the initial (start) replicase was an inverse folded (optimized) protein, the first major fitness increase is caused by the expansion of the genetic code otherwise the organisms would have to optimize the structure in a preceding step. The population is very heterogeneous at all times: A master species predominates some less fitter mutants. If a fitter variant is found, the population soon migrates towards the optimum giving rise to a new master species. But the new protein, coded by an enlarged amino acid set is target of extensive optimization afterward, nevertheless further optimization on a code level can be observed. For instance our simulations of organisms starting solely with tRNAs coding for leucine and aspartic acid expanded the amino acids via two intermediate steps to a final set of coding amino acids containing D, L, N and K.

Before a code innovation can take place, the mutation rate of the replicase is optimized. This was observed for the simplest start alphabets as well as for enlarged sets. For instance the amino acid set I, K, E, A and G showed out to be very suitable to build stable proteins, and find individuals unaccessible to mutations. This is explainable in the motivation of a population that found a decent code to build stable proteins and perform phenotypic optimization of the raw peptide

structure. Only the need for more stabilization or refined catalysis, demanded by a changing environment or alternative metabolites leads to the use of a more complex alphabet. The enhanced, correct replication could give rise to longer genomes. This finding is predicted by the hypercyclic theory of evolutionary dynamics, and well confirmed. The most important driving force for the invention of proteins was the ability to reproduce genomes more accurately. Hence increased replication mechanisms could enable longer genes and enhance neutrality. Our simulations forbid changes in sequence length since threading variable sequences onto a given target structure would be computationally to expensive for large scale simulations. Nevertheless it is an artifact of our simulations that in some examples mutation rates of zero and below are found. It might be more realistic to have an explicit lower bound μ_0 on the mutation rate since even a highly optimized machinery can never achieve error-free copying. If this physical accuracy limit is small, however, we cannot expect further innovations within the available computer time, hence our conclusions for these simulations remain valid.

We were not able to observe any optimization for error correction. In our model calculation the translation is error free, the only possible optimization for error correction can therefore act on the genomic level. However it could be shown that the selection for robustness against mutations happens by migration on neutral networks [87, 138, 142]. This effect is a second order mutation effect and hence very weak. In the case of the model organism the selection for mutational robustness was not observed because the fitness landscape was found to be very rough, i.e. neutral mutations are rare.

Summarizing the most reasonable scenario for the development of the genetic code started with a reduced set of amino acids that interacted more or less specific with nucleic acids in an RNA world or earlier. A *last universal common ancestor* used an ambiguous and small codon table that was implemented in tRNAs and RNA based aminoacyl tRNA synthetases. This enabled code optimization on a genetic level by duplication and point mutation. Because the mapping was *indirect* the specificity could change from one amino acid to another. The code was then expanded by incorporating more accessible amino acids and by chemical modification of already incorporated ones. After each expansion a long optimization

period was necessary, to fixate a change on the protein level and to optimize the genomic sequence for the new amino acid alphabet. This process was repeated until the whole set of 20 amino acids was reached and continuously takes place. The genetic code is no *frozen accident* , it is under evolutionary optimization like all other features of organisms as well.

The extension of the number of amino acid a primitive genetic code uses is possible, but not mandatory. Despite the fact that the coding schema in the simulations allowed all 20 amino acids, the code is expanded extremely slowly. The timescale of optimizing the available organisms seems to be larger by the orders of magnitude. This scenario would correspond to a prebiotic world where all 20 amino acids were available, what is not likely. However the pure availability of an amino acid is not sufficient for code changes as proposed by the co-evolution theory. This is consistent with the prediction of coding theory that information can never pass from an alphabet of higher entropy to one of lower. However this would be the case if the larger amino acid alphabet would modify the coding nucleotide words because of the occurrence of an amino acid due to a new synthesis pathway. Hence we are inclined to argue that the pattern described by the co-evolution theory is simply formed by chance in agreement with Freeland *et.al.* [54].

The basis for optimization of the code with respect to mutational impact requires the accessibility of variant codes. Our model showed that it is not likely that an extensive network of interconnected codes exists, in contrast we propose that the codon space is a very rough landscape in terms of optimization and one gets easily trapped in local minima. This makes it impossible to find an optimal code within the timescale of our simulations, although we can observe a few adaptive steps.

Our model was focused on the investigation of the *general* mechanism of code expansion, not the assignment of a particular amino acid to a codon sequence. Therefore we are not able to make any prediction for the stereochemical affinity and its theory. The acceptance of amino acids and selection of codons by them depends on the concentration of the available amino acid, therefore the composition of any primordial soup (or hot-spring environment) is unknown and even

the suggestive experiments of Miller and coworkers can not predict the exact concentration of amino acids.

The studies performed so far are promising and hold capacity for improvement. Especially the features of the RNA behavior, which are computationally cheap could provide deeper insights. The codon-anticodon interaction could be modeled more realistic by explicitly using experimental energy parameters (via the Vienna RNA package) for RNA base pairing. It is expected that according the suggestions of Eigen [41] and co-workers that a shift to GC-rich codes should happen.

Folding of mRNA onto itself makes some codons more or less accessible to the read off by tRNAs. It is expected that elaborated loop structures enhance transcription rates and this mechanism could serve as a kind of primitive transcription regulation. It should be investigated if our simple artificial life model is able to show such a behavior. Also studies to test the complementary code theory, that was recently tested by statistical analysis of retroviral mRNA [87] could be investigated by testing whether any mRNA secondary structure motives influence codon assignments.

Another enhancement with respect to a more realistic modeling of the artificial organism could be achieved by coding more proteins on the mRNA. This would change our fitness function slowly and make sure that the phage replicase does not bias our simulations.

The loading algorithm of the tRNA via the `XOR` filter is a useful simplification of what is considered important *in vivo*: A combination from structural and sequential information is recognized by the aminoacyl tRNA synthetase for proper loading. This simple filter could, for instance, be replaced by a three layer back propagated neural net that is trained using tRNAs sequences from a database in combination with their predicted secondary structure. This setup could model the aminoacyl synthetase reaction more native like.

The mode of inheritance is strictly asexual in the flow reactor as we built it. This stands in contrast to the proposal of Woese et. al [150] who proposed the universal ancestor to be more a population of genes under heavy genetic exchange,

rather than isolated individuals. The cross-over operator could act upon tRNA genes for example and find new combinations of tRNAs. This kind of sexual reproduction could enhance the optimization heuristic and shorten the plateau phases of individual adaptation.

One of the most interesting questions that could be investigated by the aid of the GCE application framework concerns the ability of the code to learn refined recognition. Is it possible to show the splitting of family blocks if a large amino acid subset (such as five polar and five hydrophobic amino acids)? Is it possible to optimize the specificity of the loading itself? These question should be faced in further computer experiments and extension to the implemented software.

APPENDIX A

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the cell*. Garland Publishing, New York and London, 1984.

[2] R. Amirnovin. An analysis of the metabolic theory of the origin of the genetic code. *J. Mol. Evol.*, 44(5):473–476, May 1997.

[3] S. G. Andersson and C. G. Kurland. Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.*, 73:775–787, 1995.

[4] A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, and P. F. Stadler. Exploring protein sequence space using knowledge based potentials. *J. Theor. Biol.*, 212:35–46, 2001.

[5] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997.

[6] R. T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lodon*, 53:370–418, 1763.

[7] D. Beazley, D. Fletcher, and D. Dumont. Perl extension building with SWIG, 1998.

[8] C. K. Biebricher, M. Eigen, and W. C. J. Gardiner. Kinetics of RNA replication: Plus-minus asymmetry and double-strand formation. *Biochemistry*, 23:3186–3194, 1984.

[9] C. K. Biebricher, M. Eigen, and W. C. J. Gardiner. Kinetics of RNA replication: Competition and selection among self-replicating RNA species. *Biochemistry*, 24:6550–6560, 1985.

[10] C. K. Biebricher and R. Luce. In vitro recombination and terminal elongation of RNA by Qβ replicase. *EMBO Journal*, 11:5129–5135, 1992.

[11] C. K. Biebricher and R. Luce. Sequence analysis of RNA species synthesized by Qβ replicase without template. *Biochemistry*, 32:4848–4854, 1992.

[12] C. K. Biebricher and R. Luce. Template-free generation of RNA species that replicate with bacteriophage T7 RNA polymerase. *EMBO Journal*, 15:3458–3465, July 1996.

[13] C. Blomberg. On the appearance of function and organisation in the origin of life. *J. Theor. Biol.*, 187(4):541–554, August 1994.

[14] C. Böhler, P. E. Nielsen, and L. E. Orgel. Template switchnig between PNA and RNA oligonucleotides. *Nature*, 376:578–581, 1995.

[15] J. U. Bowie, R. Lüthy, and D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253:164–169, 1991.

[16] C. Bradford Barber, D. P. Dobkin, and H. T. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22:469–421, 1996.
`URL: http://www.acm.org`.

[17] M. D. Brasier, O. R. Green, A. P. Jephcoat, A. K. Kleppe, M. J. Van Kranendonk, J. F. Lindsay, A. Steele, and N. V. Grassineau. Questioning the evidence for earth's oldest fossils. *Nature*, 416:76–81, March 2002.

[18] D. D. Buechter and P. R. Schimmel. Dissection of a class II tRNA synthetase: determinants for minihelix recognition are tightly associated with domain for amino acid activation. *Biochemistry*, 19:5267–5272, May 1993.

[19] A. G. Cairns-Smith and C. J. Davies. *Genetic Takeover and the Mineral Origins of Life*. Cambrige University Press, 1982.

[20] C. W. Carter jr., B. C. LeFebvre, S. A. Cammer, A. Tropsha, and M. H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.*, 311:625–638, 2001.

[21] G. Casari and M. J. Sippl. Structure-derived hydrophobic potential: Hydrophobic potentials derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, 224:725–732, 1992.

[22] T. Cech, A. Zaug, and P. Krabowski. In vitro splicing of the ribosomal RNA precursor of *tetrahymena*: involvement of a guanosin nucleotide in the excision of the intervening sequence. *Cell*, 27:487–496, 1981.

[23] G. M. Cheetham, D. Jeruzalmi, and T. A. Steitz. Structural basis for initiation of transcription from an RNA polymerase-promoter complex. *Nature*, 399:80–83, May 1999.

[24] G. D. Cody, N. Z. Boctor, T. R. Filley, R. M. Hazen, J. H. Scott, A. Sharma, and H. S. J. Yoder. Primordial carbonylated iron-sulfur compounds and the synthesis of pyruvate. *Science*, 289:1337–1340, August 2000.

[25] F. H. C. Crick. A note for the tRNA tie club. Quoted by M. B. Hoagland in *The Nucleic Acids*, E.Chargaff and J. N. Davies, Eds. (Academic Press, New York, 1960) vol 3, page 400.

[26] F. H. C. Crick. The origin of the genetic code. *J. Mol. Biol.*, 38:367–379, 1968.

[27] C. Darwin. *The Origin of Species by means of natural selection, or the preservation of favoured races in the struggle for life*. J. Murray, London, 1859.

[28] M. Delarue, O. Poch, N. Tordo, D. Moras, and P. Argos. An attempt to unify the structure of polymerases. *Protein Engineering*, 6:461–746, May 1990.

[29] M. Di Giulio. The extension reached by the minimization of the polarity distance during the evolution of the genetic code. *J. Mol. Evol.*, 29:288–293, 1989.

[30] M. Di Giulio. Reflections on the origin of the genetic code: a hypothesis. *J. Theor. Biol.*, 191:191–196, 1998.

[31] M. Di Giulio. The origin of the genetic code. *Trends in Biochemical Sciences*, 25:44, 2000.

[32] M. Di Giulio, M. Capobianco, and M. Medugno. On the optimization of the physicochemical distances between amino acids in the evolution of genetic code. *J. Theor. Biol.*, 168:43–51, 1994.

[33] M. Di Giulio and M. Medugno. Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. *J. Mol. Evol.*, 49(1):1–10, July 1999.

[34] V. Doring and P. Marliere. Reassigning cystein in the genetic code of esterichia coli. *Genetics*, 150:543–51, 1998.

[35] G. Dueck. New optimization heuristics: The Great Deluge algorithm and the Record-to-Record travel. *Journal of Computation Physics*, 104:86–92, 1993.

[36] M. Eigen. Selforganization of matter and the evolution of macromolecules. *Naturwiss.*, 58:465–523, 1971.

[37] M. Eigen, B. F. Lindemann, M. Tietze, R. Winkler-Oswatitsch, A. Dress, and A. von Haeseler. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science*, 244(4905):673–679, May 1989.

[38] M. Eigen, J. McCaskill, and P. Schuster. Molecular Quasi-Species. *J. Phys. Chem.*, 92:6881–6891, 1988.

[39] M. Eigen and P. Schuster. The hypercycle: A. Emergence of the hypercycle. *Naturwiss.*, 64:541–565, 1977.

[40] M. Eigen and P. Schuster. The hypercycle: B. The abstract hypercycle. *Naturwiss.*, 65:7–41, 1978.

[41] M. Eigen and P. Schuster. The hypercycle: C. The realistic hypercycle. *Naturwiss.*, 65:341–369, 1978.

[42] M. Eigen and R. Winkler-Oswatitsch. Transfer-RNA, an early gene? *Naturwiss.*, 68:228–292, 1981.

[43] M. Eigen and R. Winkler-Oswatitsch. Transfer-RNA: The early adaptor. *Naturwiss.*, 68:217–228, 1981.

[44] E. H. Ekland and D. P. Bartel. RNA-catalysed RNA polymerization using Nucleoside Triphosphates. *Nature*, 382:373–376, July 1996.

[45] D. A. Ellington, M. Khrapov, and C. A. Shaw. The scene of a frozen accident. *RNA*, 6:485–498, 2000.

[46] G. Eriani, M. Delarue, O. Poch, J. Gangloff, and M. D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature*, 347:203–206, September 1990.

[47] A. Eschenmoser. Chemistry of potentially prebiological natural products. *Orig. Life Evol. Biosph.*, 24:389–423, 1994.

[48] A. Eschenmoser. Towards a chemical etiology of nucleic acid structure. *Orig. Life Evol. Biosph.*, 27:535–553, 1997.

[49] M. Famulok. Molecular recognition of amino acids by RNA-aptamers: an L-citrulline binding RNA motif and its evolution into an L-arginine binder. *J. Am. Chem. Soc.*, 116:1698–1706, 1994.

[50] C. Flamm, W. Fontana, I. L. Hofacker, and P. K. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 1999.

[51] W. Fontana and P. Schuster. A computer model of evolutionary optimization. *Biophysical Chemistry*, 26:123–147, 1987.

[52] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.

[53] S. J. Freeland. Early fixation of an optimal genetic code. *J. Mol. Evol.*, 17(4):511–518, 2000.

[54] S. J. Freeland and L. D. Hurst. The genetic code is one in a million. *J. Mol. Evol.*, 47:238–248, 1998.

[55] S. J. Freeland, R. D. Knight, and L. F. Landweber. Do proteins predate DNA? *Science*, 286(5440):690–692, October 1999.

[56] S. J. Freeland, R. D. Knight, and L. F. Landweber. Measuring adaptation within the genetic code. *Trends in Biochemical Sciences*, 25:44–45, 2000.

[57] J. Friedl. *Mastering Regular Expressions*. O'Reilly and Associates,, 1997.

[58] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Publishing Company, New York, NY, 1995.

[59] G. Gamow. Possible relation between DNA and the protein structures. *Nature*, 173:318, 1954.

[60] R. Gesteland and J. Atkins (editors). *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbour, NY, USA, 1993.

[61] R. Giegé, M. Sissler, and C. Florentz. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research*, 26(22):5017–5035, 1998.

[62] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[63] D. T. Gillespie. A general method for numerical simulating the stochastic time evolution of coupled chemicla reactions. *J. Phys. Chem.*, 22:403–434, 1976.

[64] B. L. Golden, A. R. Gooding, E. R. Podell, and T. R. Cech. A preorganized active site in the crystal structure of the *tetrahymena* ribozyme. *Science*, 282(5387):259–264, October 1998.

[65] N. Goldman. Further results on error minimization in the genetic code. *J. Mol. Evol.*, 37:662–664, 1993.

[66] T. Grossman, R. Farber, and A. Lapedes. Neural net representations of empirical protein potentials. *Ismb*, 3:154–161, 1995.

[67] A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij. An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, 5:609–617, 1999.

[68] D. Haig and L. Hurst. A quantitative measurement of error minimization in the genetic code. *J. Mol. Evol.*, 33:412–417, 1991.

[69] H. Hampl, H. Schulze, and K. H. Nierhaus. Ribosomal components from escherichia coli 50S subunits involved in the reconstitution of peptidyltransferase activity. *J. Biol. Chem.*, 256(5):2284–2288, March 1981.

[70] R. M. Hazen, T. R. Filley, and G. A. Goodfriend. Selective adsorption of L- and D-amino acids on calcite: Implications for biochemical homochirality. *Proc. Natl. Acad. Sci.*, 98:5487–5490, May 2001.

[71] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models — the calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.*, 216:167–180, 1990.

[72] I. L. Hofacker. Sparse data correction for empirical potentials. personal communication, 2000.

[73] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh.Chem.*, 125(2):167–188, 1994.

[74] J. H. Holland. *Adaptation in natural artificial systems*. University of Michigan Press, Ann Arbor, 1975.

[75] J. J. Hopfield. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence and kinetic proofreading. *Proc. Natl. Acad. Sci. USA*, 75:4334–4338, 1978.

[76] M. A. Jiménez-Montaño, C. R. de la Mora-Basanez, and T. Poschel. The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions in vivo and in vitro. *BioSystems*, 39(2):117–125, 1996.

[77] J. R. Jungck. The genetic code as a periodic table. *J. Mol. Evol.*, 11:211–224, 1978.

[78] J. F. Kasting. Earth's early atmosphere. *Science*, 259:920–926, February 1993.

[79] S. A. Kauffman. Applied molecular evolution. *J. theor. Biol.*, 157:1–7, 1992.

[80] S. A. Kauffman. *The Origin of Order*. Oxford University Press, New York, Oxford, 1993.

[81] P. Khaitovich, A. S. Mankin, R. Green, L. Lancaster, and H. F. Noller. Characterization of functionally active subribosomal particles from thermus aquaticus. *Proc. Natl. Acad. Sci.*, 96:85–90, January 1999.

[82] M. Kimura (editor). *The neutral theory of evolution*. Cambridge University Press, Cambridge, 1983.

[83] R. D. Knight, S. J. Freeland, and L. F. Landweber. Selection, history and chemistry: The three faces of the genetic code. *Trends in Biochemical Sciences*, June 1999.

[84] R. D. Knight, S. J. Freeland, and L. F. Landweber. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, 2(4):1–13, March 2001.

[85] R. D. Knight and L. F. Landweber. Rhyme or reason: RNA-arginine interactions and the genetic code. *Chemistry & Biology*, 5(9):215–220, September 1998.

[86] R. D. Knight and L. F. Landweber. Guilt by association: The arginine case revisited. *RNA*, 6:499–510, 2000.

[87] J. Konecny, M. Schoniger, I. Hofacker, M. D. Weitze, and G. L. Hofacker. Concurrent neutral evolution of mrna secondary structures and encoded proteins. *J. Mol. Evol.*, 50(3):238–242, March 2000.

[88] J. Konecny, M. Schöninger, and L. G. Hofacker. Complementary coding conforms to the primeval comma-less code. *J. Theor. Biol.*, 173:263–270, 1995.

[89] L. F. Landweber. Testing ancient RNA-protein interactions. *Proc. Natl. Acad. Sci. USA*, 96:11067–11068, September 1999.

[90] L. F. Landweber, P. J. Simon, , and T. A. Wagner. Ribozyme design and early evolution. *BioSystems*, 48:94–103, 1998.

[91] W. Leinfelder, E. Zehelein, M. A. Mandrand-Berthelot, and A. Bock. Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine. *Nature*, 331:723–725, February 1988.

[92] P. A. Lohse and J. W. Szostak. Ribozyme-catalysed amino-acid transfer reactions. *Nature*, 381:442–444, May 1996.

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

[94] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[95] S. Miller and L. Orgel. *The Origin of Life on the Earth*. Prentice Hall, 1974.

[96] S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117:528–529, 1953.

[97] D. R. Mills, R. L. Peterson, and S. Spiegelman. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci.*, 58:217, 1967.

[98] T. Mizutani and T. Hitaka. The conversion of phosphoserine residues to selenocysteine residues on an opal suppressor tRNA and casein. *FEBS Letter*, 232(1):243–248, May 1988.

[99] S. J. Mojzsis, G. Arrhenius, K. McKeegan, T. M. Harrison, A. P. Nutman, and C. R. Friend. Evidence for life on earth before 3,800 million years ago. *Nature*, 384:55–59, November 1996.

[100] P. J. Munson and R. K. Singh. Statistical significance of hierarchical multi-body potentials based on delauney tessellation and their application in sequence-structure alignment. *Protein Science*, 6:1467–1481, 1997.

[101] K. Nieselt-Struwe and P. R. Wills. The emergence of genetic coding in physical systems. *J. Theor. Biol.*, 187:1–14, 1997.

[102] M. W. Nirenberg, T. Caskey, R. Marshall, R. Brimacombe, D. Kellogg, B. Doctor, D. Hattfield, J. Levin, F. Rottman, S. Pestka, M. Wilcox, and F. Anderson. The RNA code and protein synthesis. *Cold Spring Harbor Symp. Quant. Biol.*, 31:11–24, 1966.

[103] M. W. Nirenberg and P. Leder. RNA codewords and protein synthesis. *Science*, 145:1399–1407, 1964.

[104] M. W. Nirenberg and J. H. Matthaei. The dependence of cell-free protein synthesis in *e. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA*, 47:1588–1602, 1961.

[105] A. I. Oparin. *The origin of Life on earth*. Academic Press Inc., 1936.

[106] S. Osawa. *Evolution of the genetic code*. Oxford University Press, Oxford, 1995.

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

[108] P. A. Osumi-Davis, M. C. de Aguilera, R. W. Woody, and A. Y. Woody. Asp537, Asp812 are essential and Lys631, His811 are catalytically significant in bacteriophage T7 RNA polymerase activity. *J. Mol. Biol.*, 226:37–45, July 1992.

[109] J. A. Piccirilli, T. S. McConnell, A. J. Zaug, H. F. Noller, and T. R. Cech. Aminoacyl-esterase activity of the *tetrahymena* ribozyme. *Science*, 256:1420–1424, June 1992.

[110] S. A. Pizzarello and J. R. Cronina. Non-racemic amino acids in the Murray and Murchison meteorites. *Geochimica et Cosmochimica Acta*, 64(2):329–338, January 2000.

[111] C. Reid and L. E. Orgel. Synthesis of sugar in potentially prebiotic conditions. *Nature*, 216:455, 1967.

[112] L. Ribas de Pouplana and P. Schimmel. Aminoacyl-tRNA synthetases: potential markers of genetic code characterized by an active site domain that has a development. *TIBS*, 26(10):591–596, October 2001.

[113] L. Ribas de Pouplana and P. Schimmel. Two classes of tRNA synthetases suggested by sterically compatible dockings on tRNA acceptor stem. *Cell*, 104(2):191–193, January 2001.

[114] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker. Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol*, 10:805–809, October 1997.

[115] T. A. Ronneberg, L. F. Landweber, and S. J. Freeland. Testing a biosynthetic theory of the genetic code: Fact or artifact? *Proc. Natl. Acad. Sci. USA*, 97:13690–13695, December 2000.

[116] A. Roth and R. R. Breaker. An amino acid as a cofactor for a catalytic polynucleotide. *Proc. Natl. Acad. Sci.*, 95:6027–6031, May 1998.

[117] M. A. Rould, J. J. Perona, D. Soll, and T. A. Steitz. Structure of e. coli glutaminyl-tRNA synthetase complexed with tRNA(gln) and ATP at 2.8 Å resolution. *Science*, 246:1135–1142, December 1989.

[118] M. Ruff, S. Krishnaswamy, M. Boeglin, A. Poterszman, A. Mitschler, A. Podjarny, B. Rees, J. C. Thierry, and D. Moras. Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(asp). *Science*, 252:1682–1689, June 1991.

[119] M. E. Saks, J. R. Sampson, and J. Abelson. Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science*, 279:1665–1668, 1998.

[120] P. Schimmel, R. Giegé, D. Moras, and S. Yokoyama. An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl. Acad. Sci. USA*, 90:8763–8768, 1993.

[121] S. Schönauer and P. Clote. How optimal is the genetic code? In D. Frishman and H. Mewes (editors), *Computer Science and Biology, Proceedings of the German Conference on Bioinformatics (GCB'97)*, pages 65–67. September 1997.

[122] J. W. Schopf. Microfossils of the early Archean apex chert new evidence of the antiquity of life. *Science*, 260:640–646, 1993.

[123] P. Schuster, P. F. Stadler, and A. Renner. RNA structures and folding: From conventional to new issues in structure predictions. *Curr. Opinions Structural Biol.*, 7:229–235, 1997.

[124] G. Segré. The Big Bang and the genetic code. *Nature*, 404:437, March 2000.

[125] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–424, 623–656, 1948.

[126] R. Shapiro. Prebiotic cytosine synthesis: A critical analysis and implications for the origin of life. *Proc. Natl. Acad. Sci.*, 96:4396–4401, April 1999.

[127] R. K. Singh, A. Tropsha, and I. I. Vaisman. Delauney tessellation of proteins: Four body nearest neighbor propensity of amino acid residues. *J. Comp. Biol.*, 3:213–221, 1996.

[128] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force — An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883, 1990.

[129] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures. *Journal of Computer-Aided Molecular Design*, 7:473–501, 1993.

[130] M. J. Sippl. Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17:355–362, 1993.

[131] M. J. Sippl. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.*, 5:229–235, 1995.

[132] R. Sousa, Y. J. Chung, J. P. Rose, and B. C. Wang. Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. *Nature*, 364:593–599, 1993.

[133] S. Spiegelman. An approach to experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 4:213–253, 1971.

[134] S. Srinivasan. *Advanced Perl Programming*. O'Reilly & Associates, Inc., 101 Morris Street, Sebastpol, CA 95472, first edition, 1997.

[135] A. Svrcek-Seiler. personal communication, 2002.

[136] E. Szathmáry. Coding Coenzyme Handles: A hypothesis for the origin of the genetic code. *Proc. Natl. Acad. Sci. USA*, 90:9916–9920, 1993.

[137] F. J. R. Taylor and D. Coates. The code within the codons. *BioSystems*, 22:177–187, 1989.

[138] E. van Nimwegen, J. P. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci.*, 96(17):9716–9720, 1998.

[139] G. von Kiedrowski. Ein selbstreplizierendes Hexadesoxynucleotid. *Angew. Chem.*, 93:932, 1986.

[140] G. von Kiedrowski. Minimal replicator theory I: Parabolic versus exponential growth. In *Bioorganic Chemistry Frontiers, Volume 3*, pages 115–146. Springer-Verlag, Berlin, Heidelberg, 1993.

[141] G. Wächtershäuser. Origin of life. Life as we don't know it. *Science*, 289:1307–1308, August 2000.

[142] A. Wagner and P. F. Stadler. Viral rna and evolved mutational robustness. *J. Exp. Zool./ MDE*, 285:119–127, 1999.

[143] L. Wall, T. Christiansen, and S. R. L. *Programming Perl*. O'Reilly & Associates, Inc., 101 Morris Street, Sebastpol, CA 95472, third edition, 2000.

[144] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42:257–266, 1978.

[145] J. D. Watson and F. H. C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737, 1953.

[146] G. Weberndorfer, I. L. Hofacker, and P. F. Stadler. An efficient potential for protein sequence design. In *GCB '99: German Conference on Bioinformatics*. 1999.

[147] T. W. Wiegand, R. C. Janssen, and B. Eaton. Selection of RNA amide synthases. *Chem. Biol.*, 9:675–683, 1997.

[148] P. R. Wills, K. Nieselt-Struwe, and L. Henderson. The evolution of genetic coding. *NZ BioScience*, pages 7–10, May 1997.

[149] C. R. Woese. On the origin of the genetic code. *Proc. Natl. Acad. Sci. USA*, 54:1546–1552, 1965.

[150] C. R. Woese. The universal ancestor. *Proc. Natl. Acad. Sci. USA*, 95(12):6854–6859, June 1998.

[151] C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, and W. C. Saxinger. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbour Symp. Quant. Biol.*, 31:723–736, 1966.

[152] P. G. Wolynes. As simple as can be? *Nature Structural Biology*, 11:871–874, 1997.

[153] P. G. Wolynes. Computational biomolecular science. *Proc. Natl. Acad. Sci.*, 95:5848, May 1998.

[154] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 268(5204):1619–1620, March 1995.

[155] J. T. F. Wong. A Co-Evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA*, 72:1909–1912, 1975.

[156] M. Yarus and D. Schultz. Toward a theory of malleability in genetic coding. *Journal of Molecular Evolution*, 45:3–6, 1997.

[157] M. Yusupov, G. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller. Crystal structure of the ribosome at 5.5 Å resolution. *Science*, 292:883–896, 2001.

[158] B. Zhang and T. R. Cech. Peptide bond formation by in vitro selected ribozymes. *Nature*, 390:96–100, November 1997.

[159] L. Zhang and J. Skolnick. What should the z-score of native protein structures be? *Protein Science*, 7:1201–1207, 1998.

[160] W. Zheng, S. J. Cho, I. I. Vaisman, and A. Tropha. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In L. Hunter and T. Klein (editors), *Biocomputing: Proceedings of the 1997 Pacific Symposium*, pages 486–497. World Scientific Publishing Co, 1997.

[161] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[162] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull.Math.Biol.*, 46(4):591–621, 1984.

[163] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

APPENDIX B

Common Abbreviations

RNA	.....	ribonucleic acid
DNA	.....	desoxyribonucleotic acid
tRNA	.....	transfer RNA
mRNA	.....	messenger RNA

Table B.1: Biopolymers and Acronyms

abbreviation	description
A	Adenine
C	Cytosine
G	Guanine
U	Uracil
T	Thymine
R	any purine nucleotide(A, G)
Y	any pyrimidine (C, U)
N	any nucleotide (A, C, G, U)

Table B.2: Nucleotides

Amino acid name	One letter	Three letter	Mass [g/mol]
Alanine	A	Ala	71.079
Arginine	R	Arg	156.188
Asparagine	N	Asn	114.104
Aspartic acid	D	Asp	115.089
Cysteine	C	Cys	103.145
Glutamine	Q	Gln	128.131
Glutamic acid	E	Glu	129.116
Glycine	G	Gly	57.052
Histidine	H	His	137.141
Isoleucine	I	Ile	113.160
Leucine	L	Leu	113.160
Lysine	K	Lys	128.17
Methionine	M	Met	131.199
Phenylalanine	F	Phe	147.177
Proline	P	Pro	97.117
Serine	S	Ser	87.078
Threonine	T	Thr	101.105
Tryptophan	W	Trp	186.213
Tyrosine	Y	Tyr	163.176
Valine	V	Val	99.133
STOP	X	-	.

Table B.3: Names and abbrevations of the 20 standard amino acids