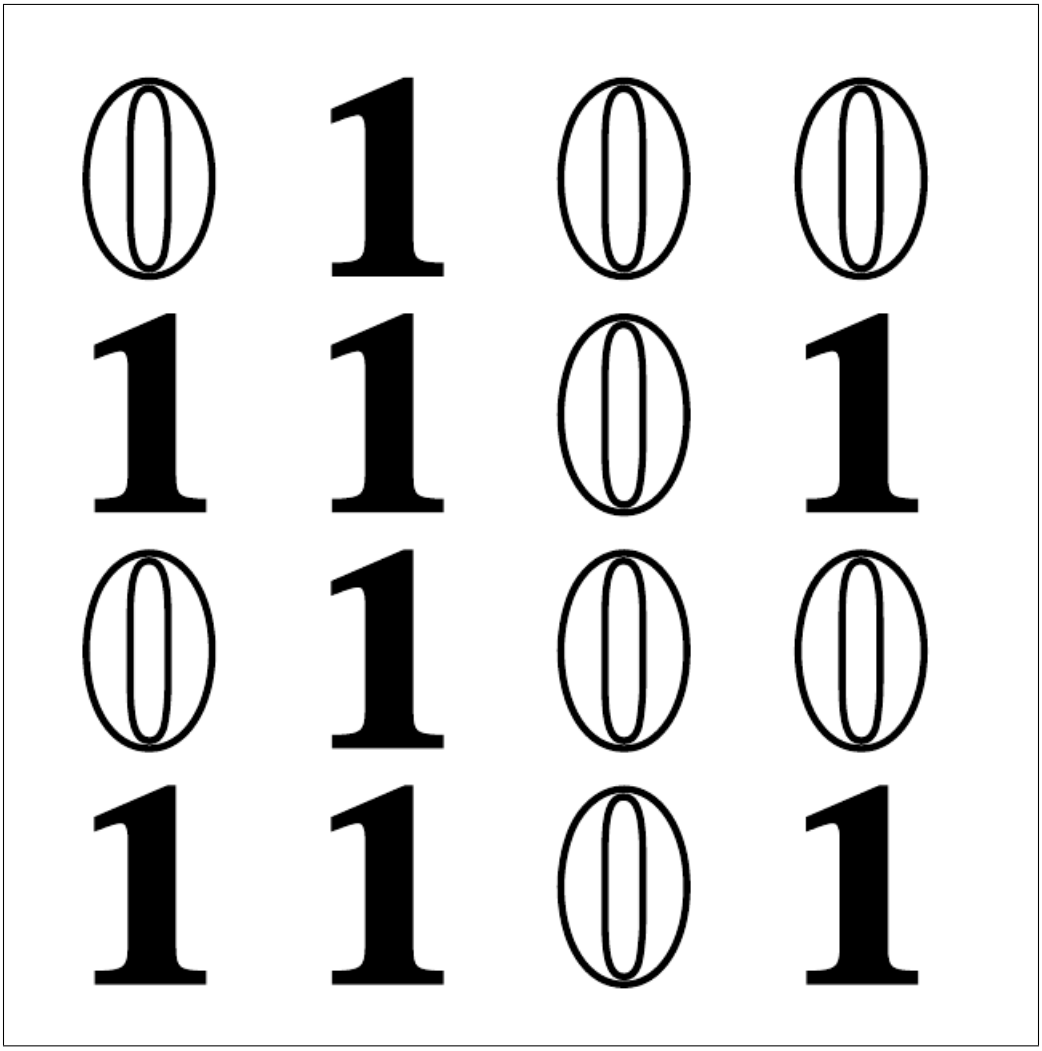
Considerations in evolutionary biochemistry

van der Gulik, P.T.S.

Considerations in Evolutionary Biochemistry



Peter T. S. van der Gulik

Considerations in Evolutionary Biochemistry

ILLC Dissertation Series DS-2019-06

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

These investigations were supported by Centrum Wiskunde & Informatica (CWI), Vici grant 639-023-302 from the Netherlands Organization for Scientific Research (NWO), and the QuSoft Research Center for Quantum Software.

Printed and bound by Ipskamp Drukkers.

ISBN: 978–94–028–1569–6

Considerations in Evolutionary Biochemistry

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 18 september 2019, te 13.00 uur

door

Petrus Theodorus Simon van der Gulik

geboren te Amsterdam.

Promotiecommissie

Promotores:	Prof. dr. H.M. Buhrman	Universiteit van Amsterdam
	Prof. dr. W.D. Hoff	Oklahoma State University
Copromotor:	Dr. D. Speijer	Universiteit van Amsterdam
Overige leden:	Prof. dr. M.A. Haring	Universiteit van Amsterdam
	Prof. dr. A.T. Groot	Universiteit van Amsterdam
	Prof. dr. L. Stougie	Vrije Universiteit Amsterdam
	Prof. dr. S.A. Massar	Université libre de Bruxelles
	Dr. C.J.M. Egas	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

dedicated to the memory of Christian de Duve

Contents

Acknowledgments

My debt to my promotor, Harry Buhrman, can hardly be exaggerated. It is very improbable that a 39-year old without a Ph.D. will get the opportunity to pursue his great scientific "queeste". I sincerely thank Harry for believing in me as a scientist, and creating the environment in which I could build my new life. Working at CWI is an exciting and rewarding experience. I vividly remember the excitement when we realized that our histograms, without the "peaks" and the "throughs" were the "real histograms", and the beautiful patterns in the published literature turned out to be artefacts. I have to admit that Harry was right when he said that suddenly the beauty of the complicated figure was gone, and it started to look awful. Since those early days we have published three articles in this field, and went through many meetings, discussions, and more casual conversations. Sometimes the thinking had to be hard and the work became very difficult. Without doubt, completing this effort was the most difficult thing I did in my life, up to now. Seeing the articles being accepted has been very rewarding. I am extremely thankful for understanding the SGC on a much deeper level than I did twenty years ago, and for functioning on a different level than I used to do. What I remember too, is the necessity to quantify things I was seeing, a necessity brought to my attention by Harry. Without this advice, I would not have thought of making the Molecular Structure Matrix we published in 2013. I am also very thankful to Harry for putting on the brake when I was close to irresponsibly neglect my RSI problems in my autistic way, and continue working towards finishing the thesis without taking summer holiday. I am happy that I followed his advice, and started late August with a fresh mind. As for my job, now, at CWI: I remember a conversation in the train from Brussels to Amsterdam, and realize I am a flower which can thrive in an environment created and maintained by Harry; and I am very thankful for that. Thank you very much, Harry!

Next, I want to thank my wonderful copromotor, Dave Speijer. Right from the start of my time at CWI, Dave was my biological reality-check support-pillar,

during the days when we established our AMC-CWI cooperation on sleeping sickness genetics. It was a huge relief for me when the "GCC" (Leen's "Genetic Code Club", Harry's "Bio-Club", Dave's "Codon Club", and my own "little RNA Tie Club") grew from five to six, and I could share the burden of being "guardian of biological reliability" (that's how I saw myself at A&C) with a fellow evolutionary biochemist. Dave was especially important during the first half of 2013, when I had trouble with writing.

I want to thank especially my second promotor: Wouter Hoff. When Harry and Dave suggested that I should write the article on nonsense suppression as a solo effort, I was faced by my inability to write in scientific style. I also had difficulty to remain in high spirit, when working all alone. I was very happy when Wouter accepted my invitation to be a co-author. Looking back to the trouble it gave us to get the message across, I am sure the article would not have been published if I had remained alone on this project. Wouter's viewpoint that, with a major expansion, the thing would be understandable and acceptable for Journal of Molecular Evolution, proved correct. My interaction with Wouter in the scientific field is not something young and recent: we were already discussing biochemistry and evolution during the eighties. I thank Wouter for keeping me on board of the ship of evolutionary biochemistry throughout the years.

I thank the members of the Doctoral Committee, Michel Haring, Astrid Groot, Leen Stougie, Serge Massar and Martijn Egas, for spending their precious time on investigation of our research. Serge and Leen are, of course, also co-authors! In the case of Leen, the work together going back all the way to my first days in the port-o-cabin (showing the *Alopochen aegyptiacus* to Harry on the computer screen...).

I thank my other co-authors, Dimitri Gilis, Steven Kelk, Gunnar Klau, Wouter Koolen, Marianne Rooman, Christian Schaffner, and Simone Severini, for taking part in this work. Science is something which you don't do alone, and I want to sincerely thank all my co-authors for working with me. Without them, this book would not have existed. I also thank Steven de Rooij for, although refraining from being a co-author, taking part in the "hunt for the minimum", which was the challenge ultimately leading to our article in TCBB.

I thank especially Maarten Dijkema, Dubravka Tepsic, Eefje Bosch, Iris Hesp, Susanne van Dam, Nada Mitrovic, Hans Hidskes and Silvia Benschop for their support work.

I thank all my colleagues from what was PNA6, used to be INS4, and is now A&C, (and is now also part of QuSoft!) in particular Farrokh, Álvaro, Tom, Koen, Joris, Sébastian, Freek, Yinan, Alex, Christian, Joran, Jan, Arjan, Bas, Yfke, Subhasree, Floor, Kareljan, Michael, Maris and Stacey. I want to thank especially Jop Briët, Fernando de Melo, Jeroen Zuiddam, Steven de Rooij, Christian Schaffner, Tim van Erven, Ronald de Wolf, Thijs van Ommen, Florian Speelman, Teresa Piovesan and Arie Matsliah for helping me with technical problems. Jop made

the beautiful cover illustration of this little book. I want to thank Florian for being part of the Bio-Club and for being my *steunpilaar* through the last months towards the defense, and I want to thank Paul Vitányi for sharing his office.
I thank Niels Nes, Michael Guravage, Erik Baquedano, Arjen de Rijke and again Maarten Dijkema and Dubravka Tepsic for expert direct ICT support. I thank the colleagues of the supporting departments of our beautiful institute: the Library (Lieke Schultze, Wouter Mettrop, Rob van Rooijen, Bikkie Aldeias and Vera Sarkol!), the Communication Department, the Personnel and Organization Department, the Secretaries, the Financial Department, the Valorization Department, the ITF Department, the janitors, and Minnie Middelberg. I thank the MT for making CWI run. I thank the colleagues of other research groups for realizing a happy working environment.
I thank the colleagues of the ILLC, especially Leen Torenvliet, Jenny Batson, Tanja Kassenaar, Marco Vervoort and Debbie Klaassen. In particular I thank the ILLC for the work on the ILLC dissertations software, and on the support towards the defense ceremony. And for making me feel an extramural part of ILLC.
Next I want to thank the people from the *directiesecretariaat* of the *Faculteit der Natuurwetenschappen, Wiskunde en Informatica* and the people from the *Bureau Pedel*, both (like the ILLC (but not the CWI!)) from the University of Amsterdam, for the correct and pleasant interaction. And I thank Jelle de Vries, Peter van Limbeek and the other people at Ipskamp Printing!
I thank the people who are involved in creating the special arrangement which makes me function in CWI despite my restrictions. Apart from Harry Buhrman, people who especially have to be mentioned in this regard are Léon Ouwerkerk from CWI's Personnel and Organization Department, Marlin van der Heijden from NWO, and Matthieu Wouters and Martin van Loenen from the Amsterdam municipality.
I thank the reviewers and editors of our articles for improving them by their comments.
I also want to thank my family, which nourishes me, and in particular: my mother, my father, my sisters, my brothers-in-law, my niece and my three nephews. And also my aunts, my uncles, my cousins, their spouses, and their children.
And I want to thank all my friends for being there, in particular Tineke Hoff-de Vries, and Marc Menon and Adelina Hasani.

Amsterdam Peter van der Gulik
July, 2019.

Chapter 1

Evolutionary Biochemistry

This chapter gives an introductory treatment of some interesting problems in evolutionary biochemistry which are amenable to a computational treatment. With "amenable to a computational treatment" I mean: the problems can be worked into a mathematical format, where meaningful computation can be done. In this thesis, the following interesting topics are considered: ancient peptide-coding sequence elements, structure of the genetic code, and linkage selection. We start with the topic of ancient peptide sequences, and thus turn to the origin of biochemistry.

Origin of biochemistry. In its ultimate goal, evolutionary biochemistry aims to understand the molecular events resulting in the current diversity of life on planet Earth, starting with the initial steps that gave rise to the origin of life. This goal faces a number of challenges. The first challenge is that the goal is to unravel biochemical events that occurred millions to billions of years ago. The Earth is approximately 4.5×10^9 years old, and life is known to have been present on Earth for at least 90% of this amount of time. The oldest signs of life are chemical footprints: organisms have a bias to use the lighter isotope of carbon during carbon fixation, and therefore the presence of life leads to isotope fractionation [Sch88]. Very old stones from Greenland carry evidence of such isotope fractionation. This biosignature is considered to be less vulnerable to misinterpretation than bacterial and archaeal fossils. Although rapid progress is made in the field of bacterial paleontology (cf. [OWD$^+$09]), microbial fossils are often hard to interpret. In many cases, there is even the possibility that the character as remains of microbes is not sure. When isotopic fractionation is considered, the age of the zircons which contain inclusions of light-carbon diamond is a controversy in the field, with some experts considering them having an age of about 3.85×10^9 years old [MAM$^+$96], while other experts maintain an age of about 3.65×10^9 years old [WK05]. Although the older age has been gaining credibility during recent years [MMH06], the whole debate is moving into obscurity with much older dates for fractionation of carbon in zircons from a different site of

origin, namely Australia [NWM+08] instead of Greenland. However, use of the light-carbon values as a unique biomarker remains controversial [NWM+08], because abiotic organic synthesis involving carbon oxides, methane, hydrogen and water could also produce this kind of values. Isotope fractionation values "should not be taken as prima-facie evidence for biological activity in the Hadean, although they do not exclude such a possibility" [NWM+08]. These new data give a staggering old age of about 4.25×10^9 years. If they do indeed derive from biological activity, I am on the safe side with the claim that life is known to have been present on Earth for at least 90% of the time that the planet exists.

Origin of amino acids. For the goal of reconstructing early steps in the origin of life, it is relevant to consider the core chemical constituents of cells. The constituents of life as we know it, are (reducing the list to its very core):

1. amino acids (the building blocks of proteins)
2. nucleotides (the building blocks of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA))
3. monosaccharides (the building blocks of sugars)
4. phospholipids (the building blocks of membranes)

Some of these constituents can be produced by rather simple chemical experiments. In this respect, the synthesis of amino acids by Stanley Miller during the 50's of the last century was a landmark achievement in the origin of life studies. The presence of amino acids, however, is not the same as the presence of life. First of all, amino acids have to be linked to large, polymeric molecules to perform the many tasks by proteins in living systems. Secondly, this polymerization has to proceed in a coded manner: a protein is a polypeptide of very specific sequence. In living systems, the information specifying these sequences is transmitted from one generation to the next in the form of nucleic acid (DNA or RNA). This highly organized system of molecular information is of course not produced during the Miller-type experiments. Thirdly, the nucleic acids and proteins are embedded in living cells, which protect them against harsh environmental circumstances, and harvest energy and building materials from the environment to keep the system functioning, making it grow, divide, and expand. The cells too, are not made during the Miller-type experiments.

Origin of proteins. A central theme in evolutionary biology is that complex phenomena have simple beginnings, and become more elaborate in a step-by-step manner. With respect to proteins, a start of the evolution of life with a situation in which simpler proteins are functioning is in line with this kind of reasoning. Several researchers have suggested that primordial proteins did not consist of twenty different kinds of amino acid, but far less. One of the proposals is that primordial proteins consisted of valine, alanine, aspartic acid, and glycine. These

amino acids are relatively small, and form a diverse set in terms of amino acid characteristics. Valine is very hydrophobic, while aspartic acid is very hydrophilic. Glycine is extremely small, and at glycine residues a protein chain can make turns not possible with other amino acid residues. Alanine is an amino acid with intermediate characteristics: larger than glycine, but smaller than valine and aspartic acid; less hydrophobic than valine, but still hydrophobic when compared to aspartic acid. With these four amino acids, many of the basic themes in protein structure should already be attainable.

The first problem which is investigated in this thesis is the question if present day proteins still contain sequence elements which are directly derived from the times when proteins possibly consisted of just valine, alanine, aspartic acid, and glycine. A search was performed for parts of contemporaneous proteins which reflect very old motifs, consisting of just the four mentioned amino acids[1]. If short enough, one can hypothesize that simple chemical processes generated such peptides: addition of clays, surfaces, metals, and cyclic environmental circumstances (e.g. hot/cold, or dry/wet) to the Miller-type experiments result in the generation of small peptides. In section 1.1, more background is given regarding our quest considering these earliest times of life: developments on the border between "just" chemistry and life.

RNA world and genetic code The molecules of heredity, DNA and RNA, were originally seen as chemically relatively inactive. The discovery of catalytically active RNA was therefore a landmark development in biochemistry. Several people, among which Alex Rich, Carl Woese, Francis Crick, and Lesley Orgel, suggested already in the 60's of the last century that it might be possible that RNA could have catalytic potential (see [BKC12] for more background on this issue). In the early 80's this was experimentally found to be true, by the research groups of Cech [KGZ$^+$82] and of Altman [GTGM$^+$83]. As a result of the realization that RNA can function both as a nucleotide sequence specifying a protein, and as a catalytically active molecule on its own, the 'RNA world' hypothesis [Gil86] was proposed. In this hypothesis, RNA was both the genetic material and the catalytic agent in a stage of life without proteins. Two problems plague this hypothesis: the difficulty of prebiotic synthesis of nucleotides and the vulnerability of RNA molecules to hydrolysis [Fis11]. Because of the instability of RNA molecules, RNA-catalysts performing RNA replication had to be very efficient, which requires large and sophisticated RNA molecules. Due to the imprecise character of replication in such an all-RNA system, the emergence of such

[1] A very interesting comparable search was performed by Sobolevsky, Frenkel, and Trifonov [SFT07]; they searched for motifs, 6 to 9 residues long which are omnipresent in the genomes of 15 fully sequenced, non-eukaryotic cellular organisms. Next to their Group Aleph (among which the Walker A motif) and Group Beth sequences, they found five other sequences: FIDEID, IDTPGHV, KMSKSL, NADFDGD, and WTTTPWT. These are all components of "central" proteins, like aminoacyl-tRNA synthetases or elongation factors; NADFDGD was also found by our study.

long, early RNA molecules seems improbable (although this is not a generally held view, see e.g. [BKC12]). In all known biology, the information specifying proteins is present as nucleic acid sequence, and, concomitantly, nucleic acid replication is performed by protein enzymes. The continuity principle might suggest that it has always been like that, and that collaboration between two kinds of catalytic biomolecules, oligopeptides and oligonucleotides, was present from the start [Fis11] (see also [Fra11, LFCJ13, CJ15, MREJR$^+$15, Wil12, BHW15]). The very short peptides generated abiotically would be responsible for enhancing RNA replication, and short RNA molecules would perform different biochemical activities. The crucial next step in the evolution of life would then be the acquisition of the power of coded synthesis of crucial peptide sequences by RNA. The establishment of a fixed assignment of short sequences of nucleotides with specific amino acids gave birth to the genetic code. Some researchers adhere to the concept that (at least part of) the genetic code is even older than coded peptides, and provided the RNA world with a "hold" on individual amino acids as prosthetic groups (the Coding-Coenzyme-Handles-hypothesis [Sza93]). An alternative view is that the very first step of the genetic code came about when a single kind of transfer RNA (tRNA) allowed homopolymerization of the first amino acid in a coded peptide (which was therefore the polymerized form of a single amino acid). At present, it is not yet possible to discard one of these two alternatives definitively: contrary to the study of present-day biochemistry, evolutionary biochemistry is still a field with many unknowns. It needs to be stressed that this state of affairs is rapidly changing: the DNA sequencing revolution has happened, and is still ongoing, spawning the science of genomics, and this is providing an exceptionally rich source of raw data for the field of evolutionary biochemistry (see also [HT13]).

Nowadays, the genetic code is large and complex: 64 sequences of three nucleotides (the sequences known as codons) specify 21 outputs (20 amino acids and the signal "stop"). It is reasonable to envisage the early genetic code as much simpler: coding for less amino acids. One of the earliest developments in the history of life is then the development of the modern "Standard Genetic Code" (SGC) from that simple early code. Why during the evolution of the SGC the particular 20 amino acids were incorporated that are now the canonical set of twenty, is an intriguing question (see [PF11]). Another fascinating question is whether the amino acids already played a role in biochemistry when they were recruited to the repertoire. Six small compounds with an adenosine part play a very important role in biochemistry: adenosine triphosphate (ATP) as an energy carrier, cyclic adenosine monophosphate (cAMP) as a messenger molecule, nicotinamide adenine dinucleotide (NAD^+) and flavin adenine dinucleotide (FAD) as redox carriers, S-adenosylmethionine (SAM) as a methyl carrier, and coenzyme A (CoA) as a carrier of many different small organic molecules. The two sulfur-containing amino acids, methionine and cysteine, are components of respectively SAM and CoA. It is an interesting question if these sulfur-containing amino acids

where part of metabolism and were then recruited to the repertoire of amino acids used in the SGC, or if it was the other way round, and that they were already protein components, and SAM and CoA did only take their place in universal biochemistry after the SGC was already complete[2]. These considerations about the growth of the repertoire of amino acids used in proteins bring us to the second major problem investigated in this thesis: the structure of the SGC. An introduction to this topic is given in section 1.2.

In evolutionary biology, one often encounters situations where a large amount of innovation takes place during a relatively short period of time. Often this is followed by long periods with evolution occuring within the boundaries of a fixed set of stable "settings". For example, after the basic aspects of metabolism were introduced, and the SGC was in place, the rest of evolution could be considered "more of the same". From a "macroscopic" viewpoint, the appearance of eukaryotic cells, of animals, of multicellular plants, and of human beings may look like big innovations, but from the viewpoint of evolutionary biochemistry, they are just variations on the theme "the cell". One could say that nothing major happened during the last three billion years, as far as the evolutionary biochemist is concerned.

This being said, and in this way contrasting chapter 6 with the preceding chapters, it remains a fact that interesting problems in evolutionary biochemistry can be found which do not relate to such basic and extremely ancient issues as the origin of metabolism or the origin of the genetic code. One of these is the remarkable genetic organization of the genome of the mitochondrion of the parasite causing sleeping sickness. An introduction to this topic is given in section 1.3.

In fact, an overarching theme of this thesis concerns "genetical errors". Allowing the environment to create useless peptides can be seen as a kind of genetical error. "Installing" replication, transcription, and translation is the strategy for not having that error, and for directing the presence of environmentally occurring amino acids into life-serving peptide activity. The genetic code is central in this strategy. The special structure of this set of codon-amino acid assignments has error-robust properties. It ensures that many "genetical errors" (in the sense of substitution mutations) have no adverse effect, even in the context of a primitive and highly vulnerable system (which lacks, for example, a DNA repair mechanism). Maybe complex life could have never gotten off the ground were it not for having these error-robust properties [Woe65a]. Finally, linkage selection (the topic of chapter 6) can be seen as a mechanism preventing another kind of "genetical error": throwing away information which is needed later on in the life cycle.

[2]Please note that these considerations about co-factors do not imply that the amino acids which are part of the cofactors are ancient compared to other members of the Set of Twenty; the relative order of appearance of the amino acids is a related, but different, issue.

In the remainder of this chapter, an overall introduction will be provided to the three main areas of evolutionary biochemistry considered in this thesis: first the search for the the first peptides at the origin of life; then the evolution and error-robustness of the genetic code; and, finally, linkage selection in trypanosomes.

1.1 The first peptides

In the approach followed here, we assume that the earliest peptides were composed of amino acids that are readily formed by abiotic chemistry. Therefore, experiments on conditions aimed to mimic chemistry under prebiotic conditions are relevant here. Amino acids are easily produced in a lot of abiotic settings. Higgs and Pudritz [HP09] emphasize that the ten proteinaceous amino acids which are seen in a diverse collection of amino acid abundance measurements (from Miller-type synthesis experiments, meteorite extraction, hydrothermal vent synthesis simulating experiments and several other chemical synthesis experiments) are surprisingly consistent. They also show a strong correlation of the relative abundance of these ten amino acids in these measurements with their free energy of formation in seawater. They conclude that thermodynamics predicts which amino acids are formed most easily, and that this probably sets the prebiotic amino acid mixture which is universally available everywhere in the Universe where circumstances allowing life to originate are present. In the view presented by Higgs and Pudritz alanine and glycine; threonine and serine; valine, leucine, and isoleucine; aspartic acid and glutamic acid; and proline; will be constituents of life everywhere. Which additional amino acids will become part of the repertoire would depend on the idiosyncracies of the particular development of metabolism (coevolution theory of genetic code: [Won75]) taking place at a certain location of origin. It should be mentioned that prebiotic production of some of the ten other proteinaceous amino acids is not entirely impossible. For example, lysine was found to be produced in experiments by Rode and co-workers [PRR06]. One of the differences between the approach of Miller and that of Rode and co-workers is that Miller studied prebiotic production in a simulated atmosphere while Rode studied prebiotic production in a simulated hot, salty ocean. This brings us to the issue of prebiotic locations. Which locations for the origin of life are considered a possibility on planet Earth?

A few environments are currently seen as interesting candidates for the cradle of life. Benner and co-workers think of a *desert valley*, with influx of rivers from borate-containing mountains [BKC12] ("a subaerial intermountain desert valley"; "serpentinizing rocks weathering with igneous borates, a CO_2 atmosphere, and rain containing abundant prebiotic HCHO and catalytic glycolaldehyde"). Mulkidjanian and colleagues speculate about *terrestrial*, anoxic, zinc-containing *geothermal fields* [MBD$^+$12] ("... shallow ponds of condensed and cooled geothermal vapor that were lined with porous silicate minerals mixed with metal sulfides

and enriched in K^+, Zn^{2+}, and phosphorous compounds"). Martin and co-workers consider *submarine*, alkaline, *hydrothermal vents* interfacing with ocean water (the vents having a non-volcanic origin: "Serpentinization occurs when rocks derived from the upper mantle (rich in olivine) are exposed to ocean water") [LM12]. Serpentinization is a geological, exothermic process in which large amounts of water are absorbed by certain rock species; these are oxidized and hydrolysed and new rock species are formed in the process, as well as hydrogen gas[3].

One can see that there is no consensus regarding life's environment of origin (and I have omitted the idea of panspermia: life arriving on early Earth from outer space, see e.g. [Cri88]). It is also a possibility that different environments generated different products, and the flux between environments brought the necessary pieces together. In this regard, Saladino and co-workers point out [SNC$^+$10] that an environment in which *formamide* replaces water as the medium (formamide has a much higher boiling point and could "be easily concentrated by simple water evaporation in lagoons and on drying beaches" [SNC$^+$10]) and in which zirconium minerals (occurring "almost everywhere ancient sediments are present" [SNC$^+$10]) are playing a catalytic role, leads to the synthesis of nucleobases (necessary for the emergence of genetics) and carboxylic acid derivatives (necessary for the emergence of metabolism), but is a destructive environment for RNA. Both the lagoons and the beaches are, as far as the formamide-based chemistry is concerned, of an ephemeral character: tide and rain can switch the system back to be a water-based one. When nucleobases, produced in environments as those envisioned by Saladino and co-workers, are exported to environments where *membrane vesicles* grow, according to the processes studied by the Szostak lab, the selectivity of membrane passage favoring ribose as compared to other sugars[SS05] could lead to RNA formation inside vesicles, as proposed by Szostak and co-workers[CS04, CRS04, MS08, MSK$^+$08, RIA$^+$10, BS11, ZZS12, ZAZS12]. With respect to the environment in which these vesicles could emerge, Szostak writes [Szo12a]: "A geothermally active region of the early earth that was generally cold could contain numerous lakes and ponds, similar to Yellowstone lake in the USA, and many other environments on the modern earth, in which hydrothermal vents release plumes of hot water into cold lake water [referring to: [MSL$^+$03]]. In such an environment, protocells would exist at low temperatures most of the time, during which template copying could occur, punctuated by short intervals at high temperature, leading to strand separation and an influx of nutrients such as nucleotides. Endorheic lakes or ponds could accumulate organic compounds to high levels, especially in geothermally active regions where fatty acids and related compounds might be synthesized by Fischer-Tropsch type chemistry, and high

[3]In the presence of carbon dioxide, methane may be produced by serpentinization. The point about serpentinization in the account of Benner et al. [BKC12] is that the reducing power and alkaline environment generated are necessary preconditions for the formose reaction to happen. At submarine alkaline hydrothermal vents, proton gradients form naturally[LAM10] and are seen by Martin and co-workers as central to the origin of life.

energy carbon-nitrogen compounds could be synthesized as a result of electrical discharges surrounding active volcanoes. Sulfurous exhalations such as COS and H_2S could be important for the synthesis of thioesters or N-carboxyanhydrides for re-activation chemistry, and for the synthesis of modified nucleosides such as 2-thio-U for improved rate and fidelity of RNA replication."

At this stage it is not possible to make a definite choice between the different options mentioned above (or other candidates which I did not highlight). We concentrate on one of the *processes* inherent to life instead. This is the proces of *polymerization*. As stated at the beginning of this chapter, the presence of amino acids is not the same as the presence of life. Life is about the production of meaningful coded polypeptides (cf. [Szo12b]). The appearance of the first coded peptides in this particular development is an enigma. Production of small, non-coded oligopeptides (e.g. dipeptides and tripeptides) in an abiotic setting is readily reproduced in the lab. Evaporation cycle experiments have shown [SLER93] that peptide bonds between single amino acids are formed under high salt concentrations as might be expected to appear in evaporation pools on the prebiotic beaches. In this Salt-Induced Peptide Formation (SIPF) reaction the hydration shells of the Na^+ ions are not completely filled, and they can be considered strong dehydrating agents. NaCl in high concentrations therefore fulfills the role of a condensation reagent. The presence of $CuCl_2$ was also essential in these experiments as a copper ion is the organizing center of the catalytic complex [RFJ07]. Clay minerals [RSSB99] and glycine and histidine [LFFR10] have additional catalytic effects in this kind of reaction. Dipeptides and tripeptides of different compositions can thus be expected to form in certain environments. The enigma is how coded peptide synthesis started in early biochemistry.

Possibly, very short peptides, just a few residues in length, could have crucial biological properties. Apart from catalytic activities, one can also think of surprising other crucial properties, e.g. being lipids ("By definition, *lipids are water-insoluble biomolecules that are highly soluble in organic solvents such as chloroform*" [BTS07a]). Zhang has found that oligopeptides like AcVVVVVVD (in which "Ac" is standing for "Acetyl", and V and D are the one-letter abbreviations of valine and aspartic acid) behave like lipids, and organize themselves in membranes [Zha12]. Other ideas about early functions of coded peptides are: RNA chaperone ("...short, possibly positively charged, chaperone-like peptides in the RNA world would increase stability and help maintain ribozyme tertiary structure" [PJP98]), enlarging the structural repertoire of RNA by binding the RNA and enforcing shapes it cannot make by itself [Nol04], a protecting function of diphenylalanine (stabilization of dinucleotides as a result of stacking interactions with FF (which is immensely thermostable, F being the one-letter abbreviation of phenylalanine) was experimentally demonstrated [CG11]), and an RNase activity [Bra08] of LKLKLKLKLK (such peptides could be excreted and break down RNA sequences of 'competitors', making the nucleosides available for own

RNA synthesis, L and K being the one-letter abbreviations of leucine and lysine). The original coded oligopeptides could also have functioned as storage oligomers. This last function allows a lot of freedom to the sequence: the variation can be used to incorporate different ratios of carbon and oxygen, and to make the storage oligomer fold in a convenient way. Another possible function of very specific peptides is the amino-acylating activity which VD and AD (A being the one-letter abbreviation of alanine) have, according to Shimizu [Shi95]. Regarding an original catalytic function, Szostak [Szo12a] points to the possibility that short random peptides rich in amino acids with acid side chains could provide the opportunity to have RNA polymerization happening at comparatively low Mg^{2+} concentrations: the peptides could bring the metal ions at the precise location needed for RNA synthesis. Concomittantly, the metal ions would be kept from destructive action against synthesized RNA. In line with this kind of ideas, the focus in chapter 2 is on catalytic activity as the first function of coded oligopeptides, but, as this short overview shows, there are many more options regarding possible first functions. At this moment, there is insufficient evidence to make a choice between them. Because lab experiments in this area are very difficult, computer simulations using e.g. molecular dynamics are currently an attractive way forward to acquire more insight into this area of interest.

1.2 The genetic code

One of the difficulties Charles Darwin encountered when he proposed the theory of evolution of the many biological species (including humans) by natural selection, was that the mechanisms by which heredity works were unknown. Since then, we made immense progress in expanding our understanding of *genetics*. One of the major steps forward was the proposal of the "One gene-one enzyme hypothesis", by Beadle and Tatum [BT41]. The new concept was that everything in the cell is governed by chemical reactions, that every individual chemical reaction is steered by an individual *enzyme*, and that every enzyme is the expression of an individual *gene*. Although we now know that the cell is an enormously complex unit of organization and there is a high degree of cross-contacts fine-tuning what reactions occur, overall the "One gene-one enzyme hypothesis" still stands, and has taken concrete form in the finding that most enzymes are proteins. The definition of enzymes has even been changed, such that now, enzymes are a *kind* of proteins. When other biomolecules are found to have enzymatic properties, we now need a new name, which has happened with ribozymes (so enzymological studies on ribozymes should be called *ribozymological* studies). The Avery-MacLeod-McCarty experiment [AMM44] was the most prominent one of a series of lab experiments which ultimately led to the view that genes consist of DNA. The "One gene-one enzyme hypothesis" thus naturally developed into the Central Dogma of Molecular Biology: "DNA makes RNA makes protein". Genes

are located on chromosomes, and are made of DNA. The genetic information is *transcribed* into RNA. The ribosome (a particle in the cell, which can be seen with a microscope) then *translates* the RNA message into protein. The structure of DNA turned out to be a double helix, as discovered due to the efforts of Franklin, Watson and Crick, and Wilkins (cf. [Olb94]). Therefore, both proteins and nucleic acids emerged to be linear molecules, which fold into three-dimensional (3D) shapes programmed by their sequence of building blocks. The eventual 3D shape (vast amounts of different individual protein forms in the case of proteins, and the double helix in the case of DNA) is associated with the biological function.

Originally, it was thought that every kind of protein had its own kind of ribosome, to produce that kind of protein. The information coming from the DNA and going to the protein would thus reside *in* the ribosome, and, to be more precise, in the *RNA component* of the ribosome (the ribosome consists of both ribosomal RNA (rRNA) and ribosomal protein). At a certain moment during the development of early molecular biology, it was realized that *another* kind of RNA carries the information. This RNA was called messenger RNA (mRNA), and it is much more ephemeral than rRNA, which is why it was missed originally. The story of the relative contributions of Brenner and Crick, of Jacob and Monod, of Watson, and of Volkin and Astrachan in the discovery of mRNA is vividly presented in [Bre01b]. The conclusive experiments demonstrating the existence of mRNA were published in 1961 [BJM61]. Summarizing: The linearly organized information of the DNA gene is thus transcribed into an (in principle linear, despite 3D peculiarities) mRNA molecule which is translated by the ribosome into a linear protein molecule with self-folding capacity.

When it had become clear that the information specifying proteins resided in DNA genes, Gamov organized the RNA Tie Club of scientists around Watson and Crick, and this group of researchers focused on the obvious problem to solve: how is the protein sequence coded in a DNA sequence? Firstly, Brenner [Bre57] performed a theoretical *tour-de-force*, in which he showed that the facts available in 1957 implied that each amino acid in a protein was coded by a separate short stretch of DNA (in jargon: the code is *non-overlapping*), coding for that specific residue in the protein. These short stretches of DNA were subsequently referred to as *codons* (cf. [Bre01a]). Secondly, Crick, Brenner, Barnett, and Watts-Tobin published "General nature of the genetic code for proteins" [CBBWT61], in which they presented experiments showing that proteins are coded in units of three DNA nucleotides. In this paper the name "genetic code" was coined, referring to the rules according to which nucleic acid sequences are translated into protein. The actual assignments of the 64 (4^3) codons were then found, not by the more theoretical approach of the RNA Tie Club (although the work of Barnett demonstrating the triplets was of course very practical), but by the more practical approach of e.g. feeding poly-U into an *in vitro* system and finding out that this leads to the production of polyphenylalanine (and therefore, UUU means Phe). Practical work along these lines was done by the groups of Nirenberg, Ochoa and

Khorana, and in 1966 the genetic code was completely known. Even before the years when the molecular biological community was frantically working to decipher the codon assignments, Crick realized that, between the genetic sequence and the amino acid an adaptor consisting of RNA had to exist. This was called transfer RNA (tRNA), and this class of molecules was found by the group of Zamecnik [HSS+58]. The part of the tRNA interacting with the codon is called the *anticodon*, and it is this stretch of nucleic acid which is responsible for implementing the coupling of certain amino acids to certain codons: the genetic code.

Similar codons code for similar amino acids. Already in articles in 1963 of the groups of Ochoa [SLB+63] and Nirenberg [NJL+63], the facts that the genetic code is not only degenerate, but that this degeneracy is taking the practical form of groups of similar codons coding for the same amino acid, and groups of these groups coding for kinds of amino acids which are similar, were reported. The order present in the genetic code assignments was subsequently highlighted by Woese, in a a series of articles in PNAS [Woe65b, Woe65a, WDSD66]. In particular, Woese pointed to the hydrophobic nature of all amino acids coded for by codons with uracil as the middle nucleotide, and to the moderate nature (in the sense of not being particularly hydrophobic nor hydrophilic) of all amino acids coded for by codons with cytosine as the middle nucleotide. Because Woese and co-workers developed an experimental scale characterizing the hydrophobicity of the amino acids (by measuring the chromatographic behaviour of the amino acids using pyridine derivatives as solvents), it was thus possible to support the claim of order with quantitative data [WDD+66]. The findings were criticized by Crick [Cri68]. Crick wondered if the patterns seen in the genetic code were random patterns, experienced by investigators as something meaningful, but generated during history in a random way. The idea being that the human mind is inclined to see patterns, even if only randomness is present. As is pointed out in subsection 3.3.3, Crick did *not* doubt the presence of similar codons coding for similar amino acids, because he expected the code to evolve by variation on more simple precursors. So, "accidental" did not mean: no order at all (and we are ignoring the similar codons coding for identical amino acids aspect here, which of course was very well known to Crick [Cri66], this was the observation leading to his formulation of the wobble rules). However, the idea that something such as all codons containing a middle U coding for hydrophobic amino acids would indeed be a special pattern (special in the sense of requiring an evolutionary biochemical explanation different from frozen accident) was something Crick wanted to have demonstrated much more explicitly before he would accept it. Having quantitative data like " 5.0, 4.9, 4.9 again, 5.3, and 5.6 have middle-U, and 7.5, 6.6, 6.6 again, and 7.0 have middle-C " on a novel, somewhat arbitrary scale was not enough. In 1991, Haig and Hurst [HH91] contributed a more explicit demonstration of the order Woese claimed. They used a function developed by

Di Giulio [DG89a] to characterize a genetic code in terms of being able to not have many large changes in hydrophobicity at substitution mutations, despite having amino acid changes (e.g. moving through the 16 middle-U codons, and changing e.g. Leu into Val), and, next, they produced randomly a large collection of genetic code variants. They showed that indeed a pattern is present: only one (cf. [HH99]) in every 10000 codes resulting from random permutation of amino acid assignments gave a lower value with their error value function. They also showed that this error robustness is mainly provided by the first and third position of the codon, which was illustrated beautifully with histograms by Freeland and Hurst in 1998 [FH98a]. In chapter 3, some mathematical refinements are added to this field of research. Firstly, the global minimum for error robustness is found, and found to be identical to a very low value already known in the field [Gol93]. Secondly, the error function is refined, and as a result is able to incorporate stop codons in the calculation. In this way, genetic code variants with reassignments involving stop codons can now be compared with the standard code. Thirdly, the space of random code variants is progressively enlarged. Computations then show the standard code to be special compared to code variants resulting from random permutation of amino acid reassignments also in the two larger spaces which can be investigated using our refined error function.

UUY Phe	UCY Ser	UAY Tyr	UGY Cys
UUA Leu	UCA Ser	UAA Ter	UGA Ter
UUG Leu	UCG Ser	UAG Ter	UGG Trp
CUY Leu	CCY Pro	CAY His	CGY Arg
CUA Leu	CCA Pro	CAA Gln	CGA Arg
CUG Leu	CCG Pro	CAG Gln	CGG Arg
AUY Ile	ACY Thr	AAY Asn	AGY Ser
AUA Ile	ACA Thr	AAA Lys	AGA Arg
AUG Met	ACG Thr	AAG Lys	AGG Arg
GUY Val	GCY Ala	GAY Asp	GGY Gly
GUA Val	GCA Ala	GAA Glu	GGA Gly
GUG Val	GCG Ala	GAG Glu	GGG Gly

Table 1.1: The standard genetic code represented as a grid of 48 entries. "Ter" indicates "Termination"; "Y" represents "pyrimidine".

Similar codons code for the same amino acid. In chapter 4, a different aspect of the amino acid assignments is investigated. There is not only error robustness with respect to similar codons coding for similar amino acids, there is also error robustness with respect to similar codons coding for *identical* amino acids. Crick suggested that part of this robustness is caused by *one* tRNA molecule recognizing *two* codons [Cri66]. A rule without exception is that when two codons differ only in the third position, and one of these codons has U while

the other one has C as the third nucleotide, they encode the same amino acid. This implies that the table showing the genetic code, which normally is presented with 64 entries, can also be presented as a table with only 48 entries (see table 1.1). This is the first of Crick's Wobble Rules: G-starting anticodons do not only read their cognate C-ending codon, but also the U-ending codon.

This line of reasoning (the concept that "neighbouring" codons coding for the same amino acid may be due to *one* molecule recognizing *several* codons) can be extended to the four-codon groups known as family boxes [LJ88]. It is regrettable that in the more recent literature the term family box is used in different ways. This term was formally introduced in 1988, and referred to the groups of four codons sharing the same first two nucleotides and coding for the same amino acid. In the standard representation of the genetic code eight such family boxes are present (please note: ***exactly*** half of the 64 codons are organized as family boxes): all codons coding for Gly, Ala, Val, Thr and Pro are present as a family box, while 2/3 of the codons (four of the six) coding for Leu, Ser and Arg are present as a family box. Many investigators have indeed argued in favour of "four-codon-wobbling" (see below) for the cases of specialized groups of genetical systems, like mitochondria and the *Mycoplasma* bacteria. In this connection, the term "four-way-wobble" was introduced by Osawa and colleagues [OJWM92], the term "hyperwobble" by Kurland [Kur92], and the term "superwobble" by Vernon and co-workers [VGC$^+$01]. Recently however, work in genomics by Higgs and co-workers showed this wobbling behaviour in the family boxes to be found in *many* bacteria [RH10]. In the same way as the orthodox 64 entries are reduced to 48 entries in table 1.1, the number of entries can be further reduced to 32 by representing each family box by just one entry, as has been done in table 1.2.

UUY Phe UUA Leu UUG Leu	UCN Ser	UAY Tyr UAA Ter UAG Ter	UGY Cys UGA Ter UGG Trp
CUN Leu	CCN Pro	CAY His CAA Gln CAG Gln	CGN Arg
AUY Ile AUA Ile AUG Met	ACN Thr	AAY Asn AAA Lys AAG Lys	AGY Ser AGA Arg AGG Arg
GUN Val	GCN Ala	GAY Asp GAA Glu GAG Glu	GGN Gly

Table 1.2: The standard genetic code represented as a grid of 32 entries. "Ter" indicates "Termination"; "Y" represents "pyrimidine". "N" represents "any of the four nucleotides".

Molecular biologists will immediately notice that all four codon boxes in which

both the first and the second nucleotide of the codon are either G or C, are family boxes. Base-pairing between G and C involves *three* hydrogen bonds, and because of that, "S" ("Strong") is used for writing "either G or C". Base-pairing between A and U involves *only two* hydrogen bonds, and so "W" ("Weak") is the convention for writing "either A or U". In table 1.3, S and W are used in the appropiate places to highlight bonding strength.

WWY Phe WWA Leu WWG Leu	WSN Ser	WWY Tyr WWA Ter WWG Ter	WSY Cys WSA Ter WSG Trp
SWN Leu	SSN Pro	SWY His SWA Gln SWG Gln	SSN Arg
WWY Ile WWA Ile WWG Met	WSN Thr	WWY Asn WWA Lys WWG Lys	WSY Ser WSA Arg WSG Arg
SWN Val	SSN Ala	SWY Asp SWA Glu SWG Glu	SSN Gly

Table 1.3: The standard genetic code represented as a grid of 32 entries. "Ter" indicates "Termination"; "Y" represents " pyrimidine"; "N" represents "any of the four nucleotides"; "S" represents "Strong" and "W" represents "Weak".

As can be clearly seen from tables 1.2 and 1.3, all 16 codons starting Strong-Strong are family boxes, while all 16 codons starting Weak-Weak are split boxes. Moreover, **the remaining 32 codons are also splitting out in two neatly separated groups of equal size: the 4 family boxes from this group are found in the *left* side of the table, while the 4 split boxes from this group are found in the *right* side.** Translated back to the actual nucleotides, this means that when a codon from this group has a middle *pyrimidine*, it belongs to a family box. The other way round: when it has a middle *purine* (an A or G), it belongs to a split box. In summary: the 64 codons are neatly splitting in four groups of 16 codons each, and two of these groups are *only* family boxes, while the other two groups are *only* split boxes (please note the **symmetry** in the table: a **black F on the right,** (consisting of the Tyr/Ter, Cys/Ter/Trp, His/Gln, Asn/Lys, Ser/Arg, and Asp/Glu split boxes) back-to-back with an **upside-down white F on the left** (consisting of the Ala, Val, Thr, Pro, Leu, and Ser family boxes, see table 1.3)). The strength of the hydrogen bonding for Strong-Strong-starting codons points to what can be behind this regularity: in the case of a family box, originally a **single** tRNA molecule could efficiently decode all four codons of that family box. Because of this, diversification of meaning within such a box was not an option: in a primitive genetic system, where the genome was very small, the total number of tRNA genes was limited and **one gene was enough**

to handle a whole family box. In the case of the split boxes, one tRNA could *not* read easily all 4 codons sharing the same first two nucleotides. **Two tRNA genes were necessary to deal with the split boxes, and diversification could and did happen.**

The first one who reported this pattern was, to the best of my knowledge, Rumer, in 1966 [Rum66]. The report was in Russian, and the molecular biology community missed the point. Next, Lagerkvist [Lag78] drew attention to the issue, in English. Because he proposed an incorrect mechanism for this pattern (hypothesizing that in family boxes the third base of the codon and the first base of the anticodon were not making contact), his work was doubted, and the pattern was put aside as being random. Here, criticism was clearly too harsh. Due to the theoretical work of Lehmann and Libchaber [LL08], the genomics work of Ran and Higgs [RH10], the biochemical work of Rogalski and co-workers [RKB08], and the molecular dynamics work of Agris and co-workers [VMA09] (showing the bridging water molecule(s) in between two unmodified pyrimidines in their figures 2B(c), 2B(d), and 4G), this basic aspect of the structure of the genetic code was re-established.

The rule that all pyrimidine-ending codons come in pairs, reported by Crick [Cri66], and the regularity of the 8 family boxes [Rum66, Lag78, LL08] are comparatively easy to see. In chapter 4, a third regularity concerning the groups of codons coding for *identical* amino acids is reported: an amino acid is never coded by a single, A-ending codon. The three regularities should be seen in the context of the wobble abilities of unmodified anticodons: G-starting anticodons can read both Y-ending codons; in the 8 family boxes, U-starting anticodons can read all 4 codons; and C-starting anticodons don't wobble.

Fixed assignments. One of the important points of the research presented in this thesis, is the realization that *if* certain amino acid assignments are not free to change because they are *fixed* by chemical rules, *then* these amino acid assignments should also not be allowed to change during the type of mathematical investigations as performed in chapter 3. This led to another approach to randomly redistributing assignments in the calculations which were performed. The most prominent consequence of this modified procedure is that the spaces of codes were no longer billions of codes in size, but only thousands of codes in size. From the viewpoint of computer science, this means that the error values of all codes can be easily calculated, and we are no longer restricted to work with approximations (for the average error value of a group of variants of the genetic code), but we can work with the exact values.

A second change was also implemented in our calculations. This change followed the approach of Freeland and Hurst in their second 1998 article [FH98b], in which they modelled a gradual growth of the amino acid repertoire, starting with amino acids like valine, alanine, aspartic acid, and glycine, and with a few steps adding larger and larger amino acids, until molecules of the size of tyro-

sine and tryptophan became part of the repertoire. The combination of both changes in the model (in addition to an update with respect to the polar requirement [MLS08]) led to a model which we described as being *realistic*. In the resulting space, *the SGC was the optimum.*

A final aspect of error robustness present in the genetic code concerns similarities in molecular structure instead of hydrophobicity. To be able to quickly compare molecular structure of amino acids, a set of values was established which reflect the similarity in structure between amino acids. Using these values, the molecular structure, an aspect of amino acids which has not been investigated computationally before, is now presented in a form allowing computations. Our calculations show that this aspect leads to only slightly less error robustness of the SGC than the polar requirement. For this aspect, the error robustness resides mainly in the *second* position of the codon; the third codon position carries no robustness at all, with respect to this aspect. In chapter 5, the realistic model is presented, and also the contrast between the codon positions responsible for error robustness when hydrophobicity characteristics are compared with molecular structure characteristics. The structure of the genetic code for this last aspect possibly reflects the gradual development of the repertoire of amino acids, starting with small amino acids encoded by G-starting codons, developing with the addition of aspartate-derived amino acids encoded by A-starting codons and, later, glutamate-derived amino acids encoded by C-starting codons, and finishing with large, aromatic amino acids encoded by U-starting codons. This error robustness differs both in possible evolutionary cause and in codon position pattern from the error robustness found with a hydrophobicity input.

Concerning the mathematics surrounding studies of the SGC, one more issue should be scrutinized here. Above, it was already mentioned that in publications of the groups of Ochoa and Nirenberg [SLB$^+$63, NJL$^+$63] certain regularities (in the sense that similar codons often code for similar amino acids) in the SGC were pointed out, and that Crick [Cri68] distrusted the ideas that remarkable patterns could be discovered along these lines. Mathematics and computer science came to the rescue of evolutionary biochemistry, and proved, with mathematical rigor, beyond reasonable doubt, that there *are* these remarkable patterns in the SGC. The first person, to my knowledge, who used computer science and mathematics to do this, was Alff-Steinberger [AS69]. He split the calculation right from the start over the three codon positions. The remarkable error robustness of the SGC compared to codes with random redistributions of the assignments were revealed in many of his figures [AS69]. Later, Wong investigated the error robustness of the SGC [Won80]. In his error function, the robustness concerning similar codons coding for *identical* amino acids was not present. The error function which was developed next, was developed by Di Giulio [DG89a] (see also [DG89b]). In taking the robustness concerning similar codons coding for *identical* amino acids also in account, it resembled the approach of Alff-Steinberger [AS69]. The big

differences with Alff-Steinberger's approach were the use of a square to amplify the effect, and the combination of all codon positions into one single value for a given code variant. Sticking to the "fixed block structure" was clearly formulated by Di Giulio too ("... those codes obtained from that code [i.e. the SGC] through random amino acid permutation, by which I use to mean all the possible permutations (20! = 2.4 x 10^{18}) of the amino acids on the synonymous codon block that remain invariant. By invariant synonymous codon block I mean that the structure of the synonymous codon blocks of the genetic code is always the same (as shown in Fig.1) - it does not change and the only thing that might vary is the position of the amino acids on the blocks" page 289 in [DG89a]). The function and block structure later used by Hurst and co-workers (see e.g. [HH91, FH98a]) were therefore introduced by Di Giulio.

1.3 Linkage selection

In molecular genetics, sometimes situations are found which are enormously complicated. A good example of this is the process by which proteins are produced in the mitochondria of the parasites causing African sleeping sickness. As we have seen, the normal way genetic information flows, is: "DNA makes RNA makes Protein". This is known as "The Central Dogma of Molecular Biology". In the mitochondria of these parasites, the first impression is that genetic information is encoded in an unknown location ("somewhere else"), and not in the DNA: large numbers of U nucleotides are sometimes absent from the sequence which is encoding the protein in the DNA. These are inserted later in a process called "RNA editing". During the last 30 years it has become clear that small RNA molecules present as small RNA-coding genes in the mitochondrial DNA are responsible for inserting the "missing U's", and occasional deleting "extra U's" during this process of RNA editing. There is therefore no violation of the Central Dogma: the U's are coming from "somewhere else", but that "somewhere else" is still located somewhere on the mitochondrial DNA. The genes for these small RNAs are scattered all over the DNA of the mitochondria. This means that the information necessary to make the protein is dispersed over the DNA molecules, and is intermingled with the information of other genes. Why is this so?

A hypothesis was proposed [Spe06, Spe07] to explain this complicated situation. In evolutionary biology, some intricate ways in which the process of natural selection works, are known by special names. "Sexual selection" for example, is the name for the process which leads to male ornamentation like the long, eye-spotted tail of a peacock. An interaction between genes producing the tail and genes producing preference of female peafowl for males with large and colourful tails leads to an evolutionary path of gradually larger and more colourful tails in the pheasant family (Phasianidae), among which the peafowl species are belonging to the most extreme. In chapter 6, a new term is introduced for another

intricate way in which natural selection may work. This new term is *linkage selection*. Like the peacock living in an evolutionary environment in which the preferences of female peafowl form a major determinant, the sleeping sickness parasite is living in an evolutionary environment in which the occurrence of intense competition during clonal growth in a host is a major determinant. As a result of this competition, deletion of mitochondrial DNA temporarily not in use is a real danger to the parasite. The consequence of such a deletion is an incapability to survive in the alternative host. The linkage between information essential on a short term and information essential in the long run protects this kind of organism against the strong advantage clonal deletion variants would otherwise have. This linkage selection model is investigated here using computational tools.

Chapter 2

Searching for primordial peptides

The content of this chapter is based on joint work with Serge Massar, Dimitri Gilis, Harry Buhrman, and Marianne Rooman [vdGMG+09].

2.1 From philosophical speculations to rigorous scientific enquiry

The development of molecular biology over the past half century has transformed the question of how life appeared from the level of philosophical speculations to the level of rigorous scientific enquiry. A number of key ideas have emerged which govern our thinking about this question, among which the synthesis of some amino acids and sugars by prebiotic synthesis [Mil87, RCOB04], the RNA-world in which RNA catalyses its own duplication [Gil86, JUL+01], and the role of lipid vesicles to limit the spatial extent of the cell precursor [Dea85, Che06, MSK+08]. At some point the controlled synthesis of proteins emerged, and proteins then took over many functions, presumably because of their great specificity and efficiency. Here, we address an important question, namely what properties did the first functional peptides and proteins have when they emerged during very early life? Answering these questions is a difficult and unsolved issue. Indeed, the smallest natural proteins that can take a stable structure by themselves, without interacting with other biomolecules, are composed of around 20 amino acids. Some synthetic constructs are even smaller: for example, chignolin, a synthetic protein of 10 amino acids, has been shown to have a stable structure in water [HYSM04]. However, it is difficult to imagine that functional proteins 10-20 amino acids long suddenly appeared out of the blue. Indeed, the sequences of proteins are very specific, and reliably making functional proteins 10-20 residues long requires an efficient code for the amino acid sequence, and an efficient translation mechanism from the code into the protein (see section 2.5).

We propose a solution to this "chicken and egg" paradox by suggesting that

specific short peptides 3 to 8 amino acids long could have served as catalysts during very early life. Longer proteins would then have gradually evolved from these early precursors. The idea that very short peptides could have had a useful role in very early life has already been put forward by Shimizu [Shi95, Shi04, Shi07] who showed that single amino acids and dipeptides could slightly enhance the rates of certain chemical reactions. But obviously some intermediate steps are required between the dipeptides of Shimizu and the smallest of today's functional enzymes.

An additional constraint on any theory of how the first functional peptides emerged is that these should be composed exclusively or almost exclusively of the amino acids that are efficiently produced by prebiotic synthesis. In most prebiotic synthesis experiments (see next section) the most efficiently produced amino acids are Gly, Ala, Val, and Asp (or G, A, V and D in one letter code) [Mil87]. Of these, the first three are neutral, and Asp is negatively charged. At first sight this is problematic: the absence of positive charges compensating the negative Asp's are likely to limit their ability to form stable structures and to carry out a catalytic activity.

We propose to resolve these conundrums by suggesting that the first peptides were composed of short chains of prebiotic amino acids bound to (one or more) positively charged metal ions. Supposing the first peptides to be bound to metal ions solves two problems at once. Namely, the metal ion(s) provide(s) an anchor around which the peptide can organize itself, thereby (at least partially) stabilizing its structure, and secondly it provides a positive charge which would be very useful for catalytic activity.

Later, as the coding and translation mechanisms improved, these very first peptides gradually lengthened, thereby improving the efficiency and specificity of their biological activity. We further conjecture that some of these first peptides, composed of prebiotic amino acids bound to metal ions, have been conserved across evolution. The fact that active sites are believed to be better conserved than all other protein regions speaks[1] in favor of this conjecture. If this idea is correct, it should be possible to find today the memory of the very first functional peptides in the active sites of some present-day proteins.

The idea of finding in present-day proteins traces of very early life is not entirely new. For instance, it has been argued that today's amino acid abundances reflect the order in which they were introduced in the genetic code (see [ZDV71, JKA$^+$05], and the criticism of [HFR06]).

To explore the validity of our conjectures, we carried out a search in the

[1] Actually, recent insights suggest that active sites are *not* the most conserved elements of proteins: residues influencing protein stability and kinetics of signaling modulation are more conserved in the PAS domain superfamily than active site residues [KKP$^+$10] (see also [PKH10]); furthermore, not residue identity but the pattern of side-chain hydrogen-bonding interactions is the characteristic which is most conserved. The idea that DNA-dependent RNA polymerase is an exception to this order (first the form, then the active site) is controversial (see e.g. [RBL16]).

protein structure database (PDB) to identify all active sites composed almost exclusively of the abundant prebiotic amino acids bound to metal ions. In the following, we review the abundances of amino acids produced in prebiotic synthesis experiments, then present our precise search methodology, and discuss the classes of prebiotic peptide candidates revealed by the search. These candidates all correspond to active sites in their host protein, and perform catalytic functions likely to be important during very early life. Furthermore, there is strong evidence that some of these active sites have been conserved at least since the Last Universal Common Ancestor (LUCA) 3.5 billion years ago. We conclude by discussing in more detail how our suggestions fit into the wider picture of the appearance of life, and how they could be tested experimentally.

2.2 State of the art: Prebiotic amino acids

In this section we review the relative abundances of amino acids that are produced in prebiotic synthesis experiments such as those originally designed by Miller [Mil55], or that are observed in meteorites, in order to determine the most plausible composition of the first peptides that appeared on the primitive earth. These amino acid abundances are summarized in table 2.1,

	Gly	Ala	Asp	Val	Ser	Glu	Other
Miller and co-workers	1	9×10^{-3} to 2	2×10^{-4} to 8×10^{-2}	0 to 4×10^{-2}	1×10^{-4} to 3×10^{-2}	0 to 2×10^{-2}	0 to 4×10^{-2}
Plankensteiner et al. (2006)	Yes	Yes	0	Yes	? to Yes	0	Yes
Botta et al. (2002)	1	0 to 7×10^{-1}	6×10^{-2} to 6×10^{-1}	? to Yes	0 to 5×10^{-1}	2×10^{-2} to 2	?

Table 2.1: Relative frequencies of proteinaceous amino acids in prebiotic synthesis experiments (Miller and co-workers, Plankensteiner et al. (2006)) and in meteorites (Botta et al. (2002)). "Miller and co-workers" refers to four articles: Miller (1955), Ring et al. (1972), Schlesinger and Miller (1983), Johnson et al. (2008).

Note that we do not review the determinations of amino acids in the interstellar medium, as this issue is under discussion [KCH$^+$03, SLH$^+$05, JCGC07, ZZDS08].

Miller and co-workers performed synthesis experiments in several environments attempting to mimic possible prebiotic circumstances, as the real environment is not agreed upon. The first three environments considered by Miller [Mil55], and recently reanalyzed by Johnson et al. [JCD$^+$08], consist of a highly reducing mixture of CH_4, NH_3, H_2O and H_2, which is sparked in three different ways. They all produced Gly, Ala, Asp, Glu, and Ser in significant amounts. Moreover, in the experimental set-up which is nowadays considered the most interesting one (see [JCD$^+$08]), Val was produced in about the same amount as Glu. In later work of Miller and co-workers [RWFM72], ammonia was only present in trace

amounts and the major source of nitrogen was N_2. In this case ten proteinaceous amino acids were produced, in order of abundance: Ala, Gly, Asp, Val, Leu, Glu, Ser, Ile, Pro and Thr.

Schlesinger and Miller [SM83] investigated the effect of changing the carbon source (CH_4, CO, or CO_2). They found that with CO or CO_2 as carbon source, Gly was almost the only amino acid found, with small amounts of Ala and Ser, and trace amounts of Asp and Glu, and, in a single case, Val. Recently, however, Miller and coworkers showed, contrary to previous reports, significant production of Ala, Ser, and Glu in neutral atmospheres [CCL$^+$08]. They conjectured that nitrite and nitrate, which are also produced by spark discharge in neutral atmospheres, destroyed the amino acids in previous work. Addition of pH buffer and oxidation inhibitor prevents this destruction. They concluded that neutral atmospheres may have been productive in prebiotic synthesis of amino acids, provided the early oceans were buffered sufficiently with respect to pH and redox balance.

Another series of experiments, performed by Plankensteiner et al. [PRR06], assumed the prebiotic atmosphere to be neutral and composed of CO_2, N_2, and H_2O. The amino acids that were systematically produced were Gly, Ala, and Val. In addition, Ser, Pro, and Lys were detected in significant amounts in several experiments. Surprisingly, Asp and Glu were not detected.

An interesting observation is the similarity between the proteinaceous amino acids produced in prebiotic synthesis experiments and those present in carbonaceous meteorites, as first noticed by Kvenvolden [Kve74] and Miller [Mil74]. A detailed analysis of the composition of several meteorites was later performed by Botta et al. [BGKB02], and showed the presence of Gly, Ala, Asp, Val, Ser, and Glu.

Clearly, some discrepancies occur among the amino acids produced, according to the experimental setups, the environments meant to represent the primitive Earth, or the type of meteorites. However, general tendencies are noted: Gly and Ala always appear as the most abundant amino acids, and Asp, Val, Ser, and Glu often appear in significant amounts.

The choice we made of which were the abundant prebiotic amino acids was motivated by the prebiotic synthesis experiments performed by Ring et al. [RWFM72]. Ten proteinaceous amino acids were found by these authors, as well as several ones which were non-proteinaceous, with a spectrum showing strong similarities with that of the Murchison meteorite. In this experiment, Gly, Ala, Asp, and Val were the four most abundant amino acids, and we chose them as the presumed components of the first active peptides.

2.3 Search for traces of prebiotic peptides

We searched for traces of prebiotic peptides in the RCSB Protein structure DataBase (PDB) [BBB$^+$02]. For the purpose of selecting all PDB entries that

contain a protein structure interacting with one metal ion at least and of identifying all interactions between ions and amino acids, we used the PDBsum summary and analysis tool [LCT05]. The metals considered are Li, Be, Na, Mg, K, Ca, Cr, Mn, Fe, Co, Ni, Cu, Zn, and Se.

The prebiotic peptide candidates we searched for in the PDB were assumed to be built almost exclusively from the amino acids that were presumably the most abundant in early times, i.e. Gly, Ala, Val, and Asp (see previous section). Moreover, these peptide candidates were required to be bound to metal ions, present in the prebiotic environment, in order to neutralize the negative charge of Asp and to allow the formation of well defined tertiary structures and catalytic functions.

The exact criteria that we used to characterize prebiotic peptide candidates are the following. Firstly, these peptides must correspond to sequence regions of maximum length 8, the longest prebiotic peptides envisaged. Note that doubling this maximum length did not change the results of the search. Secondly, they must be bound to a metal ion, and this ion may not interact with any other part of the protein. Thirdly, the ion must be bound to the abundant prebiotic amino acids only, among which three or more Asp residues. Requiring three Asp's at least increases the strength and specificity of the peptide-ion binding and thus its chance to be part of the protein active site in addition to stabilizing the structure. Indeed, when relaxing the criterion and accepting two Asp's bound to an ion, many spurious fragments were selected. Fourthly, the sequence starting at the first and ending at the last residue bound to the ion, and including start and end residues, should be composed of the abundant prebiotic amino acids for 80% at least. Lastly, if this sequence is flanked in the host protein by abundant prebiotic amino acids, these amino acids are added to the peptide.

To estimate the statistical significance of finding these prebiotic peptide candidates in the PDB, we computed E, the number of such peptides expected to occur by chance, as follows. Let E_{ij} be the expected number of sequences of length i with j residues binding an ion. We compute this as $E_{ij} = P_{ij}N_{ij}$, where P_{ij} is the probability that a motif matches our prebiotic criteria, and N_{ij} is the number of motifs of length i with j residues attached to the ion present in the PDB. The probabilities P_{ij} are estimated as:

$$\begin{aligned} P_{ij} &= \left(\begin{array}{c} i-2 \\ j-2 \end{array} \right) \sum_{k=3}^{j} \left(\begin{array}{c} j \\ k \end{array} \right) P(D|ion)^k P(A/G/V|ion)^{j-k} \\ &\quad \times \sum_{m=\lceil o.8i \rceil - j}^{i} -j \left(\begin{array}{c} i-j \\ m \end{array} \right) P(D/A/G/V|\overline{ion}^m P(\overline{D/A/G/V}|\overline{ion})^{i-j-m}, \end{aligned} \tag{2.1}$$

where '/' means 'or', and $P(y|ion)[P(y|\overline{ion})]$ is the fraction of amino acids of

type y, given that they are bound [not bound] to one of the considered ions; these fractions are computed from the full PDB. To obtain this formula, we first fixed the outer amino acids binding the ion, which leaves $\binom{i-2}{j-2}$ possibilities to choose the location of the remaining $j-2$ prebiotic amino acids that also bind the ion. Among these we have to choose at least three Asp's, giving rise to the sum from k=3 to j. Last we have to choose, out of the $i-j$ remaining positions that do not bind the ion, at least $\lceil o.8i \rceil$ prebiotic amino acids in order to have at least 80% such amino acids in the motif. The total expected number of prebiotic motifs is given by:

$$E = \sum_{i=3}^{8} \sum_{j=3}^{i} E_{ij} = \sum_{i=3}^{8} \sum_{j=3}^{i} P_{ij} N_{ij}. \tag{2.2}$$

However, this formula contains several overcountings. First, the PDB contains many highly similar, homologous, proteins. This implies that when a motif is found in one of them, it is almost automatically found in the others. To correct for this bias, we replace N_{ij} by N_{ij}^{cor}, which counts only the motifs that have at least one different amino acid, or that have at least 20% different amino acids in a sequence stretch of 10 residues around the middle residue of the motif. Another correction is due to the fact that two consecutive Asp's binding an ion are very rarely observed. Indeed, the joint probability of finding two Asp's bound to an ion at consecutive positions along the sequence is much lower than the product of the independent probabilities of finding an Asp bound to an ion; the ratio of these probabilities is only equal to 0.05. This effect is due to geometrical constraints imposed by the polypeptide chain, and the unfavorableness of having two negative charges close together. This effect is also, but to a lower extent, observed for other consecutive amino acids bound to an ion. To take this effect into account, we have to distinguish between the different ion-binding motifs, which we label by the index v. For example, for a motif of length $i = 5$ with $j = 3$ residues bound to an ion, the $\binom{i-2}{j-2} = 3$ possible motifs are `iixxi`, `ixixi` and `ixxii`, where `i` and `x` denote amino acids bound and not bound to the ion, respectively; N_{ijv}^{cor} with $i = 5$, $j = 3$ and $v = 1$ corresponds thus to the number of non-homologous matches of motifs of type `iixxi` in the PDB. This leads us to the corrected expected value E^{cor}:

$$E^{cor} = \sum_{i=3}^{8} \sum_{j=3}^{i} E_{ij}^{cor} = \sum_{i=3}^{8} \sum_{j=3}^{i} \sum_{v=1}^{\alpha_{ij}} P_{ijv}^{cor} N_{ijv}^{cor} \quad \text{with } \alpha_{ij} = \begin{pmatrix} i-2 \\ j-2 \end{pmatrix}, \tag{2.3}$$

where the corrected probability P_{ijv}^{cor} is given by:

$$P_{ijv}^{cor} = \sum_{k=3}^{j} \sum_{w=1}^{\beta_{jk}} P(D|D,ion)^{n_{vw}^{DD}} P(D|a,ion)^{n_{vw}^{Da}} P(D|ion)^{k-n_{vw}^{DD}-n_{vw}^{Da}} \times P(a|D,ion)^{n_{vw}^{aDvw}} P(a|a,ion)^{n_{vw}^{aa}} P(a|ion)^{j-k-n_{vw}^{aD}-n_{vw}^{aa}} \times \sum_{m=\lceil o.8i \rceil - j}^{i-j} \left(\begin{array}{c} i-j \\ m \end{array} \right) P(D/a|\overline{ion}^{m} P(\overline{D/a}|\overline{ion}^{i-j-m}, \quad \text{with } \beta_{jk} = \left(\begin{array}{c} j \\ k \end{array} \right). \tag{2.4}$$

where $a = A/G/V$, and $P(y|z,ion)$ is the probability that an amino acid bound to an ion is of type y, given that the next amino acid along the sequence is bound to the same ion and is of type z. The index w labels the β_{jk} motifs of type v; for example, for $j = 4$ and $k = 3$, the $\beta_{jk} = 4$ ion-binding motifs of type `iiixi` are `DDDxa`, `DDaxD`, `DaDxD`, and `aDDxD`. The integers n_{vw}^{yz} denote the number of times, in a motif of type v (e.g. `iiixi`) and subtype w (e.g. `DDaxD`), two successive amino acids bound to an ion, the first being of type y and the second of type z. In the example `DDaxD`, $n_{vw}^{DD} = 1$, $n_{vw}^{Da} = 1$, $n_{vw}^{aD} = 0$, and $n_{vw}^{aa} = 0$.

To objectively compare E (or E^{cor}), the expected number of prebiotic peptide candidates, and N (or N^{cor}), their actual number observed in the PDB, we estimate the P-value, defined as the probability of finding at least N (or N^{cor}) prebiotic peptide candidates if the null hypothesis was true. For this, we assume that E (or E^{cor}) follow a Poisson distribution, as it is a discrete probability distribution only defined for positive values. The P-value is then given by:

$$P\text{-value} = \sum_{k=N}^{\infty} \frac{E^k e^{-E}}{k!}. \tag{2.5}$$

2.4 Prebiotic peptide candidates

Using the above definition of prebiotic peptide candidates, we screened systematically the whole PDB, and found two different ion binding motifs: DxDxD and DxxxxDxD, where x stands for any amino acid. These two motifs differ in the relative positions along the sequence of the three Asp residues binding the ion. The first binding motif further divides into four groups, according to the amino acids between and flanking the Asp residues, as summarized in table 2.2.

The first two DxDxD binding motifs correspond to active sites of RNA polymerases, and are both bound to Mg^{2+}, or sometimes Mn^{2+} ions. The first motif, with strictly conserved sequence ADFDGD, or even NADFDGD if one includes

Binding motif[a]	Candidate prebiotic sequence[a]	Metal ion	N	N^{cor}	Central structure[b]	Mean Rms[c] (Å)	Rms[d] (Å)
Dx**D**x**D**	A**DFDGD**	Mg,Mn	53	1	1i6h (A480-A485)	0.63	0.67
	GG**DYDGD**	Mg	4	1	2j7n (A1005-A1011)	0.03	
	DGDGD	Mg,Zn, Ni	7	1	1p5d (X242-X246)	0.45	0.77
	DGDAD	Zn	1	1	1kfi (B308-B312)	-	
Dxxxx**D**x**D**	**D**AKVG**DGD**	Mg	2	1	1un9 (A380-A387)	-	-

Table 2.2: Prebiotic peptide candidates found in the protein structure database.

[a]The residues in bold are bound to the metal ion.

[b]The central structure of the group is that for which the rms deviation with respect of all other members of the group is minimum.

[c]The rms deviation is computed between all heavy atoms for the main chain and side chain and the metal ion, after coordinate superimposition; the mean is taken between all pairs in the class

[d]The rms deviation is computed between the equivalent parts of the central structures of two classes, in particular between DFDGD and DYDGD (considering common atoms in F and Y), and between DGDGD and DGDAD (considering the common atoms in G and A).

the preceding, non-prebiotic residue N, is found in DNA-dependent RNA polymerase, the normal RNA polymerase engaged in protein-coding gene expression. More precisely, it is found in the second domain of RNA polymerase Rpb1, using the nomenclature of the protein families database Pfam [FTM$^+$08].

It is worthwhile to linger a few moments on the age of this NADFDGD sequence. It corresponds to the active site in the mRNA-producing enzyme of all animals, such as *Homo sapiens*, the frog *Xenopus laevis*, and the nematode worm *Caenorhabditis elegans*, but also in the yeast *Saccharomyces cerevisiae* and the plant *Arabidopsis thaliana* [IKA03]. This sequence seems thus to be present in all eukaryotes, which populate the earth since two billion years at least. Moreover, it is present in other organisms, which share a common ancestor with eukaryotes in an even more remote past, in particular in the bacterial genera *Thermus*, *Deinococcus*, *Escherichia*, *Agrobacterium*, *Mesorhizobium*, *Helicobacter*, *Ureaplasma*, *Bacillus*, *Thermotoga*, *Synechocystis*, *Aquifex*, *Chlamydia*, *Mycoplasma*, *Mycobacterium*, and *Treponema*, and the archaeal genera *Sulfolobus*, *Aeropyrum*, *Methanosarcina*, and *Halobacterium*.

The orthodox way in molecular biology to interpret the almost completely universal appearance of some characteristic in a class of organisms, is that their common ancestor already[2] did have this characteristic.

[2]This argument becomes less convincing as the sequences become shorter. In this case, some

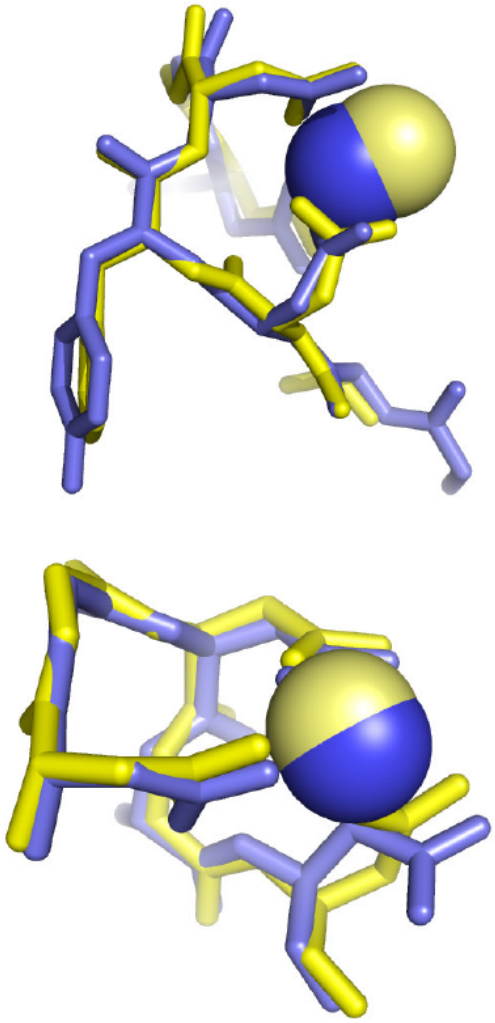


Figure 2.1: The resemblance between metal ion binding prebiotic peptide candidates suggested by the analysis of polymerases and mutases. The figures are drawn using PyMol (http://pymol.sourceforge.net). On top: D(F/Y)DGD motif in RNA polymerases. Yellow: PDB entry 1i6h A480-A485 bound to a Mg^{2+} ion; blue: PDB entry 2j7n A1005-A1011 bound to a Mg^{2+} ion. Bottom: DGD(G/A)D motif in phospho(gluco/manno)mutase. Yellow: PDB entry 1p5d X242-X246 bound to a Zn^{2+} ion; blue: PDB entry 1kfi B308-B312 bound to a Zn^{2+} ion.

The sequence NADFDGD is found in the mRNA-synthesizing RNA polymerase of all living cellular organisms, and is thus assumed to be present in their common ancestor (the Last Universal Common Ancestor, or LUCA), which lived

may consider the sequence short enough to have originated by convergent evolution. To make the case for common ancestry, the "continuity principle" might be needed. Another viewpoint is, that "chemical determinism" might be the cause behind sequence similarities, and that these "do not reflect an evolutionary continuity" [RBL16].

approximately 3.5 billion years ago. This is thus the minimum age of the sequence NADFDGD, given that there is still a long history from the origin of life up to the time of LUCA.

The second motif, with sequence GGDYDGD, corresponds to the active site of a cell-encoded RNA-dependent RNA polymerase of *Neurospora crassa*, which is part of the RdRP family according to the Pfam database. Its role is to produce double-stranded RNA from aberrant single-stranded RNA as part of an RNA silencing response [SKM+06]. Although this kind of protein is not present in insects, vertebrates and the yeast *Saccharomyces cerevisiae*, it is a kind of protein that is a normal constituent of an eukaryotic cell. For example, related sequences are known from the nematode worm *Caenorhabditis elegans*, the plants *Arabidopsis thaliana*, *Nicotiana tabacum* and *Oryza sativa*, and the unicellular eukaryotes *Dictyostelium discoideum* and *Giardia intestinalis* [IKA03]. The GGDYDGD motif is not totally conserved among these sequences, but has the following variations: G(G/S/A)D(Y/F/M/L)DGD, where the '/' symbol means 'or'; it is thus very close to the NADFDGD motif.

The catalytic domains of these two types of RNA polymerases, Rpb1 domain 2 and RdRP, are structurally similar, whereas their other domains are completely different. The geometries of their active sites bound to Mg^{2+} or Mn^{2+} ions are particularly alike: they present a root mean square (rms) deviation of 0.67 Å (see table 2.2), and their resemblance is clearly illustrated in Fig. 2.1.

Binding motif[a]	Candidate prebiotic sequence[a]	Biological function	Consensus sequence[b]
Dx**D**x**D**	A**DF**D**GD**	DNA-directed RNA polymerase Rpb1 domain 2	**D**(F/Y/M/L)**DGD**
	GG**DYDGD**	RNA-dependent RNA polymerase	
	DGDGD	Phosphoglucomutase, phosphomannomutase	**DGD**(G/A/F)**D**
	DGDAD	Phosphoglucomutase	
Dxxxx**D**x**D**	**DAKVGDGD**	Dihydroxyacetone kinase	**DAKVGDGD**

Table 2.3: Biological functions and consensus sequences of the considered motifs.

[a]The residues in bold are bound to the metal ion.

[b]The consensus sequence is obtained from sequence alignments in homologous protein families

This leads us to merge these two motifs and to define a single RNA polymerase motif, of consensus sequence D(F/Y/M/L)DGD, with an invariant Gly and three invariant Asp residues (see Table 2.3). Note that the only non abundant prebiotic residue of the motif, F/Y/M/L, is not well conserved. Moreover, it points towards

the protein core rather than to the ions. It has thus probably a structural role, but is unimportant for the catalytic function; this leads to the conjecture that this position was occupied by an abundant prebiotic residue A, G, or V in early times.

The three conserved Asp's of the motif are involved in the binding of one or two Mg^{2+} or Mn^{2+} ions, which coordinate the phosphates of the nucleotide triphosphate (NTP) that is going to be built into the RNA polymer, as illustrated in Fig. 2.2.

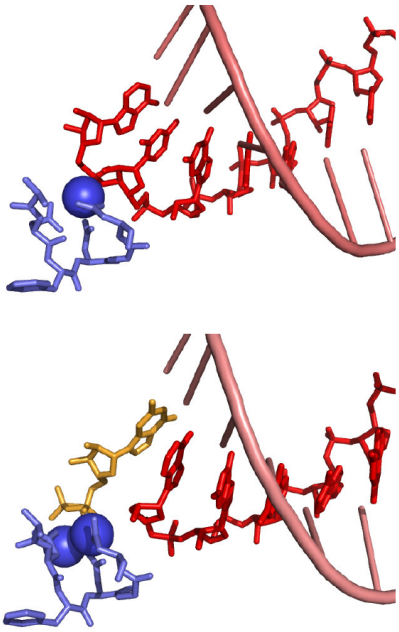


Figure 2.2: Biological function of the first prebiotic peptide candidate. The figures are drawn using PyMol (http://pymol.sourceforge.net). On top: DFDGD motif: PDB entry 1i6h A480-A485 bound to a Mg^{2+} ion (blue), interacting with an RNA strand (red), iself in interaction with a DNA strand (pink). Bottom: DFDGD motif: PDB entry 2e2h A480-A485 bound to two Mg^{2+} ions (blue), interacting with a GTP molecule (orange) and an RNA strand (red), itself in interaction with a DNA strand (pink).

What is the evolutionary relationship between these two RNA polymerase active sites, which share the same geometry, the same mechanism and the same Asp residues and Mg^{2+} or Mn^{2+} ions? Iyer et al. [IKA03] tend to view these common characteristics as a sign of very ancient common ancestry, dating from more than 3.5 billion years ago. In this view, the RNA-dependent RNA polymerase would originally have been the mRNA-producing enzyme of an organism that lived before the last common ancestor of all contemporary cells, and would have found its way to the eukaryote by a pathway of horizontal gene transfer.

The Asp residues and the Gly residue in the ancestral RNA polymerase con-

sensus sequence xxDxDGD can be expected to play such an important role in the polymerase function that they cannot be changed, Asp because of its specific interaction with the metal ions, and Gly because of its ability to make main chain turns impossible[3] with other amino acids. This does not exclude other ways to make an RNA polymerase; however, when starting from this particular sequence, function is lost upon mutation. Note that the extreme conservation of the other amino acids in the NADFDGD motif of DNA-dependent RNA polymerase cannot be attributed to the conservation of the function, but is probably related to their interaction with other parts of the protein and other parts of the transcription machinery. Following this line of thought, xxDxDGD can be viewed as one of the most profound molecular fossils in life, and possibly one of the first coded peptide sequences. In these very early times, more than 3.5 billion years ago, they were possibly built from Val, Ala, Asp, and Gly residues only, and might have been originally composed of amino acids with a prebiotic rather than a biogenic origin.

The last two DxDxD motifs presented in Table 2.2 correspond to the two sequences DGDAD and DGDGD, bound to Mg^{2+}, Zn^{2+}, or Ni^{2+} ions. They have very similar structures, as shown in Fig. 2.1; their rms deviation is equal to 0.77Å. They are also part of the same homologous family according to Pfam [FTM$^+$08], named PGM_PMM_II, with consensus sequence DGD(G/A/F)D (see Table 2.3). These motifs correspond to the active sites in enzymes with mutase activity (see Fig 2.3).

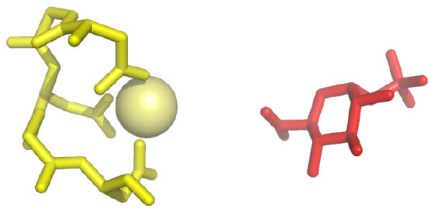


Figure 2.3: Biological function of the second prebiotic peptide candidate. The figure is drawn using Pymol (http://pymol.sourceforge.net). DGDGD motif: PDB entry 1p5d X242-X246 bound to a Zn^{2+} ion (yellow), and close to a G1P molecule (red).

These enzymes can usually bind several metal ions, but Mg^{2+} yields systematically the maximum activity [RTB02]. Some, like rabbit phosphoglucomutase,

[3]This means that the Phi and Psi angles of these residues would place these residues in a region of the Ramachandran plot that is sterically inaccessible to the backbone of other amino acids. Although the NADFDGD sequence has been known for a long time [DHLDL$^+$95], to the best of my knowledge, this has not been checked, although the structure of this active site is a major focus of interest (see e.g. [ZS12, SZS$^+$05, Nud09, MRC13, LBK13] and references therein).

show a slight activity with Ni^{2+} [RTB02], whereas Zn^{2+} is usually observed to inhibit the enzyme [RSB06b]. As an example, phosphomannomutase/ phosphoglucomutase of the bacterium *Pseudomonas aeruginosa* is changing the position on which a phosphorylated hexose sugar molecule is carrying the phosphate group. The Mg^{2+} ion, which is essential for catalysis [SRB04], is coordinated by the three Asp's. The catalytic process, in which a biphosphorylated sugar intermediate has to engage in a dramatic, 180° reorientation, is described in detail in [RSB06a]. Of course, a complex reaction mechanism like this is a long way from simple prebiotic circumstances. Nevertheless, the proteins of the α-D-phosphohexomutase enzyme superfamily could have inherited the simple DGD(G/A)D Mg^{2+} ion-binding motif from a more remote past, possibly from a very early biotic age, where this motif fulfilled simplified catalytic activities.

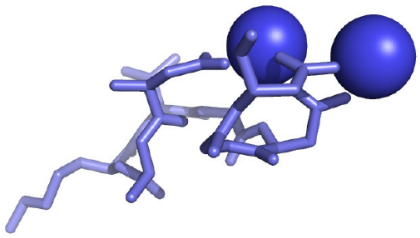


Figure 2.4: The metal binding prebiotic peptide candidate suggested by the analysis of a kinase: DAKVGDGD motif in dihydroxyacetone kinase. Blue: PDB entry 1un9 A380-A387 bound to two Mg^{2+} ions. The figure is drawn using PyMol (http://pymol.sourceforge.net.

The DxxxxDxD motif is represented in our search by just one sequence, DAKVGDGD, which binds to two Mg^{2+} ions and turned out to be the active site of dihydroxyacetone (DHA) kinase (Fig. 2.4). Because this enzyme was producing the glycolytic intermediate dihydroxyacetone-phosphate (DHAP), we consider this enzymatic activity as interesting from the viewpoint of the transition of prebiotic to early biotic. Like the situation we came across with the RNA polymerases, the active site consists of two Mg^{2+} ions, coordinated by three Asp residues, and these Asp residues and Mg^{2+} ions are essential for catalysis [SAGA$^+$03]. As with the RNA polymerase active sites, the phosphate groups of an NTP are complexed with the Mg^{2+} ions (Fig. 2.5). The γ-phosphate is used to phosphorylate the substrate of this kinase enzyme. The substrate spectrum may be broader than the name DHA kinase suggests, because the activity of flavin mononucleotide (FMN) cyclase was found to be identical to that of human DHA kinase [ESC$^+$06].

The structure of the active site of DHA kinase is different from that of the RNA polymerase and mutase classes (please compare Fig. 2.4 to Fig. 2.1). The catalytic action presents similarities with that of RNA polymerases, but differs

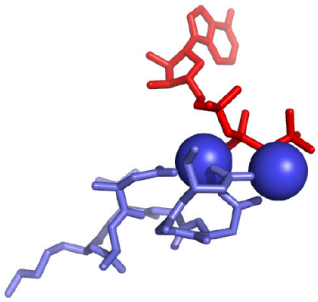


Figure 2.5: Biological function of the third prebiotic peptide candidate. The figure is drawn using Pymol (http://pymol.sourceforge.net). DAKVGDGD motif: PDB entry 1un9 A380-A387 bound to two Mg^{2+} ions (blue) and interacting with an ATP molecule (red).

when examined in detail: in this case the bond between the second and the third phosphate is broken, while in the case of the RNA polymerases, it is the bond between the first and the second phosphate that is being broken.

To assess the statistical significance of finding these sequence motifs in present-day proteins and proposing them as prebiotic peptide candidates, we compute the number E of such motifs that are expected to be found in the PDB by chance, as defined by equations 2.1 and 2.2. We found $E = 30$, whereas the number of motifs indeed present in the PDB is equal to $N = 67$ (Table 2.2); the associated P-value (see equation 2.5) is 4×10^{-9}. When taking into account all biases, in particular the presence of homologous proteins in the PDB, the geometrical constraints that tend to prevent successive residues to be bound to an ion, and the preference for an Asp to be bound to an ion rather than to be close but not bound to it, we find using equations 2.3 and 2.4 that the expected number of occurrences is $E^{cor} = 1.8$, whereas its actual number is $N^{cor} = 5$. Thus, we found 2.8 times more prebiotic peptide candidates than expected by chance, leading to a P-value of 3×10^{-2}. We would moreover like to stress that all these peptides are part of the active site of their host protein, although this was not among the selection criteria. This strongly improves the statistical significance of our results, even though this improvement cannot be estimated quantitatively at the moment.

As described in section 2.2, it is not easy to determine exactly which were the most abundant amino acids on the primitive earth, and therefore our choice of Ala, Val, Asp, and Gly as prebiotic amino acids may give rise to criticisms. Indeed, Asp and Val are not formed in all prebiotic synthesis experiments (Table 2.1). Asp cannot be removed because its negative charge is indispensable for binding tightly to the metal ions, but Val could be dropped. However, this choice only removes a single motif from the prebiotic peptide candidates, i.e. DxxxxDxD. The Ala-Gly-Asp choice of abundant prebiotic amino acids thus does not change

the statistical significance level of our results by much, as we have $E^{cor} = 1.3$ and $N^{cor} = 4$, with a P-value of 5×10^{-2}.

As an additional test of our procedure, we investigate the effect of replacing Asp by His among the abundant prebiotic amino acids, given that histidines also bind very efficiently to metal ions. This led to detect $N^{cor} = 2$ prebiotic peptide candidates, whereas the expected number is $E^{cor} = 0.1$. In these two hits, however, the structure of the host protein is incomplete: the coordinates of 5 and 11 residues, respectively, are lacking in the immediate vicinity of the metal ion. These matches have thus to be considered as spurious. Finally, when considering as prebiotic amino acids all 10 amino acids produced in the experiments of Ring et al. [RWFM72], i.e. Val, Gly, Asp, Ala, Ile, Pro, Ser, Thr, Leu and Glu (Table 2.1), the number of expected and observed peptide candidates is of course higher: $N^{cor} = 22$ and $E^{cor} = 20$. However, given that this set of amino acids contain the four abundant prebiotic amino acids Val, Gly, Asp, and Ala, for which we found $N^{cor} = 5$ and $E^{cor} = 1.8$, we may conclude that the six additional amino acids yield less peptide candidates than expected by chance.

In conclusion, the most statiscally significant results come out when considering Gly, Ala and Asp, and possibly Val, as abundant prebiotic amino acids. This can be viewed as an *a posteriori* indication that these are really the true abundant prebiotic amino acids.

2.5 The origin of life and the first peptides

How do our results fit into a wider picture of the appearance of life? The present-day picture of how life appeared is punctuated by a number of established facts, in a background of unknown or dubious stages and links. We make here an attempt to review the results that are well verified or supported, in their tentative order of occurrence:

(1) Many essential molecules for life can be synthesized in a prebiotic environment: these include amino acids [Mil87], but also pentoses [RCOB04], nitrogenous bases [CNM06], and vesicle-producing organic compounds [Dea85].

(2) Still in a prebiotic environment, more complex molecules can be built from these basic building blocks: adenine and ribose can be condensed into a nucleoside by the presence of borate in aqueous solution [Pri01]; polyphosphate, likely to be a source of energy and phosphate for producing nucleotides, can be produced by volcanic processes [YWSN91]; montmorillonite clay has been discovered to catalyze oligonucleotide formation [Fer06]. Of particular importance to the present article, dipeptides can be produced in a dehydrating salt solution [SR89], and longer peptides can be synthesized in the presence of montmorrillonite [RSSB99].

(3) Vesicles, the ancestors of cells, have been shown to have a prebiotic origin [Dea85] as far as their lipid constituents are concerned, and vesicle formation

has been shown to be also catalyzed by montmorillonite [Fer06]. The fact that vesicles containing oligonucleotides [4] have a higher fitness as compared with empty vesicles [CRS04], and the fact that vesicles have a ribose-collecting power [SS05] are indications of the early involvement of vesicles in the process of the origin of life. Also important is the fact that a membrane potential is not necessarily coupled to cells, but can be found in much simpler vesicles [CS04].

(4) Homochiral molecules[5] were selected out of the racemic mixture, by a mechanism which is still not elucidated.

(5) At some stage RNA and peptides became part of the same biochemical system. However, which of the RNA or peptides came first is highly controversial. The arguments in favor of the RNA world include the existence of ribozymes, the essential role of RNA in transcription and translation, and the fact that amino acids are activated by ATP before they are coupled to their cognate tRNAs. An argument against the RNA world is the huge size of the RNA-dependent RNA-polymerizing ribozyme [JUL$^+$01]. The arguments in favor of the peptide world include the fact that peptides are more easily synthesized and more resistant to degradation [PRR05]. Moreover some peptides have been shown to catalyze their own synthesis [LGM$^+$96, SLM$^+$98, SYSG01, UBL$^+$09]. It has also been shown that catalytic dipeptides have been formed with the substrate as a molecular template [KBNF97, FYEBN05]; in this case a condensing agent was added to the experimental mixture. A hybrid picture between these two extremes of an RNA world and a peptide world would be an RNA-peptide world, in which both types of molecules evolved together.

(6) The coding of genetic information in DNA was probably, given the difficulty of synthesizing DNA, one of the last steps[6] before the appearance of life as we know it (see e.g. [FKL99]).

---

[4] If one adds nucleic acids to membrane vesicles and exposes this mixture to wet-dry cycles, vesicles are formed that contain nucleic acids [ADM02]. This physical-chemical phenomenon could be relevant for prebiotic chemistry. It also leads to a question: does the same occur when one mixes short peptides with vesicles and exposes them to wet-dry cycles? The enzymes catalase [ADM02] and polynucleotide phosphorylase [WGM$^+$94] can become encapsulated in vesicles and remain enzymatically active. See also the work of Szostak and co-workers as referred to in section 1.1 and the work of Black and Blosser [BB16].

[5] A possible experiment would be: what happens to the catalytic activity of short peptides and short enzymes if one chemically synthesizes them from racemic mixtures? How does that affect their catalytic activity? However, it should be noted that selection of L-amino acids could be established by short RNAs consisting of D-nucleotides only (see e.g. [Tam15] and references therein); seen in this way, the right question to ask is not "Why are there only L-amino acids?" but is: "Why are there only D-nucleotides?".

[6] One could argue that as a seventh stage the long evolutionary path from the first DNA organism to the last universal common ancestor should be added, but I think this would not be right; this list follows the developments from the presence of prebiotic amino acids to a cell as we know it, with DNA as genetic material. This is the picture of the appearance of life, and the evolution from the first DNA-containing prokaryote to LUCA is another story. This viewpoint is not uncontested: see [WSM$^+$16], but see also [BB16].

Exactly how these different more or less well established facts are related is the subject of much debate and speculation. Here we addressed the issue of the form and functionality of the first peptides. We believe it should be possible to answer the latter question to a large extent independently of how points (1) to (6) fit together. It can also to some extent be addressed independently of how the first peptides were synthesized and how their sequences were coded, and of whether a peptide, RNA, or RNA-peptide world appeared first. Concerning the last point there are two conflicting results which cannot be resolved at present: on the one hand small peptides can be synthesized in prebiotic environments as mentioned in point (2) above, but on the other hand peptide synthesis as it occurs in the cell is deeply connected to the RNA world, and the traces of this connection can still be seen in the essential role in translation of mRNA, rRNA, and tRNA.

Two main facts can be stated with confidence about the first functional peptides: first of all the first peptides were presumably made of the amino acids that are efficiently produced in prebiotic synthesis, as these are the chemically simplest amino acids. A supporting argument is that the present genetic code, with its block structure, could easily have evolved[7] from a more primitive code that used fewer amino acids.

Secondly, the first peptides probably appeared when a coding mechanism for the amino acid sequence was inexistent, or at least highly imperfect and error prone. Indeed the present (highly complex) coding, with its essentially error free mechanisms for duplication, transcription and translation, was almost certainly not present initially, but presumably evolved in response to the increasingly critical role of complex proteins in early life. This is supported by the fact that the genetic code seems to have been optimized against transcription and translation errors [FH98a].

A first question to address when trying to understand what were the first functional peptides is to decide whether they had lengths comparable to present-day protein chains (about 20 to 10000 amino acids long), or whether they were much shorter. We believe that the first possibility is highly unlikely. Indeed long random peptides will not have a well defined 3-dimensional structure and certainly will not have any useful functionality. On the other hand random synthesis of very short peptides could have produced useful ones with reasonably high yield. For instance the common submotif of our prebiotic peptide candidates, DGD, could be one of those randomly selected peptides. In this case the success rate of the synthesis would have been $4^{-3} = 1/64$ (assuming that only the four abundant prebiotic residues were used and were equally abundant). The fraction of active

[7]Similar codons code for similar amino acids (e.g. UUR encodes Leu and UUY encodes Phe), and so an evolutionary development in which codons were transferred from simpler amino acids (e.g. Leu) to more complex amino acids (e.g. Phe) could be envisioned (e.g. originally all UUN codons could have encoded Leu). See also e.g. [Woe65a, Cri68, Won75, LJ88, KFL99, Ike02, Mas06, DG08, Hig09, Fra13] and chapters 3, 4, and 5 from this thesis.

versus possible peptides drops very rapidly as the length increases (at least if all positions have to be occupied by specific residues): for 8 residues it is as small as 4^{-8}. However, if the synthesis occurred in the presence of metal ions[8], which tend to attract free Asp's, the amount of meaningful peptides could be much higher. The prebiotic active sites of 5-8 residues long seem to form the limit between randomly synthesized and template-directed peptides.

To discover what the original functional peptides were, we have adopted here a theoretical approach. We have identified, using the method described above, active sites of present-day proteins that are made of prebiotic amino acids, carry

[8]The importance of metal ions in RNA structure, and also in the function of the oligopeptide sequences contemplated in this chapter, makes one wonder if the presence of metal ions would change the outcome of Miller-type experiments. Traditionally, it is mainly the character of the nitrogen source or the carbon source which has been the focus of interest (see e.g. the literature discussed in section 2.2) in such experiments. It is important to recognize that sulfur and metal ions are essential elements too, of the living cell (see e.g. [FdSW91]; and also [RFJ07]: "[...] it should be pointed out that most of the scientists discussing chemical evolution have taken into consideration only the organic substances [...] and completely neglected the inorganic materials [...]"). When no sulfur has been added in a Miller-type experiment, it is not strange that not all 20 amino acids come out: a sulfur atom is a **part** of some of the twenty amino acids (to be precise: of Cys and of Met). No sulfur, no twenty amino acids. With metal ions, it is not *that* straightforward, but ignoring their possible effects on the reactions should not be the basic attitude. It is interesting that K^+ (see [Fra13]) and Fe^{2+} (see [HCO^+13]) probably have played a more important role in primitive biochemistry than they do in modern biochemistry. Na^+ too, is a metal ion: "A large number of salts are soluble in water, and, therefore, they should have formed a major component of the primordial sea. For example, it is estimated that the content of sodium chloride (NaCl) of the primitive ocean was approximately the same as of nowadays' oceans" [RFJ07]. Apart from the non-reducing circumstances, the experiments of Rode and co-workers, mentioned in [RFJ07], differed from the classical Miller design in a further three fundamental ways: the temperature was higher, the spark was going **into** the liquid phase by placing a copper electrode *under* the liquid surface, and, in some experiments, the water was salt. This is because Rode and co-workers modelled lightning striking into a hot, salty ocean, while Miller modelled lightning between clouds in a reducing atmosphere. An important point is the presence of Cu^{2+} in the hot, salty ocean experiments. Copper was judged to be an abundant component of the prebiotic milieu, based on geological evidence: "[...] inside [...] precambrian rocks, large, so-called 'green zones' were found, containing mostly the copper minerals azurite and malachite" [RFJ07]. In the conclusions of the article reporting the actual hot, salty ocean experiments [PRR06], the following sentences can be found: "Amino acid production works with and without the presence of sodium chloride in the simulated prebiotic ocean, therefore all kinds of water bodies can be considered a relevant site for prebiotic synthesis of amino acids. No real preference for certain conditions could be detected by looking at the yields of amino acids analyzed so far, which shows a very general applicability of this type of prebiotic amino acid formation" and "Altogether it is of importance to note that under up-to-date prebiotic Earth conditions simple and also somewhat more complex amino acids can be formed in a very short amount of time on the geological timescale, and therefore must have been readily available in vast amounts on the prebiotic Earth [...]" Remarkable is that one of the 'more complex' amino acids in these experiments, is lysine. Please note that Higgs and Pudritz too [HP09], pointed out that diverse environments (e.g. a mixture of atmospheric gases or hydrothermal vent circumstances or interstellar gas) are remarkably consistent about which amino acids formed, and that the order of abundance can be predicted by thermodynamics.

out functionalities that are essential for early life, and can be traced back to LUCA. We then take the bold step of supposing that these sequences in fact go back much earlier, and constitute some of the earliest useful peptides. This is of course highly conjectural. If we accept this conjecture, we obtain a remarkably precise picture of the chemical role, the first, very short, oligopeptides could have played during very early life. In all cases we have identified, phosphate groups are manipulated, mainly by Mg^{2+} ions, coordinated by oligopeptide motifs consisting of three Asp residues and at least one Gly residue. Remarkable is the fact that ribozymes also need specifically Mg^{2+} ions for both structure and function [GBB02, TIK04, RS07]. Mg^{2+} ions seem thus to have a particularly important role in handling phosphate groups and therefore in the development of life.

Moreover, in the peptides we identified, the biological functions performed by these active sites, i.e. RNA polymerization, glycolytic intermediate production and sugar phosphorylation, can be assumed as prebiotic, or at the very least, very early biotic. The existence of these early peptides would of course have required the existence of a stable environment, protected from toxic chemical species. This environment could have been provided by the vesicles that correspond to the ancestors of cells (point (3) above). At present we do not have enough information to decide whether these oligopeptide motifs appeared before, during or after the RNA world. In the latter case they could have constituted an interesting improvement to the phosphate chemistry available in the RNA world.

More generally, we can conjecture that, once small random peptides became useful to early life, mechanisms emerged to select for the active peptides. Thus gradually a coding and translation mechanism appeared, which in turn would allow for the synthesis of longer peptides, and for the exploration of the space of possible amino acid sequences, progressively selecting more active and specific oligopeptides and proteins. Therefore, the appearance of prebiotic oligopeptides is tightly connected[9] with the appearance of the genetic code.

[9]The concept is that the earliest genetic codes efficiently produced oligopeptides that were also produced prebiotically, and which played an important role in the first biochemistry. During the development of the genetic code, the peptides could gradually become longer, and the repertoire of consisting amino acids could gradually become richer. In the start the role of the genetic code is supposed to have been a method for more reliable, more focused and more massive delivery of oligopeptides which were also produced by prebiotic processes. Proteins consisting of alfa-helices and beta-strands are elements of a later world, because a peptide needs to be large to make a precise structure with these structure components. Things do not start large, but do start small, and thus the exploration of protein space started with the production of oligopeptides.

2.6 How to validate our findings?

We identified in present-day protein structures five short segments bound mainly to Mg^{2+} ions, and built for 80 % at least from the four amino acids Asp, Gly, Val, and Ala, thought to be the most abundant amino acids in prebiotic times. These motifs appear as statistically significant as they are 2.8 times more frequent than expected by chance, with a P-value of 3×10^{-2}. Moreover, they all form the active sites of their host protein: NADFDGD and GGDYDGD in RNA polymerases, DGDGD and DGDAD in mutases, and DAKVGDGD in dihydroxyacetone kinase, where they manipulate phosphate groups, thought to be an important biological function in the very first stages of life. We thus conjecture that these motifs could correspond to the first functional peptides, at the earliest stages of life, before or at the beginning of the setup of the genetic coding mechanism, and after the purely RNA world or during the RNA-peptide world. These peptides could be viewed as transitional fossils, constituting a missing link between the prebiotic amino acids and the coded proteins.

How could our suggestions be tested experimentally? The easiest point to analyze is whether the short prebiotic peptide candidates are soluble in aqueous solution, and whether they form more or less unique and stable structures with Mg^{2+} ions. The second point to test is whether these short, metal ion-complexed peptides do still have some kind of enzymatic activity, even very inefficient, when they are isolated from their normal protein context. This would possibly require introducing other environments, present in the prebiotic era, which would favor the assembly of the different partners required for enzymatic function. Finally, we could consider peptides built from the abundant prebiotic residues only, for example ADVDGD and DAAVGDGD, which could represent even earlier peptides, and test if their properties are comparable to those of ADFDGD and DAKVGDGD[10]. A further area of research is the interaction of these kinds of oligopeptides with RNA molecules.

[10]In this line, one could also search for precursors for the omnipresent motifs of [SFT07], e.g. one could test VVDDVD and FIDEID, in the case of the omnipresent motifs of course as parts of larger molecules.

Chapter 3

Error minimization in the genetic code

The content of this chapter is based on joint work with Harry Buhrman, Steven Kelk, Wouter Koolen, and Leen Stougie [BvdGK+11].

3.1 Mathematical formulation of genetic code spaces

Although a few small variations on the standard genetic code (SGC) are known (especially in mitochondrial systems), the set of rules is essentially the same for all organisms. The genetic code is therefore one of the most fundamental aspects of biochemistry. The pattern of codon assignments in the genetic code appears to be organized in some way (Table 3.1). First, there is codon similarity for codons encoding the same amino acid. The underlying biochemical reason [Cri66] is (partly) that tRNA molecules often recognize more than one codon. A second phenomenon is that similar amino acids are often specified by similar codons. One way to quantify amino acid similarity is to use the values of polar requirement introduced by Woese et al. [WDD+66]. According to this measure amino acids with a polar side chain like glutamate and aspartate have a high value (12.5 and 13.0, respectively), while hydrophobic amino acids like leucine and valine have a low value (4.9 and 5.6, respectively). An example of similar codons coding for similar amino acids is asparagine, specified by codons AAU and AAC with a polar requirement of 10.0 and lysine, specified by AAA and AAG, with a polar requirement of 10.1. Although one may suspect that similar codons code for similar amino acids may also be present in a random grouping [Cri68], Haig and Hurst [HH91, HH99] showed that this is not the case. Random codes do not have this property to the same extent as the standard genetic code.

Haig and Hurst [HH91] generated by computer a large number of alternative genetic codes, in which the blocks coding for amino acids in the SGC, e.g. the UCU, UCC, UCA, UCG, AGU, AGC block encoding serine, were kept the

UUU Phe (5.0)	UCU Ser (7.5)	UAU Tyr (5.4)	UGU Cys (4.8)
UUC Phe (5.0)	UCC Ser (7.5)	UAC Tyr (5.4)	UGC Cys (4.8)
UUA Leu (4.9)	UCA Ser (7.5)	UAA STOP	UGA STOP
UUG Leu (4.9)	UCG Ser (7.5)	UAG STOP	UGG Trp (5.2)
CUU Leu (4.9)	CCU Pro (6.6)	CAU His (8.4)	CGU Arg (9.1)
CUC Leu (4.9)	CCC Pro (6.6)	CAC His (8.4)	CGC Arg (9.1)
CUA Leu (4.9)	CCA Pro (6.6)	CAA Gln (8.6)	CGA Arg (9.1)
CUG Leu (4.9)	CCG Pro (6.6)	CAG Gln (8.6)	CGG Arg (9.1)
AUU Ile (4.9)	ACU Thr (6.6)	AAU Asn (10.0)	AGU Ser (7.5)
AUC Ile (4.9)	ACC Thr (6.6)	AAC Asn (10.0)	AGC Ser (7.5)
AUA Ile (4.9)	ACA Thr (6.6)	AAA Lys (10.1)	AGA Arg (9.1)
AUG Met (5.3)	ACG Thr (6.6)	AAG Lys (10.1)	AGG Arg (9.1)
GUU Val (5.6)	GCU Ala (7.0)	GAU Asp (13.0)	GGU Gly (7.9)
GUC Val (5.6)	GCC Ala (7.0)	GAC Asp (13.0)	GGC Gly (7.9)
GUA Val (5.6)	GCA Ala (7.0)	GAA Glu (12.5)	GGA Gly (7.9)
GUG Val (5.6)	GCG Ala (7.0)	GAG Glu (12.5)	GGG Gly (7.9)

Table 3.1: The standard genetic code. Assignment of the 64 possible codons to amino acids or stop signals, with polar requirement of the amino acids indicated in brackets.

same, but their assignment to an amino acid was randomly redistributed (a procedure generally called "swapping"). We will refer to this as the *fixed block model* [Gol93]. Note that the use of the word "block" is different from the use in studies such as [NWK07, SYH07]. We use the word "block" as in [Gol93], [FH98a] and [FKLH00]: in the sense of the collection of all codons specifying the same amino acid or chain termination ("STOP" in Table 3.1). We will call the collection of all codons sharing the same first and second nucleotide "box". The space of codes which is created as a result of random code generation under the fixed block model, denoted as Space 0, contains exactly 20! ($\approx 2.433 \times 10^{18}$) codes.

As a measure for the quality of a code the change in polar requirement caused by one step point mutations in the codons was proposed. Each codon has nine codons to which it can mutate in one step: e.g. for the UCU serine codon, these are UCC, UCA, UCG (these three remain coding for serine in the actual code), UUU (coding for phenylalanine, a 2.5 difference in polar requirement), UAU (coding for tyrosine, a 2.1 difference), UGU (coding for cysteine, a 2.7 difference), CCU (coding for proline, a 0.9 difference), ACU (coding for threonine, also a 0.9 difference), and GCU (coding for alanine, a 0.5 difference). The quality of the code is then measured by averaging over all squared differences: MS_0. In this calculation, Haig and Hurst [HH91] ignored the three "stop codons" which are coding for chain termination. In this way, 263 connections between adjacent codons contribute equally to MS_0.

To facilitate the mathematical formulation of MS_0 we introduce an undirected

graph $G = (V, E)$ that has the 61 codons as its vertices and an edge between any two codons if they differ in only one position, yielding 263 edges. Let $G^S = (V^S, E^S)$ be the graph obtained by adding the 3 stop codons to V, yielding 288 edges. A code F maps each codon c to exactly one amino acid $F(c)$. We denote by $r(F(c))$ the polar requirement of the amino acid that codon c encodes for with respect to code F. The error function of code F is then given by

$$MS_0(F) = \frac{1}{263} \sum_{\{c,c'\} \in E} \left(r(F(c)) - r(F(c')) \right)^2 .$$

Using MS_0 as a quality measure of a genetic code (please see also the last paragraph of section 1.2) Haig and Hurst found that only 1 out of 10,000 random codes performs better, i.e. has a lower MS_0, than the standard genetic code [HH99]. This shows that in the standard genetic code not only identical amino acids are encoded by similar codons, but also similar amino acids are encoded by similar codons. Originally, Haig and Hurst [HH91, HH99] investigated three other characteristics beside polar requirement (like e.g. the isoelectric point), but the correspondence between codon assignments and error robustness with respect to polar requirement was most striking. It may be interesting to find other measures which perform equally well, or better. However, the measure has to be independent from the genetic code (this point has been made in connection with the use of values derived from replacement mutations known from sequence data). We have to be careful not to artificially create a measure that is based on the genetic code itself. To keep results comparable to the work of Haig and Hurst, use of polar requirement is preferable.

The work of Haig and Hurst was soon followed by the work of Goldman [Gol93], who found a code using a heuristic method that has a lower MS_0 value than any of the codes generated before. In Section 3.2.1 we verify that Goldman's code is in fact the global optimum in the fixed block model.

Freeland and Hurst [FH98a] presented four histograms to visualise the particular error robustness, in the sense of Haig and Hurst [HH91], of the standard genetic code. They reported that with respect to the MS_0 value, 114 codes out of the 1,000,000 random codes had a lower value than the standard genetic code. They also reported similar results with respect to the MS measure restricted to point mutations in the first, second and third codon, respectively denoted by MS_1, MS_2 and MS_3. To define them we partition the edge set E in the graph representation $G = (V, E)$ of the adjacency structure of codons, depending on the position in which two adjacent codons differ: E_1 is the set of edges between two codons that differ only in the first position, E_2 the set of edges between two codons that differ only in the second position, and E_3 the set of edges between two codons that differ only in the third position. Clearly these sets are disjoint and $E = E_1 \cup E_2 \cup E_3$. Then for $p = 1, 2, 3$,

$$MS_p(F) = \frac{1}{|E_p|} \sum_{\{c,c'\} \in E_p} \left(r(F(c)) - r(F(c'))\right)^2,$$

where $|X|$ denotes the cardinality of X i.e. the number of elements in X. In fact, $|E_1| = 87$, $|E_2| = 88$ and $|E_3| = 88$. The results of Freeland and Hurst show that there is not much error robustness for mutations in the middle position of the codon; the third position, however, is extremely robust against changes in polar requirement.

Subsequent research following this approach has concentrated on nuancing the error function [FH98a, Ard98, FKLH00, GMCR01, Hig09] or taking a parameter different from polar requirement as an amino acid characteristic [Ard98, FKLH00, GMCR01, Hig09]. The common theme in most of these approaches is the code space from which random alternative codes are generated; in [FKLH00] this space is referred to as "possible code space" and we denote this space as Space 0. Remarkably, known genetic code variations lie *outside* Space 0. In code variations certain individual codons are *reallocated* from one block to another. The fixed block structure of the standard genetic code is thereby replaced by an alternative, slightly different, fixed block structure. In Section 3.2.3 we construct four progressively larger code spaces (denoted Space 1, Space 2, Space 3 and Space 4), which encompass successively more known genetic code variations next to the standard genetic code. To be able to compare the genetic code with respect to alternative codes sampled randomly from Spaces 1 and 2, we nuance the MS measure such as to accommodate values of polar requirement for stop codons. In this paper, we aim at refining several points in the seminal work by Haig and Hurst [HH91, HH99], Goldman [Gol93] and Freeland and Hurst [FH98a]. Apart from determining the global minimum, the refinements concern the code space structure and the kind of conclusions assumed to be possible to draw based on the research. We do not intend to change the characteristic taken to represent the amino acid (which is polar requirement in the work of Haig and Hurst [HH91, HH99] and Goldman [Gol93]) or to weigh the three positions of the codons differently in the error function (as is done in the second part of [FH98a] and most subsequent work). We only intend to enlarge the space from which random codes are sampled, and find out how they relate to [FH98a].

3.2 The global minimum and four larger spaces

3.2.1 Goldman's best solution is the global minimum

Goldman [Gol93] applied a heuristic algorithm for finding the best code under the *fixed block model*. The best solution he found had an MS_0 value of 3.489, which

was well below the value of 5.194 reported by Haig and Hurst [HH91, HH99] for the standard genetic code. A heuristic does not guarantee that the code found is optimal. We designed an exact method for finding the optimal code by formulating the minimization problem as a Quadratic Assignment Problem (QAP) [Cel98] and solved it using the exact QAP-solver QAPBB [QAP]. An intuitive formulation of QAP is as follows. We are given two sets of objects V_1 and V_2 of equal size. We are to match each object from V_1 to exactly one object from V_2 such that all objects of V_2 are matched as well; as a result we get a perfect matching (pairing) of the objects of V_1 and V_2. In the ordinary (linear) assignment problem, there is a cost for assigning object i from V_1 to object k from V_2 and we wish to find the assignment that minimizes total cost. In QAP the cost is dependent on pairs of assignments: there is a cost for assigning object i from V_1 to object k from V_2 *and* object j from V_1 to object ℓ from V_2. Again we wish to minimize the total assignment cost.

If we consider the set of objects V_1 to be the 20 blocks in Table 3.1, and the set of objects V_2 to be the 20 amino acids, then we can model the minimization of MS_0 by letting the cost of assigning one amino acid to one block and another amino acid to another block be given by the difference of their polar requirements times the number of point mutations between the two blocks.

Small instances of QAP can be solved effectively using an exhaustive enumeration technique known as *branch and bound* [LdABN$^+$07]. This searches (implicitly) through the entire space of solutions, keeping note of the best solution found so far, and ignoring parts of the solution space that could not possibly lead to a better solution. Even with branch and bound it is in general not feasible to use the QAP model for finding a code with minimum MS_0 value in any reasonable time when we leave the *fixed block model*. However, we could find the global minimum MS_0 value in Space 0. We found the same solution as Goldman, certifying that his solution was in fact the optimal one.

3.2.2 Incorporating stop codons

Leaving the fixed block model required us to nuance the MS measure and attach a value of polar requirement to the stop signal. Chain termination is produced by Release Factors (RFs), which are proteins, and therefore most probably later elements of the coding system than tRNAs. This is an argument which can also be found in e.g. [Hig09] ("... I do not want to assume that there were stop codons in the current positions from the beginning, because it is more likely that stop codons were a late addition to the code, after the main layout of most of the codons was already established"). Genetic codes lacking stop codons are not impossible. During the evolutionary development of the genetic code, mRNAs could have been short, and the last sense codon of a message could have been the end of the mRNA. After attaching the last amino acid of the polypeptide, the primordial

ribosome could move further along the mRNA, and both the polypeptide and the mRNA could lose the association with the ribosome, as the tape leaving the tape recorder in the classical analogy. The more sophisticated mechanism with Release Factors could have evolved later, to make things run more smoothly. When this is the scenario of evolution of chain termination we follow, we want the stop codons to have the smallest influence on our calculations possible.

How to assign values to stop codons?

There are at least four possible ways to deal with the stop codons. In the work described in Section 3.1 the stop codons were ignored and no value was assigned to them. A second way to deal with stop codons is to assign a fixed value to a stop codon. A third way would be to assign a fixed value to the mutation to a stop codon, which would be the same for all amino acids. The last way to deal with the problem would be to mimic the natural process of suppression.

Assigning no value to stop codons

Ignoring stop codons in the calculation as has been done until now [HH91, Gol93, FH98a] is not the way in which their influence is the smallest possible. This is because they eliminate a lot of the edges from G^S. For the UCA serine codon, in the previous treatment only the edges to UCU, UCC, UCG, UUA, CCA, ACA and GCA take part in the calculation. The edges to UAA and UGA are ignored, which means in fact that they behave towards serine as if those codons were encoding serine. Due to this effect, the four alanine codons have a stronger influence on the calculation than the four glycine codons. Thus ignoring stop codons artificially favors certain amino acids. This effect will even become more pronounced when we enlarge the space of possible codes. For example, if we allow codes to have as many as four stop codons (like our mitochondrial code), or to have stop codons in unusual places (like the UUA and UUG stop codons of the mitochondria of *Pycnococcus provasolii* [TOL10]).

Assigning a fixed value to a node (i.e. give the stop codon a fixed value)

If we were to reason that a mutation to a stop codon would lead to truncation of messages, we might be inclined to attach a very large value to a stop codon (because truncated proteins would be non-functional and the mutation therefore lethal). To model "lethal", we could use the value "infinity". This makes our calculation useless. We could also attach a polar requirement of 1,000,000 to a stop codon. In this case the stop codons are going to dominate the calculation and this is exactly what we didn't want to begin with.

Assigning a fixed value to an edge (i.e. give the mutation to a stop codon a fixed value)

There is another way to model the concept that a mutation to a stop codon is worse than a mutation to a sense codon. One could assign a fixed penalty to a mutation to "stop", no matter which amino acid is mutated to stop. One relatively large value which could be given as a penalty is the difference in polar requirement between the two most dissimilar amino acids. The disadvantage of this approach is again the domination of the calculation by the stop mutations. Although less dominating than the very high fixed values suggested for the stop nodes, this approach still has the stop codons dominating the calculation, and possibly obscuring the phenomenon we want to see.

The suppression approach

What would happen if there is a mRNA with a codon which does not have a tRNA? In such a case, one possibility is that decoding is performed by the tRNA which, among the tRNA repertoire present in the system in consideration, is the most similar to the one which would be needed to decode the codon regularly. This phenomenon is called "suppression" in molecular biology [Lew08a]. In the living cell, the cognate tRNA or RF competes with several different potential suppressor tRNAs for decoding a codon [KF07, SSV12]. By using in the calculation the value which would be there in case of the most probable suppression[1], a value is attached to a stop codon which results in a relatively small influence of the stop codons in the calculation. The most probable suppression for a stop codon ending on A, is by the tRNA which recognizes the sense codon ending on G from the same box. This is reflected by genetic code variants: apparently suppressing tRNAs often evolve towards full recognition. We can illustrate this with the UGA codon, which can be found in the top right-hand corner of Table 3.1. Because the most probable suppression for UGA is by the tRNA which normally reads UGG as tryptophan, genetic code variants in which both UGA and UGG encode tryptophan evolved multiple times. Although there exists an organism in which UGA is encoding cysteine, the more frequent reassignment for UGA is to tryptophan. The same phenomenon is found for AUA, which can be found towards the bottom left-hand side of Table 3.1. AUA has been reassigned several times to methionine. Suppression of AUA codons in protein coding sequences by the tRNA which is normally reading the AUG codons has apparently been followed by the evolution of full recognition of the AUA codon by this tRNA. Assigning to a stop codon ending on a purine (A or G) the value of polar requirement of the amino acid specified by the other purine-ending codon in the box is therefore a possible way to deal with stop codons. This obviously can not be done when both purine-ending

[1]It should be realized that Alff-Steinberger [AS69] did already in 1969 used the suppression approach for Ochre and Amber.

codons in a box are stop codons. Genetic code variants suggest an approach also in this case. In bilaterian mitochondria the tRNA which decodes AGA and AGG (recall Table 3.1, the AGA and AGG codons can be found towards the bottom right-hand side) as arginine in the standard code is not present. The tRNA which decodes AGU and AGC as serine usually takes over the function of decoding AGA and AGG by reading them as serine [SYH07]. This suggests the approach: if in one box both purine-ending codons are stop codons, the value of polar requirement of the amino acid specified by the codons ending on a pyrimidine (U or C) in that box can be assigned to them. This is always a single amino acid because the two pyrimidine-ending codons in the same box always code for the same amino acid. Until now, no genetic code variants are discovered with pyrimidine-ending stop codons, so our approach is to develop only a way to deal with stop codons ending on purines.

How to modify the MS measure?

By treating the stop codons as sense codons according to the suppression approach, we simplified the MS measure. In the notation introduced before,

$$MS_0^S(F) = \frac{1}{|E^S|} \sum_{\{c,c'\}\in E^S} \left(r(F(c)) - r(F(c'))\right)^2,$$

and similarly with respect to the three positions $p = 1, 2, 3$ of the codons

$$MS_p^S(F) = \frac{1}{|E_p^S|} \sum_{\{c,c'\}\in E_p^S} \left(r(F(c)) - r(F(c'))\right)^2.$$

In this way, all 64 codons contribute equally to the error measure. Note that $|E^S| = 288$ and that $|E_1^S| = |E_2^S| = |E_3^S| = 96$. It should be realized that by using MS_0^S or MS_p^S we do not necessarily start working in a space larger than Space 0. We can use MS_0^S and MS_p^S when we generate random codes from Space 1 or Space 2 (see Section 3.2.3) but we can also use MS_0^S and MS_p^S when we generate random codes from Space 0.

We investigate how the new measure reflects the nature of Space 0 when used as a background to study the SGC (Table 3.1). We produce four plots as in [FH98a]. The plots (Figure 3.1) have the same general shape as the four plots in [FH98a]. In particular, the prominent shoulder at the left side is present in both the MS_3^S (Figure 3.1.(d)) and the MS_3 [FH98a] frequency distributions. The spikes present in the plots in [FH98a] are not present. They are an artifact of rounding errors in both the data and the bin borders of the histograms. The combination of MS values rounded to two digits after the decimal point and bin border values which are repetitive binary fractions rounded by the histogram software, are probably the source of the spikes in [FH98a].

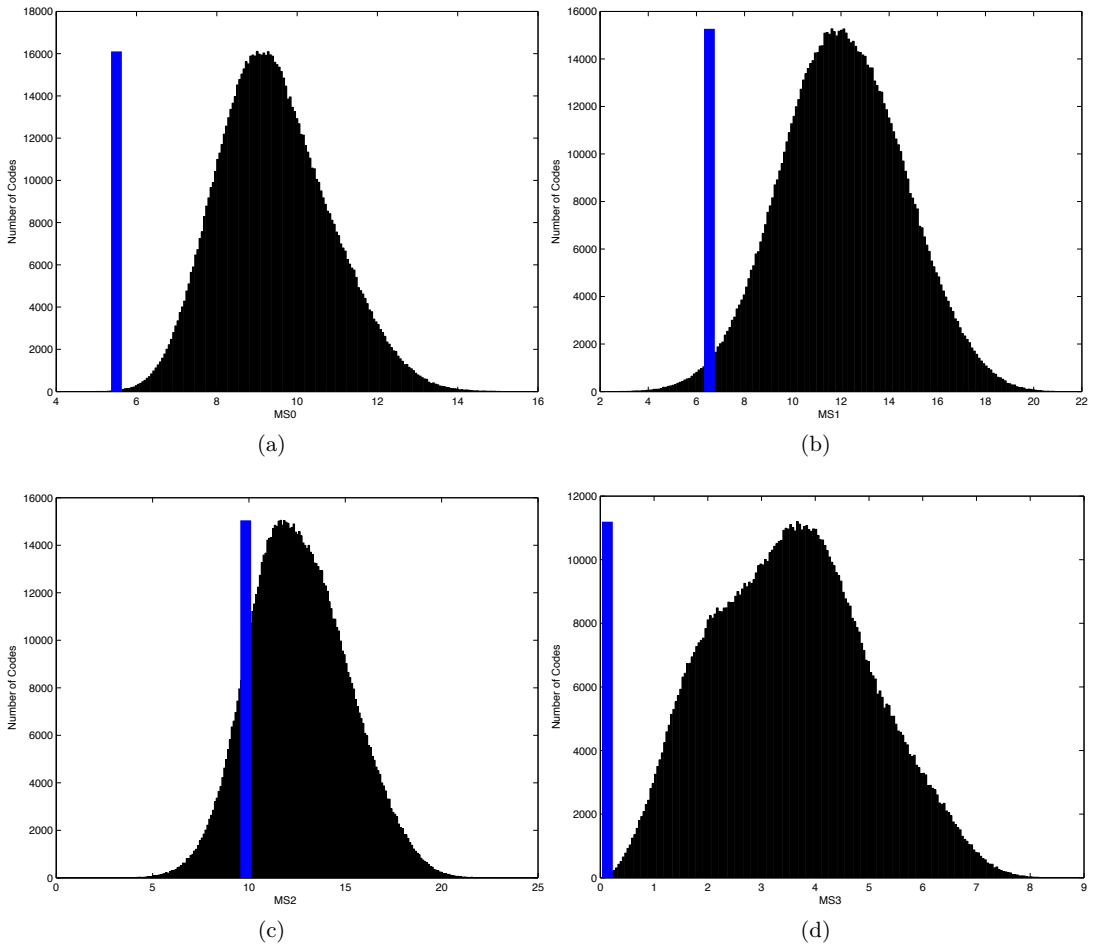


Figure 3.1: Histograms for the MS values obtained from codes randomly sampled from Space 0. MS value of the standard genetic code indicated by the blue bar. 10^6 samples. The MS measure was slightly modified in comparison to earlier work. The modification does not change the basic characteristics found there. (a) MS_0^S (b) MS_1^S (c) MS_2^S (d) MS_3^S.

The global minimum code in Space 0 for the MS_0^S measure was also found using the quadratic assignment approach described in Section 3.2.1. We calculated the average of both MS_0 and MS_0^S of 1,000,000 randomly generated codes as well as the global minimum in Space 0 with respect to both measures (Table 3.2).

Clearly, both measures give similar results. We also studied the proportions of random codes better than the standard genetic code with the MS_0^S measure. Out of 1,000,000 random codes 156 codes had a lower MS_0^S-value than the standard genetic code, resulting in a proportion P_0^S of 0.000156. This was also investigated for $p = 1, 2$ and 3 (Table 3.3). Again the MS and MS^S measures give similar

	MS_0	MS_0^S
Mean of random codes	9.41 ± 1.51	9.43 ± 1.89
Standard genetic code (rounded)	5.194	5.501
Global minimum code (rounded)	3.489	3.946

Table 3.2: Comparison of MS_0 and MS_0^S. Values were calculated for 10^6 randomly sampled codes from Space 0. The averages and variance are shown; MS_0 is taken from [FH98a].

results (as can be detected also from the plots of Figure 3.1).

MS	MS^S
$P_0 = 0.000114$	$P_0^S = 0.000156$
$P_1 = 0.002964$	$P_1^S = 0.012369$
$P_2 = 0.221633$	$P_2^S = 0.129075$
$P_3 = 0.000088$	$P_3^S = 0.000078$

Table 3.3: Comparison of proportions of "better codes" for MS and MS^S.

We conclude that it is acceptable to replace MS by MS^S to study the character of the standard genetic code compared to randomly generated ones. MS^S gives the same results in all essential aspects, and can be used to investigate larger spaces and spaces with different codons used as chain termination signal.

3.2.3 Enlarging the "possible code space"

Space 0 has a fixed block structure. It is possible to leave this fixed block structure and generate randomly genetic codes, without relaxing all biochemical constraints. In this section we develop a method to enlarge the space from which codes are sampled randomly, by specification of allowed subdivision of boxes.

Space 0 ("possible code space") does not even cover all existing genetic codes: the only existing genetic code present in Space 0 is the SGC. By studying code variants general rules with respect to the possible ways to construct a genetic code can be found. Using these rules, we enlarge the code space progressively. Genetic code variants are derived from the SGC, as can be concluded by studying the codon assignments of close relatives. For mitochondrial code variants this is recently described in [SYH07]. The number of code variants apart from mitochondria is very small and it is nowadays believed that they all are derived from the standard code (although this was less clear when the very first variants were discovered). Although these variants probably emerged after the SGC, we use the larger spaces because they contain possible ways for constructing genetic codes with the system found in living organisms on Earth.

In the SGC, the box in the top left-hand corner (see Table 3.1) shows one of several ways in which a box can be subdivided according to the codon-anticodon pairing patterns allowed by the simple wobble rules [Cri66, BTS07b]. The codons UUU and UUC are assigned to one amino acid, and the codons UUA and UUG to another. Recognition of both pyrimidine-ending codons by one tRNA molecule is the wobbling behavior of G in the first position of the anticodon as proposed by Crick [Cri66]. Modification of U (in the first position of the anticodon) to thio-U restricts the wobbling behavior of the tRNA molecule to recognition of both purine-ending codons [Tak06, NIF$^+$06]. A second pattern of subdivision is presented by a box towards the bottom left-hand side of Table 3.1. In this box AUU, AUC and AUA are assigned to one amino acid and AUG is assigned to another. The existence is known of tRNA molecules which recognize all three codons in the top of a box [JEH$^+$08]. Recognition of the G-ending codon only, is the wobbling behavior of C in the first position of the anticodon as proposed by Crick [Cri66]. Therefore, this pattern of subdivision of a box can be understood by the pairing characteristics of tRNA molecules. In eight boxes of Table 3.1 all four codons are assigned to one amino acid, as in the box in the bottom right-hand corner: GGU, GGC, GGA and GGG are assigned to the same amino acid. Recognition of all four codons of a box seems to be the wobbling behavior of a tRNA molecule with unmodified U in the first anticodon position: "An unusual situation exists in (at least) mammalian mitochondria, however, in which there are only twenty-two different tRNAs. How does this limited set of tRNAs accomodate all codons? The critical feature lies in a simplification of codon-anticodon pairing, in which one tRNA recognizes all four members of a codon family. [...] In all eight codon families, the sequence of the tRNA contains an unmodified U at the first position of the anticodon." [Lew08b]. Takai [Tak06] provides more information on these tRNAs which recognize all four members of a box: "... many tRNAs with U(34) from mitochondria and mycoplasmas recognize all of the four different codons with the first two bases in common without discriminating the third bases [AYMO89, BAB$^+$80, BBC$^+$80, HSAD$^+$80, SABL83]. This type of codon reading was first considered to be dependent on the 'two-out-of-three mechanism', by which the N(III) bases were ignored when the interaction of the first two positions of the codon with the last two positions of the anticodon is strong enough to support codon recognition by itself [Lag78, Lag81]. However, significant level of discrimination was later observed even in such ambiguous codon reading [IKB$^+$95]." In summary, the wobbling behavior of tRNA molecules with unmodified anticodons allows subdivision of boxes with only sense codons in three ways: no subdivision, division in a pyrimidine-ending pair and a purine-ending pair, and division in a set of three codons in the top of a box, and a single codon at the bottom. Although extensive modifications of anticodons in contemporary organisms can lead to much more complex patterns of wobbling behavior [MG02, TY03, Tak06, JEH$^+$08], for the purpose of enlarging Space 0 we do not take these aspects of the wobble phenomenon in account.

These modifications are produced by proteins, and therefore were probably not present during the development of the coding system. To allow the modifications of U to thio-U (enabling the exclusive recognition of purine-ending codons) and A to I (enabling the recognition of three codons by one tRNA molecule) is already pushing the limit concerning capacities credibly attributable to a very early living system.

Further subdivisions of boxes are possible when stop codons are added to the possibilities in a box. Because stop codons ending on pyrimidines are not discovered yet, we restrict the possibilities to purine-ending stop codons only. This adds four further ways to subdivide a box. The upper two codons assigned to an amino acid, and the lower two codons being stop codons is the first. The upper three codons assigned to one amino acid, and the bottom codon being a stop codon is the second. The upper two codons assigned to one amino acid, the third codon being a stop codon, and the last codon assigned to an amino acid, but different from the amino acid assigned to the upper two codons, is the third possibility. The last possibility again has the third codon being a stop codon, but the three remaining codons are assigned to the same amino acid in this case. Taken together with the three possibilities for subdivision with only sense codons presented in the previous paragraph, we arrive at seven possible ways to subdivide a box according to the simple wobbling behaviour without extensive anticodon modification. This is summarized in Table 3.4. We generate block structures uniformly at random according to the rules described in Table 3.4, the block structures consist of 21 blocks.

Box	Meaning
AAAA	All 4 codons recognized by the same tRNA (or by several tRNAs carrying the same amino acid).
AAAB	NNU, NNC, NNA recognized by one tRNA, NNG recognized by another tRNA carrying a different kind of amino acid.
AAAS	NNU, NNC, NNA recognized by a tRNA, NNG by a Release Factor (RF).
AABB	NNU, NNC recognized by one tRNA, NNA, NNG by another tRNA carrying a different kind of amino acid.
AASA	NNU, NNC recognized by one tRNA, NNA by a RF, NNG by another tRNA, but carrying the same amino acid.
AASB	NNU, NNC recognized by one tRNA, NNA by a RF, NNG by another tRNA, carrying a different kind of amino acid.
AASS	NNU, NNC recognized by one tRNA, NNA, NNG recognized by a RF.

Table 3.4: Possible types of boxes. a = amino acid. b = amino acid, different from a. s = stop.

In our first extension, the "stop block" consists of three stop codons, as in the SGC. However, their location is free, under the condition that they do not end in U or C. The number of codons allocated to any amino acid is free, as long as each amino acid is encoded by at least one codon. In this way we obtain a first enlarged space, Space 1, that is more realistic than Space 0. Space 1 is, with approximately 5.908×10^{45} possible codes, much larger than Space 0 (with approximately 2.433×10^{18} codes). A genetic code variant which is present in Space 1, and not in Space 0, is the code variant with seven Ser codons, and only five Leu codons [SN99].

To include most existing genetic code variations, which differ in the number of stop codons, we enlarged Space 1 to Space 2, by allowing the codes to have $0 - 4$ stop codons. A genetic code variant which is present in Space 2, but not in Space 0 or Space 1, is the code variant with two Trp codons, and only two stop codons (in [KFL01] the several independent emergences of this variant are beautifully presented in their Fig. 2).

For completeness, we also define two more spaces[2] but we will not use them in our calculations. In some bacteria some codons are not used: neither tRNAs nor release factors to recognize them (without suppression) are present. To include these code variations too we in addition add a new block "unassigned" to our block structure, allowing the number of unassigned codons to range between 0 and 40 (Space 3). Every codon is allowed to be unassigned, with the restriction that codons ending on U or C are either both assigned or both unassigned. Space 3 contains all existing natural genetic code variations. A genetic code variant which is present in Space 3, but not in Spaces 0, 1, or 2, is the variant present in certain *Micrococcus* bacteria, in which two A-ending codons are not present [KOAO93].

Finally (Space 4) we also include codes with fewer or more than 20 amino acids. In many speculations on the origin of the genetic code, codes with less than 20 amino acids play a role; Jukes suggested such an evolutionary pathway already in 1966 [Juk66]. With the extreme of just one codon in use, the number of unassigned codons ranges from 0 to 63. The size of Space 4 is approximately 1.120×10^{50} codes. New code variants developed in synthetic biology (see e.g. [LS10]) will normally also be part of Space 4.

The sizes of Spaces 0 - 4 are presented in Table 3.5. The presence of unassigned codons in Spaces 3 and 4 causes the function MS^S to be ill-defined. Therefore we could not investigate the nature of these spaces, as we will do for Spaces 1 and 2.

Figure 3.2 shows four plots (as in Figure 3.1) of MS^S-values, but of codes

[2]Because the fourfold degenerate codon boxes of the SGC are an ancient characteristic (as will become clear in the next chapter), one more space is important: the one in which these 8 codon boxes are always fourfold degenerate while the variation shown in Table 3.4 is allowed in the 8 remaining codon boxes. The size of this space was calculated by Wouter Koolen (in 2018) to be 1.737×10^{30} codes, or to be precise: 1737152665468773186346352640000. We called this space Space $\frac{1}{2}$.

Space	Characteristics of codes	Approximate size of space
Space 0	21 blocks, 20 amino acids, 3 stop codons, 0 unassigned codons	2.433×10^{18}
Space 1	21 blocks, 20 amino acids, 3 stop codons, 0 unassigned codons, free block structure	5.908×10^{45}
Space 2	20-21 blocks, 20 amino acids, 0-4 stop codons, 0 unassigned codons, free block structure	1.932×10^{46}
Space 3	20-22 blocks, 20 amino acids, 0-4 stop codons, 0-40 unassigned codons, free block structure	8.635×10^{48}
Space 4	2-34 blocks, 1-32 amino acids, 0-4 stop codons, 0-63 unassigned codons, free block structure	1.120×10^{50}

Table 3.5: Sizes and characteristics of the five progressively larger spaces. Number of codes present in Spaces $0-4$. The block structure of Spaces $1-4$ is free, except for the constraints imposed by adherence to the Wobble Rules, and the specifications listed under "Characteristics of codes".

sampled from Space 1 rather than Space 0. We notice the great similarity with the plots in Figure 3.1. Despite the fact that Space 1 is about 2×10^{27} times larger than Space 0, the mean MS_0^S-value is still about 10. The frequency distributions have the same general nature, and the position of the frequency distribution relative to that of the SGC has not changed. We also notice that the prominent shoulder at the left side of the MS_3^S frequency distribution in Figure 3.1 has disappeared. We conjecture that the particular block structure of the SGC is responsible for this shoulder.

Figure 3.3 shows the same four plots for Space 2. It is hard to find differences with Figure 3.2. The genetic code seems a bit more special against the background with progressively larger spaces: the number of "better codes" found with a million randomly generated codes decreased from 156 in Space 0, via 7 in Space 1, to just a single one (Table 3.6) in Space 2.

3.3 Implications for genetic code evolution

We now compare five published possible scenarios concerning the evolution of the genetic code and show they are not inconsistent with low MS values.

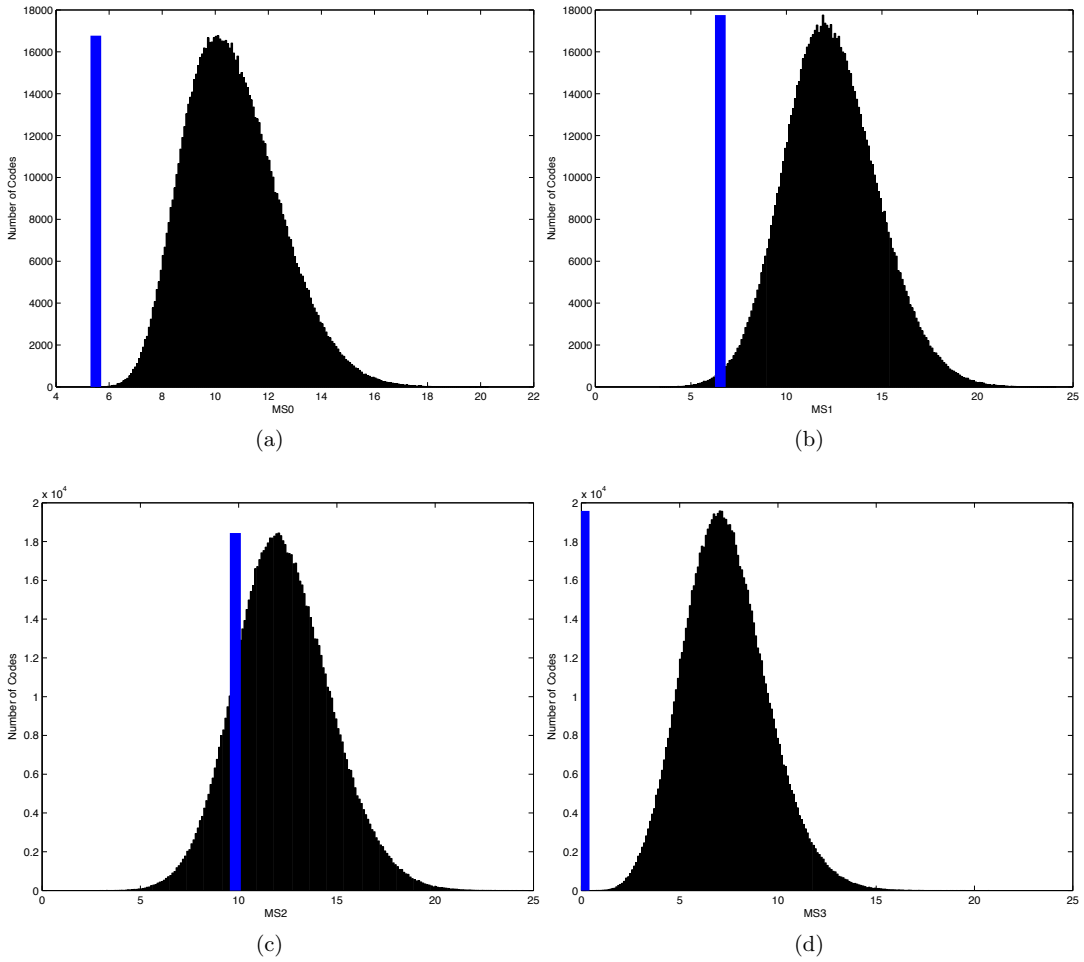


Figure 3.2: Histograms for the MS values obtained from codes randomly sampled from Space 1. MS value of the standard genetic code indicated by the blue bar. 10^6 samples. The modified MS measure was used to calculate a MS value because the random redistribution of the three stop codons made the use of the MS measure from earlier work impossible. The distribution of randomly generated codes is more regular with respect to changes in the third codon position compared with that distribution resulting from codes sampled from Space 0 (shown in Figure 3.1). (a) MS_0^S (b) MS_1^S (c) MS_2^S (d) MS_3^S.

3.3.1 Selection for error minimization

The concept that the codon assignments are a feature of living organisms which protects them against damage to the genetic information and which is, as such, specifically selected for by natural selection, was first published by Sonneborn [Son65]. Woese [Woe65a] elaborated on this concept by pointing out that it is much more probable that translation errors instead of mutations in the genome were the errors against which the system in which the genetic code was developed

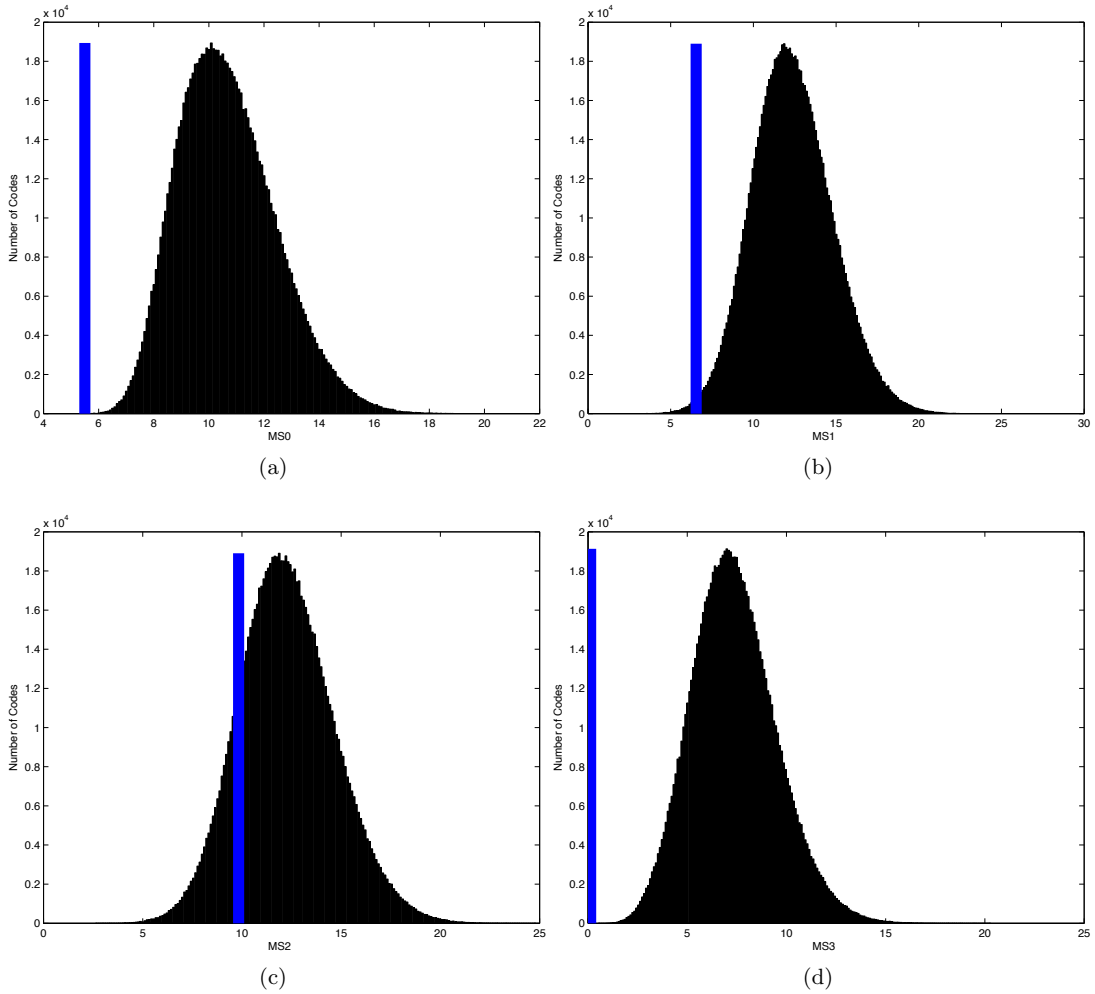


Figure 3.3: Histograms for the MS values obtained from codes randomly sampled from Space 2. MS value of the standard genetic code indicated by the blue bar. 10^6 samples. The modified MS measure was used to calculate a MS value because random redistribution of $0-4$ stop codons made the use of the MS measure from earlier work impossible. The behavior of the distributions is virtually the same as that sampled from Space 1 and shown in the previous figure. (a) MS_0^S (b) MS_1^S (c) MS_2^S (d) MS_3^S.

had to be protected. The concept and first experiment of producing computer-generated random codes to compare with the genetic code was published by Alff-Steinberger [AS69]. This author points out that the differences found depending on the codon position suggest translation errors rather than mutations as responsible for determining (in part!) the structure of the code. Haig and Hurst [HH91] developed the MS measure and were able to generate much more random codes than Alff-Steinberger. They again found differences depending on codon position, but left the possibility open, that "... the code acquired its major features before

Measure	Space 0	Space 1	Space 2
Mean ± variance			
MS_0^S	9.426 ± 1.89	10.663 ± 3.13	10.665 ± 3.12
MS_1^S	12.100 ± 6.37	12.362 ± 5.88	12.368 ± 5.86
MS_2^S	12.627 ± 6.33	12.270 ± 5.79	12.278 ± 5.79
MS_3^S	3.550 ± 2.09	7.358 ± 4.51	7.348 ± 4.49
Proportion of better codes found			
P_0^S	0.000156	0.000007	0.000001
P_1^S	0.012369	0.004853	0.004864
P_2^S	0.129075	0.151506	0.150269
P_3^S	0.000078	0.000000	0.000000

Table 3.6: Basic descriptive statistics of Space 0, Space 1 and Space 2. From each space 10^6 codes were randomly sampled.

the evolution of proteins" [HH91], implying that selection for protection against errors in protein-coding messages maybe played no role in the evolution of the genetic code. Freeland and Hurst [FH98a] elaborated on the work of Haig and Hurst, and presented the code as "one in a million": "We thus conclude not only that the natural genetic code is extremely efficient at minimizing the effects of errors, but also that its structure reflects biases in these errors, as might be expected were the code the product of selection" [FH98a]. The extreme version of the "Error Minimization Hypothesis" would be that all possible codes were tested by natural selection, and the standard genetic code was the best. With a measure which would be a good model for the errors against which the genetic code was optimized, the standard genetic code would then be found to be the global minimum code. There probably are no scientists who adhere to such an extreme variant of the "Error Minimization Hypothesis". It is, however, tempting to see the low MS_0 value as an indication that specific selection for error minimization was a major determinant of the codon assignments in the standard genetic code (e.g. [FWK03]).

3.3.2 The Sequential "2-1-3" Model

Figures 3.1, 3.2, and 3.3 show that the main result of [FH98a] remains valid when Space 0 is enlarged to Space 1, and subsequently to Space 2: the MS value of the SGC is *better* than the MS value of the average code when point mutations in the *second* position are considered; it is *much better* when point mutations in the *first* position are considered; and it is *so much* better when point mutations in the *third* position are considered that better codes in this respect are not visible in the graphs. This could point to the chronological order in which the codon positions

acquired coding information. Massey [Mas06, Mas08, Mas10] published a series of papers in which the sequential acquisition of coding information by the second, then the first, and finally the third codon position is the major determinant of the codon assignments in the standard genetic code. According to this "2-1-3" model, the genetic code started with full degeneracy in the side positions. The amino acid repertoire would originally have been limited to four amino acids, and coding information was carried by the middle position. Subsequently the amino acid repertoire was expanded by assigning coding information to the first position. Because the code expansion would be "...facilitated by duplication of the genes encoding adaptor molecules and charging enzymes" [Mas08], amino acids of similar properties would be assigned to codons with the same middle nucleotide. Selection on error minimization plays a limited role in the "2-1-3" model in so far that code expansion via duplication of adaptor molecules followed by mutation of the middle position of the anticodon is selected against. Hence: "... amino acids of similar properties were selectively assigned to codons separated from one another by a single potential mutation" [Mas06]. Finally, a further expansion was possible by assigning coding information to the third codon position. A consideration of the structure of the tRNA anticodon leads Massey to conclude that the third codon position is intrinsically the most error-prone. Therefore it is logical that distinguishing codons unambiguously on the third position is only possible when protein biochemistry has already progressed beyond the initial stages. Massey states that his analyses "...demonstrate that a substantial proportion of error minimization is likely to have arisen neutrally, simply as a consequence of code expansion, facilitated by duplication of the genes encoding adaptor molecules and charging enzymes. This implies that selection is at best only partly responsible for the property of error minimization" [Mas08]. The concept of a genetic code in which coding information was carried by the middle position only, has been around since the sixties (e.g. with Crick: "For example, only the middle base of a triplet may have been recognized, a U in that position standing for any of a number of hydrophobic amino acids, an A for an acidic one etc." [Cri68]). The "2-1-3" model, however, goes further than that: it presents the chronological order in which the codon positions acquired coding information as the major determinant of the error minimization present in the code. The low MS_0 value is not incompatible with the "2-1-3" model; to the contrary, the "2-1-3" model is based on the low MS_0 value.

3.3.3 The Frozen Accident Theory

A third scenario is the Frozen Accident Theory of Crick [Cri68]. In this scenario, "... the actual allocation of amino acid to codons is mainly accidental and yet related amino acids would be expected to have related codons" [Cri68]. This is because there "...are several reasons why one might expect [...] a substitution of one amino acid for another to take place between structurally similar amino

acids. First, [...] such a resemblance would diminish the bad effects of the initial substitution. Second, the new tRNA would probably start as a gene duplication of the existing tRNA for those codons. Moreover, the new activating enzyme might well be a modification of the existing activating enzyme. This again might be easier if the amino acids were related. Thus, the net effect of a whole series of such changes would be that *similar amino acids would tend to have similar codons*, which is just what we observe in the present code" [Cri68]. Please note that in text preceding this fragment the possibility has been raised that "... the primitive tRNA was its own activating enzyme" [Cri68], which is a description of a ribozyme *avant la lettre*. At a certain moment the system would reach a stage in which "... more and more proteins would be coded and their design would become more sophisticated until eventually one would reach a point where no new amino acid could be introduced without disrupting too many proteins. At this stage the code would be frozen" [Cri68]. Please note that on the very first page of the paper the possibility is mentioned that the genetic code is not exactly identical for all organisms, although for widely different organisms it had been found to be very similar. Therefore the word "frozen" was probably from the start meant to be interpreted with a small degree of flexibility. The concept "relatedness" of amino acids is not rigorously defined in the paper, but Crick presents three examples of what he considers to be groups of related amino acids. "All codons with U in the second place code for hydrophobic amino acids". The polar requirements of this specific group of hydrophobic amino acids are 5.0, 4.9, again 4.9, 5.3, and 5.6. A second group of "related" amino acids is described in: "The basic and acidic amino acids are all grouped near together towards the bottom right-hand side ..." The polar requirements of this group of charged (and thus hydrophilic) amino acids are 10.1, 9.1, 13.0, and 12.5. The third example is the group of aromatic amino acids: "Phenylalanine, tyrosine and tryptophan all have codons starting with U". The polar requirements of these are 5.0, 5.4, and 5.2. Because "related" amino acids according to Crick tend to share a similar polar requirement, the low MS_0 value is not incompatible with the "Frozen Accident Theory". A clear difference between the "2-1-3" model and the "Frozen Accident Theory" is the presence of pairs of "related" amino acids with a second position difference in the latter: e.g. lysine-arginine, and phenylalanine-tyrosine. In this respect, it is relevant to observe that the MS value of the genetic code is lower than the MS value of the average code when point mutations in the second position are considered. Both the "2-1-3" model and the "Frozen Accident Theory" are scenarios in which the genetic code is basically a piece of historical information. Differences between these two scenarios are a lack of emphasis on sequential acquisition of coding information for the different codon positions in Crick's scenario; and a "refusal" by Crick to have a role for specific selection for error minimization in the scenario: "There is no reason to believe, however, that the present code is the best possible, and it could have easily reached its present form by a sequence of happy accidents. In other words, it may not be the result of trying all possible codes and selecting

the best. Instead, it may be frozen at a local minimum which it has reached by a rather random path" [Cri68].

3.3.4 The Stereochemical Theory

A fourth scenario is what Crick named "The Stereochemical Theory" [Cri68]. According to this scenario there is a physico-chemical relationship between certain nucleic acid triplets and certain amino acids. The first such proposal was published by Gamow [Gam54]. Woese spent a lot of effort collecting evidence for the support of the Stereochemical Theory [Woe65b, Woe65a, WDSD66, WDD$^+$66, Woe67]. Orgel described this scenario as follows: "The simplest theory suggests that the role of tRNA's was originally filled by a set of much shorter polynucleotides, perhaps the anticodon trinucleotides themselves. In this form, the theory postulates that trinucleotides have a selective affinity for the amino acid coded by their complementary trinucleotide. Of course, the selectivity must have been limited in the first place, but it is argued that it might have been sufficient to produce primitive activating enzymes in the presence of a suitable messenger RNA. Then the system could have perfected itself by the "bootstrap" principle, [...]. If this type of theory is correct the code is not arbitrary; if life were to start again, certain features of the code would be reproduced because the physical interactions on which it is based are unchanging" [Org68]. Exactly these kind of unchanging physical interactions are found in a number of recently published experiments ([CYK05, YCK05, YWK09] and references therein). Anticodons like GAA, GUA, GUG, and CCA are part of RNA molecules which bind respectively phenylalanine, tyrosine, histidine, and tryptophan. Again, phenylalanine and tyrosine form a group of amino acids coded by codons with U in the first position (contributing to a low MS_0 value), but in this scenario the formation of the group is due to a straightforward binding affinity of a GAA-containing RNA for phenylalanine, and another one of a GUA-containing RNA for tyrosine. Earlier experimental work pointed to a stereochemical relationship between the anticodons GCC, AGC and GAC and the simple amino acids glycine, alanine and valine respectively [Shi95]. The same author published models in which e.g. asparagine and lysine were shown binding their cognate anticodons [Shi82]. If the major determinant for the codon assignments in the standard genetic code is stereochemical affinity between triplets and amino acids as reported in these publications, this implies a low MS_0 value. Therefore, the Stereochemical Theory is not incompatible with a low MS_0 value.

3.3.5 A Four-Column Theory

The four scenarios discussed above share the characteristic that one factor (either "minimization", "history" or "stereochemistry") is the major determinant of the codon assignments in the standard genetic code. They share this characteristic

with the scenarios published by Wong [Won75] and by Ikehara [Ike02]. Other scenarios are present in which all three factors are major determinants [KFL99, vdG07]. As a last scenario, we discuss the four-column theory published by Higgs [Hig09]. Like the scenario proposed by Massey, the earliest genetic code according to the four-column theory is encoding a repertoire of four amino acids. Higgs is very detailed on the amino acids and the codon assignments in this earliest genetic code: the sixteen codons with U in the middle originally encoded valine, the sixteen middle-C codons alanine, the sixteen middle-A codons aspartate, and the sixteen middle-G codons glycine. Later amino acids were added to this code by a process of subdivision of these 16-codon blocks, in which a subset of the codons assigned to an early amino acid were reassigned to a later amino acid. In the four-column theory, codons with a certain middle position are reassigned to amino acids similar to the one originally assigned to codons with this middle position because this is the least disturbing to already existing protein sequences. The driving force for the reassignment is the "positive selection for the increased diversity and functionality of the proteins that can be made with a larger amino acid alphabet" [Hig09]. An intermediate code is presented, with Leu, Ile and Val coded by middle-U codons, Ser, Pro, Thr and Ala coded by middle-C codons, Asp and Glu coded by middle-A codons, and all middle-G codons coding Gly. At this stage, the total of protein-coding sequences starts to influence the further development of the code even more strongly (code-message coevolution, as in the series of papers by Sella and Ardell [AS01, SA02, AS02, SA06]) because, as a consequence of their function in proteins, glycine codons become rare codons. The consequence of this is that the constraint to reassign them is relaxed. The final result is that amino acids which are not similar to glycine, but which are associated with strong positive selection because they bring radical new functions for proteins (cysteine, tryptophan and arginine) are found coded by middle-G codons. Although Higgs emphasizes that the driving force during the process of expansion of the amino acid repertoire is not the minimization of translational error, the four-column theory is not as "neutral" as the "2-1-3" model, because the "minimal disruption to the proteins already encoded by the earlier code" by adding "...later amino acids into positions formerly occupied by amino acids with similar properties" is such an important component of the scenario.

Like the other discussed scenarios, the four-column theory is compatible with a low MS_0 value. All five discussed scenarios agree that error robustness due to codon assignments is present in the SGC. The scenarios differ in the way they propose the error robustness has been built.

3.3.6 Consequence of the error robustness

The consequence of the error robustness is an enormous potential to evolve. A variation in an RNA sequence can have different kinds of consequences in the

protein sequence. At the one end of the spectrum, the different codon does not lead to a different amino acid. Slightly more effect would be that a different codon would lead to a different amino acid, but this would be so similar to the original amino acid that no difference in protein structure is the consequence. Most important would be the effect that there is a difference in protein structure, but so small that natural selection can use it as a slight step along an evolutionary path. At the far end of the spectrum, finally, we find the lethal mutations. Because of this graded intensity of evolutionary effect, the nature of the relationship between RNA sequence and protein sequence (i.e. the SGC) gives biochemistry an enormous evolvability [Wag05, ZF06]. This not necessarily implies that the phenomenon itself is built by direct optimizing selection for the error minimizing aspects. Exactly the same argument holds for the aspects of stop codons allowing additional information to be encoded in protein-coding sequences as described by Itzkovitz and Alon [IA07].

Chapter 4

Unassigned codons in the genetic code

The content of this chapter is based on joint work with Wouter Hoff [vdGH11]

4.1 Potential lethality of unassigned codons

The origin of the genetic code can be envisioned as starting with a single primordial tRNA, which gave rise to the full complement of tRNAs by a complex series of gene duplication and diversification events. This view of tRNA genes as paralogues pervades thinking about the origin and evolution of the genetic code [Cri68, FU87, OJWM92]. While many aspects of tRNA evolution have been considered (cf. [DG06, RS08, FSK$^+$09, SBBG10, RSR11]), gene duplication and diversification are common themes during the evolutionary development of tRNA sets. Presumably, during this diversification process additional amino acids were incorporated one by one into the developing genetic code. This consideration leads to an important problem facing possible scenarios for the evolution of the code. In very early stages of the development of the SGC most codons were unassigned, leading to a situation in which many mutations in an early protein-encoding nucleic acid sequence would result in the introduction of an unassigned codon [SLB$^+$63, Son65, Cri68].

One can envision two general approaches to this problem of potentially lethal unassigned codons. The first option is that as soon as a small set of amino acids started to be transferred by tRNAs, rapid tRNA gene duplication and mutation of the anticodon resulted in a situation in which all codons were assigned to this initial set of amino acids. An important consequence of this scenario is that subsequent incorporation of novel amino acids into the expanding code requires reassignments of the meaning of codons. A second approach is that the code evolved more slowly, and that for extended periods of evolutionary time indeed many codons were not assigned [LJ88, Ike02, Fra11]. The introduction of novel amino acids could then proceed without codon reassignment. However, this scenario requires the non-lethality of nonsense mutations during the early evolution

of the code. Thus, formulating specific molecular scenarios for the evolution of the genetic code requires a choice: either numerous codon reassignment events or the prolonged existence of unassigned codons. Current thinking strongly favors the first option (e.g. [AVG07, Hig09, GdCLM10], see also comments by Higgs on page 15 of [DG08]).

Here we examine the strength of the evidence supporting this choice, and use biochemical knowledge regarding nonsense suppression in existing organisms [BG01, KF07] to support the viability of the second scenario. In addition, we use knowledge on tRNA wobble rules [Cri66, TY03, AVG07, GdCLM10, RH10] and the biochemistry of tRNA anticodon modifications [MNN$^+$88, MKS$^+$10, IKN$^+$10] to examine when tRNA anticodon modifications were introduced into the developing genetic code. These considerations lead to a novel scenario for the development of the SGC. All such scenarios are faced with the issue of the temporal order of and interplay between three key developments: (i) the assignment of unassigned codons, (ii) the incorporation of all 20 canonical amino acids into the code, and (iii) the introduction of tRNA anticodon modifications. We present an analysis of relevant available biochemical information that supports a model that contrasts with most published models with respect to the relative order of these three processes. This analysis supports the viability of scenarios involving the persistence of unassigned codons until all 20 amino acids were included in the code, and the incorporation of anticodon modifications at a relatively late stage in the evolution of the code.

4.2 Unassigned codons and suppression

The highly deleterious nature of nonsense codons was vividly described in an influential 1965 paper by Tracey Sonneborn:

"A nonsense mutation resulting in nontranslation of all codons distal to it would as a rule be enormously more detrimental (and therefore more rapidly eliminated) than a sensible (or mis-sensible) mutation which permits translation of the entire message. Hence, neutralizing the detriment of a nonsense mutation by a second mutation or a genic recombination is very much less likely. In short, such nonsense mutations would with high probability have no evolutionary future, and they would by virtue of their detriment be prime targets for elimination by natural selection. On the other hand, mis-sense mutations could sometimes have relatively little detrimental effect and therefore a relatively long persistence and correspondingly greater chance to enter into a lucky genic combination by further mutation or recombination."

This early view on the highly lethal nature of nonsense mutations and the relatively benign character of missense mutations has been solidly incorporated

into thinking about the evolution of the genetic code (e.g. [Cri68, AVG07]). As a result, the persistence of unassigned codons during most of the evolution of the SGC has not been considered as a viable possibility, while codon reassignments during this process are viewed as realistic and unproblematic. This view has been developed in detail in an important recent paper [Hig09]. While the deleterious effect of nonsense mutations stands unchallenged [Son65, Cri68, AVG07, Hig09], here we want to re-investigate its implications for early stages of the genetic code. Specifically, we will examine both the presumed level of lethality of nonsense mutations and the presumed likelihood of codon reassignments in the light of current knowledge of existing organisms. A significant body of data is available regarding the translational fate of mRNA molecules containing nonsense mutations [BG01, CKIV$^+$04, DB06, LMAK09]. This work has revealed that a significant level of translational readthrough across stop codons occurs. As a result, nonsense mutations even in essential genes often are non-lethal. Such nonsense suppression can involve mutations in tRNAs as in the amber, ochre, and opal suppressor tRNAs. However, natural nonsense suppression through the reading of stop codons by normal cellular tRNAs, which are called natural suppressors, has also been well documented [BG01]. In general, a view of translation has emerged in which the meaning of a codon is always a balance between the affinities of several different tRNAs for that codon, and the affinity of release factors for that codon [KF07, SSV12]. The current translational machinery in general exhibits a very low error rate. Thus, the amount of full-length protein that is produced in the presence of a stop codon in a coding sequence is significantly reduced, but in a number of cases (e.g. [LBZ$^+$07, MMZ10]) has been found to allow for viability of the organism. The degree to which the use of formally unassigned codons diminishes the translational efficiency of an organism will depend on its codon usage. In some organisms the usage of certain codons can be extremely low (see e.g. [UHLW04]), and inefficient translation of these codons will therefore only affect the synthesis of a small number of proteins. A central factor affecting codon usage is the abundance of the tRNA involved: tRNAs that are rare in the cellular tRNA pool tend to translate codons that are also rare, particularly in highly expressed proteins, presumably to optimize translational efficiency [Aka01]. If such rare codons were to become formally unassigned, this event would be expected to result in relatively mild detrimental effects. Indeed, formally unassigned codons are known in current organisms [OAMO91, KOAO93], providing a powerful argument against the supposed lethality of unassigned codons due to their introduction into the genome by mutations.

4.3 Suppression in primordial organisms

The experimental work on natural nonsense suppression discussed above has been obtained using contemporary organisms. What to expect in the case of primordial

organisms? The first critical consideration is that it appears likely that the fidelity of the early translational system was considerably lower. Thus, the "meaning" of a codon would be determined by its relative affinities for various tRNAs, and would thus be translated as a weighted mixture of various amino acids. Such "statistical proteins" were introduced by Woese [Woe65a], and have also been considered in later work [SA06, Hig09]. Reduced translational fidelity implies a level of readthrough (and therefore non-lethality) that is higher than that observed in current organisms. The presence of "inaccurate decoding" does not necessarily mean lethality: the acquisition of new evolutionary potentialities as a result of production of "statistical proteins" can even confer growth advantage. This has been experimentally demonstrated using mutants in which the editing function of isoleucinyl-tRNA synthetase was impaired, resulting in the low-level incorporation of non-canonical amino acids like norvaline into the proteome and an increased growth yield [PMH$^+$04].

The second critical consideration is that the modern system of release factors provides a rapid and high-fidelity system for recognizing stop codons. The introduction of a dedicated system for the recognition of stop codons during the evolution of the genetic code in general has not received much attention. The most primitive system for handling a stop codon would be that the ribosome stalls when it reaches an unassigned codon and eventually dissociates from the mRNA. In this view all unassigned codons would have stop codon activity. The actual translation of unassigned codons in such an early translational system would then be a balance between the rate of natural nonsense suppression and spontaneous ribosome dissociation.

Thus, we arrive at a situation in which early translational systems combine a relatively high translational error rate, resulting in the frequent translation of formally unassigned codons, with the absence of an efficient system dedicated to recognizing stop codons. This line of thought thus predicts that formally unassigned codons could be translated either as a stop codon (through spontaneous ribosome dissociation) or as a sense codon (through nonsense suppression). The relative frequency of these events would be open to optimization through molecular evolution of the components of the early translational system. The essence here is that "unassigned codons" in effect were to a significant extent not unassigned. The introduction of such codons would thus have likely been somewhat detrimental but not lethal.

Genome size is a third consideration with respect to the proposed process of rapid tRNA gene duplication and mutation to assign all codons to a small set of initial amino acids during an early stage of the evolution of the genetic code. The early genome replication machinery can reasonably be expected to have had limited fidelity. Thus, these early systems would be at considerable risk of facing an error catastrophe in which the chance of deleterious mutations per replication event would overwhelm the rate at which natural selection can purge deleterious mutations [ES77]. This effect would result in a strong selection for organisms with

very small genomes. Thus, it is not clear if systems in which the development of the genetic code has just started had sufficient genome replication fidelity to allow for a substantial number of different tRNAs. Based on these considerations, we conclude that it is reasonable to consider scenarios for the evolution of the genetic code in which many formally unassigned codons persisted throughout most of the evolutionary development of the code.

In summary, point mutations can introduce formally unassigned codons into the genome of early organisms. Because of the existence of natural nonsense suppression, such mutations will tend to reduce translational efficiency but will often not be lethal. This selective pressure against the use of such unassigned codons will cause these codons to remain rare in early organisms. Thus, the persistence of formally unassigned codons during the evolution of the genetic code is biochemically entirely plausible.

4.4 Codon reassignments are difficult

The SGC is nearly universal. Most code variants are known from mitochondria (see [SYH07] for an up-to-date treatment of mitochondrial codes and the mechanisms which lead to their emergence), which have an extremely small genome: less than a hundred protein-coding genes. Apart from mitochondria, code variants are extremely rare. Organisms as different as an elephant and an *Escherichia coli* bacterium have exactly the same 64 codon assignments, as stressed in early molecular biology. Apart from mitochondria only one sense reassignment is known: the 7 serine codons code of certain yeasts [SGS$^+$11]. A handfull of code variants with stop codon reassignments are known, among them the 4 glutamine codons code of certain ciliates [HKSB86], the 3 cysteine codons code of other ciliates [MSP$^+$91] and the 2 tryptophan codons code of *Mycoplasma* bacteria [YMK$^+$85]. Despite enormous genomics efforts during the last decade, no new non-mitochondrial codon reassignments have emerged[1].

Several code variants are known to have emerged multiple times, both in the group where they were discovered the first time (e.g. the 4 glutamine codons code in the ciliates, cf. [LKL01]) and in other groups (e.g. the 4 glutamine codons code in diplomonads [KD96], in certain green algae [SLY89] and oxymonads [KL03]). This shows that certain taxonomic groups (e.g. the ciliates) are prone to reach the rare situation in which codon reassignment can occur (see also [CGL$^+$10]). It also shows that certain particular codon reassignments (e.g. the reassignment of UAR from stop to Gln) are prone to happen in widely different taxa (but please note that all taxa with the UAR reassignment are eukaryotes; again, see [CGL$^+$10]). Taken together, this extensive body of work on codon reassignments in current organisms shows that reassignment events are very rare,

[1] The exceptions proving the rule are UAA coding for glutamate in certain ciliates [SSVMT03] and UGA coding for glycine in certain bacteria [COC$^+$13]

which implicates that codon reassignments are very difficult. This observation contrasts sharply with the ease with which codons are reassigned in origin of the SGC scenarios (e.g. [Cri68, Hig09]).

The functional impact of codon reassignments during the development of the genetic code can be expected to strongly depend on the degree of evolutionary optimization of the proteins in these early systems. On one end of the spectrum one can envision organisms using statistical proteins with a low level of structure-function optimization. In such a system the detrimental effects caused by introduction of a substantial number of mutations because of a codon reassignment may be limited. However, it is also possible that the genetic code evolved slowly, and that the functional properties of the proteins in early systems were already quite advanced, with highly optimized amino acid sequences. In that case most codon reassignments would be expected to have devastating effects on the proteome function. In recent work the fitness cost of codon reassignment events was modeled [Hig09]. This analysis focused on the presumably rare sites in proteins at which the reassignment will benefit the protein, while the likely damage to protein function caused by the reassignment was not considered. However, a body of recent work regarding the extrapolated amino acid composition of organisms predating the last universal common ancestor (LUCA) has provided support for the presence of a highly optimized proteome [BFS04, JKA$^+$05, FG10].

The analysis of trends in amino acid composition for sets of resurrected ancient proteins offers an interesting approach to explore the proteome of organisms predating the LUCA. A number of independent analyses following different bioinformatics strategies have revealed that amino acids that are often considered to have been added during a late stage of the evolution of the SGC (such as the aromatic amino acids and cysteine) were underrepresented in the LUCA [BFS04, JKA$^+$05, FG10]. This result implies that the functions of the proteins in these early systems were already sufficiently evolved to leave detectable traces in the proteins of current organisms. This conclusion suggests that the protein world was already fairly well developed before all 20 amino acids were incorporated. If this inference is correct, then codon reassignment during the evolution of the SGC would have been very difficult.

The above analysis indicates that in current scenarios of the evolution of the SGC the degree of lethality of nonsense mutations tends to be overestimated, while the difficulties associated with codon reassignments are generally underestimated. We therefore conclude that scenarios in which many unassigned codons persisted throughout most of the evolutionary development of the code should be considered. Such scenarios have the advantage that they do not require codon reassignments. In addition, they allow the developing code to function with a relatively small number of tRNAs, which is attractive in view of the error catastrophe threat in early systems with limited genome replication fidelity.

What properties would be expected for such small tRNA sets during the early stages of the development of the SGC? In general, nonsense suppression

relieves the need for the developing translational system to contain tRNAs for the formal assignment of all codons. A second important aspect of the SGC in current organisms is the widespread use of anticodon modifications to achieve the correct assignment of all codons. Did this highly sophisticated system of base modifications develop concomitant with the assignment of codons in the developing code? Or is it biochemically plausible that anticodon modifications were incorporated at a late stage, after the incorporation of all 20 amino acids into the code? Below we provide support for the latter possibility, leading to a view in which a small set of tRNAs with unmodified anticodons capable of nonsense suppression allowed the effective functioning of early systems encoding all 20 amino acids. In this scenario the lack of modifications in the tRNAs specifically regards the three nucleotides in the anticodon. It is entirely possible (however not necessarily probable) that other regions of these tRNAs did contain modified bases[2].

4.5 Role of anticodon modifications in the SGC

Many tRNA anticodon modifications have been identified. A in the first position of the anticodon is nearly always deaminated to inosine, as already discussed by [Cri66]. The effect of this is that the tRNA readily recognizes three codons instead of two (with the complicating factors that the exact effects are different in each codon box [JEH$^+$08] and may be taxonomically diverse). U in the first position of the anticodon is nearly always modified, which can occur in various ways. 2-Thiolation results in recognition of both purine-ending codons (e.g. [NIF$^+$06, PH10]). G in the first position of the anticodon can be modified in various different and complex ways, often resulting in increased specificity for the recognition of pyrimidine-ending codons. Modifications in the other positions of the anticodon also occur. A pseudouridine in the second position enlarges the capability of a $tRNA^{Tyr}$ to also recognize UAG, which is counteracted by a first position modification [GdCLM10]. Furthermore, modifications of other residues of the anticodon-loop, and in other parts of the tRNA molecule, can influence the readout properties of the tRNA (see e.g. [BG01, JEH$^+$08]). In summary, anticodon modifications in the tRNA molecules of contemporary organisms are widespread, and usually substantially alter the readout properties of the tRNA.

Since anticodon modifications alter the readout properties of tRNAs, the issue of when these tRNA anticodon modifications arose during the development

[2]The discovery that the tRNAs of C. Riesia pediculicola lack modification except for the stretch 34-40 (the anticodon and the 4 nucleotides following the anticodon) as reported by de Crécy − Lagard and co-workers [dCLMG12] suggests that these tRNAs did *not* contain modified bases; this also has implications for our judgement of the timing of introduction of modifications in the rRNAs; arguing for relatively late introduction is strengthened by the presence of different modification systems in Bacteria and Archaea (ribozymatic versus protein enzymatic systems).

of the SGC is important. Despite the large body of information on the effects of anticodon modifications on the translational properties of tRNA [TY03, AVG07, JEH+08, GdCLM10], this question has not received much attention in the literature regarding the evolution of the SGC. Below we explore the possibility that the machinery to perform anticodon modifications evolved after the 20 amino acids were already incorporated into the developing genetic code.

4.6 Unmodified anticodon wobble rules

When anticodon modifications are taken into account, the wobble rules are complex (see e.g. [AVG07]). But when anticodon modifications are ruled out, the wobble rules are simple: C recognizes one codon, G recognizes two codons, U recognizes four codons (but see subsection 4.6.1!), and A is not used[3]. The wobble behavior of tRNAs with anticodons starting with unmodified G or unmodified C was already described in 1966 [Cri66]. Regarding the wobble behavior of tRNAs with anticodons starting with unmodified U significant progress has recently been made, as summarized below. Based on this information we deduce the predicted properties of tRNA sets containing only unmodified anticodons. As discussed below, in this analysis we take the approach that the wobble rules operational during early stages of the evolution of the genetic code were the same as the wobble rules that apply to contemporary organisms.

4.6.1 Wobble rules and family boxes

The boxes of 4 codons in the genetic code table which differ only in the third position and which all encode the same amino acids (e.g. the GCN codons encoding alanine) are referred to as "family boxes" [LJ88]. Here we use the expression"codon box" as a more general concept for collections of 4 codons which only differ in the third position (e.g. the GAN codons are a codon box which is not a family box).

The factor causing the distribution of family boxes in the SGC is a long-standing question in the field [Lag78]. Recently a molecular mechanism was reported explaining this pattern based on hydrogen bonding interactions [LL08]. When the first two nucleotides of a codon form six hydrogen bonds with the anticodon, the codon box is a family box (codons CCN, CGN, GCN, and GGN). When the first two nucleotides of a codon make only four hydrogen bonds with the anticodon, the codon box is not a family box (codons UUN, UAN, AUN, and AAN). When the first two nucleotides of a codon are able to make five hydrogen bonds with the anticodon, the codon box is a family box only if the middle base of the codon is a pyrimidine (codons UCN, CUN, ACN, and GUN). This is caused

[3]Lehman an Jukes [LJ88] suggest an explanation why GNN anticodons developed in evolution preferentially to ANN anticodons, even when pairing with NNU codons.

by the stabilization of the position of the purine that forms the middle base of the anticodon by a long-range intramolecular hydrogen bond from U33 [LL08].

For the resulting eight family boxes the codon-anticodon complex is sufficiently strong to allow the recognition of all three non-cognate nucleotides in the third position of the codon by wobble. Recent experimental results have demonstrated the *in vivo* importance of this phenomenon in chloroplasts: their ribosomes allow "superwobbling", in which an anticodon with unmodified U in the first position can read all 4 codons in the glycine family box [RKB08]. A recent analysis of the tRNA sets present in bacterial genomes shows that in many bacteria "superwobbling" is widely used [RH10].

This information allows the conclusion that a set of 8 tRNAs with the anticodons UGA, UAG, UGG, UCG, UGU, UAC, UGC, and UCC, all starting with unmodified U, suffices to read the 32 codons of the family boxes (table 4.1).

<table>
<tr><td></td><td rowspan="2">UCN Ser</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td rowspan="2">CUN Leu</td><td rowspan="2">CCN Pro</td><td></td><td rowspan="2">CGN Arg</td></tr>
<tr><td></td></tr>
<tr><td rowspan="2"></td><td rowspan="2">ACN Thr</td><td></td><td></td></tr>
<tr><td></td><td></td></tr>
<tr><td rowspan="2">GUN Val</td><td rowspan="2">GCN Ala</td><td></td><td rowspan="2">GGN Gly</td></tr>
<tr><td></td></tr>
</table>

Table 4.1: Coding by tRNAs with anticodons starting with an unmodified U. The codons read by a set of 8 tRNAs with unmodified-U-starting anticodons as based on the wobble rules are indicated. The specific codon sets were selected to reflect the family boxes in the SGC.

4.6.2 Unmodified-G-starting anticodons

The first two codons in a codon box in the SGC always encode the same amino acid. The molecular basis for this pattern is that a single tRNA with an anticodon starting with unmodified G recognizes both Y-ending codons [Cri66]. The C-ending codon is the cognate codon, and the U-ending codon is recognized by

wobble.

This pattern implies that a set of 8 tRNAs with the anticodons GAA, GUA, GCA, GUG, GAU, GUU, GCU, and GUC, all starting with unmodified G, suffices to read the 16 Y-ending codons of the codon boxes which are not family boxes (table 4.2).

UUY Phe		UAY Tyr	UGY Cys
		CAY His	
AUY Ile		AAY Asn	AGY Ser
		GAY Asp	

Table 4.2: Coding by tRNAs with anticodons starting with an unmodified G. The codons read by a set of 8 tRNAs with unmodified-G-starting anticodons as based on the wobble rules are indicated. The specific codon sets were selected to reflect the Y-ending codons in the SGC that are not part of family boxes.

4.6.3 Unmodified-C-starting anticodons

Anticodons starting with unmodified C do not wobble [Cri66]. Thus, a set of 7 tRNAs with the anticodons CAA, CCA, CUG, CAU, CUU, CCU, and CUC, all starting with unmodified C, suffices to read the seven G-ending sense codons in the codon boxes which are not family boxes (table 4.3). UAG is a stop codon in the SGC. For the present purpose, we ignore the UAG codon and focus on the seven G-ending sense codons of the non-family boxes of the SGC.

4.6.4 Wobble rules in early evolution

The degree to which the wobble rules already operated during early stages of the evolution of the genetic code is difficult to ascertain definitively. A specific example is that, based on their work on tRNA sets, Tong and Wong have proposed that

UUG Leu			UGG Trp
		CAG Gln	
AUG Met		AAG Lys	AGG Arg
		GAG Glu	

Table 4.3: Coding by tRNAs with anticodons starting with an unmodified C. The codons read by a set of 7 tRNAs with unmodified-C-starting anticodons as based on the wobble rules are indicated. The specific codon sets were selected to reflect the G-ending sense codons of the codon boxes which are not family boxes in the SGC.

the superwobble was a relatively late development that took place in the bacterial domain [TW04]. This would not alter the main conclusions of our manuscript, because the 20 canonical amino acids can be coded, with the canonical assignments, by a small set of codons read by G-starting and C-starting anticodons only. However, the following two arguments provide support for the approach taken here, in which current wobble rules apply to the first stages of the evolution of the SGC. First, the wobble rules are a direct consequence of the physical chemistry of codon-anticodon hydrogen-bonding interactions, and thus would be expected to apply as soon as the first codons and anticodons started to interact. Secondly, two classic regularities in the genetic code are readily interpreted as being direct results of the operation of the wobble rules.

First, the fact that, without exception, both Y-ending codons in a codon box encode the same amino acid is most easily explained as a result of the wobble behavior of unmodified G in the first position of the anticodon. Second the fact that, also without exception, the 32 codons which form the most stable codon-anticodon pairs are organized as family boxes is most easily explained as a result of the superwobble. These regularities are consistent with the view that whenever a single tRNA could read several codons with a reasonable level of efficiency, diversification of the meaning of these codons was blocked. Natural selection favored the appearance of anticodons starting with unmodified U for reading the codons of the 8 family boxes because a minimal number of tRNAs in this way

could read a maximal number of codons. The basic structure of the SGC (8 quartets and 8 pairs) can therefore be seen as a reflection of the wobble rules for anticodons starting with unmodified U for the family boxes, and anticodons with unmodified G for the other codon boxes.

These considerations leave ample room for the further development of various aspects of the genetic code, such as those considered by Tong and Wong [TW04], since the first organism in which all 20 amino acids were encoded likely was an earlier and more primitive organism than the Last Universal Common Ancestor (LUCA). However, such developments do not affect the main conclusions reached here.

4.7 Small sets without anticodon modifications

<table>
<tr><td>UUY Phe</td><td rowspan="3"></td><td>UAY Tyr</td><td>UGY Cys</td></tr>
<tr><td></td><td rowspan="2"></td><td></td></tr>
<tr><td></td><td>UGG Trp</td></tr>
<tr><td rowspan="3">CUN Leu</td><td rowspan="3">CCN Pro</td><td>CAY His</td><td rowspan="3"></td></tr>
<tr><td></td></tr>
<tr><td>CAG Gln</td></tr>
<tr><td>AUY Ile</td><td rowspan="3">ACN Thr</td><td>AAY Asn</td><td>AGY Ser</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>AUG Met</td><td>AAG Lys</td><td>AGG Arg</td></tr>
<tr><td rowspan="3">GUN Val</td><td rowspan="3">GCN Ala</td><td>GAY Asp</td><td rowspan="3">GGN Gly</td></tr>
<tr><td></td></tr>
<tr><td>GUG Glu</td></tr>
</table>

Table 4.4: The coding behavior of a set of 20 tRNAs with unmodified anticodons. The set of 20 tRNAs is selected from the first three tables of this chapter. This set is an example in which the UCN (Ser in the SGC), CGN (Arg in the SGC), and UUG (Leu in the SGC) codons were unassigned. Together, these 20 tRNAs without anticodon modifications can translate all 20 canonical amino acids.

From the 23 tRNAs listed above (8 U-starting anticodon tRNAs, 8 G-starting anticodon tRNAs, and 7 C-starting anticodon tRNAs, all with unmodified anticodons), various sets of twenty can be picked such that all twenty canonical amino acids are encoded. This observation leads to the conclusion that no anticodon modification is needed to specifically encode all 20 amino acids. This conclusion

is a direct consequence of the wobble rules that is relevant for possible scenarios for the development of the SGC.

The coding capabilities of one possible set of 20 tRNAs derived above (table 4.4) are compared with those of the SGC (table 4.5). In the above set of 23 tRNAs, the amino acids Ser, Arg, and Leu are translated by two distinct tRNAs. Here we describe one specific example of a set of 20 tRNAs encoding all 20 amino acids. A very similar description applies to other 20 tRNA set variants. The main feature of the depicted 20-tRNA code is a striking similarity to the SGC. A few small but systematic deviations are present. First, the three stop codons in the SGC are not assigned in the 20-tRNA code. Secondly, in the SGC Ile is encoded by three codons, while in the 20-tRNA code this is reduced to two codons. For Lys, Arg, Gln, and Glu the SGC contains two adjacent codons; in the 20-tRNA code these are each reduced to a single (G-ending) codon.

<table>
<tr><td>UUY Phe</td><td rowspan="2">UCN Ser</td><td>UAY Tyr</td><td>UGY Cys</td></tr>
<tr><td rowspan="2">UUR Leu</td><td rowspan="2">UAR STOP</td><td>UGA STOP</td></tr>
<tr><td>UGG Trp</td></tr>
<tr><td rowspan="2">CUN Leu</td><td rowspan="2">CCN Pro</td><td>CAY His</td><td rowspan="2">CGN Arg</td></tr>
<tr><td>CAR Gln</td></tr>
<tr><td rowspan="2">AUH Ile</td><td rowspan="2">ACN Thr</td><td>AAY Asn</td><td>AGY Ser</td></tr>
<tr><td>AUG Met</td><td>AAR Lys</td><td>AGR Arg</td></tr>
<tr><td rowspan="2">GUN Val</td><td rowspan="2">GCN Ala</td><td>GAY Asp</td><td rowspan="2">GGN Gly</td></tr>
<tr><td>GAR Glu</td></tr>
</table>

Table 4.5: The SGC. "H" is IUPAC nucleotide code for "U or C or A".

The key conclusion is that sets of 20 tRNAs that do not contain anticodon modifications can encode all 20 canonical amino acids in a pattern that is highly similar to that of the SGC. This analysis shows the biochemical feasibility of scenarios for the development of the SGC in which anticodon modifications were introduced only after all 20 canonical amino acids were already incorporated into the developing code. We would like to stress that this finding does not constitute proof for such a relative late development of tRNA anticodon modifications. In addition, it also does not necessarily imply that such a set of 20 tRNAs existed at a

specific stage of the evolution of the SGC. For example, for the tRNAs transferring Arg it is entirely possible that two isoacceptors already existed (one reading the codons of the CGN family box, the other reading the AGG codon) before Cys and the aromatic amino acids were added to the amino acid repertoire, and similar considerations apply to Leu and Ser. However, it does demonstrate that this option is biochemically feasible and thus should be considered, since current knowledge does not allow a firm identification of the stage of the development of the SGC at which anticodon modifications were introduced. Similarly, with the above series of tRNAs we do not wish to imply that this sequence of events occurred during the evolution of the SGC. Our conclusion is that small sets of 20-23 tRNAs with unmodified anticodons are capable of encoding all twenty canonical amino acids. In view of the relative simplicity of these tRNA sets and their biochemical plausibility, we propose that scenarios for the evolution of the SGC incorporating such a tRNA set should be considered as a viable possibility.

This view of the evolution of the SGC presents two novel possibilities: (i) that nonsense suppression is an important feature of the developing code, and (ii) that tRNA anticodon modifications were not introduced until after all 20 amino acids were encoded. In this scenario, eight A-ending codons remained unassigned far longer than generally assumed. It should be noted that this does not mean that these codons were not used at all in early protein-coding genes, as they could be read by sense suppression. The key attraction of such scenarios for the evolution of the SGC is the relative simplicity of the tRNA set that would allow for the translation of all 20 canonical amino acids. The minimal number of 20-23 tRNAs would be able to perform this translational task in the absence of any machinery for introducing tRNA anticodon modifications. The set of 23 can be reached by a relatively straightforward series of steps involving tRNA gene duplication / anticodon mutation / mutation in tRNA amino acid charging specificity, and (as discussed further below) can be refined by the subsequent incorporation of tRNA anticodon modifications.

It has been argued that the tRNA set of the archaeon *Methanopyrus kandleri* reflects a relatively early stage of development and resembles that in the LUCA ([TW04, WCM+07], but see [BFG04, GBA06]). In accord with the scenario developed here, the tRNA set of *M. kandleri* resembles the 20 tRNA set depicted in table 4.4. The tRNA set of *M. kandleri* shows a certain "simplicity" [TW04]. In all 8 family boxes of the tRNA set of *M. kandleri* two isoacceptors exist, one in which the anticodon starts with G and another in which the anticodon starts with U. The resemblance with the 20 tRNA set depicted in table 4.4 resides in the fact that these 16 tRNAs could have developed from a primordial set of 8 tRNAs with anticodons starting with unmodified U. In the 5 codon boxes which are not family boxes and which are considered "standard boxes" by Tong and Wong (i.e. the UUN, CAN, AAN, GAN, and AGN codon

boxes), the Y-ending codons are read by a tRNA with a G-starting anticodon, and the R-ending codons are read by a tRNA with an U-starting anticodon. This resembles the 20 tRNA set depicted in table 4.4 in the sense that these 10 tRNAs could have developed from a primordial set of 10 tRNAs in which the G-ending codons were read by a tRNA with an anticodon starting with unmodified C, and the A-ending codons were unassigned. The "uniform GU coding" concept of Tong and Wong could in this way be a next step from a more primordial situation in which a more restricted set of codons was read by a set of tRNAs like the one depicted in table 4.4. To make this step, anticodon modification would need to be introduced. An alternative way to look to the tRNA set of *M. kandleri* is to consider the organism as having returned (cf. [BFG04]) to a simpler set of tRNAs, coming from the more elaborate "uniform GUC coding" [TW04] predominant in archaea. In that case, *M. kandleri*, like vertebrate mitochondria in a different aspect (superwobbling), used the potential for "simplicity", a potential which was present in the system as a trace of the past. Seen in this light, these simplicities are not entirely new "discoveries", but potentials lurking in the system, because the system had evolved from these simplicities. The resemblance of the tRNA sets of archaea in combination with the proposed resemblance to the LUCA is in excellent agreement with the scenario described here, both when *M. kandleri* is considered as a living fossil, and when *M. kandleri* is seen as a case of return to simpler stage.

4.8 No codon reassignments required

Earlier, we argued that codon reassignments are very rare. Since tRNA modifications alter anticodon readout properties, the introduction of these modifications at a late stage in the development of the genetic code faces the possible problem of highly deleterious changes in the meaning of codons that are used to encode proteins. Below we provide a scenario in which the late introduction of tRNA modifications can proceed without perturbing protein-coding gene sequences.

The introduction of the enzyme that adds a sulfur atom to U-starting anticodons [NIF$^+$06] also containing U on the second position allows for the appearance of duplicates of the tRNAs with C-starting anticodons for Gln, Lys and Glu, followed by C-to-U mutations at the first anticodon positions. In this way, the collection of codons specifying e.g. Lys increases from one (AAG) to two (AAA and AAG). Since in the scenario proposed here these A-ending codons had thus far remained unassigned, no codon reassignments are involved, and no deleterious changes in the existing proteome result from the introduction of the anticodon modification systems. This process is part of a proposed final stage of the process of tRNA repertoire expansion which leads to a situation in which all codons are efficiently and unambiguously encoded. With a pattern of codon assignments as presented in table 4.4 as starting point, anticodon modifications

can be introduced without the concomitant introduction of assignment changes of codons used in the protein-coding part of the genome. Similar scenarios can result in the incorporation of the remaining codons in the SGC.

The observation that tRNA anticodon modifications as observed in the SGC can be introduced into the early 20-tRNA set proposed here without deleterious codon reassignments adds to the plausibility of this scenario.

4.9 Agmatidine and Lysidine

Tong and Wong [TW04] used the analysis of tRNA sets to deduce that the introduction of the inosine modification of A in the first position of the anticodon was a relatively late evolutionary development. In general, such a relatively late introduction of tRNA anticodon modifications lends support to the scenario presented here. Very recently, detailed biochemical data have become available which imply that modification of the anticodon responsible for decoding the AUA codon occurred after the LUCA.

If the incorporation of tRNA anticodon modifications indeed occurred after all 20 amino acids were incorporated into the developing code, it is possible that this modification system was not yet fully developed in the LUCA. In that case, one would expect differences in the tRNA anticodon modification machinery in the three domains of life. Recent reports on tRNA anticodon modifications in bacteria and archaea indeed provide support for the view that at least some tRNA anticodon modifications were not yet present in the LUCA. Bacteria use the modified nucleoside lysidine to translate AUA as Ile without concomitantly translating AUG as Ile [MNN$^+$88]. Archaea use another modification, agmatidine [MKS$^+$10], and another type of modification enzyme [IKN$^+$10]. This implies that Bacteria and Archaea independently evolved both the modified anticodon nucleoside and the modification enzyme. They presumably had a common ancestor in which this anticodon modification was not yet present.

These results indicate that the tRNA anticodon modification machinery is a valuable source of information on the development of the genetic code [GdCLM10]. The analysis reported here provides a natural framework for understanding this emerging taxonomic diversity in tRNA anticodon modifications: divergent evolution from an earlier translation system lacking these tRNA anticodon modifications. Future studies along these lines should take into account the complicating possibility of inter-domain lateral gene transfer of tRNA anticodon modification enzymes. Inter-domain lateral gene transfer has been documented for the aminoacyl-tRNA synthetases [WOIS00]. This line of research promises to reveal at which stage of the evolution of the SGC the various tRNA anticodon modifications were introduced.

4.10 A novel regularity in the genetic code

In summary, nonsense suppression can permit the persistence of unassigned codons throughout the evolution of the genetic code, resulting in a small but functional tRNA set, and small sets of tRNAs with unmodified anticodons can efficiently encode all 20 amino acids. These findings allow for a relatively simple early genetic code, specifying all twenty canonical amino acids, in the absence of tRNA anticodon modifications. This proposal appears to be compatible with the main features of influential ideas on the evolution of the SGC [Cri68, Won05, YCK05, DG08, Hig09]. Future studies on the taxonomic distribution of tRNA anticodon modifications offer a viable avenue to further explore the properties of the genetic code in organisms predating the last universal common ancestor.

The analysis described here reveals a novel regularity in the genetic code, expanding upon known regularities. In the nineteen sixties it was realized that, without exception, all pairs of Y-ending codons sharing a codon box encode the same amino acid [Cri66], and that the middle-U codons are all encoding hydrophobic amino acids while the middle-C codons are all encoding amino acids of comparable value of polar requirement [WDSD66]. Subsequently, it was pointed out that amino acids encoded by A-starting codons tend to have aspartate as a biosynthetic precursor while amino acids encoded by C-starting codons tend to have glutamate as a biosynthetic precursor [Won75] and that, without exception, all 32 codons which form the most stable codon-anticodon pairs are organized as family boxes [Lag78]. Here we report that no canonical amino acid is encoded by one single A-ending codon only, and that this regularity, in combination with the known wobble behavior of tRNAs with G-starting and C-starting anticodons, has implications for the likely primordial tRNA sets which existed *before* the LUCA.

Chapter 5

Aptamers and the genetic code

The content of this chapter is based on joint work with Harry Buhrman, Gunnar Klau, Christian Schaffner, Dave Speijer, and Leen Stougie [BvdGK+13]

5.1 The three "faces" of the genetic code

The genetic code probably evolved by a process of gradual evolution from a proto-biological stage, via many intermediary stages, to its present form (see e.g. [Cri68, LJ88, VWG06]). During this process, error robustness was built into the code (see e.g. [Ard98, VWG06, Hig09, Cri68, IOAH02, FWK03, CYK05, Won05, WK07, Mas08, DG08]). As already mentioned earlier in this thesis, two different kinds of error robustness can be observed [VWG06] by even the most superficial inspection of the Standard Genetic Code (SGC). On one hand, codons assigned to the same amino acid are almost always similar, see Table 5.1. As an example, all codons ending with a pyrimidine (U or C) in a codon box (the four codons sharing first and second nucleotides) are without exception assigned to the same amino acid (e.g. UAU and UAC both code for Tyr). On the other hand, similar codons are mostly assigned to similar amino acids, e.g. codons with U in the second position are all assigned to hydrophobic amino acids [Woe65b, WDD+66, WDSD66]. This is illustrated in Table 5.1, when looking at the values of polar requirement: overall, low values of polar requirement correspond to hydrophobic amino acids.

Three main approaches exist to explain the emergence of this robustness of the code: specific selection for robustness (see e.g. [HH91, FH98a, VWG06]), amino acid-RNA interactions leading to assignments (see e.g. [Woe65b, YWK09]), or a slow growth process of assignment patterns reflecting the history of amino acid repertoire growth (see e.g. [Cri68, Won75, Mas06, DG08]). The concept that all three competing hypotheses are important has also been brought forward [KFL99]; the three different facets of code evolution have been called "the three faces of the genetic code" in this connection. In this chapter, adjustments

UUU Phe (4.5)	UCU Ser (7.5)	UAU Tyr (7.7)	UGU Cys (4.3)
UUC Phe (4.5)	UCC Ser (7.5)	UAC Tyr (7.7)	UGC Cys (4.3)
UUA Leu (4.4)	UCA Ser (7.5)	UAA STOP	UGA STOP
UUG Leu (4.4)	UCG Ser (7.5)	UAG STOP	UGG Trp (4.9)
CUU Leu (4.4)	CCU Pro (6.1)	CAU His (7.9)	CGU Arg (8.6)
CUC Leu (4.4)	CCC Pro (6.1)	CAC His (7.9)	CGC Arg (8.6)
CUA Leu (4.4)	CCA Pro (6.1)	CAA Gln (8.9)	CGA Arg (8.6)
CUG Leu (4.4)	CCG Pro (6.1)	CAG Gln (8.9)	CGG Arg (8.6)
AUU Ile (5.0)	ACU Thr (6.2)	AAU Asn (9.6)	AGU Ser (7.5)
AUC Ile (5.0)	ACC Thr (6.2)	AAC Asn (9.6)	AGC Ser (7.5)
AUA Ile (5.0)	ACA Thr (6.2)	AAA Lys (10.2)	AGA Arg (8.6)
AUG Met (5.0)	ACG Thr (6.2)	AAG Lys (10.2)	AGG Arg (8.6)
GUU Val (6.2)	GCU Ala (6.5)	GAU Asp (12.2)	GGU Gly (9.0)
GUC Val (6.2)	GCC Ala (6.5)	GAC Asp (12.2)	GGC Gly (9.0)
GUA Val (6.2)	GCA Ala (6.5)	GAA Glu (13.6)	GGA Gly (9.0)
GUG Val (6.2)	GCG Ala (6.5)	GAG Glu (13.6)	GGG Gly (9.0)

Table 5.1: The standard genetic code. Assignment of the 64 possible codons to amino acids or stop signals, with updated polar requirement [MLS08] values indicated in brackets.

are made to earlier mathematical work in this field (see e.g. [HH91, FH98a, BvdGK$^+$11]) which integrate the three concepts into a single mathematical procedure. We will now (in the three following subsections), one by one, introduce these three adjustments. In section 5.2 we show that with the three adjustments combined, the SGC is optimal in the resulting space. The biological implications of this mathematical fact are discussed in section 5.3. Next, we show in section 5.4 that with a measure for the molecular structure as input, again a remarkable error robustness is found, but in a different way than with polar requirement as input. Finally, some last considerations about the SGC, in particular concerning the question why *these* twenty amino acids are found in the SGC, are brought forward in section 5.5.

5.1.1 Polar Requirement

The polar requirement [WDD$^+$66] is not just a measure related to hydrophobicity. Several different measures of hydrophobicity exist, each focusing on different aspects of it. Polar requirement specifically focuses on the nature of the interaction between amino acids and nucleic acids. Stacking interactions between e.g. the planar guanidinium group of arginine and the planar purine ring systems and pyrimidine ring systems of RNA is an example of that. Woese choose to chemically model the nucleotide rings by using pyridine as the solvent system in the measurements leading to the polar requirement scale [Woe65b, WDSD66, WDD$^+$66,

Woe67, Woe73]. This interaction between amino acids and nucleic acids has been stressed as an especially important aspect of early protein chemistry because one possibility for the very first function of coded peptides was suggested [Nol04] to be the enlargement of the number of conformations accessible for RNA (realized by the binding of small, oligopeptide cofactors). Thus polar requirement could have been among the most important aspects of an amino acid during early stages of genetic code evolution.

The remarkable character of polar requirement as a measure of amino acids in connection to the genetic code was found again and again throughout the years. Firstly, Woese found that distinct amino acids coded by codons differing only in the third position are very close in polar requirement, despite differences in general character [WDSD66]. The pair cysteine and tryptophan nicely exemplifies this. Secondly, Haig and Hurst [HH91] discovered that polar requirement showed the SGC to be special to a much larger degree than another scale of hydrophobicity (the hydropathy scale of Kyte and Doolittle [KD82]). Thirdly, when Mathew and Luthey-Schulten updated the values of polar requirement [MLS08] by *in silico* methods (the most important change was believed to be due to a cellulose-tyrosine interaction artefact in the original experiments), the SGC showed a further factor 10 increase [BGMLS09] in error robustness calculations. In all these developments the expectation that polar requirement would behave in a special way, as interaction between nucleotides and amino acids is biochemically important, was more than borne out by the results. One of the adjustments we introduce in our work compared to our earlier calculations [BvdGK$^+$11] is that in the present work we use the new, updated values of polar requirement (see Table 5.1).

5.1.2 Aptamers

Oligonucleic-acid molecules that bind to a specific target molecule (e.g. a specific amino acid) are called aptamers [ES90]. Over the last two decades, many results have been obtained regarding specific binding of amino acids by RNA aptamers, mainly by Yarus and co-workers [MY94, IY02, YWK09]. For several amino acids, codons and anticodons were found in binding sites, in quantities higher than would be expected to occur by chance [YWK09]. In Table 5.2, a list of occurrences of anticodons in binding sites of RNA sequences is given, together with the articles in which these sequences were reported. Please note that the definition of anticodons used in these articles is: triplets complementary to codons. These anticodons are therefore not necessarily identical to the triplets found in tRNA molecules which are normally meant with the word "anticodon". As an example: the triplet AUG is considered as an His anticodon because it is complementary to the His codon CAU. In tRNAs, however, the anticodon recognizing CAU is GUG (see [JEH$^+$08, GdCLM10] for reviews on codon-anticodon interaction). We summarize published details on the aptamers for seven amino acids, and subsequently formulate a conclusion regarding the implications of the existence of

these molecules for genetic-code error-robustness calculations. This conclusion is based on reasoning presented by the Yarus group concerning the existence of specific relationships between certain triplets and certain amino acids. These relationships could have led to evolutionary conserved assignments of these amino acids to these triplets, e.g. by a mechanism as presented in [YWK09]. Another mechanism has been suggested in [vdG07] (page 3).

For Ile, Trp, and His, three binding motifs were described, respectively named the "UAUU-motif" [LCMY03], the "CYA-motif" [MY05, MCM$^+$10], and the "histidine-motif" [MPY05]. As can be seen from the names, the anticodons UAU for Ile, and CCA for Trp, are characteristic for the motifs ("CYA" stands for "CUA or CCA"). In the case of His, both GUG and AUG (the anticodons for the two His codons CAC and CAU) are found in quantities higher than would be expected by chance [MPY05].

Amino Acid	Anticodon	References
Ile	UAU	[YWK09, pages 415-419]
Trp	CCA	[MCM$^+$10, page 1918]
His	GUG, AUG	[YWK09, pages 413-414]
Phe	GAA, AAA	[YWK09, page 420]
Tyr	GUA, AUA	[YWK09, page 423]
Arg	CCU, UCU, ACG, GCG, UCG, CCG	[JWKY10, page 2]
Leu	CAA, GAG, UAG	[YWK09, page 420]

Table 5.2: The occurence of anticodons in binding sites of the RNA sequences of amino-acid binding aptamers, and the references in which the actual RNA sequences can be found.

Although binding sites for Phe and Tyr have so far not been studied as extensively as those for Ile, Trp, and His, the analysis of Yarus and co-workers [YWK09] shows that the anticodons (GAA and AAA for Phe, and GUA and AUA for Tyr) are present in the binding sites more often than would be expected on a random basis.

Both the CCU anticodon [JWKY10] and the UCG anticodon [YWK09] are present in Arg binding sites more often than would be expected on a random basis. A physico-chemical background compatible with: (1) Arg having more than 4 codons, and (2) all 6 Arg codons sharing the same middle nucleotide, was thus observed.

A similar observation can be made for the other amino acid which is encoded by six codons all having the same middle nucleotide: Leu. For this amino acid, however, only a single RNA sequence was found binding the amino acid with specificity [YWK09]. Inspection of this sequence shows anticodons UAG, GAG, and CAA to be present in its binding parts.

Taking the combined results of Yarus and co-workers into consideration, we

propose to fix assignments of Ile, Trp, His, Phe, Tyr, Arg, and Leu for calculations using random variants of the SGC.

5.1.3 Gradual Growth

In section 5.2 we present our approach in detail. We use Haig and Hurst's "mean square" measure, (as first proposed in [HH91], but please see also the last paragraph of section 1.2) to quantify the error robustness of a given code. With this measure, a relatively error-robust code gets a low value when compared to the average value of a large set of codes produced by random allocation of amino-acid assignments (see [BvdGK$^+$11] for a more in-depth treatment of the approach). The space of codes allowed to exist by the allocation procedure can be large (in the original work of Haig and Hurst [HH91] the space has a size of exactly 20! codes, which is $\approx 2.433 \cdot 10^{18}$ codes). We call a code optimal if it reaches the minimum in error robustness calculations among all possible codes in a particular setting.

In 1975, Wong proposed the coevolution theory of the genetic code [Won75]. According to this proposal, SGC codons assigned to an amino acid biosynthetically derived from another amino acid, were originally assigned to that "precursor" amino acid. As an example: Pro is biosynthetically derived from Glu. According to coevolution theory, the four Pro codons (CCN) would have originally encoded Glu. Without embracing all details of the original coevolution theory, or modern refinements of the theory [Won07, DG08], something remarkable can be noted as a result of this way of looking at the SGC. Shikimate-derived amino-acids (Phe, Tyr, and Trp) all have U in the first position of the codon (Phe: UUY; Tyr: UAY; and Trp: UGG). Glu-derived amino-acids (Pro, Gln, and Arg) almost always have C in the first position of the codon (Pro: CCN; Gln: CAR, which stands for "CAA or CAG"; and Arg: AGR and CGN, where N stands for all 4 nucleotides). Asp-derived amino-acids (Ile, Met, Thr, Asn, and Lys) all have A in the first position of the codon (Ile: AUY and AUA; Met: AUG; Thr: ACN; Asn: AAY; and Lys: AAR). Codons with G in the first position all code for amino acids produced in Urey-Miller experiments[1] (Val: GUN; Ala: GCN; Asp: GAY; Glu: GAR; and Gly: GGN). This "layered structure" of the SGC was first pointed out explicitly by Taylor and Coates [TC89]. It may indeed suggest a sequential development of the repertoire of amino acids specified in the developing code, and a possibly sequential introduction of use of G, A, C, and U as first nucleotide in codons. The "layered structure" of the SGC is a different regularity from the well-known error-robust distribution of polar requirement [HH91], which is pronounced in the first and the third, but not the second position of the codon (please note: having, as a group, all the same nucleotide in the *first* position,

[1] For a recent update on prebiotic synthesis see [PCD$^+$11] and references therein.

gives error robustness for the group character to changes in the *second* and *third* position). As is shown in section 5.4, it is possible to prove the presence of the "layered structure" quantitatively, when the appropiate set of values is developed and used as input.

Freeland and Hurst [FH98b] followed the concept of Taylor and Coates, and formally divided the 20 amino acids in four groups of five amino acids each: Gly, Ala, Asp, Glu, and Val in a first group which could be called "the prebiotic group"; a second group of amino acids with codons starting with A (Ile, Met, Thr, Asn, and Lys); a third group with codons mainly starting with C (Leu, Pro, His, Gln, and Arg); and, finally, a group with codons mainly starting with U (Phe, Ser, Tyr, Cys, and Trp). Division of the set of twenty in these four subsets was subsequently incorporated in the calculations on code error robustness [FH98b]. This approach reduced the size of the space from which codes could be sampled randomly in a drastic way: from a size of about $2 \cdot 10^{18}$ codes (see above) to a size of $(5!)^4$ codes (which is exactly $2.0736 \cdot 10^8$ codes). This space was called the "historically reasonable" set of possible codes [FH98b]. By sampling from the historically reasonable set of possible codes, we incorporate in the current study the notion of a chronologically-determined, layered structure of the SGC.

5.1.4 Integration of assumptions

We have found that if: (1) the updated values for polar requirement are used as amino-acid attributes; (2) the assignments of seven amino-acids to codons are fixed following the rationale given above; and (3) the subdivision leading to the historically reasonable set of possible codes is used to define the space of code variations (which is also reduced in size by (2)), then the SGC is optimal. It is important to note that the constraints applied drastically reduce the size of the space: with applying both (2) and (3), the "realistic space" has a size of 11520 codes.

5.2 Optimality of the genetic code

We use the method developed by Allf-Steinberger [AS69], Wong [Won80], and Di Giulio [DG89a], and used by Haig and Hurst [HH91] (please see also the last paragraph of 1.2). For the mathematical formulation, we follow the approach of [BvdGK$^+$11] and consider the undirected graph $G = (V, E)$ that has the 61 codons[2] as its vertices and an edge between any two codons if they differ in only one position, yielding 263 edges. A code F maps each codon c to exactly one amino acid $F(c)$. We denote by $r_{F(c)}$ the polar requirement of the amino acid

[2]In the original calculation, Haig and Hurst ignored the three "stop codons" encoding chain termination.

that codon c encodes in the code F and by $\mathbf{r}$ the full vector of 20 values. The mean-square error function of code F is then given by

$$MS_0^{\alpha,\mathbf{r}}(F) = \frac{1}{N} \sum_{\{c,c'\}\in E} \alpha_{c,c'} \left(r_{F(c)} - r_{F(c')}\right)^2$$

where the $\alpha_{c,c'}$ are the weights of the different mutations that can occur (corresponding to edges of the graph) and $N = \sum_{\{c,c'\}\in E} \alpha_{c,c'}$ is the total weight. Following Haig and Hurst [HH91], we use a subscript 0 to indicate the overall measure. If we set all 263 weights $\alpha_{c,c'}$ to 1, we get the original function described by [HH91] which we simply denote by $MS_0(F)$. We also consider the following set of weights introduced by Freeland and Hurst [FH98a] which differentiates between transition errors (i.e. U to C, C to U, A to G, G to A) and transversion errors and the position where they occur in the codon:

- $\alpha_{c,c'} = 0.5$ if (c, c') is a transversion in the first position or a transition in the second position,
- $\alpha_{c,c'} = 0.1$ if (c, c') is a transversion in the second position,
- $\alpha_{c,c'} = 1$ otherwise.

Using weights for different codon positions implies the existence of a tRNA with a triplet anticodon during the process of code evolution. As we consider a process of gradual expansion of the repertoire of amino acids during the evolution of the SGC (see e.g. [Cri68, LJ88, Ard98]) as the most likely mechanism - with duplication of tRNA genes, and subsequent divergence (cf. [Ohn70]) of their sequences and functions - we think this assumption is acceptable. This assumption does not necessarily imply the existence of protein aminoacyl-tRNA synthetases during all or part of the process of code evolution, as there could originally have been ribozymes which fulfilled their function. The value of error-robustness of a code F using the set of weights introduced above will be denoted by $MS_0^{FH}(F)$.

In principle, there are at least three ways in which one can improve the model of [HH91] to reflect biological reality more accurately. The first possibility is to change how the level of error robustness is measured, e.g. by introducing weighting factors as described above. Variations of the weighting factors used in the calculation show an even higher error robustness of the SGC, as noticed by e.g. [FH98a, GMCR01, BGMLS09]. The rationale behind changing weighting factors is improved reflection of natural selection pressures. It is, however, difficult to decide which weighting factors adequately reflect the natural selection pressures operating during the early evolution of the genetic code (see comment 4 of Ardell in [NWK07] and the exchange of thoughts with respect to "column 4" in [Hig09]).

The second way to improve the model is to change the set of values representing amino-acid properties used as input in the error-robustness calculation.

For instance, one can use the values of hydropathy from [KD82], or the matrix of [GMCR01] instead of the polar requirement scale. In this chapter, the values of the 2008 update of polar requirement by *in silico* methods [MLS08] given in Table 5.1 are used.

The third way to improve the model is to change the size of the space from which random codes are sampled [BvdGK$^+$11]. The incentive to enlarge that space (as was done in [BvdGK$^+$11]) is the wish to work from a space that encompasses all possible codes, or at least, all known codes. As indicated in [BvdGK$^+$11], larger spaces are increasingly difficult to work with. The frequency distributions obtained by sampling from the larger spaces in [BvdGK$^+$11] highly coincide with the frequency distribution obtained from the original space (as presented in [HH91]). From this viewpoint, working in the original space is acceptable as a simplification. In the current study, we *shrink* the size of the space, based on considerations of fixed assignments of certain codons, and combining this with the constraint of the historically reasonable set of possible codes of [FH98b], as outlined in section 5.1.

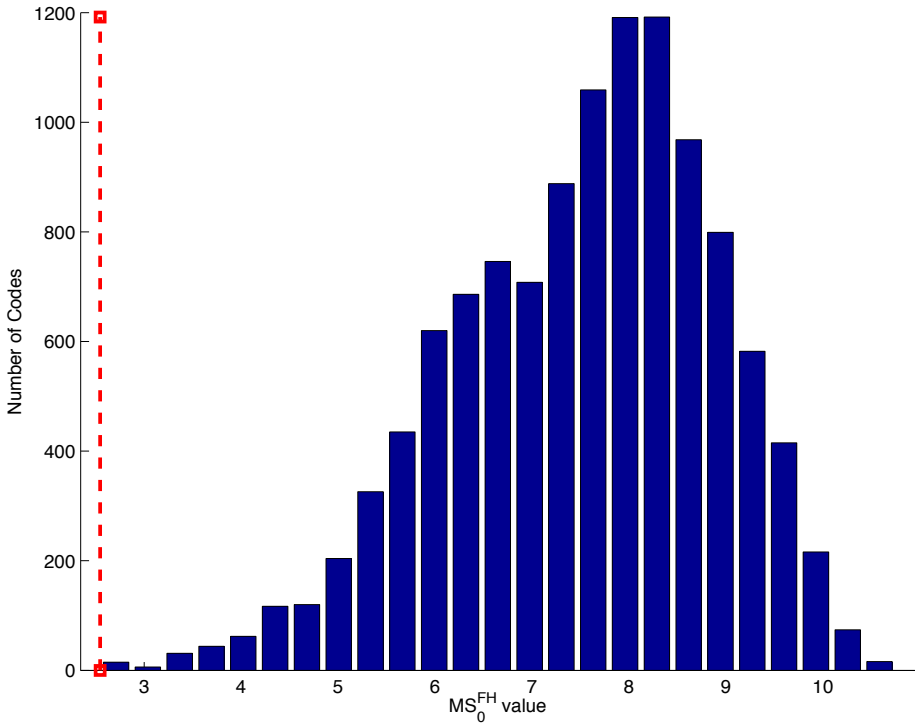


Figure 5.1: Histogram of MS_0^{FH}-values when using the historically reasonable set of possible codes, and fixing Phe, Tyr, Trp, His, Leu, Ile, Arg. Standard genetic code (indicated by dashed red line) is optimal.

Among all genetic codes (in this particular setting of the problem), the SGC is optimal in terms of error-robustness if:

1. We use the updated values of polar requirement [MLS08].

2. We use fixation for Phe, Tyr, Trp, His, Leu, Ile, and Arg, based on aptamer experiments [YWK09, JWKY10].

3. We use the historically reasonable set of possible codes [FH98b].

Figure 5.1 shows a histogram of $MS_0^{FH}(F)$-values resulting from this procedure. When, the original error function $MS_0(F)$ from [HH91] is used, the result is essentially the same: the SGC is the optimal code. We wondered if by fixation of just one or two more assignments, the SGC would be optimal in the space resulting from the combination of these fixations with the random permutations of amino acid assignments according to the method used by Haig and Hurst [HH91], *without* the constraint of the historically reasonable set of possible codes [FH98b]. This was not the case.

5.3 Different stages of code development

What is the biological relevance of the mathematical result presented, if any? Can we indeed conclude that natural selection steered the translation system toward better and better variants of the assignments (in terms of error-robustness) within realistic boundaries? Stated differently, when making a model, should one respect that seven assignments are *fixed*, and that the system evolved *gradually* (as reflected by using the historically reasonable set of possible codes), until the optimal code (within these boundaries) was reached? Or is it rash to arrive at such a conclusion, and could one imagine positive selection for error-robustness to be an illusion?

The space of codes resulting from the constraints imposed on the calculations is a space of very limited size: only 11520 codes ($2! \cdot 2! \cdot 4! \cdot 5!$). The fact that the SGC is optimal in this space is impressive, but of a different order of magnitude than the near-optimalities in significantly larger spaces presented in earlier studies (e.g. [FH98a, FKLH00, GMCR01, BGMLS09, BvdGK$^+$11]). The impact of the different fixed assignments varies: for the MS_0-values, it would theoretically suffice to fix the three assignments of Phe, Trp, and Arg (or any set containing them) in order to find the SGC to be optimal in the resulting space.[3] In this way, the SGC can be thought of as the global optimum in a space of $3! \cdot 4! \cdot 5! \cdot 5! = 2073600$ codes. We further refrain from presenting it thus, because in doing so we would abandon the physico-chemical facts which were the starting point for our calculations with fixed assignments.

It is also possible to *increase* the number of fixed assignments (and in this way *decrease* the size of the space of random code variants) even further. A recent

[3]When using the Freeland and Hurst weights (and hence the MS_0^{FH}-values), it is possible to fix another set of three amino-acids Phe, His, Trp in order to make the SGC optimal.

article [JW10] suggests that more than the seven assignments (listed in Table 5.2) are fixed.

The logical extreme of fixing assignments is that *all* assignments of the SGC are fixed, as argued recently by Erives [Eri11]. In his theory, a kind of RNA cage (pacRNA: proto-anti-codon RNA) is presented, in which different amino acids are bound by different kinds of "walls", which are exposing anticodons to the different amino acids. Although this model combines elegant explanations for several aspects of present-day tRNA functioning, it is very hard to get an objective measure for the specificity of amino acid-anticodon interactions in this model. In particular, the different possibilities allowed by "breathing" of the cage cast doubt on interaction specificity. Some objections can also be raised regarding the tRNA activation mechanism. Yarus and co-workers recently reported a very small ribozyme (only five nucleotides in length) which was experimentally shown to aminoacylate certain small RNAs using aminoacyl-NMPs as activated precursors [TCY10, Yar11]. Such an early activation mechanism, using NTPs as source of energy, is different from the one in Erives' model, where the $5'$ end of the pacRNA is performing this role.

Taking all considerations sketched above into account, it is possible to draw a tentative picture of genetic code evolution which is compatible with the indications concerning which aspects of code evolution are important. Code evolution probably followed classical [Lew51, Ohn70, Fan12] mechanisms of gene duplication and subsequent diversification (here of 'tRNA' genes and genes involved in aminoacylation). Evolution would be mainly by stop-to-sense reassignments [LJ88], with occasional reassignments in only slightly different new or developing uses of codons (cf. [Ard98, VWG06]), not yet massively present in protein-coding sequences (cf. the frozen accident concept [Cri68]). In a proto-biological stage, RNA would be absent while very small peptides could have been synthesized, e.g. by the Salt-Induced Peptide Formation (SIPF) reaction [SR89, RSSB99]. Under prebiotic conditions especially Ala and Gly would be expected to be present in relatively large amounts (see e.g. [HP09, PF11]). Asp-containing peptides could possibly play a role in the origin of RNA, as they could position Mg^{2+} ions in the correct orientation to help polymerize nucleotides, and, concomitantly, keep these ions from stimulating RNA hydrolysis [Szo12a]. Asp content of peptides could be enriched in the presence of carboxyl-group binding montmorillonite surfaces [RSSB99].

In the first stages of coded peptide synthesis, GCC and GGC probably were the only codons in mRNAs [ES78], and coded peptides would consist of Ala and Gly. The remaining codons effectively would be stop codons [LJ88], although functioning without release factors: water would break bonds between tRNA and peptide whenever codons stayed unoccupied for too long. The "single-step biosynthetic distance" between Ala and pyruvate suggests a carbon storage role for these peptides; Gly allowing folding of such molecules. A mRNA/tRNA system func-

tioning without a ribosome has been proposed by several authors [CBBWT61, Woe73, LJ88]. The first rRNA could then have been functioning in improved termination (see above). At this stage the proposal that coded peptides enlarge the possible range of RNA conformations should be taken into account [Nol04].

In the next stage of *coded* peptide synthesis, Asp and Val could have been added to the repertoire (see e.g. [ES78, Ard98, Ike02, Hig09, vdGMG$^+$09]). This would have been a crucial step: enabling *directed* production of the important Asp-containing peptides [Szo12a, vdGMG$^+$09] (see chapter 2 of this thesis) as well as formation of something resembling protein structure, characterized by hydrophobic cores (Val) and hydrophilic exteriors (Asp). The emerging *polypeptides* could have functioned in carbon storage, as mentioned above (the hydrophilic exteriors making the storage molecules soluble in the cytosol). Having started with trinucleotide codons, this aspect was retained, not because four nucleotide codons are in principle impossible, but because this system allowed a further robust development (cf. [VWG06]). Depletion of prebiotic pools of either Ala, Gly, Asp, or Val could have led to the biosynthetic routes involving Gly, Val, Asp, Ala, and pyruvate. In this way the lack of an amino acid could in principle be resolved by use of the other three (cf. the hypothesized carbon storage function of coded peptides). In this respect, it is interesting that during amino acid breakdown [VV95a], Thr is split in Gly and an acetyl-moiety (which is transferred to CoA). Thr itself is synthesized in five biochemical steps [VV95b] from Asp, and therefore, Thr might originally just have been an intermediary on the way from Asp to Gly. In the same way, Ser can be seen as an intermediary on the way from Gly to Ala.

In a further stage, Ser, and Asp-derived amino acids (like Asn, Thr, and Ile) would be added to the repertoire. Asn would be the first amino acid with an entirely biosynthetic origin (it is relatively unstable, and does not accumulate prebiotically). The production of Asn is known to be originally linked to enzymatic conversion of Asp to Asn on a tRNA (see e.g.[Won07]). This is taken as a very old mechanism of biochemistry (i.e. the biochemical conversion *on* the tRNA) but in this thesis the viewpoint is maintained that this is not a justification for assuming that *all* early amino acid biosynthesis happened *on* a tRNA (as seems to be concluded in e.g. [DG08]). One of the problems with this view is what the identity determinants (de Duve's *second genetic code...*) would be of tRNAs with widely different anticodons, but aminoacylated with one amino acid at the base of a biosynthetic family (see the second comment of Higgs on page 14 in [DG08]). An interesting aspect of the Ile biosynthesis routes (cf. [HPO$^+$99]) is the role of alpha-keto-butyrate in these routes. When instead of two molecules of pyruvate, one molecule of pyruvate and one molecule of alpha-keto-butyrate are fed into the Val biosynthesis pathway, Ile is produced instead. Therefore, when both Thr and Val biosynthesis are present, the evolution of just one enzyme (making alpha-keto-butyrate from Thr) suffices for the emergence of Ile. Aptamers can handle this amino acid, and these two factors (easy development

from existing biochemistry and easy manipulation by RNA) could be responsible for the "choice" of Ile (cf. [PF11]).

Larger amino acids like His and Gln would have appeared in a later stage of code development than Asp-derived amino acids like Asn and Thr. The reactions catalyzed by the few enzymes in the Leu biosynthesis which are not enzymes involved in Val biosynthesis (apart from leucine aminotransferase) are reminiscent of the first three reactions of the citric acid cycle [VV95c]. Jensen [Jen76] hypothesized that originally enzymes would have had much broader substrate specificity. With the citric acid cycle being 'old', as well as important for bio-energetic reasons, and Val biosynthesis being present, the system could have produced an excess of Leu. Again, aptamers would be able to "handle" Leu. Existing biochemistry and aptamer potential would thus answer the question why Ile and Leu are part of the Set of Twenty, and e.g. norleucine and alpha-amino-butyric acid are not (cf. [PF11]). Linked to the citric acid cycle and important in nitrogen management are Glu and Gln. A further expansion of the repertoire with a Glu-derived amino acid is the expansion with Arg. Two of the enzymes of the urea (nitrogen management) cycle are related to pyrimidine synthesis enzymes, two others to purine synthesis enzymes [BTS07c]. The last enzyme in the cycle is arginase. This suggests an ancient accumulation of Arg as a side effect of RNA synthesis, upon Glu becoming a major cell component. Arginase could function in bringing the Arg concentration down to acceptable levels. Aptamers could also have evolved to manipulate Arg levels, allowing Arg to become part of the Set of Twenty. Again Jensen's concept of primordial broad substrate specificity [Jen76] is essential to get a possible answer to the "Why these 20?" question: Arg could be part of the set, rather than ornithine and citrulline, because Arg accumulates, and Arg can be manipulated by aptamers.

Fani and co-workers suggest that the AAA pathway of lysine synthesis is older than the DAP pathway of lysine synthesis [FBE$^+$07]; this has as important implication that lysine should be considered as an amino acid belonging to the glutamate family rather than the aspartate family. A further finding is that the first enzymes of this AAA pathway are paralogues of the same citric acid cycle enzymes as the Leu biosynthesis enzymes mentioned above [FBE$^+$07, VLL02, IB98, SC66]. A single set of enzymes would thus originally have been responsible for a large part of converting Asp to Glu, a large part of converting Val to Leu, and the initial part of converting Glu to Lys. Interestingly, enzymes further down the AAA pathway are paralogues of the ornithine-synthesizing part of Arg synthesis [FBE$^+$07]. The picture which emerges (of a limited set of enzymes being responsible for a large amount of enzymatic reactions happening; and of several compounds sharing "highways" of the same sequential biochemical conversions) confirms Jensen's concept of metabolism emergence [Jen76]. Please note the common aspect with the ideas of tRNA repertoire enlargement presented in chapter 4 of this thesis: in the primordial cell, genetic memory space was probably a scarce commodity.

In an advanced stage of code development aromatic amino acids would be added to the repertoire, and release factors would evolve. Van der Gulik and Hoff [vdGH11] (see chapter 4 of this thesis) have argued that codons UUA, AUA, UAA, CAA, AAA, GAA, UGA, and AGA could not function unambiguously until the anticodon modification machinery was developed, which is seen by them as the last development leading to the full genetic code. Because archaea and bacteria have different solutions for the "AUA problem" (agmatidinylation vs. lysidinylation [vdGH11]), unambiguous sense assignment of AUA must have been late indeed.

The SGC has probably evolved in a genetic environment characterized by rampant horizontal gene-flow [VWG06]. The interaction between genetic systems with slightly different, still-evolving codes, is thought to have caused both universality and optimality of the SGC [VWG06]. Universality, because the genetic code functioned as an innovation sharing protocol [VWG06]. Optimality, because competition allowed selection for the ability to translate the genetic information accurately [VWG06]. The work presented in this chapter illuminates constraints within which this process of genetic code development took place. Both the step-by-step increasing complexity of biochemistry, and the stereochemical relationship between at least some amino acids and triplets, are factors which have to be taken into account.

In summary, although there are at least two different lines of research suggesting a greater number of fixed assignments than the seven given in Table 5.2 (based on the work of Yarus and co-workers [YWK09, JWKY10]), for now it is not clear that more (or even all [Eri11]) assignments are fixed. Thus, the observed error-robustness still needs explanation. It is possible that the optimality of the SGC we found results from positive selection for error-robustness, though starting within a more restricted set of possibilities than previously thought.

5.4 Molecular Structure Matrix

Polar requirement is just one physico-chemical aspect of amino acids. The discovery that only 1 in 10000 random codes has a lower error-robustness value than the SGC when polar requirement is used as an amino-acid characteristic [HH91, HH99] is compelling evidence that error robustness is present in the SGC. When a conservative attitude is taken, and a phenomenon is considered noteworthy only when the probability to encounter it as a random effect is less than 0.1 %, the SGC is clearly noteworthy. If one considers the error-robustness values for the three positions separately [HH91, FH98a] (please refer to [BvdGK$^+$11] or chapter 3 for details) the results in the left column of Figure 5.2 are obtained. The third position is in the less than 0.1 % category, the first position is in the less than 1 percent category, while the second position, with about 22 %, is not even in the less than 5 % category, and can thus not be

considered special.

This result is not entirely satisfactory, because the codons of several pairs of similar amino acids are related by second position changes. For instance, a change from phenylalanine (Phe) to tyrosine (Tyr) is clearly a conservative change from a biological viewpoint. The fact that several conservative changes of this type (e.g. Thr-Ser, His-Arg) can be quickly spotted, urges to a more in-depth investigation into second-position error-robustness. To develop a measure for this kind of amino-acid relatedness (e.g. the Phe-Tyr similarity), we introduce a new way of measuring amino-acid similarity by one-atom changes which yields a measure of similarity in terms of molecular structure. We should stress that this measure does *not* reflect actual chemical reactions/steps. As an example, we compute the distance between Phe and Tyr to be 3 as follows: the hydrogen atom at the end of the side chain of Phe is taken off as a first step. An oxygen atom is placed on the position which the hydrogen atom had before as a second step. The Tyr molecule is completed by addition of an hydrogen atom on top of this oxygen atom, producing the hydroxyl group at the end of the side chain of Tyr, and this is the third and final step. Generally, the distance between two molecules is defined to be the minimal number of "allowed one-atom changes" to transform one molecule into the other, where the allowed one-atom changes are the following:

- taking off or attaching an arbitrary single atom,
- creating or destroying a single bond (thereby possibly opening or closing a ring structure),
- changing a single bond to a double bond or *vice versa*.

It is not hard to see that an algorithmic way of computing the distance between two molecules m_1 and m_2 is to find the maximal common sub-graph m_c of their molecular structure and to sum up how many steps are required to go from m_1 to m_c and from m_2 to m_c. The distance matrix between the 20 amino acids in Table 5.3 has been obtained in this way, using the Small Molecule Subgraph Detector (SMSD) toolkit [RBH$^+$09] to find the maximal common subgraph and post-processing this information with a python script.

In order to perform the error-robustness calculations, we followed the procedure by Haig and Hurst [HH91] and considered the squared distances. In this way, the zeroes in the diagonal remain zero. The values for small changes become slightly larger (so the edge from Phe to Tyr gets a value 9) while the values for large changes (like going from Gly to Tyr) become considerably larger (in the case of Gly to Tyr 20 becomes 400). Large changes thus get stronger emphasis [DG89a]. Whether squaring is the right way to make these kind of calculations has been discussed elsewhere [Ard98, FKLH00]; we just want to compare molecular structure as an input to characteristics like polar requirement, hydropathy,

	Phe	Leu	Ile	Met	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Trp	Arg	Gly
Phe	0																			
Leu	15	0																		
Ile	21	10	0																	
Met	21	14	14	0																
Val	22	15	5	11	0															
Ser	17	12	14	10	11	0														
Pro	17	8	8	10	11	10	0													
Thr	20	13	9	9	6	5	9	0												
Ala	16	11	13	9	10	3	9	8	0											
Tyr	3	16	22	22	23	18	18	21	17	0										
His	18	15	17	17	18	13	13	16	12	19	0									
Gln	20	13	13	11	12	11	9	10	10	21	12	0								
Asn	19	14	16	12	13	8	12	11	7	20	13	13	0							
Lys	17	12	12	14	15	14	8	13	13	18	17	13	16	0						
Asp	18	13	15	11	12	7	11	10	6	19	14	12	5	15	0					
Glu	19	12	12	10	11	10	8	9	9	20	15	5	12	12	11	0				
Cys	17	12	14	10	11	4	10	9	3	18	13	11	8	14	7	10	0			
Trp	12	23	27	27	28	23	23	26	22	15	18	22	25	23	24	25	23	0		
Arg	24	15	15	17	18	17	11	16	16	25	10	12	19	15	18	15	17	24	0	
Gly	19	14	14	12	11	6	12	9	5	20	15	13	10	16	9	12	6	25	19	0

Table 5.3: Molecular structure matrix. The entry in row i and column j denotes the number of steps required to transform the ith amino acid into the jth. (The gray tones are for ease of reading only, they do not carry special meaning.)

volume and isoelectric point, as studied by [HH91].The histograms of the error-robustness in terms of molecular structure are shown in the right column of Figure 5.2.

Polar requirement is producing a result in the less than 0.1 % category [HH91]: only 1 in 10000 codes has a lower value than the SGC [HH99]. When molecular structure is used as input, the result is *not* in the less than 0.1 % category. However, it is still remarkable that the SGC is, with 0.151 %, in the less than 1 % category when molecular structure is used as input. This means that this matrix is performing better than volume or the hydropathy scale of hydrophobicity in the work of Haig and Hurst [HH91]. Even more remarkable, the error robustness comes mainly from the second position, using this measure (Figure 5.2).

One of the options to explain *this* kind of error robustness is that it is caused by the gradual growth of the amino acid repertoire, from small amino acids like alanine and glycine, via aspartate-derived amino acids to larger amino acids (like leucine and histidine), and finally aromatic amino acids. This trend can also be seen when one takes the arithmetic mean of the molar mass of amino acids sharing the same first nucleotide: from 112.31 g mol^{-1} (for Val, Ala, Asp, Glu,

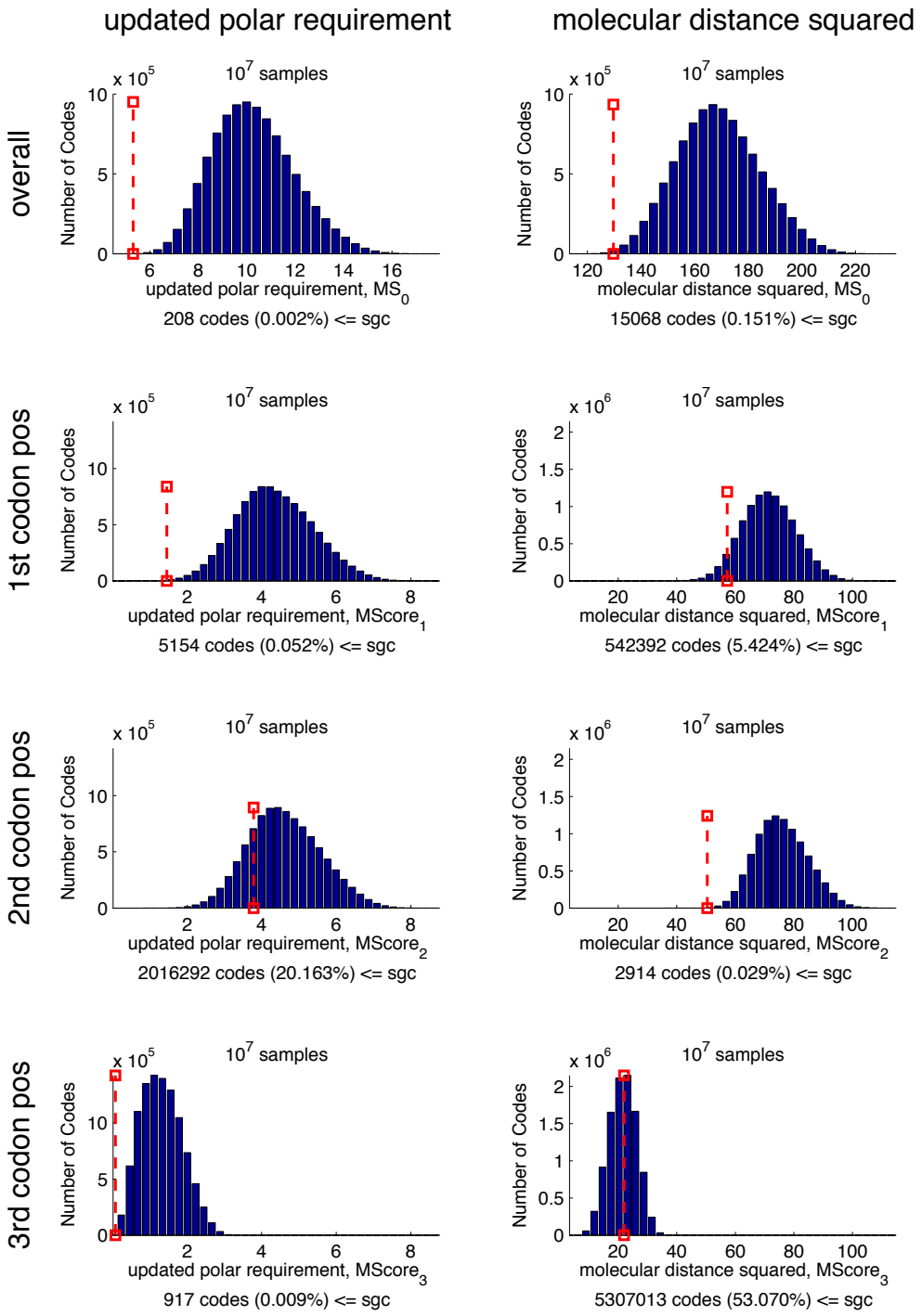


Figure 5.2: Histograms of the MS-values of 10 million random samples using updated polar requirement [MLS08] (4 histograms on the left) and molecular-structure distances from Table 5.3 squared (4 histograms on the right). The top row shows the MS_0 value, the second row is the component from the first codon position (MScore_1), third and forth row the components from the middle (MScore_2) and last (MScore_3) codon position. In contrast to the original definition [HH91] of MS_i for $i \geq 1$, we have chosen to normalize MScore_i with the same constant as MS_0 so that $MS_0 = \sum_{i=1}^{3} \text{MScore}_i$. The dashed red line indicates the value of the SGC.

and Gly) via 136.73 g mol^{-1} (for Ile, Met, Thr, Asn, Lys, Ser, and Arg) to 144.36 g mol^{-1} (for Leu, Pro, His, Gln, and Arg), and finally 151.34 g mol^{-1} (for Phe, Leu, Ser, Tyr, Cys, and Trp).

5.5 Why these twenty?

Fascinating questions can be asked concerning the basic characteristics of the protein synthesis mechanism. Why these twenty amino acids? Why twenty amino acids, and not 64? Why not 16, or less? Why four kinds of nucleotides? Why these 4 kinds? Why codons consisting out of three nucleotides, and not four, five, six, or two? Why only amino acids (strictly speaking, proline is an imino acid), and not hydroxy acids, or thio acids, or other organic molecules? Why carbon compounds and not silicium compounds? Some answers to these broad questions can be found in recent publications [CI10, MXW10, PF11], and some slightly older ones [VWG06, HP09]. A basic aspect with respect to this kind of questions, however, is illustrated by proline: because biological development is characterized by growth from simpler beginnings, all twenty "amino acids" resemble each other. Although not being an amino acid, proline has the same L configuration as 18 of the other 19 (in a sense, glycine is too small to have an L configuration). *The "twenty amino acids" are variations on one theme, rather than being twenty amino acids.* One can therefore try to sketch a broad panorama of how this evolutionary development probably unfolded, roughly (see section 5.3 for a slightly more detailed panorama). Starting with just one or two tRNA molecules, the SGC developed from an "early-biological" stage, via a number of intermediary stages to the contemporary stages (there are domain-specific aspects of the tRNA sets [NPEPRDP12], there are "extra amino acids" like selenocysteine in several lineages [YOA+10, GZGCK11], and there are code variants [KFL01, SYH07, JVvLG09, TOL10]). During this process, compounds were added which (1) were available; (2) could fit into the translation process; and (3) were innovations to the amino acid repertoire whose addition had a positive feedback on the system. So, starting with molecules present in the prebiotic mixture of interacting organic molecules (like glycine and alanine), the repertoire was slowly expanded, probably first with valine and aspartic acid, then with molecules derived from aspartic acid (like isoleucine), next with leucine and glutamic acid, followed by molecules derived from glutamic acid (like proline). Finally, relatively large amino acids, like tyrosine and tryptophan became part of the repertoire. All these expansions are L-amino acids (or something resembling an L-alpha-amino acid very much: proline) *because they fit into the same "production line"* as glycine and L-alpha-alanine.

A fascinating aspect of this growth process is that it has seemed to occur in a fashion in which error robustness was built into the coding rules. If the process started with only one or two tRNAs present, most of the codons will originally

have been stop codons [LJ88, vdGH11], not in the sense that they would be recognized by a release factor (we do not expect release factors to exist in the early stages of the genetic code), but in the sense that decoding would take so long that hydrolysis would happen earlier than suppression (see section 5.3). The first rRNA can be seen as a molecule allowing this process to happen faster. There would be strong negative selection going on against the use of such "stop codons" inside protein coding sequences, but their frequency of occurrence would not be zero. The suppression by the tRNA most prone to suppression would give the stop codons a kind of "amino acid character", leading to a "four-column" structure as proposed by Higgs [Hig09]. Introduction of a new tRNA transferring a new amino acid in a way that the repertoire expansion respects this "character" would be minimally disruptive to the system [HP09]. The consequence of this way of evolutionary development (and of the fact that in many cases one tRNA molecule recognizes several codons) is a degree of error robustness to substitution mutations.

Francis [Fra13] points to a second kind of relationship which can be found between amino acids in Higg's four-column structure. Citing work of Nevskaya et al. [NTV$^+$06], he points out the improvement in function realized by introducing lysine on a position where originally a glutamate was functioning, in cooperation with a K^+ ion. After the introduction of a $tRNA^{Lys}_{CUU}$, GAG codons encoding a glutamate involved in such a K^+ coordinating role are prone to change in AAG, in the cases in which the change is an improvement. This is a much smoother evolutionary development than the one from the modelling approach of [Hig09], see also section 4.4 in chapter 4. Of course, Francis' reasoning is not restricted to columns: relationships like those between phenylalanine and tyrosine, and between aspartate and glutamate can be seen in his way too.

One aspect which should not be left unmentioned, is the role of some "amino acids" and their derivatives (glutamate, glycine betaine, and proline betaine) in osmoprotection. K^+, glutamate and glycine betaine are osmoprotectants in all three domains of life [RGW09, Rob04]. Despite the fact that proline was thought to have no role in archaeal osmoprotection [Rob04], proline betaine was found to be dealt with in much the same way as glycine betaine in the archaeon *Archaeoglobus fulgidus* [SHD$^+$04]. Contrary to the ideas of convergent evolution and horizontal gene transfer, the possibility of a role in osmoprotection of all three amino acids (Gly, Glu, *and* Pro) in the LUCA should be considered. Being an osmolyte (directly, or in a di- or tri-methylated form) has as a consequence that the molecule is optionally present in large amounts in the cell (like storage compounds). In some ways, there is thus a relationship between roles in carbon storage, and roles in osmoprotection. Maybe an osmoregulatory role for glutamate had as an effect large amounts of this molecule being present in the cell. This could be followed by formation of glutamine (think again of Jensen's low substrate specificity [Jen76], see section 5.3, and keep in mind that an asparagine-

producing activity is present in the cell[4]), and as a side effect of accumulation of glutamic gamma-semialdehyde (an intermediary between glutamate and glutamine), an accumulation of proline. This molecule could then very well have started as an osmoprotectant (being thus present in large amounts) before having entered the set of twenty. The synthetase enzymes have proofreading sites which destroy misacylated tRNAs. One can envision that with the accumulation of proline, severe misacylation of $tRNA^{Thr}$ with proline was occurring (remember the resemblance of serine and threonine with an intramolecular hydrogen bond to the proline ring-structure as mentioned in [WOIS00]). The proofreading sites of synthetases are thought to have been independent soluble proteins before they became part of the synthetase protein chain [AKIS03]. Developing this kind of proteins [AMF04, CYS08] could have been the answer to the misacylation. Developing a new kind of tRNA, and a new kind of proofreading enzyme subsequently could establish proline as a *bona fide* member of the set of twenty. The same series of events could have led to four more serine codons with middle C[5]. We therefore see that such evolutionary sequences of events can suggest an answer to both the question "Why these twenty?" and the question "What is the mechanism behind error robustness development?". Of course we have already seen similar answers concerning isoleucine, leucine, and arginine (see section 5.3). With the codons of Phe, Tyr, His, and Trp fixed by aptamer considerations, there is not much room left for specific selection on error robustness in polar requirement. It is mainly the sulfur-containing amino acids and lysine whose assignments are left to explain (their polar requirement values being so close to the ones of Ile, Trp, and Asn). Much of the error robustness of the SGC might very well be an inherent characteristic of our universe, being due to stereochemical interactions between amino acids and nucleotide sequences or to constrained pathways of biochemical development.

[4]An asparagine-producing activity would be present in the cell if asparagine is considered to be one of the earliest members of the "Set of Twenty".

[5]That is, if one considers the AGY codons to be the older codons encoding serine, and the UCN codons later codons encoding serine, being added in a later stage of code development.

Chapter 6

The danger of losing information

The content of this chapter is based on joint work with Harry Buhrman, Simone Severini, and Dave Speijer (arXiv:1307.1163[q-bio.PE]).

6.1 Shrinking pressure and large deletions

In this chapter, we investigate the behavior of genetic material which is under pressure to shrink. A very generally occurring aspect of cells is that not all genes are needed at the same time. Because of this, we have regulation of gene expression in biology. That we do not express all genes in a genome under a specific growth condition leads to the situation that not all genes in a genome are needed under that condition. The genes not needed during that growth condition are "momentarily" not under any selective pressure at all. The risk therefore is that these genes become irreparably damaged during that period. This is a real "problem" for the cell, because when conditions alter these damaged genes may become essential. This problem is exacerbated under conditions where (1) there is intense competition to reduce the genome to its smallest possible size because this increases growth rate; and (2) the time interval between the need for different sets of genes is relatively long. Both of these conditions are present in the case of the mitochondrial DNA of the parasite *Trypanosoma brucei* (known because it is the parasite causing African sleeping sickness). We suggest that the unusual way in which this DNA is organized, is actually functioning as a counter-measure against the danger of losing information [Spe06, Spe07]. The problem of losing genes that are "currently not under evolutionary pressure but that will become very important in the relatively near future" is a general problem in biological systems. We will focus here, as a particular example, on the tendency of the mitochondrial DNA of *T. brucei* to suffer large deletions. We begin by describing the unusual organization of this mitochondrial DNA.

6.2 *Trypanosoma* mitochondrial DNA

Two types of DNA molecules are present in the mitochondria of *T. brucei*: the minicircles and the maxicircles [MAGH02]. On the DNA of the several thousand minicircles, sequences coding for guide RNAs are found. The "normal" genes found in mitochondria (genes encoding the rRNAs of the mitochondrial ribosomes and genes encoding several mitochondrial proteins) are located on the maxicircles, of which a few dozen are present. Some guide RNAs are also encoded on the maxicircles [BBS90]. A type of gene *not* found in these mitochondria, is genes coding for mitochondrial tRNA: all tRNAs, necessary for mitochondrial translation, are imported from the cytosol. All circles together form a highly concatenated network in *T. brucei*. The function of the guide RNAs is the processing of RNA transcripts of protein-coding genes: many U's are not present in the sequences of several of these genes. Functioning as messengers in the protein production process of the *T. brucei* mitochondrion only becomes possible *after* guide RNAs have processed the mRNAs. This processing occurs in an obligate order, starting at the $3'$ end of the mRNA and progressing backwards to the $5'$ start of the message. The phenomenon of changes made to the transcribed sequence is known as editing of RNA [BvdBB86]. In the case that editing is so extensive that sophisticated sequence recognizing software can no longer recognize the unprocessed sequences, the name pan-editing is used for the process. In *T. brucei*, pan-editing of several protein-coding genes is found.

A few remarks can be made when the pan-editing of *T. brucei* is considered. The first remark is that laboratory lineages of parasitic cells maintained outside the natural hosts tend to lose the phenomenon of pan-editing and "return" to a much more normal organization of the mitochondrial translation process. This provides a dramatic demonstration that editing indeed is maintained because it is under selective pressure that is absent in the lab. The second remark is that the information necessary to make a protein which is produced via pan-editing is, in effect, distributed over a much larger area of the mitochondrial DNA than would have been the case when no editing would have been present. The third remark is that this information is *mixed* with the information of other protein-coding genes which are also subject to pan-editing. The fourth remark is that such a way of running the protein production, and of running the maintenance of the sequence information underlying that protein production, is necessarily biochemically very costly. A few isolated cases of RNA editing could be considered as frivolous behavior of evolution; the massive pan-editing present in *T. brucei* must involve such a significant part of the energy budget of the parasitic cell, that a *bona fide* advantage from an evolutionary biological viewpoint *has* to be present. Insight into the evolutionary advantage of pan-editing is the goal of the research presented in this chapter.

Two features of the life cycle of *T. brucei* are of particular importance to

the argument made here: intense clonal selection in the bloodstream and the involvement of two biochemically distinct hosts of the parasite. The human victim is infected with the parasite after being bitten by a Tsetse fly (genus *Glossina*). The parasite starts to live and multiply in the blood of the victim, and feeds on the abundantly present glucose in the blood. Clonal reproduction follows, and very high numbers of parasites can accumulate in the blood. An important aspect is that the feeding situation is so rich that several parts of the mitochondrial DNA are no longer under selective pressure. To close the life cycle and get on to the next phase in *T. brucei* biology, the victim (lying exhausted on the bed because being robbed of his/her energy by the millions of parasite cells) needs to be bitten by a Tsetse fly again. In the salivary glands of the fly, the parasite goes through other stages; important is that in these stages the parasite needs the aspects of the proteome encoded by the mitochondrial DNA that were not under selective pressure in the human blood stream. The last stage in the fly's salivary glands is the stage which can infect a new human victim. Please note that nothing is mentioned about *T. brucei* sexuality: our understanding of parasitology of even long-studied and economically-important organisms is still fragmentary and incomplete.

If a shorter genome leads to a faster replication time, a large deletion in the genome might be advantageous, during part of the life-cycle. The take-home-message of the description of this parasite's life cycle is that a parasite which has got a deletion in mitochondrial DNA early in the infection process in the human host, is in an advantageous position compared to the parasite cells without that deletion from the same "batch of infection". As long as the deletion is in the part of the DNA temporarily not under selective pressure, the multiplication time of this cell and its progeny will be shorter than that of its clonal competitors. Small differences of this kind can have large effects in the long run: the example of the amount of rice on a chessboard when the amount is doubled when the next field is entered is the classic one to keep in mind. We draw attention to the fact that the amount of progeny of these doomed cells (doomed because they will not be able to have progeny going through the *Glossina* salivary gland stages and close the life cycle) possibly will swamp the numbers of cells with a complete mitochondrial genome. Very likely the danger is real that when the fly bites, all parasite cells entering the fly will be dead-end "winners". These "winners" would no longer be able to infect the fly and would thus be evolutionary dead ends destined to die together with the person they infected without having the possibility of infecting other people.

If the parts of the parasite mitochondrial DNA under selective pressure in the human host and the parts of the parasite mitochondrial DNA not under selective pressure in the human host would be thoroughly mixed, gaining advantage of a large deletion would simply be impossible. Every large deletion would not only affect DNA temporarily not under selective pressure, but also DNA of direct vital

importance. Swamping would be no issue, even progeny with a large deletion in the mitochondrial DNA would be no issue. With this concept [Spe06, Spe07] published, and accepted as not being improbable (reply to [Spe08]), we want to go one step further, and establish in a mathematical rigorous manner that the argumentation, which is intuitively appealing, can also stand the test of being able to be mathematically modeled. Surely, constructions which sound intellectually acceptable when presented in a general way, are sometimes found to be false, when approached with mathematical rigour. It is our ambition to demonstrate that the concept is a fertile one, even when being put to the test of facing strict mathematics.

6.3 Modeling *Trypanosoma* mitochondrial DNA

We now describe the mathematical model we have made to represent the genetic organization of *T. brucei* mitochondria. With A and B we denote the two environments of *T. brucei*. In principle, we could distinguish three kinds of mitochondrial DNA in the parasite. First, there are genes which lead to products that are needed both in environment A and environment B. Second, there are genes which lead to products that are needed in environment A, but not in environment B. Third, there are genes which lead to products which are needed in environment B, but not in environment A. We do not expect that significant parts of the *T. brucei* mitochondrial genome are not needed in both environments, because, compared to the bacterial chromosome, mitochondrial DNA has a very small size, and appears to be as small as possible. Our first simplification is that we reduce these three kinds of DNA to just two kinds. With n_b we denote a sequence of mitochondrial DNA needed in environment B, but not in environment A. This is the third kind of the three kinds just mentioned before. With n_a we denote the rest of the sequence of mitochondrial DNA. The total length of the mitochondrial DNA will thus be $n = n_a + n_b$ base pairs. We have chosen n_a to be 66 % and n_b to be 34 %.

Here we examine the question how fragmenting the sequence of mitochondrial DNA needed in environment B, but not in environment A, and mixing this genetic information with the rest of the mitochondrial DNA protects the integrity of the mitochondrial genetic information. We study the distribution of size classes of mitochondrial DNA in environment A, when deletions take away parts of the sequence, and competition favours parasites with smaller mitochondrial DNA. Fragmenting n_b (which is temporarily relieved from being necessary for survival) and mixing the resulting B-blocks with likewise fragmented DNA from sequence n_a protects the vulnerable information.

Our second simplification is that we ignore that in *T. brucei* many circles of DNA are physically concatenated. We model the *T. brucei* mitochondrial DNA

as one circular molecule, and subsequences of n_a and n_b are intertwined such as to have a sequence $a_1b_1a_2b_2...a_kb_k$. We call k the level of fragmentation; in this way, $k = 1$ when both n_a and n_b are not fragmented and not mixed (the total sequence being thus a_1b_1), $k = 2$ when both are broken in two and mixed (such as to get a total sequence $a_1b_1a_2b_2$), $k = 3$ when both are broken in 3 pieces and mixed (producing $a_1b_1a_2b_2a_3b_3$), and so on. Because the molecule is circular, b_k connects with a_1. In our model, all fragments of n_a have the same length, and also all fragments of n_b have the same length. The length of one block a_i is n_a/k, and likewise the length of one block b_i is n_b/k; please note that the B blocks do not have the same length as the A blocks because $n_a : n_b = 66 : 34$. The mitochondrial DNA of an individual surviving a round of replication is represented by the sequence $a'_1b'_1a'_2b'_2...a'_kb'_k$. When deletions are going to start damaging the B-blocks, these will of course start to differ in length: $b'_i \leq b_i$. All deletions affecting n_a are instantaneously lethal, and so: $a'_i = a_i = n_a/k$. Note that b'_i may become 0. During every round of replication, we model a probability $p = 10^{-6}$ with which a deletion of mitochondrial DNA with a random length at a random position may occur. The random position is modeled by picking uniformly at random a point *pos* for which $1 \leq pos \leq n$; the random length is modeled by cutting out an uniformly random chosen length l for which $1 \leq l \leq n$. Surviving individuals replicate such, that the smaller their total mitochondrial DNA, the faster they replicate; this advantage is modeled by a function $r(x)$, see below.

The process of development of a population of parasites with different amounts of mitochondrial DNA is modeled using a Markov chain approach. The states of the Markov chain represent the different classes of mitochondrial DNA. The entries of the Markov chain will not be probabilities since they can be larger than 1. At level k we have $((n_b/k) + 1)^k$ states in our Markov chain. The extra state (the "+1") is the state in which a B-block b_i is still complete. The quotient n_b/k gives the series of possible states with a deletion in a B-block b_i (when e.g. $n_b = 6000$ and $k = 3$, each B-block b_i is 2000 base pairs in length and 2000 different deletions can happen (and survive in environment A), the largest of them removing the complete B-block b_i). The power k is present because all different *combinations* of deletions in *different B-blocks* have to be present in our Markov chain. During the first round of replication no combination will be reached; but from the second round of replication onwards, combinations of different deletions can be the result of deletion.

6.3.1 The replication advantage function

When (part of) n_b is removed, the individual will replicate faster, which is modeled as follows. The total size of the vulnerable part of the mitochondrial DNA (the B-part of sequence $x = a_1b_1a_2b_2...a_kb_k$) is $s(x) = \sum_{i=1}^{k} b_i$. We now set two factors, $\max_r$ and $\min_r$; individuals which lost all their vulnerable mitochondrial DNA

replicate with $\max_r$, and individuals which still have this DNA intact, replicate with $\min_r$. The function that models the replication advantage is

$$r(x) = \max_r - s(x) * ((\max_r - \min_r)/n_b) \tag{6.1}$$

This function linearly interpolates between $\max_r$ and $\min_r$, depending on the size of the DNA. When there is no vulnerable mitochondrial DNA left, $s(x) = 0$, and $r(x) = \max_r$. We set the values for $\max_r$ and $\min_r$ as respectively 3 and 2, and therefore, when the vulnerable mitochondrial DNA is still complete (i.e. when $s(x) = n_b$), $r(x)$ will be $\max_r - (\max_r - \min_r)$ which is 3 - (3 - 2), which is 2.

6.3.2 The graph of the Markov chain

The Markov chain can be seen as a graph G. The nodes are the different living states of the parasite. The edges are the events of deletion where the new sequence can appear from the old sequence; the events of "non-deletion" (the great majority of events during replication) are edges which are self loops in the graph. The edges are labeled with the probability that the event occurs: this is p for the deletions and $1 - p$ for the self loops. The edges are directed: the parasites can lose information, but can not get it back. There is a directed edge from node x to node y in G when sequence y can be obtained from sequence x via a *single* deletion.

The graph of the Markov chain is called $G(E, N)$, where the set E is the set of edges (deletion events *and* self loops), and the set N is the set of nodes corresponding to all the possible states of the parasite being alive (the sequences $a_1 b'_1 a_2 b'_2 ... a_k b'_k$, with $b'_i \leq b_i$). Next, a matrix D is constructed from the Markov chain, with a size $|N| \cdot |N|$. When an entry $D(x, y)$ is not an edge in the graph, the entry is zero. If the edge exists in our graph, the entry is equal to the label of the edge (x, y).

The starting state of our process corresponds to all the individuals that have all their DNA still present. This corresponds to the unit vector v_0, with the entry $v_0(1) = 1$ and the entries $v_0(i) = 0$ for $2 \leq i \leq |N|$. In other words: we fix the first entry of our vector to correspond to the state where all the DNA is still present. Likewise entry x corresponds to the fraction of individuals in state x.

In order to model the replication part of the process we define the diagonal matrix $R(x, x) = r(x)$. Multiplication with R corresponds to replicating state x with replication factor $r(x)$. A single deletion step followed by a replication step is now simply the matrix $M = RD$.

The vector $v' = Mv_0$ corresponds to our population after a deletion and replication step of our process. Note that the vector v' does not have L_1 norm 1 anymore[1]. We now need to take into account the boundary conditions induced

[1]The L_1 norm of v, $|v|_1 = \sum_{i=1}^{|N|} |v(i)|$

by the maximum population size of the parasites as follows. We would like to view v_0 and v' as the probability distributions over the state space. Initially all the probability mass is on the full DNA state and progressively this mass flows to other states. We can then interpret the multiplication of $v_t(i)$ with the size of the maximum population s_p, $s_p v_t(i)$ as the *expected* number of individuals that have DNA corresponding to state i after t generations. This means that we have to renormalize our vector: $v_1 = v'/|v'|_1$ in order to make it a probability. This completes one full generation step of our process, and in general:

$$v_t = \frac{M v_{t-1}}{|M v_{t-1}|_1}$$

Since M is a linear map, we may renormalize at the end, so that:

$$v_t = \frac{M^t v_0}{|M^t v_0|_1}$$

Note that implicit in M is the value of k, the fragmentation level, which we have omitted in our notation so far for simplicity. Note that when k grows so does the size of the Markov chain. Keeping track of this parameter we define for each k:

$$v_{t,k} = \frac{M_k^t v_{0,k}}{|M_k^t v_{0,k}|_1}$$

Let $t_{max}(k)$ be the maximum t such that $s_p v_{t,k}(1) \geq 1$. The value $t_{max}(k) + 1$ tells us the expected number of generations until there are no individuals left that have their full DNA, at a fragmentation level k. We are interested in the growth rate of this function $t_{max}(k)$.

6.3.3 State Space Reduction

As mentioned before, the number of states in our Markov chain is $((n_b/k) + 1)^k$. With realistic values of n_b and k (n_b being about 10^4 base pairs and k being around 200) this has as a consequence that the Markov chain becomes too big: simulation becomes computationally expensive, to the point where work becomes impractical. However, we can reduce the number of states in our simulation in two different ways. First we can exploit some symmetries in our problem. Second we can simplify our model further by introducing the concept of *confidence level*. We will explain both ways of reducing state space, and then explain how this affects the transition probabilities in our Markov chain.

Exploiting Symmetries

We are only interested in the number of generations which it takes until no parasite cells are present with a complete n_b. This means that the only entry of the

tuple giving the number of individuals present in each state which is of interest to us, is the first entry. As we do not need the information of the states $a_1b'_1...a_kb'_k$ with $b'_i \leq b_i$, we only need to keep track of how *many* blocks we have of size $0 \leq s \leq n_b/k$. The starting state corresponds to the tuple with $((n_b/k)+1)$ entries $(k, 0, ..., 0)$, the first entry indicating how many blocks we have of size n_b/k (the full size with not yet a deletion damage), the second entry indicating how many blocks we have of a size one base pair less, and so on. The last entry then indicates how many blocks we have of size 0. Following this convention, the tuple describing the fully depleted state becomes $(0, 0, ..., k)$.

These are precisely the states $(c_1, c_2, ..., c_{((n_b/k)+1)})$ such that $\sum_{i=1}^{((n_b/k)+1)} c_i = k$. This corresponds exactly to the number of multisets of cardinality k with elements taken from a finite set of cardinality $((n_b/k)+1)$. This multiset coefficient is equal to $\binom{(n_b/k)+k}{k}$ and can be bound from below by $(((n_b/k)+k)/k)^k$. This second representation is significantly smaller than our initial set-up, but still too large for the range of parameters we are interested in. We therefore have to simplify our process further.

Further Simplification

We modeled a deletion of DNA by randomly picking a position *pos* in the (circular) DNA and then remove a piece starting at *pos* of random length l. Individuals survived this deletion whenever only DNA from within a B-block i was deleted. We now simplify this as follows. Fix a parameter d, $1 \leq d < n_b$, which we call the *confidence level*. We will only keep track for each block i when it has size $a * n_b/(k * d)$, with $0 \leq a \leq d$. Whenever a random deletion left us with a block size

$$a * n_b/(k * d) \leq b'_i < (a+1) * n_b/(k * d),$$

we set the block size $b'_i = a * n_b/(k * d)$, while keeping the probability of this event the same[2]. We thus give slightly more probability to deleting larger parts within block i. We will see later that this change is not very significant. For example, setting $d = 1$, models that whenever a deletion falls within block i, we completely remove block i (i.e. it will have size 0). On the other hand for $d = n_b/k$ we get back our old process. Confidence level d thus allows us to interpolate smoothly between the simplified process and the original process.

For confidence level d, the states of our Markov chain will be $(d+1)$-tuples $(c_1, ..., c_{d+1})$ such that $\sum_{i=1}^{d+1} c_i = k$. The first entry indicates the number of blocks that have size $n_b/k = (n_b/(k * d)) * d$, the second entry describes the number of blocks that have size $(n_b/(k*d)) * (d-1)$, and the last entry the number of blocks that have size $0 = (n_b/(k * d)) * 0$.

[2]Strictly speaking we should write $\lceil a * n_b/(k * d) \rceil$, we approximate this and assume that n_b is divisible by $d * k$.

The number of states we have in our Markov chain, with fragmentation k at confidence level d is equal to the number of multisets of cardinality k out of a finite set of cardinality $d+1$, which is $\binom{k+d}{k}$.

Transition Probabilities

Here we will make precise the transition probabilities between any pair of states in our Markov chain. Given any state $x = (c_1, ..., c_{d+1})$ such that $\sum_{i=1}^{d+1} c_i = k$. The transition from x to x (i.e. no deletion occurred) is labeled with $(1-p)*r(x)$, where $r(x)$ is taken as in equation 6.1 with $s(x)$ the size function for these simplified states:

$$s(x) = n_a + \sum_{i=0}^{d} c_{i+1} * \frac{n_b(d-i)}{k*d}$$

Transition from $x = (c_1, ..., c_{d+1})$ to $x' = (c'_1, ..., c'_{d+1})$ is only possible if there is an $i < j$ such that $c_i = c'_i + 1$ and $c_j = c'_j - 1$, and for all the other indices i' the states are the same: $c'_i = c'_{i'}$. This guarantees that exactly one B-block of size corresponding to $i : n_b(d+1-j)/(k*d)$ transforms, by means of a deletion, to a block of size corresponding to $j : n_b(d+1-j)/(k*d)$. The probability that this transition occurs turns out to be:

$$\frac{c_i * m(j,k,d)}{s(x)^2} * p \tag{6.2}$$

where $m(j,k,d)$ is the number of ways one can transform a block of length corresponding to i, i.e. of length $n_b*(d+1-i)/(k*d)$, to a block of length corresponding to j, using the rule of rounding down described in subsubsection "Further Simplification". Note that this number only depends on j, k, and d and not on i. For example if $i = 1$ and $j = 2$ this corresponds to the number of ways one can delete a sequence of length 1 up-to $n_b/(k*d)$ in a sequence of length n_b/k, which is equal to

$$(n_b/k) + ((n_b/k) - 1) + ... + (((n_b/k) - (n_b/(k*d))) + 1).$$

In general this becomes

$$m(j,k,d) = \sum_{i'=\frac{n_b*(d-(j-1))}{k*d}+1}^{\frac{n_b*(d+1-(j-1))}{k*d}} i'$$

In equation 6.2 we divide by $s(x)^2$ because each possible deletion has probability $s(x)^2$ to occur at a fixed position and is of a fixed length. Finally we multiply in equation 6.2 with p, the probability that a deletion occurs.

6.3.4 Results

In a first simulation, we compared the number of generations until only one parasite cell is left with the vulnerable part of the mitochondrial genome intact (the $t_{max}(k)$) under $k = 1$ (no split: a_1b_1) and $k = 2$ (one split: $a_1b_1a_2b_2$). The simulation was run with the following parameters: maximum population size $s_p = 10^{10}$ parasite cells, size of mitochondrial DNA $n = 2.6 \cdot 10^4$ base pairs, size of vulnerable DNA 34%, $p = 10^{-6}$, $\max_r = 3$, and $\min_r = 2$. The l varied from 0 to the full size n.

Going from $k = 1$ to a situation with $k = 2$ means going from an unmixed situation to the simplest mixed situation. Splitting and mixing increased the $t_{max}(k)$ with 48 generations ($t_{max}(1) = 108$ generations and $t_{max}(2) = 156$ generations). This comparison between $t_{max}(1)$ and $t_{max}(2)$ was also done with maximum population sizes of 10^8 and 10^{12} parasite cells, as is presented in Table 6.1. *In all three cases, the allowed generation time under periods of partially relaxed selective pressure has been extended by more than 40 %.* This shows that the concept published in [Spe06, Spe07] stands the test of being mathematically modeled.

	population size		
k	10^8	10^{10}	10^{12}
1	96	108	119
2	146	156	169

Table 6.1: Increase of $t_{max}(k)$ by splitting and mixing.

We were also interested in the behavior of the model with higher values of k. The actual k in the mitochondrial genome of *T. brucei* was estimated to be $k = 200$. To study the increase of t_{max} with higher values of k, we had to introduce a simplification to our model, as described in subsection 6.3.3. In our first approach, we simulated the process with a confidence level of 1. The state space has in that case a size $k + 1$, and the simulation matrix which corresponds to this state space has a size $(k + 1)^2$. The simulation was run with a maximum population size of 10^{10} parasite cells, and the further parameters the same as stated above. For values of k ranging from 1 to 200, the function $t_{max}(k)$ shows a nearly perfect line, as can be seen in Figure 6.1 when looking at the line for d = 1. This simulation gave a surprising result: we obtained a direct, quasi-linear correlation between the level of fragmentation (k) and the number of generations after which a complete population loses ecological competence. In other words, *the function $t_{max}(k)$ has an almost linear growth rate.*

We next further investigated the quasi-linearity. We were able to show rigorously that in the simple case of $d = 1$ the fragmentation advantage can never be more than linear, that is, we were able to show a linear upper bound on the

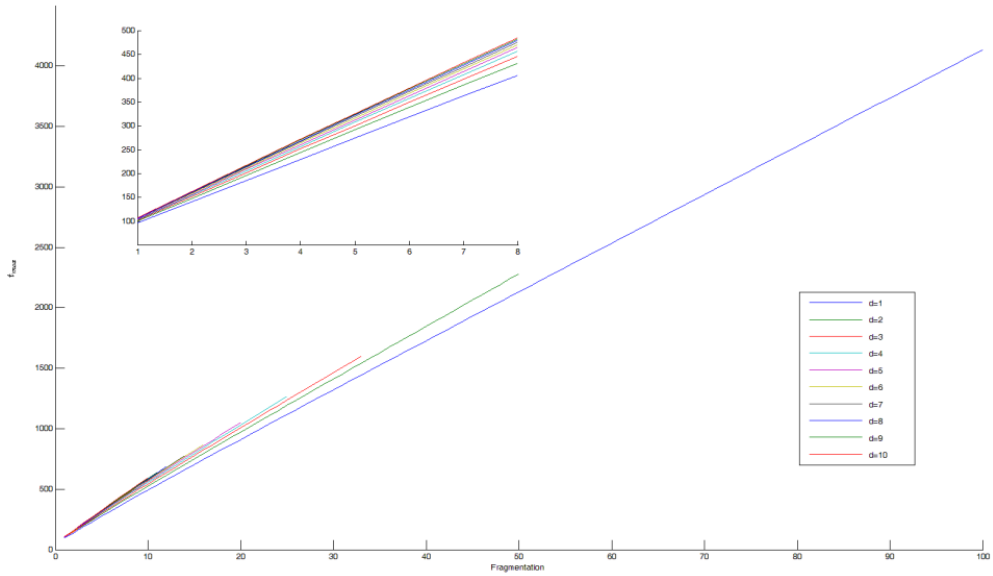


Figure 6.1: Quasi-linear correlation between amount of fragments (x-axis, parameter k of our model) and maximum number of generations upon which the population loses ecological competence (y-axis, $t_{max}(k)$ of our model). Inset: close-up of fragmentation levels 1 - 8; inset on the right: colour code for different confidence levels (d). The higher d, the better the approximation (see text).

function $t_{max}(k)$. This was done by studying the spectrum of a simplified 2×2 Markov chain.

Finally, we simulated the same process with increasing confidence values d. These results show that in each case, for these parameters, we get almost straight lines, each one with a slightly steeper slope, as can be seen in Figure 6.1. However, for successive values of d, the *increase* of the slope appears to be halving each time. This suggests that already for a small value of d, we have a reasonably good approximation of our original Markov chain.

6.4 Linkage selection and batch selection

We introduce the term "linkage selection" to describe the mechanism of natural selection as we propose it is manifested in the case of mitochondrial DNA of *T. brucei*. The linkage we are pointing out, is the physical linkage of fragments of different genes which protects "vulnerable" genes by means of the close presence of "vital" genes. The mixing forces the organism to make large deletions improbable. As a result of this organization, life stages during which certain parts of the genome are not expressed can be passed without loss of (in the long run!) essential DNA. We do not propose RNA editing originated under linkage selection. We see

linkage selection as the factor which made the massive expansion of RNA editing possible. We thus propose that *pan-editing* continues to exist because of linkage selection.

The genetic organization of the mitochondria of *Plasmodium falciparum*, the parasite causing malaria, came to our attention as a possible second case where linkage selection might have played a role. In these mitochondria, the rRNAs are split in pieces [FHL$^+$12]. The mitochondrial genome of *Plasmodium* can roughly be described (see e.g. [HKT13]) as:

$$(\text{Cox1-Cob-first pile of rRNA fragments-Cox3-second pile of rRNA fragments})_{\text{n}}$$

Because of this, the case of *Plasmodium* is, in terms of our model, only having $k = 2$. The protein-coding genes are not broken in small pieces at all. There is no mixing of pieces of rRNA (vital information) with pieces of protein (vulnerable information), and so the situation is not comparable with the one in *Trypanosoma*. Fragments of the large subunit rRNA are mixed with fragments of the small subunit rRNA, but because both ribosomal subunits are vital, this is not a case of linkage selection.

In *Plasmodium*, the pieces of rRNA are not ligated into large rRNAs after transcription: the rRNA pieces combine (together with ribosomal proteins) into a ribosomal subunit as a fragmented collection of pieces. The facts that the ribosomal RNAs can be split in smaller pieces, and that these pieces have the power to self-organize into a functional ribosome suggest that, originally, the ribosome might have been a multi-ribozyme assembly. If this is indeed the case, then we are seeing, in *Plasmodium* mt rDNA, a phenomenon in evolutionary biochemistry which we have already encountered in chapter 4. Reverting to a ribosome built up from more than a dozen pieces of rRNA would then be the fulfillment of a potential for simplicity lurking in the system, and present in the system as a trace of the past, because the system had evolved from such a ribosome (see section 4.7 for this potential for simplicity in connection to the superwobble).

In connection to pan-editing, Speijer [Spe06, Spe07] drew attention to the fact that the mitochondrion of *T. brucei* is not encoding any tRNA genes (as originally reported by [HH90]). In other organisms, the evenly distributed location of tRNA genes over mitochondrial genomes (see e.g. [GPDC$^+$89]) suggests that these genes might be playing the role of protecting actors during the occurrence of linkage selection, as we have suggested for guide RNAs in *T. brucei* (see above, [Spe06, Spe07]). If this is indeed the case, then linkage selection is a very widespread phenomenon of general biological relevance. Of course, other evolutionary reasons for the evenly distributed location of tRNA genes over a genome should be taken into consideration, e.g. the role they can play in genome organization and rearrangement [MW09]. Citing the work of Gordon et al. ([GBW09]),

McFarlane and Whitehall write: "The study of evolutionarily related yeast species has demonstrated that sites of gene gain and evolutionary breakpoints are associated with tRNA genes, inferring that tRNA genes were associated with the recombination events which resulted in stable, and presumably advantageous, evolutionary changes to the structure of evolving yeast genomes." While recognizing that the evolutionary environment of eukaryotic nuclear DNA is a totally different one from that of mitochondrial DNA, it is prudent to keep in mind that the sequence similarities among tRNA genes might be involved in genome reorganizations. Involvement in genome (re)organization might be a totally different reason why tRNA genes can be evenly distributed over genomes.

Dead-end "winners", leading their whole clonal batch to doom, could be a more general phenomenon beyond parasites. Pan-editing is also known from the free-living kinetoplastids (see [Spe07] and references therein). The central idea of linkage selection is that fragmentation and mixing of genes leads to protection against large-scale deletions. These large DNA deletions would give the entities harboring them an advantage over their competitors from the same clonal batch. Such a clonal batch is a characteristic of infection processes in parasite life cycles, but can also be encountered in ecological situations where a new location, rich in nutrients, is invaded by unicellular organisms. Examples of these situations are algal blooms, or the development of a population of predatory ciliates living on a large bacterial "feeding ground". In algal blooms, clonal reproduction often continues until nutrients become limiting; upon sensing nutrient limitation, a sexual process is triggered (see e.g. [vdEvdBL$^+$92]), ensuring many different genetic combinations are created to try starting a new bloom. The thick-walled zygote formed by freshwater algae is able to survive desiccated conditions (see e.g. [Agr12]); dispersal happens when wind is blowing around this "algal dust". When a new freshwater environment is entered, we see the same conditions of intense intraspecific clonal competition as we have presented for *T. brucei* upon entering the mammalian blood. The "run" towards very high cell numbers seen in parasite infection of the blood or in algal blooms in nutrient-rich waters, is also an aspect of growing populations of microbial predators, like e.g. *Noctiluca* blooms, or of mass development of ciliates. It is a possibility that the very strong evolutionary pressures of intense, clonal, intraspecific competition have led to further remarkable phenomena, apart from linkage of genes. One possible case could be the unusual micronucleus/macronucleus organization of the nuclear genome of ciliate protozoa (see e.g. [MLA$^+$13, VGML13]), as this organization possibly could enable fast clonal reproduction. I propose the term "batch selection" for this more general case. Linkage selection is thus a special case of batch selection.

A final consideration in connection to linkage selection, is that many copies of the mitochondrial genome of *T. brucei* are present in the mitochondrion. A large deletion taking away several minicircles could, in the view presented here,

take away such a significant part of the amount of a certain guide RNA, that the functioning of the parasite harmed by this deletion would be damaged enough to lead to the evolutionary phenomenon of linkage selection. This multi-copy aspect however, (of the *T. brucei* mitochondrial genome) is not a part of our simulation, and new work would be needed to be able to discuss this aspect in a more thorough manner.

Bibliography

[ADM02] C. L. Apel, D. W. Deamer, and M. N. Mautner. Self-assembled vesicles of monocarboxylic acids and alcohols: conditions for stability and for the encapsulation of biopolymers. *Biochim Biophys Acta*, 1559:1–9, 2002.

[Agr12] S. C. Agrawal. Factors controlling induction of reproduction in algae. *Folia Microbiologica*, 57(5):387–407, 2012.

[Aka01] H. Akashi. Gene expression and molecular evolution. *Curr Opin Gen Dev*, 11(6):660–666, 2001.

[AKIS03] I. Ahel, D. Korencic, M. Ibba, and D. Söll. Trans-editing of mischarged tRNAs. *Proc Natl Acad Sci USA*, 100(26):15422–15427, 2003.

[AMF04] S. An and K. Musier-Forsyth. Trans-editing of Cys-tRNA$^{\mathrm{Pro}}$ by *Haemophilus influenzae* YbaK protein. *J Biol Chem*, 279(41):42359–42362, 2004.

[AMM44] O. T. Avery, C. M. MacLeod, and M McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. *Journal of Experimental Medicine*, 79(2):137–158, 1944.

[Ard98] D. H. Ardell. On error minimization in a sequential origin of the standard genetic code. *J Mol Evol*, 47(1):1–13, 1998.

[AS69] C. Alff-Steinberger. The genetic code and error transmission. *Proc Natl Acad Sci USA*, 64(2):584–591, 1969.

[AS01] D. H. Ardell and G. Sella. On the evolution of redundancy in genetic codes. *J Mol Evol*, 53:269–281, 2001.

[AS02] D. H. Ardell and G. Sella. No accident: Genetic codes freeze in error-correcting patterns of the standard genetic code. *Phil Trans R Soc B*, 357(1427):1625, 2002.

[AVG07] P. F. Agris, F. A. P. Vendeix, and W. D. Graham. tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol*, 366(1):1–13, 2007.

[AYMO89] Y. Andachi, F. Yamao, A. Muto, and S. Osawa. Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *Mycoplasma capricolum*: Resemblance to mitochondria. *J Mol Biol*, 209(1):37–54, 1989.

[BAB+80] B. G. Barrell, S. Anderson, A. T. Bankier, M. H. de Bruijn, E. Chen, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young. Different pattern of codon recognition by mammalian mitochondrial tRNAs. *Proc Natl Acad Sci USA*, 77:3164–3166, 1980.

[BB16] R. A. Black and M. C. Blosser. A self-assembled aggregate composed of a fatty acid membrane and the building blocks of biological polymers provides a first step in the emergence of protocells. *Life (Basel)*, 6:article number 33, 2016.

[BBB+02] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58:899–907, 2002.

[BBC+80] S. G. Bonitz, R. Berlani, G. Coruzzi, M. Li, G. Macino, F. G. Nobrega, M. P. Nobrega, B. E. Thalenfeld, and A. Tzagoloff. Codon recognition rules in yeast mitochondria. *Proc Natl Acad Sci USA*, 77:3167–3170, 1980.

[BBS90] B. Blum, N. Bakalara, and L. Simpson. A model for RNA editing in kinetoplastid mitochondria: 'Guide' RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell*, 60(2):189–198, 1990.

[BFG04] C. Brochier, P. Forterre, and S. Gribaldo. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. *Genome Biol*, 5(3):article number R17, 2004.

[BFS04] D. J. Brooks, J. R. Fresco, and M. Singh. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics*, 20(14):2251–2257, 2004.

[BG01] H. Beier and M. Grimm. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Res*, 29(23):4767–4782, 2001.

[BGKB02] O. Botta, D. P. Glavin, G. Kminek, and J. L. Bada. Relative amino acid concentrations as a signature for parent body processes of carbonaceous chondrites. *Orig Life Evol Biosph*, 32(2):143–163, 2002.

[BGMLS09] T. Butler, N. Goldenfeld, D. Mathew, and Z. Luthey-Schulten. Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement. *Phys Rev E*, 79:article number 060901(R), 2009.

[BHW15] J. C. Bowman, N. V. Hud, and L. D. Williams. The ribosome challenge to the RNA world. *J Mol Evol*, 80:143–161, 2015.

[BJM61] S. Brenner, F. Jacob, and M. Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581, 1961.

[BKC12] S. A. Benner, H. J. Kim, and M. A. Carrigan. Asphalt, water, and the prebiotic synthesis of ribose, ribonucleosides, and RNA. *Acc Chem Res*, 45(12):2025–2034, 2012.

[Bra08] A. Brack. From Interstellar Amino Acids to Prebiotic Catalytic Peptides: A Review. 2008. In: P. Herdewijn and M. V. Kisakurek (eds.) *Origin of Life: Chemical Approach*, pp. 215-229, Zurich: Verlag Helvetica Chimica Acta.

[Bre57] S. Brenner. Impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proc Natl Acad Sci USA*, 43(8):687–694, 1957.

[Bre01a] S. Brenner. Deciphering the genetic code. 2001. In: E. C. Friedberg and E. Lawrence (eds.) *My Life in Science: Sydney Brenner; As told to Lewis Wolpert*, pp. 88-106, London: BioMed Central Limited.

[Bre01b] S. Brenner. Discovering messenger RNA. 2001. In: E. C. Friedberg and E. Lawrence (eds.) *My Life in Science: Sydney Brenner;*

As told to Lewis Wolpert, pp. 62-87, London: BioMed Central Limited.

[BS11] I. Budin and J. W. Szostak. Physical effects underlying the transition from primitive to modern cell membranes. *Proc Natl Acad Sci USA*, 108(13):5249–5254, 2011.

[BT41] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci USA*, 27(11):499–506, 1941.

[BTS07a] J. M. Berg, J. L. Tymoszko, and L. Stryer. *Biochemistry*. 2007. Sixth edition, p. 329, W.H. Freeman and Company.

[BTS07b] J. M. Berg, J. L. Tymoszko, and L. Stryer. *Biochemistry*. 2007. Sixth edition, p. 875, W.H. Freeman and Company.

[BTS07c] J. M. Berg, J. L. Tymoszko, and L. Stryer. *Biochemistry*. 2007. Sixth edition, p. 664, W.H. Freeman and Company.

[BvdBB86] R. Benne, J. van den Burg, and J. P. J. Brakenhoff. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, 46(6):819–826, 1986.

[BvdGK+11] H. Buhrman, P. T. S. van der Gulik, S. M. Kelk, W. M. Koolen, and L. Stougie. Some mathematical refinements concerning error minimization in the genetic code. *IEEE/ACM Trans Comput Biol Bioinf*, 8(5):1358–1372, 2011.

[BvdGK+13] H. Buhrman, P. T. S. van der Gulik, G. W. Klau, C. Schaffner, D. Speijer, and L. Stougie. A realistic model under which the genetic code is optimal. *J Mol Evol*, 77:170–184, 2013.

[CBBWT61] F. H. C. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192(4809):1227–1232, 1961.

[CCL+08] H. J. Cleaves, J. H. Chalmers, A. Lazcano, S. L. Miller, and J. L. Bada. A reassessment of prebiotic organic synthesis in neutral planetary atmospheres. *Orig Life Evol Biosph*, 38(2):105–115, 2008.

[Cel98] E. Cela. *The Quadratic Assignment Problem: Theory and Algorithms*. Kluwer Academic Publishers, 1998.

[CG11] O. Carny and E. Gazit. Creating prebiotic sanctuary: Self-Assembling Supramolecular Peptide Structures Bind and Stabilize RNA. *Orig Life Evol Biosph*, 41(2):121–132, 2011.

[CGL+10] E. Cocquyt, G. H. Gile, F. Leliaert, H. Verbruggen, P. J. Keeling, and O. De Clerck. Complex phylogenetic distribution of a non-canonical genetic code in green algae. *BMC Evol Biol*, 10(1):article number 327, 2010.

[Che06] I. A. Chen. The emergence of cells during the origin of life. *Science*, 314(5805):1558–1559, 2006.

[CI10] H. J. Cleaves II. The origin of the biologically coded amino acids. *J Theor Biol*, 263(4):490–498, 2010.

[CJ15] C. W. Carter Jr. What RNA world? Why a peptide/RNA partnership merits renewed experimental attention. *Life (Basel)*, 5:294–320, 2015.

[CKIV+04] S. Chabelskaya, D. Kiktev, S. Inge-Vechtomov, M. Philippe, and G. Zhouravleva. Nonsense mutations in the essential gene SUP35 of *Saccharomyces cerevisiae* are non-lethal. *Mol Genet Genomics*, 272(3):297–307, 2004.

[CNM06] H. J. Cleaves, K. E. Nelson, and S. L. Miller. The prebiotic synthesis of pyrimidines in frozen solution. *Naturwissenschaften*, 93(5):228–231, 2006.

[COC+13] J. H. Campbell, P. O'Donoghue, A. G. Campbell, P. Schwientek, A. Sczyrba, T. Woyke, D. Söll, and M. Podar. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci USA*, 110(14):5540–5545, 2013.

[Cri66] F. H. C. Crick. Codon-anticodon pairing: The wobble hypothesis. *J Mol Biol*, 19(2):548–555, 1966.

[Cri68] F. H. C. Crick. The origin of the genetic code. *J Mol Biol*, 38(3):367–379, 1968.

[Cri88] F. H. C. Crick. *What Mad Pursuit: A personal view of scientific discovery*. Weidenfeld and Nicolson, 1988.

[CRS04] I. A. Chen, R. W. Roberts, and J. W. Szostak. The emergence of competition between model protocells. *Science*, 305(5689):1474–1476, 2004.

[CS04] I. A. Chen and J. W. Szostak. Membrane growth can generate a transmembrane pH gradient in fatty acid vesicles. *Proc Natl Acad Sci USA*, 101(21):7965–7970, 2004.

[CYK05] J. G. Caporaso, M. Yarus, and R. Knight. Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J Mol Evol*, 61(5):597–607, 2005.

[CYS08] Y. E. Chong, X. L. Yang, and P. Schimmel. Natural homolog of tRNA synthetase editing domain rescues conditional lethality caused by mistranslation. *J Biol Chem*, 283(44):30073–30078, 2008.

[DB06] V. A. Doronina and J. D. Brown. When nonsense makes sense and vice versa: Noncanonical decoding events at stop codons in eukaryotes. *Mol Biol*, 40(4):654–663, 2006.

[dCLMG12] V. de Crécy-Lagard, C. Marck, and H. Grosjean. Decoding in Candidatus *Riesia pediculicola*, close to a minimal tRNA modification set? *Trends Cell Molec Biol*, 7:11–34, 2012.

[Dea85] D. W. Deamer. Boundary structures are formed by organic components of the Murchison carbonaceous chondrite. *Nature*, 317(6040):792–794, 1985.

[DG89a] M. Di Giulio. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J Mol Evol*, 29(4):288–293, 1989.

[DG89b] M. Di Giulio. Some aspects of the organization and evolution of the genetic code. *J Mol Evol*, 29(3):191–201, 1989.

[DG06] M. Di Giulio. The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA). *J Theor Biol*, 240(3):343–352, 2006.

[DG08] M. Di Giulio. An extension of the coevolution theory of the origin of the genetic code. *Biol Direct*, 3:article number 37, 2008.

[DHLDL+95] G. Dieci, S. Hermann-Le Denmat, E. Lukhtanov, P. Thuriaux, M. Werner, and A. Sentenac. A universally conserved region of the largest subunit participates in the active site of RNA polymerase III. *EMBO J*, 14(15):3766–3776, 1995.

[Eri11] A. Erives. A model of proto-anti-codon RNA enzymes requiring L-amino acid homochirality. *J Mol Evol*, 73:10–22, 2011.

[ES77] M. Eigen and P. Schuster. The hypercycle. A principle of natural self organization. Part A: Emergence of the hypercycle. *Naturwissenschaften*, 64(11):541–565, 1977.

[ES78] M. Eigen and P. Schuster. The hypercycle. A principle of natural self organization. Part C: The realistic hypercycle. *Naturwissenschaften*, 65(7):341–369, 1978.

[ES90] A. D. Ellington and J. W. Szostak. *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, 346:818–822, 1990.

[ESC+06] B. Erni, C. Siebold, S. Christen, A. Srinivas, A. Oberholzer, and U. Baumann. Small substrate, big surprise: Fold, function and phylogeny of dihydroxyacetone kinases. *Cell Mol Life Sci*, 63(7-8):890–900, 2006.

[Fan12] R. Fani. The Origin and Evolution of Metabolic Pathways: *Why* and *How* did Primordial Cells Construct Metabolic Routes? *Evo Edu Outreach*, 5:367–381, 2012.

[FBE+07] M. Fondi, M. Brilli, G. Emiliani, D. Paffetti, and R. Fani. The primordial metabolism: An ancestral interconnection between leucine, arginine, and lysine biosynthesis. *BMC Evol Biol*, 7(suppl. 2):article number S3, 2007.

[FdSW91] J. J. R. Fraústo da Silva and R. J. P. Williams. *The Biological Chemistry of the Elements*. Clarendon Press, 1991.

[Fer06] J. P. Ferris. Montmorillonite-catalysed formation of rna oligomers: The possible role of catalysis in the origins of life. *Phil Trans R Soc B*, 361(1474):1777–1786, 2006.

[FG10] G. P. Fournier and J. P. Gogarten. Rooting the ribosomal tree of life. *Mol Biol Evol*, 27(8):1792–1801, 2010.

[FH98a] S. J. Freeland and L. D. Hurst. The genetic code is one in a million. *J Mol Evol*, 47(3):238–248, 1998.

[FH98b] S. J. Freeland and L. D. Hurst. Load minimization of the genetic code: History does not explain the pattern. *Proc R Soc B Biol Sci*, 265(1410):2111–2119, 1998.

[FHL+12] J. E. Feagin, M. I. Harrell, J. C. Lee, K. J. Coe, B. H. Sands, J. J. Cannone, G. Tami, M. N. Schnare, and R. R. Gutell. The fragmented mitochondrial ribosomal RNAs of Plasmodium falciparum. *PLoS ONE*, 7(6):article number e38320, 2012.

[Fis11] M. Fishkis. Emergence of self-reproduction in cooperative chemical evolution of prebiological molecules. *Orig Life Evol Biosph*, 41(3):261–275, 2011.

[FKL99] S. J. Freeland, R. D. Knight, and L. F. Landweber. Do proteins predate DNA? *Science*, 286(5440):690–692, 1999.

[FKLH00] S. J. Freeland, R. D. Knight, L. F. Landweber, and L. D. Hurst. Early fixation of an optimal genetic code. *Mol Biol Evol*, 17(4):511–518, 2000.

[Fra11] B. R. Francis. An alternative to the RNA world hypothesis. *Trends Evol Biol*, 3(1):2–11, 2011.

[Fra13] B. R. Francis. Evolution of the genetic code by incorporation of amino acids that improved or changed protein function. *J Mol Evol*, 77:134–158, 2013.

[FSK+09] K. Fujishima, J. Sugahara, K. Kikuta, R. Hirano, A. Sato, M. Tomita, and A. Kanai. Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. *Proc Natl Acad Sci USA*, 106(8):2683–2687, 2009.

[FTM+08] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The pfam protein families database. *Nucleic Acids Res*, 36(suppl. 1):D281–D288, 2008.

[FU87] W. M. Fitch and K. Upper. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb Symp Quant Biol*, 52:759–767, 1987.

[FWK03] S. J. Freeland, T. Wu, and N. Keulmann. The case for an error minimizing standard genetic code. *Orig Life Evol Biosph*, 33(4-5):457–477, 2003.

[FYEBN05] G. Fleminger, T. Yaron, M. Eisenstein, and A. Bar-Nun. The structure and synthetic capabilities of a catalytic peptide formed by substrate-directed mechanism - implications to prebiotic catalysis. *Orig Life Evol Biosph*, 35(4):369–382, 2005.

[Gam54] G. Gamow. Possible relation between deoxyribonucleic acid and protein structures. *Nature*, 173(4398):318, 1954.

[GBA06] S. Gribaldo and C. Brochier-Armanet. The origin and evolution of Archaea: A state of the art. *Phil Trans R Soc B*, 361(1470):1007–1022, 2006.

[GBB02] M. E. Glasner, N. H. Bergman, and D. P. Bartel. Metal ion requirements for structure and catalysis of an RNA ligase ribozyme. *Biochemistry*, 41(25):8103–8112, 2002.

[GBW09] J.L. Gordon, K.P. Byrne, and K.H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genetics*, 5(5):article number e1000485, 2009.

[GdCLM10] H. Grosjean, V. de Crécy-Lagard, and C. Marck. Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. *FEBS Lett*, 584(2):252–264, 2010.

[Gil86] W. Gilbert. Origin of life: The RNA world. *Nature*, 319(6055):618, 1986.

[GMCR01] D. Gilis, S. Massar, N. J. Cerf, and M. Rooman. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol*, 2:article number R49, 2001.

[Gol93] N. Goldman. Further results on error minimization in the genetic code. *J Mol Evol*, 37(6):662–664, 1993.

[GPDC+89] G. Gadaleta, G. Pepe, G. De Candia, C. Quagliariello, E. Sbisa, and C. Saccone. The complete nucleotide sequence of the rattus norvegicus mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. *J Mol Evol*, 28(6):497–516, 1989.

[GTGM+83] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[GZGCK11] M. A. Gaston, L. Zhang, K. B. Green-Church, and J. A. Krzycki. The complete biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. *Nature*, 471(7340):647–650, 2011.

[HCO+13] C. Hsiao, I.-C. Chou, C. D. Okafor, J. C. Bowman, E. B. O'Neill, S. S. Athavale, A. S. Petrov, N. V. Hud, R. M. Wartell, S. C. Harvey, and L. D. Williams. RNA with iron(II) as a cofactor catalyses electron transfer. *Nat Chem*, 5:525–528, 2013.

[HFR06] L. D. Hurst, E. J. Feil, and E. P. C. Rocha. Protein evolution: Causes of trends in amino-acid gain and loss. *Nature*, 442(7105):E11–E12, 2006.

[HH90] K. Hancock and S. L. Hajduk. The mitochondrial tRNAs of *Trypanosoma brucei* are nuclear encoded. *J Biol Chem*, 265(31):19208–19215, 1990.

[HH91] D. Haig and L. D. Hurst. A quantitative measure of error minimization in the genetic code. *J Mol Evol*, 33(5):412–417, 1991.

[HH99] D. Haig and L. D. Hurst. Erratum: A quantitative measure of error minimization in the genetic code. *J Mol Evol*, 49(5):708, 1999.

[Hig09] P. G. Higgs. A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct*, 4:article number 16, 2009.

[HKSB86] N. Hanyu, Y. Kuchino, N. Susumu, and H. Beier. Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* $\text{tRNAs}^{\text{Gln}}$. *EMBO J*, 5:1307–1311, 1986.

[HKT13] K. Hikosaka, K. Kita, and K. Tanabe. Diversity of mitochondrial genome structure in the phylum Apicomplexa. *Molecular and Biochemical Parasitology*, 188(1):26–33, 2013.

[HP09] P. G. Higgs and R. E. Pudritz. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*, 9(5):483–490, 2009.

[HPO+99] M. Hochuli, H. Patzelt, D. Oesterhelt, K. Wuthrich, and T. Szyperski. Amino acid biosynthesis in the halophilic archaeon Haloarcula hispanica. *J Bacteriol*, 181(10):3226–3237, 1999.

[HSAD+80] J. E. Heckman, J. Sarnoff, B. Alzner-DeWeerd, S. Yin, and U. L. RajBhandary. Novel features in the genetic code and codon reading patterns in *Neurospora crassa* mitochondria based on sequences of six mitochondrial tRNAs. *Proc Natl Acad Sci USA*, 77:3159–3163, 1980.

[HSS+58] M. B. Hoagland, M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamecnik. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem*, 231:241–257, 1958.

[HT13] M. J. Harms and J. W. Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*, 14:559–571, 2013.

[HYSM04] S. Honda, K. Yamasaki, Y. Sawada, and H. Morii. 10-residue folded peptide designed by segment statistics. *Structure*, 12(8):1507–1518, 2004.

[IA07] S. Itzkovitz and U. Alon. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, 17(4):405–412, 2007.

[IB98] S. D. Irvin and J. K. Bhattacharjee. A unique fungal lysine biosynthesis enzyme shares a common ancestor with tricarboxylic acid cycle and leucine biosynthetic enzymes found in diverse organisms. *J Mol Evol*, 46(4):401–408, 1998.

[IKA03] L. M. Iyer, E. V. Koonin, and L. Aravind. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struc Biol*, 3:article number 1, 2003.

[IKB+95] Y. Inagaki, A. Kojima, Y. Bessho, H. Hori, T. Ohama, and S. Osawa. Translation of synonymous codons in family boxes by *Mycoplasma capricolum* tRNAs with unmodified uridine or adenosine at the first anticodon position. *J Mol Biol*, 251(4):486–492, 1995.

[Ike02] K. Ikehara. Origins of gene, genetic code, protein and life: Comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. *J Biosci*, 27(2):165–186, 2002.

[IKN+10] Y. Ikeuchi, S. Kimura, T. Numata, D. Nakamura, T. Yokogawa, T. Ogata, T. Wada, T. Suzuki, and T. Suzuki. Agmatidine-conjugated cytidine in a tRNA anticodon is essential for AUA decoding in archaea. *Nature Chem Biol*, 6:277–282, 2010.

[IOAH02] K. Ikehara, Y. Omori, R. Arai, and A. Hirose. A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. *J Mol Evol*, 54(4):530–538, 2002.

[IY02] M. Illangasekare and M. Yarus. Phenylalanine-binding RNAs and genetic code evolution. *J Mol Evol*, 54(3):298–311, 2002.

[JCD+08] A. P. Johnson, H. J. Cleaves, J. P. Dworkin, D. P. Glavin, A. Lazcano, and J. L. Bada. The Miller volcanic spark discharge experiment. *Science*, 322(5900):404, 2008.

[JCGC07] P. A. Jones, M. R. Cunningham, P. D. Godfrey, and D. M. Cragg. A search for biomolecules in Sagittarius B2 (LMH) with the Australia telescope compact array. *Monthly Notices of the Royal Astronomical Society*, 374(2):579–589, 2007.

[JEH+08] M. J .O. Johansson, A. Esberg, B. Huang, G. R. Bjork, and A. S. Bystrom. Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Mol Cell Biol*, 28(10):3301–3312, 2008.

[Jen76] R. A. Jensen. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, 30:409–425, 1976.

[JKA+05] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov, and S. Sunyaev. A universal trend of amino acid gain and loss in protein evolution. *Nature*, 433(7026):633–638, 2005.

[Juk66] T. H. Jukes. *Molecules and Evolution*. Columbia Press, 1966.

[JUL+01] W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner, and D. P. Bartel. RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science*, 292(5520):1319–1325, 2001.

[JVvLG09] J. E. Jacob, B. Vanholme, T. van Leeuwen, and G. Gheysen. A unique genetic code change in the mitochondrial genome of the parasitic nematode Radopholus similis. *BMC Research Notes*, 2:article number 192, 2009.

[JW10] D. B. F. Johnson and L. Wang. Imprints of the genetic code in the ribosome. *Proc Natl Acad Sci USA*, 107(18):8298–8303, 2010.

[JWKY10] T. Janas, J. J. Widmann, R. Knight, and M. Yarus. Simple, recurring RNA binding sites for L-arginine. *RNA*, 16(4):805–816, 2010.

[KBNF97] E. Kochavi, A. Bar-Nun, and G. Fleminger. Substrate-directed formation of small biocatalysts under prebiotic conditions. *J Mol Evol*, 45(4):342–351, 1997.

[KCH+03] Y. J. Kuan, S. B. Charnley, H. C. Huang, W. L. Tseng, and Z. Kisiel. Interstellar glycine. *Astrophysical Journal Letters*, 593:848–867, 2003.

[KD82] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, 1982.

[KD96] P. J. Keeling and W. F. Doolittle. A non-canonical genetic code in an early diverging eukaryotic lineage. *EMBO J*, 15(9):2285–2290, 1996.

[KF07] E. B. Kramer and P. J. Farabaugh. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA*, 13(1):87–96, 2007.

[KFL99] R. D. Knight, S. J. Freeland, and L. F. Landweber. Selection, history and chemistry: The three faces of the genetic code. *Trends Biochem Sci*, 24(6):241–247, 1999.

[KFL01] R. D. Knight, S. J. Freeland, and L. F. Landweber. Rewiring the keyboard: Evolvability of the genetic code. *Nat Rev Genet*, 2(1):49–58, 2001.

[KGZ+82] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–157, 1982.

[KKP+10] M. Kumauchi, S. Kaledhonkar, A. F. Philip, J. Wycoff, M. Hara, Y. Li, A. Xie, and W. D. Hoff. A conserved helical capping hydrogen bond in PAS domains controls signaling kinetics in the superfamily prototype Photoactive Yellow Protein. *J. Am. Chem. Soc.*, 132:15820–15830, 2010.

[KL03] P. J. Keeling and B. S. Leander. Characterisation of a non-canonical genetic code in the oxymonad Streblomastix strix. *J Mol Biol*, 326(5):1337–1349, 2003.

[KOAO93] A. Kano, T. Ohama, R. Abe, and S. Osawa. Unassigned or nonsense codons in Micrococcus luteus. *J Mol Biol*, 230(1):51–56, 1993.

[Kur92] C. G. Kurland. Evolution of mitochondrial genomes and the genetic code. *BioEssays*, 14(10):709–714, 1992.

[Kve74] K. A. Kvenvolden. Amino and fatty acids in carbonaceous meteorites. 1974. In: K. Dose, S. W. Fox, G. A. Deborin, and T. E. Pavlovskaya (eds.) *The origin of life and evolutionary biochemistry*, pp.301-309, New York: Plenum Press.

[Lag78] U. Lagerkvist. 'Two out of three': An alternative method for codon reading. *Proc Natl Acad Sci USA*, 75(4):1759–1762, 1978.

[Lag81] U. Lagerkvist. Unorthodox codon reading and the evolution of the genetic code. *Cell*, 23(2):305–306, 1981.

[LAM10] N. Lane, J. F. Allen, and W. Martin. How did LUCA make a living? Chemiosmosis in the origin of life. *BioEssays*, 32(4):271–280, 2010.

[LBK13] X. Liu, D. A. Bushnell, and R. D. Kornberg. RNA polymerase II transcription: Structure and mechanism. *Biochim Biophys Acta*, 1829:2–8, 2013.

[LBZ+07] D. G. Longstaff, S. K. Blight, L. Zhang, K. B. Green-Church, and J. A. Krzycki. In vivo contextual requirements for UAG translation as pyrrolysine. *Mol Microbiol*, 63(1):229–241, 2007.

[LCMY03] C. Lozupone, S. Changayil, I. Majerfeld, and M. Yarus. Selection of the simplest RNA that binds isoleucine. *RNA*, 9(11):1315–1322, 2003.

[LCT05] R. A. Laskowski, V. V. Chistyakov, and J. M. Thornton. PDBsum more: New summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res*, 33:D266–D268, 2005.

[LdABN+07] E. M. Loiola, N. M. M. de Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido. A survey for the quadratic assignment problem. *European J Operational Research*, 176(2):657–690, 2007.

[Lew51] E. B. Lewis. Pseudoallelism and gene evolution. *Cold Spring Harb Symp Quant Biol*, 16:159–174, 1951.

[Lew08a] B. Lewin. *Genes IX*. Jones and Bartlett Publishers, 2008. pp. 208-209.

[Lew08b] B. Lewin. *Genes IX*. Jones and Bartlett Publishers, 2008. p. 198.

[LFCJ13] L. Li, C. Francklyn, and C. W. Carter Jr. Aminoacylating urzymes challenge the RNA world hypothesis. *J Biol Chem*, 288(37):26856–26863, 2013.

[LFFR10] F. Li, D. Fitz, D. G. Fraser, and B. M. Rode. Catalytic effects of histidine enantiomers and glycine on the formation of dileucine and dimethionine in the salt-induced peptide formation reaction. *Amino Acids*, 38(1):287–294, 2010.

[LGM+96] D. H. Lee, J. R. Granja, J. A. Martinez, K. Severin, and M. R. Ghadiri. A self-replicating peptide. *Nature*, 382(6591):525–528, 1996.

[LJ88] N. Lehman and T. H. Jukes. Genetic code development by stop codon takeover. *J Theor Biol*, 135(2):203–214, 1988.

[LKL01] C. A. Lozupone, R. D. Knight, and L. F. Landweber. The molecular basis of nuclear genetic code change in ciliates. *Curr Biol*, 11(2):65–74, 2001.

[LL08] J. Lehmann and A. Libchaber. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. *RNA*, 14(7):1264–1269, 2008.

[LM12] N. Lane and W. F. Martin. The origin of membrane bioenergetics. *Cell*, 151(7):1406–1416, 2012.

[LMAK09] N. T. Lao, A. P. Maloney, J. F. Atkins, and T. A. Kavanagh. Versatile dual reporter gene systems for investigating stop codon readthrough in plants. *PLoS ONE*, 4(10):article number e7354, 2009.

[LS10] C. C. Liu and P. G. Schultz. Adding new chemistries to the genetic code. *Annu Rev Biochem*, 79:413–444, 2010.

[MAGH02] S. Madison-Antenucci, J. Grams, and S.L. Hajduk. Editing machines: The complexities of trypanosome RNA editing. *Cell*, 108(4):435–438, 2002.

[MAM+96] S. J. Mojzsis, G. Arrhenius, K. D. McKeegan, T. M. Harrison, A. P. Nutman, and C. R. L. Friend. Evidence for life on Earth before 3,800 million years ago. *Nature*, 384(6604):55–59, 1996.

[Mas06] S. E. Massey. A sequential "2-1-3" model of genetic code evolution that explains codon constraints. *J Mol Evol*, 62(6):809–810, 2006.

[Mas08] S. E. Massey. A neutral origin for error minimization in the genetic code. *J Mol Evol*, 67(5):510–516, 2008.

[Mas10] S. E. Massey. Searching of code space for an error-minimized genetic code via codon capture leads to failure, or requires at least 20 improving codon reassignments via the ambiguous intermediate mechanism. *J Mol Evol*, 70(1):106–115, 2010.

[MBD+12] A. Y. Mulkidjanian, A. Yu. Bychkov, D. V. Dibrova, M. Y. Galperin, and E. V. Koonin. Origin of first cells at terrestrial, anoxic geothermal fields. *Proc Natl Acad Sci USA*, 109(14):E821–E830, 2012.

[MCM+10] I. Majerfeld, J. Chocholousova, V. Malaiya, J. Widmann, D. McDonald, J. Reeder, M. Iyer, M. Illangasekare, M. Yarus, and R. Knight. Nucleotides that are essential but not conserved; a sufficient L-tryptophan site in RNA. *RNA*, 16(10):1915–1924, 2010.

[MG02] C. Marck and H. Grosjean. tRNomics: Analysis of tRNA genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, 8(10):1189–1232, 2002.

[Mil55] S. L. Miller. Production of some organic compounds under possible primitive Earth conditions. *J Am Chem Soc*, 77(9):2351–2361, 1955.

[Mil74] S. L. Miller. The atmosphere of the primitive Earth and the prebiotic synthesis of amino acids. *Orig Life*, 5(1-2):139–151, 1974.

[Mil87] S. L. Miller. Which organic compounds could have occurred on the prebiotic earth? *Cold Spring Harb Symp Quant Biol*, 52:17–27, 1987.

[MKS+10] D. Mandal, C. Kohrer, D. Su, S. P. Russell, K. Krivos, C. M. Castleberry, P. Blum, P.A. Limbach, D. Söll, and U.L. RajBhandary. Agmatidine, a modified cytidine in the anticodon of archaeal tRNA(Ile), base pairs with adenosine but not with guanosine. *Proc Natl Acad Sci USA*, 107(7):2872–2877, 2010.

[MLA+13] D. W. Morgens, K. M. Lindbergh, M. Adachi, A. Radunskaya, and A. R. O. Cavalcanti. A model for the evolution of extremely fragmented macronuclei in ciliates. *PLoS ONE*, 8(5):article number e64997, 2013.

[MLS08] D. C. Mathew and Z. Luthey-Schulten. On the physical basis of the amino acid polar requirement. *J Mol Evol*, 66(5):519–528, 2008.

[MMH06] C. E. Manning, S. J. Mojzsis, and T. M. Harrison. Geology, age and origin of supracrustal rocks at Akilia, West Greenland. *American Journal of Science*, 306(5):303–366, 2006.

[MMZ10] O. A. Murina, S. E. Moskalenko, and G. A. Zhouravleva. Overexpression of genes encoding $tRNA^{Tyr}$ and $tRNA^{Gln}$ increases the viability of *Saccharomyces cerevisiae* strains with nonsense mutations in the SUP45 gene. *Mol Biol*, 44(2):268–276, 2010.

[MNN+88] T. Muramatsu, K. Nishikawa, F. Nemoto, Y. Kuchino, S. Nishimura, T. Miyazawa, and S. Yokoyama. Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. *Nature*, 336(6195):179–181, 1988.

[MPY05] I. Majerfeld, D. Puthenvedu, and M. Yarus. RNA affinity for molecular L-histidine; genetic code origins. *J Mol Evol*, 61:226–235, 2005.

[MRC13] F. W. Martinez-Rucobo and P. Cramer. Structural basis of transcription elongation. *Biochim Biophys Acta*, 1829:9–19, 2013.

[MREJR+15] L. Martinez-Rodriguez, O. Erdogan, M. Jimenez-Rodriguez, Gonzalez-Rivera K., T. Williams, L. Li, V. Weinreb, M. Collier, S. N. Chandrasekaran, X. Ambroggio, B. Kuhlman, and C. W. Carter Jr. Functional Class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. *J Biol Chem*, 290(32):19710–19725, 2015.

[MS08] S. S. Mansy and J. W. Szostak. Thermostability of model protocell membranes. *Proc Natl Acad Sci USA*, 105(36):13351–13355, 2008.

[MSK+08] S. S. Mansy, J. P. Schrum, M. Krishnamurthy, S. Tobé, D. A. Treco, and J. W. Szostak. Template-directed synthesis of a genetic polymer in a model protocell. *Nature*, 454(7200):122–125, 2008.

[MSL+03] L. A. Morgan, W. C. Shanks, D. A. Lovalvo, S. Y. Johnson, W. J. Stephenson, K. L. Pierce, S. S. Harlan, C. A. Finn, G. Lee, M. Webring, B. Schulze, J. Dühn, R. Sweeney, and L. Balistrieri. Exploration and discovery in Yellowstone Lake: Results from

high-resolution sonar imaging, seismic reflection profiling, and submersible studies. *J Volcanol Geotherm Res*, 122(3-4):221–242, 2003.

[MSP+91] F. Meyer, H. J. Schmidt, E. Plumper, A. Hasilik, G. Mersmann, H. E. Meyer, A. Engstrom, and K. Heckmann. UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus. *Proc Natl Acad Sci USA*, 88(9):3758–3761, 1991.

[MW09] R. J. McFarlane and S. K. Whitehall. tRNA genes in eukaryotic genome organization and reorganization. *Cell Cycle*, 8(19):3102–3106, 2009.

[MXW10] W. K. Mat, H. Xue, and J. T. F. Wong. Genetic code mutations: The breaking of a three billion year invariance. *PLoS ONE*, 5(8):article number e12206, 2010.

[MY94] I. Majerfeld and M. Yarus. An RNA pocket for an aliphatic hydrophobe. *Nat Struc Biol*, 1(5):287–292, 1994.

[MY05] I. Majerfeld and M. Yarus. A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res*, 33(17):5482–5493, 2005.

[NIF+06] T. Numata, Y. Ikeuchi, S. Fukai, T. Suzuki, and O. Nureki. Snapshots of tRNA sulphuration via an adenylated intermediate. *Nature*, 442(7101):419–424, 2006.

[NJL+63] M. W. Nirenberg, O. W. Jones, P. Leder, B. F. C. Clark, W. S. Sly, and S. Pestka. On the coding of genetic information. *Cold Spring Harb Symp Quant Biol*, 28:549–557, 1963.

[Nol04] H. F. Noller. The driving force for molecular evolution of translation. *RNA*, 10(12):1833–1837, 2004.

[NPEPRDP12] E. M. Novoa, M. Pavon-Eternod, T. Pan, and L. Ribas De Pouplana. A role for tRNA modifications in genome structure and codon usage. *Cell*, 149(1):202–213, 2012.

[NTV+06] N. Nevskaya, S. Tishchenko, S. Volchkov, V. Kljashtorny, E. Nikonova, O. Nikonov, A. Nikulin, C. Kohrer, W. Piendl, R. Zimmermann, P. Stockley, M. Garber, and S. Nikonov. New insights into the interaction of ribosomal protein L1 with RNA. *J Mol Biol*, 355(4):747–759, 2006.

[Nud09] E. Nudler. RNA polymerase active center: the molecular engine of transcription. *Annu Rev Biochem*, 78:335–361, 2009.

[NWK07] A. S. Novozhilov, Y. I. Wolf, and E. V. Koonin. Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct*, 2:article number 24, 2007.

[NWM+08] A. A. Nemchin, M. J. Whitehouse, M. Menneken, T. Geisler, R. T. Pidgeon, and S. A. Wilde. A light carbon reservoir recorded in zircon-hosted diamond from the Jack Hills. *Nature*, 454(7200):92–95, 2008.

[OAMO91] T. Oba, Y. Andachi, A. Muto, and S. Osawa. CGG: An unassigned or nonsense codon in Mycoplasma capricolum. *Proc Natl Acad Sci USA*, 88(3):921–925, 1991.

[Ohn70] S. Ohno. *Evolution by Gene Duplication*. Springer, 1970.

[OJWM92] S. Osawa, T. H. Jukes, K. Watanabe, and A. Muto. Recent evidence for evolution of the genetic code. *Microbiol Rev*, 56(1):229–264, 1992.

[Olb94] R. C. Olby. *The path to the double helix: the discovery of DNA*. Dover Publications Inc., 1994.

[Org68] L. E. Orgel. Evolution of the genetic apparatus. *J Mol Biol*, 38(3):381–393, 1968.

[OWD+09] F. Orange, F. Westall, J. R. Disnar, D. Prieur, N. Bienvenu, M. Le Romancer, and C. Defarge. Experimental silicification of the extremophilic archaea *Pyrococcus abyssi* and *Methanocaldococcus jannaschii*: Applications in the search for evidence of life in early earth and extraterrestrial rocks. *Geobiology*, 7(4):403–418, 2009.

[PCD+11] E. T. Parker, H. J. Cleaves, J. P. Dworkin, D. P. Glavin, M. Callahan, A. Aubrey, A. Lazcano, and J. L. Bada. Primordial synthesis of amines and amino acids in a 1958 Miller H_2S-rich spark discharge experiment. *Proc Natl Acad Sci USA*, 108(14):5526–5531, 2011.

[PF11] G. K. Philip and S. J. Freeland. Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology*, 11(3):235–240, 2011.

[PH10] E. M. Phizicky and A. K. Hopper. tRNA biology charges to the front. *Genes Dev*, 24(17):1832–1860, 2010.

[PJP98] A. M. Poole, D. C. Jeffares, and D. Penny. The path from the RNA world. *J Mol Evol*, 46(1):1–17, 1998.

[PKH10] A. F. Philip, M. Kumauchi, and W. D. Hoff. Robustness and evolvability in the functional anatomy of a PER-ARNT-SIM (PAS) domain. *Proc Natl Acad Sci USA*, 107:17986–17991, 2010.

[PMH+04] V. Pezo, D. Metzgar, T. L. Hendrickson, W. F. Waas, S. Hazebrouck, V. Doring, P. Marliere, P. Schimmel, and V. De Crécy-Lagard. Artificially ambiguous genetic code confers growth yield advantage. *Proc Natl Acad Sci USA*, 101(23):8593–8597, 2004.

[Pri01] B. E. Prieur. Etude de l'activite prebiotique potentielle de l'acide borique. *Comptes Rendus de l'Academie des Sciences - Series IIc: Chemistry*, 4(8-9):667–670, 2001.

[PRR05] K. Plankensteiner, H. Reiner, and B. M. Rode. Prebiotic chemistry: The amino acid and peptide world. *Current Organic Chemistry*, 9(12):1107–1114, 2005.

[PRR06] K. Plankensteiner, H. Reiner, and B. M. Rode. Amino acids on the rampant primordial Earth: Electric discharges and the hot salty ocean. *Mol Divers*, 10(1):3–7, 2006.

[QAP] QAPLIB. http://www.opt.math.tu-graz.ac.at/qaplib/,2011.

[RBH+09] S. Rahman, M. Bashton, G. Holliday, R. Schrader, and J. Thornton. Small molecule subgraph detector (SMSD) toolkit. *Journal of Cheminformatics*, 1(1):doi 10.1186/1758–2946–1–12, 2009.

[RBL16] L. Raggi, J. L. Bada, and A. Lazcano. On the lack of evolutionary continuity between prebiotic peptides and extant enzymes. *Phys Chem Chem Phys*, 18:20028–20032, 2016.

[RCOB04] A. Ricardo, M. A. Carrigan, A. N. Olcott, and S. A. Benner. Borate minerals stabilize ribose. *Science*, 303(5655):196, 2004.

[RFJ07] B. M. Rode, D. Fitz, and T. Jakschitz. The first steps of chemical evolution towards the origin of life. *Chemistry and Biodiversity*, 4(12):2674–2702, 2007.

[RGW09] T. Romantsov, Z. Guan, and J. M. Wood. Cardiolipin and the osmotic stress responses of bacteria. *Biochim Biophys Acta*, 1788(10):2092–2100, 2009.

[RH10] W. Ran and P. G. Higgs. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol Biol Evol*, 27(9):2129–2140, 2010.

[RIA+10] S. Rajamani, J. K. Ichida, T. Antal, D. A. Treco, K. Leu, M. A. Nowak, J. W. Szostak, and I. A. Chen. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. *J Am Chem Soc*, 132(16):5880–5885, 2010.

[RKB08] M. Rogalski, D. Karcher, and R. Bock. Superwobbling facilitates translation with reduced tRNA sets. *Nat Struct Mol Biol*, 15(2):192–198, 2008.

[Rob04] M. F. Roberts. Osmoadaptation and osmoregulation in archaea: update 2004. *Frontiers in bioscience : a journal and virtual library*, 9:1999–2019, 2004.

[RS07] M. P. Robertson and W. G. Scott. The structural basis of ribozyme-catalyzed RNA assembly. *Science*, 315(5818):1549–1553, 2007.

[RS08] L. Randau and D. Söll. Transfer RNA genes in pieces. *EMBO Rep*, 9(7):623–628, 2008.

[RSB06a] C. Regni, A. M. Schramm, and L. J. Beamer. The reaction of phosphohexomutase from *Pseudomonas aeruginosa*: Structural insights into a simple processive enzyme. *J Biol Chem*, 281(22):15564–15571, 2006.

[RSB06b] C. Regni, G. S. Shackelford, and L. J. Beamer. Complexes of the enzyme phosphomannomutase/phosphoglucomutase with a slow substrate and an inhibitor. *J Acta Cryst F*, 62(8):722–726, 2006.

[RSR11] A. S. Rodin, E. Szathmary, and S. N. Rodin. On origin of genetic code and tRNA before translation. *Biol Direct*, 6:article number 14, 2011.

[RSSB99] B. M. Rode, H. L. Son, Y. Suwannachot, and J. Bujdak. The combination of salt induced peptide formation reaction and clay catalysis: A way to higher peptides under primitive earth conditions. *Orig Life Evol Biosph*, 29(3):273–286, 1999.

[RTB02] C. Regni, P. A. Tipton, and L. J. Beamer. Crystal structure of PMM/PGM: An enzyme in the biosynthetic pathway of *P. aeruginosa* virulence factors. *Structure*, 10(2):269–279, 2002.

[Rum66] I. B. Rumer. On codon sytematization in the genetic code. *Dokl Akad Nauk SSSR*, 167:1393–1394, 1966.

[RWFM72] D. Ring, Y. Wolman, N. Friedmann, and S. L. Miller. Prebiotic synthesis of hydrophobic and protein amino acids. *Proc Natl Acad Sci USA*, 69(3):765–768, 1972.

[SA02] G. Sella and D. H. Ardell. The impact of message mutation on the fitness of a genetic code. *J Mol Evol*, 54(5):638–651, 2002.

[SA06] G. Sella and D. H. Ardell. The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J Mol Evol*, 63(3):297–313, 2006.

[SABL83] T. Samuelsson, T. Axberg, T. Boren, and U. Lagerkvist. Unconventional reading of the glycine codons. *J Biol Chem*, 258(21):13178–13184, 1983.

[SAGA+03] C. Siebold, I. Arnold, L. F. Garcia-Alles, U. Baumann, and B. Erni. Crystal structure of the *Citrobacter freundii* dihydroxyacetone kinase reveals an eight-stranded alpha-Helical Barrel ATP-binding domain. *J Biol Chem*, 278(48):48236–48244, 2003.

[SBBG10] S. Shaul, D. Berel, Y. Benjamini, and D. Graur. Revisiting the operational RNA code for amino acids: Ensemble attributes and their implications. *RNA*, 16(1):141–153, 2010.

[SC66] M. Strassman and L. N. Ceci. Enzymatic formation of cis-homoaconitic acid, an intermediate in lysine biosynthesis in yeast. *J Biol Chem*, 241(22):5401–5407, 1966.

[Sch88] M. Schidlowski. A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature*, 333(6171):313–318, 1988.

[SFT07] Y. Sobolevsky, Z. M. Frenkel, and E. N. Trifonov. Combinations of ancestral modules in proteins. *J Mol Evol*, 65(6):640–650, 2007.

[SGS+11] M. A. S. Santos, A. C. Gomes, M. C. Santos, L. C. Carreto, and G. R. Moura. The genetic code of the fungal CTG clade. *Comptes Rendus - Biologies*, 334(8-9):607–611, 2011.

[SHD+04] A. Schiefner, G. Holtmann, K. Diederichs, W. Welte, and E. Bremer. Structural basis for the binding of compatible solutes by prox from the hyperthermophilic archaeon Archaeoglobus fulgidus. *J Biol Chem*, 279(46):48270–48281, 2004.

[Shi82] M. Shimizu. Molecular basis for the genetic code. *J Mol Evol*, 18(5):297–303, 1982.

[Shi95] M. Shimizu. Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: Origin of the genetic code. *J Biochem*, 117(1):23–26, 1995.

[Shi04] M. Shimizu. Histidine and its anticodon GpUpG are similar metabolic reaction rate enhancers: Molecular origin of the genetic code. *J Phys Soc Jp*, 73(2):323–326, 2004.

[Shi07] M. Shimizu. Amino acid and anticodon enhance metabolic reaction rates weakly but specifically: Genetic code world. *J Phys Soc Jp*, 76(5):article number 053801, 2007.

[SKM+06] P. S. Salgado, M. R. Koivunen, E. V. Makeyev, D. H. Bamford, D. I. Stuart, and J. M. Grimes. The structure of an RNAi polymerase links RNA silencing and transcription. *PLoS BIOL*, 4:article number 12, 2006.

[SLB+63] J. F. Speyer, P. Lengyel, C. Basilio, A. J. Wahba, R. S. Gardner, and Ochoa S. Synthetic polynucleotides and the amino acid code. *Cold Spring Harb Symp Quant Biol*, 28:559–567, 1963.

[SLER93] S. Saetia, K. R. Liedl, A. H. Eder, and B. M. Rode. Evaporation cycle experiments - A simulation of salt-induced peptide synthesis under possible prebiotic conditions. *Orig Life Evol Biosph*, 23(3):167–176, 1993.

[SLH+05] L. E. Snyder, F. J. Lovas, J. M. Hollis, D. N. Friedel, P. R. Jewell, A. Remijan, V. V. Ilyushin, E. A. Alekseev, and S. F. Dyubko. A rigorous attempt to verify interstellar glycine. *Astrophysical Journal Letters*, 619:914–930, 2005.

[SLM+98] K. Severin, D. H. Lee, J. A. Martinez, M. Vieth, and M. R. Ghadiri. Dynamic error correction in autocatalytic peptide networks. *Angewandte Chemie - International Edition in English*, 37(1-2):126–128, 1998.

[SLY89] S. U. Schneider, M. B. Leible, and X. P. Yang. Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. *Mol Gen Genet*, 218(3):445–452, 1989.

[SM83] G. Schlesinger and S. L. Miller. Prebiotic synthesis in atmospheres containing CH_4, CO, and CO_2. I. Amino acids. *J Mol Evol*, 19(5):376–382, 1983.

[SN99] T. Sugita and T. Nakase. Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida. *Systematic and Applied Microbiology*, 22(1):79–86, 1999.

[SNC+10] R. Saladino, V. Neri, C. Crestini, G. Costanzo, M. Graciotti, and E. Di Mauro. The role of the formamide/zirconia system in the synthesis of nucleobases and biogenic carboxylic acid derivatives. *J Mol Evol*, 71(2):100–110, 2010.

[Son65] T. M. Sonneborn. Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications. 1965. In: V. Bryson and H. J. Vogel (eds.) *Evolving Genes and Proteins*, pp. 377-397, New York: Academic Press.

[Spe06] D. Speijer. Is kinetoplastid pan-editing the result of an evolutionary balancing act? *IUBMB Life*, 58(2):91–96, 2006.

[Spe07] D. Speijer. Evolutionary aspects of RNA editing. 2007. In: H. U. Goringer (ed.) *RNA Editing*, pp. 199-227, Berlin Heidelberg: Springer-Verlag.

[Spe08] D. Speijer. Making sense of scrambled genomes. *Science*, 319(5865):901, 2008.

[SR89] M. G. Schwendinger and B. M. Rode. Possible role of copper and sodium in prebiotic evolution of peptides. *Anal Sci*, 5:411–414, 1989.

[SRB04] G. S. Shackelford, C. A. Regni, and L. J. Beamer. Evolutionary trace analysis of the alpha-d-phosphohexomutase superfamily. *Protein Science*, 13(8):2130–2138, 2004.

[SS05] M. G. Sacerdote and J. W. Szostak. Semipermeable lipid bilayers exhibit diastereoselectivity favoring ribose. *Proc Natl Acad Sci USA*, 102(17):6004–6008, 2005.

[SSV12] L. Samhita, S. Shetty, and U. Varshney. Unconventional initiator tRNAs sustain Escherichia coli. *Proc Natl Acad Sci USA*, 109(32):13058–13063, 2012.

[SSVMT03] R. Sanchez-Silva, E. Villalobo, L. Morin, and A. Torres. A new noncanonical nuclear genetic code: Translation of UAA into glutamate. *Curr Biol*, 13(5):442–447, 2003.

[SYH07] S. Sengupta, X. Yang, and P. G. Higgs. The mechanisms of codon reassignments in mitochondrial genetic codes. *J Mol Evol*, 64(6):662–688, 2007.

[SYSG01] A. Saghatelian, Y. Yokobayashi, K. Soltani, and M. R. Ghadiri. A chiroselective peptide replicator. *Nature*, 409:797–801, 2001.

[Sza93] E. Szathmary. Coding coenzyme handles: A hypothesis for the origin of the genetic code. *Proc Natl Acad Sci USA*, 90(21):9916–9920, 1993.

[Szo12a] J. W. Szostak. The eightfold path to non-enzymatic RNA replication. *Journal of Systems Chemistry*, 3:article number 2, 2012.

[Szo12b] J.W. Szostak. Attempts to define life do not help to understand the origin of life. *J Biomol Struct Dyn*, 29(4):599–600, 2012.

[SZS+05] V. Sosunov, S. Zorov, E. Sosunova, A. Nikolaev, I. Zakayeva, I. Bass, A. Goldfarb, V. Nikiforov, K. Severinov, and A. Mustaev. The involvement of the aspartate triad of the active center in all catalytic activities of multisubunit RNA polymerase. *Nucleic Acids Res*, 33(13):doi 10.1093/nar/gki688, 2005.

[Tak06] K. Takai. Classification of the possible pairs between the first anticodon and the third codon positions based on a simple model assuming two geometries with which the pairing effectively potentiates the decoding complex. *J Theor Biol*, 242(3):564–580, 2006.

[Tam15] K. Tamura. Origins and early evolution of the tRNA molecule. *Life (Basel)*, 5:1687–1699, 2015.

[TC89] F. J. R. Taylor and D. Coates. The code within the codons. *BioSystems*, 22(3):177–187, 1989.

[TCY10] R. M. Turk, N. V. Chumachenko, and M. Yarus. Multiple translational products from a five-nucleotide ribozyme. *Proc Natl Acad Sci USA*, 107(10):4585–4589, 2010.

[TIK04] Y. Takagi, Y. Ikeda, and Taira K. Ribozyme Mechanisms. 2004. In: J.-P. Majoral (ed.) *New Aspects in Phosphorus Chemistry IV*, pp. 213-251, Berlin Heidelberg: Springer-Verlag (*Topics in Current Chemistry* Series, Volume 232).

[TOL10] M. Turmel, C. Otis, and C. Lemieux. A deviant genetic code in the reduced mitochondrial genome of the picoplanktonic green alga Pycnococcus provasolii. *J Mol Evol*, 70(2):203–214, 2010.

[TW04] K. L. Tong and J. T. F. Wong. Anticodon and wobble evolution. *Gene*, 333(suppl.):169–177, 2004.

[TY03] K. Takai and S. Yokoyama. Roles of 5-substituents of tRNA wobble uridines in the recognition of purine-ending codons. *Nucleic Acids Res*, 31(22):6383–6391, 2003.

[UBL+09] Y. Ura, J. M. Beierle, L. J. Leman, L. E. Orgel, and M. R. Ghadiri. Self-assembling sequence-adaptive peptide nucleic acids. *Science*, 325(5936):73–77, 2009.

[UHLW04] D. W. Ussery, P. F. Hallin, K. Lagesen, and T. M. Wassenaar. Genome update: tRNAS in sequenced microbial genomes. *Microbiology*, 150(6):1603–1606, 2004.

[vdEvdBL+92] H. van den Ende, M. L. van den Briel, R. Lingeman, P. van der Gulik, and T. Munnik. Zygote formation in the homothallic green alga *Chlamydomonas monoica* Strehlow. *Planta*, 188(4):551–558, 1992.

[vdG07] P. T. S. van der Gulik. Three phases in the evolution of the standard genetic code: How translation could get started. Technical report, Centrum Wiskunde & Informatica, 2007. Arxiv0711.0700.

[vdGH11] P. T. S. van der Gulik and W. D. Hoff. Unassigned codons, nonsense suppression, and anticodon modifications in the evolution of the genetic code. *J Mol Evol*, 73(3-4):59–69, 2011.

[vdGMG+09] P. van der Gulik, S. Massar, D. Gilis, H. Buhrman, and M. Rooman. The first peptides: The evolutionary transition between prebiotic amino acids and early proteins. *J Theor Biol*, 261(4):531–539, 2009.

[VGC+01] D. Vernon, R. R. Gutell, J. J. Cannone, R. W. Rumpf, and Birky C. W. Accelerated evolution of functional plastid rRNA and elongation factor genes due to reduced protein synthetic load after the loss of photosynthesis in the chlorophyte alga Polytoma. *Mol Biol Evol*, 18(9):1810–1822, 2001.

[VGML13] A. Vogt, A. D. Goldman, K. Mochizuki, and L. F. Landweber. Transposon domestication versus mutualism in ciliate genome rearrangements. *PLoS Genetics*, 9(8):article number e1003659, 2013.

[VLL02] A. M. Velasco, J. I. Leguina, and A. Lazcano. Molecular evolution of the lysine biosynthetic pathways. *J Mol Evol*, 55(4):445–459, 2002.

[VMA09] F. A. P. Vendeix, A. M. Munoz, and P. F. Agris. Free energy calculation of modified base-pair formation in explicit solvent: A predictive model. *RNA*, 15(12):2278–2287, 2009.

[VV95a] D. Voet and J. G. Voet. *Biochemistry*. 1995. Second edition, p. 736, John Wiley and Sons, Inc.

[VV95b] D. Voet and J. G. Voet. *Biochemistry*. 1995. Second edition, p. 771, John Wiley and Sons, Inc.

[VV95c] D. Voet and J. G. Voet. *Biochemistry*. 1995. Second edition, p. 773, John Wiley and Sons, Inc.

[VWG06] K. Vetsigian, C. Woese, and N. Goldenfeld. Collective evolution and the genetic code. *Proc Natl Acad Sci USA*, 103(28):10696–10701, 2006.

[Wag05] A. Wagner. *Robustness and Evolvability in Living Systems*. Princeton University Press, 2005.

[WCM+07] J. T. F. Wong, J. Chen, W. K. Mat, S. K. Ng, and H. Xue. Polyphasic evidence delineating the root of life and roots of biological domains. *Gene*, 403(1-2):39–52, 2007.

[WDD+66] 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 Harb Symp Quant Biol*, 31:723–736, 1966.

[WDSD66] C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre. The molecular basis for the genetic code. *Proc Natl Acad Sci USA*, 55(4):966–974, 1966.

[WGM+94] P. Walde, A. Goto, P.-A. Monnard, M. Wessicken, and P. L. Luisi. Oparin's reactions revisited: Enzymic synthesis of poly(adenylic acid) in micelles and self-reproducing vesicles. *J Am Chem Soc*, 116:7541–7547, 1994.

[Wil12] L. D. Williams, 2012. http://astrobiology2.arc.nasa.gov/focus-groups/current/origins-of-life/articles/yin-and-yang-polypeptide-and-polynucleotide/.

[WK05] M. J. Whitehouse and B. S. Kamber. Assigning dates to thin gneissic veins in high-grade metamorphic terranes: A cautionary tale from Akilia, southwest Greenland. *J Petrol*, 46(2):291–318, 2005.

[WK07] Y. I. Wolf and E. V. Koonin. On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biol Direct*, 2:article number 14, 2007.

[Woe65a] C. R. Woese. On the evolution of the genetic code. *Proc Natl Acad Sci USA*, 54(6):1546–1552, 1965.

[Woe65b] C. R. Woese. Order in the genetic code. *Proc Natl Acad Sci USA*, 54(1):71–75, 1965.

[Woe67] C. R. Woese. *The Genetic Code*. Harper and Row, 1967.

[Woe73] C. R. Woese. Evolution of the genetic code. *Naturwissenschaften*, 60(10):447–459, 1973.

[WOIS00] C. R. Woese, G. J. Olsen, M. Ibba, and D. Söll. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev*, 64(1):202–236, 2000.

[Won75] J. T. Wong. A co-evolution theory of the genetic code. *Proc Natl Acad Sci USA*, 72(5):1909–1912, 1975.

[Won80] J. T. Wong. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc Natl Acad Sci USA*, 77(2 II):1083–1086, 1980.

[Won05] J. T. F. Wong. Coevolution theory of genetic code at age thirty. *BioEssays*, 27(4):416–425, 2005.

[Won07] J. T. Wong. Question 6: Coevolution theory of the genetic code: A proven theory. *Orig Life Evol Biosph*, 37(4-5):403–408, 2007.

[WSM+16] M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, and W. F. Martin. The physiology and habitat of the last universal common ancestor. *Nat Microbiol*, page article number 16116, 2016.

[Yar11] M. Yarus. The meaning of a minuscule ribozyme. *Phil Trans R Soc B*, 366(1580):2902–2909, 2011.

[YCK05] M. Yarus, J. G. Caporaso, and R. Knight. Origins of the genetic code: The escaped triplet theory. *Annu Rev Biochem*, 74:179–198, 2005.

[YMK+85] F. Yamao, A. Muto, Y. Kawauchi, M. Iwami, Y. Azumi, and S. Osawa. UGA is read as tryptophan in Mycoplasma capricolum. *Proc Natl Acad Sci USA*, 82(8):2306–2309, 1985.

[YOA+10] J. Yuan, P. O'Donoghue, A. Ambrogelly, S. Gundllapalli, R. L. Sherrer, S. Palioura, M. Simonović, and D. Söll. Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. *FEBS Lett*, 584(2):342–349, 2010.

[YWK09] M. Yarus, J. J. Widmann, and R. Knight. RNA-amino acid binding: A stereochemical era for the genetic code. *J Mol Evol*, 69(5):406–429, 2009.

[YWSN91] Y. Yamagata, H. Watanabe, M. Saitoh, and T. Namba. Volcanic production of polyphosphates and its relevance to prebiotic evolution. *Nature*, 352(6335):516–519, 1991.

[ZAZS12] T.F. Zhu, K. Adamala, N. Zhang, and J.W. Szostak. Photochemically driven redox chemistry induces protocell membrane pearling and division. *Proc Natl Acad Sci USA*, 109(25):9828–9832, 2012.

[ZDV71] E. Zuckerkandl, J. Derancourt, and H. Vogel. Mutational trends and random processes in the evolution of informational macromolecules. *J Mol Biol*, 59(3):473–490, 1971.

[ZF06] W. Zhu and S. Freeland. The standard genetic code enhances adaptive evolution of proteins. *J Theor Biol*, 239(1):63–70, 2006.

[Zha12] S. Zhang. Lipid-like self-assembling peptides. *Acc Chem Res*, 45(12):2142–2150, 2012.

[ZS12] Y. Zhang and D. Salahub. A theoretical study of the mechanism of the nucleotidyl transfer reaction catalyzed by yeast RNA polymerase II. *Sci China Chem*, 55:1887–1894, 2012.

[ZZDS08] D. A. M. Zaia, C. T. B. V. Zaia, and H. De Santana. Which amino acids should be used in prebiotic chemistry studies? *Orig Life Evol Biosph*, 38(6):469–488, 2008.

[ZZS12] N. Zhang, S. Zhang, and J. W. Szostak. Activated ribonucleotides undergo a sugar pucker switch upon binding to a single-stranded RNA template. *J Am Chem Soc*, 134(8):3691–3694, 2012.

Samenvatting

Computationele methoden bieden een krachtige manier om moeilijke problemen in de evolutionaire biochemie te onderzoeken. Een duidelijk voorbeeld hoe computationele methoden nieuwe en degelijke kennis kunnen opleveren in dit vakgebied, is de geschiedenis van het onderzoek naar regelmatigheden in de structuur van de genetische code. Na een eerste blik op de tabel die de translatie regels geeft van nucleinezuur naar eiwit, merkten verschillende onderzoekers op dat op elkaar lijkende aminozuren vaak gecodeerd werden door op elkaar lijkende codons. Een voorbeeld: Carl Woese merkte in het midden van de zestiger jaren op dat codons met C in de middelste positie zonder uitzondering aminozuren codeerden die niet echt groot en niet echt hydrofoob waren, maar zeker niet hydrofiel. Francis Crick plaatste een kritische kanttekening bij dit soort waarnemingen: de twintig aminozuren van de genetische code lijken allemaal veel op elkaar en voor het menselijk brein is het heel moeilijk om ergens géén patroon in te zien, zelfs als het een random presentatie betreft. Computationele methoden boden een uitweg bij dit meningsverschil. Door het produceren van een set van random herverdelingen van de codon toewijzingen, in combinatie met het gebruik van een kwantitatieve schaal van aminozuur karakteriseringen (ontwikkeld door Carl Woese en zijn collega's) en het gebruik van een functie om de "error robustness" van de genetische code en de door random herverdeling daarop geproduceerde varianten weer te geven, werd het mogelijk te bewijzen dat Carl Woese het bij het rechte eind had toen hij het genoemde patroon onder de aandacht bracht. Reeds in 1969 publiceerde Alff-Steinberger genoemde benadering, maar pas in de negentiger jaren werd het feit algemeen geaccepteerd, na werk van Hurst en zijn collega's.

Één van de gebieden waarop we computationele methoden hebben toegepast bij een probleem in de evolutionaire biochemie, was het raadsel van de oorspronkelijke peptiden. Wat was de functie van de eerste gecodeerde peptiden? Welke sequence fragmenten in eiwit-coderende genen zijn het alleroudst van al de

eiwit-coderende informatie? Wij benaderden deze vragen door aan te nemen dat oeroude biologische systemen een kleiner repertoire van aminozuren gebruikten in hun eiwitten. Om precies te zijn: we hebben aangenomen dat, in een bepaald stadium van het leven, eiwitten uit slechts vier soorten aminozuren bestonden: valine, alanine, aspartaat en glycine. Vervolgens hebben we in de PDB (Protein Data Bank) gezocht naar stukken eiwit die uit slechts deze vier aminozuren bestonden, met één positie vrij als een uitzonderingspositie om latere adaptatie van oude motieven niet geheel uit te sluiten. Opmerkelijk genoeg vonden we eiwitsoorten die fundamenteel zijn voor het leven: polymerases, mutases en kinases. Mutases en kinases spelen een rol in de glycolyse, wat een biochemische route van centraal belang is. De sequence "alanine-aspartaat-phenylalanine-aspartaat-glycine-aspartaat" in RNA polymerase is de active site van het enzym dat mRNA maakt in alle levende cellen. Wij trokken de conclusie dat onze procedure inderdaad sequence fossielen in bestaande eiwitten aan het licht bracht. Bovendien concludeerden we ook dat stukken eiwit die glycine en aspartaat bevatten, en tweewaardig positieve magnesiumionen manipuleren, tot de alleroudste coderende eiwit sequences behoord kunnen hebben. Mogelijkerwijs werden peptiden als "aspartaat-glycine-aspartaat" oorspronkelijk door een prebiotische omgeving gegenereerd, en misschien behoorden hun concentraties tot de eerste aspecten van het milieu waar het leven vat op kreeg.

Een ander gebied waarop we computationele methoden gebruikt hebben bij problemen in de evolutionaire biochemie, was de puzzel van de structuur van de genetische code. Zoals hierboven beschreven, gebruikte Hurst en zijn collega's een functie voor de "error robustness" van de genetische code om te laten zien dat op elkaar lijkende aminozuren in het algemeen gecodeerd worden door codons die op elkaar lijken. Carl Woese's "polar requirement" werd gebruikt om het op elkaar lijken van de verschillende aminozuren te kwantificeren, en aangetoond werd dat de "error robustness", die een resultaat is van de specifieke verdeling van codon-aminozuur toewijzingen, voornamelijk in de eerste en derde positie van het codon zit. Wij raakten door dit werk gefascineerd, en besloten om bepaalde mathematische aspecten ervan te verfijnen. Ten eerste wilden we het globale optimum van de error functie in de ruimte, die door de randomiserings-procedure gedefinieerd werd, weten. Een waarde (gevonden door Goldman met gebruikmaking van een zoekprocedure gebaseerd op heuristiek) werd in het vakgebied gebruikt alsof het het globale minimum was. Deze waarde was voor zover bekend de laagste waarde die in die ruimte bestond, maar het was niet bekend of het inderdaad het globale minimum was. Vanuit een wiskundig gezichtspunt is dat een uitermate onbevredigende situatie. Wij hebben het optimum gezocht, en bewezen dat de waarde die door Goldman is gemeld het globale minimum is. Terwijl we ons met die materie bezig hielden, merkten we dat wij bij onze berekeningen altijd een geleidelijke verdeling kregen van de waarden in de histogrammen, in tegenstelling tot wat we

in de literatuur zagen, waar de histogrammen een serie pieken en dalen vertoonden, waarvan men dacht dat deze een resultaat waren van de combinatie van de in groepjes gedistribueerde verdeling van waarden van aminozuur "polar requirement" en de patronen van de sets van codons wat betreft de eerste en derde positie. Omdat wij deze pieken niet reproduceerden, concludeerden wij dat deze gedachtengang op een misverstand moest berusten. Wij probeerden toen te vinden wat de oorzaak van de pieken moest zijn geweest, en kwamen tot de slotsom dat ze een artefact waren voortvloeiend uit de combinatie van afrondingsfouten in zowel de gegevens als in de begrenzingen van de bins van de histogrammen. Een ander facet van het werk waar we niet helemaal gelukkig mee waren, betrof de procedure waarmee random variante codes werden gegenereerd. Door het simpel herverdelen van aminozuur toewijzingen werd een ruimte gemaakt (waaraan wij de naam "Space 0" besloten te geven) die allerlei bekende genetische code varianten niet bevatte. Deze code varianten bestaan daadwerkelijk in vreemde uithoeken van het leven (en, wat betreft mitochondriën, delen van het leven die, in zekere zin, helemaal geen vreemde uithoeken zijn). Door het opzetten van een nieuwe procedure om random code varianten te genereren, maakten wij achtereenvolgens ruimtes die ook codes met "sense-to-sense reassignments", "stop-to-sense reassignments" en niet-in-gebruik-zijnde codons bevatten. We noemden deze ruimtes "Space 1", "Space 2" en "Space 3", en we definieerden tevens een "Space 4" welke ook hypothetische voorloper codes met minder dan 20 aminozuren bevat, en synthetische codes, gemaakt tijdens wetenschappelijke experimenten, waarbij aminozuren die door onderzoekers werden uitgezocht co-translationeel in eiwitten werden gezet. Met Space 1 en Space 2 konden we berekeningen uitvoeren, waarbij we vonden dat de belangrijkste aspecten van de relatie tussen de genetische code en de gemiddelde code niet wezenlijk veranderden, ondanks de (aanzienlijke) vergroting van de ruimte. Een verdere verfijning die we aan het vakgebied bijdroegen, was een kritische beschouwing van de gevolgtrekkingen die in het vakgebied getrokken werden, gebaseerd op berekeningen zoals hierboven beschreven. In het bijzonder werd tegen het licht gehouden hoe het concept "Frozen Accident" werd gebruikt. Ook de neiging om uit een lage waarde van de genetische code ten opzichte van de gemiddelde code te concluderen dat zeer grote hoeveelheden codes moeten zijn gescreened door natuurlijke selectie om tot de genetische code te komen zoals we die kennen, bleek niet de enige manier te zijn hoe men zo een resultaat kan interpreteren. Scenario's van code evolutie verschillen niet zozeer in het aantonen dat "error robustness" door codon toewijzingen aanwezig is in de code, maar in de manier waarop die scenario's voorstellen dat die "error robustness" tot stand is gekomen. De verfijningen die wij op deze manier aan het vakgebied hebben bijgedragen zijn in detail beschreven in het derde hoofdstuk.

In het vierde hoofdstuk beschrijven we een resultaat met betrekking tot een ander soort "error robustness" in de genetische code. Het is niet alleen zo dat op elkaar lijkende aminozuren vaak door op elkaar lijkende codons worden gecodeerd; codons die voor hetzelfde aminozuur coderen, lijken bijna altijd op elkaar (het

aminozuur dat, gedeeltelijk, een uitzondering vormt op deze regel, is serine). Een voorbeeld van deze "error robustness": alle arginine codons hebben een G in de middelste positie. Tijdens het overdenken van deze vorm van "error robustness" werd een verbluffend punt plotseling ontdekt. Niet-gemodificeerde anticodons kunnen slechts op een beperkt aantal manieren paren met verschillende codons. Een anticodon met G op de eerste positie paart met beide codons die op een pyrimidine eindigen, en een anticodon met C op de eerste positie paart alleen met een codon dat op G eindigt. De implicatie van deze meest basale "Wobble Rules" is dat een set tRNA's zonder anticodon-modificaties in staat is alle twintig aminozuren van de genetische code in eiwitten in te bouwen. Een complex modificatie-apparaat is niet nodig voor het functioneren van een vroege biochemie, wat precies is wat je verwacht als het systeem geëvolueerd is vanuit een relatief simpele toestand. Dit alles suggereert dat oorspronkelijk een aantal codons niet in gebruik was, omdat het eenvoudige systeem niet in staat was deze codons ondubbelzinnig te herkennen. Negatieve selectie heeft deze codons op een bijzonder laag niveau van aanwezigheid gehouden in de vroege eiwit-coderende sequences. Om precies te zijn: de exacte redenering achter de gevolgtrekking dat UUA, UAA, UGA, CAA, AUA, AAA, AGA en GAA codons zijn die niet in gebruik waren in een stadium van de ontwikkeling van de genetische code waarin alle twintig aminozuren wel al onderdeel uitmaakten van het aminozuur repertoire is uitgewerkt in het vierde hoofdstuk.

Het werk aan de genetische code is nog verder uitgewerkt in het vijfde hoofdstuk. Terwijl het derde hoofdstuk verfijning aanbracht in een reeds bestaande benadering, en het vierde hoofdstuk een tot nog toe over het hoofd geziene regelmatigheid belichtte, integreert het vijfde hoofdstuk verschillende aspecten, die allen als belangrijk in de evolutie van de genetische code worden gezien, in één mathematische procedure. Het belangrijkste punt in de redenering is dat àls bepaalde codon toewijzingen vastliggen dankzij stereochemische interacties tussen triplet en aminozuur (hetgeen gesuggereerd wordt door experimenteel werk in het vakgebied) die codon toewijzingen óók vast moeten liggen gedurende de randomiserings-procedure waarmee code varianten worden gegenereerd. Naast dit aspect is er een ander aspect wat eveneens in het model moet worden geïntegreerd, namelijk het concept van een geleidelijke groei van het repertoire, startend met valine, alanine, aspartaat en glycine, en gradueel ontwikkelend naar een twintig-aminozuren-code. Dit kan worden verwezenlijkt door het gebruik van een procedure om random codes te genereren die ontwikkeld is door Freeland en Hurst. Door deze verschillende aspecten in één model te verenigen komt volgens ons een realistisch model tot stand, voor de ruimte die beschikbaar was voor het vroege leven om code varianten te onderzoeken. In deze (kleine!) ruimte is de standaard genetische code optimaal. Als onderdeel van dit werk werd de gelijkenis in moleculaire structuur van de aminozuren onderzocht. Gebruik makend van de procedure uit het derde hoofdstuk om de positie-afhankelijkheid van "error robustness" te onderzoeken, vonden we dat met de "Molecular Structure Matrix",

die we ontwikkeld hadden, als input, de eerste en *tweede* codon positie error robustness bleken te bezitten (terwijl dit in het geval van "polar requirement" als input, de eerste en *derde* codon positie waren). Gesuggereerd wordt dat deze regelmatigheid het gevolg is van de geleidelijke uitbreiding van het aminozuurrepertoire van eenvoudige naar meer complexe aminozuren, in combinatie met een geleidelijke uitbreiding van het codon repertoire, startend met codons beginnend met purines (eerst G, later A) en uitbreidend naar codons beginnend met pyrimidines (waarbij codons beginnend met U de laatste zijn die aan het repertoire werden toegevoegd).

Het laatste hoofdstuk gaat over een ander probleem in evolutionaire biochemie dat onderzocht kan worden met computationele methoden. De expressie van het mitochondriaal erfelijk materiaal van de slaapziekte parasiet *Trypanosoma brucei* is zeer complex. Informatie die noodzakelijk is om te zorgen dat tal van uridine nucleotiden op de juiste plaatsen in het mRNA aanwezig zullen zijn, is, feitelijk, volstrekt verspreid over het mitochondriaal genoom. In de wetenschappelijke literatuur zijn reeds vele suggesties naar voren gebracht met betrekking tot de evolutionaire achtergrond van deze complexe organisatie; één van deze suggesties is dat deze organisatie een bescherming biedt tegen verlies van informatie als gevolg van intense competitie binnen de soort in combinatie met een complexe levenscyclus. Het zesde hoofdstuk geeft onze inspanningen weer om dit concept een mathematische onderbouwing te geven.

Abstract

Computational methods offer a powerful way to investigate difficult problems in evolutionary biochemistry. A clear example how computational methods can provide new and thorough knowledge in this area, is the history of the investigation of regularities in the structure of the genetic code. Upon visual inspection of the table which gives the translation rules of nucleic acid to protein, several investigators noted that similar codons often encoded similar amino acids. As an example: Carl Woese noted in the mid sixties that codons with C at the middle position encoded without exception amino acids that are not really large and not really hydrophobic, but certainly not very hydrophilic. Francis Crick placed a critical note to these kind of observations: all twenty canonical amino acids resemble each other and it is very difficult for the human mind to not think to see patterns, even in a random presentation. Computational methods offered a way out of this disagreement. By producing a set of random redistributions of the codon assignments, coupled with the use of a quantitative scale of amino acid characterizations (developed by Carl Woese and co-workers) and the use of a function to reflect the error robustness of the genetic code and its variants-by-random-redistribution, it became possible to prove that Carl Woese was right in recognizing the pattern. Already in 1969 Alff-Steinberger published this approach, but only in the nineties the fact was generally accepted, after work of Hurst and co-workers.

One of the areas in which we applied computational methods to problems in evolutionary biochemistry, was the enigma of the primordial peptides. What was the function of the first coded peptides? Which sequence fragments in protein-coding genes are the very oldest in all protein-coding information? We adressed these questions by assuming that ancient biological systems used a smaller repertoire of amino acids in their proteins. To be specific: we assumed that at a certain stage of life, proteins consisted of just four amino acids: valine, alanine, aspartic acid and glycine. We next searched the PDB (Protein Data Bank) for stretches of protein which consisted of just these four amino acids, with one position left as

an exception to not totally exclude later adaptation of old motives. Interestingly, we found types of proteins, which are fundamental to life: polymerases, mutases and kinases. The mutases and kinases play roles in glycolysis, which is a central pathway in biochemistry. The sequence "Alanine-Aspartic acid-Phenylalanine-Aspartic acid-Glycine-Aspartic acid" in RNA polymerase is the active site of the enzyme which produces mRNA in all living cells. We conjectured that our procedure indeed pinpointed sequence fossils in existing proteins. Furthermore, we concluded that protein stretches containing glycine and aspartic acid, and manipulating magnesium dications, may have been among the very first coded peptide sequences. Maybe, peptides like "Aspartic acid-Glycine-Aspartic acid" were originally produced by a prebiotic environment, and maybe their concentrations were among the first aspects of the environment which life managed to get under control.

Another area in which we applied computational methods to problems in evolutionary biochemistry, was the riddle of the structure of the genetic code. As pointed out above, Hurst and co-workers used a function for the error robustness of the genetic code to show that similar codons in the genetic code encode, in general, similar amino acids. Carl Woese's polar requirement was used to quantify the similarity of different amino acids, and it was shown that the error robustness which is a result of this distribution of assignments resides mainly in the first and the third position of the codon. We were fascinated by this work, and decided to refine some mathematical aspects of it. First of all, we wanted to characterise the global optimum of the error function in the space defined by the randomization procedure. In the field, a value (found by Goldman using a heuristic search procedure) was used as if it was the global minimum. This value was the lowest value known to exist in this space, but it was not known to be the global optimum. From a mathematical viewpoint, such a situation is very dissatisfying. We searched for the optimum and proved that the value reported by Goldman was the global minimum. During our work in this area, we noted that we always obtained a smooth distribution of values in our histograms, while the published histograms were characterised by spikes, which were thought to result from the combination of the discrete, clumped distribution of amino acid polar requirement and the patterns of codon blocks in the first and third bases. Because we did not obtain these spikes, we concluded that this line of reasoning had to be false. We tried to determine what the cause of the spikes must have been, and decided they are an artifact resulting from the combination of rounding errors in both the data and the bin borders of the histograms. Another facet of the work with which we were not completely happy, was the procedure with which random variant codes were generated. By swapping the amino acid assignments, a space was created in this field (a space which we decided to give the name 'Space 0') which did not contain variant codes which were in fact known to exist in remote corners of biology (and,

in the case of mitochondria, even corners of biology which, in a certain sense, are not at all remote). By devising a new procedure to generate random code variants, we enlarged the code space to successively contain known sense-to-sense reassignments, known stop-to-sense reassignments, and known unused codons. The resulting spaces were called 'Space 1', 'Space 2', and 'Space 3', and we also devised a 'Space 4' which contains hypothetical precursor codes with less than 20 amino acids, and experimental synthetic codes with amino acids selected to be co-translationally incorporated in proteins by researchers. We could perform calculations with Space 1 and Space 2, and found that the basics of the relationship between the genetic code and the average code did not change, despite the (considerable) enlargement of the space. A further refinement we contributed to the field was a critical examination of the conclusions drawn in the field based on calculations as described above. In particular the use of the concept "Frozen Accident" was examined. Also the tendency to conclude from a low error value of the genetic code as compared with the average code that very large amounts of codes must have been screened by natural selection to arrive at the genetic code as we know it, was shown to not be the only way one can interprete this low value. Scenarios of code development do not so much differ in showing that error robustness due to codon assignments is present in the genetic code, but in the way these scenarios propose this error robustness has been built. The refinements we contributed in this way to the field are described in detail in the third chapter.

In the fourth chapter, a result is reported concerning another kind of error robustness in the genetic code. Not only are similar amino acids often encoded by similar amino acids; identical amino acids are nearly always (serine being the exception) encoded by similar codons. As an example: all arginine codons share middle G. During contemplation of this kind of error robustness, a stunning fact was suddenly discovered. Unmodified anticodons are known to pair in a limited set of ways with several codons. An anticodon starting with guanine pairs with both codons ending with a pyrimidine, and an anticodon starting with cytosine pairs only with a codon ending on guanine. The implication of these most basic wobble rules is that a set of tRNAs without anticodon modifications is able to transfer all twenty amino acids of the canonical genetic code. No complex modification apparatus is needed in early biochemistry, which is exactly what would be expected if the system evolved from a simple origin. This observation suggests that originally certain codons would not have been in use because the simple system was not able to recognize them unambiguously. Negative selection would keep these codons on an extremely low level of presence in protein-coding sequences. To be precise: in the fourth chapter the exact argument for the conjecture that UUA, UAA, UGA, CAA, AUA, AAA, AGA, and GAA were unused codons in a stage of code development in which all twenty amino acids were already part of the amino acid repertoire is laid down.

The work on the genetic code is further elaborated in the fifth chapter. While the third chapter contributed refinements of an existing approach and the fourth

chapter highlighted a missed regularity, the fifth chapter integrates different aspects, which are all considered important in the evolution of the genetic code, in a single mathematical procedure. The main line of reasoning is, that if certain assignments are determined by stereochemical interactions between triplet and amino acid (which is indicated by experimental work in the field), these assignments should not be allowed to vary in the randomization procedure which provides code variants. Another aspect which should be integrated into the model, is the concept of a gradual growth of the repertoire, starting with valine, alanine, aspartic acid and glycine, and gradually developing towards a twenty amino acid code. This can be done by using a randomization procedure developed by Freeland and Hurst. Taken together, the different aspects provide a realistic model of the space available for early life to probe code variations. In this (limited!) space, the standard genetic code is optimal. During this work, the similarity of amino acids in molecular structure was investigated. Using the procedure of the third chapter to study the position dependence of error robustness, we found that with the Molecular Structure Matrix developed in this work as input data, error robustness was found to reside in the first and second position of the codon (as contrasted to the first and third position as found for polar requirement). It is suggested that this regularity derives from a gradual expansion of the amino acid repertoire from simple to complex amino acids combined with a gradual expansion of the codon repertoire from codons starting with purines (first guanine, later adenine) to codons starting with pyrimidines (codons starting with uracil being the last to be added to the set).

The last chapter deals with another problem in evolutionary biochemistry which can be investigated with computational methods. The expression of the mitochondrial genetical material of the sleeping sickness parasite *Trypanosoma brucei* is very complex. Information necessary to ensure that many uridine nucleosides are present in the right places in mRNA, is, in fact, scattered over the mitochondrial genome. Many different suggestions about the evolutionary background of this complex organization are brought forward in the scientific literature; one of these is that this organization provides a protection against loss of information due to intense intraspecific competition in combination with a complex life cycle. The sixth chapter presents our efforts to give this concept a mathematical foundation.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry