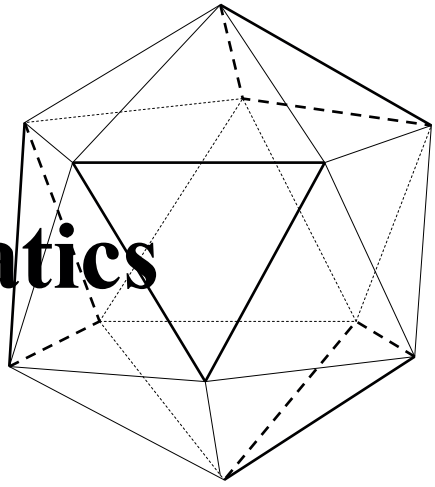
Wolf-Michael Wendler

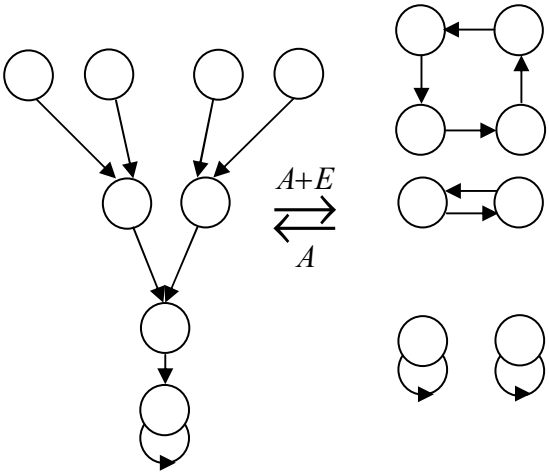
atggaacaacgcataaccctgaaagattat
gcaatgcgctttgggcaaaccaagacagc
taaagat ... taa



# Mathematics and Codons

$$O(3;\mathbb{F}(4)) = SO(3;\mathbb{F}(4))$$

$$GUG = (CCU)^*$$




$$(r,\omega) \in \mathbb{F}(16)$$

$$D_z(\beta) = \begin{pmatrix} \beta & \beta^2 & 0 \\ \beta^2 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\Rightarrow \quad D_z(\beta) \begin{pmatrix} \beta^2 \\ 0 \\ 0 \end{pmatrix}_{(CAA)^T} = \begin{pmatrix} 1 \\ \beta \\ 0 \end{pmatrix}_{(UGA)^T}$$

$$(TAAAAT)_r$$

SHAKER VERLAG

$$\underline{c} = \underline{r} + \underline{s}$$

Berichte aus der Mathematik

**Wolf-Michael Wendler**

# Mathematics and Codons

Second Edition

Shaker Verlag
Düren 2021

**Bibliographic information published by the Deutsche Nationalbibliothek**
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.



Printed in Germany.

ISBN 978-3-8440-7858-9
ISSN 0945-0882

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9
Internet: www.shaker.de • e-mail: info@shaker.de

# Preface to the First Edition

It is felt appropriate to start with a brief introduction to the six Chapters of this book.

The first Chapter essentially gives the basic mathematics used throughout the book, that is elementary number theory, elementary linear algebra, and in particular algebra. The latter comprises groups, rings, and fields with an emphasis on finite fields. Such fields are basic within the scope of coding theory and cryptography; see for instance [Friedrichs], [Pretzel], [Lidl, Niederreiter], [Niederreiter, Xing].

Finite fields are denoted by $\mathbb{F}(q), q = p^k, k > 0,$ where $p$ is a prime number. It may be shown that such fields exist for all prime numbers and all powers of primes, the number of elements is finite, namely $q$. To finite fields the characteristic $p$ is assigned. On the contrary, the field of real numbers $\mathbb{R}$ by definition is of characteristic 0, and so is the field of complex numbers $\mathbb{C}$. These fields contain uncountable infinite elements. Moreover, there is a crucial difference between finite fields and the field of real numbers. The field of real numbers contains, besides the usual laws of a field, laws for the ordering of its elements, that is relations "less than" or "greater than", that is $a < b$ or $a > b$. Such relations are not defined in finite fields, so an ordering is not possible, which however is true for the field of complex numbers in characteristic 0, too. Predominantly, we work on finite fields of characteristic $p = 2$.

Chapter 1 also gives a guide and help for readers not acquainted with this elementary mathematics. A certain understanding is necessary to follow the rest of the Chapters.

Chapter 2 is titled as representations of Codons and the protein code, so we have to say a bit on what the terms Codon, protein code, and genetic code mean. Starting with the term genetic code, this means the complete code written on the DNA of any organism. The DNA consists of two strands coiled around each other to form a double helix [Watson, Crick]. These strands in turn are built up by four distinct bases only, which in the chemical sense are adenine (A), cytosine (C), guanine (G), and thymine (T). We may as well say that the DNA is written in the alphabet of these four symbols. There is one particularity of this code, namely that we see a base pairing of the form (A, T) as well as (G, C). Each type of the base on one strand bonds with just one base on the other strand, for instance A bonds with T only, and vice versa, T bonds with A only. This is called complementary base pairing. We come back to this complementary behaviour below and give it a precise mathematical meaning.

By the term protein code, one means those parts of the genetic code or DNA, which one has found out that they are responsible for the coding of proteins. In case of proteins, three con-

secutive bases on one strand define a Codon. Because we have four bases A, C, G, T, there exist $4^3 = 64$ distinct Codons. Three of them are Stop Codons, which do not participate in the coding of amino acids. Therefore, we have $63 - 3 = 61$ Codons for coding 20 proteinogenic amino acids, at least in the usual case. Therefore, there exists an access of Codons. It has been found out, that there exist 20 multiplets, like singlets, doublets, quartets, sextets of Codons, which then code 20 amino acids. Codons of one multiplet are called synonym Codons where we add that this redundancy is chemical in nature, but certainly not in the mathematical sense. The amino acids in turn then are responsible for the synthesis of proteins. At the time being there are over 60,000 proteins known, making up roughly 2% of the code of the DNA [wiki$_2$]. Thus, we may speak of the genetic code in the general sense and the protein code in the special sense. Within this scope, repetition codes are also part of the genetic code making up roughly about 5% of the DNA.

In Chapter 2 we identify the four bases with the numbers 0, 1, $\beta$, $\beta^2$ of the finite field $\mathbb{F}(4)$ which is of characteristic $p = 2$. This is nothing new but is known from coding theory and bioinformatics, such codes are called DNA-codes and are used for storage of information. Then the numbers 0, 1 and $\beta$, $\beta^2$ are complementary with respect to each other. This is obvious for the former and holds for $\beta$, $\beta^2$ too, because in this field the equation $\beta^2 = \beta + 1$ is valid. This gives the complementary behaviour a more precise mathematical description. Moreover, our 64 Codons then may be viewed simply as a 3-dimensional vector space over the field $\mathbb{F}(4)$, which we then call the Codon space and which contains 64 3-dimensional vectors over $\mathbb{F}(4)$. In fact in characteristic $p = 2$ any number of a given field is complementary to exactly one of the other numbers. For the complementary behaviour of the bases, we refer also to [Feynman]. By doing so, we gain that the bases now are a matter of algebra, analysis, and in particular geometry. This opens up a huge variety of methods to be applied [Wendler$_3$].

In Section 2 of Chapter 2 we treat first representations of Codons, like the power and polynomial representation. They are basic in order to carry out calculations in the fields. Furthermore, we introduce the complex representation and work out a variety of geometrical aspects of the 3-dimensional Codon space, yielding a geometrical representation or a partition. While originally the term Codon is hooked on the 3-dimensional case, we give it a generalization to the $n$-dimensional case, which means that we working over $\mathbb{F}^n(4)$. In particular, we see that the 4-dimensional Codon space is of some importance for our purposes. The $n$-dimensional Codon space appears to be suitable in the geometrical sense, for the description of repetition codes, and in systems theory.

Section 3 of this Chapter contains the assignment of the Codon multiplets, such as singlets, doublets, triplets, quartets, and sextets, to the amino acids. These multiplets are defined in detail showing that they obey an inner structure. Based on the solutions of Diophantine equations for the cases of 20, 21, and 22 amino acids to be coded, we show that we obtain unique solutions by applying information theoretical methods, namely the maximum of the information entropy. These solutions reproduce the numbers of multiplets given in the literature. We point out that these numbers represent experimental facts, which seem not to be detected as such yet. Moreover, the basic scheme of the 64 Codons consisting of 16 quartets then defines the origin of the information entropy.

In preparation of Chapter 4 where the structure of proteins is treated, the first Section of Chap-

ter 3, comprises the differential geometry of space curves and the formulae of Frenet over the field $\mathbb{F}(4)$. This relies on the pictorial idea that the DNA and in particular a protein may be viewed as a space curve within the Codon space. The integration of the formulae of Frenet yields what is called a motion, which in turn is an element of the Euclidean group of motions. The position vector then is of the form $\underline{r}(s) = A\underline{\Omega}(s) + \underline{b}$. Thereby, $A$ is an element of the special orthogonal group SO(3; $\mathbb{F}(4)$), $\underline{\Omega}(s)$ is a helix vector and $\underline{b}$ a translation. This gives the guideline for the next two Sections. In Section 2 the special orthogonal group is investigated, which in characteristic $p = 2$ coincides with the orthogonal group O(3; $\mathbb{F}(4)$). We first derive the number of elements, their eigenvalues, and its subgroups. In the 3-dimensional case, this group turns out to be the icosahedron group. Section 3 contains the analysis of the helix vector with respect to analytical, algebraic, and geometric properties.

In Section 4 some supplements to the Euclidean group of motions as well as group operations are given. The latter throw a different glimpse on the properties of a group in general.

Chapter 4 treats elementary protein structure and similar codes. While in Section 1 the primary structure of proteins is defined by a space curve in the Codon space, Section 2 gives a simple test for localizing elementary secondary structures, which is felt instructive for what follows. It is based on the ad hoc assumption that the determinant of the Codon matrix, comprising three consecutive Codons of the code, is non-vanishing for a helical part. Section 3 then gives a detailed analysis of these secondary structures, where at the very beginning two Postulates are given. By means of these Postulates, we are able to retrieve the helical parts with parameters $(r, \omega)$ or a translation from the code. This is accomplished not directly by evaluating the Codons, but evaluating the Codon matrices of three consecutive Codons as well as the Codon length vectors. The latter comprise the lengths of the Codons.

Section 4 gives the comparison with the experimental data of the cro- and cii-protein of the enteriobacteria phage lambda. At the very beginning some preliminaries are given such as the algebra of the pairs $(r, \omega)$, which turn out to be elements of the field $\mathbb{F}(16)$. This allows for the assignment of these pairs and length vectors to the code of the protein. Within the frame of the evaluation of the code, definitions of Codon triple positions, patterns of concatenation, helices within the code, and some set theoretical considerations are given.

Section 5 treats the tertiary and quaternary structure, which is done rudimentary.

Finally, in Section 6 similar codes are investigated. This is not done by means of the alphabet of amino acids, but on the Codon level where we employ rotations, given by the elements of the orthogonal group, and translation to a given secondary structure in the Codon space. This also is done by similarity transformations known from linear algebra.

Chapter 5 and 6 are somewhat interweaved. In Chapter 5 an introduction to linear systems theory on finite fields is given, where an emphasis is put on characteristic $p = 2$. Given an arbitrary time-invariant system matrix $A$ of format $(n, n)$ the corresponding state space equations may be solved in the homogeneous and autonomous case by algebraic means. Thus, we are able to explain the structure of the state space, which is governed by cycles, counted as stable states, and unstable states. For the autonomous case, we prove that under certain precautions a period doubling takes place. Furthermore, we detect self-replicating structures. Both phenomena are found in characteristic 0 usually by nonlinear equations, only. While for multiple peri-

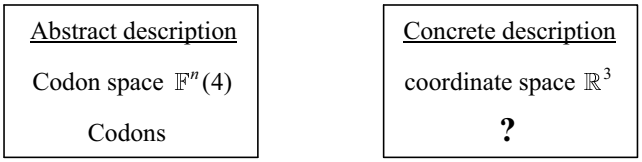
ods we refer to any textbook on nonlinear systems [Jetschke], [Guckenheimer, Holmes], for self-replication within hypercycles we refer to [Eigen, Schuster]. Introducing certain assumptions, we are also able to solve the general linear state space equations for time-variant inputs to the system. This gives the Chapter additional interpretations.

In Chapter 6, Section 1, repetition codes are investigated as well as finite sum transform is introduced and investigated. Moreover, we give three simple but elaborate examples of the prominent codes $(CA)_r$, $(TAG)_r$, $(CATC)_r$. It is shown that by applying certain reading frames we obtain all ingredients for constituting a linear system in the homogeneous or autonomous case as well as the inputs.

In Section 2 of this Chapter we treat first nonlinear systems in characteristic 0 and then carry this over to characteristic $p = 2$. Thereby, we show that the term nonlinear in finite fields does have a considerable different meaning as compared to characteristic 0. If for instance the function $f(x) = x^2$ is taken as the most simple nonlinear function in characteristic 0, in characteristic $p = 2$ this amounts merely to a reordering of the values of the range of this function, so there exists a one-one mapping to the function $f(x) = x$ which is linear. This does not remain true in characteristic $p > 2$. In particular, nonlinear functions over the field $\mathbb{F}(4)$ essentially do not exist. This resolves the situation.

Before we go on, we remark that Chapters 1 and 5 essentially are taken from the book of the author [Wendler$_1$] that is written in German. However, most of the contents were published over the years in the Proceedings of the International Workshop on Boolean Problems, which is are cited in [Wendler$_2$]. Furthermore, in Chapters 2 and 4 as well as the Appendices we give numerous Tables in order to make our findings as explicit as possible and in particular reproducible to the reader.

After this introduction to the book, we wish to give a somewhat different view on what we do and attempt. For this purpose, we give the following Figure, which is a slight generalization of the Figure given in Chapter 4, Section 1, and Definition 4.1.




On one hand, we have the DNA consisting of Codons of arbitrary but finite dimension within the Codon space. Because the DNA is written over the field $\mathbb{F}(4)$ we may speak of an abstract description of all genetic instructions used in growth, development, functioning, and reproduction of any living organism. However, these organisms appear in the usual 3-dimensional coordinate space as indicated on the right hand side. The question mark is supposed to indicate that there must be some counterparts of the Codons in the coordinate space.

We give a variety of examples. The most prominent example is given in Chapter 4, where we give the above Figure but the question mark is concretized to be amino acids. Once more, Codon multiplets code the alphabet of amino acids, which in turn are responsible for protein synthesis, so we end up in the 3-dimensional coordinate space.

Another type of example consists of the works of [Eigen, Schuster]. They prove that based on certain nonlinear equations in characteristic 0 there exist self-replicating structures. They also give matrices, which are adjoined to these graphs. The matrix elements are defined by certain exponents of variables entering the theory and are either 0 or 1.

Viewing these graphs in the frame of graph theory, these matrices appear as what is called adjacency matrices of the corresponding graphs. Moreover, these matrices may be interpreted as a repetition code arising from shifting the word by one to the left as we do in Chapter 6 and Section 1, which amounts to a certain but usual reading frame.

With respect to the above Figure, we have passed from the right hand side to the left. Still one has to be aware of what is proved. It is proved by [Eigen, Schuster] that by solving certain nonlinear equations over characteristic 0, one obtains a matrix of a self-replicating process, which in turn may be interpreted in terms of a repetition code. In addition, it is proved that we obtain self-replicating structures by solving linear equations over $\mathbb{F}(4)$. This does not prove that it exists on the DNA, but it makes it possible.

Another example of this type is the partitions of the 3-dimensional Codon space into different icosahedrons, which are geometrical objects usually appearing in characteristic 0. However, we have proved that such objects also exist over the field $\mathbb{F}(4)$, where again we do not know at which place and form of the DNA, if at all, they appear, but again it is possible. However, the DNA must contain a lot of geometrical information too. Within this line our findings of period doubling or even multiple periods may be associated with the huge number of control processes running in any living organism.

Finally, in [Schrödinger] it is stated that he has enhanced a new interest of physics to investigate the physical structure of the genetic information [Fischer, therein]. In addition, it is claimed that one should be open to new physical laws describing life processes. On the contrary, we apply mathematics, like number theory, algebra, linear algebra, geometry, analysis, graph theory, and systems and information theory to minor pieces of the DNA. To be sincere the author is quite astonished and delighted that this is possible at all, because a priori it may not be expected that the DNA offers this. Thereby, when we speak of an application we take advantage of the fact that the term application is not well defined.

Braunschweig, Spring 2017

Wolf-Michael Wendler

# Preface to the Second Edition

The second edition gains essentially two new Chapters, namely Chapter 5 and 6, so that the former Chapters now become Chapter 7 and 8. In Chapter 5, we present an alternative approach to the elementary structure of proteins based on the usage of affine geometry. Within this scope, the primary protein structure is described by the eigenvalues of the Codon matrix $C$ which allow for an assignment to affine mappings. These are introduced in Section 1. In Section 2 the usual 2-dimensional canonical form of affinities are generalized to the 3-dimensional case, which is not unique in general. Nevertheless, we see from Chapter 4, that the description by $(r, \omega)$-pairs is not unique as well.

In Section 3 on the secondary structure of proteins, we use the fixed points of the Codon matrix $C$, where the fixed-point equation $\underline{x} = C\underline{x} + \underline{t}$ within the frame of affine geometry is formally alike to what we know from motions, see above. However, we are not hooked on orthogonal matrices but may use the Codon matrices directly, which in general are not orthogonal. For the translational part, we employ by Postulate 5.1 the Start-Codon, which gives it a mathematical meaning. Postulate 5.2 requires that the equality of fixed points for some distinct Codon matrices indicate a secondary structure or part of it. In addition, we use accumulated fixed points, which pictorially are total directions. While the fixed points give the basic level, the accumulated fixed points define a first level. For the secondary structure, we use in addition a second level consisting of a sum over the first level.

In Chapter 4, Section 6, we introduce for the tertiary structure of proteins the terms lines and axes associated with secondary structures, which originally relies on the Codon quartets of the form GGZ (glycine) and CCZ′ (proline). The former allow for different conformations of the polypeptide chain while the latter may appear at the beginning or the end of a secondary structure, only. Thereby, we define axes such that no point of a line coincides with the Codons belonging to the corresponding secondary structure. Within this frame, we consider intersections of axes, which give a unique value for the scalar product of the axes under consideration.

This may be carried over to affine lines, which are investigated thoroughly in Section 4, Chapter 5. Therein, we find a variety of further methods for the determination of the value of the scalar product of axes or in this sense of affine lines. These rely all on the linear fractional transformation derived from the scalar product of affine lines. While the value of the scalar product naturally is an element of $\mathbb{F}(4)$, the comparison with experimental data is qualitative only. Nevertheless, by using the cardinalities of the sets of values for the scalar product, we may derive by means of the conformation or information entropy a funnel-shaped behaviour.

Chapter 6 gives a comparison of both approaches developed in Chapters 4 and 5. This is structured such that we go through primary, secondary, tertiary structure of proteins. In Section 2 of this Chapter, we treat similar protein codes, which in the first edition appear in Chapter 4. Section 3 contains mutations of the code within both approaches. While in Subsection 3.2, we investigate under which conditions a piece of code remains invariant, we also investigate point mutations within a doublet, a triplet and a quartet. In Subsection 3.3, a point mutation of a doublet is given as well as an interesting relation of the original and mutated code, namely a branching. This is somewhat similar to what we find in Chapter 5, Subsection 3.2.3, in conjunction with the $\alpha_2$-helix of the cii-protein of bacteriophage lambda. Finally, in Section 4 we gather some results throughout the book with respect to redundancy.

We mention that for Chapters 5 and 6, we added two Appendices A5 and A6.

In Chapter 7 on linear systems theory, we include in Section 3 two supplements. The first is concerned with the investigation of the states of a system rather than there structure. There we see certain regimes, like constant, periodic, linear, and exponential behaviour, which tentatively could be used within the control of a system. The second supplement consists of a relation of the homogeneous and autonomous systems.

Of course, this edition contains a variety of minor changes, clarifications, and supplements. We mention only one of them, namely that in Chapter 4 to the length vector UGC + permutations the radii $r = 1, \beta, \beta^2$ are assigned erroneously, while it now is given by $r = 0$. Besides, it is felt that partly the separation of secondary structures is improved.

Braunschweig, Winter 2021

Wolf-Michael Wendler

# Contents

# List of Figures, Tables, and Examples

## Figures

## Tables

**Examples**