BioSystems 204 (2021) 104392

Contents lists available at ScienceDirect

BioSystems

journal homepage: http://www.elsevier.com/locate/biosystems

Infinite combinatorics in mathematical biology

Saharon Shelah [a,b], Lutz Strüngmann [c,*]

[a] Einstein Institute of Mathematics, The Hebrew University of Jerusalem[1], 9190401, Jerusalem, Israel
[b] Department of Mathematics, Rutgers University, Piscataway, NJ, 08854-8019, USA
[c] Institute of Mathematical Biology, Faculty of Computer Sciences, Mannheim University of Applied Sciences, 68163, Mannheim, Germany

ARTICLE INFO

Keywords:
Forcing
Genetic code
Circular codes
Trees
Forests

ABSTRACT

Is it possible to apply infinite combinatorics and (infinite) set theory in theoretical biology? We do not know the answer yet but in this article we try to present some techniques from infinite combinatorics and set theory that have been used over the last decades in order to prove existence results and independence theorems in algebra and that might have the flexibility and generality to be also used in theoretical biology. In particular, we will introduce the theory of forcing and an algebraic construction technique based on trees and forests using infinite binary sequences. We will also present an overview of the theory of circular codes. Such codes had been found in the genetic information and are assumed to play an important role in error detecting and error correcting mechanisms during the process of translation. Finally, examples and constructions of infinite mixed circular codes using binary sequences hopefully show some similarity between these theories - a starting point for future applications.

1. Introduction

Nature has its own rules that very often can be modelled using mathematical theories. However, the description of nature's processes and structures in terms of mathematical functions or differential equations mostly involving a large number of parameters is usually a very complex task. Prominent examples are Fibonacci's rabbits whose population growth was described by the Fibonacci numbers or the modelling of cancer cells' life cycles. Another important property that appears in nature is symmetry. It is one of the central concepts in many mathematical theories and it is also one of the most visible patterns that can be observed in nature. Crystals, plants, animals and also the human body show very symmetric structures (just think of the human face with two ears, eyes etc.) and in the form of fractals even self-symmetry can be frequently detected in nature. Another important Example of nature's favour for symmetry is given by the standard genetic code table that offers several symmetry properties with respect to transformations. Thus geometry is also used for modelling and explaining structures and processes in nature. Obviously, also statistical investigations and tools from (bio-)informatics play an important role for the study of nature's secrets.

However, there exist many other beautiful theories in mathematics that offer a tremendous machinery and very interesting and powerful results that so far lack an application in (theoretical) biology, admitting that in some cases such applications may never be possible. Finite combinatorics is, of course, used all over the place but what about infinite combinatorics? Or infinite set theory? To the authors' knowledge there is only little (or even no) application of these fields in (theoretical) biology. In this article we start the attempt to make some aspects of infinite combinatorics available to (mathematical) biologists having only little background in this area. The main focus lies on the theory of forcing that was mainly developed by Paul Cohen in the 60s and later substantially extended by the first-named author. This technique allows one to construct new models of set theory that extend our daily life model based on the axioms of Zermelo and Fraenkel. In such extension models questions that are undecidable in our world have an answer. The most prominent Example is the Continuum Hypothesis asking for the precise cardinality of the Continuum $|\mathbb{R}|$. It was shown by Kurt Gödel and Paul Cohen that this question cannot be decided unless we pass to another model of set theory. Details, references and historical remarks will be given in Section 2 but let us mention in this introduction that the method of forcing allows one to construct new infinite binary sequences (infinite sequences of zeros and ones) that have prescribed properties. These sequences were used for the Continuum Hypothesis but also in many algebraic settings in order to construct structures like

* Corresponding author.
E-mail addresses: shelah@math.huji.ac.il (S. Shelah), l.struengmann@hs-mannheim.de (L. Strüngmann).
[1] https://shelah.ma.huji.ac.il/

https://doi.org/10.1016/j.biosystems.2021.104392
Received 17 November 2020; Received in revised form 11 February 2021; Accepted 18 February 2021
Available online 14 March 2021

groups or rings that are based on equations reflecting the combinatorics of those sequences. This is demonstrated in Section 3.

We then pass to a more biological setting and give an overview of the theory of circular codes (see Section 4). The theory of circular codes is based on a statistical finding by Arquès and Michel from 1996. They identified a set X of 20 codons in a large statistical analysis on genes of bacteria, archae, eukaryotes, plasmids and viruses that had the highest preference to the correct reading frame (frame 0) when compared to the two shifted frames (frame 1 and 2 shifted by 1 respectively 2 nucleotides in the $5'$-$3'$ direction). Later it was then shown that this set of codons X is a subcode of the genetic code that has remarkable error-detecting properties in the sense that circular code motifs (sequences of codons from the code X) allow one to retrieve, maintain and synchronise the correct reading frame in genes. This is called circularity. Several mathematical approaches have been developed to investigate and understand the theory of circular codes but most interestingly from a biological point of view is that circular code motifs are enriched in the reading frame of extant genes, tRNA sequences as well as in important functional regions of rRNA involved in the mRNA translation process. Also in the decoding center of the ribosome, circular code motifs can be detected disproportionately. Thus the circular code seems to be enriched in the main actors involved in translation, underlining its importance and possible role in the evolution of the genetic code. Details, references and historical remarks on the theory of circular codes will be given in Section 4.

In the final section we then give some easy examples and construction methods for infinite mixed circular codes which are probably known but which are also based on infinite binary sequences and their properties. The hope is that this shows some similarities - and not more - between the above theories, i.e. between the theory of forcing and its application to algebra as well as the theory of circular codes. The paper closes with a short conclusion.

We do not intend to give any evidence or proof for a proper use of infinite combinatorics in (theoretical) biology but just want to do a first step towards a novel approach to theoretical biology - may it be successful or not. Thus the goal of this work is to provide an easy access to methods from infinite combinatorics and set-theory to biologists and similarly to introduce a part of theoretical biology (the theory of circular codes) to mathematicians thereby hopefully initiating future research in this direction.

The paper is structured as follows. In Section 2 we explain the method of forcing in an easy way to non-mathematicians including a subsection on the history of forcing with the Continuum Hypothesis used as an Example of its application. In Section 3 we then show some construction methods from algebra, based on trees and forests, that are used for showing the existence of large complicated algebraic structures. Finally, in Sections 4 and 5 we give an introduction to the theory of circular codes and their role in biology, as well as some basic examples and results about infinite mixed circular codes.

2. Forcing - an easy approach

In this chapter we will try to present an easy approach to the method of forcing - a powerful method that was developed by Paul Cohen (Cohen, 1963, 2002) and led to the proof of many beautiful independence theorems in logic and set theory but most interestingly also in commutative group theory. In particular, the first-named author has developed the theory of forcing tremendously and has solved many long-standing problems in algebra and other fields using this method (see (Shelah, 1974) or (Shelah and Strüngmann, 2003)). Before we go into mathematical details we will start by some remarks about the fundamentals of mathematics. These fundamentals are by no means as uncontroversial as a layman might perhaps assume. There are certainly controversial views of this technique, but they hardly ever come up for discussion in everyday practical mathematics. Only in set theory (or proof theory) are such considerations of importance. Before 1930, the formalism founded by Hilbert had been dominant in mathematics and the natural sciences since 1880. The desire to eliminate or clarify certain contradictions and inconsistencies in the mathematics of the preceding decades was prevalent. Such inconsistencies as the paradox of Russell (Russell, 1903) needed an explanation. Georg Cantor also struggled with such a problem called the Continuum Hypothesis - the problem of an exact description of the Continuum, i.e. of the size of the set of real numbers $\mathbb{R}$. Of course, one has to agree first on how to define the size, or cardinality, of an infinite set and therefore Cantor (Cantor, 1879, 1895) developed the theory of ordinal and cardinal numbers (measures for infinite sets). However, he was not able to prove or disprove the Continuum Hypothesis, but he believed in its validity. Around the turn of the century mathematics could be described in frames of logical calculus, where a calculus is a formal system closed in itself. The idea of formalism - in short - is to get the certainty about mathematical statements by deriving them from previously fixed axioms by logically correct transformations. The underlying theory of our set theory axiom system are the axioms set up by Zermelo and Fraenkel (ZF) together with the Axiom of Choice (C) called ZFC - see (Zermelo, 1908). However, Kurt Gödel, probably one of the most famous mathematicians and logicians of the 20th century, published in 1931 (Gödel, 1931) his incompleteness theorem. This showed that every possible axiom system is always insufficient to prove all statements within this system and hence contains (an infinite number of) statements, which can neither be proved nor disproved in this calculus. Later Gödel and Cohen (using the method of forcing) showed that the Continuum Hypothesis belongs to such undecidable statements - see (Cohen, 1963) and (Gödel, 1938). This was the starting point for a far-reaching development, in which questions from various fields of mathematics were checked for their (un)decidability in ZFC. The basic idea of Cohen's proof was to add *new real numbers* to our model of set theory and we will see later how real numbers can be used in describing properties of the genetic code and the translation process of the ribosome.

Obviously, resources, time and energy are limited like in nature, hence it is impossible to explain in detail all mathematical notions and theories needed for the method of forcing. However, the authors have tried to make the following paragraphs as accessible as possible also for non-mathematicians by first giving some historical remarks about the birth of forcing and then considering the Continuum Hypothesis as an application.

2.1. Some historical remarks about the evolution of forcing

It was Georg Cantor (1845–1918) who developed the *Lehre der Mengen* (theory of sets - (Cantor, 1895)) and the associated new concept of infinity in mathematics and he is certainly considered to be one of the founders of set theory today. In particular, Cantor developed his theory of cardinal numbers and, probably of even greater importance, the theory of ordinal numbers. He regarded the arithmetic of transfinite numbers as one of his greatest achievements and, although set theory was initially ignored and was shaken by antinomies, it is – in its present axiomatic shape – one of the foundations of mathematics. The cornerstone of Cantor's theory was the notion of a *cardinal number*. To every set M Cantor assigned a power $|M|$, its *cardinality*. With the help of his axiom of well-ordering, in which he demanded that every set could be well-ordered (i.e. there is a total order with the property that every non-empty subset has a least element in this ordering), he was able to arrange the cardinals in ascending order. He used the first letter of the Hebrew alphabet $\aleph$, e.g. $|M| = \aleph_{17}$, as a designation for his cardinal numbers. The first infinite cardinal number $\aleph_0$ corresponds to the size of the natural numbers $\mathbb{N}$, which we call countable. By means of a diagonal argument Cantor was able to show $|\mathbb{N}| = |\mathbb{Q}|$ and later to prove $|\mathbb{R}| > |\mathbb{N}|$. However, Cantor's joy at the discovery of transfinite numbers was marred by his inability to answer the most obvious question:

How does the Continuum $2^{\aleph_0} = |\mathbb{R}|$ fit into the sequence of the cardinal numbers ?

Cantor conjectured the equality $2^{\aleph_0} = \aleph_1$, today known as *Continuum Hypothesis* (*CH*). David Hilbert later included the Continuum Hypothesis (that the Continuum corresponds to the first uncountable cardinal number) in his famous list of problems (Hilbert, 1901). Today, we know that the Continuum Hypothesis *is undecidable* in ordinary set theory, which is given by the well known Zermelo-Fraenkel axioms. What does this mean? In 1908 Zermelo (Zermelo, 1908) published the list of set theory axioms known to us. Together with Cantor's axiom of well-ordering (in equivalent form as axiom of choice *C*) the Zermelo-Fraenkel axioms form the today generally accepted axiom system *ZFC* of set theory. But the axioms are neither independent from each other, nor can their consistency be proved, as the incompleteness theorem of Gödel (Gödel, 1931) shows. In 1915 Löwenheim published (Löwenheim, 1915) a groundbreaking work, in which he dealt with formal systems and the predicate calculus. Löwenheim's work was difficult to read but was revised by Skolem (Skolem, 1920, 1922) and today the Löwenheim-Skolem theorem on formal languages and models of these languages is known to most mathematicians:

> *Each model $\mathcal{M}$ of a finite language has a countable submodel in which the same sentences are true as in $\mathcal{M}$.*

Today, Löwenheim is considered the father of model theory and although the Löwenheim-Skolem theorem is so easy to formulate, it initially raises confusing questions. When applied to the axioms of set theory, it yields a countable subset $\mathcal{M}$ of the class of all sets, so that all axioms of set theory are still valid if we consider only the sets in $\mathcal{M}$ and ignore all other sets. This seems to be a paradox at first sight, because Cantor had already shown, that uncountable sets exist. But already Skolem recognised, that *to be uncountable* simply means that there is no count in $\mathcal{M}$, i.e., no bijection to $\mathbb{N}$. Thus, a countable set can very well play the *role* of an uncountable set in $\mathcal{M}$ provided we consider it only within $\mathcal{M}$.

We will end this historical paragraph by explaining ordinal and cardinal numbers in a bit more detail, because later we will mainly deal with the Continuum Hypothesis. Cantor described two sets *A* and *B* as being of *equal size*, if there is a bijection between *A* and *B*. As already mentioned, Cantor was able to show that the power set of a set *M* is always *greater* than the power of the set *M* itself. To order the infinite numbers, he used the Cantor-Bernstein theorem (originally proved by Dedekind): *If there is a injection of A to B and an injection of B to A, then there is also a bijection between A and B.* So what Cantor had to show was that given two sets *A* and *B*, is it always the case that the power of one is less than or equal to the power of the other. For two arbitrary sets, this seems to be an insoluble problem, because how should one define an (injective) mapping from one set to the other, if nothing about the two sets is known? The way out, which Cantor pointed out, consisted in the notion of well-ordering. A *well-ordering* of *A* is an order in which every non-empty subset of *A* has a smallest element. Now it is not difficult to show, that of two well-ordered sets one can be mapped to an initial piece of the other and thus is (not necessarily strictly) smaller. Cantor defined an equivalence class of well-orderings as an *ordinal number*, which are well ordered again. The first ordinal numbers are, for Example, $0 = \varnothing$, $1 = \{0\}$, $2 = \{1, 0\}, \ldots, \omega = \{0, 1, 2, \ldots\}, \omega + 1 = \omega \cup \{\omega\}, \ldots$. If we accept the principle of well-ordering, i.e. every set can be well ordered, then to every set *M* there exists an ordinal number of equal power. We call the smallest of these ordinal numbers the *cardinal number* belonging to *M*. The cardinals are themselves well ordered again and are numbered by the $\aleph$'s. The first uncountable cardinal number is $\aleph_1$, which corresponds to the set of all countable ordinal numbers. According to the Continuum Hypothesis $\aleph_1$ corresponds to the size of the power set of $\aleph_0$ or the size of all real numbers $\mathbb{R}$, the Continuum. After Zermelo had shown the equivalence of the axiom of choice and the axiom of well-ordering and Hilbert had described the Continuum Hypothesis as one of the most important problems of mathematics, today it is known that the Continuum Hypothesis is undecidable in *ZFC*.

2.2. The Continuum Hypothesis and the birth of forcing

After Cantor and Hilbert were not able to prove or disprove the Continuum Hypothesis $2^{\aleph_0} = \aleph_1$, a new era of mathematics began in 1937. More specifically, on June 14, 1937, Kurt Gödel proved (Gödel, 1938) the consistency of *CH* with *ZFC*. Gödel constructed *Gödel's universe* - a model of set theory, in which both the Zermelo-Fraenkel axioms and the axiom of well-ordering apply, and the Continuum Hypothesis holds. The explicit construction is of no importance for us, but the construction only allowed sets which could be constructed using already known sets. The essential idea is as follows: Even if we take the point of view that every property defines a set, it is important that when defining *new* sets, those sets, about which the property speaks, are already defined or constructed. For a quite complete biography of Gödel we refer to the article by Dawson (Dawson, 1983). After Gödel had shown the consistency of *ZFC* and *CH* by the method of *internal models*, it remained open to prove or disprove the independence of *ZFC* and *CH*. Until 1963 such a proof was not forthcoming, although some partial results were obtained by Gödel himself.

In 1962, Paul Cohen finally began working to prove the independence of the Continuum Hypothesis. Cohen himself describes in a very nice way his thoughts and approaches in Cohen (1963), so that it would be presumptuous to report from his perspective. We will therefore only mention here the cornerstones of his thought experiment. Initially, Cohen concentrated on working on the consistency of the Zermelo-Fraenkel axioms and the negation of the axiom of well-ordering. His original idea was based on showing, by an induction proof, that any proof of the axiom of well-ordering can be shortened. In consequence no such proof could exist. However, Cohen was initially denied success and he changed his approach. Instead of turning to Proof Theory, he looked at formulas and *standard models* of set theory, read Gödel's work, and recognised that it is impossible to prove the existence of an uncountable model of *ZFC* in which *CH* is not satisfied. Once again, these considerations showed that countable models played a special role and eventually led Cohen to deal with countable models $\mathcal{M}$ of *ZF*. His intuition was that by adding *new elements* to $\mathcal{M}$, a new model $\mathcal{N}$ could be obtained. Just as Gödel constructed an *inner model* by ignoring sets, Cohen strove to construct *extension models*. In analogy to Gödel, who did not remove ordinal numbers from the model, Cohen made the decision not to add new ordinal numbers to $\mathcal{M}$. Since the integers are *absolute* (independent of the model), it was obvious to first add a *set of integers*, i.e. a real number, to $\mathcal{M}$. To adjoin an element from $\mathcal{M}$ itself does not present any difficulty, but does not provide a truly new model. Cohen recognised that it was intuitively best to add a set *G* that did not share any *special property* with $\mathcal{M}$, as is done in field theory in the case of field extensions by transcendental elements. If this were possible, Cohen could adjoin many such sets and thus construct a wide range of new models. Cohen called such elements *G generic* and his hope was that the *generic extensions* $\mathcal{M}[G]$ would again provide models of set theory. The last hurdle now was to specify what generic means and which statements about the extension model can be derived. One of Cohen's first ground breaking findings was that $\mathcal{M}[G]$ contains sets that cannot be constructed in the sense of Gödel. However, the adjunction of a new set to $\mathcal{M}$ was to be treated with caution for the following reason: Since $\mathcal{M}$ is countable, there exists an ordinal number α (outside of $\mathcal{M}$), which is larger than all ordinal numbers in $\mathcal{M}$. Now α is countable and therefore can be expressed by a set *A* of natural numbers. So if we now try to add the set *A* to $\mathcal{M}$, the newly created model $\mathcal{N}$ will contain the new ordinal number α - contrary to Cohen's decision not to add new ordinal numbers to $\mathcal{M}$. Once again, Cohen was looking for a way out.

Cohen's solution was to define *generic* to be inductive. Thus, the new set *A is not completely described*, but properties of *A* are defined due to *incomplete (partial) information* about *A*. This incomplete information forms the set of *forcing conditions P*. To illustrate this, let us take the Example of the set *A* of natural numbers again. The conditions $p \in P$ give us finite information about *A*, i.e. they determine for a finite set of

natural numbers whether they are contained in A or not. Thus, if A' is a set of $\mathscr{M}$ and $p \in P$, then we can force $A' \neq A$ by choosing a natural number n, which has not yet been defined by the *finite* condition p. So then we extend p and thus force $n \in A$ or $n \notin A$ - depending on whether $n \in A'$ or not. This fundamental idea shows that A is a really new (not constructable) set, but is far from being completely described. Finally, let us turn to the Continuum Hypothesis to illustrate Cohen's principle once again. Assuming we want to show that *CH* is violated, the simplest method is to add many sets of natural numbers A_α $(\alpha < \aleph_2)$ to $\mathscr{M}$. Due to the *generic* properties it follows that all these sets A_α must be different. Thus, in the extension model $\mathscr{N}$, *CH* seems to be violated, but this should also be treated with caution, because $\neg CH$ means that the Continuum does not correspond to the first uncountable cardinal number in $\mathscr{N}$. However, the ordinal number $\aleph_2$ from $\mathscr{M}$ was used in the construction. Thus, it must be ensured that this ordinal number also corresponds to the second uncountable cardinal number in $\mathscr{N}$ in order to really refute *CH*. Cohen had to overcome this last hurdle by adding the *countable chain condition* to the set of forcing conditions, as will be explained in the next paragraph.

Before we specify forcing mathematically, we would like to end this paragraph with a quote from Paul Cohen from Cohen (2002). The second-named author met Cohen in June 2001 at a conference on Abelian groups in Hawaii. He gave a talk on the discovery of forcing. In the related article (Cohen, 2002), Cohen concludes by asking the question whether - regardless of the undecidability results - the Continuum Hypothesis is false or true. Here is his view:

> *…. I think the consensus will be that CH is false. The intuition that pleases me most strongly is the following: The axiom of separation, or replacement, and the axiom of the power set are in some sense orthogonal to each other. No process for describing a cardinal by a property of the type used in the replacement axiom (here I must be vague) can adequately describe the size of the Continuum. Thus I feel that the Continuum is greater than* $\aleph_2$*, etc.*

2.3. The mathematics of forcing

In this paragraph we want to explain the basics of forcing and demonstrate simple applications based on the Continuum Hypothesis. For more details and unexplained notations we refer the reader to the books by Kunen (Kunen, 1980) or Burgess (Burgess, 1977). First we need the corresponding notation.

Let $P = (P, \leq)$ be a partially ordered set together with a unique element 1_P that is maximal with respect to the order $\leq$. However, it is assumed that P does not contain any minimal elements with respect to $\leq$. We will use the following notation for $p, q \in P$:

- p and q are **comparable** if $p \leq q$ or $q \leq p$;
- p and q are **compatible** if there is some $r \in P$ such that $r \leq p$ and $r \leq q$;
- p and q are **incompatible** if they are not compatible.

A subset $A \subseteq P$ is called

- **open** if for all $p \in A$ and $q \leq p$ $(q \in P)$ it follows that $q \in A$; i.e. A is closed downwards;
- **dense** if for all $p \in P$ there is some $q \in A$ such that $q \leq p$;
- **dense below** $p \in P$ if for all $p' \leq p$ $(p' \in P)$ there is some $q \in A$ such that $q \leq p'$;
- **an antichain** if any pair of different elements from A is incompatible;
- **a maximal antichain** if A is an antichain and for all $p \in P$ there exists a compatible $q \in A$.

In order to construct Cohen's extension model we shall need the definition of *generic filter* as described in Paragraph 2.2.

Definition 2.1. Let $\mathfrak{D}$ be a family of dense subsets of P. A set $G \subseteq P$ is called a $\mathfrak{D}$**-generic filter** if the following conditions are satisfied:

(i) G is **upward closed**, i.e. if $p \in G$ and $p \leq q \in P$ then $q \in G$;
(ii) each pair of elements from G is compatible in G, i.e. for $p, q \in G$ there exists $r \in G$ such that $r \leq p, r \leq q$;
(iii) $G \cap D \neq \emptyset$ for all $D \in \mathfrak{D}$ which are dense in P.

If $\mathfrak{D} = \mathscr{P}(P)$ (the power set of P) then we call G a P**-generic filter**.

By definition the $\mathfrak{D}$-generic filter tell us something about the dense subsets of P which are in $\mathfrak{D}$. If we choose $\mathfrak{D}$ to be the set of all dense subsets of P then an easy argument shows that a $\mathfrak{D}$-generic filter even has non-trivial intersection with all sets that are dense below some element $p \in P$. However, we waive the proof here. A $\mathfrak{D}$-generic filter need not always exist, however, they do exist for sets $\mathfrak{D}$ which are countable. In fact, each element $p \in P$ can be embedded in such a $\mathfrak{D}$-generic filter.

Lemma 2.2. *Let* $\mathfrak{D}$*be a countable family of dense subsets of P.* If $p \in P$ then there exists a $\mathfrak{D}$-generic filter G with $p \in G$.

Proof. Let $\mathfrak{D} = \{D_n : n \in \omega\}$ be an enumeration of $\mathfrak{D}$ and $p \in P$. Inductively we define elements $p_n \in P$ $(n \in \omega)$ as follows:

- let $p_0 = p$;
- if p_n has been chosen, then there exists $p_{n+1} \in D_n$ such that $p_{n+1} \leq p_n$ since D_n is a dense subset of P.

Now put $G = \{q \in P | \exists n \in \omega$ such that $p_n \leq q\}$. An easy calculation shows that G is a $\mathfrak{D}$-generic filter that contains p. ∎

At this point the following remark should be made: If $\mathscr{M}$ is a model of set theory, then there exists according to the theorem of Löwenheim-Skolem (see Section 2.1) a countable submodel $\mathscr{M}'$ with the same theory. However, the enumeration of $\mathscr{M}'$ does not apply inside $\mathscr{M}'$ but outside, namely in $\mathscr{M}$. If P is now a partially ordered set in $\mathscr{M}'$, then the set of all dense subsets of P is countable (in $\mathscr{M}$). According to the above Lemma 2.2 a P-generic filter G exists. However, this filter is initially outside $\mathscr{M}'$ and not an element of $\mathscr{M}'$ but of $\mathscr{M}$. In fact, it is even almost always the case, that the generic filter does not contain any element of the model. For Example, if P is subject to the condition that for every $p \in P$ there are incompatible elements $q, r \in P$ with $q \leq p$ and $r \leq p$. In fact, the incompatibility condition is essentially equivalent to the non-existence of generic filters.

An axiom often used in algebra (Martin's axiom) requires the existence of such generic filters (in $\mathscr{M}'$) for sets $\mathfrak{D}$ of power $< 2^{\aleph_0}$. If the Continuum Hypothesis applies, then Martin's axiom is fulfilled according to Lemma 2.2. To illustrate this, here is another Example, showing that this demand is best possible when the Continuum Hypothesis is negated. Recall that for any function f by $dom(f)$ we mean the domain of f, i.e. the set of elements on which f is defined.

Example 2.3. *Let* $P = \{f : \mathrm{dom}(f) \to \{0,1\} | \mathrm{dom}(f) \subseteq \omega$ is finite $\}$. *Let the partial order P be given by* $p \leq q$ *if and only if p is an extension* (*as function*) *of q. For* $A \subseteq \omega$ *let* $D_A = \{p \in P | \exists n \in dom(p)$ such that $p(n) = 0$ if and only if $n \in A\}$. *If* $\mathfrak{D} = \{D_A : A \subseteq \omega\}$ *then there is no* $\mathfrak{D}$*-generic filter.*

Proof. By easy calculations it can be shown that the sets D_A $(A \subseteq \omega)$ are dense subsets of P and are pairwise different. Consequently, we have $2^{\aleph_0} = |D|$. Assume that there is a $\mathfrak{D}$-generic filter G. Since each two elements of G are G compatible, there exists $\tilde{g} = \bigcup_{g \in G} g : \omega \to \{0,1\}$. Now put $B = \{n \in dom(\tilde{g}) | \tilde{g}(n) = 1\}$. Since G is $\mathfrak{D}$-generic there exists a $p \in G \cap D_B$. However, this is a contradiction as we can easily see. ∎

Together with Lemma 2.2 we derive as an immediate corollary the

inequality $|\mathbb{R}| > |\mathbb{N}|$ which had already been observed by Cantor himself.

Corollary 2.4. *We have* $2^{\aleph_0} > \aleph_0$.

Cohen's idea was to construct new models of set theory using generic filters. Let $\mathscr{M}$ be a model in which the Zermelo-Fraenkel axioms and the axiom of choice apply. According to the theorem of Löwenheim-Skolem , there exists a countable submodel of $\mathscr{M}$ with the same theory. So we can limit ourselves to countable models without restriction. By Lemma 2.2 thus exists a *P*-generic filter *G* for each partially ordered set P of $\mathscr{M}$. The first main theorem about forcing says that $\mathscr{M}$ and G can be combined to a new model $\mathscr{M}[G]$ of *ZFC*. We do not want to give the exact construction here, because it would go beyond the scope of this work. Intuitively, however, $\mathscr{M}[G]$ can be derived from all sets that can be created with the help of *G* and elements from $\mathscr{M}$.

Theorem 2.5. *Let $\mathscr{M}$ be a countable model of ZFC and let $P \in \mathscr{M}$ be a partially ordered set. If G⊆P is a P-generic filter, then there exists a countable model $\mathscr{N} = \mathscr{M}[G]$ of ZFC with the following properties:*

(i) $\mathscr{M} \subseteq \mathscr{N}$;
(ii) $G \in \mathscr{N}$;
(iii) *$\mathscr{M}$ and $\mathscr{N}$ contain the same ordinal numbers.*

Two questions are now obvious if we want to decide whether the Continuum Hypothesis is fulfilled in $\mathscr{M}[G]$. First, we need to know how the elements look like in $\mathscr{M}[G]$ and second, we need a way to check the satisfiability of a statement in $\mathscr{M}[G]$. But how shall we describe the elements of $\mathscr{M}[G]$, let alone prove statements, if we do not know the generic filter *G* ? We even know that *G* is outside $\mathscr{M}$. The key to this lies in the construction of $\mathscr{M}[G]$. Cohen associated with each object $T = T[G]$ from $\mathscr{M}[G]$ a construction description $\widetilde{T}$ in $\mathscr{M}$ (independent of *G*). This description is called the *P-name* of *T*. *P*-names are defined inductively and determine the object $T[G]$ as *G*-interpretation of $\widetilde{T}$ using the generic filter.

Definition 2.6. Let P be a partially ordered set. A ***P*-name** τ is a relation that satisfies the following properties:

$$\langle \sigma, p\rangle \in \tau \quad \Rightarrow \quad \sigma \text{ is a } P-\text{name and } p \in P.$$

If *G* is a *P*-generic filter then the ***G*-interpretation** of τ is given by $I_G(\tau) = \{I_G(\tau')|\langle \tau', p\rangle \in \tau \text{ for some } p \in G\}$.

Simple examples of P names are $0 = \varnothing$ or $\tau_p = \{\langle 0, p\rangle\}$ for each $p \in P$. The corresponding G interpretations are then $I_G(0) = 0$ and $I_G(\tau_p) = 0$, if $p \notin G$ and $I_G(\tau_p) = \{0\}$, if $p \in G$. So as we can see, the interpretations of a P-name can be different depending on the choice of the generic filter. By construction, $\mathscr{M}[G]$ consists of all G interpretations of P names.

Theorem 2.7. *Let $\mathscr{M}$ be a countable model of ZFC and $P \in \mathscr{M}$ a partially ordered set. If G is a P-generic filter, then $\mathscr{M}[G] = \{I_G(\tau)|\tau \text{ a } P-\text{name }\}$. If $M \in \mathscr{M}$, then there exists a P-name τ such that $M = I_G(\tau)$.*

So now we have a description of the elements of $\mathscr{M}[G]$ and we can focus on the second question: which formulas apply in $\mathscr{M}[G]$? Let us say that an element $p \in P$ forces a statement φ , if φ is fulfilled in every model $\mathscr{M}[G]$, provided $p \in G$ is valid. Formally, this means:

Definition 2.8. *Let $\mathscr{M}$ be a countable model of ZFC, $P \in \mathscr{M}$ a partially ordered set, $p \in P$, $\tau_1, \cdots, \tau_n$ P-names in $\mathscr{M}$ and $\varphi(x_1, \cdots, x_n)$ a formula.*

Then P ***forces*** *$\varphi(\tau_1, \cdots, \tau_n)$ if $\varphi(I_G(\tau_1), \cdots, I_G(\tau_n))$ is satisfied in $\mathscr{M}[G]$ for any P-generic filter G such that $p \in G$.*

Since a generic filter is upward closed, larger elements from *P* force less than smaller elements. We therefore call the elements from *P* conditions, so that $p \leq q$ means that *q* is a weaker condition than *p*. It can easily be calculated that the operation of forcing respects the usual logical operations ∧, ∨ etc. Now we have a formalism to make statements in $\mathscr{M}[G]$, but this formalism talks about **all** *P*-generic filters - a further hurdle. But there is also a way out and the following is provable:

- Each statement that is satisfied in some extension model $\mathscr{M}[G]$ is forced by some condition $p \in G$;
- if a condition $p \in \mathscr{P}$ forces a statement φ then this can be decided within $\mathscr{M}$ i.e., we can decide within the model $\mathscr{M}$ if a statement φ is true in an extension model $M[G]$.

The proof of the last statement is complicated and already for simple formulas of the form $x_1 \in x_2$ it is very complex. Therefore, we will refrain from a more detailed explanation at this point (for details see the book by Kunen (Kunen, 1980)).

Let us now turn to the Continuum Hypothesis and first consider a simple Example of cardinal collapsing. It should be noted that certain properties are absolute, i.e. independent of the chosen model, such as the natural numbers, the property to being finite or to being an ordinal number. Thus ω is absolute, but a cardinal number λ, which is uncountable, does not necessarily have to be absolute.

Example 2.9. Let $\mathscr{M}$ be a countable model of *ZFC*. Let $\lambda^{\mathscr{M}}$ be an uncountable cardinal number in $\mathscr{M}$. Then there exists an extension model $\mathscr{N}$ of $\mathscr{M}$ in which $|\lambda^{\mathscr{M}}| = \aleph_0$ holds.

Proof. Let $P = P(\aleph_0, \lambda, \aleph_0) = \{f : \{0, \cdots, n-1\} \to \lambda^{\mathscr{M}} | n \in \mathbb{N}$ *and* f *injective*$\}$ be partially ordered by reversed inclusion. By Lemma 2.2 there exists a *P*-generic filter *G*. We put $F = \bigcup_{f \in G} f \in \mathscr{M}[G]$ and claim that *F* is a bijection between ω and $\lambda^{\mathscr{M}}$. Therefore, let

$$D_n = \{p \in P | n \in dom(p)\} \text{ for } n \in \omega$$

and

$$E_\alpha = \{p \in P | \alpha \in range(p)\} \text{ for } \alpha < \lambda^{\mathscr{M}}.$$

An easy calculation shows again that $D_n \in \mathscr{M}$ and $E_\alpha \in \mathscr{M}$ are dense subsets of P for all $n \in \omega$ and $\alpha < \lambda^{\mathscr{M}}$. Hence $G \cap D_n \neq \varnothing$ and $G \cap E_\alpha \neq \varnothing$ for all $n \in \omega$ and $\alpha < \lambda^{\mathscr{M}}$. Thus $\mathrm{dom}(F) = \omega$ and $\mathrm{range}(F) = \lambda^{\mathscr{M}}$. Since $F \in M[G]$ we conclude that $\lambda^{\mathscr{M}}$ must be countable in $\mathscr{M}[G]$. ∎

Following this Example, we want to construct a model of *ZFC* in which the Continuum Hypothesis is violated. We are looking for a model $\mathscr{N}$ in which an embedding $F : \omega_2^{\mathscr{N}} \to \mathscr{P}^{\mathscr{N}}(\omega)$ exists. The naive approach is to start with a countable model $\mathscr{M}$ of *ZFC* and then, as in Example 2.9, construct an extension model $\mathscr{N}$ of $\mathscr{M}$ in which an embedding $F : \omega_2^{\mathscr{M}} \to \mathscr{P}(\omega)$ exists. However, to really violate the Continuum Hypothesis, we still make sure that $\omega_2^{\mathscr{M}}$ does not collapse as in Example 2.9. For this we need the following definition:

Definition 2.10. Let *P* be a partially ordered set. *P* fulfills the **countable chain condition** if there is no uncountable anti-chain in *P*.

The following theorem shows that a partially ordered set $P \in \mathscr{M}$, which satisfies the countable chain condition, preserves cardinals and their cofinalities. The cofinality $\mathrm{cf}(\kappa)$ of a cardinal number κ is the smallest cardinal number γ, so that κ can be written as the limit of γ many smaller cardinal numbers. For Example, $\mathrm{cf}(\aleph_1) = \aleph_1$, but $\mathrm{cf}(\aleph_\omega) = \mathrm{cf}(\lim_{n\in\omega}\aleph_n) = \aleph_0$. A cardinal number κ with $\mathrm{cf}(\kappa) = \kappa$ is called *regular*, otherwise *singular*. We use the following theorem without proof.

Theorem 2.11. *Let P be a partially ordered set. If P satisfies the countable chain condition, then P preserves cofinalities and cardinal numbers.*

Thus the countable chain condition ensures that for each *P*-generic filter *G* and each cardinal $\kappa \in \mathscr{M}$, *k* is again a cardinal in $\mathscr{M}[G]$ with the same cofinality. Specifically, ω_2 cannot collapse. We are now ready to derive the theorem proved by Cohen.

Theorem 2.12. (*Cohen*). *There exists a model of set theory where the Continuum Hypothesis is violated.*

Proof. Let $\mathscr{M}$ be a countable model of *ZFC* and $\kappa^{\mathscr{M}} \in \mathscr{M}$ a cardinal number $\geq \aleph_2$. We choose

$P = P(\kappa \times \omega, 2, \aleph_0) = \{f : \mathrm{dom}(f) \subseteq \kappa \times \omega \to \{0,1\} || \mathrm{dom}(f)| < \aleph_0\}$.

Let G be a P-generic filter and put $F = \bigcup_{f \in G} f$. Analogous to Example 2.9 it follows that $F : \kappa \times \omega \to \{0,1\} \in \mathscr{M}[G]$. Moreover, P satisfies the countable chain condition (as we can easily see) and hence $\kappa^{\mathscr{M}} = \kappa^{\mathscr{M}[G]} = \kappa$. We now put

$F_\alpha : \omega \to \{0,1\}$ via $F_\alpha(n) = F(\alpha, n)$

for $\alpha < \kappa$. Using the dense sets

$D_{\alpha,\beta} = \{p \in P | \exists n \in \omega$ with $(\alpha, n), (\beta, n) \in \mathrm{dom}(p)$ and $p(\alpha, n) \neq p(\beta, n)\}$

it can easily be shown that $F_\alpha \neq F_\beta \in \mathscr{M}[G]$ is true for all $\alpha \neq \beta < \kappa$. Therefore in $\mathscr{M}[G]$ there exist at least κ many functions from ω to $\{0,1\}$ which shows $\aleph_2 \leq \kappa \leq |\mathscr{P}(\omega)|$. It follows that $2^{\aleph_0} \geq \aleph_2$. ∎

So the above theorem shows the consistency of *ZFC* and $\neg CH$. In particular, we have shown that even a model of ZFC and *CH* can be extended so that *CH* is violated. Of course, this is only the tip of the iceberg and, as mentioned in the first paragraph, the Continuum can be any cardinal which does not contradict König's result. However, we want to be content with what we have said so far and now we want to follow Gödel's path and construct a model of ZFC, in which the Continuum Hypothesis is fulfilled. Analogous to the countable chain condition, we need another property of partially ordered sets.

Definition 2.13. Let P be a partially ordered set and $\kappa \geq \omega$ a cardinal number. P is called **κ-distributive** if the intersection of κ many open and dense sets of P is again dense in P.

In parallel to Theorem 2.11 we need (without proof):

Theorem 2.14. *Let $\mathscr{M}$ be a countable model of ZFC, $P \in \mathscr{M}$ a partially ordered set and $\kappa \in \mathscr{M}$ a cardinal. If P is κ-distributive and G is a P-generic filter, then the following hold:*

(i) *Each cardinal $\leq \kappa^+$ is preserved in $\mathscr{M}[G]$;*
(ii) $\mathscr{P}^{\mathscr{M}[G]}(\kappa) = \mathscr{P}^{\mathscr{M}}(\kappa)$.

The consistency of *ZFC* and *CH* now is an easy consequence.

Theorem 2.15. (*Gödel*). *There exists a model of ZFC in which the Continuum Hypothesis holds.*

Proof. Let $\mathscr{M}$ be a countable model of *ZFC*. We choose

$P = \{f : \mathrm{dom}(f) \to \mathscr{P}(\omega) || \mathrm{dom}(f)| \leq \aleph_0, \mathrm{dom}(f) \subseteq \omega\}$.

where P is, as before, partially ordered. Let G be a P-generic filter and, as before,

$F = \bigcup_{f \in G} f : \omega_1^{\mathscr{M}} \to \mathscr{P}^{\mathscr{M}}(\omega) \in \mathscr{M}[G]$

is a surjective function. An easy calculation shows that P is ω-distributive and therefore Theorem 2.14 implies that $\omega_1^{\mathscr{M}} = \omega_1^{\mathscr{M}[G]}$ as well as $\mathscr{P}^{\mathscr{M}}(\omega) = \mathscr{P}^{\mathscr{M}[G]}(\omega)$. Since $F \in \mathscr{M}[G]$ it follows that $2^{\aleph_0} = |\mathscr{P}^{\mathscr{M}[G]}(\omega)| = \aleph_1$ in $\mathscr{M}[G]$. ∎

Gdel's statement was, of course, stronger than Theorem 2.15, since under $(V = L)$ even the Generalised Continuum Hypothesis is valid, but the above theorems show relatively simple ways to influence cardinal arithmetic by skilful choices of a partially ordered set. There are much more complex forcing methods than those described here, e.g. Easton's forcing showing that the size of power sets can be almost arbitrary, but these are sufficient for the goal of this paper, namely to illustrate the basic idea of forcing. The authors therefore refer to (Kunen, 1980) or (Shelah, 1998) for further studies.

3. Algebraic objects defined by trees and forests

As we have seen in the previous section, the method of forcing allows one to construct models of ZFC that have - among many other nice properties - in particular new real numbers that did not exist before. These real numbers, viewed as infinite binary sequences $\eta : \omega \to \{0,1\}$ can be used to define algebraic objects like groups, modules or rings with prescribed properties. In this section we would like to demonstrate such a construction in order to present the idea behind the method and refer to the books by Göbel and Trlifaj (Göbel and Trlifaj, 2012) or Eklof and Mekler (Eklof and Mekler, 1990) for further such constructions. The reader will see that infinite binary sequences (real numbers) are cleverly interlocked with finite ones. We will see in Chapter 5 how a similar construction (in the sense that also infinite binary sequences are used) will help to construct codes that play a role in the theory about the evolution of the genetic code.

Let $\mathbb{Z}$ be the set of integers and choose $S = \{p_n : n \in \omega\} = \{2, 3, 5, 7, 11, 13, \cdots\}$ the infinite set of natural prime numbers. It is well-known that for any pair of different prime numbers p, q there is a representation of 1 as $1 = rp + sq$ for some $r, s \in \mathbb{Z}$, e.g. $1 = 2 \cdot 3 - 1 \cdot 5$. Consequently, we have $p_n\mathbb{Z} + p_m\mathbb{Z} = \mathbb{Z}$ for all $n \neq m$. We now want to define a so-called *topology* on $\mathbb{Z}$ using the set S and therefore choose a sequence of elements

$$q_0 = 1 \text{ and } q_{n+1} = p_n q_n \text{ for all } n \in \omega, \tag{1}$$

e.g. $q_1 = 2, q_2 = 2 \cdot 3, q_5 = 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13$. These elements form a *divisor chain* and hence the descending chain $q_n\mathbb{Z}\ (n \in \omega)$ satisfies $\bigcap_{n \in \omega} q_n\mathbb{Z} = 0$ which means that they generate a so-called *Hausdorff S-topology* on $\mathbb{Z}$. Once one has such a topology it is possible to form the *S-adic completion* of $\mathbb{Z}$ consisting of *limits* of sequences of integers from $\mathbb{Z}$ such that differences between members of the sequences become more and more divisible by the elements from S. We do not go into further detail here but we will see later that this allows one to formally build infinite sums of integers from $\mathbb{Z}$. The S-adic completion of $\mathbb{Z}$ is denoted by $\widehat{\mathbb{Z}}$ and satisfies $q_n\mathbb{Z} = q_n\widehat{\mathbb{Z}} \cap \mathbb{Z}$ for all $n \in \omega$.

Now our infinite binary sequences (i.e. the real numbers) come into play. Let $T = {}^{\omega >}\{0,1\}$ denote the *tree* of all finite *branches* $\tau : n \to \{0,1\}$ $(n \in \omega)$ where we identify n with the set $n = \{0, 1, \cdots, n-1\}$. Moreover, ${}^{\omega}2 = \mathrm{Br}(T)$ denotes all infinite branches $\eta : \omega \to \{0,1\}$. (See Fig. 1) Clearly, the restriction of any such infinite branch η to a subset $\{0, 1, \cdots, n-1\}$ will be a finite branch then, hence $\eta{\restriction}_n \in T$ for all $\eta \in \mathrm{Br}(T)$ $(n \in \omega)$. If $\eta \neq \mu \in \mathrm{Br}(T)$ then

$\mathrm{br}(\eta, \mu) = \inf\{n \in \omega : \eta(n) \neq \mu(n)\}$

denotes the *branch point* of η and μ.

The aim is now to incorporate the combinatorics of infinite branches into algebraic objects. Therefore we need to collect certain subtrees of T depending on their *domain*. Recall that the *length* of the finite branch $\tau : n \to \{0,1\}$ is denoted by $l(\tau) = n$. For $C \subset \omega$ we now define

$T_C = \{\tau \in T :$ if $e \in l(\tau) \setminus C$ then $\tau(e) = 0\}$

which collects all branches of T that are non-trivially (i.e. $\neq 0$) defined only on elements from C. Similarly,

$\mathrm{Br}(T_C) = \{\eta \in \mathrm{Br}(T) :$ if $e \in \omega \setminus C$ then $\eta(e) = 0\}$

and hence $\eta{\restriction}_n \in T_C$ for all $\eta \in \mathrm{Br}(T_C)$ $(n \in \omega)$.

As in nature, usually one tree is not enough to give shelter and hence we now construct a *forest* of trees. Let $\lambda \leq 2^{\aleph_0}$ be a suitable cardinal (recall that $2^{\aleph_0} = |\mathbb{R}|$). It is a well-known result that then there exists a family $\mathfrak{S} = \{C_\alpha \subset \omega : \alpha < \lambda\}$ of pair-wise *almost disjoint* infinite subsets of ω, i.e. $C_\alpha \cap C_\beta$ is finite for any pair $\alpha \neq \beta$. We let $T_\alpha = T_{C_\alpha}$ for $\alpha < \lambda$ which is a tree whose branches have support contained in C_α. Recall that the support of a function is the subset of its domain where the function operates non-trivially ($\neq 0$). We may even assume that each tree T_α is *perfect* for $\alpha < \lambda$, i.e. if $n \in \omega$ then there is at most one finite branch $\eta{\restriction}_n$ such that $\eta{\restriction}_{n+1} \neq \mu{\restriction}_{n+1}$ for some $\mu \in T_\alpha$. Finally, we are able to plant our

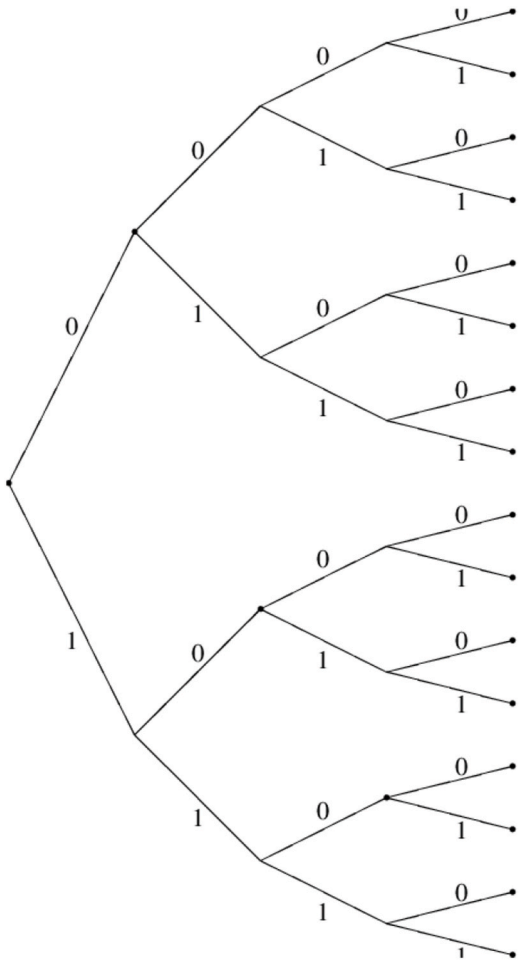

Fig. 1. The figure displays the set of all finite branches of length 4. Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Binary-tree-for-representing-all-binary-sequences-of-length-4-An-ascending-branch-fig1-225533425.

forest

$$T_\Lambda = \bigcup_{\alpha<\lambda} T_\alpha.$$

This forest will now be used to define our free *base group* as $B_\Lambda = \bigoplus_{\tau\in T_\Lambda} \mathbb{Z}e_\tau$ which sits nicely inside its S-adic completion $\widehat{B_\Lambda}$ taken in the S-topology on B_Λ.

Since $\lambda \leq 2^{\aleph_0} = |\mathrm{Br}\,(T_{C_\alpha})|$ we can easily choose a family $\{V_\alpha \subseteq \mathrm{Br}\,(T_{C_\alpha}) : \alpha < \lambda\}$ of subsets V_α of $\mathrm{Br}\,(T_{C_\alpha})$ such that $|V_\alpha| = \lambda$ for $\alpha < \lambda$. The important observation here is that for $\alpha \neq \beta < \lambda$ the infinite branches from V_α and V_β branch at almost disjoint sets since $C_\alpha \cap C_\beta$ is finite. Therefore, the pairs V_α, V_β are disjoint and we may assume that for any $m \in \omega$, λ pairs of branches in V_α branch above m.

The following definition is crucial now and shows how the properties of the infinite binary sequences (branches) are encoded into algebraic elements.

Definition 3.1. Let $x \in \widehat{B_\Lambda}$ be any element in the completion of the base group B_Λ. Moreover, choose any infinite branch $\eta \in V_\alpha$ with $\alpha < \lambda$. Then for any $n \in \omega$ the element

$$y_{\eta n x} := \sum_{i\geq n} \frac{q_i}{q_n}(e_{\eta\restriction_i}) + x\sum_{i\geq n}\frac{q_i}{q_n}\eta(i)$$

is called a **branch like element** and is a well-defined element from $\widehat{B_\Lambda}$. Note that each element $y_{\eta n x}$ formally is an infinite sum and connects an infinite branch $\eta \in \mathrm{Br}\,(T_{C_\alpha})$ with finite branches from the tree T_α. Moreover, the element $y_{\eta n x}$ encodes the infinite branch η into an element of $\widehat{B_\Lambda}$. The following equation is the crucial observation that relates two elements $y_{\eta n x}$ and $y_{\eta(n+1)x}$ and is used in the algebraic setting to ensure properties of the desired groups, modules or rings:

$$y_{\eta n x} = s_{n+1}y_{\eta(n+1)x} + z_{\eta\restriction_n} + x\eta(n) \text{ for all } \alpha<\lambda, \eta \in V_\alpha. \tag{2}$$

For the convenience of the reader we give a short argument that proves the above equation.

Proof (of equation (2)) We calculate the difference

$$q_n y_{\eta n x} - q_{n+1}y_{\eta(n+1)x}$$

$$= \sum_{i\geq n} q_i(z_{\eta\restriction_i}) + x\sum_{i\geq n} q_i\eta(i) - \sum_{i\geq n+1} q_i(z_{\eta\restriction_i}) - x\sum_{i\geq n+1} q_i\eta(i) = q_n z_{\eta\restriction_n} + q_n x\eta(n).$$

Dividing by q_n yields $y_{\eta n x} = s_{n+1}y_{\eta(n+1)x} + z_{\eta\restriction_n} + x\eta(n)$. ∎

To conclude this section we finally define a group using carefully chosen elements $x_\alpha \in B_\Lambda$:

$$G = \langle B_\Lambda, y_{\eta n x_\alpha} : \eta \in V_\alpha, \alpha<\lambda, n\in\omega\rangle. \tag{3}$$

Doing detailed calculations it is then possible to prove desired properties of G using the combinatorics of the infinite branches. For instance, a classical result claiming the existence of arbitrarily large commutative groups G with endomorphism ring $\mathrm{End}_{\mathbb{Z}}(G) = \mathbb{Z}$ being equal to the ring of integers. Once more we refer to the book by Göbel and Trlifaj (Göbel and Trlifaj, 2012) for further details.

After we have explained the method of forcing in Section 2 and how infinite binary sequences could be used to build complicated algebraic objects, we will show some similarities to these techniques that arise in a biological setting, namely the properties of the genetic code and its evolution. As we pointed out before in the introduction the intention is just to present a biological setting in which also infinite binary sequences are used/appear in the hope that there will be possible applications of deeper infinite combinatorics in the future.

4. Circular codes: error detecting mechanisms in the genetic code

In this section we will give an overview of/introduction to the theory of *circular codes* that are assumed to play an essential role in maintaining the correct reading frame during the translational process in the ribosome. Moreover, variants of circular codes hypothetically existed in ancient genetic codes. The motivation is that in the next section we will present some basic construction methods for (infinite) mixed circular codes using infinite binary sequences or equivalently real numbers. This hopefully shows that also in this biological setting such infinite sequences are helpful - similar to the way it was described in the algebraic setting in Section 3 - in the sense that they are encoded in e.g. codes that nature seems to use.

First we give some background on the biological setting and the theory of circular codes.

The discovery of the DNA double helix by Crick and Watson (Crick and Watson, 1953) in 1953 was undoubtedly a breakthrough in deciphering the origin of all life on earth. The discovery that the genome of (human) life is based on sequences of four so-called *nucleotide bases* only, pointed to a natural mechanism that underpins a connection between biology (genetics) and the theory of coding as well as mathematics. The biochemical process of genetic coding could be interpreted as an abstract problem of symbol manipulation. Motivated by this, researchers mainly focused on experimental aspects such as the sequencing of genomes of various organisms after the discovery of the standard genetic code. The basic idea of deciphering the origin and facets of life from pure knowledge of the structure of the genome led to numerous scientific projects, e.g. resulting in the sequencing of the complete human genome (Human-Genome). However, it is still not possible to explain or predict

the spatial structure and thus above all the function of the encoded protein from a genetic sequence of bases alone. This is certainly one of the core problems of modern research in this field. Similarly, the very complex translation of the genetic information into proteins - called *protein synthesis* - is sufficiently understood from a biological point of view, however, its robustness against potential errors is still a miracle. In Chapter 4.1 and Chapter 4.2 we will discuss one possible mechanism that nature has implemented in order to avoid so-called *frame-shift errors*. Before that we recall some basic facts about the genetic code.

The genetic code is based on four nucleic bases: adenine (A), uracil (U), guanine (G) and cytosine (C). Each triplet of such bases encodes one of the 20 amino acids or a *stop signal* and is called a *codon*. Moreover, the codon *AUG* also serves as a *start codon* in the translation process performed by the ribosome (see the next Section 4.1). This assignment of a codon to an amino acid is what we call the *genetic code*. As mentioned before, each codon encodes exactly one amino acid, however, the converse is not true. Apart from methionine and tryptophan (note that the codon *UGA* also encodes the stop signal), all amino acids are coded by several codons which is why the code is called *degenerated*. From a combinatorial point of view this also follows immediately from the fact that each codon consists of three nucleotides which themselves can be one of the four bases. Thus, there exist $4^3 = 64$ possibilities to encode amino acids. However, there are only 21 amino acids (including the stop signal) that are encoded and so it follows that the code is degenerated. Fig. 2 displays the genetic code as a wheel that is read from the inner circle to the outer one.

4.1. Biological background

In 1953, Watson and Crick (Crick and Watson, 1953) published the first fundamental findings on the properties of the structure of *deoxyribonucleic acid* (DNA): they discovered that it consists of two nucleotide strands that form a so-called *double helix*. The nucleotides are the four chemical bases adenine (A), thymine (T), guanine (G) and cytosine (C), which are glued together by hydrogen bonds working efficiently only between adenine and thymine (two hydrogen bonds) and between guanine and cytosine (three hydrogen bonds). Any other combination between nucleobases would prevent the two strands from forming a double helix. The *complementary bases* - adenine and thymine or guanine and cytosine - are always opposite each other (see Fig. 3). While adenine and guanine are purines, the two bases thymine and cytosine are pyrimidines. The former are larger than the latter and both kinds of molecules can only base pair with the opposing type of nucleobase. DNA plays a particularly important role in two processes - *replication* and *protein synthesis* (see Fig. 3).

During replication, the DNA is copied in order to be able to pass on the information it contains to daughter cells or descendants. The DNA strands are first separated from each other and then reassembled with complementary, newly synthesised strands so that two double helices are present at the end. During protein synthesis, the information contained in the DNA is read, understood and converted into amino acids, which then form a protein. This information is contained in so-called *genes*. The protein synthesis comprises the following steps. During transcription (see Fig. 3), ribonucleic acid (RNA) polymerase converts the genetic information of a gene into messenger RNA (mRNA). The mRNA is then read by the ribosome - a complex molecule - in reading frames that focus mainly on three nucleotides (codons) simultaneously. (Note that when passing from DNA to RNA the base thymine is replaced by uracil; however, in the sequel we will stay with thymine when talking about circular codes in the next sections). In the process, the ribosome assigns a matching transfer RNA (tRNA) molecule to each trinucleotide (codon) - see Fig. 4. Each of these molecules consists of an *anti-codon*, which is complementary to a specific codon, and a region in which an amino acid can bind. Recall that the anti-codon of a codon $N_1N_2N_3$ is built by reading it in reversed order and replacing each base N_i by its complementary base, e.g. the anti-codon of *ACG* is *CGT*. Clearly, if we read a codon on one strand of the DNA, then the anti-codon will be on the opposite strand (read in the opposite direction). After the ribosome has matched the codon of the mRNA strand with the anti-codon of the tRNA, the amino acid is taken off the molecule and attached to the amino acid chain already present. It is then moved to the next triplet on the mRNA strand. Finally, a *polypeptide chain* - a protein - is obtained. This process is called *translation*. The translation process is quite error-prone because it is usually very fast and is a highly complex mechanism. A potential error that can occur during translation occurs when the ribosome is inadvertently shifted by a few nucleotides, thereby changing the reading frame. If it shifts less or more than three nucleotides, different amino acids are combined to synthesise a different protein. This can lead to illness or disorders. We call this the *frame-shift problem*. In reality, however, these errors appear much less frequent than

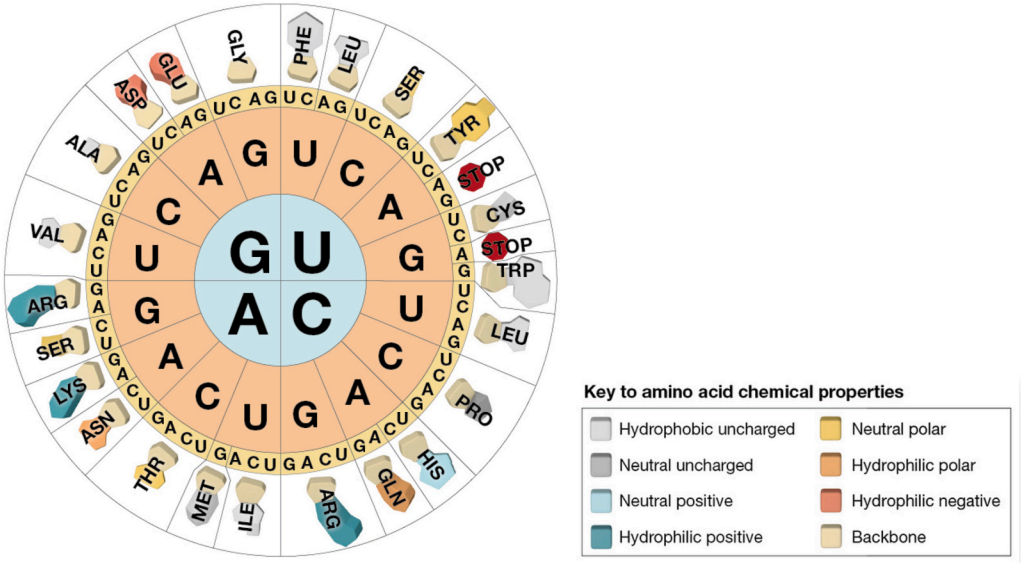

Fig. 2. The Genetic Code Table assigns to each codon a unique amino acid. The circles are to be read from in to out while the colours of amino acids refer to chemical properties as indicated in the table. (The picture is taken from http://learn.genetics.utah.edu/content/basics/dnacodes/). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

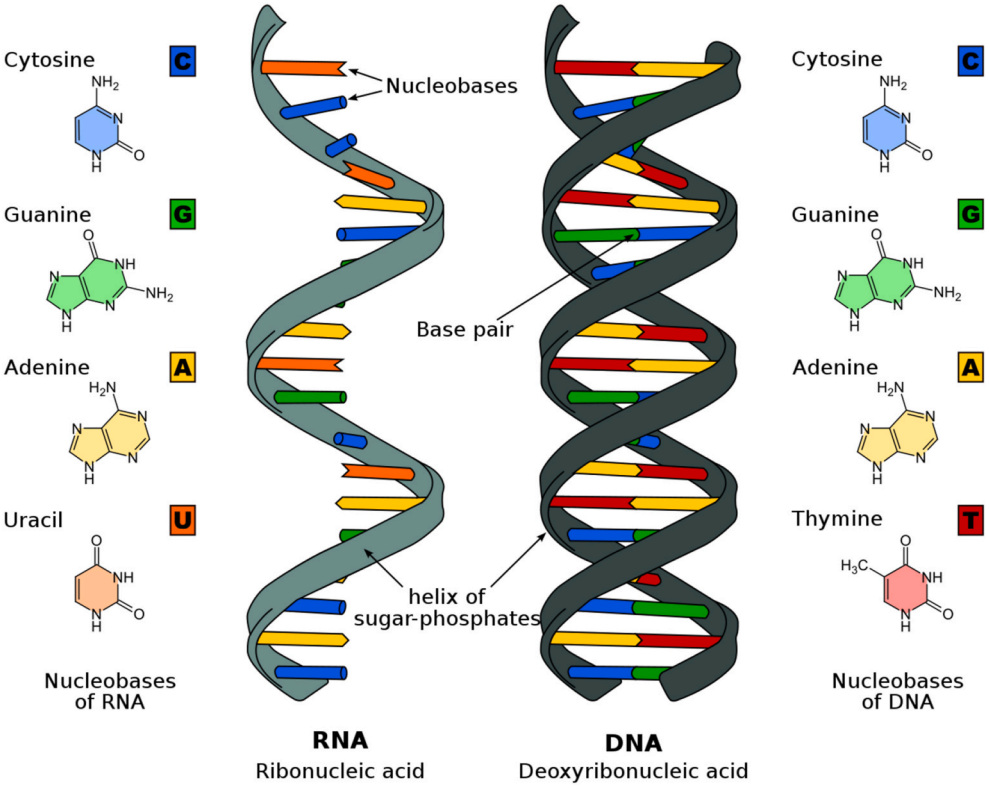

Fig. 3. Comparison of a single-stranded RNA and a double-stranded DNA with their corresponding nucleobases. Structure of the nitrogenous bases: adenine, guanine, cytosine, thymine, and uracil. (Picture taken from https://commons.wikimedia.org/wiki/File:Difference-DNA-RNA-EN.svg, license https://creativecommons.org/licenses/by-sa/3.0/deed.en).

they should theoretically occur (see (Johansson et al., 2008) or (Schmeing and Ramakrishnan, 2009)). Therefore, it is very likely that nature has implemented some mechanisms to avoid or even correct frame-shift errors. One of these mechanisms was demonstrated by Gupta and Singh (Gupta and Raj Singh, 2013), who provided evidence that stop codons occur more frequently in sequences modified by frame shifts, so that the translation is aborted prematurely. This saves energy and resources and prevents the formation of potentially toxic substances.

For the convenience of the reader Fig. 5 displays a possible frame-shift. In the correct reading frame the codons CTG, ATC, CAC, GAC and TTC are read. By the genetic code wheel (see Fig. 2) they correspond to the amino acids *Leu*, *Ile*, *His*, *Glu* and *Phe*. However, if the reading frame is shifted by one nucleotide the codons TGA, TCC, ACG, AGT are read and therefore the different sequence of amino acids *Stop*, *Ser*, *Thr* and *Ser* are obtained which shows the substantial impact and consequences of frame shifts.

4.2. Background on circular codes

As we have explained in the previous sections, a frame shift during the translation process of the genetic information by the ribosome seems to be a fatal error. The result would most likely be a non-functional protein that is a waste of energy and resources that nature normally tries to avoid. Already more than 60 years ago Crick et al. (Crick et al., 1957) proposed a solution to this problem when he suspected that only 20 trinucleotides (codons) among the 64 possible trinucleotides $\{AAA, .., TTT\}$ encode the 20 amino acids. His assumption was that only these codons appear in the appropriate reading frame - the *comma-free* property. This implied two main conditions for this *code* of 20 codons: Obviously the *periodic* trinucleotides AAA, CCC, GGG, TTT must be excluded, because their concatenation, e.g. AAA,AAA,AAA, allows them to appear in the reading frame 1 or 2, e.g. A,AAA,AAA,AA. In addition, at least one of two non-periodically *permuted* codons (i.e. two codons connected by a *circular permutation*, e.g. *ACG* and *CGA*) must also be excluded from such a code. The reason for this is that a concatenation of e.g. *ACG* with itself does not allow to retrieve the correct reading raster: there are two possible decompositions, namely $\cdots ACG, ACG, ACG \cdots$ in the correct frame and $\cdots A, CGA, CGA, CG \cdots$ in the first frame. Consequently, for each non-periodic codon $N_1N_2N_3$ at most one of the *circular permutation class* $\{N_1N_2N_3, N_2N_3N_1, N_3N_1N_2\}$ can be a member of a comma-free code. So if we exclude the four periodic codons and divide the remaining 60 trinucleotides into circular permutation classes of 20, we see that a comma-free code can have a size of 20 at most. Surprisingly, this number is identical to the amino acid number, and so the scientific community was excited about a comma-free code that assigns exactly one trinucleotide per amino acid without ambiguity. However, no comma-free trinucleotide code was statistically identified in the genes, and in the early 1960s Nirenberg and Matthaei (Nirenberg and Matthaei, 1961) discovered that the periodic codon *TTT* - which is excluded from a comma-free code - encodes phenylalanine (see Nirenberg and Matthaei (1961)). This led to the rejection of Crick's theory of comma-free codes as nature's method of finding the correct reading frame. Later, comma-free codes were taken up again (as codes over arbitrary alphabets and with any word length - see (Golomb et al., 1958; Scholtz, 1969; Tang et al., 1987; Levenshtein, 2004; Michel et al., 2008)), also in a biological context, especially in the context of the interaction between mRNA and tRNA (see (Golomb et al., 1958), (Crick et al., 1976) or (Shepherd, 1981) and (Eigen and Schuster, 1978) and many others.

It took until 1996 before a new theory for frame retrieval based on codes was developed. The theory of *circular* codes proposes that the ribosome uses a mechanism based on a *circular code* consisting of 20 trinucleotides for retrieving, maintaining and synchronising the reading frame as well as for coding amino acids (Arquès and Michel, 1996). As before, a circular code is able to determine the correct reading frame, however, not immediately but eventually after the ribosome has read a

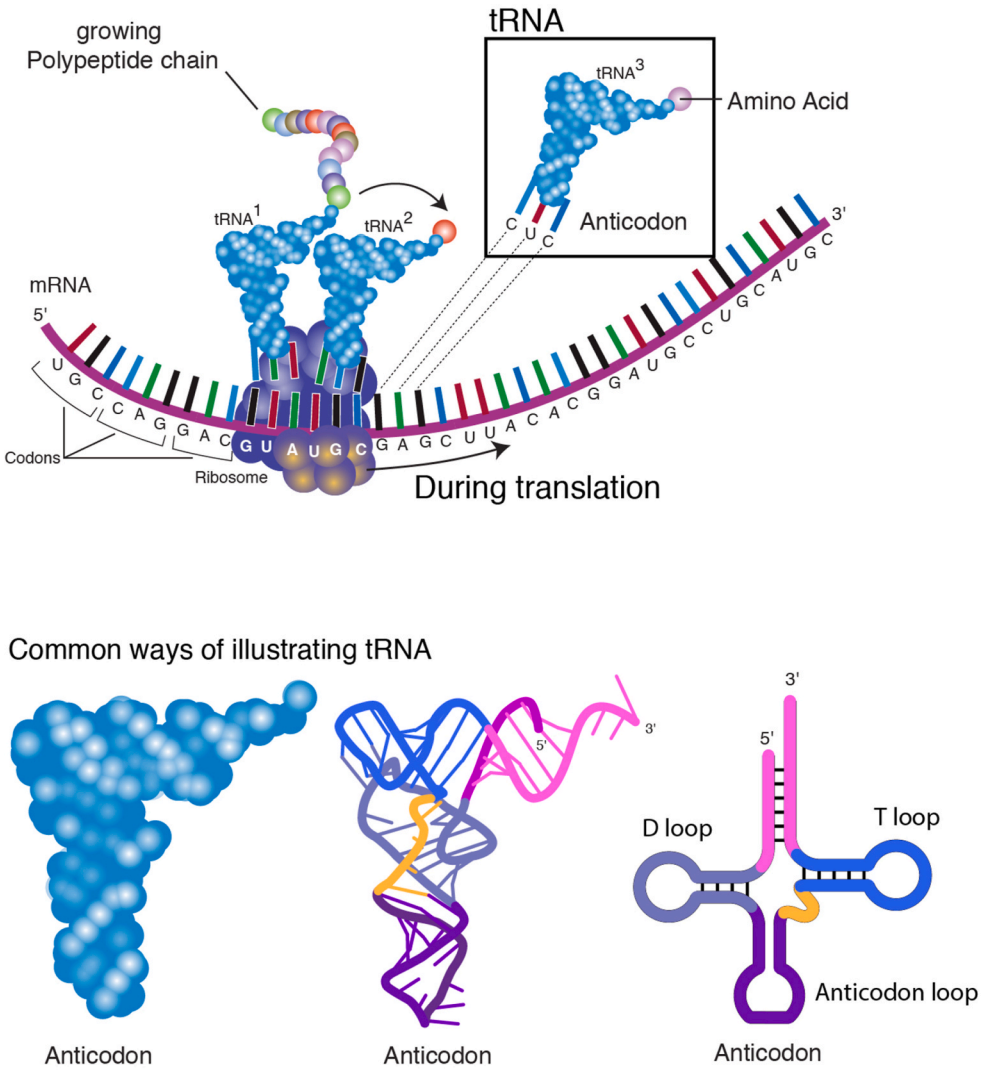

Fig. 4. The process of protein translation together with the structure of tRNAs. The latter attach to the mRNA via anticodon-codon interaction, carrying the amino acid coded by the codon in the mRNA. (Picture taken from https://www.genome.gov/genetics-glossary/Transfer-RNA, Courtesy: National Human Genome Research Institute).

sequence of at most 4 codons from the code. We will give the precise definition below but let us mention at this stage that obviously comma-freeness is the stronger property than circularity and that again the maximal size of a circular code can be at most 20 due to the same reasoning as above.

Definition 4.1. A trinucleotide code $X \subseteq \mathscr{B}^3$ is circular if any word over the alphabet $\mathscr{B} = \{A, C, G, T\}$ written on a circle has at most one decomposition into words from X. Here, written on a circle means that after the last letter the word starts again from its first letter.

For the convenience of the reader we first give an Example of a maximal circular code which is the first and most important example in the biological context (see Arquès and Michel (1996)).

Example 4.2. Let

$$X = \{AAC, AAT, ACC, ATC, ATT, CAG, CTC, CTG, GAA, GAC,$$

$$GAG, GAT, GCC, GGC, GGT, GTA, GTC, GTT, TAC, TTC\}$$

Then X is a maximal circular code.

Example 4.3. Let

$$X = \{CGT, ACG, TAC, GTA\}$$

Then X is not circular since the sequence $CGTACGTACGTA$ has two decompositions over X when read on a circle, namely $CGT|ACG|TAC|GTA$ and $C|GTA|CGT|ACG|TA$. Note that $TAC \in X$.

The code given in Example 4.2 even has the additional property that it does not only detect the correct reading frame in frame 0 but also in frames 1 and 2. This is to say that the shifted versions X_1 and X_2 of X are again circular codes. By shifted we mean that each codon of the code X is circular permuted by one or two bases:

$$X_1 = \{ACA, ATA, CCA, TCA, TTA, AGC, TCC, TGC, AAG, ACG,$$

$$AGG, ATG, CCG, GCG, GTG, TAG, TCG, TTG, ACT, TCT\}$$

and

$$X_2 = \{CAA, TAA, CAC, CAT, TAT, GCA, CCT, GCT, AGA, CGA,$$

$$GGA, TGA, CGC, CGG, TGG, AGT, CGT, TGT, CTA, CTT\}$$

This property is called the C^3-property. Moreover, X from Example 4.2 is self-complementary which means that with each codon the code also contains the corresponding anti-codon - a biologically important property since it reflects the structure of the DNA double-helix. The theory of circular codes initiated by Arques and Michel in 1996 (Arquès

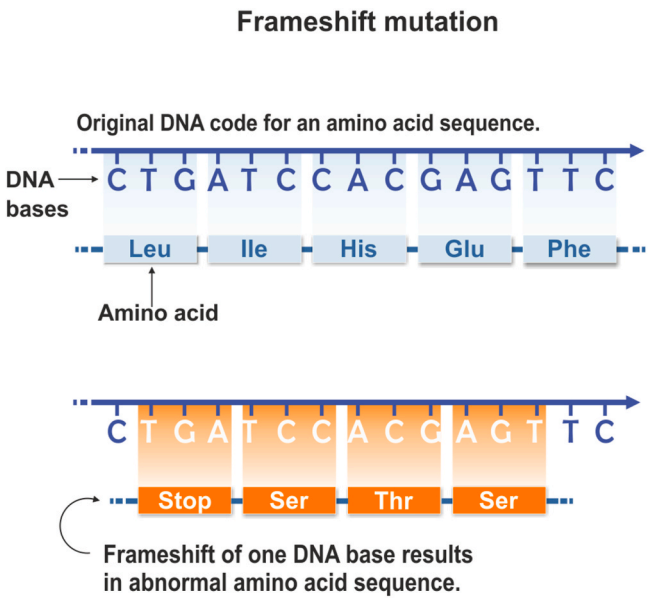

Fig. 5. Graphical representation of frameshifts (picture taken from Fimmel and Strüngmann (2018)). Codons in the correct reading-frame are marked in blue, the ones of the shifted reading-frame are displayed in orange. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and Michel, 1996) is based on the code X from Example 4.2. In fact, the set of 20 codons forming the code X from Example 4.2 was statistically identified in genes of bacteria, archaea, eukaryotes, plasmids and viruses (see Arquès and Michel, 1996; Michel, 2015, 2017) by collecting those codons that have the highest preference for the correct reading frame compared to the other two frames. Moreover, the finding of X circular code motifs (sequences of codons from X) in tRNAs and rRNAs and especially in the ribosome decoding center (see Michel, 2012; El Soufi and Michel, 2014, 2015) and in the genomes of eukaryotes (see El Soufi and Michel, 2016, 2017) supported the importance of circular codes in this biological setting. Finally, the universally conserved nucleotides A1492 and A1493 and the conserved nucleotide G530 in the ribosome decoding center are part of X circular code motifs. The recent paper (Michel, 2020) by Michel gives a nice summary of the statistical discovery of circular codes in genes (see also Michel, 2008).

By now there is an extensive literature on circular codes and their role in biology and especially genetic coding (see Fimmel and Strüngmann (2018) for a good survey). It turned out that there are exactly 216 maximal self-complementary circular C^3-codes, i.e. codes over the genetic alphabet $\{A, C, G, T\}$ having the same error-detecting properties as the code X from Example 4.2. These codes have been studied deeply using group theory, combinatorics and recently also graph theory. For instance, the 216 have been classified into 27 equivalence classes of size 8 each by applying a subgroup L of the symmetric group S_4 isomorphic to the symmetry group D_4 of a square. The permutations in L are exactly those that preserve the double-helix structure of the DNA (see Fimmel et al., 2014). Moreover, by assigning a graph to (circular) codes a beautiful and very comprehensive theory of circular codes over arbitrary finite alphabets and with finite word lengths has been developed (see e.g. Fimmel et al. (2019)).

Definition 4.4. Let $X \subseteq \mathscr{B}^3$ be a trinucleotide code. We define a directed graph $G(X) = (V(X), E(X))$ with set of vertices $V(X)$ and set of edges $E(X)$ as follows:

- $V(X) = \{\{N_1, N_2N_3, N_1N_2, N_3\} : N_1N_2N_3 \in X\}$
- $E(X) = \{\{[N_1, N_2N_3], [N_1N_2, N_3]\} : N_1N_2N_3 \in X\}$.

The graph $G(X)$ is called **graph associated** to X.

The very useful result states that $G(X)$ is *acyclic*, i.e. does not contain any circles, if and only if the code X is circular. For the convenience of the reader we close this section by displaying the graph associated to the important circular code X from Example 4.2 in Fig. 6.

5. Circular codes build on trees and forests

As we have seen in the previous sections the structure and theory of circular codes is of great interest in theoretical biology, in particular the theory of evolution of the genetic code and translation of the genetic information in the ribosome, since circular codes are supposed to play a significant role in maintaining the correct reading frame during the translation process. In this section it is therefore our aim to construct (infinite) circular codes over a finite alphabet Σ. The reader will see that we are going to use infinite binary sequences as they are used in forcing and also in the algebraic method introduced in Section 3. However, all results presented here are very basic, probably well-known and are not supposed to show evidence for a proper application of infinite combinatorics in biology; however, it is our intention to show that, due to the similarity, perhaps, in the future, it might be possible to apply methods from infinite combinatorics, set-theory and algebra to shed more light on processes and mechanisms that nature uses, e.g. in genetic coding.

Throughout this section let Σ be a finite alphabet and Σ^n the set of words over Σ of length n where $n \in \mathbb{N}$ is a natural number. Moreover, Σ^* is the set of all words of finite length over Σ including the empty word ε.

Recall the following definition which generalises Definition 4.1 to codes of arbitrary finite alphabets and to arbitrary word length. In particular it covers the case of *mixed* circular codes, i.e. codes where the words can have different lengths. Such mixed circular codes over the genetic alphabet $\mathscr{B} = \{A, C, G, T\}$ have been considered in Fimmel et al. (2019). In fact, in recent years several hypotheses about the origin of the current genetic code were formulated in which it was postulated that ancient amino acids were first encoded by dinucleotides or tetranucleotides rather than trinucleotides (see e.g. Baranov et al. (2009); Gonzalez et al. (2012), Patel (2005), Seligmann (2014), Wilhelm and Nikolajewa (2004), Wu et al. (2005). Since the number of amino acids needed to be encoded increased during evolution, such a coding process was necessary. However, at the same time, the genetic code still had to be robust against errors as e.g. point mutations. The precise mechanism

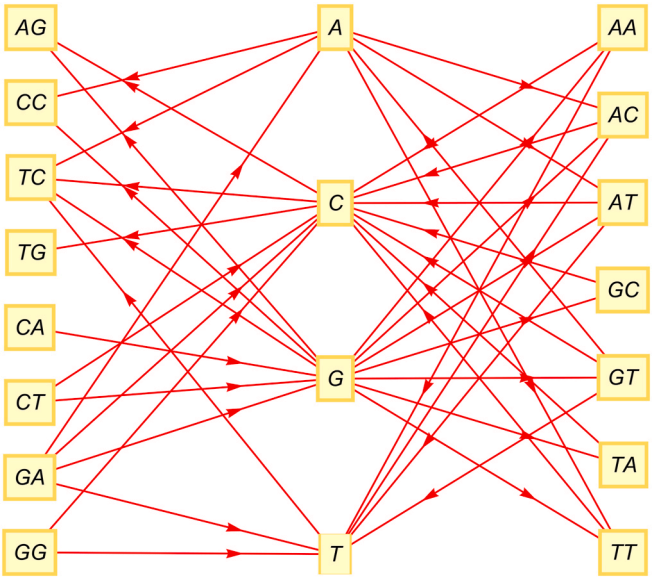

Fig. 6. Graph $G(X)$ of the maximal trinucleotide circular code X observed in genes of bacteria, eukaryotes, plasmids and viruses from Example 4.2. Picture taken from Fimmel et al. (2016).

for reading oligonucleotides can only be hypothesised, but in a transition phase, mixed circular codes which allowed the reading frame to retrieve words of different sizes, could have existed. A first obstacle is that in the mixed case it is not always true that any set of words forms a code as in the case of fixed word length.

Definition 5.1. A subset $X \subseteq \Sigma^*$ is called a **circular code** if it satisfies the following two conditions:

(i) X is a **code**, i.e. if $x_1 \cdots x_n = x_1' \cdots x_m'$ for some $x_i, x_j' \in X$ and $n,m \in \mathbb{N}$, then $n = m$ and $x_i = x_i'$ for all $i \leq n$;
(ii) X is **circular**, i.e. any concatenation $x_1 \cdots x_n \in X^n$ of words from X has no second decomposition over X when read on a circle.

Clearly, any set $X \subseteq \Sigma^n$ for some fixed n is a code but not necessarily circular (see the previous sections). However, there are sets $X \subseteq \Sigma^*$ that are not even a code. For Example, the mixed set $X = \{ACG, TGA, AC, GT, GA\}$ is not a code since the word $ACGTGA$ has two different decompositions into words from X, namely $ACG|TGA = AC|GT|GA$.

We are interested in constructing mixed (infinite) circular codes over Σ and ask the following question:

Question 5.2. Can we construct (in ZFC or some model of ZFC) an ω-sequence $\eta : \omega \to \Sigma$ such that the set of initial segments of η forms a circular code, i.e. such that $X = \{\eta \restriction_n : n \to \Sigma\}$ is a circular code? Can we even construct a system of such sequences so that the union of the sets of their initial segments is circular?

We will utilise so called ω-sequences over Σ in order to construct large mixed circular codes. As it is usual in cardinal arithmetic (see Section 3) we identify a natural number $n \in \mathbb{N}$ with its set of predecessors, i.e. we think of n as the set $n = \{0, \cdots, n-1\}$. A finite sequence $a_0 a_1 \cdots a_{n-1}$ of elements from Σ can therefore be thought of as a function $\eta : n \to \Sigma$ with $\eta(i) = a_i$ for any $i \in n$. The length $l(\eta)$ of such a sequence is the size of its domain which is n (see Section 3). By ${}^{<\omega}\Sigma$ we denote the set of all finite length sequences over Σ. Moreover, an infinite sequence of elements from Σ can be represented by an ω-sequence, i.e. by a function $\eta : \omega \to \Sigma$ where $\omega = \{0, 1, \cdots\}$. Given such an ω-sequence η we may restrict it to any element in its domain

$$\eta \restriction_n : n \to \Sigma$$

and call this function an initial segment of η (compare Section 3).

We now start with an Example in order to illustrate the idea.

Example 5.3. Let $\Sigma = \{0, 1\}$ be the binary alphabet and define the ω-sequence $\eta : \omega \to \Sigma$ by $\eta(0) = 1$ and $\eta(i) = 0$ for all $i \geq 1$. Finally, put

$$\Delta' = \{\eta \restriction_n | n \in \mathbb{N}\}.$$

Then $\Delta = \{Im(\mu) : \mu \in \Delta'\} \subseteq \Sigma^*$ is an infinite mixed circular code where $Im(\mu)$ is the image of the function μ.

Proof. First note that Δ is a mixed circular code if and only if Δ' is such a code viewed inside Σ^*. Clearly, Δ' consists of all the initial segments of η and hence is an infinite set consisting of exactly one finite function $\eta_n = \eta \restriction_n : n \to \Sigma$ for every $n \in \mathbb{N}$. Note that all these initial segments start with 1 and then continue with 0s. Thus, in any sequence of words from Δ' the positions of 1s determine exactly the unique decomposition of the word over Δ'. This shows that Δ' is a code. Moreover, the same holds for sequences of words over Δ' when read on a circle. Thus Δ' is a circular code as well. ∎

Obviously, the circular code constructed in the above Example 5.3 is very restricted since it contains exactly one element from each Σ^n. However, we can do better.

Proposition 5.4. *Let Σ be a finite alphabet of size at least 2. Then there is a mixed circular code $\Delta \subseteq \Sigma^*$ such that for every $n \in \mathbb{N}$ we have $|\Delta \cap \Sigma^n| = (|\Sigma| - 1)^{n-1}$.*

Proof. Since Σ has size at least 2 we may choose an element $a \in \Sigma$ and construct Δ as the following set

$$\Delta' = \{\eta \in {}^{<\omega}\Sigma | \eta(0) = a \text{ and } \eta(i) \neq a \text{ for all } 0 \neq i < lg(\eta)\}$$

Clearly, as in the Example 5.3 above it follows that any concatenation of words from Δ' is uniquely determined by the positions of the a's (no matter if considered on the line or on a circle). Hence Δ' is a circular code. Let $\Delta = \{Im(\mu) : \mu \in \Delta'\}$. By definition $\Delta \cap \Sigma^n$ is the set of all images of functions $\eta : n \to \Sigma$ that start with an a and that have no further a. Clearly, this set has size $(|\Sigma| - 1)^{n-1}$. ∎

In our final construction we try to improve the above construction in order to get mixed circular codes such that their intersection with Σ^n converges to the maximum for all $n \in \mathbb{N}$ large enough.

Theorem 5.5. *Let Σ be a finite alphabet of size at least 2 and let $1 > \varepsilon > 0$ be a real number. Then there is a mixed circular code $\Delta_\varepsilon \subseteq \Sigma^*$ such that for every $n \in \mathbb{N}$ we have $|\Delta_\varepsilon \cap \Sigma^n| \geq |\Sigma|^{n(1-\varepsilon)}$.*

Proof. Let Σ and ε be given as stated. Choose $a \in \Sigma$. We fix k big enough such that $\frac{1}{k} < \varepsilon$. For $n > k$ we define

$$\Delta'_n = \{\eta \in {}^n\Sigma | \eta(i) = a \text{ for all } i < k \text{ and } \eta(k) \neq a \text{ and}$$

$$\eta \text{ does not contain any other sequence of consecutive } a\text{'s of length } k\}$$

Let $\Delta'_\varepsilon = \bigcup_{n > k} \Delta'_n$. We claim that $\Delta_\varepsilon = \{Im(\mu) : \mu \in \Delta'_\varepsilon\}$ is as required.

First we need to show that Δ_ε is a circular code. The argument is as in Example 5.3 and Proposition 5.4. Any word in Δ_ε has a unique subword consisting of consecutive a's only that has length k. This subword can be seen as a marker and uniquely determines any decomposition of a concatenation of words from Δ_ε no matter if read on a line or a circle. Thus Δ_ε is a code and also circular. It remains to show that Δ_ε satisfies $\left|\Delta_\varepsilon \cap \Sigma^n\right| \geq \left|\Sigma\right|^{n(1-\varepsilon)}$. In order to show this, we define a subset of Δ'_n that gives a lower bound for the size of Δ'_n. Let

$$\Delta_n'' = \{\eta \in {}^n\Sigma | \eta(i) = a \text{ for all } i < k \text{ and } \eta(i) \neq a \text{ for all } k | i\}$$

Clearly, $\Delta_n'' \subseteq \Delta'_n$ and

$$\left|\Delta_n''\right| = \left|\Sigma\right|^{n-k} \cdot \left(1 - \frac{1}{|\Sigma|}\right)^{\frac{n}{k}}$$

The latter is approximately $\left|\Sigma\right|^{n \cdot \left(1 - \frac{1}{k}\right)} \geq \left|\Sigma\right|^{n \cdot (1-\varepsilon)}$ which finishes the proof. ∎

As we have seen the special choice of sequences $\eta : \omega \to \Sigma$ having certain combinatorial properties like a unique subword of length k consisting of a fixed letter only, can be utilised to construct large mixed circular codes - a fact/method that is not surprising to mathematicians or computer scientists. However, this is just a very basic construction but very likely, improving the combinatorial properties by using may lead to more complicated and useful codes.

6. Conclusion

In this paper we have asked if it might be possible to combine techniques from a rather sophisticated area of mathematics, the theory of forcing, with theoretical biology. We have merely made a first step on what may, or may not, be a successful journey, but hopefully this work will cause interested scientists from (mathematical or theoretical biology) or mathematics to sit up and think about a totally novel approach to some aspects of theoretical biology.

The main idea was to describe a technique called *forcing* from set theory, which allows one to construct new models of set theory in which all the standard axioms ZFC apply (Zermelo-Fraenkel axioms plus the

axiom of choice). The most famous Example of the use of forcing was Paul Cohen's proof that the Continuum Hypothesis is undecidable (see Section 2). The core of his proof was to construct new real numbers, i.e. infinite binary sequences, in a set-theoretical expansion model. In addition, it is possible to force these binary sequences to have prescribed combinatorial properties which can be useful in applications. Outside set theory, the powerful method of forcing has implied many independence results, including in Abelian group theory, and we have presented a construction technique for large complex algebraic objects based on trees and forests (see Section 3). The latter are constructed using binary sequences which could be obtained, for example, from a forcing extension.

Our intention was to make these theories easily accessible to readers coming from biology and who might not have any background in mathematics. Thus our focus was on presenting the material in an accessible fashion, illustrating theories with examples and historical background.

In the last two sections we have given some overview of the theory of circular codes used in genetic information and again have tried to make this accessible to non-expert readers - now coming from the mathematical side.

Circular codes have been observed statistically and are important in the system of genetic information processing because they describe a class of error-detecting codes that are able to maintain the correct reading frame during the translation process in the ribosome. We have given an introduction to the theory of circular codes (see Section 4) and in particular have shown how large mixed circular codes can be constructed using infinite binary sequences (see Section 5). The goal was to show that there are similarities between the three theories - forcing, complex Abelian groups and circular codes - in the sense that they all use binary sequences and their combinatorial properties. Thus it is imaginable that deeper (infinite) combinatorics might be applicable to problems arising from real questions from biology.

The authors hope that this paper may serve as a starting point for the application of infinite combinatorics and set theory in the field of theoretical biology and, beyond that, for the construction of an extended mathematical framework for the description of certain processes and mechanisms in nature - at least when it comes to some aspects of signal processing in nature.

Conflicts of interest

There is no conflict of interest.

Acknowledgement

This is paper 1209 in the first author's publication list. First author's research partially supported by Israel Science Foundation (ISF) grant no: 1838/19 and by NSF grant no: DMS 1833363.

References

Arquès, D.G., Michel, C.J., 1996. A complementary circular code in the protein coding genes. J. Theor. Biol. 182, 45–58.
Baranov, P.V., Venin, M., Provan, G., 2009. Codon size reduction as the origin of the triplet genetic code. PloS One 4, e5708.
Burgess, J.P., 1977. *Forcing*, Handbook of Math. Logic. North-Holland, pp. 404–452.
Cantor, G., 1879. Über unendliche lineare Punctmannichfaltigkeiten. Math. Ann. 15.
Cantor, G., 1895. Beiträge zur Begründung der transfiniten Mengenlehre. Math. Ann. 46, 481–512.
Cohen, P., 1963. The independence of the Continuum hypothesis. Proc. Natl. Acad. Sci. U.S.A. 50, 1143–1148.
Cohen, P., 2002. The discovery of forcing. Rocky Mt. J. Math. 32, 1071–1100.
Crick, F.H.C., Watson, J.D., 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171, 737738. https://doi.org/10.1038/171737a0.
Crick, F.H.C., Griffith, J.S., Orgel, L.E., 1957. Codes without commas. Proc. Natl. Acad. Sci. U.S.A. 43, 416421.
Crick, F.H., Brenner, S., Klug, A., Pieczenik, G., 1976. A speculation on the origin of protein synthesis. Orig. Life 7, 389397. https://doi.org/10.1007/BF00927934.
Dawson Jr., J.W., 1983. The published work of Kurt Gödel: an annotated bibliography. Notre Dame J. Formal Logic 24, 255–284.
Eigen, M., Schuster, P., 1978. The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. Naturwissenschaften 65, 341369. https://doi.org/10.1007/BF00439699.
Eklof, P.C., Mekler, A., 1990. Almost Free Modules; Set–Theoretic Methods. North–Holland.
El Soufi, K., Michel, C.J., 2014. Circular code motifs in the ribosome decoding center. Comput. Biol. Chem. 52, 917.
El Soufi, K., Michel, C.J., 2015. Circular code motifs near the ribosome decoding center. Comput. Biol. Chem. 59, 158176.
El Soufi, K., Michel, C.J., 2016. Circular code motifs in genomes of eukaryotes. J. Theor. Biol. 408, 198212.
El Soufi, K., Michel, C.J., 2017. Unitary circular code motifs in genomes of eukaryotes. Biosystems 153, 45–62.
Fimmel, E., Giannerini, S., Gonzalez, D.L., Strüngmann, L., 2014. Circular codes, symmetries and transformations. J. Math. Biol. 70, 16231644.
Fimmel, E., Michel, C.J., Strüngmann, L., 2016. n-nucleotide circular codes in graph theory. Phil. Trans. Roy. Soc. Lond. 374, 20150058.
Fimmel, E., Strüngmann, L., 2018. Mathematical fundamentals for the noise immunity of the genetic code. Biosystems 164, 186–198.
Fimmel, E., Michel, C.J., Pirot, F., Sereni, J.-S., Strngmann, L., 2019. Mixed circular codes. Math. Biosci. 317, 108–231. https://doi.org/10.1016/j.mbs.2019.108231.
Göbel, R., Trlifaj, J., 2012. Approximations and endomorphism algebras of modules. De Gruyter Expo. Math. 41 https://doi.org/10.1515/9783110218114.
Gödel, K., 1931. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. Monatshefte für Mathematik und Physik 38, 173–198.
Gödel, K., 1938. Consistency of the axiom of choice and the generalized Continuum hypothesis. Proc. Acad. Nat. Sci. 24, 556–557.
Golomb, S.W., Gordon, B., Welch, L.R., 1958. Comma-free codes. Can. J. Math. 10, 202209.
Gonzalez, D., Giannerini, S., Rosa, R., 2012. On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. Nat. Proc. https://doi.org/10.1038/npre.2012.7136.
Gupta, A., Raj Singh, T., 2013. SHIFT: server for hidden stops analysis in frame-shifted translation. BMC Res. Notes 6. https://doi.org/10.1186/1756-0500-6-68.
Hilbert, D., 1901. Mathematische Probleme. Arch. Math. Phys. 1, 44–63.
Human Genome Project Completion: Frequently Asked Questions. National Human Genome Research Institute (NHGRI). www.genome.gov/human-genome-project.
Johansson, M., Bouakaz, E., Lovmar, M., Ehrenberg, M., 2008. The kinetics of ribosomal peptidyl transfer revisited. Mol. Cell 30, 589598.
Kunen, K., 1980. *Set Theory - an Introduction to Independent Proofs*, Studies in Logic and the Foundations of Mathematics. North Holland, p. 102.
Levenshtein, V.I., 2004. Combinatorial problems motivated by comma-free codes. J. Combin. Des. 12 (3), 184196.
Löwenheim, L., 1915. ber Mglichkeiten im Relativkalkl. Math. Ann. 76, 447–470.
Michel, C.J., 2008. A 2006 review of circular codes in genes. Comput. Math. Appl. 55, 984–988.
Michel, C.J., 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput. Biol. Chem. 37, 2437.
Michel, C.J., 2015. The maximal C^3self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J. Theor. Biol. 380, 156–177.
Michel, C.J., 2017. The maximal C^3self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. Life 7 (20), 1–16.
Michel, C.J., 2020. The maximality of circular codes in genes statistically verified. Biosystems 197, 104201.
Michel, C.J., Pirillo, G., Pirillo, M.A., 2008. Varieties of comma free codes. Comput. Math. Appl. 55, 989–996.
Nirenberg, M.W., Matthaei, J.H., 1961. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. U.S.A. 47, 15881602. https://doi.org/10.1073/pnas.47.10.1588.
Patel, A., 2005. The triplet genetic code had a doublet predecessor. J. Theor. Biol. 233, 527–532.
Russell, B., 1903. The Principles of Mathematics, p. 100. Cambridge.
Shelah, S., 1974. Infinite abelian groups. Whitehead problem and some constructions. Isr. J. Math. 18, 243–325.
Shelah, S., 1998. *Proper and Improper Forcing*, Perspectives in Mathematical Logic. Springer Verlag.
Shelah, S., Strüngmann, L., 2003. It is consistent with ZFC that -B_1 groups are not ,B_2. Forum Math. 15, 507–524.
Seligmann, H., 2014. Putative anticodons in mitochondrial tRNA sidearm loops: pocketknife tRNAs? J. Theor. Biol. 340, 155–163.
Schmeing, T.M., Ramakrishnan, V., 2009. What recent ribosome structures have revealed about the mechanism of translation. Nature 461, 12341242.
Scholtz, R., 1969. Maximal and variable word-length comma-free codes. IEEE Trans. Inf. Theor. 15 (2), 300306.
Shepherd, J.C.W., 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. U.S.A. 78, 1596–1600.

Skolem, T., 1920. *Logisch-kombinatorische Untersuchungen über die Erfüllbarkeit oder Beweisbarkeit mathematischer Sätze nebst einem theorem über dichte Mengen*, Videnskapsselskapets skrifter, I. Matematisk-naturvidenskabelig klasse.

Skolem, T., 1922. *Einige Bemerkungen zur Begründung der Mengenlehre*, Matematikerkongressen i. Helsingsfors den 4-7 Juli 1922 Den femte skandinaviska matematikerkongressen, pp. 217–232. Redogoerelse.

Tang, B., Golomb, S.W., Graham, R.L., 1987. A new result on comma-free codes of even word-length. Can. J. Math. 39 (3), 513526.

Wilhelm, T., Nikolajewa, S., 2004. A new classification scheme of the genetic code. J. Mol. Evol. 59, 598–605.

Wu, H.L., Bagby, J.M., van den Elsen, S., 2005. Evolution of the genetic triplet code via two types of doublet codons. J. Mol. Evol. 61, 54–64.

Zermelo, E., 1908. Untersuchungen über die Grundlagen der Mengenlehre. I. Math. Ann. 65, 261–281.