GPH-International Journal of Mathematics
(Online ISSN 3050-9645)
Volume 09 | Issue 10 | 2026
Published by: Global Publication House
DOI: https://doi.org/10.5281/zenodo.24032800
Journal homepage: https://gphjournal.org/index.php/m/index

Skew-Cyclic and Constacyclic Codes over Finite Chain Rings for DNA-Based Data Storage and Error Correction

Peter Guchu Muiruri
Zetech University
peter.muiruri@zetech.ac.ke

Abstract:

DNA has considerable potential for long-term data storage, but the process of synthesising, preserving, and sequencing DNA strands can introduce substitutions, insertions, and deletions. In addition, biochemical requirements restrict the sequences that can be used reliably. This paper develops a structured framework for constructing error-correcting DNA codes from skew-cyclic and skew-constacyclic codes over finite chain rings. We first select a chain ring whose additive structure can represent the four DNA nucleotides and define a nucleotide map under which Watson–Crick complementation corresponds to addition by a fixed ring element. We then introduce the skew polynomial ring $R[x;\theta]$ and describe skew λ-constacyclic codes as left ideals of $R[x;\theta]/\langle x^n-\lambda\rangle$. Their generator representations and cardinalities are presented alongside sufficient conditions for reversibility, reverse-complement closure, and balanced GC-content. Error performance is considered through Lee weight and a distance-preserving Gray map. Finally, we outline an end-to-end DNA storage workflow and discuss illustrative constructions: a length-four code over $\mathbb{F}_2+u\mathbb{F}_2$ and a genuinely skew construction over $\mathbb{F}_4+u\mathbb{F}_4$ that demonstrates non-unique factorisation. The examples clarify the algebraic framework; practical biochemical performance remains a matter for further testing.

Keywords:

DNA codes; skew-cyclic codes; skew-constacyclic codes; finite chain rings; Gray map; Lee distance; DNA data storage

1 Introduction

DNA is being investigated as a medium for archival storage because of its potential storage density and longevity. Demonstrations of storing digital information in synthesised DNA have established the feasibility of the general approach [5, 6, 7]. However, practical systems must contend with base substitutions, strand loss (erasures), and sequence-dependent biochemical constraints. Algebraic coding theory offers a systematic way to construct code families with analysable minimum distance and structured encoding procedures.

Cyclic codes over GF(4), as well as codes over rings such as $\mathbb{F}_2[u]/\langle u^2\rangle$, have been used in DNA-code constructions because reversal and reverse-complement conditions can often

© 2026 Author(s). Published by Global Publication House (GPH).
This is an Open Access article distributed under the terms of the CC BY 4.0 License.

be expressed algebraically [8, 9, 10]. Skew polynomial rings extend the range of available constructions. Their non-commutative multiplication can produce factorisations of $x^n - \lambda$ that are unavailable in the corresponding commutative setting [2, 3]. This paper brings these approaches together in a reproducible construction framework.

The paper makes four main contributions:

(1) It defines a nucleotide map for which complementation is represented by addition of a fixed ring element.

(2) It summarises generator representations and cardinality formulas for skew λ-constacyclic codes over finite chain rings.

(3) It discusses algebraic conditions and practical filters for reversibility, reverse-complement closure, GC-content, and run length.

(4) It sets out an encode–synthesise–sequence–decode workflow and illustrates the method with worked examples.

2. Preliminaries

2.1 The chain ring and nucleotide map

A finite chain ring is a finite commutative ring whose ideals are linearly ordered by inclusion. One commonly studied family is $R = \mathbb{F}_{p^m}[u]/\langle u^k \rangle$. To represent four nucleotides with four ring elements, we use $R = \mathbb{F}_2 + u\mathbb{F}_2$, where $u^2 = 0$ and $|R| = 4$.

The nucleotide correspondence is

$$0 \leftrightarrow \text{A}, \quad 1 \leftrightarrow \text{C}, \quad u \leftrightarrow \text{G}, \quad 1 + u \leftrightarrow \text{T}$$

Under this map, adding $1 + u$ implements Watson–Crick complementation: $0 + (1 + u) = 1 + u$ (A→T), $1 + (1 + u) = u$ (C→G), $u + (1 + u) = 1$ (G→C), and $(1 + u) + (1 + u) = 0$ (T→A). Thus, for $\mathbf{c} \in R^n$, complementation is the translation $\mathbf{c} \mapsto \mathbf{c} + (1 + u)\mathbf{1}$, where $\mathbf{1}$ denotes the all-ones vector.

For a larger alphabet, consider $R = \mathbb{F}_4 + u\mathbb{F}_4$, which has 16 elements. As an additive group, it is isomorphic to $(\mathbb{F}_2 + u\mathbb{F}_2)^2$. Write $a = a_0 + a_1\omega$ and $b = b_0 + b_1\omega$, where $\omega^2 = \omega + 1$. The element $a + ub$ can then be mapped to the dinucleotide $(a_0 + ub_0, a_1 + ub_1)$. Under this representation, complementation of a dinucleotide corresponds to addition of $(1 + u)\omega^2 = (1 + u)(1 + \omega)$.

2.2 Lee weight and the Gray map

Define the Gray map $\varphi: R \to \mathbb{F}_2^2$ by $\varphi(a + ub) = (b, a + b)$. It maps 0, 1, u, and $1 + u$ to 00, 01, 11, and 10, respectively. The Lee weight is the Hamming weight of the corresponding Gray image: $\text{wt}_L(0) = 0$, $\text{wt}_L(1) = 1$, $\text{wt}_L(u) = 2$, and $\text{wt}_L(1 + u) = 1$. Applied coordinatewise, φ preserves distance between R^n with Lee distance and $\mathbb{F}_2^{2n}$ with Hamming distance, and maps R-linear codes to binary linear codes.

3. Skew polynomial rings and skew-constacyclic codes

3.1 The ring $R[x;\theta]$

Let θ be an automorphism of R. The skew polynomial ring $R[x;\theta]$ consists of polynomials in x with coefficients in R. Addition is defined as usual, while multiplication is governed by $xa = \theta(a)x$. The ring is non-commutative whenever θ is not the identity.

The ring choice matters. The only automorphism of $\mathbb{F}_2 + u\mathbb{F}_2$ is the identity: an automorphism fixes $\mathbb{F}_2$ and must send u to a non-zero square-zero element, which can only be u. Consequently, a genuinely skew construction requires a larger ring, for example $R = \mathbb{F}_{2^m} + u\mathbb{F}_{2^m}$ with $m \geq 2$, and an automorphism such as a power of the Frobenius map extended by $\theta(u) = u$. The smaller ring remains useful for the commutative examples considered below.

3.2 Definition

Let λ be a unit of R satisfying $\theta(\lambda) = \lambda$, and let n be a positive integer. If the order of θ divides n, then $\langle x^n - \lambda\rangle$ is a two-sided ideal of $R[x;\theta]$, so the quotient $R[x;\theta]/\langle x^n - \lambda\rangle$ is well-defined.

Definition 1. A *skew λ-constacyclic code* of length n over R is a left ideal of $R[x;\theta]/\langle x^n - \lambda\rangle$. When $\lambda = 1$, the code is *skew-cyclic*. Equivalently, the code is an R-submodule of R^n that is invariant under the skew constacyclic shift:

$$(c_0, \dots, c_{n-1}) \mapsto \big(\theta(\lambda c_{n-1}), \theta(c_0), \dots, \theta(c_{n-2})\big)$$

Because $x^n = \lambda$ in the quotient, multiplication by x makes $\theta(c_{n-1})\lambda$ the new constant coefficient. Since $\theta(\lambda) = \lambda$, this agrees with $\theta(\lambda c_{n-1})$.

4. Structure and generators

Factor $x^n - \lambda$ in $R[x;\theta]$. In the skew setting, factorisation need not be unique; different right divisors can therefore generate distinct codes of the same length.

Reduce the factorisation modulo u and lift the resulting factors using a Hensel-type procedure.

Generator representation. The manuscript uses the form $C = \langle g_0, ug_1\rangle$, with $g_1 \mid g_0 \mid (x^n - \lambda)$. The precise meaning of divisibility should be stated consistently in the skew-polynomial setting.

Cardinality. For a single monic generator g of degree r dividing $x^n - \lambda$ over $\mathbb{F}_2 + u\mathbb{F}_2$, the stated formula is $|C| = 4^{n-r}$. For $C = \langle g_0, ug_1\rangle$ over $\mathbb{F}_2 + u\mathbb{F}_2$, the stated formula is $|C| = 2^{2n-\deg(g_0)-\deg(g_1)}$. More generally, a monic right divisor of degree r generates a free code of size $|R|^{n-r}$.

Generator matrix. A generator matrix can be formed from the coefficient vectors of the appropriate shifts of $g(x)$, namely $g(x), xg(x), \dots, x^{n-r-1}g(x)$. In a skew setting, the shift convention and coefficient action must be applied consistently.

5. Incorporating DNA constraints

Reversibility. A code is reversible if reversing any codeword produces another codeword in the same code. A common route is to impose a suitable (skew) self-reciprocal condition on the generator, up to multiplication by a unit. The exact criterion depends on θ and λ, so each proposed construction should be checked directly, for example by examining its generator matrix.

Reverse-complement closure. If a code is reversible and contains the all-$(1+u)$ vector, then its additive structure makes it closed under complementation. Together, these properties imply closure under reverse-complementation. This is an important design condition for DNA sequences because it helps control interactions between a strand and reverse complements in the code library.

GC-content and sequence filters. Under the nucleotide map above, GC-content is the proportion of coordinates whose symbols are 1 or u, corresponding to C or G. A target such as 50% GC-content can be enforced by retaining only codewords with the required GC-count. The resulting subset is generally not linear, although it may still be closed under reverse-complementation, as illustrated in Example 1.

Additional filters should remove sequences with long homopolymer runs and sequences likely to form problematic secondary structures, including hairpins. These checks are applied to the DNA sequences obtained from the algebraic codewords.

6. Error-correcting capability

Determine the minimum Lee distance $d_L(C)$. Under the Gray map, this equals the minimum Hamming distance of the binary image.

Estimate or calculate the minimum distance using a BCH-type bound for skew-constacyclic codes, a Singleton-type bound, or exhaustive enumeration for parameter sizes that are computationally manageable. Tools such as Magma or a short Python program may be used for enumeration.

If the minimum distance is d, the code corrects up to $\lfloor (d-1)/2 \rfloor$ errors, detects up to $d-1$ errors, and corrects up to $d-1$ erasures. Erasure correction is particularly relevant to DNA storage, where a strand or individual positions may be lost or unreadable.

7. DNA storage workflow

(1) **Convert the source data into ring symbols.** Split the binary stream into two-bit blocks and use $00 \mapsto 0$, $01 \mapsto 1$, $10 \mapsto u$, and $11 \mapsto 1+u$.

(2) **Encode each block.** Represent the $k = n - r$ message symbols by a polynomial $a(x)$ and form $c(x) = a(x)g(x)$, producing a codeword of length n.

(3) **Map the codeword to DNA.** Convert each ring symbol to its nucleotide, apply the sequence filters described in Section 5, and add indexing and primer regions as required by the synthesis platform.

(4) **Synthesise** the DNA strands, store them, and sequence them when retrieval is required.

(5) **Decode the reads.** Align the reads and form a consensus, correct errors using syndrome decoding or handle missing positions through erasure decoding, and convert the recovered ring symbols back into bits.

8. Application examples

Example 1: A length-four reverse-complement code over $\mathbb{F}_2 + u\mathbb{F}_2$

Let $R = \mathbb{F}_2 + u\mathbb{F}_2$, $n = 4$, $\theta = \mathrm{id}$, and $\lambda = 1$. Over this ring, $x^4 - 1 = (x+1)^4$, and $g(x) = (x+1)^2 = x^2 + 1$ is a monic divisor of degree two. Consider $C = \langle g \rangle$.

Size and structure. The code has $|C| = 4^{4-2} = 16$ codewords. Every codeword can be written as $a(x)(x^2+1)$, where $a(x) = a_0 + a_1 x$. Thus

$$C = \{(a, b, a, b) : a, b \in R\}$$

Distance. Choosing $a = 1$ and $b = 0$ gives the codeword $(1, 0, 1, 0)$, whose Lee weight is two. No non-zero codeword has smaller Lee weight, so $d_L = 2$; the Hamming distance is also two. The binary image is a $[8, 4, 2]$ code. It detects any single substitution and corrects any single erasure, but it cannot correct a substitution by itself. This example is intended to illustrate the construction rather than serve as a practical storage code.

Reversibility and complementation. Reversing (a, b, a, b) gives (b, a, b, a), which is also in C, so the code is reversible. The all-$(1+u)$ vector is a multiple of g and therefore belongs to C. Consequently, the code contains TTTT and is closed under complementation and reverse-complementation.

GC-content. A word (a, b, a, b) has 50% GC-content when one of a and b represents C or G and the other represents A or T. The resulting eight words are CACA, CTCT, GAGA, GTGT, ACAC, TCTC, AGAG, and TGTG. This subset is closed under reversal and complementation and contains no homopolymer runs. It provides eight sequences, equivalent to three bits per four-nucleotide oligo if the subset is used as the codebook.

Example 2: Encoding, substitution detection, and dropout

Encoding. The message bits 01 00 correspond to $a_0 = 1$ (C) and $a_1 = 0$ (A). The resulting codeword is $(1, 0, 1, 0)$, which maps to CACA. It has 50% GC-content and no homopolymer run. Its reverse complement is TGTG; the relationship between this sequence and the selected balanced subset should be checked carefully when specifying the library's hybridisation constraints.

Substitution detection. If sequencing returns CCCA, the repeated-coordinate condition $c_0 = c_2$ is violated, so the discrepancy is detected. Since $d = 2$, however, the error cannot necessarily be corrected uniquely.

Erasure correction. If the third position is missing, the read is CA?A. Because each codeword satisfies $c_0 = c_2$, the missing symbol is C and the original sequence CACA is recovered, giving the message 01 00. This illustrates erasure recovery even though the code does not correct a single substitution.

A less suitable message. The message 01 10 maps to $(1, u, 1, u)$, or CGCG. Although it is a valid word in C, it has 100% GC-content and is self-reverse-complementary under the stated mapping. This example illustrates why algebraic membership alone is insufficient: sequence-level filters are also required.

Example 3: A genuinely skew construction over $\mathbb{F}_4 + u\mathbb{F}_4$

Let $R = \mathbb{F}_4 + u\mathbb{F}_4$, where $\mathbb{F}_4 = \{0, 1, \omega, \omega^2\}$ and $\omega^2 = \omega + 1$. Define θ by $\theta(\omega) = \omega^2$ and $\theta(u) = u$. This automorphism has order two, so choose $n = 2$ and $\lambda = 1$; then $\langle x^2 - 1 \rangle$ is a two-sided ideal.

Non-unique factorisation. In $R[x; \theta]$,

$$(x + a)(x + b) = x^2 + (\theta(b) + a)x + ab$$

Choosing $a = \theta(b)$ makes the coefficient of x equal to zero, since the characteristic is two. For $b \in \mathbb{F}_4^{\times}$, the constant term is $\theta(b)b = b^3 = 1$. Hence

$$x^2 + 1 = (x + 1)(x + 1) = (x + \omega^2)(x + \omega) = (x + \omega)(x + \omega^2)$$

The three degree-one right divisors $x + 1$, $x + \omega$, and $x + \omega^2$ therefore give three distinct skew-cyclic codes of this length.

One code. Take $g = x + \omega$ and $C = \langle g \rangle$. This is a free code of size $|R|^{2-1} = 16$, with codewords $a \cdot (x + \omega) = (a\omega, a)$, for $a \in R$. The skew shift $(c_0, c_1) \mapsto \big(\theta(c_1), \theta(c_0)\big)$ sends $(a\omega, a)$ to $\big(\theta(a), \omega^2\theta(a)\big)$, which can be written as $(a'\omega, a')$ with $a' = \omega^2\theta(a)$. Thus C is invariant under the shift. Every non-zero codeword has two non-zero coordinates, giving Hamming distance two; as a free rank-one code, it meets the Singleton bound $n - k + 1 = 2$.

DNA representation. Under the dinucleotide map in Section 2.1, each ring coordinate is represented by a two-nucleotide block. The resulting oligo has length four nucleotides and represents four bits. Longer even lengths $n = 2m$, with $m \geq 2$, provide a setting in which reversibility and GC-balance constraints can be investigated in more detail.

Potential application settings

The following are proposed application directions rather than experimentally demonstrated outcomes:

- **Archival storage:** inner skew-constacyclic codes could protect individual strands, while an outer code across strands could address complete strand loss.
- **Hybridisation-aware oligo libraries:** reverse-complement closure and controlled GC-content could help produce libraries with more consistent melting behaviour and reduced cross-hybridisation, including for DNA computing and barcode design.
- **Sequencing with dropout:** erasure correction can recover missing positions when the number of erasures is within the code's capability, as illustrated in Example 2.
- **Choice of metric:** Lee distance may be useful when the error model assigns particular significance to substitutions associated with nearby Gray-map images.

9. Discussion and limitations

A genuinely skew construction requires a non-identity automorphism. The ring must therefore be larger than $\mathbb{F}_2 + u\mathbb{F}_2$, and the code length must satisfy the condition involving the order of θ.

Reversibility is more delicate for skew codes than for ordinary cyclic codes. It should be verified for each generator rather than inferred from the code family alone.

The Gray map provides a useful treatment of substitution errors, but it does not by itself address insertions and deletions. Depending on the synthesis and sequencing technology, synchronisation codes and sequence-alignment methods may also be needed.

GC-balanced subsets are generally non-linear. A decoder for the full algebraic code may therefore need a further membership test to ensure that the recovered word satisfies the required DNA constraints.

The analysis is algebraic. The biochemical behaviour of the proposed sequences must be assessed separately through suitable experiments or simulations.

10. Conclusion

This paper has presented a framework for constructing DNA codes from skew-cyclic and skew-constacyclic codes over finite chain rings. The framework connects the choice of ring and nucleotide map with code generators, distance analysis, DNA-specific constraints, and a practical storage workflow. The examples illustrate both a small commutative construction and a genuinely skew construction with non-unique factorisation. Further work should classify reversible skew codes over $\mathbb{F}_{2^m} + u\mathbb{F}_{2^m}$, identify codes with minimum distance at least three that also satisfy the relevant DNA constraints, and develop methods that address insertion and deletion errors.

References

1. H. Q. Dinh and S. R. López-Permouth, Cyclic and negacyclic codes over finite chain rings, *IEEE Transactions on Information Theory*, 50(8), 2004.
2. D. Boucher, W. Geiselmann, and F. Ulmer, Skew-cyclic codes, *Applicable Algebra in Engineering, Communication and Computing*, 18, 2007.
3. D. Boucher and F. Ulmer, Coding with skew polynomial rings, *Journal of Symbolic Computation*, 44, 2009.
4. I. Siap, T. Abualrub, and N. Aydin, Skew cyclic codes of arbitrary length, *International Journal of Information and Coding Theory*, 2, 2011.
5. G. M. Church, Y. Gao, and S. Kosuri, Next-generation digital information storage in DNA, *Science*, 337, 2012.
6. N. Goldman et al., Towards practical, high-capacity, low-maintenance information storage in synthesized DNA, *Nature*, 494, 2013.
7. Y. Erlich and D. Zielinski, DNA Fountain enables a robust and efficient storage architecture, *Science*, 355, 2017.
8. P. Gaborit and O. D. King, Linear constructions for DNA codes, *Theoretical Computer Science*, 334, 2005.
9. T. Abualrub, A. Ghrayeb, and X. Zeng, Construction of cyclic codes over GF(4) for DNA computing, *Journal of the Franklin Institute*, 343, 2006.
10. B. Yildiz and I. Siap, Cyclic codes over $\mathbb{F}_2[u]/(u^4 - 1)$ and applications to DNA codes, *Computers & Mathematics with Applications*, 63, 2012.