

* Archival DNA Storage Model

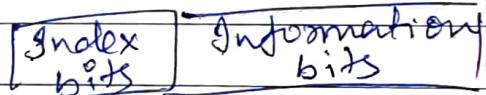
Let $\Sigma_{DNA} = \{A, T, G, C\}$

$C_{DNA}(n, M, d) \subseteq \Sigma_{DNA}^n$ as DNA code.

Objective : • Optimal DNA code with max min dist d (to ~~more~~ correct errors).

- Max storage capacity achieving codes.
- Design chunk architecture of DNA blocks

→ run-length coding?

Chunk Architecture : 

Length of chunk = 15

No. of chunks = 54898

Error Correction = No

quadratic residues $x^2 \equiv q \pmod{n}$

Page No.:
Date: / /

eg : Refer to slides.

* Goldman Model



chunk

Architecture.

- Ternary Huffman codes were used.
- map such that no two consecutive A, C, G, T repeats occurs.

* DNA Golay Subcode Model → Can we use this?

$(11, 256, 5)_3$ was used (n, M, d)
ternary Golay subcode

- 2 bit flip error correction.

= 115 EB per gram of DNA achieved.

Extended Golay code $[24, 12, 8]$

0 1 2 3 4 5 6 7 8 9 10

$G_1 = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$

g where
quadratic
residues
cyclic
shifts

11 times

0 1

Ternary Golay Code $G = [I_6 | B]$

$$B = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 & 2 & 1 \\ 1 & 1 & 0 & 1 & 2 & 2 \\ 1 & 2 & 1 & 0 & 1 & 2 \\ 1 & 2 & 2 & 1 & 0 & 1 \\ 1 & 1 & 2 & 2 & 1 & 0 \end{bmatrix}_3$$

Punchure one column we get $[11, 6, 5]_3$

729 code words

$\rightarrow x_{\text{DNA}} \in C_{\text{DNA}}$ sent

y_{DNA} received

error

If $d_H(x_{\text{DNA}}, y_{\text{DNA}}) = t$ & corresponding

Ternary code obtained using Goldman's base Table 10 are x & y then

$$t < d_H(x, y) \leq 2t$$

Eg: $x_{\text{DNA}} = \text{GTCTCGTCGTC}$ $x = 10111001001$

$y_{\text{DNA}} = \text{GACTCGTATGTC}$ $y = 11011000101$

$$d_H(x_{\text{DNA}}, y_{\text{DNA}}) = 2 \quad d_H(x, y) = 4$$

$x_{\text{DNA}} \leftarrow c_{\text{DNA}}$ sent

$y_{\text{DNA}} \leftarrow c_{\text{DNA}}$ received

Case 1: 1 bit flip in DNA $d_H(x_{\text{DNA}}, y_{\text{DNA}}) = 1$

$\Rightarrow d_H(x, y) = 2 \rightarrow$ can be resolved

Case 2: $d_H(x_{\text{DNA}}, y_{\text{DNA}}) = 2$

$\Rightarrow d_H(x, y) \leq 4$

Let $w_{\text{DNA}} \neq x_{\text{DNA}}$ be c_{DNA}

Claim: $d_H(y_{\text{DNA}}, w_{\text{DNA}}) \geq 2$

Hence we can correct 2 bit flip in DNA
 $\neq 4$ in bits.

* Constrained Coding for DNA Data Storage

→ # of consecutive 0s & 1s is called run length.

$$x = (x_1, x_2, \dots, x_n) \in \{0, 1\}^n$$

→ A dk-seq :

d - const:

1111...0000...01111...1111

at least d zeros

K - const: zeros at most K.

→ RLL (run length limited) sequence

$$\bar{y} = (y_1, y_2, \dots, y_n)$$

can be constructed from dk sequence

$$y_i = y_{i-1} \oplus x_i, \quad y_0 = 1$$

$$\bar{x} = (0, 1, 0, 0, 1, 0, 0, 0, 1) \rightarrow \text{dk sequence}$$

$$\bar{y} = (1, 0, 0, 0, 1, 1, 1, 1, 0)$$

→ Run length of RLL seq lies between d+1 and K+1

→ Capacity of constrained channels

$$C = \lim_{T \rightarrow \infty} \frac{1}{T} \log_2 N(T)$$

$N(T) = \# \text{ of allowed signal seq in the time interval } T$

$N_d(n) = \# \text{ of d seq of length } n$

$$\rightarrow N_d(n) = C \lambda^n \quad n \geq 1, C = \text{const}$$

$\lambda = \text{growth factor}$

$$1 \leq \lambda \leq 2$$

→ is the largest real root of the characteristic eqⁿ. $\lambda^{d+1} - \lambda^d - 1 = 0$

$$\rightarrow d\text{-const channel} \quad C_d = \log_2 \lambda$$

$$\text{Code eff} = \frac{\text{code rate}}{\text{capacity}}$$

→ For DNA Storage,

→ $(8, 289, 3)_4$ was used.

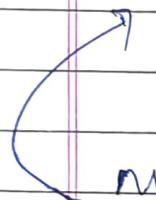
→ 166.997 EB per gram of DNA

→ Variable chunk length.

$$B(n, u) = \sum_{y=0}^{v-1} 2^{2v+1-2y} \binom{v-1}{y} \binom{n-v}{v-y}$$

+ \dots

$$\sum_{y=0}^{v-2} 2^{2v-1-2y} \binom{v-1}{y} \binom{n-v-1}{v-y-2}$$



$v = \min(u, n-u)$
number of codewords of GC-weight ~~the~~ with no repetition

→ Net Information = no. of input info bits
density no. of bases in resulting
DNA string

* Hamming Codes

Parity check matrix

$$G = \begin{bmatrix} I_{26} & -A \end{bmatrix}$$

26x31

$$H = \begin{bmatrix} A \\ I_5 \end{bmatrix}$$

31x5

Generator
matrix

$d=3$, so can correct 1 error.

Step 1 \rightarrow hello \rightarrow ASCII $\in \mathbb{Z}_4$ - 20 quads

Step 2 \rightarrow SHA-256, take rightmost 6 quads - 26 quads

Step 3 $\rightarrow b = aG$ 31 quads

Step 4 \rightarrow add parity 32 quads

Step 5 \rightarrow Convert c to DNA

$$\{0, 1, 2, 3\} \rightarrow \{A, G, C, T\}$$

* Reed Solomon Codes

$$GF(q) = \{0, 1, \alpha, \dots, \alpha^{q-2}\} \quad q = p^m$$

$$\beta (\neq 0) \in GF(q)$$

Order of β = smallest +ve integer m s.t $\beta^m = 1$

$$\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$$

order(2) = 4 $\rightarrow 2 \neq 3$ are primitive elements

ord(3) = 4 \therefore ord = 5 - 1 \downarrow
generators

→ All elements of field having ord $q-1$ are called primitive elements.

→ # of elements in $\text{GF}(q)$ of order t is $\phi(t)$

$$\phi(t) = t \prod_{p|t} \left(1 - \frac{1}{p}\right)$$

$p \neq t$ $p = \text{prime}$

$$\phi(6) = \phi(2 \cdot 3) = 2 \quad \phi(15) = 8$$

→ In $\text{GF}(q)$: # of p.e. = $\phi(q-1)$

→ A primitive polynomial is a polynomial having primitive elements as its roots.

Eg $\text{GF}(8)$

$$p(x) = x^3 + x + 1$$

Roots are: $\alpha, \alpha^2, \alpha^4 \rightarrow \text{p.e.}$

$$(\alpha + \alpha) (\alpha + \alpha^2) (\alpha + \alpha^4) = \alpha^3 + \alpha + 1$$

→ If α is primitive element, so is α^{-1} :

$$\alpha^{-1} = \alpha^{7-1} = \alpha^6$$

$$(\alpha + \alpha^6)(\alpha + \alpha^5)(\alpha + \alpha^3)$$

$$\alpha^{-2} = \alpha^5$$

$$= \alpha^3 + \alpha^2 + 1$$

$$\alpha^{-4} = \alpha^3$$

↳ P.P

→ If $p(x)$ is primitive, so $p^*(x) = x^m p(x^{-1})$

Reciprocal polynomial

→ No. of p.p of degree m is $\left\lfloor \frac{\phi(q-1)}{m} \right\rfloor$

Note

$$GF(8) = GF(q^m)$$

$$q=2, m=3$$

Conjugacy class

$$\{0\}$$

$$\alpha$$

poly of lowest degree
minimal polynomial

$$\{1\}$$

$$\alpha + 1$$

$$\{\alpha, \alpha^2, \alpha^4\}$$

$$\alpha^3 + \alpha + 1$$

$$\{\alpha^3, \alpha^6, \alpha^5\}$$

$$\alpha^3 + \alpha^2 + 1$$

out of these whose order is $q-1$ are primitive elements of p.p respectively.

→ The partition of powers of α by the conjugacy classes is called the set of cyclotomic cosets

→ For $GF(8)$: $\{0\}, \{1, 2, 4\}, \{3, 6, 5\}$

~~General~~Set $\sum_{p=1}^{m-1}$. Multiply by p

$$C_s = \{s, ps, p^2s, p^3s, \dots, p^{m_s-1}s\}$$

 m_s = smallest +ve integer

$$p^{m_s}s \equiv s \pmod{p^{m_s-1}}$$

~~eg~~ ~~eg~~ mod 15

$$p=2; m=4$$

$$C_0 = \{0\} \quad \text{Trivial}$$

$$C_1 = \{1, 2, 4, 8\}$$

$$C_3 = \{3, 6, 12, 9\}$$

$$C_5 = \{5, 10\}$$

$$C_7 = \{7, 14, 13, 11\}$$

cyclotomic cosets.

→ message bits : $(m_0, m_1, m_2, \dots, m_{k-1}) \in [GF(q)]^k$

$$p(x) = m_0 + m_1 x + \dots + m_{k-1} x^{k-1}$$

$$\bar{C} \in RS = \left\{ \bar{C} = c_0 + c_1 x + \dots + c_{k-1} x^{k-1} = P(0)P(x)P(x^2) \dots P(x^{k-1}) \right\}$$

$$\rightarrow c_i = P(x^i), \quad x^i \in GF(q)$$

$$GF(q) = \{0, 1, \alpha, \alpha^2, \dots, \alpha^{q-2}\}$$

$$\boxed{|RS| = 2^k} \rightarrow \dim RS = k \\ \rightarrow RS \text{ is a linear code}$$

$$P(0) = m_0$$

$$P(\alpha) = m_0 + m_1 \alpha + \dots + m_{k-1} \alpha^{k-1}$$

$$P(\alpha^{k-1}) = m_0 + m_1 \alpha^{k-1} + \dots + m_{k-1} \alpha^{(k-1)(k-1)}$$

→ Any k of these eqns $\begin{bmatrix} n \\ q \end{bmatrix}, [n, k, d]$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & & \\ 1 & \alpha & \dots & \dots & \dots & \alpha^{k-1} & \\ \vdots & & & & & & \\ 1 & \alpha^{k-1} & \dots & \dots & \dots & \alpha^{(k-1)(k-1)} & \end{bmatrix} \begin{bmatrix} m_0 \\ m_1 \\ \vdots \\ m_{k-1} \end{bmatrix} = \begin{bmatrix} P(0) \\ P(\alpha) \\ \vdots \\ P(\alpha^{k-1}) \end{bmatrix}$$

- non-singular matrix
- unique solution.

* Vandermonde Matrices

$$V = \begin{bmatrix} 1 & \gamma_1 & \gamma_1^2 & \cdots & \gamma_1^n \\ 1 & \gamma_2 & \gamma_2^2 & \cdots & \gamma_2^n \\ 1 & \gamma_3 & \gamma_3^2 & \cdots & \gamma_3^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \gamma_n & \gamma_n^2 & \cdots & \gamma_n^n \end{bmatrix}$$

$\gamma_i \in GF(q)$
distinct non-zero elements.

→ columns are linearly independent over $GF(q)$

* What γ_i so that we get matrix shown on prev. page?

→ Suppose that t of codeword coordinates are corrupted out of q^k .

• we can construct all possible of distinct sys of k -eqs

$$\# \text{ of ways} = \binom{q}{k}$$

→ $\binom{k+t-1}{k}$ of this will give incorrect solution.

Majority among all the solutions,

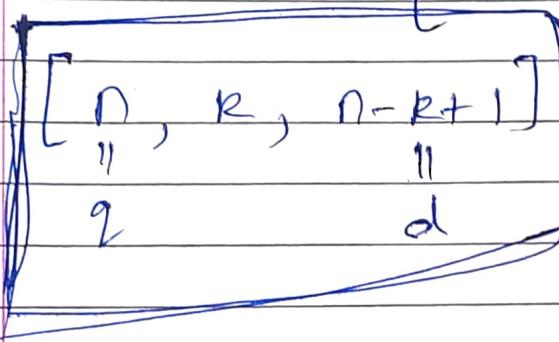
$$\binom{t+k-1}{k} < \binom{q-t}{k}$$

why?

$$\Rightarrow t+k-1 < q-t$$

$$\Rightarrow 2t < q-k+1$$

$$\Rightarrow t = \left\lfloor \frac{q-k+1}{2} \right\rfloor$$



Reed Solomon Codes

$$\begin{cases} d = 2t+1 \\ d = n-k+1 \end{cases} \quad n = q-1 \text{ (now)}$$

$$\rightarrow m(x) = m_0 + m_1 x + \dots + m_{R-1} x^{R-1}$$

$$c(x) = m(x)g(x) = \langle g(x) \rangle$$

$$g(x) = \prod_{j=1}^{2t} (x - \alpha^j) \quad (x - \alpha^d), \quad n = q-1$$

α = primitive element