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 ~~no~~ correct more errors).
- Max storage capacity achieving codes.
- Design chunk architecture of DNA blocks

→ Run-length coding?

Chunk Architecture : [Index bits] [Information bits]

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

Info bits	Index bits	Parity

chunk Architecture.

- Ternary Huffman codes were used.
- maps 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 bits flip error correction.
- 115 EB per gram of DNA achieved.

Extended Golay code $[24, 12, 8]_2$

$G = [\, I_{11\times 11} \;\; \begin{matrix} 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{matrix} \;\; \begin{matrix} 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & & & & & & & & & & \\ 0 & & & & & & & & & & \\ \vdots & & & & & & & & & & \\ 0 & & & & & & & & & & \end{matrix} \,]$

0 1 2 3 4 5 6 7 8 9 10

1 where quadratic residues cyclic shifts

11 times

0 1

Page No.:
Date: / /

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

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

$[12, 6, 6]_3$

Puncture one colum we get $[11, 6, 5]_3$

729 code words

$\longrightarrow$ $x_{DNA} \in C_{DNA}$ sent

y_{DNA} received

Lemma

If $d_H(x_{DNA}, y_{DNA}) = t$ & corresponding ternary code obtained using Goldman's base Table IV are x & y then

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

Eg: $x_{DNA} = GTCTCGTCCGTC$ $x = 10111001001$

$y_{DNA} = GACTCGTAGTC$ $y = 11011000101$

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

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

$y_{DNA} \in C_{DNA}$ received

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

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

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

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

Let $w_{DNA} \neq x_{DNA} \in C_{DNA}$

Claim: $d_H(y_{DNA}, w_{DNA}) > 2$

Hence we can correct 2 bit flip in DNA & 4 in bits.

* Constrained Coding for DNA Data storage

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

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

→ A dk- seq:

d- const:

1111 ... 0000 ... 0111 ...

at least d zeros

K - const: zeros at most K.

→ RLL (run length limited) sequence

$\bar{y} = (y_1, y_2, \ldots y_n)$ can be constructed from dk sequence

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

$\bar{x} = (0,0010001)$ → dk sequence

$\bar{y} = (1000111 10)$

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

→ Capacity of constrained channels

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

$N(T)$ = # of allowed signal seq in the time interval T.

$N_d(n)$ = # of d seq of length n

→ $N_d(n) = C\lambda^n$ $\quad n \gg 1$, C = const, λ = growth factor

$1 \leq \lambda \leq 2$

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

→ d-const channel $C_d = \log_2 \lambda$

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

→ For DNA Storage,

→ $(8, 289, 3)_4$ was used.

→ 166.977 EB per gram of DNA

→ Variable chunk length.

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

$v = \min(u, n-u)$

number of codewords of GC-weight u with no repitions

→ **Net Information density** = $\frac{\text{no. of input into bits}}{\text{no. of bases in resulting DNA string}}$

Page No.:
Date: / /

* Hamming Codes

$G = [I_{26} \mid -A]_{26 \times 31}$ — Generator matrix

parity check matrix $H = \begin{bmatrix} A \\ I_5 \end{bmatrix}_{31 \times 5}$

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

Step 1 → hello → Ascii → Z_4 — 20 quads

Step 2 → SHA-256, take rightmost 6 quads — 26 quads

Step 3 → $b = aG$ 31 quads

Step 4 → add parity 32 quads

Step 5 → Convert ∈ to DNA.

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

* Reed Solomon Codes

$GF(q) = \{0, 1, \alpha_2, \ldots, \alpha_{q-2}\}$ $q = p^m$

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

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

$Z_5 = \{0, 1, 2, 3, 4\}$

order (2) = 4 → 2 & 3 are primitive elements

ord (3) = 4 ∵ ord = 5 − 1 ↓ generators

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

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

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

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

→ In GF(q): # of p.e = $\phi(q-1)$

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

Eg GF(8)

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

Roots are: $\alpha, \alpha^2, \alpha^4$ → p.e

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

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

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

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

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

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

$= x^3 + x^2 + 1$

↳ 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 $\boxed{\phi(q-1)/m}$

Note: $GF(8) = GF(q^m)$ $\quad q=2, m=3$

Conjugacy class	minimal polynomial (poly of lowest degree)
$\{0\}$	x
$\{1\}$	$x+1$
$\{\alpha, \alpha^2, \alpha^4\}$	x^3+x+1
$\{\alpha^3, \alpha^6, \alpha^5\}$	x^3+x^2+1

out of these whose order is $q-1$ are primitive elements & 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 Z_{p^m-1} . Multiply by p

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

m_s = smallest +ve integer

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

Eg mod 15

$p = 2$; $m = 4$

$C_0 = \{0\}$ 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 : $(m_0\ m_1, m_2 \dots m_{k-1}) \in [GF(q)]^k$

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

$$\bar{c} \in RS = \{\bar{c} = c_0 c_1 \dots c_{q-1} = P(0)P(\alpha)P(\alpha^2) \dots P(\alpha^{q-1})\}$$

→ $c_i = P(\alpha^i)$, $\alpha^i \in GF(q)$

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

$|RS| = 2^k$ → dim RS = k
→ RS is a linear code

$$P(0) = m_0$$
$$P(\alpha) = m_0 + m_1 \alpha + \dots m_{k-1} \alpha^{k-1}$$
$$\vdots$$
$$P(\alpha^{q-1}) = m_0 + m_1 \alpha^{q-1} + \dots + m_{k-1} \alpha^{(k-1)(q-1)}$$

→ Any k of these eqns [n, k, d]
n = q

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

- non-singular matrix
- unique solution.

Vandermonde Matrices

$$V = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & & \lambda_n^2 \\ 1 & & & \\ \vdots & & & \\ \lambda_1^n & \lambda_2^n & & \lambda_n^n \end{bmatrix}$$

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

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

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

→ Suppose that t of codeword coordinates are corrupted out of q

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

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

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

Page No.:
Date : / /

Majority among all the solutions,

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

why?

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

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

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

$$[n, k, n-k+1]$$
$$\| \quad \| \quad \|$$
$$q \quad d$$

Reed Solomon Codes

$d = 2t+1$ $\quad n = q-1$ (now)

$d = n-k+1$

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

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

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

α = primitive element.