

DNA ストレージに対する誤り訂正符号

金子晴彦

東京工業大学 情報理工学院

2022 年電子情報通信学会総合大会 (AT-1-2)

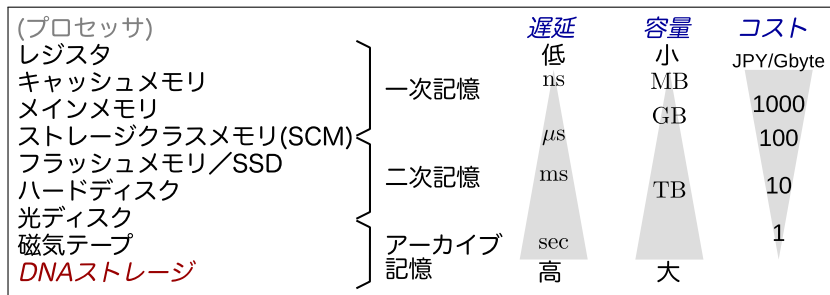
- DNA ストレージに対する誤り訂正符号に関連する研究の動向を紹介
- DNA ストレージ：次世代アーカイブ用ストレージとして期待
 - + 高記憶密度： 1.1×10^{23} bytes/g [1]
 - + 高耐久性：数百年
 - 高い誤り率，同期（挿入削除）誤り
 - 記録パターンに制約（cf. 眞田先生のご講演）
 - 低スループット（書き込み／読み出し）
- 本発表の内容
 - DNA ストレージの概要
 - 挿入削除誤り訂正符号（2元符号）
 - 複数トレース復号

- 1 DNA ストレージの概要
- 2 挿入削除誤り訂正符号
- 3 複数トレース復号
- 4 まとめと今後の課題

- 1 DNA ストレージの概要
- 2 挿入削除誤り訂正符号
- 3 複数トレース復号
- 4 まとめと今後の課題

DNA ストレージの位置づけ

- アーカイブ用ストレージ
 - 高記憶密度 : 1.1×10^{23} bytes/g [1]
(G: 10^9 , T: 10^{12} , P: 10^{15} , E: 10^{18})
 - 高耐久性 [2]
 - 低スループット, 高遅延
- コンピュータシステムの記憶階層



DNA ストレージの実装例 [2]

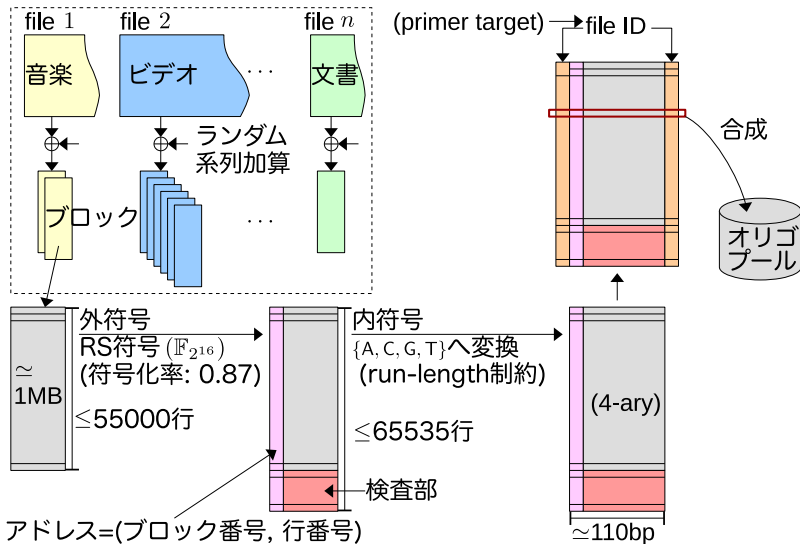
	capacity (MB)	oligo pool size	oligo length	cover- age	RA (*1)	SQ (*2)	rate (bit/nt)	ECC (*3)
Church et al. (2012)	0.655	54898	115	3000	no	IH	0.83	(none)
Goldman et al. (2013)	0.739	153335	117	52	no	IH	0.33	overlap, PC, MV
Yazdi et al. (2015)	0.017	32	1000	n/a	yes	SG	n/a	prefix- synchronization
Grass et al. (2015)	0.083	4991	117	372	no	IM	1.14	concatenated RS code
Bornholt et al. (2016)	0.151	45652	120	122	yes	IM	0.85	tunable redundancy, PC
Blawat et al. (2016)	22	900000	190	161	no	IH	1.03	BCH/RS code, PC, MV
Erlich & Zielinski (2017)	2.14	72000	152	10	no	IM	1.57	fountain code, RS code
Yazdi et al. (2017)	0.004	17	1000	392	yes	NP	1.72	consensus alignment, MV
Organick et al. (2018)	200.2	13448372	110	5	yes	IN	1.10	randomization, concatenatd code,
	0.033	2130	n/a	36	yes	NP	1.10	clustering, majority alignment

*1:random access capability

*2:sequencer (IH:Illumina HiSeq, IM:Illumina MiSeq, IN:Illumina NextSeq, SG:Sanger, NP:Nanopore MinION)

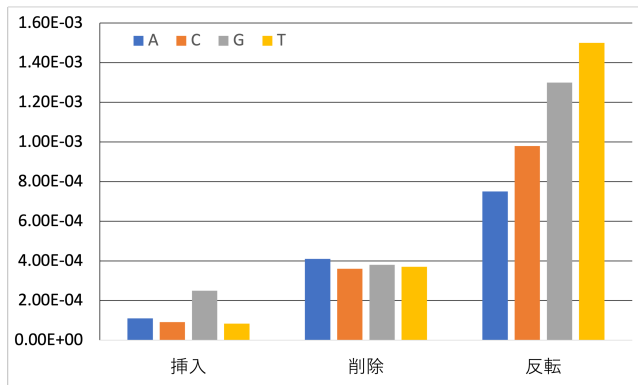
*3:error control code (PC:parity-check code, MV:majority voting)

ランダムアクセス [3]



誤り率の例

- Illumina NextSeq における挿入削除反転 (IDS) 誤り率の例 [3]



- 反転誤り (substitution) が最も多い
- Nanopore シーケンサの誤り率: Illumina の数倍～十倍程度 [3]

DNA ストレージで求められる符号化

(1) IDS 誤り訂正符号

- 通信路誤り率： 10^{-2} から 10^{-3} 程度（同期誤りを含む）
- 符号化／復号計算量の制約： 緩い（∵ アーカイブ用ストレージ）

(2) 制約符号

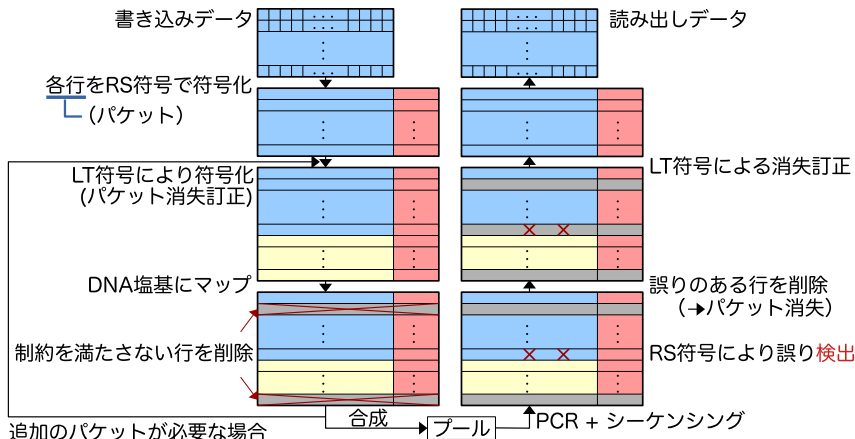
- オリゴ生成，シーケンシング，等に由来する制約
- run-length 制約： 同一シンボルの連続数を最大 ℓ シンボルとする
- GC バランス制約： 系列中の GC 成分の割合を約半分とする

(3) 複数トレース復号

- 読み出し時にオリゴを PCR により増幅（複数のコピーを生成）
- 単一の符号語に対して複数の受信語（トレース）が得られる
- 複数トレースを用いた復号が有効

DNA ストレージに対する符号化の例：DNA fountain [5]

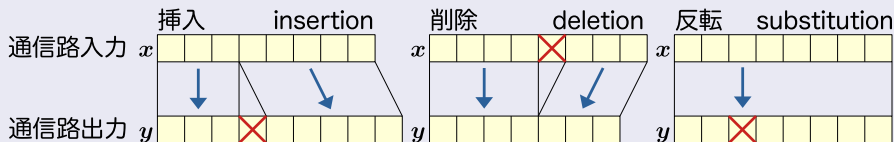
- IDS 誤り訂正：RS 符号による誤り検出 + LT 符号によるパケット消失訂正
- 制約符号化：非充足パケット破棄 + LT 符号によるパケット生成



- 1 DNA ストレージの概要
- 2 挿入削除誤り訂正符号
- 3 複数トレース復号
- 4 まとめと今後の課題

挿入削除誤り訂正符号

挿入／削除／反転 (IDS) 誤り



IDS 誤り訂正符号

- t 個の IDS 誤りを訂正する符号
 - (例) Varshamov-Tenengolts(VT) 符号 ($t = 1$) [6,7]

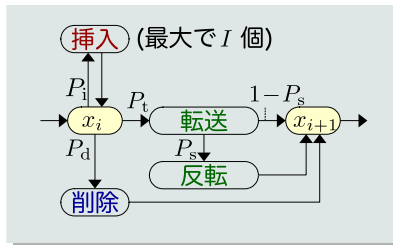
$$\mathcal{C} = \left\{ (u_1, u_2, \dots, u_n) \in \mathbb{F}_2^n \mid \sum_{i=1}^n i u_i \equiv a \pmod{2n} \right\}$$

- 確率的 IDS 通信路に対する符号

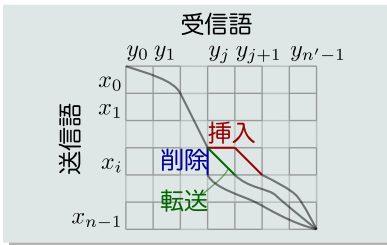
(※以降では 2 元符号について述べる)

確率的 IDS 通信路モデル

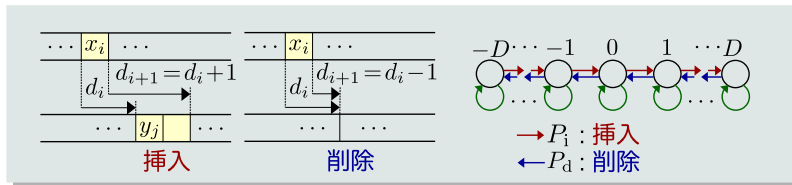
- 誤り確率: 挿入 P_i , 削除 P_d , 反転 P_s
- 通信路入力: $\mathbf{x} = (x_0, \dots, x_{n-1})$, 通信路出力: $\mathbf{y} = (y_0, \dots, y_{n'-1})$



(a) 状態遷移図

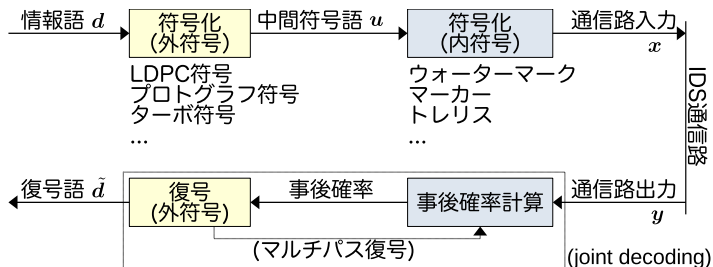


(b) 格子

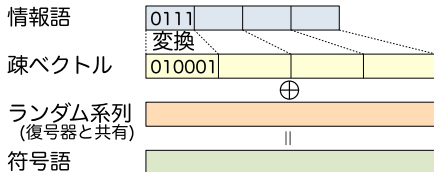


(c) ドリフト値

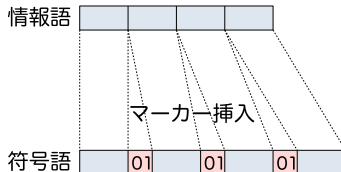
確率的IDS通信路に対する符号化：連接符号化



ウォーターマーク [10]



マーカー [12,13]

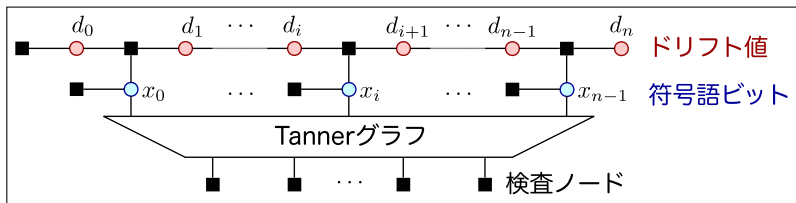


確率的IDS通信路に対する符号化：単一の符号

● LDPC 符号

- 空間結合符号 [15]
- 非正則 LDPC 符号 [16]
- 復号： $p(\mathbf{y}, \mathbf{x}, \mathbf{d})$ のファクターグラフ上の sum-product アルゴリズム
 - $\mathbf{x}$: 送信語, $\mathbf{y}$: 受信語, $\mathbf{d}$: ドリフトベクトル, $\mathcal{C} \subset \mathbb{F}_2^n$: LDPC 符号

$$\begin{aligned} p(\mathbf{y}, \mathbf{x}, \mathbf{d}) &= p(\mathbf{y}|\mathbf{x}, \mathbf{d})p(\mathbf{x})p(\mathbf{d}) \\ &= p(\mathbf{x})p(d_0) \prod_{i=0}^{n-1} p(\mathbf{y}_{i+d_i}^{i+d_{i+1}} | x_i, d_i, d_{i+1})p(d_{i+1}|d_i) \\ &= \mathbb{1}[\mathbf{x} \in \mathcal{C}] \times p(d_0) \prod_{i=0}^{n-1} p(\mathbf{y}_{i+d_i}^{i+d_{i+1}} | x_i, d_i, d_{i+1})p(d_{i+1}|d_i) \end{aligned}$$



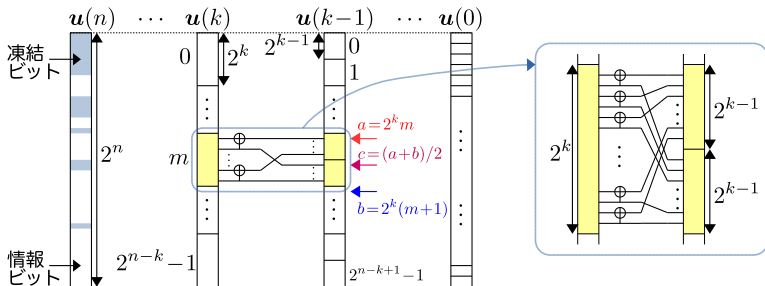
● Polar 符号 [17,18]

Polar 符号による IDS 誤り訂正 [18] : 符号化

- 符号長: $N = 2^n$
- 凍結ビットのインデックス集合: $\mathcal{A}^c$
- 情報ベクトル: $\mathbf{u}_0^{N-1} = \mathbf{u}(n)_0^{N-1}$ (情報ビットと凍結ビットから構成)
- 符号語: $\mathbf{x}_0^{N-1} = \mathbf{u}(0)_0^{N-1}$
- 再帰的符号化: レベル $k \rightarrow$ レベル $k-1$

$$\mathbf{u}(k-1)_a^{c-1} = \text{even}(\mathbf{u}(k)_a^{b-1}) \oplus \text{odd}(\mathbf{u}(k)_a^{b-1})$$

$$\mathbf{u}(k-1)_c^{b-1} = \text{odd}(\mathbf{u}(k)_a^{b-1})$$



Polar 符号による IDS 誤り訂正：逐次除去復号

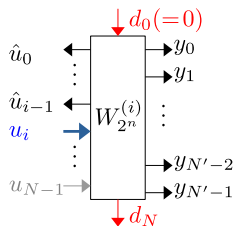
- 受信語： $\mathbf{y}_0^{N'-1} = (y_0, y_1, \dots, y_{N'-1})$
 - 受信語長： $N' = N + d_N$ (既知とする)
- 復号語： $\hat{\mathbf{u}}_0^{N'-1} = (\hat{u}_0, \hat{u}_1, \dots, \hat{u}_{N'-1})$
- 第 i ビット の 値： ($i = 0$ から順に計算)

$$\hat{u}_i = \begin{cases} u_i & (i \in \mathcal{A}^c : \text{凍結ビット}) \\ h_i(\mathbf{N}', \mathbf{y}_0^{N'-1}, \hat{\mathbf{u}}_0^{i-1}) & (i \in \mathcal{A} : \text{情報ビット}) \end{cases}$$

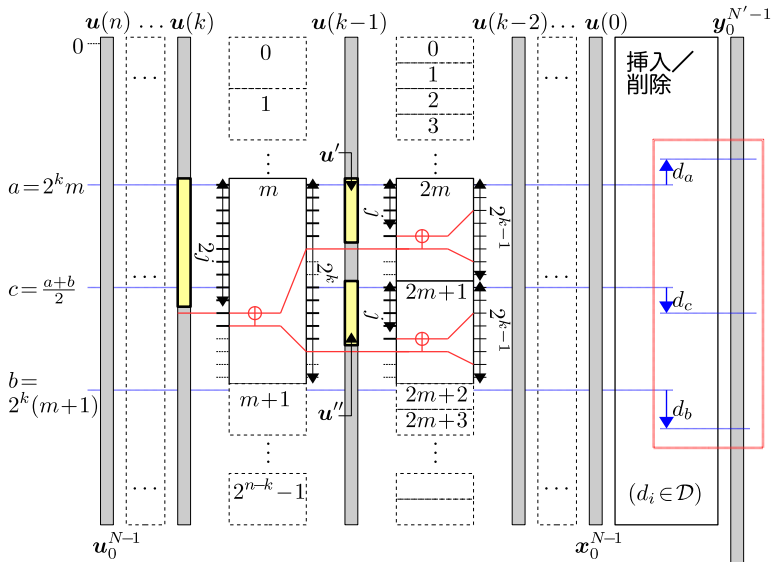
$$h_i(\mathbf{N}', \mathbf{y}_0^{N'-1}, \hat{\mathbf{u}}_0^{i-1}) = \begin{cases} 0 & \left(\frac{p(\mathbf{d}_N, \mathbf{y}_0^{N'-1}, \mathbf{u}_0^{i-1} = \hat{\mathbf{u}}_0^{i-1} | u_i = 0)}{p(\mathbf{d}_N, \mathbf{y}_0^{N'-1}, \mathbf{u}_0^{i-1} = \hat{\mathbf{u}}_0^{i-1} | u_i = 1)} \geq 1 \right) \\ 1 & (\text{otherwise}) \end{cases}$$

- Polar ビット 通信路:

$$W_{2^n}^{(i)} \left(\mathbf{d}_N, \mathbf{y}_0^{N'-1}, \mathbf{u}(n)_0^{i-1} = \hat{\mathbf{u}}_0^{i-1} \middle| \mathbf{d}_0 = 0, \mathbf{u}(n)_i = \alpha \right)$$

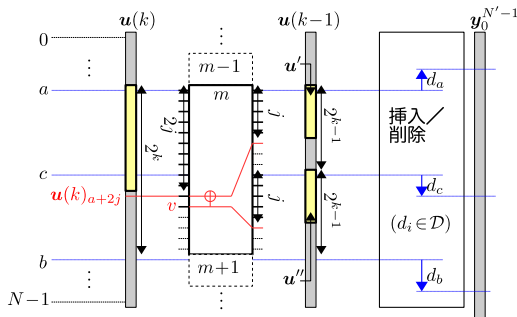


Polar 符号による IDS 誤り訂正：復号の流れ



Polar 符号による IDS 誤り訂正：偶数インデックス

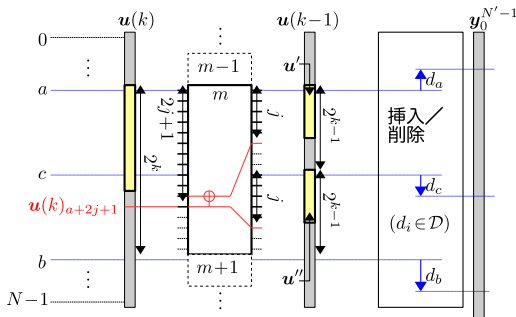
$$\begin{aligned}
 & W_{2^k}^{(2j)} \left(\mathbf{d}_b, \mathbf{y}_{a+\mathbf{d}_a}^{b+\mathbf{d}_b-1}, \mathbf{u}(k)_a^{a+2j-1} \middle| \mathbf{d}_a, \mathbf{u}(k)_{a+2j} \right) = \\
 & \frac{1}{2} \sum_{\mathbf{d}_c \in \mathcal{D}} \sum_{v \in \mathbb{B}} \left\{ W_{2^{k-1}}^{(j)} \left(\mathbf{d}_c, \mathbf{y}_{a+\mathbf{d}_a}^{c+\mathbf{d}_c-1}, \mathbf{u}(k-1)_a^{a+j-1} = \mathbf{u}' \middle| \mathbf{d}_a, \mathbf{u}(k-1)_{a+j} = \mathbf{u}(k)_{a+2j} \oplus v \right) \right. \\
 & \quad \left. \times W_{2^{k-1}}^{(j)} \left(\mathbf{d}_b, \mathbf{y}_{c+\mathbf{d}_c}^{b+\mathbf{d}_b-1}, \mathbf{u}(k-1)_c^{c+j-1} = \mathbf{u}'' \middle| \mathbf{d}_c, \mathbf{u}(k-1)_{c+j} = v \right) \right\}
 \end{aligned}$$



$$\mathbf{u}' = \text{even} \left(\mathbf{u}(k)_a^{a+2j-1} \right) \oplus \text{odd} \left(\mathbf{u}(k)_a^{a+2j-1} \right), \quad \mathbf{u}'' = \text{odd} \left(\mathbf{u}(k)_a^{a+2j-1} \right).$$

Polar 符号による IDS 誤り訂正：奇数インデックス

$$\begin{aligned}
 & W_{2^k}^{(2j+1)} \left(\mathbf{d}_b, \mathbf{y}_{a+d_a}^{b+d_b-1}, \mathbf{u}(k)_a^{a+2j} \middle| \mathbf{d}_a, \mathbf{u}(k)_{a+2j+1} \right) = \\
 & \frac{1}{2} \sum_{d_c \in \mathcal{D}} \left\{ W_{2^{k-1}}^{(j)} \left(\mathbf{d}_c, \mathbf{y}_{a+d_a}^{c+d_c-1}, \mathbf{u}(k-1)_a^{a+j-1} = \mathbf{u}' \middle| \mathbf{d}_a, \mathbf{u}(k-1)_{a+j} = \mathbf{u}(k)_{a+2j} \oplus \mathbf{u}(k)_{a+2j+1} \right) \right. \\
 & \quad \left. \times W_{2^{k-1}}^{(j)} \left(\mathbf{d}_b, \mathbf{y}_{c+d_c}^{b+d_b-1}, \mathbf{u}(k-1)_c^{c+j-1} = \mathbf{u}'' \middle| \mathbf{d}_c, \mathbf{u}(k-1)_{c+j} = \mathbf{u}(k)_{a+2j+1} \right) \right\}
 \end{aligned}$$



$$\mathbf{u}' = \text{even} \left(\mathbf{u}(k)_a^{a+2j-1} \right) \oplus \text{odd} \left(\mathbf{u}(k)_a^{a+2j-1} \right), \quad \mathbf{u}'' = \text{odd} \left(\mathbf{u}(k)_a^{a+2j-1} \right).$$

Polar 符号による IDS 誤り訂正 : レベル 0

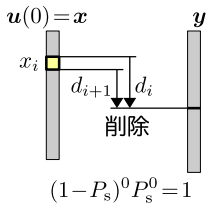
$$W_{2^0}^{(i)} \left(d_{i+1}, \mathbf{y}_{i+d_i}^{i+d_{i+1}} \mid d_i, \mathbf{u}(0)_i \right) = p \left(\mathbf{y}_{i+d_i}^{i+d_{i+1}} \mid d_i, d_{i+1}, x_i \right) p(d_{i+1} \mid d_i)$$

$$p \left(\mathbf{y}_{i+d_i}^{i+d_{i+1}} \mid d_i, d_{i+1}, x_i \right) = (1 - P_s)^{|\mathcal{I}(i, d_i, d_{i+1})| - \delta} \times P_s^\delta$$

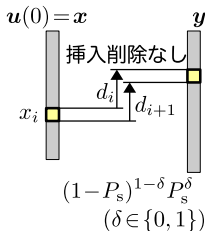
$$\mathcal{I}(i, d_i, d_{i+1}) = \{i' \mid i + d_i \leq i' < i + 1 + d_{i+1}\}$$

$$\delta = \sum_{i' \in \mathcal{I}(i, d_i, d_{i+1})} (x_i \oplus y_{i'}) : \text{反転誤りの数}$$

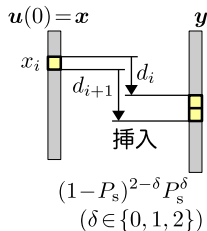
(i) $d_{i+1} = d_i - 1$



(ii) $d_{i+1} = d_i$



(iii) $d_{i+1} = d_i + 1$



Polar 符号による IDS 誤り訂正：凍結ビットの選択

- 符号長： $N = 2^n$
- 凍結ビット数： $M = |\mathcal{A}^c|$
- 凍結ビットのインデックス集合 $\mathcal{A}^c$ ：
 - 対称通信路容量 $I(W_N^{(i)})$ が最も小さい M 個のインデックス i の集合

$$I(W_N^{(i)}) = \frac{1}{2^N} \sum_{\mathbf{u}} \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{u}) \tilde{I}(d_N, \mathbf{y}, \mathbf{u}_0^{i-1} | u_i)$$

$$\tilde{I}(d_N, \mathbf{y}, \mathbf{u}_0^{i-1} | u_i) = \log \frac{p(d_N, \mathbf{y}, \mathbf{u}_0^{i-1} | u_i)}{\frac{1}{2}p(d_N, \mathbf{y}, \mathbf{u}_0^{i-1} | u_i) + \frac{1}{2}p(d_N, \mathbf{y}, \mathbf{u}_0^{i-1} | \bar{u}_i)}$$

- (本研究では) シミュレーションにより推定
- 修正 SC 復号により N_f 個の受信語を復号
- 各 $i \in \{0, 1, \dots, N-1\}$ に対して $\tilde{I}(d_N, \mathbf{y}, \mathbf{u}_0^{i-1} | u_i)$ の平均を求める
 - u_i の復号において $\mathbf{u}_0^{i-1}$ は正しく復号されていると仮定

Polar 符号による IDS 誤り訂正 : シミュレーション

通信路・符号パラメータ

- ドリフト値（絶対値）の最大値 : $D = 4$
- 誤り確率 : $P_i = P_d, P_s = 0$
- 符号長 : $N = 4096$
- 符号化率 : $R = 0.80$

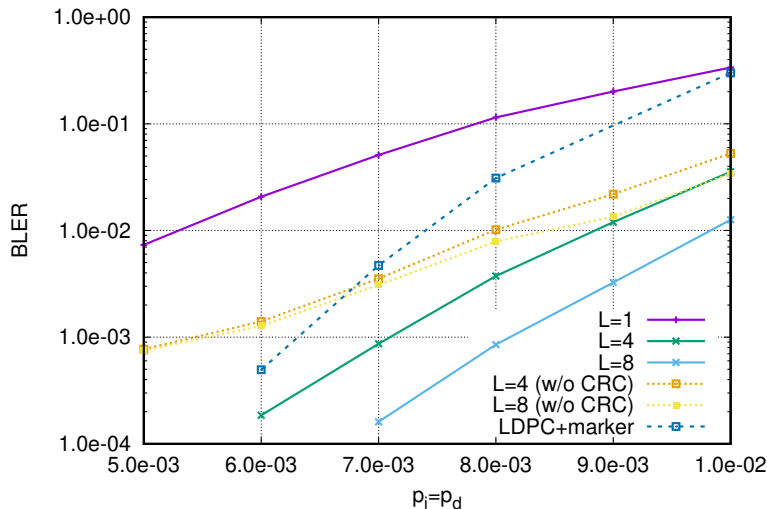
Polar 符号

- リストサイズ : $L \in \{1, 4, 8\}$
- CRC: $g(x) = x^8 + x^7 + x^6 + x^4 + x^2 + 1$

LDPC 符号+マーカー符号

- 外符号 : 2元正則 LDPC 符号 (列重み 3)
- 内符号 : マーカー挿入 (間隔 : $\lambda \in \{6, 20\}$)

Polar 符号による IDS 誤り訂正：ブロック誤り率の例



- 1 DNA ストレージの概要
- 2 挿入削除誤り訂正符号
- 3 複数トレース復号**
- 4 まとめと今後の課題

DNA ストレージと複数トレース復号

通信路入力 (書き込みデータ)

x C T A G T A C A G T A G T C

通信路出力 (複数トレース)

z^1 C T A G T A C A T G T A G T C

挿入

z^2 C T A G T A A G T A G T C

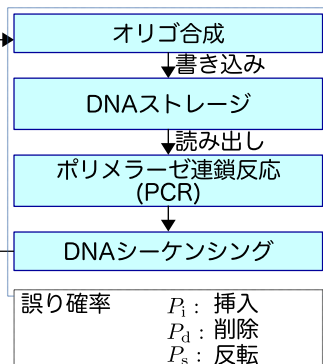
削除

z^3 C G A G T A C A G T A G T C

反転

$\vdots$

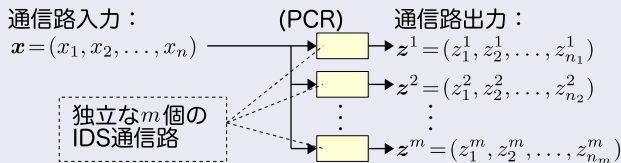
z^m C T A G T A C A G T A G T C



複数トレース通信路

複数トレース IDS 通信路

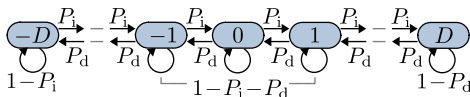
単一の送信語 $\Rightarrow m$ 個の IDS 通信路 $\Rightarrow m$ 個の受信語



- 完全再構成： $z^1, \dots, z^m$ から z を復元
- 送信シンボルの事後確率 $p(x_i | z^1, \dots, z^m)$ の計算
 - 削除通信路を対象 [19]
 - 演算量の削減 [20]
 - **IDS 通信路を対象 [21]**

通信路モデル

- 誤り確率 : P_i (挿入), P_d (削除), P_s (反転)
- 入力 : $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{Z}_q^n$
- 出力 : $\mathbf{Z} = \begin{pmatrix} z^1 \\ z^2 \\ \vdots \\ z^m \end{pmatrix} = \begin{pmatrix} (z_1^1, z_2^1, \dots, z_{n_1}^1) \\ (z_1^2, z_2^2, \dots, z_{n_2}^2) \\ \vdots \\ (z_1^m, z_2^m, \dots, z_{n_m}^m) \end{pmatrix}$
 - $\mathbf{z}^k = (z_1^k, z_2^k, \dots, z_{n_k}^k) \in \mathbb{Z}_{q_k}^{n_k}$: k 番目のトレース (長さ n_k)
- 入出力シンボル間の最大ドリフト値 : D
- ドリフト値の集合 : $\mathcal{D} = \{-D, \dots, -1, 0, 1, \dots, D\}$
- ドリフトベクトル : $\mathbf{d}^k = (d_1^k, d_2^k, \dots, d_n^k, d_{n+1}^k) \in \mathcal{D}^{n+1}$
- マルコフ過程として定義 (初期状態 $d_1^k = 0$)



ドリフト値行列

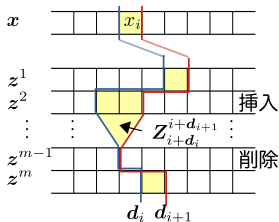
$$D = \begin{bmatrix} d^1 \\ d^2 \\ \vdots \\ d^m \end{bmatrix} = \begin{bmatrix} d_1^1 & d_2^1 & \cdots & d_{n+1}^1 \\ d_1^2 & d_2^2 & \cdots & d_{n+1}^2 \\ \vdots & \vdots & & \vdots \\ d_1^m & d_2^m & \cdots & d_{n+1}^m \end{bmatrix}$$

$$= \begin{bmatrix} d_1 & d_2 & \cdots & d_{n+1} \end{bmatrix} \in \mathcal{D}^{m \times (n+1)}$$

- 第 k 行 d^k : k 番目のトレース z^k のドリフトベクトル
- 第 i 列 d_i : 入力シンボル x_i に対応するドリフト値

Z の第 i セグメント

$$Z_{i+d_i}^{i+d_{i+1}} = \begin{pmatrix} (z_{i+d_i}^1, \dots, z_{i+d_{i+1}}^1) \\ (z_{i+d_i}^2, \dots, z_{i+d_{i+1}}^2) \\ \vdots \\ (z_{i+d_i}^m, \dots, z_{i+d_{i+1}}^m) \end{pmatrix}$$



ファクターグラフの導出 (1/2)

- 事後確率 $p(x_i | \mathbf{Z})$:
同時確率 $p(\mathbf{Z}, \mathbf{x}, \mathbf{D})$ に対するファクターグラフ上で計算

$$\begin{aligned} p(\mathbf{Z}, \mathbf{x}, \mathbf{D}) &= p(\mathbf{Z} | \mathbf{x}, \mathbf{D}) p(\mathbf{x}, \mathbf{D}) = p(\mathbf{Z} | \mathbf{x}, \mathbf{D}) p(\mathbf{D}) p(\mathbf{x}) \\ &= p(\mathbf{d}_1) \prod_{i=1}^n p\left(\mathbf{Z}_{i+\mathbf{d}_i}^{i+\mathbf{d}_{i+1}} \mid x_i, \mathbf{d}_i, \mathbf{d}_{i+1}\right) p(\mathbf{d}_{i+1} | \mathbf{d}_i) p(x_i) \end{aligned}$$

$$p(\mathbf{d}_1) = \prod_{k=1}^m p(d_1^k) = \begin{cases} 1 & (\mathbf{d}_1 = (0, \dots, 0)) \\ 0 & (\text{otherwise}) \end{cases}$$

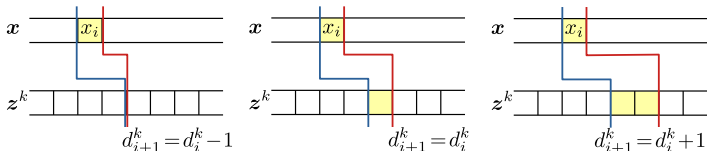
$$p\left(\mathbf{Z}_{i+\mathbf{d}_i}^{i+\mathbf{d}_{i+1}} \mid x_i, \mathbf{d}_i, \mathbf{d}_{i+1}\right) = \prod_{k=1}^m p\left((\mathbf{z}^k)_{i+\mathbf{d}_i^k}^{i+\mathbf{d}_{i+1}^k} \mid x_i, d_i^k, d_{i+1}^k\right)$$

$$p(\mathbf{d}_{i+1} | \mathbf{d}_i) = \prod_{k=1}^m p(d_{i+1}^k | d_i^k)$$

ファクターグラフの導出 (2/2)

- k 番目のトレースにおける尤度 = (既存の)IDS 通信路の尤度

$$p \left((z^k)^{i+d_{i+1}^k} \middle| x_i, d_i^k, d_{i+1}^k \right) = \begin{cases} 1 & (d_{i+1}^k = d_i^k - 1) \\ f \left(x_i, z_{i+d_i^k}^k \right) & (d_{i+1}^k = d_i^k) \\ f \left(x_i, z_{i+d_i^k}^k \right) f \left(x_i, z_{i+1+d_i^k}^k \right) & (d_{i+1}^k = d_i^k + 1) \end{cases}$$



- 反転誤り確率 :

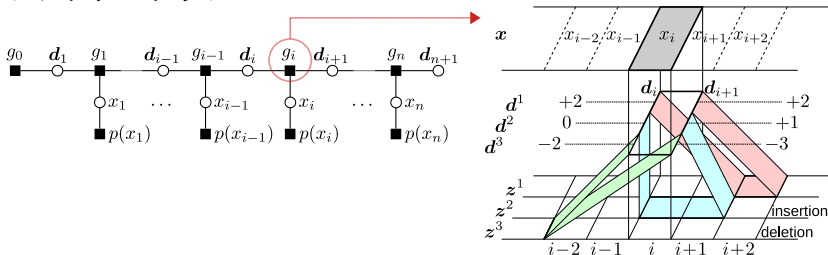
$$f(x, z) = \begin{cases} 1 - P_s & (x = z) \\ P_s / (q - 1) & (x \neq z) \end{cases}$$

ファクターグラフ

- 同時確率の因数分解：

$$p(\mathbf{Z}, \mathbf{x}, \mathbf{D}) = p(d_1) \prod_{i=1}^n p\left(\mathbf{Z}_{i+d_i}^{i+d_{i+1}} \mid x_i, d_i, d_{i+1}\right) p(d_{i+1} \mid d_i) p(x_i)$$

- ファクターグラフ：

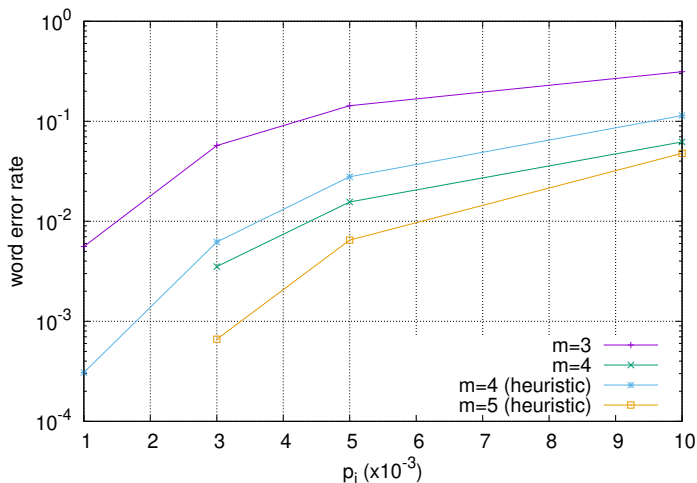


- 事後確率 $p(x_i | \mathbf{Z})$ の計算：

- ファクターグラフ上で sum-product アルゴリズムを実行
- MAP 推定： $\tilde{x}_i = \arg \max_{x_i \in \mathbb{Z}_q} p(x_i | \mathbf{Z})$

シミュレーションによる数値例

- ブロック長: $n = 152$, 最大ドリフト値: $D = 4$
- 誤り率: $P_d = 5 \times P_i$, $P_s = 10^{-3}$



- 1 DNA ストレージの概要
- 2 挿入削除誤り訂正符号
- 3 複数トレース復号
- 4 まとめと今後の課題

まとめと今後の課題

まとめ

DNA ストレージに対する符号化技術

- ① IDS 誤り訂正符号 (例: Polar 符号の逐次除去復号法)
- ② 複数トレース復号 (例: ファクターグラフを用いた MAP 推定)
- ③ 制約符号化: Run-length 制約, GC バランス制約

今後の課題

- 上記 (1)~(3) を統合した符号化・復号法
- Nanopore シーケンサ
 - 信号検出アルゴリズム (basecaller) と復号アルゴリズムの関係
 - 読み取り信号 → basecaller (RNN) → 塩基配列 → 誤り訂正 → 復号語
 - 読み取り信号 → **basecaller (RNN) + 誤り訂正** → 復号語