DNAストレージ

- 情報をDNA塩基配列（A,T,G,Cの）として記憶
- 次世代アーカイブストレージとして研究開発が進展

【利点】
アーカイブストレージとしての特性
- 高記憶密度：～3400PB/g
- 高耐久性：　～750年（@10℃）
- 電磁ノイズ耐性（非磁性体）
（↔HDD，磁気テープ）

【課題点】
既存のストレージ（HDD，磁気テープ，等）と異なる性質
- 高誤り率：シンボル誤り，同期（挿入削除）誤り
（例）誤り率＝10^{-4} ～ 10^{-2}
- 記録可能な系列に制約：連長（run-length）制約，バランス制約
（オリゴ生成／保存過程，読み出し機構，等に起因）
- 低スループット，高遅延，高コスト

応用・実装 ← → **理論・学術**

オリゴ合成・シーケンサー技術分野
＋ 実験に基づくデータ
＋ 正確な誤りモデル
＋ 具体的な制約条件
＋ ソースコード公開
－ 基礎的な符号化復号技術
－ 誤り率解析の不足

情報システム技術分野
－システムモデル不足
・アクセス方式
・インターフェース
・ソフトウェア

p_i：挿入誤り確率
p_d：削除誤り確率

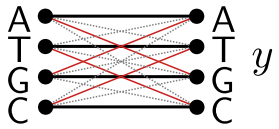
本研究
・DNA通信路モデル
非対称シンボル誤り
非対称同期誤り：$p_\mathrm{i} \neq p_\mathrm{d}$
オリゴ消失
ヘッダ／アドレス部誤り
・符号化法
同期誤り訂正＋制約符号
（連長制約，GCバランス）
連接符号化
・復号法
soft-input復号(FASTQ形式)
・実用的な誤り率 $\leq 10^{-15}$
・アーカイブ用途に適した
ファイルアクセス方式
・オープンソースライブラリ構築

符号理論分野
＋ 高効率な符号化／復号
LDPC符号，polar符号，
プロトグラフ，
sum-product復号，
ファクターグラフ，
複数トレース復号，…
＋ 理論的解析
通信路容量，誤り率，
符号語数，…
－ 単純化された通信路モデル
対称シンボル誤り
対称同期誤り：$p_\mathrm{i} = p_\mathrm{d}$
－ 制約符号化の欠落
－ 符号パラメーターの妥当性
－ 多くがソースコード非公開

研究計画

【符号理論】

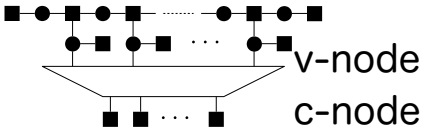
通信路モデル
- 非対称同期誤り： $p_{\mathrm{i}} \neq p_{\mathrm{d}}$
- 非対称シンボル誤り： $p(y|x) \neq p(y'|x)$
- オリゴ消失，ヘッダ／アドレス部誤り
- 補助情報出力（信頼度パラメータ）



性能評価
- 復号語誤り率
- エラーフロア
- 符号化率
- 計算量

成果発表
- 国際会議
- ジャーナル

復号アルゴリズム
(例) 確率伝搬法
- ファクターグラフ
- soft-input復号
- マルチパス復号
- 拡大アルファベット

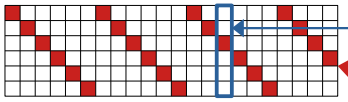


符号機能
通信路符号化： 非対称同期誤り
非対称シンボル誤り
符号語消失

＋

制約符号化： 連長制約
GCバランス制約
motif回避

符号設計
(例) 連接符号化
外符号：LDPC符号，polar符号
内符号：非線形符号(計算機探索)



シミュレーション環境

シーケンサモデル調査
サンプルデータ収集／生成
- MESA: DNA storage simulator
- Oxford Nanopore: base caller
- MSR experimental data
  など

プログラム作成
提案手法： 通信路モデル
符号探索
符号化／復号

比較対象： DNA-Aeon
Hedges
DNA fountain

最適化
- ルックアップテーブル
- 近似計算
- 並列化／GPU

システム設計
- ファイルアクセス方式
- データ構造／ヘッダ情報

ライブラリ公開

【実データ・実装・応用】