堅牢かつ有用なDNAストレージシステム設計のための データ変換法の比較

小久保 剛基　　清水 佳奈

早稲田大学 基幹理工学部

1　はじめに

DNA をストレージとして利用するシステムは, データを高密度で書き込むことができ, また適切な環境下であれば長期間の安定保存が可能であるという長所を持っていて, 既存のストレージに代わるシステムとして大きな可能性を秘めている. しかし, 実用化するために解決すべき大きな課題が二つあり, 一つ目は DNA 合成時と DNA 読み取り時にかかるコストの問題である. たとえば 454 FLX Titanium という次世代シーケンサを用いた場合, 一回の読み取りで $6200 かかる [2]. また既存の機器では合成時と読み取り時にエラーが起きてしまい, 正しいデータが得られないことがある. この二つの問題点は機器の性能の向上と共に改善されるものではあるが, エラーを訂正するための冗長性を付与しつつ, コンパクトな DNA データを作成できる変換法が研究されてきた. 従来手法ではデータ密度とエラー訂正能力とのオーバーヘッドがあり, どちらかを犠牲にする必要があったが提案手法ではこの問題点を改善し, 一定のエラー訂正能力を維持しつつデータ密度を高い値までコントロールする方法を開発した. 本研究では関連研究で提案されている既存手法を用いて予備実験を行い, 提案手法での実験結果と比較をして有用性の検証を行った.

2　関連研究

2.1　基本の変換法

あらゆるデジタルデータは 0 と 1 のバイナリデータとして表現でき, それに対して DNA データはアデニン (A), チミン (T), グアニン (G), シトシン (C) の四種類の塩基で表すことができるが, 単純にマッピングをすると同じ塩基が続くことがあり, DNA 合成時と読み取り時に機器の原理的にエラーが起きやすくなってしまう. そのためバイナリデータを Huffman 符号化し, 3 進数で表現した後に一つ前の塩基と自身の数値から一定の規則に従って塩基の連続が起きないように DNA データへと変換する.

2.2　Goldman, N. の手法

2.1 の変換法で得られた DNA データを重複部分ができるように多数のセグメントに分割し, 分割したセグメントに順序情報やパリティ情報を付与して出力する変換法であり, 復元時には, 重複箇所を比較することでエラーを訂正する [3]. この手法は 2.3 で述べる手法のベースとなる変換法であり, 一度の読み取りで得られた DNA データに対してのエラー訂正能力の高さが長所であるが, そのために多量の冗長性を付与する必要があるという短所がある.

2.3　XOR を用いた手法

2.1 の変換法で得られた DNA データを多数のセグメントに分割し, ペアを作って各セグメントの XOR をとって新たなセグメントを生成していき, 2.2 の手法と同様に順序情報やパリティ情報を付与する変換法である [1]. 復元時には XOR の性質を利用してエラーが発生したセグメントとは別のペアからエラーを訂正していく。ベースとなるデザインは Goldman, N の手法を採用し, エラー訂正の仕組みを変えたこの手法は, エラー訂正能力では Goldman, N. の手法に劣るものの付与する冗長性の量を大幅に抑えることができるという長所がある.

3　提案手法

図 1: 提案手法の生成塩基列

提案手法では図 1 のような三つの塩基列を生成する. まず 1：元ゲノムシーケンスに対して一塩基単位で階層的に XOR を取り, 2：エラー訂正用ゲノムシーケンスを生成する. さらに先頭の塩基を自身以降の塩基に XOR をした 3：エラー判定用ゲノムシーケンスを生成する. 元データ復元時にはまず 1 と 3 の DNA 配列の各桁で XOR 計算をして 1 のエラー部分を判定し, エラー部分の情報を元に 1 と 2 の DNA 配列で XOR 計算をしていくことでエラーを訂正することができる.

Comparison of encoding methods for robust and efficient DNA storage system
†Yoshiki Kokubo
‡Waseda University

4 実験

提案手法を実装し，ファイルサイズ 7.2Kbyte の JPEG 画像を DNA 配列に変換し，データ密度と復元率を計測した．データを復元する際に生じるシークエンシングエラーをシミュレートするために，MetaSim Genome Simulator を利用した [4]．ここでは illumina シークエンサーのエラーパラメータを設定し, 実際のシークエンシングエラーを想定したデータを生成した．結果を表 1, 表 2 に示す．

表 1: 提案手法によるデータ変換結果

分割幅 [nt]	DNA の長さ [nt]	密度 [bits/nt]
50	209643	0.39183
150	174849	0.46980
250	168210	0.48834

表 1 は, 分割するセグメント幅を 50nt から実際のシーケンサでの読み取り可能な最大の長さを想定した 250nt まで変化させて実験を行った結果である．セグメントの幅が長くなるにつれてデータ密度が上昇し, 一定の値へと収束しているのが分かる．50nt でのセグメント分割ですでに Goldman, N. の手法のデータ密度 0.33711 を超えていて最大で 0.4883 程度までデータ密度を上げることが可能だという結果が得られた．

表 2: 提案手法と Goldman, N. の手法の比較

手法	データ密度	エラー生起率	復元率
Goldman	0.33711	0.00682	0.99989
提案手法	0.39183	0.01393	0.99879

次に表 2 は実装した Goldman, N. の手法と, 提案手法のセグメントの幅 50nt で, 対象画像ファイルに対して実験を行った結果を比較したものである．シーケンス機器の性質上, 長い DNA 配列はよりエラーが起こりやすい [1]．エラー訂正能力を測定するために, 最もエラー生起率が高くなると予想されるセグメント幅が 50nt のものを選択し, 実験を行った．データ密度は元データ [bits] を DNA 配列の長さ [nt] で除算したもので Goldman, N. の手法では 0.33711 となり, シーケンシングエラー生起率が 0.00682 の DNA 配列に対して復元を行うと 0.99989 の精度まで復元することができた．一方, 提案手法では対象とする DNA 配列の塩基のパターンによって Goldman, N. のエラーの生起率と差が現れてしまったが, Goldman, N. の手法と比較するとより高密度なデータを生成することに成功し, かつ多量のエラーを含んだ DNA データに対して 0.99879 の精度まで復元することができた．

5 考察

提案手法はエラー訂正時には一塩基単位で計算を行うため, セグメントの幅によらず一定のエラー訂正能力を発揮できるのが利点である．そのため大きなセグメントに分割することでより高密度に, またシーケンサに合わせた無駄のない適切なセグメント幅に設定することができる．本来セグメントの幅を大きくすることで順序情報やパリティといった付与する冗長性の量を減らすことができるため他の手法でも大きなセグメントに分割するのが理想だが, Goldman, N. の手法ではそれ以上に重複させる箇所が増えてしまいデータ密度が高くなってしまう．また XOR 法においてセグメントを大きく分割するということは XOR 計算によるエラー訂正に使えない, エラーを含んでしまったセグメントの数が増加するということであり, その部分を再び読み取る際には余計なコストが必要になってしまうという短所がある．提案手法は一塩基単位でエラーを判定する仕組みを備えており, 無駄を最小限に抑えてエラー訂正処理を行うことができるという点で優れていると言える．例えば 454 FLX+e というシーケンサは 1 度に 700nt と広域を読み取ることができ高性能だが, 稼働に $6200, 1Mb の DNA 配列を読むのに $7 とコストがかかる [2]．提案手法ならばこういったシーケンサに適した無駄なく低コストな DNA 配列の設計が可能である．

6 まとめ

本研究では DNA ストレージシステムにおける従来手法と提案手法を実装して比較した．問題点の一部を改善することで従来手法を超えるデータ密度と, 十分なエラー訂正能力を実現できたと考えられる．[1] で述べられているが, よりコンパクトな DNA 配列を生成することでシーケンサによる読み取り精度が上昇するため, そもそものエラーが起こる確率を下げることができる．また DNA の合成と読み取りのコストを考えるとよりデータ密度の高いコンパクトなデータ列を設計することが重要視される．データ密度の大きい提案手法はシーケンス機器の性能向上と共にさらに実用的なものになると考えられる．今後はより正確なエラー判定と適切な計算手順によりエラー訂正能力を上げられるように提案手法を改善していくつもりである．

参考文献

[1] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss. A dna-based archival storage system. *SIGPLAN Not.*, 51(4):637–649, Mar. 2016.

[2] T. C. GLENN. Field guide to next-generation dna sequencers. *MOLECULAR ECOLOGY RESOURCES*, May 2011.

[3] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. Leproust, B. Sipos, and E. Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *NATURE*, 494(7435):77–80, February 2013.

[4] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson. Metasim—a sequencing simulator for genomics and metagenomics. *PROS*, October 2008.