様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １６ 日現在

機関番号：１３１０２
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１１８３３

研究課題名（和文）DNAストレージに適した符号化方式の離散数学的検討

研究課題名（英文）Discrete mathematical approach for suitable coding scheme on DNA strorage

研究代表者
眞田　亜紀子（Manada, Akiko）

長岡技術科学大学・工学研究科・准教授

研究者番号：２０６３１１３８

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：DNAを用いた記録媒体であるDNAストレージは，長寿命と高密度の観点で既存の記録媒体よりはるかに優れており次世代記録媒体として注目されている．一方で，実用化に向けては「低コストでのデータの保存・読み出し」や，「データの誤りや欠落をなくす」ことを考慮した，低コストで誤りなくデータの読み書きが可能な手法が必須となる．
本研究では，DNA系列として適切でない系列を排除したりデータ圧縮を予め行うことで，低コストで信頼性の高い符号化方式を提案した．具体的には，制約符号の概念からグラフ理論を用いて提案手法の符号化率の解析を行なったり，循環系列を用いた新たなデータ圧縮法を提案し，その有効性を確かめた．

研究成果の学術的意義や社会的意義
DNAストレージの実用化に向けての「低コスト」かつ「信頼性」の観点から，主としてDNAの系列に関した研究成果をあげている．具体的には，DNAに適したデータ系列を考慮し，その系列にどのように符号化するか着目した成果である．つまり，系列の観点から「効率的に誤り耐性を持たせる」ことを論じているため，単体として使用した場合はもちろんのこと，ランダムに起きた誤りを修正する「誤り訂正」と融合させることで更なる有効性が期待できる．このような基礎的研究を基盤としてDNAストレージが実用化されれば，ビッグデータをコンパクトに保存することが現実的となり，社会的貢献は非常に高いと言える．

研究成果の概要（英文）：DNA storage media, storage media consisting of DNA strands, are promising data storage media in the next generation from the perspective of ultra long lifespan and high density. On the other hand, “the low cost in reading/writing data” and “error robustness performances” are important issues, and suitable methods to approach these issues are highly demanded for the practical use of the media.
In this study, we proposed suitable coding schemes by concerning constraints on data sequences and data compression schemes. More precisely, we utilized the concepts of constrained coding and data compression based on circular strings, and evaluated their effectiveness.

研究分野：情報理論基礎

キーワード：DNAストレージ　グラフ理論　データ系列制約　キャパシティ　データ圧縮

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

DNA（デオキシリボ核酸）は A（アデニン），C（シトシン），G（グアニン），T（チミン）の４つの塩基から成る遺伝子情報を保存する高分子生体物質であり，DNA ストレージは A，G，C，T から成る系列とバイナリ系列（0-1 列）とを対応させ（例：00=A，01=C，10=G，11=T），それらを繋げた DNA 鎖を系列とすることでデータの保存・読み出しを行う記録メディアである．DNA ストレージは，その「記録密度の高さ」や「長期保存可能」という観点で既存の磁気・電気・光学記録メディアよりもはるかに優れており（引用文献[1]），次世代記録媒体して昨今注目されている．特に Goldman ら（引用文献[2]）が DNA ストレージの可能性について明確に論じてから，実用化に向けての研究が盛んに行われている．

DNA ストレージは，長寿命と高密度の観点で非常に優位である一方で，実用化に向けては「データの保存・読み出しが困難である」ことや，「データの誤りや欠落をなくす」ことを考慮する必要がある．実際，DNA 鎖を生成するためにかかる金額や，データ読み出しには特別な装置が必要になることは金銭的・時間的コスト面で大きな問題である．また，長年保存する上で交叉や突然変異による誤りが発生することは無視できない．つまり，低コストで誤りなくデータの読み書きが可能な手法が必須である．

DNA ストレージに適した符号化方式として，研究開始当初は主として誤り訂正符号（ある閾値以下の誤り数であれば誤りを訂正できる符号）の観点から議論されてきた．しかし，DNA 鎖において誤りが起こりやすい系列や作成が難しい系列なども実験でわかっている．よって，そのような系列を排除してあげることで，より信頼性が高く効率の良い DNA ストレージ作成への貢献ができるのではと考えた．また，研究申請者らのデータ圧縮の知見も融合させることで，コスト面において有効性を高められないかと考えた．

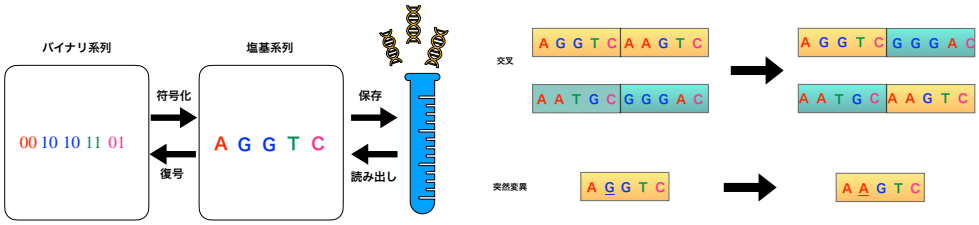

図 1 DNA ストレージのイメージ

２．研究の目的

本研究では，DNA ストレージ実用化に向けて，「データ系列の制約」及び「データ系列の圧縮」の観点から DNA 鎖に適した系列にエンコードするための低コストと高信頼性を兼ね備えた符号化を提案することを目的とした．具体的には，

（１）　突然変異を起こしやすい記号列を含まない記号列への制約符号化

（２）　書き込みと読み出し時間を短縮するための低コストなデータ圧縮法

を離散数学（主としてグラフ理論）を用いてモデル化し，適した符号化を数学的に解析することで，DNA ストレージ研究の符号化に関する基礎を固めることを試みた．

３．研究の方法

本研究で用いた主な研究方法は次の３つである．離散数学，特にグラフ理論を研究の核として研究に取り組んだ．

(1) 制約符号の概念とその拡張

制約符号とは，系列の中にある特定の系列（禁止語）が現れないように符号化する方法で，HDD や DVD などの記録媒体に広く活用されている．禁止語が与えられたとき，「経路に沿って辺のラベルを読むことで生成される系列は禁止語を含まない」ラベル付き有向グラフ（例として図２参照）を作成することができる．このグラフは制約符号の研究の核であり，禁止語が現れないようにする符号化方式や，その最大符号化率に関する情報などが得られる（例：引用文献[3]参照）．

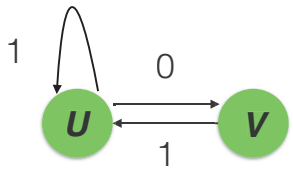

図 2　00 を含まないバイナリ系列を生成するグラフ

DNA ストレージに用いられる DNA 鎖は，

①　同じ塩基は高々k回連ねることができる制約（最大連長制約）

②　A と T の割合と C と G の割合を同じにする制約（GC バランス制約）

の両方を兼ね備える必要がある．最大連長制約は「同じ塩基を $k+1$ 回連ねたもの」を禁止語として考えることができ，禁止語からラベル付き有向グラフを作成することで，最大符号化率など主要な結果を得ることができる．一方で，GC バランス制約では禁止語を考えることができずグラ

フを作成できない．つまり，最大連長制約と GC バランス制約を満たす系列は制約符号のみでは扱うことができない．よって，制約符号の考えと Knuth のアルゴリズム（下記(2)）を融合させて本研究に取り組んだ．

(2) バランス状態にするアルゴリズム

Knuth（引用文献[4]）は，系列長が $2m$ のバイナリ系列 $\boldsymbol{b}=b_1b_2\cdots b_{2m}$ は，1 ビット目から j ビットまでをフリップ（0→1，1→0）することで，0 と 1 の数を等しくすること（バランス状態にすること．例として図３参照）が可能であることを示している．DNA 鎖における GC バランス制約は A と T を 0，G と C を 1 と考えたバイナリ系列をバランス状態にすることと同値である．よって，Knuth の方法を応用して，GC バランス制約を満たした DNA 鎖を生成することを試みた．

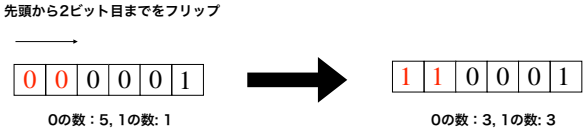

図３　バランス状態にするアルゴリズムの例

(3) Compression by Substring Enumeration（CSE 法）の応用

Compression by Substring Enumeration（CSE 法）は 2010 年に Dubé と Beaudoin（引用文献[5]）によって提案されたユニバーサルな無ひずみデータ圧縮方法（ユニバーサル符号）である．ユニバーサル符号とは，圧縮対象の情報源について，事前の知識（記号の出現確率など）がなくても，最良の圧縮率が得られるものである．CSE 法では，圧縮したい環状系列（例として図４参照）の部分系列の出現回数や極小禁止語（一番前，もしくは一番後ろのシンボルを削ると系列中に現れる禁止語）を用いてデータを圧縮している．

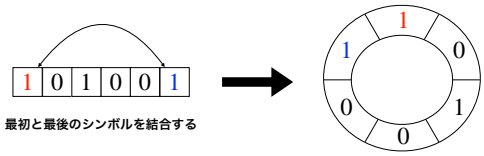

図 ４　環状系列の例

CSE 法に関して，Dubé らはバイナリ系列（0-1 系列）に対する CSE 法（2 値 CSE 法）を議論していたが，DNA など 3 個以上のシンボルの場合への CSE 法の拡張（多値 CSE 法）は，共同研究者らにより行われている（引用文献[6]）．一方で，多値 CSE 法は，2 値 CSE 法に比べて，実用上では圧縮率が悪い課題があった．本研究では，極小禁止語について更に知識を深めることで，多値 CSE 法について，圧縮率を改善し，より良いデータ圧縮に貢献することを試みた．

４．研究成果

本研究で得られた主な成果は次の４つである．(1)，(2)は「データ系列の制約」の観点から，(3)，(4)は「データ系列の圧縮」の観点からの結果である．また，以下の文中で用いる（論○），（学○）は，それぞれ「５．主な研究論文等」における「雑誌論文」の先頭から○番目，「学会発表」の先頭から○番目を意味する．

(1) Bond が存在するための必要十分条件（論４，学８）

制約を満たす系列が２つ与えられたとき，それらを直接結合して生成した新たな系列が制約を満たすかどうか保証はない．そこで結合する際に特別な系列（Bond）を毎回間に挟むことで制約を保持することができないか考えた．具体的には，Bond が存在するため必要十分条件を「禁止語の最初と最後に使われているシンボル数」と「実際に使用するシンボル数」との不等式で与えるとともに，Bond の長さの上限値についても議論した．また，Bond の作成アルゴリズムの複雑性について考察した．それらの結果を元に，DNA 鎖など連長制約を考慮した系列への応用についても議論した．

図５　11 を禁止語とした場合の Bond の例（黄色の 0 が Bond）

(2) 最大連長制約と GC バランス制約を兼ね備えた場合の最大符号化率（論２，学２）

DNA 鎖は「最大連長制約」と「GC バランス制約」の２つ制約を満たす必要があるが，GC バランス制約はグラフで表現することができない．つまり，２つの制約を満たす DNA 鎖へ符号化する際の最大符号化率はグラフを用いて求めることができない．しかし，最大連長制約を満たす制約符号と Knuth らの方法を融合させることで，「２つの制約を満たした際の最大符号化率」と「最大連長制約のみを満たした場合の最大符号化率」が一致することを示した．

最大連長制約を満たした系列はグラフで表現可能のため，グラフから最大符号化率を求めることが可能となる．それに関連して，グラフの隣接行列から最大連長制約の最大符号化率を効率的に求めるための式も導出した．

(3) 環状系列が復元できるための必要十分条件（論１，学３）

環状系列 $\boldsymbol{x}$ が与えられたとき，

- 頂点：$\boldsymbol{x}$ の部分系列（頻度情報含む．極小禁止語の情報も利用）
- 有向辺とそのラベルの付け方の基本操作（必要に応じて削除も行う）：

  V から U へラベル s の有向辺を引く．

  ⇔系列 Vs の接尾辞（Vs の末尾に現れる部分系列）のうち，頂点の中で最長のものが U である．

とするラベル付き有向グラフを考える（例として，図６参照）．このグラフにおいて，頂点を訪れる際に頻度を１減らす操作を行う．このとき，「部分系列と頻度，及び極小禁止語を用いて（環状系列として）$\boldsymbol{x}$ を一意に復元可能であること」と「グラフにおける一筆書きが一意にできること」が必要十分条件であることを示した．

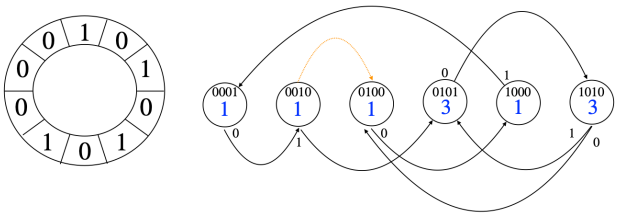

図６　環状系列（左）と対応するラベル付き有向グラフ（右）．黄色の辺は途中で削除される．

(4) ランダムアクセスを可能にするための位置情報埋め込み（論３，論５，学７）

DNA ストレージシステムでは，原データを非常に短い(塩基数で 100～1000 程度)固定長の DNA 鎖に分けて保存を行う．保存する際にその鎖が元々のデータの何番目の位置であったかを示す必要があるため，データそのものだけでなく，位置を示すためのヘッダーを必要とする．一方で，DNA 鎖にヘッダーが占める割合が 10%以上となるため課題がある．

本研究では，データハイディングの手法を応用して，位置情報をデータに埋め込めるか検討を行った．その結果，バイナリ系列を塩基列に変換するテーブルを拡張することにより，位置情報をデータに埋め込む手法を確立し，一つの DNA 鎖に格納するデータ長を増やすことを可能とした．データ削減の観点で言い換えると，ヘッダーのオーバヘッドを削減し，与えられたデータを保存するための DNA 鎖の総数を削減した．また，埋め込み可能な位置情報の長さについても評価を行った．

本研究では「データ系列の制約」と「データ圧縮」の観点から中心に行ったが，より信頼性を高めるためには，交叉や突然変異など DNA に起こり得る誤りも考慮した符号化が必要になる．今後は，「データ系列の制約」と「データ圧縮」だけでなく「誤り訂正」も融合した符号化の提案を目指して研究に勤しむ．

＜引用文献＞

[1] A. Extance, “How DNA could store all the world’s data,” Nature, vol. 537, pp. 22-25, September 2016.

[2] 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, vol. 494, pp. 77-80, February 2013.

[3] D. Lind and B. Marcus, “An Introduction to Symbolic Dynamics and Coding,” Cambridge University Press, 1995.

[4] D. E. Knuth, “Efficient Balanced Codes,” IEEE Trans. Inform. Theory, vol. IT-32, no. 1, pp. 51-53, January 1986.

[5] D. Dubé and V. Beaudoin, “Lossless data compression via substring enumeration,” Proc. of the Data Compression Conference, pp.229-238, March 2010

[6] T. Ota and H. Morita, “A Compact Tree Representation of an Antidictionary,” IEICE Trans. Fundamentals, vol. E100-A, No. 9, September 2017.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Takahiro Ota and Akiko Manada	４．巻 1
２．論文標題 A Necessary and Sufficient Condition for Reconstruction of Circular Binary String based on the Lengths of Substrings with Weights	５．発行年 2022年
３．雑誌名 Proc. of 2022 International Symposium on Information Theory and Its Applications	６．最初と最後の頁 118-122
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Akiko Manada, Takahiro Ota and Hiroyoshi Morita	４．巻 1
２．論文標題 The Maximum Run-Length Constrained Balanced Codes for Random-Access DNA Storage	５．発行年 2022年
３．雑誌名 Proc. of 2022 International Symposium on Information Theory and Its Applications	６．最初と最後の頁 123-127
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Takahiro Ota and Akiko Manada	４．巻 1
２．論文標題 Addressing Information Using Data Hiding for DNA-based Storage Systems	５．発行年 2020年
３．雑誌名 Proc. of 2020 International Symposium on Information Theory and Its Applications	６．最初と最後の頁 509-513
掲載論文のDOI（デジタルオブジェクト識別子） 10.34385/proc.65.C07-1	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Akiko Manada, Takahiro Ota and Hiroyoshi Morita	４．巻 1
２．論文標題 Bonds of Constrained Systems and Their Characteristics	５．発行年 2020年
３．雑誌名 Proc. of 2020 International Symposium on Information Theory and Its Applications	６．最初と最後の頁 240-244
掲載論文のDOI（デジタルオブジェクト識別子） 10.34385/proc.65.B05-1	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Takahiro Ota, Hiroyoshi Morita, Akiko Manada	4．巻 Vol. E103-A, No. 6
2．論文標題 Compression by Substring Enumeration Using Sorted Contingency Tables	5．発行年 2020年
3．雑誌名 IEICE Trans. Fundamentals	6．最初と最後の頁 829-835
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transfun.2019EAP1063	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計8件（うち招待講演　2件／うち国際学会　1件）

1．発表者名 Akiko MANADA
2．発表標題 Introduction on Constrained Coding for DNA Storage
3．学会等名 The 12th Vietnam Japan Scientific Exchange Meeting（招待講演）（国際学会）
4．発表年 2022年

1．発表者名 安納 直毅，眞田 亜紀子，山内 陸，太田 隆博
2．発表標題 q元連長制約符号に特化した特性方程式について
3．学会等名 情報理論研究会
4．発表年 2023年

1．発表者名 太田隆博，眞田亜紀子
2．発表標題 反辞書符号化法における一意復号可能な重み付き部分列の十分条件
3．学会等名 第44回情報理論とその応用シンポジウム
4．発表年 2021年

１．発表者名
Akiko Manada

２．発表標題
Graph Theoretical Reviews on Constrained Coding for Data Storage Media

３．学会等名
2022年電子情報通信学会総合大会　企画セッション（招待講演）

４．発表年
2022年

１．発表者名
眞田亜紀子, 太田隆博

２．発表標題
A survey on coding for DNA storage

３．学会等名
第９回誤り訂正符号のワークショップ

４．発表年
2020年

１．発表者名
Akiko Manada, Takahiro Ota, Hiroyoshi Morita

２．発表標題
Maximum Sum-Rates of Input-Constrained Repeatable WOM Codes

３．学会等名
第４２回情報理論とその応用シンポジウム

４．発表年
2019年

１．発表者名
太田 隆博, 眞田 亜紀子

２．発表標題
DNAストレージのデータ圧縮に関する検討

３．学会等名
第４２回情報理論とその応用シンポジウム

４．発表年
2019年

１．発表者名 Akiko Manada, Takahiro Ota, Hiroyoshi Morita
２．発表標題 On the Existence of Bonds for Constrained Systems
３．学会等名 情報理論研究会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	太田　隆博 (Ota Takahiro) (60579001)	専修大学・ネットワーク情報学部・教授 (32634)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	森田　啓義 (Morita Hiroyoshi)		
研究協力者	安納　直毅 (Annou Naoki)		
研究協力者	山内　陸 (Yamauchi Riku)		

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	金　範洙 (Kim Byum Soo)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関