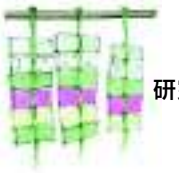
完全線形符号のDNA配列解析への応用



研究ノート

竹　中　要　一*

Perfect Linear Code for DNA sequence analysis

Key Words：Bioinformatics Perfect Linear Code DNA analysis

1．はじめに

生物であれば人も微生物もその体を形成する細胞内にゲノムを持っています．ゲノムにはその生物の設計図がデオキシリボ核酸（DNA）という物質で書き込まれています．そのうち最も解析が進んでいるのは，タンパク質を規定するDNA配列である遺伝子です．タンパク質は生物の構造を形作っていたり，酵素として生命現象を担っていたりします．分かりやすい酵素としては，食物の消化に使われる消化酵素が挙げられます．このタンパク質は生物種によって持っている種類や数が異なり，同じ役割を果たすタンパク質であっても生物種によって形状や効率が異なる事が知られています．また，同じ生物種であったとしてもタンパク質は同じというわけではなく，少しずつ異なります．例えば人間であっても人それぞれ異なる事が，血液型や目の色，髪の色の違いに繋がったり，ある疾病に罹りやすくなったりといった現象を導きます．このタンパク質の差異は，その設計図である遺伝子の差異でもあります．そこで遺伝子が書き込まれているDNAを読む事によってタンパク質の差異を明らかにし，これに基づき生物学や生物工学の研究を発展させ，疾病の解明による医学や治療法の発展に繋げる営みが特に2000年以降頻繁に行われる事になりました．

DNAとは，五炭糖とリン酸，塩基から構成される核酸であり，これがホスホジエステル結合によって一本鎖を形成します．一般には2つのDNA一本鎖が水素結合によって螺旋状に並んだ二重螺旋構造が有名だと思います．DNAの塩基は，アデニン，シトシン，グアニン，チミンの4種類が用いられており，それぞれアルファベット一文字A, C, G, Tで略記されます．遺伝子はこの4文字からなる単語（文字列）として扱われます．この表現法により，遺伝子の違いは文字列の違いと言い換える事ができます．図1にABO式血液型を決定する遺伝子において，

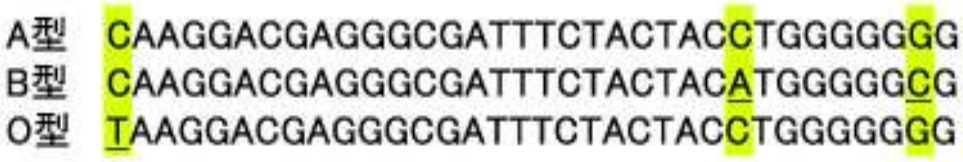


図1　ABO式血液型を決める遺伝子の一部．A型，B型，O型で異なる塩基を着色．

血液型によって異なるDNA配列部分を抜粋します[1]．遺伝子のDNA配列の違いは，血液型だけではなく他の遺伝的特徴や遺伝性疾患に見受けられます．これらの解析は，ACGTからなる文字列で表現された遺伝子群の，文字列の違いを探す事で行う事ができます．文字列比較は計算機が得意とする処理であり，生物の情報を主に解析する学問として生物情報学が誕生しました．

* Yoichi TAKENAKA

1973年6月生
大阪大学　大学院基礎工学研究科　情報数理系計算機科学分野　博士後期課程
現在、大阪大学　情報科学研究科　バイオ情報工学専攻　ゲノム情報工学講座
准教授　博士（工学）　生物情報学
TEL：06-6879-4391
FAX：06-6879-4394
E-mail：takenaka@ist.osaka-u.ac.jp

DNA配列の文字列を比較する研究は比較的古く，2つの遺伝子の類似度を動的計画法で計算するアルゴリズムの論文が1970年に上梓されています[2]．それ以降も高速化を主目的とする多数のアルゴリズムが提案されてきました[4-6]．この理由の一つとしては，データベースに登録される遺伝子のDNA配列数，すなわちデータ量の増加が，計算機の速度向上度を大きく上回っている事が挙げられます．図2に計算機の速度とDNA配列を読む実験器具（DNAシーケンサー）の速度を比較したグラフを記します．縦軸は，計算機が一秒間に計算できる浮動小数点演算数とDNAシーケンサーが1回の実験で決定する事ができる塩基配列の数を対数で示しています．この図より，DNAシーケンサーの速度向上度が計算機の速度向上度を大幅に上回っている事がわかります．この傾向は今後も続く事が予想されるため，より一層効率的で高速な計算アルゴリズムが求められています．私は情報通信を行うための方式の一つである完全線形符号を，アルゴリズムの効率化と高速化に役立てる研究を行ってきました．本研究ノートではその基礎部分をお伝えしたいと思います．

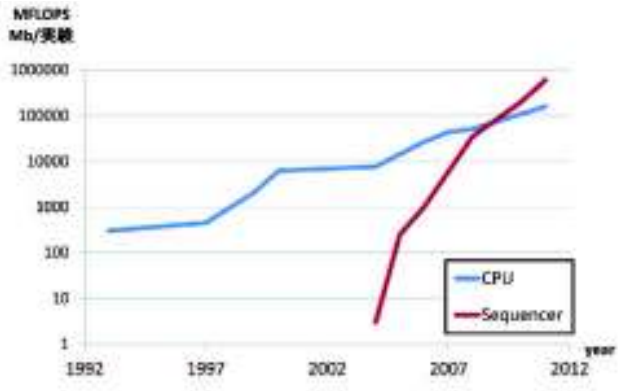


図2　計算機のCPUの速度とDNA配列を読む実験器具（DNAシーケンサー）の速度比較

2．完全線形符号

線形符号は情報を通信する際に用いられます．送信した情報の一部が通信中に誤って受信された場合に，一部の誤りを検出したり，誤りを訂正したりする事ができます．図3を用いて説明します．今，送りたい情報が2種類あり，それを真と偽だとします．これを通信するために，01のビット列との対応を取ります．ここでは，真を0000，偽を1111とします．この0000と1111の事を符号と呼びます．この符号を送受信に利用します．例えば，真という情報を送りたい場合，当方で0000を送信します．先方では0000を受信した後，送りたい情報と符号の対応表より，真という情報が送信されたと解釈します．符号を送信する場合，ノイズの影響により送信側と受信側とで符号が異なる事があります．例えば，0000を送信したのに，0100を受信した，といった具合です．この場合，0100は送りたい情報と符号の対応表から1111より0000の方に似ている（値が同じビットの数が0000の方が多い）ため，0000が送られたと判断され，誤り訂正が行われます．この結果，真が送信されたと解釈されます．また，0110を受信した場合，0000と1111のいずれからも2ビット異なるため，どちらが送信されたか判断できません．ただし，通信途上で誤りが発生した事がわかるため，これを誤り検出と呼びます．線形符号は，1) 送信に用いる符号と2) 受信された符号から送信された符号を推定する2つの役割を行う方法の一種です．生成行列と呼ばれる行列が1つ与えられ，行ベクトルの線形和を符号とする事から線形符号と呼ばれます．そして全ての受信語について最も似ている符号が1つしかない場合，完全線形符号と呼びます．

図3　通信路と符号，誤り訂正，誤り検出の例

AAAAA	GCAGA	TTTAA	CGTGA	GGGAA	ATGGA	CCCAA	TACGA
TCAAT	CAAGT	AGTAT	GTTGT	CTGAT	TGGGT	GACAT	ACCGT
GTAAG	AGAGG	CATAG	TCTGG	ACGAG	GAGGG	TGCAG	CTCGG
CGAAC	TTAGC	GCTAC	AATGC	TAGAC	CCGGC	ATCAC	GGCGA
TGATA	CTACA	ACTTA	GATCA	CAGTA	TCGCA	GTCTA	AGCCA
ATATT	GGACT	TATTT	CCTCT	GCGTT	AAGCT	CGCTT	TTCCT
CCATG	TAACG	GGTTA	ATTCG	TTGTG	CGGCG	AACTG	GCCCG
GAATC	ACACC	CTTTC	TGTCC	AGGTC	GTGCC	TCCTC	CACCC

図4　符号長5の4元完全線形符号の符号語．ただしDNA配列として表現

3．ガロア拡大体GF(4)

DNA配列は4つの塩基で構成されるため，私の研究では前節のような0,1の2元ではなく，4元の完全線形符号を用います．完全線形符号では，各元が加法と乗法の2つの演算が定義されている必要があります．これはガロア拡大体GF(4)の要素である事を意味します．4つの要素を$0, 1, \alpha, \alpha^2$とした場合，$1+\alpha+\alpha^2=0$となります．以降では，GF(4)の要素と塩基が$(0, 1, \alpha, \alpha^2)=(\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T})$と対応しているとします．

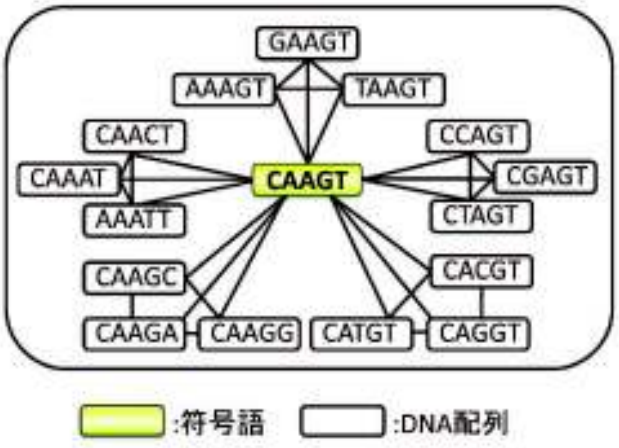


図5　符号語へと誤り訂正されるDNA配列の例

4．DNA配列の完全線形符号

道具がそろいましたので，DNA配列を完全線形符号として扱いたいと思います．ここで符号長5のGF(4)上の完全線形符号を使用します．符号長5なので，DNA配列の長さは5となります．図4に示す64個のDNA配列が符号語となります．長さ5のDNA配列の総数は，塩基種類4の5乗=1024種類あります． DNA配列を受信語として扱い誤り訂正を行う事により，1024個のDNA配列は64個のDNA配列へと復号されます．なお，復号前と後のDNA配列は高々一塩基の違いしかなく，また符号語DNA配列と一塩基違いのDNA配列は必ずその符号語に復号される事が保証されます．そのため，一つの符号語に復号されるDNA配列の数が符号語自身を含め16個（=1024/64）となります．図5に符号語CAAGTに復号される全DNA配列を記します．図中ではDNA配列を四角で表し，一塩基違いの関係にあるDNA配列間に辺をひいています．

5．類似DNA　配列探索への利用と探索空間の削減

第一節で述べましたが，データベースに登録されるDNA配列のデータ量は膨大です．このデータベースの使用法の一つに，手元にあるDNA配列と類似したDNA配列をデータベースから取り出す事が挙げられます．完全一致のDNA配列を見つけ出す事は容易なのですが，類似したDNA配列を取り出す事になるとその難易度は跳ね上がります．例えば「類似」の定義をDNA配列中の塩基が一つ異なるとします．長さnのDNA配列と完全一致するDNA配列は1種類ですが，類似したDNA配列は$1+3n$種類となります．同様に「類似」の定義をm

箇所異なるとした場合，類似したDNA配列は，$1+nCm\cdot 3^m$となります．コンピュータで類似したDNA配列を見つけ出すためには，上記のDNA配列の種類数を全て探索する必要があるため，特にmが大きい場合には膨大な計算時間を必要とします．逆にいうと，探索する必要のあるDNA配列数を減らす事ができれば，計算時間の削減，すなわち探索ソフトウェアの高速化を実現する事ができます．私の研究は，完全線形符号を適用する事で探索する必要のあるDNA配列数を削減可能である事を論理的に示す事にあります．

与えられたDNA配列を5文字ごとに区切り，完全線形符号の受信語として誤り訂正を行います．これによりDNA配列を完全線形符号の符号列に置き換える事ができます．この置換により探索する必要のある類似したDNA配列の種類数を大幅に削減する事ができます．仮に「類似」の定義を塩基が1箇所異なるとした場合，探索する必要のあるDNA配列数は$22.2+0.00879n$となります．従来の方法と比較した場合，定数項は大きくなっていますが，nの係数は340分の1程度になります．この2つの関係をグラフとして図6に示します．ここに示すように，DNA配列を完全線形符号の受信語とみなして誤り訂正を行い，符号語に基づき探索する事によって類似するDNA配列の探索の高速化が可能となる事がわかります．これが私の研究の成果です．

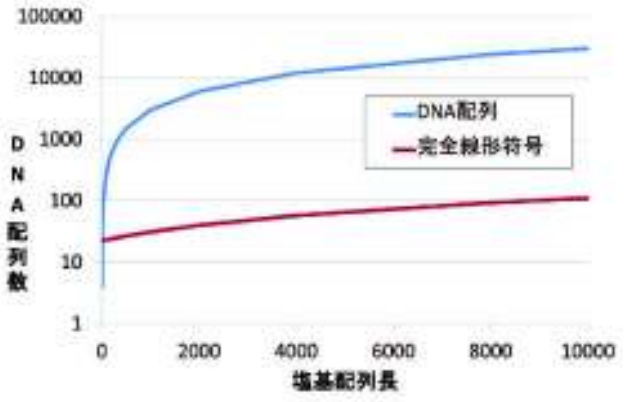


図6　塩基が1カ所異なるDNA配列を探索する場合に調べる必要のあるDNA配列数

6．おわりに

生物情報学の解析対象はDNA配列以外にも多く存在します．しかし基本且つ主流なのはDNA配列だと考えています．そのDNA配列の扱い方に関する本研究は，多くのアルゴリズムに影響を与える事ができます．本研究の成果が広まるよう努力していきたいと思っています．

参考文献

[1] Yip S. P., Sequence variation at the human ABO locus, Ann. Hum. Genet, **66**, 1-27, (2002)

[2] Needleman S. B., Wunsch C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biology, **48**, 443-453 (1970)

[3] Smith T. F., Waterman M. S., Identification of Common Molecular Subsequences, J. Mol. Biology, **147**, 195-197 (1981)

[4] Lipman D. J., Pearson W. R., Rapid and Sensitive Protein Similarity Searches, Science **227**, 1435-1441 (1985)

[5] Altschul S F. et al, Basic Local Alignment Search Tool, J. Mol. Biology, **215**, 403-410 (1990)

[6] Langmead B, Trapnell C, Pop M, Salzberg S. L., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol., **10**:R25 (2009)

[7] Takenaka T., Seno S., Matsuda S., Perfect Hamming Code with a hash table for faster genome mapping, BMC Genomics **12**:S8 (2011)