様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　７　日現在

機関番号：　１４４０１
研究種目：　挑戦的萌芽研究
研究期間：　2012～2015
課題番号：　２４６５０１５５

研究課題名（和文）完全線形符号に基づくＤＮＡの符号化によるゲノムマッピングの高速化

研究課題名（英文）Faster genome mapping method using 4-ary Perfect Linear Code

研究代表者
竹中　要一（takenaka, yoichi）

大阪大学・情報科学研究科・准教授

研究者番号：００３２４８３０

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）：高速なＤＮＡシーケンサー技術の発展により，生物のゲノム，遺伝子のDNA断片の読み取りを大量，かつ安価に取得する事が可能となった．読み取り速度の向上に伴うデータ産生速度は，計算機の計算速度向上度を大幅に上回っており，新たな解析アルゴリズムを必要としている．本研究は4元完全線形符号が有効である事を明らかにする事を目的とした．４元完全線形符号によって類似塩基配列を探索する際の解空間が大幅に削減される事を明らかにし，ゲノムマッピング，メタゲノム解析に有効である事を示した．

研究成果の概要（英文）：The emerging of the next generation DNA sequencers enables us to get an extraordinary amount of DNA short sequences and read their bases. The data size comes bigger and bigger year by year and the increase ratio overwhelms the Moore’s law. This requires new algorithms and mechanism to manipulate the short read faster. This research proposed that 4-ary perfect linear code meets the demand. The DNA sequences are coded to one of the code words of the perfect linear code and we proved that it reduces the search space when we try to find DNA subsequences that is one mismatch from the query DNA sequence. Then we have shown that the perfect linear code is useful to genome mapping and metagenome analyses.

研究分野：　生物情報学

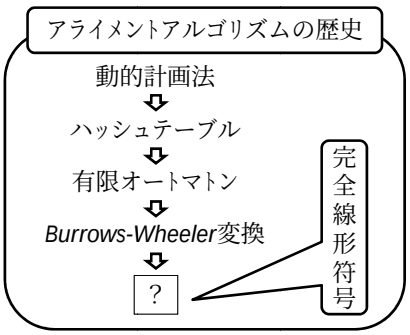
キーワード：　ゲノム　完全線形符号　次世代シークエンサー

１．研究開始当初の背景

(1) 高速なＤＮＡシーケンサー技術の発展により，生物のゲノム，遺伝子の DNA 断片の読み取りを大量，かつ安価に取得する事が可能となった．この DNA 断片の解析を行う為には，対象となるゲノム配列や遺伝子配列，メタゲノム由来の場合には多種多様なゲノム配列のいずれかにアライメントを行う必要がある．しかし，DNA シーケンサーから産生されるデータ量は年々指数関数的に増加しており，かつムーアの法則で表される計算機の計算速度向上度を大幅に凌駕している．そのため，常に高速なアライメントアルゴリズムの開発が求められている．これまで，動的計画法，ハッシュテーブル，有限オートマトン，Burrows-Wheeler 変換を用いた手法が実用に供されてきた．

(2)その中でも Burrows-Wheeler 変換を用いたアルゴリズムは，クエリとして与えられたＤＮＡ配列と完全一致するゲノム上の位置を特定するのに有効である．しかし，完全には一致しない配列の探索は不得手としており，Ｎ箇所の塩基が異なると計算時間がおおよそ「クエリの DNA 配列長×３Ｎ」倍になってしまう．そのため，塩基が異なる場合でも高速なゲノムマッピング手法が求められている．

２．研究目的



(1) 本研究の目的は，Burrows-Wheeler 変換に続くアライメントアルゴリズムの基礎を確立する事である．そして，そのアルゴリズムの基礎がゲノムマッピングに有効である事を示すとともに，ゲノムマッピング以外，例えば RNA-seq やメタゲノムにも応用可能である事を示す事である．

(2)そして，新しいアルゴリズムの基礎として，DNA の４種類の塩基を元とする完全線形符号が有効である事を示す事である．下に ACGT を $0, \alpha, \alpha^2, 1$ にコード化した場合の４元(5,3)完全線形符号の生成行列と検査行列を記す．また，4 元の演算にも記す．

生成行列

$$G = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & \alpha & 0 & 1 & 0 \\ 1 & \alpha^2 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} T & T & T & A & A \\ T & G & A & T & A \\ T & C & A & A & T \end{pmatrix}$$

検査行列

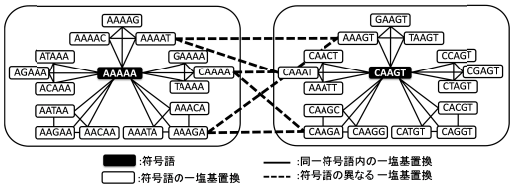
$$H = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & \alpha & \alpha^2 \end{pmatrix}$$

+	0	1	α	α^2
0	0	1	α	α^2
1	1	0	α^2	α
α	α	α^2	0	1
α^2	α^2	α	1	0

×	0	1	α	α^2
0	0	0	0	0
1	0	1	α	α^2
α	0	α	α^2	1
α^2	0	α^2	1	α

３．研究の方法

(1) まず DNA 配列を 4 元完全線形符号で表現するところから始める．本研究では，4 元(5,3)完全線形符号，及び 4 元(21,18)完全線形符号を対象とする．この完全線形符号とは，言うなれば，長さ５ないしは１８の塩基配列のクラスタリングである．ただし，各クラスタに所属する塩基配列は，中心となる一塩基配列から一塩基置換の関係にある（図参照）．



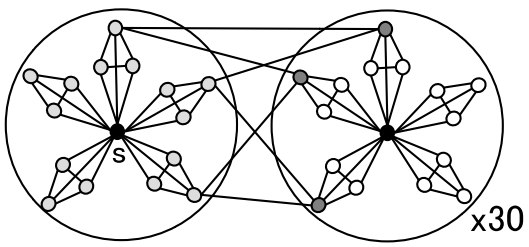
中心に位置する塩基配列は，この完全線形符号の符号語である．これを用いる事で完全一致しない DNA リードをゲノムへとマッピングする際の手間（計算手順）が削減される事を明らかにする．

(2)次に提案アルゴリズムを用いたゲノムマッピングの実装を行い，当該アルゴリズムの有効性を実証する．

(3)そして，ゲノムマッピング以外にも提案する完全線形符号の利用が有効である事を明らかとする．

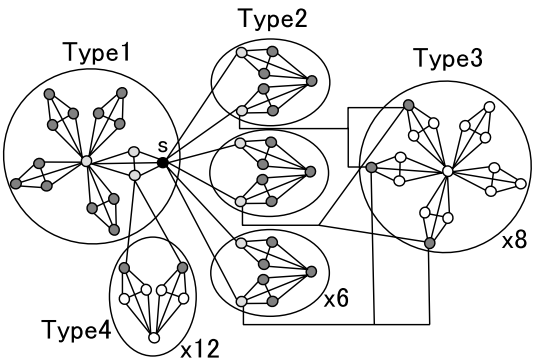
４．研究成果

(1)4 元完全線形符号の符号語，すなわち中心となる塩基配列の間にある関係性の分類を行い，これが完全一致ではないゲノムへ DNA リードをマッピングする際に有効である事を理論的に示した．具体的には，長さ N の DNA 配列にはどこか一箇所の塩基が異なる，すなわち一塩基置換配列が 3N 個ある．４元完全線形符号を用いる事により，この全 1+3N 個の配列を 22.2+0.00879N 個の符号語で表現する事が可能になる．これらの数値は類似 DNA



配列を探索する際の探索空間の広さに直結している．そのため，長さ 8 以上の DNA 配列の類似配列を探索する際に，完全線形符号が探索空間の削減に有効である事を明らかにした，

(2)次に，４元完全線形符号上の DNA 配列間の関係性を明らかにした．例えば，4 元(5,3)完全符号において，ある DNA 配列と２塩基置換の関係にある塩基配列との関係を図示する．ある DNA 配列が符号語ある場合，２塩基置換の関係にある全 DNA 配列は，隣接する 30 個のクラスタに位置する（前ページ下図）．ある DNA 配列が符号語ではない場合，２塩基



置換の関係にある DNA 配列の関係は，４種類のクラスタに分類される事を明らかにした．（下図）

(3) 当該手法を BWT を用いたゲノムマッピングソフトウェア Bowtie のラッパーとして実装を行うその性能評価を行った．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　4 件）

[1] Naoki Matsushita, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Metagenome fragment classification based on multiple motif-occurrence profiles, PeerJ (2 : e559). (201

[2] Ryo Araki, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, An estimation method for a cellular-state-specific gene regulatory network along tree-structured gene expression profiles, Gene(518), pp.17-25 (2013).

[3] 渡邉之人，瀬尾茂人，竹中要一，松田秀雄，複数時系列遺伝子発現プロファイルを利用した遺伝子制御ネットワーク推定の精度向上手法,情報処理学会論文誌 数理モデル化と応用，Vol.6，No.3，pp.151-162.(2013)

[4] Yoichi Takenaka, Shigeto Seno, Hideo Matsuda, Detecting shifts in gene regulatory networks during time-course experiments at single-time-point temporal resolution J. Bioinformatics and Computational Biology, vol. 13 Issue 5 (2015. Sep.)

〔学会発表〕（計　36 件）

[1]Yoichi Takenaka, All the 1+3n one-mismatch sequences of n-mer DNA are involved in 22.2+0.00879n strings of Perfect Linear Code words on DNA, International Conference on Intelligent Systems for Molecular Biology (ISMB2012), 2012 年 07 月 17 日，CA USA.

[2] Yoichi Takenaka, Perfect linear code reduces the solution space of genome mapping from 1+3n to 22.2 + 0.00879n to find one-mismatch for n-mer short reads, Special Interest Group Meetings of High-Throughput Sequencing at International Conference on Intelligent Systems for Molecular Biology (SIG-ISMB2012), 2012 年 07 月 13 日- 2012 年 07 月 14 日，CA USA.

[3] 竹中要一，ゲノムマッピング: Burrows-Wheeler transform の次のアルゴリズム，NGS 現場の会 第二回研究会，2012 年 05 月 25 日，大阪 日本

[4] Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Transcript-Type Dependent Normalization of Expression Levels in RNA-Seq Data for Non-Coding RNA Analysis, Joint Conference on Informatics is Biology, Medicine and Pharmacology (2012,Oct)

[5] Masakazu Sugiyama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Comparison of Gene Expressions measured by RNA-seq and Microarray for Transcriptome Analysis of Adipose Tissues, Joint Conference on Informatics is Biology, Medicine and Pharmacology (2012, Oct)

[6] Sho Ohsuga, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Two-stage method to infer gene regulatory network utilizing Link Prediction, Joint Conference on Informatics is Biology, Medicine and Pharmacology (2012, Oct)

[7] Tomoyoshi Nakayama, Yoshiyuki Kido, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Estimation of Dynamic Gene Regulatory Networks for Cell Differentiation by Splitting Time Course Data, Joint Conference on Informatics in Biology, Medicine and Pharmacology (2012, Oct.)

[8] Yukito Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Bayes-based

inference of gene regulatory network for multiple time series gene expression profile, Joint Conference on Informatics in Biology, Medicine and Pharmacology (2012, Oct.)
[9] Ryo Araki, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, An estimation method for a cellular-state-specific gene regulatory network along tree-structured gene expression profiles, Proceedings of the 2012 International Conference on Genome Informatics (GIW2012)
[10] Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, A Method for Isoform Prediction from RNA-Seq Data by Iterative Mapping,情報処理学会研究報告第 29 回バイオ情報学研究発表会 Vol.2012-BIO-29, No.13, pp.1-7 (2012, June)
[11] Sho Ohsuga, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda
Two-stage method to infer gene regulatory network utilizing Link Prediction
生命医薬情報学連合大会, B73_132,
[12] Masakazu Sugiyama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Comparison of Gene Expressions measured by RNA-seq and Microarray for Transcriptome Analysis of Adipose Tissues,生命医薬情報学連合大会, B72_115 (2012, Oct.)
[13] Tomoshige Ohno, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Transcript-Type Dependent Normalization of Expression Levels in RNA-Seq Data for Non-Coding RNA Analysis, 生命医薬情報学連合大会, B71_74 (2012, Oct.)
[14] Yukito Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Bayes-based inference of gene regulatory network for multiple time series gene expression profile, 生命医薬情報学連合大会, JSBi-29 (B70_29) (2012, Oct.)
[15] Tomoshige Ohno, Shigeto Seno, Hiromi Daiyasu, Yoichi Takenaka, Hideo Matsuda
Measuring Transcript-Type Dependent Expression Levels of ncRNAs in RNA-Seq Analysis, 第３５回分子生物学会年会 (2012, Dec.)
[16] 奥田華代, 竹中要一, 大野朋重, 瀬尾茂人, 松田秀雄, Improvement of the Accuracy of Mapping by Composing Alleles 情報処理学会 第75回全国大会, IB-6 (2013, Mar.)
[17] Tomoyoshi Nakayama, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Reconstruction of Dynamic Gene Regulatory Networks for Cell Differentiation by Separation of Time-course Data, The 2013 World Congress in Computer Science Computer Engineering and Applied Computing (2013 Jul.)
[18] Kensuke Suzuki, Daisuke Ueta, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, A Method of Sequence Analysis for High-Throughput Sequencer Data Based on Shifted Short Read Clustering, The 2013 World Congress in Computer Science Computer Engineering and Applied Computing (2013 Jul.)
[19] Yuta Okuma, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Gene Set Enrichment Analysis for a Long Time Series Gene Expression Profile, The 2013 World Congress in Computer Science Computer Engineering and Applied Computing (2013,Jul.)
[20] Yuta Okuma, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Gene Set Enrichment Analysis for Time-Series Gene Expression Profile, 35th Annual International IEEE EMBS Conference, 2013 IEEE EMBC, Short Papers, 3225 (2013 Jul.)
[21] Naoki Matsushita, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Multiplication of Motif Occurrence Profiles for Metagenome Fragment Classification, 35th Annual International IEEE EMBS Conference, 2013 IEEE EMBC, Short Papers, 3055 (2013, Jul.)
[22] Tomoshige Ohno, Shigeto Seno, Hiromi Daiyasu, Yoichi Takenaka, Hideo Matsuda
Integrative prediction of miRNA-mRNA interactions from high-throughput sequencing data, RECOMB/ISCB Conference on Regulatory and Systems Genomics, with DREAM Challenges 2013 (2013 Nov.)
[23] 竹中要一, 瀬尾茂人, 松田秀雄, 長いリードを計算機で扱うのに完全線形符号が効果的だと思うのです, 第三回 NGS 現場の会, 7-13-A (2013 Sep.)
[24] 河田愛明, 竹中要一, 瀬尾茂人, 松田秀雄, RNA-seq データを用いた Differential Alternative Splicing の検出ツールの比較考察, 第三回 NGS 現場の会, 3-28-A (2013 Sep.)
[25] 吉田拓真, 竹中要一, 瀬尾茂人, 松田秀雄,多次元尺度構成法によるリード分類結果の視覚化, 第三回 NGS 現場の会, 5-19-B (2013 Sep.)
[26] Tomoyoshi Nakayama, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Estimation method of large-scale dynamic gene regulatory networks for cell differentiation by separation of time-course data, 2013 年日本バイオインフォマティクス学会年会, 47 (2013 Oct.)
[27] 荒木 嶺, 瀬尾 茂人, 竹中 要一, 松田 秀雄, 時系列発現プロファイルを用いた時期特異的に機能する PPI サブネットワークの探索手法 第 36 回日本分子生物学会年会,

2P-1053 (2013 Dec.)
[28] Yoichi Takenaka, Shigeto Seno and Hideo Matsuda, Chronological analysis of regulatory strength on gene regulatory networks, 13rd European Conference on Computational Biology (ECCB2014), B04 (2014 Jul.)
[29] 津田絢子, 瀬尾茂人, 竹中要一, 松田秀雄, ベイジアンネットワークによる遺伝子制御ネットワーク推定結果の反復構築のための計算速度向上手法,情報処理学会研究報告, 第 98 回数理モデル化と問題解決研究会 2014-MPS-98-9 (2014 Jun.)
[30] 吉田 拓真, 瀬尾茂人, 竹中要一, 松田秀雄, リードの由来階級の既知・未知予測に基づくメタゲノム配列の系統分類手法, 情報処理学会研究報告, 第３９回バイオ情報学研究会 (2014 Sep.)
[31] Yoichi Takenaka, Shigeto Seno, Hideo Matsuda, Detecting the shifts of gene regulatory networks during time-course experiments with a single time point temporal resolution, Proceeding of GIW/InCoB2015, J13 (2015 Sep.)
[32] Yoichi Takenaka,Identification of the state-change-time-point of gene regulations from time-course experiments, The second workshop on Advanced Methodologies for Bayesian Networks (2015 Nov.)
[33] 上木怜, 瀬尾茂人, 竹中要一, 松田秀雄, ダイナミックベイジアンネットワークを用いた遺伝子制御ネットワーク推定の部分問題化による近似解法, 情報処理学会研究報告, 第 42 回バイオ情報学研究会 (2015 Jun.)
[34] 吉田拓真, 瀬尾茂人, 竹中要一, 松田秀雄, 共通 k-mer 種別数に基づくメタゲノム分類手法, 第 4 回 NGS 現場の会(2015 Jul.)
[35] 竹中要一, 風が吹くといつ桶屋は儲かるの？　遺伝子発現制御の時間変化解析, 第 4 回 NGS 現場の会 (2015 Jul.)
[36] 齋藤和正, 瀬尾茂人, 竹中要一, 松田秀雄, 小サンプル数データに対するベイジアンネットワークのスコアベース構造学習改善手法, 第 45 回バイオ情報学研究発表会 2016-BIO-45-1 (2016 Mar.)

６．研究組織

(1)研究代表者

竹中 要一 (TAKENAKA, Yoichi)
大阪大学・大学院情報科学研究科
・准教授
研究者番号：00324830

(2)研究分担者　なし

(3)連携研究者　なし