様式 C-19

# 科学研究費補助金研究成果報告書

平成 23 年 5 月 1 日現在

機関番号：11301
研究種目：若手研究（B）
研究期間：2009～2010
課題番号：21710207
研究課題名（和文）　ベクトル量子化による塩基配列の高速検索に関する研究

研究課題名（英文）　Research on Fast Search for DNA Sequence Using Vector Quantization

研究代表者
陳　キュウ（CHEN QIU）
東北大学・未来科学技術共同研究センター・准教授
研究者番号：00400292

研究成果の概要（和文）：
現状では、膨大なゲノムの配列データが GenBank、EMBL、DDBJ などのデータベースに蓄積されている。しかも、まだ遺伝子データベースのデータ量が指数関数的に増加している。ホモロジー検索は、進化・系統分類の解析、蛋白質の機能解析などを目的とした配列解析の最も基本的な手法の一つとなっている。現在最も頑健なアルゴリズムとして、Smith-Waterman (SW) アルゴリズムがあるが、その計算を行うことは時間的に現実的ではない。遺伝子データベースのデータ量が急速に増えている現状を考えると、さらに実行時間の大幅な増加を意味する。現状では、精度と検索速度が両立できる塩基配列の高速検索法はまだ実現されていない。本研究では、必要最小限の SW アルゴリズムによるアライメント処理と組み合わせたベクトル量子化による高精度かつ高速な塩基配列の検索手法を実現した。

研究成果の概要（英文）：
The enormous quantity of DNA sequence data has been accumulated in the database like GenBank, EMBL, and DDBJ, etc. Moreover, the volume of data still increases in exponential. Homology search of DNA sequences is the most important task in the life science area. In this research, we propose an efficient hierarchical DNA sequence search method to improve the search speed while the accuracy is being kept constant. For a given query DNA sequence, firstly, a fast local search method using histogram features is used as a filtering mechanism before scanning the sequences in the database. A large number of DNA sequences with low similarity will be excluded for latter searching. The Smith-Waterman algorithm is then applied to each remainder sequences. Experimental results using GenBank sequence data show the proposed method combining histogram information and Smith-Waterman algorithm is more efficient for DNA sequence search.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2009 年度 | 1,300,000 | 390,000 | 1,690,000 |
| 2010 年度 | 1,100,000 | 330,000 | 1,430,000 |
| 年度 | | | |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 2,400,00 | 720,000 | 3,120,000 |

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・システムゲノム科学
キーワード：バイオインフォマティクス、ベクトル量子化、塩基配列、高速検索、
データベース、ヒストグラム特徴

１．研究開始当初の背景

2003年4月、ヒトゲノムの30億塩基の全配列決定に国際協力で取り組んでからほぼ15年、その解読が遂に完了になった。これは生命科学のアポロ計画と呼ばれたヒトゲノム計画の最重要目標の達成である。現在、各国の研究者らが、シーケンスの意味付け、たんぱく質の構造と機能解析、遺伝子およびたんぱく質ネットワークの解明にしのぎを削っており、日々、新しい遺伝子配列が明らかになっていて、膨大なデータがデータベースに蓄積されている。しかも、まだ遺伝子データベースのデータ量が指数関数的に増加している。

ゲノムの配列データは、DNA を構成する４種類の塩基 G(グアニン)、A(アデニン)、T(チミン)、C(シトシン)で表現される文字列である。遺伝子 A と遺伝子 B の間のホモロジー(相同性) が高いというのは、一般的に両者が共通の祖先遺伝子から由来している可能性が高いことを意味する。遺伝子 A が遺伝子 B と高いホモロジーをもつことが分かれば、遺伝子 A の機能が遺伝子 B の機能とよく似ていることが推測できる。なので、ホモロジー検索は、進化・系統分類の解析、蛋白質の機能解析などを目的とした配列解析の最も基本的な手法の一つとなっている。

２つの配列の類似度を計算するときは、進化の過程における塩基やアミノ酸の挿入・欠失に対応したギャップを考慮する必要があるので、さまざまなアラインメントを探索して、スコアが最大となるものを用いる。スコアを最大にする最適なアラインメントはダイナミックプログラミング (DP) と呼ばれる手法で計算することができる。そのなか、現在最も頑健なアルゴリズムとして、Smith-Waterman (SW)アルゴリズムがある。概要としては相同性の指標として重み付きの編集距離にギャップペナルティー値を加えたものを用い、その元で相同性をダイナミックプログラミングで計算する。この計算時間は比較する2 つ配列のサイズをN、Mとした場合、積O(NM)となるため、データベースの配列のすべてに対して一つ一つこの手法を適用すると膨大な時間を要する。それに、上述したように、遺伝子データベースのデータ量が急速に増えている現状を考えると、さらに実行時間の大幅な増加を意味する。その計算を行うことは時間的に現実的ではない。

このため、実際には近似手法が用いられ、BLASTとFASTAがその代表的なものである。たとえば、NCBIが開発したBLASTアルゴリズムは、配列を固定長の断片 (ワード) に区切り、ワード単位で類似する断片を検索し、これらを類似度が最大になるまで両方向に伸ばして、局所的なアラインメントを行う領域を決定し、最後に、その部分だけのローカルなアラインメントはダイナミックプログラミングを用いて行う手法である。同じような近似手法であるFASTAより高速である。しかし、BLASTとFASTAなどの手法は高速化のために対象データの一部を省略するヒューリスティックな処理を行い、ギャップを入れていないため、精度が高く要求される部分的な相同性の比較ではエラーが発生する可能性が高い。

以上述べたように、現状では、精度と検索速度が両立できる塩基配列の高速検索法はまだ実現されていない。

我々はこれまでの処理手法とまったく異なる新しい概念の「ベクトル量子化コードブック空間情報処理」手法を提案し、それによる顔画像人物認識技術を開発してきた。ベクトル量子化は、圧縮分野での有名なアルゴリズムの一つであり、複数の標本値をまとめて量子化することにより、標本値間の冗長度を情報圧縮に利用したものであり、標本値を一つ一つ量子化するスカラ量子化において生じる冗長性を低減することができる。ベクトル量子化処理により、入力情報を圧縮してヒストグラム情報を顔画像の特徴情報として抽出することが可能になる。この原理に基づいて、我々は「ベクトル量子化コードブック空間情報処理」による大規模顔データベース FERET で Top1 認識率 97.4%を実現し、簡単かつ高い認識率を持つ顔認識技術を実現した。一枚の顔画像に対する全認識処理時間は、汎用なパーソナルコンピュータを用いた場合 31ms であり、ビデオレートでの高速認識を実現している。

これらの研究成果は、上述塩基配列の高速検索技術実現に向けてその基礎となる重要な成果である。

２．研究の目的

本研究では、①「ベクトル量子化コードブック空間情報処理」を用いた認識技術、②我々が提案した顔画像を31ｍｓのビデオレートで認識できる高速認識手法に基づき、4種類の塩基からなる塩基配列を小さい配列(例えばACT、CGGなど)に分割し、一つの配列を一つの3次元のベクトルと見なす。それから、ベクトル量子化処理することにより高速に塩基配列のヒストグラム特徴を生成し、それを塩基配列の特徴量として使う。検索するときに、同じようにベクトル量子化処理が施されたデータベース中のすべての塩基配列のヒストグラム特徴との類似度を計算し、あらかじめ設定した閾値と比較し、閾値を上回る塩基配列のみに対して必要最小限のSWアルゴリズムによるアライメント処理を行って、スコアを計算する。アライメント処理の範囲が限定されるため、検索精度を保ちながら、高速に検索を実現することが可能である。

従来の高速検索手法の代表として BLAST アルゴリズムは固定長（DNA では 11）の全ての類似単語のリストを生成し、ある閾値以上の単語ペアを探し、それをもとに両側に伸長させる手法である。この手法では、塩基配列の順番を考えなければならないので、ギャップは入らないため、精度が高く要求される部分的な相同性の比較ではエラーが発生する可能性が高い。それに対して、ベクトル量子化処理で生成されたヒストグラム特徴による検索手法は塩基配列の順番を利用しないため、ギャップの影響をほとんど無視できる。この利点から、提案手法は BLAST アルゴリズムにより精度の高い検索ができると予想される。また、ベクトル量子化処理によるヒストグラム特徴の生成およびヒストグラム特徴同士の類似度計算は非常に高速に実行できるため、BLAST アルゴリズムに上回る速度で検索できる手法になると考えられる。

また、我々は「ベクトル量子化コードブック空間情報処理」による大規模顔データベース FERET で Top1 認識率 97.4%を実現し、簡単かつ高い認識率を持つ顔認識技術の実現により、ベクトル量子化ヒストグラム特徴のロバスト性が証明された。対象を 2 次元の顔画像から一次元の配列に拡大して適用し、必要最小限の SW アルゴリズムによるアライメント処理と組み合わせる検索手法は SW アルゴリズムだけ利用する検索手法と比較しても、遜色のない精度で検索できると予想される。

## ３．研究の方法

### (1) 塩基配列の特徴量抽出手法

本研究では、①「ベクトル量子化コードブック空間情報処理」を用いた認識技術、②我々が提案した顔画像を31ｍｓのビデオレートで認識できる高速認識手法に基づき、対象を2次元の顔画像から１次元の配列に適用する。図1に示すように、まず、未知の塩基配列を入力し、ｎ個の部分に空間分割する。これは塩基配列の順序を考慮するためである。ベクトル量子化によるヒストグラム特徴はギャップの影響を無視できるのは配列の順序を利用しないためである。一方、ある程度塩基配列の順序情報を入れれば、よりロバスト性の高い特徴量が抽出できると考えられる。分割された各々の塩基配列に対して、4種類の塩基からなる塩基配列を小さい配列(例えばACT、CGGなど)に分割し、一つの配列を一つの3次元のベクトルと見なす。この処理は全配列に亘って、オーバーラップして行う。

それから、ベクトル量子化処理を行う。従来のベクトル量子化と違い、あらかしめコードブックを用意する必要せずに、3 次元のベクトルと 0～63 までのインデックス番号との対応付けでベクトル量子化が済む。あらかじめ 64 という小さいサイズの 3 次元塩基ベクトルとインデックス番号の参照テーブルを用意しておけば、ベクトル量子化は Look－Up 法で非常に高速で実行できる。それから、各インデックス番号の使用頻度を数えれば、簡単に塩基配列のヒストグラム特徴を生成し、それを塩基配列の特徴量として使う。ｎ個の部分に分割された塩基配列はｎ個のヒストグラムが生成される。

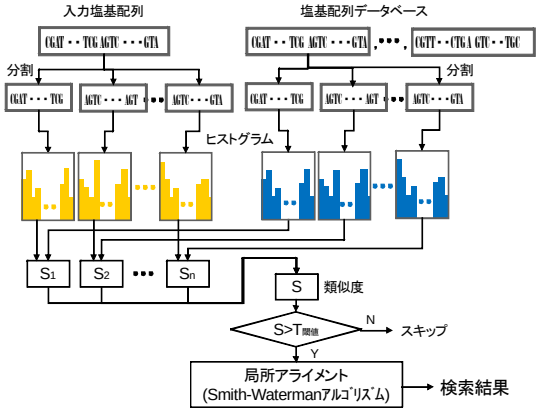


図 1.　塩基配列高速検索手法の原理図

### (2) 塩基配列の高速検索手法

入力塩基配列各部分のヒストグラム特徴が生成されたあと、データベース中のすべての塩基配列に対して、同じようにｎ個の部分に分割し、ベクトル量子化処理により各分割部分のヒストグラムを作成する。続いて、データベース中の塩基配列のヒストグラム特徴との類似度を計算する。

各分割部分のヒストグラム同士の類似度を計算し、得られた各々の類似度を組み合わせて配列の総合類似度にする。類似度指標として重なり率が使われる。得られた総合類似度をあらかじめ設定した閾値と比較し、閾値を上回る塩基配列のみに対して、必要最小限のアライメント処理を行う。

現在最も頑健なアルゴリズムとして SW アルゴリズムは相同性の指標として重み付きの編集距離にギャップペナルティー値を加えたものを用い、その元で相同性をダイナミックプログラミングで計算する手法である。それによるアライメント処理を行って、スコアを計算する。アライメント処理の範囲が限定されるため、検索精度を保ちながら、高速に検索を実現する。

## ４．研究成果

本研究では、必要最小限のSWアルゴリズムによるアライメント処理と組み合わせたベクトル量子化による高精度かつ高速な塩基配列の検索手法を試みた。本研究の研究成果は以下のようになる。

高速検索するための塩基配列の特徴量抽出手法を開発した。塩基配列のヒストグラム

特徴を生成するため、従来のベクトル量子化と違い、あらかじめ 64 という小さいサイズの 3 次元塩基ベクトルとインデックス番号の参照テーブルを用意することにより、ベクトル量子化は非常に高速で実行できた。それから、各インデックス番号の使用頻度を数え、簡単に塩基配列のヒストグラム特徴を生成できた。

配列の長さ不一致の対応策として、ローカル検索手法を導入し検討した。入力塩基配列を小さいサイズの塩基配列に分割し、各々の部分配列はデータベース中の塩基配列に最も似ている部分を探し出し、そこから類似度が上がらないまで左右伸長させ、部分配列の類似度を計算する。得られた各々の類似度を組み合わせて配列の総合類似度にし、あらかじめ設定した閾値と比較し、閾値を上回る塩基配列のみに対して、必要最小限のアライメント処理を行う。ローカル検索によって、配列間の似ている部分だけを照合させるので、検索のロバスト性が図れる。提案手法の有効性を検証するため、検索範囲を減らす実験を行った。世界的な公共の塩基配列データベースである GenBank のサブデータベースを利用し、全体 853,825 個(配列長 400～2000)の塩基配列を全検索するのと比べ、提案手法を使って、約 0.269%に当たる 2301 個の塩基配列を検索するだけで同じ結果が得られた。SW 手法を使って塩基配列の全体検索でかかった約 2 時間で、提案手法は約 0.52%に当たる約 37.4 秒で同じ結果を得ることができた。また、代表的な高速手法である BLAST により 2.78 倍速くなった。

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 2 件）


[1] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “A Fast Search Method for DNA Sequence Database Using Histogram Information”, *International Journal of Bioinformatics Research*, Vol. 3, Issue 1, pp. 161-166, 2011. (査読有)

[2] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “A Codebook Design Method for Robust VQ-based Face Recognition Algorithm”, *Journal of Software Engineering and Applications*, Vol. 3, No. 2, pp. 119-124, 2010. (査読有)


〔学会発表〕（計 4 件）


[1] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “An Improved Fast Search Method Using Histogram Features for DNA Sequence Database,” *Proc. of the Int’l Conf. on Computer and Information Science (ICCIS 2010)*, pp. 237-240, Amsterdam, Netherlands, Sep. 29, 2010. (査読有)

[2] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “A Local Search Method Using Histogram Features for Fast Retrieval of DNA Sequences”, *Proc. of the 2010 Int’l Conf. of Information Engineering (ICIE 2010)*, pp. 395-398, London, U.K., Jul. 2, 2010. (査読有)

[3] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “Facial Image Recognition Using VQ Histogram in the DCT Domain”, *2010 Int’l Conf. on Digital Image Processing (ICDIP 2010)*, edited by K. Jusoff, Y. Xie, *Proc. of SPIE*, Vol. 7546, 75460J, Singapore, Feb. 27, 2010. (査読有)

[4] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “A Fast Retrieval of DNA Sequences Using Histogram Information”, *Proc. of 2009 Int’l Conf. on Future Information Technology and Management Engineering (FITME 2009)*, pp. 529-532, Sanya, China, Dec. 13, 2009. (査読有)


〔図書〕（計 1 件）


[1] Qiu Chen, Koji Kotani, Feifei Lee, and Tadahiro Ohmi, “Face Recognition Using Self-Organizing Maps”, *Self-Organizing Maps*, Edited by George K. Matsopoulos, ISBN 978-953-307-074-2, pp. 277-288, In-Tech, Vienna, Austria, 2010.


６．研究組織

(1)研究代表者


陳　キュウ（CHEN QIU）
東北大学・未来科学技術共同研究センター・准教授
研究者番号：00400292


(2)研究分担者

（　　）

研究者番号：

(3)連携研究者

（　　）

研究者番号：