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　　５　月　１９　日現在

機関番号：　１０１０１
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2015
課題番号：　２５２８００７９

研究課題名（和文）巨大シーケンス内の類似繰り返し構造の分析

研究課題名（英文）Analysis of Repetition Structure in Huge Sequences

研究代表者
中村　篤祥（Nakamura, Atsuyoshi）

北海道大学・情報科学研究科・准教授

研究者番号：５０３４４４８７

交付決定額（研究期間全体）：（直接経費）　　6,400,000円

研究成果の概要（和文）：DNAなどの巨大シーケンス内に存在する散在反復配列を抽出する方法として，頻出近似文字列パターンを列挙するアルゴリズムを開発し，列挙されたパターンの出現領域を抽出する方式を考案した．提案手法のパターンは出現領域の境界がはっきりしているため，同じ領域を重ねて数えることが少ない．また，提案列挙アルゴリズムは高速かつ省メモリーで動作する．ヒトゲノムの２１番染色体に適用したところ，著名な散在反復配列であるAlu配列の出現領域とされる領域の約５０％を，１００個の代表パターンの出現領域として抽出することに成功した．

研究成果の概要（英文）：We developed an algorithm for enumerating frequent approximate string patterns, and proposed a method of extracting occurrence regions of the enumerated patterns as a method of extracting interspersed repetitive elements in a huge sequence like a DNA sequence. Patterns of proposed methods have occurrences of clear boundaries, so there is little chance to count essentially the same region more than once. Furthermore, our enumeration algorithm runs very fast and with small memory. According to our empirical results using human chromosome 21, a half of the known Alu regions, which are famous interspersed repetitive elements, is extracted as occurrence regions of 100 representative patterns that were selected from enumerated frequent approximate patterns.

研究分野：　知能情報学

キーワード：　知識発見とデータマイニング　シーケンスマイニング　ゲノム情報処理　頻出パターンマイニング

１．研究開始当初の背景

(1) 生物の DNA 配列内には散在反復配列が多数存在することが知られている．それらの散在反復配列を発見する方法として，最初に類似部分列のペアを見つけてそれらをクラスタリングして分類する手法が用いられてきたが，組織的な探索ではないため探索漏れが生じている可能性があった．

(2) 頻出パターンマイニングの分野でも，固定サイズのウィンドウを決めてパターン部分列が出現するウィンドウ数をカウントする方法やハミング距離がある基準を満たす同じ長さの連続部分列をカウントする方法が開発されていたが，前者にはウィンドウサイズを超える長さのパターンが抽出できないという問題があり，後者には長さの異なる類似部分列が多くあるパターンを抽出できないという問題があった．

２．研究の目的

(1) シーケンス内に頻出する近似パターンを組織的に探索する有効な手法を開発する．

(2) 巨大シーケンスを扱う実際の問題に適用可能な高速なアルゴリズムを開発する．

(3) 実際の問題に適用し，新たな頻出パターンやその出現領域を発見する．

３．研究の方法

(1) 近似文字列パターンの出現境界に曖昧性がない，頻出近似文字列パターンマイニング問題への定式化を行う．

(2) 定義した頻出近似文字列パターンマイニング問題を効率的に解くアルゴリズムを開発する．DNA シーケンスに適用し，実際に巨大シーケンスが扱えるかチェックする．

(3) DNA シーケンスから既知の散在反復配列の抽出ができるかチェックすると共に，新たな散在反復配列の発見を試みる．

４．研究成果

(1)近似文字列パターンにおいて，従来の出現の定義は，出現境界に曖昧性があり，出現領域を抽出するのに不向きであった．それは，類似文字列の領域は左右に何文字かずらしても類似しているという事実に起因する．そこで，パターンとの局所アライメントにおいて局所最適な部分文字列のみを考え，さらに極小性を満たすもののみを出現とする方式を提案した．図１は，一致していたら１，不一致ならば−１というスコア関数を用いて文字列“□abacb”の局所最適な出現を列挙した結果である．極小性を満たすのは s[1..4]と s[11..16]のみであり，この２つの出現領域に重なりはなく，本質的に同じ領域を重ねて抽出する可能性は減っている．

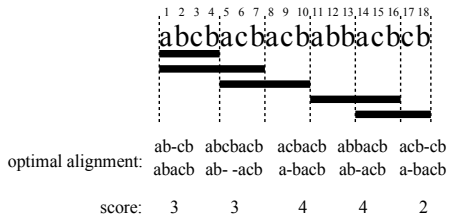


図１：*パターン文字列 abacb の文字列 s における局所最適な出現文字列．S[1..4]=“□abcb”および s[11..16]=“□abbacb”のみが極小性を満たす．*

これにより散在反復配列の抽出を，発見された近似文字列パターンの出現を抽出することにより行えるようになった．

(2)極小局所最適な出現のみをカウントする頻出近似文字列パターンマイニング問題を定義し，与えられた回数以上出現する近似パターンを列挙するアルゴリズム ESFLOO を提案した．ESFLOO は，パターンの候補としては高速性のために対象文字列内に出現する部分文字列のみを考え，接尾辞木を用いて重複なく生成する．文字列の長さ n に対し，ESFLOO は$O(n^3)$時間$O(n^3)$空間で動作する．ESFLOO は１０万程度の長さの文字列に対しては動作したが，１００万以上の長さ文字列ではメモリー不足で適用できなかった．そこで k 個までしかギャップを許さないアライメントでの局所最適な出現を定義し，その定義を用いて頻出パターンを列挙する$O(k^2n^3)$時間$O(k^2n^2)$空間アルゴリズム ESFLOO-kG を開発した．さらに，最悪ケースでは同じ時間・空間計算量だが，実問題では高速かつ省メモリー

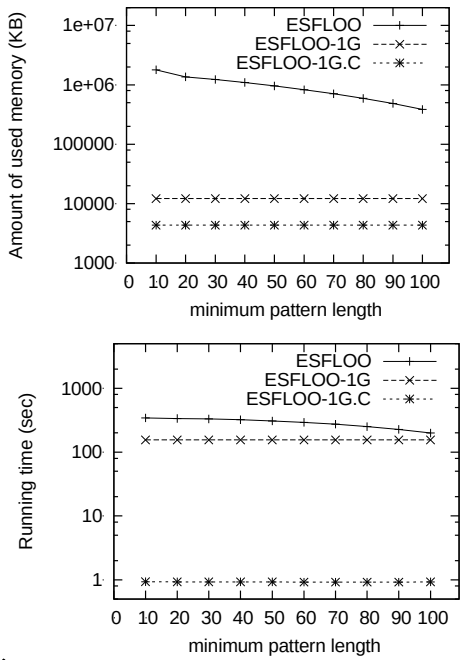


図２：*開発した３つのアルゴリズム ESFLOO, ESFLOO-kG および ESFLOO-kG.C の出力パターン最小長に対するメモリー使用量(上)と計算時間(下)の変化．*

である候補ベースのアルゴリズム ESFL00-kG.C を開発した.

図２に長さが約 20100 で長さ 100 のパターンの類似文字列が約 200 間隔で 100 回埋め込まれた４種類の文字からなるランダム文字列対し，３つのアルゴリズムを適用したときのメモリー使用量と計算時間を示す．横軸はパラメータとして与える出力パターン最小長である. ESFL00 では両端の局所最適性をチェックするために，片側が最適なものをメモリーに記憶しておく必要があり，そのため出力パターン最小長を小さくすると，メモリー使用量が増加する．アライメントのギャップの数を１つまでに限定した残り２つのアルゴリズム ESFL00-1G と ESFL00-1G.C は，両端の局所最適性を一度に行うことができるため，記憶しておく必要がなく，出力パターン最小長にメモリー使用量は依存しない．出力最小パターン長が 100 のところで比較すると，ESFL00 と比べて ESFL00-1G，ESFL00-1G.C はそれぞれメモリー使用量が 31.6 倍，89.5 倍少なく，計算時間は 169 倍, 216 倍速かった．文字列の長さに対する計算時間の増加の割合を実験的に調べると ESFL00 が $\Theta(n^{2.07})$ であるのに対し ESFL00-1G が $\Theta(n^{1.14})$，ESFL00-1G.C が $\Theta(n^{0.877})$ であり，k が小さいとき ESFL00-kG.C は非常に高速なアルゴリズムであることが確認できた.

実際にヒトゲノムの 21 番染色体(長さ約 4700 万塩基対(連続する N 以外の有効部分は 3510 万塩基対)に ESFL00-kG.C 適用し，３０回以上出現するパターンをメモリー45GB のコンピュータで求めた結果，k=0,1,2 でそれぞれ 43 分，９時間 17 分，45 時間 55 分で動作し，長い DNA にも適用可能な速度であることが確認できた.

(3) ヒトゲノムの 21 番染色体に ESFL00-kG.C を k=0,1,2 に対して適用し，長さが 100 以上で 30 回以上出現するパターンを抽出した．比較のため，完全一致する部分文字列パターンを抽出するアルゴリズム(Exact)も実行した.

表１に抽出されたパターンの数を示す．完全一致パターン(Exact)は３０万しか抽出されなかったが近似パターン(ESFL00-kG.C)は 5000 万以上抽出されている. 抽出されたパターンの内 0.5-寛容閉パターン (0.5-tolerance closed pattern: 包含するパターンで頻度が 0.5 倍以上のものがないパターン)のみ選びそれらを頂点とし，編集距離ベースの類似度を用いて重み付けした辺からなるグラフに対し，正規化スペクトラルクラスタリングを用いて頂点を 100 のクラスタに分類し，それぞれのクラスタから中心に最も近いパターンを抽出した.

表１: *抽出パターンの数と長さ*

アルゴリズム	パターン数	パターン長 (クラスタリング後)	
		[min,max]	平均
Exact	306,662	[12,25]	13
ESFL00-0G.C	55,250,250	[100,383]	152
ESFL00-1G.C	87,952,197	[100,292]	175
ESFL00-2G.C	105,466,125	[100,369]	191

表２：*抽出パターンの出現の数と長さの総計*

アルゴリズム	パターン数	出現数	出現長総計
Exact	100	18,510	252,710
ESFL00-0G.C	100	8,472	1,138,398
ESFL00-1G.C	100	9,768	1,617,957
ESFL00-2G.C	100	9,377	1,633,905
RepeatMasker-All	29,163	59,720	16,606,741
RepeatMasker-Alu	137	12,561	3,311,757

100 パターンの長さの最小，最大および平均を表１に示す．完全一致パターンの長さの平均は 13 しかなく，最大でも 25 のものしか取れてきていない．それに対し，近似パターンは平均の長さが 150 以上で最大 290 以上のものが取れてきている.

抽出されたパターンの出現を元の DNA シーケンスから取り出した結果と Repbase Update (20130422 版)に登録されている反復パターンの出現を RepeatMasker で取り出した結果を比べてみた．比較対象の方式はパターンの抽出を行っておらず，すでに登録されているパターンを用いていることに注意されたい.

表２にそれぞれのアルゴリズムにより抽出されたパターンの出現数と出現長総計を示す. 出現長総計では完全一致パターンが 25 万程度であるのに対し，近似パターンでは 100 万以上となってる．RepeatMasker-All では，Repbase Update に登録されているパターンをすべて使って出現を抽出たものであり，Repbase-Alu では，Alu 配列のみをパターンとして使って出現を抽出している．ESFL00-1G.C および ESFL00-2G.C では，Repbase Update に登録されているパターンの出現領域と比べて, 1/10 程度の領域が抽出できており，Alu 配列パターンのみの出現領域と比べると約半分程度の領域が抽出できている.

長さだけでなく実際に既知のパターンの出現領域が提案法で抽出できているかを示したのが表３である. RepeatMasker により抽出された出現領域に対し，提案法で抽出したパターンの出現領域が含まれる率(包含率)と覆っている率(被覆率)を求めた．完全一致パターンは RepeatMasker で抽出された領域以外の領域も含んでいるが，ESFL00-kG.C は

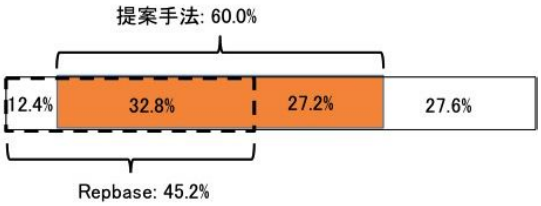
表３：*既知のパターンの出現領域の包含率と被覆率*

アルゴリズム	RepeatMasker-All		RepeatMasker-Alu	
	包含率	被覆率	包含率	被覆率
Exact	71.4%	1.1%	42.4%	3.2%
ESFL00-0G.C	100.0%	6.9%	96.4%	33.1%
ESFL00-1G.C	100.0%	9.7%	100.0%	48.8%
ESFL00-2G.C	100.0%	9.8%	94.1%	46.4%

抽出したすべての領域が RepeatMasker で抽出された領域に含まれている．また Alu 配列に限れば，既知の Alu 配列パターンの出現領域の半分近くが提案法により抽出できていることがわかる．これは，提案法が新しい散在反復配列発見に使うことができる可能性を示した結果といえる．

(4) 最小サポートを指定する頻出パターンマイニングの欠点は，パターン長により出現のしやすさが違うのに頻出基準として同じ最小サポートを用いることである．そこでランダム文字列における出現頻度分布での上側 100α%点を最小サポートとし，パターン長毎に計算した最小サポートを用いて頻出パターンマイニングを行う方法を提案した．
ヒトゲノムの 21 番染色体に対し，$\alpha=1.0\times10^{-23}$ に設定して ESFL00-0G.C を適用した．その結果長さ 7～3180 のパターンが抽出された．図３に抽出されたパターンの出現により覆われた領域の情報を示す．Repbese Update の全登録パターンの出現により覆われる領域は全体の 45.2%であるのに対し，提案手法では 60.0%の領域を覆った．Repbase Update で覆われた領域の 72.6%をカバーしているが，覆われていない領域の 49.6%を抽出している．それらの領域には，何らか意味があるのかどうか今後調査する予定である．

図３：*ランダムシーケンスにおける出現頻度分布での上側 100α%点を最小サポートに用いる方法で抽出されたパターンの出現により覆われる領域*



(5) 逐次意思決定問題はオンライン学習の問題であり，時系列シーケンスにおいて繰り返し起こるパターンの学習とみることもできる．バンディット問題はそのような逐次意思決定問題で，自分が選んだところしか観測できない部分情報問題である．バンディット問題において，各時刻にマッチングを選択する問題のリグレット上界の改善や，全情報と部分情報を繋ぐ理論の構築，常に損失のない選択肢がある場合のリグレット下界の証明等の成果を得た．

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　9 件）

① Atsuyoshi Nakamura, Ichigaku Takigawa, Hisashi Tosaka, Mineichi Kudo, Hiroshi Mamitsuka, “□Mining approximate patterns with frequent locally optimal occurrences”, Discrete Applied Mathematics 200, 査読有, 2016, 123-152,
DOI: 0.1016/j.dam.2015.07.002

② Ryo Watanabe, Atsuyoshi Nakamura, Mineichi Kudo, “An improved upper bound on the expected regret of UCB-type policies for a matching-selection bandit problem”, Operations Research Letters 43(6), 査読有, 2015, 558-563,
DOI: 10.1016/j.orl.2015.08.008

③Koji Ouchi, Atsuyoshi Nakamura, Mineichi Kudo, “An efficient construction and application usefulness of rectangle greedy covers”, Pattern Recognition 47(3), 査読有, 2014, 1459-1468,
DOI: 10.1016/j.patcog.2013.09.008

〔学会発表〕（計　9 件）

①中村武憲, 中村篤祥, 工藤峰一, “ランダム仮説下での出現回数分布を利用した散在反復配列の発見”, 第 8 回データ工学と情報マネジメントに関するフォーラム, 2016/03/01, ヒルトン福岡シーホーク(福岡県・福岡市).

②中村篤祥, “Hedge と Exp3 アルゴリズム間の橋渡し”, 第 21 回情報論的学習理論と機械学習研究会, 2015/06/24, 沖縄先端科学技術大学院大学(沖縄県・国頭郡恩納村).

③中村篤祥, “インターネットにおけるオンライン学習”, 第 12 回情報科学技術フォーラム, 2013/09/04, 鳥取大学(鳥取県・鳥取市).

６．研究組織
(1)研究代表者
中村　篤祥（NAKAMURA, Atsuyoshi）
北海道大学・大学院情報科学研究科・准教授
研究者番号：５０３４４４８７

(2)研究分担者
工藤　峰一（KUDO, Mineichi）
北海道大学・大学院情報科学研究科・教授
研究者番号：６０２０５１０１

瀧川　一学（TAKIGAWA, Ichigaku）
北海道大学・大学院情報科学研究科・准教授
研究者番号: １０３７４５９７

(3)連携研究者
馬見塚　拓（MAMITSUKA, Hiroshi）
京都大学・化学研究所・教授
研究者番号：００３４６１０７

喜田　拓也（KIDA, Takuya）
北海道大学・大学院情報科学研究科・准教授
研究者番号：７０３４３３１６

大久保　好章（OKUBO, Yoshiaki）
北海道大学・大学院情報科学研究科・助教
研究者番号：４０２７１６３９