Vol.2022-BIO-72 No.6
2022/11/29

Inexact seedを用いた散在反復配列検出手法の開発

武田 淳志[1,3,a)]　野中 大輔[2]　今津 裕太[2]　福永 津嵩[1]　浜田 道昭[1,3,b)]

概要：多くの真核生物のゲノムに大量に存在する散在反復配列を, データベースなしで高感度に検出するソフトウェア, REPrise(REPeat Recognition using Inexact_Seed-and-Extend) を開発した. REPrise は, 既存手法 RepeatScout をベースに, 以下の三点の改良を施した. (1)Seed-and-extend algorithm において, seed に文字の置換を許容する inexact seed を採用した. (2)Extension alignment に affine gap を導入した. (3) 一度検出した repeat に含まれた seed を除去する masking step において,seed を除去されにくくした. ベンチマークの結果, REPrise は RepeatScout よりも高感度に散在反復配列を検出することを示した.

Interspersed repeat detection using inexact seed

1. 背景

2022 年 4 月, T2T コンソーシアムがヒト完全長ゲノム配列の解読を発表した [1]. 今後様々な種で完全長のゲノム配列が読まれる中で, 次に必要な処理は読まれたゲノムのアノテーションである. 散在反復配列は, 多くの真核生物のゲノムに大量に存在し, 例えばヒトでは全ゲノムの内 54% を占める [2]. 従って, 散在反復配列の検出はゲノムのアノテーションの最上流の解析と言える.

散在反復配列検出で現在ゴールデンスタンダードとされる RepeatMasker[3] は, 十分な散在反復配列データベースを持たない種には適用ができない. そこで, 多くのデータベースを利用しない散在反復配列検出手法が考案されてきた [4]. 特に, BLAST 様の seed-and-extend アルゴリズム [5] によって散在反復配列を検出する RepeatScout[6] は, 高感度なソフトウェアの一つとして報告されている.

多くの散在反復配列は転移因子由来とされるが, 転移因子由来領域は, ゲノム中で変異を受けることが知られている [7]. これらの検出には変異に頑健な散在反復配列検出ソフトウェアが必要であるが, これまでのデータベースなし散在反復配列検出ソフトウェアは対応していなかった [8].

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan
[2] 東京大学
University of Tokyo
[3] 産総研・早稲田大学 CBBD-OIL
AIST・Waseda CBBD-OIL
[a)] atsushi17@fuji.waseda.jp
[b)] mahamada@waseda.jp

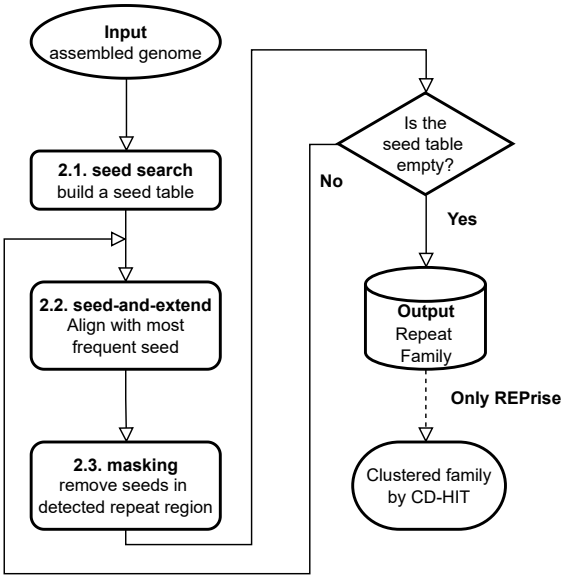

図 1　REPrise 及び RepeatScout のアルゴリズム概要

私たちは, RepeatScout をベースに, 変異に頑健でより高感度な散在反復配列検出を可能にするソフトウェア, REPrise を開発した.

2. 手法

図 1 に REPrise 及び RepeatScout のフローチャートを示す. アセンブリされたゲノム配列を入力とし, 散在反復配列のライブラリを出力するソフトウェアである. 実際に, 散在反復配列をマスキングするためには, 得られたライブラリとゲノム配列を入力として, RepeatMasker の利用が

必要となる.

2.1 Inexact seed

inexact seed S が配列 T と match する地点 i の集合 U_d を

$$U_d = \{i | h(S, T[i, j]) \leq d\}$$

と定義する. ここで, $h(X,Y)$ は文字列 X と Y のハミング距離, d は許容する置換数であり, $d = 0$ であれば exact match seed と同値である.

2.2 Seed-and-extend

散在反復配列のコンセンサス配列 Q を求める問題を考える. ここで, コンセンサス配列を extension alignment により一塩基ずつ伸長していくこととすれば, t 回目の伸長で得られた Q_t が存在した時, Q_{t+1} は,

$$Q_{t+1} = Q_t \cdot \underset{N \in \{A,G,C,T\}}{\operatorname{argmax}} A(Q_t \cdot N; S_1, S_2, \ldots, S_n)$$

と定式化される. ここで, $A(Q_t \cdot N; S_1, S_2, \ldots, S_n)$ は配列 $Q_t \cdot N$ と配列 $S_1, S_2, \ldots, S_n$ のペアアワイズアラインメントスコア $a(Q_t \cdot N, S_l)$ の総和であり,

$$A(Q_t \cdot N; S_1, S_2, \ldots, S_n) = \sum_{l}^{n} \max\left(a(Q_t \cdot N, S_l), 0\right)$$

と表せる. コンセンサス配列伸長の起点 Q_0 は inexact seed S に対応しており, さらにアラインメントされる領域 $S_1, S_2, \ldots, S_n$ は集合 U_d に含まれる地点と対応している.

REPrise は, この extension alignment に affine gap を導入した. つまり, RepeatScout では連続するギャップ数 l に対して, 線形なギャップコストであった $\gamma(l) = l \cdot g$ に対して, REPrise では

$$\gamma(l) = g + l \cdot e \qquad \text{但し } e < g$$

を採用する.

2.3 Masking

Masking は, 散在反復配列の検出に利用した seed を除去することで, 出力ファミリの冗長性を排除や, 計算リソースの効率な利用を目的としている. RepeatScout は, seed-and-extend で検出したコンセンサス配列に含まれた seed を起点に, コンセンサス配列をゲノムに再マッピングして, 重なった領域にそれ以降の seed match が起きないようにする. しかしながら, この方法では別のファミリの散在反復配列が同時に持つ seed を見逃すことになる.

REPrise は, masking のルールを緩くすることによって, 複数の散在反復配列が seed を共有していた場合の, 散在反復配列検出感度の低下を防ぐ. seed-and-extend により得られたコンセンサス配列を, 集合 U_d の領域のみに再度マッピングする.

表 1 REPrise と RepeatScout の性能評価

	Sensitivity	Specifiity	F1-Score
d0	0.9253	0.9430	0.9309
d1	0.9264	0.9410	0.9311
d2	0.9321	0.9336	0.9296
RepeatScout	0.8969	0.9561	0.9224

3. 結果

Ou らが散在反復配列検出ソフトウェアのベンチマーク用にアノテーションしたイネゲノム [9] を用いて, REPrise($d = 0, d = 1, d = 2$) と RepeatScout の性能比較を行なった (**表 1**). なお, 緩い masking によりファミリの冗長性が上がる問題に対応するために, REPrise は適用後に配列クラスタリングソフトウェア CD-HIT[10] を利用している.

REPrise$d = 0$ が RepeatScout よりも高感度であることから, affine gap 及び masking の改良の効果を確認することができた. また, inexact seed の許容する置換数 d を増加させることにより, sensitivity が向上することを示した.

参考文献

[1] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V. et al.: The complete sequence of a human genome, *Science*, Vol. 376, No. 6588, pp. 44–53 (2022).
[2] Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G., Gershman, A. et al.: From telomere to telomere: The transcriptional and epigenetic state of human repeat elements, *Science*, Vol. 376, No. 6588, p. eabk3112 (2022).
[3] Smit, A.F.A.and Green, P.: Repeatmasker.
[4] Zeng, C., Takeda, A., Sekine, K., Osato, N., Fukunaga, T. and Hamada, M.: Bioinformatics Approaches for Determining the Functional Impact of Repetitive Elements on Non-coding RNAs, *piRNA*, Springer, pp. 315–340 (2022).
[5] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.: Basic local alignment search tool, *Journal of Molecular Biology*, Vol. 215, No. 3, pp. 403–410 (1990).
[6] Price, A. L., Jones, N. C. and Pevzner, P. A.: De novo identification of repeat families in large genomes, *Bioinformatics*, Vol. 21, No. suppl_1, pp. i351–i358 (2005).
[7] Wells, J. N. and Feschotte, C.: A field guide to eukaryotic transposable elements, *Annual review of genetics*, Vol. 54, p. 539 (2020).
[8] Rodriguez, M. and Makałowski, W.: Software evaluation for de novo detection of transposons, *Mobile DNA*, Vol. 13, No. 1, pp. 1–14 (2022).
[9] Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A. et al.: Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline, *Genome Biology*, Vol. 20, No. 1, p. 275 (2019).
[10] Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W.: CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, Vol. 28, No. 23, pp. 3150–3152 (2012).