複数サンプルハイスループットシークエンスデータからの未観測遺伝子系図を用いたマイクロサテライトリピート数推定

東北大学 東北メディカル・メガバンク機構　小島 要
東北大学 東北メディカル・メガバンク機構　河合 洋介
カリフォルニア大学サンディエゴ校 ゲノム医学研究所　成相 直樹
東北大学 東北メディカル・メガバンク機構　三森 隆広
東北大学 東北メディカル・メガバンク機構　長谷川 嵩矩
東北大学 東北メディカル・メガバンク機構　長崎 正朗

DNA ハイスループットシークエンシング技術の発展により、現実的な時間とコストにより各個人の全ゲノムシークエンシングが可能となっている。こうしたハイスループットシークエンス (HTS) データから、全ゲノム規模において 1000 人以上の集団の単塩基変異の高精度な検出が可能となった。しかしながら、ゲノムの挿入、欠失による変異、マイクロサテライトにおける変異、コピー数変異などの構造変異の検出に関しては特に低深度のシークエンスデータにおいてその精度は十分でない。構造変異の中でマイクロサテライトは数塩基を単位とする配列のリピートからなる領域であり、個人間でリピート数の異なるリピート数変異があることが知られている。リピート数変異は、Huntingtin 遺伝子における CAG リピート数がハンチントン病と関連など、多くの疾患との関連が報告されており [7]、HTS データからのリピート数変異の解析が疾患関連変異の同定の観点からも重要である。

HTS データからのリピート数解析には、リードデータを各生物種の代表ゲノム配列であるリファレンスゲノムにアラインメントされたデータを元に推定が行われるが、大きく分けて二種類の解析方法が考えられている。一つは、対象となるマイクロサテライトにアラインメントされたシークエンスリードから直接リピート数を数え上げる方法であり、lobSTR [2] や RepeatSeq [3] といった手法が提案されている。もう一つは、対象となるマイクロサテライト周辺にアラインメントされたペアドエンドリードから推定されたインサートサイズの変化を元にした方法である。元のゲノム上のマイクロサテライト領域の長さがリファレンスゲノム内のマイクロサテライト領域の長さと異なる場合、その分、実際のインサートサイズは推定されたインサートサイズと異なる。インサートサイズの分布は、通常のゲノム領域において推定されたインサートサイズの分布から取得可能であり、これを用いて統計的にマイクロサテライト領域長の違い、すなわちリピート数の変化量を推定する。ペアドエンドリードを利用したものとしては STRViper [1] といった手法が提案されている。前者の手法は直接リピート数を数え上げるため推定結果の精度が非常に高い利点があるが、リピート数が推定可能なマイクロサテライトの長さがシークエンスリード長以下に限られる問題がある。一方、後者の方法では、インサートサイズの長さ前後までのマイクロサテライトについて推定が可能であるが、マイクロサテライト周辺にアラインメントされたリードに限定して、インサートサイズの分布が推定されるため、各個人単位で推定を行う場合、シークエンスリードの量に限りがあり、推定精度が低い問題があった。

そこで、複数人の遺伝子系図を周辺のフェージングされた遺伝子型をもとに Coalescent 理論 [4] により表現し、各個人の情報をマイクロサイトリピート数変異モデルを用いて自然に統合したモデルを提案している [5]。提案モデルでは、遺伝子系図が複数人のゲノムについて時間を遡って共通祖先へ結合していく過程を記述した coalescent 木の分布を用いて表現される。Coalescent 木によりマイクロサテライトの遺伝子系図上におけるリピート数の変化を自然に扱うことができることから、より精度の高いリピート数推定結果が期待される。Coalescent 木の分布は解析対象のマイクロサテライトの周辺のフェーズされた遺伝子型からのマルコフ連鎖モンテカルロ法によるサンプル分布を用いる。提案モデルからのリピート数の推定は周辺 MAP を拡張した問題に帰着できるが、周辺 MAP 問題自身が NP 困難であることが知られている。そこで、本発表では周辺 MAP 問題の近似解法である Mixed-product Belief Propagation [6] を拡張した手法を提案し、問題解決を行う。提案手法の有効性の検証として、合成的に得られた複数サンプルに対する HTS データによるシミュレーションと 1000 ゲノムプロジェクトからの 33 人の日本人サンプルの実 HTS データによる評価により、提案手法と既存手法における推定精度を比較する。

References

[1] Cao, M.D. *et al.*: Inferring short tandem repeat variation from paired-end short reads. Nucleic Acids Research 42(3) (2014)

[2] Gymrek, M. *et al.*: lobSTR: A short tandem repeat profiler for personal genomes. Genome Research 6, 1154–1162 (2012)

[3] Highnam, G. *et al.*: Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Research 4(1) (2013)

[4] Kingman, J.F.C.: On the genealogy of large populations. Journal of Applied Probability 19(A), 27–43 (1982)

[5] Kojima, K. *et al.*: Short tandem repeat number estimation from paired-end sequence reads by considering unobserved genealogy of multiple individuals , Proceedings of the 11th International Symposium on Bioinformatics Research and Applications, 422–423, (2015)

[6] Liu, Q., Ihler, A.: Variational algorithms for marginal MAP. Journal of Machine Learning Research 14, 3165–3200 (2013)

[7] Walker, F.O.: Huntington's disease. Lancet 369(9557), 2185–2228 (2007)