平成28年度　　京都大学化学研究所　スーパーコンピュータシステム　利用報告書

三角不等式を制約として考慮した整数計画法による中央文字列と中心文字列の厳密解法の高速化

On triangle inequality constraints to integer linear programming for finding median and center strings for a probability distribution on a set of strings

京都大学化学研究所バイオインフォマティクスセンター数理生物情報　林田守広

研究成果概要

本研究ではこれまで開発してきた，文字列の集合上の確率分布に対する中央文字列および中心文字列の整数計画法による厳密解法について，高速化を目的とした改良を行った．

有限個の文字からなるアルファベットを A とし，文字列の集合 A*上の確率分布を p(s)とする．二つの文字列 s, t の間の距離を d(s, t) とするとき，中央文字列 m はすべての文字列 s に対して p(s)d(m, s)の和を最小とする A*の元である．また中心文字列 c はすべての文字列 s における p(s)d(c, s)の最大値を最小化する文字列であると定義される．距離としてレーベンシュタイン距離を選択するとき，中央文字列および中心文字列を求める問題はともに NP 困難となるので，整数計画法による厳密解法 ILPMed, ILPCen をそれぞれ提案した．

本研究では，レーベンシュタイン距離が三角不等式を満たすことを利用する．$p(s)>0$ となる文字列 s_1, s_2 と中央文字列あるいは中心文字列 t に三角不等式を適用することで，$d(s_1, s_2)+d(s_2, t) \geq d(s_1, t)$ あるいは，$d(s_1, t)+d(s_2, t) \geq d(s_1, s_2)$ が得られる．ここで $d(s_1, s_2)$ は定数であり，$d(s_1, t)$ は整数計画問題における変数の線形式として表現される．そこでこれらの線形不等式を制約条件として ILPMed，ILPCen に付加した整数計画法による解法をそれぞれ ILPMedTri, ILPCenTri として提案する．

DNA あるいは RNA の塩基配列は 4 種類の文字からなっていることより$|A|=4$としていくつかの A*上の確率分布を用いて計算時間を複数回計測しその平均と分散を ILPMed，ILPMedTri，ILPCen，ILPCenTri それぞれについて計算した．その結果，$p(s)>0$ となる文字列 s の数 N が 10 本程度で，それぞれの文字列の大きさ $|s|$ が 10 程度のとき，三角不等式を制約条件として加えた ILPMedTri, ILPCenTri の方が，元の ILPMed, ILPCen よりもそれぞれ約 1/10 程度の実行時間となった．三角不等式は常に満たされるものであり，厳密解であることを保ちつつ実行の高速化を実現した．

発表論文（謝辞あり）

発表論文（謝辞なし）

Hayashida, M, and Koyano, H, Finding median and center strings for a probability distribution on a set of strings under Levenshtein distance based on integer linear programming, *submitted*.