数学と情報科学で解き明かす多様な対象の数理構造と活用
2019 年度採択研究者

2021 年度
年次報告書

稲永　俊介

九州大学 大学院システム情報科学研究院
准教授

文字列学的手法によるシーケンシャルデータ解析

§1. 研究成果の概要

本研究では，文字列，生物学的配列，時系列などの多様な系列データに加え，ラベル付き木・グラフをも包含する広義文字列を対象とするアルゴリズム開発を行った．n を入力サイズ，σ をアルファベットサイズ，d をスライド窓サイズとする．

(a) 動的文字列データ処理：DNA・RNA 配列などにおいて重要な役割を果たす回文構造，および，PCR プライマー設計などに動機を持つ極小ユニーク部分文字列を，それぞれ O(n log σ) 時間・O(d)領域で計算するオンラインアルゴリズムを示した [1]．いずれにおいても，接尾辞データ構造の組合せ的性質を巧みに利用することが処理効率化の鍵であった．

(b) 広義文字列の数理とアルゴリズム：文字列集合のコンパクト表現であるラベル付き木に含まれる反復構造を O(n loglog n)時間・O(n)領域で求めるアルゴリズムを示した [2]．さらに，時系列データ処理のためのデカルト木照合をラベル付き木上で実行可能な世界初の索引データ構造を開発した．加えて，文字列構造を反映したパラメタ化照合に対する現在最高速の索引データ構造を開発した [3]．デカルト木照合とパラメタ化照合は，ともに SCER と呼ばれるパターン照合の枠組みに属する近似照合の一種であり，共通するアルゴリズム的テクニックを駆使して両成果を導いた．

(c) 文字列反復性指標と圧縮：圧縮感度という新指標を導入した．これは，(1)圧縮率，(2)圧縮・展開速度，(3)検索可能性という従来の圧縮アルゴリズムの評価指標に続く，第4の評価指標である．文法圧縮，LZ 系圧縮，双方向スキーム，文字列アトラクタ，部分文字列複雑性などに対して，感度の非自明な上界と下界を与えた [4]．その多くは，上界と下界が合致するタイトな結果となっている．加えて，定数感度の文法圧縮 GCIS に着目し，圧縮パターン照合のための GCIS-index 圧縮索引構造を開発した [5]．

【代表的な原著論文情報】

1) Computing Minimal Unique Substrings for a Sliding Window, Takuya Mieno, Yuta Fujishige, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Algorithmica, August 2021.

2) Efficiently computing runs on a trie, Ryo Sugahara, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Theoretical Computer Science, 887:143-151, October 2021.

3) The Parameterized Suffix Tray, Noriki Fujisato, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda, Proc. 12th International Conference on Algorithms and Complexity (CIAC 2021), LNCS 12701, pp. 258-270, May 2021.

4) Sensitivity of string compressors and repetitiveness measures, Tooru Akagi, Mitsuru Funakoshi, Shunsuke Inenaga, arXiv, abs/2107.08615, 2021.

5) Grammar Index By Induced Suffix Sorting, Tooru Akagi, Dominik Köppl, Yuto Nakashima, Shunsuke Inenaga, Hideo Bannai, Masayuki Takeda, Proc. 28th International Symposium on String Processing and Information Retrieval (SPIRE 2021), October 2021.