公開

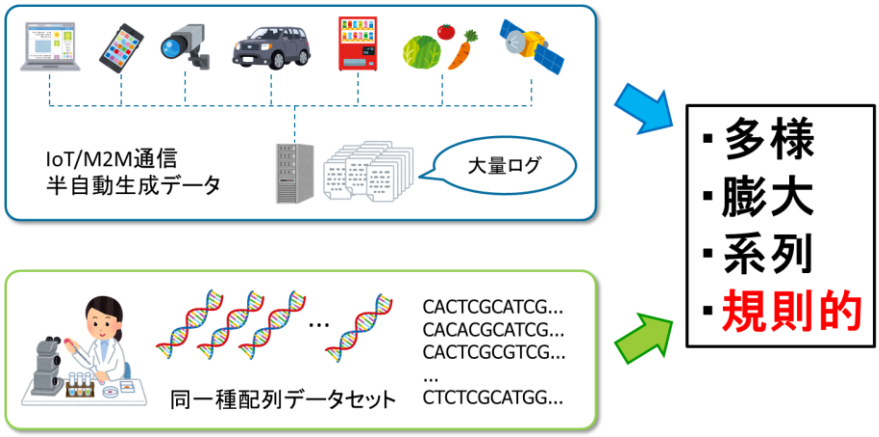
研 究 終 了 報 告 書

「文字列学的手法によるシーケンシャルデータ解析」

研究期間： 2019 年 10 月～2023 年 3 月
研 究 者： 稲永 俊介

1. 研究のねらい

本研究では，IoT や M2M 通信による半自動生成データ，あるいは，各種生物学データベースに蓄積される配列データセットなどの，多様かつ大規模な系列データ（いわゆるビッグデータ）に対する高速処理アルゴリズムと，それに基づくデータ解析基盤の確立を目標としている．
これらのデータは，一見，それぞれがまったく異なる特徴を有しているように思われがちであるが，一方で，いずれも「反復性」あるいは「規則性」を多く含むという，組合せ的な性質を有している（下図参照）．そこで，本研究では，文字列組合せ論とアルゴリズム技術を融合させる「文字列学的手法」によって，系列ビッグデータの高速処理を実現することを目標とする．



図：研究背景

また，一般に，文字列とは記号の一本鎖のことを指すが，本研究では，枝分かれ・合流を有するラベル付き木・グラフや，2次元文字列などにもその対象を拡大し，これらの反復性・規則性の解明，および高速処理アルゴリズムの開発も行った．加えて，時系列などの離散数列データに対するアルゴリズムを提案した．

2. 研究成果

(1) 概要

本研究課題では，以下の3つテーマを柱とする研究開発を行った．

- 研究テーマ A「動的文字列データ処理」
- 研究テーマ B「広義文字列の数理とアルゴリズム」
- 研究テーマ C「文字列反復性指標と圧縮」

研究テーマ A では，逐次入力あるいは動的編集される動的系列データを実時間高速処理するアルゴリズムの開発を行った．特に，実応用データ処理の場面で頻出するスライド窓処理アルゴリズムの開発に注力した．ここで，スライド窓処理は，先頭文字削除＋末尾文字追加という 2 つの編集操作の組として定式化される．

研究テーマ B においては，数列，時系列，木，グラフ，2 次元列（画像）などを広義文字列と総称し，文字列学のこれまでの研究蓄積を拡張することによって，広義文字列の数理構造解明と高速データ処理を実現した．主な成果として，木型文字列（辺ラベル付き木）上での情報検索クエリに高速応答可能な省領域データ構造，木型文字列中の反復構造・回文構造の高速発見アルゴリズム，時系列データの高速比較アルゴリズムなどを開発した．

研究テーマ C では，数多ある既存の圧縮アルゴリズム・反復性指標の理論的性能評価を徹底的に行い，反復性指標の性能の「見取り図」を作成した．加えて，データの編集に対する圧縮アルゴリズムの頑強さを定量化した新指標「圧縮感度」を提案し，主要な圧縮手法を感度の観点から再評価した．さらには，圧縮符号化したままパターンの高速検索が可能なデータ構造の開発にも成功した．

A，B，C はそれぞれ独立したテーマではなく，相互に有機的に絡み合いながら研究開発を行った．例えば，A の動的データ高速処理を実現するためには，C のデータ圧縮で用いられる文字列の周期性を活用することが必須であるし，C の圧縮感度は A と深く関連する．さらに，B の木型文字列中の反復構造・回文構造計算には，C の周期性が深く関わっている．

（2）詳細

研究テーマ A「動的文字列データ処理」

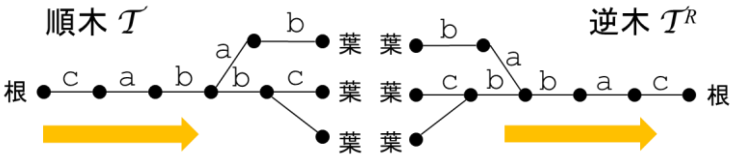
本テーマでは，文字の挿入・削除・置換といった編集操作に逐次対応可能な文字列データ構造の開発を行った．特に，実応用データ処理の場面で頻出するスライド窓処理アルゴリズムの開発に注力した．ここで，スライド窓処理は，先頭文字削除＋末尾文字追加という 2 つの編集操作の組として定式化される．特に，データが逐次的に送られてくるストリーミングデータは，一般にその長さ n が無制限に巨大であるため，データをすべて記憶領域に保存する方法は現実的ではないため，スライド窓による処理が効果的かつ実用的である．本テーマにおける主要な結果を以下に示す．

- **スライド窓に対する極小ユニーク部分文字列の高速計算：**文字列 w が文字列 T の極小ユニーク部分文字列 (Minimal Unique Substring, MUS) であるとは，w が T 中にちょうど 1 回出現し，かつ w の部分文字列は T 中に2回以上出現することをいう．MUS は，生物学的配列に対する PCR プライマー設計などに動機を有する，文字列中の数理構造である．本テーマに関し，さきがけ研究初年度では，まず固定長窓幅 d のスライド窓中の MUS を O(nlogσ)時間・O(d)領域で計算する手法を与えていた．ここで，σ はアルファベットサイズである．当該成果をまず，2019 年度に査読付き国際会議で発表した．2021 年度は，さらに，このアルゴリズムに可変長窓幅に対応可能な機能を追加した．このアップデータした成果をまとめてフルペーパー化し，Algorithmica 誌にて発表した．

研究テーマ B「広義文字列の数理とアルゴリズム」

本テーマでは，グラフ／木／時系列／画像などを含む広義の意味での文字列を対象とし，これらに内在する数理的性質を解明することによって，広義文字列の効率的処理技法を確立することを目的としている．本テーマの主要結果を以下に概説する．

- **木に対する索引構造**：本研究では，木型文字列に対する，接尾辞木，接尾辞配列，DAWG，CDAWG (Compact DAWG) といった代表的な索引構造のサイズに関するタイトな上界・下界を与えた（下記表参照）



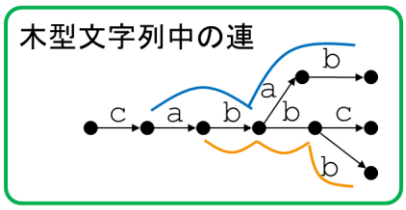
図：左）根から葉へ読むトライ，右）葉から根へ読むトライ

中でも特筆すべき成果として，DAWG はトライを根から葉，または葉から根へとどちらから読んだとしても $O(n^2)$個の辺数を要してしまうにも拘わらず，O(n)領域でコンパクトに表現可能であることを証明した．ここで n は入力トライ中の総文字数である．本成果は，まず国際会議 LATIN2021 にダイジェスト版が採択され [主要な学会発表 2]，続けてフルバージョンが情報処理学会創立 60 周年記念論文を受賞した [受賞 1，代表的な論文 1].

	順木$\mathcal{T}$		逆木$\mathcal{T}^R$	
索引	頂点数	辺数	頂点数	辺数
DAWG	$2n-3$	$O(n^2)$	$O(n^2)$	$O(n^2)$
CDAWG	$2n-3$	$O(n^2)$	$2n-3$	$2n-4$
Suffix Tree	$O(n^2)$	$O(n^2)$	$2n-3$	$2n-4$
Suffix Array	$O(n^2)$		$n+1$	

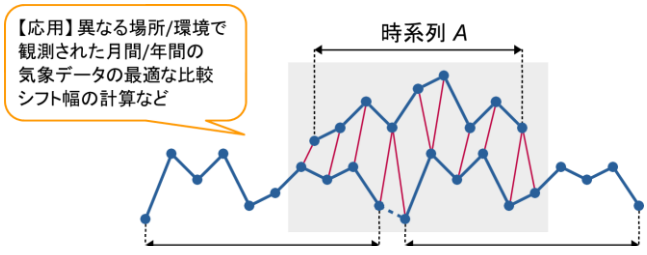
表：トライに対する索引構造のサイズ

- **木型文字列中の反復構造・回文構造計算**：abab, babab のような反復構造，あるいは abba，aacaa のような回文構造は，生物学的配列の特徴的な部分区間に出現することが知られている．本研究では，辺数 n の木型文字列中に出現するすべての反復構造を O(n loglog n)時間・O(n)領域で計算するアルゴリズムを与えた．木型文字列を文字列集合に分解してから個別に反復を計算する既存手法は，$O(n^2)$時間を要する．すなわち，本提案アルゴリズムによって O(n/loglog n)倍の大幅な高速化を実現した．



さらに，本研究では，辺数 n の木型文字列中に出現するすべての回文構造を O(n)最適時間・領域で計算するアルゴリズムを与えた [主要な学会発表 4]．本成果は，O(n log n)時間を要する既存手法とはまったく異なる，独創的なアプローチによって得られた．

- **時系列データ比較アルゴリズム**：Dynamic Time Warping (DTW) は，株価や音声波形データなどの時系列データの比較に幅広く用いられる手法である．本研究では，気象データなどへの応用を考慮した発展問題である循環 DTW，周期的 DTW，スクエア DTW をそれぞれ $O(n^2\,\mathrm{polylog}(n))$ 時間で計算するアルゴリズムを開発し，その成果は国際会議 ISAAC2020 に採択された [主要な学会発表]．提案手法は，これらの発展問題に $O(n^3)$ 時間を要する既存の動的計画法よりも優れている．続けて，この手法の計算量オーダー表



図：循環 DTW（A と B の最適比較シフトを求める）

記に隠れている定数項を大幅に削減した新手法を開発し，Algorithmica 誌で発表した [代表的な論文 3].

研究テーマ C「文字列反復性指標と圧縮」

本テーマでは，可逆圧縮とデータの反復性に関する研究を行った．旧来は，反復度≒圧縮率のように捉えられてきたが，近年の研究で，必ずしもそのような近似は成り立たないことが示唆され始めた．そこで，新テーマでは，この近似率を数学的に厳密に評価する研究を追加で実施した．また，実用的な圧縮アルゴリズムの開発，および簡潔データ構造や擬似周期性に関する研究成果も得ることができた．

- **文字列アトラクタと実用的辞書式圧縮のサイズ比**：直感的には，入力データ（文字列）中に反復構造が多いほど，より圧縮できる． 2018 年に Kempa と Prezza によって提案された最小文字列アトラクタのサイズ γ は，圧縮可能性を上手く捉えうる反復性指標の有力な候補として高く評価されてきた．一方，実用の場面においては，文字列の貪欲分割に基づく LZ77 圧縮が広く用いられている．LZ77 圧縮は，zip や png などの圧縮プログラムのコア・アルゴリズムである．本研究では，最小文字列アトラクタのサイズ γ と，LZ77 圧縮サイズ z の間に $\log n$ 倍の差存在する文字列が存在することを示した．ここで n は圧縮前の文字列長を表す．本結果はすなわち，文字列アトラクタは圧縮可能性を上手く捉えることができていない，ということを理論的に証明したものある [受賞 3].
- **圧縮感度**：さきがけ同期の吉田氏の研究からインスパイアを受け，圧縮アルゴリズムの感度に関する研究に着手した．すなわち，入力文字列を1文字編集する前と後とで，圧縮率にどれ程の差が生じうるかということを，圧縮データサイズの比と差分の観点から数学的に評価した．具体的には，zip や png などに幅広く用いられる LZ77 系圧縮法，およびその一般化である双方向スキーム，各種文法圧縮法，バイオ情報学の解析ツールとして注目を集める連長 BW 圧縮などに対して，感度の非自明な上下界を与えた [代表的な論文 2, 受賞 2]. 特筆すべき点として，多くの圧縮アルゴリズム・反復性指標に対して，それらの感度のタイトな上下界を導出することに成功した．

3. 今後の展開

本研究では，実社会の多様なビッグデータをシーケンシャルデータという枠組みのもとで統一的に取り扱った．本研究の成果は，リアルタイム処理でストレスフリーに動作し，圧縮処理によって安価なデバイスにも実装可能な，大規模データ利活用プラットフォーム構築のための礎となりうる．

本研究課題は文字列組合せ論とアルゴリズム技術に関する基礎研究であるため，得られた成果が将来的な社会実装に繋がるための展開やタイムスパンを述べることは難しい．現在は，提案アルゴリズムをベースとした時系列データ比較ソフトウェアの開発を行うなど，実用化に向けた活動を地道に進めているところである．一方，当該分野では特許申請は一般的ではなく，オープンソースとして公開するやり方が主流であるため，本研究の成果に基づく特許申請を行う予定はない．

今後は，テーマ A, B, C をより深く結びつけた研究を進めていく予定である．例えば，

- 時系列ストリーミングデータに対するパターン照合（A+B）

- 広義文字列に対する動的データ構造（A+B）
- 圧縮感度と圧縮率の両立（C+A）
- 広義文字列の反復性指標（B+C）

などが，今後の自身の研究テーマの候補として挙げられる．

4. 自己評価

研究目的の達成状況：本さきがけ研究の当初目標は，「文字列組み合わせの高度理論と最先端文字列処理アルゴリズム技術，および数論・代数学の知見を有機的に融合させることで，次世代の社会情報基盤たる多様シーケンシャルデータ解析プラットフォームを確立する」ことであった．ここでいうプラットフォームとは，言葉の通り土台のことである．すなわち，多様なシーケンシャルデータの解析を文字列処理という統一的な枠組みの中で取り扱うためのアルゴリズム研究開発が，本さきがけ研究課題の中核であった．このことを踏まえれば，当初の目的を十分に達成できたと自己評価している．

研究の進め方：研究期間の大半がコロナ禍の中であったが，このことを逆に利用して，自らの研究時間をより多く確保し，単著論文を複数執筆できたことはプラスに評価できる．また，コロナ禍の後半には，オンラインツールを利用して，欧州の研究者との共同研究打合せを頻繁に行い，極東という地理的不利を打開できたことも新たな経験だった．学内においては，テクニカルスタッフと RA を雇用し，研究補助を担ってもらうことで，自らの研究時間の確保に活用した．また，さきがけ研究費によって，PI 人件費やバイアウトなどの新制度を積極的に活用することができた．コロナ禍が落ち着いた最終年度には，国内・海外出張を再開し，自らのさきがけ成果に関する広報活動と，今後の研究活動に繋がる議論を活発に行った．

研究成果の科学技術及び社会・経済への波及効果：3つの研究成果が受賞したことは，科学技術分野への波及効果が少なからずあったと自己評価する．また，圧縮感度などの新指標や，圧縮アルゴリズムの見取り図の作成など，当該学術分野への貢献は高いと評価できる．研究成果の社会・経済への波及効果については，「3. 今後の展開」で述べた通りである．

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：41件

1. Shunsuke Inenaga: Towards a complete perspective on labeled tree indexing: new size bounds, efficient constructions, and beyond, Journal of Information Processing, 29:1-13, January 2021.
XML や BLAST などのデータは，各辺が文字でラベル付けされた木型文字列（トライ）としてモデル化できる．また，トライは文字列集合のコンパクト表現としての応用がある．本論文では，サイズ n の入力木型文字列に対して，双方向パターン照合を線形 O(n)領域で実行可能な初の索引構造を提案した．既存手法では，$O(n^2)$領域が必要であった．双方向パターン照合は，RNA2次構造照合や，自然言語処理に応用がある．
2. Tooru Akagi, Mitsuru Funakoshi, and Shunsuke Inenaga: Sensitivity of string compressors

and repetitiveness measures, Information and Computation, Volume 291, March 2023, 104999 (available online).
本論文では, 圧縮アルゴリズムの新評価指標「圧縮感度」を提案した. ここで, 圧縮感度とは, 入力文字列に1文字編集操作を行った際の, 圧縮サイズの増分のことである. この新しい指標によって, 圧縮法のエラーやデータ編集に頑強性の定量化が可能となった. 本研究では, LZ 圧縮や文法圧縮といった, 実社会で実装・活用されている主要な圧縮アルゴリズムの感度のタイトな上界・下界を与え, データ圧縮分野の未踏領域を開拓した.
3. Yoshifumi Sakai and Shunsuke Inenaga: A faster reduction of the dynamic time warping distance to the longest increasing subsequence length, Algorithmica, 84:2581-2596, May 2022.
DTW（時間伸縮法）は, 時系列データをはじめとする離散数列の比較指標として, 幅広く利用されている. 本論文では, 既存研究では十分に議論されてこなかった DTW の理論研究に切り込んだ. 具体的には, 気象データの比較などに動機を有する, スライド DTW 問題や循環 DTW 問題を, $O(n^2 \mathrm{polylog}(n))$時間で解く高速アルゴリズムを提案した. 既存手法では, これらの問題を解くのに $O(n^3)$時間を要する.

(2)特許出願

なし

(3)その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

主要な学会発表

1. Shunuke Inenaga: Combinatorial algorithms for grammar-based text compression, Tutorial on Combinatorial Methods for String and Graph, March 2020. (招待講演)
2. Shunsuke Inenaga: Suffix Trees, DAWGs, and CDAWGs for Forward and Backward Tries, 14th Latin American Theoretical Informatics Symposium (LATIN 2020), December 2020.
3. Shunsuke Inenaga: Pointer-Machine Algorithms for Fully-Online Construction of Suffix Trees and DAWGs on Multiple Strings, Prague Stringology Conference 2020 (PSC 2020), August-September 2020.
4. Takuya Mieno, Mitsuru Funakoshi, and Shunsuke Inenaga: Computing palindromes on a trie in linear time, 33rd International Symposium on Algorithms and Computation (ISAAC 2022), December 2022.
5. Yoshifumi Sakai and Shunsuke Inenaga: A reduction of the dynamic time warping distance to the longest increasing subsequence length, 31st International Symposium on Algorithms and Computation (ISAAC 2020), December 2020.

受賞

1. 情報処理学会 60 周年記念論文
2. 国際会議 SOFSEM 2021 Best Paper Award
3. 国際会議 SPIRE 2020 Best Paper Award

著作物（編著）

1. “Combinatorial Methods for String Processing”, Special Issue of Algorithms, 2021.