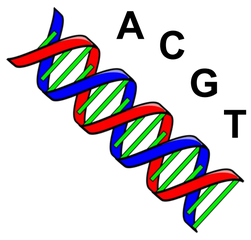
文字列のアルゴリズムと数理に関する研究

情報学部門 数理情報講座 文字列学研究室
教授 稲永俊介　助教 中島祐人

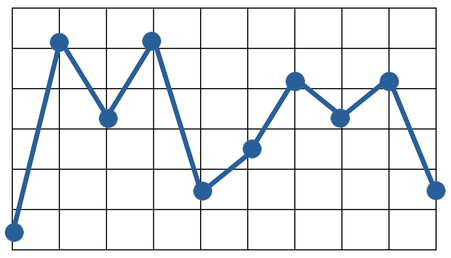
文字列学とは

■ コンピュータ可読な多くのデータは
文字列（記号の列）とみなすことができる.

テキストデータ



生物学的データ



時系列データ

➔ **文字列のアルゴリズムと数理に関する研究**

圧縮文字列処理アルゴリズム

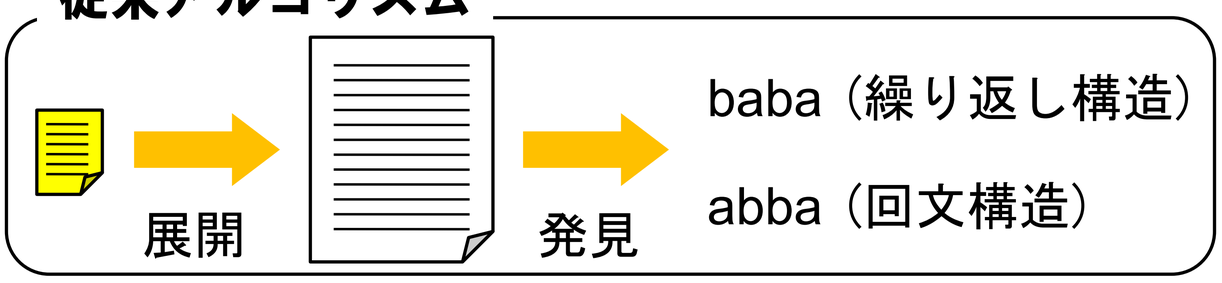
■ 大規模な文字列データを圧縮しておくことで,
記憶領域削減と通信高速化を実現できる.

■ 圧縮の例）abababababbbaaa = $(ab)^5\,b^1\,a^3$

■ しかし，データの中身にアクセスしたいときに
逐一データを展開していたのでは，せっかく
データを圧縮しておいた利点が失われてしまう.

■ 【本研究】圧縮データを陽に展開することなく,
圧縮したまま高速処理するアルゴリズムを開発.

従来アルゴリズム



提案アルゴリズム

従来手法に対して**劇的な高速化**を達成！
➔ 従来は圧縮サイズ n の指数時間かかって
いたものを，n の低次多項式時間に削減した.

ところで，なぜ圧縮したまま
処理すると高速化できるのだろう？

「ラーメン早食い競争」に例えると分かりやすい

展開してから法　展開しながら法　圧縮したまま法

文字列中の繰り返し構造

■ 文字列中に現れる繰り返し構造は,
文字列の**特徴抽出**，**頻出パターン発見**,
データ圧縮などと深い関係がある.

■ 文字列中の極大な繰り返し構造を連と呼ぶ.

この文字列は８つの連を含む

■ 長さ N の任意の文字列が含む連の最大個数を
$\rho(N)$ と表す.

■ 1999年，Kolpakov と Kucherov は $\rho(N) = O(N)$
を示し，さらに $\rho(N) < N$ と予想した（**連予想**）.

$\rho(N) < cN$	Kolpakov と Kucherov, 1999
$\rho(N) < 5N$	Rytter, 2006
$\rho(N) < 3.48N$	Puglisi ら, 2006
$\rho(N) < 1.6N$	Crochemore と Ilie, 2008
$\rho(N) < 1.029N$	Crochemore ら, 2011
$\rho(N) < N$	**我々の研究グループ，2017**

熾烈な国際競争を制し，$\rho(N) < N$ の証明に成功！
➔「**連定理**」の誕生

文字列学の研究に関連深い講義など

■ 形式言語とオートマトン
■ データ構造とアルゴリズムI・同演習
■ データ構造とアルゴリズムII
■ その他，**プロコン経験者**も大歓迎！

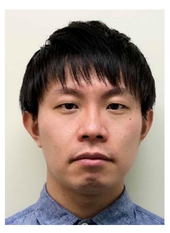
研究体制


稲永教授


中島助教


坂内教授
（東京医科歯科大）


Koeppl助教
（東京医科歯科大）


三重野助教
（電気通信大学）

■ 稲永教授・中島助教で文字列学研究室を共同運営
■ 坂内教授・Koeppl助教（東京医科歯科大）
三重野助教（電通大）らの学外研究協力者と
オンラインゼミなどを通じて緊密に連携
■ 学生数: 15名（2022年度）
内訳 B4: 5名, M1: 3名, M2: 5名, D:2名