グラフの中の多様な文字列と最長共通部分列の発見

Finding Diverse Strings and Longest Common Subsequences in a Graph

有村 博紀，北海道大学 大学院情報科学研究院
志田 祐仁*、小林靖明（北海道大学），*Giulia Punzi (NII, Uni Pisa)*，
宇野毅明（*NII*）との共同研究
*1) 2024.3MC修了

Yuto Shida, Giulia Punzi, Yasuaki Kobayashi, Hiroki Arimura, "Finding Diverse Strings and Longest Common Subsequences in a Graph," manuscript under submission for a conference:
https://www.dropbox.com/t/gMLYjniIcFdeUUbs

背景： 最適化における望ましい解とは

- 伝統的に，最適化では単一の最適解を追求
- 複数の多様な解の発見に興味がもたれている
  - 医療，配送，生産，運用の計画

- 理由：
  - 最適化問題の仕様が完璧でない（目的関数+制約条件）
  - 複数の最適解がある（アルゴリズム依存）
  - 人間が関与したい ＝ “Human-in-the-Loop”

複数の多様な解の発見

複数解を計算するために，過去にいろいろな手法が試されてきた

- 乱択 – 解をランダムに生成
- 列挙 – 解を網羅的に生成
- top-K – 解を目的関数の降順で生成

どの方法も，現在の目的には今ひとつ十分でない．．

背景： 最適化における多様な解

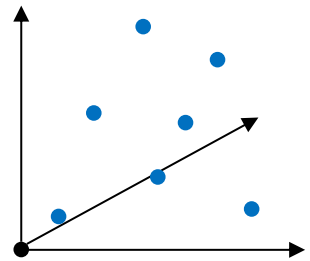
Diverse-X プログラム: 明示的に，多様な解を見つける問題を解こう！ (Baste, Fellows+, AIJ ’22)

Michael R. Fellows
U. Bergen, Norway

多様性最大化問題

X = Problem Name
Path, MST, Matching, LCS,
Decision Trees, … DNN?

- 最適化問題のN個の全ての解から，多様性尺度Dを最大化するK個の解の集合$X \subseteq Sol$を見つける
- 点集合版は，1970年代から研究されている：
  - 最大施設配置問題(facility location)
  - K分散問題（K-dispersion）



J.Baste, M.R.Fellows, L.Jaffke, T.Masařík, M.de Oliveira Oliveira, G.Philip, and F.A.Rosamond. Diversity of solutions: An exploration through the lens of fixed-parameter tractability theory. Artificial Intelligence, 303:103644, 2022.

多様性最大化問題の計算複雑さ

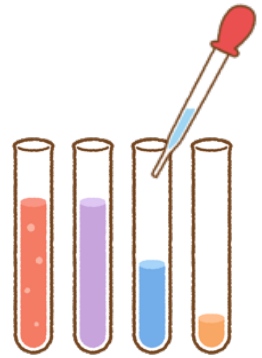
- **簡単な場合**：距離空間のN個の点集合
  - $O(N^K\ t_{\mathrm{dist}})$時間： 全てのK個の組合せを試す
- **難しい場合**：離散構造上の最適化問題の解空間
  - グラフに含まれるK個の全域木
  - 有向グラフに含まれるK個の最短路

…

  - 機械学習やデータマイニングに現れる最適化問題
    - K個の予測モデル, K本の最長共通部分列, . . .

本研究の目的

- 生命情報解析における重要な問題において，多様解問題の計算複雑さを明らかにする



- **最大共通部分列問題 (LCS)**
  - M本の入力文字列のすべてに共通して含まれる長さLの(不連続な)部分列の一つを求める問題
  - 情報科学における最も基礎的な問題の一つ
  - 50年以上にわたって理論と応用で研究される
  - 計算量：Mが定数なら，多項式時間で計算可能

Mが入力ならNP困難; MパラメタでW[t]困難; LパラメタでW[2]困難;

解集合＝最長共通部分列 (LCS)の全体

共通部分列
(common subsequences)

ε , A,B,C,D,E

AA, AB, AC, AD, AE, BA, ..., CD, CE, DD, EE,

ABA, ABB, ABC, ABD, ..., CEE,

ABAD, ABAE, ABBD, . . . , BCEE,

ABADD, ABAEE, ABBDD,

ABBEE, ABCDD, ABCEE

入力文字列
集合S

X = ABABCDDEE

Y = ABCBAEEDD

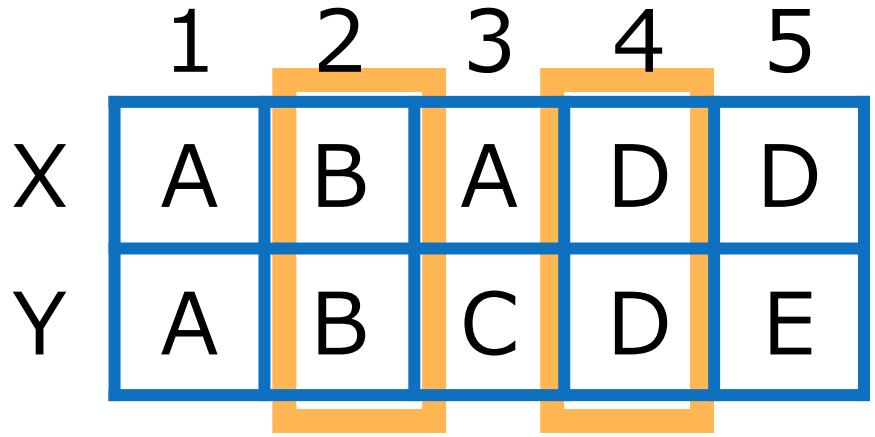
最長共通部分列
(longest common subsequences)

文字列どうしの距離

ハミング距離 (Hamming distance)

同じ長さ n をもつ文字列 X と Y の間のハミング距離は，文字どうしが異なる位置の総数.

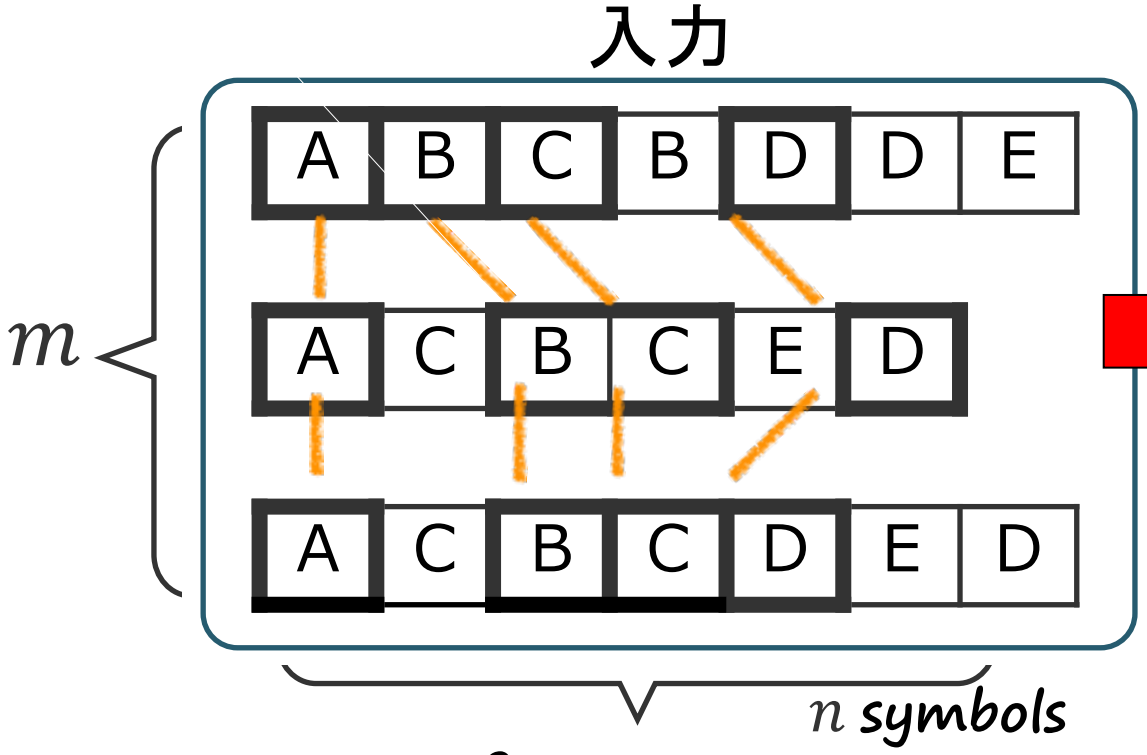
$$d_{HD}(X,Y) = \sum_{i=1}^{n} \mathbb{1}\{X[i] \neq Y[i]\}$$

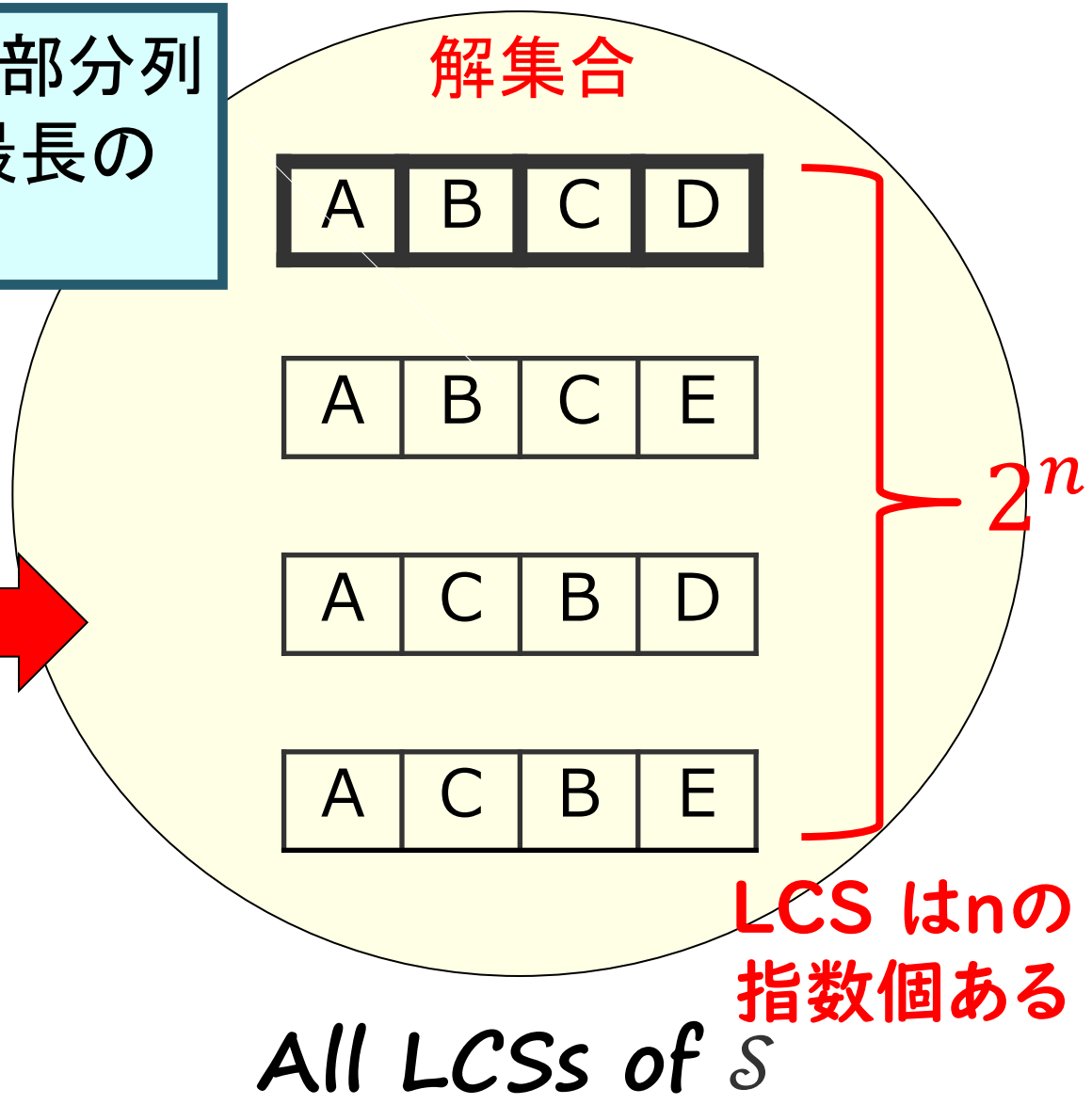


$$d_{HD}(X,Y) = 2$$

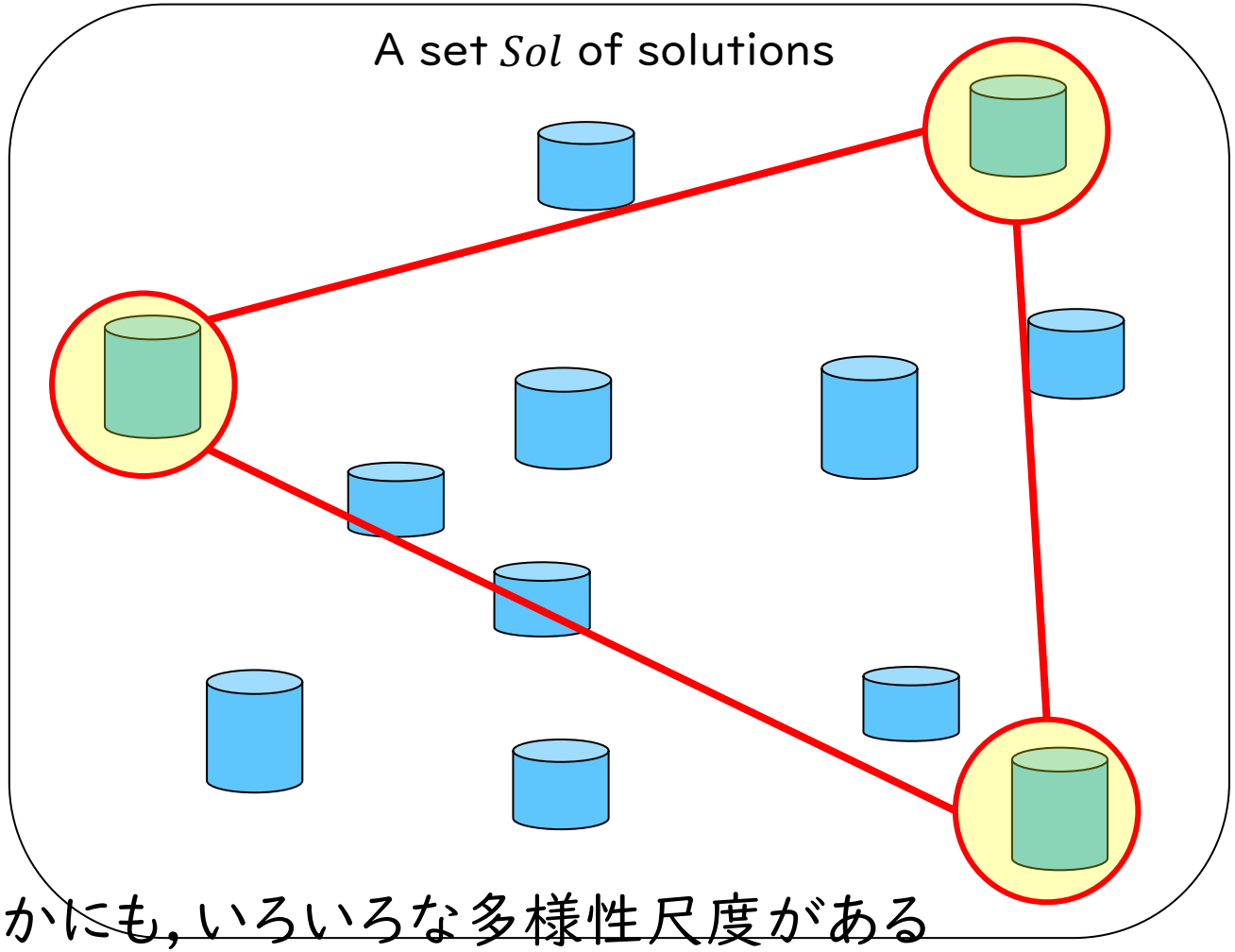
最大共通部分列問題 (LCS)

m 本の文字列の集合 Sから，最長共通部分列（共通して含まれる非連続な部分列で最長のもの）を求めよ.





問題：Max-Sum多様性最大化



Given: A set of solutions Sol, distance function $d: Sol^2 \to \mathbb{R}_+$ integers $K \geq 1, \Delta \geq 0$

Task: Find a subset $X = \{x_1, \ldots, x_K\} \subseteq Sol$ such that

1. $|X| \leq K$
2. $D(X) \geq \Delta$

Max-Sum多様性

$$D(X) := \sum_{i<j} d(x_i, x_j)$$

ほかにも，いろいろな多様性尺度がある
(Max-Min, Tree, …)

結果のまとめ: Max-Sum多様なLCS問題の計算量

解数Kが定数のとき

多項式時間計算可能

Proof: 動的計画法

- 本頁の全ての結果は，本研究で示した
- PTAS以外は， Max-Min版についても同じ結果が成立する

解数Kが入力のとき

NP困難

PTAS

任意誤差で多項式時間近似可能

Proof: ローカル探索法 (Cevallos+ 2019)

解数Kと入力長さrがパラメタのとき

FPT

固定パラメタ計算容易

Proof: 色符号化技法

解数Kがパラメタのとき

W[1]困難

固定パラメタ計算困難

Proof: p-クリークからのFPT帰着

結果1：解数Kが限定されたとき

- **定理**：Kが定数のとき，Max-Sum最大多様LCSs問題は，多項式時間計算可能である (入力文字列数m ≥ 2は定数)．
  Max-Minについても同じ結果

解法：DAG上の動的計画法を用いて，より一般的な次の結果を示すことで，定理を示している

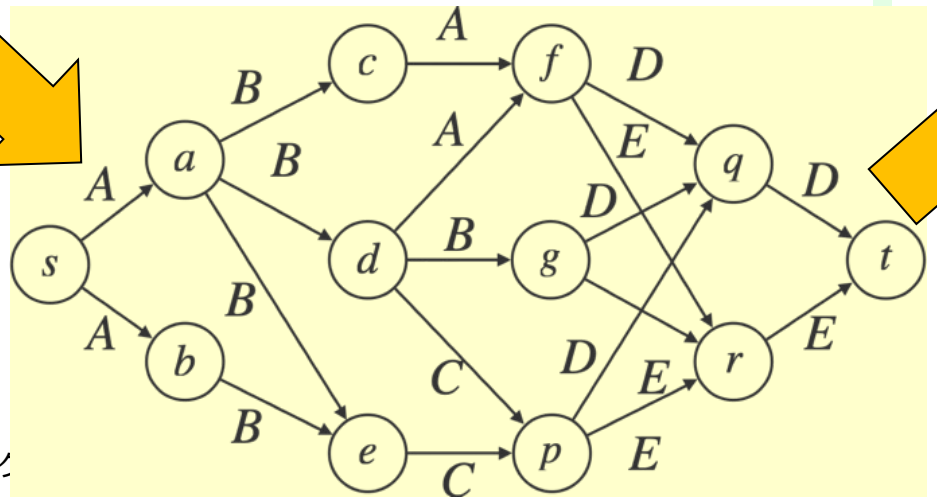
- **補題**：Kが定数のとき，任意の等長文字列の集合Lについて，それを表す非巡回有向グラフ(Σ-DAG)が与えられたとき，Max-SumとMax-Min版の最大多様LCSs問題は，多項式時間計算可能である

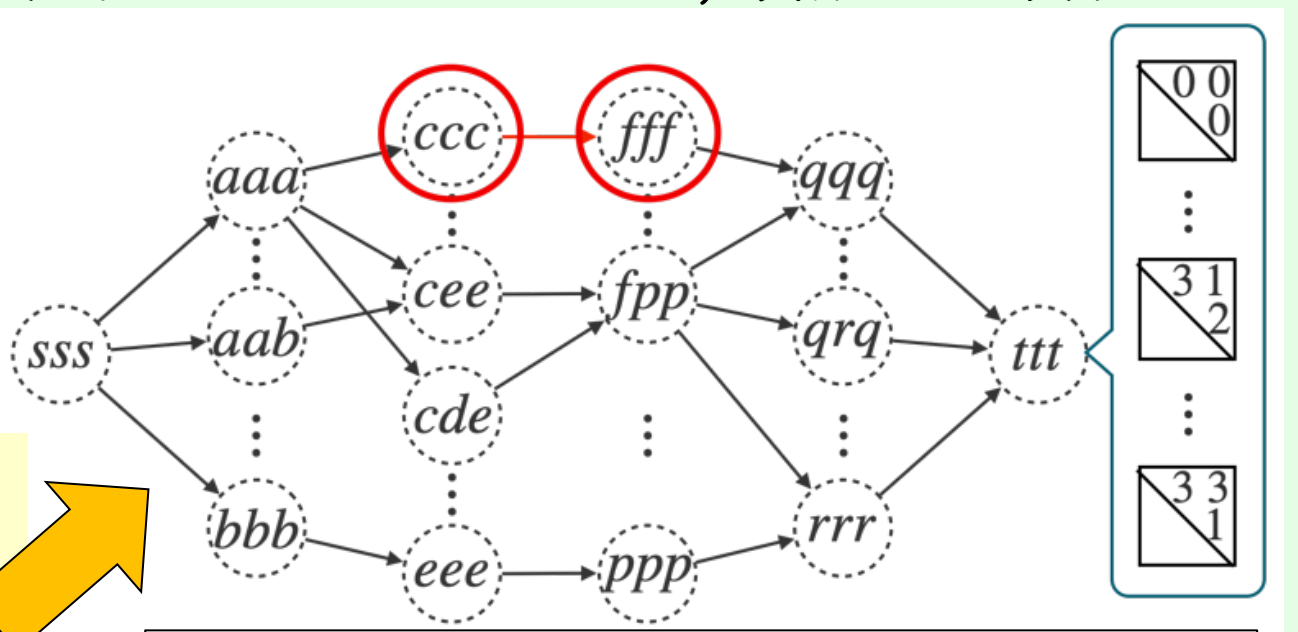
結果1：解数Kが限定されたとき

- **観察**：定数個の入力文字列のLCS全体は指数個あるが，多項式サイズのDAGに格納できる．
- **多項式時間アルゴリズム**：多項式サイズのDAG上で，頂点のK項組の上で動的計画法を実行する

LCS問題の入力文字列

X_1 = ABABCDDEE
X_2 = ABCBAEEDD





- K本のパスは，対毎のハミング距離を表すKxKの整数行列を定める．
- 頂点のK項組すべてに，根から到達可能なK本のパスすべての組合せの重み行列のリストを保持させる．

結果2：解数Kが非限定のとき

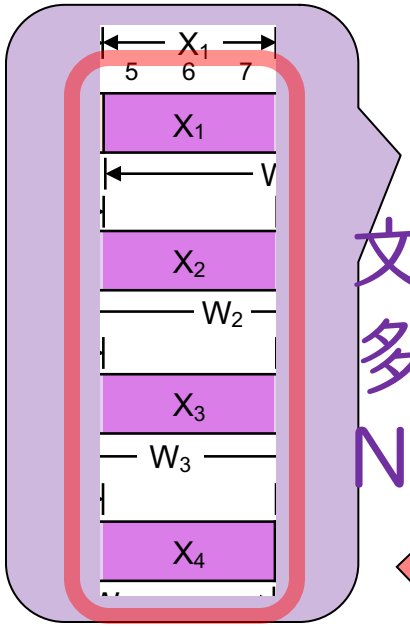
- **定理3**：Kが入力のとき，Max-Sum最大多様LCSs問題は，NP困難である．(入力文字列数が定数m ＝ 2でも)．Kがパラメタなら**W[1]困難**．
  Max-Minについても同じ結果

方針：より基本的な「文字列集合の多様性問題」の困難性を示し，もとの問題に帰着する

- **補題1**：Kが入力のとき，文字列集合のMax-Sum最大多様性問題はNP困難である．Kがパラメタなら**W[1]困難**である．

- **補題2**：文字列集合のMax-Sum多様性問題を，LCS集合のMax-Sum多様性問題に帰着できる．

アイディア：文字列多様性問題から，2本の文字列のLCS多様性問題への還元

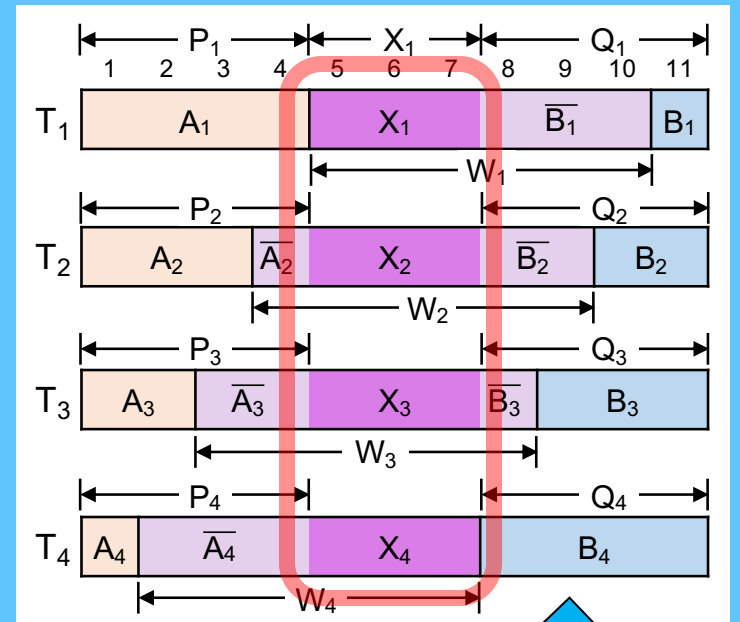


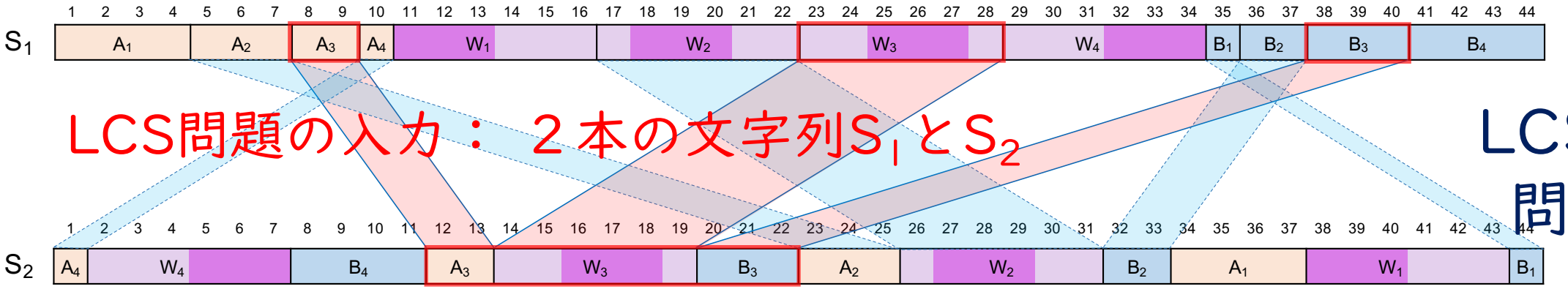
文字列集合の
多様性問題の入力：
N本の文字列

polytime
reduction

LCS
computa
-tion

LCS問題の解：
復元された元の
N本の文字列



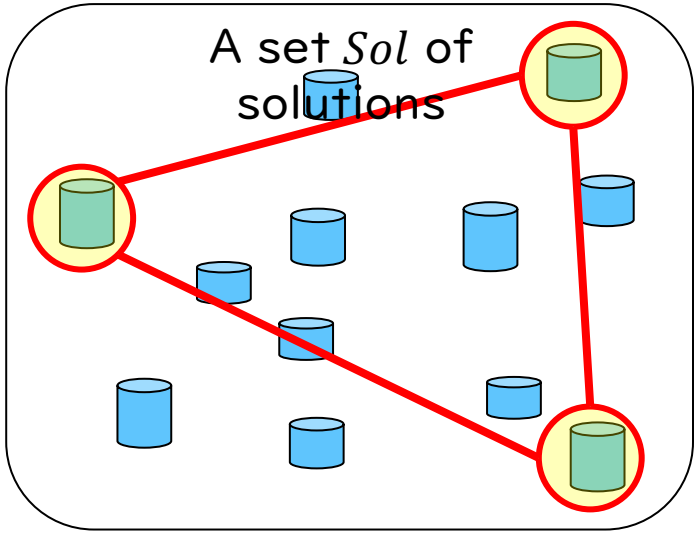


LCS問題の入力：２本の文字列S_1とS_2

LCS多様性
問題の解

結果3：解数Kが非限定のとき（その2）

- **定理**：Kが入力のとき，Max-Sum最大多様LCSs問題は，**PTASをもつ**．（任意の近似誤差 $\varepsilon > 0$に対して，多項式時間で近似可能である）



Select
K = 3
solutions

結果3：解数Kが非限定のとき（その2）

解法：Hanakaら(AAAI'23)にしたがい，次の定理を使う

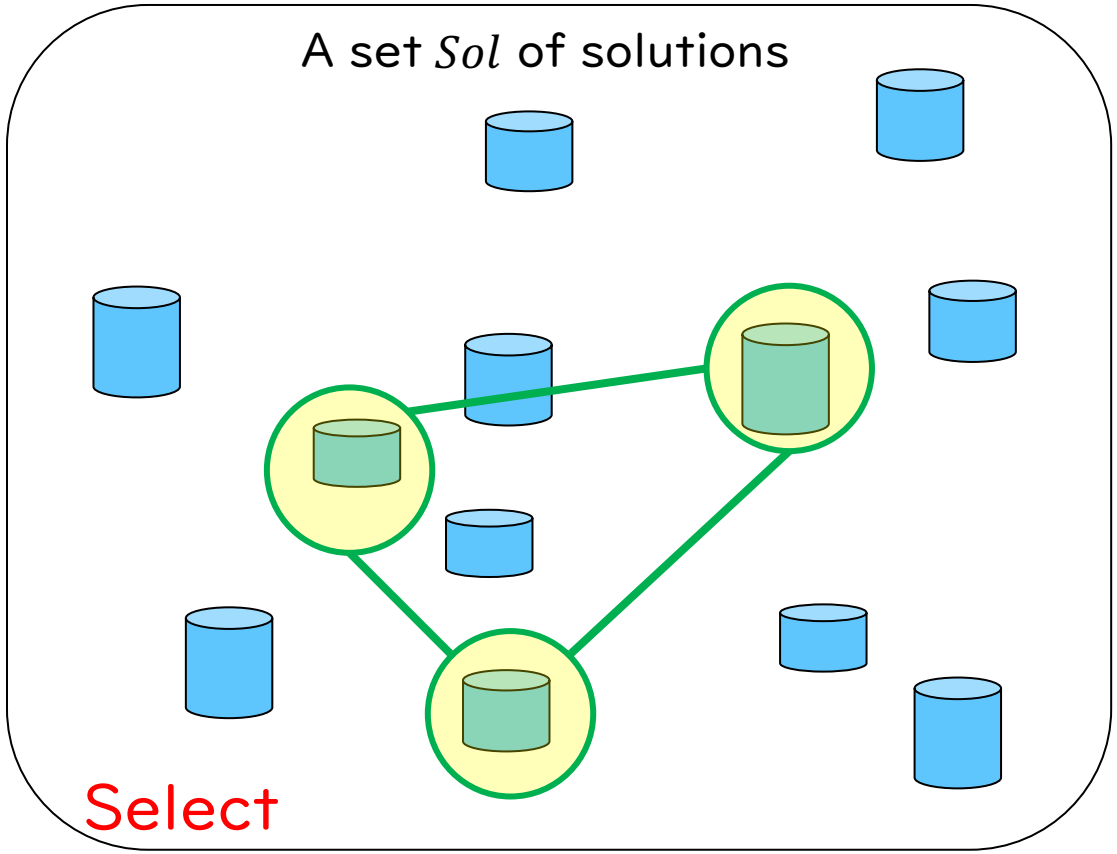
- **定理**（Cevallos, Eisenbrand, Zenklusenら，2019）：
  距離関数が「負値型不等式」を満たし，「最遠点問題」が多項式時間計算可能と仮定する．Kが入力のとき，Max-Sum版の点集合の最大多様性問題はPTASをもつ．

- 補題：文字列間のハミング距離は，「負値型不等式」を満たす準距離関数
- 補題：Kが限定されたときの動的計画法を用いて，最遠点問題を多項式時間計算可能

Tesshu Hanaka, Masashi Kiyomi, Yasuaki Kobayashi, Yusuke Kobayashi, Kazuhiro Kurita, and Yota Otachi: "A framework to design approximation algorithms for finding diverse solutions in combinatorial problems," *AAAI 2023*.

Alfonso Cevallos, Friedrich Eisenbrand, and Rico Zenklusen. "An improved analysis of local search for max-sum diversification," *Mathematics of Operations Research*, 44(4):1494–1509, 2019.

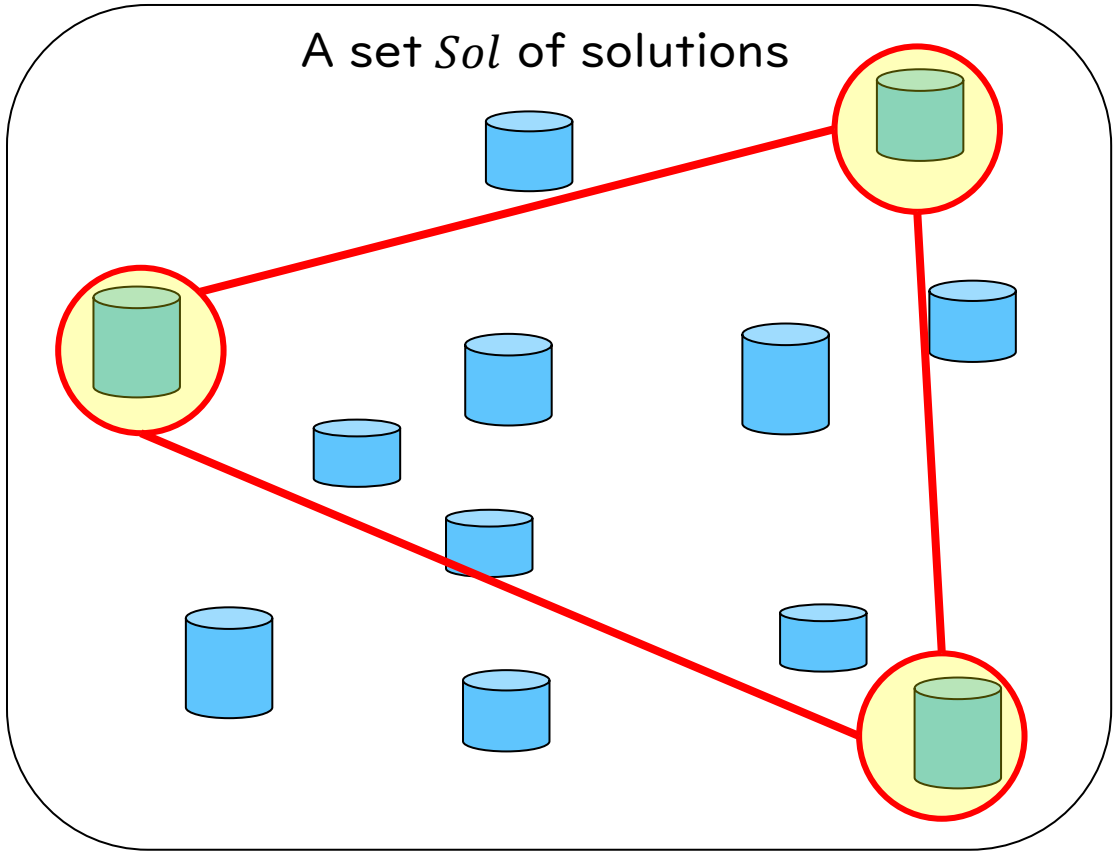
結果3：解数Kが非限定のとき（その2）



アイディア：次の局所探索（local search）法を用いる.

- 最初に，任意の解集合 S = {X_1, …, X_K}を選ぶ
- 次の処理をO(K^2 log K)回くり返す
  - 更新後の多様性D$(S - \{X\}) \cup \{Y\})$を最大にするように，Sから解Xを取り除き，代わりに解Yを入れる.
- 解集合Sを返す

結果3：解数Kが非限定のとき（その2）



K = 3

アイディア：次の局所探索（local search）法を用いる.

- 最初に，任意の解集合 S = {X_1, …, X_K}を選ぶ
- 次の処理をO(K^2 log K)回くり返す
  - 更新後の多様性$D((S - \{X\}) \cup \{Y\})$を最大にするように，Sから解Xを取り除き，代わりに解Yを入れる.
- 解集合Sを返す

まとめ：LCS問題の多様な解の計算量

Problem	Type	K: const	K: param	K: input
Max-Sum Diverse String &				ard on Σ-DAG 3:const rem 6.1) ard on LCS lary 6.1) rem 4.2)
Max-Min Diverse String &				ard on Σ-DAG 3:const rem 6.1) ard on LCS lary 6.1)
		or FPT	(Theorem 5.1)	

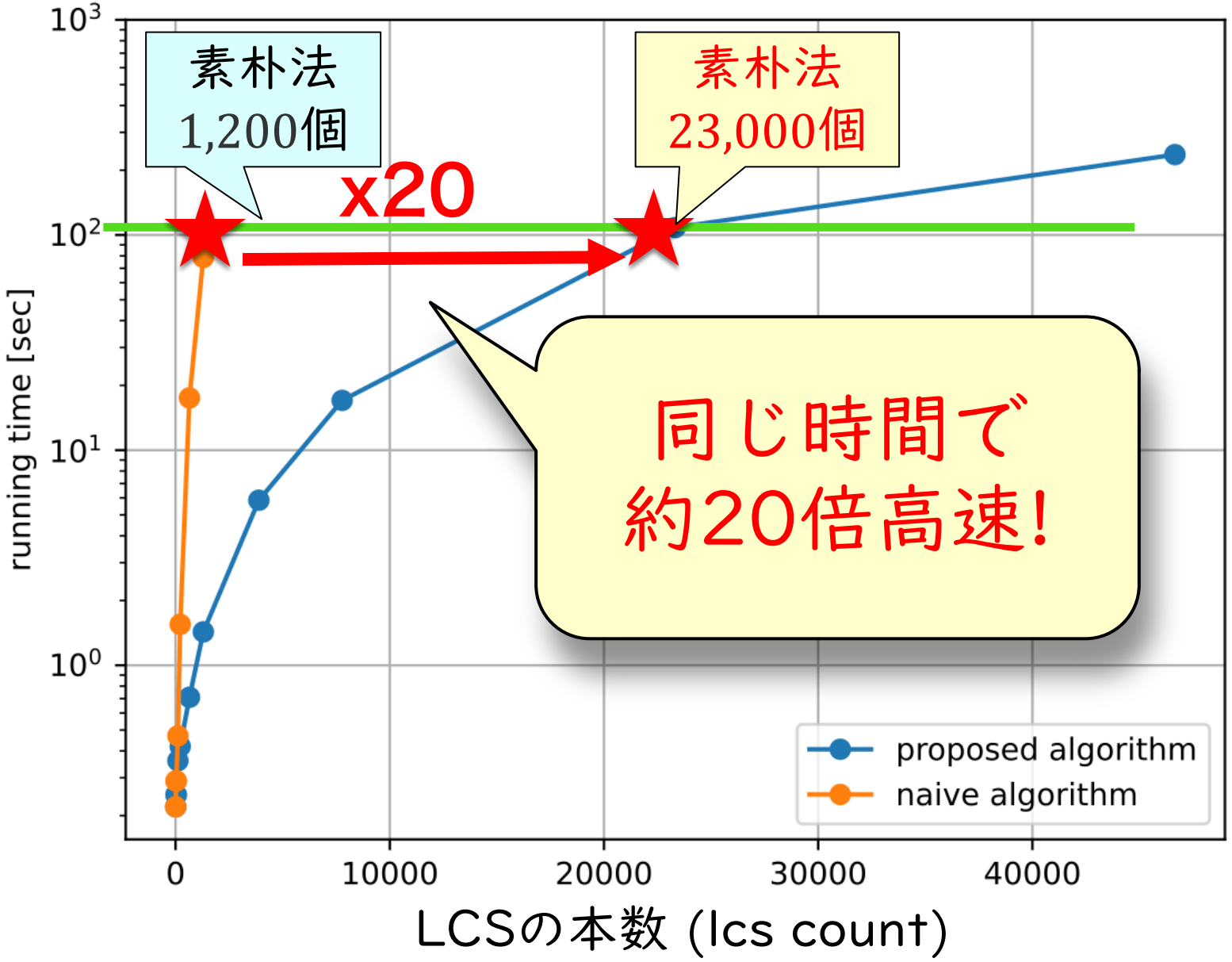
- **解の数Kが限定されたとき：**
  多項式時間計算可能.
- **Kが非限定のとき：**
  計算困難(NP困難)だが，任意の近似誤差の多項式時間近似が可能(PTAS).
- **Kがパラメタのとき：**
  固定パラメタ計算困難(W[1]困難).
  一方，Kとrがパラメタのとき，
  固定パラメタ計算容易.

実験：

志田 祐仁さん
北大情報科学院

- 次の2つの手法を比較した
  - 生成検査法：入力長nの指数時間
  - 提案手法：入力長nの多項式時間
- データ：次の2つの入力文字列（$m = 1,2,3, \ldots$）
  - $S_1 = (ABABCDDEE)^m(ABAB)^{\{0,1\}}$
  - $S_2 = (ABCBAEEDD)^m(ABCB)^{\{0,1\}}$
- 実験環境
  - PC(CPU Intel® CoreTMi5-1038NG7 2.00GHz, メモリ 16 GB 3733 MHz LPDDR4X,OS macOS 14.0 (23A344)), Python 3.9.12

実験1：規模耐性 (Scalability)



- 問題サイズ（実行解数＝LCSの総数）を増加させて計算時間を計測
- 解数 $k = 2$ で，制限時間$t = 110$ (sec)の入力サイズを比較した

まとめ: Max-Sum多様なLCS問題の計算量

解数Kが定数のとき

多項式時間計算可能

Proof: 動的計画法

- 本頁の全ての結果は，本研究で示した
- PTAS以外は， Max-Min版についても同じ結果が成立する

解数Kが入力のとき

NP完全

PTAS

任意誤差で多項式時間近似可能

Proof: ローカル探索法 (Cevallos+ 2019)

解数Kと入力長さrがパラメタのとき

FPT

固定パラメタ計算容易

Proof: 色符号化技法

解数Kがパラメタのとき

W[1]困難

固定パラメタ計算困難

Proof: p-クリークからのFPT帰着

今後の課題：機械学習における多様な最適モデルの学習

Our laboratory members

2024.3.25, IKN Laboratory
IST School, Hokkaido University

Thank you!