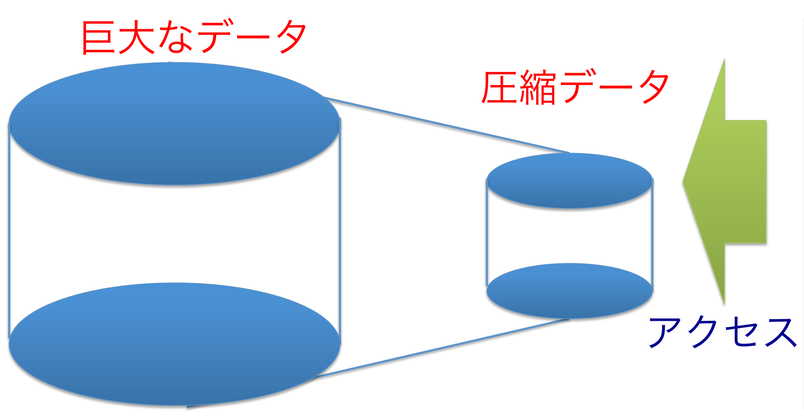
圧縮情報処理ユニット　　田部井 靖生

Succinct Information Processing Unit

Yasuo Tabei

目標：大規模データ解析のため圧縮情報処理技術の開発
利点：IoT機器などの限られた計算リソースでの大規模データ解析が可能



データ解析技術
- 類似度検索
- 線形モデルの学習
- カーネル法
- 文字列処理

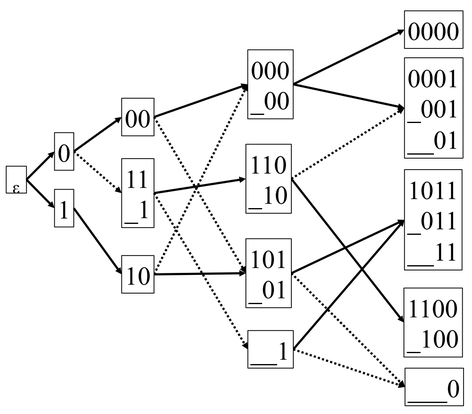
etc

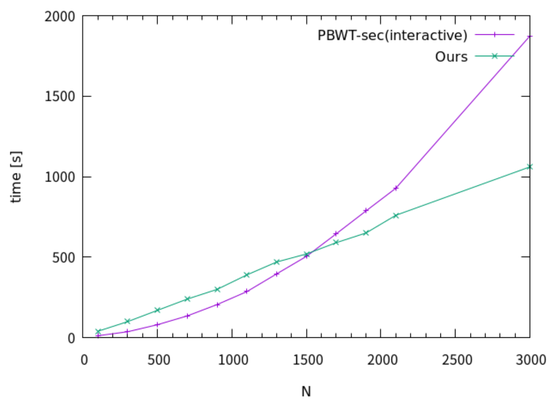
ゲノム秘匿検索 (IDASH コンペで3位入賞)

背景：ゲノムを暗号化したまま他所のデータベースに送信して分析したい（秘匿計算）
問題設定：clientのゲノムとデータベースのゲノム集合で最長一致している各部分文字列を返す (SMM)
問題点：従来手法は何度も通信するので非効率 [Shimizu et al.'16]

アイデア
- SMM問題を一度に解くオートマトンをブール回路で実装
- 完全準同型暗号上で任意のブール回路を評価できる
  高速暗号化演算TFHEを利用し、暗号化上での高速検索を実現

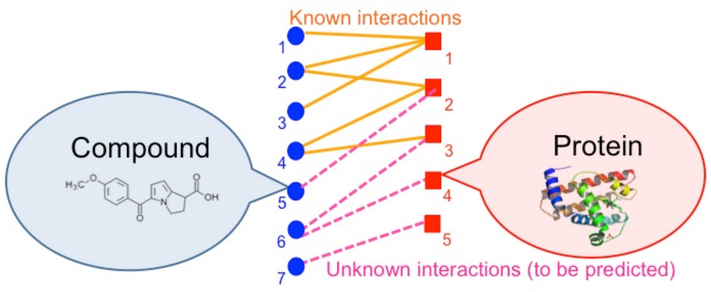
利点： 一回の通信のみで検索可能
成果：ゲノムの秘匿検索において、従来法よりも高速な検索が可能になった



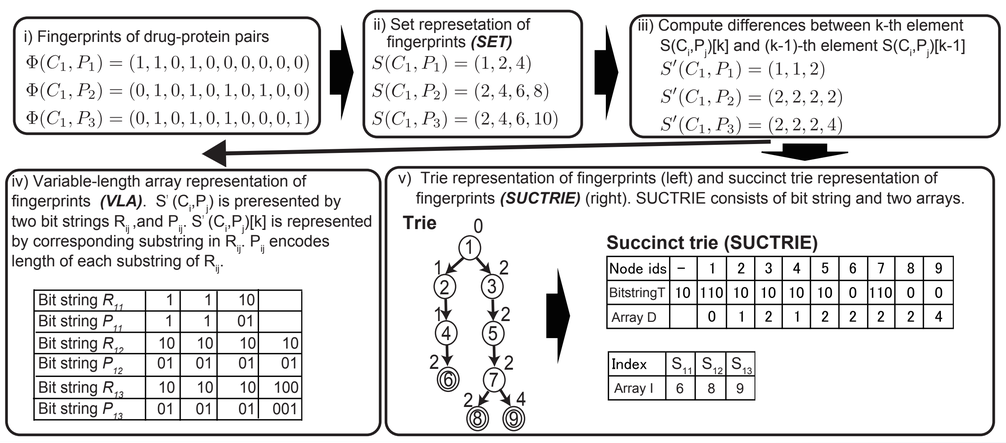


大規模バーチャルスクリーニングのためのドラッグとタンパク質の相互作用予測 (APBC'19)

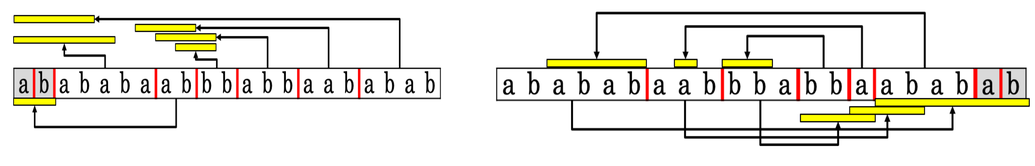
問題：相互作用既知のドラッグとタンパク質ペアーから予測モデルを学習し、ドラッグとタンパク質ペアーが相互作用するかどうかを予測
問題点:学習データ数がタンパク質数と化合物数の積となる
→ 大規模学習問題



提案手法：フィンガープリントで表現されたドラッグとタンパク質を圧縮し、圧縮されたデータ上で予測モデルをSVMで学習



LZRR: 双方向参照を利用したデータ圧縮法 (DCC'19)

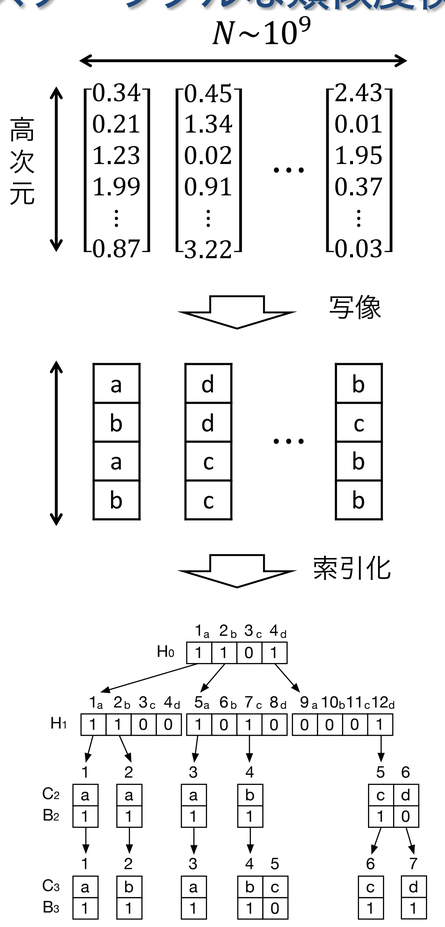


LZ_{77}（前方参照）　　提案手法（双方向参照）

問題：反復文字列と呼ばれるゲノムなどの文字列の圧縮
問題点：現在主流の可逆圧縮アルゴリズムはLZ77に基づく
　　LZ77（従来法）は前方参照のみで圧縮
　　後方参照を用いればより高圧縮を期待できる
アイデア：双方向参照可能なフレーズで貪欲に分解
結果:フレーズ数が常にLZ77以下であることを証明
　　実験ではLZ77とほぼ同じ圧縮速度と作業領域で圧縮率を5%小さくすることに成功

実験データ	フレーズ数		実行時間 [sec]		使用メモリ [MB]	
	LZRR	LZ77	LZRR	LZ77	LZRR	LZ77
fib41 (267MB)	**5**	22	113	99	11,978	6,542
rs.13 (216MB)	**51**	52	110	80	9,654	5,292
tm29 (268MB)	**31**	56	142	108	11,797	6,554
einstein.de.txt (92MB)	**31,798**	34,287	27	24	3,808	2,266
einstein.en.txt (467MB)	**83,368**	89,437	147	130	19,196	11,418
world leaders (46MB)	**165,626**	175,670	16	8	1,939	1,148
influenza (154MB)	**720,282**	779,213	51	42	6,351	3,781
kernel (257MB)	**741,556**	794,938	88	71	10,602	6,299
cere (461MB)	**1,597,657**	1,695,631	500	131	18,925	11,263
coreutils (205MB)	**1,359,606**	1,441,384	68	56	8,453	5,013
Escherichia Coli (112MB)	**1,962,013**	2,078,869	46	32	4,632	2,752
para (429MB)	**2,205,032**	2,338,919	203	125	17,609	10,481

スケーラブルな類似度検索のための簡潔なTrie索引の開発

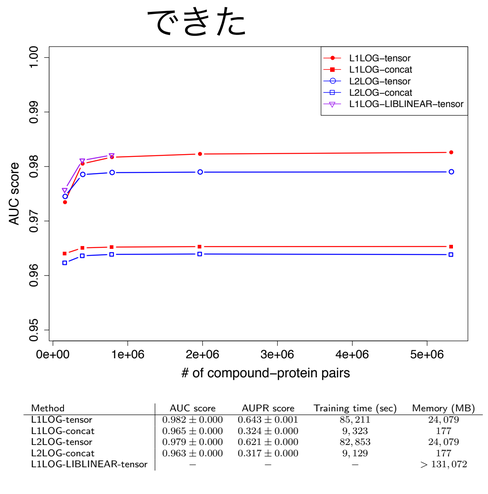


背景
Webページ、画像、分子データなどの高次元なベクトルデータに対し類似度検索をしたい
高次元なベクトルデータは低次元な整数ベクトルに写像することで効率的に類似度検索できるしかし、そのデータベースが大規模な場合、素朴な検索では時間が掛かりすぎる！
→ 効率的な索引技術が必要！

提案手法
Trie木を用いて索引を構築
枝刈りなどを応用し、高速に類似度検索
簡潔データ構造とヒューリスティックを組み合わせることで高いメモリ効率を実現

実験結果
10億オーダーの高次元ベクトルデータベースに対する索引をたった10GBで構築可能！
→ 既存の索引技術は30GB〜40GBを要する既存手法と比べ、最大で10倍以上高速に検索可能！

結果：530万の学習データ(131GB)を24GBに圧縮した状態でSVMでモデル学習可能。精度も高精度。
予測に有効な特徴量も抽出可能であり、創薬に有効な新規特徴が抽出できた



Method	AUC score	AUPR score	Training time (sec)	Memory (MB)
L1LOG-tensor	0.982 ± 0.000	0.643 ± 0.001	85, 211	24, 079
L1LOG-concat	0.965 ± 0.000	0.324 ± 0.000	9, 323	177
L2LOG-tensor	0.979 ± 0.000	0.621 ± 0.000	82, 853	24, 079
L2LOG-concat	0.963 ± 0.000	0.317 ± 0.000	9, 129	177
L1LOG-LIBLINEAR-tensor	–	–	–	$>131,072$