

2002年 3月 5日

# PrefixSpan法を用いた モチーフ発見システム

広島市立大学 情報科学部

蒲原朋樹, 森康真, 北上始, 黒木進

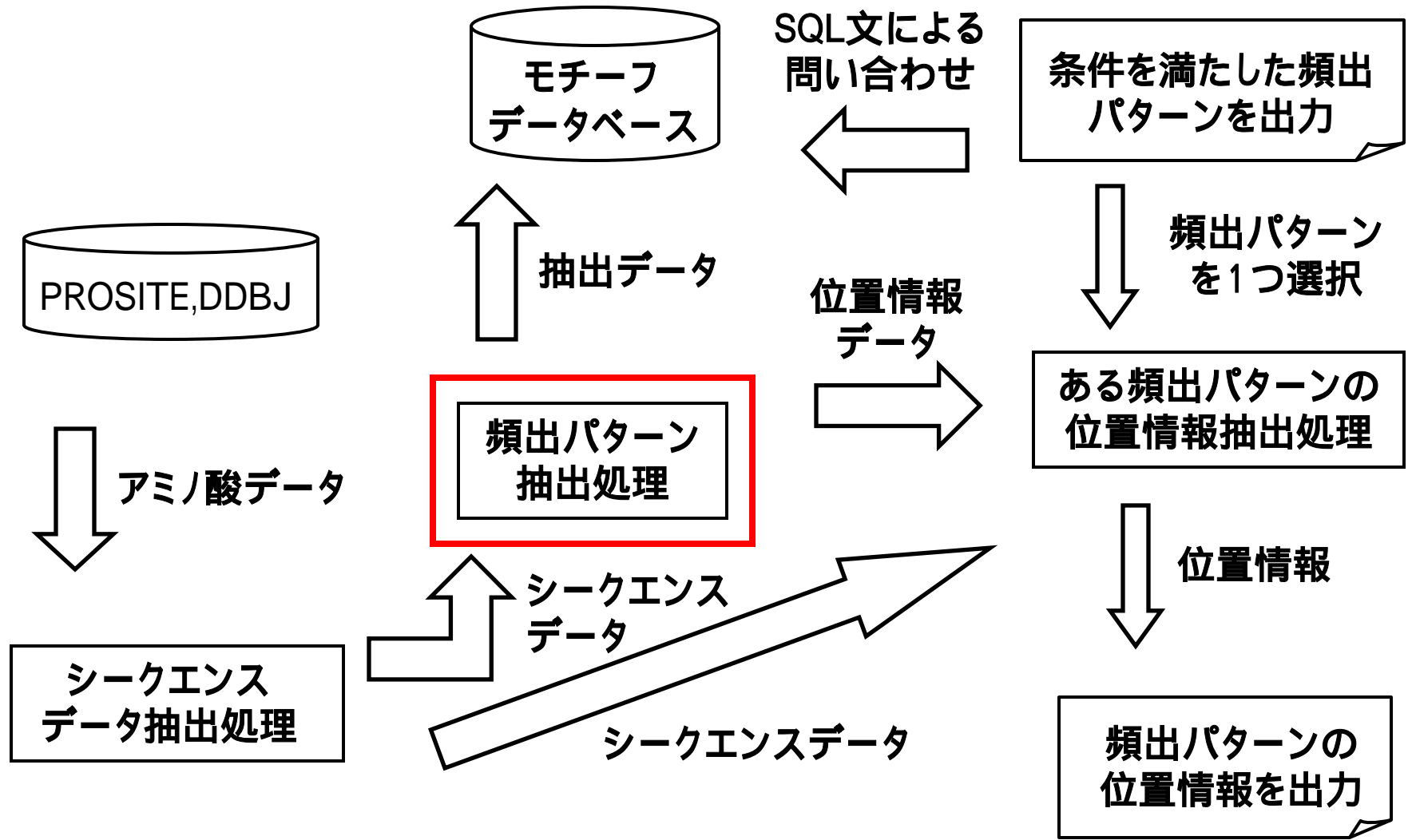
# 背景

- 大規模な分子生物学データベースの利用
  - DNA配列データ, アミノ酸配列データ等
- モチーフの発見
  - マルチプルアライメントの使用
  - 最適に並び変えることによる共通パターンの抽出
  - 全ての共通パターンを求めることが困難

# 研究内容

- PrefixSpan法の改良
  - 可変ギャップ法
  - 固定ギャップ法
- インタフェースの作成
  - モチーフ発見の支援

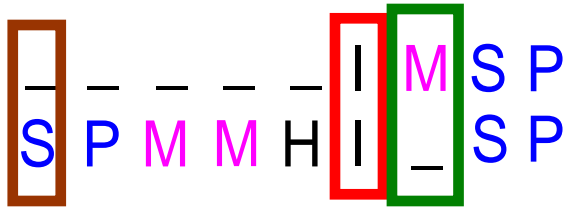
# システムの処理手順



# マルチプルアライメント

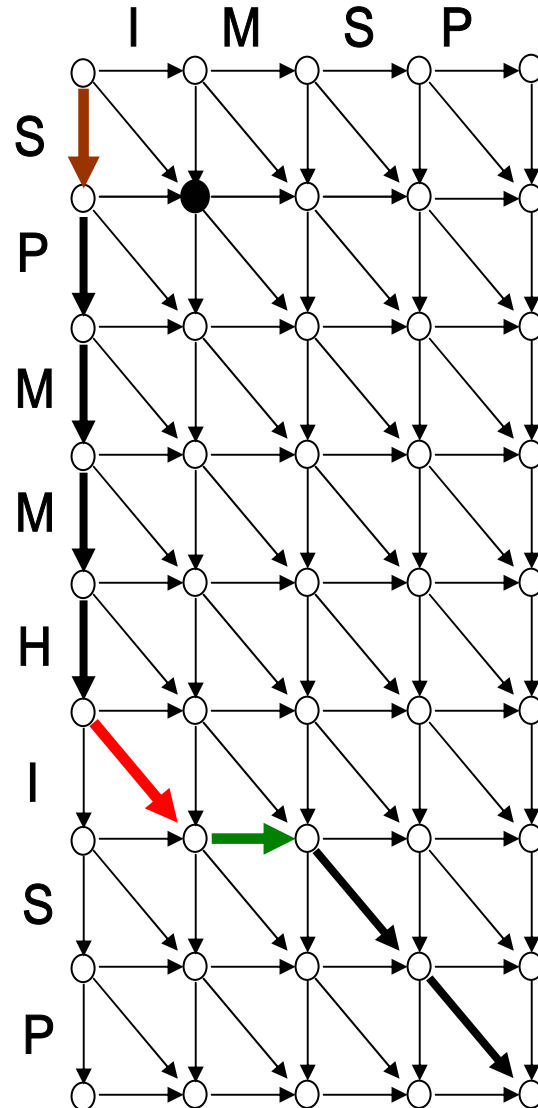
2つの配列(シーケンス)を対象とするDP法を考える

DP法によって, 2つのシーケンスを最適に並びかえる(ギャップを挿入する)



解(共通パターン)を全て求めることができない

3つ以上の配列を対象とする場合, 2つの配列を比較した結果を利用することにより求める



# PrefixSpan法 その1

最小支持率を満たした共通パターンを短いパターンから抽出する(頻出パターン)

MFKALRTIPVILNMNKDSKL

MSPNPTNIHTGKTLR

最小支持率100%を満たす頻出パターンM

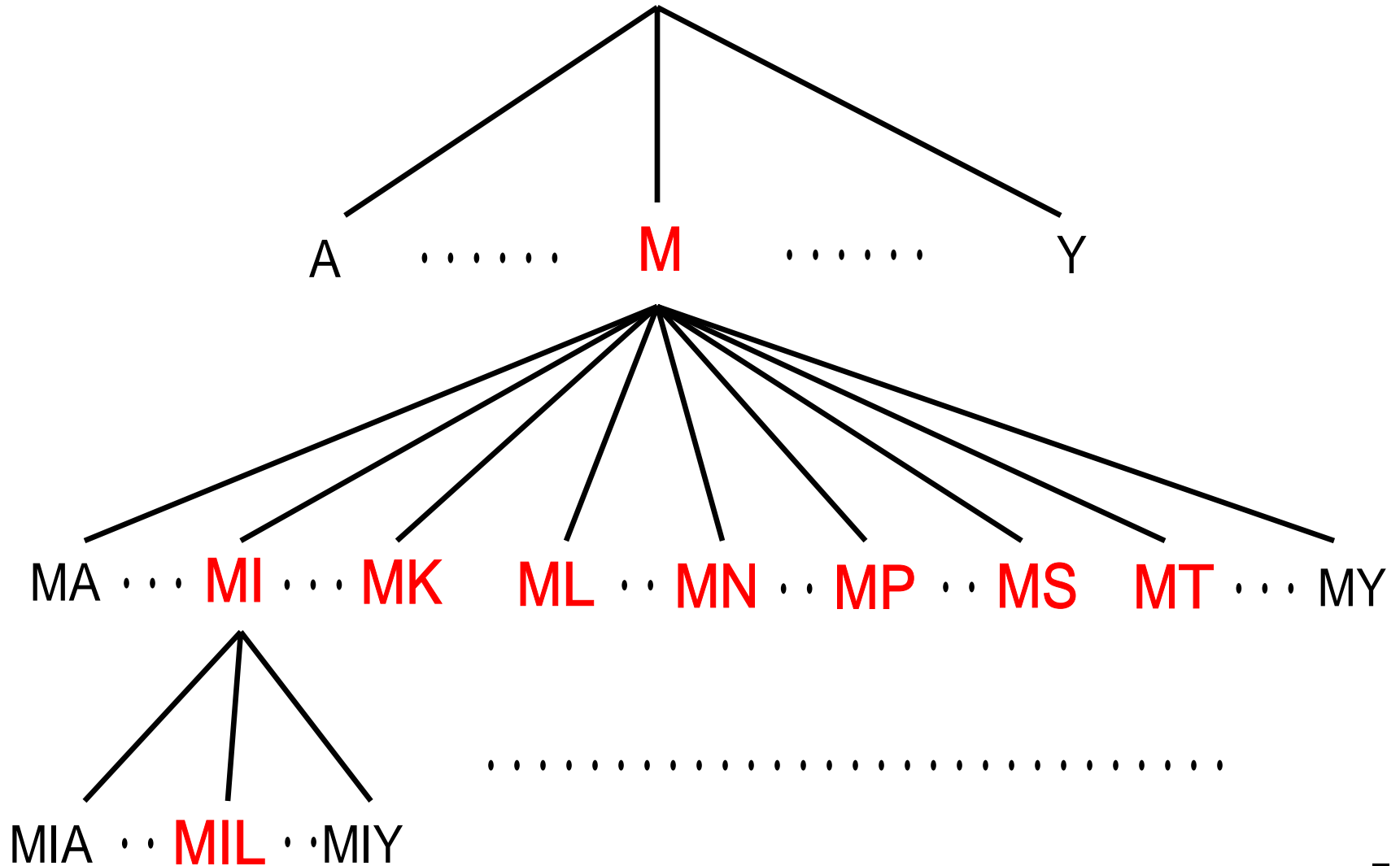
MFKALRTIPVILNMNKDSKL → M3L

MSPNPTNIHTGKTLR → M12L 数字はギャップ数

頻出パターンMを用いて, 接頭辞Mを持つ長さ2の頻出パターンを抽出する

(Projected Databaseを作成する)

# PrefixSpan法 その2



# PrefixSpan法の改良

- ギャップ制限を導入する
- 抽出される頻出パターンに対して, 2種類の改良方法を提案する
  - 可変ギャップ法: 異なるギャップ数を持つパターンは同じパターンと見なす方法である
  - 固定ギャップ法: ギャップ数が異なる場合, 異なるパターンと見なす方法である



# 可変ギャップ法 その1

ギャップ制限を導入することにより, 頻出パターンを絞る (最小支持率100%, 最大ギャップ数3)

MFKALRTIPVILNMNKDSKL

MSPNPTNIHTGKTLR

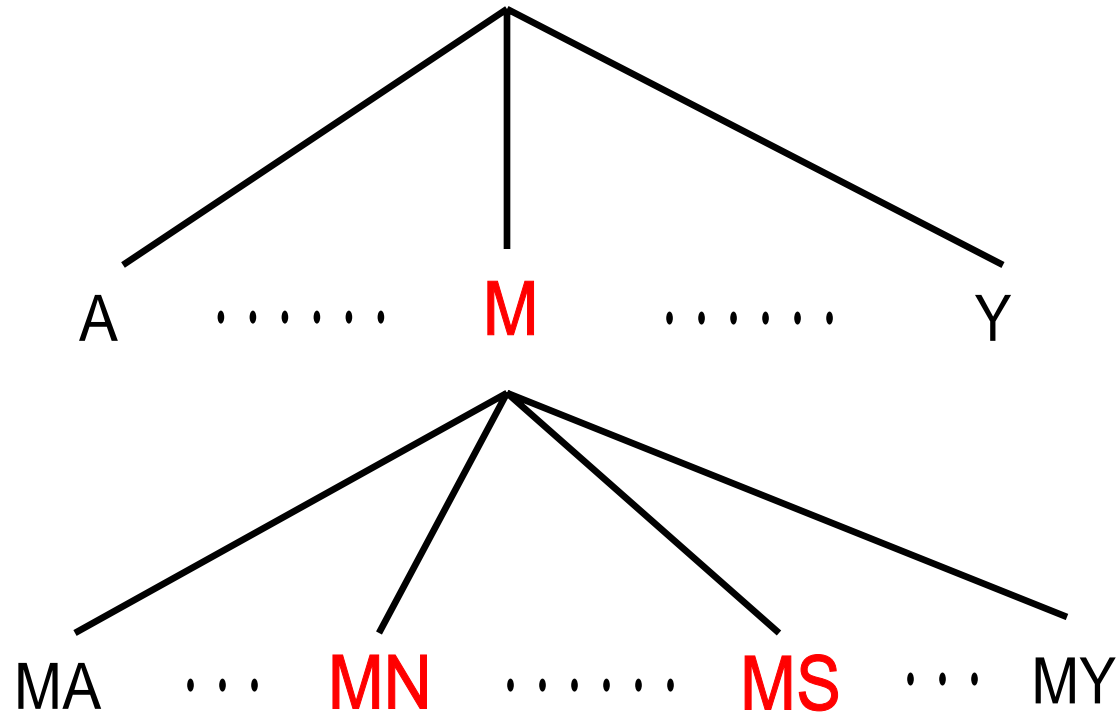
1つのシーケンスに対し, 複数の頻出パターンを考慮する

MFKALRTIPVILNMNKDSKL

MSPNPTNIHTGKTLR

Projected Databaseの範囲について, 最大ギャップ数を考慮する

# 可変ギャップ法 その2



長いシーケンスやギャップ数を大きく設定した場合、  
頻出パターンが多く抽出される

# 固定ギャップ法

最大ギャップ数を増やした場合の頻出パターンの絞込み



ギャップ数ごとに区別して、頻出パターンを抽出

例) MSALDSL  
MLBLAAL

最小支持率100% , 最大ギャップ数5

MAとMLを抽出することを考えてみる

## 可変ギャップ法の場合

MA,MLともに抽出される(MとA,Lの間隔が5文字以内)。

## 固定ギャップ法

MAに関しては、同じギャップ数を持つMAがないため、抽出されない。

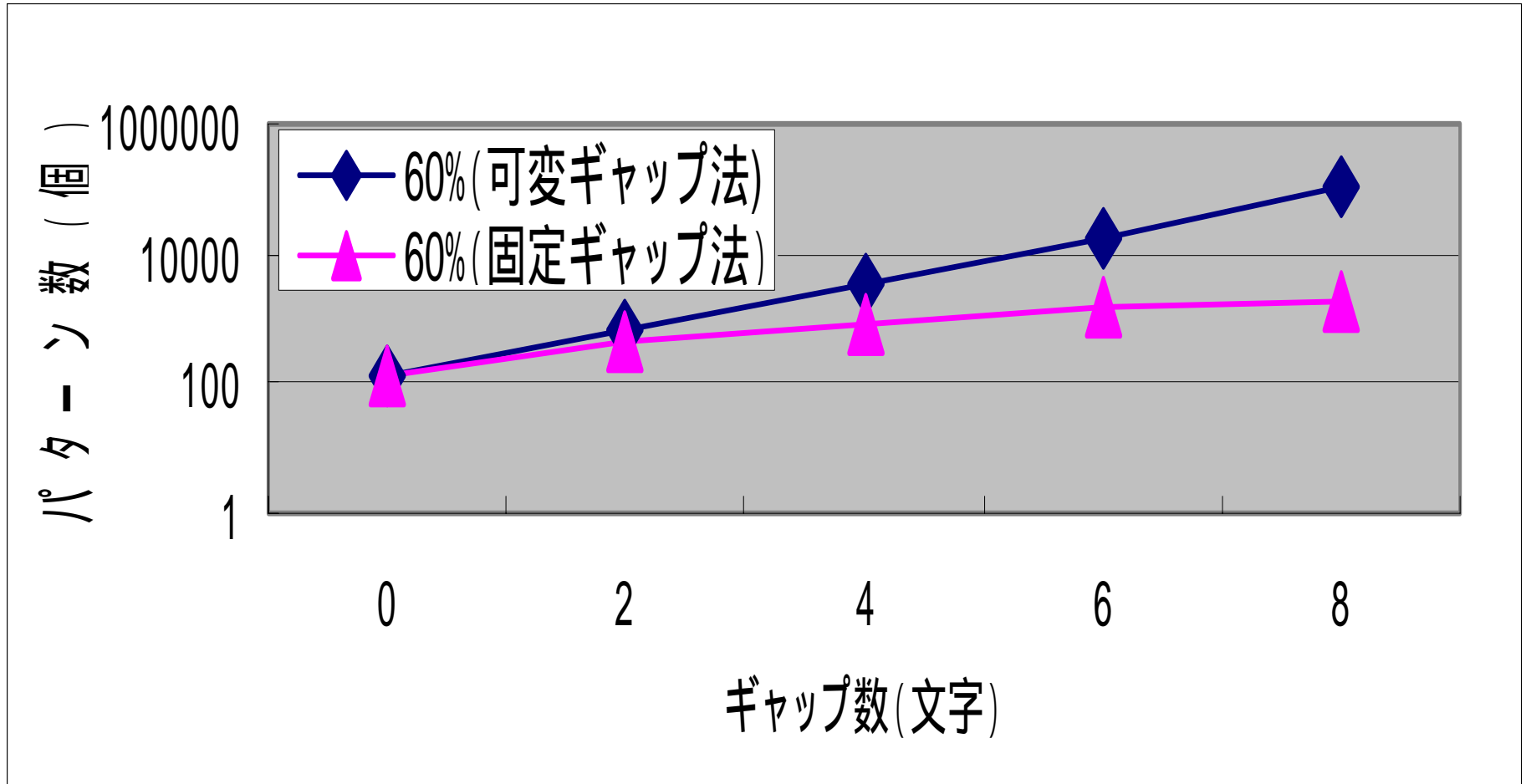
MLに関しては、M2LとM5Lが抽出される。

# モチーフ抽出の例

データセット	モチーフ	モチーフの支持率 (%)
データセット1	H3H7C2C (Zinc Finger)	69
データセット2	F3GC6[FY]5C (Kringle)	94 ( <b>F</b> の方)

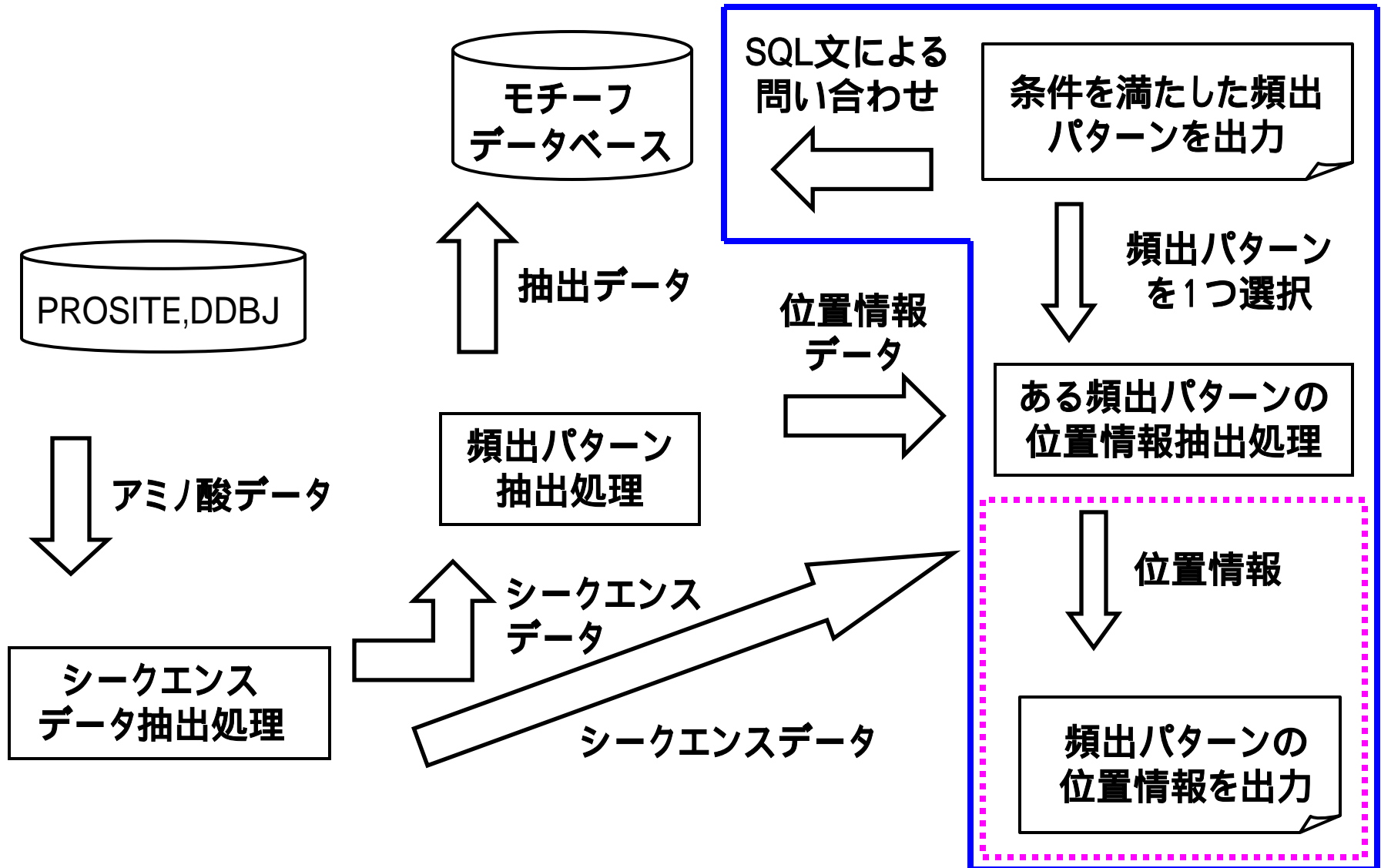
Kringleのパターン[FY]の部分は、FかYのどちらかという意味

# 測定結果



データセット1について, 行った(使用データのスライド参照)

# システムの処理手順



# インタフェース(位置情報)

Zinc Finger のモチーフである  
H3H7C2Cの位置情報

右図より, パターンの位置や数に関係なく, パターンが抽出されている

```
http://www.db.its.hiroshima-cu.ac.jp/~lanbara/syuka/en/cgi-bin/resu#2.cgi?pattern=H3H7C2C&inc2 - Microsoft Internet Explorer
ファイル 編集 表示 お気に入り ツール ヘルプ
戻る 進む 中止 更新 ホーム 検索 お気に入り 履歴 メール サイズ 印刷
アドレス http://www.db.its.hiroshima-cu.ac.jp/~lanbara/syuka/en/cgi-bin/resu#2.cgi?pattern=H3H7C2C&inc2=TRUE
Zinc finger パターン:H3H7C2C
シークエンス番号:1 アクセション番号:013089 ファイルサイズ:522
NEMEEAQMESQMPGRDSPPPNDVSEENDEAMP I PEDLSASSNLQHNRRGDKKEGLACNIKVEARCDENGL
AIDMMMNGEEEECAEDLRVLDASGAKVNGSHAGGPDSKGPYSSAGGIRLPNGKCLKDCIGIVCIGPNVL
NVHKRSHTGERPFQCTQCGASFTQKGNLLRH I KLHSGEKPFKCHLCNYACRRRDALSGHLRTHSVGKPHK
CAYCGRSYKQRSSLEEHKERCHNYLQCMGLQNS I YTVVYKESNQNEQREDLSOMGSKRALVLDRLANNVA
KRKSTMPQK FVGEKRFSN I SFEGGPGELMOPHY I DQAINSA I NYLGAESLRPL I OTSPTSSDMGVMGSMY
PLHKPPAEGHGLSAKDSAAENLLLLAKSKSASSEKDGSPSHSGQDSTDTESNNEEKAGVGASGL I YLTNH
ITSGVRNGVLPVKEEQROYEAMRASIE I ASEGFKVLSGEGEQVRAYRCEHCRILFLDHVMT I IHMGCH
GFRDPFECNL CGHRSQDRYEFSSHMT RGEHRY
シークエンス番号:2 アクセション番号:P17034 ファイルサイズ:58
YECSECGKSFRQRSL I QHRR LHTGERPYECSECGKSFSQSASL I QHQRVHTGERP
シークエンス番号:4 アクセション番号:P51508 ファイルサイズ:337
RDEKLYICTKCGKAF I QNSEL I NHEKHTHTREKPYKNECGKSFFQVSSLLRHQTHTTGEKLFECSECGK
FSI NSAI NTHDK I HTGFRHHKCSFCGKAFTOKSTI RMHOR I HTGFRSY I CTQCGQAF I QKAHI I AHDRTH
ページが表示されました インターネット
```

# まとめ

- PrefixSpan法の改良
  - 可変ギャップ法, 固定ギャップ法を提案することにより, 頻出パターンを絞ることを可能にした
- インタフェースの作成
  - 固定ギャップ法により, 抽出された頻出パターンをさらに細かく出力することが可能である
  - マルチプルアライメントと比べて, 抽出パターンの取りこぼしがない



# 今後の課題

- 複数通りのパターンを持つ頻出パターンの抽出
  - 前提知識を用いたパターン発見
    - 例) アミノ酸FとYは構造が似ているため, 同一のものともみなす (KringleのF3GC6 [FY]5C:ただし, 先頭のFは例外)
- インタフェース部分の拡張
  - 複数のパターンの位置情報を同時に表示する
  - モチーフの包含関係が見つかりやすくなる

# 分子生物学データベース

- DNA配列データ
  - アデニン、グアニン、シトシン、チミンと呼ばれる4つの塩基が鎖状に結合した配列構造(アルファベットのA, T, G, Cで表現)
- アミノ酸配列データ
  - 20種類のアミノ酸が鎖状に結合した配列構造(B, J, O, U, X, Zを除いた20のアルファベットで表現)

# モチーフ

- Zinc Finger
  - 植物の成長因子
- Kringle
  - アルツハイマー病にかかる可能性がある人に見られるモチーフ

# 開始位置の記憶方法(長さ1)

行: アミノ酸を表す20の  
アルファベット(固定)

列: シークエンスデータ  
の件数(データ数により  
変動)

開始位置に関しては,  
複数存在するため,  
線形リストに格納する

Nは, NULL文字

	支持率	10			20
A	50%	5	N		N
C	0%	N			N
⋮	⋮	⋮	⋮	⋮	⋮
M	100%	2	15	N	2
⋮	⋮	⋮	⋮	⋮	⋮

# 開始位置の記憶方法(長さ2)

MのProjected Database

番号 10

MFKALRTIPVILNMNKD  
SKL

番号 20

MSPNPTNIHTGKTLR

	支持率	10		20	
:	:	:	:	:	:
D	50%	18	N	N	
:	:	:	:	:	:
N	100%	16	N	5	N
:	:	:	:	:	:
S	100%	19	N	2	N
:	:	:	:	:	:

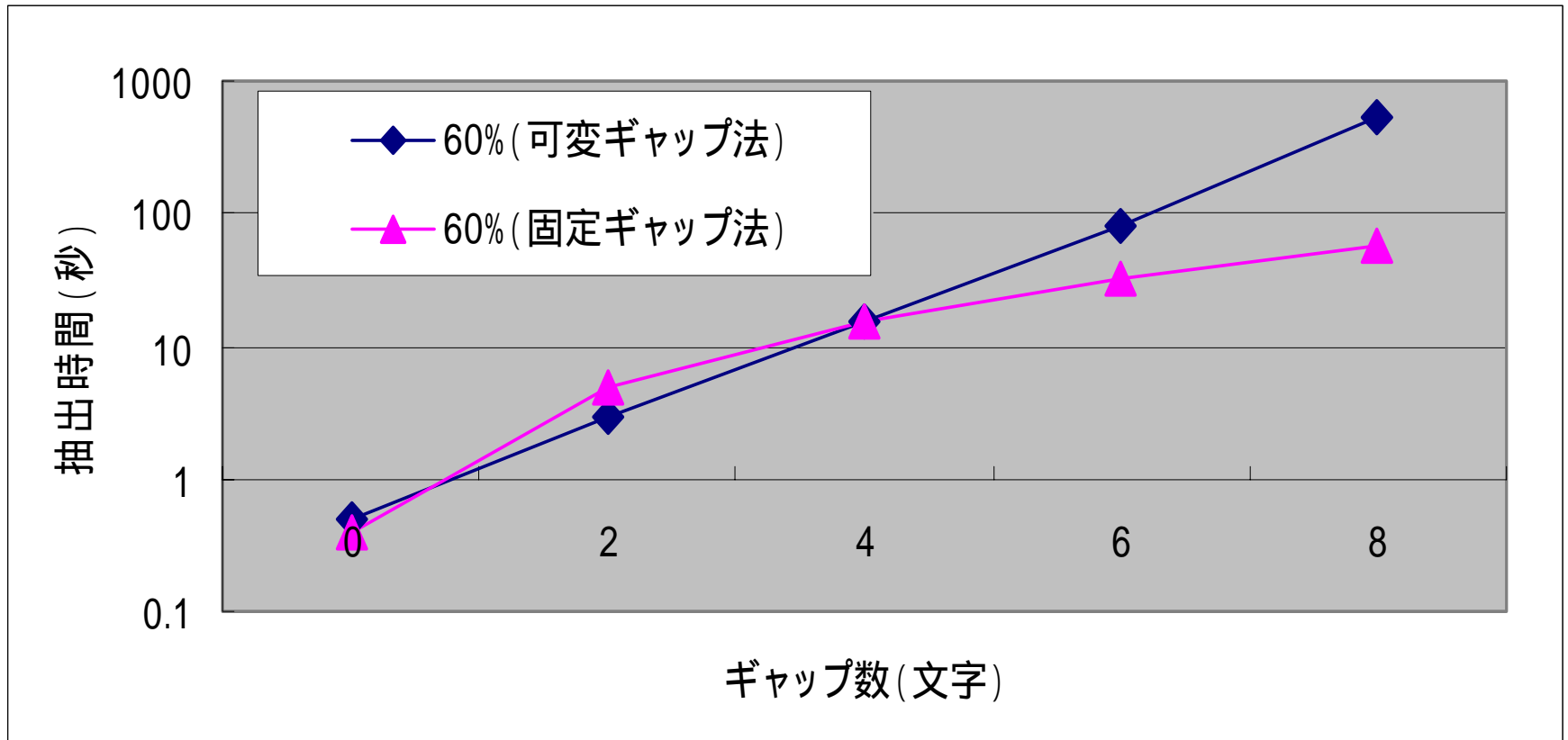
# 使用データの詳細(データ)

データセット	データ件数(件)	平均長(byte)	最大長(byte)	最小長(byte)
データセット1	467	525	4036	34
データセット2	70	334	3176	53

# 使用データの詳細(時間)

データセット	抽出時間 (s)	最小支持 率(%)	最大ギャッ プ数(文字)
データセット1	96.2	60	10
データセット2	60.7	60	10

# 測定結果(抽出時間)



データセット1について, 行った(「モチーフ抽出の例」のスライド参照)

PrefixSpan法に関しては, 支持率95%, 90%において, 約12分, 5時間以上かかった



# インタフェース(入力画面)

ソートする項目の選択  
(支持率, 最大ギャップ  
数, 最小ギャップ数,  
ギャップを含まない長さ,  
ギャップを含む長さ: 第3  
キーまで選択可能)

The screenshot shows a web browser window with the URL <http://www.db.hiroshima-u.ac.jp/~labana/nyuhan/cgi-bin/epu2.html>. The page title is "入力画面" (Input Screen). The interface includes several sections highlighted with red circles:

- 配列データの選択** (Array Data Selection): A dropdown menu showing "cnc (mem-047) PROSITE".
- 出力したい項目をチェックしてください** (Please check the items you want to output): A row of checkboxes for "パターン名", "支持率", "最大ギャップ数", "最小ギャップ数", "ギャップを含まない長さ", and "ギャップを含む長さ".
- 各項目に対して、検索条件を選択してください** (Please select search conditions for each item): A grid of dropdown menus for "支持率", "最大ギャップ数", "最小ギャップ数", "ギャップを含まない長さ", and "ギャップを含む長さ".
- ソートする項目を選んでください(最大第3キーまで)** (Please select the item to sort (up to the 3rd key)): A table with columns for "第1キー", "第2キー", and "第3キー", each with a dropdown menu.

At the bottom, there are buttons for "実行" (Execute) and "リセット" (Reset).

# インタフェース(頻出パターン)

Cytochrome Cを選択  
した場合の頻出パ  
ターン出力結果

入力画面の条件を  
満たした頻出パ  
ターンを出力

以下にパターンを入力すると、そのパターンの配列上の位置が表示されます。

pattern	support	max_gap	min_gap	pat_len	all_len
CH	98	0	0	2	2
C2C	98	2	2	2	4
C3H	98	3	3	2	5
C2CH	98	2	0	3	5

(4 rows affected)

ページが表示されました

インターネット