PrefixSpan 法を用いたモチーフ発見システム

論文番号 C4-3
蒲原朋樹　　森康真　　北上始　　黒木進
広島市立大学大学院情報科学研究科

1. はじめに

モチーフは，アミノ酸配列上における特徴的なパターンであり，生物の進化の過程で保存されてきた蛋白質の機能に関係していると考えられている．そこで，分子生物学分野の様々な専門家が，自分達の研究に関連する配列を多数集め，集められた多数の配列の長さを等しくするマルチプルアライメント処理により，様々なモチーフを発見した．それらは，PROSITE[1]や DDBJ[2]といったモチーフライブラリーとして整備されている．

しかしながら，一般にマルチプルアライメント処理[3]により見つけ出される共通パターンは，関連する配列のどの 2 つを比較しても，配列上でほぼ同じ位置にあるものに限定されている．このため，配列によって大きく異なる位置にある共通パターンを見つけることが難しいという問題がある．さらに，DNA データやアミノ酸データを十分に調べようとするとき，任意に選択された大変多くの配列を対象とする．その場合に，マルチプルアライメント処理だけで特定の配列に共通なパターンを全て見つけ出すには大変な手間がかかるという問題がある．

本研究では，これらの問題を解決するために有用な Prefix Span 法[4]を改良し，多数のアミノ酸配列からモチーフ発見を支援するシステムを提案する．また, 部分試作したプロトタイプシステムを用いて，Zinc Finger，Cytochrome C，Leucine Zipper，Kringle などを含むアミノ酸配列データベースからモチーフ発見を試みる．なお，以後では，配列をシークエンスと呼ぶことにする．

2. マルチプルアライメント

塩基データやアミノ酸データの類似性解析の中でも基本的なもののひとつは，複数のシークエンスの類似する部分を縦に揃えて並べ合わせる操作で，これをマルチプルアライメントと呼ぶ．この方法は，分子生物学分野において，多用されている．

また，マルチプルアライメントを行うために用いる方法に，DP（ダイナミック・プログラミング）がある[5]．DP 法は，基本的には 2 つのシークエンスが与えられたときに，適当な場所にギャップを入れてずらすことにより，両者で対応する文字の一致数が最大になるように，並べる方法である．

例として，２つのシークエンス「IMSP」と「SPMMHISP」を考え，これらのシークエンスから頻出パターン SP を抽出することを考える．

また，DP 法は，比較するシークエンスを横方向と縦方向に並べた行列の形で説明することができる．

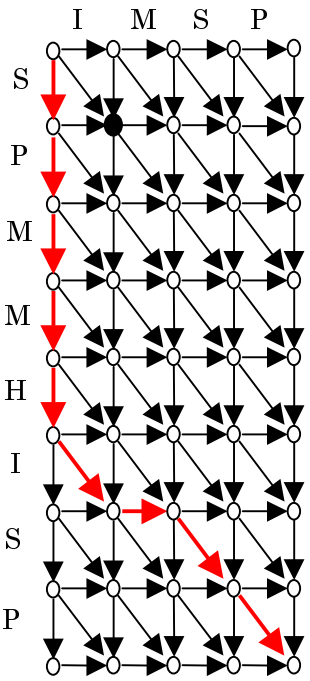

図 1 行列によるアライメント処理

図 1 では，「IMSP」と「SPMMHISP」を比較して，対応する文字の一致数が最大になるようなルートを求めている．

２つのシークエンス「IMSP」と「SPMMHISP」を比較する場合，1）両者のシークエンスから 1 文字もってきて並べる場合，2）「IMSP」からのみ，1 文字もってきて並べる（「SPMMHISP」にギャップを入れる）場合，3）「SPMMHISP」からのみ，1 文字もってきて並べる（「IMSP」にギャップを入れる）場合の 3 つがある．これは，図 1 の丸印から 3 つの矢印が出ていることに相当する．斜め方向は 1）の場合，横方向は 2）の場合，縦方向は 3）である．

従って，「IMSP」と「SPMMHISP」を比較する場合，図 1 の左上端の丸印から右下端の丸印まで，最も斜め方向に進んだ回数が多いルートが，比較結果となり，図 1 のようになる．また，実際にギャップ（ _ をギャップとする）を入れた結果を以下に示す．

_ _ _ _ _ I _ S P
SPMMHI _ S P

結果より，2 つのシークエンスの右側にある SP は並べて比較できる（アライメント）．しかし，「SPMMHISP」の左側にある SP はアライメントされていない．従って，全ての SP を見つけることができないといえる．

また，マルチプルアライメントでは，3 本以上のシークエンスを対象としている．例えば，3 本のシークエンスを比較する場合，2 本のシークエンスの比較結果を 3 本目のシークエンスと比較することによって，求める．4 本以上の場合も同様である．

3. 頻出パターン抽出のアルゴリズム

既存の PrefixSpan 法 を直接利用すると不要なパターンが数多く抽出されるので，これにギャップ数を制限する仕組みを導入することにより，計算時間の短縮や余分な頻出パターンを削減している．

また，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．

(1) 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．ギャップ長を無限大と設定すれば，既存の PrefixSpan 法と同じになる．
(2) 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．

3．1 既存の PrefixSpan 法

まず，既存の PrefixSpan 法について，説明する．表 1 に例として, 2 つのシークエンスデータを挙げ, PrefixSpan 法を用いた頻出パターン抽出について説明する．

PrefixSpan 法では, 抽出する頻出パターンの条件として, 最小支持率を設定する. 最小支持率は, 100%（番号 10，20 のうち 2 箇所に同じパターンが出現していることを示す）とする．

表 1 シークエンスデータベース S

シークエンス番号	シークエンス
10	MFKALRTIPVILNMNKD SKL
20	MSPNPTNIHTGKTLR

Prefix Span 法では，短い頻出パターンから求めていく．従って，まず長さ 1 の文字から求める．条件を満たす長さ 1 の文字は，「I，K，L，M，N，P，R，S，T」である．

そして，長さ k（k≧2）の頻出パターンは，長さ k-1 の頻出パターンを用いて，求めることが可能である．ここでは, 長さ 1 の頻出パターン M を用いて, M から始まる全ての頻出パターンを求めるところまで説明する．

頻出パターン M から始まる頻出パターンは, 図 2 のような深さ優先探索を用いることにより，抽出することができる（太線は，条件を満たした頻出パターンを示す）. また、多くの頻出パターンが抽出されたため，図 2 は省略した部分がある．例えば，長さ 2 の頻出パターンは，図 2 の他に「MP，MR，MS，MT」が抽出される．

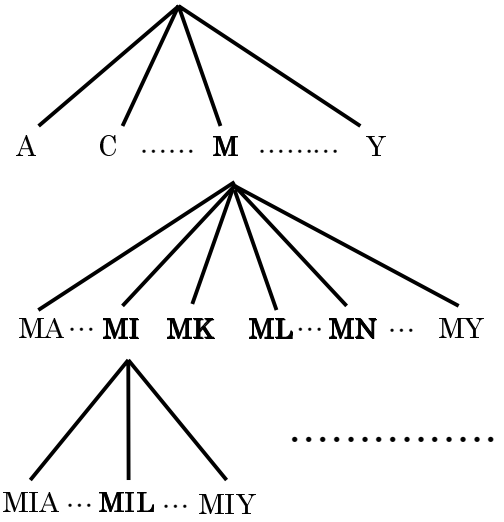

図 2 深さ優先探索

次に，実際に順をおって，より長い頻出パターンを求める方法を説明する．

まず，長さ 1 の頻出パターン M を用いて，長さ 2 の頻出パターンを求める処理を図 2 に示す．

10 **M** FKALRTIPVILNMNKDSKL
20 **M** SPNPTNIHTGKTLR

図 3 Projected Database（長さ1：M）

長さ k-1 の頻出パターンから，長さ k の頻出パターンを求める場合に，条件を満たす k 番目の文字の候補を集める．その集合を Projected Database と呼ぶ．図 3 では，矩形の部分が長さ 1 の頻出パターン M の Projected Database に相当する．

また，図 3 では，文字 M の 2 文字目の候補についてのみ，Projected Database を載せた．実際には，アミノ酸塩基に対応する 20 種類のアルファベット全てについての Projected Database を作成し，各 Projected Database の開始位置を記憶する必要がある．その全ての開始位置の記憶をまとめると表 2 のようになる．

表 2 開始位置の記憶（2文字目の候補）

	支持率	番号 10	番号 20
A	50%	5	
C	0%		
D	50%	18	
E	0%		
F	50%	3	
G	50%		12
H	50%		10
I	100%	12	9
K	100%	4	13
L	100%	6	15
M	100%	2	2
N	100%	14	8
P	100%	10	6
Q	0%		
R	100%	7	16
S	100%	19	3
T	100%	8	11
V	50%	11	
W	0%		
Y	0%		

表 2 には，各文字の支持率と各文字の Projected Database が各シークエンスに存在する先頭からの位置を格納している．図 3 で示した M の Projected Database の開始位置は，表 2 の太線の部分に相当する．

図 3 の Projected Database を走査することによって，長さ 2 の頻出パターン「MI，MK，ML，MN，MP，MR，MS，MT」を抽出する．このうち，「MI，MK」の Projected Database を求めた結果を図 4，図 5 に示す．

10 **M**FKALRT**I** PVILNMNKDSKL
20 **M**SPNPTN**I** HTGKTLR

図 4 Projected Database（長さ 2：MI）

10 **M**F**K** ALRTIPVILNMNKDSKL
20 **M**SPNPTNIHTG**K** TLR

図 5 Projected Database（長さ 2：MK）

長さ 1 の場合と同様に，20 種類全て「MA，MC，MD，ME，MF，MG，MH，MI，MK，ML，MM，MN，MP，MQ，MR，MS，MT，MV，MW，MY」の Projected Database の開始位置をまとめると表 3 のようになる．

表 3 開始位置の記憶（3 文字目：接頭辞 M）

	支持率	番号 10	番号 20
A	50%	5	
C	0%		
D	50%	18	
E	0%		
F	50%	3	
G	50%		12
H	50%		10
I	100%	12	9
K	100%	4	13
L	100%	6	15
M	50%	19	
N	100%	14	8
P	100%	10	6
Q	0%		
R	100%	7	16
S	100%	19	3
T	100%	8	11
V	50%	11	
W	0%		
Y	0%		

長さ 3 以上の頻出パターンについても，最小支持率 100%を満たしているアルファベットに対して，同様の処理を行うことによって，接頭辞 M を持った頻出パターンを全て抽出することができる．

3．2 可変ギャップ法

固定ギャップ法について，表 1 の例を用いて，説明する．固定ギャップ法も，PrefixSpan 法と同様に短い頻出パターンから，求める．

固定ギャップ法では，抽出する頻出パターンの条件として，最小支持率とギャップ数を設定する．ギャップ数とは，抽出される頻出パターンにおける文字と文字の間隔を示す． また，最小支持率は 100%，最大ギャップ数は 3 とする．

頻出パターン M から始まる頻出パターンを，図 6 に載せる．PrefixSpan 法と比較して，抽出される頻出パターンが削減していることが分かる．

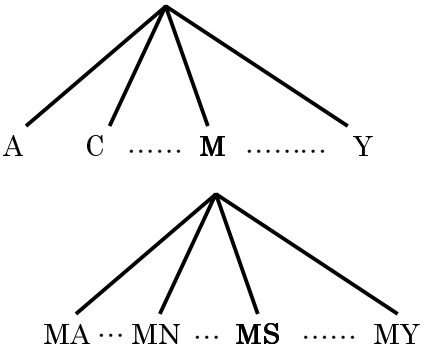

図 6 深さ優先探索（可変ギャップ法）

次に，PrefixSpan 法と同様に，長さ 2 の頻出パターンを求める処理を図 7 に示す．

PrefixSpan 法では，1 つのシークエンスから，最初に発見した頻出パターンのみ抽出する．しかし，図 7 のように，可変ギャップ法では，複数の頻出パターンを抽出する．また，最大ギャップ数を 3 と設定したため，Projected Database の範囲は 4 となる．

10 **M**FKALRTIPVILN **M**NKDSKL
20 **M**SPNPTNIHTGKTLR

図 7 Projected Database（長さ 1：M：可変）

PrefixSpan 法と同様に，2 文字目の候補となる Projected Database の開始位置を記憶し，表 4 にまとめる．

図 7 や表 4 のように，可変ギャップ法の場合，Projected Database は 1 つのシークエンスに対して，1 つも存在しない場合もあれば，3 つも存在する場合がある．従って，Projected Database の開始位置は，線形リストに記憶する．

表 4 開始位置の記憶(2文字目:可変)

	支持率	番号 10			番号 20		
A	50%	5					
C	0%						
D	50%	18					
E	0%						
F	50%	3					
G	50%				12		
H	50%				10		
I	100%	12			9		
K	100%	4	17	20	13		
L	100%	6	13	21	15		
M	100%	2	15		2		
N	100%	14	16		5	8	
P	100%	10			4	6	
Q	0%						
R	100%	7			16		
S	100%	19			3		
T	50%				7	11	14
V	50%	11					
W	0%						
Y	0%						

例えば，図 7 の番号 10 のシークエンスにおける M の Projected Database は図 8 のように記憶される．

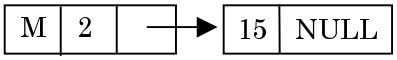

図 8 線形リスト

図 7 の Projected Database を走査することによって，長さ 2 の頻出パターン「MN：2，MS：2」を抽出する．次に，先程求めた MN，MS の Projected Database を求める（図 9，図 10）．

図 9 Projected Database(長さ 2:MN)

図 10 Projected Database(長さ 2:MS)

PrefixSpan 法と同様に，20 種類全て「MA，MC，MD，ME，MF，MG，MH，MI，MK，ML，MM，MN，MP，MQ，MR，MS，MT，MV，MW，MY」の Projected Database の開始位置をまとめると表 5 のようになる．

表 5 開始位置の記憶(3 文字目:可変)

	支持率	番号 10		番号 20	
A	50%	5			
C	0%				
D	50%	18			
E	0%				
F	50%	3			
G	0%				
H	0%				
I	0%				
K	50%	4	17		
L	50%	6			
M	0%				
N	100%	16		5	
P	50%			4	6
Q	0%				
R	0%				
S	100%	19		3	
T	0%				
V	0%				
W	0%				
Y	0%				

可変ギャップ法では，M から始まる長さ 3 以上の頻出パターンは条件を満たさないため，抽出されない．従って，PrefixSpan 法と比較して，頻出パターンを削減できているといえる．

3．3 固定ギャップ法

可変ギャップ法でも，モチーフの候補となり得ない頻出パターンが多く抽出される．

そこで，固定ギャップ法を提案する．長さ 1 の頻出パターン M を用いて，長さ 2 の頻出パターン MA，ML を抽出する例を示す（図 11）．ここでは，最小支持率を 100%とし，最大ギャップ数を 5 とする．

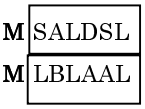

図 11 M の Projected Database

まず，可変ギャップ法の場合，MA，ML ともに 2 つのシークエンスに含まれているため，抽出可能である．また，固定ギャップ法の場合，1 つ目のシークエンスからは，M1A（M と A の間に，任意の 1 文字が含まれているパターン），M2L，M5L が抽出される．2 つ目のシークエンスからは，M3A，M4A，M0L，M2L，M5L が抽出される．この中で，2 つのシークエンスに存在する頻出パターンは，M2L と M5L となる．

4. システムの処理手順

図 11 に，本研究のシステムの処理手順をデータフローダイヤグラムで示す．図 12 では 2 箇所程，点線で囲んでいる部分がある．

１つ目は，先程 3 章で説明した固定ギャップ法を表している．２つ目は，抽出したパターンを Web 上で問合せすることができるように，インタフェース化した部分である．利用者はこのインタフェースを用いて，モチーフ発見を行う．

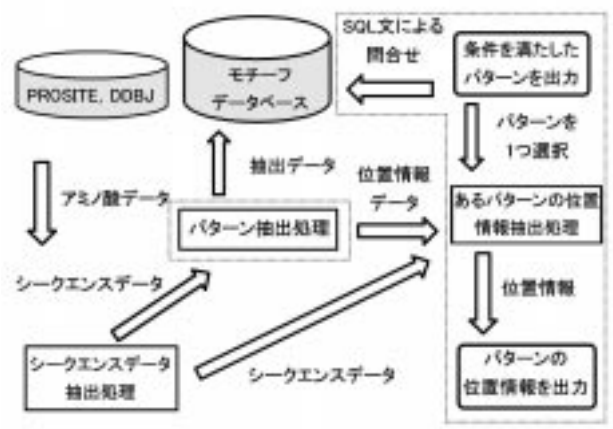

図 12 データフローダイヤグラム

5. 評価実験

評価実験において，既存の PrefixSpan 法を応用した 2 通りの手法の評価と作成したインタフェースの評価を行った．

評価実験に使用したマシンは，CPU が Pentium Ⅲ500MHz でメモリを 256M バイト搭載したものを使用した．また，固定ギャップ法の評価実験に使用したデータを表 6，7，8 にまとめた．

表 6 のモチーフ中に存在する数字は，ギャップ数を示している．また，Kringle のモチーフでは，[FY] となっている部分がある．これは，F と Y のどちらであっても同一のモチーフとみなすということを意味する．

表 6 使用データの詳細1

	データ元	モチーフ
Kringle	PROSITE	F3GC6[FY]5C
Zinc Finger	PROSITE	H3H7C2C
Leucine Zipper	DDBJ	L6L6L6L
Cytochrome C	PROSITE	C2CH

表 7 使用データの詳細2

	データ件数（件）	総長（byte）	平均長（byte）
Kringle	70	23385	334
Zinc Finger	467	245595	525
Leucine Zipper	370	139422	376
Cytochrome C	783	341927	436

表 8 使用データの詳細3

	最大長（byte）	最小長（byte）	支持率（%）	抽出時間(s)
Kringle	3176	53	94（F）	60.7
Zinc Finger	4036	34	69	96.2
Leucine Zipper	3224	3	41	79.8
Cytochrome C	4196	25	98	64.3

表 8 の支持率と抽出時間については，固定ギャップ法を用いた結果を載せている．また，Kringle の支持率 94(F) とは，Kringle のモチーフは，F3GC6F5C と F3GC6Y5C の 2 通り存在するが，F3GC6F5C が全シークエンス中, 94%存在していたことを示す．

また，この測定で使用したデータは，表 6，7，8 で使用した Zinc Finger 467 件である．

次に，固定ギャップ法の有効性を確かめるために可変ギャップ法と比較を行った．両手法の測定結果を図 13，図 14 に載せた．図 13，14 はいずれも，横軸はギャップ数を示している．縦軸に関しては，図 13 では，頻出パターンの抽出時間を示しており，図 14 では，抽出された頻出パターンの数を示している．

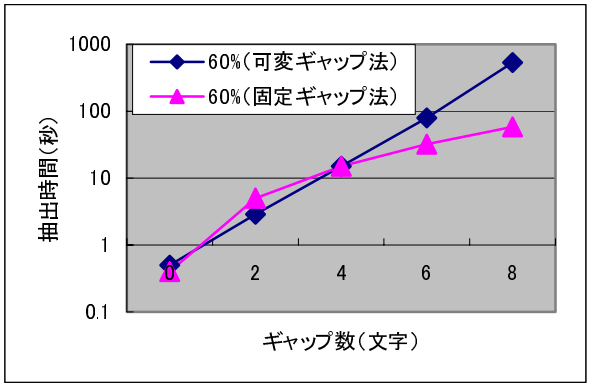

図 13 測定結果(抽出時間)

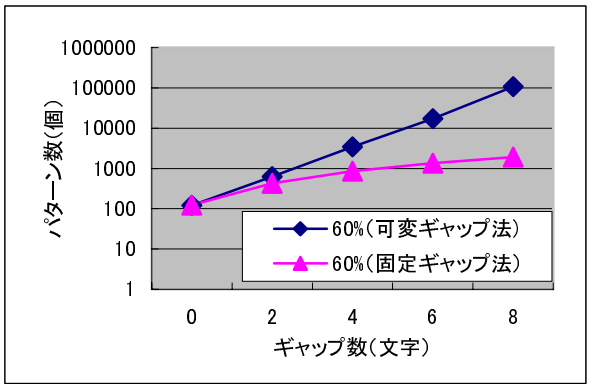

図 14 測定結果(パターン数)

図 13 より，ギャップ数が多くなった場合，固定ギャップ法が可変ギャップ法と比較して，抽出パターン数が 1/50 に減っている．そして，図 13 の抽出時間は，図 14 の抽出パターン数にほぼ比例している．

最後に，試作したインタフェースについて説明する．このインタフェースは，図 12 の右側の部分に相当する．SQL 文による問い合わせの画面が，図 15 である．ここでは，アミノ酸配列データの選択や支持率，ギャップ数に関する条件設定を行うことができる．

この画面の下にある実行ボタンをクリックすると，図 16 の例に示されるように条件を満たした抽出パターンの出力画面が表示される．この例は，Cytochrome C を含むアミノ酸配列から，頻出パターン抽出を行った結果に対するある検索結果が表示されている．幸いにも，Cytochrome C のモチーフである C2CH が上位にあることがわかった．

次に，図 16 の上部にパターンを入力することにより，そのパターンの各シークエンス上での位置情報を出力させてみよう．図 17 では，パターン C2CH の位置情報が出力されている．また，^ は C2CH が存在している位置を示している．

これにより，マルチプルアライメントとは違い，各シークエンスの任意の位置に頻出パターンが存在しても，抽出できることができていることがわかる．

図 15 入力画面

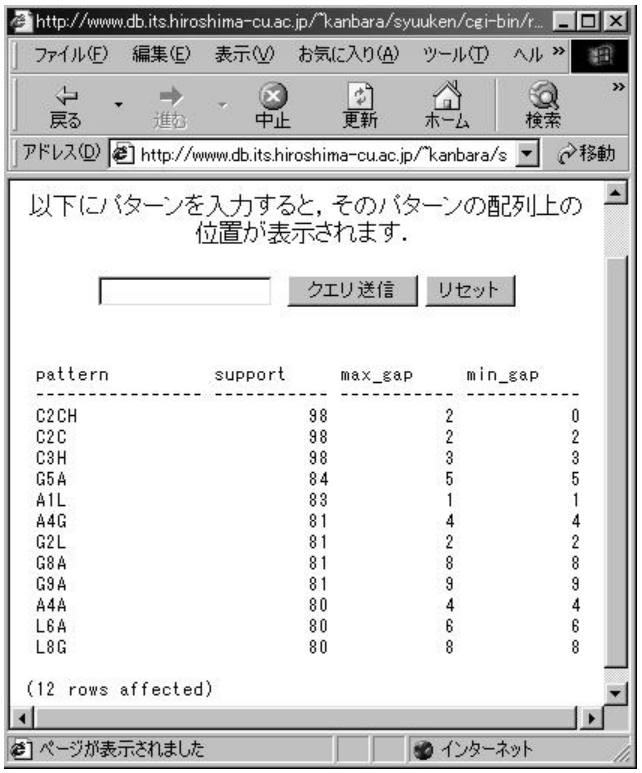

図 16 出力画面(抽出パターンの出力)

図 17 出力画面(位置情報の出力:C2CH)

6. おわりに

本研究では，従来の PrefixSpan 法にギャップ数を制限した上で，固定ギャップ法を用いることにより，頻出パターンの抽出時間や数を減少させることを可能にした．また，生物学の専門家に新しいモチーフを発見してもらえるようなインタフェースも部分作成し，その有効性を確認した．

しかし，問題点として，モチーフによって，容易に発見できる場合とできない場合がある．モチーフは大きく分けて，１通りの表現しかないパターン（例：Zinc Finger H3H7C2C）と，複数通りの表現があるパターン（例：Kringle F3GC6[FY]5C ：[FY] は F か Y のどちらかを示す）が存在するといえる．固定ギャップ法では，前者の場合しか抽出できない．

従って，今後の課題として，固定ギャップ法を拡張し，部分的にギャップ数やアルファベットが異なっていても同一のモチーフだと判断できるようにする必要がある．

謝辞

本研究において，抽出された頻出パターンに対して，有益なコメントをいただいた国立遺伝学研究所の山崎由紀子助教授，池尾一穂助手に深く感謝致します．

参考文献

[1] Dan Gusfield: Algorithms on Strings, Trees, and Sequences, Cambridge University Press, 1997.

[2] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth, Proc. of International Conference on Data Engineering (ICDE 2001), IEEE Computer Society Press, p215-p224, 2001.

[3] http://au.expasy.org/prosite/ PROSITE

[4] http://www.ddbj.nig.ac.jp/ DDBJ

[5] 金久實：ゲノム情報への招待，共立出版社，p96-98，1996.