バイオインフォマティクス基礎講座
配列解析

川端　猛

奈良先端科学技術大学院大学・情報科学研究科・准教授

2009.9.12

バイオインフォマティクス技術者認定試験について

- 試験日：平成21年11月29日（日）
- 申込期間：平成21年9月1日（火）～10月15日（木）
- 試験会場：全国6都市（札幌、仙台、東京、長浜、大阪、福岡）
- 試験方法：　分子生物学、情報科学、バイオインフォマティクスの各分野における基礎的な知識と理解度を測る。
  □試験時間：13時30分～15時30分（120分）
  □解答方法：4者択一式
  □出題数　：80問
- http://www.jsbi.org/modules/jsbi/index.php/nintei/H21/H21_info.html

出題範囲主要キーワード

生命科学分野、情報科学分野、バイオインフォマティクスの三つの分野からなる。

バイオインフォマティクス

分子生物学データベース	文献DB（PubMed）、ゲノムDB、核酸配列DB、アミノ酸DB、モチーフDB（モチーフライブラリー）、立体構造DB、代謝パスウェイDB、多型DB、発現DB、アノテーション、遺伝子オントロジー（Gene ontology）
配列解析	アライメント（動的計画法（dynamic programming）、スコアテーブル、ギャップペナルティ、ローカルアライメント、グローバルアライメント、Smith-Waterman法、ペアワイズアライメント、マルチプルアライメント、累進法（ツリーベース法）、ClustalW、HMM（隠れマルコフモデル））、相同性検索（FASTA、ハッシング、BLAST、有限オートマトン、PSI-BLAST、位置特異的スコア行列（PSSM）、プロファイル比較）、モチーフ解析（正規表現、重み行列）、分子系統解析（オーソログ、パラログ、距離行列法、UPGMA、近隣結合法（N-J法）、最節約法、最尤法、同義置換、非同義置換）、タンパク質機能予測（膜貫通部位予測、細胞内局在部位予測）、RNA二次構造予測
タンパク質立体構造解析	立体構造表現（コンタクトマップ、ラマチャンドランマップ）、構造比較（重ね合わせ、RMSD、構造アライメント、構造モチーフ、構造分類）、タンパク質二次構造予測、立体構造予測（ホモロジーモデリング、フォールド認識、スレッディング、3D-1D法）
ゲノム解析・ゲノム遺伝学	遺伝子発見（ORF（open reading frame）、スプライシング解析、プロモータ解析、偽遺伝子）、ゲノム特徴抽出（繰り返し配列発見、転写因子、SSR（simple sequence repeat）、GC含量、コドン使用頻度）、ゲノム比較（ゲノムアライメント、編集距離、系統プロファイル法、ロゼッタストーン法、遺伝子並び順の保存、遺伝子の水平伝搬、多型マーカー（SNP、マイクロサテライト、VNTR、RFLP、HLAタイプ）
トランスクリプトーム解析・プロテオーム解析	遺伝子発現クラスタリング、遺伝子ネットワーク推定（ブーリアンネットワーク、ベイジアンネットワーク）、タンパク質相互作用解析
パスウェイ解析・システム生物学	ネットワーク解析（スケールフリー、ハブ、ネットワークモチーフ）、動的シミュレーションとシステム解析（微分方程式、ロバストネス、フィードバック、フィードフォワード、感度解析、安定性解析、代謝流束解析）

http://www.jsbi.org/modules/jsbi/index.php/nintei/keyword_ver1_6.pdfから引用

「配列解析」のキーワード（１）ペアワイズアライメント

- アライメント(動的計画法　dynamic programing）
- スコアテーブル
- ギャップペナルティ
- ローカルアライメント
- Smith & Waterman法
- ペアワイズアライメント

分子生物学のセントラルドグマ

atg acg gac aaa
ttg acc tcc ctt
cgt cag tac acc
acc gta gtg gcc
gac act ggg gac

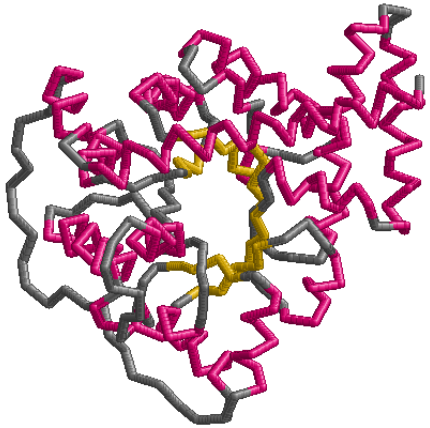
DNA配列

情報

M T D K
L T S L
R Q Y T
T V V A
D T G D

アミノ酸配列

もの



立体構造

かたち

化学反応を触媒（酵素）
酸素を運ぶ（ヘモグロビン）
異物を排除（免疫グロブリン）

分子機能

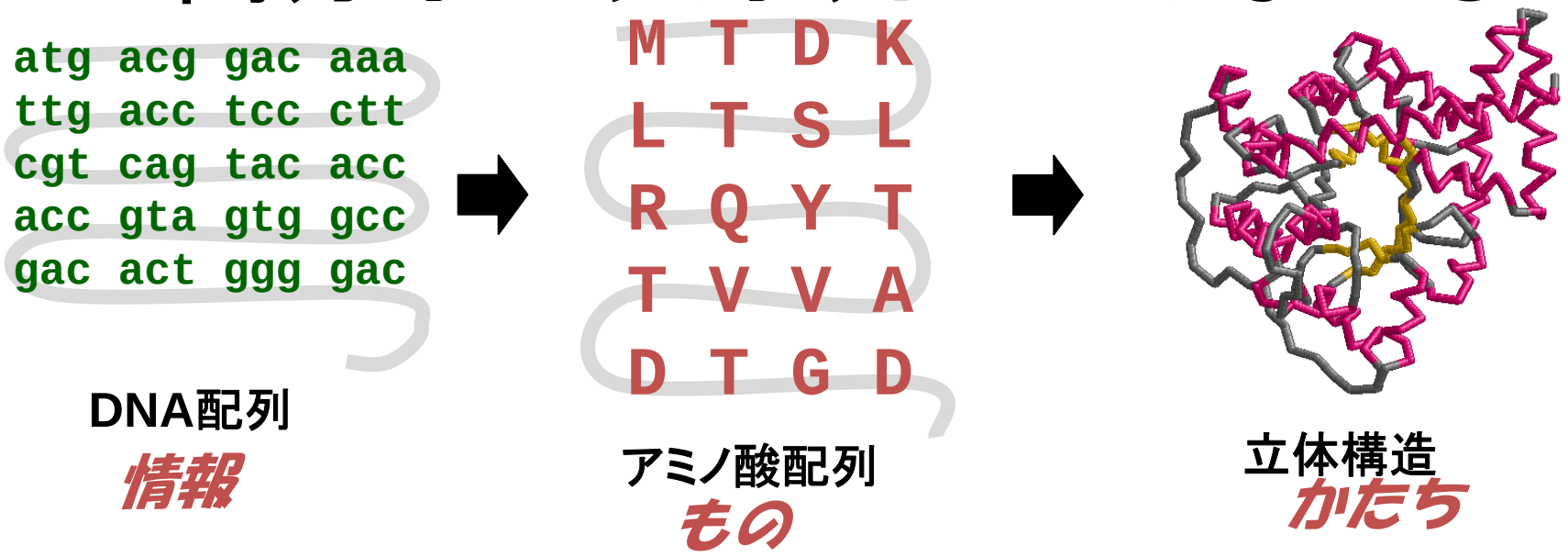
はたらき

細胞

個体

進化！

高分子は文字列だとみなせる



DNAもタンパク質もユニットが一列に並んだ高分子

ユニット: DNAは4種の核酸(atgc)、タンパク質は20種のアミノ酸(ACDEFGH…)

atgacggacaaattgacctcccttcgtcagtacaccaccgtagtggccga

M T D K L T S L R Q Y T T V V A D T G D

→単なる文字列だとみなして処理をしてもある種の本質は失われない

「進化」とはDNAという文字列が変化すること

atgacggacaaattgacctcccttcgtcagtacacc

M T D K L T S L R Q Y T

atgacgaacaaattgacctcccttcgtcagtacacc

M T N K L T S L R Q Y T

より正確には、個体のDNAが変化したあとに、その変異がその種の集団において定着する「集団遺伝学」的な過程が必要

①個体のDNAに変異が生じる

②その変異が子孫に継承され、

③中立か正の淘汰が働けば、同じ変異を持った子孫が種の集団内で多数を占める

違う生物の同じ機能のタンパク質のアミノ酸配列

トリオースリン酸異性化酵素（ Triosephosphate isomerase (EC 5.3.1.1) (TIM,TPIS)）

>TPIS_HUMAN ヒト "Triosephosphate isomerase (EC 5.3
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_RABIT ウサギ "Triosephosphate isomerase (EC 5
APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALSEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

違う生物の同じ機能のタンパク質のアミノ酸配列

トリオースリン酸異性化酵素（ Triosephosphate isomerase (EC 5.3.1.1) (TIM,TPIS)）

>TPIS_HUMAN ヒト "Triosephosphate isomerase (EC 5.3

APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_YEAST 酵母 "Triosephosphate isomerase (EC 5.

ARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATY
LDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWV
ILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKKA
GKTLDVVERQLNAVLEEVKDWTNVVVAYEPVWAIGTGLAATPEDA
QDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADV
DGFLVGGASLKPEFVDIINSRN

違う生物の同じ機能のタンパク質のアミノ酸配列

トリオースリン酸異性化酵素（ Triosephosphate isomerase (EC 5.3.1.1) (TIM,TPIS)）

>TPIS_HUMAN ヒト "Triosephosphate isomerase (EC 5.3
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_ECOLI 大腸菌 "Triosephosphate isomerase (EC 5
MRHPLVMGNWKLNGSRHMVHELVSNLRKELAGVAGCAVAIAPPEM
YIDMAKREAEGSHIMLGAQNVDLNLSGAFTGETSAAMLKDIGAQY
IIIGHSERRTYHKESDELIAKKFAVLKEQGLTPVLCIGETEAENE
AGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP
AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQP
DIDGALVGGASLKADAFAVIVKAAEAAKQA

進化的なイベント： 置換 と 削除･挿入

トリオースリン酸異性化酵素（ Triosephosphate isomerase (EC 5.3.1.1) (TIM,TPIS)）の場合

ヒト(TPIS_HUMAN)とウサギ(TPIS_RABIT)の比較

```
HUMAN  1:APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA:60
         ******************  *****  ********************************
RABIT  1:APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA:60
```

TPIS_HUMAN 248 vs TPIS_RABIT 248 SeqID 98.4 %

置換(substitution) ： アミノ酸･核酸の変化

ヒト(TPIS_HUMAN)と大腸菌(TPIS_ECOLI)の比較

```
HUMAN  4:RKFFVGGNWKMNGRKQSLGELIGTLNAAKVP-ADTEVVCAPPTAYIDFARQKLD-PKIAV:61
         *   * **** **      **   *     *   *  ***  *** *     * *
ECOLI  2:RHPLVMGNWKLNGSRHMVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIML:61
```

TPIS_HUMAN 248 vs TPIS_ECOLI 255 SeqID 45.9 %

挿入･欠失(insertion, deletion ; indel)

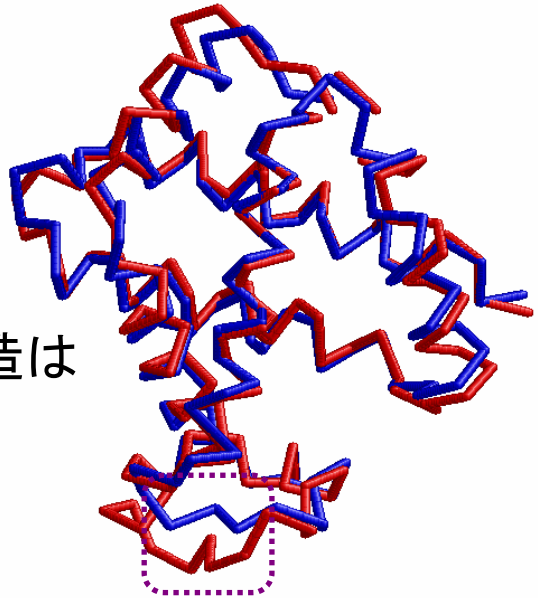
配列の類似と立体構造の類似

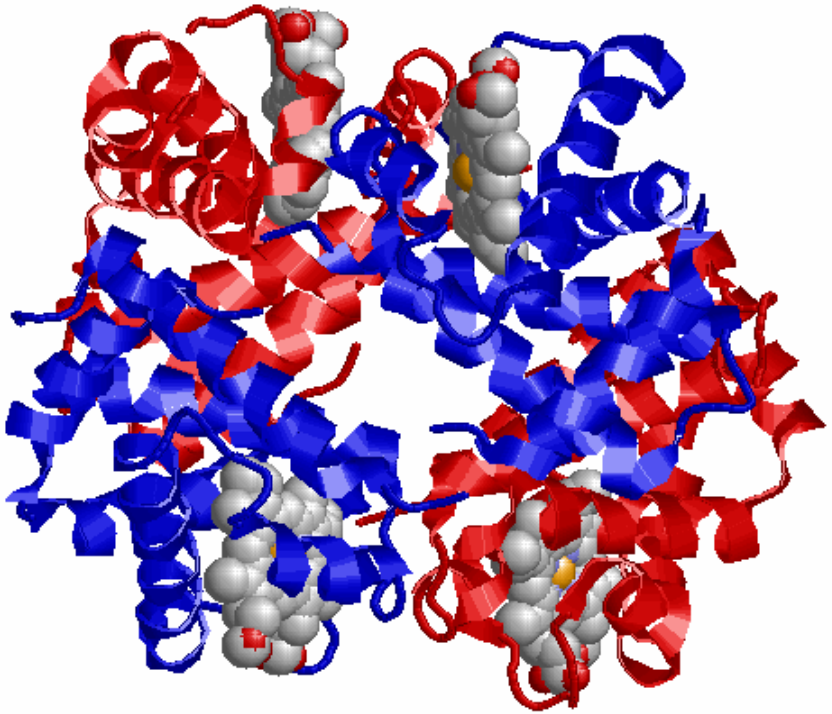
ヒトのヘモグロビンの α 鎖と β 鎖　(SeqID 46.0%)

```
Alpha  2:LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQV:55
           * *   *   * * ****      * * *** *      * *    *   * ***       *    *
Beta   3:LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV:60

Alpha 56:KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA:11
           * *****   *      ** *          ** **   ** *** ** **    *    ** *
Beta  61:KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK:120

Alpha 116:EFTPAVHASLDKFLASVSTVLTSKY:140
            **** * *    *   * *    *   **
Beta  121:EFTPPVQAAYQKVVAGVANALAHKY:145
```





機能や立体構造は
よく似ている

配列の類似を知ることは立体構造予測につながる

2つの配列を比較するには？

1. 類似性のスコア関数の定義

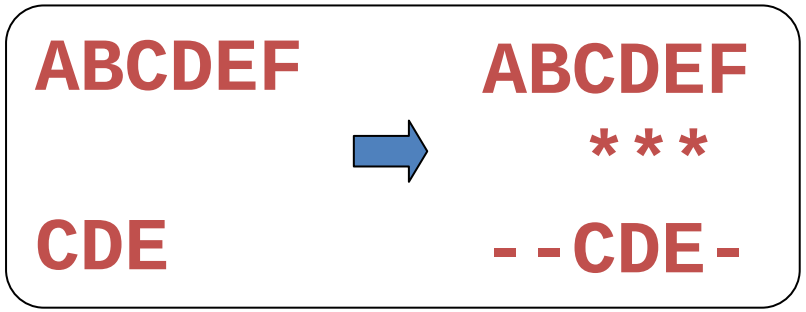
文字の間の類似性をどうやって定量するか？

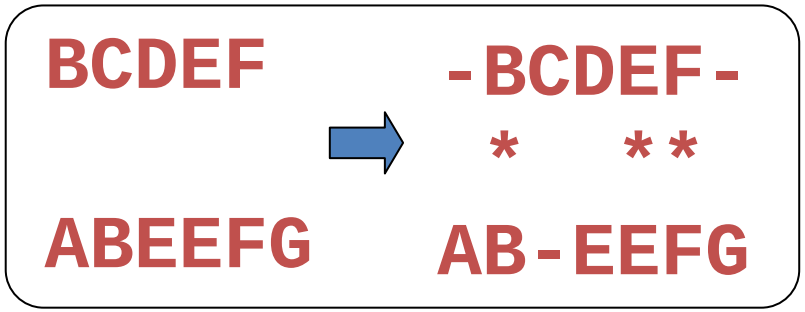
```
ACFDE
**  *
ACEEE
```

3つ同じだから3点？
FとEの対応とDとEの対応は等価だろうか？

2. アライメント

どうやって文字と文字を対応づけるか？





もっと長いときはどうやって計算する？

置換スコア関数（行列）の定義

(1)一致・不一致スコア

$$S(A,B)=\begin{cases}\alpha & A=B\\ \beta & A\neq B\end{cases}$$

もっとも簡単。DNAの場合によく使われる。
BLASTの核酸のデフォルトは、$\alpha=1, \beta=-3$

	A	T	G	C
A	1	−3	−3	−3
T	−3	1	−3	−3
G	−3	−3	1	−3
C	−3	−3	−3	1

＃問題点：文字列間の類似性を捉えられない。

L(ロイシン,疎水性) → V(バリン、疎水性) ：起こりやすい

L(ロイシン,疎水性) → E(グルタミン酸、－荷電) ：起こりにくい

(2)対数オッズスコア(log odds score)

$$S(A,B) = \log \frac{q(A,B)}{p(A)\,p(B)}$$

2つの異なるタンパク質のあるサイトのアミノ酸がA,Bであったとき、

Protein1 : XXXXAXXXX
Protein2 : XXXXBXXXX

$q(A,B)$: 進化的な関係からAとBの対応が生じた確率

$p(A)$: 偶然にAが生じた確率。

$p(A)\cdot p(B)$: 偶然にAとBの対応が生じた確率。

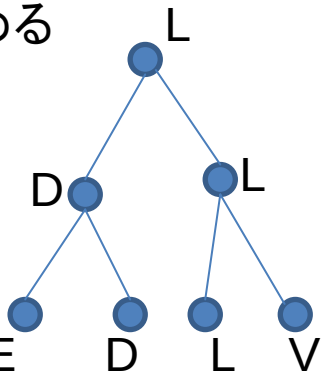
BLOSUM62（blastpのデフォルトで使われている置換スコア行列）

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

(2-1)PAMスコア行列　(Dayhoff et al.,1978)

(1)極めて近縁のよく似た蛋白質を集め、系統樹を作成。祖先配列も求める
(2)系統樹の枝間で起こった置換の回数を数え、変異確率M_{AB}を求める

$$\Pr(A \to B) = M_{AB} = \frac{q(A,B)}{p(A)}$$



ここで、M_{AB}を100個に1個のアミノ酸が置換起こるように調整する。この進化距離のことを1PAM (Accepted Point Mutation)と呼ぶ。

(3)より遠い進化は、行列MをN回累乗することで得る（マルコフ連鎖による進化モデル）

$$\Pr(A \to B; N) = (M^N)_{AB}$$

最終的なスコアは以下のような形式となる。

$$S(A,B) = \log \frac{q(A,B)}{p(A)\,p(B)} = \log \frac{(M^N)_{AB}}{p(B)}$$

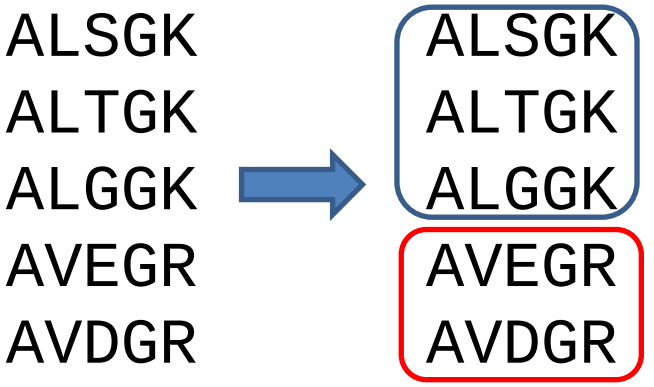
PAMスコア行列の名称、PAM30, PAM70, PAM250などの数字はこの乗算した回数Nを指す。**この数が大きいほど、遠縁の進化を表している。**

(2-2)BLOSUMスコア行列　(Henikoff & Henikoff.,1992)

(1)　マルチプルアライメントされた配列群を用意

短い長さのマルチプルアライメントのデータベース
BLOCKS (http://blocks.hfcrc.org/blocks/)を使用

SeqID=60
でクラスタリング



(2)配列一致率(Sequence Identity)がある値以上の配列をクラスタリングし、サブファミリーを作成する

(3)サブファミリー間の置換を数えて、確率$q(A,B)$を推定する

$$p(A) = q(A,A) + \sum_{B \neq A} q(A,B)/2$$

$$S(A,B) = \log \frac{q(A,B)}{p(A)\,p(B)}$$

BLOSUMスコア行列の名称、BLOSUM45, BLOSUM62, BLOSUM80などの数字はこのサブファミリーにクラスタリングするときのsequence identityを示している。
この数が大きいほど、近縁の進化を表している。

H19　問55

配列データ解析の一つである置換スコア行列に関する次の説明文の中で不適切なものはどれか、一つ選べ。

1. 通常の置換スコア行列では、進化的に置換の起こり難い組み合わせに正の数が付けられている。
2. PAMスコア行列は、タンパク質の変異による進化モデルに基づいている。
3. 進化的に遠縁の配列を比較する場合は、PAM60より、PAM120を用いたほうがよい。
4. BLOSUMスコア行列は、BLOCKSデータベースを元に作成されている。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19　問55

配列データ解析の一つである置換スコア行列に関する次の説明文の中で不適切なものはどれか、一つ選べ。

1. 通常の置換スコア行列では、進化的に置換の起こり難い組み合わせに正の数が付けられている。
2. PAMスコア行列は、タンパク質の変異による進化モデルに基づいている。
3. 進化的に遠縁の配列を比較する場合は、PAM60より、PAM120を用いたほうがよい。
4. BLOSUMスコア行列は、BLOCKSデータベースを元に作成されている。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

スコアの計算例

AFDC	S(A,A) +	S(F,E)	S(D,E) +	S(C,C) = 12	
AEEC	4	-3	2	9	

ギャップがある場合はギャップのスコア（ギャップペナルティ）を設定する

AFDGC	S(A,A) +	S(F,E) +	S(D,E) +	gap +	S(C,C) = 10	
AEE-C	4	-3	2	-2	9	

H20　問48

下記の二本のアミノ酸配列のアライメントについて、BLOSUM62スコア行列（下記）を用いてスコアを計算したい。スコアとして適切な値を、選択肢の中から一つ選べ。

```
DDDGW
|   ||
DEEGW
```

1. 35
2. 27
3. 23
4. 22

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0
R	−1	5	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3
N	−2	0	6	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3
D	−2	−2	1	6	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3
C	0	−3	−3	−3	9	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1
Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2
E	−1	0	0	2	−4	2	5	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2
G	0	−2	0	−1	−3	−2	−2	6	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3
H	−2	0	1	−1	−3	0	0	−2	8	−3	−3	−1	−2	−1	−2	−1	−2	−2	2	−3
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4	2	−3	1	0	−3	−2	−1	−3	−1	3
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4	−2	2	0	−3	−2	−1	−2	−1	1
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5	−1	−3	−1	0	−1	−3	−2	−2
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5	0	−2	−1	−1	−1	−1	1
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6	−4	−2	−2	1	3	−1
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7	−1	−1	−4	−3	−2
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4	1	−3	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5	−2	−2	0
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	2	−3
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	−1
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20　問48

下記の二本のアミノ酸配列のアライメントについて、BLOSUM62スコア行列（下記）を用いてスコアを計算したい。スコアとして適切な値を、選択肢の中から一つ選べ。

DDDGW
|　||
DEEGW

6+2+2+6+11=27

1. 35
2. 27
3. 23
4. 22

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

アライメント

スコア関数（ギャップを含む）を最大にするような文字の対応つけを探す

1. ギャップなしアライメント
2. ギャップありアライメント

ギャップなし

```
AFDC
AEEC
```

ギャップあり

```
AFAED-C
A--EEGC
```

a. グローバルアライメント (ClustalW)

b. ローカルアライメント (FASTA, BLAST)

```
ACDEFGHKLM
AFGHKKL
```

➡

```
ACDEFGHK-LM
A---FGHKKL-
```

グローバル

```
FGHK-L
FGHKKL
```

ローカル

動的計画法というアルゴリズムで解く。
そのイメージをつかむためには**ドットマトリックス法**が有効

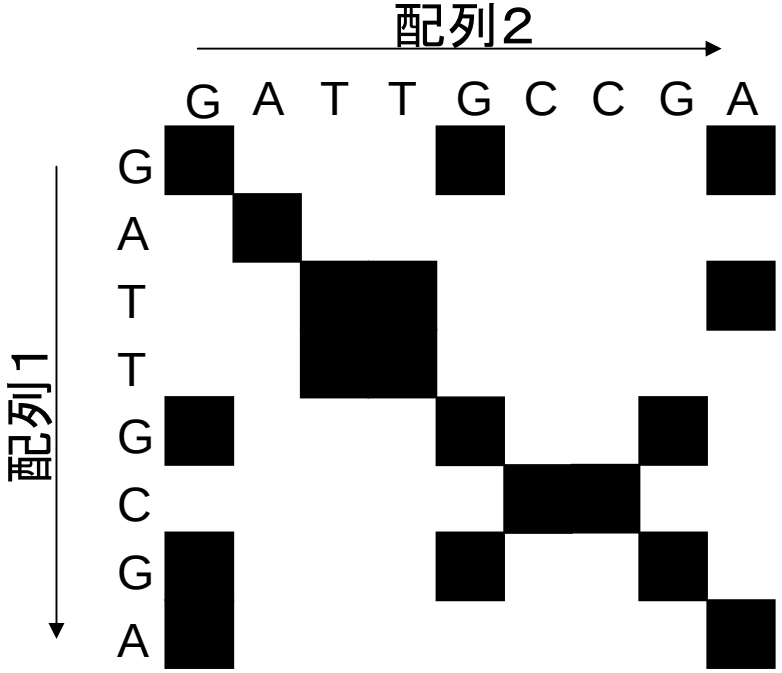
ドットマトリックス法

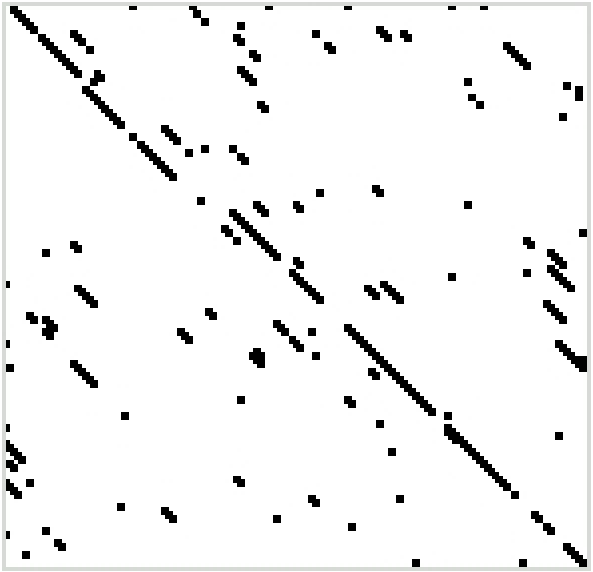
- 比較する配列を二次元の格子の縦横に並べ、一致している文字のペアを黒く塗った、グラフィカルな表示法
- 対応する部分は、連続する対角線として表示される

※考案者Robert Harrにちなみ**ハー・プロット**とも呼ばれる。

※ゲノムレベルの非常に長い配列の比較にも対応

※部分一致、繰り返しなど特殊なケースにも対応できる。





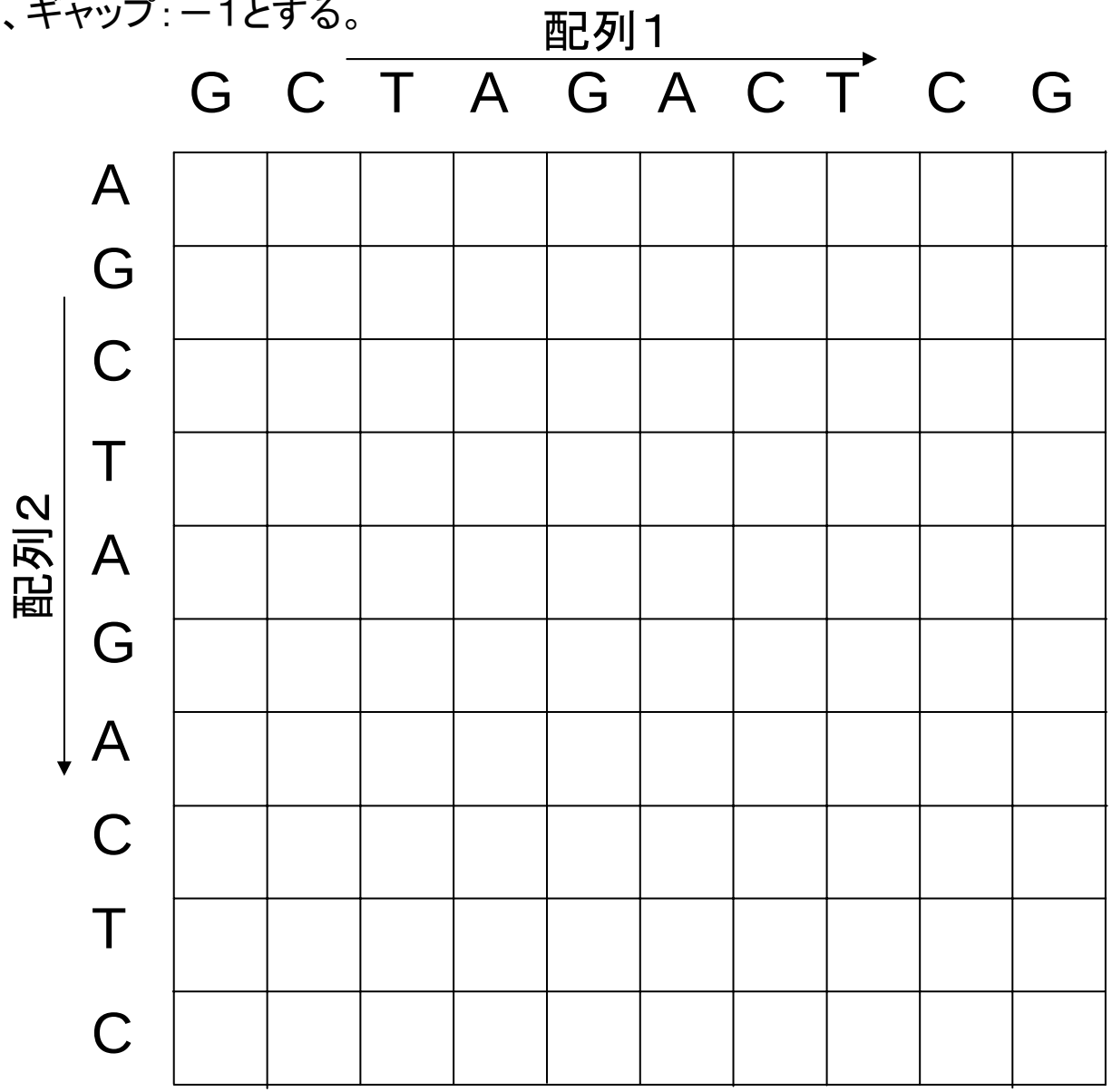
ドットマトリックス　：　例1　(1)

※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

1：**GCTAGACTCG**

2：**AGCTAGACTC**

(1)配列1、配列2を
横と縦に並べる



ドットマトリックス　：　例1　(2)

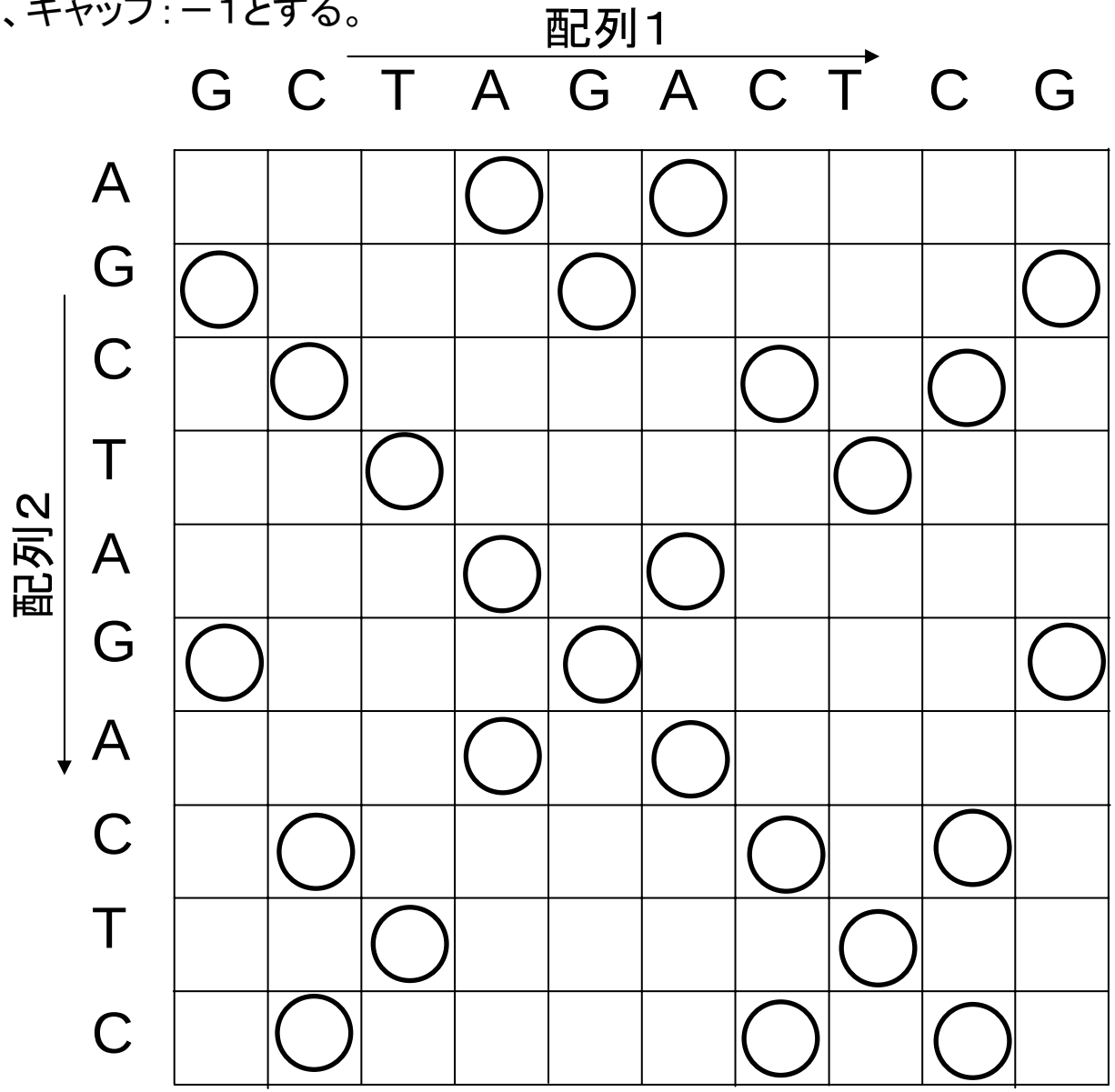
※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

1：GCTAGACTCG

2：AGCTAGACTC

(1)配列1、配列2を横と縦に並べる

(2)文字が一致するマスに〇を描く



ドットマトリックス　：　例1　(3)

※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

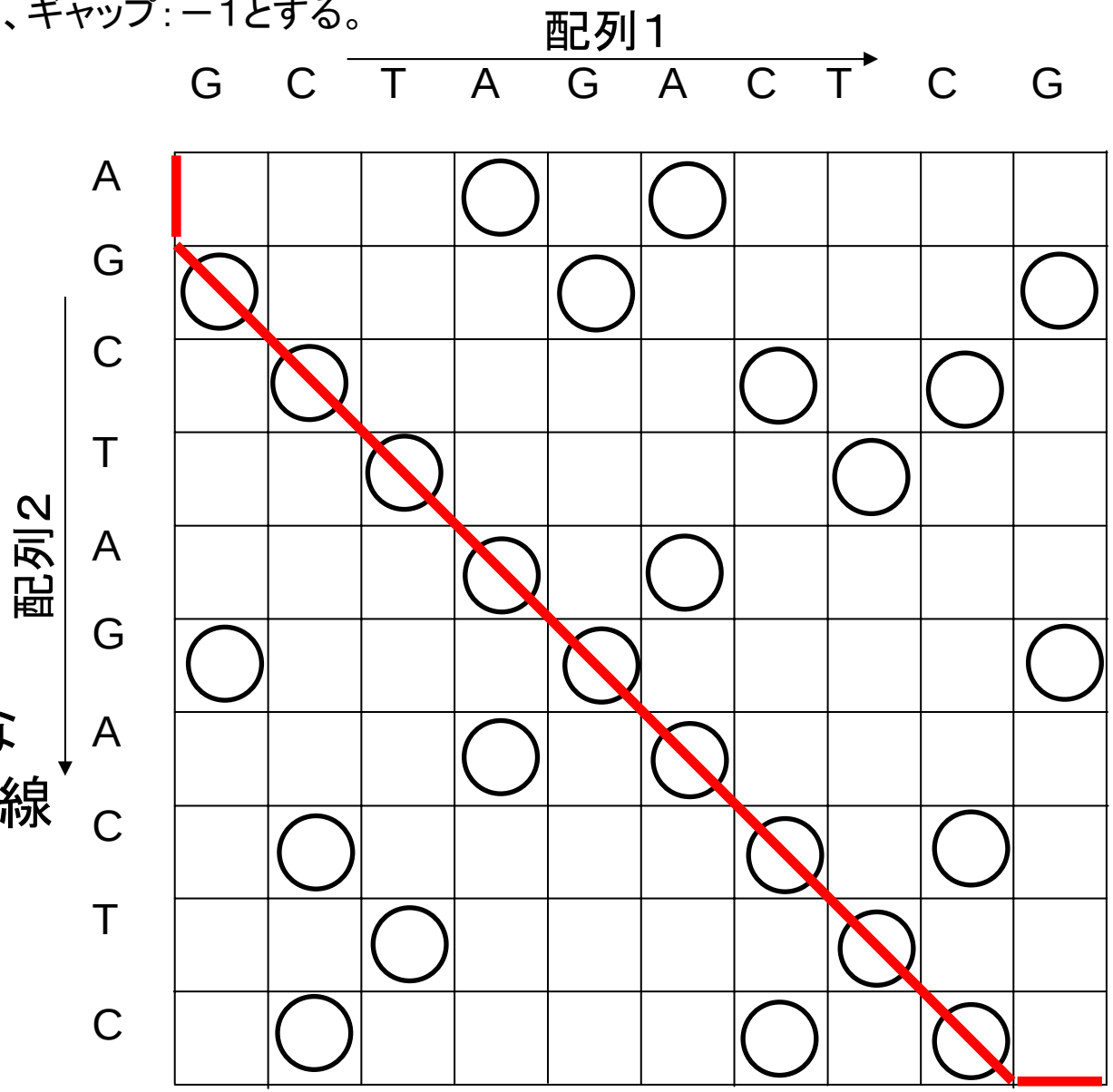
1：**GCTAGACTCG**

2：**AGCTAGACTC**

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線



ドットマトリックス　：　例1　(4)

※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

1:GCTAGACTCG
2:AGCTAGACTC

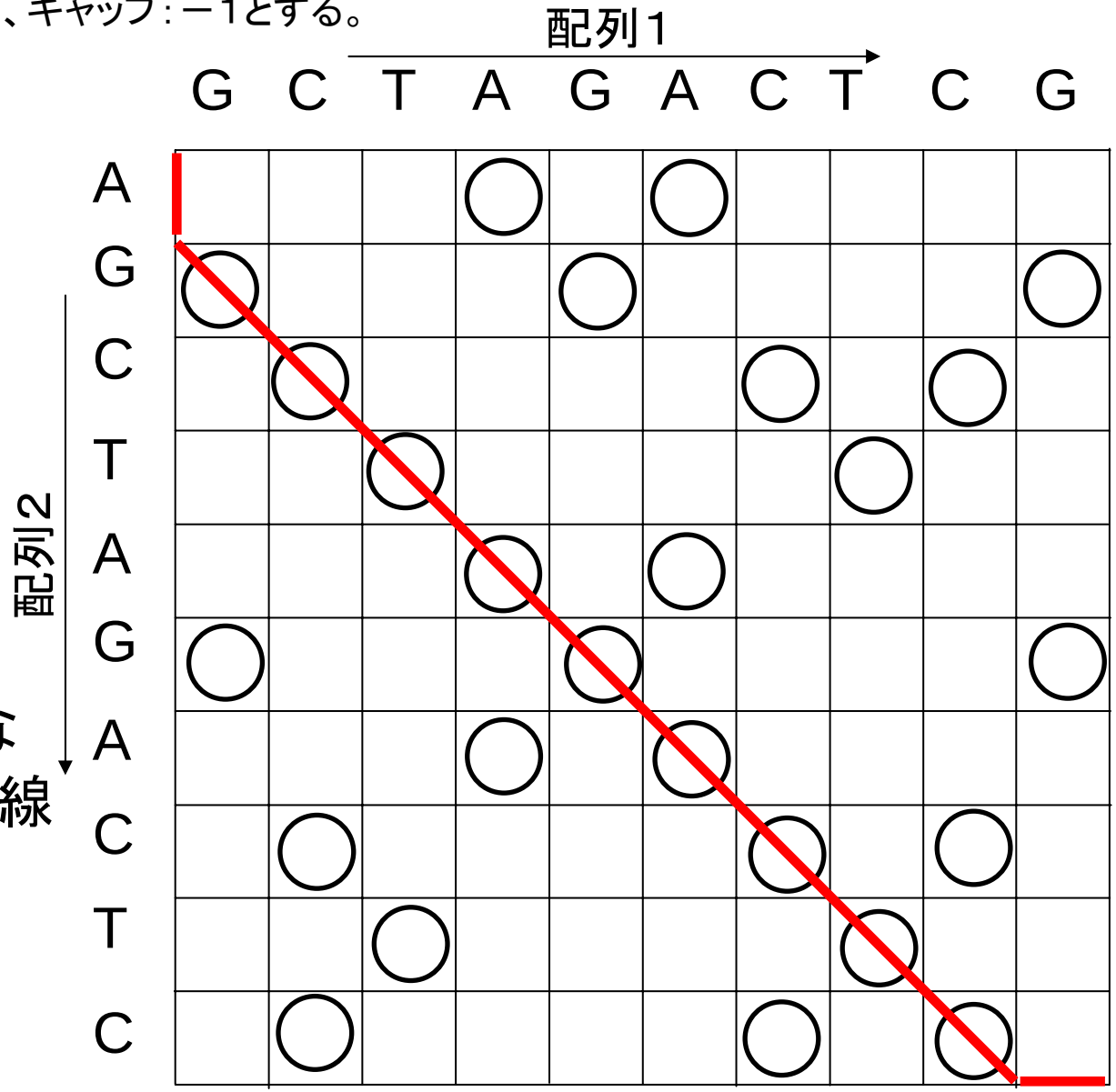
(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線

(4)アライメント

1:-GCTAGACTCG
****
2:AGCTAGACTC-

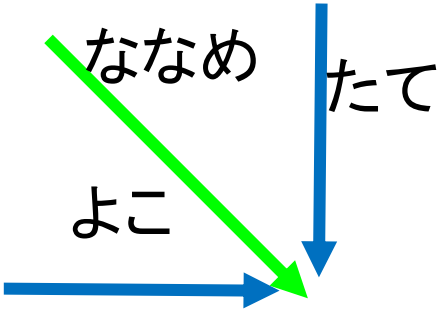


スコア：一致（＋1）×9＋不一致（0）×0＋ギャップ（－1）×2＝7

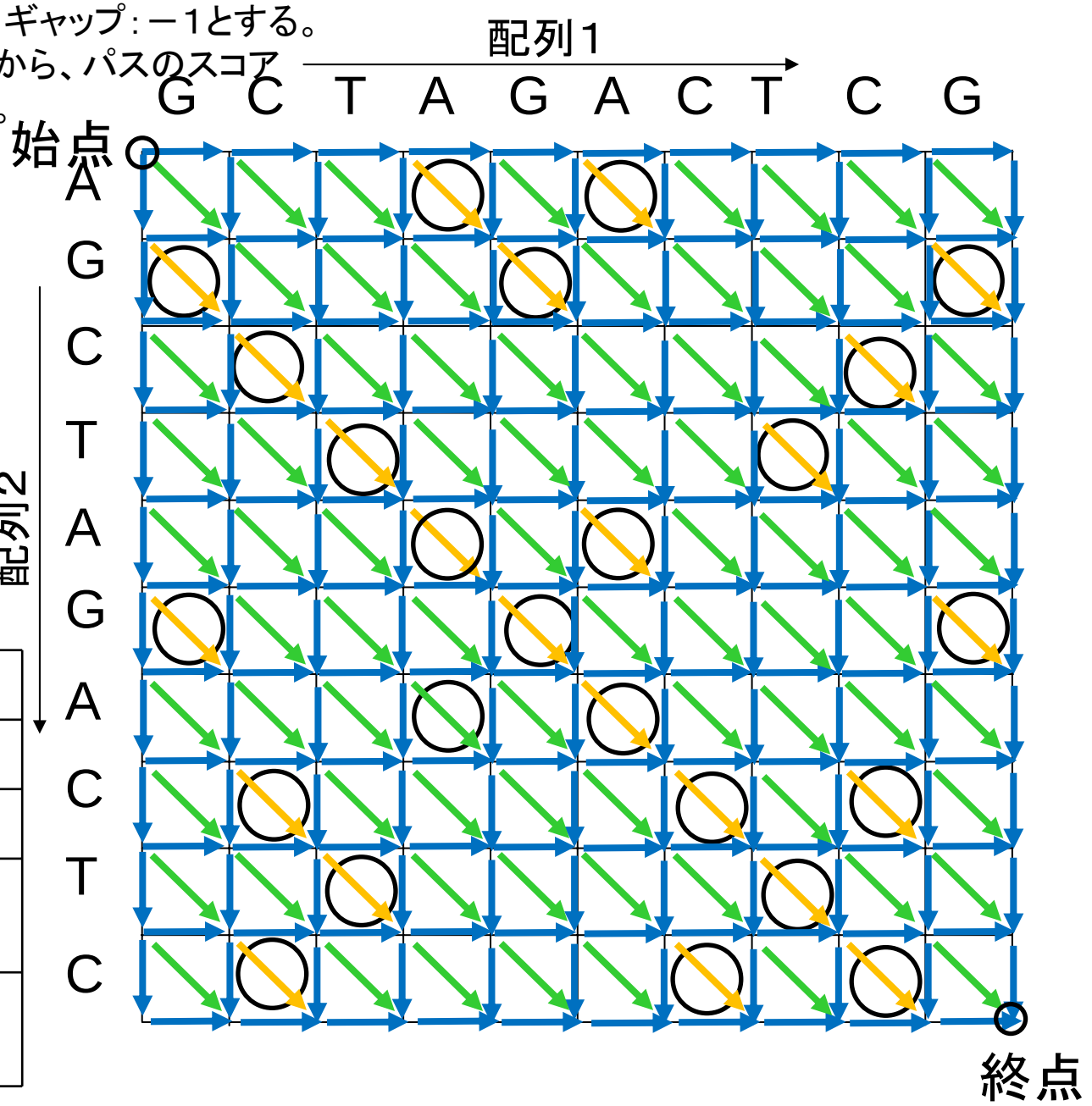
ドットマトリックスのパスの引き方の詳細

※スコア：一致：＋１、不一致：０、ギャップ：－１とする。
始点から終点を結ぶパスのなかから、パスのスコアの合計が最大になるパスを選ぶ。

進む方向は3通り



	点数	アライメント
たて	－1	配列１が“-”
よこ	－1	配列２が“-”
ななめ	0	文字が一致しない対応
○にななめ	＋1	文字が一致する対応



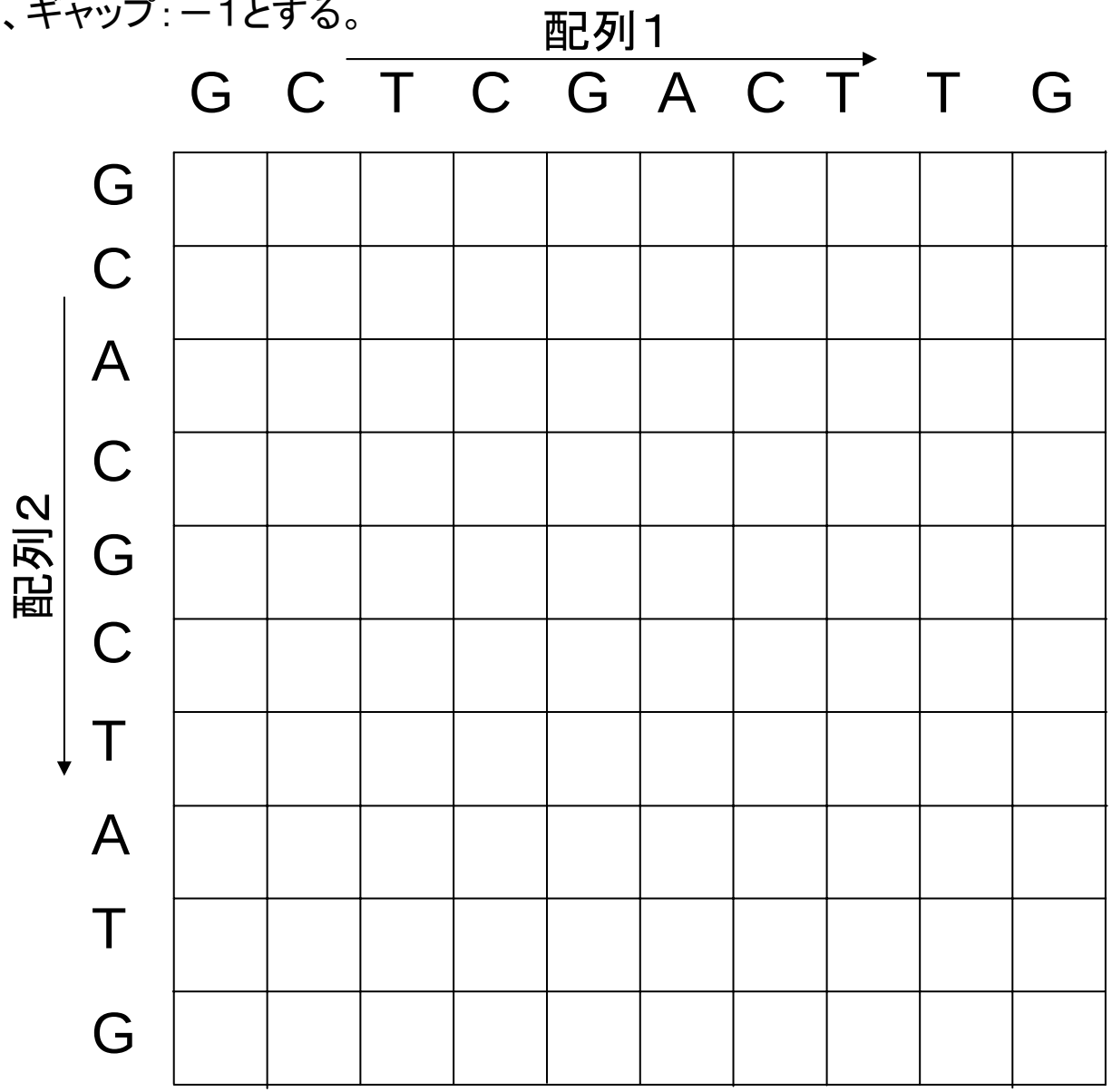
ドットマトリックス　：　例2　(1)

※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

配列1：**GCTCGACTTG**

配列2：**GCACGCTATG**

(1)配列1、配列2を
横と縦に並べる



ドットマトリックス　：　例2　(2)

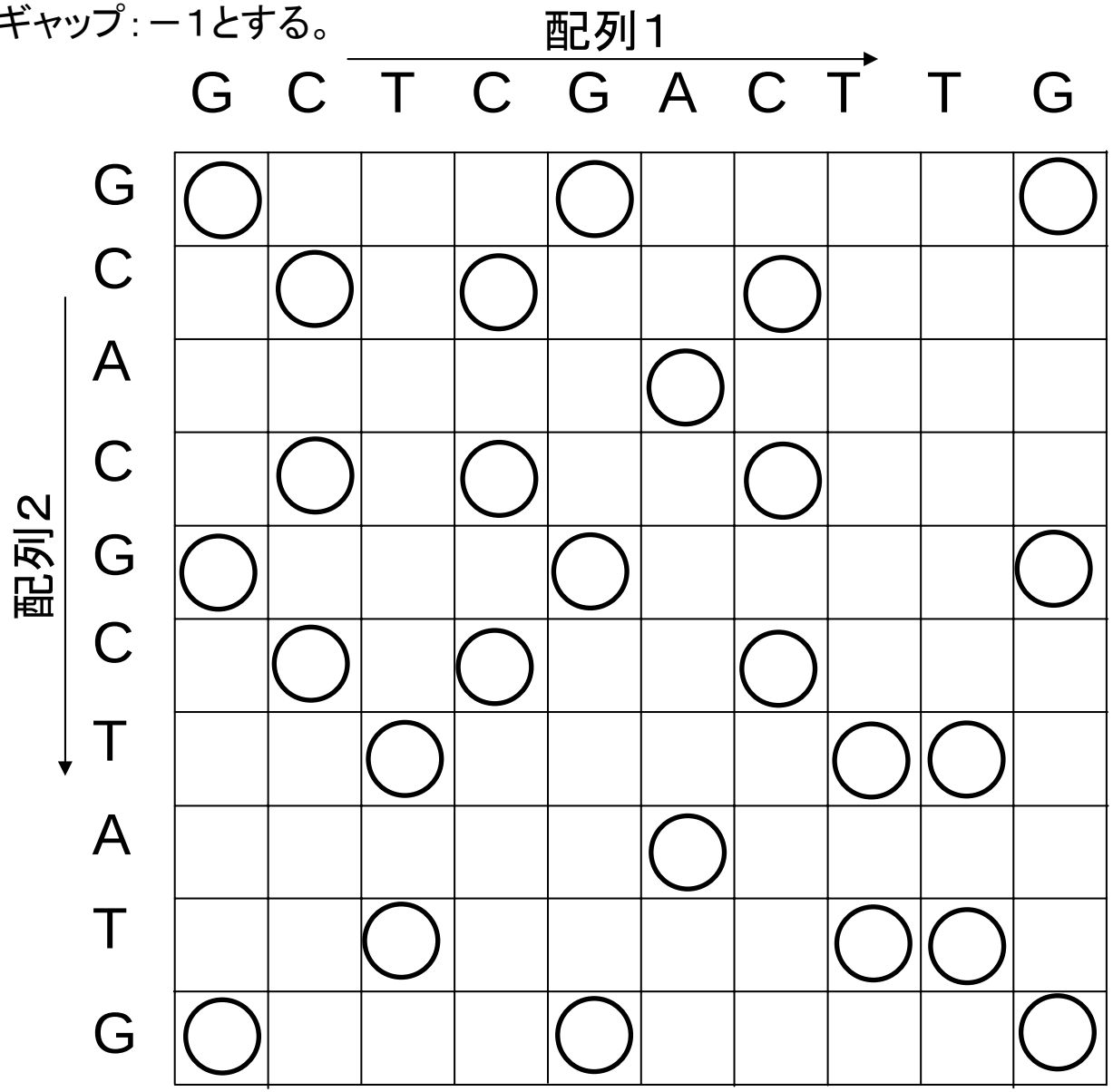
※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

配列１：**GCTCGACTTG**

配列２：**GCACGCTATG**

(1)配列１、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く



ドットマトリックス　：　例2　(3)

※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

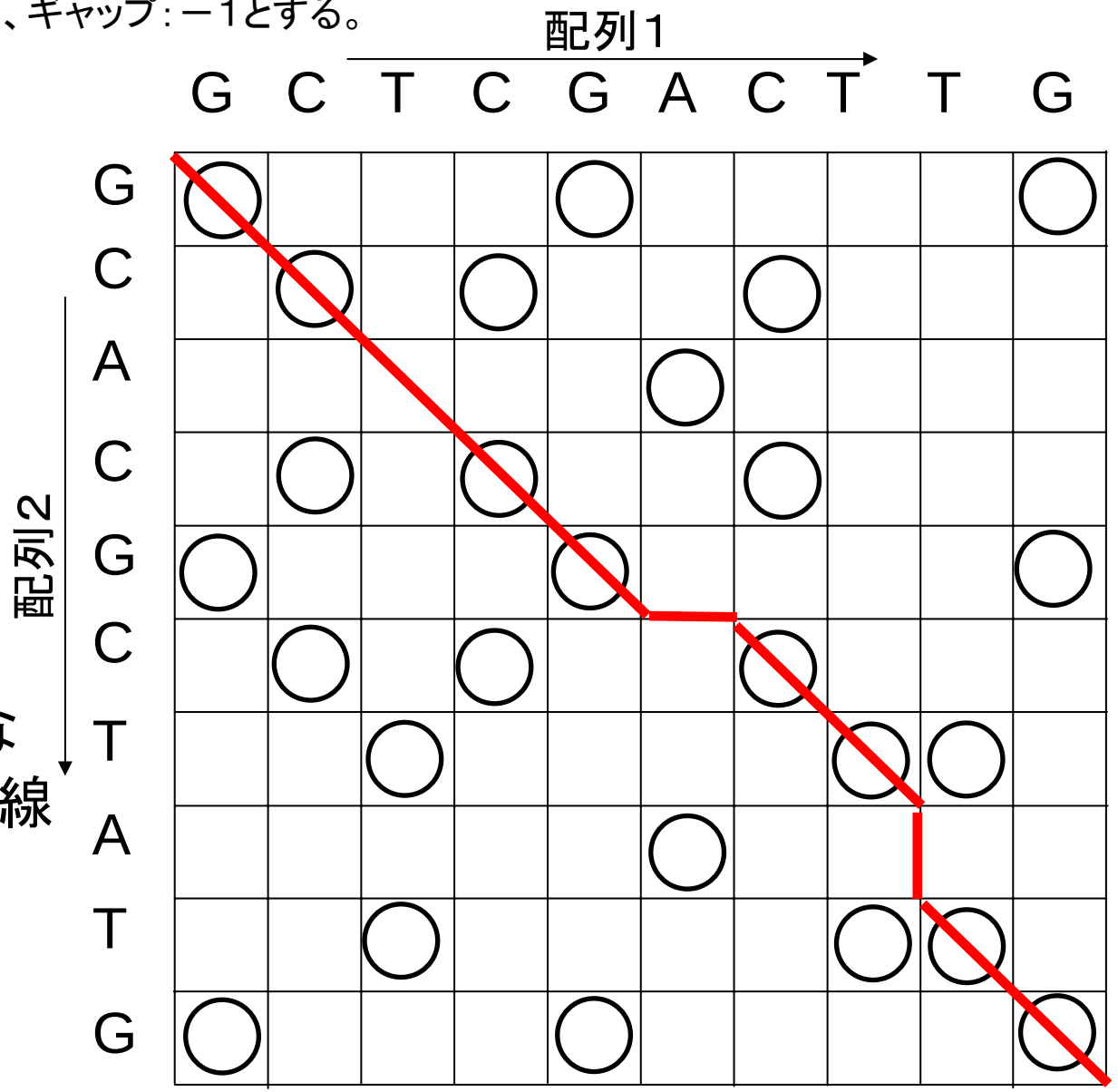
配列1：**GCTCGACTTG**

配列2：**GCACGCTATG**

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線



ドットマトリックス　：　例2　(4)

※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

配列1:**GCTCGACTTG**

配列2:**GCACGCTATG**

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

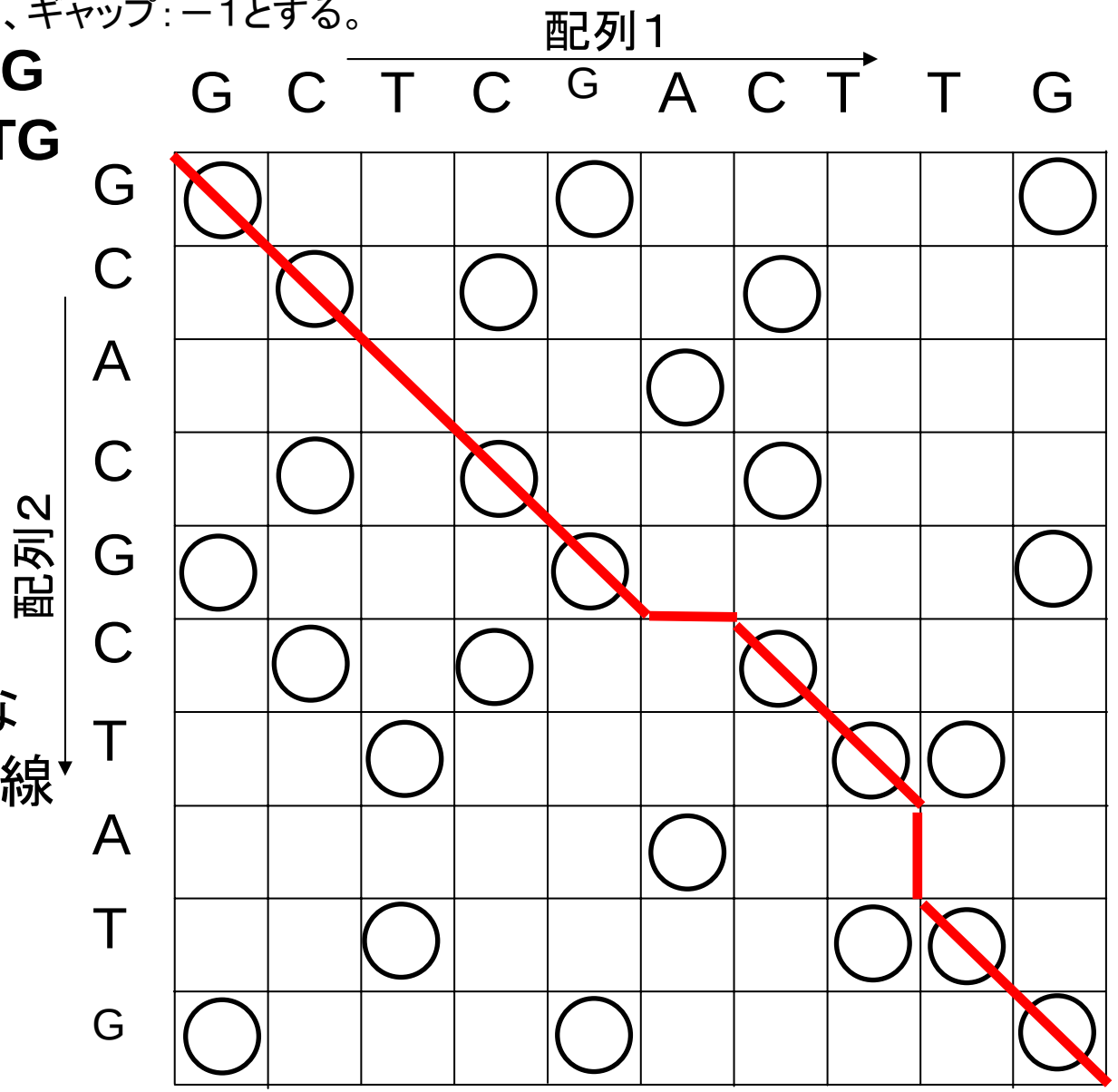
(3)多くの〇を通るような
左上と右下を結ぶ折れ線

(4)アライメント

1:GCTCGACT-TG

****　**　**　****

2:GCACG-CTATG



スコア：一致（＋1）×8＋不一致（0）×1＋ギャップ（－1）×2＝6

H20 問50

以下の2本の塩基配列において両配列間で対応する塩基数が最大となるように、ギャップの挿入を許すアライメントを行う。塩基が対応するとは、A-A,T-T,G-G,C-Cというように塩基が完全に一致することである。簡単のために、ギャップペナルティ、塩基配列の不一致については考慮しない。アライメントした両配列の塩基が一致する最大数でもっとも適切なものを選択肢の中から一つ選べ。

ATGCATGC
AATCAACG

1. 3, **2.** 4, **3.** 5, **4.** 6

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20 問50

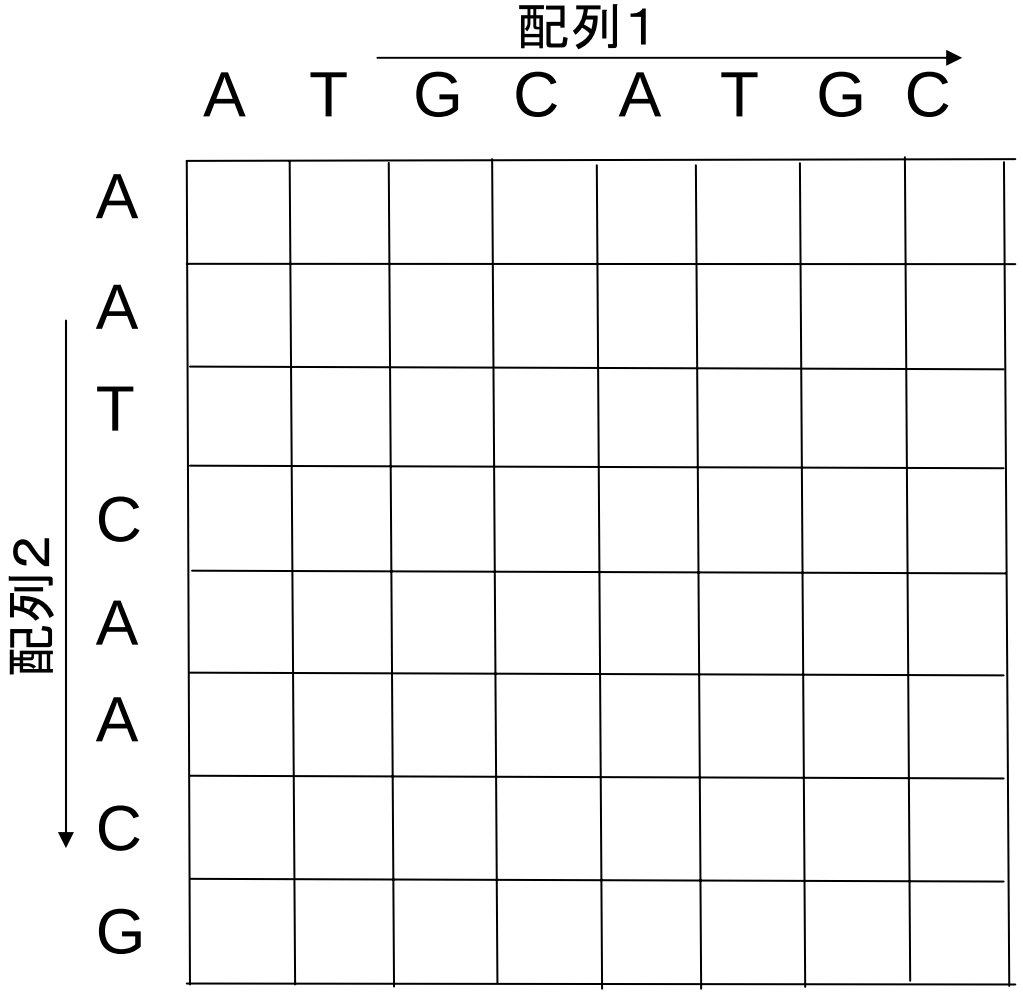
※スコア：一致：＋1、不一致：0、ギャップ：0とする。

ATGCATGC
AATCAACG

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線



H20 問50

※スコア：一致：＋1、不一致：0、ギャップ：0とする。

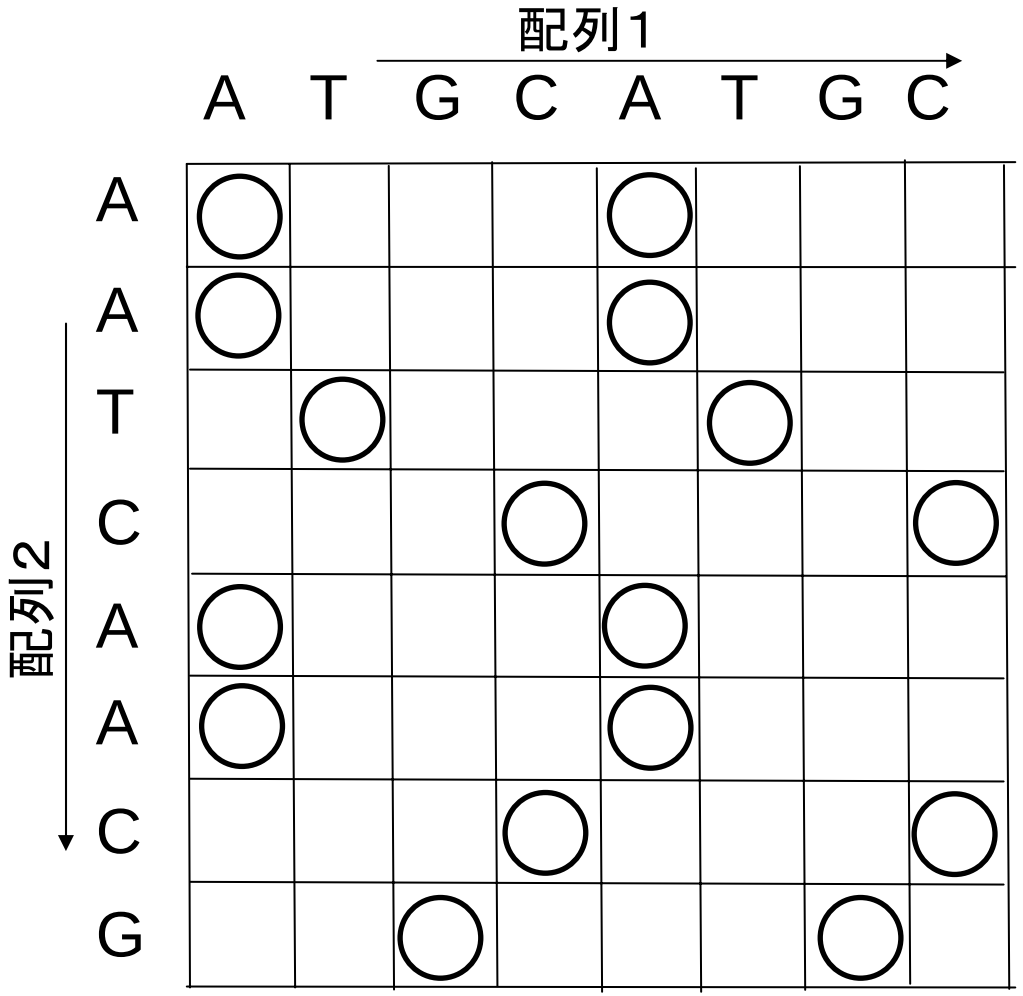
ATGCATGC

AATCAACG

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線



H20 問50

※スコア：一致：＋1、不一致：0、ギャップ：0とする。

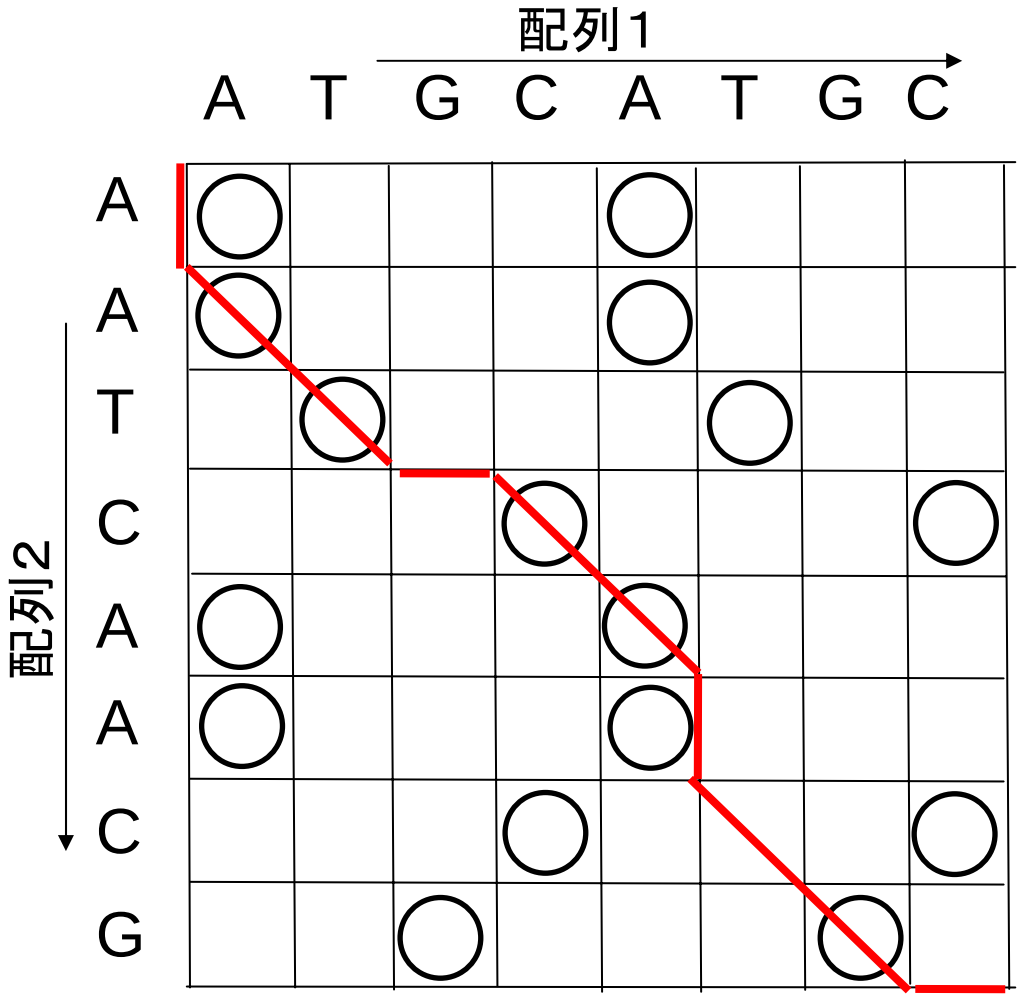
ATGCATGC

AATCAACG

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線



-ATGCA-TGC
**** ** ***
AAT-CAACG-

この場合、解は何通りもあるが、いずれも一致する残基数は5

H20 問50

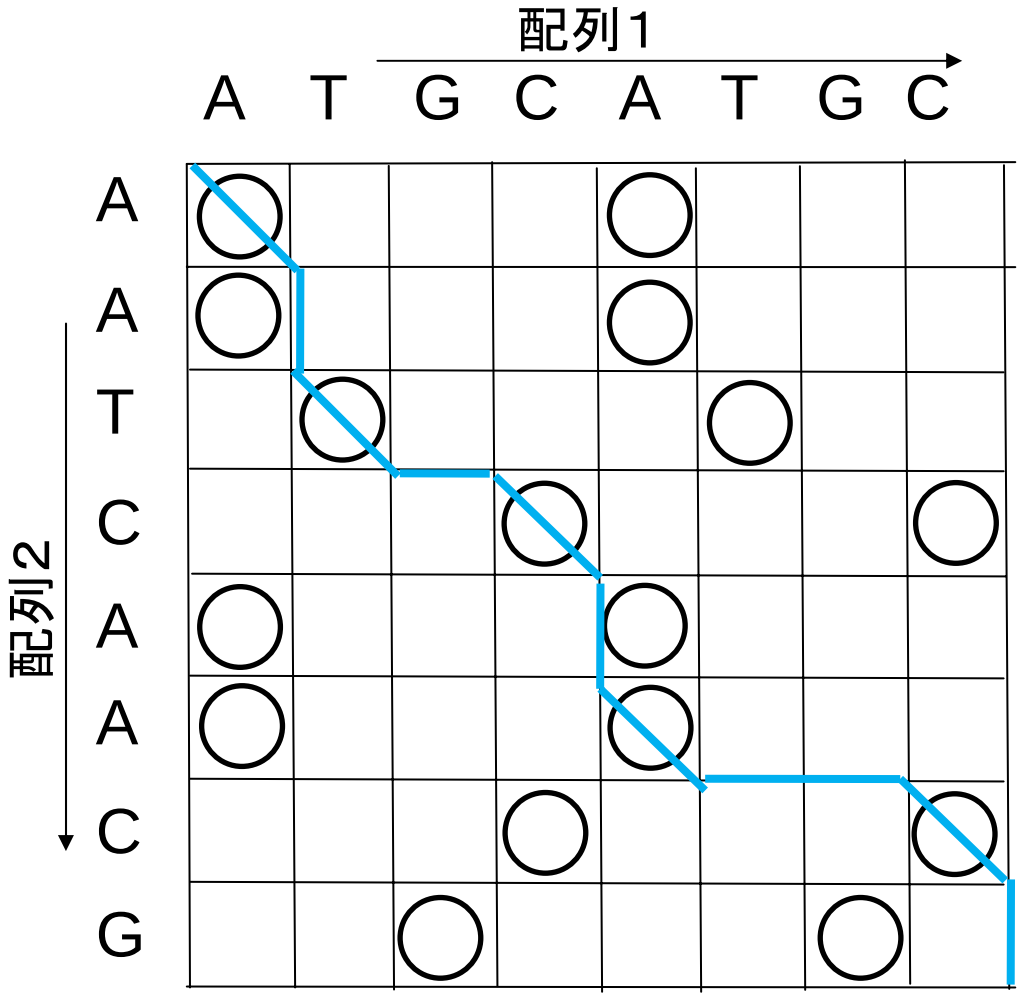
※スコア：一致：＋1、不一致：0、ギャップ：0とする。

ATGCATGC
AATCAACG

(1)配列1、配列2を
横と縦に並べる

(2)文字が一致する
マスに〇を描く

(3)多くの〇を通るような
左上と右下を結ぶ折れ線



A-TGC-ATGC-
*** * * * ***
AAT-CAA--CG

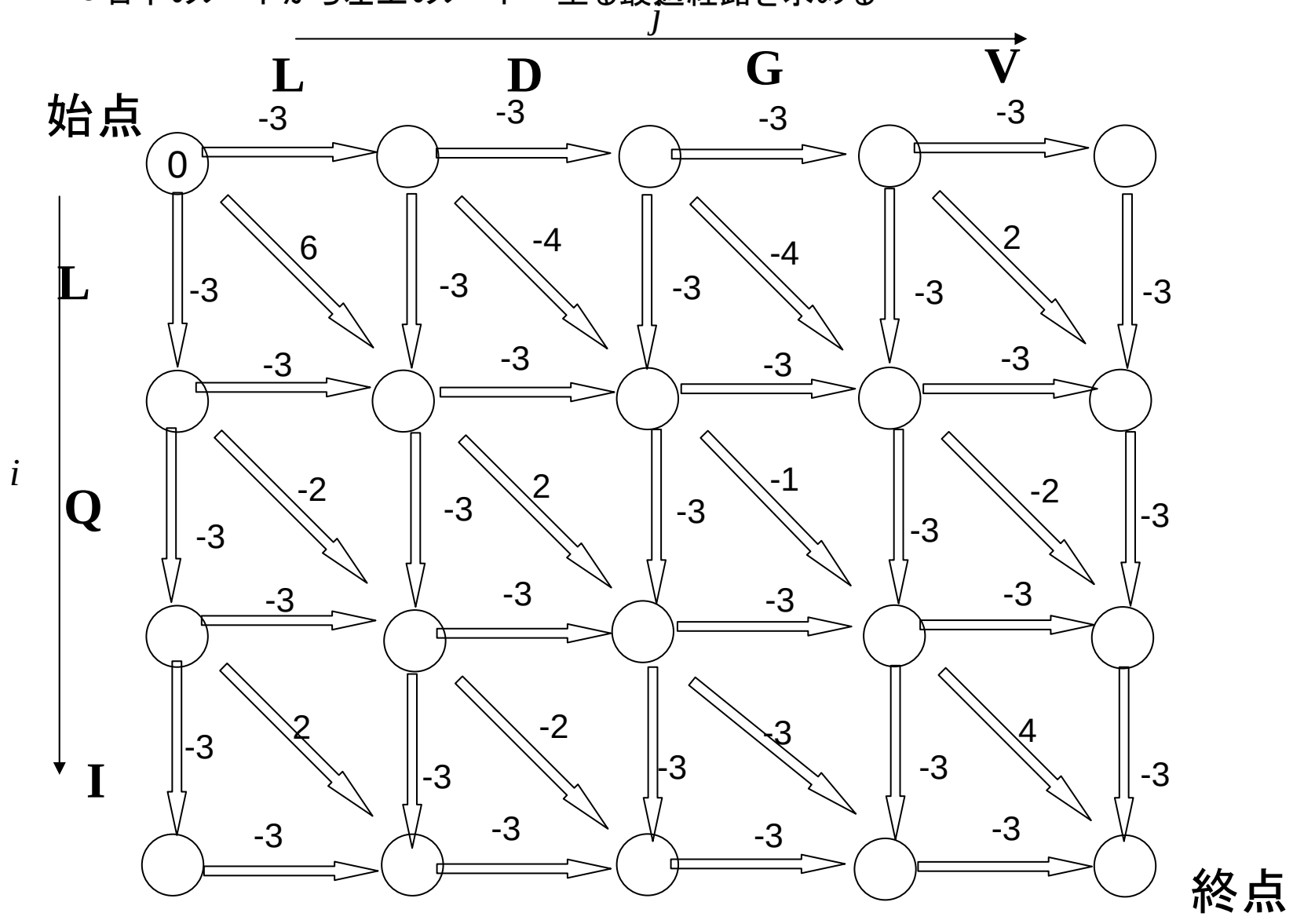
この場合、解は何通りもあるが、いずれも一致する残基数は5

動的計画法によるアライメント

- アライメント問題は、有向グラフの最適経路問題と等価
- 有向グラフの最適経路問題は動的計画法（Dynamic Programming)と呼ばれるアルゴリズムで解ける。
- *O(NM)*の計算量（文字列長の積に比例）

動的計画法によるグローバルアライメントの解法

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める

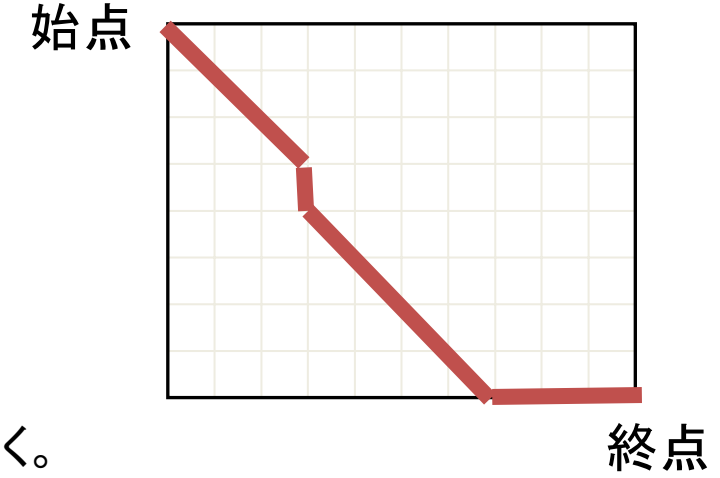


動的計画法によるグローバル・アライメントの解法 (Needleman & Wunsh,1970)

※$D(i,j)$は始点$(0,0)$から格子点(i,j)までのスコアの和の最大値

(0)準備

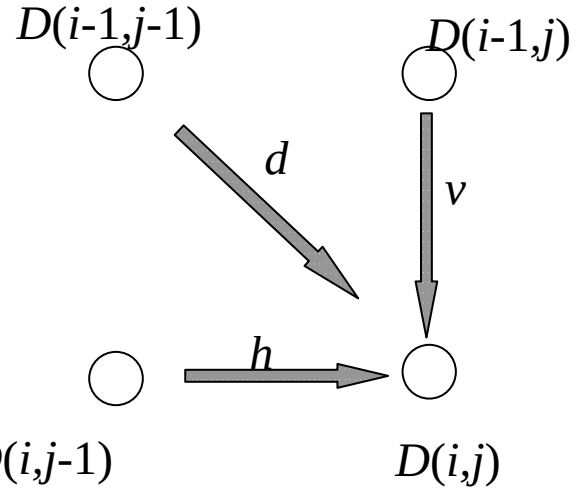
始点の格子点のスコア$D(0,0)$を0に設定



(1)前向きステップ

i=1,j=1から、開始し、iとjを一つずつ
大きくしながら、以下の式に従って、$D(i,j)$を決めていく。
そのとき、使用した矢印をマークする。

$$D(i,j)=\max\begin{cases} D(i-1,j-1)+s(i,j) & \text{対角}(d) \\ D(i-1,j)-Gap & \text{鉛直}(v) \\ D(i,j-1)-Gap & \text{水平}(h) \end{cases}$$



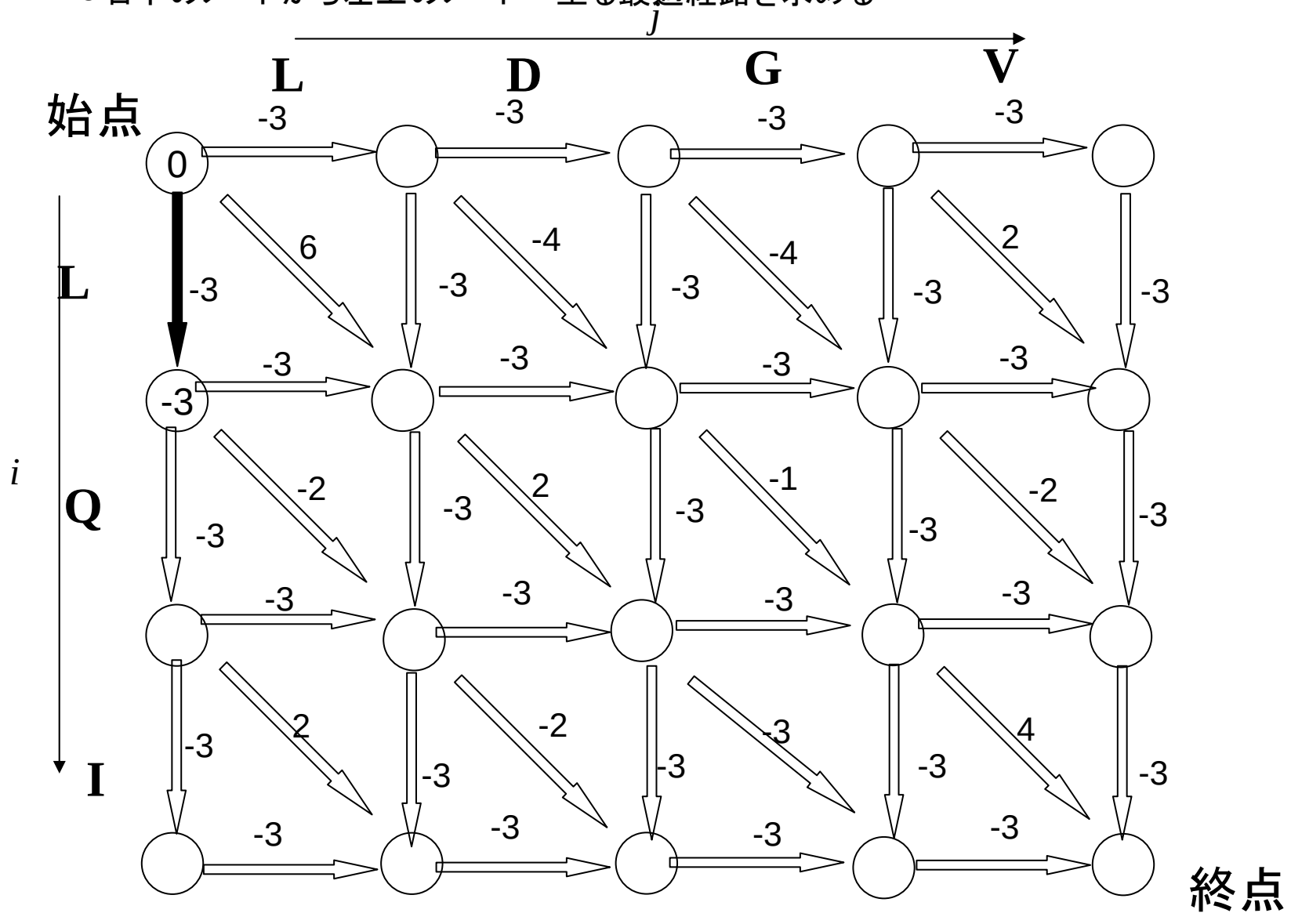
$s(i,j)$は配列1のi番目と配列2のj番目の文字がマッチしたときのスコア

(2)後ろ向きステップ

終点を起点にして、マークした矢印を逆向きにたどる。終点に到着したら終了。

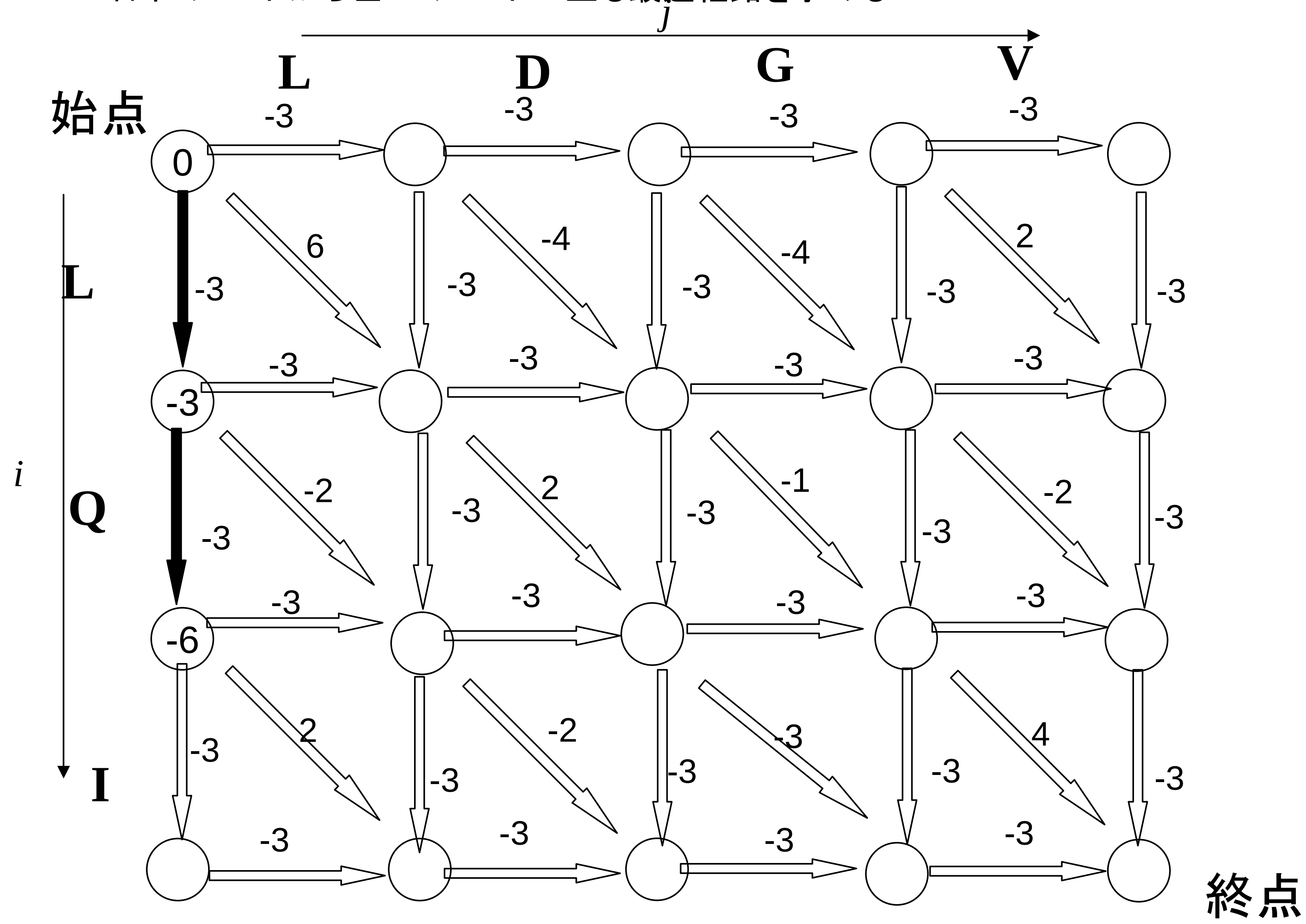
左端と上端の$D(i,j)$をまず、決めていく

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



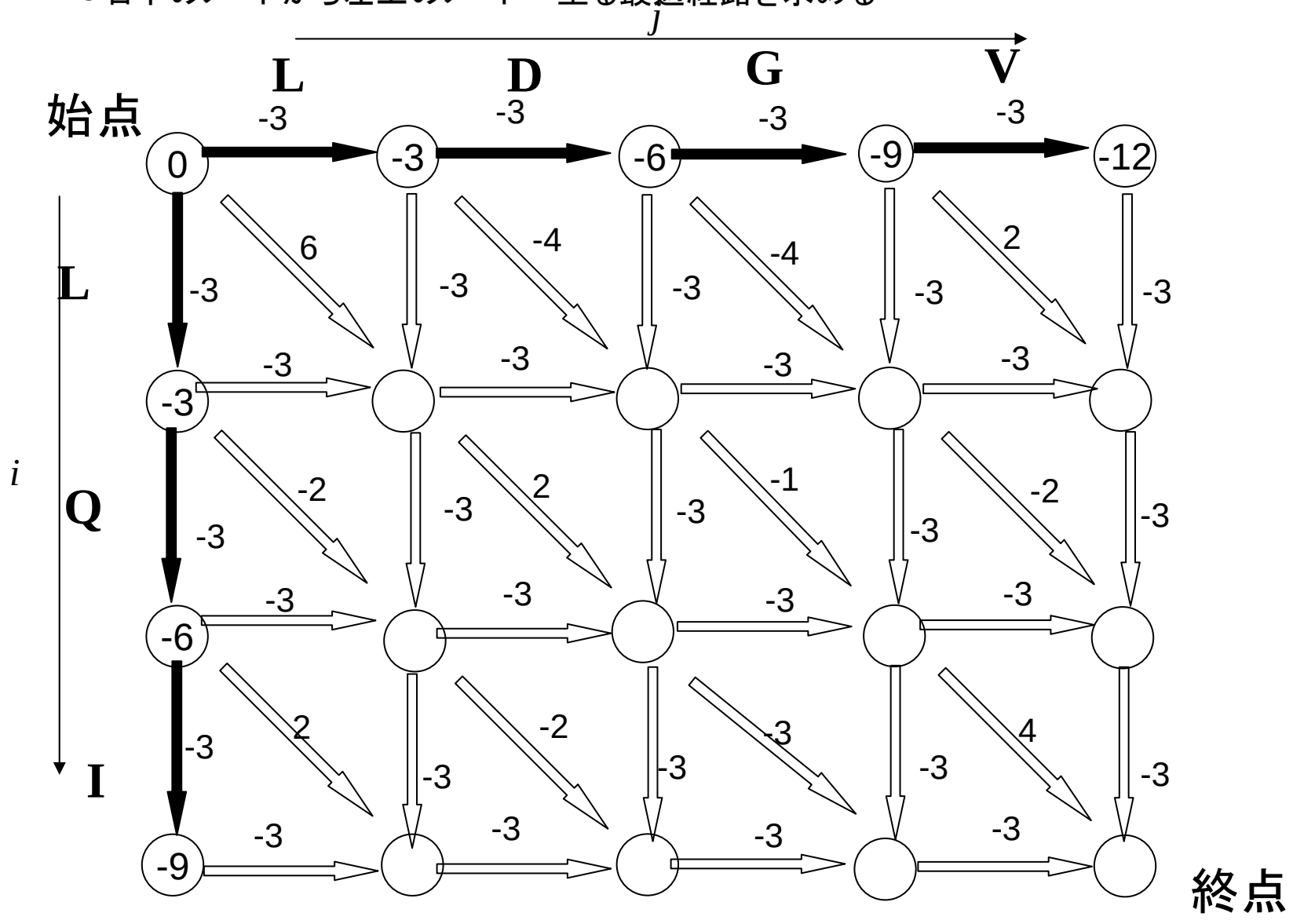
左端と上端の$D(i,j)$をまず、決めていく

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



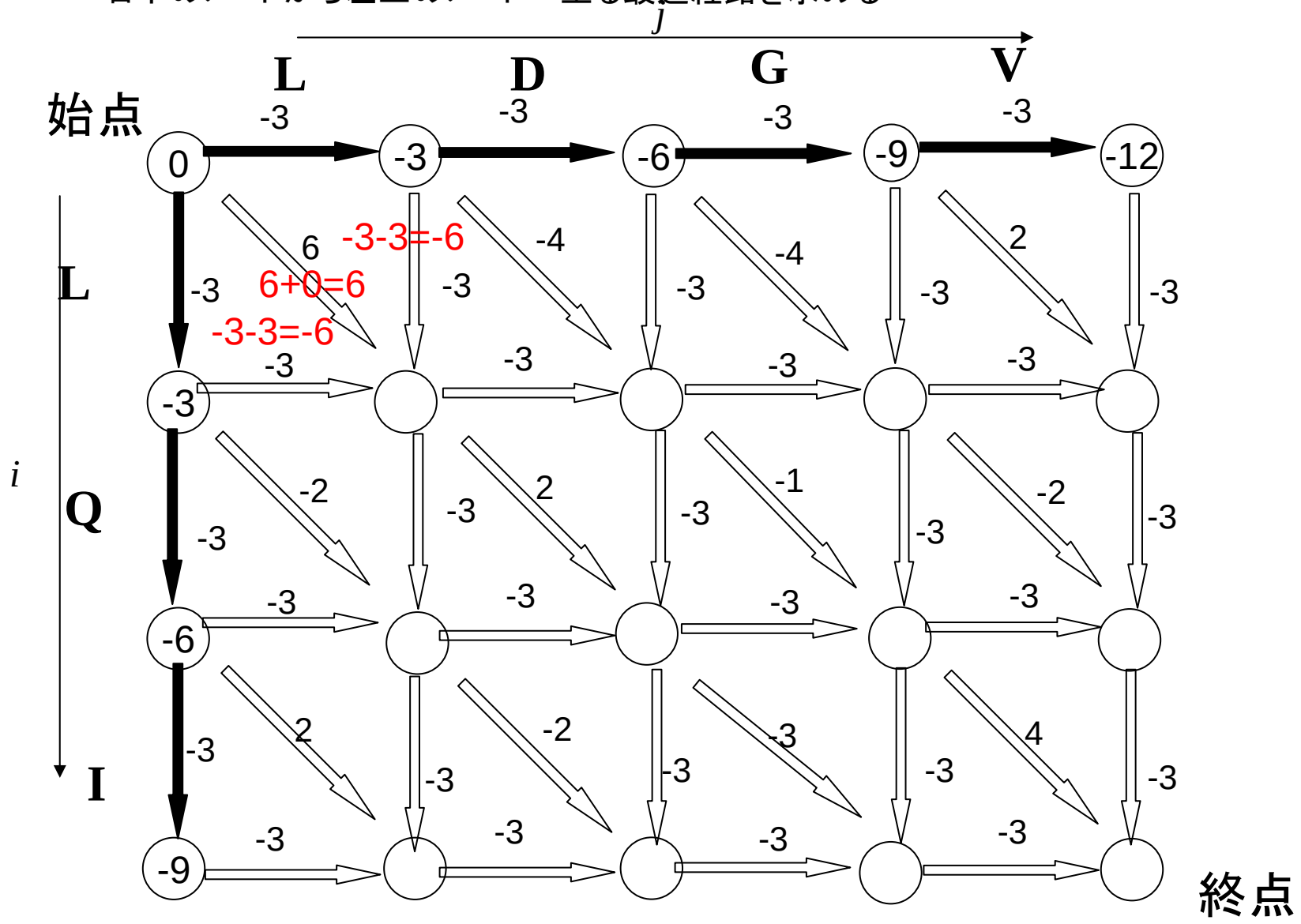
左端と上端の$D(i,j)$をまず、決めていく

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



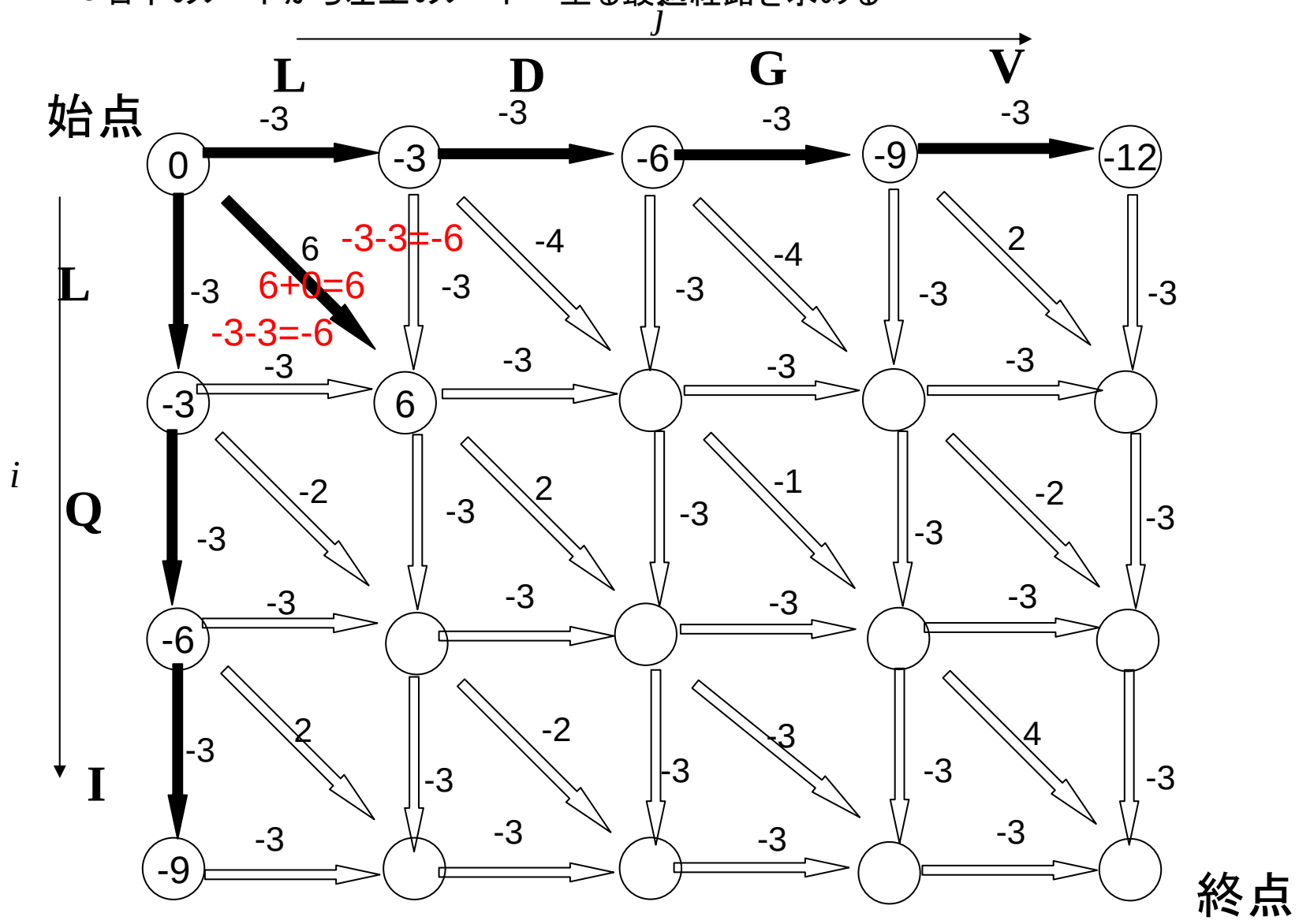
（1）前向きステップ：たて、よこ、ななめのスコアを比べる

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



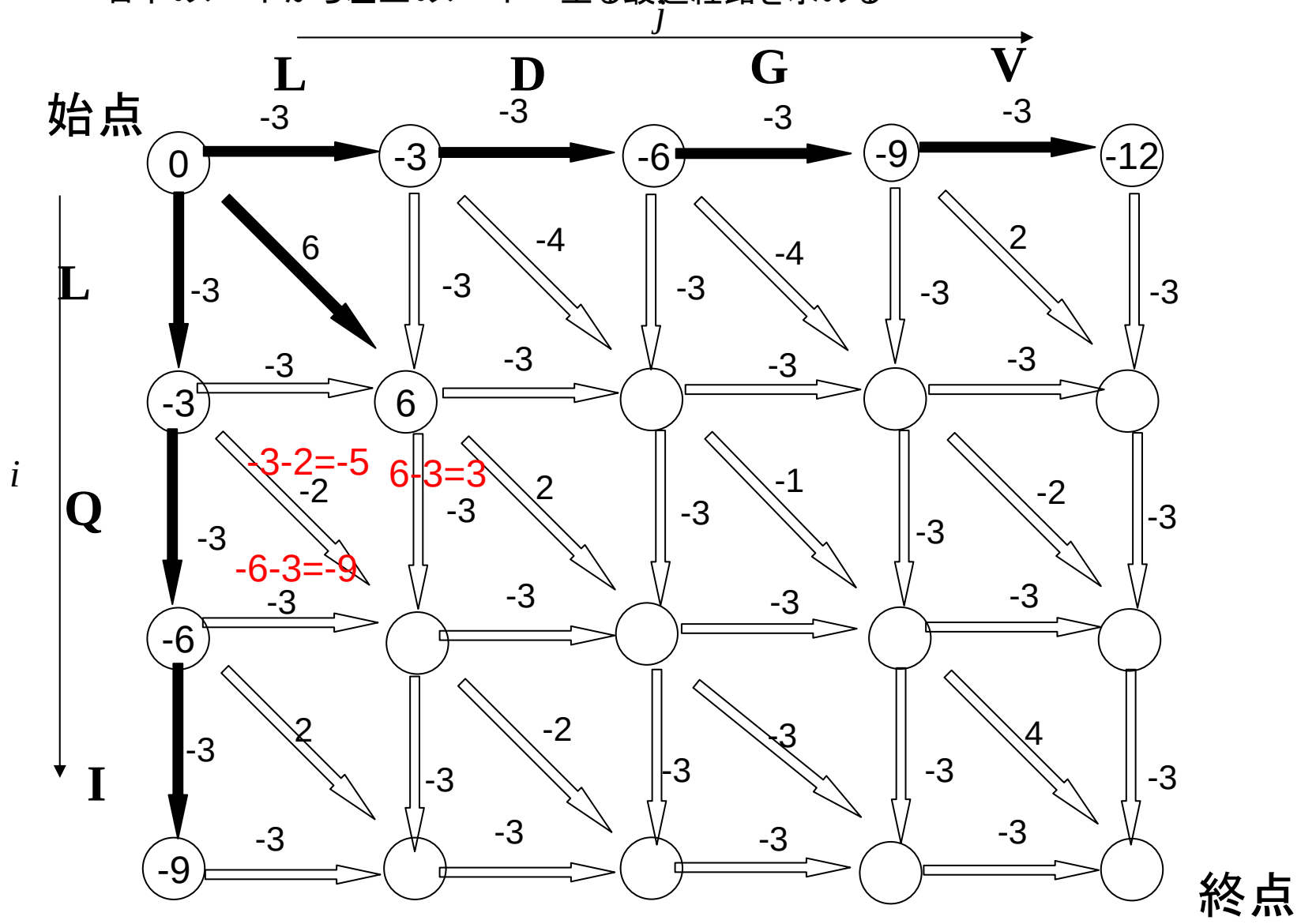
（1）前向きステップ：たて、よこ、ななめのスコアを比べる

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



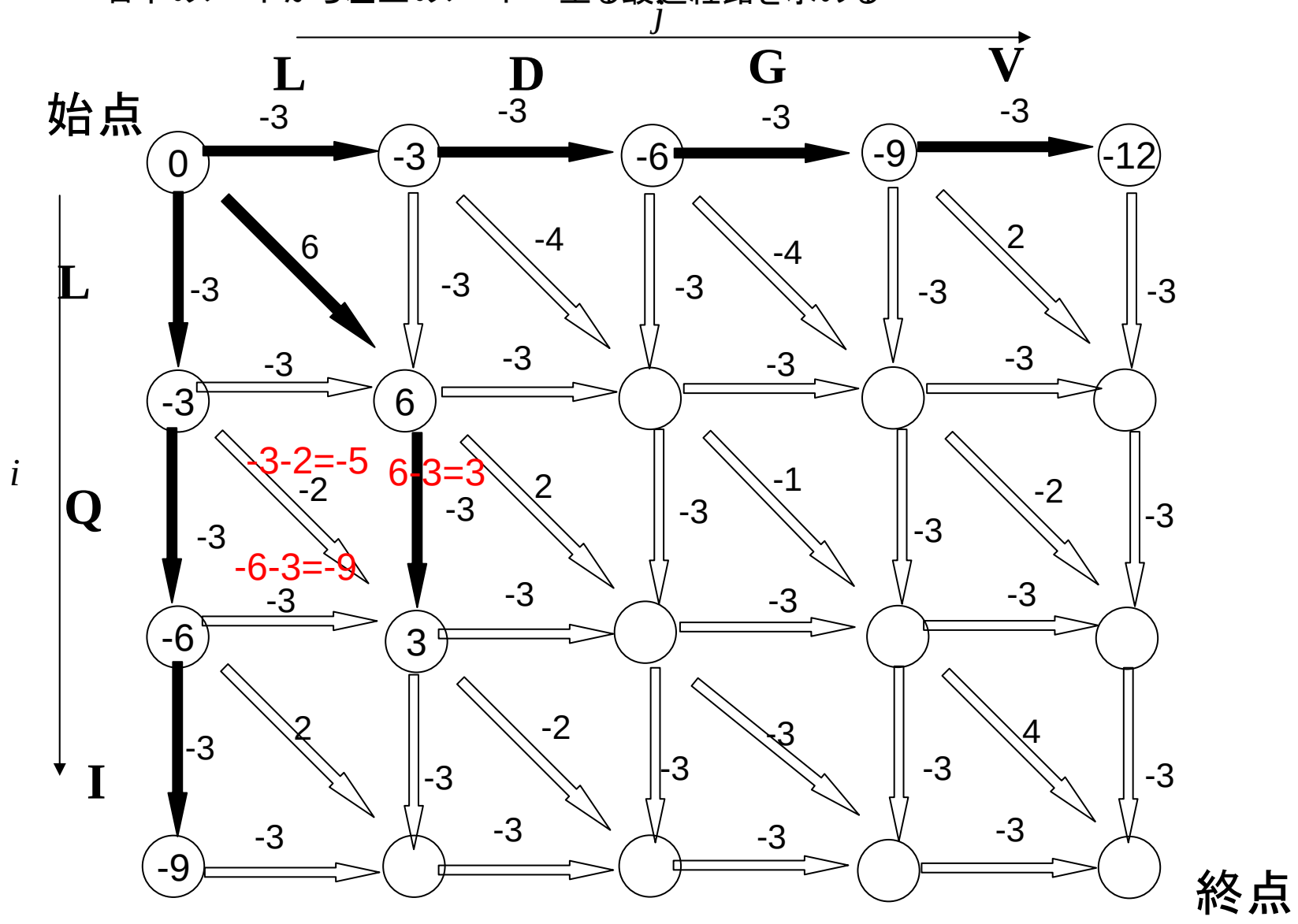
（1）前向きステップ：たて、よこ、ななめのスコアを比べる

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



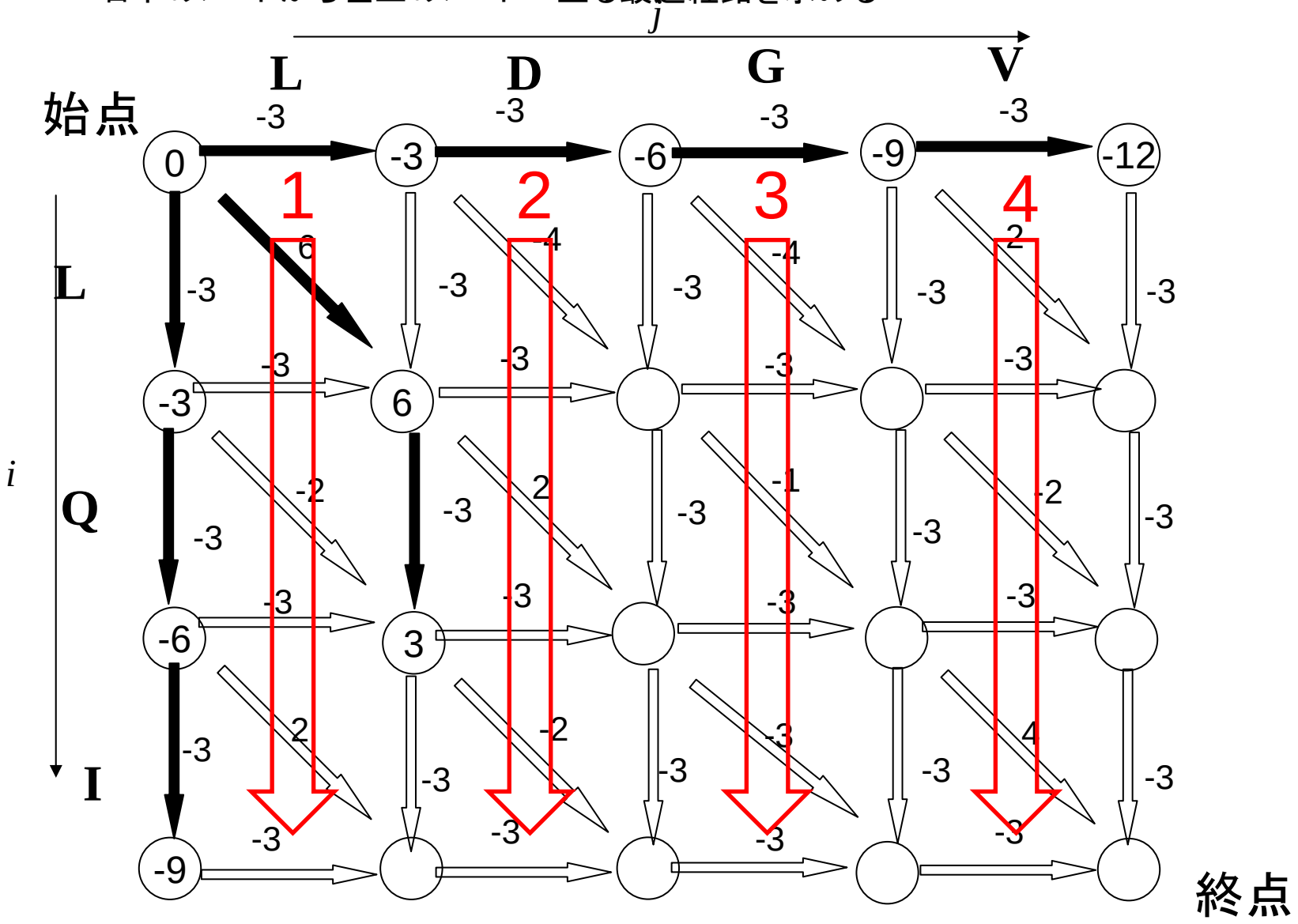
（1）前向きステップ：たて、よこ、ななめのスコアを比べる

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



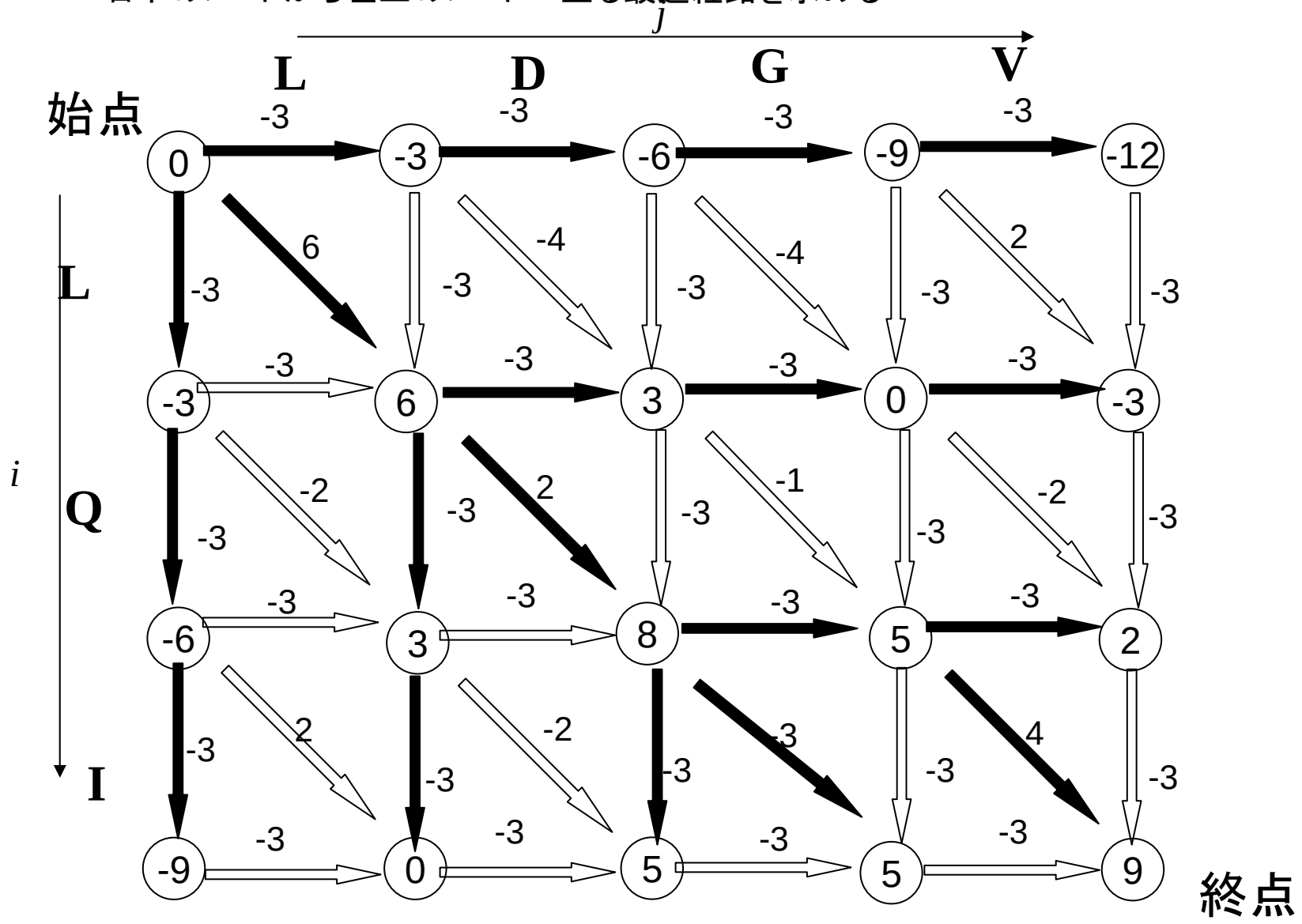
（1）前向きステップ：たて、よこ、ななめのスコアを比べる

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



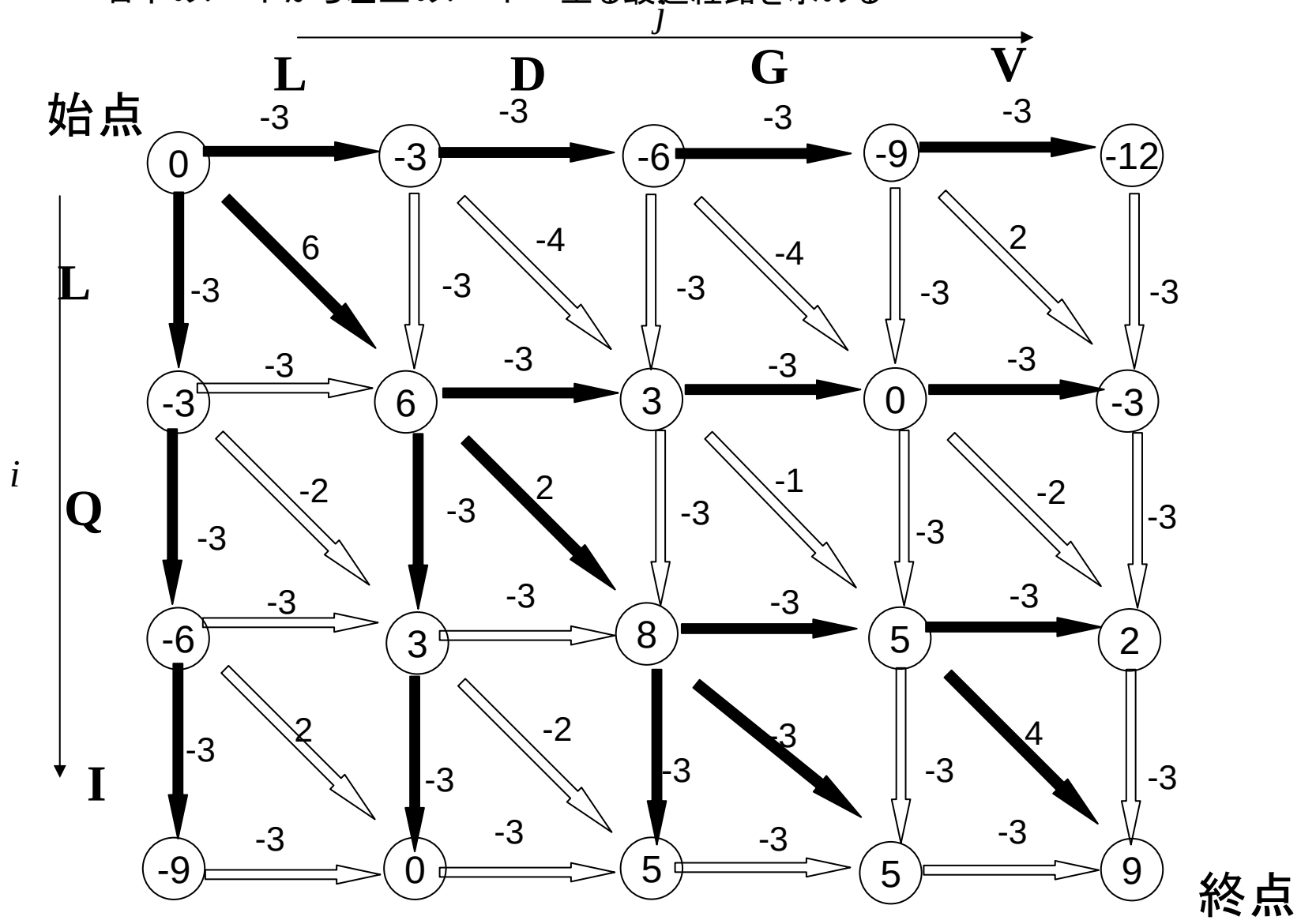
（1）前向きステップ：たて、よこ、ななめのスコアを比べる

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



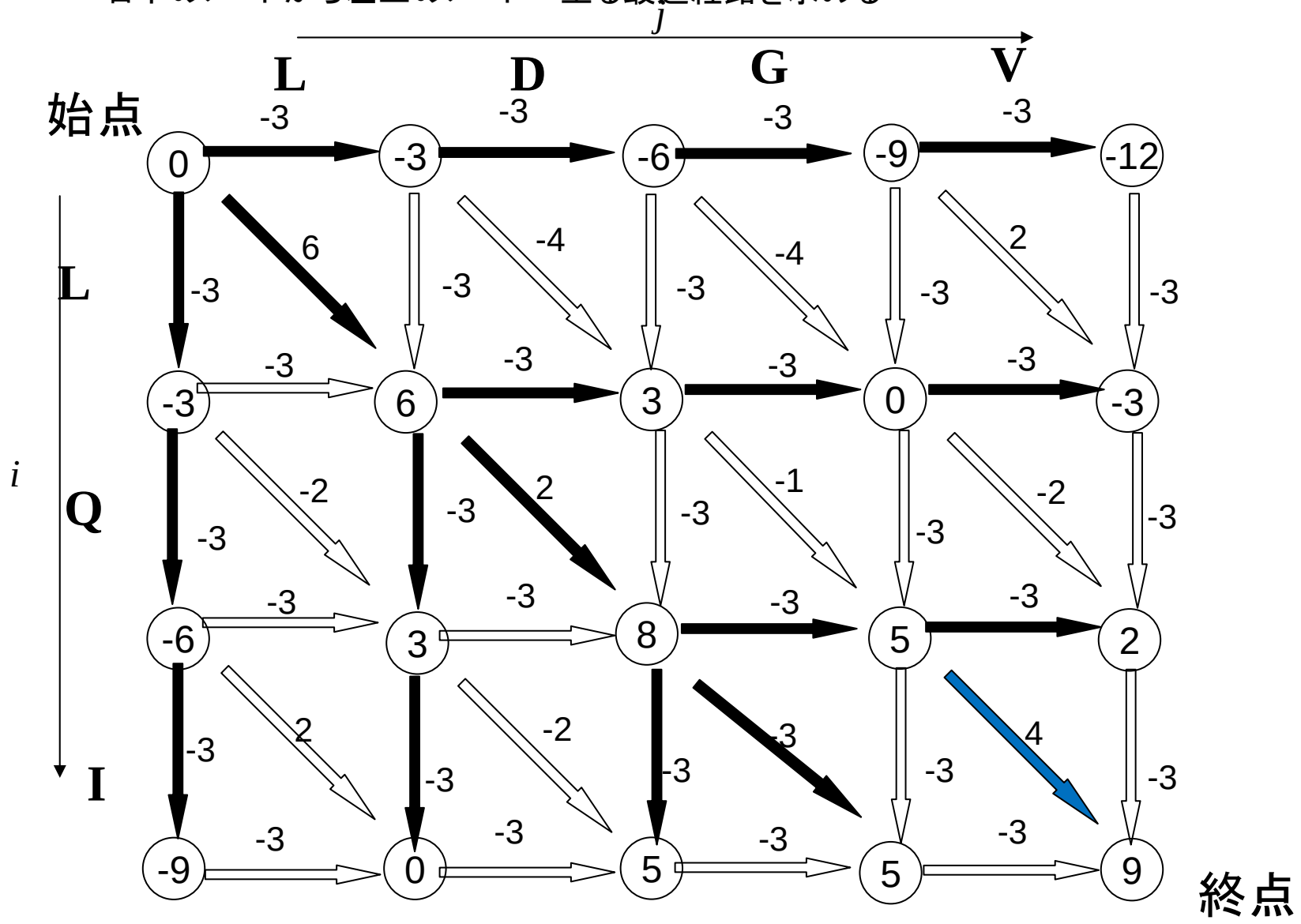
（2）後ろ向きステップ：マークした矢印を終点から

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



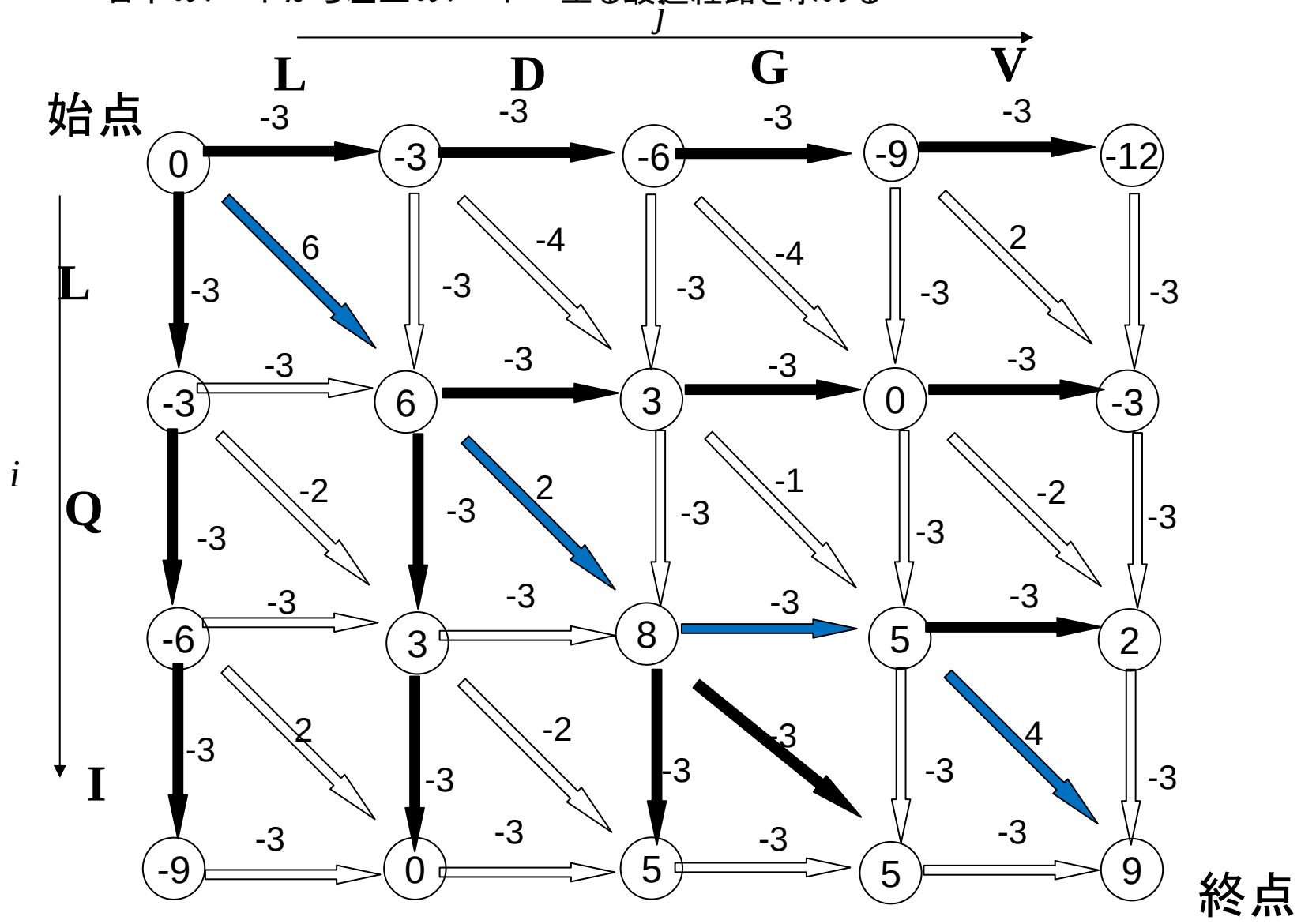
（2）後ろ向きステップ：マークした矢印を終点から

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



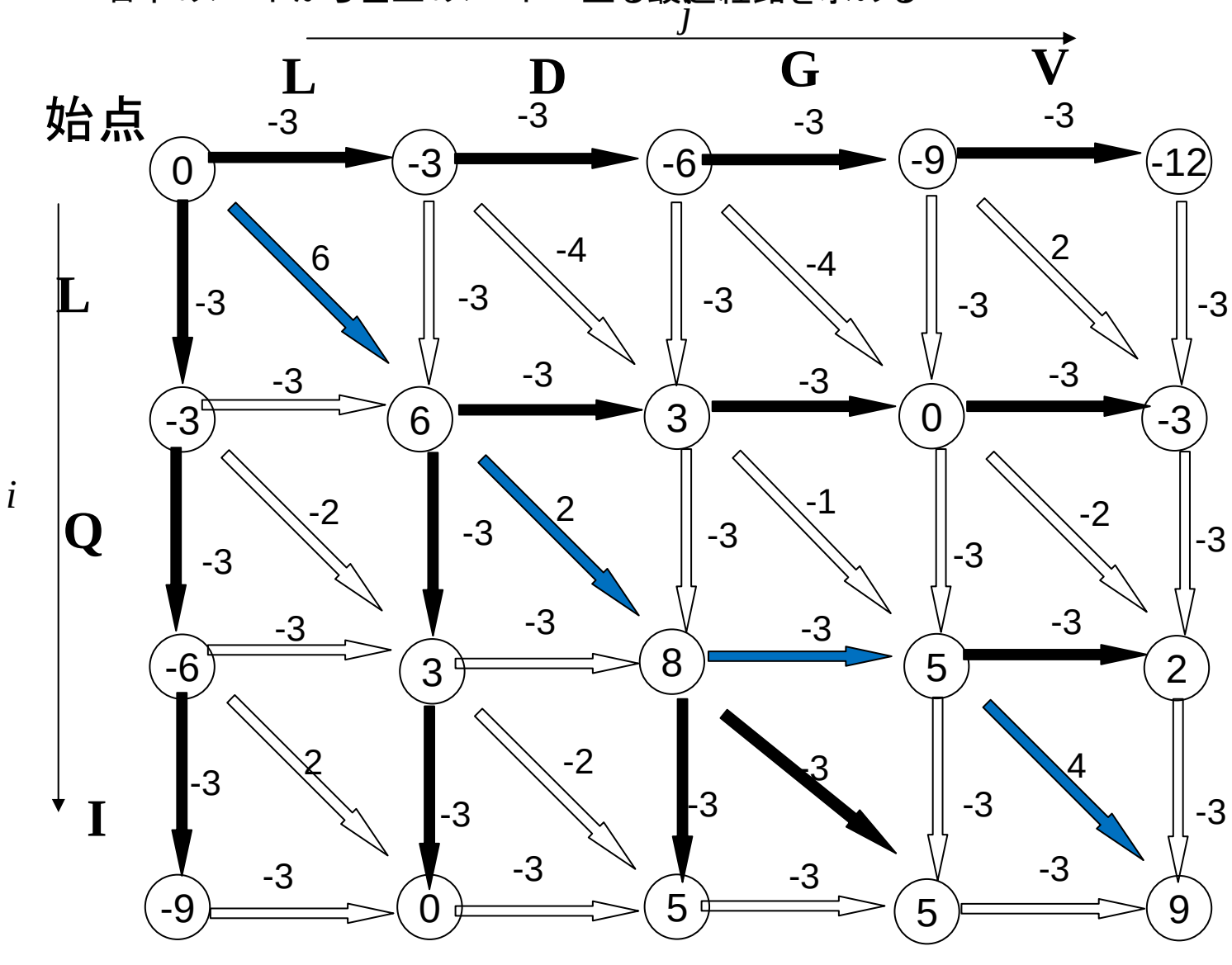
（2）後ろ向きステップ：マークした矢印を終点から

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



(2)後ろ向きステップ:マークした矢印を終点から

- 鉛直、水平に比較したい文字列を並べる
- 対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- 右下のノードから左上のノードへ至る最適経路を求める



LDGV

LQ-I

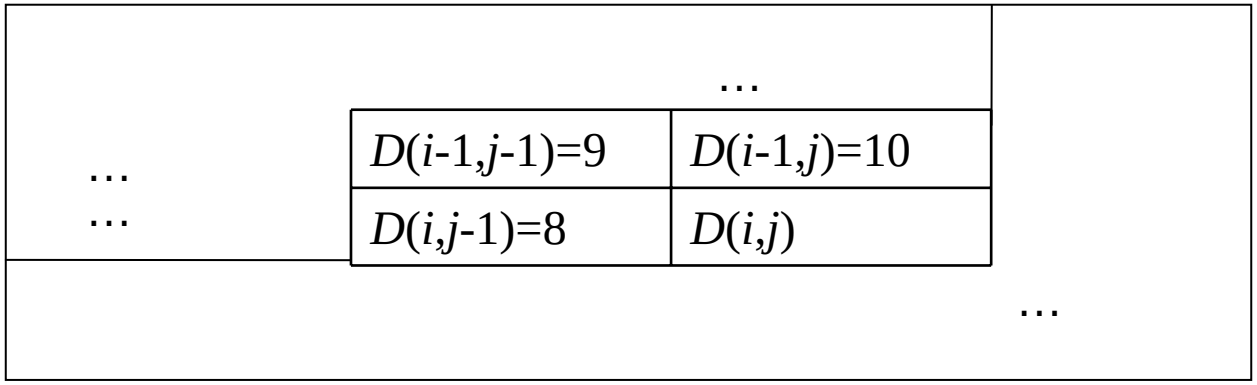
スコア:9点

H20　問51

DNA塩基配列2本のグローバルアライメントを動的計画法を用いて作成する。動的計画法の漸化式は、

$$D(i, j) = Max\begin{cases} D(i-1, j-1) + s(i, j) \\ D(i-1, j) - p \\ D(i, j-1) - p \end{cases}$$

とする。ここで、$s(i,j)$は、第一の配列のi番目の塩基と第二の配列のj番目の塩基が一致していれば1、不一致であれば0の値をとる。pはギャップペナルティであり、正の値2をとる。漸化式を5'から解き、$D(i\text{-}1,j\text{-}1),D(i\text{-}1,j),D(i,j\text{-}1)$は図のように既に　求まっているとする。一方の配列のi番目の塩基はG,他方の配列のj番目の塩基はTとする。このとき、$D(i,j)$の値を選択肢の中から一つ選べ。



1:7,　**2**: 8,　**3**: 9,　**4**:10

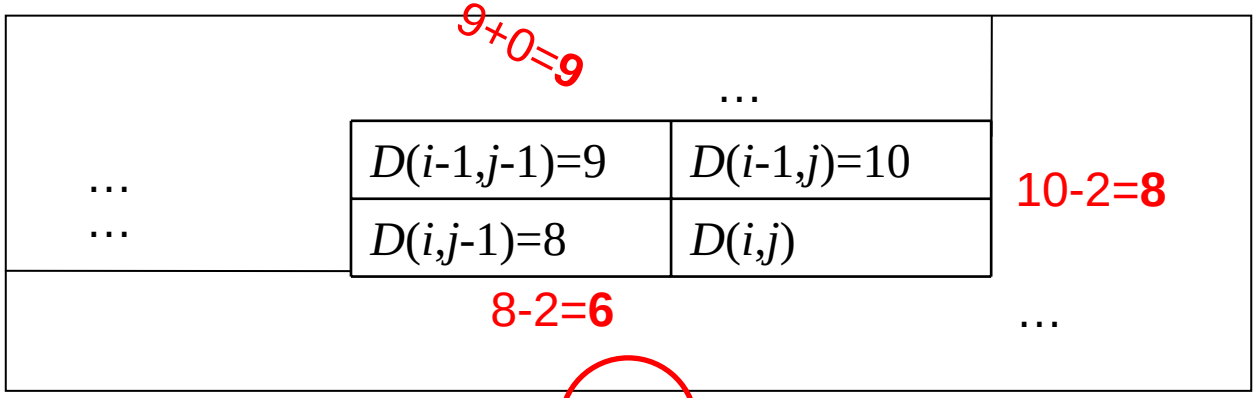
平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20　問51

DNA塩基配列2本のグローバルアライメントを動的計画法を用いて作成する。動的計画法の漸化式は、

$$D(i,j)=Max\begin{cases} D(i-1,j-1)+s(i,j) \\ D(i-1,j)-p \\ D(i,j-1)-p \end{cases}$$

とする。ここで、$s(i,j)$は、第一の配列のi番目の塩基と第二の配列のj番目の塩基が一致していれば1、不一致であれば0の値をとる。pはギャップペナルティであり、正の値2をとる。漸化式を5'から解き、$D(i\text{-}1,j\text{-}1)$,$D(i\text{-}1,j)$,$D(i,j\text{-}1)$は図のように既に　求まっているとする。一方の配列のi番目の塩基はG,他方の配列のj番目の塩基はTとする。このとき、$D(i,j)$の値を選択肢の中から一つ選べ。



1:7,　**2**: 8,　**3**: 9,　**4**:10

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

グローバルとローカルの格子上の違い

ACDEFGHKLM
AFGHKKL

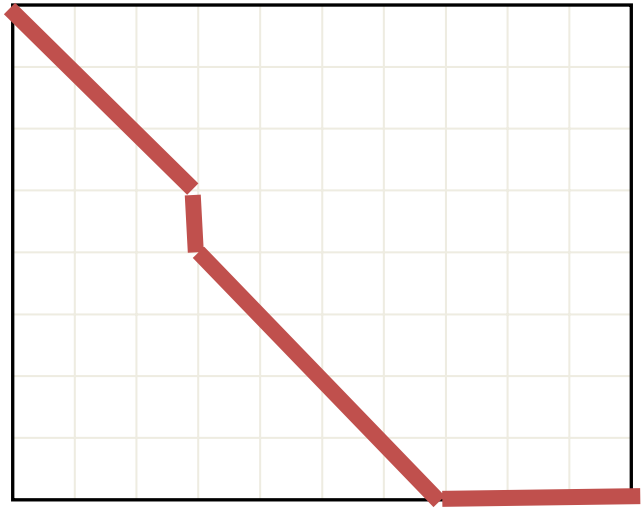
➡

ACDEFGHK-LM
A---FGHKKL-

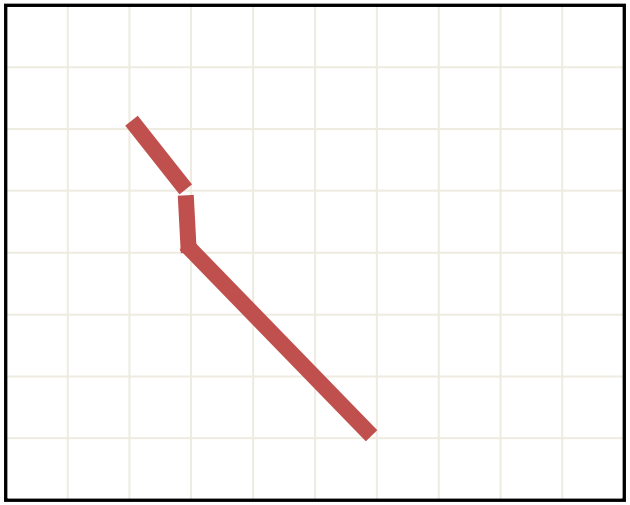
グローバル

FGHK-L
FGHKKL

ローカル



グローバル

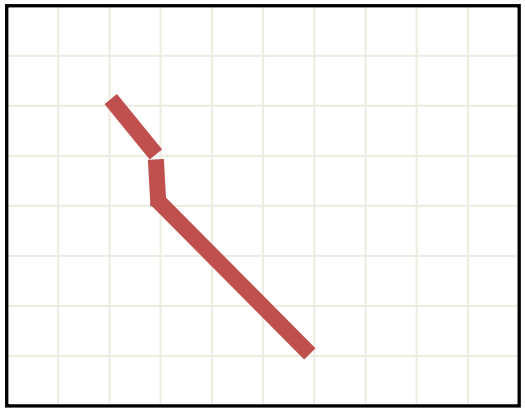


ローカル

ローカルアライメントの解法
(Smith & Waterman,1981)

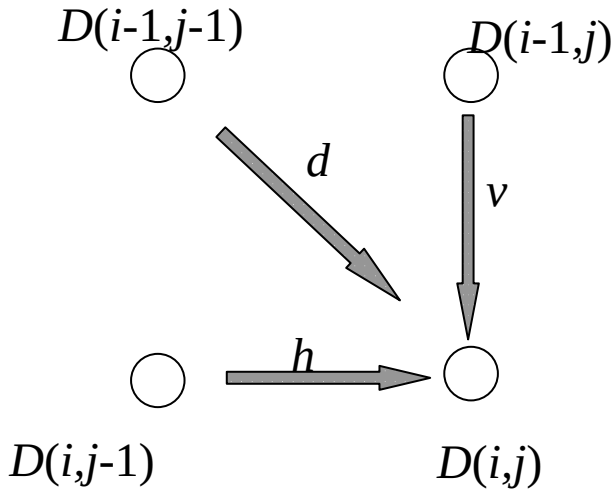
(0)準備

格子の端のスコアを0に設定



(1)前向きステップ

$$D(i,j)=\max\begin{cases} D(i-1,j-1)+s(i,j) & \text{対角}(d) \\ D(i-1,j)-Gap & \text{鉛直}(v) \\ D(i,j-1)-Gap & \text{水平}(h) \\ 0 & \text{終結}(0) \end{cases}$$



(2)後ろ向きステップ

最大のスコアのノードを探し、そのノードを起点にして辿る。パス'0'が現れたら終了

「配列解析」のキーワード（マルチプルアライメント）

- マルチプルアライメント
- 累進法（ツリーベース法）
- ClustalW

マルチプルアライメント（多重配列整列）とは

3本以上の配列を進化的な対応関係に従って並べること

>1nshA
SRPTETERCIESLIAVFQKYAGKDGHSVTLSKTEFLSFMNTELAAFTKNQKDPGVLDRMMKKLDLNSDGQLDFQEFL
NLIGGLAVAESFVKAAPPQKRF
>1j55A
MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLDAVDKLLKDLDANGDAQVDFSEFIVFVAAITS
ACHKYFEKAL
>1ig5A
KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ
>1qx2A
MKSPEEIKGAFEVFAAKEGDPNQISKEELKLVMQTLGPSLLKGMSTLDEMIEEVDKNGDGEVSFEEFLVMMKKISQ

```
CLUSTAL W (1.83) multiple sequence alignment

1nshA      SRPTETERCIESLIAVFQKYAGKDGHSVTLSKTEFLSFMNTELAAFTKNQKDPGVLDRMM
1j55A      --MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLD------AVDKLL
1ig5A      -------KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKG---PSTLDELF
1qx2A      ------MKSPEEIKGAFEVFAAKEGDPNQISKEELKLVMQTLGPSLLKG---MSTLDEMI
                     .   :   *. ::..:* .  ::* *:  .::.  ..: .      .:*.::

1nshA      KKLDLNSDGQLDFQEFLNLIGGLAVACHESFVKAAPPQKRF
1j55A      KDLDANGDAQVDFSEFIVFVAAITSACHKYFEKAGL-----
1ig5A      EELDKNGDGEVSFEEFQVLVKKISQ----------------
1qx2A      EEVDKNGDGEVSFEEFLVMMKKISQ----------------
           :.:* *.*.::.*.**  ::  ::
```

マルチプルアライメントの目的

```
1nshA      SRPTETERCIESLIAVFQKYAGKDGHSVTLSKTEFLSFMNTELAAFTKNQKDPGVLDRMM
1j55A      --MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLD------AVDKLL
1ig5A      -------KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKG---PSTLDELF
1qx2A      ------MKSPEEIKGAFEVFAAKEGDPNQISKEELKLVMQTLGPSLLKG---MSTLDEMI
                     .   :   *. ::..:* .  ::* *:  .::.  ..: .      .:*.::
```

- ファミリ内の機能的重要部位の検出
- ファミリを特徴付けるモチーフの発見
- プロフィール法による遠縁のホモログ発見
- 分子系統樹を作成するための第一ステップとして不可欠
- 進化的追跡法(evolutionary trace method)など、発展的な機能部位予測にも重要

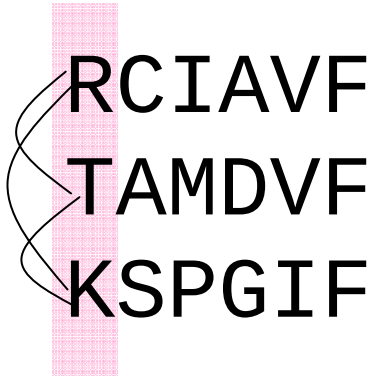
多重整列のスコア

（1）SP（sum-of-pairs)スコア

複数の文字列間のスコアを
ペアワイズのアミノ酸置換スコア$s(a,b)$の和で表す

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

m_i^k : k 番目の配列 の i番目の文字



$S(m_1)$ = s(R,T) + s(T,K) + s(R,K)

理論的にはおかしい： $S(A,B)+S(B,C)+S(A,C) = \log\frac{P(A,B)P(B,C)P(A,C)}{P(A)^2P(B)^2P(C)^2} \neq \log\frac{P(A,B,C)}{P(A)P(B)P(C)}$

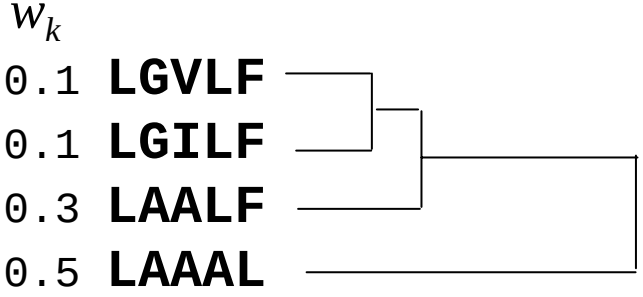
BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

多重配列のスコア（続き）

（2）配列への重み付きのSum-of-pair関数　(ClustalW)

$$S(m_i) = \sum_{k<l} w_k \cdot w_l \cdot s(m_i^k, m_i^l)$$



（3）エントロピー関数の最小化

各サイトのアミノ酸の頻度$p_i(a)$を推定し、そのエントロピーの和を求める

$$S(m_i) = -\sum_a p_i(a) \log p_i(a)$$

12345
LGVLF
LGILF
LAALF
LAAAL

サイト	*Pi(a)*	$S(m_i)$
1	$P_1(L)=1.0,$	*0.00*
2	$P_2(G)=0.5$,$P_2(A)=0.5$	*0.69*
3	$P_3(V)=0.25$, $P_3(I)=0.25$, $P_3(A)=0.5$	*1.04*

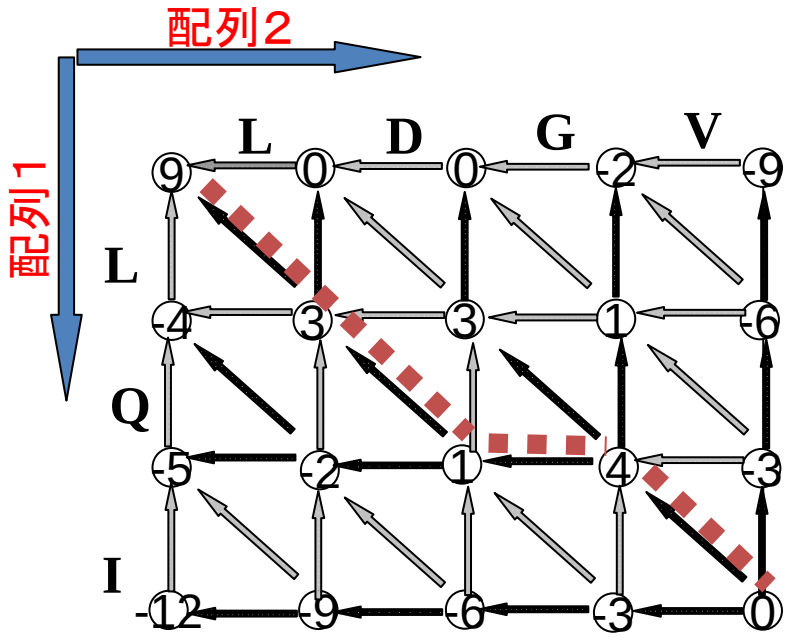
（4）対アライメントライブラリの重複による部位特異的スコア

(T-COFFEE)

どうやって並べるか？

多次元DPによる多重配列の厳密解

2本の配列のアライメント



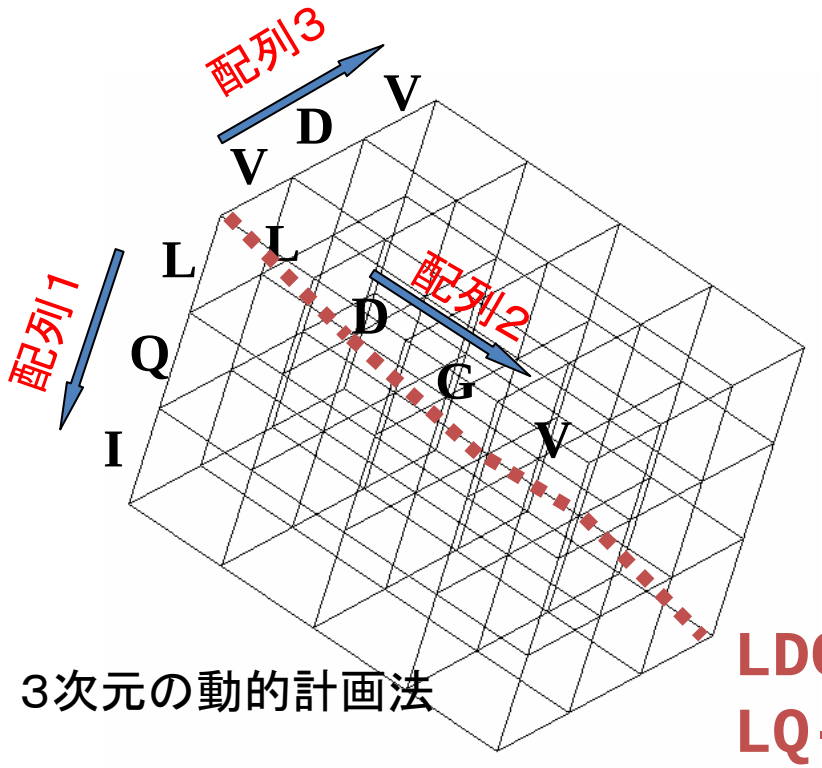
2次元の動的計画法

メモリ・計算時間　$O(L^2)$

LDGV

LQ-I

3本の配列のアライメント



3次元の動的計画法

メモリ・計算時間　$O(L^3)$

LDGV

LQ-I

VD-V

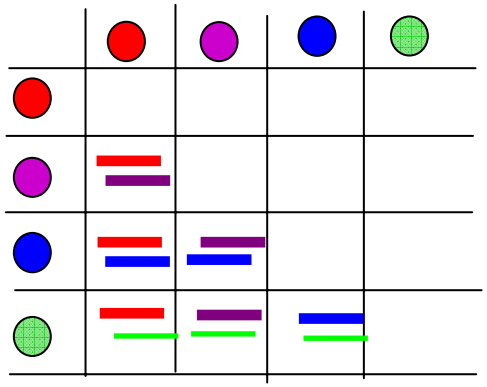
長さLの N本の配列のアライメントのメモリ・計算時間は$O(L^N)$

（　[配列の長さ]の[配列の本数]乗　に比例）　⇒　非現実的

長さ100の2本のアライメントが1秒でできても、10本に増やすと100^8 秒かかる！

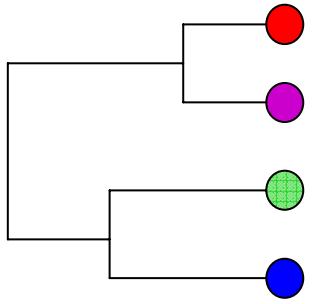
累進法
(progressive alignment, ツリーベース法)

Feng and Doolittle (1987)

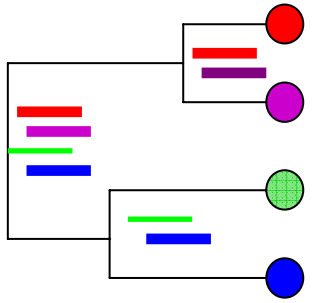


（1）全ての配列ペアのペアワイズアライメントを計算する

（2）ペアワイズアライメントによる距離行列を計算し、樹形図を計算する。



（3）樹形図の葉から、ペアワイズアライメントを組み上げていく



※ステップ1に最も計算時間がかかる。
全体の計算量は[配列の本数]^2×[配列の長さ]にほぼ比例

ClustalW / ClustalX

UNIX/Windows/Mac版：ftp://ftp.ebi.ac.uk/pub/software/clustalw2
WEBサーバ：http://www.ebi.ac.uk/Tools/clustalw2

・現在、最も一般的な多重整列のプログラム

・アルゴリズムは累進法。ペアワイズアライメントはグローバルアライメントを用い、
ガイド木はNJ法で　作成。スコアは配列の重みを導入したSum-of-pairs。
置換スコア行列の選択、ギャップペナルティ等に様々な経験的な工夫が見られる。

・CUI版はClustalW, GUI版はClustalX.
UNIX, Windows, MACでも動作する。

・NJ法による系統樹計算機能付き。

Thompson, J.D., Higgins, D.G., Gibson T.J. “CLUSTALW : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. Nucleic Acids Reseach, 1994, 22, 4673-4680.

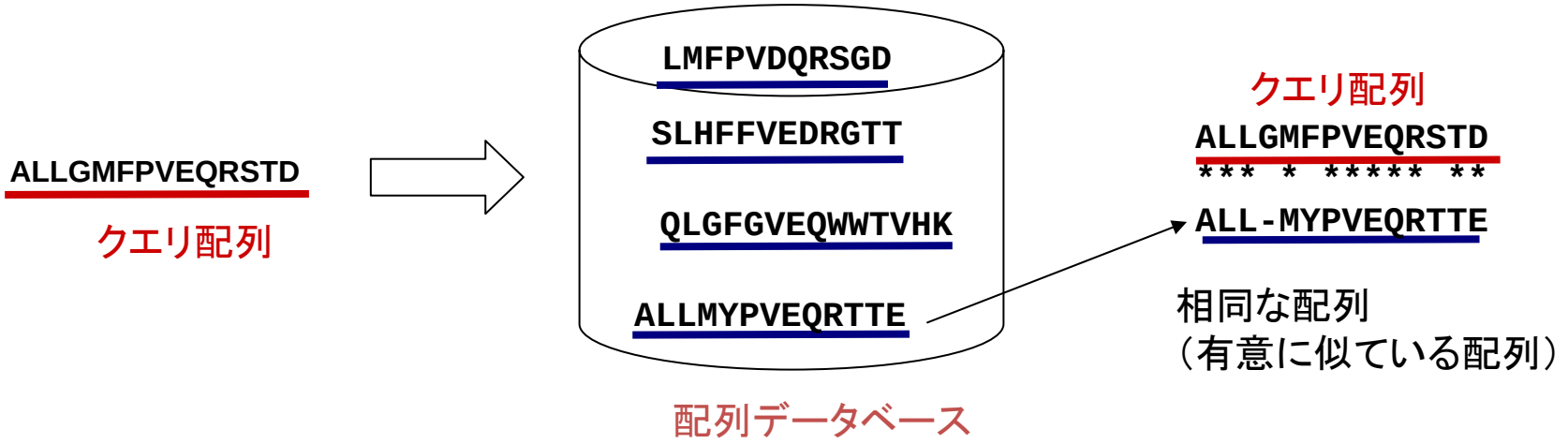
主要なマルチプルアライメントのプログラム

	WEBサイト	アルゴリズム	特徴
ClustalW・ClustalX	http://www.ebi.ac.uk/Tools/clustalw2	累進法。重み付きSPスコアを使用。 置換スコア行列の選択、ギャップペナルティ等に様々な工夫	もっとも広く使われている標準的なプログラム
T-COFFEE	http://www.ebi.ac.uk/t-coffee/	ペアワイスアライメントをローカル、グローバル、進展を用いて多数生成。それらの集合から、位置特異的スコアを作成し、累進法を実行する。	計算時間がかかるが精度は高い。**配列の本数が100本以下**の場合に向いている。
MAFFT	http://align.bmr.kyushu-u.ac.jp/mafft/online/server/	高速フーリエ変換(FFT)を用いて、高速にペアワイズアライメントを実装、それを利用して、累進法、あるいは反復改善法を実行する。	計算時間は高速なので、**配列の本数が100～500本程度**でも、計算可能。

「配列解析」のキーワード（相同性検索）

- 相同性検索
- FASTA
- ハッシング
- BLAST
- 有限オートマトン

配列相同性検索
(Sequence Homology Search)
→クエリ配列を配列データベースと比較、相同な配列を探す



- 機能未知遺伝子の機能予測(アノテーション)
  機能既知の配列との類似→機能の類似を示唆
- 立体構造予測
  構造既知の配列との類似→構造の類似を示唆
- 遺伝子発見
  既知遺伝子と類似している領域の発見→遺伝子の存在を示唆

配列相同性検索の基本動作原理

①2つの　DNA　/　アミノ酸　の文字列が似ている

↓

②進化的に関係がある（相同）から似ている

↓

③進化的に関係があるなら、他の生物学的な性質(機能、立体構造など)も似ているはず

相同性の発見により、他の生物学的な性質を予測できる

類似(similarity)

相同（homology）:進化的な原因によるもの。祖先を共有。

（進化史の中である時点まで同じであったから似ている）

相似（analogy） :それ以外の原因によるもの

配列データベースの中からクエリ配列と類似したエントリを見つけるには？

→　動的計画法を繰り返し実行すればよい

1. いかに高速に計算を実行するか

   動的計画法は$O(NM)$の計算時間

   1,000～100,000配列の検索には時間がかかる

   →　高度なヒューリスティック解法の導入

2. どれだけ似ていれば意味があるのか？

   何をもって類似性の指標とするのか

   同一残基率(%)、スコア？

   →統計的有意性の判断の導入

BLASTのアライメントアルゴリズム

動的計画法を使わず、独自のヒューリスティックアルゴリズムを開発

ヒューリスティック（発見的解法）
：常に正しい解を返すわけではないが、多くの場合まあまあの解を返すことが経験的に知られているアルゴリズム

計算時間の比較

153残基のクエリ配列を54,457配列のデータベースと比較
クアッドコアIntel Xeon X5355(2.66GHz)でシングルCPUで計算

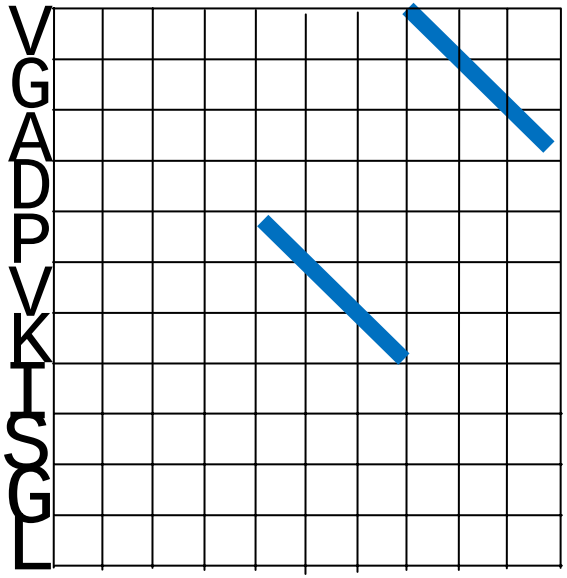
	説明	計算時間
私が書いたDP	Smith & WatermanをCで素朴に実装	144.97 sec
SSEARCH35	FASTAの開発グループが実装した Smith & Waterman	15.01 sec
FASTA35	ヒューリスティックアルゴリズムを使用	2.36 sec
BLASTP	ヒューリスティックアルゴリズムを使用	0.38 sec

BLASTの発見的アルゴリズム

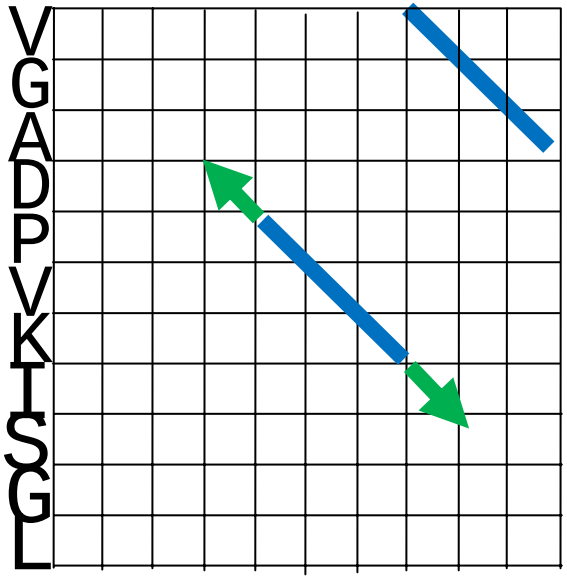
目標：Smith&Watermanのローカルアライメントの DPの近似解

1. クエリの各wordに対し、スコアの高い類縁wordのリストを作成。クエリについてハッシュ表を作る。
2. 類縁wordリストのハッシュ表を用いてデータベースを検索
3. ヒットしたwordをungapで伸展(HSP)
4. 動的計画法を行いgap入りアライメントでさらに伸展

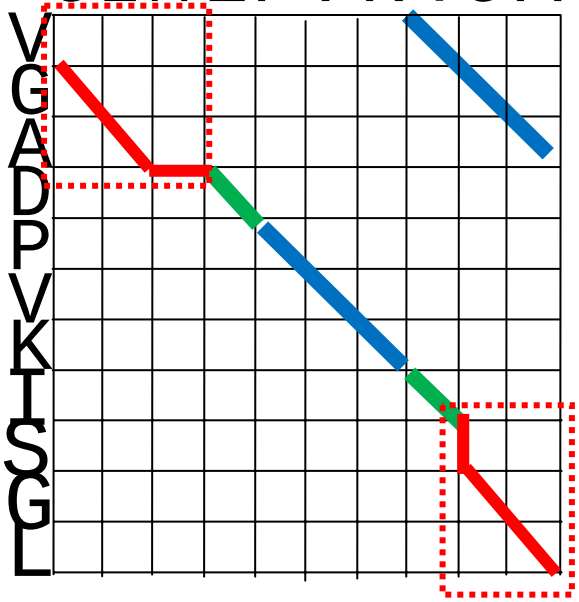
ステップ2



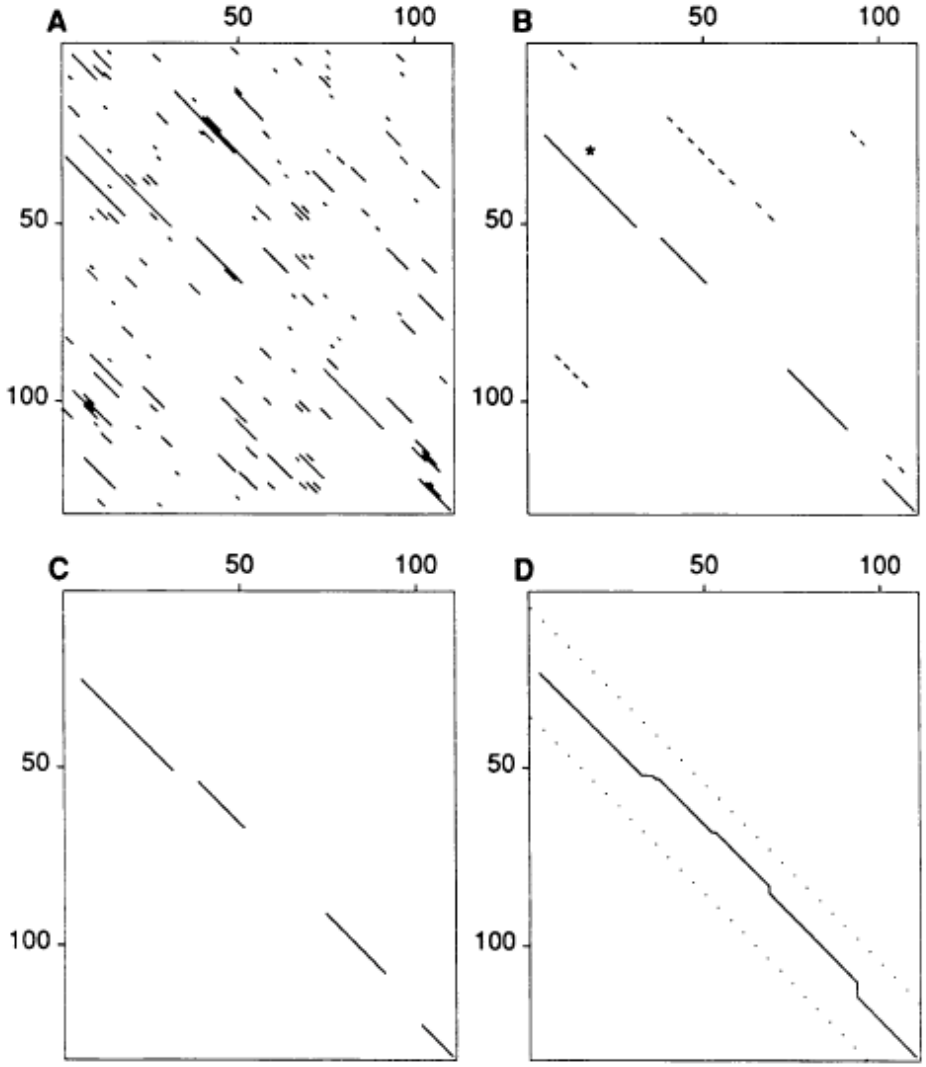
ステップ3



ステップ4



FASTAの発見的アルゴリズム



A) 連続する長さkの同一のwordを抽出。(このkをktupという)ハッシュ表を使用

B) スコア行列を用いて、最適な初期領域を絞り込む

C) 初期領域を接続する

D) 領域内で動的計画法を実行、アライメントを得る

Pearson WR, Lipman DJ. PNAS, 85,2444-2448 (1988)

H20　問52

FASTAに関する記述について、不適切なものを選択肢の中から一つ選べ。

1. FASTAは、DNAの塩基配列やタンパク質のアミノ酸配列のデータベース検索を行うためのソフトウエアである。
2. 塩基配列検索のとき、タプル(tuple)のサイズをkからk+2にすると、検索速度は32倍速くなる。
3. FASTAは、部分一致文字列の検出にハッシュ表を用いている。
4. タプルのサイズが大きくなるとホモロジー検索速度は向上するが、検索の感度は低下する傾向がある。

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20　問52

FASTAに関する記述について、不適切なものを選択肢の中から一つ選べ。

1. FASTAは、DNAの塩基配列やタンパク質のアミノ酸配列のデータベース検索を行うためのソフトウエアである。
2. 塩基配列検索のとき、タプル(tuple)のサイズをkからk+2にすると、検索速度は32倍速くなる。
3. FASTAは、部分一致文字列の検出にハッシュ表を用いている。
4. タプルのサイズが大きくなるとホモロジー検索速度は向上するが、検索の感度は低下する傾向がある。

※　タプルの種類が4^kから4^{k+2}個になるので、4^2=16倍、タプルの種類が増える。
よって、タプルがヒットする数もおよそ1/16になり、計算時間は約16倍速くなると考えられる。

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

BLAST(blastp)の出力例(1)

```
BLASTP 2.2.16 [Mar-25-2007]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= RECA_BACSU Protein recA [Bacillus subtilis]
         (347 letters)

Database: 40scop1.75nm.fasta
           9671 sequences; 1,701,902 total letters


                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

1u94A1 [c.37.1.11] RECA PROTEIN (1 A 6 268 )                        259   2e-70
1u94A2 [d.48.1.1] RECA PROTEIN (1 A 269 328 )                        61   1e-10
1rypG [d.153.1.4] 20S PROTEASOME                                     29   0.36
1p9rA [c.37.1.11] GENERAL SECRETION PATHWAY PROTEIN E                29   0.47
1n0wA [c.37.1.11] DNA REPAIR PROTEIN RAD51 HOMOLOG 1                 28   1.1
1uq5A [d.165.1.1] RICIN                                              27   1.8
1rypB [d.153.1.4] 20S PROTEASOME                                     27   2.4
1wg7A [b.55.1.1] DEDICATOR OF CYTOKINESIS PROTEIN 9                  26   3.1
1ji0A [c.37.1.12] ABC TRANSPORTER                                    26   4.0
1xx7A [a.211.1.1] OXETANOCIN-LIKE PROTEIN                            25   5.3
1ec7A1 [c.1.11.2] GLUCARATE DEHYDRATASE (1 A 138 446 )               25   6.9
1otkA [a.25.1.2] PHENYLACETIC ACID DEGRADATION PROTEIN PAAC          25   9.0

>1u94A1 [c.37.1.11] RECA PROTEIN (1 A 6 268 )
          Length = 243
```

```
1ec7A1 [c.1.11.2] GLUCARATE DEHYDRATASE (1 A 138 446 )                  25   6.9
1otkA [a.25.1.2] PHENYLACETIC ACID DEGRADATION PROTEIN PAAC             25   9.0
```

BLAST(blastp)の出力例(2)

```
>1u94A1 [c.37.1.11] RECA PROTEIN (1 A 6 268 )
          Length = 243

 Score =  259 bits (662), Expect = 2e-70
 Identities = 143/263 (54%), Positives = 176/263 (66%), Gaps = 21/263 (7%)

Query: 4   RQAALDMALKQIEKQFGKGSIMKLGEKTDTRISTVPSGSLALDTALGIGGYPRGRIIEVY 63
           +Q AL  AL QIEKQFGKGSIM+LGE     + T+ +GSL+LD ALG GG P GRI+E+Y
Sbjct: 1   KQKALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIALGAGGLPMGRIVEIY 60

Query: 64  GPESSGKTTVALHAIAEVQQQ-RTSAFIDAEHALDPVYAQKLGVNIEELLLSQPDTGEQA 122
           GPESSGKTT+ L  IA  Q++ +T AFIDAEHALDP+YA+KLGV+I+ LL SQPDTGEQA
Sbjct: 61  GPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTGEQA 120

Query: 123 LEIAEALVRSGXXXXXXXXXXXXXPKAEIEGDMGDSHVGLQARLMSQALRKLSGAINKS 182
           LEI +AL RSG               PKAEIEG+      GL AR+MSQA+RKL+G + +S
Sbjct: 121 LEICDALARSGAVDVIVVDSVAALTPKAEIEGE------GLAARMMSQAMRKLAGNLKQS 174

Query: 183 KTIAIFINQIREKVGVMFGNPETTPGGRALKFYSSVRLEVRRAEQLKQGNDVMGXXXXXX 242
            T+ IFINQ               T GG ALKFY+SVRL++RR   +K+G +V+G
Sbjct: 175 NTLLIFINQ--------------TTGGNALKFYASVRLDIRRIGAVKEGENVVGSETRVK 220

Query: 243 XXXXXXAPPFRTAEVDIMYGEGI 265
                 A PF+ AE  I+YGEGI
Sbjct: 221 VVKNKIAAPFKQAEFQILYGEGI 243


>1u94A2 [d.48.1.1] RECA PROTEIN (1 A 269 328 )
```

BLAST(blastp)の出力例(3)

```
>1u94A2 [d.48.1.1] RECA PROTEIN (1 A 269 328 )
          Length = 60

 Score = 60.8 bits (146), Expect = 1e-10
 Identities = 23/54 (42%), Positives = 42/54 (77%)

Query: 269 GEIIDLGTELDIVQKSGSWYSYEEERLGQGRENAKQFLKENKDIMLMIQEQIRE 322
           GE++DLG +  +++K+G+WYSY+ E++GQG+ NA  +LK+N +    I++++RE
Sbjct: 4   GELVDLGVKEKLIEKAGAWYSYKGEKIGQGKANATAWLKDNPETAKEIEKKVRE 57


>1rypG [d.153.1.4] 20S PROTEASOME
          Length = 244

 Score = 29.3 bits (64), Expect = 0.36
 Identities = 13/37 (35%), Positives = 24/37 (64%)

Query: 275 GTELDIVQKSGSWYSYEEERLGQGRENAKQFLKENKD 311
           G  L +++ SGS++ Y+    G+GR++AK  L++  D
Sbjct: 141 GAHLYMLEPSGSYWGYKGAATGKGRQSAKAELEKLVD 177


>1p9rA [c.37.1.11] GENERAL SECRETION PATHWAY PROTEIN E
          Length = 378

 Score = 28.9 bits (63), Expect = 0.47
 Identities = 23/77 (29%), Positives = 36/77 (46%), Gaps = 3/77 (3%)

Query: 7   ALDMALKQIEKQFGKGSIMKLGEKTDTRISTVPSGSLAL--DTALGIGGYPRGRIIEVYG 64
           A+D+ +  +    G+  +M+L +K  TR+     G  A   D    +   P G II V G
Sbjct: 89  AVDVRVSTMPSSHGERVVMRLLDKNATRLDLHSLGMTAHNHDNFRRLIKRPHG-IILVTG 147
```

類似性の指標
どれだけ似ていれば意味があるのか？

- 同一残基率(Sequence Identity) [%]

直感的にわかりやすい。一般に30%ぐらいがしきい値とされる。感度が低く、アライメントの長さや不一致ペアの類似性に鈍感。

```
SLKA
*   *
SELA
```

4/8 = 50 %
Score = 4

```
SLKALLNKCKTFGWGAQ
*   **     **  *  **
SIRALDRRCKSFAWGKE
```

8/16 = 50 %
Score = 55

- スコア

同一残基率より感度は高いが、比較する配列の長さに依存。長いほど高いスコアになる。

- E-value

スコアの統計的有意性。

ランダムな配列を比較した場合に、そのスコアが生じる可能性を見積もる。

E-value

E-value (expectation value)

ランダムな配列データベースを検索したときに、
そのスコアS以上の値になるアライメントの本数の期待値

ランダムな配列とは：アミノ酸がランダムな順序に並んだ配列。ただし、
アミノ酸の組成　→　平均的な値に従うとする
アミノ酸の長さ　→　比較したアミノ酸の同じにする。

論理の流れ

ランダムな配列では起こりえないスコア
→　偶然では起こりえないスコア　→　進化的に関係がある類似性に違いない

値の大きさ

単位は本。小さいほどよく似ている。 必ず0以上の値になる。

しきい値

原理的には1。経験的には0.0001から0.01ぐらい。

E-valueの計算に必要なパラメータ

$$E(S) = Kmn \cdot e^{-\lambda S}$$

- パラメータ定数K, λ →スコア行列とギャップに依存
- m:クエリの残基長
- n:データベースの残基長

データベースに含まれる全ての配列を一つにつなげた場合の長さ

ビットスコアS'を以下のように定義すると、E-valueはより簡単な式で計算できる。

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad E(S') = mn \cdot 2^{-S'}$$

- クエリ配列長とデータベースの大きさにE-valueは比例
- 比較した配列が同じでも、データベースのほかの配列の数が変わると、E-valueも変わってしまう。

```
1ec7A1 [c.1.11.2] GLUCARATE DEHYDRATASE (1 A 138 446 )                  25   6.9
1otkA [a.25.1.2] PHENYLACETIC ACID DEGRADATION PROTEIN PAAC             25   9.0
```

BLAST(blastp)の出力例(2)

```
>1u94A1 [c.37.1.11] RECA PROTEIN (1 A 6 268 )
          Length = 243
```

ビットスコア　スコア

```
 Score =  259 bits (662), Expect = 2e-70
 Identities = 143/263 (54%), Positives = 176/263 (66%), Gaps = 21/263 (7%)

Query: 4   RQAALDMALKQIEKQFGKGSIMKLGEKTDTRISTVPSGSLALDTALGIGGYPRGRIIEVY 63
           +Q AL  AL QIEKQFGKGSIM+LGE     + T+ +GSL+LD ALG GG P GRI+E+Y
Sbjct: 1   KQKALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIALGAGGLPMGRIVEIY 60

Query: 64  GPESSGKTTVALHAIAEVQQQ-RTSAFIDAEHALDPVYAQKLGVNIEELLLSQPDTGEQA 122
           GPESSGKTT+ L  IA  Q++ +T AFIDAEHALDP+YA+KLGV+I+ LL SQPDTGEQA
Sbjct: 61  GPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTGEQA 120

Query: 123 LEIAEALVRSGXXXXXXXXXXXXXPKAEIEGDMGDSHVGLQARLMSQALRKLSGAINKS 182
           LEI +AL RSG               PKAEIEG+      GL AR+MSQA+RKL+G + +S
Sbjct: 121 LEICDALARSGAVDVIVVDSVAALTPKAEIEGE------GLAARMMSQAMRKLAGNLKQS 174

Query: 183 KTIAIFINQIREKVGVMFGNPETTPGGRALKFYSSVRLEVRRAEQLKQGNDVMGXXXXXX 242
            T+ IFINQ               T GG ALKFY+SVRL++RR   +K+G +V+G
Sbjct: 175 NTLLIFINQ--------------TTGGNALKFYASVRLDIRRIGAVKEGENVVGSETRVK 220

Query: 243 XXXXXXAPPFRTAEVDIMYGEGI 265
                 A PF+ AE  I+YGEGI
Sbjct: 221 VVKNKIAAPFKQAEFQILYGEGI 243


>1u94A2 [d.48.1.1] RECA PROTEIN (1 A 269 328 )
```

BLAST(blastp)の出力例(4)

```
  Database: 40scop1.75nm.fasta
    Posted date:  Sep 11, 2009  9:01 AM
  Number of letters in database: 1,701,902
  Number of sequences in database:  9671

Lambda     K      H
   0.314    0.133    0.364

Gapped
Lambda     K      H
   0.267   0.0410    0.140

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 9671
Number of Hits to DB: 995,144
Number of extensions: 36844
Number of successful extensions: 87
Number of sequences better than 10.0: 12
Number of HSP's gapped: 84
Number of HSP's successfully gapped: 12
Length of query: 347
Length of database: 1,701,902
Length adjustment: 84
Effective length of query: 263
Effective length of database: 889,538
Effective search space: 233948494
Effective search space used: 233948494
Neighboring words threshold: 11
Window for multiple hits: 40
X1: 16 ( 7.2 bits)
```

H20　問54

配列データベースに対して相同性検索を行ったとき、あるしきい値Xよりも高いスコアを持つヒットが何個くらい得られるかについてはKarlin-Altschulの理論がある。すなわち、ギャップ無しの局所アライメントに関しては、得られるヒット数の期待値Eは下式で与えられる。

$$E(S) = mn \cdot 2^{-S}$$

ただし、mは入力した問い合わせ配列の長さ、nはデータベース側の配列の全長、Sはしきい値Xをビットスコアと呼ばれるスコアに換算した値である。
ここで、長さ400残基の配列を、全長25億残基(2.5×10^9)のデータベースに対して、検索をしたとき、ビットスコアS=30以上のスコアのヒットはおおよそ何個得られるか。もっとも適切なものを選択肢の中から一つ選べ。ただし$\log_{10}2$=0.3010である。

1: およそ10個、2:およそ100個、3:およそ1000個、4:およそ10,000個

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20　問54

配列データベースに対して相同性検索を行ったとき、あるしきい値Xよりも高いスコアを持つヒットが何個くらい得られるかについてはKarlin-Altschulの理論がある。すなわち、ギャップ無しの局所アライメントに関しては、得られるヒット数の期待値Eは下式で与えられる。

$$E(S) = mn \cdot 2^{-S}$$

ただし、mは入力した問い合わせ配列の長さ、nはデータベース側の配列の全長、Sはしきい値Xをビットスコアと呼ばれるスコアに換算した値である。
ここで、長さ400残基の配列を、全長25億残基(2.5×10^9)のデータベースに対して、検索をしたとき、ビットスコアS=30以上のスコアのヒットはおおよそ何個得られるか。もっとも適切なものを選択肢の中から一つ選べ。ただし$\log_{10}2$=0.3010である。

1: およそ10個、2:およそ100個、3:およそ1000個、4:およそ10,000個

E-valueの公式に値を代入して計算していけばよい。
m=400, n=2.5x10^9、S=30を代入すると、
　E(S)=mn2^{-S}=4.0x10^2 x 2.5 x 10^9 x 2^{-30}=10x10^{11}x2^{-30}=10^{12}x2^{-30}
ここで、10の対数をとると以下のようになる。
　$\log_{10}$E(S)=$\log_{10}$(10^{12}x2^{-30})=12$\log_{10}$10-30$\log_{10}$2=12-30*0.3010=2.97
よって、E(S)は$10^{2.97} \fallingdotseq 10^3$=1000となる。

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19　問49

ある塩基配列に対してBLASTを用いて相同性検索を行った結果、Score=150, Expect=3e-20という結果が得られた。この結果の解釈としてもっとも適切なものを選択肢の中から一つ選べ。

1. 150以上のスコアが偶然に出る確率は、およそ$3x10^{-20}$である。
2. 150以下のスコアが偶然に出る確率は、およそ$3x10^{-20}$である。
3. 150以上のスコアが偶然に出る確率は、およそ$1-3x10^{-20}$である。
4. 150以下のスコアが偶然に出る確率は、およそ$1-3x10^{-20}$である。

⇨

1. 150以上のスコアの偶然に生じるアライメントの本数の期待値は、およそ$3x10^{-20}$である。
2. 150以下のスコアの偶然に生じるアライメントの本数の期待値は、およそ$3x10^{-20}$である。
3. 150以上のスコアの偶然に生じるアライメントの本数の期待値は、およそ$1-3x10^{-20}$である。
4. 150以下のスコアの偶然に生じるアライメントの本数の期待値は、およそ$1-3x10^{-20}$である。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19　問49

ある塩基配列に対してBLASTを用いて相同性検索を行った結果、Score=150, Expect=3e-20という結果が得られた。この結果の解釈としてもっとも適切なものを選択肢の中から一つ選べ。

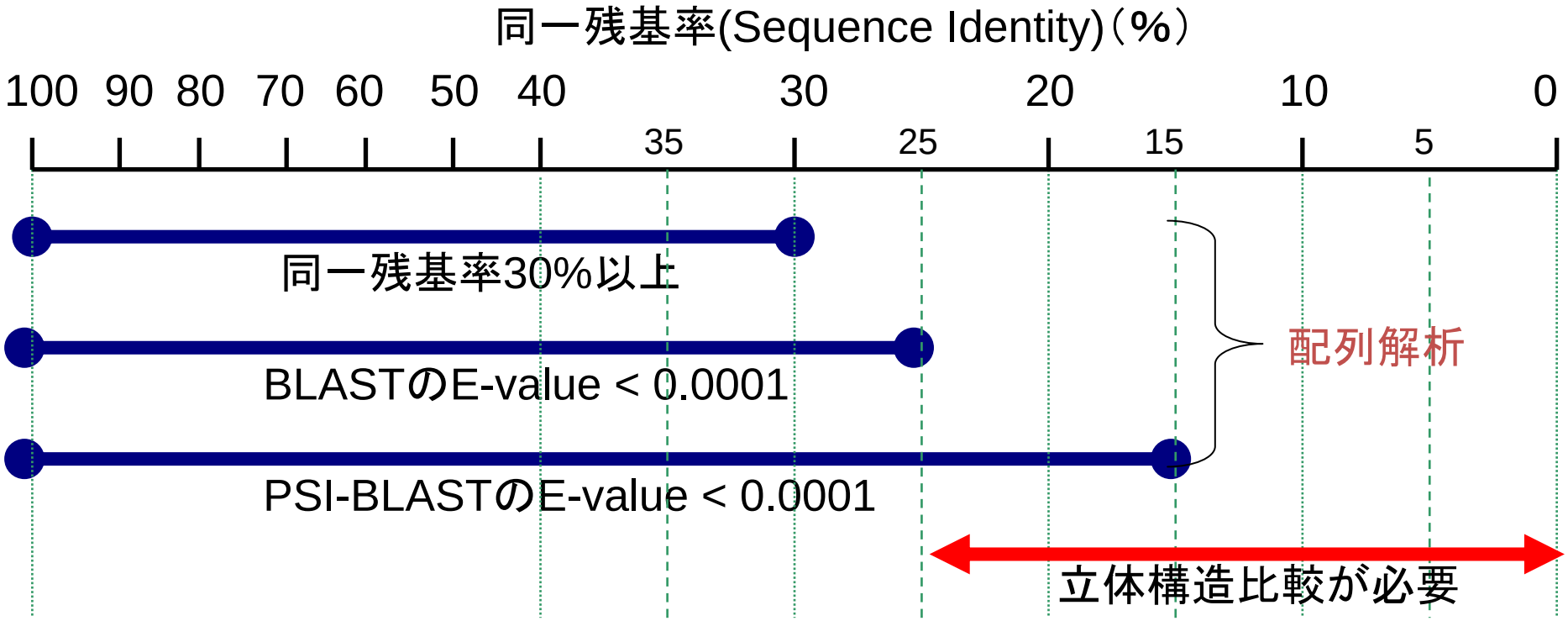
1. 150以上のスコアが偶然に出る確率は、およそ$3x10^{-20}$である。
2. 150以下のスコアが偶然に出る確率は、およそ$3x10^{-20}$である。
3. 150以上のスコアが偶然に出る確率は、およそ$1-3x10^{-20}$である。
4. 150以下のスコアが偶然に出る確率は、およそ$1-3x10^{-20}$である。

⇨

1. 150以上のスコアの偶然に生じるアライメントの本数の期待値は、およそ$3x10^{-20}$である。
2. 150以下のスコアの偶然に生じるアライメントの本数の期待値は、およそ$3x10^{-20}$である。
3. 150以上のスコアの偶然に生じるアライメントの本数の期待値は、およそ$1-3x10^{-20}$である。
4. 150以下のスコアの偶然に生じるアライメントの本数の期待値は、およそ$1-3x10^{-20}$である。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

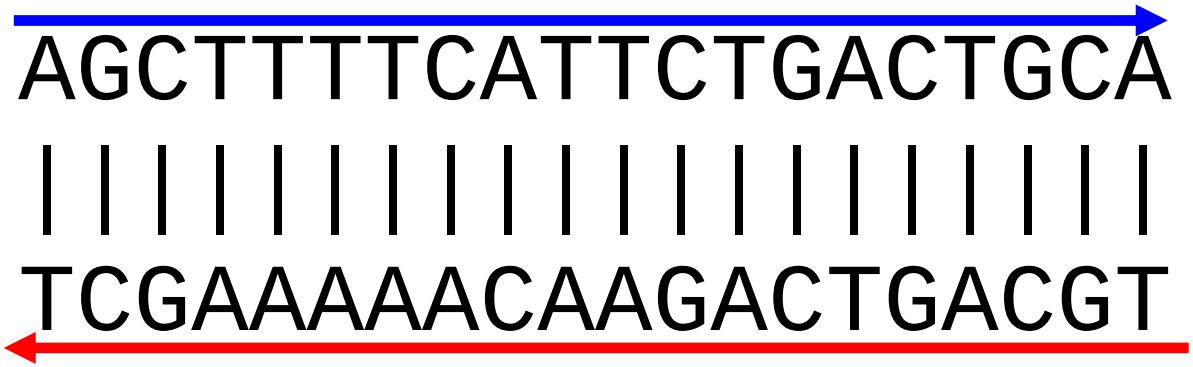
タンパク質の相同性の判断基準



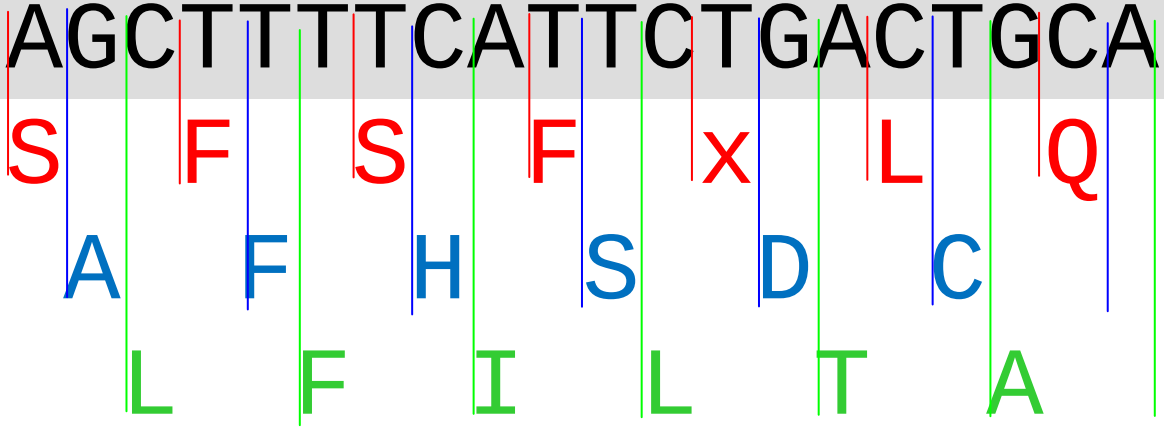
BLASTのプログラムの種類

	クエリ配列	データベース配列	比較回数	典型的な使用目的
blastn	核酸	核酸	2回 相補鎖にしたDB配列とも比較	ゲノムDNAのアノテーション、cDNAのゲノムへのマッピング、非コーディング領域の比較
blastp	アミノ酸	アミノ酸	1回	タンパク質配列からの比較的遠縁のホモログの発見
blastx	核酸（を翻訳したアミノ酸）	アミノ酸	6回 クエリから6通りのアミノ酸配列を生成して比較	ゲノムDNAから遺伝子（タンパク質をコードしている領域）を発見する
tblastn	アミノ酸	核酸（を翻訳したアミノ酸）	6回 クエリから6通りのアミノ酸配列を生成して比較	あるタンパク質をコードしているゲノムの領域を発見する
tblastx	核酸（を翻訳したアミノ酸）	核酸（を翻訳したアミノ酸）	36回 クエリ、DBとも6通りのアミノ酸配列を生成して比較	やや遠縁の生物種のゲノムを、その中にコードされたタンパク質で比較。DBに登録されていない遺伝子の発見を期待。

DNAには相補鎖があり、それぞれ3つのアミノ酸の読み枠がある



DNAは二重らせん構造を
作っているため、
A⇔T、G⇔Cに入れ替えて、
向きを逆にした相補鎖が
あるはず。



3つの核酸が1つのアミノ酸に
翻訳されるので、読み枠を
ずらせば一本の核酸配列から
3本のアミノ酸配列を
作ることができる

※核酸よりアミノ酸で比較したほうがより遠縁のホモログを認識可能

H19　問48

相同性検索に用いられるBLASTには、クエリ配列と対象データベースのデータの種類によって使い分けられるいくつかの異なるバージョンがある。BLASTに含まれるblastnプログラムでの、クエリ配列と対象データベースの組み合わせは、どのようなものか。適しているものを選択肢の中から一つ選べ。

	クエリ配列	対象データベース
1	DNA配列	DNA配列
2	DNA配列	タンパク質（アミノ酸）
3	タンパク質（アミノ酸）	タンパク質（アミノ酸）
4	タンパク質（アミノ酸）	DNA配列

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19　問48

相同性検索に用いられるBLASTには、クエリ配列と対象データベースのデータの種類によって使い分けられるいくつかの異なるバージョンがある。BLASTに含まれるblastnプログラムでの、クエリ配列と対象データベースの組み合わせは、どのようなものか。適しているものを選択肢の中から一つ選べ。

	クエリ配列	対象データベース
1	DNA配列	DNA配列
2	DNA配列	タンパク質（アミノ酸）
3	タンパク質（アミノ酸）	タンパク質（アミノ酸）
4	タンパク質（アミノ酸）	DNA配列

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

「配列解析」のキーワード（プロフィール法）

- 位置特異的スコア行列(PSSM)
- プロファイル比較
- HMM（隠れマルコフモデル）
- モチーフ解析（正規表現、重み行列）

モチーフ･プロフィールを用いた類似性

→相同な配列群のマルチプルアライメントから、このファミリーに特徴的なパターンを見出したい

```
5p21-   MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSY
1ctqA   MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSY
1c1yA   MREYKLVVLGSGGVGKSALTVQFVQGIFVEKYDPTIEDSY
1kao-   MREYKVVVLGSGGVGKSALTVQFVTGTFIEKYDPTIEDFY
1huqA   --QFKLVLLGESAVGKSSLVLRFVKGQFHEYQESTIGAAF
1g16A   ----KILLIGDSGVGKSCLLVRFVE----DKFNPI--DFK
1ek0A   VTSIKLVLLGEAAVGKSSIVLRFVSNDFAENKEPTIGAAF
3rabA   ---FKILIIGNSSVGKTSFLFRYADDSFTPAFVSTVGIDF
1mh1-   ----KCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNY
2ngrA   MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNY
1tx4B   ----KLVIVGDGACGKTCLLIVNSKDQF---YVPTVFENY
```

サイトごとに保存の度合いに差がある。
サイトごとにアミノ酸の出現傾向に差がある

[AG]-x(4)-G-K-[ST]

モチーフ解析

- 正規表現風のパターンで、局所的な配列のパターンを表現。

  PROSITE(http://www.expasy.ch/prosite/)が有名

1.進化的に保存している局所配列パターン

・マルチプルアライメント由来

・保存しているサイト→機能的に重要なサイト→活性部位

2.機能的な局所配列パターン

・リン酸化サイト、N-ミリストイル化サイトなど

PROSITEのモチーフの記述法

（例）

ATP_GTP_A : [AG]-x(4)-G-K-[ST]

2FE2S FERREDOXIN:

C-{C}-{C}-[GA]-{C}-C-[GAST]-{CPDEKRHFYW}-C

ZINC_FINGER_C2H2_1:

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

x	：任意のアミノ酸
x(n)	：n個の任意のアミノ酸
x(n,m)	：nからm個の任意のアミノ酸
[ACD]	：AかCかDのいずれかのアミノ酸
{ACD}	：AでもCでもDでもないアミノ酸

x :任意のアミノ酸
x(n) :n個の任意のアミノ酸
x(n,m):nからm個の任意のアミノ酸
[ACD] :AかCかDのいずれかのアミノ酸
{ACD} :AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P C Q L C G R S P L G E A P P G T P P
C R L C C P A T A P Q E A P A P E G R A L G E E E E E E P P R A G E G R P
A G R E E E E E E E E E E G T Y H C T E C E D S F D N L G E L H G H F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x ：任意のアミノ酸
x(n) ：nの任意のアミノ酸
x(n,m)：nからm個の任意のアミノ酸
[ACD] ：AかCかDのいずれかのアミノ酸
{ACD} ：AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P C Q L C G R S P L G E A P P G T P P
C R L C C P A T A P Q E A P A P E G R A L G E E E E E E P P R A G E G R P
A G R E E E E E E E E E E G T Y H C T E C E D S F D N L G E L H G H F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x ：任意のアミノ酸
x(n) ：n個の任意のアミノ酸
x(n,m)：nからm個の任意のアミノ酸
[ACD] ：AかCかDのいずれかのアミノ酸
{ACD} ：AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P **C** Q L C G R S P L G E A P P G T P P
C R L **C** **C** P A T A P Q E A P A P E G R A L G E E E E E E P P R A G E G R P
A G R E E E E E E E E E E G T Y **H** **C** T E **C** E D S F D N L G E L **H** G **H** F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x ：任意のアミノ酸
x(n) ：nの任意のアミノ酸
x(n,m)：nからm個の任意のアミノ酸
[ACD] ：AかCかDのいずれかのアミノ酸
{ACD} ：AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P C Q L C G R S P L G E A P P G T P P
C R L C C P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E G T Y H C T E C E D S F D N L G E L H G H F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x ：任意のアミノ酸
x(n) ：nの任意のアミノ酸
x(n,m)：nからm個の任意のアミノ酸
[ACD] ：AかCかDのいずれかのアミノ酸
{ACD} ：AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P C Q L C G R S P L G E A P P G T P P
C R L C C P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E G T Y H C T E C E D S F D N L G E L H G H F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x ：任意のアミノ酸
x(n) ：nの任意のアミノ酸
x(n,m)：nからm個の任意のアミノ酸
[ACD] ：AかCかDのいずれかのアミノ酸
{ACD} ：AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P C Q L C G R S P L G E A P P G T P P
C R L C C P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E G T Y H C T E C E D S F D N L G E L H G H F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

H19 問54

塩基配列やアミノ酸配列において、特定の機能を持った配列は進化の過程で多少の変化を起こしながらも種間で保存されている。このような配列をモチーフ配列と呼びパターンの表現方法の一つには正規表現がある。次に示した正規表現で表わされるアミノ酸配列として適切なものを選択肢の中から一つ選べ。

正規表現: **C-x(2,4)-C-[LIV]-H**

ここで、正規表現の記号の意味は次の通りである。

[ ]は、[ ]内に並べられた文字のうちいずれか1文字が選択される。

x(a,b)は、任意の文字がa個以上b個以下挿入されることを表す。

-は文字の連結を表す。

1: **CPKRLH**

2: **CPKRCLVH**

3: **CPKRGCIH**

4: **CPKRGKCVH**

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19 問54

塩基配列やアミノ酸配列において、特定の機能を持った配列は進化の過程で多少の変化を起こしながらも種間で保存されている。このような配列をモチーフ配列と呼びパターンの表現方法の一つには正規表現がある。次に示した正規表現で表わされるアミノ酸配列として適切なものを選択肢の中から一つ選べ。

正規表現： **C-x(2,4)-C-[LIV]-H**

ここで、正規表現の記号の意味は次の通りである。

[ ]は、[ ]内に並べられた文字のうちいずれか1文字が選択される。

x(a,b)は、任意の文字がa個以上b個以下挿入されることを表す。

-は文字の連結を表す。

1: **CPKRLH**

2: **CPKRCLVH**

3: **CPKRGCIH**

4: **CPKRGKCVH**

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

ProSiteモチーフの問題点

False positiveが多く、ファミリの認識能力は高くない。

```
          [AG]-x(4)-G-K-[ST]
5p21-   MTEYKLVVVGAGGVGKSAL
1ctqA   MTEYKLVVVGAGGVGKSAL
1c1yA   MREYKLVVLGSGGVGKSAL
1kao-   MREYKVVVLGSGGVGKSAL
1huqA   --QFKLVLLGESAVGKSSL
1g16A   ----KILLIGDSGVGKSCL
1ek0A   VTSIKLVLLGEAAVGKSSI
3rabA   ---FKILIIGNSSVGKTSF
1mh1-   ----KCVVVGDGAVGKTCL
2ngrA   MQTIKCVVVGDGAVGKTCL
1tx4B   ----KLVIVGDGACGKTCL
1i2mA   --QFKLVLVGDGGTGKTTF
2efgA   -RLRNIGIAAHIDAGKTTT
            . . .  ...
           .
```

1. パターンの表現能力の限界
2. 客観的にパターンを生成するのが難しい。
3. もっと大域的な領域も淡く似ているはず

プロフィール法

プロフィール法

マルチプルアライメントから**サイトごとのスコア行列**を作成。
これに対して動的計画法等を用いて配列をアライメント。

サイトごとのスコア行列
↓
プロフィール(Profile)
位置特異的スコア行列
（PSSM; Position Specific Score Matrix)

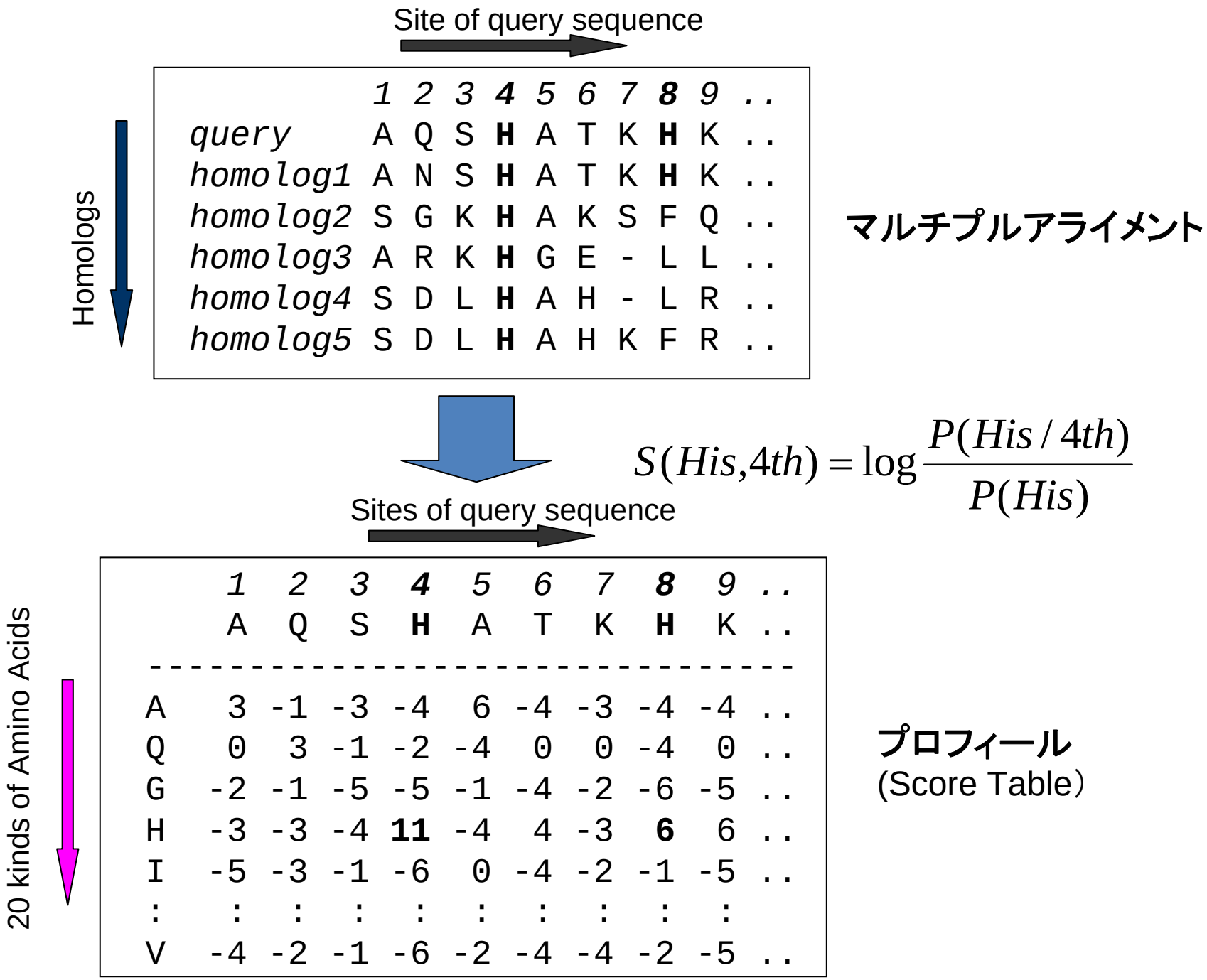
	1	*2*	*3*	***4***	*5*	*6*	*..*
A	3	-1	-3	-4	6	-4	..
Q	0	3	-1	-2	-4	0	..
H	-3	-3	-4	**11**	-4	4	..
:	:	:	:	:	:	:	
V	-4	-2	-1	-6	-2	-4	..

HMMer

マルチプルアライメントを入力とする。隠れマルコフモデル(HMM)を使用しているため、表現力はPSI-BLASTより高いはずだが、計算速度は遅い。PfamはHMMerを採用している。

PSI-BLAST

BLASTの拡張版。反復的にデータベース検索を行うことで、厚いマルチプルアライメントを生成する。


Site of query sequence
Homologs
1 2 3 4 5 6 7 8 9 ..
query A Q S H A T K H K ..
homolog1 A N S H A T K H K ..
homolog2 S G K H A K S F Q ..
homolog3 A R K H G E - L L ..
homolog4 S D L H A H - L R ..
homolog5 S D L H A H K F R ..
マルチプルアライメント
$S(His,4th) = \log \frac{P(His/4th)}{P(His)}$
Sites of query sequence
20 kinds of Amino Acids
1 2 3 4 5 6 7 8 9 ..
A Q S H A T K H K ..
A 3 -1 -3 -4 6 -4 -3 -4 -4 ..
Q 0 3 -1 -2 -4 0 0 -4 0 ..
G -2 -1 -5 -5 -1 -4 -2 -6 -5 ..
H -3 -3 -4 11 -4 4 -3 6 6 ..
I -5 -3 -1 -6 0 -4 -2 -1 -5 ..
V -4 -2 -1 -6 -2 -4 -4 -2 -5 ..
プロフィール
(Score Table)

位置特異的スコア行列
Position Specific Score Matrix ; PSSM

$$S_i(a) = \log \frac{p_i(a)}{q(a)}$$

$p_i(a)$:i番目のサイトのアミノ酸aの確率

$q(a)$:アミノ酸aの背景確率(background probability)

※ $S_i(a)>0.0$（ $p_i(a)>q(a)$ ）のとき、このファミリーに属することを示唆

$Si(a)<0.0$（ $p_i(a)<q(a)$ ）のとき、このファミリーに属さないことを示唆

※ $p_i(a)=0$だと$S_i(a)=-\infty$ になってしまう。すべてのaについて$p_i(a)>0$となるような補正が必ず必要。

PSSMの計算例

$$S_i(a) = \log \frac{p_i(a)}{q(a)}$$

（5） 以下の10本の配列からなるマルチプルアライメントから計算されたサイトごとの確率$p_i(a)$を用いて、対数オッズスコアのPSSM $S_i(a)=\log_2(p_i(a)/q(a))$を求めよ。$q(a)$は5種のアミノ酸で同じ値$q(a)=1/5=0.2$とする。また、空欄の$p_i(a)$は$p_i(a)=0$とすること。

$\log_2(0.1/0.2)=\log_2(1/2)=$ -1.0, $\log_2(0.2/0.2)=\log_2(1)=0.0$
$\log_2(0.3/0.2)=\log_2(3/2)=$ 0.6, $\log_2(0.4/0.2)=\log_2(2)=1.0$,
$\log_2(0.6/0.2)=\log_2(3)=$ 1.6, $\log_2(0.8/0.2)=\log_2(4)=2.0$,
$\log_2(1.0/0.2)=\log_2(5)=$ 2.3, $\log_2(0)=$ -∞とする。

i	配列	確率 $p_i(a)$					PSSM $S_i(a)=\log_2(p_i(a)/q(a))$				
		A	E	G	H	L	A	E	G	H	L
1	**HHHHHHHEHE**		***0.2***		***0.8***		−∞	***0.0***	−∞	***2.0***	−∞
2	**AAHGHGLLEE**										
3	**HHHHHHHHHH**										
4	**HLLHLHLHHH**										
5	**HLLEHLHHHH**										
6	**AHAAHAHGHG**										
7	**AEHAHEHHGL**										

PSSMの計算例

$$S_i(a) = \log \frac{p_i(a)}{q(a)}$$

（5） 以下の10本の配列からなるマルチプルアライメントから計算されたサイトごとの確率$p_i(a)$を用いて、対数オッズスコアのPSSM $S_i(a)=\log_2(p_i(a)/q(a))$を求めよ。$q(a)$は5種のアミノ酸で同じ値$q(a)=1/5=0.2$とする。また、空欄の$p_i(a)$は$p_i(a)=0$とすること。

$\log_2(0.1/0.2)=\log_2(1/2)=$ -1.0, $\log_2(0.2/0.2)=\log_2(1)=0.0$
$\log_2(0.3/0.2)=\log_2(3/2)=$ 0.6, $\log_2(0.4/0.2)=\log_2(2)=1.0$,
$\log_2(0.6/0.2)=\log_2(3)$ = 1.6, $\log_2(0.8/0.2)=\log_2(4)=2.0$,
$\log_2(1.0/0.2)=\log_2(5)$ = 2.3, $\log_2(0)=$ -∞とする。

i	配列	確率 $p_i(a)$					PSSM $S_i(a)=\log_2(p_i(a)/q(a))$				
		A	E	G	H	L	A	E	G	H	L
1	**HHHHHHHEHE**		**0.2**		**0.8**		−∞	**0.0**	−∞	**2.0**	−∞
2	**AAHGHGLLEE**	**0.2**	**0.2**	**0.2**	**0.2**	**0.2**					
3	**HHHHHHHHHH**				**1.0**						
4	**HLLHLHLHHH**				**0.6**	**0.4**					
5	**HLLEHLHHHH**		**0.1**		**0.6**	**0.3**					
6	**AHAAHAHGHG**	**0.4**		**0.2**	**0.4**						
7	**AEHAHEHHGL**	**0.2**	**0.2**	**0.1**	**0.4**	**0.1**					

PSSMの計算例

$$S_i(a) = \log \frac{p_i(a)}{q(a)}$$

（5） 以下の10本の配列からなるマルチプルアライメントから計算されたサイトごとの確率$p_i(a)$を用いて、対数オッズスコアのPSSM $S_i(a)=\log_2(p_i(a)/q(a))$を求めよ。$q(a)$は5種のアミノ酸で同じ値$q(a)=1/5=0.2$とする。また、空欄の$p_i(a)$は$p_i(a)=0$とすること。

$\log_2(0.1/0.2)=\log_2(1/2)=$ -1.0, $\log_2(0.2/0.2)=\log_2(1)=0.0$
$\log_2(0.3/0.2)=\log_2(3/2)=$ 0.6, $\log_2(0.4/0.2)=\log_2(2)=1.0$,
$\log_2(0.6/0.2)=\log_2(3)$ = 1.6, $\log_2(0.8/0.2)=\log_2(4)=2.0$,
$\log_2(1.0/0.2)=\log_2(5)$ = 2.3, $\log_2(0)=$ -∞とする。

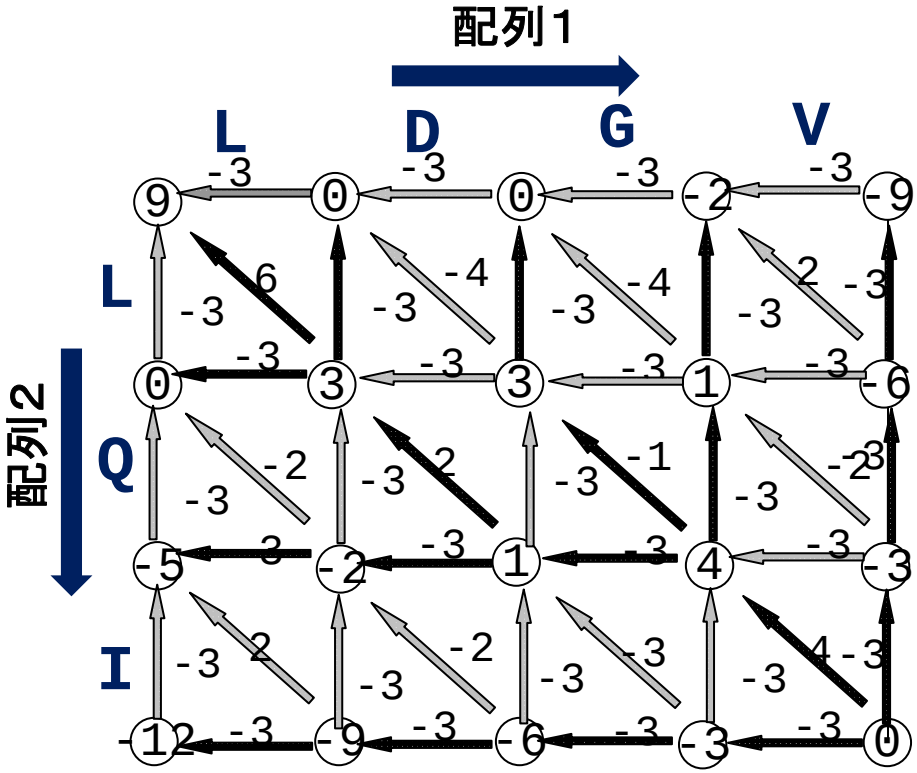
i	配列	確率 $p_i(a)$					PSSM $S_i(a)=\log_2(p_i(a)/q(a))$				
		A	E	G	H	L	A	E	G	H	L
1	**HHHHHHHEHE**		**0.2**		**0.8**		−∞	**0.0**	−∞	**2.0**	−∞
2	**AAHGHGLLEE**	**0.2**	**0.2**	**0.2**	**0.2**	**0.2**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
3	**HHHHHHHHHH**				**1.0**		−∞	−∞	−∞	**2.3**	−∞
4	**HLLHLHLHHH**				**0.6**	**0.4**	−∞	−∞	−∞	**1.6**	**1.0**
5	**HLLEHLHHHH**		**0.1**		**0.6**	**0.3**	−∞	**-1.0**	−∞	**1.6**	**0.6**
6	**AHAAHAHGHG**	**0.4**		**0.2**	**0.4**		**1.0**	−∞	**0.0**	**1.0**	−∞
7	**AEHAHEHHGL**	**0.2**	**0.2**	**0.1**	**0.4**	**0.1**	**0.0**	**0.0**	**-1.0**	**3.0**	**-1.0**

BLOSUM62（blastpのデフォルトで使われている置換スコア行列）

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

動的計画法によるアライメント

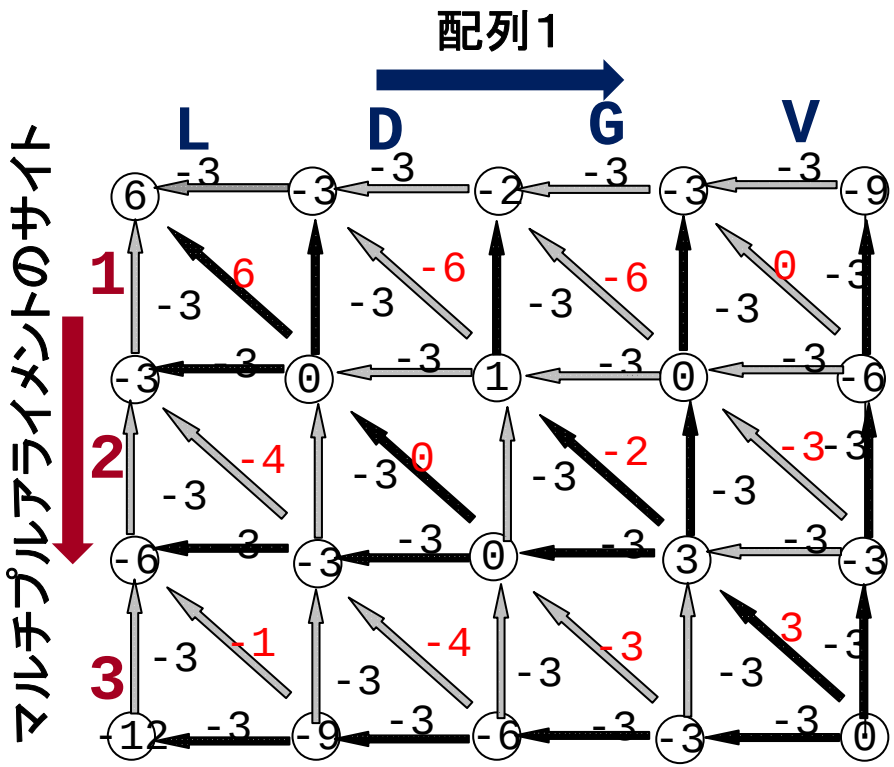
通常のペアワイズアライメント



LDGV

LQ-I

PSSMを用いたアライメント



LDGV

12-3

PSI-BLASTにより計算されたアミノ酸頻度

Myoglobin (1a6m/MYG_PHYCA、クジラ)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	V	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	75
2	L	0	0	0	0	0	0	0	0	0	0	95	0	1	4	0	0	0	0	0	0
3	S	1	0	2	0	0	0	0	1	0	0	0	0	1	0	0	61	34	0	0	0
4	E	35	0	0	32	0	1	8	4	1	0	0	2	0	0	8	8	1	0	0	0
5	G	30	0	2	6	0	5	19	14	1	0	0	12	0	0	0	4	5	0	0	1
6	E	0	0	2	43	0	0	54	0	0	0	0	0	0	0	0	0	0	0	0	0
7	W	0	14	0	0	1	0	1	0	0	0	0	61	2	4	0	0	1	15	0	1
8	Q	24	0	4	6	1	16	7	2	1	0	0	12	0	0	0	11	16	0	0	1
9	L	21	2	15	0	0	4	1	0	6	4	26	6	3	1	0	0	6	0	0	5
10	V	0	0	0	0	0	0	0	0	0	42	1	0	0	0	0	0	0	0	0	57
11	L	3	8	6	0	1	7	0	1	0	2	20	28	0	0	0	3	18	0	0	4
12	H	22	0	8	6	0	7	0	11	**5**	0	1	9	0	0	0	19	10	0	0	0
:																					
24	H	2	0	3	1	4	0	0	1	**16**	18	7	0	2	7	4	2	4	0	19	10
:																					
36	H	0	0	1	0	1	0	0	0	**20**	0	0	0	0	24	0	1	3	0	50	0
:																					
64	H	0	0	0	0	0	5	0	1	**92**	0	0	0	0	0	0	0	0	0	0	2
:																					
93	H	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0

PSI-BLASTにより計算されたスコア

Myoglobin (1a6m/MYG_PHYCA、クジラ)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	V	-2	-4	-4	-5	-2	-3	-4	-5	-4	1	0	-3	5	-2	-4	-3	-2	-4	-3	6
2	L	-4	-4	-6	-6	-3	-4	-5	-6	-5	0	6	-5	1	1	-5	-5	-3	-4	-3	-1
3	S	-1	-3	-1	-3	-3	-2	-3	-2	-3	-4	-4	-3	-2	-5	-3	5	5	-5	-4	-3
4	E	4	-3	-2	5	-4	-2	1	-1	-1	-5	-5	-1	-3	-5	1	1	-2	-5	-4	-4
5	G	3	-3	-1	1	-4	1	3	2	-1	-5	-5	2	-4	-5	-3	0	0	-5	-4	-3
6	E	-4	-3	0	6	-6	-1	6	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
7	W	-3	3	-3	-4	-2	-1	-2	-4	-3	-5	-4	6	-1	0	-4	-3	-3	7	-3	-3
8	Q	3	-2	0	0	-1	3	1	-2	-2	-4	-4	2	-3	-5	-3	1	2	-5	-4	-3
9	L	2	-2	3	-4	-2	0	-2	-4	2	0	2	0	1	-2	-4	-2	0	-5	-3	0
10	V	-3	-5	-6	-6	-3	-5	-5	-6	-6	5	-1	-5	-1	-3	-5	-4	-3	-5	-4	6
11	L	-1	1	1	-3	-2	1	-3	-3	-3	-2	2	4	-2	-4	-4	-1	3	-5	-4	-1
12	H	3	-2	2	0	-3	1	-2	1	**2**	-4	-4	1	-4	-5	-4	3	1	-5	-4	-3
:																					
24	H	-2	-4	-1	-4	2	-3	-4	-4	**5**	3	0	-4	0	2	-1	-2	-1	-2	5	1
:																					
36	H	-4	-4	-2	-5	-3	-3	-4	-5	**6**	-4	-3	-4	-3	5	-5	-3	-1	-1	7	-4
:																					
64	H	-4	-2	-2	-3	-5	1	-2	-3	**10**	-5	-5	-3	-4	-4	-5	-3	-4	-5	-1	-3
:																					
93	H	-4	-2	-2	-3	-5	-2	-2	-4	**11**	-6	-5	-3	-4	-4	-5	-3	-4	-5	0	-6

BLASTにより発見されたホモログ

Myoglobin (1a6m/MYG_PHYCA、クジラ)をクエリとしてPDBを検索

```
BLASTP 2.2.16 [Mar-25-2007]
Query= 1a6mAA  (151 letters)
Database: 40pdb09Jan8
```

Sequences producing significant alignments:	Score (bits)	E Value
*2nrlA [a.1.1 (101mA)] MYOGLOBIN	114	4e-27
*2dc3A [a.1.1 (1umoA)] CYTOGLOBIN	85	4e-18
*1irdA [a.1.1] HEMOGLOBIN ALPHA CHAIN	46	2e-06
*1c7cA [a.1.1 - a.1.1] PROTEIN (DEOXYHEMOGLOBIN (ALPHA CHAIN))	46	2e-06
*1it2A [a.1.1] HEMOGLOBIN	44	6e-06
*1mbaA [a.1.1] MYOGLOBIN	40	1e-04
*1x3kA [x.x.x] HEMOGLOBIN COMPONENT V	37	0.001
1hlbA [a.1.1] HEMOGLOBIN (DEOXY)	35	0.003
2c0kA [x.x.x] HEMOGLOBIN	35	0.004
2z8aA [a.1.1 (1hbiA)] GLOBIN-1	34	0.006
2olpA [x.x.x] HEMOGLOBIN II	32	0.024
1x46A [x.x.x] HEMOGLOBIN COMPONENT VII	32	0.031
2bk9A [x.x.x] CG9734-PA	27	0.99
1un7A [b.92.1 - c.1.9] N-ACETYLGLUCOSAMINE-6-PHOSPHATE DEACETYLASE	27	1.3
1zx5A [b.82.1] MANNOSEPHOSPHATE ISOMERASE, PUTATIVE	26	2.2
1nh1A [e.45.1] AVIRULENCE B PROTEIN	26	2.2
1q1fA [a.1.1] NEUROGLOBIN	25	2.9
2dy1A [c.37.1 - b.43.3 - d.58.11 - d.14.1 - d.58.11 (1wdtA)] ELO...	25	2.9
1b0bA [a.1.1] HEMOGLOBIN	25	3.8
1vbiA [x.x.x] TYPE 2 MALATE/LACTATE DEHYDROGENASE	24	6.4
2rd9A [x.x.x] BH0186 PROTEIN	24	6.4

PSI-BLASTにより発見されたホモログ

Myoglobin (1a6m/MYG_PHYCA、クジラ)をクエリとしてPDBを検索

```
BLASTP 2.2.16 [Mar-25-2007]
Query= 1a6mAA         (151 letters)
Database: 40pdb09Jan8
```

Sequences producing significant alignments:	Score (bits)	E Value
1c7cA [a.1.1 - a.1.1] PROTEIN (DEOXYHEMOGLOBIN (ALPHA CHAIN))	222	9e-60
1irdA [a.1.1] HEMOGLOBIN ALPHA CHAIN	222	1e-59
2dc3A [a.1.1 (1umoA)] CYTOGLOBIN	169	1e-43
2nrlA [a.1.1 (101mA)] MYOGLOBIN	156	8e-40
1it2A [a.1.1] HEMOGLOBIN	111	5e-26
***1cg5B [a.1.1] PROTEIN (HEMOGLOBIN)**	**103**	**8e-24**
***1hlbA [a.1.1] HEMOGLOBIN (DEOXY)**	**66**	**2e-12**
***2c0kA [x.x.x] HEMOGLOBIN**	**57**	**7e-10**
***1q1fA [a.1.1] NEUROGLOBIN**	**53**	**2e-08**
1x3kA [x.x.x] HEMOGLOBIN COMPONENT V	51	5e-08
***2z8aA [a.1.1 (1hbiA)] GLOBIN-1**	**51**	**5e-08**
1mbaA [a.1.1] MYOGLOBIN	50	1e-07
***2olpA [x.x.x] HEMOGLOBIN II**	**49**	**2e-07**
***2bk9A [x.x.x] CG9734-PA**	**49**	**3e-07**
***1jf3A [a.1.1] MONOMER HEMOGLOBIN COMPONENT III**	**48**	**4e-07**
***1x46A [x.x.x] HEMOGLOBIN COMPONENT VII**	**45**	**3e-06**
***1gdjA [a.1.1] LEGHEMOGLOBIN (DEOXY)**	**41**	**6e-05**
***2zs0C [a.1.1 (1x9fA)] EXTRACELLULAR GIANT HEMOGLOBIN MAJOR GLOBIN**	**40**	**1e-04**
***1b0bA [a.1.1] HEMOGLOBIN**	**39**	**2e-04**
***1cqxA [a.1.1 - b.43.4 - c.25.1] FLAVOHEMOPROTEIN**	**38**	**6e-04**
1ecaA [a.1.1] ERYTHROCRUORIN (AQUO MET)	35	0.004

BLASTにより発見されたホモログ

```
>1x3kA [x.x.x] HEMOGLOBIN COMPONENT V ユスリカのヘモグロビン Length = 152
 Score = 37.0 bits (84), Expect = 0.001
 Identities = 24/102 (23%), Positives = 42/102 (41%), Gaps = 1/102 (0%)

Query: 2   LSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDL 61
           LS+ E +LV   WA +  D+ G    +     K +P   +KF+  K    + E+K + +
Sbjct: 5   LSDSEEKLVRDAWAPIHGDLQGTANTVFYNYLKKYPSNQDKFETLKGHPLD-EVKDTANF 63

Query: 62  KKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKY 103
           K     + T     +K  G+ +   K +A          PI +
Sbjct: 64  KLIAGRIFTIFDNCVKNVGNDKGFQKVIADMSGPHVARPITH 105
```

PSI-BLASTにより発見されたホモログ

```
>1cqxA [a.1.1 - b.43.4 - c.25.1] FLAVOHEMOPROTEIN 微生物のフラボヘム蛋白質 Length = 403
 Score = 37.6 bits (87), Expect = 6e-04,   Method: Composition-based stats.
 Identities = 26/148 (17%), Positives = 51/148 (34%), Gaps = 21/148 (14%)

Query: 1   VLSEGEWQLVLHVWAKVEADVAGHGQDIL----IRLFKSHPETLEKF--DRFKHLKTEAE 54
           +L++    +V    A V   +A HG DI+     R+F++HPE    F     +  + +
Sbjct: 1   MLTQKTKDIVKAT-APV---LAEHGYDIIKCFYQRMFEAHPELKNVFNMAHQEQGQQQA 56

Query: 55  MKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHV 114
           +              + A    ++      A LK +A  HA    +  +    + E ++
Sbjct: 57  L----------ARAVYAYAENIEDPNSLMAVLKNIANKHA-SLGVKPEQYPIVGEHLLAA 105

Query: 115 LHSRHPGDFGADAQGAMNKALELFRKDI 142
           +          D   A  +A       +
Sbjct: 106 IKEVLGNAATDDIISAWAQAYGNLADVL 133
```

マルチプルアライメント

```
          1 2 3 4 5 6 7 8 9 ..
query     A Q S H A T K H K ..
homolog1  A N S H A T K H K ..
homolog2  S G K H A K S F Q ..
homolog3  A R K H G E - L L ..
homolog4  S D L H A H - L R ..
```

良質のマルチプルアライメントを作るには淡い相同性の配列を集め、アラインする必要がある。それには、よいプロフィール
が不可欠

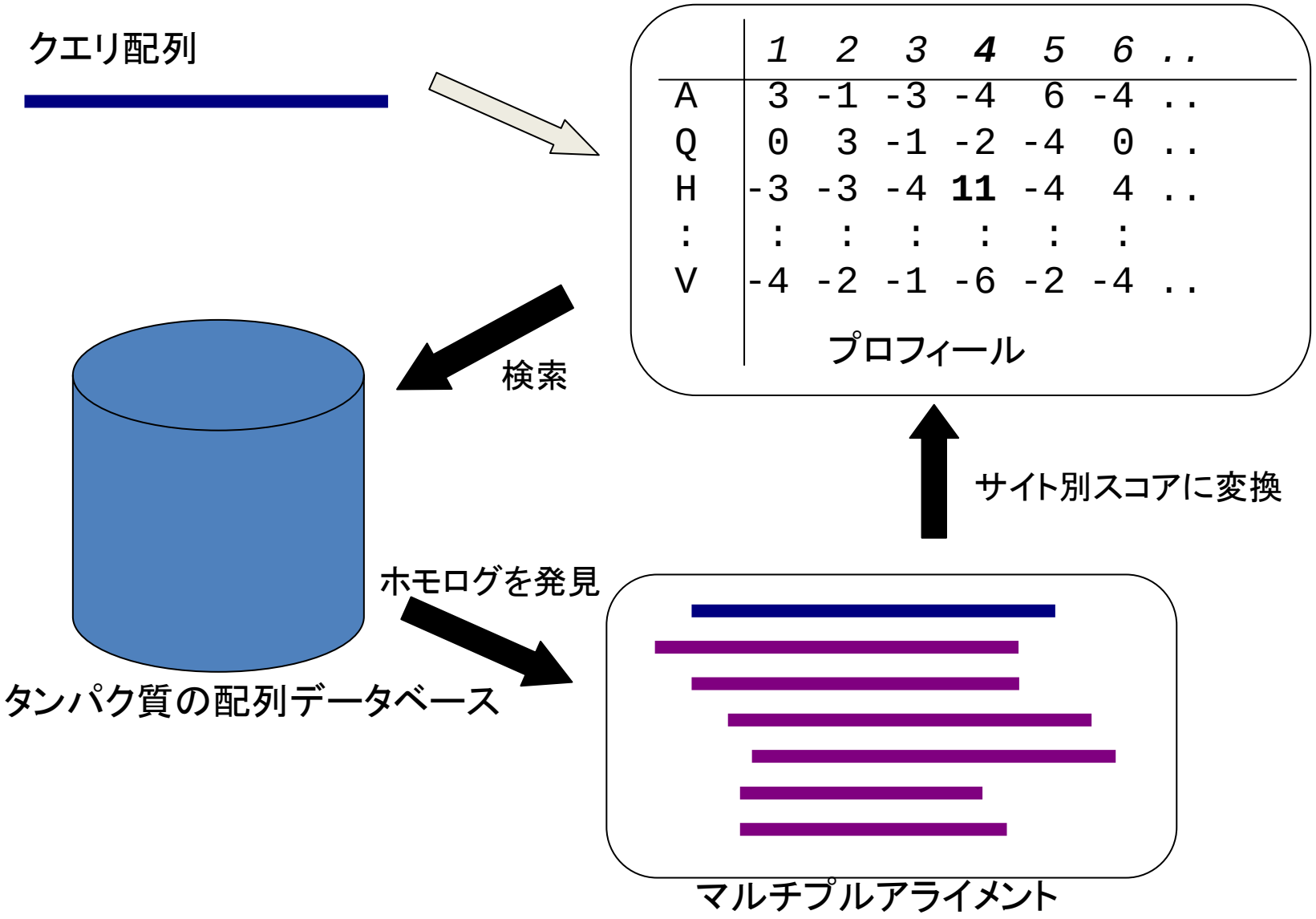
プロフィール

```
    1  2  3  4  5  6  7  8  ..
    A  Q  S  H  A  T  K  H  ..
-----------------------------
A   3 -1 -3 -4  6 -4 -3 -4  ..
G  -2 -1 -5 -5 -1 -4 -2 -6  ..
H  -3 -3 -4 11 -4  4 -3  6  ..
:   :  :  :  :  :  :  :  :
V  -4 -2 -1 -6 -2 -4 -4 -2  ..
```

良質のプロフィールを作るには
できるだけ多くの配列を集めた
マルチプルアライメントが必要

堂々巡りの関係

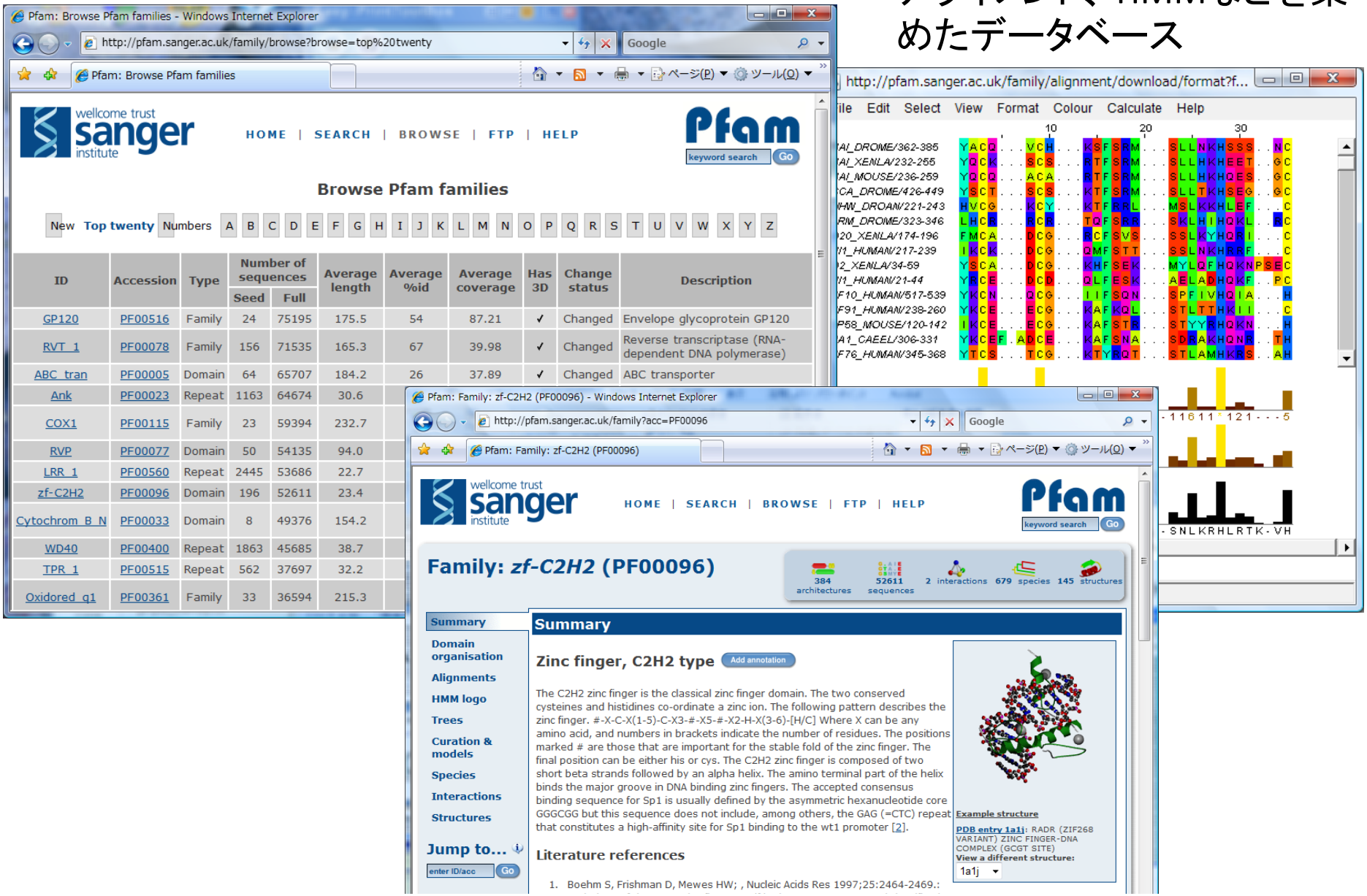
PSI-BLASTの手続き



Pfam：蛋白質ファミリのデータベース

http://pfam.sanger.ac.uk

各蛋白質ファミリのマルチプルアライメント、HMMなどを集めたデータベース



H19 問50

相同性検索に用いられるツールの一つに、PSI-BLASTがある。このPSI-BLASTでは位置特異的スコア行列(PSSM)を利用している。次に示した説明文の中で、PSI-BLASTとそこで用いられるPSSMについての記述として不適切なものはどれか。一つ選べ。

1. 一般的に、通常のBLASTに比べて感度が高い。
2. 一度作成されたPSSMを用いて検索を行い、その結果を用いてPSSMを再構築する処理を繰り返す。
3. PSI-BLASTはDNA配列しか取り扱えない。
4. PSI-BLASTでは、ギャップを取り扱うことができる。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19 問50

相同性検索に用いられるツールの一つに、PSI-BLASTがある。このPSI-BLASTでは位置特異的スコア行列(PSSM)を利用している。次に示した説明文の中で、PSI-BLASTとそこで用いられるPSSMについての記述として不適切なものはどれか。一つ選べ。

1. 一般的に、通常のBLASTに比べて感度が高い。
2. 一度作成されたPSSMを用いて検索を行い、その結果を用いてPSSMを再構築する処理を繰り返す。
3. PSI-BLASTはDNA配列しか取り扱えない。
4. PSI-BLASTでは、ギャップを取り扱うことができる。

※プロフィール法の考え方自体はDNAでもタンパク質でも適用可能だが、PSI-BLASTはアミノ酸配列しか取り扱うことができない。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19　問53

以下に示すような位置特異的スコア行列(PSSM)がある。このPSSMを利用してスコアを付けた結果、もっとも高いスコアを示す配列を選択肢の中から選べ。

	位置				
	1	**2**	**3**	**4**	**5**
A	6	-3	-3	0	-3
C	-9	0	-5	-3	6
G	-3	7	-4	-7	0
T	2	-3	0	0	-3

1．AGTAC
2．CACGA
3．TCTTG
4．TGTTC

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H19　問53

以下に示すような位置特異的スコア行列(PSSM)がある。このPSSMを利用してスコアを付けた結果、もっとも高いスコアを示す配列を選択肢の中から選べ。

	位置				
	1	**2**	**3**	**4**	**5**
A	**6**	-3	-3	**0**	-3
C	-9	0	-5	-3	**6**
G	-3	**7**	-4	-7	0
T	2	-3	**0**	**0**	-3

1. AGTAC　=6+7+0+0+6=19

2. CACGA　=-9+6-6-7-3=-19

3. TCTTG　=2+0+0+0+0=2

4. TGTTC　=2+7+0+0+6=15

※この問題はDNA配列のPSSMを扱っている。DNAのPSSMは遠縁のホモログの発見よりは、転写調節領域のパターンを記述するのによく使われる。

平成19年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20　問47

4塩基からなる塩基配列のモチーフを、次のような重み行列で表現した。

	Position 1	Position 2	Position 3	Position 4
A	10	-21	-11	-10
T	1	-22	-15	23
G	-20	13	12	-21
C	-20	-22	3	-15

この重み行列を用いて、7塩基の長さの配列、AGAGGTCを検索した時に、最も高いスコアを示す部分配列はどれか。選択肢の中から選べ。

1．AGAG
2．GAGG
3．AGGT
4．GGTC

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用

H20　問47

4塩基からなる塩基配列のモチーフを、次のような重み行列で表現した。

	Position 1	Position 2	Position 3	Position 4
A	**10**	-21	-11	-10
T	1	-22	-15	**23**
G	-20	**13**	**12**	-21
C	-20	-22	3	-15

この重み行列を用いて、7塩基の長さの配列、AGAGGTCを検索した時に、最も高いスコアを示す部分配列はどれか。選択肢の中から選べ。

1．AGAG
2．GAGG
③．AGGT
4．GGTC

各ポジションで最大のスコアをとる塩基を並べるとAGGTとなる。

平成20年度バイオインフォマティクス技術者認定試験
（日本バイオインフォマティクス学会主催）問題から引用