近畿大学・農学部・生命情報学

ペアワイズアライメントと配列相同性解析

2009年4月14日（火）

奈良先端大･情報･蛋白質機能予測学講座
川端　猛
takawaba@is.naist.jp

http://isw3.naist.jp/IS/Kawabata-lab/home-ja.html

平成21年度「生命情報学＆生命情報学実習」講義日程

2009.4.14

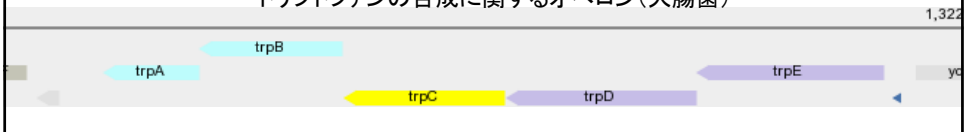
	講義	生命情報学	演習	生命情報学演習
4/7	川端1	配列決定とバイオインフォマティクス		
4/14	川端2	ペアワイズアライメントと配列相同性検索	川端	主要WEBデータベースの使用法(BLAST)
4/21	川端3	マルチプルアライメントとその応用	中村	ChemOfficeを用いた計算化学演習
4/28	川端4	分子系統学基礎		
5/12	川端5	蛋白質の物理化学的性質と配列解析	中村	系統樹作成演習(ClustalX)
5/19	川端6	蛋白質立体構造データの情報解析	川端	蛋白質立体構造データの可視化(RasMol)
5/26	川端7	>>試験<<		
6/2	金谷1	ポストゲノム解析入門(トランスクリプトーム解析)		
6/9	金谷2	ポストゲノム解析入門(インタラクトーム解析)	金谷1	発現プロフィール解析演習
6/16	金谷3	ポストゲノム解析（統合解析）	金谷2	インタラクトローム・代謝物解析演習
6/23	金谷4	メタボローム解析（その1）		
6/30	金谷5	メタボローム解析（その2）		
7/7	金谷6	メタボローム解析（その3）		
7/14	金谷7	>>試験<<		

先週のゲノムデータベースの話題の補足のスライド

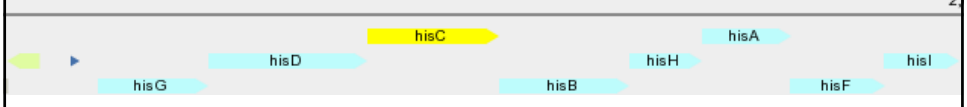
バクテリアのオペロン構造

オペロン：ゲノム上、遺伝子群が隣接して同じ方向にコードされた領域。
多くの場合、それらはまとめて一度に転写され翻訳される。

トリプトファンの合成に関するオペロン（大腸菌）

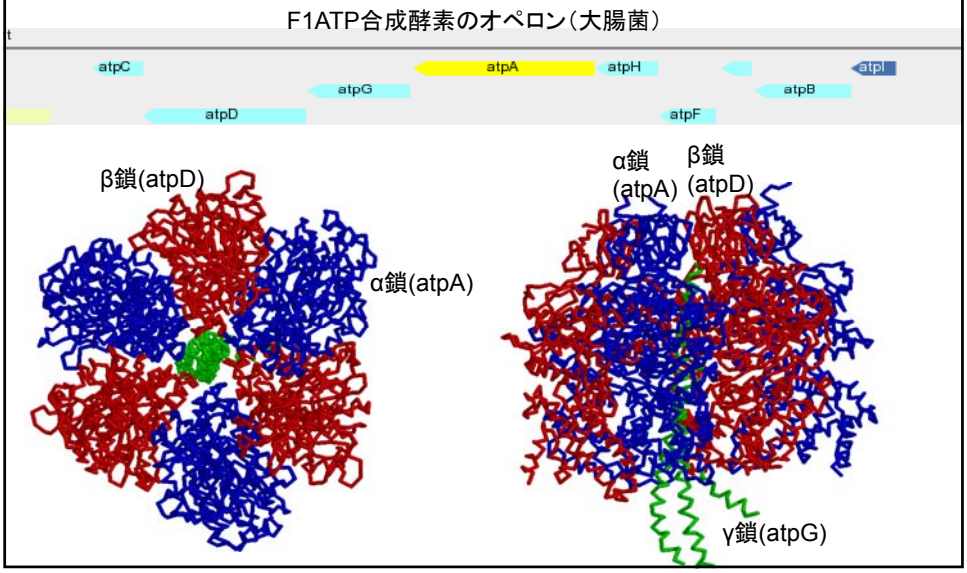


ヒスチジンの合成に関するオペロン（大腸菌）



※オペロンにコードされる遺伝子群はある生物学的機能を担うのに必要な遺伝子群であることが多い。アミノ酸合成、細胞外からの分子の取り込み（トランスポータ）、リボゾームのタンパク質、鞭毛のタンパク質などが、オペロンを構成することが多い。
※オペロンがあるのは原核生物のみ。真核生物はオペロンを持たない。

オペロン構造をなす遺伝子群が一つの複合体を形成する例

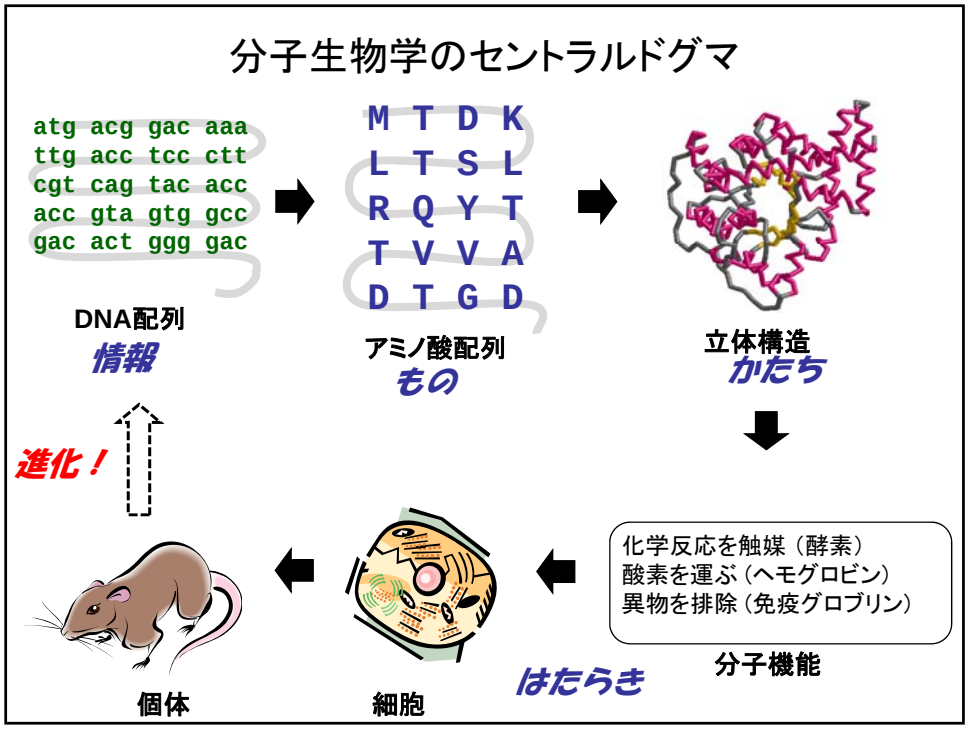


エキソン・イントロンの境界に現れやすい配列

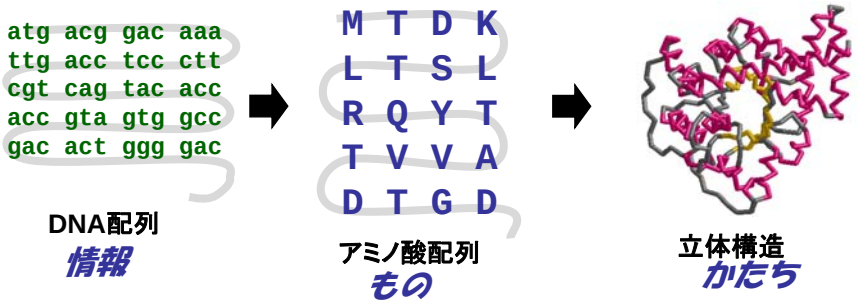
ヒトの場合、エキソンの直後のイントロンの配列は”GT”であることが多い

最初のエキソンの先頭は開始コドン(ATG)、
最後のエキソンの末尾は終止コドンになる

ペアワイズアライメント



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

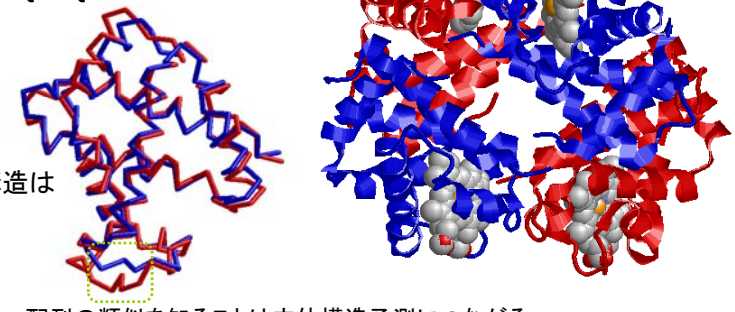
```
HUMAN 4:RKFFVGGNWKMNGRKQSLGELIGTLNAAKVP-ADTEVVCAPPTAYIDFARQKLD-PKIAV:61
        *   * **** **      **   *      *   *  ***  *** *     *  *
ECOLI 2:RHPLVMGNWKLNGSRHMVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIML:61

TPIS_HUMAN 248 vs TPIS_ECOLI 255 SeqID 45.9  %
```

削除･挿入(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

配列比較(配列相同性検索)の基本論理

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)

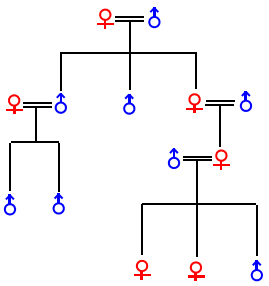
相同(homology):進化的な原因によるもの。祖先を共有。
(進化史の中である時点まで同じであったから似ている)

相似(analogy) :それ以外の原因によるもの

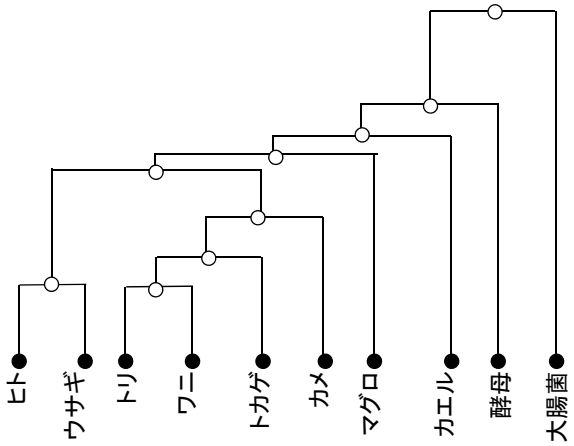
進化のイメージ：系統樹

対象物が生成される過程（歴史、進化史）を木構造で示したもの

家系図



生物種の系統図



2つの配列を比較するには？

1. <u>類似性のスコア関数の定義</u>

   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. <u>アライメント</u>

   どうやって文字と文字を対応づけるか？

```
ABCDEF      ABCDEF
        ⇒    ***
CDE         --CDE-
```

```
BCDEF       -BCDEF-
        ⇒    *  **
ABEEFG      AB-EEFG
```

もっと長いときはどうやって計算する？

スコア関数の定義

(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)

$$S(A,B)=\log\frac{P_{evo}(A,B)}{P_{rand}(A)P_{rand}(B)}$$

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

$P_{evo}(A,B)$: 進化的な関係からAとBの対応が生じた確率

$P_{rand}(A)\cdot P_{rand}(B)$: 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

スコアの計算例

AFDC
AEEC

S(A,A) + S(F,E) S(D,E) + S(C,C) = 12
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
AEE-C

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

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

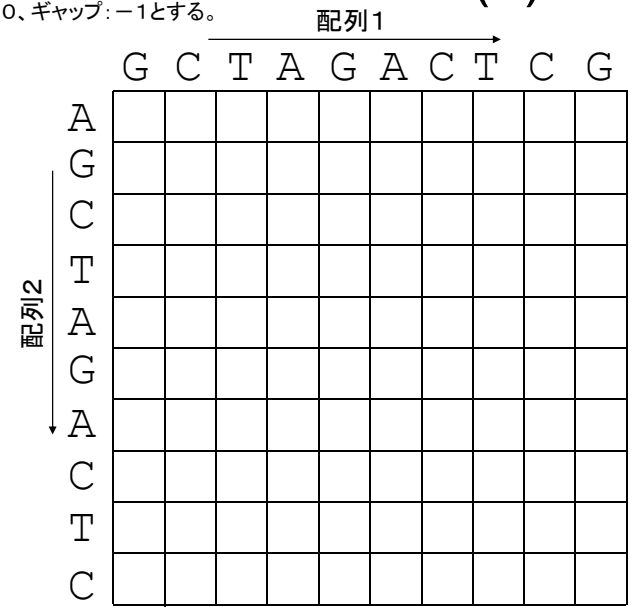
動的計画法というアルゴリズムで解く。
そのイメージをつかむためには**ドットマトリックス法**が有効

ドットマトリックス　：　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



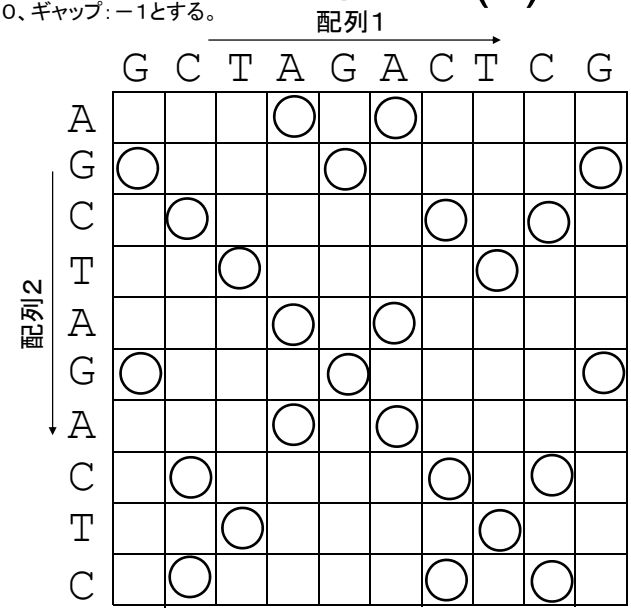
ドットマトリックス　：　例１　(2)

※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)

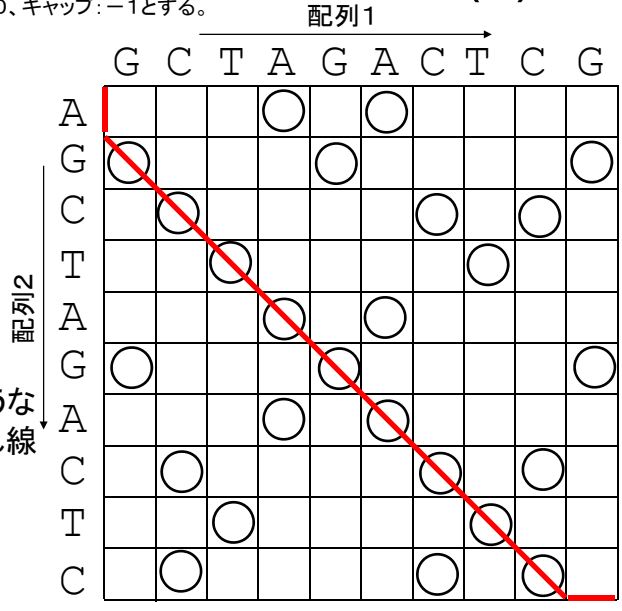
※スコア：一致：＋１、不一致：０、ギャップ：－１とする。

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

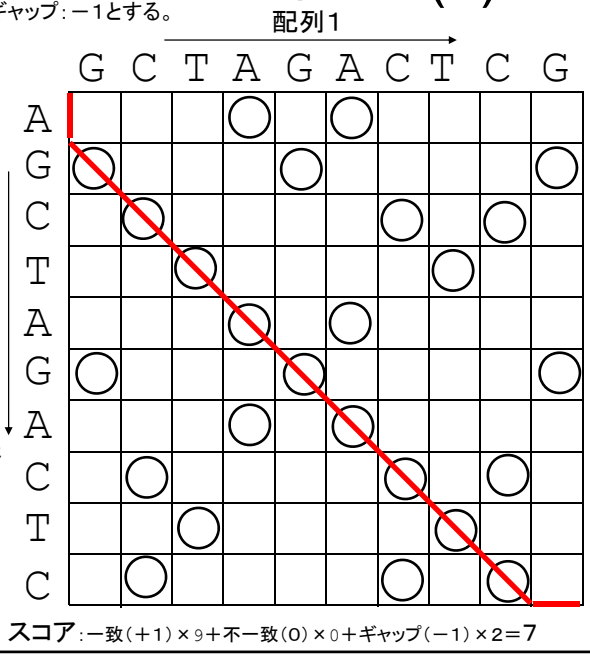
(1) 配列1、配列2を横と縦に並べる

(2) 文字が一致するマスに○を描く

(3) 多くの○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-

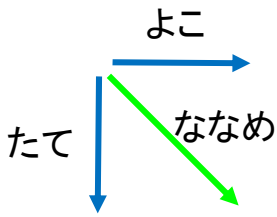


スコア：一致(＋1)×9＋不一致(0)×0＋ギャップ(－1)×2＝7

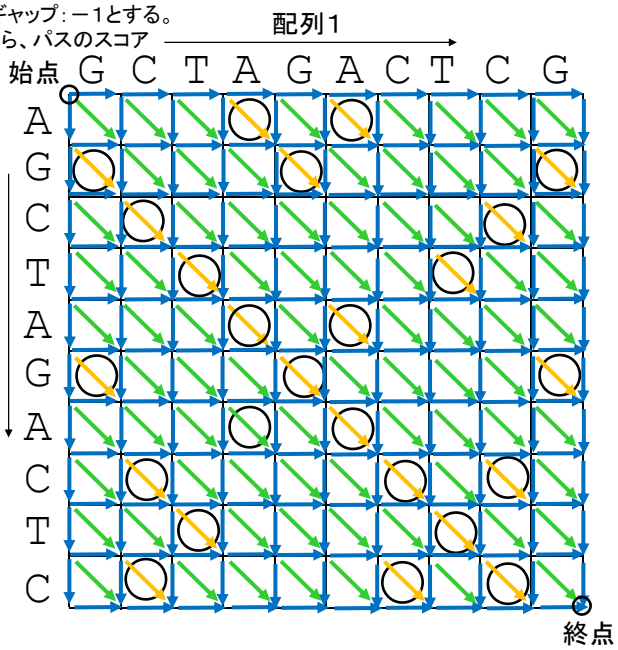
ドットマトリックスのパスの引き方の詳細

※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

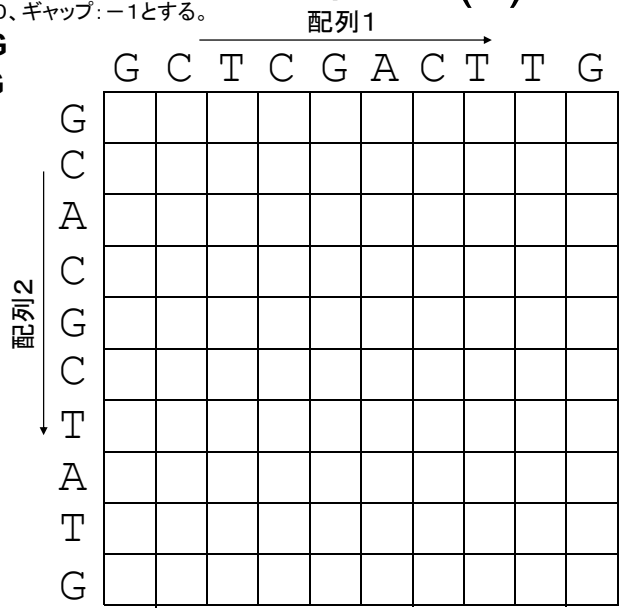


ドットマトリックス　：　例2　(1)

※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



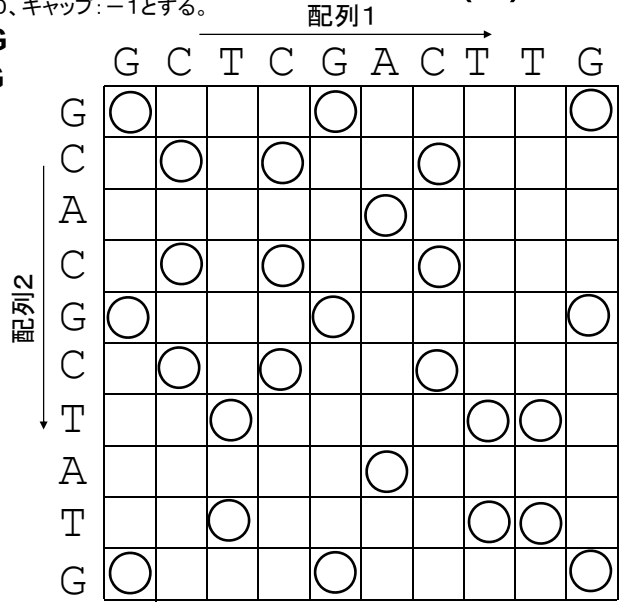
ドットマトリックス　：　例2　(2)

※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○を描く



ドットマトリックス　：　例2　(3)

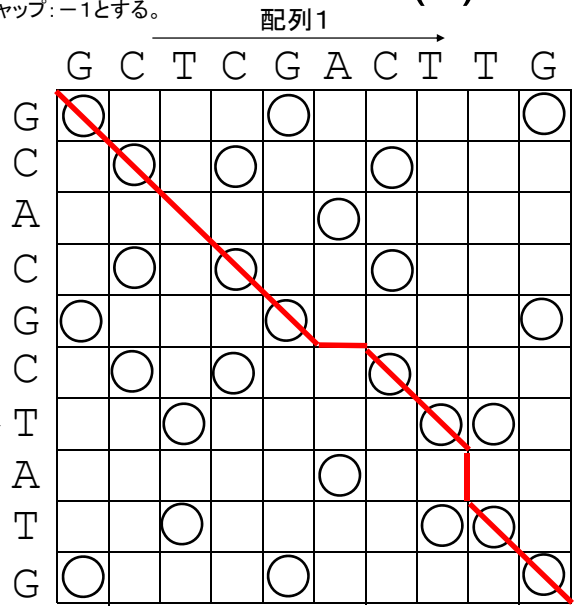
※スコア：一致：＋1、不一致：0、ギャップ：－1とする。

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**

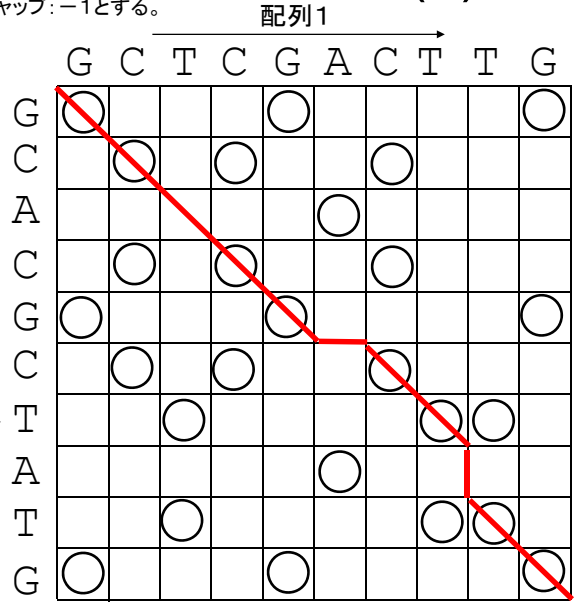
(1) 配列1、配列2を横と縦に並べる

(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

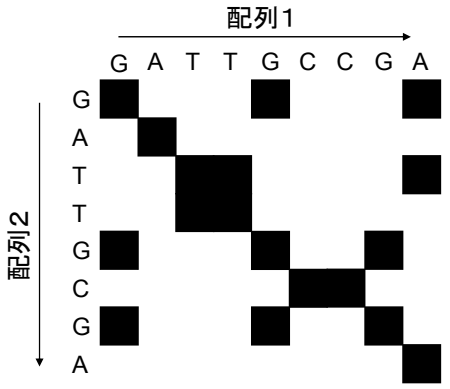


スコア：一致（＋1）×8＋不一致（0）×1＋ギャップ（－1）×2＝6

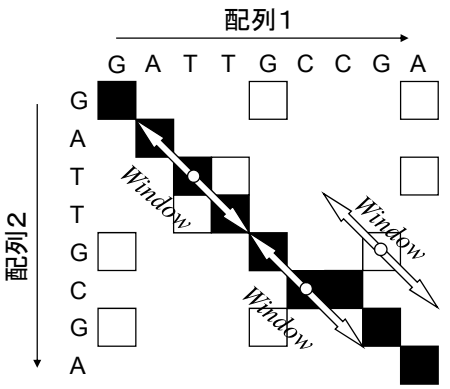
対角上の平均化によるスムージング

配列1：GATTGCCGA

配列2：GATTGCGA

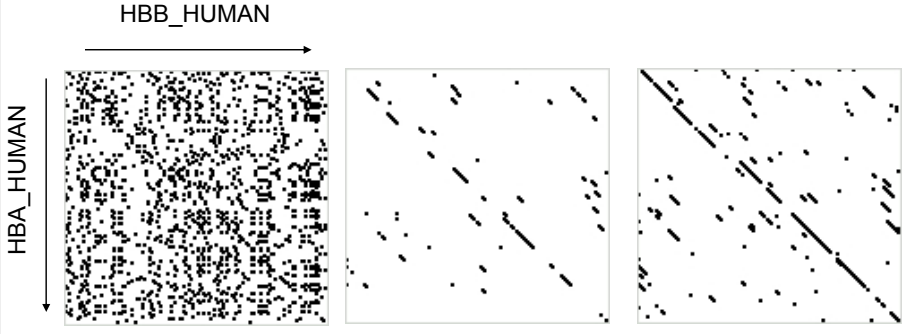


（1）単純に一致している座標を黒く塗る
Window=1, *Threshold*=1に相当



（2）長さ*Window*の連続したペアが比較し、
一致度が*Threshold*以上であれば黒く塗る
Window=3, *Threshold*=2の場合

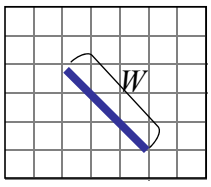
ドットマトリックスの例



Matrix=ID,W=1,T=1

Matrix=ID,W=5,T=3

Matrix=BLOSUM62,
W=7,T=10

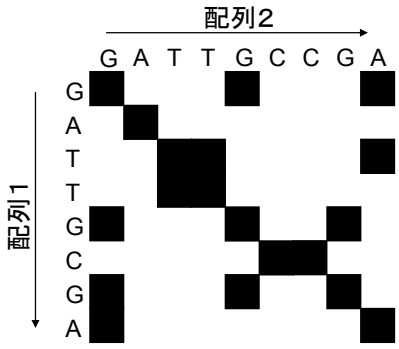


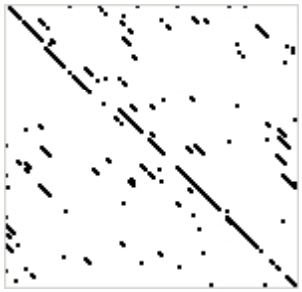
文字が一致しているペアを黒く塗る
→ 長さW のwordの総スコアがT 以上なら中心を黒く塗る

スコアは、最も簡単には一致・不一致スコア
対数オッズスコアを使うとより高感度になる

ドットマトリックス法の特徴

- アルゴリズムが平易
- 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。
- あくまでグラフィカルな対応なので、具体的な文字列対応（アライメント）は与えない。



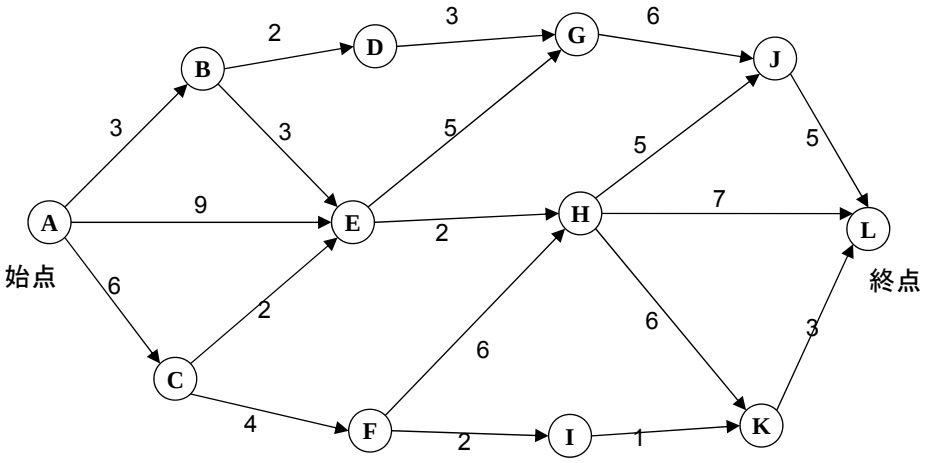


動的計画法によるアライメント

- 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）

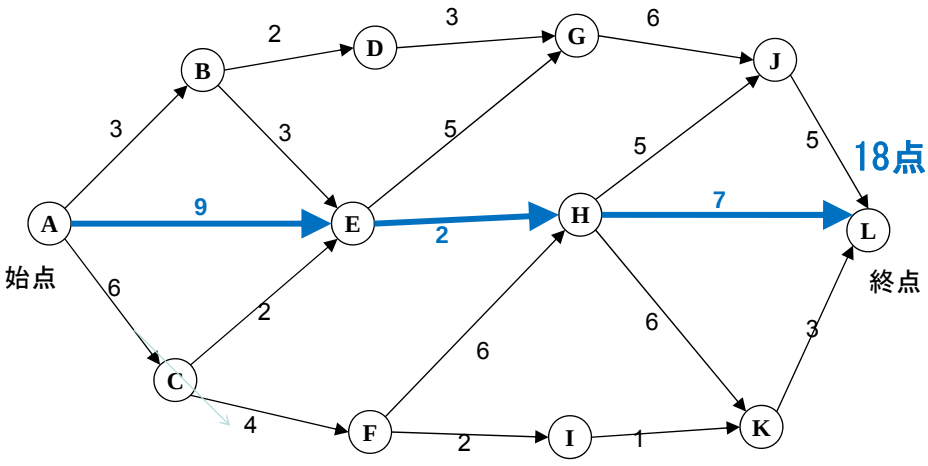
最適経路問題

始点Aから終点Lにいたるエッジの得点の合計が
最大となる経路を探す



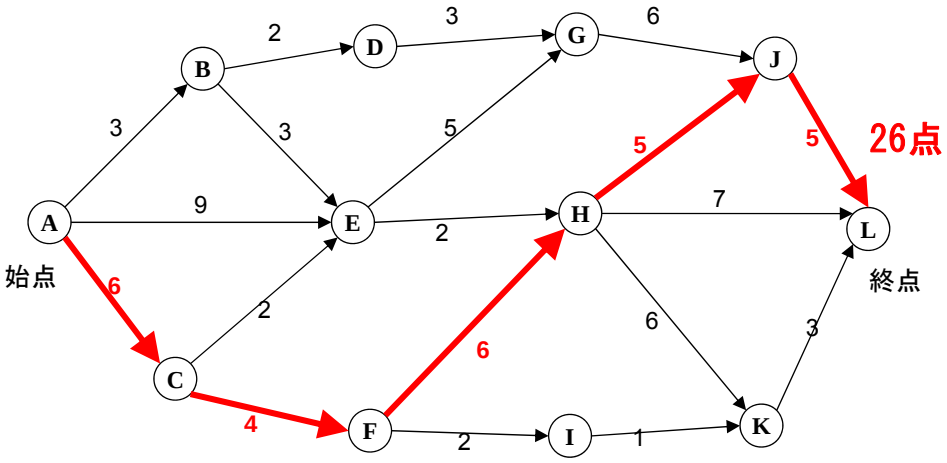
最適経路問題

始点Aから終点Lにいたるエッジの得点の合計が
最大となる経路を探す



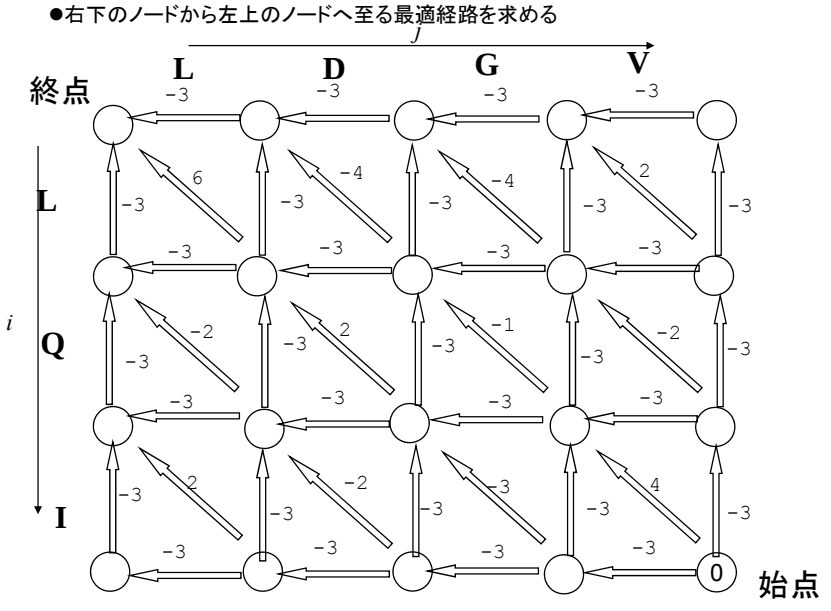
最適経路問題

始点Aから終点Lにいたるエッジの得点の合計が
最大となる経路を探す



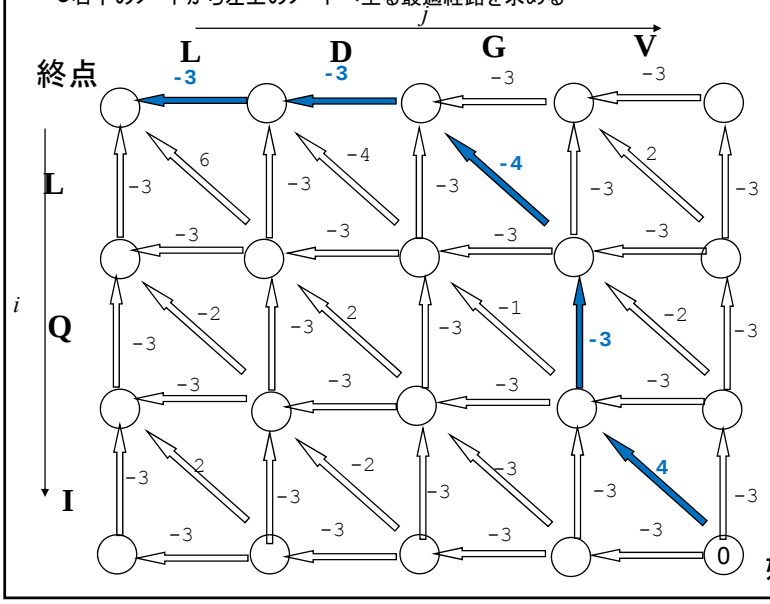
アライメントを最適経路問題として考える

- 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



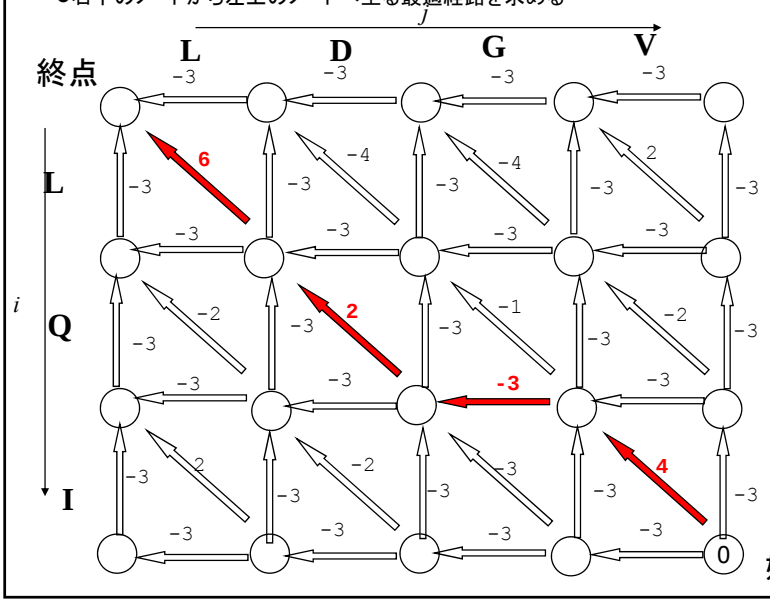
アライメントを最適経路問題として考える

- ●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



アライメントを最適経路問題として考える

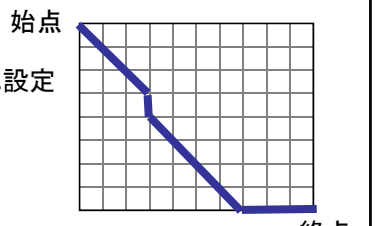
- ●鉛直、水平に比較したい文字列を並べる
- ●対角線のエッジには一致スコア、鉛直水平のエッジにはギャップスコアを書き込む
- ●右下のノードから左上のノードへ至る最適経路を求める



グローバル・アライメントの解法
(Needleman & Wunsh,1970)

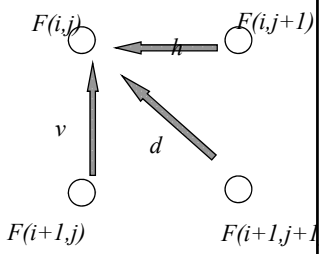
(0)準備

右端の列、下端の行の格子点のスコアを0に設定



(1)前向きステップ

$$F(i,j)=\max\begin{cases}F(i+1,j+1)+S(x_i,y_j) & \text{対角}(d)\\ F(i+1,j)+Gap & \text{鉛直}(v)\\ F(i,j+1)+Gap & \text{水平}(h)\end{cases}$$

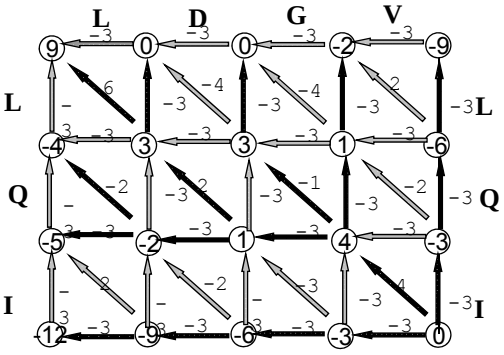


(2)後ろ向きステップ

始点を起点にして辿る。終点に到着したら終了。

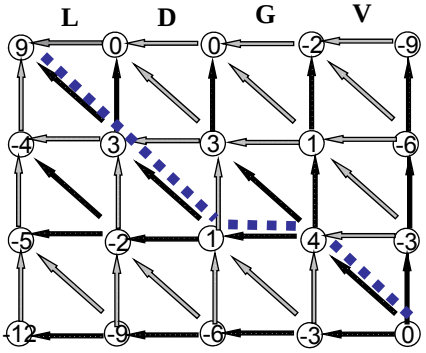
動的計画法の手続き

(1)前向き（Forward）



O(NM)

(2)後ろ向き（TraceBack）



LDGV
LQ-I

グローバルとローカルの格子上の違い

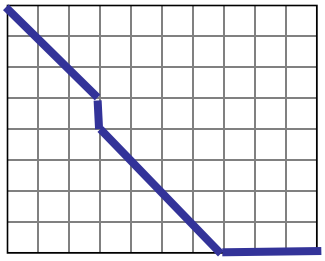
ACDEFGHKLM
AFGHKKL

ACDEFGHK-LM
A---FGHKKL-

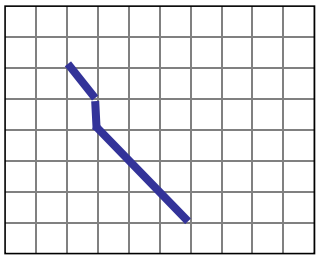
グローバル

FGHK-L
FGHKKL

ローカル

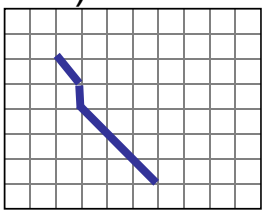


グローバル



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

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

(2)後ろ向きステップ

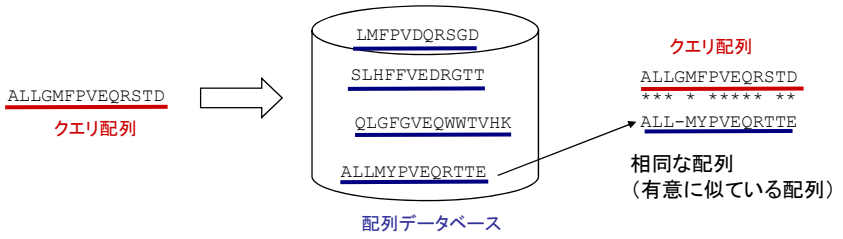
最大のスコアのノードを探し、そのノードを起点にして辿る。パス’0’が現れたら終了

配列相同性検索

ー BLASTを中心として ー

配列相同性検索

→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：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

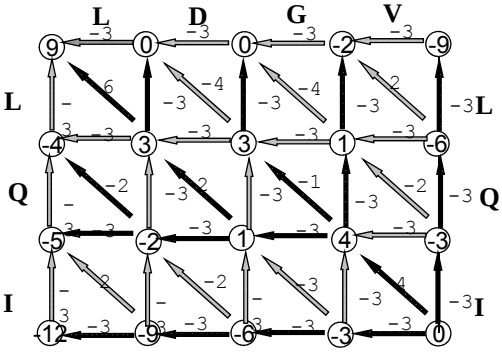
計算時間の比較

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

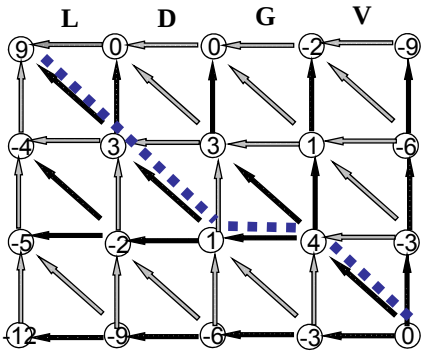
	説明	計算時間
私が書いたDP	Smith & WatermanをCで素朴に実装	144.97 sec
SSEARCH35	FASTAの開発グループが実装した Smith & Waterman	15.01 sec
FASTA35	ヒューリスティックアルゴリズムを使用	2.36 sec
BLASTP	ヒューリスティックアルゴリズムを使用	0.38 sec

動的計画法の復習

(1)Forward



(2)TraceBack



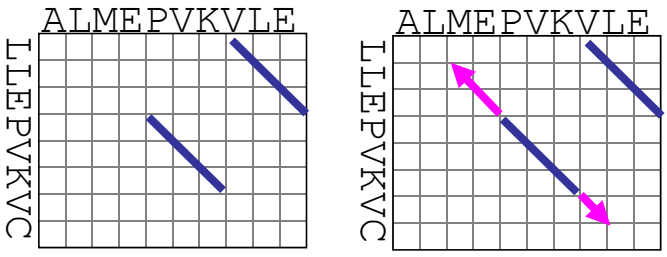
O(NM)

LDGV
LQ-I

BLASTのヒューリスティックス

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に対し近隣wordのリストを作成
2. 近隣wordリスト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. さらにgap入りアライメントで伸展



BLASTの出力例(1)

```
BLASTP 2.2.1 [Apr-13-2001]


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_ECOLI "RecA protein (Recombinase A)"
         (352 letters)

Database: 40scop1.59nm
           3886 sequences; 705,110 total letters

Searching.....done


                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

2reb-1 [c.37.1.11] RECA PROTEIN (E.C.3.4.99.37)                     448  e-127
1g18A2 [d.48.1.1] RECA PROTEIN                                       70  9e-14
1g0uF [d.153.1.4] PROTEASOME COMPONENT C1                            32  0.020
1byrA [d.136.1.1] ENDONUCLEASE                                       28  0.29
1g3qA [c.37.1.10] CELL DIVISION INHIBITOR                            28  0.38
1ct5A [c.1.6.2] YEAST HYPOTHETICAL PROTEIN, SELENOMET                28  0.49
1g0uD [d.153.1.4] PROTEASOME COMPONENT PUP2                          27  1.1
1e32A2 [c.37.1.13] P97                                               26  1.4
1g0uA [d.153.1.4] PROTEASOME COMPONENT Y7                            26  1.9
1cp2A [c.37.1.10] NITROGENASE IRON PROTEIN                           26  1.9
1f3oA [c.37.1.12] HYPOTHETICAL ABC TRANSPORTER ATP-BINDING PROTEIN   25  2.4
1qj2B2 [d.133.1.1] CARBON MONOXIDE DEHYDROGENASE                     25  3.2
1dgyA [c.72.1.1] ADENOSINE KINASE                                    25  3.2
1skyB3 [c.37.1.11] F1-ATPASE                                         25  3.2
1g6oA [c.37.1.13] CAG-ALPHA                                          25  4.2
1cmxA [d.3.1.6] UBIQUITIN YUH1-UBAL                                  24  7.1
8abp- [c.93.1.1] L-*ARABINOSE-BINDING PROTEIN (MUTANT WITH MET 1...  24  7.1
2tpsA [c.1.3.1] THIAMIN PHOSPHATE SYNTHASE                           24  7.1
```

BLASTの出力例(2)

```
1bi3A [c.2.1.2] ALCOHOL DEHYDROGENASE                               23  9.3
1pmi- [b.82.1.3] PHOSPHOMANNOSE ISOMERASE                           23  9.3

>2reb-1 [c.37.1.11] RECA PROTEIN (E.C.3.4.99.37)
          Length = 243

 Score =  448 bits (1152), Expect = e-127
 Identities = 243/266 (91%), Positives = 243/266 (91%), Gaps = 23/266 (8%)

Query: 3   DENKQKALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIALGAGGLPMGRIV 62
           DENKQKALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIALGAGGLPMGRIV
Sbjct: 1   DENKQKALAAALGQIEKQFGKGSIMRLGEDRSMDVETISTGSLSLDIALGAGGLPMGRIV 60

Query: 63  EIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTG 122
           EIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTG
Sbjct: 61  EIYGPESSGKTTLTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPDTG 120

Query: 123 EQALEICDALARSGAVDVIVVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNL 182
           EQALEICDALARSGAVDVIVVDSVAALTPKAEIE         GLAARMMSQAMRKLAGNL
Sbjct: 121 EQALEICDALARSGAVDVIVVDSVAALTPKAEIE---------GLAARMMSQAMRKLAGNL 172

Query: 183 KQSNTLLIFINQIRMKIGVMFGNPETTTGGNALKFYASVRLDIRRIGAVKEGENVVGSET 242
           KQSNTLLIFINQ               TGGNALKFYASVRLDIRRIGAVKEGENVVGSET
Sbjct: 173 KQSNTLLIFINQ---------------TGGNALKFYASVRLDIRRIGAVKEGENVVGSET 217

Query: 243 RVKVVKNKIAAPFKQAEFQILYGEGI 268
           RVKVVKNKIAAPFKQAEFQILYGEGI
Sbjct: 218 RVKVVKNKIAAPFKQAEFQILYGEGI 243


>1g18A2 [d.48.1.1] RECA PROTEIN
          Length = 60

 Score = 70.1 bits (170), Expect = 9e-14
 Identities = 30/56 (53%), Positives = 44/56 (78%)

Query: 272 GELVDLGVKEKLIEKAGAWYSYKGEKIGQGKANATAWLKDNPETAKEIEKKVRELL 327
           G L+D+GV + LI K+GAW++Y+GE++GQGK NA  +L +N + A EIEKK++ L
```

BLASTの
出力例(3)

どれだけ似ていれば意味があるのか？
類似性の指標

- **同一残基率(%)**
  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ペナルティに依存
- 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

・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。

```
BLASTP 2.2.1 [Apr-13-2001]


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_ECOLI "RecA protein (Recombinase A)"
         (352 letters)

Database: 40scop1.59nm
           3886 sequences; 705,110 total letters

Searching........done


                                                                  Score     E
Sequences producing significant alignments:                       (bits)  Value

2reb-1 [c.37.1.11] RECA PROTEIN (E.C.3.4.99.37)                     448  e-127
1g18A2 [d.48.1.1] RECA PROTEIN                                       70  9e-14
1g0uF [d.153.1.4] PROTEASOME COMPONENT C1                            32  0.020
1byrA [d.136.1.1] ENDONUCLEASE                                       28  0.29
1g3qA [c.37.1.10] CELL DIVISION INHIBITOR                            28  0.38
1ct5A [c.1.6.2] YEAST HYPOTHETICAL PROTEIN, SELENOMET                28  0.49
1g0uD [d.153.1.4] PROTEASOME COMPONENT PUP2                          27  1.1
1e32A2 [c.37.1.13] P97                                               26  1.4
1g0uA [d.153.1.4] PROTEASOME COMPONENT Y7                            26  1.9
1cp2A [c.37.1.10] NITROGENASE IRON PROTEIN                           26  1.9
1f3oA [c.37.1.12] HYPOTHETICAL ABC TRANSPORTER ATP-BINDING PROTEIN   25  2.4
1qj2B2 [d.133.1.1] CARBON MONOXIDE DEHYDROGENASE                     25  3.2
1dgyA [c.72.1.1] ADENOSINE KINASE                                    25  3.2
```

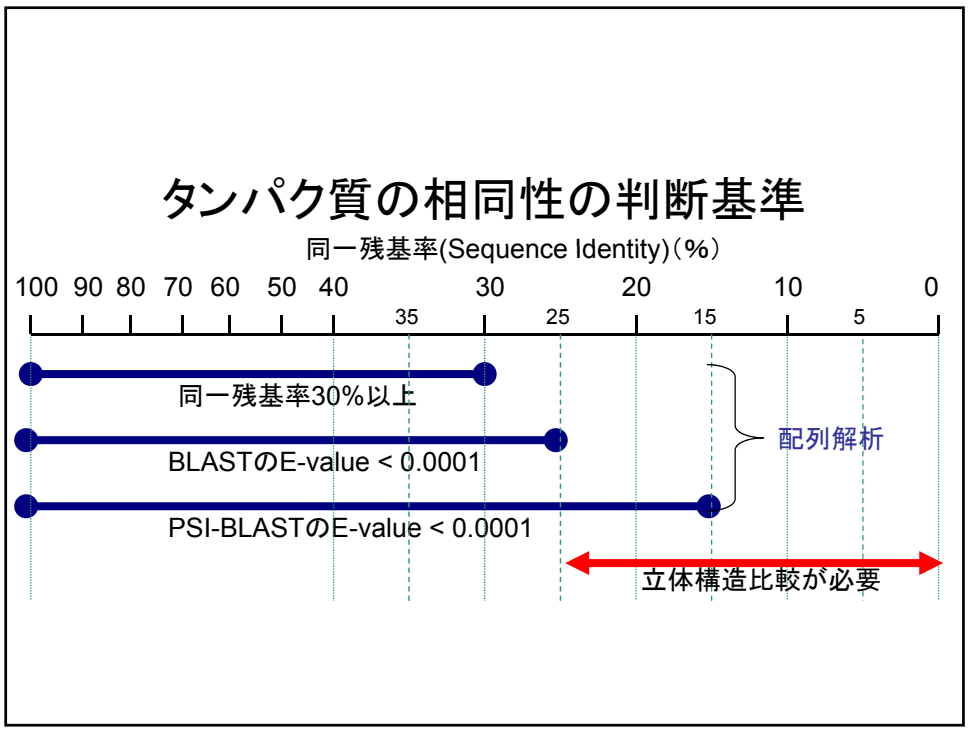
```
Query: 123 EQALEICDALARSGAVDVIVVDSVAALTPKAEIEGEIGDSHMGLAARMMSQAMRKLAGNL 182
           EQALEICDALARSGAVDVIVVDSVAALTPKAEIE        GLAARMMSQAMRKLAGNL
Sbjct: 121 EQALEICDALARSGAVDVIVVDSVAALTPKAEIE--------GLAARMMSQAMRKLAGNL 172

Query: 183 KQSNTLLIFINQIRMKIGVMFGNPETTTGGNALKFYASVRLDIRRIGAVKEGENVVGSET 242
           KQSNTLLIFINQ               TGGNALKFYASVRLDIRRIGAVKEGENVVGSET
Sbjct: 173 KQSNTLLIFINQ---------------TGGNALKFYASVRLDIRRIGAVKEGENVVGSET 217

Query: 243 RVKVVKNKIAAPFKQAEFQILYGEGI 268
           RVKVVKNKIAAPFKQAEFQILYGEGI
Sbjct: 218 RVKVVKNKIAAPFKQAEFQILYGEGI 243
```

Bit Score

Raw Score

```
>1g18A2 [d.48.1.1] RECA PROTEIN
          Length = 60

 Score =  70.1 bits (170), Expect = 9e-14
 Identities = 30/56 (53%), Positives = 44/56 (78%)

Query: 272 GELVDLGVKEKLIEKAGAWYSYKGEKIGQGKANATAWLKDNPETAKEIEKKVRELL 327
           G L+D+GV + LI K+GAW++Y+GE++GQGK NA  +L +N + A EIEKK++E L
Sbjct: 4   GSLIDMGVDQGLIRKSGAWFTYEGEQLGQGKENARNFLVENADVADEIEKKIKEKL 59


>1g0uF [d.153.1.4] PROTEASOME COMPONENT C1
          Length = 242

 Score = 32.3 bits (72), Expect = 0.020
 Identities = 25/88 (28%), Positives = 47/88 (53%), Gaps = 9/88 (10%)

Query: 271 YGELVDLGVKEKLIEKAGAWYSYKGEKIGQGKANATAWLK----DNPE--TAKEIEKKVR 324
```

```
    Database: 40scop1.59nm
      Posted date:  Jun 22, 2002  3:06 PM
    Number of letters in database: 705,110
    Number of sequences in database:  3886
  Lambda     K      H
     0.314    0.134    0.369
  Gapped
  Lambda     K      H
     0.267   0.0410    0.140
  Matrix: BLOSUM62
  Gap Penalties: Existence: 11, Extension: 1
  Number of Hits to DB: 469,543
  Number of Sequences: 3886
  Number of extensions: 18494
  Number of successful extensions: 65
  Number of sequences better than 10.0: 17
  Number of HSP's better than 10.0 without gapping: 13
  Number of HSP's successfully gapped in prelim test: 4
  Number of HSP's that attempted gapping in prelim test: 50
  Number of HSP's gapped (non-prelim): 17
  length of query: 352
  length of database: 705,110
  effective HSP length: 79
  effective length of query: 273
  effective length of database: 398,116
  effective search space: 108685668
  effective search space used: 108685668
```

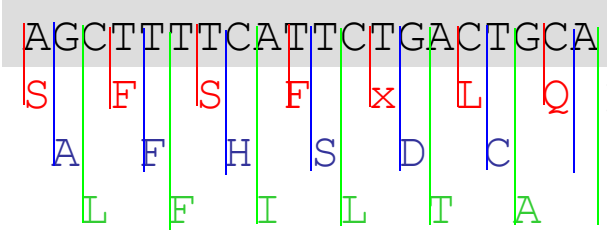


BLASTのプログラムの種類

	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
tblastn	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

blastp(アミノ酸対アミノ酸)によるタンパク質の機能予測

クエリ:*T.thermophius*のタンパク質, データベース:大腸菌の全タンパク質

```
BLASTP 2.2.3 [May-13-2002]
Query= X07 AAS80531.1 tthe0          (144 letters)
Database: ecoli_aa            4237 sequences; 1,350,094 total letters

                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

infC NP_416233.1 "protein chain initiation factor IF-3" NC_000913     137   2e-34
rhsD NP_415030.1 "RhsD protein in RhsD element" NC_000913              28   0.19
pta NP_416800.1 "phosphotransacetylase" NC_000913                      25   2.0
prsA NP_415725.1 "phosphoribosylpyrophosphate synthetase" NC_000913    25   2.7
yiaK NP_418032.1 "2,3-diketo-L-gulonate dehydrogenase, NADH-depe...    24   3.5
ffh NP_417101.1 "4.5S-RNP protein, GTP-binding export factor, pa...    24   4.6
ybdR NP_415141.1 "putative dehydrogenase, NAD(P)-binding" NC_000913    24   4.6
ydfG NP_416057.1 "putative oxidoreductase" NC_000913                   23   7.8

>infC NP_416233.1 "protein chain initiation factor IF-3" NC_000913
          Length = 180

 Score =  137 bits (346), Expect = 2e-34
 Identities = 72/139 (51%), Positives = 92/139 (65%), Gaps = 1/139 (0%)

Query: 4   REALRLAQEMDLDLVLVGPNADPPVARIMDYSKWRYEQQMXXXXXXXXXXXTEVKSIKFR 63
           REAL  A+E  +DLV + PNA+PPV RIMDY K+ YE+              +VK IKFR
Sbjct: 40  REALEKAEEAGVDLVEISPNAEPPVCRIMDYGKFLYEKSKSSKEQKKKQKVIQVKEIKFR 99

Query: 64  VKIDEHDYQTKLGHIKRFLQEGHKVKVTIMFRGREVAHPELGERILNRVTEDLKDLAVVE 123
              DE DYQ KL  + RFL+EG K K+T+ FRGRE+AH ++G  +LNRV +DL++LAVVE
Sbjct: 100 PGTDEGDYQVKLRSLIRFLEEGDKAKITLRFRGREMAHQQIGMEVLNRVKDDLQELAVVE 159
```

blastp(アミノ酸対アミノ酸)の適用例)

ORFのアノテーション:*H.influenzae*のORF対大腸菌のORF

```
Query= HI0078 hinf0 AAC21753.1
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

cysS ecol0 AAC73628.1 "cysteine tRNA synthetase"                    730   0.0
metG ecol0 AAC75175.1 "methionine tRNA synthetase"                   39   5e-04
ileS ecol0 AAC73137.1 "isoleucine tRNA synthetase"                   39   0.001
leuS ecol0 AAC73743.1 "leucine tRNA synthetase"                      30   0.25
yidW ecol0 AAC76718.1 "regulator protein for dgo operon"             28   1.3
```

→ HI0078はcysteine tRNA syntetase

```
Query= HI0083 hinf0 AAC21762.1
         (71 letters)                                             Score    E
Sequences producing significant alignments:                       (bits) Value

ispB ecol0 AAC76219.1 "octaprenyl diphosphate synthase"              23   3.1
lplA ecol0 AAC77339.1 "lipoate-protein ligase A"                     22   6.9
nlpA ecol0 AAC76684.1 "lipoprotein-28"                               22   6.9
b1372 ecol0 AAC74454.1 "putative membrane protein"                   22   6.9
mdaA ecol0 AAC73938.1 "modulator of drug activity A"                 22   9.0
```

→ HI0083は大腸菌にはホモログがない

参考文献

- 金久實　著 「ポストゲノム情報への招待」 (2001) 共立出版
- 中村保一他編 「バイオデータベースとウェブツールの手とり足とり活用法　改訂第2版」 (2007) 羊土社
- Arthur M.Lesk(岡崎康司、坊農秀雄　監訳)「バイオインフォマティクス基礎講義 一歩進んだ発想をみがくために」(2003), メディカル・サイエンス・インターナショナル
- D.W.Mount著、岡崎康司、坊農秀雄 監訳「バイオインフォマティクス – ゲノム配列から機能解析へ　－」 第2版　メディカル・インターナショナル、2005年、11500円
- 阿久津達也 「バイオインフォマティクスの数理とアルゴリズム」(2007) 共立出版
- R.Durbin　他著、阿久津達也他訳 「バイオインフォマティクス - 確率モデルによる遺伝子解析」医学出版、2001年、9800円
- BLAST WEB page　http://www.ncbi.nlm.nih.gov/BLAST/