

ペアワイズアライメントと 配列相同性解析

2010年4月20日(火)

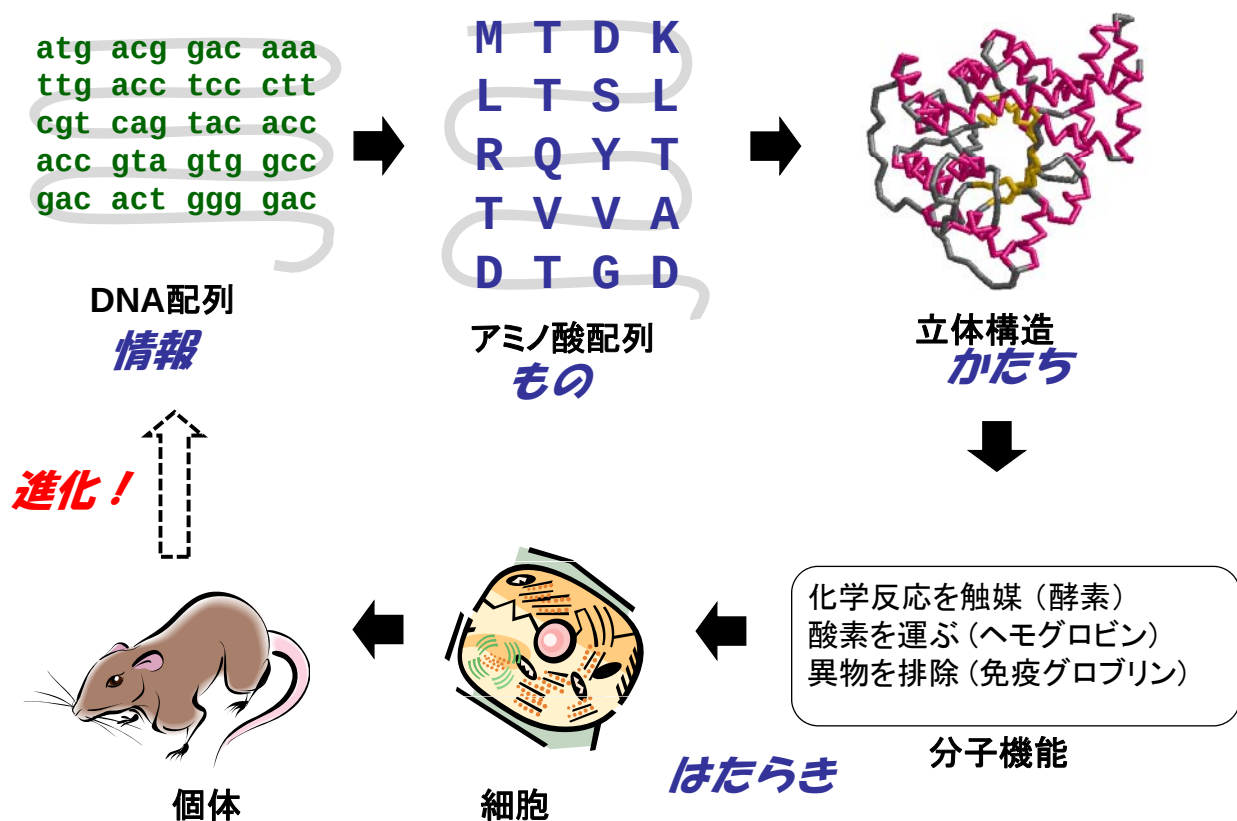
奈良先端大・情報・蛋白質機能予測学講座

川端 猛

takawaba@is.naist.jp

<http://isw3.naist.jp/IS/Kawabata-lab/lec-ja.html>

分子生物学のセントラルドグマ



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行に並んだ高分子

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...)

atgacggacaaaattgacctcccttcgtcagtacaccaccgtagtggccga
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

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

↓

atgacgaacaaaattgacctcccttcgtcagtacacc
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 5.3.1.1) (TIM, TPIS)
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESEDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_RABIT ウサギ "Triosephosphate isomerase (EC 5.3.1.1) (TIM, TPIS)
APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESEDELIGQKVAHALSEGLGVIACIGEKLDERE
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 5.3.1.1) (TIM, TPIS)
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESEDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_YEAST 酵母 "Triosephosphate isomerase (EC 5.3.1.1) (TIM, TPIS)
ARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATY
LDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWV
ILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKKA
GKTLDVVERQLNAVLEEVKDWTNVVVAYEPVWAIGTGLAATPEDA
QDIHASIRKFLASKLGDKAASELRILYGGSSANGSNAVTFKDKADV
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 5.3.1.1)
APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESEDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_ECOLI 大腸菌 "Triosephosphate isomerase (EC 5.3.1.1)
MRHPLVMGNWKLNGSRHMHVHELVSNLRLKELAGVAGCAVAIAPPEM
YIDMAKREAEGSHIMLGAQNVDLNLGAFTGETSAAMLKDIGAQY
IIIGHSERRTYHKESDELIAKKFAVLKEQGLTPVLCIGETEAENE
AGKTEEVCAQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP
AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAELFAQP
DIDGALVGGASLKADAFVIVKAAEAAKQA

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: APSRKFFVGGNWKMNGRKKNLGEITTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKIA: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: RHPLVMGNWKLNGSRHMHVHELVSNLRLKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIML: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: LSPADKTNVKAAGWKVGAHAGEYGAELERMFLSFPTTKTYFPHF-DLS-----HGSAQV:55
* *

Beta 3: LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV:60
* *

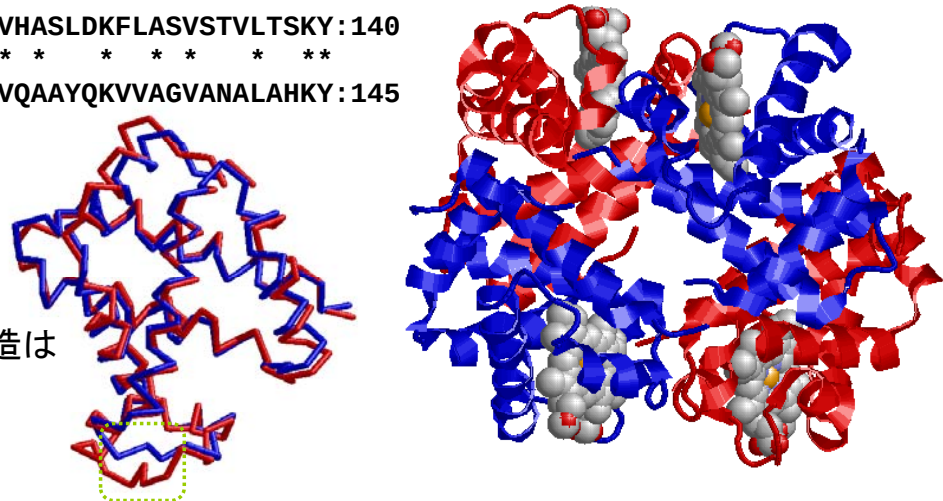
Alpha 56: KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA:11
* *

Beta 61: KAHGKKVLGAFSDGLAHLNLIKGTFFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK: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

ペアワイズアライメント

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

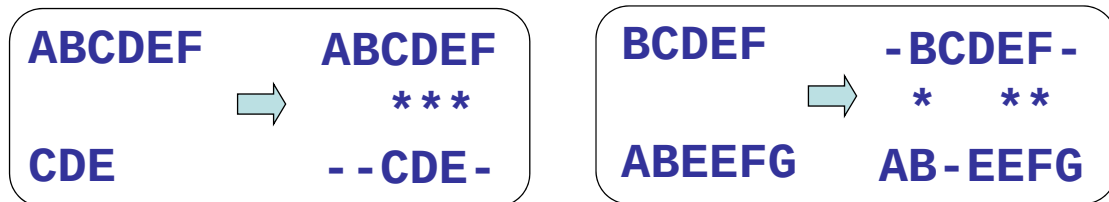
文字の間の類似性をどうやって定量するか？

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？

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、一荷電) : 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 : XXXXA XXXX

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。

[illegible]

スコアの計算例

$$\begin{array}{l} \text{AFDC} \\ \text{AEEC} \end{array} \quad \begin{array}{cccc} S(A,A) & + & S(F,E) & S(D,E) + S(C,C) = 12 \\ 4 & & -3 & 2 \quad 9 \end{array}$$

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

$$\begin{array}{l} \text{AFDGC} \\ \text{AEE-C} \end{array} \quad \begin{array}{cccccc} S(A,A) & + & S(F,E) & + & S(D,E) & + \text{gap} + S(C,C) = 10 \\ 4 & & -3 & & 2 & -2 \quad 9 \end{array}$$

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	ギャップあり	AFAED-C
	AEEC		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	➡	ACDEFGHK-LM	FGHK-L
AFGHKKL		A---FGHKKL-	FGHKKL
		グローバル	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。

配列1

1: GCTAGACTCG

2: AGCTAGACTC

(1) 配列1、配列2を
横と縦に並べる

		G	C	T	A	G	A	C	T	C	G
配列2 ↓	A										
	G										
	C										
	T										
	A										
	G										
	A										
	C										
	T										
	C										

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。

配列1

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

		G	C	T	A	G	A	C	T	C	G
配列2 ↓	A				○		○				
	G	○				○					○
	C		○					○		○	
	T			○					○		
	A				○		○				
	G	○				○					○
	A				○		○				
	C		○					○		○	
	T			○					○		
	C		○					○		○	

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。

配列1

1: GCTAGACTCG

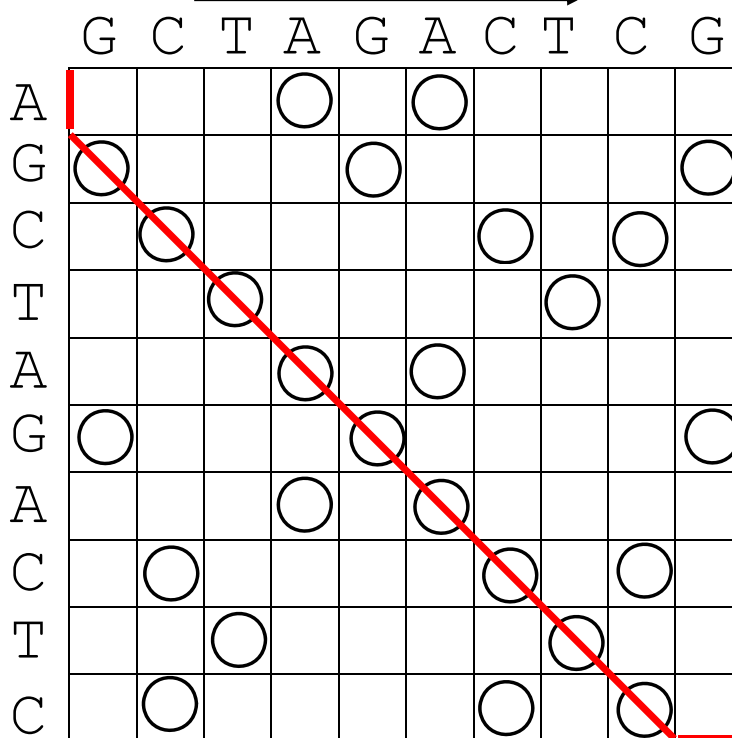
2: AGCTAGACTC

(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

配列2



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。

配列1

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

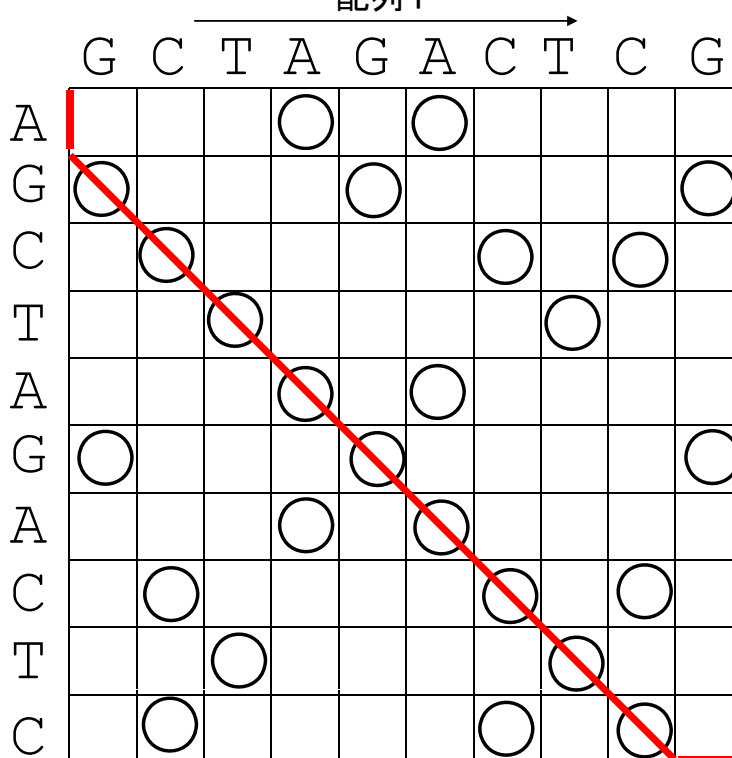
(3) 多くの○を通るような左上と右下を結ぶ折れ線

(4) アライメント

1: -GCTAGACTCG

2: AGCTAGACTC-

配列2

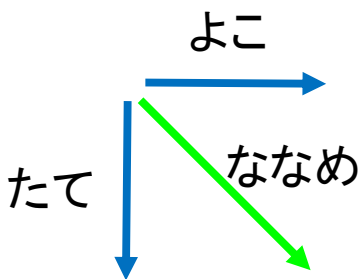


スコア: 一致(+1) × 9 + 不一致(0) × 0 + ギャップ(-1) × 2 = 7

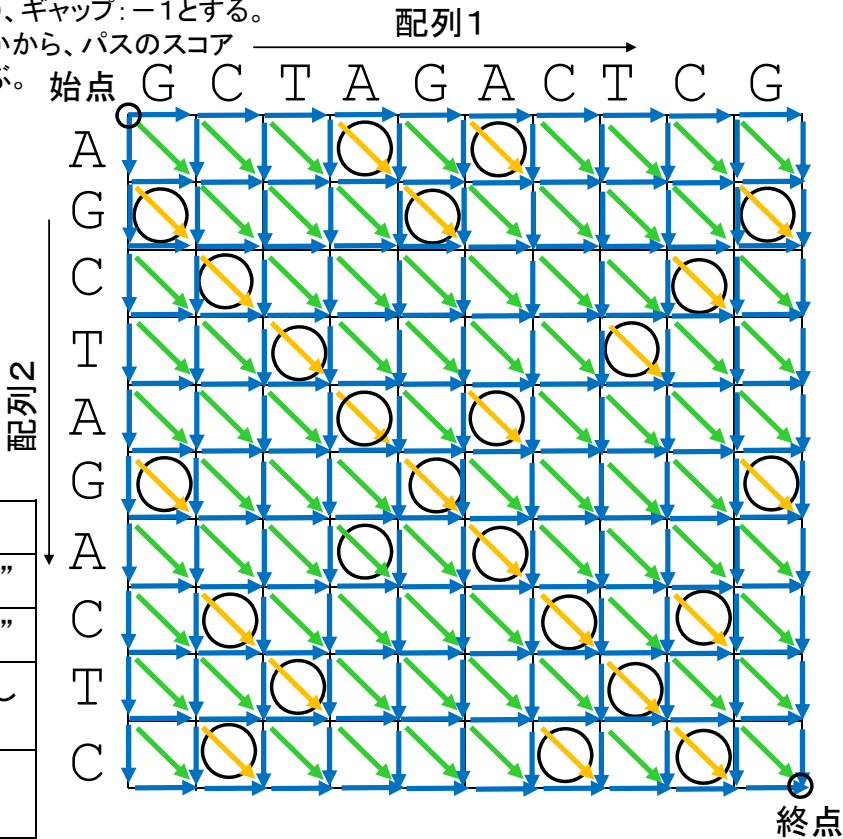
ドットマトリックスのパスの引き方の詳細

※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

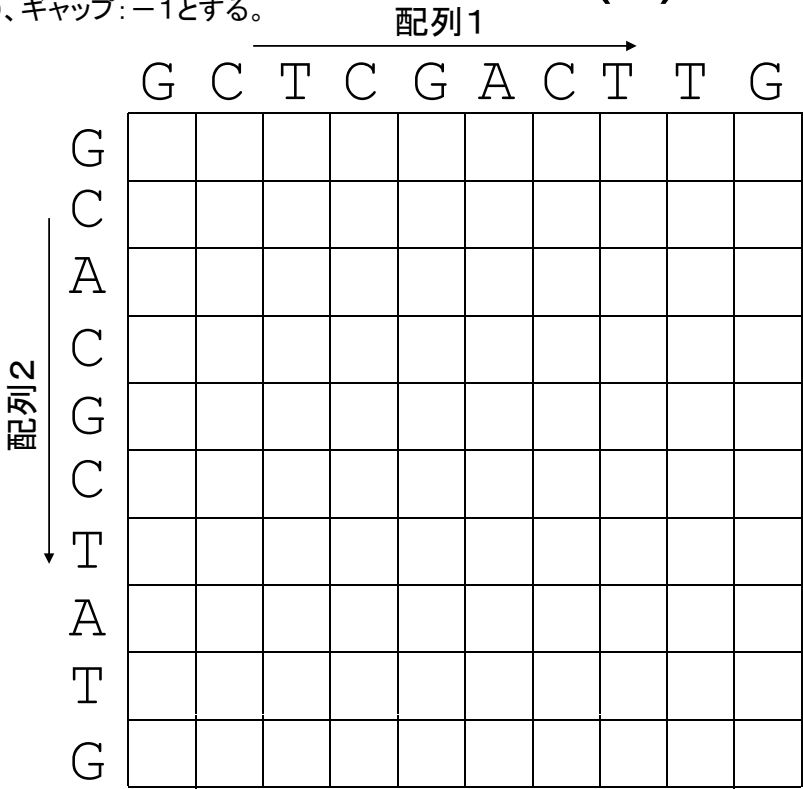


ドットマトリックス：例2 (1)

※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 (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○を描く

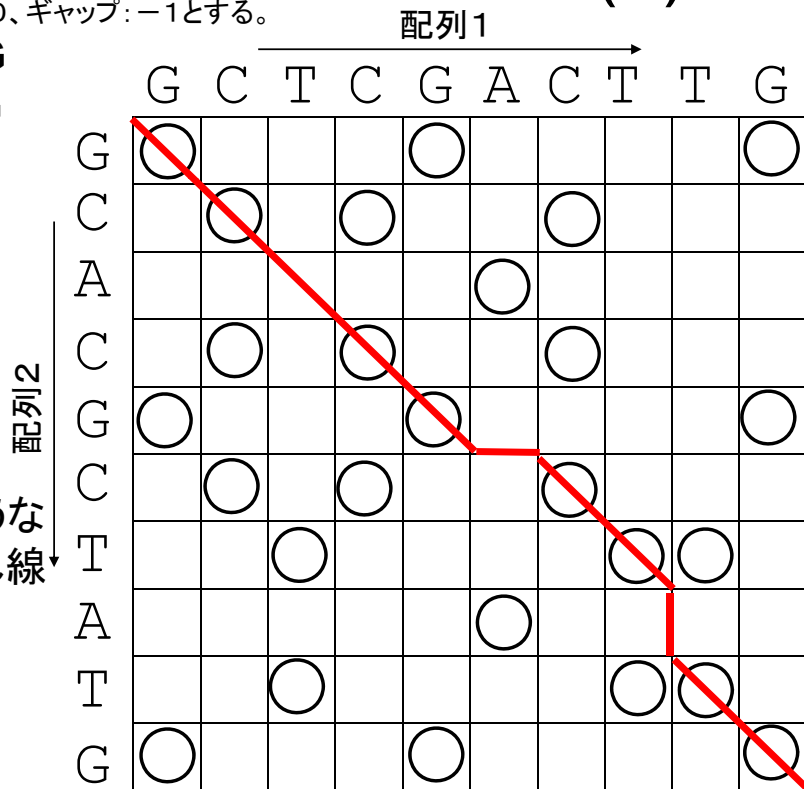
(3) 多くの○を通るような左上と右下を結ぶ折れ線

(4) アライメント

1: GCTCGACT-TG

** ** *

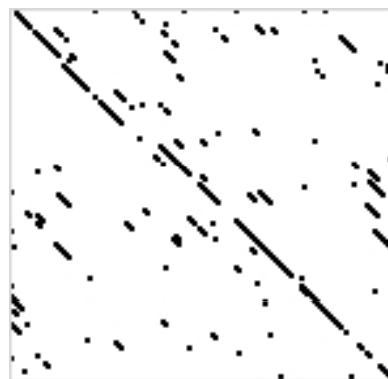
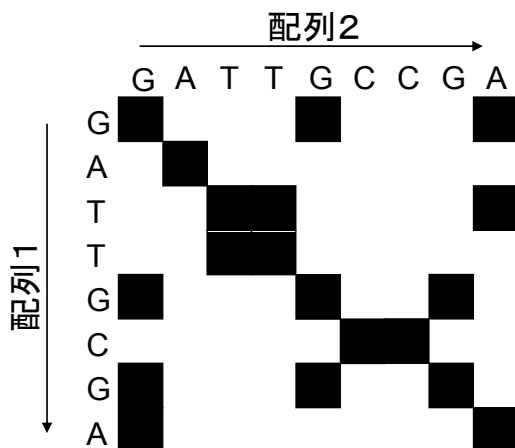
2: GCACG-CTATG



スコア: 一致(+1) × 8 + 不一致(0) × 1 + ギャップ(-1) × 2 = 6

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。

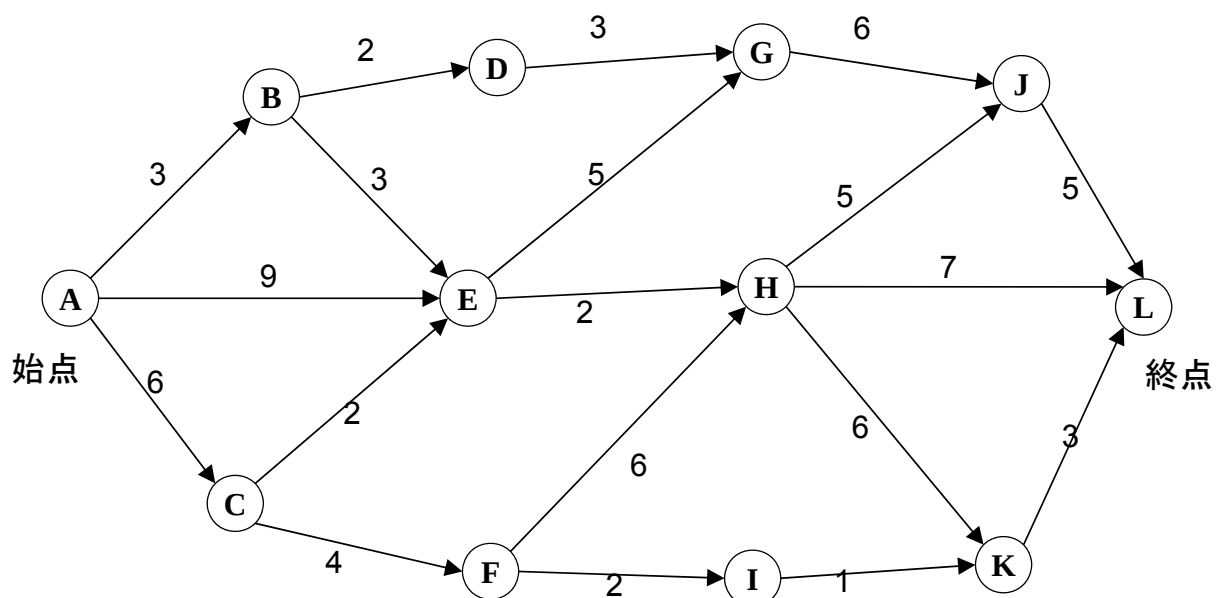


動的計画法によるアライメント

- 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)

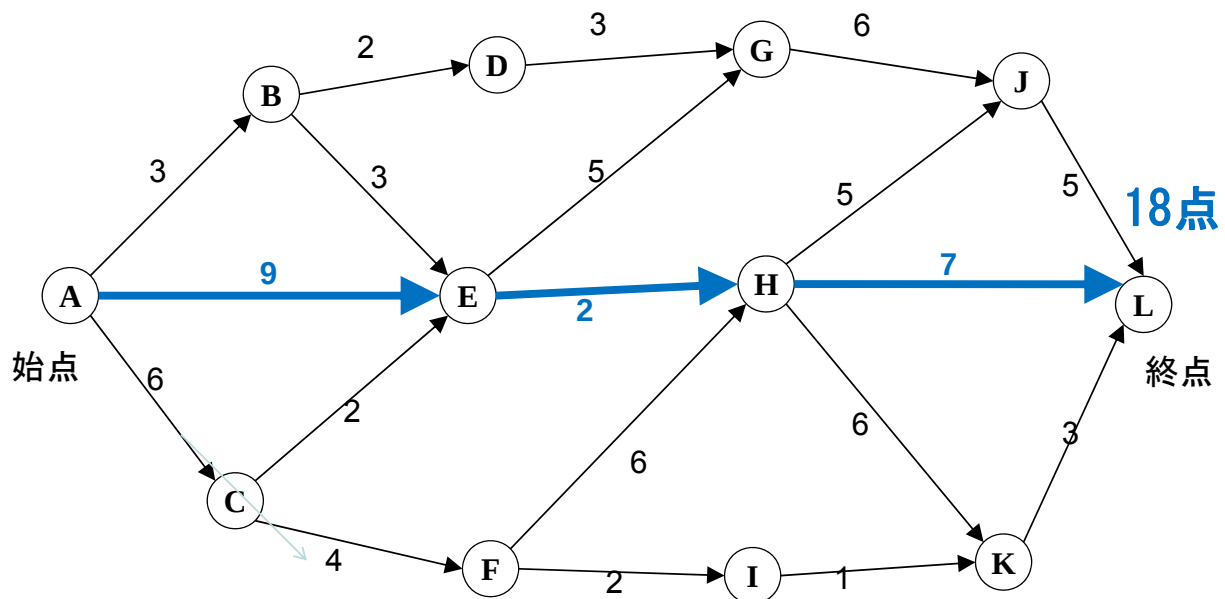
最適経路問題

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



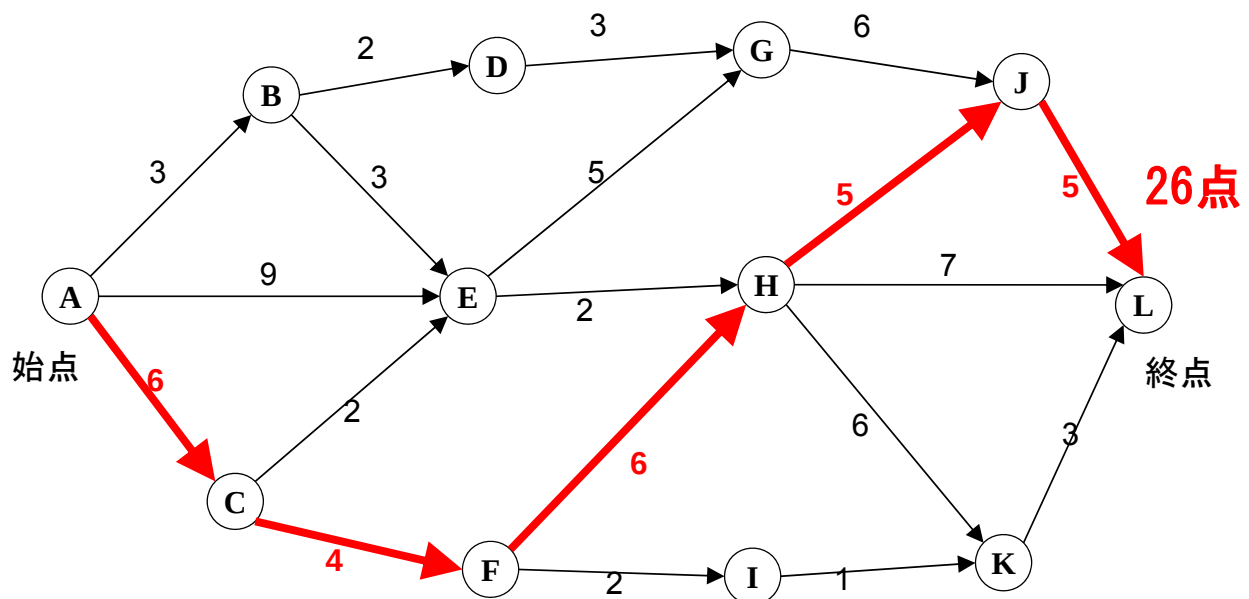
最適経路問題

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



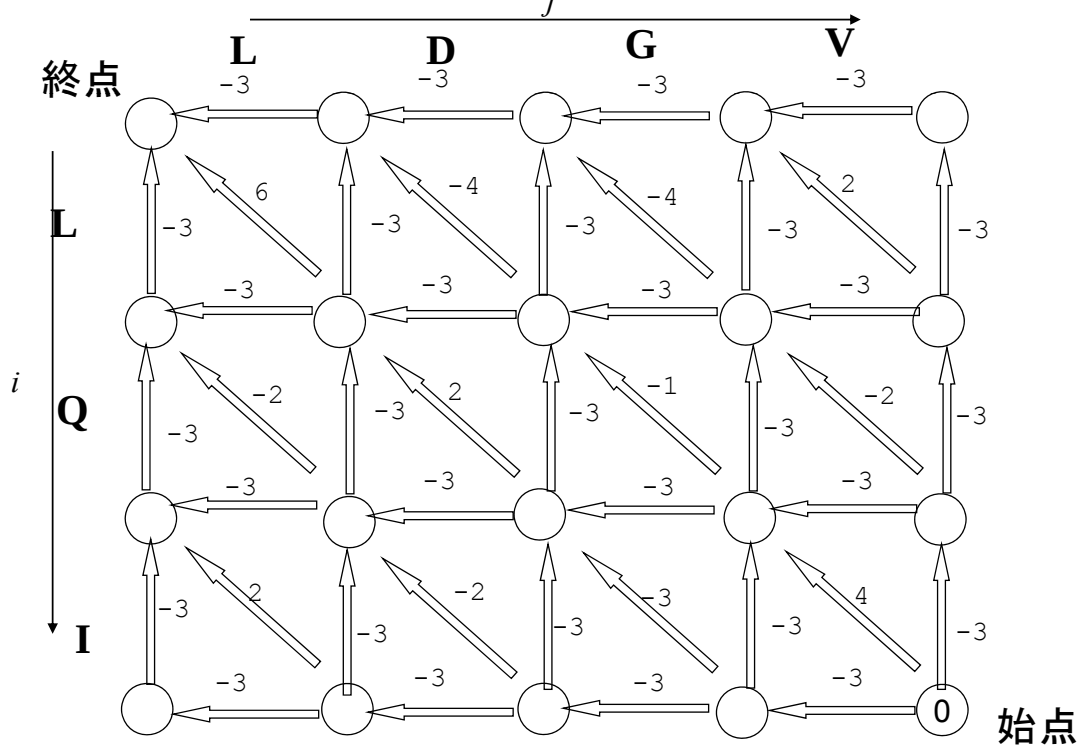
最適経路問題

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



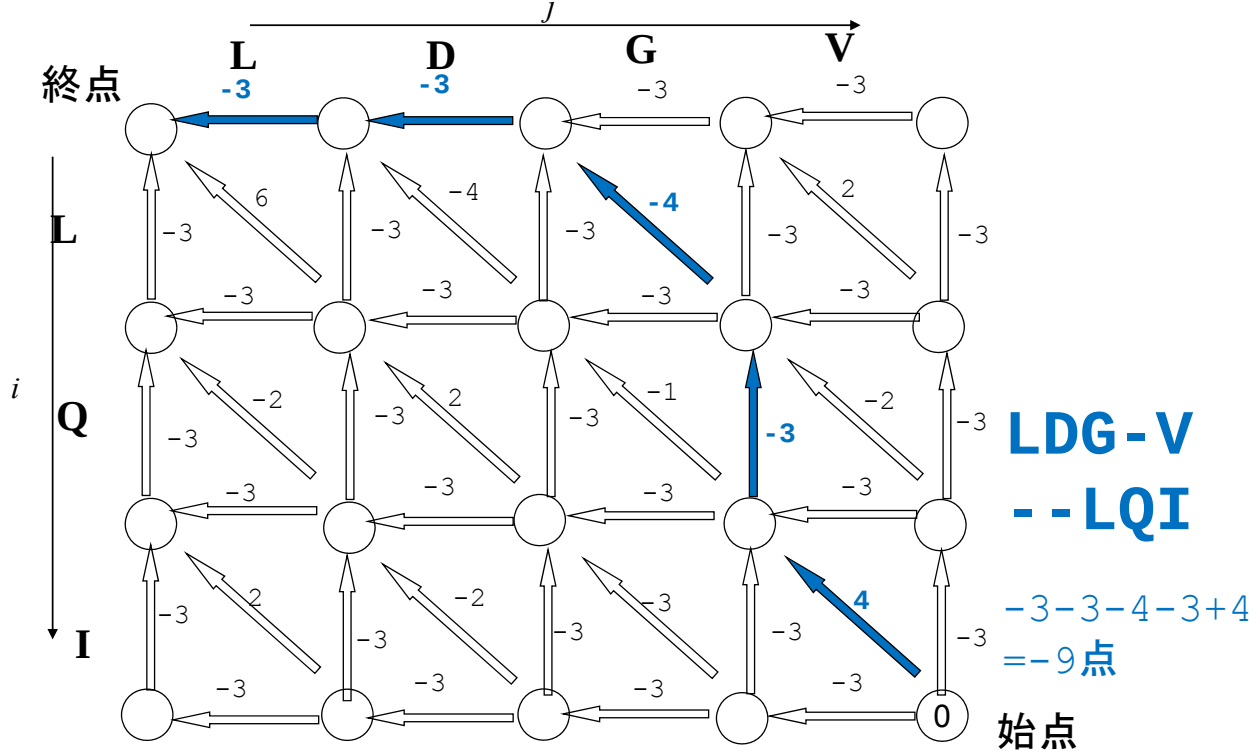
アライメントを最適経路問題として考える

- 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



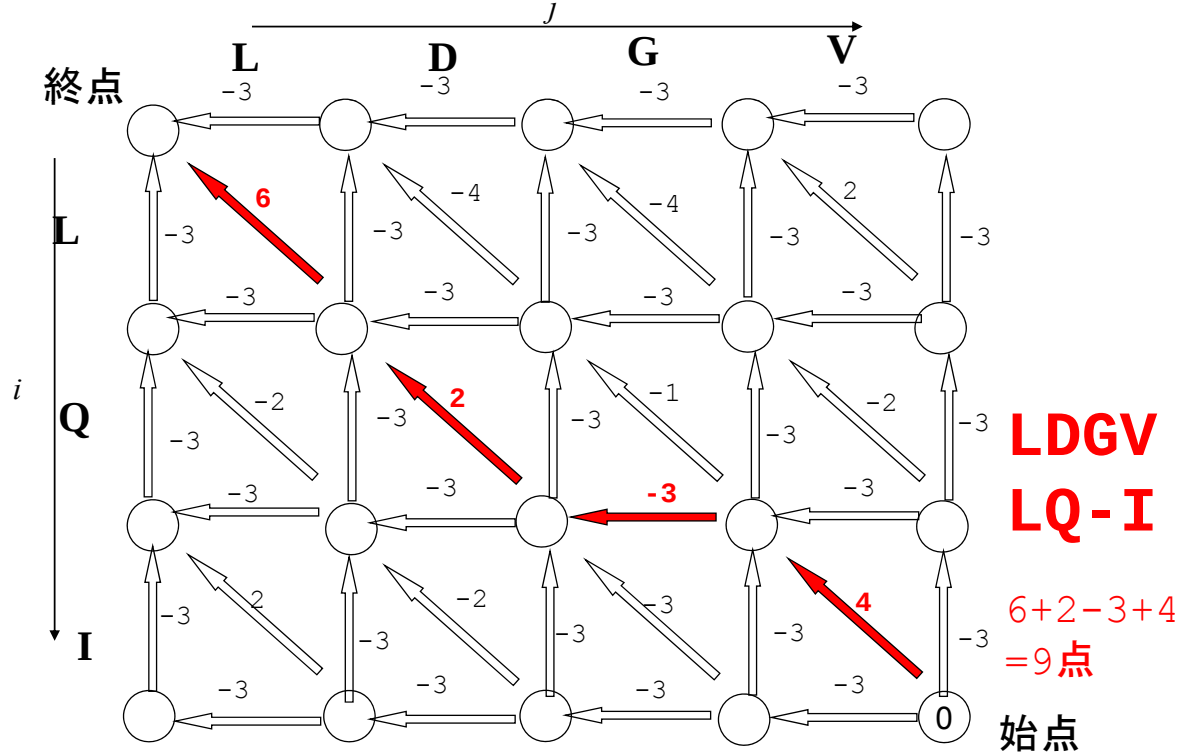
アライメントを最適経路問題として考える

- 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 & Wunsch, 1970)

(0) 準備

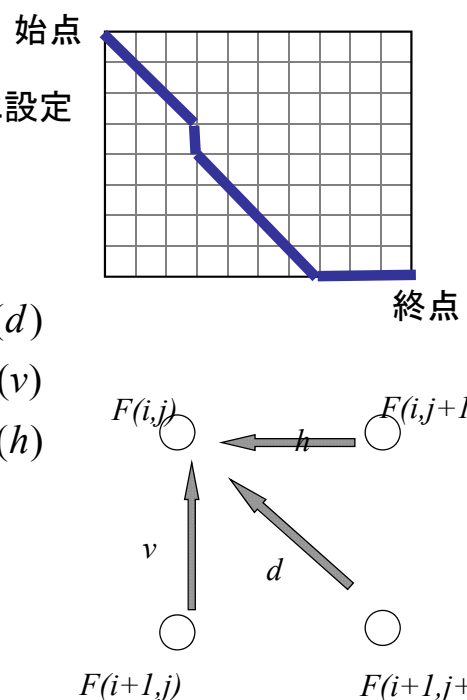
右端の列、下端の行の格子点のスコアを0に設定

(1) 前向きステップ

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

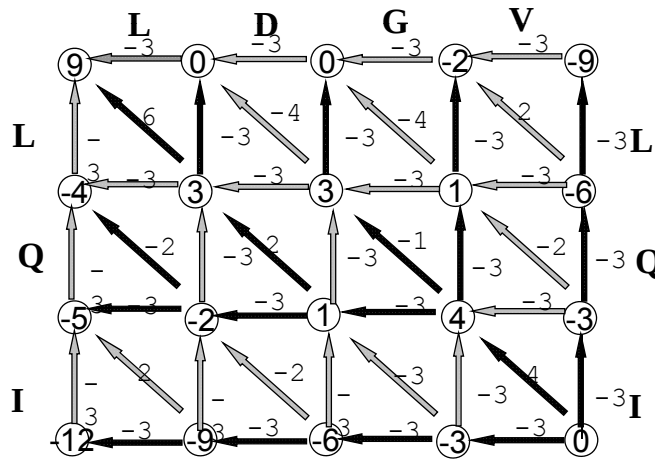
(2) 後ろ向きステップ

始点を起点にして辿る。終点に到着したら終了。



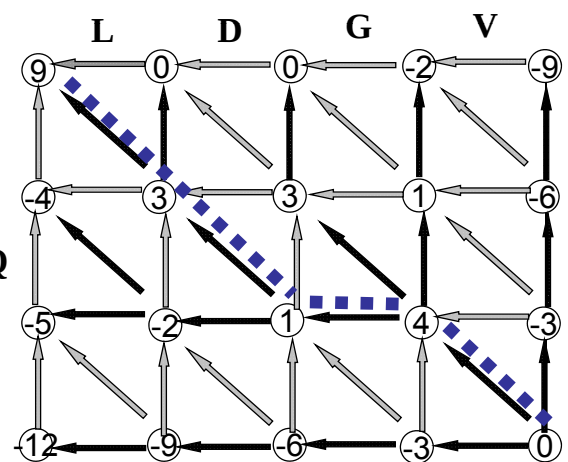
動的計画法の手続き

(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

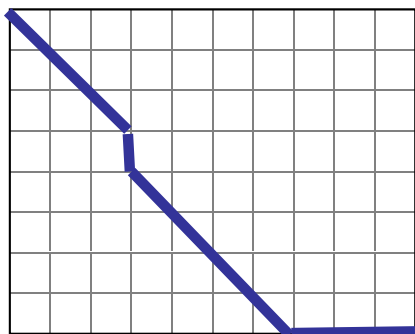


ACDEFGHK-LM
A---FGHKKL-

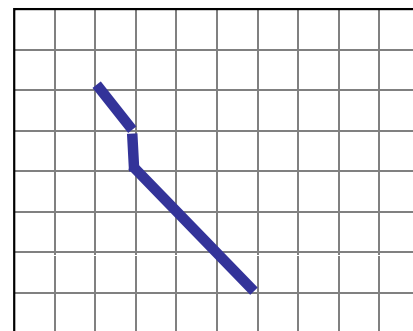
グローバル

FGHK-L
FGHKKL

ローカル



グローバル

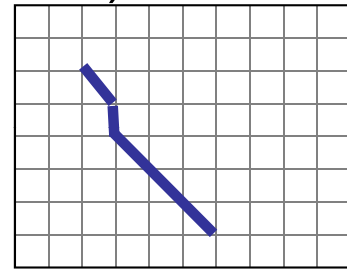


ローカル

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

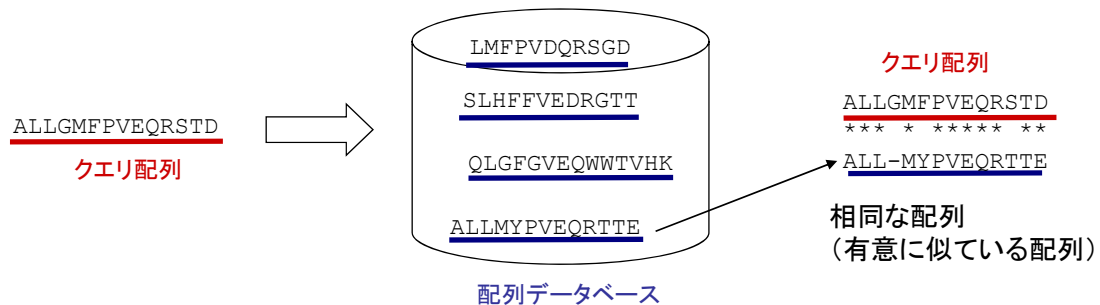
最大のスコアのノードを探し、そのノードを起点にして辿る。パス'0'が現れたら終了

配列相同性検索

— BLASTを中心として —

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

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

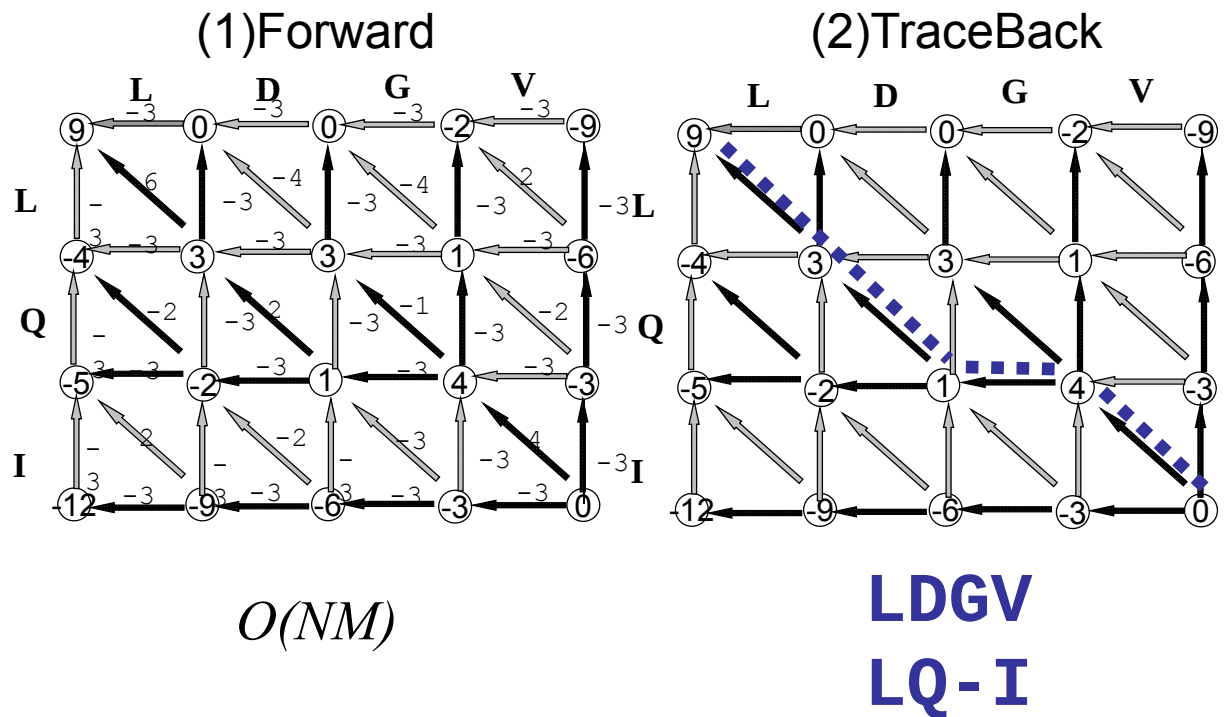
ヒューリスティック: 常に正しい解を返すわけではないが、多くの場合まあまあの解を返すことが経験的に知られているアルゴリズム

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

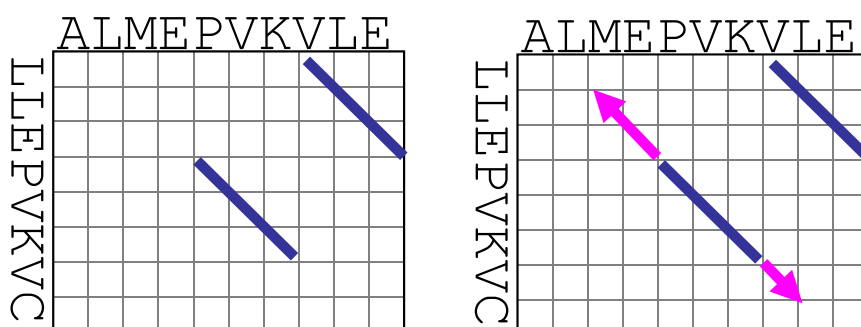
動的計画法の復習



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)

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

Searching.....done

Sequences producing significant alignments:

	Score (bits)	E 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

1b1bA [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 EYGPRESSGKTTTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPD TG 122
EYGPRESSGKTTTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPD TG
Sbjct: 61 EYGPRESSGKTTTLQVIAAAQREGKTCAFIDAEHALDPIYARKLGVDIDNLLCSQPD TG 120

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

Query: 183 KQNTLLIFINQIRMKIGVMFGNPETTGGNALKFYASVRLDIRRIGAVKEGENVVGSET 242
KQNTLLIFINQ TGGNALKFYASVRLDIRRIGAVKEGENVVGSET
Sbjct: 173 KQNTLLIFINQ-----TGGNALKFYASVRLDIRRIGAVKEGENVVGSET 217

Query: 243 RVKVVKNKIAAPFKQAEFQILYGE GI 268
RVKVVKNKIAAPFKQAEFQILYGE GI
Sbjct: 218 RVKVVKNKIAAPFKQAEFQILYGE GI 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 GELVDLGVKEKLIKAGAWYSYKGEKIGQKANATAWLKDNPETAKEIEKKVRELL 327
G L+D+GV + LI K+GAW++Y+GE++GQGK NA +L +N + A EIEKK++E L

BLASTの
出力例(2)

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

BLASTの 出力例(3)

```
>lg18A2 [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 GELVDLGVKEKLEKAGAWYSYKGEKIGQGANATAWLKDNPETAKEIEKKVRELL 327
      G L+D+GV + LI K+GAW++Y+GE++GQGK NA +L +N + A EIEKK++E L
Sbjct: 4   GSLIDMGVDQGLIRKSGAWFTYEGEQLGQKKNARNFLVENADVADIEKKIKEKL 59
```

```
>lg0uF [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 YGELVDLGVKEKLEKAGAWYSYKGEKIGQGANATAWLK----DNPE--TAKEIEKKVR 324
      +G +   G   ++E +G+++ YKG   G+G+ +A A L+   +PE +A+E K+
Sbjct: 132 FGGVDKNGAHLMLPSGSYWGKYKAATGKGRQSAKAELEKLVDHHPGLSAREAVKQAA 191
```

```
Query: 325 EL--LLSNPNSTPDFSVDDSE-GVAETN 349
      ++ L   N   DF ++ S   ++ETN
Sbjct: 192 KIIYLAHEDNKEKDFELEISWCSLSETN 219
```

```
>lbyrA [d.136.1.1] ENDONUCLEASE
      Length = 152
```

```
Score = 28.5 bits (62), Expect = 0.29
Identities = 28/102 (27%), Positives = 46/102 (44%), Gaps = 19/102 (18%)
```

```
Query: 65 YGESSCKTTITIOVIAAOPCKTCATF----DAFHAI DPTVAPKICVDIDNII CSQPD 120
```

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

* * 4/8 = 50 %

SELA

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。

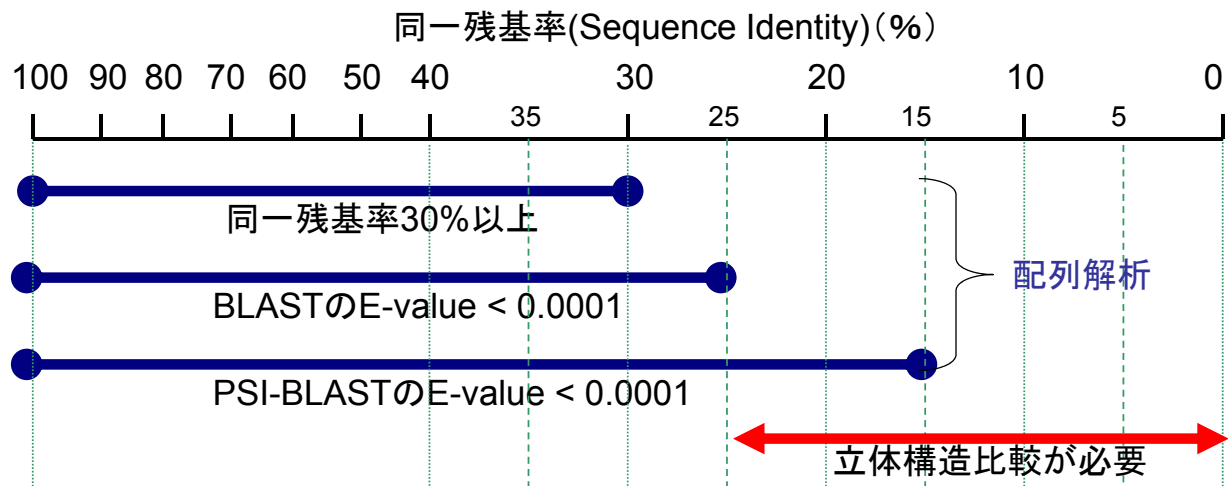
E-valueの計算に必要なパラメータ

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

- ・ 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。

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。

blastxやtblastnにおけるアミノ酸と核酸配列の比較

AGCTTTTTCATTCTGACTGCA
| | | | | | | | | | | | | | | | |
TCGAAAAACAAGACTGACGT

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

AGCTTTTTCATTCTGACTGCA
S F S F x L Q
A F H S D C
L F I L T A

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

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かくために」(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ーゲノム配列から機能解析へー」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/>