2014年度「理論分子生物学」講義予定表

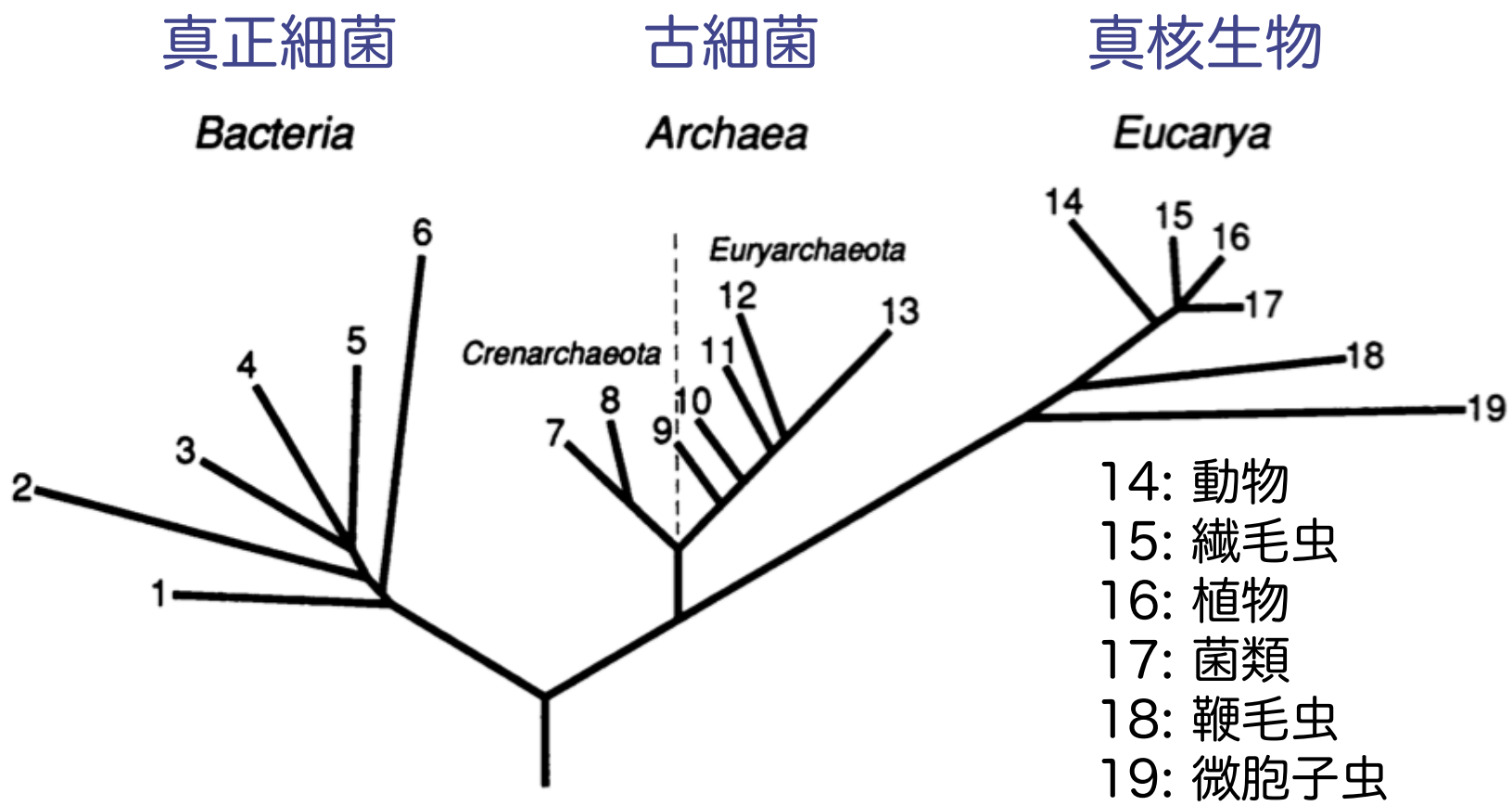
- ゲノム解析・オミックス解析とバイオインフォマティクス
- 配列アライメント：ダイナミックプログラミング法
- ホモロジー検索：FASTA、BLASTアルゴリズム
- マルチプルアライメント、系統樹解析
- 配列モチーフ
- 二次構造予測、膜貫通部位予測、立体構造予測
- 遺伝子の機能アノテーション、比較ゲノム解析
- メタゲノム解析
- ネットワーク解析
- 分子生物学データベース
- 演習

- http://goto.kuicr.kyoto-u.ac.jp/lecture/bioinfo.html

配列アライメント

- 配列アライメント（sequence alignment）
  - ２つのタンパク質または遺伝子の配列を並べて、進化的な関連があるかどうかを調べること
- ２つの遺伝子が進化的に関連があるか？
  - 異なる生物種間で同じ機能を持つ遺伝子
  - 一つの生物種内で類似した機能を持つ遺伝子
- 異なる生物種で同じ機能の遺伝子を比較する
  - 生物種間の進化的な関係

生物種の系統関係



14: 動物
15: 繊毛虫
16: 植物
17: 菌類
18: 鞭毛虫
19: 微胞子虫

ウーズらによる進化系統樹
分子レベル（配列レベル）の情報：16S rRNA

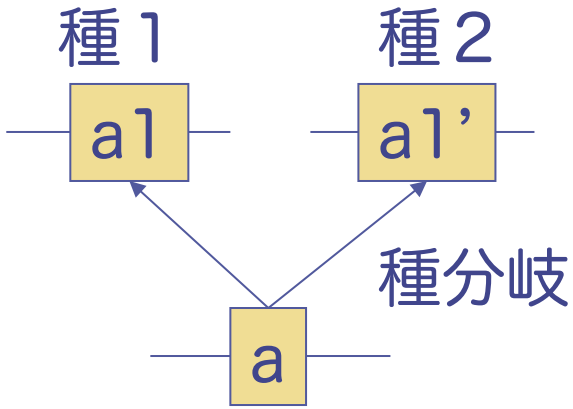
オーソログとパラログ

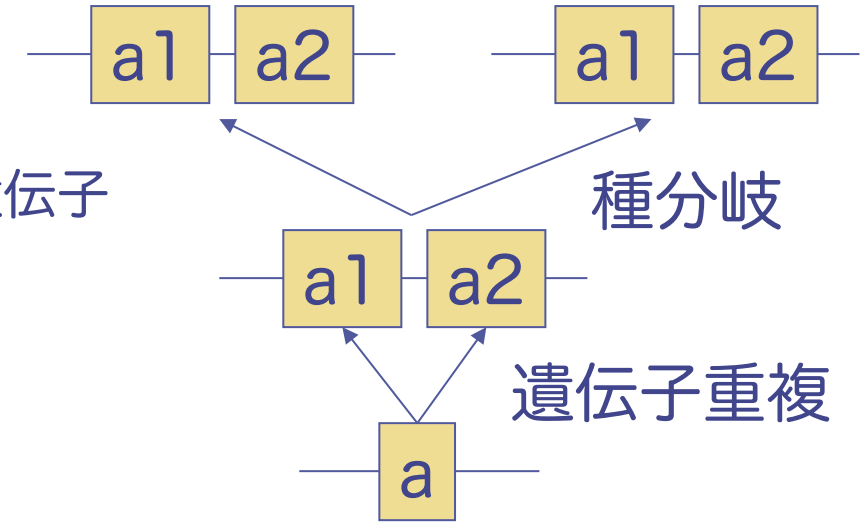
ホモログ（Homolog）
　進化的な起源を同じくする遺伝子

オーソログ（Ortholog）
　種分岐の際に同じ遺伝子だったもの
　通常同じ機能を持つ

パラログ（Paralog）
　遺伝子重複によってできた類似遺伝子
　通常異なる機能を持つ

ゼノログ（Xenolog）
　水平移動によって得られた類似遺伝子





配列アライメント

- 配列アライメント（alignment）
  - ２つのタンパク質または遺伝子の配列を並べて、ホモログ（相同）かどうかを調べること
  - 実際には類似性を調べる
  - 文字の一致（マッチ）、不一致（ミスマッチ）、挿入、欠失を考慮する

- アライメントのキーポイントは
  - アライメントの種類
  - アライメントの方法・アルゴリズム
  - アライメントを評価するためのスコア
  - スコアの重要性を評価するための統計的基準

配列アライメントの種類

◆グローバルアライメント

- 配列全体を並べる

```
LGPSSKQTGKGW-SRIWDN
|    +|   |||    |+ |
LN-ITKSAGKGAIMRLGDA
```

マッチ（+, |）
ミスマッチ
ギャップ・挿入（-）

◆ローカルアライメント

- 局所的によく似た部分を探す

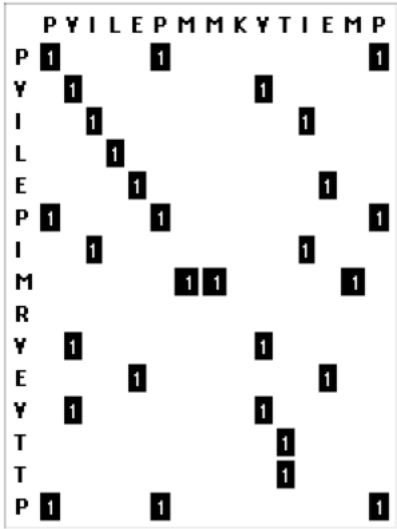
```
--------TGKG---------
        |||
--------AGKG---------
```

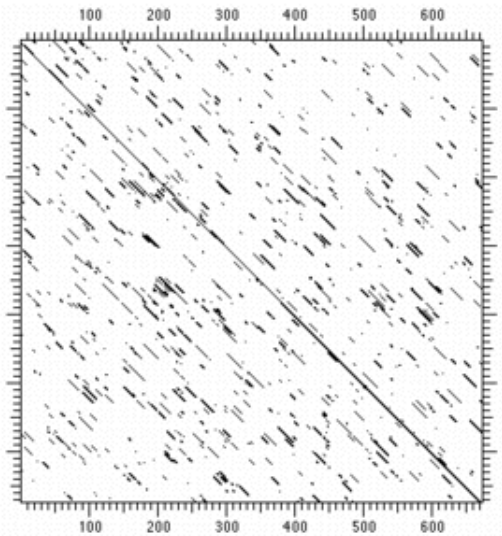
アライメント方法の種類

- ２本の配列のアライメント
  - ペアワイズアライメント
    - ドットマトリックス
    - ダイナミックプログラミング
    - ワードを使った方法
- ３本以上の配列のアライメント
  - マルチプルアライメント
    - 累進法
    - 反復改善法

ドットマトリックス

- 比較する配列を行列の行と列に並べて一致する文字に対応する行列成分にドットをうつ
- 類似部分をマトリックスの対角成分として表示
- 同一配列をマトリックス表示するとリピート部分も抽出できる

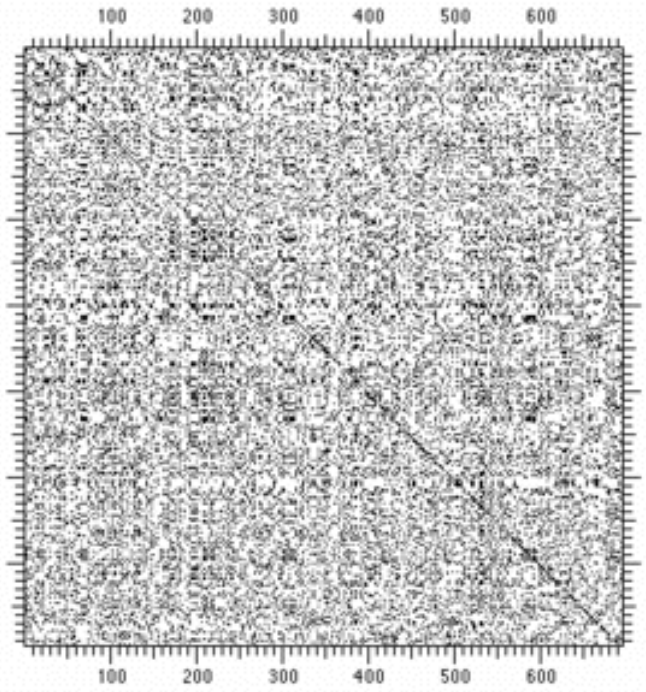




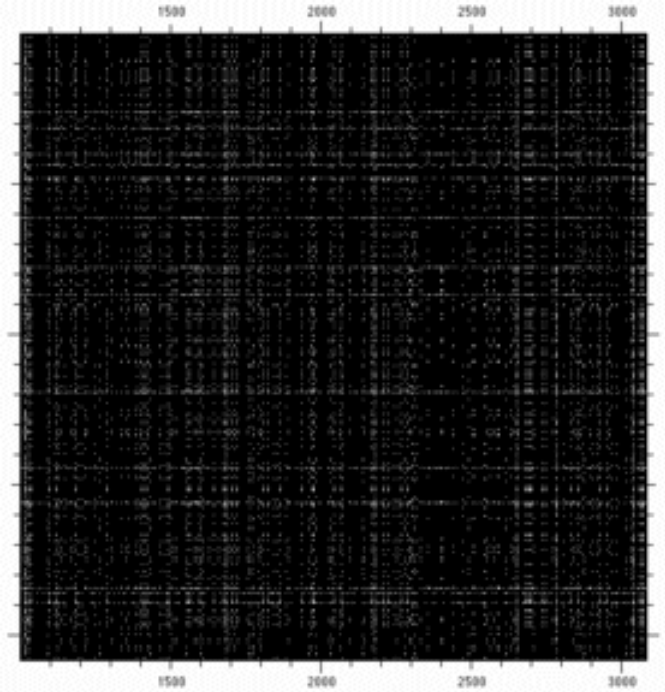
http://www.tulane.edu/~biochem/lecture/723/seqalign.html

ドットマトリックス

二つのバクテリアの EF-G（転写伸長因子）の例
S. aureus（黄色ブドウ球菌）と *T. thermophilus*（好熱菌）



タンパク質アミノ酸配列



DNA塩基配列

http://www.tulane.edu/~biochem/lecture/723/seqalign.html

KEGG **Staphylococc**

Entry	SA0505
Gene name	fus
Definition	elongation fa
Orthology	KO: K02355 e
Class	BRITE hierarchy
SSDB	Ortholog Para
Motif	Pfam: GTP_EFT PROSITE: EFAC Motif
Other DBs	NITE: SA0505 NCBI-GI: 1592 NCBI-GeneID: UniProt: P687
LinkDB	PDB All D
Position	588492..59057
AA seq	693 aa AA se MAREFSLEKTRNI ITITSAATTAAWE RQATTYGVPRIVF MKCFKYTNDLGTE KEAIRQATTNVEF AKADDSAEFAALA NSRQEIDTVYSGD DKMTQALVKLQEE SYRETFKSSAQVQ VEAGLKDAMENGV LEPMMKVTIEMPE NTQGRGTYTMYFD

KEGG **Thermosynechococcus elongatus: tlr1749**

Help

Entry	tlr1749 CDS T.elongatus
Gene name	fus
Definition	translation elongation factor EF-G [EC:3.6.5.3]
Orthology	KO: K02355 elongation factor EF-G
Class	BRITE hierarchy
SSDB	Ortholog Paralog Gene cluster
Motif	Pfam: GTP_EFTU MMR_HSR1 Miro GTP_EFTU_D2 EFG_IV EFG_C PROSITE: EFACTOR_GTP Motif
Other DBs	CYORF: tlr1749 Kazusa: tlr1749 NCBI-GI: 22299292 NCBI-GeneID: 1011972 UniProt: Q8DI43
LinkDB	PDB All DBs
Position	1829639..1831714 Genome map
AA seq	691 aa AA seq DB search MARTTPLERVRNIGIAAHIDAGKTTTTERILFYSGVVHKIGEVHEGTTVTDWMEQERERG ITITAAAISTSWRDHQINIIDTPGHVDFTIEVERSMRVLDGVIAVFCSVGGVQPQSETVW RQADRYSVPRIVFVNKMDRTGANFYKVHDQIRDRLRANAVPIQLPIGAEDQFKGIVDLVR MRAKIYKDDLGKEIEDTEIPAEMTELAQEYRTKLIEAVAETDDALMEKYFEGEELTEEEI RAALRKGTIAGTIVPMLCGSAFKNKGVQLLLDAVVDYLPAPIDIPAIKGRLPDGTEVERA ADDDQPLAALAFKIMSDPYGRLTFVRVYSGVLKKGSYVLNATKGKKERISRLIVLKADER IEVDELRAGDLGAALGLKETFTGDTLCDESSPVILESLYIPEPVISVAVEPKTKQDMEKL SKALQALSEEDPTFRVSVDPETNQTVIAGMGELHLEILVDRMQREYKVEANIGQPQVAYR ETIRKPVRAEGKFIRQSGGKGQYGHVVIEVEPAEPGTGFEFVSKIVGGVVPKEYIPPAEQ GMKEACESGILAGYPVIDLKVTLVDGSYHEVDSSEMAFKIAGSIAIKEAVMKANPVLLEP MMKVEVEVPEEFLGTVMGDLIARRGQIEGQTVENGIAKVTAKVPLERMFGYATDIRSNTQ GRGIFSMEFSHYEEVPRNVAEAIIAKNKGNA

NT seq

2082 nt NT seq +upstream 0 nt +downstream 0 nt

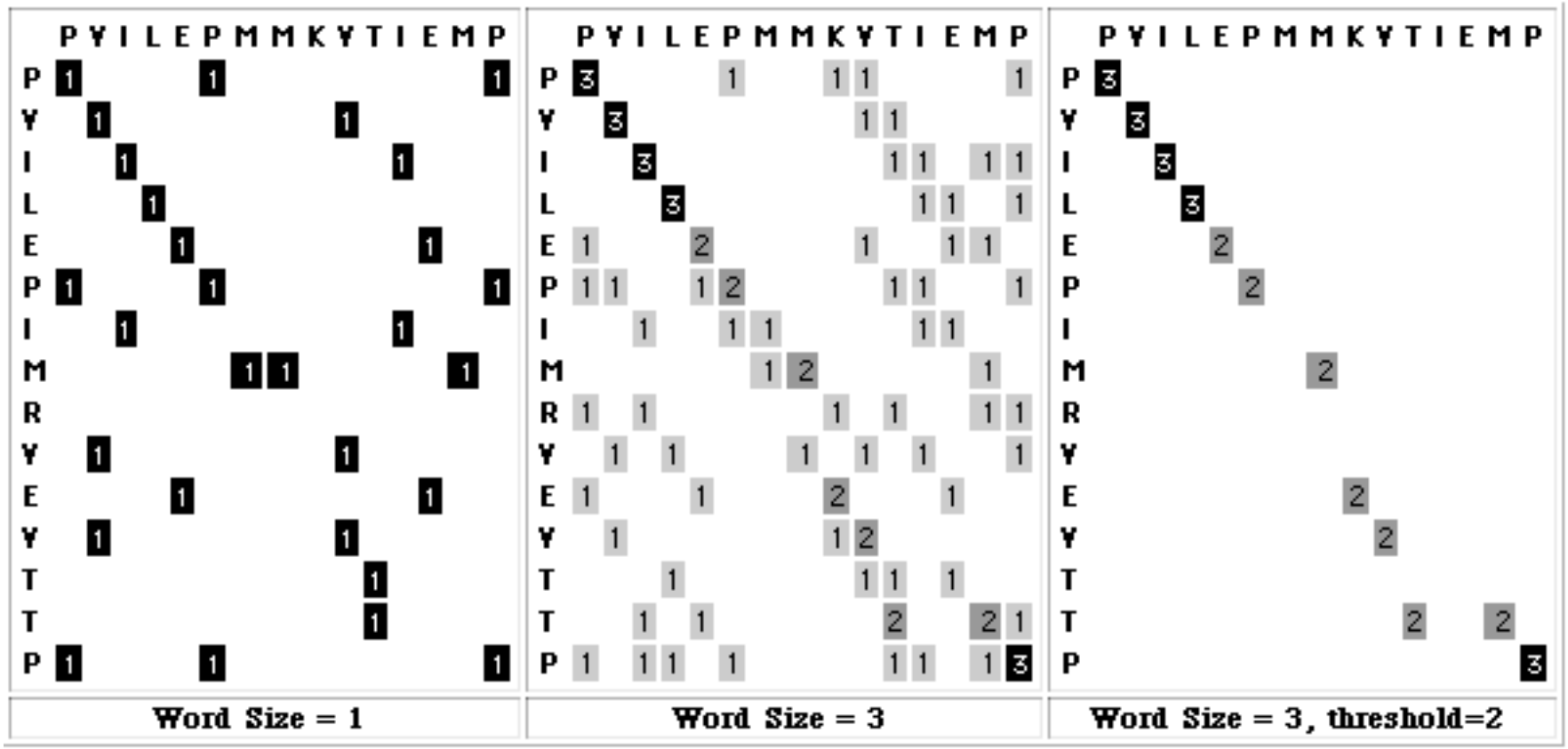
```
atggctagagaatt
gctggtaaaacgac
ggtgaaacacacga
attactatcacatc
gatacacctggaca
ggagcagttacagt
cgtcaggctacaac
ggtgctaacttcga
ccaatccaattacc
atgaaatgtttcaa
gaagaccacttaga
actagcgacgaatt
aaagaagctatccg
gctttcaaaacaa
ccactagacgttaa
gcgaaagcagacga
tatgttggtaaatt
gttaagaactctac
aactcacgtcaaga
aaagatacaggtac
atggaattcccaga
gataaaatgactca
actgacgaagaaac
ttagtagaccgtat
tcatatcgtgaaac
ggtggtcgtggtca
ggtttcgaattcga
gtagaagctggtct
gatgttaaagctaa
ttcaaaattgctgc
ttagaaccaatgat
ggtgacgtaacatc
gttgttaatgctta
aacactcaaggtcg
tcaatcgctgaaga
```

NT seq

2076 nt NT seq +upstream 0 nt +downstream 0 nt

```
gtggcacggacgaccccgctagagcgagtgcgaaatatcgggattgccgctcacattgat
gcgggtaaaacaaccacaactgaacgtattctcttctattccggtgtggtgcacaaaatt
ggtgaagtgcacgaggggaccacggttaccgactggatggaacaggagcgagagcggggc
atcaccattacagcagccgccattagcacctcttggagagatcaccaaatcaacattatt
gacactcccggccacgtggacttcacgattgaggtggagcgctccatgcgggttctcgat
ggggtgattgcggtattctgctccgttggtggtgtgcagccccaatcagaaaccgtgtgg
cggcaagccgatcgctacagcgttccccgcatcgtttttgtcaacaagatggatcgcacg
ggggcaaacttctacaaagtccatgaccagattcgcgatcgcctacgggcgaacgccgtg
ccgattcagttgcccattggtgctgaagatcaattcaaaggcatcgttgatctggtgcgc
atgcgcgccaaaatctacaaagacgaccttggcaaggaaatcgaagacactgaaattcct
gccgagatgactgaactggcacaggagtatcgcaccaagctcatcgaggctgttgccgaa
accgacgacgccctcatggaaaaatacttcgagggggaagaactcacagaagaggaaatc
cgtgcggctctacgcaaagggacgatcgctggcaccattgtgcccatgctttgcggctct
gccttcaaaaataaaggcgtgcaattactcctggatgccgtggtggactacctaccggcg
cccattgacattcctgccattaaaggtcgtctgcccgatggcacagaagtggagcgggct
gccgatgacgatcaacccctggcggcactcgcttttaagatcatgtcggatccctatggc
cgccttacctttgtgcgtgtttattccggtgttctgaaaaaaggtagctatgtgctcaat
gccaccaaaggcaagaaagagcgcatttctcgtctaatcgtgctgaaggccgatgaacgg
attgaggtggatgaattgcgagctggcgatctaggagctgccttaggtctcaaggaaacc
ttcacaggggataccctctgcgatgaaagctcgcccgtgattcttgagtcgctgtatatt
ccagagccagttatttccgtggccgtagaacccaaaactaaacaggacatggaaaagctc
tccaaggctctgcaagccctctctgaggaggaccccaccttccgcgtcagcgttgatcct
gaaacgaaccaaaccgtgattgcgggaatgggggaactacacctcgagattcttgtggat
cggatgcagcgggagtacaaagtggaagccaacatcggccaaccccaagtggcttaccgc
gaaaccattcgcaaacccgtccgggctgaaggcaagtttatccgtcaaagtggtggtaaa
ggtcagtatggtcacgttgttattgaagtagaacctgctgaacccggtactgggtttgaa
tttgtctccaaaattgtcggtggtgtcgtacccaaagagtacattccacccgctgagcag
gggatgaaggaagcctgtgaatcggggattttggcagggtacccagtgattgacctcaaa
gtcacactagtggatggctcctaccacgaggtggactcctcagagatggctttcaaaatt
gctggctccatcgccattaaagaggcagtcatgaaagccaatccagtactgctggagcca
atgatgaaggtcgaggtcgaggttccagaagaattccttgggacagtgatgggagatcta
attgcccgtcgcggtcaaattgagggacaaaccgtagaaaatggcattgccaaagtgacg
gcaaaagttcctctcgaacggatgtttggttatgccactgacattcgctcgaatacccaa
ggtcggggaattttctcgatggaatttagccattacgaggaagtcccccgtaatgtggct
gaggcgatcattgccaaaaataaagggaacgcatag
```

ワードサイズによる評価

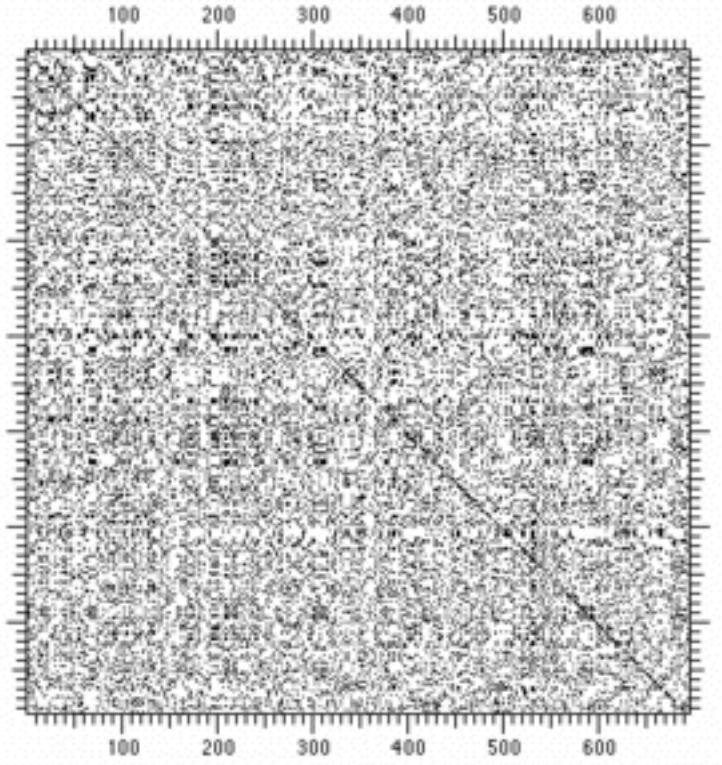


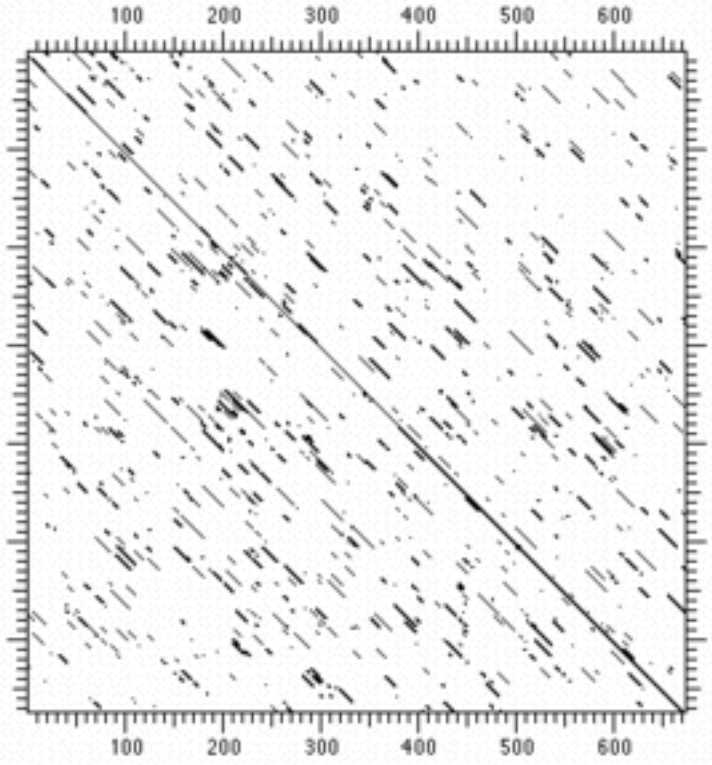
1文字ずつ比較して、一致するところにドット

3文字ずつ比較して、完全一致するところにドット

３文字ずつ比較して、そのうち２文字一致するところにドット

ワードサイズによる評価

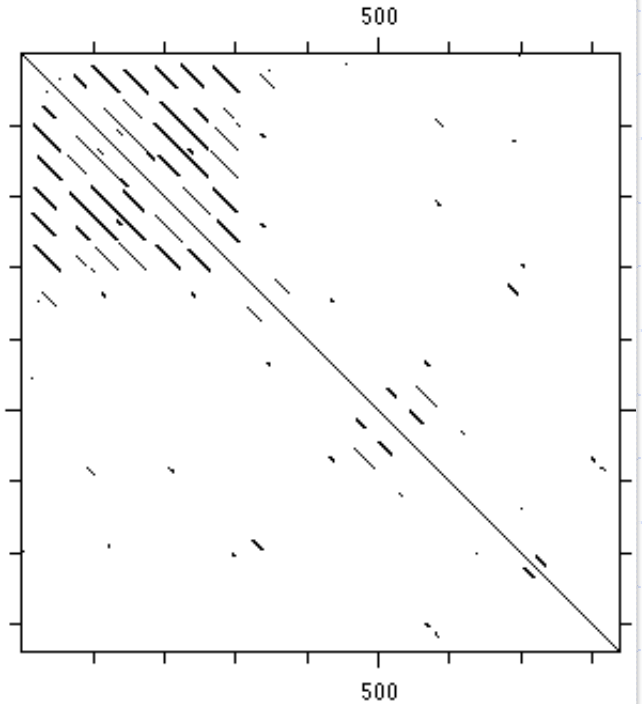




ワードサイズ 23、閾値 5（23文字のうち5文字） の
マッチでドット

ドットマトリックスによるリピート同定

human LDL (low density lipoprotein) receptor



EGF_CA:

calcium binding EGF domain

タンパク質間相互作用に必要

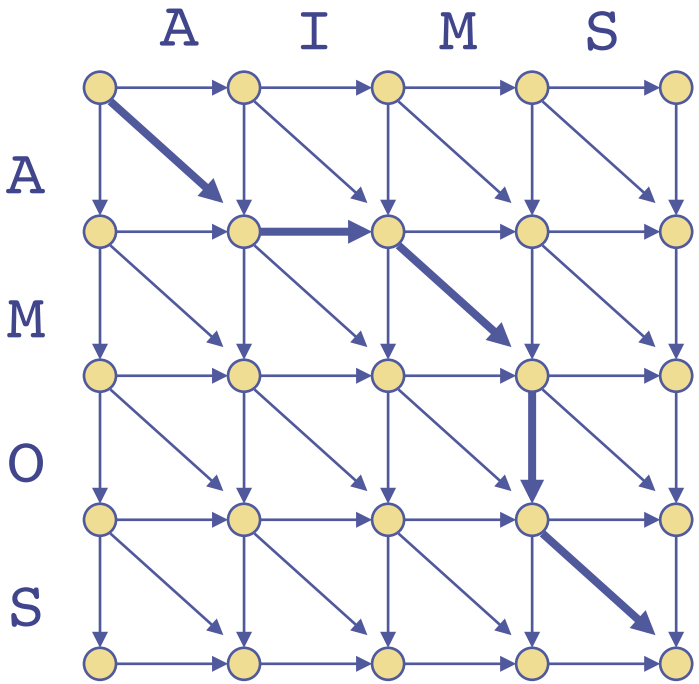
http://www.blc.arizona.edu/courses/bioinformatics/matrix-lect.html

アライメント方法の種類

- ２本の配列のアライメント
  - ペアワイズアライメント
    - ドットマトリックス
    - ダイナミックプログラミング
    - ワードを使った方法
- ３本以上の配列のアライメント
  - マルチプルアライメント
    - 累進法
    - 反復改善法

ダイナミックプログラミング（DP）

ドットマトリックスでは対角線成分だけ
→　ギャップも考慮したい（塩基の挿入・欠失）

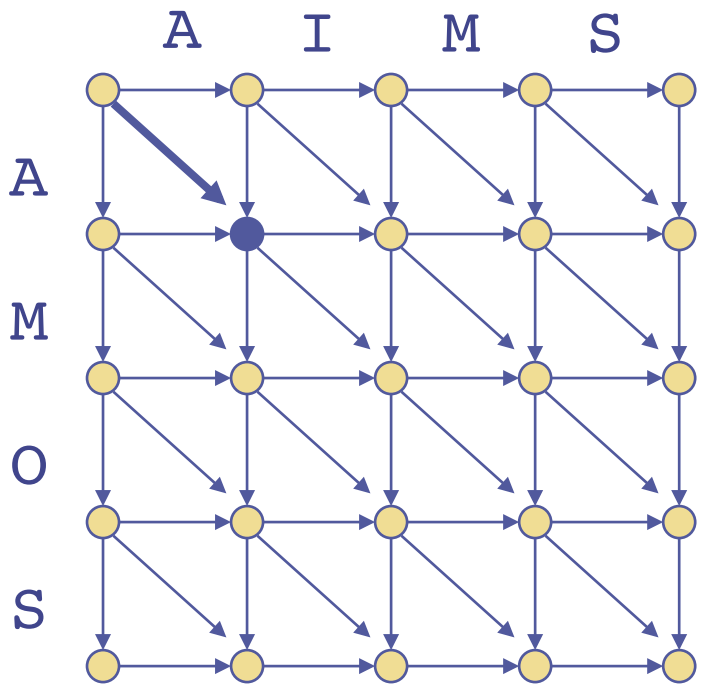


全パターンの並べ方から最適なものを探し出す問題

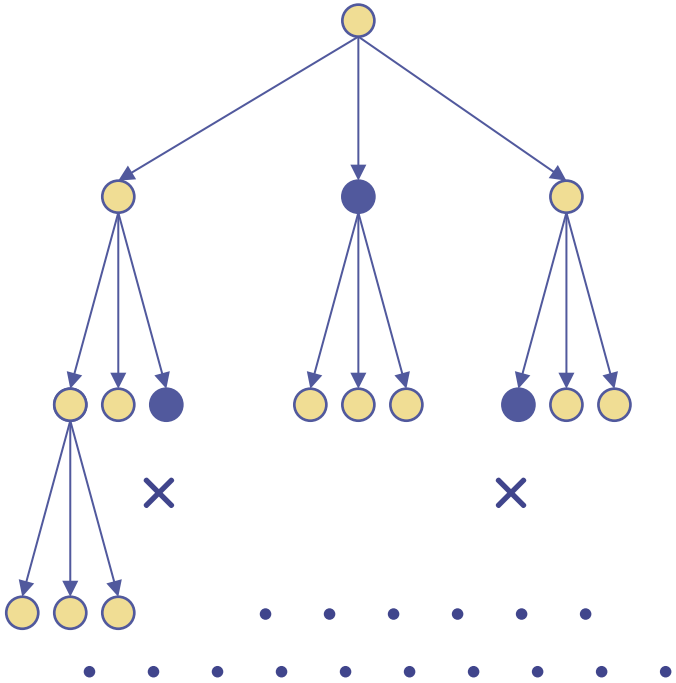
全パターン
　左上から右下まで辿る経路
　縦と横の矢印がギャップ

アライメント　AIM-S
A-MOS

ダイナミックプログラミング（DP）

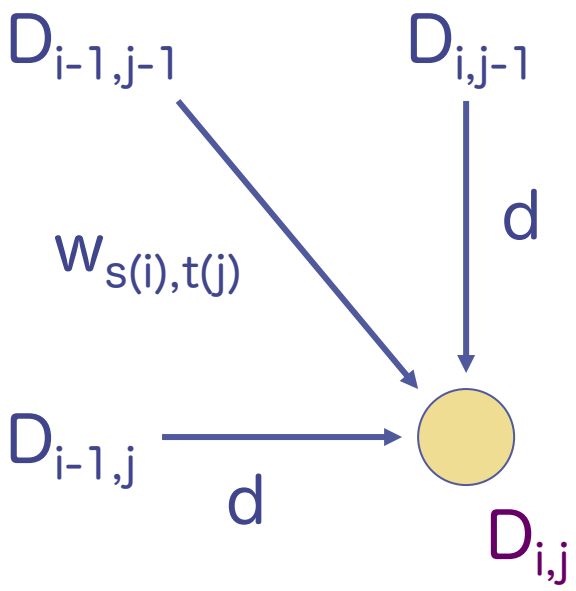


DP計算用の行列D



評価関数による枝刈り

評価関数（スコア）の計算方法



$$D_{i,j} = \max\left(D_{i-1,j-1} + w_{s(i),t(j)},\ D_{i-1,j} + d,\ D_{i,j-1} + d \right)$$

	1	2	3	4
文字列 s	A	I	M	S
文字列 t	A	M	O	S

$w_{s(i),t(j)}$:

アミノ酸の場合、s(i) と t(j) の類似度

塩基配列の場合、通常マッチ１，ミスマッチ０

d: ギャップペナルティ

グローバルアライメント

ニードルマンとブンシュによる評価関数の計算方法

$D_{0,0} = 0$

$D_{i,0} = i\ d$

(for i = 1 to n)

$D_{0,j} = j\ d$

(for j = 1 to m)

$D_{i,j} = \max\ (D_{i-1,j-1} + w_{s(i),t(j)},\ D_{i-1,j} + d,\ D_{i,j-1} + d)$

		A	T	A	T	G	C	C	T	T
	0	−1	−2	−3	−4	−5	−6	−7	−8	−9
A	−1	1	0	−1	−2	−3	−4	−5	−6	−7
A	−2	0	1	1	0	−1	−2	−3	−4	−5
G	−3	−1	0	1	1	1	0	−1	−2	−3
C	−4	−2	−1	0	1	1	2	1	0	−1
C	−5	−3	−2	−1	0	1	2	3	2	1
T	−6	−4	−2	−2	0	0	1	2	4	3
G	−7	−5	−4	−2	−1	1	0	1	3	4
T	−8	−6	−4	−3	−1	0	1	0	2	4

ここでは d = -1 で計算

max の計算で使った経路を覚えておく

グローバルアライメント

ニードルマンとブンシュによる評価関数の計算方法

$D_{0,0} = 0$

$D_{i,0} = i\ d$

(for i = 1 to n)

$D_{0,j} = j\ d$

(for j = 1 to m)

$D_{i,j} = \max\ (D_{i-1,j-1} + w_{s(i),t(j)},\ D_{i-1,j} + d,\ D_{i,j-1} + d)$

			A	T	A	T	G	C	C	T	T
		0	−1	−2	−3	−4	−5	−6	−7	−8	−9
A	−1	1	0	−1	−2	−3	−4	−5	−6	−7
A	−2	0	1	1	0	−1	−2	−3	−4	−5
G	−3	−1	0	1	1	1	0	−1	−2	−3
C	−4	−2	−1	0	1	1	2	1	0	−1
C	−5	−3	−2	−1	0	1	2	3	2	1
T	−6	−4	−2	−2	0	0	1	2	4	3
G	−7	−5	−4	−2	−1	1	0	1	3	4
T	−8	−6	−4	−3	−1	0	1	0	2	4

ATATGCCT−T

A−A−GCCTGT

右下（右端）から覚えていた経路を辿って戻る（バックトラック）

ローカルアライメント

スミスとウォーターマンによる評価関数の計算方法

$D_{0,0} = 0$

$D_{i,0} = 0$

(for i = 1 to n)

$D_{0,j} = 0$

(for j = 1 to m)

$D_{i,j} = \max($
$D_{i-1,j-1} + w_{s(i),t(j)},$
$D_{i-1,j} + d,$
$D_{i,j-1} + d,$
$0)$

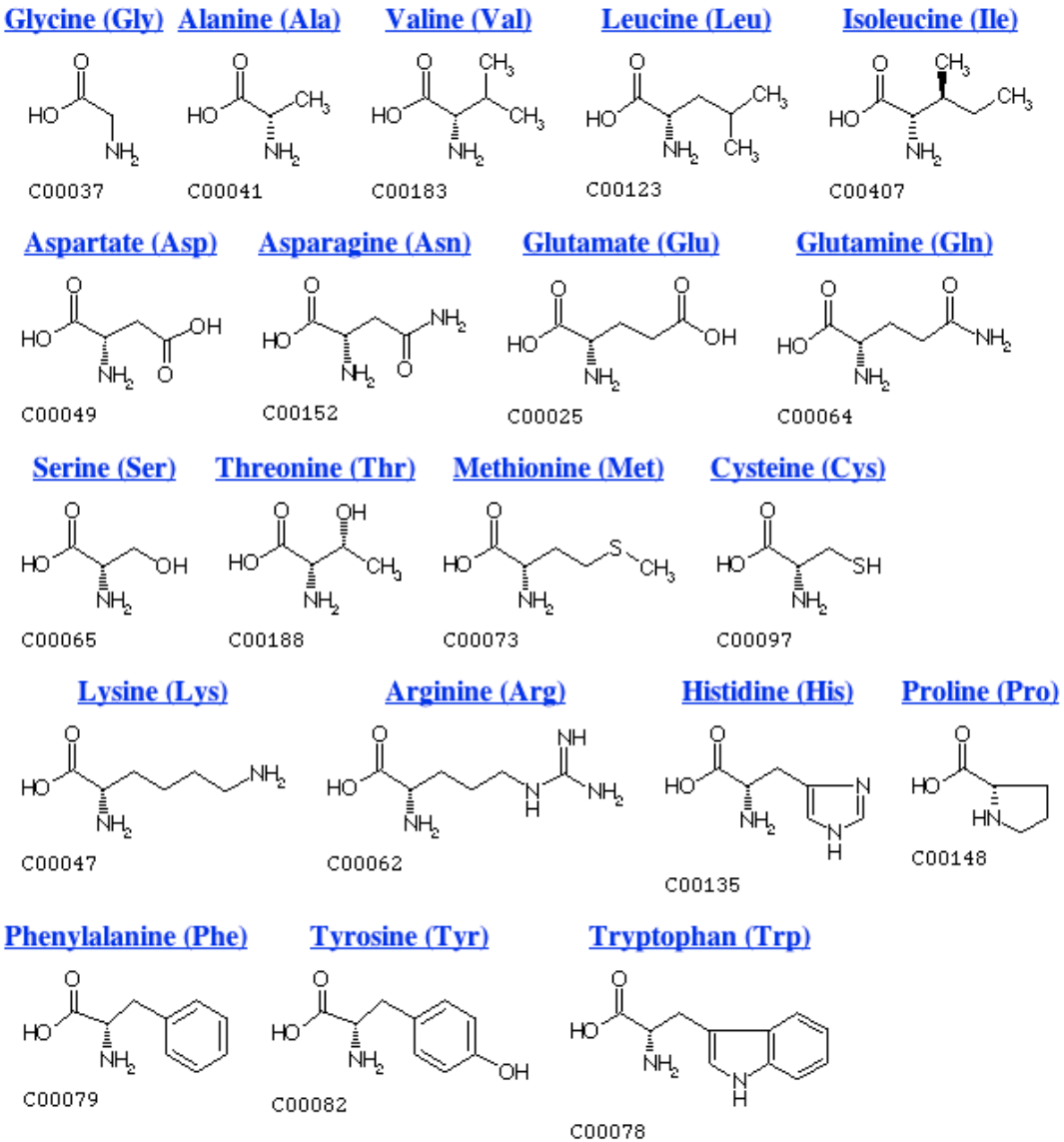
		A	T	A	T	G	C	C	T	T
	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	0	0	0
A	0	0	1	1	1	0	0	0	0	0
G	0	0	0	1	1	2	1	0	0	0
C	0	0	0	0	1	1	3	2	1	0
C	0	0	0	0	0	1	2	4	3	2
T	0	0	1	0	1	0	1	2	5	4
G	0	0	0	1	0	2	1	1	4	5
T	0	0	1	0	2	1	2	1	3	5

ATGCCT−T

AAGCCTGT

スコア最大の箇所から覚えていた経路を辿って戻る（バックトラック）

アミノ酸の構造と類似度



- 疎水性、親水性
- 電荷の有無
- 残基のサイズ
- コドン使用パターン

アミノ酸置換行列

BLOSUM50

BLOck SUbstitution Matrices

参考：AAindexデータベース

http://www.genome.jp/dbget/aaindex.html

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	1																			
Arg	-5	7																		
Asn	-5	-1	7																	
Asp	-5	-2	2	8																
Cys	-5	-4	-2	-4	13															
Gln	-5	1	0	0	-3	7														
Glu	-5	0	0	2	-3	2	6													
Gly	-5	-3	0	-1	-3	-2	-3	8												
His	-5	0	1	-1	-3	1	0	-2	10											
Ile	-5	-4	-3	-4	-2	-3	-4	-4	-4	5										
Leu	-5	-3	-4	-4	-2	-2	-3	-4	-3	2	5									
Lys	-5	3	0	-1	-3	2	1	-2	0	-3	-3	6								
Met	-5	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7							
Phe	-5	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8						
Pro	-5	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10					
Ser	-5	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5				
Thr	-5	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5			
Trp	-5	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15		
Tyr	-5	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	
Val	-5	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

PAM (Accepted Point Mutations)

- デイホフらによる定式化（1978年）
  - 71 グループの配列：類似度が 85% 以上
  - 1,572 の置換をカウント
  - 各アミノ酸置換の確率を1,572 の置換の割合で求める
  - 対数尤度（Log odds） 比で表現
    - $\log_{10}(q_{ij}/p_ip_j)$ に比例したスコア
    - p_i はその文字の出現率、q_{ij}は同時出現率
  - PAM 1
  - PAM 250 は PAM 1 を250 回掛け合わせたもの
    - 20 % の類似度に相当する

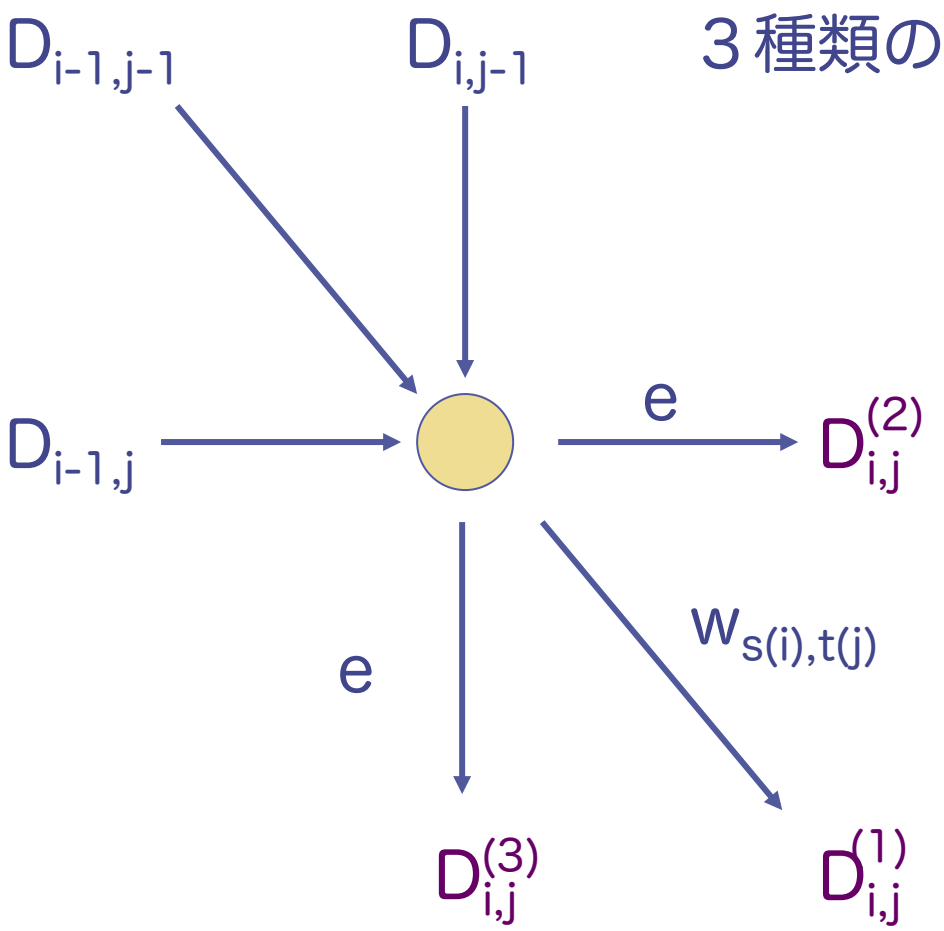
BLOSUM (*BLOck SUbstitution Matrices*)

- ヘニコフらによる定式化（1992年）
  - ギャップなしのローカルアライメントのデータベース（ブロック）から計算
  - すべての置換頻度と非置換頻度（すべてのアミノ酸ペア 20x19/2 = 210 パターン）をカウント
  - 対数尤度（Log odds） 比を使うのはPAMと同じ
    - $\log_2(q_{ij}/p_ip_j)$ に比例したスコア
    - p_i はその文字の出現率、q_{ij}は同時出現率
  - 70% 一致、60% 一致など配列一致度ごとに作成したブロックを用いてそれぞれの置換行列を計算

ギャップペナルティの種類

- 線形スコア
  - 長さ g のギャップペナルティ
  - $\gamma(g) = -gd$
- Affine スコア
  - 長さ g のギャップペナルティ
  - $\gamma(g) = -d - (g - 1)e$
  - d : gap open penalty
  - e : gap extension penalty
  - 通常は $d > e$

アフィンスコアを使った DP



３種類の $D_{i,j}$ を用意する

スコアの重要性を評価するための統計的基準

- ランダム配列との比較
  - Z-score

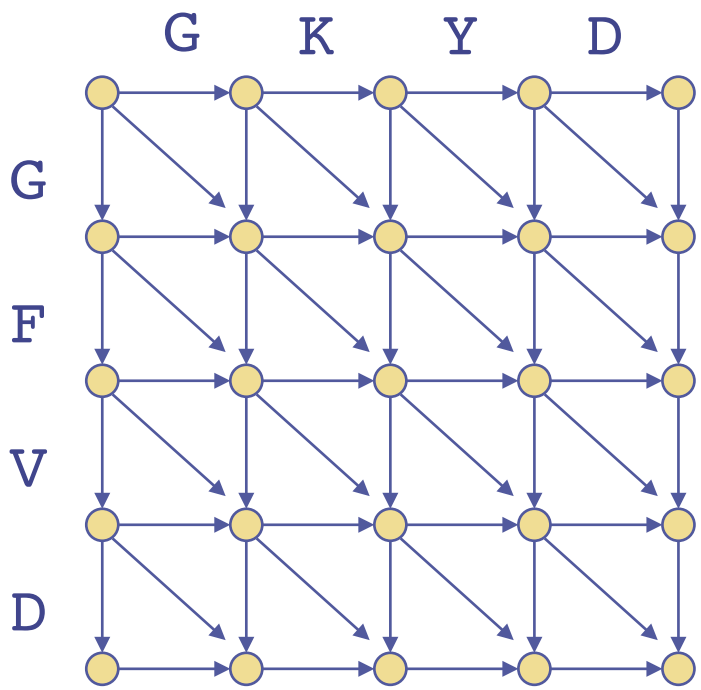
$$Z = \frac{S - \mu}{\sigma}$$

S ：スコア
μ ：平均
σ ：分散

練習問題

アミノ酸配列 GKYD と GFVD のローカルアライメントを、ギャップペナルティに２（d=-2）を、アミノ酸置換行列にBLOSUM50 を使って作成しなさい。

$$D_{i,j} = \max\left(\begin{array}{l} D_{i-1,j-1} + w_{s(i),t(j)}, \\ D_{i-1,j} + d, \\ D_{i,j-1} + d, \\ 0 \end{array}\right)$$



文献

- Woese, C. R., Kandler, O. and Wheelis, M. L. (1990) Towards a natural system of organisms: Proposal for the domains, Archaea, Bacteria and Eukarya. *Proc. Natl. Sci. USA,* 87, 4576-4579.
- Gibbs, A. J. and McIntyre, G. A. (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.*, 16, 1-11.
- Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48, 443-453.
- Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, 147, 195-197.
- Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162, 705-708.
- Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3, pp. 345-352.
- Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89, 10915-10919.