バイオインフォマティクス（配列検索）
＆
ケモインフォマティクス（構造検索）

統合薬学教育開発分野
奥野恭史

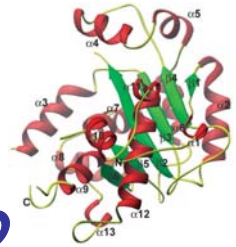
創薬におけるインフォマティクス

ゲノム情報 → ゲノム基盤研究 → ターゲット探索 → ターゲットバリデーション → 創薬リード探索 → 創薬リード最適化 → 前臨床研究 → 臨床研究 → 創 薬

ゲノム情報
（～2万2千遺伝子）

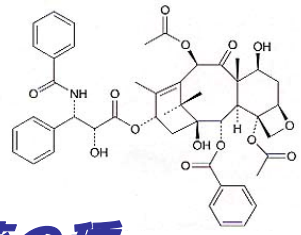
バイオインフォマティクス

疾患の
原因遺伝子の同定



化合物ライブラリー
（10^60 化合物）

ケモインフォマティクス

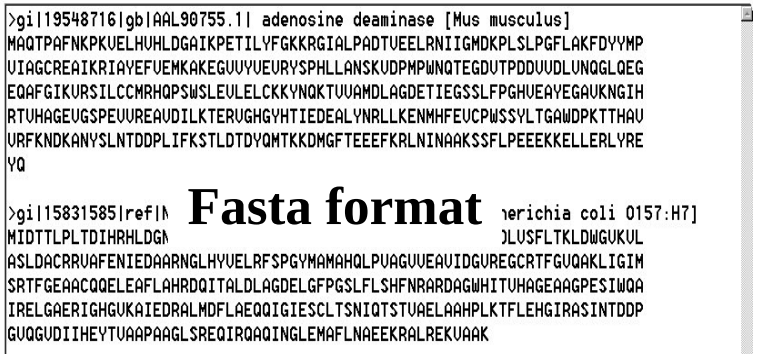


薬の種
リード化合物の選択

医薬品最適化
&
臨床試験

バイオインフォマティクス
配列解析

Sequences information

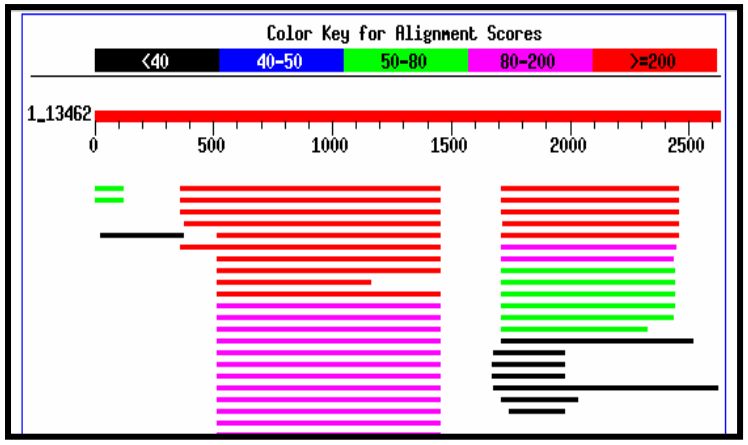


Fasta format

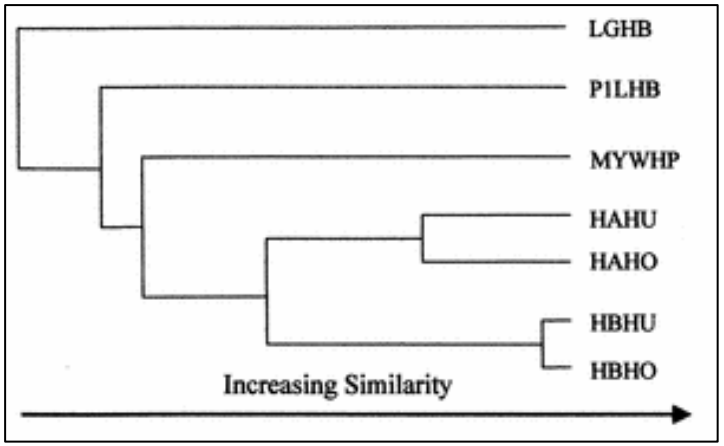
Similarity matrix

	HAHU	HBHU	HAHO	HBHO	MYWHP	P1LHB	LGHB
HAHU							
HBHU	21.1						
HAHO	32.9	19.7					
HBHO	20.7	**39.0**	20.4				
MYWHP	11.0	9.8	10.3	9.7			
P1LHB	9.3	8.6	9.6	8.4	7.0		
LGHB	7.1	7.3	7.5	7.4	7.3	4.3	

Alignment (ex. Blast…)



Classification



ケモインフォマティクス
構造解析

Structure

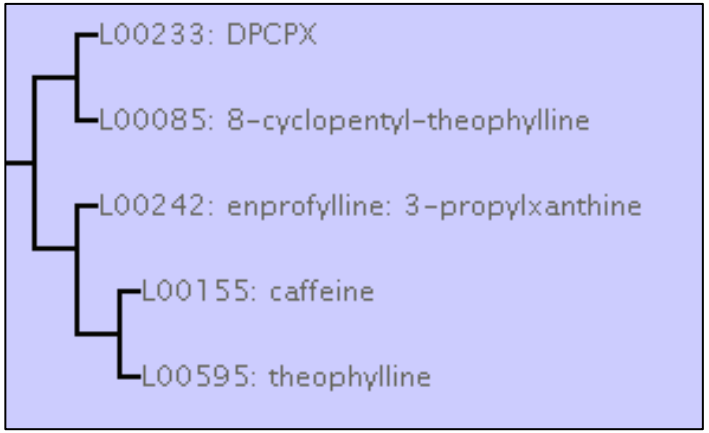
OC(=O)C(N)CC1=CC=C(O)C=C1

```
5  4  0     0  0               999 V2000
   -0.1276    0.2621    0.0000 C   0  0  0  0  0  0
    0.5552   -0.1862    0.0000 C   0  0  0  0  0  0
   -0.8552   -0.1483    0.0000 O   0  0  0  0  0  0
   -0.1552    1.0931    0.0000 O   0  0  0  0  0  0
    0.5793   -1.0207    0.0000 N   0  0  0  0  0  0
  1  2  1  0     0  0
  1  3  1  0     0  0
  1  4  2  0     0  0
  2  5  1  0     0  0
M  END
```

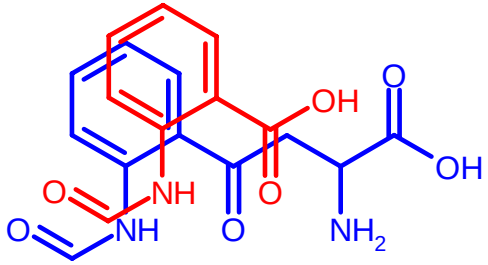
Distance matrix

	I1	I2	I3	I4	I5	I6
I2	1					
I3	7	5				
I4	13	6	16			
I5	10	0	7	5		
I6	9	9	12	13	12	
I7	11	20	10	9	14	8

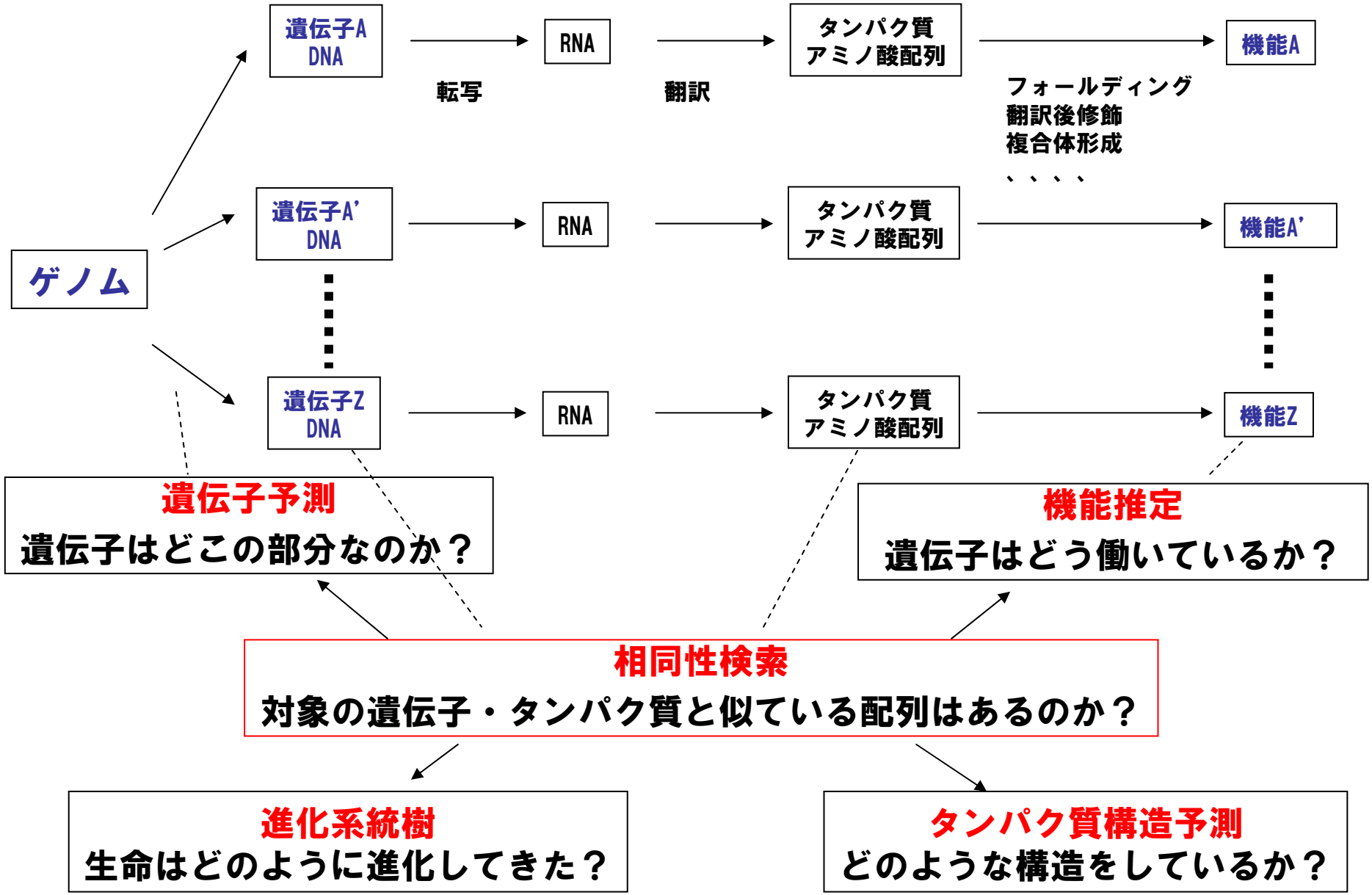
Classification



Structure comparison

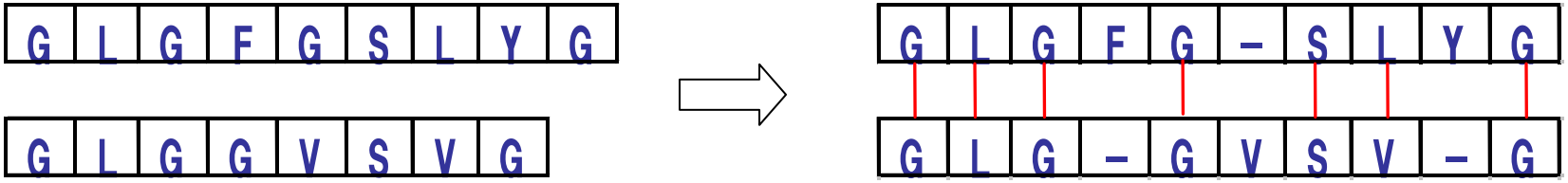


配列解析とは


遺伝子A
DNA
RNA
転写
翻訳
タンパク質
アミノ酸配列
フォールディング
翻訳後修飾
複合体形成
、、、、
機能A
ゲノム
遺伝子A'
DNA
RNA
タンパク質
アミノ酸配列
機能A'
遺伝子Z
DNA
RNA
タンパク質
アミノ酸配列
機能Z
遺伝子予測
遺伝子はどこの部分なのか？
機能推定
遺伝子はどう働いているか？
相同性検索
対象の遺伝子・タンパク質と似ている配列はあるのか？
進化系統樹
生命はどのように進化してきた？
タンパク質構造予測
どのような構造をしているか？

配列アライメント

- 配列が類似しているかをみるためには並べて比較すれば良い。
  配列中で同じ並び方をしている配列パターンを探すために、配列を並べる操作をアライメントと呼ぶ
- ２つの配列に対するアライメントはペアワイズアライメント、３つ以上の場合マルチプルアライメントという
- 文字の一致を最大限にするためにギャップ記号（挿入、欠失に対応）を挿入する

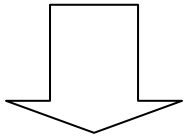


最適アライメントを求める
（最も類似していると思われる配列の並べ方）

スコア：　同じ文字は1点、異なる文字は－３点、ギャップは－２点

-10	-2	2	-12
		最適アライメント	
A G C T -	A G - C T	A - G C T	- A G C - - T
A C G C T	A C G C T	A C G C T	A C - - G C T

アライメント：　並べ方



つまり、類似性スコアの選択と並べる手順（方法）によって、
最適アライメントは影響を受ける

アライメントの方法（アルゴリズム）

【２つの考え方】

- グローバルアライメント

  配列全体の類似性を調べたいのか？

- ローカルアライメント

  局所的に、類似性の高い部分を調べたいのか？

＊例えば、顔が似ている、体格が似ている、どっちが似ているの？

【有名なアルゴリズム】

- ドットマトリックス法　（グローバル　＆　ローカル）
- 動的計画法　－　Needleman – Wunschアルゴリズム（グローバル）
  Smith – Wathermanアルゴリズム（ssearch）（ローカル）
- 近似的な方法　－　Blast（ローカル）
  Fasta（ローカル）

＊計算時間がかかっても、厳密にアライメントをしたいか？
多少厳密で無くても、速く結果を手にしたいか？でアルゴリズムが選択される。

スコア行列（アミノ酸配列）

PAM行列：先祖の共通タンパク質ファミリから多数のタンパク質を集め、置換の頻度を調べて分子進化学的に求めたもの

BLOSUM行列：配列の一致度が高いところで、マルチプルアライメントをとり特に保存性の高いところでのアミノ酸の変異を解析して求めたもの

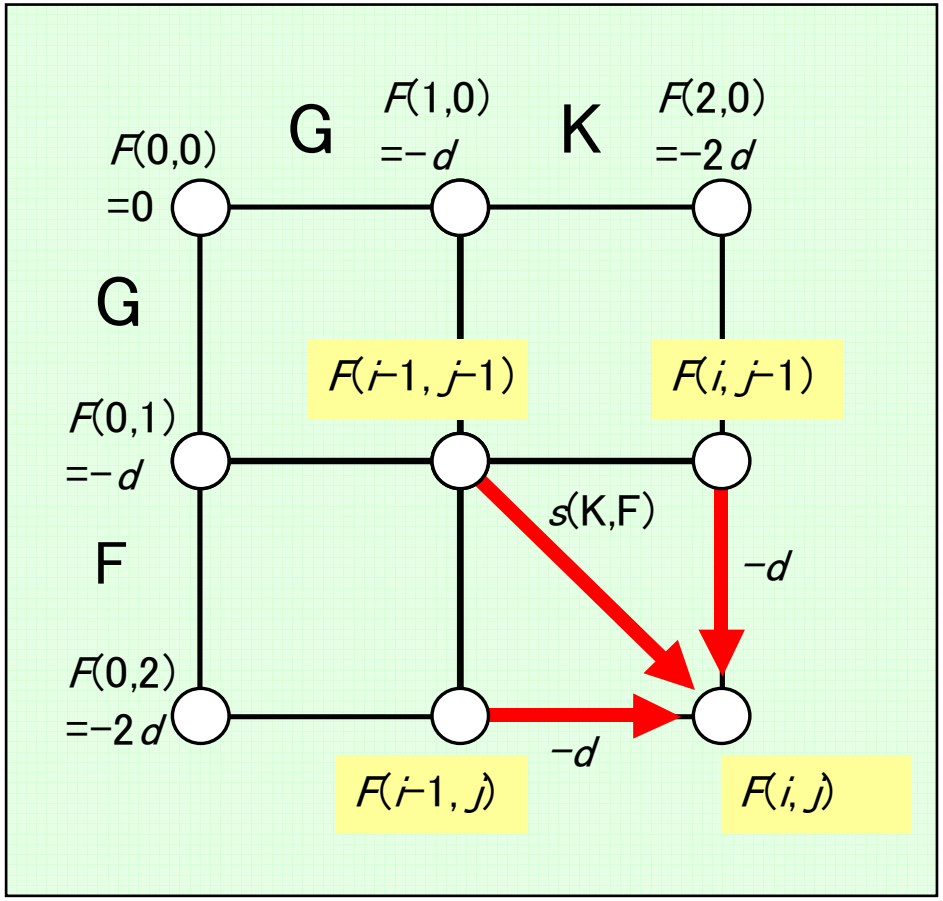
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**2**	−2	0	0	−2	0	0	1	−1	−1	−2	−1	−1	−4	1	1	1	−6	−3	0
R	−2	**6**	0	−1	−4	1	−1	−3	2	−2	−3	3	0	−4	0	0	−1	2	−4	−2
N	0	0	**2**	2	−4	1	1	0	2	−2	−3	1	−2	−4	−1	1	0	−4	−2	−2
D	0	−1	2	**4**	−5	2	3	1	1	−2	−4	0	−3	−6	−1	0	0	−7	−4	−2
C	−2	−4	−4	−5	**12**	−5	−5	−3	−3	−2	−6	−5	−5	−4	−3	0	−2	−8	0	−2
Q	0	1	1	2	−5	**4**	2	−1	3	−2	−2	1	−1	−5	0	−1	−1	−5	−4	−2
E	0	−1	1	3	−5	2	**4**	0	1	−2	−3	0	−2	−5	−1	0	0	−7	−4	−2
G	1	−3	0	1	−3	−1	0	**5**	−2	−3	−4	−2	−3	−5	−1	1	0	−7	−5	−1
H	−1	2	2	1	−3	3	1	−2	**6**	−2	−2	0	−2	−2	0	−1	−1	−3	0	−2
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	**5**	2	−2	2	1	−2	−1	0	−5	−1	4
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	**6**	−3	4	2	−3	−3	−2	−2	−1	2
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	**5**	0	−5	−1	0	0	−3	−4	−2
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	**6**	0	−2	−2	−1	−4	−2	2
F	−4	−4	−4	−6	−4	−5	−5	−5	−2	1	2	−5	0	**9**	−5	−3	−3	0	7	−1
P	1	0	−1	−1	−3	0	−1	−1	0	−2	−3	−1	−2	−5	**6**	1	0	−6	−5	−1
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	**2**	1	−2	−3	−1
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	**3**	−5	−3	0
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	**17**	0	−6
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	**10**	−2
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	**4**

PAM250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**

BLOSUM50

動的計画法によるグローバルアライメント
Needleman-Wunschアルゴリズム



スコア値の計算式

$$F(0,\ j) = -jd, \quad F(i,0) = -id$$

$$F(i,\ j) = \max \begin{cases} F(i-1,\ j-1) + s(x_i,\ y_j) \\ F(i-1,\ j) - d \\ F(i,\ j-1) - d \end{cases}$$

$s(x_i, y_j)$ ： 置換行列の要素

d ： ギャップペナルティ (>0)

行列からの経路の復元は、
$F(m, n)$ からmaxで＝となっている
$F(i, j)$ を逆にたどることに行う
（トレースバック）

F(i−1, j−1), F(i, j−1), F(i−1, j)の３つが決まれば、F(I, j)が決まる

Needleman-Wunschアルゴリズムによる計算例

ＨＥＡとＰＡＷをアライメントする場合

		H	E	A
	0	−8	−16	−24
P	−8	−2	−9	−17
A	−16	−10	−3	−4
W	−24	−18	−11	−6

スコア値の計算式

$F(0, j) = -jd, \quad F(i, 0) = -id$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) & -2+(-1) \\ F(i-1, j) - d & -9+(-8) \\ F(i, j-1) - d & -10+(-8) \end{cases}$$

$s(x_i, y_j)$：置換行列の要素　E／A: -1

d: ギャップペナルティ(>0)　8

置換行列：　BLOSUM50　　**リニアスコアギャップ**：d = 8

スコア行列：ＢＬＯＳＵＭ５０

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	7	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	7	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	8	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	13	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	7	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	6	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	8	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	10	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	5	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	5	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	6	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	7	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	8	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	10	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	5	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	5	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	15	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	8	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	5

Needleman-Wunschアルゴリズムによる計算例

置換行列： BLOSUM50　　**リニアスコアギャップ：** d = -8

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

得られる結果

H	E	A	G	A	W	G	H	E	-	E
-	-	P	-	A	W	-	H	E	A	E

進化系統樹

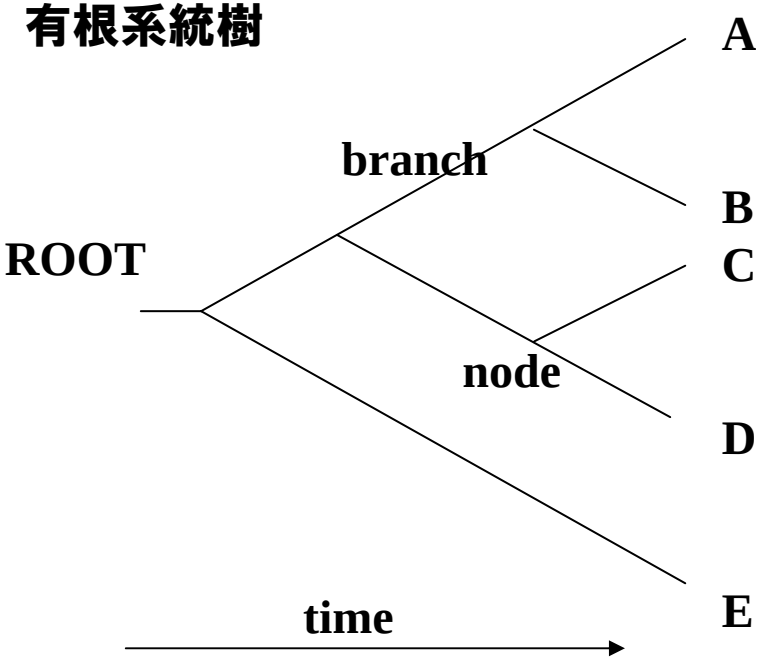
異なる生物種に由来する遺伝子・タンパク質の配列が類似している

⇩

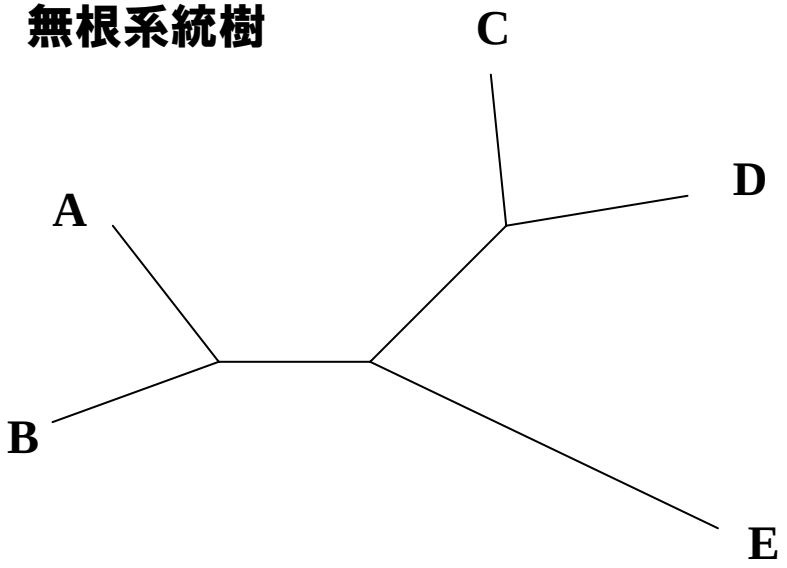
それらの遺伝子・タンパク質が共通祖先を持つ可能性が高い

配列相同性と進化的距離の関連がある

有根系統樹



無根系統樹



Brachの長さ＝進化的距離

進化系統樹の作成方法

- 距離行列法
- 最大節約法
- 最尤法

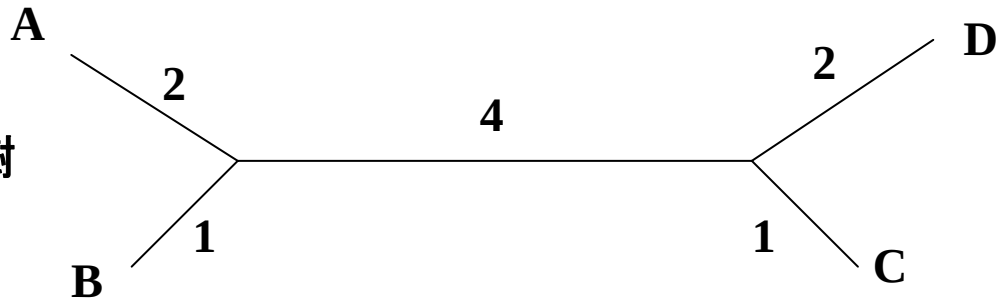
配列

配列A :ACGCGTTGGGCGATGGCAAC
配列B :ACGCGTTGGGCGACGGTAAT
配列C :ACGCATTGAATGATGATAAT
配列D :ACACATTGAGTGATAATAAT

配列間の距離
（置換数）

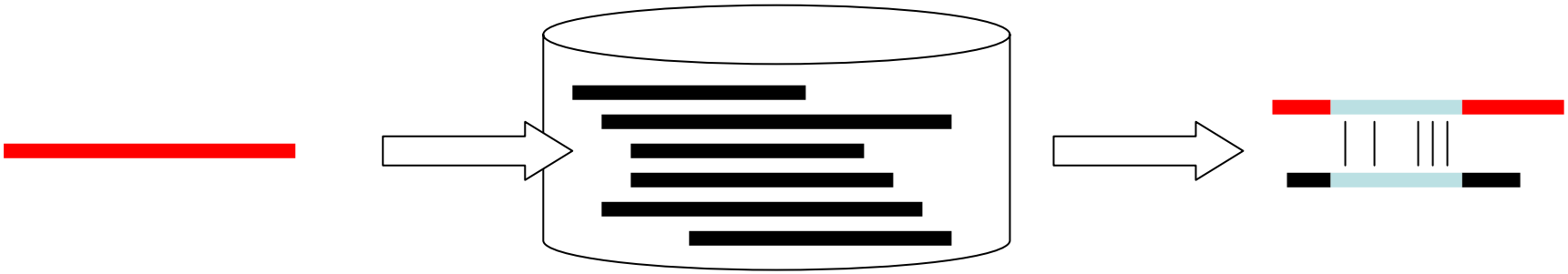
	A	B	C	D
A	-	3	7	8
B	-	-	6	7
C	-	-	-	3
D	-	-	-	-

系統樹



ホモロジーサーチ（相同性検索）

- 相同性検索は対象となる配列と類似の配列が配列データベースに存在するかどうかを検索する手法である。
- 検索する配列（クエリー配列）とデータベース中の配列の間でアライメントを作成し、その中からよく類似した配列を選び出す。



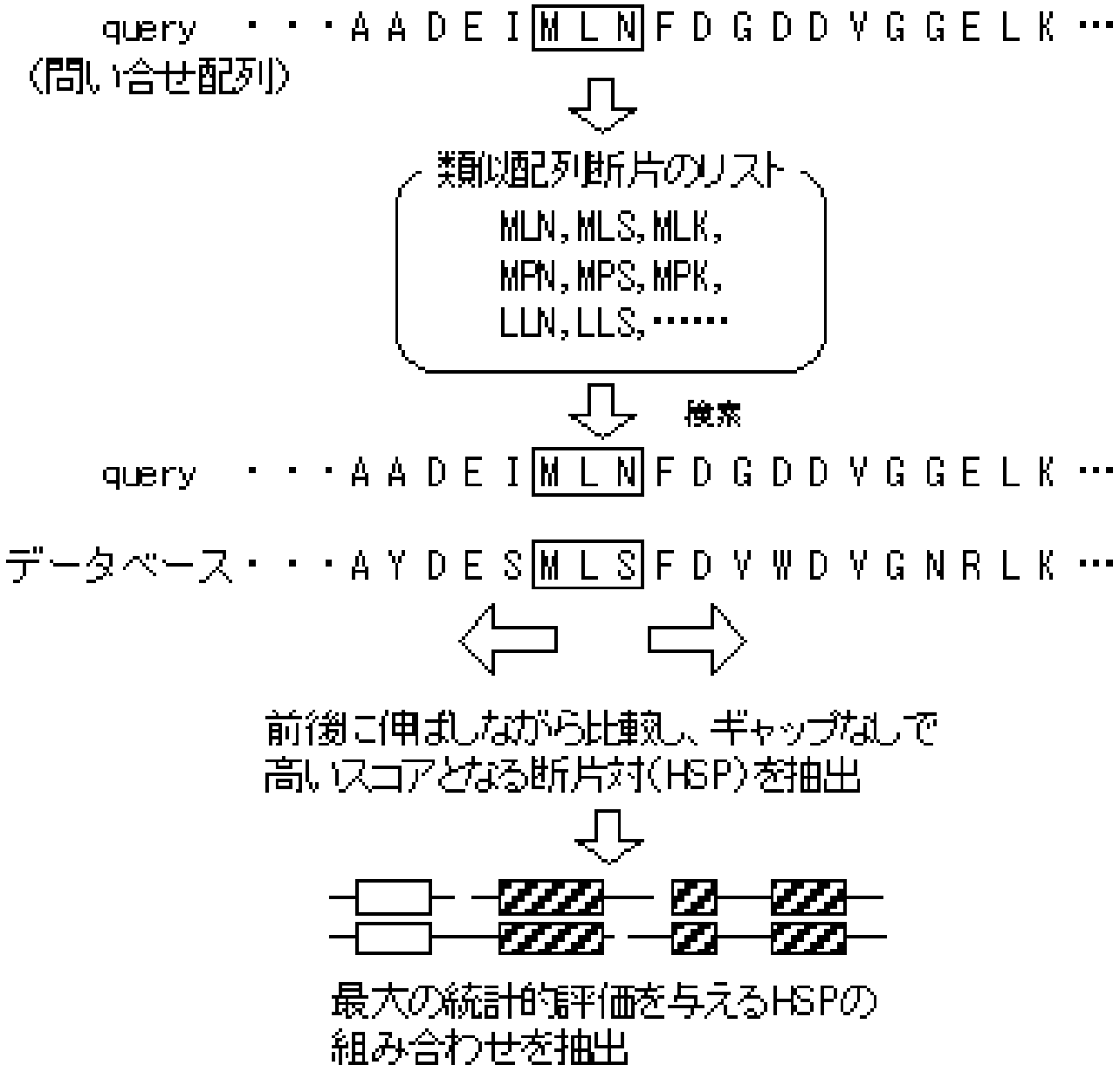
ホモロジーサーチに用いられるプログラム

スコアを最大にする最適なアライメントは**動的計画法（ssearch）**により計算できるが、データベースの配列全てに対して１つ１つこの手法を適用すると膨大な時間がかかる。

実際には近似手法が用いられている

BLAST、FASTAといったホモロジー検索プログラムが用いられている

Blastのアルゴリズム



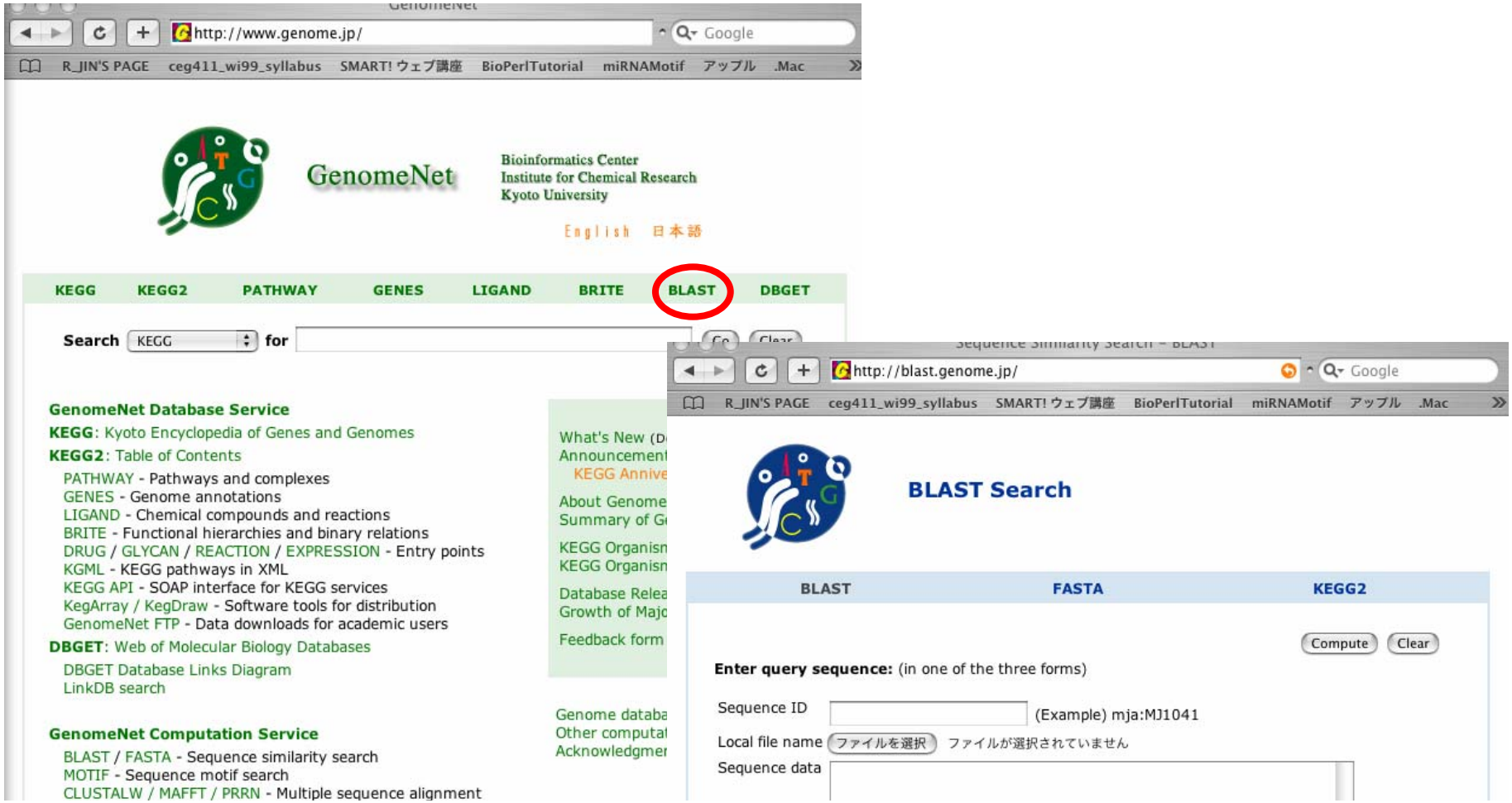
実際にBlast検索する（１）

http://blast.genome.jp/

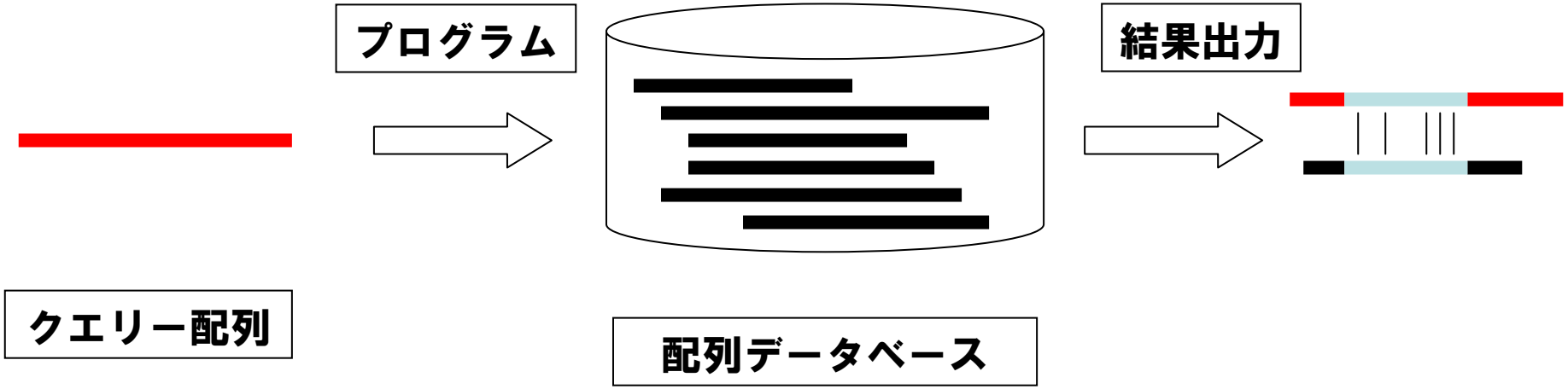
京大・化学研究所・バイオインフォマティクスセンター

http://www.ncbi.nlm.nih.gov/BLAST/

米国・The National Center for Biotechnology Information (NCBI)



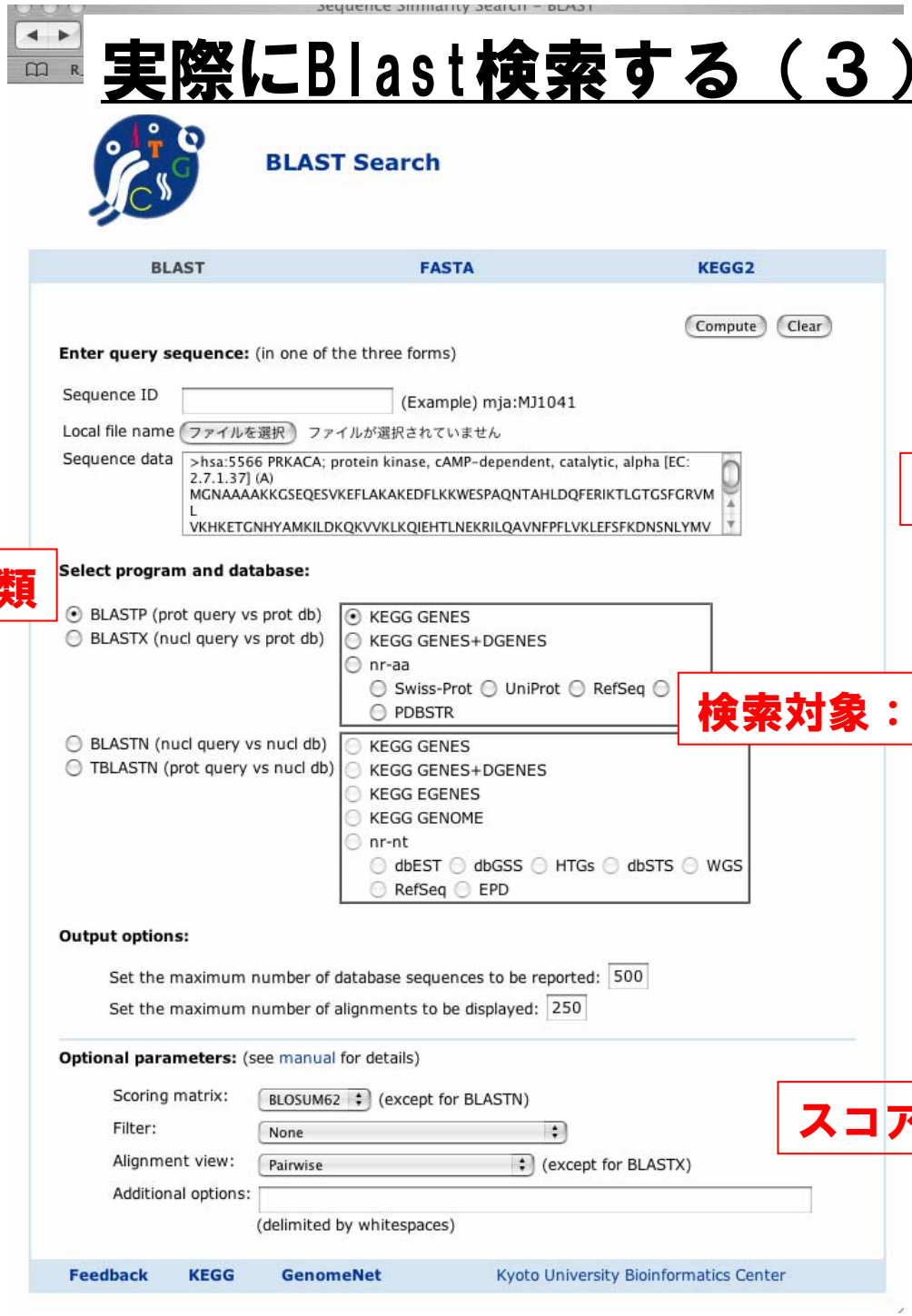
実際にBlast検索する（２）



クエリー配列を用意する：　FASTA形式の配列

```
>hsa:5566 PRKACA; protein kinase, cAMP-dependent, catalytic, alpha [EC:2.7.1.37] (A)
MGNAAAAKKGSEQESVKEFLAKAKEDFLKKWESPAQNTAHLDQFERIKTLGTGSFGRVML
VKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMV
MEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGY
IQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFF
ADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWFAT
TDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF
```

実際にBlast検索する（３）



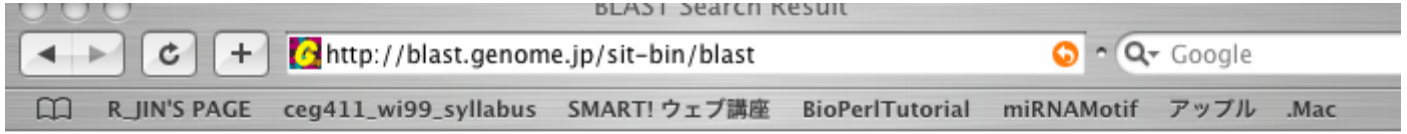
プログラムの種類

クエリー配列を入力

検索対象：データベースの種類

スコア行列の選択

実際にBlast検索する（4）



出力結果

BLASTP Search Result

Database: genes

Protein sequence database entries related to hsa:5566 - 500 hits

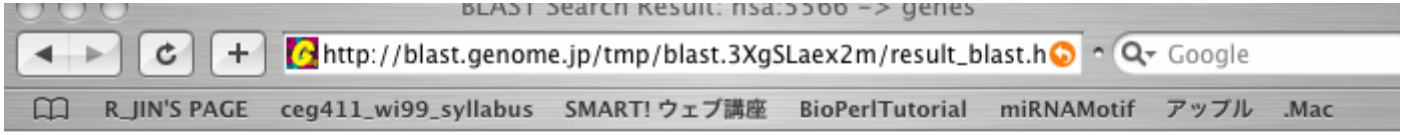
Entry	bits	E-val
Top 10 / Clear / Select operation / Exec		
hsa:5566 PRKACA; protein kinase, cAMP-dependent, catalytic, alph...	717	0.0
bta:282322 PRKACA; protein kinase, cAMP-dependent, catalytic, alpha	713	0.0
cfa:403556 PRKACA; protein kinase, cAMP-dependent, catalytic, alpha	706	0.0
rno:25636 Prkaca; protein kinase, cAMP-dependent, catalytic, alp...	706	0.0
mmu:18747 Prkaca; protein kinase, cAMP dependent, catalytic, alp...	704	0.0
xla:446502 prkacb-prov; protein kinase, cAMP-dependent, catalyti...	677	0.0
xla:380388 kin-1-prov; protein kinase, cAMP-dependent, catalytic...	670	0.0
bta:282323 PRKACB; protein kinase, cAMP-dependent, catalytic, beta	668	0.0
hsa:5567 PRKACB; protein kinase, cAMP-dependent, catalytic, beta...	667	0.0
rno:293508 LOC293508; similar to protein kinase, cAMP dependent,...	661	0.0
mmu:18749 Prkacb; protein kinase, cAMP dependent, catalytic, bet...	661	0.0
dre:445076 zgc:91856	657	0.0
gga:424542 LOC424542; similar to cAMP-dependent protein kinase c...	650	0.0
cfa:479975 LOC479975; similar to cAMP-dependent protein kinase, ...	644	0.0
ptr:469367 LOC469367; similar to cAMP-dependent protein kinase c...	644	0.0
cel:ZK909.2f kin-1; Hypothetical protein ZK909.2f [EC:2.7.1.37] ...	613	e-174
cel:ZK909.2a kin-1; Hypothetical protein ZK909.2a [EC:2.7.1.37] ...	612	e-174
hsa:5568 PRKACG; protein kinase, cAMP-dependent, catalytic, gamm...	612	e-174
cel:ZK909.2h kin-1; Hypothetical protein ZK909.2h [EC:2.7.1.37] ...	612	e-174
cel:ZK909.2g kin-1; Hypothetical protein ZK909.2g	610	e-174
cel:ZK909.2l kin-1; Hypothetical protein ZK909.2l [EC:2.7.1.37] ...	610	e-173
cel:ZK909.2m kin-1; Hypothetical protein ZK909.2m	610	e-173
ptr:472944 LOC472944; similar to protein kinase, cAMP-dependent,...	608	e-173

遺伝子

スコア

実際にBlast検索する（5）

出力結果



>hsa:5586 PKN2, PRKCL2; protein kinase N2 [EC:2.7.1.37] [KO:K06071] Top

Length = 984

Score = 251 bits (640), Expect = 2e-65
Identities = 125/297 (42%), Positives = 180/297 (60%), Gaps = 7/297 (2%)

```
Query: 40  HLDQFERIKTLGTGSFGRVMLVKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAV 99
           +L  F     LG G FG+V+L ++K T   +A+K L K  +V   +++  + EKRI + V
Sbjct: 653 NLQDFRCCAVLGRGHFGKVLLAEYKNTNEMFAIKALKKGDIVARDEVDSLMCEKRIFETV 712

Query: 100 N---FPFLVKLEFSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFE 156
           N    PFLV L   F+   ++  VMEY  GG++  H+     FSEP A FYAA +VL  +
Sbjct: 713 NSVRHPFLVNLFACFQTKEHVCFVMEYAAGGDLMMHIHT-DVFSEPRAVFYAACVVLGLQ 771

Query: 157 YLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKG---RTWTLCGTPEYLAPEIILS 213
           YLH   ++YRDLK +NLL+D +G++++ DFG  K   G   RT T CGTPE+LAPE++
Sbjct: 772 YLHEHKIVYRDLKLDNLLLDTEGFVKIADFGLCKEGMGYGDRTSTFCGTPEFLAPEVLTE 831

Query: 214 KGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNL 273
             Y +AVDWW LGVLIYEM  G  PF  D   ++++ IV+ +VR+P   S++   ++R L
Sbjct: 832 TSYTRAVDWWGLGVLIYEMLVGESPFPGDDEEEVFDSIVNDEVRYPRFLSTEAISIMRRL 891

Query: 274 LQVDLTKRFGNLKNGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDD 330
           L+ +  +R G  +    D+K H +F   DW A+  +KV+ PFIP  +G  D SNFDD
Sbjct: 892 LRRNPERRLGASEKDAEDVKKHPFFRLIDWSALMDKKVKPPFIPTIRGREDVSNFDD 948
```

>rno:81749 Prkch; protein kinase C, eta [EC:2.7.1.-] [KO:K06068] Top

Length = 683

Score = 250 bits (639), Expect = 2e-65
Identities = 129/301 (42%), Positives = 191/301 (63%), Gaps = 6/301 (1%)

```
Query: 41  LDQFERIKTLGTGSFGRVMLVKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQ-AV 99
           +D FE I+ LG GSFG+VML + KETG  YA+K+L K  +++   +E T+ EKRIL  A
Sbjct: 352 IDNFEFIRVLGKGSFGKVMLARIKETGELYAVKVLKKDVILQDDDVECTMTEKRILSLAR 411

Query: 100 NFPFLVKLEFSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLH 159
           N PFL +L   F+    L+ VME+V GG++  H+++  RF E  ARFYAA+I+    +LH
Sbjct: 412 NHPFLTQLFCCFQTPDRLFFVMEFVNGGDLMFHIQKSRRFDEARARFYAAEIISALMFLH 471
```

アライメント

実際にBlast検索する（７）

遺伝子情報

KEGG **Rattus norvegicus (rat): 81749**

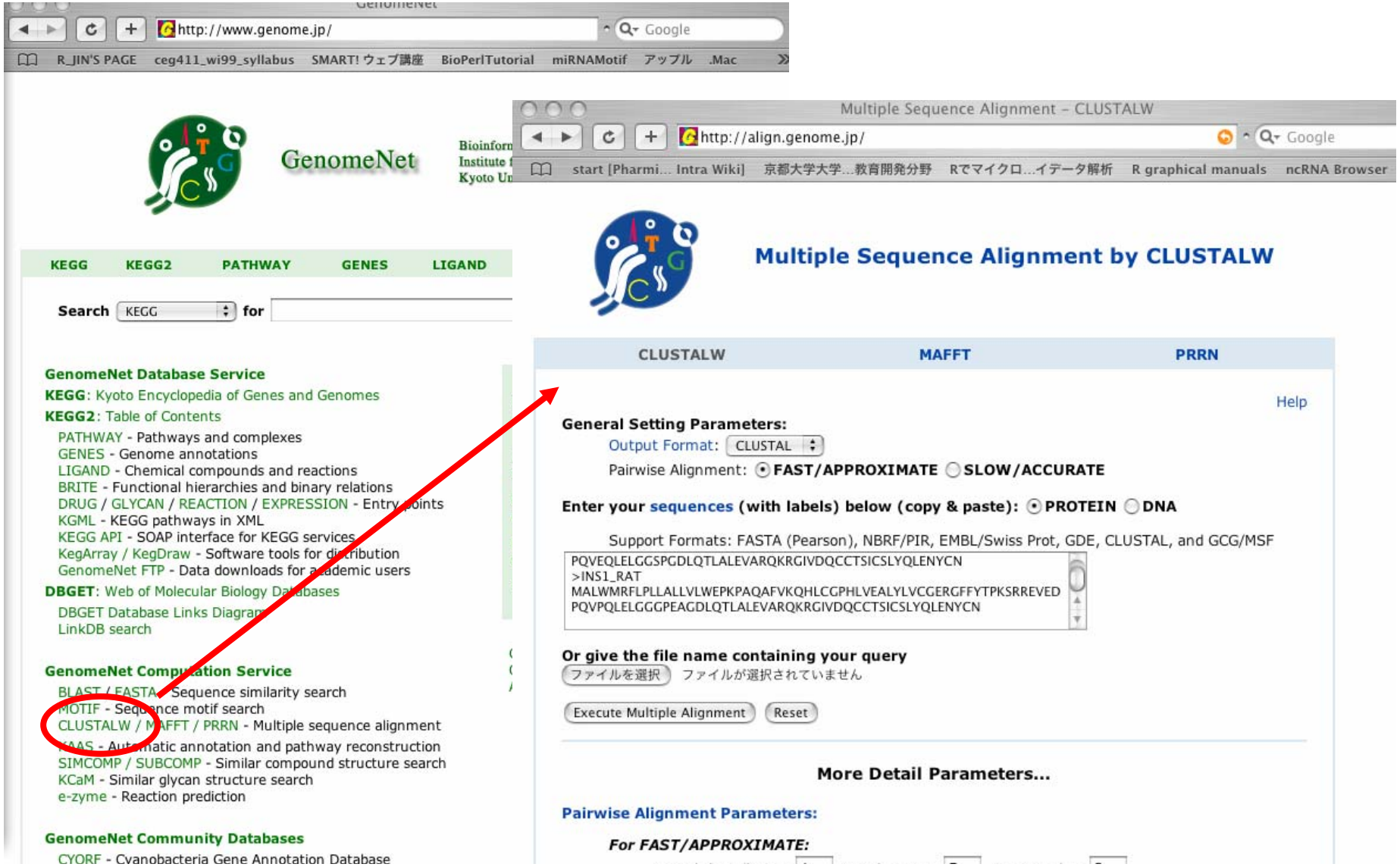
Help

Entry	81749 CDS R.norvegicus
Gene name	Prkch
Definition	protein kinase C, eta [EC:2.7.1.-]
KO	KO: K06068 novel protein kinase C OC search OC viewer
Pathway	PATH: rno04530 Tight junction
Class	Gene catalog
SSDB	Ortholog Paralog Gene cluster
Motif	Pfam: C2 C1_1 Pkinase Pkinase_C PROSITE: PROTEIN_KINASE_ST PROTEIN_KINASE_ATP DAG_PE_BIND_DOM_1 DAG_PE_BIND_DOM_2 C2_DOMAIN_2 PROTEIN_KINASE_DOM Motif
Other DBs	RGD: 621888 NCBI-GI: 13592027 NCBI-GeneID: 81749 UniProt: Q64617
LinkDB	PDB All DBs
Position	6q24
AA seq	683 aa AA seq DB search MSSGTMKFNGYLRVRIGEAVGLQPTRWSLRHSLFKKGHQLLDPYLTVSVDQVRVGQTSTK QKTNKPTYNEEFCTNVSDGGHLELAVFHETPLGYDHFVANCTLQFQELLRTAGTSDTFEG WVDLEPEGKVFVVITLTGSFTEATLQRDRIFKHFTRKRQRAMRRRVHQVNGHKFMATYLR QPTYCSHCREFIWGVFGKQGYQCQVCTCVVHKRCHHLIVTACTCQNNINKVDAKIAEQRF GINIPHKFNVHNYKVPTFCDHCGSLLWGIMRQGLQCKICKMNVHIRCQANVAPNCGVNAV ELAKTLAGMGLQPGNISPTSKLISRSTLRRQGKEGSKEGNGIGVNSSSRFGIDNFEFIRV LGKGSFGKVMLARIKETGELYAVKVLKKDVILQDDDVECTMTEKRILSLARNHPFLTQLF CCFQTPDRLFFVMEFVNGGDLMFHIQKSRRFDEARARFYAAEIISALMFLHEKGIIYRDL KLDNVLLDHEGHCKLADFGMCKEGICNGVTTATFCGTPDYIAPEILQEMLYGPAVDWWAM GVLLYEMLCGHAPFEAENEDDLFEAILNDEVVYPTWLHEDATGILKSFMTKNPTMRLGSL

実際に系統樹を作成する（１）

http://align.genome.jp/

京大・化学研究所・バイオインフォマティクスセンター



実際に系統樹を作成する（２）

MultiFASTA形式

>INS_HUMAN

MALWMRLLPLLALLALWGPDPAAAFVNQH
PKTRREAED

LQVGQVELGGGPGAGSLQPLALEGSLQKRG

>INS_BOVIN

MALWTRLRPLLALLALWPPPPARAFVN QHL
KARREVEG

PQVGALELAGGPGAGGLEGPPQKRGIVEQC

>INS_PIG

MALWTRLLPLLALLALWAPAPAQAFVNQHI
KARREAEN

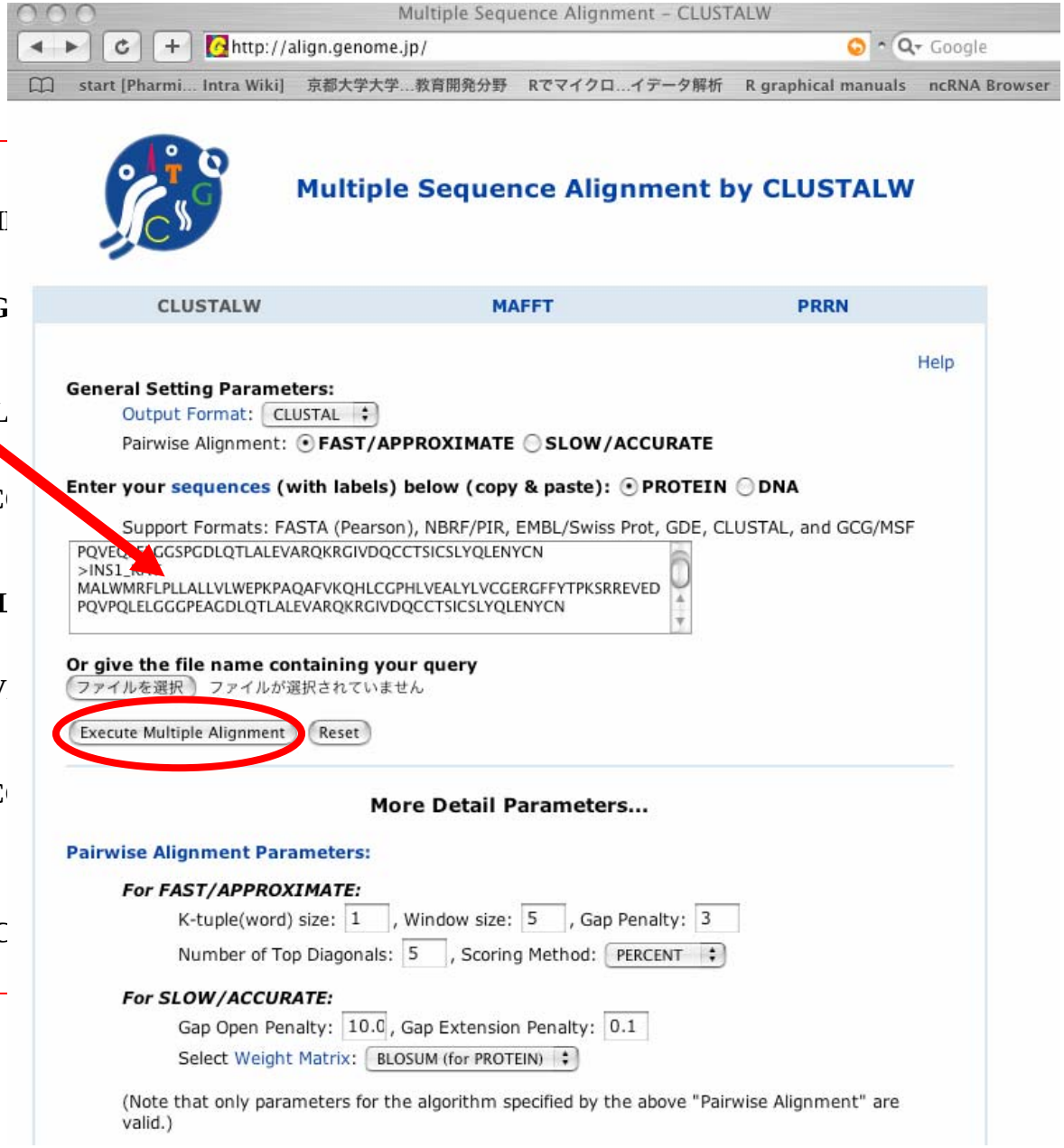
PQAGAVELGGGLGGLQALALEGPPQKRGIV

>INS_CYPCA

MAVWIQAGALLFLLAVSSVNANAGAPQHLC
RDVDPPLG

>INS_CHICK

MALWIRSLPLLALLVFSGPGTSYAAANQHLC
ARRDVEQ



実際に系統樹を作成する（３）

CLUSTALW Result

http://align.genome.jp/sit-bin/clustalw

start [Pharmi... Intra Wiki] 京都大学大学...教育開発分野 Rでマイクロ...イデータ解析 R graphical manuals ncRNA Browser

```
Group 2: Sequences:   2      Score:1799
Group 3: Sequences:   2      Score:1687
Group 4: Sequences:   4      Score:1605
Group 5: Sequences:   6      Score:1641
Group 6: Sequences:   7      Score:925
Group 7: Sequences:   2      Score:921
Group 8: Sequences:   9      Score:459
Alignment Score 13137
CLUSTAL-Alignment file created  [clustalw.aln]
```

clustalw.aln

```
CLUSTAL W (1.83) multiple sequence alignment


INS_BOVIN       MALWTRLRPLLALLALWPPPPARAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVEG
INS_PIG         MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN
INS_HUMAN       MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
INS_CERAE       MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
INS1_MOUSE      MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED
INS1_RAT        MALWMRFLPLLALLVLWEPKPAQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED
INS_CHICK       MALWIRSLPLLALLVFSGPGTSYAAANQHLCGSHLVEALYLVCGERGFFYSPKARRDVEQ
INS_CYPCA       MAVWIQAGALLFLLAVSSVN-ANAGAPQHLCGSHLVDALYLVCGPTGFFYNPK--RDVDP
INS_BRARE       MAVWLQAGALLVLLVVSSVS-TNPGTPQHLCGSHLVDALYLVCGPTGFFYNPK--RDVEP
                **:  :  .** **..       . . *****.***:*******  ****.**  *:.:

INS_BOVIN       PQVGALELAGGP--G---AGGLEGPPQKRGIVEQCCASVCSLYQLENYCN
INS_PIG         PQAGAVELGGGL--GGLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN
INS_HUMAN       LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
INS_CERAE       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
INS1_MOUSE      PQVEQLELGGSP--GDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
INS1_RAT        PQVPQLELGGGPEAGDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN
INS_CHICK       ----------------------------------------------------
INS_CYPCA       PLG-------------------------------------------------
INS_BRARE       LLG-------------------------------------------------
```

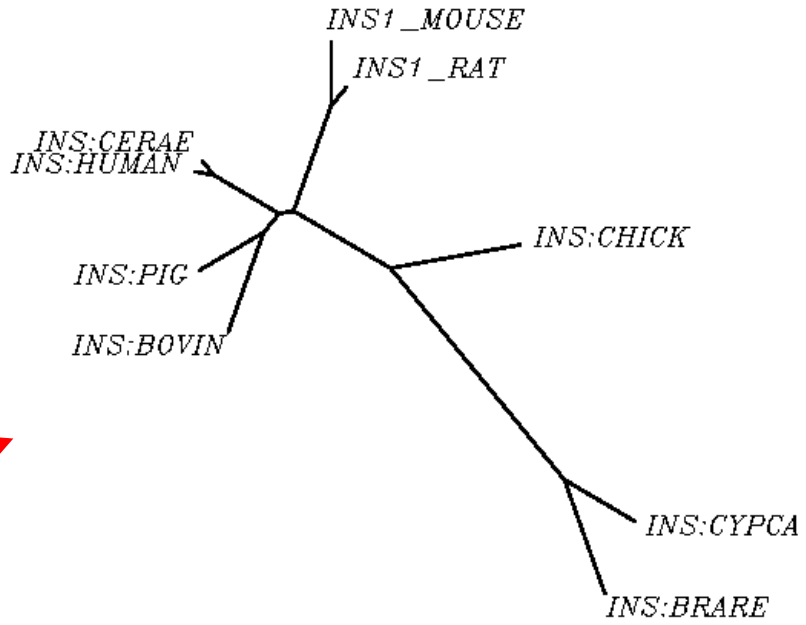
マルチプルアライメント結果

clustalw.dnd

```
(
(
INS_HUMAN:0.01410,
INS_CERAE:0.01317)
:0.06149,
(
(
INS_BOVIN:0.09228,
INS_PIG:0.06962)
:0.01412,
(
(
INS_CYPCA:0.06667,
INS_BRARE:0.08333)
:0.30208,
INS_CHICK:0.11458)
:0.05909)
:0.01184,
(
INS1_MOUSE:0.06405,
INS1_RAT:0.02855)
:0.07837);
```

Select tree menu | Exec

Generate profile HMM



ケモインフォマティクス
構造解析

Structure

OC(=O)C(N)CC1=CC=C(O)C=C1

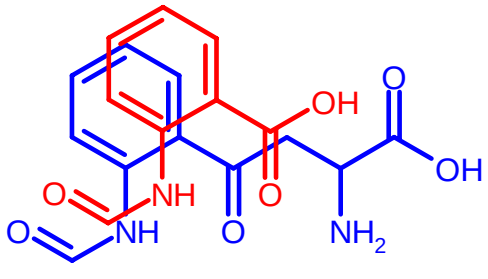
```
5  4  0     0  0               999 V2000
   -0.1276    0.2621    0.0000 C   0  0  0  0  0  0
    0.5552   -0.1862    0.0000 C   0  0  0  0  0  0
```

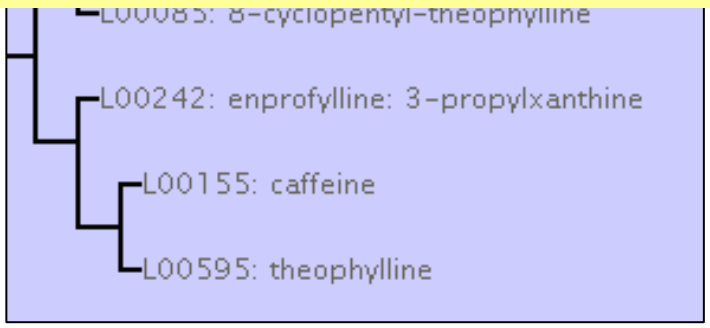
Distance matrix

I2	1		
I3	7	5	
I4	13	6	16

- ◆ 化学物質(分子)の情報学的表現
- ◆ 分子比較
- ◆ 化合物データベース
- ◆ 分子の特徴抽出、化学量定義

Structure comparison



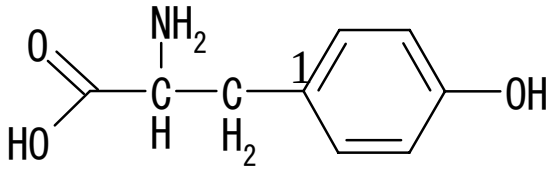


化学物質(分子)の情報学的表現

1. Line notation : represent structures as compact linear string of alphanumeric symbols

SMILES (Simplified Molecular Input Line Entry System) : developed by Daylight

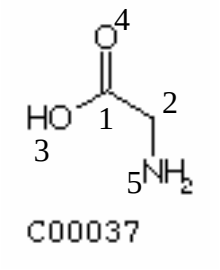
OC(=O)C(N)CC1=CC=C(O)C=C1



2. Connection Table

KCF (KEGG Chemical Format)

```
ENTRY       C00037                  Compound
NODE        5
            1 C6a C  -0.12760 0.2621
            2 C1b C  0.55520 -0.1862
            3 O6a O  -0.85520 -0.1483
            4 O6a O  -0.15520 1.0931
            5 N1a N  0.57930 -1.0207
EDGE        4
            1   1   2 1
            2   1   3 1
            3   1   4 2
            4   2   5 1
///
```



MDL CT format

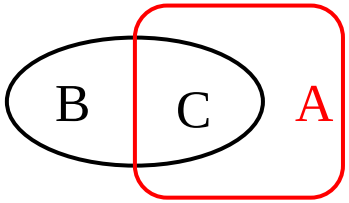
```
ISISHOST03020323002D 1   1.00000     0.00000     37

  5  4  0     0  0               999 V2000
   -0.1276    0.2621    0.0000 C   0  0  0  0  0  0           0  0  0
    0.5552   -0.1862    0.0000 C   0  0  0  0  0  0           0  0  0
   -0.8552   -0.1483    0.0000 O   0  0  0  0  0  0           0  0  0
   -0.1552    1.0931    0.0000 O   0  0  0  0  0  0           0  0  0
    0.5793   -1.0207    0.0000 N   0  0  0  0  0  0           0  0  0
  1  2  1  0     0  0
  1  3  1  0     0  0
  1  4  2  0     0  0
  2  5  1  0     0  0
M  END
```

⟶ Graph representation

分子比較（化合物類似性）

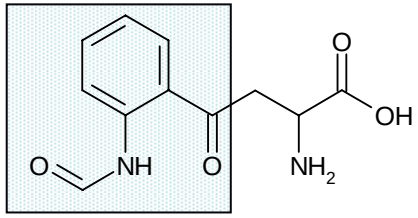
Tanimoto coefficient



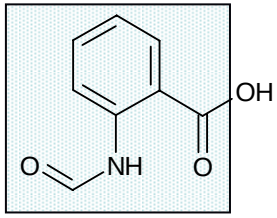
a: size of mol_A
b: size of mol_B
c: size of overlap

Tanimoto coefficient
= c / (a+b-c)

structure



formylkynurenine



formylanthranilate

a=17 c=11 b=12
-> 11 / (17+12-11) = 0.61

fingerprint

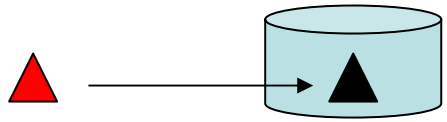
Mol A: 0101011001010000100100
Mol B: 0000101010010010000100

a=8 c=3 b=6
-> 3 / (8+6-3) = 0.27

構造検索（データベースサーチ）

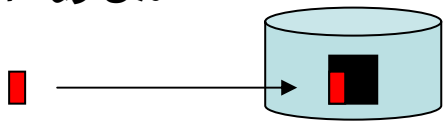
1. Full structure search

問い合わせ分子と全く同じ構造をもつ分子がDB中にあるか？



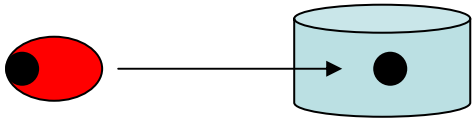
2. Substructure search

問い合わせ構造を部分構造として含む分子がDB中にあるか？



3. Superstructure search

問い合わせ分子中の部分構造と一致する分子がDB中にあるか？



4. Similar structure search

問い合わせ分子と、或る閾値以上の類似性を示す分子がDB中にあるか？

→ 類似度の定義が必要

5. Reaction search

molecular alignment (atom-atom matching)

Maximal common subgraph search

6. 3D substructure search

分子の特徴抽出、化学量定義
Chemical descriptors

“Chemical property correlates with chemical strucutre”

Chemical property
Molecular weight
Number of rotatable bonds
Number of potential hydrogen-bond donors/acceptors
Solubility
Acid dissociation constant
Standard gibbs free energy
Octanol-water distribution coefficient

Can be estimated by Chemical descriptors

Public available Chemical database