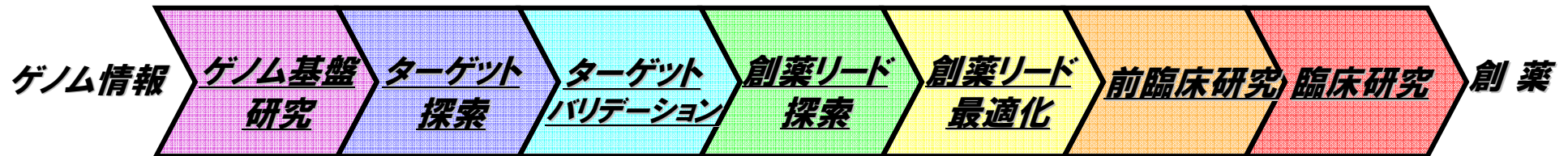


**バイオインフォマティクス(配列検索)**  
**&**  
**ケモインフォマティクス(構造検索)**

**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

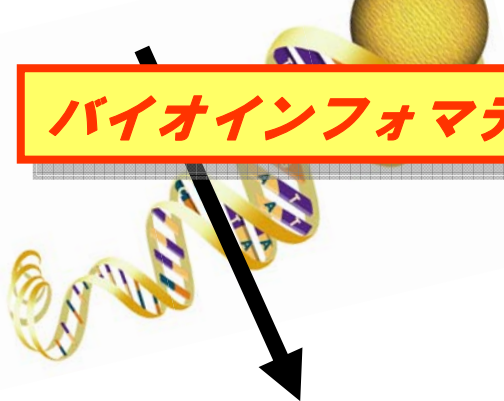


ゲノム情報  
(~2万2千遺伝子)

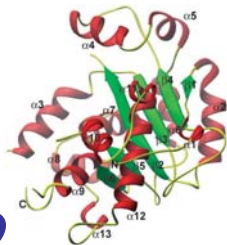
化合物ライブラリー  
( $10^{60}$  化合物)

**バイオインフォマティクス**

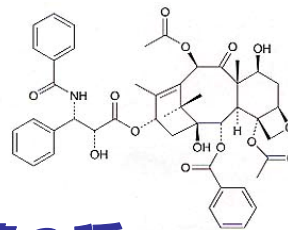
**ケモインフォマティクス**



疾患の  
原因遺伝子の同定



薬の種  
リード化合物の選択



医薬品最適化  
&  
臨床試験



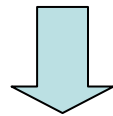
# バイオインフォマティクス 配列解析

## Sequences information

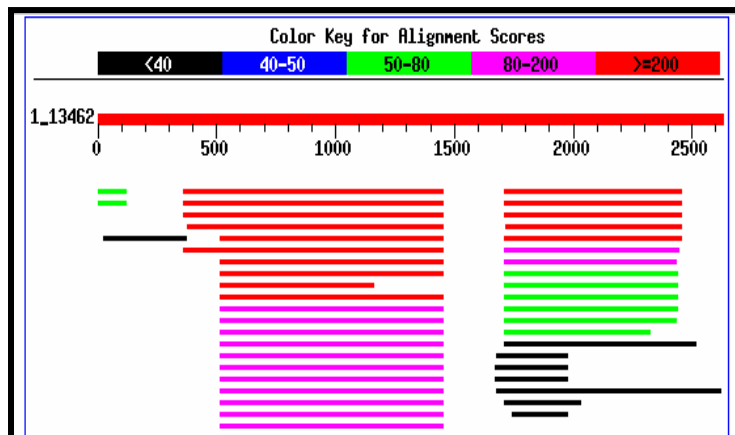
```
>gi119548716|gb|AAL90755.1| adenosine deaminase [Mus musculus]
MAQTPAFNPKPKUELHUHLDGAIKPETILYFGKKRGIALPADTUEELRNIGMDKPLSLPGFLAKFDVYMP
UIAGCREAIKRIAYEFUEMKAKEGUUYUEURYSPLLANSKUDPMPWNQTEGDUTDDUUDLUNQGLQEG
EQAFGIKURSILCCMRHQPSWSLEULELCKKYNQKTUVAAMLADGETIEGSSLFPGHUEAYEGAUKNGIH
RTUHAGEUGSPUEVUREAUDILKTERUGHGVTIEDEALYNRLKENMHFEUCPWSSYLTGAWDPKTTTHAU
URFKNDKANYSLNTDDPLIFKSTLTDYQHTKKDMGFTEEEFKRLNINAAKSSFLPEEEKELLERLYRE
YQ

>gi115831585|ref|Hs15831585| verichia coli 0157:H7]
MIDTTLPLTDIHRHLDGN
ASLDACRRUAFENIEDAARNGLHYUELRFSPGYMAMAHQLPUAGUVEAUIDGUREGCRTFGUQAKLIGIM
SRTFGEAACQEQLEAFLAHRDQITALDLAGDELGPGLSLFSLHFNARDAGWHITUHAGEAAGPESIWQA
IRELGAERIGHGUKAIEDRALMDFLAEQQIGIESCLTSNIQTSTUAEAAHPLKTFLEHGIRASINTDDP
GUQGUDIIEHYTUAAPAAGLSREQIRQAQINGLEMAFLNAEEKRALREKUAAK
```

### Fasta format

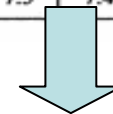


## Alignment (ex. Blast...)

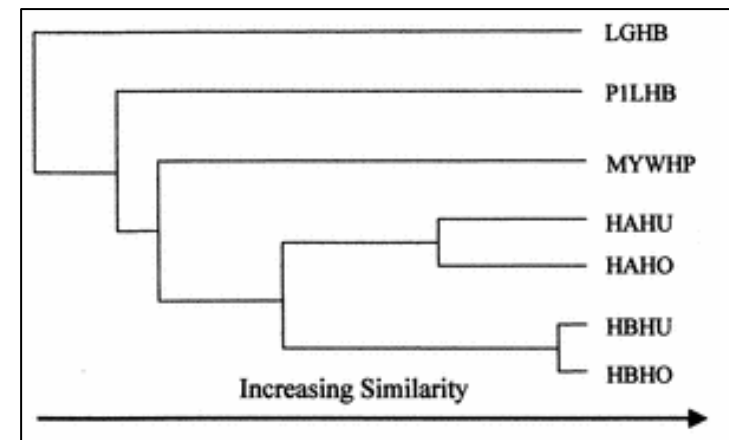


## Similarity matrix

	HAHU	HBHU	HAHO	HBHO	MYWHP	PILHB	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			
PILHB	9.3	8.6	9.6	8.4	7.0		
LGHB	7.1	7.3	7.5	7.4	7.3	4.3	



## Classification



# ケモインフォマティクス

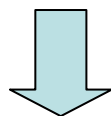
## 構造解析

### Structure

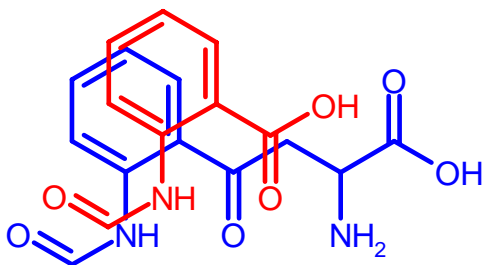
OC(=O)C(N)CC1=CC=C(O)C=C1

```

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
    
```

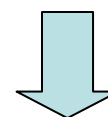


### Structure comparison

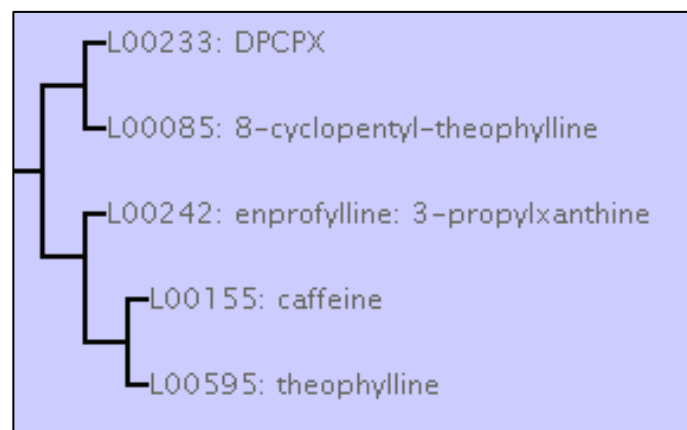


### Distance matrix

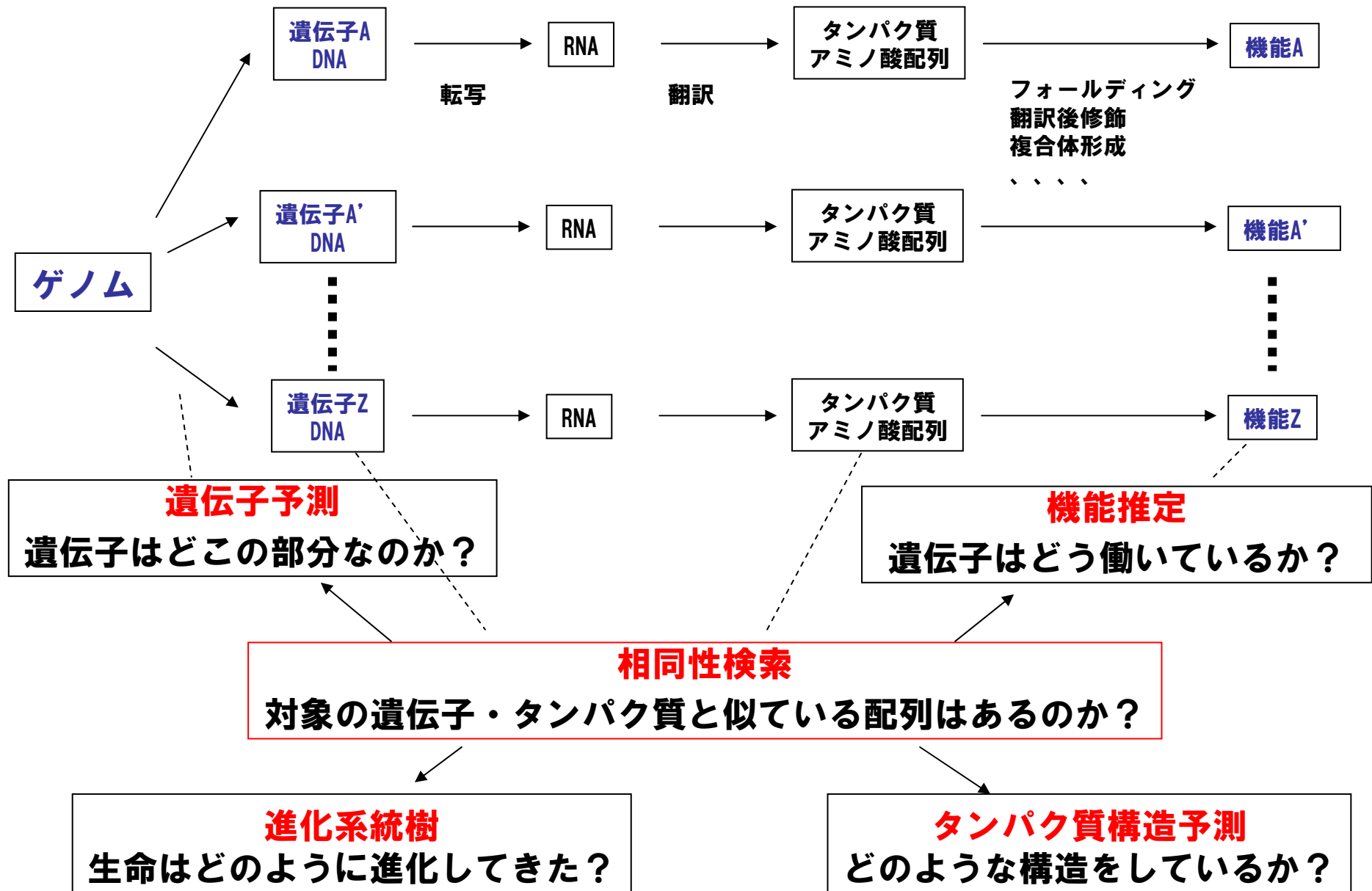
I2	1					
I3	7	5				
I4	13	6	16			
I5	10	0	7	5		
I6	9	9	12	13	12	
I7	11	20	10	9	14	8
	I1	I2	I3	I4	I5	I6



### Classification

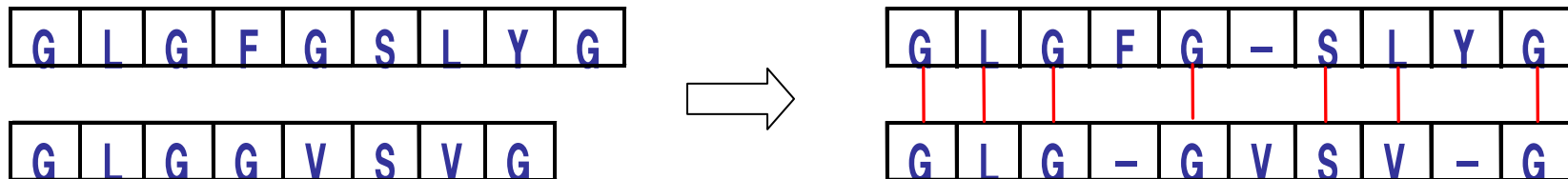


# 配列解析とは



# 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

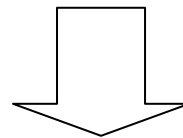


# 最適アライメントを求める (最も類似していると思われる配列の並べ方)

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 – Waterman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ー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

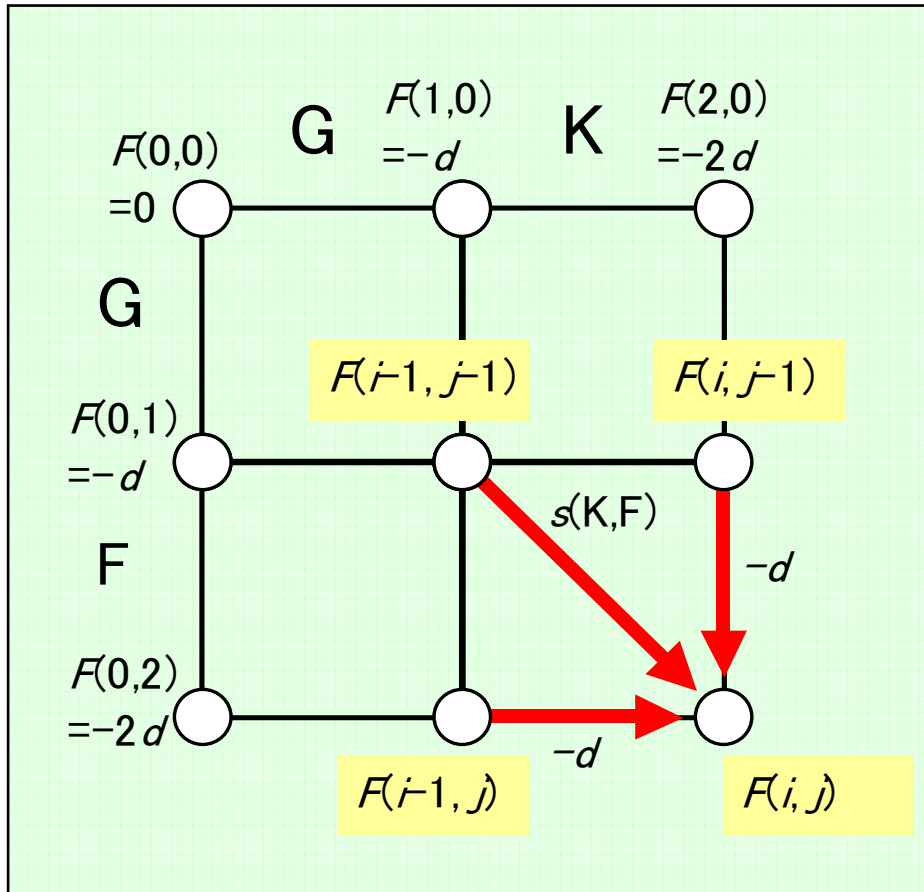
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-4	1	1	1	-6	-3	0
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-4	-1	1	0	-4	-2	-2
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	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	-2	-3	-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	-2	-3	-1	1	-4	-4	-3	15	2	-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
P	0	-8	-16	-24
A	-8	-2	-9	-17
W	-16	-10	-3	-4
	-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 : B L O S U M 5 0

	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 :  $d = -8$

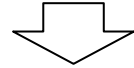
	H	E	A	G	A	W	G	H	E	E	
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12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

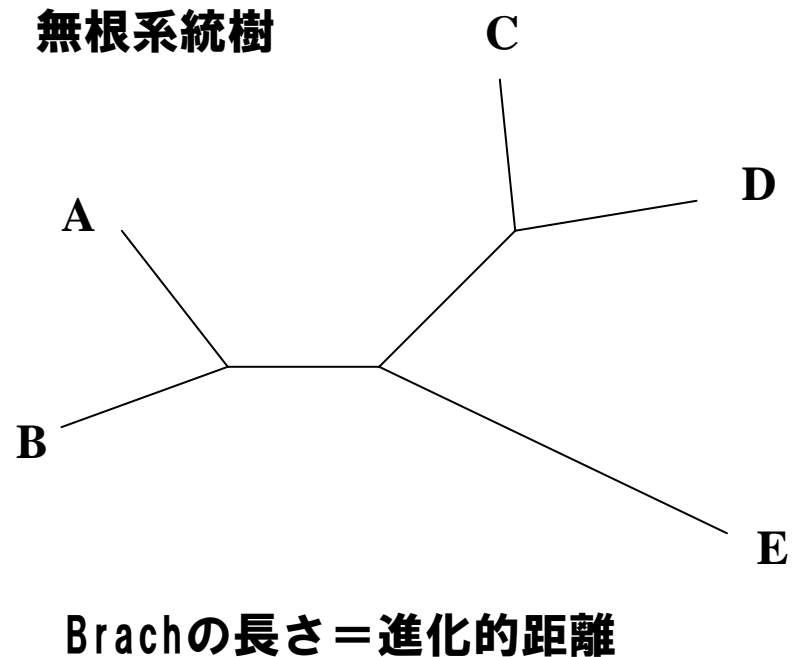
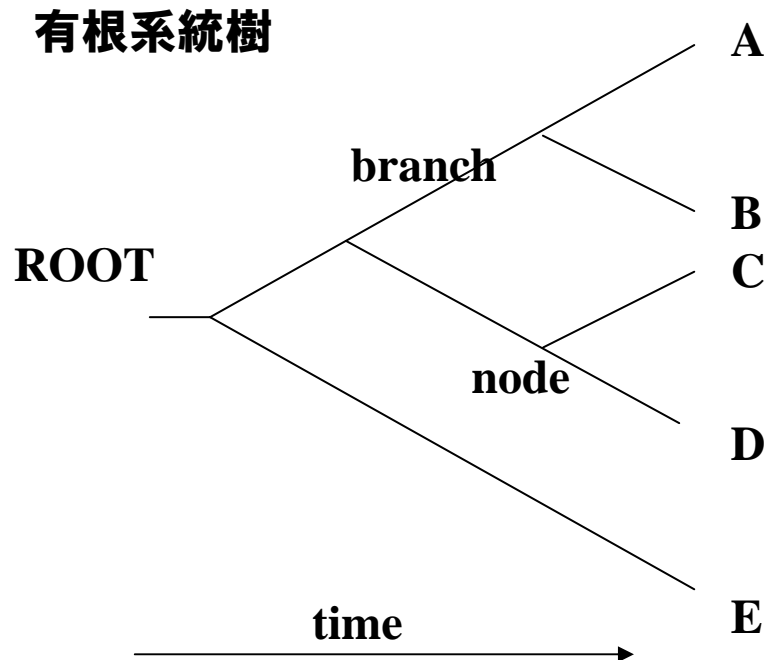
# 進化系統樹

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**



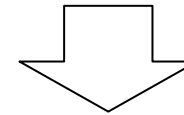
# 進化系統樹の作成方法

• 距離行列法

• 最大節約法

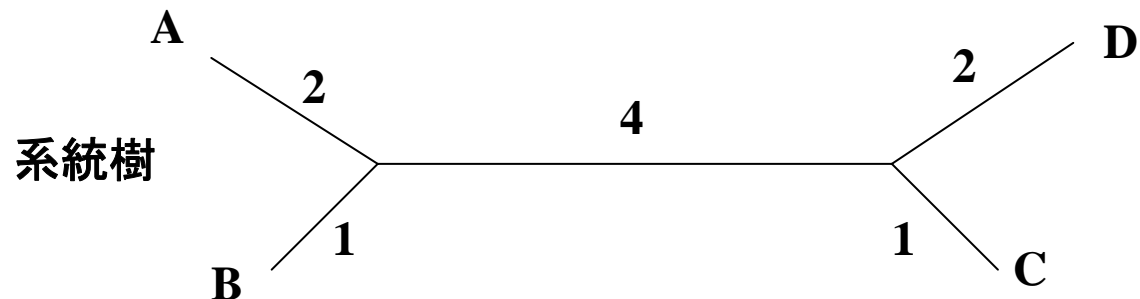
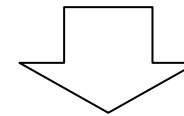
• 最尤法

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	-	-	-	-



# ホモロジーサーチ（相同性検索）

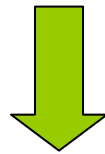
- 相同性検索は対象となる配列と類似の配列が配列データベースに存在するかどうかを検索する手法である。
- 検索する配列（クエリー配列）とデータベース中の配列の間でアライメントを作成し、その中からよく類似した配列を選び出す。



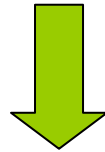


# ホモロジーサーチに用いられるプログラム

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

query . . . A A D E I **M L N** F D G D D V G G E L K ...  
(問い合わせ配列)



類似配列断片のリスト

MLN, MLS, MLK,  
MPN, MPS, MPK,  
LLN, LLS, .....



検索

query . . . A A D E I **M L N** F D G D D V G G E L K ...

データベース . . . A Y D E S **M L S** F D V W D V G N R L K ...



前後に伸ばしながら比較し、ギャップなしで  
高いスコアとなる断片対(HSP)を抽出



最大の統計的評価を与えるHSPの  
組み合わせを抽出

# 実際にBlast検索する（１）

<http://blast.genome.jp/>

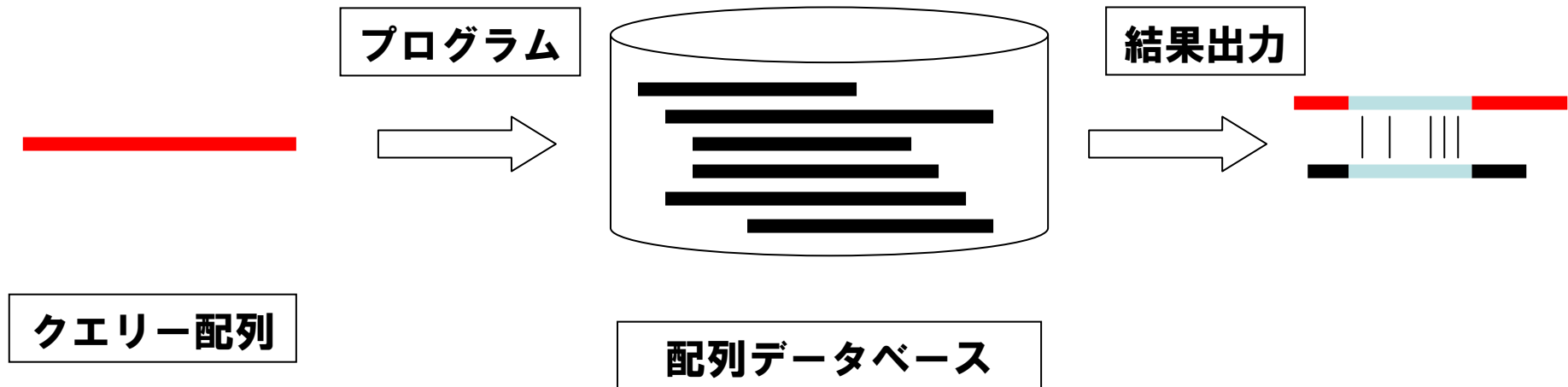
京大・化学研究所・バイオインフォマティクスセンター

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

米国・The National Center for Biotechnology Information (NCBI)

The screenshot shows a web browser window displaying the GenomeNet website. The browser's address bar shows <http://www.genome.jp/>. The website header includes the GenomeNet logo, the text "Bioinformatics Center Institute for Chemical Research Kyoto University", and language options "English" and "日本語". A navigation bar contains links to KEGG, KEGG2, PATHWAY, GENES, LIGAND, BRITE, **BLAST** (highlighted with a red circle), and DBGET. Below the navigation bar is a search form with a dropdown menu set to "KEGG" and a text input field. The main content area is divided into two columns. The left column lists "GenomeNet Database Service" (KEGG, KEGG2, PATHWAY, GENES, LIGAND, BRITE, DRUG / GLYCAN / REACTION / EXPRESSION, KGML, KEGG API, KegArray / KegDraw, GenomeNet FTP) and "GenomeNet Computation Service" (BLAST / FASTA, MOTIF, CLUSTALW / MAFFT / PRRN). The right column contains a "What's New" section and a "Genome database Other computational Acknowledgmer" section. An inset window shows the "BLAST Search" page, which has tabs for "BLAST", "FASTA", and "KEGG2". The "BLAST" tab is active, showing a form to "Enter query sequence: (in one of the three forms)". The form includes fields for "Sequence ID" (with an example "mja:MJ1041"), "Local file name" (with a "ファイルを選択" button and a message "ファイルが選択されていません"), and "Sequence data". There are "Compute" and "Clear" buttons at the top right of the form.

## 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)
MGNAAAARKGSEQESVKEFLAKAKEDFLKKWESPAQNTAHL DQFERIKTLGTGSFGRVML
VKHKETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMV
MEYVPGGEMFSLRRI GRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGY
IQVTDFGFAKRVKGRTWTL CGTPEYLAPEIILSKGYNKAVDWWALGVL IYEMAAGYPPFF
ADQPIQIYEKIVSGKVRFP SHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWFAT
TDWIAIYQRKVEAPFI PKFKGPGDTSNFDDYEEEEIRVSI NEKCGKEFSEF
```



# 実際にBlast検索する（3）



BLAST Search

BLAST FASTA KEGG2

Compute Clear

Enter query sequence: (in one of the three forms)

Sequence ID  (Example) mja:MJ1041

Local file name  ファイルが選択されていません

Sequence data

Select program and database:

☒ BLASTP (prot query vs prot db)  
☐ BLASTX (nucl query vs prot db)

☐ BLASTN (nucl query vs nucl db)  
☐ TBLASTN (prot query vs nucl db)

☒ KEGG GENES  
☐ KEGG GENES+DGENES  
☐ nr-aa  
☐ Swiss-Prot ☐ UniProt ☐ RefSeq ☐ PDBSTR

☐ KEGG GENES  
☐ KEGG GENES+DGENES  
☐ KEGG EGENES  
☐ KEGG GENOME  
☐ nr-nt  
☐ dbEST ☐ dbGSS ☐ HTGs ☐ dbSTS ☐ WGS  
☐ RefSeq ☐ EPD

Output options:

Set the maximum number of database sequences to be reported:

Set the maximum number of alignments to be displayed:

Optional parameters: (see [manual](#) for details)

Scoring matrix:  (except for BLASTN)

Filter:

Alignment view:  (except for BLASTX)

Additional options:

(delimited by whitespaces)

Feedback KEGG GenomeNet Kyoto University Bioinformatics Center

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

BLAST Search Result

Database: genes

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

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

```
BLAST Search Result: hsa:5586 -> genes
http://blast.genome.jp/tmp/blast.3XgSLaex2m/result_blast.h
R_JIN'S PAGE ceg411_wi99_syllabus SMART! ウェブ講座 BioPerlTutorial miRNAMotif アップル .Mac

>hsa:5586 PKN2, PRKCL2; protein kinase N2 [EC:2.7.1.37] [KO:K06071]
Length = 984

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

Query: 40 HLDQFERIKTLGTGSFGRVMLVKHKETGNHYAMKILDQKQVVKLKQIEHTLNEKRILQAV 99
+L F LG G FG+V+L ++K T +A+K L K +V +++ + EKRI + V
Sbjct: 653 NLQDFRCCAVLGRGHFGKVLAEYKNTNEMFAIKALKKGDIVARDEVDLSMCEKRIFETV 712

Query: 100 N---FPFLVKLEFSFKDNSNLYMVEYVPGGEMFSLRRIGRFSEPHARFYAAQIVLTFE 156
N PFLV L F+ ++ VMEY GG++ H+ FSEP A FYAA +VL +
Sbjct: 713 NSVRHPFLVNLFACFQTKHEVCFVMEYAAGGDLMMHIHT-DVFSEPRAVFYAACVVVLGLQ 771

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

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

Query: 274 LQVDLTKRFGNLKNGVNDIKNHKWFATDWDIAIYQRKVEAPFIPKFKGPGDTSNFDD 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]
Length = 683

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

Query: 41 LDQFERIKTLGTGSFGRVMLVKHKETGNHYAMKILDQKQVVKLKQIEHTLNEKRILQ-AV 99
+D FE I+ LG GSFG+VML + KETG YA+K+L K +++ +E T+ EKRI L A
Sbjct: 352 IDNFEFIRVLGKGSFGKVMARIKETGELYAVKVLKKDVILQDDDDVECTMTTEKRILSLAR 411

Query: 100 NFPFLVKLEFSFKDNSNLYMVEYVPGGEMFSLRRIGRFSEPHARFYAAQIVLTFFEYLH 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

 <b>Rattus norvegicus (rat): 81749</b> <span>Help</span>	
<b>Entry</b>	81749 CDS <a href="#">R.norvegicus</a>
<b>Gene name</b>	Prkch
<b>Definition</b>	protein kinase C, eta [EC:2.7.1.-]
<b>KO</b>	KO: <a href="#">K06068</a> novel protein kinase C <a href="#">OC search</a> <a href="#">OC viewer</a>
<b>Pathway</b>	PATH: <a href="#">rno04530</a> Tight junction
<b>Class</b>	<a href="#">Gene catalog</a>
<b>SSDB</b>	<a href="#">Ortholog</a> <a href="#">Paralog</a> <a href="#">Gene cluster</a>
<b>Motif</b>	Pfam: <a href="#">C2</a> <a href="#">C1_1</a> <a href="#">Pkinase</a> <a href="#">Pkinase_C</a> PROSITE: <a href="#">PROTEIN_KINASE_ST</a> <a href="#">PROTEIN_KINASE_ATP</a> <a href="#">DAG_PE_BIND_DOM_1</a> <a href="#">DAG_PE_BIND_DOM_2</a> <a href="#">C2_DOMAIN_2</a> <a href="#">PROTEIN_KINASE_DOM</a> <a href="#">Motif</a>
<b>Other DBs</b>	RGD: <a href="#">621888</a> NCBI-GI: <a href="#">13592027</a> NCBI-GeneID: <a href="#">81749</a> UniProt: <a href="#">Q64617</a>
<b>LinkDB</b>	<a href="#">PDB</a> <a href="#">All DBs</a>
<b>Position</b>	6q24
<b>AA seq</b>	683 aa <a href="#">AA seq</a> <a href="#">DB search</a> MSSGTMKFNGYLVRVIGEAVGLQPTRWSLRHSLFKKGHQLLDPYLTVSVDQVRVGQTSTK QKTNKPTYNEEFCTNVSDGGHLELAVFHETPLGYDHFVANCTLQFQELLRTAGTSDTFEG WVDLEPEGKVFFVITLTGSFTEATLQDRIFKHFTKRQRAMRRRVHQVNGHKFMATYLR QPTYCSHCREFIWGVFGKQGYQCQVCTCVVHKRCHHLIVTACTCQNNINKVDKIAEQRF GINIPHKFNVHNYKVPTFCDHCGSLWLGIMRQGLQCKICKMNVHIRCQANVAPNCGVNAV ELAKTLAGMGLQPGNISPTSKLISRSTLRRQKKEGSKEGNGIGVNSSSRFGIDNFEFIRV LGKGSFGKVMLARIKETGELYAVKVLKDDVILQDDDVECTMTEKRILSLARNHPFLTQLF CCFQTPDRLFFVMEFVNGGDLMFHIQKSRRFDEARARFYAAEIIISALMFLHEKGIIYRDL KLDNVLLDHEGHCKLADFGMCKEGICNGVTTATFCGTPDYIAPEILQEMLYGPAVDWWAM GVLLYEMLCGHAPFEAENEDDLFEAILNDEVVYPTWLHEDATGILKSFMTKNPTMRLGSL TOGGEHETLRHPFFKETDWOI.NHROI.EPPFRPRTKSREDVSNFDPDFTKKEPVI.TPTDE



# 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

The image shows a screenshot of a web browser displaying the GenomeNet website and the Multiple Sequence Alignment - CLUSTALW interface. The browser's address bar shows the URL <http://www.genome.jp/>. The GenomeNet website has a navigation menu with links to KEGG, KEGG2, PATHWAY, GENES, and LIGAND. A search bar is present with the text "Search KEGG for". The main content area lists various services, including "GenomeNet Database Service" and "GenomeNet Computation Service". The "GenomeNet Computation Service" section is circled in red, and a red arrow points from it to the CLUSTALW interface. The CLUSTALW interface has a header with the title "Multiple Sequence Alignment by CLUSTALW" and a sub-header "CLUSTALW". It includes a "General Setting Parameters" section with options for "Output Format" (CLUSTAL), "Pairwise Alignment" (FAST/APPROXIMATE), and "Enter your sequences (with labels) below (copy & paste):". The "Enter your sequences" section contains a text area with the following sequence: 

```
PQVEQLLEGGSPGDLQTLALEVARQKRGIVDQCCTICSILYQLENYCN
>INS1_RAT
MALWMRFLPLLALLVLWEKPAQAFVKQHLGPHLVEALYLVCGERGFFYTPKSRREV
PQVPELLEGGPEAGDLQTLALEVARQKRGIVDQCCTICSILYQLENYCN
```

 Below the text area are buttons for "Execute Multiple Alignment" and "Reset". The interface also includes a "More Detail Parameters..." section and a "Pairwise Alignment Parameters:" section.

# 実際に系統樹を作成する（２）

## MultiFASTA形式

```
>INS_HUMAN
MALWMRLLPLLALLALWGPDPAAAFVNQH
PKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRG
>INS_BOVIN
MALWTRLRPLLALLALWPPPPARAFVNQHL
KARREVEG
PQVGALELAGGPGAGGLEGPPQKRGIVEQC
>INS_PIG
MALWTRLRPLLALLALWAPAPAQAFVNQHI
KARREAEN
PQAGAVELGGGLGGLQALALEGPPQKRGIV
>INS_CYPCA
MAVWIQAGALLFLLAVSSVNANAGAPQHLC
RDVDPPLG
>INS_CHICK
MALWIRSLPLLALLVFSGPGTSYAAANQHLC
ARRDVEQ
```



## Multiple Sequence Alignment by CLUSTALW

CLUSTALW      MAFFT      PRRN

General Setting Parameters:

Output Format:

Pairwise Alignment: ☒ FAST/APPROXIMATE ☐ SLOW/ACCURATE

Enter your sequences (with labels) below (copy & paste): ☒ PROTEIN ☐ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

PQVEQ...AGGSPGDLQTLALEVARQKRGIVDQCCTICSISLYQLENYCN  
>INS1...  
MALWMRLLPLLALLLWEPKPAQAFVKQHLCCGPHLVEALYLVCGERGFFYTPKSRREVED  
PQVPQLELGGGPEAGDLQTLALEVARQKRGIVDQCCTICSISLYQLENYCN

Or give the file name containing your query

ファイルが選択されていません

More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size:  , Window size:  , Gap Penalty:

Number of Top Diagonals:  , Scoring Method:

For SLOW/ACCURATE:

Gap Open Penalty:  , Gap Extension Penalty:

Select Weight Matrix:

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

# 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 MALWTRLRPLLALLALWPPPPARAFVNQHLGSHLVEALYLVCGERGFFYTPKARREVEG  
INS\_PIG MALWTRLRPLLALLALWAPAPAQAFVNQHLGSHLVEALYLVCGERGFFYTPKARREAEN  
INS\_HUMAN MALWMRLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED  
INS\_CERAE MALWMRLPLLALLALWGPDPVFAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED  
INS1\_MOUSE MALLVHFLPLLALLALWEPKFTQAFVNQHLGSHLVEALYLVCGERGFFYTPKSRREVED  
INS1\_RAT MALWMRFLPLLALLLWEPKPAQAFVNQHLGSHLVEALYLVCGERGFFYTPKSRREVED  
INS\_CHICK MALWIRSLPLLALLVFSGGPGTSYAAANQHLGSHLVEALYLVCGERGFFYSPKARRDVEQ  
INS\_CYPCA MAVWIQAGALLFLLAVSSVN-ANAGAPQHLGSHLVDALYLVCGPTGFFYNPK--RDVDP  
INS\_BRARE MAVWLQAGALLVLLVSSVS-TNPGTPQHLGSHLVDALYLVCGPTGFFYNPK--RDVEP  
\*\* : .\*\* \*\*.. . . \*\*\*\*\*:\*\*\*\*\* \*\*\*,\*\* \*:.

INS\_BOVIN PQVGALELAGGP--G---AGGLEGPQKRGIQCCASVCSLYQLENYCN  
INS\_PIG PQAGAVELGGGL--GGLQALALEGPPQKRGIQCCSICSLYQLENYCN  
INS\_HUMAN LQVGQVELGGGPGAGSLQPLALEGSLQKRGIQCCSICSLYQLENYCN  
INS\_CERAE PQVGQVELGGGPGAGSLQPLALEGSLQKRGIQCCSICSLYQLENYCN  
INS1\_MOUSE PQVEQLELGGSP--GDLQTLALEVARQKRGIQCCSICSLYQLENYCN  
INS1\_RAT PQVPQLELGGGPEAGDLQTLALEVARQKRGIQCCSICSLYQLENYCN  
INS\_CHICK PLG-----  
INS\_CYPCA LLG-----  
INS\_BRARE

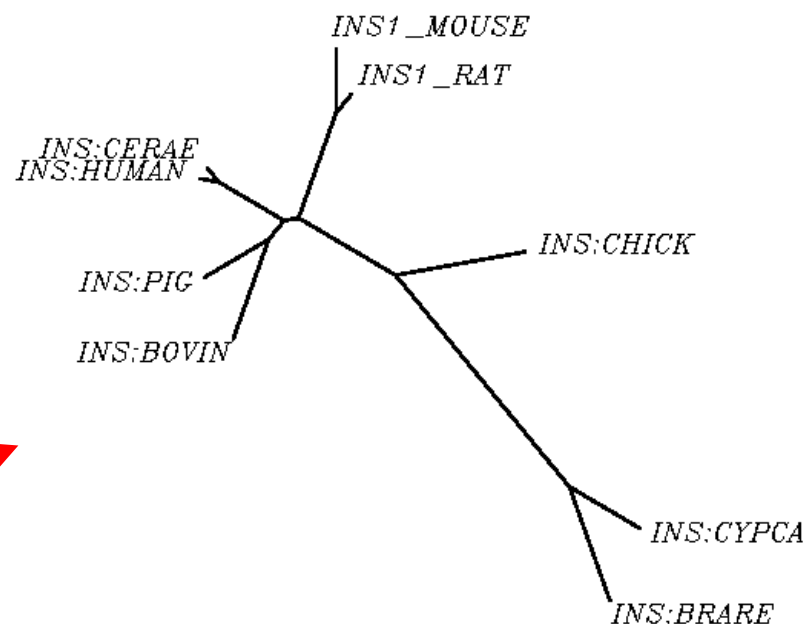
clustalw.dnd

```
(
(
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

## マルチプルアライメント結果



# ケモインフォマティクス

## 構造解析

### Structure

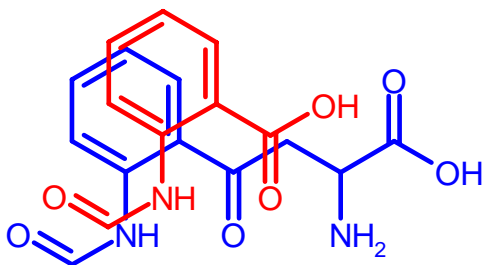
OC(=O)C(N)CC1=CC=C(O)C=C1

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

### Distance matrix

I2	1			
I3	7	5		
I4	13	6	16	
I5	10	0	7	5

- ◆ 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



-L00085: 8-cyclopentyl-theophylline

-L00242: enprofylline: 3-propylxanthine

-L00155: caffeine

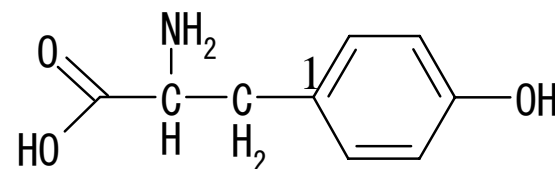
-L00595: theophylline

# 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

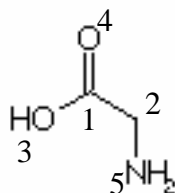


## 2. Connection Table

### 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
```

///



C00037

### MDL CT format

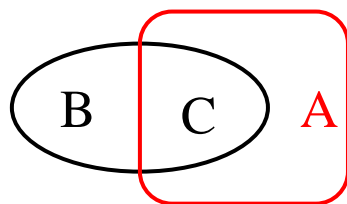
```
ISISHOST03020323002D 1 1.00000 0.00000 37

5 4 0 0 0 999 V2000
-0.1276 0.2621 0.0000 C 0 0 0 0 0 0 0 0 0 0
0.5552 -0.1862 0.0000 C 0 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
-0.1552 1.0931 0.0000 O 0 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 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

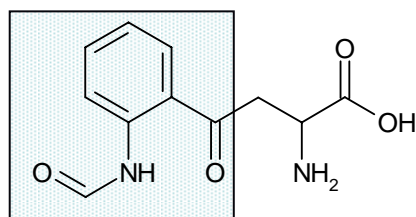
b: size of mol\_B

c: size of overlap

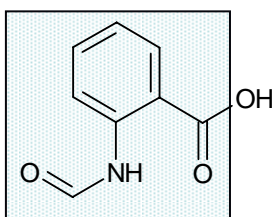
Tanimoto coefficient

$$= c / (a+b-c)$$

### structure



formylkynurenine



formylanthranilate

$$\begin{aligned} a &= 17 & c &= 11 & b &= 12 \\ \rightarrow 11 / (17 + 12 - 11) &= 0.61 \end{aligned}$$

### fingerprint

Mol A: 0101011001010000100100

Mol B: 0000101010010010000100

$$\begin{aligned} a &= 8 & c &= 3 & b &= 6 \\ \rightarrow 3 / (8 + 6 - 3) &= 0.27 \end{aligned}$$

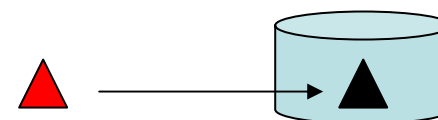
# 構造検索(データベースサーチ)

## 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？

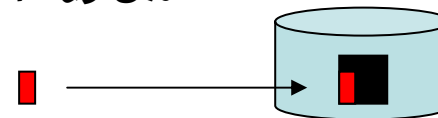
## 2. Substructure search

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

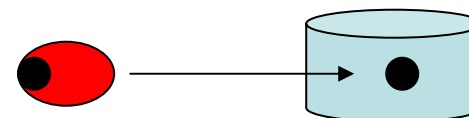
問い合わせ分子中の部分構造と一致する分子がDB中にあるか？



## 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 structure”

### 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*

The screenshot shows a web browser window with the address bar displaying `http://pubchem.ncbi.nlm.nih.gov/`. The browser's search bar contains the word "pubchem". The page header features the NCBI logo, the PubChem logo, and the National Library of Medicine (NLM) logo. Below these logos is a navigation menu with links: HOME, SEARCH, SITE MAP, PubMed, Entrez, Structure, GenBank, PubChem, and Help. The main content area is titled "PubChem Text Search" and includes a search form with a dropdown menu set to "PubChem Compound", a text input field, and a "GO" button. Below the search form, a paragraph states: "PubChem provides information on the biological activities of small molecules. It is a component of NIH's [Molecular Libraries Roadmap Initiative](#). If you would like to learn more about how to use the PubChem resources, please go to our [help page](#)." Two announcement boxes follow, each starting with a "New" star icon. The first announcement says: "BioAssay data from [University of Pittsburgh Molecular Library Screening Center](#) are now available in PubChem." The second announcement says: "Structures and BioAssay data from [MTDP \(Molecular Targets Development Program\)](#) are now available in PubChem." Below these announcements is a link: "More PubChem announcements ...". At the bottom, there are two informational boxes. The first, titled "PubChem Compound:", explains that users can search unique chemical structures using names, synonyms, or keywords, and that links to biological property information are provided for each compound. The second, titled "PubChem Substance:", explains that users can search deposited chemical substance records using... The browser's status bar at the bottom shows the "インターネット" (Internet) icon.

アドレス(D) `http://pubchem.ncbi.nlm.nih.gov/`

Google pubchem 検索 PageRank 14 をブロックしました ABC チェック オプション pubchem

NCBI PubChem National Library of Medicine NLM

HOME SEARCH SITE MAP PubMed Entrez Structure GenBank PubChem Help

### PubChem Text Search

PubChem Compound  GO

PubChem provides information on the biological activities of small molecules. It is a component of NIH's [Molecular Libraries Roadmap Initiative](#). If you would like to learn more about how to use the PubChem resources, please go to our [help page](#).

**New** BioAssay data from [University of Pittsburgh Molecular Library Screening Center](#) are now available in PubChem.

**New** Structures and BioAssay data from [MTDP \(Molecular Targets Development Program\)](#) are now available in PubChem.

[More PubChem announcements ...](#)

**PubChem Compound:** Search unique chemical structures using names, synonyms or keywords. Links to available biological property information are provided for each compound.

**PubChem Substance:** Search deposited chemical substance records using

インターネット