平成22年度・近畿大学・農学部・生命情報学

マルチプルアライメントと その応用

2010年4月27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

平成22年度「生命情報学＆生命情報学実習」講義日程

2010.3.25

	講義	生命情報学	演習	生命情報学演習
4/13	川端1	分子生物学の基礎と配列データベース		
4/20	川端2	ペアワイズアライメントと配列相同性検索	川端	主要WEBデータベースと配列相同性検索
4/27	川端3	マルチプルアライメントとその応用		
5/11	川端4	分子系統学基礎	川端	マルチプルアライメントと系統樹作成演習
5/18	川端5	蛋白質の物理化学的性質とアミノ酸配列解析		
5/25	川端6	蛋白質立体構造データの情報解析	川端	蛋白質立体構造データの可視化
6/1	川端	>>試験（川端　担当分）<<		
6/8	中村1	化学構造データと計算化学基礎I		
6/15	中村2	化学構造データと計算化学基礎II	中村	ChemOfficeを用いた計算化学演習
6/22	中村	>>試験（中村担当分）<<		
6/29	金谷1	トランスクリプトーム解析		
7/6	金谷2	インタラクローム解析	金谷	発現プロフィール解析演習
7/13	金谷3	メタボローム解析	金谷	インタラクトローム・代謝物解析演習
7/20	金谷	>>試験（金谷担当分）<<		

マルチプルアライメント

（multiple sequence alignment
多重配列整列）

マルチプルアライメント（多重配列整列）とは

3本以上の配列を進化的な対応関係に従って並べること

```
>1nshA
SRPTETERCIESLIAVFQKYAGKDGHSVTLSKTEFLSFMNTELAAFTKNQKDPGVLDRMMKKLDLNSDGQLDFQEFL
NLIGGLAVAESFVKAAPPQKRF
>1j55A
MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLDAVDKLLKDLDANGDAQVDFSEFIVFVAAITS
ACHKYFEKAL
>1ig5A
KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ
>1qx2A
MKSPEEIKGAFEVFAAKEGDPNQISKEELKLVMQTLGPSLLKGMSTLDEMIEEVDKNGDGEVSFEEFLVMMKKISQ
```

```
CLUSTAL W (1.83) multiple sequence alignment

1nshA      SRPTETERCIESLIAVFQKYAGKDGHSVTLSKTEFLSFMNTELAAFTKNQKDPGVLDRMM
1j55A      --MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLD------AVDKLL
1ig5A      -------KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKG---PSTLDELF
1qx2A      ------MKSPEEIKGAFEVFAAKEGDPNQISKEELKLVMQTLGPSLLKG---MSTLDEMI
                  .   :   *. ::..:* .  ::* *:  .::.  ..: .      .:*.::

1nshA      KKLDLNSDGQLDFQEFLNLIGGLAVACHESFVKAAPPQKRF
1j55A      KDLDANGDAQVDFSEFIVFVAAITSACHKYFEKAGL-----
1ig5A      EELDKNGDGEVSFEEFQVLVKKISQ----------------
1qx2A      EEVDKNGDGEVSFEEFLVMMKKISQ----------------
           :.:* *.*.::.*.**  ::  ::
```

マルチプルアライメントの目的

```
1nshA      SRPTETERCIESLIAVFQKYAGKDGHSVTLSKTEFLSFMNTELAAFTKNQKDPGVLDRMM
1j55A      --MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLD------AVDKLL
1ig5A      -------KSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKG---PSTLDELF
1qx2A      ------MKSPEEIKGAFEVFAAKEGDPNQISKEELKLVMQTLGPSLLKG---MSTLDEMI
                  .   :   *. ::..:* .  ::* *:  .::.  ..: .      .:*.::
```

- ファミリ内の機能的重要部位の検出
- ファミリを特徴付けるモチーフの発見
- プロフィール法による遠縁のホモログ発見
- 分子系統解析の第一ステップとして不可欠
- 進化的追跡法(evolutionary trace method)

多重整列のスコア

（1）SP（sum-of-pairs)スコア

複数の文字列間のスコアを
ペアワイズのアミノ酸置換スコア$s(a,b)$の和で表す

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

m_i^k : k 番目の配列 の i番目の文字

RCIAVF
TAMDVF
KSPGIF

$S(m_1) = s(\mathrm{R},\mathrm{T}) + s(\mathrm{T},\mathrm{K}) + s(\mathrm{R},\mathrm{K})$

理論的にはおかしい： $S(A,B)+S(B,C)+S(A,C) = \log\frac{P(A,B)P(B,C)P(A,C)}{P(A)^2P(B)^2P(C)^2} \neq \log\frac{P(A,B,C)}{P(A)P(B)P(C)}$

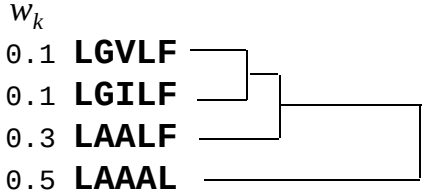
BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

多重配列のスコア（続き）

（2）配列への重み付きのSum-of-pair関数 (ClustalW)

$$S(m_i) = \sum_{k<l} w_k \cdot w_l \cdot s(m_i^k, m_i^l)$$



（3）エントロピー関数の最小化

各サイトのアミノ酸の頻度$p_i(a)$を推定し、そのエントロピーの和を求める

$$S(m_i) = -\sum_a p_i(a) \log p_i(a)$$

12345
LGVLF
LGILF
LAALF
LAAAL

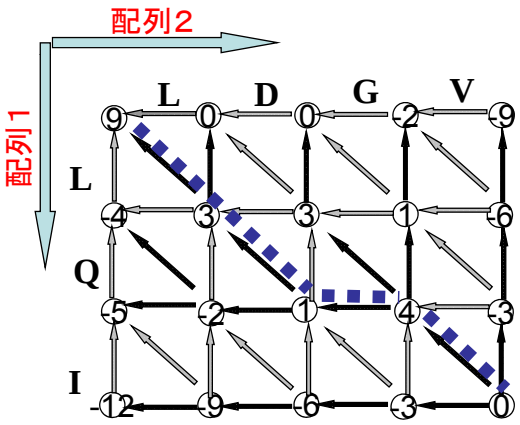
サイト	$Pi(a)$	$S(m_i)$
1	$P_1(L)=1.0$,	0.00
2	$P_2(G)=0.5$,$P_2(A)=0.5$	0.69
3	$P_3(V)=0.25$, $P_3(I)=0.25$, $P_3(A)=0.5$	1.04

（4）対アライメントライブラリの重複による部位特異的スコア (T-COFFEE)

どうやって並べるか？

多次元DPによる多重配列の厳密解

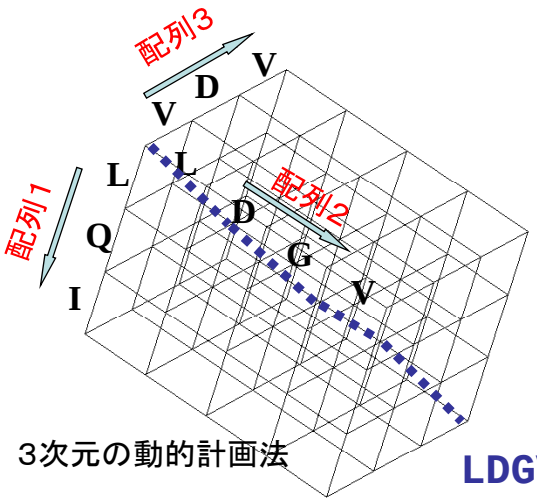
2本の配列のアライメント



2次元の動的計画法　LDGV / LQ-I

メモリ・計算時間　$O(L^2)$

3本の配列のアライメント



3次元の動的計画法　LDGV / LQ-I / VD-V

メモリ・計算時間　$O(L^3)$

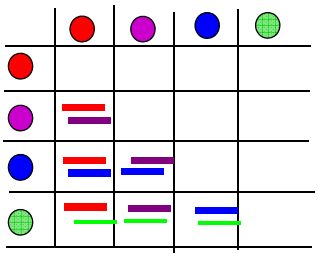
N本の配列のアライメントのメモリ・計算時間は$O(L^N)$→非現実的

長さ100の2本のアライメントが1秒でできても、10本に増やすと100^8 秒かかる。

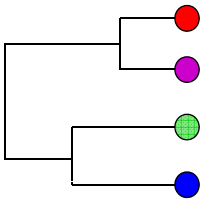
プログレッシブ・アライメント
(progressive alignment, 累進法)

Feng and Doolittle (1987)

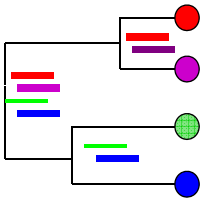
（1）全ての配列ペアのペアワイズアライメントを計算する



（2）ペアワイズアライメントによる距離行列を計算し、樹形図を計算する。



（3）樹形図の葉から、ペアワイズアライメントを組み上げていく



※ステップ1に最も計算時間がかかる。
全体の計算量は[配列の本数]^2×[配列の長さ]にほぼ比例

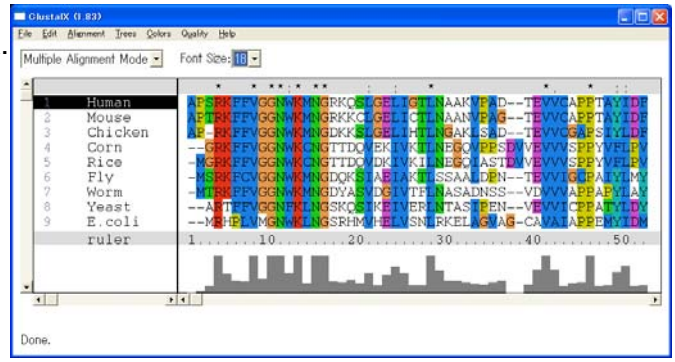
ClustalW / ClustalX

UNIX/Windows/Mac版：ftp://ftp.ebi.ac.uk/pub/software/clustalw2
WEBサーバ：http://www.ebi.ac.uk/Tools/clustalw2

・現在、最も一般的な多重整列のプログラム
・アルゴリズムは累進法。ペアワイズアライメントはグローバルアライメントを用い、ガイド木はNJ法で　作成。スコアは配列の重みを導入したSum-of-pairs。置換スコア行列の選択、ギャップペナルティ等に様々な経験的な工夫が見られる。

・CUI版はClustalW, GUI版はClustalX. UNIX, Windows, MACでも動作する。

・NJ法による系統樹計算機能付き。



Thompson, J.D., Higgins, D.G., Gibson T.J. “CLUSTALW : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”. Nucleic Acids Reseach, 1994, 22, 4673-4680.

主要なマルチプルアライメントのプログラム

	WEBサイト	アルゴリズム	特徴
ClustalW・ClustalX	http://www.ebi.ac.uk/Tools/clustalw2	累進法。重み付きSPスコアを使用。 置換スコア行列の選択、ギャップペナルティ等に様々な工夫	もっとも広く使われている標準的なプログラム
T-COFFEE	http://www.ebi.ac.uk/t-coffee/	ペアワイスアライメントをローカル、グローバル、進展を用いて多数生成。それらの集合から、位置特異的スコアを作成し、累進法を実行する。	計算時間がかかるが精度は高い。**配列の本数が100本以下**の場合に向いている。
MAFFT	http://align.bmr.kyushu-u.ac.jp/mafft/online/server/	高速フーリエ変換(FFT)を用いて、高速にペアワイズアライメントを実装、それを利用して、累進法、あるいは反復改善法を実行する。	計算時間は高速なので、**配列の本数が100～500本程度**でも、計算可能。

サイトの保存度による
機能部位予測

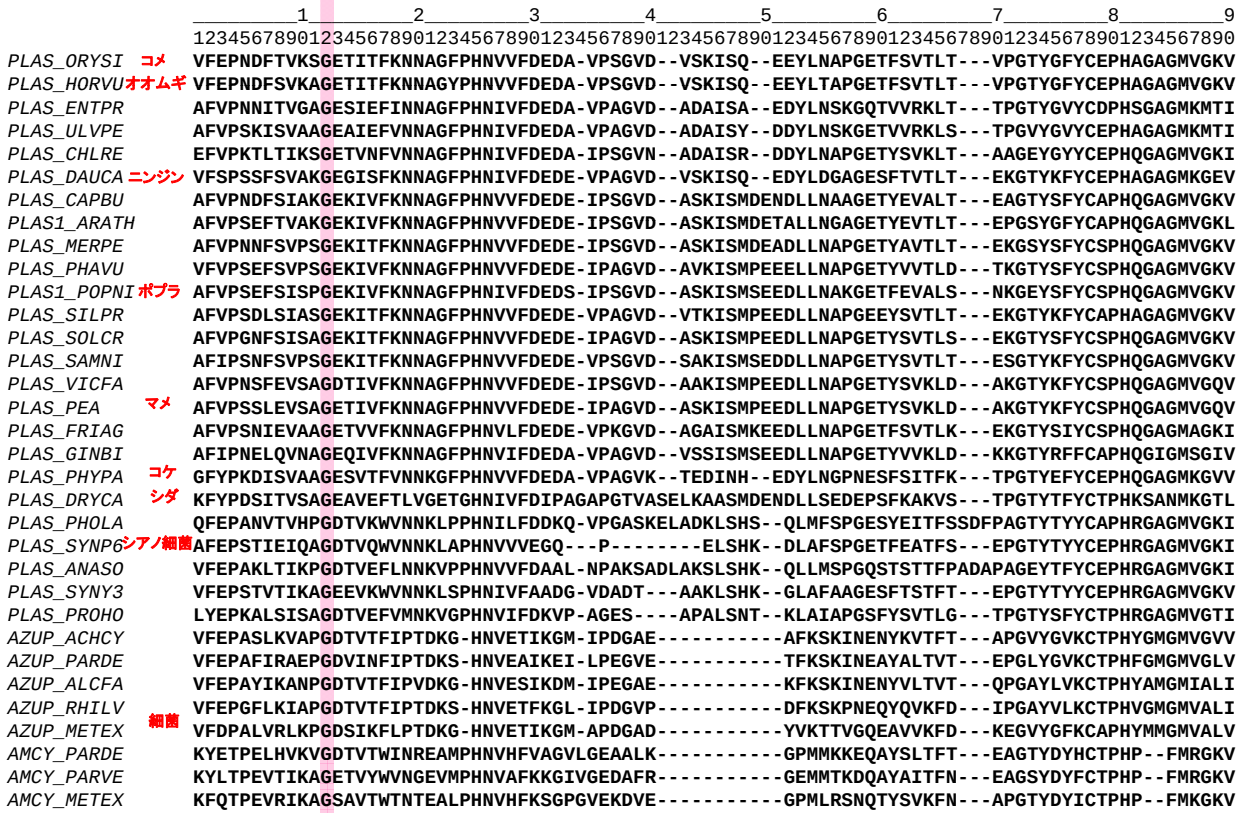
サイトごとに保存の度合いに差がある

よく保存しているサイト → そのファミリーにとって重要なサイト
→機能上重要なサイトである可能性が高い

```
5p21-   MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSY
1ctqA   MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSY
1c1yA   MREYKLVVLGSGGVGKSALTVQFVQGIFVEKYDPTIEDSY
1kao-   MREYKVVVLGSGGVGKSALTVQFVTGTFIEKYDPTIEDFY
1huqA   --QFKLVLLGESAVGKSSLVLRFVKGQFHEYQESTIGAAF
1g16A   ----KILLIGDSGVGKSCLLVRFVE----DKFNPI--DFK
1ek0A   VTSIKLVLLGEAAVGKSSIVLRFVSNDFAENKEPTIGAAF
3rabA   ---FKILIIGNSSVGKTSFLFRYADDSFTPAFVSTVGIDF
1mh1-   ----KCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNY
2ngrA   MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNY
1tx4B   ----KLVIVGDGACGKTCLLIVNSKDQF---YVPTVFENY
```

サイトごとに保存の度合いに差がある。
サイトごとにアミノ酸の出現傾向に差がある

[AG]-x(4)-G-K-[ST]



```
                    _________1_________2_________3_________4_________5_________6_________7_________8_________9
                    123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
PLAS_ORYSI  コメ    VFEPNDFTVKSGETITFKNNAGFPHNVVFDEDA-VPSGVD--VSKISQ--EEYLNAPGETFSVTLT---VPGTYGFYCEPHAGAGMVGKV
PLAS_HORVUオオムギ  VFEPNDFSVKAGETITFKNNAGYPHNVVFDEDA-VPSGVD--VSKISQ--EEYLTAPGETFSVTLT---VPGTYGFYCEPHAGAGMVGKV
PLAS_ENTPR          AFVPNNITVGAGESIEFINNAGFPHNIVFDEDA-VPAGVD--ADAISA--EDYLNSKGQTVVRKLT---TPGTYGVYCDPHSGAGMKMTI
PLAS_ULVPE          AFVPSKISVAAGEAIEFVNNAGFPHNIVFDEDA-VPAGVD--ADAISY--DDYLNSKGETVVRKLS---TPGVYGVYCEPHAGAGMKMTI
PLAS_CHLRE          EFVPKTLTIKSGETVNFVNNAGFPHNIVFDEDA-IPSGVN--ADAISR--DDYLNAPGETYSVKLT---AAGEYGYYCEPHQGAGMVGKI
PLAS_DAUCA ニンジン VFSPSSFSVAKGEGISFKNNAGFPHNIVFDEDE-VPAGVD--VSKISQ--EDYLDGAGESFTVTLT---EKGTYKFYCEPHAGAGMKGEV
PLAS_CAPBU          AFVPNDFSIAKGEKIVFKNNAGFPHNVVFDEDE-IPSGVD--ASKISMDENDLLNAAGETYEVALT---EAGTYSFYCAPHQGAGMVGKV
PLAS1_ARATH         AFVPSEFTVAKGEKIVFKNNAGFPHNVVFDEDE-IPSGVD--ASKISMDETALLNGAGETYEVTLT---EPGSYGFYCAPHQGAGMVGKL
PLAS_MERPE          AFVPNNFSVPSGEKITFKNNAGFPHNVVFDEDE-IPSGVD--ASKISMDEADLLNAPGETYAVTLT---EKGSYSFYCSPHQGAGMVGKV
PLAS_PHAVU          VFVPSEFSVPSGEKIVFKNNAGFPHNVVFDEDE-IPAGVD--AVKISMPEEELLNAPGETYVVTLD---TKGTYSFYCSPHQGAGMVGKV
PLAS1_POPNI ポプラ  AFVPSEFSISPGEKIVFKNNAGFPHNIVFDEDS-IPSGVD--ASKISMSEEDLLNAKGETFEVALS---NKGEYSFYCSPHQGAGMVGKV
PLAS_SILPR          AFVPSDLSIASGEKITFKNNAGFPHNVVFDEDE-VPAGVD--VTKISMPEEDLLNAPGEEYSVTLT---EKGTYKFYCAPHAGAGMVGKV
PLAS_SOLCR          AFVPGNFSISAGEKITFKNNAGFPHNVVFDEDE-IPAGVD--ASKISMPEEDLLNAPGETYSVTLS---EKGTYSFYCSPHQGAGMVGKV
PLAS_SAMNI          AFIPSNFSVPSGEKITFKNNAGFPHNVVFDEDE-VPSGVD--SAKISMSEDDLLNAPGETYSVTLT---ESGTYKFYCSPHQGAGMVGKV
PLAS_VICFA          AFVPNSFEVSAGDTIVFKNNAGFPHNVVFDEDE-IPSGVD--AAKISMPEEDLLNAPGETYSVKLD---AKGTYKFYCSPHQGAGMVGQV
PLAS_PEA    マメ    AFVPSSLEVSAGETIVFKNNAGFPHNVVFDEDE-IPAGVD--ASKISMPEEDLLNAPGETYSVKLD---AKGTYKFYCSPHQGAGMVGQV
PLAS_FRIAG          AFVPSNIEVAAGETVVFKNNAGFPHNVLFDEDE-VPKGVD--AGAISMKEEDLLNAPGETFSVTLK---EKGTYSIYCSPHQGAGMAGKI
PLAS_GINBI          AFIPNELQVNAGEQIVFKNNAGFPHNVIFDEDA-VPAGVD--VSSISMSEEDLLNAPGETYVVKLD---KKGTYRFFCAPHQGIGMSGIV
PLAS_PHYPA  コケ    GFYPKDISVAAGESVTFVNNKGFPHNVVFDEDA-VPAGVK--TEDINH--EDYLNGPNESFSITFK---TPGTYEFYCEPHQGAGMKGVV
PLAS_DRYCA  シダ    KFYPDSITVSAGEAVEFTLVGETGHNIVFDIPAGAPGTVASELKAASMDENDLLSEDEPSFKAKVS---TPGTYTFYCTPHKSANMKGTL
PLAS_PHOLA          QFEPANVTVHPGDTVKWVNNKLPPHNILFDDKQ-VPGASKELADKLSHS--QLMFSPGESYEITFSSDFPAGTYTYYCAPHRGAGMVGKI
PLAS_SYNP6シアノ細菌 AFEPSTIEIQAGDTVQWVNNKLAPHNVVVEGQ---P--------ELSHK--DLAFSPGETFEATFS---EPGTYTYYCEPHRGAGMVGKI
PLAS_ANASO          VFEPAKLTIKPGDTVEFLNNKVPPHNVVFDAAL-NPAKSADLAKSLSHK--QLLMSPGQSTSTTFPADAPAGEYTFYCEPHRGAGMVGKI
PLAS_SYNY3          VFEPSTVTIKAGEEVKWVNNKLSPHNIVFAADG-VDADT---AAKLSHK--GLAFAAGESFTSTFT---EPGTYTYYCEPHRGAGMVGKV
PLAS_PROHO          LYEPKALSISAGDTVEFVMNKVGPHNVIFDKVP-AGES----APALSNT--KLAIAPGSFYSVTLG---TPGTYSFYCTPHRGAGMVGTI
AZUP_ACHCY          VFEPASLKVAPGDTVTFIPTDKG-HNVETIKGM-IPDGAE----------AFKSKINENYKVTFT---APGVYGVKCTPHYGMGMVGVV
AZUP_PARDE          VFEPAFIRAEPGDVINFIPTDKS-HNVEAIKEI-LPEGVE----------TFKSKINEAYALTVT---EPGLYGVKCTPHFGMGMVGLV
AZUP_ALCFA          VFEPAYIKANPGDTVTFIPVDKG-HNVESIKDM-IPEGAE----------KFKSKINENYVLTVT---QPGAYLVKCTPHYAMGMIALI
AZUP_RHILV          VFEPGFLKIAPGDTVTFIPTDKS-HNVETFKGL-IPDGVP----------DFKSKPNEQYQVKFD---IPGAYVLKCTPHVGMGMVALI
AZUP_METEX  細菌    VFDPALVRLKPGDSIKFLPTDKG-HNVETIKGM-APDGAD----------YVKTTVGQEAVVKFD---KEGVYGFKCAPHYMMGMVALV
AMCY_PARDE          KYETPELHVKVGDTVTWINREAMPHNVHFVAGVLGEAALK----------GPMMKKEQAYSLTFT---EAGTYDYHCTPHP--FMRGKV
AMCY_PARVE          KYLTPEVTIKAGETVYWVNGEVMPHNVAFKKGIVGEDAFR----------GEMMTKDQAYAITFN---EAGSYDYFCTPHP--FMRGKV
AMCY_METEX          KFQTPEVRIKAGSAVTWTNTEALPHNVHFKSGPGVEKDVE----------GPMLRSNQTYSVKFN---APGTYDYICTPHP--FMKGKV
```

（1）完全に保存しているサイト番号は：*12G,*

（2）そのうち銅イオンの結合に関与するサイト番号は：

金属イオンの結合に関与するアミノ酸

（1）マイナスの電荷を持つアミノ酸

Glu(E), Asp(D)

（2）硫黄原子を含むアミノ酸

Cys(C), Met(M)

(3)ヒスチジン

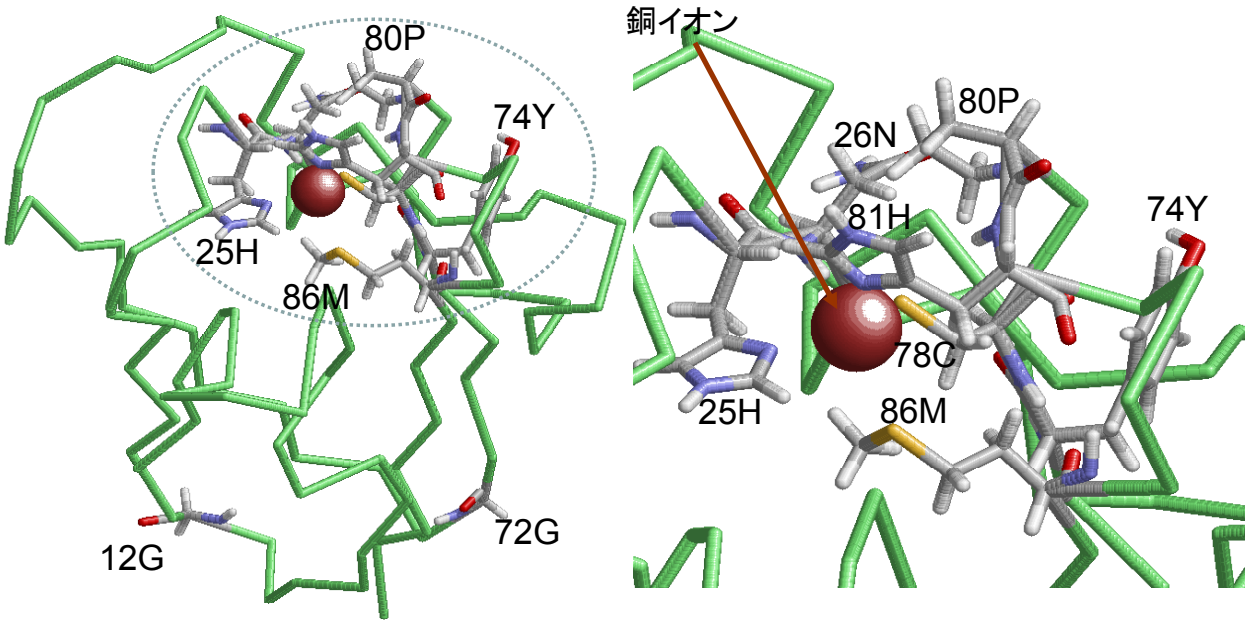
His(H)

```
                    _________1_________2_________3_________4_________5_________6_________7_________8_________9
                    123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890
PLAS_ORYSI  コメ    VFEPNDFTVKSGETITFKNNAGFPHNVVFDEDA-VPSGVD--VSKISQ--EEYLNAPGETFSVTLT---VPGTYGFYCEPHAGAGMVGKV
PLAS_HORVUオオムギ  VFEPNDFSVKAGETITFKNNAGYPHNVVFDEDA-VPSGVD--VSKISQ--EEYLTAPGETFSVTLT---VPGTYGFYCEPHAGAGMVGKV
PLAS_ENTPR          AFVPNNITVGAGESIEFINNAGFPHNIVFDEDA-VPAGVD--ADAISA--EDYLNSKGQTVVRKLT---TPGTYGVYCDPHSGAGMKMTI
PLAS_ULVPE          AFVPSKISVAAGEAIEFVNNAGFPHNIVFDEDA-VPAGVD--ADAISY--DDYLNSKGETVVRKLS---TPGVYGVYCEPHAGAGMKMTI
PLAS_CHLRE          EFVPKTLTIKSGETVNFVNNAGFPHNIVFDEDA-IPSGVN--ADAISR--DDYLNAPGETYSVKLT---AAGEYGYYCEPHQGAGMVGKI
PLAS_DAUCA ニンジン VFSPSSFSVAKGEGISFKNNAGFPHNIVFDEDE-VPAGVD--VSKISQ--EDYLDGAGESFTVTLT---EKGTYKFYCEPHAGAGMKGEV
PLAS_CAPBU          AFVPNDFSIAKGEKIVFKNNAGFPHNVVFDEDE-IPSGVD--ASKISMDENDLLNAAGETYEVALT---EAGTYSFYCAPHQGAGMVGKV
PLAS1_ARATH         AFVPSEFTVAKGEKIVFKNNAGFPHNVVFDEDE-IPSGVD--ASKISMDETALLNGAGETYEVTLT---EPGSYGFYCAPHQGAGMVGKL
PLAS_MERPE          AFVPNNFSVPSGEKITFKNNAGFPHNVVFDEDE-IPSGVD--ASKISMDEADLLNAPGETYAVTLT---EKGSYSFYCSPHQGAGMVGKV
PLAS_PHAVU          VFVPSEFSVPSGEKIVFKNNAGFPHNVVFDEDE-IPAGVD--AVKISMPEEELLNAPGETYVVTLD---TKGTYSFYCSPHQGAGMVGKV
PLAS1_POPNI ポプラ  AFVPSEFSISPGEKIVFKNNAGFPHNIVFDEDS-IPSGVD--ASKISMSEEDLLNAKGETFEVALS---NKGEYSFYCSPHQGAGMVGKV
PLAS_SILPR          AFVPSDLSIASGEKITFKNNAGFPHNVVFDEDE-VPAGVD--VTKISMPEEDLLNAPGEEYSVTLT---EKGTYKFYCAPHAGAGMVGKV
PLAS_SOLCR          AFVPGNFSISAGEKITFKNNAGFPHNVVFDEDE-IPAGVD--ASKISMPEEDLLNAPGETYSVTLS---EKGTYSFYCSPHQGAGMVGKV
PLAS_SAMNI          AFIPSNFSVPSGEKITFKNNAGFPHNVVFDEDE-VPSGVD--SAKISMSEDDLLNAPGETYSVTLT---ESGTYKFYCSPHQGAGMVGKV
PLAS_VICFA          AFVPNSFEVSAGDTIVFKNNAGFPHNVVFDEDE-IPSGVD--AAKISMPEEDLLNAPGETYSVKLD---AKGTYKFYCSPHQGAGMVGQV
PLAS_PEA    マメ    AFVPSSLEVSAGETIVFKNNAGFPHNVVFDEDE-IPAGVD--ASKISMPEEDLLNAPGETYSVKLD---AKGTYKFYCSPHQGAGMVGQV
PLAS_FRIAG          AFVPSNIEVAAGETVVFKNNAGFPHNVLFDEDE-VPKGVD--AGAISMKEEDLLNAPGETFSVTLK---EKGTYSIYCSPHQGAGMAGKI
PLAS_GINBI          AFIPNELQVNAGEQIVFKNNAGFPHNVIFDEDA-VPAGVD--VSSISMSEEDLLNAPGETYVVKLD---KKGTYRFFCAPHQGIGMSGIV
PLAS_PHYPA  コケ    GFYPKDISVAAGESVTFVNNKGFPHNVVFDEDA-VPAGVK--TEDINH--EDYLNGPNESFSITFK---TPGTYEFYCEPHQGAGMKGVV
PLAS_DRYCA  シダ    KFYPDSITVSAGEAVEFTLVGETGHNIVFDIPAGAPGTVASELKAASMDENDLLSEDEPSFKAKVS---TPGTYTFYCTPHKSANMKGTL
PLAS_PHOLA          QFEPANVTVHPGDTVKWVNNKLPPHNILFDDKQ-VPGASKELADKLSHS--QLMFSPGESYEITFSSDFPAGTYTYYCAPHRGAGMVGKI
PLAS_SYNP6シアノ細菌 AFEPSTIEIQAGDTVQWVNNKLAPHNVVVEGQ---P--------ELSHK--DLAFSPGETFEATFS---EPGTYTYYCEPHRGAGMVGKI
PLAS_ANASO          VFEPAKLTIKPGDTVEFLNNKVPPHNVVFDAAL-NPAKSADLAKSLSHK--QLLMSPGQSTSTTFPADAPAGEYTFYCEPHRGAGMVGKI
PLAS_SYNY3          VFEPSTVTIKAGEEVKWVNNKLSPHNIVFAADG-VDADT---AAKLSHK--GLAFAAGESFTSTFT---EPGTYTYYCEPHRGAGMVGKV
PLAS_PROHO          LYEPKALSISAGDTVEFVMNKVGPHNVIFDKVP-AGES----APALSNT--KLAIAPGSFYSVTLG---TPGTYSFYCTPHRGAGMVGTI
AZUP_ACHCY          VFEPASLKVAPGDTVTFIPTDKG-HNVETIKGM-IPDGAE----------AFKSKINENYKVTFT---APGVYGVKCTPHYGMGMVGVV
AZUP_PARDE          VFEPAFIRAEPGDVINFIPTDKS-HNVEAIKEI-LPEGVE----------TFKSKINEAYALTVT---EPGLYGVKCTPHFGMGMVGLV
AZUP_ALCFA          VFEPAYIKANPGDTVTFIPVDKG-HNVESIKDM-IPEGAE----------KFKSKINENYVLTVT---QPGAYLVKCTPHYAMGMIALI
AZUP_RHILV          VFEPGFLKIAPGDTVTFIPTDKS-HNVETFKGL-IPDGVP----------DFKSKPNEQYQVKFD---IPGAYVLKCTPHVGMGMVALI
AZUP_METEX  細菌    VFDPALVRLKPGDSIKFLPTDKG-HNVETIKGM-APDGAD----------YVKTTVGQEAVVKFD---KEGVYGFKCAPHYMMGMVALV
AMCY_PARDE          KYETPELHVKVGDTVTWINREAMPHNVHFVAGVLGEAALK----------GPMMKKEQAYSLTFT---EAGTYDYHCTPHP--FMRGKV
AMCY_PARVE          KYLTPEVTIKAGETVYWVNGEVMPHNVAFKKGIVGEDAFR----------GEMMTKDQAYAITFN---EAGSYDYFCTPHP--FMRGKV
AMCY_METEX          KFQTPEVRIKAGSAVTWTNTEALPHNVHFKSGPGVEKDVE----------GPMLRSNQTYSVKFN---APGTYDYICTPHP--FMKGKV
```

（1）完全に保存しているサイト番号は：*12G, 25H, 26N, 72G, 74Y, 78C, 80P, 81H, 86M*

（2）そのうち銅イオンの結合に関与するサイト番号は：*25H, 78C, 81H, 86M*

実際の金属イオン結合サイト



PLAS_ORYSIを1plsAを鋳型にモデリングした構造

より定量的な保存度の計算法

より細やかにサイトの保存性を抽出するには？

```
PLAS_ORYSI    PNDFTVKSGETITFKNNAGFPHNVVFDEDA
PLAS_MERPE    PNNFSVPSGEKITFKNNAGFPHNVVFDEDE
PLAS_DAUCA    PSSFSVAKGEGISFKNNAGFPHNIVFDEDE
PLAS_SAMNI    PSNFSVPSGEKITFKNNAGFPHNVVFDEDE
PLAS_VICFA    PNSFEVSAGDTIVFKNNAGFPHNVVFDEDE
PLAS_CUCPE    PNDFSVAAGEKIVFKNNAGFPHNVVFDEDE
PLAS1_ARATH   PSEFTVAKGEKIVFKNNAGFPHNVVFDEDE
PLAS_PEA      PSSLEVSAGETIVFKNNAGFPHNVVFDEDE
PLAS_FRIAG    PSNIEVAAGETVVFKNNAGFPHNVLFDEDE
PLAS_PHYPA    PKDISVAAGESVTFVNNKGFPHNVVFDEDA
PLAS_ULVPE    PSKISVAAGEAIEFVNNAGFPHNIVFDEDA
PLAS_ANASO    PAKLTIKPGDTVEFLNNKVPPHNVVFDAAL
PLAS_SYNP6    PSTIEIQAGDTVQWVNNKLAPHNVVVEGQP
PLAS_DRYCA    PDSITVSAGEAVEFTLVGETGHNIVFDIPA
AZUP_RHILV    PGFLKIAPGDTVTFIPTDK-SHNVETFKGL
AMCY_METEX    TPEVRIKAGSAVTWTNTEALPHNVHFK---
AMCY_PARDE    TPELHVKVGDTVTWINREAMPHNVHF----
AMCY_PARVE    --EVTIKAGETVYWVNGEVMPHNVAFKKGI
              PNDFSVKSGETVVWKNNAGFPHNVVFDEDE
              TSNLEIAA EKITFVLVKAPG  IHTEAAA
              -KSIT PK DA S IPTEVAS   E FGQL
               AKVK  P  G E T  GLT    A KIPP
               DT    V    Q    D -      -KGI
```

・配列の本数が多い場合、完全保存サイトは置きにくくなる。

・完全保存サイトではなくてもより相対的に保存が良いサイトはある。

⇒より定量的に保存性を評価する必要がある

⇒サイトごとのアミノ酸頻度を計算する必要性

サイトごとの保存度の計算法

（1）最も多いアミノ酸の頻度

$$p_{\max}(i) = \max_{a \in A}[p_i(a)]$$

※大きいほど保存が高い。値の範囲: $0.0 < p_{max}(i) < 1.0$

（2）エントロピー

$$Entropy(i) = -\sum_{a \in A} p_i(a) \cdot \log_2[p_i(a)]$$

※小さいほど保存が高い。値の範囲: $0.0 \leqq Entropy(i) \leqq \log_2|A|$

H21　生命情報学 2010.4.27

学籍番号_____　名前_____

（4）:　以下の5本の配列からなるマルチプルアライメントから、度数*Ci*(*a*)を計算し、それを本数5で割ることで（単純）頻度*fi*(*a*)=*Ci*(*a*)/*N*を計算せよ。さらに最大の頻度を保存度*pmax*として記入せよ。

サイト	配列	度数 *Ci*(*a*)								頻度 *fi*(*a*)								保存度 *pmax*
		A	D	E	G	H	K	L	S	A	D	E	G	H	K	L	S	
1	**AALLA**	**3**						**2**		**0.6**						**0.4**		**0.6**
2	**SSLLS**																	
3	**HHHHH**																	
4	**SSDSS**																	
5	**GALSE**																	
6	**ADEEG**																	
7	**DDDEK**																	
8	**KKKKH**																	

もっとも保存がよいサイトは_____番目、最も保存が悪いサイトは_____番目

H21　生命情報学 2010.4.27

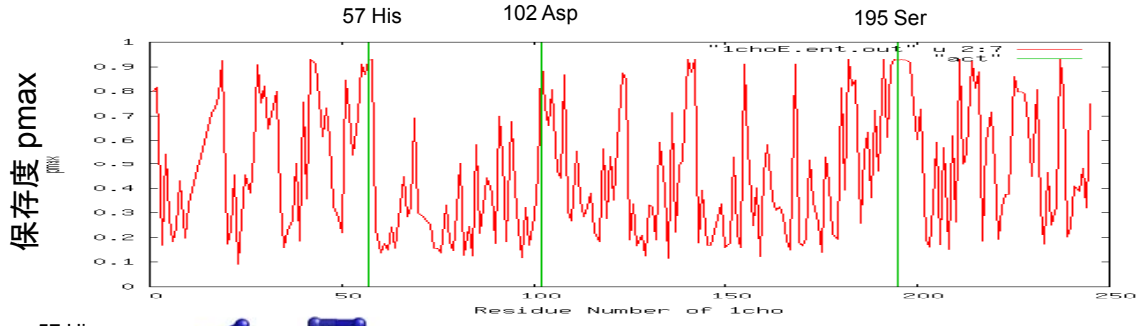
学籍番号＿＿＿＿＿＿　名前＿＿＿＿＿＿＿＿＿＿＿＿＿＿

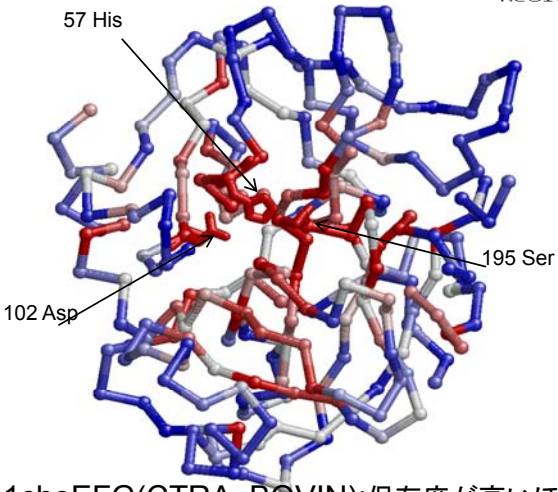
（4）：　以下の5本の配列からなるマルチプルアライメントから、度数$Ci(a)$を計算し、それを本数5で割ることで（単純）頻度$fi(a)=Ci(a)/N$を計算せよ。さらに最大の頻度を保存度$pmax$として記入せよ。

サイト	配列	度数 $Ci(a)$								頻度 $fi(a)$								保存度 $pmax$
		A	D	E	G	H	K	L	S	A	D	E	G	H	K	L	S	
1	**AALLA**	**3**						**2**		**0.6**						**0.4**		**0.6**
2	**SSLLS**							**2**	**3**							**0.4**	**0.6**	**0.6**
3	**HHHHH**					**5**								**1.0**				**1.0**
4	**SSDSS**		**1**						**4**		**0.2**						**0.8**	**0.8**
5	**GALSE**	**1**		**1**	**1**			**1**	**1**	**0.2**		**0.2**	**0.2**			**0.2**	**0.2**	**0.2**
6	**ADEEG**	**1**	**1**	**2**	**1**					**0.2**	**0.2**	**0.4**	**0.2**					**0.4**
7	**DDDEK**		**3**	**1**			**1**				**0.6**	**0.2**			**0.2**			**0.6**
8	**KKKKH**					**1**	**4**							**0.2**	**0.8**			**0.8**

もっとも保存がよいサイトは＿＿**3**＿＿番目、最も保存が悪いサイトは＿＿**5**＿＿番目

キモトリプシンの保存度と活性部位





キモトリプシン(Chymotrypsin) :他のタンパク質を加水分解する酵素

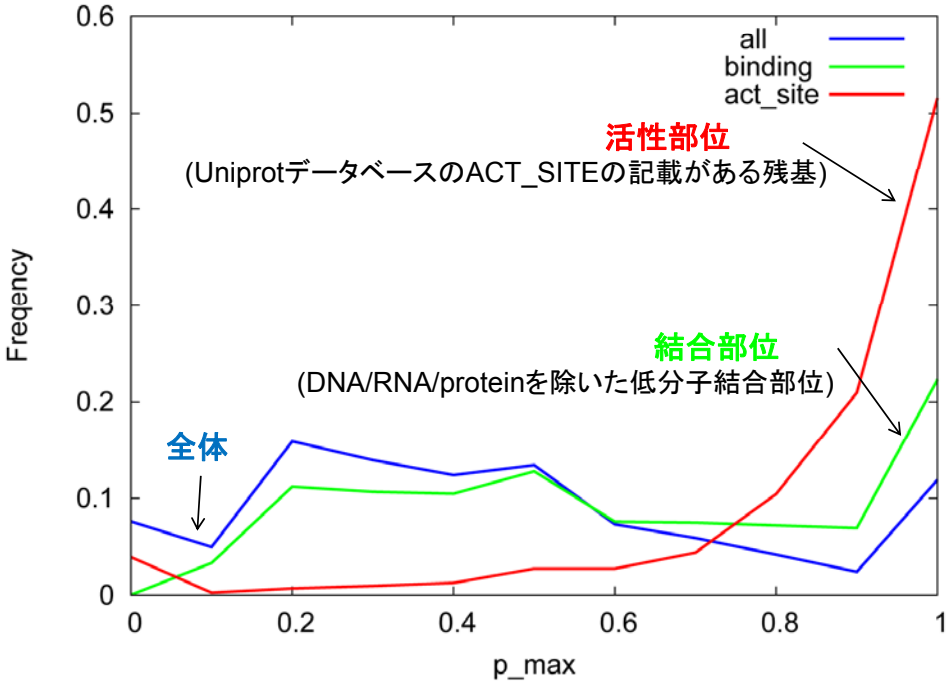
三つのアミノ酸(57His,102Asp,195Ser)が反応に必須である（活性部位）

分子内部、特に活性部位の保存度が高い。

1choEFG(CTRA_BOVIN):保存度が高いほど赤く色づけ

結合部位・活性部位の保存度

SCOP 1.73の40%の代表蛋白質 7315鎖　の統計解析
Uniprot 56.0のアミノ酸配列からホモログを収集



モチーフ解析

モチーフ･プロフィールを用いた類似性

より大きなグループ（スーパーファミリー）にまとめようとした場合、弱い相同性をより正確に認識できる類似性を採用する必要

→近縁の配列群のマルチプルアライメントから、このファミリーの本質的な特徴を見出したい

```
5p21-   MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSY
1ctqA   MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSY
1c1yA   MREYKLVVLGSGGVGKSALTVQFVQGIFVEKYDPTIEDSY
1kao-   MREYKVVVLGSGGVGKSALTVQFVTGTFIEKYDPTIEDFY
1huqA   --QFKLVLLGESAVGKSSLVLRFVKGQFHEYQESTIGAAF
1g16A   ----KILLIGDSGVGKSCLLVRFVE----DKFNPI--DFK
1ek0A   VTSIKLVLLGEAAVGKSSIVLRFVSNDFAENKEPTIGAAF
3rabA   ---FKILIIGNSSVGKTSFLFRYADDSFTPAFVSTVGIDF
1mh1-   ----KCVVVGDGAVGKTCLLISYTTNAFPGEYIPTVFDNY
2ngrA   MQTIKCVVVGDGAVGKTCLLISYTTNKFPSEYVPTVFDNY
1tx4B   ----KLVIVGDGACGKTCLLIVNSKDQF---YVPTVFENY
```

サイトごとに保存の度合いに差がある。
サイトごとにアミノ酸の出現傾向に差がある

[AG]-x(4)-G-K-[ST]

モチーフ解析

- 正規表現風のパターンで、局所的な配列のパターンを表現。

  PROSITE(http://www.expasy.ch/prosite/)が有名

1.進化的に保存している局所配列パターン

･マルチプルアライメント由来

･保存しているサイト→機能的に重要なサイト→活性部位

2.機能的な局所配列パターン

･リン酸化サイト、N-ミリストイル化サイトなど

PROSITEのモチーフの記述法

（例）　**ATP_GTP_A : [AG]-x(4)-G-K-[ST]**

2FE2S FERREDOXIN:
C-{C}-{C}-[GA]-{C}-C-[GAST]-{CPDEKRHFYW}-C

ZINC_FINGER_C2H2_1:
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

x	:任意のアミノ酸
x(n)	:n個の任意のアミノ酸
x(n,m)	:nからm個の任意のアミノ酸
[ACD]	:AかCかDのいずれかのアミノ酸
{ACD}	:AでもCでもDでもないアミノ酸

x :任意のアミノ酸
x(n) :n個の任意のアミノ酸
x(n,m):nからm個の任意のアミノ酸
[ACD] :AかCかDのいずれかのアミノ酸
{ACD} :AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1)　[AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2)　C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P C Q L C G R S P L G E A P P G T P P
C R L C C P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E G T Y H C T E C E D S F D N L G E L H G H F M L
H A R G E V

3)　[GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x :任意のアミノ酸
x(n) :n個の任意のアミノ酸
x(n,m) :nからm個の任意のアミノ酸
[ACD] :AかCかDのいずれかのアミノ酸
{ACD} :AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P **C** Q L C G R S P L G E A P P G T P P
C R L **C** **C** P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E E G T Y **H** **C** T E **C** E D S F D N L G E L **H** G **H** F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x :任意のアミノ酸
x(n) :n個の任意のアミノ酸
x(n,m) :nからm個の任意のアミノ酸
[ACD] :AかCかDのいずれかのアミノ酸
{ACD} :AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-
M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN
R G G P S R R A P R A A Q P P A Q P **C** Q L C G R S P L G E A P P G T P P
C R L **C** **C** P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E E G T Y **H** **C** T E **C** E D S F D N L G E L **H** G **H** F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI
V F E P N D F T V K S G E T I T F K N N A G F P H N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y C E P
H A G A G M V G K V T V N

x	:任意のアミノ酸
x(n)	:n個の任意のアミノ酸
x(n,m)	:nからm個の任意のアミノ酸
[ACD]	:AかCかDのいずれかのアミノ酸
{ACD}	:AでもCでもDでもないアミノ酸

(3)以下のPROSITEのモチーフに適合する箇所を□で囲め

1) [AG]-x(4)-G-K-[ST]

>5p21-

M T E Y K L V V V G A G G V G K S A L T I Q L I Q N H F V D E Y D P T I
E D S Y R K Q V V I D G E T C L L D I L D T A G Q E E Y S A M R D Q Y M
R T G E G F L C V F A I N N T K S F E D I H Q Y R E Q I K R V K D S D D
V P M V L V G N K C D L A A R T V E S R Q A Q D L A R S Y G I P Y I E T
S A K T R Q G V E D A F Y T L V R E I R Q H

2) C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

>ZN428_HUMAN

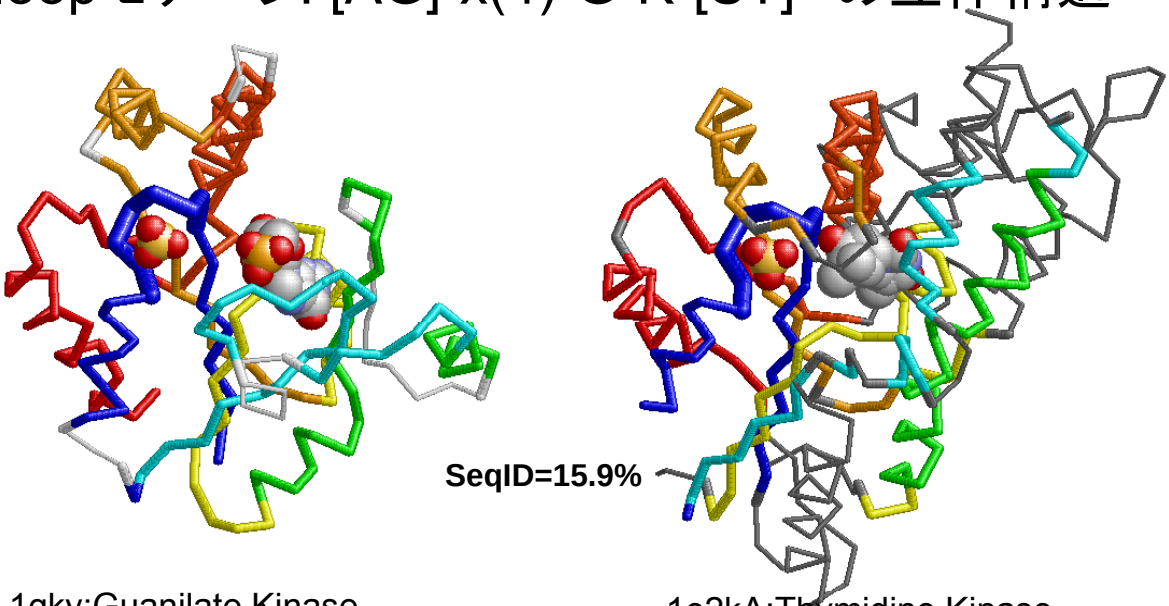
R G G P S R R A P R A A Q P P A Q P **C** Q L C G R S P L G E A P P G T P P
C R L **C** **C** P A T A P Q E A P A P E G R A L G E E E E E P P R A G E G R P
A G R E E E E E E E E E G T Y **H** **C** T E **C** E D S F D N L G E L **H** G **H** F M L
H A R G E V

3) [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ]

>PLAS_ORYSI

V F E P N D F T V K S G E T I T F K N N A G F P **H** N V V F D E D A V P S
G V D V S K I S Q E E Y L N A P G E T F S V T L T V P G T Y G F Y **C** E P
H A G A G M V G K V T V N

P-loopモチーフ: [AG]-x(4)-G-K-[ST] の立体構造



1gky:Guanilate Kinase
(8-15:GPSGTGKS)

1e2kA:Thymidine Kinase
(56-63:GPHGMGKT)

・P-loopモチーフは、ヌクレオチドのリン酸基結合サイトに対応
・モチーフ以外の領域も、立体構造は似ている

ProSiteモチーフの問題点

False positiveが多く、ファミリの認識能力は高くない。

```
          [AG]-x(4)-G-K-[ST]
5p21-   MTEYKLVVVGAGGVGKSAL
1ctqA   MTEYKLVVVGAGGVGKSAL
1c1yA   MREYKLVVLGSGGVGKSAL
1kao-   MREYKVVVLGSGGVGKSAL
1huqA   --QFKLVLLGESAVGKSSL
1g16A   ----KILLIGDSGVGKSCL
1ek0A   VTSIKLVLLGEAAVGKSSI
3rabA   ---FKILIIGNSSVGKTSF
1mh1-   ----KCVVVGDGAVGKTCL
2ngrA   MQTIKCVVVGDGAVGKTCL
1tx4B   ----KLVIVGDGACGKTCL
1i2mA   --QFKLVLVGDGGTGKTTF
2efgA   -RLRNIGIAAHIDAGKTTT
            . . .  ...
           .
```

1. パターンの表現能力の限界
2. 客観的にパターンを生成するのが難しい。
3. もっと大域的な領域も淡く似ているはず

プロフィール法

プロフィール法

マルチプルアライメントから**サイトごとのスコア行列**を作成。
これに対して動的計画法等を用いて配列をアライメント。

サイトごとのスコア行列

↓

プロフィール(Profile)
位置特異的スコア行列
（PSSM; Position Specific Score Matrix）

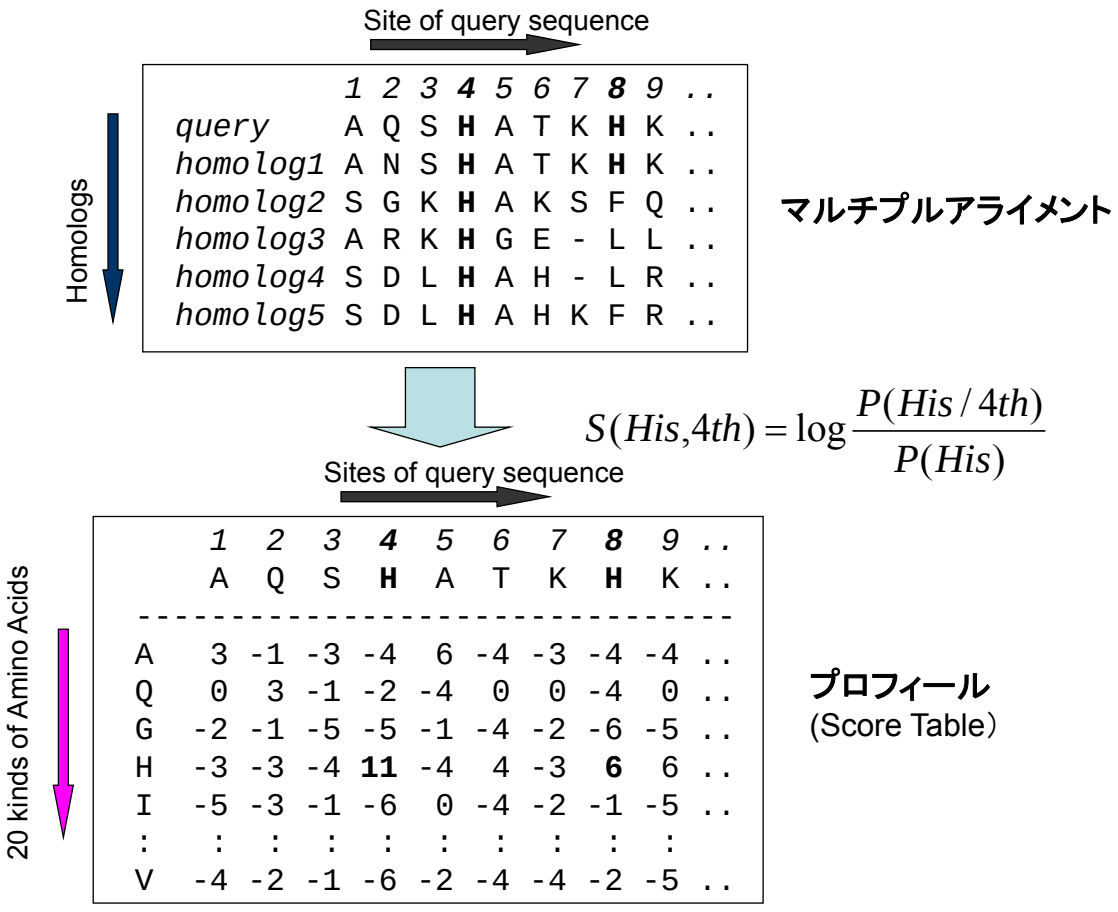
	1	*2*	*3*	***4***	*5*	*6*	..
A	3	-1	-3	-4	6	-4	..
Q	0	3	-1	-2	-4	0	..
H	-3	-3	-4	**11**	-4	4	..
:	:	:	:	:	:	:	
V	-4	-2	-1	-6	-2	-4	..

HMMer

マルチプルアライメントを入力とする。隠れマルコフモデル(HMM)を使用しているため、表現力はPSI-BLASTより高いはずだが、計算速度は遅い。PfamはHMMerを採用している。

PSI-BLAST

BLASTの拡張版。反復的にデータベース検索を行うことで、厚いマルチプルアライメントを生成する。



位置特異的スコア関数(PSSM)

$$S_i(a) = \log \frac{p_i(a)}{q(a)}$$

$p_i(a)$:i番目のサイトのアミノ酸aの確率

$q(a)$:アミノ酸aの背景確率(background probability)

※ $S_i(a)>0.0$ ($p_i(a)>q(a)$) のとき、このファミリーに属することを示唆

$Si(a)<0.0$ ($p_i(a)<q(a)$) のとき、このファミリーに属さないことを示唆

※ $p_i(a)=0$だと$S_i(a)=-\infty$ になってしまう。すべてのaについて$p_i(a)>0$となるような補正が必ず必要。

PSSMスコアの計算例

マルチプルアライメント

```
        12345678
seq1  AHHSGADK
seq2  ALHSADDK
seq3  LLHDLEEK
seq4  LLHSSEEK
seq5  AAHSEGEH
```

Laplaceの方法で推定された確率 $p_i(a)$

	A	D	E	G	H	K	L	S
1	.16	.04	.04	.04	.04	.04	.12	.04
2	.08	.04	.04	.04	.08	.04	.16	.04
3	.04	.04	.04	.04	.24	.04	.04	.04
4	.04	.12	.04	.04	.04	.04	.04	.20
5	.08	.04	.08	.08	.04	.04	.08	.08

PSSMスコア $\log[p_i(a)/q(a)]$ $q(a)=1/20$とした。

	A	D	E	G	H	K	L	S
1	1.7	-0.3	-0.3	-0.3	-0.3	-0.3	1.3	-0.3
2	0.7	-0.3	.0.3	-0.3	0.7	-0.3	1.7	-0.3
3	-0.3	-0.3	-0.3	-0.3	2.3	-0.3	-0.3	-0.3
4	-0.3	1.3	-0.3	-0.3	-0.3	-0.3	-0.3	2.0
5	0.7	-0.3	0.7	0.7	-0.3	-0.3	0.7	0.7

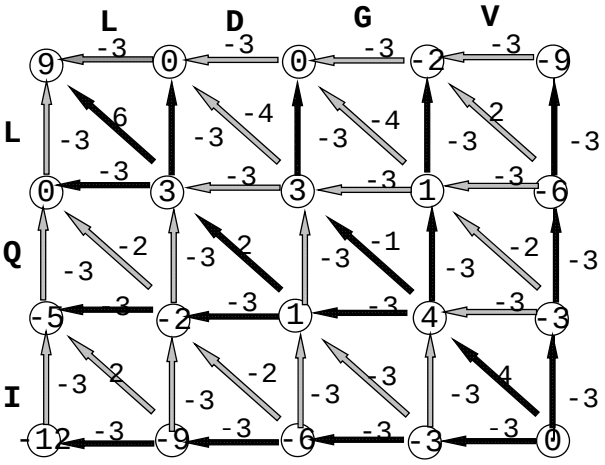
※Laplaceの方法:
アミノ酸頻度が0になってしまうのを避けるために、アミノ酸の観察数$Ci(a)$にすべて1を加えてから、頻度を計算する方法。

BLOSUM62（blastpのデフォルトで使われている置換スコア行列）

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

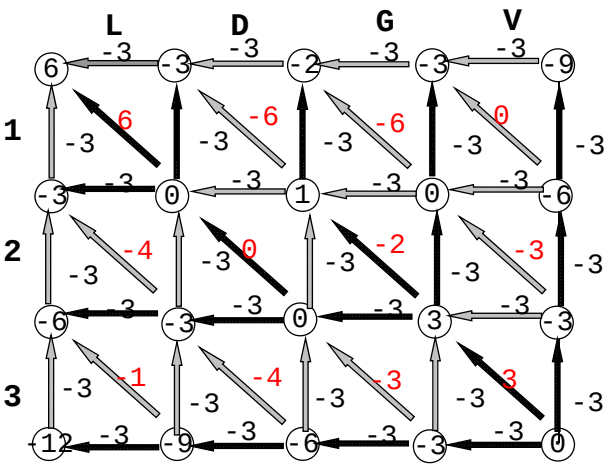
動的計画法によるアライメント

通常のペアワイズアライメント



LDGV
LQ-I

PSSMを用いたアライメント



LDGV
12-3

PSI-BLASTにより計算されたアミノ酸頻度

Myoglobin (1a6m/MYG_PHYCA、クジラ)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	V	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	75
2	L	0	0	0	0	0	0	0	0	0	0	95	0	1	4	0	0	0	0	0	0
3	S	1	0	2	0	0	0	0	1	0	0	0	0	1	0	0	61	34	0	0	0
4	E	35	0	0	32	0	1	8	4	1	0	0	2	0	0	8	8	1	0	0	0
5	G	30	0	2	6	0	5	19	14	1	0	0	12	0	0	0	4	5	0	0	1
6	E	0	0	2	43	0	0	54	0	0	0	0	0	0	0	0	0	0	0	0	0
7	W	0	14	0	0	1	0	1	0	0	0	0	61	2	4	0	0	1	15	0	1
8	Q	24	0	4	6	1	16	7	2	1	0	0	12	0	0	0	11	16	0	0	1
9	L	21	2	15	0	0	4	1	0	6	4	26	6	3	1	0	0	6	0	0	5
10	V	0	0	0	0	0	0	0	0	0	42	1	0	0	0	0	0	0	0	0	57
11	L	3	8	6	0	1	7	0	1	0	2	20	28	0	0	0	3	18	0	0	4
12	H	22	0	8	6	0	7	0	11	5	0	1	9	0	0	0	19	10	0	0	0
:																					
24	H	2	0	3	1	4	0	0	1	16	18	7	0	2	7	4	2	4	0	19	10
:																					
36	H	0	0	1	0	1	0	0	0	20	0	0	0	0	24	0	1	3	0	50	0
:																					
64	H	0	0	0	0	0	5	0	1	92	0	0	0	0	0	0	0	0	0	0	2
:																					
93	H	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0

PSI-BLASTにより計算されたスコア

Myoglobin (1a6m/MYG_PHYCA、クジラ)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	V	-2	-4	-4	-5	-2	-3	-4	-5	-4	1	0	-3	5	-2	-4	-3	-2	-4	-3	6
2	L	-4	-4	-6	-6	-3	-4	-5	-6	-5	0	6	-5	1	1	-5	-5	-3	-4	-3	-1
3	S	-1	-3	-1	-3	-3	-2	-3	-2	-3	-4	-4	-3	-2	-5	-3	5	5	-5	-4	-3
4	E	4	-3	-2	5	-4	-2	1	-1	-1	-5	-5	-1	-3	-5	1	1	-2	-5	-4	-4
5	G	3	-3	-1	1	-4	1	3	2	-1	-5	-5	2	-4	-5	-3	0	0	-5	-4	-3
6	E	-4	-3	0	6	-6	-1	6	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
7	W	-3	3	-3	-4	-2	-1	-2	-4	-3	-5	-4	6	-1	0	-4	-3	-3	7	-3	-3
8	Q	3	-2	0	0	-1	3	1	-2	-2	-4	-4	2	-3	-5	-3	1	2	-5	-4	-3
9	L	2	-2	3	-4	-2	0	-2	-4	2	0	2	0	1	-2	-4	-2	0	-5	-3	0
10	V	-3	-5	-6	-6	-3	-5	-5	-6	-6	5	-1	-5	-1	-3	-5	-4	-3	-5	-4	6
11	L	-1	1	1	-3	-2	1	-3	-3	-3	-2	2	4	-2	-4	-4	-1	3	-5	-4	-1
12	H	3	-2	2	0	-3	1	-2	1	2	-4	-4	1	-4	-5	-4	3	1	-5	-4	-3
:																					
24	H	-2	-4	-1	-4	2	-3	-4	-4	5	3	0	-4	0	2	-1	-2	-1	-2	5	1
:																					
36	H	-4	-4	-2	-5	-3	-3	-4	-5	6	-4	-3	-4	-3	5	-5	-3	-1	-1	7	-4
:																					
64	H	-4	-2	-2	-3	-5	1	-2	-3	10	-5	-5	-3	-4	-4	-5	-3	-4	-5	-1	-3
:																					
93	H	-4	-2	-2	-3	-5	-2	-2	-4	11	-6	-5	-3	-4	-4	-5	-3	-4	-5	0	-6

BLASTにより発見されたホモログ

Myoglobin (1a6m/MYG_PHYCA、クジラ)をクエリとしてPDBを検索

```
BLASTP 2.2.16 [Mar-25-2007]
Query= 1a6mAA  (151 letters)
Database: 40pdb09Jan8
```

Sequences producing significant alignments:	Score (bits)	E Value
*2nrlA [a.1.1 (101mA)] MYOGLOBIN	114	4e-27
*2dc3A [a.1.1 (1umoA)] CYTOGLOBIN	85	4e-18
*1irdA [a.1.1] HEMOGLOBIN ALPHA CHAIN	46	2e-06
*1c7cA [a.1.1 - a.1.1] PROTEIN (DEOXYHEMOGLOBIN (ALPHA CHAIN))	46	2e-06
*1it2A [a.1.1] HEMOGLOBIN	44	6e-06
*1mbaA [a.1.1] MYOGLOBIN	40	1e-04
*1x3kA [x.x.x] HEMOGLOBIN COMPONENT V	37	0.001
1hlbA [a.1.1] HEMOGLOBIN (DEOXY)	35	0.003
2c0kA [x.x.x] HEMOGLOBIN	35	0.004
2z8aA [a.1.1 (1hbiA)] GLOBIN-1	34	0.006
2olpA [x.x.x] HEMOGLOBIN II	32	0.024
1x46A [x.x.x] HEMOGLOBIN COMPONENT VII	32	0.031
2bk9A [x.x.x] CG9734-PA	27	0.99
1un7A [b.92.1 - c.1.9] N-ACETYLGLUCOSAMINE-6-PHOSPHATE DEACETYLASE	27	1.3
1zx5A [b.82.1] MANNOSEPHOSPHATE ISOMERASE, PUTATIVE	26	2.2
1nh1A [e.45.1] AVIRULENCE B PROTEIN	26	2.2
1q1fA [a.1.1] NEUROGLOBIN	25	2.9
2dy1A [c.37.1 - b.43.3 - d.58.11 - d.14.1 - d.58.11 (1wdtA)] ELO...	25	2.9
1b0bA [a.1.1] HEMOGLOBIN	25	3.8
1vbiA [x.x.x] TYPE 2 MALATE/LACTATE DEHYDROGENASE	24	6.4
2rd9A [x.x.x] BH0186 PROTEIN	24	6.4

PSI-BLASTにより発見されたホモログ

Myoglobin (1a6m/MYG_PHYCA、クジラ)をクエリとしてPDBを検索

```
BLASTP 2.2.16 [Mar-25-2007]
Query= 1a6mAA         (151 letters)
Database: 40pdb09Jan8
```

Sequences producing significant alignments:	Score (bits)	E Value
1c7cA [a.1.1 - a.1.1] PROTEIN (DEOXYHEMOGLOBIN (ALPHA CHAIN))	222	9e-60
1irdA [a.1.1] HEMOGLOBIN ALPHA CHAIN	222	1e-59
2dc3A [a.1.1 (1umoA)] CYTOGLOBIN	169	1e-43
2nrlA [a.1.1 (101mA)] MYOGLOBIN	156	8e-40
1it2A [a.1.1] HEMOGLOBIN	111	5e-26
***1cg5B [a.1.1] PROTEIN (HEMOGLOBIN)**	**103**	**8e-24**
***1hlbA [a.1.1] HEMOGLOBIN (DEOXY)**	**66**	**2e-12**
***2c0kA [x.x.x] HEMOGLOBIN**	**57**	**7e-10**
***1q1fA [a.1.1] NEUROGLOBIN**	**53**	**2e-08**
1x3kA [x.x.x] HEMOGLOBIN COMPONENT V	51	5e-08
***2z8aA [a.1.1 (1hbiA)] GLOBIN-1**	**51**	**5e-08**
1mbaA [a.1.1] MYOGLOBIN	50	1e-07
***2olpA [x.x.x] HEMOGLOBIN II**	**49**	**2e-07**
***2bk9A [x.x.x] CG9734-PA**	**49**	**3e-07**
***1jf3A [a.1.1] MONOMER HEMOGLOBIN COMPONENT III**	**48**	**4e-07**
***1x46A [x.x.x] HEMOGLOBIN COMPONENT VII**	**45**	**3e-06**
***1gdjA [a.1.1] LEGHEMOGLOBIN (DEOXY)**	**41**	**6e-05**
***2zs0C [a.1.1 (1x9fA)] EXTRACELLULAR GIANT HEMOGLOBIN MAJOR GLOBIN**	**40**	**1e-04**
***1b0bA [a.1.1] HEMOGLOBIN**	**39**	**2e-04**
***1cqxA [a.1.1 - b.43.4 - c.25.1] FLAVOHEMOPROTEIN**	**38**	**6e-04**
1ecaA [a.1.1] ERYTHROCRUORIN (AQUO MET)	35	0.004

BLASTにより発見されたホモログ

```
>1x3kA [x.x.x] HEMOGLOBIN COMPONENT V ユスリカのヘモグロビン Length = 152
 Score = 37.0 bits (84), Expect = 0.001
 Identities = 24/102 (23%), Positives = 42/102 (41%), Gaps = 1/102 (0%)

Query: 2   LSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDL 61
           LS+ E +LV   WA +  D+ G    +     K +P   +KF+  K    + E+K + +
Sbjct: 5   LSDSEEKLVRDAWAPIHGDLQGTANTVFYNYLKKYPSNQDKFETLKGHPLD-EVKDTANF 63

Query: 62  KKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKY 103
           K     + T     +K  G+ +   K +A          PI +
Sbjct: 64  KLIAGRIFTIFDNCVKNVGNDKGFQKVIADMSGPHVARPITH 105
```

PSI-BLASTにより発見されたホモログ

```
>1cqxA [a.1.1 - b.43.4 - c.25.1] FLAVOHEMOPROTEIN 微生物のフラボヘム蛋白質 Length = 403
 Score = 37.6 bits (87), Expect = 6e-04,   Method: Composition-based stats.
 Identities = 26/148 (17%), Positives = 51/148 (34%), Gaps = 21/148 (14%)

Query: 1   VLSEGEWQLVLHVWAKVEADVAGHGQDIL----IRLFKSHPETLEKF--DRFKHLKTEAE 54
           +L++    +V    A V   +A HG DI+     R+F++HPE    F     +  + +
Sbjct: 1   MLTQKTKDIVKAT-APV---LAEHGYDIIKCFYQRMFEAHPELKNVFNMAHQEQGQQQQA 56

Query: 55  MKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHV 114
           +             + A    ++      A LK +A  HA    +  +    + E ++
Sbjct: 57  L----------ARAVYAYAENIEDPNSLMAVLKNIANKHA-SLGVKPEQYPIVGEHLLAA 105

Query: 115 LHSRHPGDFGADAQGAMNKALELFRKDI 142
           +          D   A  +A       +
Sbjct: 106 IKEVLGNAATDDIISAWAQAYGNLADVL 133
```

マルチプルアライメント

	1	2	3	**4**	5	6	7	**8**	9	..
query	A	Q	S	**H**	A	T	K	**H**	K	..
homolog1	A	N	S	**H**	A	T	K	**H**	K	..
homolog2	S	G	K	**H**	A	K	S	F	Q	..
homolog3	A	R	K	**H**	G	E	-	L	L	..
homolog4	S	D	L	**H**	A	H	-	L	R	..

良質のマルチプルアライメントを作るには淡い相同性の配列を集め、アラインする必要がある。それには、よいプロフィール
が不可欠

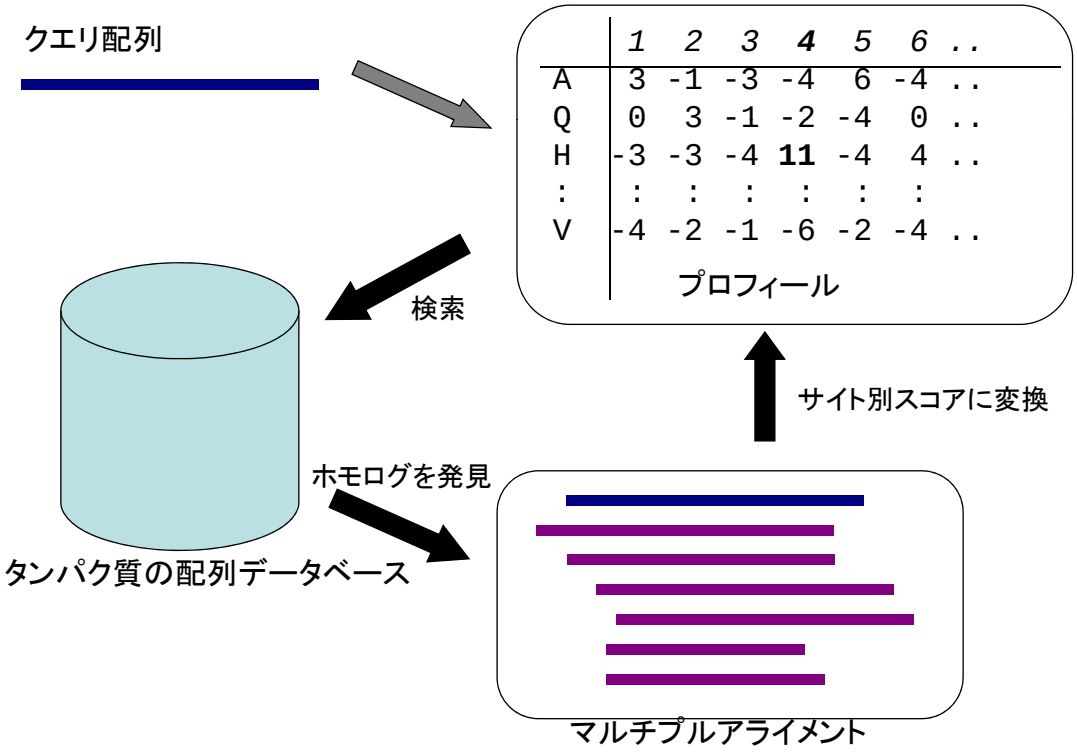
プロフィール

	1	2	3	**4**	5	6	7	**8**	..
	A	Q	S	**H**	A	T	K	**H**	..
A	3	-1	-3	-4	6	-4	-3	-4	..
G	-2	-1	-5	-5	-1	-4	-2	-6	..
H	-3	-3	-4	**11**	-4	4	-3	**6**	..
:	:	:	:	:	:	:	:	:	
V	-4	-2	-1	-6	-2	-4	-4	-2	..

良質のプロフィールを作るには
できるだけ多くの配列を集めた
マルチプルアライメントが必要

堂々巡りの関係

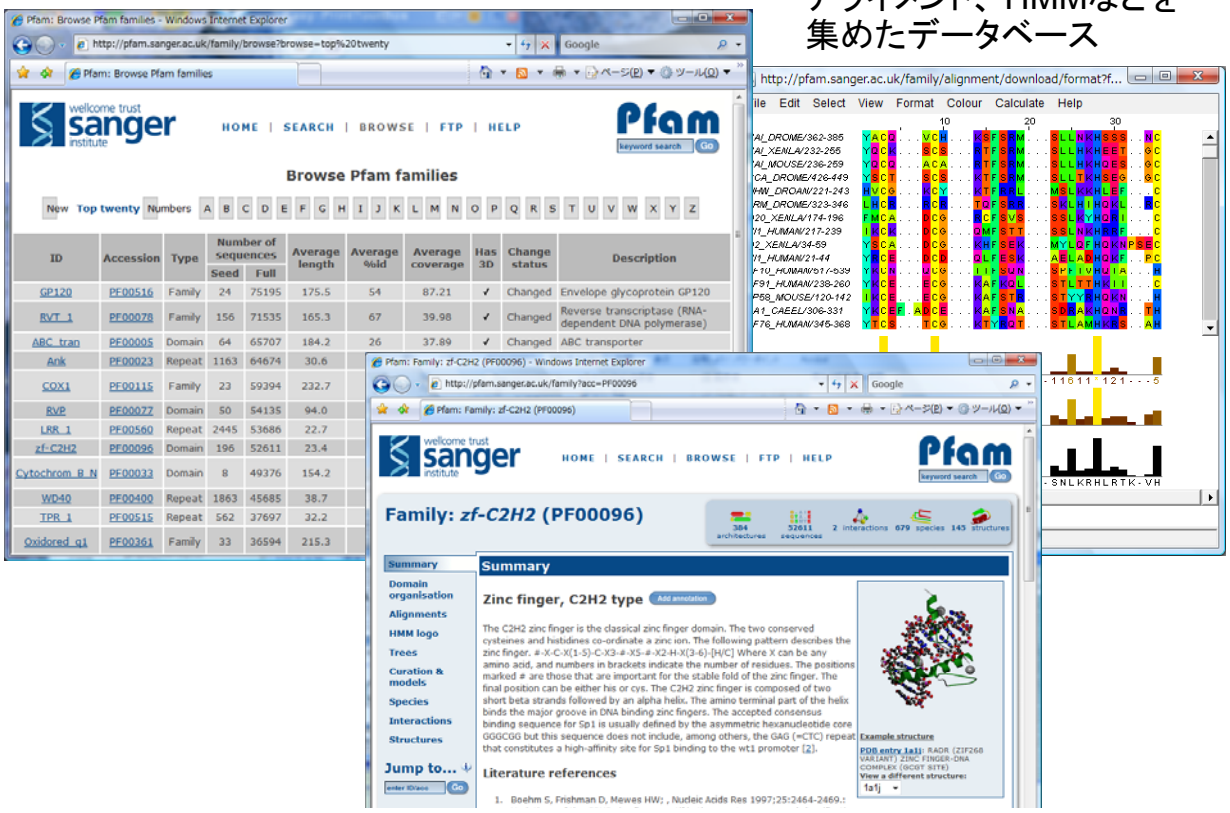
PSI-BLASTの手続き



Pfam：蛋白質ファミリのデータベース

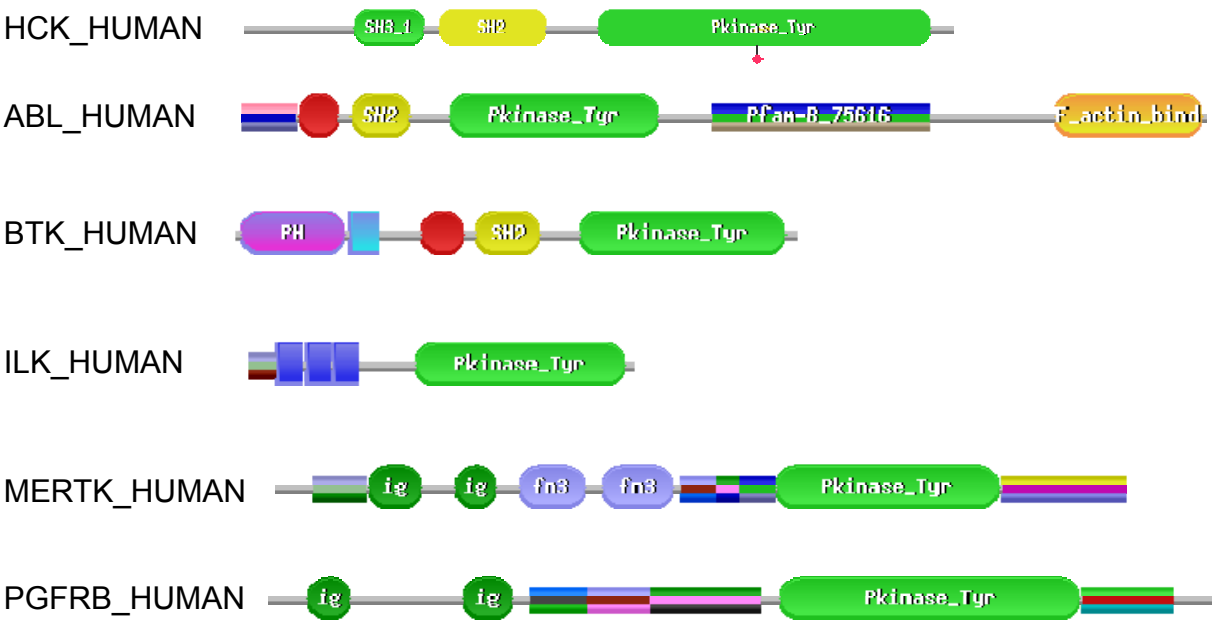
http://pfam.sanger.ac.uk

各蛋白質ファミリのマルチプルアライメント、HMMなどを集めたデータベース



Pkinase_Tyrドメインをもつタンパク質の例

Family : Pkinase_Tyr (PF07714) : Protein tyrosine kinase



Pfamデータベース（http://pfam.sanger.ac.uk/Software/Pfam/）からの引用

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
- Arthur M.Lesk(岡崎康司、坊農秀雄　監訳)「バイオインフォマティクス基礎講義 一歩進んだ発想をみがくために」(2003), メディカル・サイエンス・インターナショナル
- 長谷川政美、岸野洋久 「分子系統学」 岩波書店（1996）
- 根井正利、S.クマー「分子進化と分子系統学」（2006）培風館
- 斎藤成也 「ゲノム進化学入門」（2007） 共立出版
- Durbin R.,Eddy.S.,Krogh A.,Mitchson,G. “Biological Sequence analysis”,Cambridge University Press, 1998.Chapter 7,8.
- 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