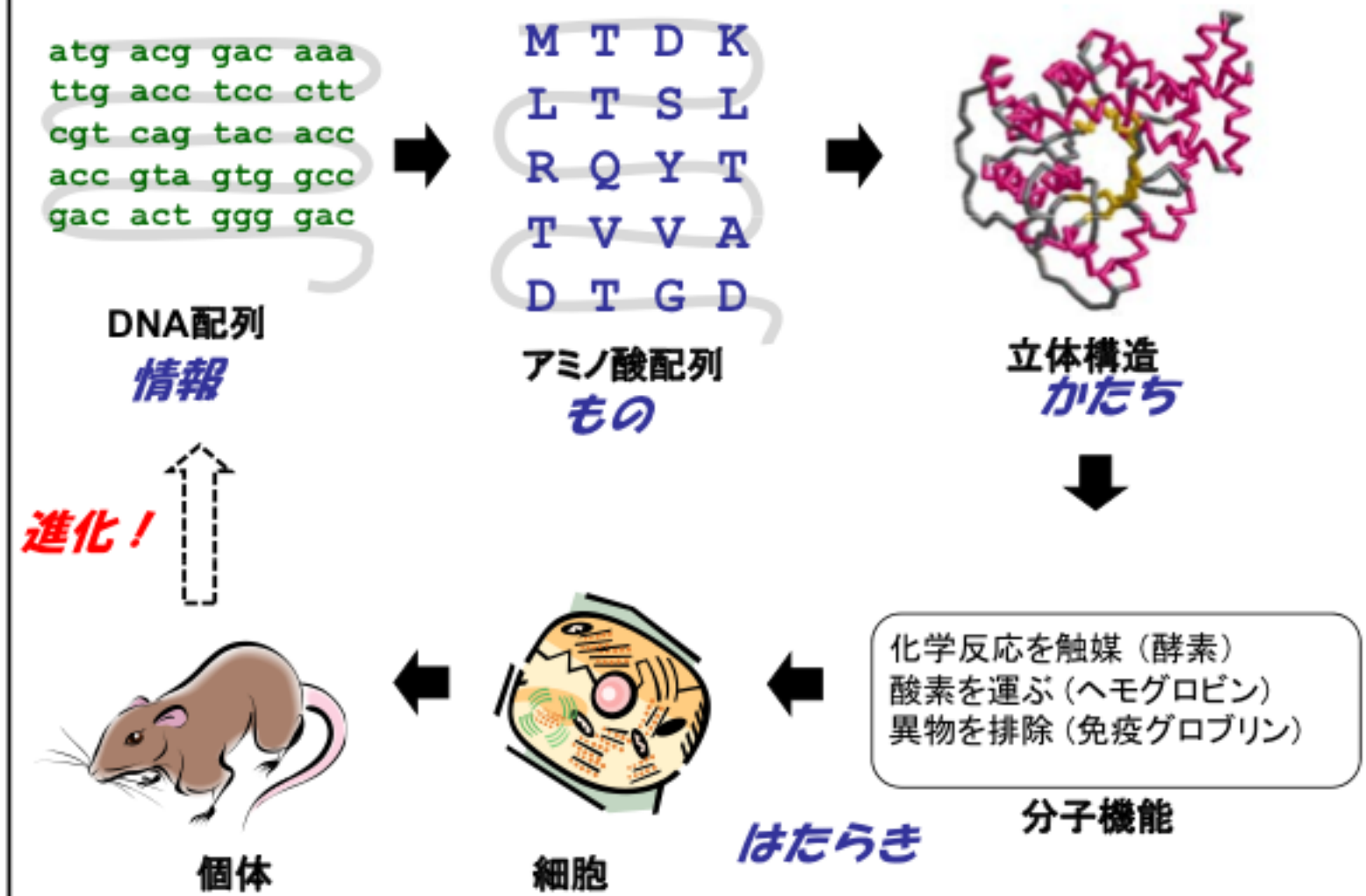


ペアワイズアライメント

分子生物学のセントラルドグマ



高分子は文字列だとみなせる

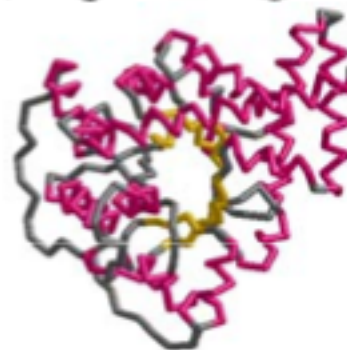
atg acg gac aaa
ttg acc tcc ctt
cgt cag tac acc
acc gta gtg gcc
gac act ggg gac

DNA配列
情報



M T D K
L T S L
R Q Y T
T V V A
D T G D

アミノ酸配列
もの



立体構造
かたち

DNAもタンパク質もユニットが一行に並んだ高分子

ユニット: DNAは4種の核酸(atgc)、タンパク質は20種のアミノ酸(ACDEFGH...)

atgacggacaaattgacctcccttcgtcagtacaccaccgtagtgccga

M T D K L T S L R Q Y T T V V A D T G D

→単なる文字列だとみなして処理をしてもある種の本質は失われない

「進化」とはDNAという文字列が変化すること

atgacggacaaattgacctcccttcgtcagtacacc

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



atgacg**a**caaaattgacctcccttcgtcagtacacc

M T **N** K L T S L R Q Y T

より正確には、個体のDNAが変化したあとに、その変異がその種の集団において定着する「集団遺伝学」的な過程が必要

- ①個体のDNAに変異が生じる
- ②その変異が子孫に継承され、
- ③中立か正の淘汰が働けば、同じ変異を持った子孫が種の集団内で多数を占める

違う生物の同じ機能のタンパク質のアミノ酸配列

トリオースリン酸異性化酵素 (**Triosephosphate isomerase** (EC 5.3.1.1) (TIM, TPIS))

>TPIS_HUMAN ヒト "Triosephosphate isomerase (EC 5.3

APSRKFFVGGNWKMNGRKRQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAI GTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDI INAKQ

>TPIS_RABIT ウサギ "Triosephosphate isomerase (EC 5

APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALSEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAI GTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDI INAKQ

違う生物の同じ機能のタンパク質のアミノ酸配列

トリオースリン酸異性化酵素 (**Triosephosphate isomerase** (EC 5.3.1.1) (TIM, TPIS))

>TPIS_HUMAN ヒト "Triosephosphate isomerase (EC 5.3

APSRKFFVGGNWKMNQRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDIINAKQ

>TPIS_YEAST 酵母 "Triosephosphate isomerase (EC 5.

ARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVICPPATY
LDYSVSLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWV
ILGHSERRSYFHEDDKFIADKTKFALGQGVGVIILCIGETLEEKKA
GKTLDVVERQLNAVLEEVKDWTNVVVAYEPVWAIGTGLAATPEDA
QDIHASIRKFLASKLGDKAASELRILYGGSSANGSNAVTFKDKADV
DGFLVGGASLKPEFVDIINSRN

違う生物の同じ機能のタンパク質のアミノ酸配列

トリオースリン酸異性化酵素 (**Triosephosphate isomerase** (EC 5.3.1.1) (TIM, TPIS))

>TPIS_HUMAN ヒト "Triosephosphate isomerase (EC 5.3

APSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPT
AYIDFARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATW
VVLGHSERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDERE
AGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAI GTGKTATPQQ
AQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPD
VDGFLVGGASLKPEFVDI INAKQ

>TPIS_ECOLI 大腸菌 "Triosephosphate isomerase (EC 5

MRHPLVMGNWKLNGSRHMHVHELVSNLRLKELAGVAGCAVAIAPPEM
YIDMAKREAEGSHIMLGAQNVDLNLSGAFTGETSAAMLKDIGAQY
IIIGHSERRTYHKESEDELI AKKFAVLKEQGLTPVLCIGETEAENE
AGKTEEV CARQIDAVLKTQGAAAFEGAVIAYEPVWAI GTGKSATP
AQAQAVHKFIRDHIAKVDANIAEQV I IQYGGSVNASNAELFAQP
DIDGALVGGASLKADAFVIVKAAEAAKQA

ペアワイズアラインメントを行うホームページ、たとえば

<http://www.ebi.ac.uk/Tools/emboss/align/index.html>

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- EMBOSS-Align Help

Similar Applications

- ClustalW2
- Kalign
- MAFFT
- MUSCLE
- T-Coffee

Emboss Programmatic Access

EBI > Tools > Sequence Analysis > EMBOSS

EMBOSS Pairwise Alignment Algorithms

This tool is used to compare 2 sequences. When you want an alignment that covers the whole length of both sequences, use [needle](#). When you are trying to find the best region of similarity between two sequences, use [water](#).

Method: EMBOSS::needle (global) | Gap Open: 10.0 | Gap Extend: 0.5 | Molecule: Protein | Matrix: Blosum62

Sequence 1: paste Sequence in any format OR upload a file: [Help](#)

```
>sp|P60174|TPIS_HUMAN Triosephosphate isomerase OS=Homo sapiens GN=TP11 PE=1 SV=2
MAPSRKFFVGGNWKMNCRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQLDPKI
AVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGEDELIGQKVAHALAE
GLCVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWKVVLAYEPVWAIKGTATPQ
QAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPE
FVDIINAKQ
```

Seq. 1 Upload a file: [ファイルを選択](#) ファイルが選...ていません

Sequence 2: paste Sequence in any format OR upload a file: [Help](#)

```
>sp|P00939|TPIS_RABIT Triosephosphate isomerase OS=Orctolagus cuniculus GN=TP11 PE=1 SV=1
APSRKFFVGGNWKMNCRKKNLGLITLNAKVPADTEVVCAPPTAYIDFARQLDPKIA
VAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGEDELIGQKVAHALSE
LCVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWKVVLAYEPVWAIKGTATPQQ
QAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPE
VDIINAKQ
```

Seq. 2 Upload a file: [ファイルを選択](#) ファイルが選...ていません **Run** **Reset**

If you plan to use these services during a course please [contact us](#).

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2009. EBI is an Outstation of the European Molecular Biology Laboratory.

UNIPROTのホームページで
TPIS_HUMAN, TPIS_RABITの
配列をFASTA形式で得て
それぞれ貼り付ける

Runを押すと結果が表示される

ヒトとウサギ

```
#####  
Program: needle# Runday: Tue Nov 10 02:40:01 2009  
#####
```

```
=====  
# 1: TPIS_HUMAN# 2: TPIS_RABIT  
# Matrix: EBLOSUM62  
# Gap_penalty: 10.0  
# Extend_penalty: 0.5  
# Length: 249  
# Identity: 244/249 (98.0%)  
# Score: 1270.0  
=====
```

Sequence Identity : 配列相同性

配列の類似性を評価する指標の一つ
一致した箇所の数(244)を
配列の全長(249)で割り
百分率で表す(98%)

```
TPIS_HUMAN      1  MAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYID      50  
TPIS_RABIT      1  APSRKFFVGGNWKMNGRKKNLDELITTLNAAKVPADTEVVCAPPTAYID      49  
  
TPIS_HUMAN      51  FARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERR      100  
TPIS_RABIT      50  FARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERR      99  
  
TPIS_HUMAN     101  HVFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKV      150  
TPIS_RABIT     100  HVFGESDELIGQKVAHALSEGLGVIACIGEKLDEREAGITEKVVFEQTKV      149  
  
TPIS_HUMAN     151  IADNVKDWSKVVLAYEPVWAI GTGKTATPQQAQEVHEKLRGWLKSNVSDA      200  
TPIS_RABIT     150  IADNVKDWSKVVLAYEPVWAI GTGKTATPQQAQEVHEKLRGWLKSNVSDA      199  
  
TPIS_HUMAN     201  VAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDI INAKQ      249  
TPIS_RABIT     200  VAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDI INAKQ      248
```

アラインメントをせずに頭から単純に並べると、ほとんど一致しない

TPIS_HUMAN	1	MAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYID	50
TPIS_RABIT	1	APSRKFFVGGNWKMNGRKKNLGELITTLNAAKVPADTEVVCAPPTAYIDF	50
TPIS_HUMAN	51	FARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERR	100
TPIS_RABIT	50	ARQKLDPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRH	100
TPIS_HUMAN	101	HVFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKV	150
TPIS_RABIT	100	VFGESDELIGQKVAHALSEGLGVIACIGEKLDEREAGITEKVVFEQTKVI	150
TPIS_HUMAN	151	IADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSDA	200
TPIS_RABIT	150	ADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSDAV	200
TPIS_HUMAN	201	VAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ	249
TPIS_RABIT	200	AQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ	249

ヒトとイースト

Length: 250
Identity: 131/250 (52.4%)
Similarity: 168/250 (67.2%)
Gaps: 3/250 (1.2%)
Score: 687.0

TPIS_HUMAN	1	MAPSRKFFVGGNWKMN	GRKQSLGELIGTLNAAKVPADTEVVCAPPTAYID	50
TPIS_YEAST	1	MARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLD		48
TPIS_HUMAN	51	FARQKL-DPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSER		99
TPIS_YEAST	49	YSVSLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSER		98
TPIS_HUMAN	100	RHVFGESEDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTK		149
TPIS_YEAST	99	RSYFHEDDKFIADKTKFALGQGVVILCIGETLEEKKAGKTLDVVERQLN		148
TPIS_HUMAN	150	VIADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSNVSD		199
TPIS_YEAST	149	AVLEEVKDWTNVVVAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGD		198
TPIS_HUMAN	200	AVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPEFVDIINAKQ		249
TPIS_YEAST	199	KAASELRILYGGSSANGSNAVTFKDKADVVDGFLVGGASLKPEFVDIINSRN		248

だいたい半分ぐらいが一致

ヒトと大腸菌

Length: 260
Identity: 114/260 (43.8%)
Similarity: 149/260 (57.3%)
Gaps: 16/260 (6.2%)
Score: 487.0
##=====

TPIS_HUMAN	1	MAPSRKFFVGGNWKMNGRKQSLGELIGTL--NAAKVPADTEVVCAPPTAY	
TPIS_ECOLI	1	MRHPLVMGNWKLNGSRHMHVHELVSNLRKELAGV-AGCAVAIAPPEMY	46
TPIS_HUMAN	49	IDFARQKLD-PKIAVAQNICYKVTNGAFTGEISPGMIKDCGATWVVLGHS	97
TPIS_ECOLI	47	IDMAKREAEGSHIMLGAQNVDLNLGSAFTGETSAAMLKDIGAQYIIIGHS	96
TPIS_HUMAN	98	ERRHVFGESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQ	147
TPIS_ECOLI	97	ERRTYHKESDELIAKKFAVLKEQGLTPVLCIGETEAEENEAGKTEEVCARQ	146
TPIS_HUMAN	148	TKVI--ADNVKDWSKVVLAYEPVWAIIGTGKTATPQQAQEVHEKLRGWLKS	195
TPIS_ECOLI	147	IDAVLKTQGAAAFEGAVIAYEPVWAIIGTGKSATPAQAQAVHKFIRDHI-A	195
TPIS_HUMAN	196	NVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPE-FVDI	244
TPIS_ECOLI	196	KVDANIAEQVIIQYGGSVNASNAELFAQPDIDGALVGGASLKADAFVI	245
TPIS_HUMAN	245	INAKQ	249
TPIS_ECOLI	246	VKAAEAAKQA	255

半分弱が一致

いずれもトリオースリン酸異性化酵素、
同じ機能、同じタンパク質由来 = ホモログ

ホモログの間の配列相同性はどの程度低いものがあるか？
トリオースリン酸異性化酵素では、ヒトと大腸菌で44%だった
全くランダムなアミノ酸配列をランダムに比べた場合、
20種のアミノ酸なので $1/20$ つまり5%の一致となる、

したがって、5%程度の場合は、ホモロジーは全くない

一般にアラインメントにより**20%以上**の配列相同性が認められる場合
ホモロジーがあると考えてよい、

一方、15%でもホモロジーがある場合があり、15%~20%程度の領域を
Twilight zone(夕暮でものがはっきり見えない時間帯)とよぶことがある。

注意しなければならない違いとしては
存在するタンパク質のアラインメントの結果として50%の一致があれば、
ほぼ同じ構造で同じ機能のタンパク質となるが、
あるタンパク質の配列をたとえば10%任意に入れ替えて90%相同な
アミノ酸配列を作ったとしても同じ形にフォールドして機能するタンパク質
とはならない。

配列の類似と立体構造の類似

ヒトのヘモグロビンの α 鎖と β 鎖 (SeqID 46.0%)

Alpha 2: LSPADKTNVKAAWGKVGGAHAGEYGAELERMFLSFPTTKTYFPHF-DLS-----HGSAQV: 55

* *

Beta 3: LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLS TPDAVMGNPKV: 60

Alpha 56: KGHGKKVADALTNVAHAVDDMPNALSALSDDLHAHKL RVPVNFKLLSHCLLVTLAAHLPA: 11

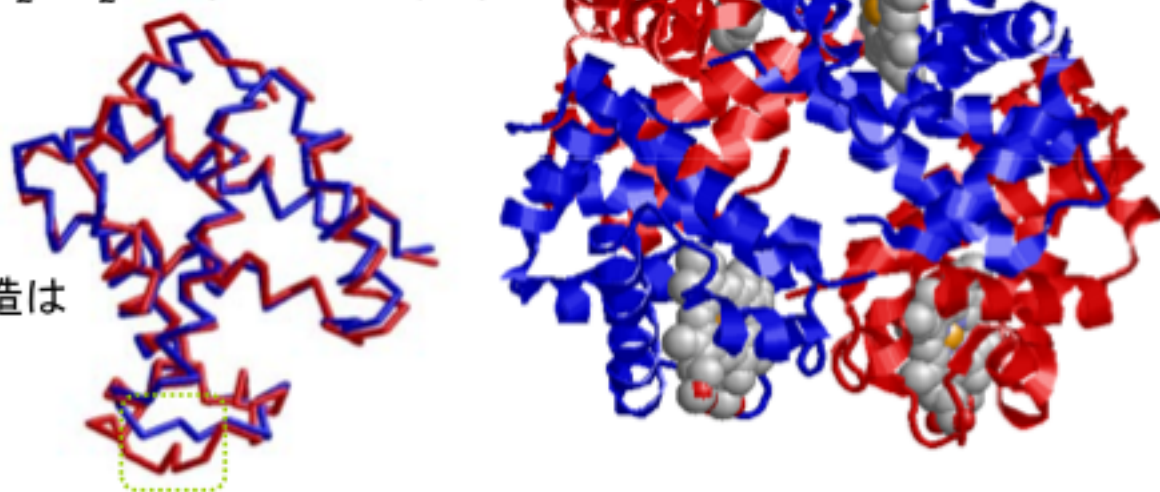
* *

Beta 61: KAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRL LGNVLVCVLAH HFGK: 120

Alpha 116: EFTP AVHASLDKFLASVSTVLT SKY: 140

* *

Beta 121: EFT PPVQAAYQKV VAGVANALAHKY: 145



機能や立体構造はよく似ている

配列の類似を知ることは立体構造予測につながる

配列比較(配列相同性検索)の基本論理

①2つの DNA / アミノ酸 の文字列が似ている



②進化的に関係がある(相同)から似ている



③進化的に関係があるなら、他の生物学的な性質(機能、立体構造など)も似ているはず

相同性の発見により、他の生物学的な性質を予測できる

類似(similarity)

相同 (homology):進化的な原因によるもの。祖先を共有。

(進化史の中である時点まで同じであったから似ている)

相似 (analogy):それ以外の原因によるもの

2つの配列を比較するには？

1. 類似性のスコア関数の定義

文字の間の類似性をどうやって定量するか？

ACFDE

** *

ACEEE

3つ同じだから3点？

FとEの対応とDとEの対応は等価だろうか？

2. アライメント

どうやって文字と文字を対応づけるか？

ABCDEF



ABCDEF

CDE

--CDE--

BCDEF



-BCDEF-

* **

ABEEFG

AB-EEFG

もっと長いときはどうやって計算する？

スコア関数の定義

(1)一致・不一致スコア

$$S(A, B) = \begin{cases} \alpha & A = B \\ \beta & A \neq B \end{cases}$$

もっとも簡単。DNAの場合によく使われる。
BLASTの核酸のデフォルトは、 $\alpha=1, \beta=-3$

	A	T	G	C
A	1	-3	-3	-3
T	-3	1	-3	-3
G	-3	-3	1	-3
C	-3	-3	-3	1

問題点: 文字列間の類似性を捉えられない。

L(ロイシン, 疎水性) → V(バリン, 疎水性) : 起こりやすい

L(ロイシン, 疎水性) → E(グルタミン酸, 一荷電) : 起こりにくい

(2)対数オッズスコア(log odds score)

$$S(A, B) = \log \frac{P_{evo}(A, B)}{P_{rand}(A)P_{rand}(B)}$$

2つの異なるタンパク質のあるサイトのアミノ酸がA,Bであったとき、

Protein1 : XXXXA XXXX

Protein2 : XXXXB XXXX

$P_{evo}(A, B)$: 進化的な関係からAとBの対応が生じた確率

$P_{rand}(A) \cdot P_{rand}(B)$: 偶然にAとBの対応が生じた確率。

E-value

配列相同性のもう一つの指標

後で用いる配列相同性検索プログラム: BLASTで用いられる

ランダムな配列の比較で、偶然にスコアSが生じる可能性

0~1で、低いほど、相同性が高いと考えられる

ひとつの目安として、 $0.0001 = 10^{-4}$ より小さければホモロジーが有ると考える

BLASTの出力では、指数表記で表されるので注意

例えば、 10^{-4} は $1.0e-4$ と表記される

0.24 は $2.4e-1$, 0.0000000000098 は、 $9.8e-11$ と表記される