

計算生命科学の基礎10
2023年10月11日

タンパク質の機能の進化と その情報解析

藤 博幸

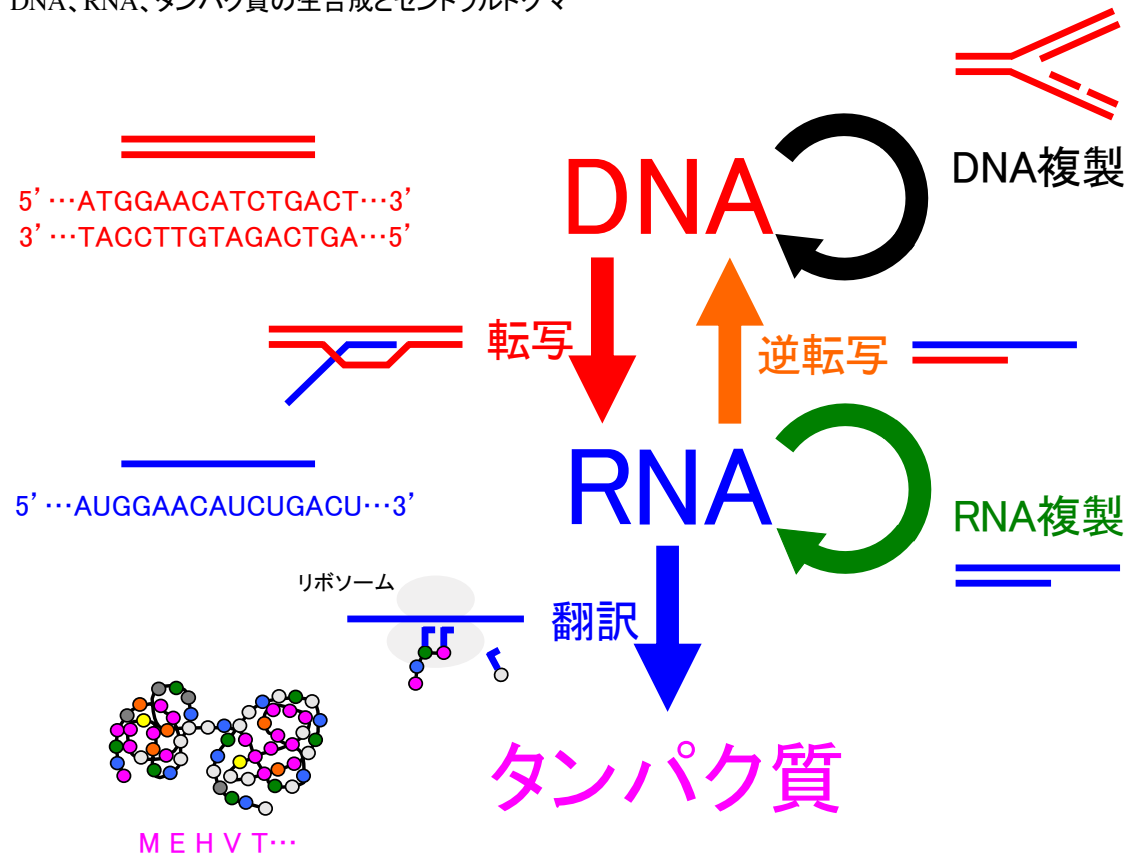
関西学院大学

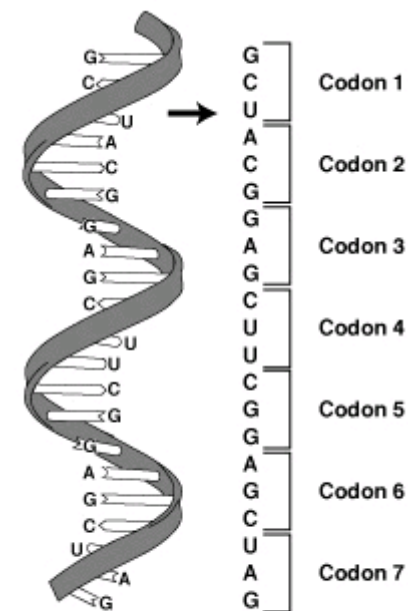
生命環境学部生物科学科

本日の予定

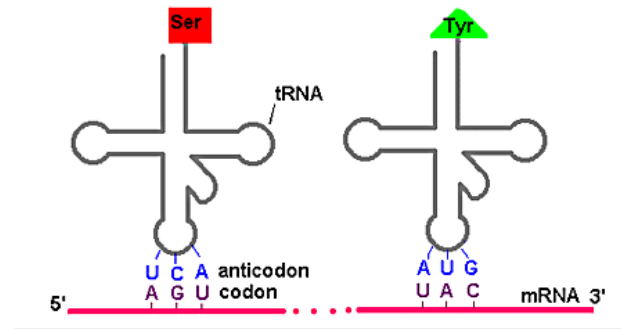
1. タンパク質
2. タンパク質の機能
3. 情報解析に使用するタンパク質のデータ
4. 配列解析
5. 構造解析
6. ネットワーク解析

図1.2 DNA、RNA、タンパク質の生合成とセントラルドグマ





Ribonucleic acid



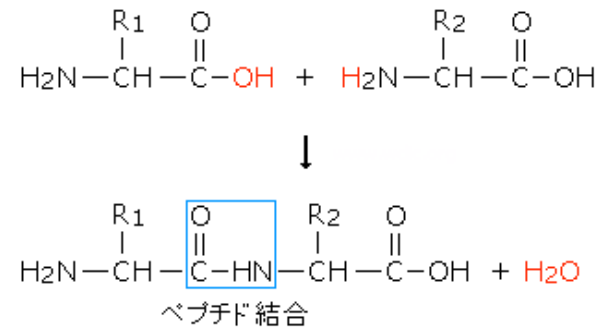
2nd base in codon

	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

1st base in codon

3rd base in codon

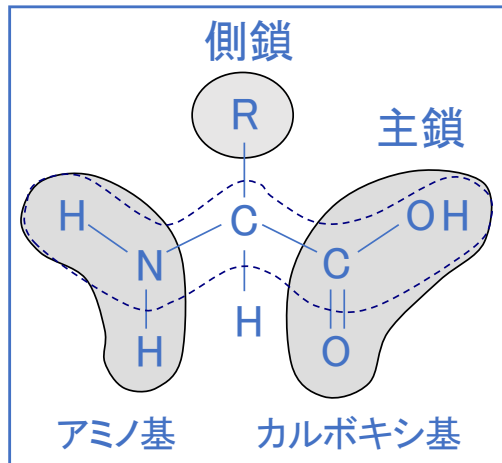
The Genetic Code



<http://www.wdic.org/w/SCI/ペプチド結合>

タンパク質は20種類のアミノ酸がペプチド結合でつながった紐状の分子
 各アミノ酸を1文字で表現すると、一つのタンパク質は、20種類のアルファベットで構成された文字列として表すことができる。

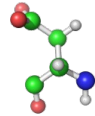
アミノ酸



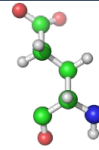
基本構造

アミノ酸は
”大文字”
で表記する

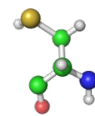
Asp (D)



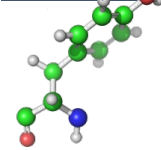
Glu (E)



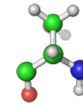
Cys (C)



Tyr (Y)

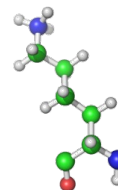


Ser (S)

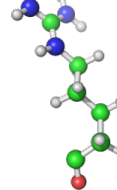


解離性アミノ酸

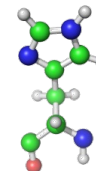
Lys (K)



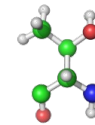
Arg (R)



His (H)



Thr (T)



親水性
アミノ酸

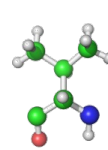
Gly (G)



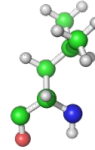
Ala (A)



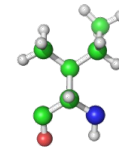
Val (V)



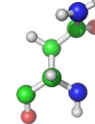
Leu (L)



Ile (I)

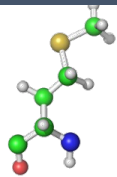


Asn (N)

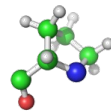


疎水性アミノ酸

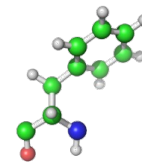
Met (M)



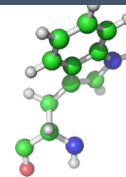
Pro (P)



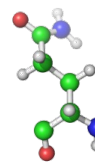
Phe (F)

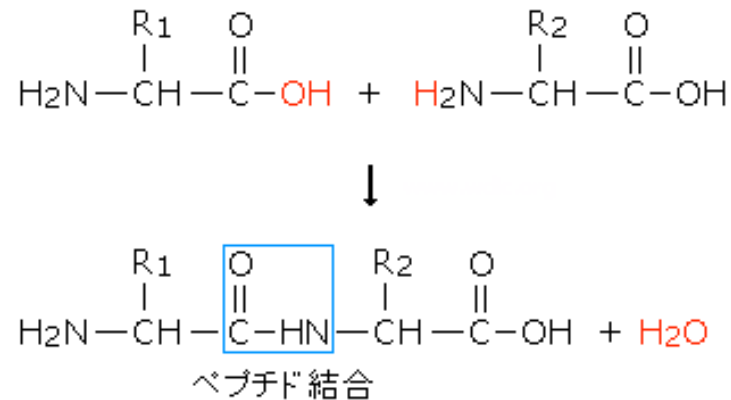


Trp (W)



Gln (Q)





<http://www.wdic.org/w/SCI/ペプチド結合>

タンパク質は20種類のアミノ酸がペプチド結合でつながった紐状の分子
 各アミノ酸を1文字で表現すると、一つのタンパク質は、20種類のアルファベットで構成された文字列として表すことができる。

リゾチームのアミノ酸配列

>LYC_HUMAN ヒト Lysozyme C

MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRST
DYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVR
QYVQGC

>LYC1_BOVIN ウシ Lysozyme C 1

MKALIILGFLFLSVAVQKVFERCELARTLKRLGLDGYKGVSLANWLCLTKWESSYNTKATNYPGSEST
DYGIFQINSKWWCNDGKTPNAVDGCHVSCSELMENDIAKAVACAKQIVSEQGITAWEVAVKSHCRDHDVSS
YVEGCTL

>LYC_CHICK ニワトリ Lysozyme C

MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTD
YGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAGKIVSDGNGMNAWVAWRNRCKGTDVQA
WIRGCRL

>LYC2_ONCMY マス Lysozyme C II

MRAVVVLLLVAVASAKVYDRCELARALKASGMDGYAGNSLPNWWCLSKWESSYNTQATNRNTDGSTDYGI
FQINSRYWCDDGRTPGAKNVCGIRCSQLLTADLTVAIRCAKRVVLDPNGIGAWVAWRLHCQNQDLRSYVA
GCGV

>LYC_BOMMO カイコ Lysozyme

MQKLIIFALVVLVGVSEAKTFTRCGLVHELKRGHGFEEENLMRNWVCLVEHESSRDTSKTNTNRNGSKDYGL
FQINDRYWCSKGASPGKDCNVKCSDLLTDDITKAACAKKIYKRHRFDWYGWKNHCQGSPLDISSC

>LYSP_DROME ハエ Lysozyme P

MKAFLVICALTLTAVATQARTMDRCSLAREMSKLGVPDRQLAKWTCIAQHESSEFRTGVVGPANSNGSNDY
GIFQINNKYWCKPADGRFSYNECGLSCNALLTDDITNSVKCARKIQRQOGWTAWSTWKYCSGSLPSINSC
F

FASTA フォーマット

```
>LYC_HUMAN ヒト Lysozyme C
```

```
MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWES  
GYNTRATNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIA  
DAVACAQRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGC
```

“>” で始まる注釈行（通常、配列名）と
改行後に、1文字表記で塩基あるいはアミノ酸の配列を記す書式を
FASTAフォーマットといい、分子系統解析ばかりでなく、バイオインフォ
マティクス分野では配列の書式としてよく使われている。

前ページのように、一つのファイルに複数の**FASTA**フォーマットの
配列がおさめられている場合は、マルチ**FASTA**フォーマットとよぶ。

配列データベース

塩基配列データベース

INSDC (International Nucleotide Sequence Database Collaboration)

米国 NCBIの**GenBank**

欧州 EMBL-EBIの**ENA**

日本 国立遺伝学の**DDBJ**

3者が連携して国際的に維持管理

タンパク質のデータベース

UniProt

以下の3者のコンソーシアムとして統合

スイスバイオインフォマティクス研究所(SIB)の**Swiss-Prot**

EBIの**TrEMBL**

ジョージタウン大学の**PIR**

塩基配列データベースのエントリ: **GenBank形式** (タンパク質の場合: **GenPept形式**)

LOCUS AH003092 6907 bp DNA linear ROD 01-AUG-2016
DEFINITION Nannospalax ehrenbergi clone pScr-1 crystallin (CRY-alpha-A) gene,
complete cds, alternatively spliced.
ACCESSION AH003092 M17247 M17248 M17249 Y00464
VERSION AH003092.2
KEYWORDS alpha-crystallin; crystallin.
SOURCE Nannospalax ehrenbergi (Ehrenberg's mole-rat)
ORGANISM Nannospalax ehrenbergi
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
Muridae; Spalacidae; Spalacinae; Nannospalax.
REFERENCE 1 (bases 2161 to 2880; 3841 to 4080; 5592 to 6431)
AUTHORS Hendriks,W., Leunissen,J., Nevo,E., Bloemendal,H. and de Jong,W.W.
TITLE The lens protein alpha A-crystallin of the blind mole rat, Spalax
ehrenbergi: evolutionary change and functional constraints
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 84 (15), 5320-5324 (1987)
PUBMED 3474658

配列の定義
(学名、遺伝子名)

アクセッション番号

キーワード

学名(生物資料の由来)

分類

文脈

FEATURES
source Location/Qualifiers
1..6907
/organism="Nannospalax ehrenbergi"
/mol_type="genomic DNA"

コード領域 CDS
join [2349..2537,2736..2804,3859..3981,5715..5924]
/gene="CRY-alpha-A"
/codon_start=1
/product="alpha-A-ins-crystallin"
/protein_id="AAA66165.1"
/translation="MDVTIQHPWFKHALGFPYPSRLFDQFFQGGLFYEDLLPFLSSTI
SPYYRQTLRLTVLDSICSLMTHREWEVHOPHAGNEENELKVRSDRDKFVIFLDVKH
FSPEDLTVKVLEDFVEIHGKHNERQDDHGYISREFHRRYRLPSSVDQSALSCSLADG
MLTFSGPKVQSLDAGHSERAIQVSEKPKSSAPLF"

アミノ酸配列
(ins あり)

コード領域 CDS
join [2349..2537,3859..3981,5715..5924]
/gene="CRY-alpha-A"
/codon_start=1
/product="alpha A-crystallin"
/protein_id="AAA66166.1"
/translation="MDVTIQHPWFKHALGFPYPSRLFDQFFQGGLFYEDLLPFLSSTI
SPYYRQTLRLTVLDSICSEVRSRDKFVIFLDVKHFSPELTVKVLEDFVEIHGKHNE
RQDDHGYISREFHRRYRLPSSVDQSALSCSLADGMLTFSGPKVQSLDAGHSERAIQ
VSEKPKSSAPLF"

アミノ酸配列
(ins なし)

ORIGIN 1 bp upstream of EcoRI site.
1 gaattctctc ttgtggagca ggccttaaat ccaagcagaa agtggtttct caccceccata
2281 gctcccccact ccagcggccc ctgcccgact caccaccagc cttoagtcoc tgcgttgcca
2341 cgtggaacat ggaagtcacc atccagcacc cctggttcaa gcaagccctg ggtccctctt
2401 atccacgccc gctgttcgac cagtttttcg gccaggccct ctttgagtac gatctcgtgc
2461 ccttcctgtc ctccacatcc agccctattt accgccagac cctcctccgc acagtgtgtg
2521 actcctgcac ctctgaggtt agatggcggt ggtcagaggg gctcagggcc ttggagaggt
2581 cagtagagga ttctctgagg gccacctgcc tegtccctgc ctctccagc ctcggacctc
2641 ggacctcgtc gcggctgggt ggaagtcgcc ccgccacagc gaetgggaat ggggtttgaa
2701 gagaagggcc gactcgtgtc ttcttcctcc atcatctcat gaaccatagg tagttttgac
2761 cgaacacacc acatgctgga aaccccagaa acaaccccat caagcgaagt tctgtggcaa
2821 aggcgagcgt cctggaccgc ctgtccctct tgcctttctg ccgggtgcgg gtgcggtgat
3781 accatgctgt tgcacctccc acccccggtt ctggagctgt tagctctctg gagctccaca
3841 tgaaggtctt ggttctaggt cctgacccgc cggagcaaat cctgcatctt cctgcatctg
3901 aagcactctt cctctgagga cctcaacgtg aaggtgtgga aggaactcgt ggagatccat
3961 ggcaagcaca acagagagca ggtgagctcg atggggacct caccatgcgg gcatcctgta
4021 tcccttgaga ggaagcaagg ggaagtgcca ctggctcagc tgcctcgtc agaaggagcc
5641 ggtctttttc cagttgggga tggcagtcct gtgtctcacc atgcctgacc atgcctcttc
5701 ctggtttgtc ccagagacag cacygetaca ttctccgtga gtccacagc cgtacacccc
5761 tgccttccag tggagacag tctgcctctt cctgcctcct gtctgcagat ggcctgtctg
5821 ccttctctgg ccccaagatc caactctggc tggatgctgg ccacagttaa agggacattc
5881 ctgtgtctca ggaggagaag cccagctcgg cacccttggt ctacgcagcg ctacacttag
5941 ctgcacctg aggcacatga cccatcaagg tcagggacca cagcaaaagg tcttctcttc
6841 acacacagc atacacatga cacatatact actgcgcatg cacagcacac acacactgat
6901 caaccaca

イントロン1

イントロン2

イントロン3

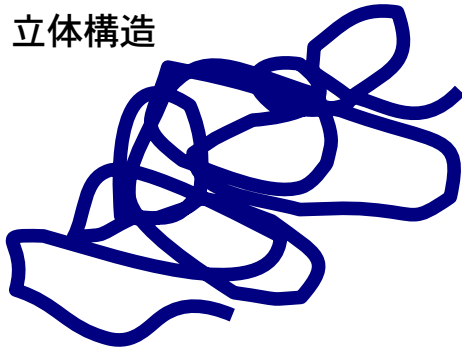
イントロン 1-ins-2

//

アミノ酸配列



立体構造



ヒト・リゾチームのアミノ酸配列

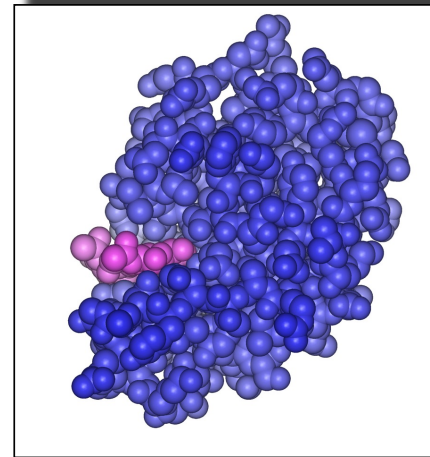
(一文字表記)

>LYC_HUMAN

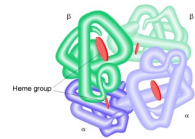
MKALIVLGLVLLSVTVQGKVFERCELARTL
KRLGMDGYRGISLANWMCLAKWESGYNTRA
TNYNAGDRSTDYGIFQINSRYWCNDGKTPG
AVNACHLSCSALLQDNIADAVACAKRVVRD
PQGIRAWVAWRNRCQNRDVRQYVQGCGV

148 残基

ヒト・リゾチームの立体構造

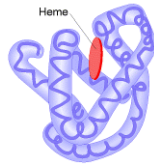


タンパク質立体構造の階層性



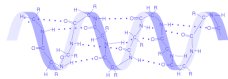
四次構造→

立体構造をとる二本以上のペプチド鎖から形成される構造
(具体例: ヘモグロ빈は4つのペプチド鎖が集まり機能 (α サブユニット $\times 2 + \beta$ サブユニット $\times 2$))

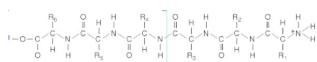


三次構造→

ドメイン構造 -->



二次構造→



一次構造→

立体構造

立体構造

α ヘリックス

β シート

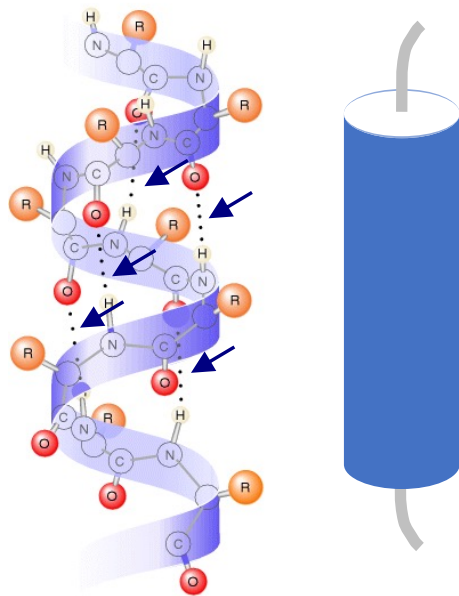
ターン

ループ

アミノ酸配列

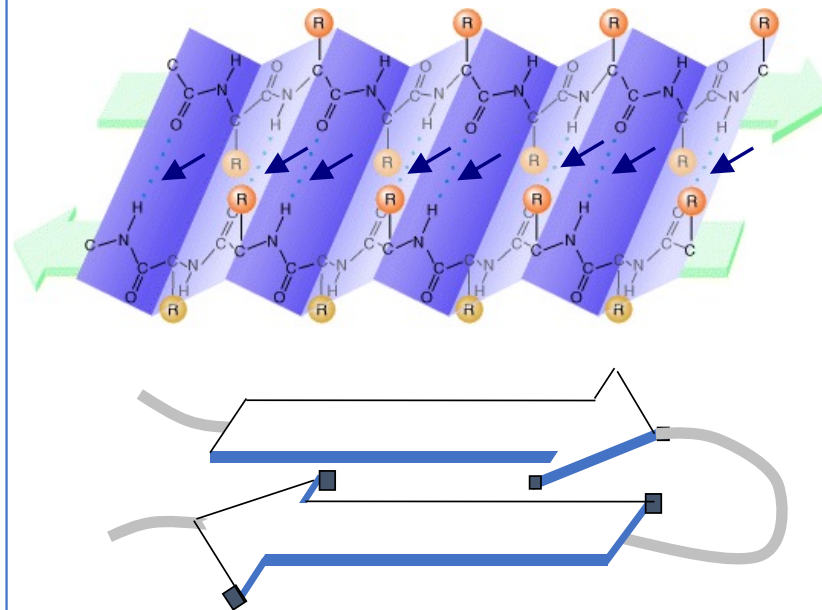
二次構造 規則的な水素結合のパターンを持つ

α ヘリックス



4残基先のアミノ酸と水素結合を作る

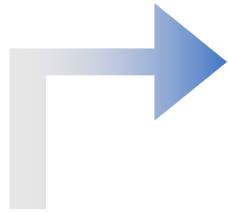
β シート



配列上離れた位置にアミノ酸と水素結合を作る

『MOLECULAR CELL BIOLOGY』より引用

三次構造

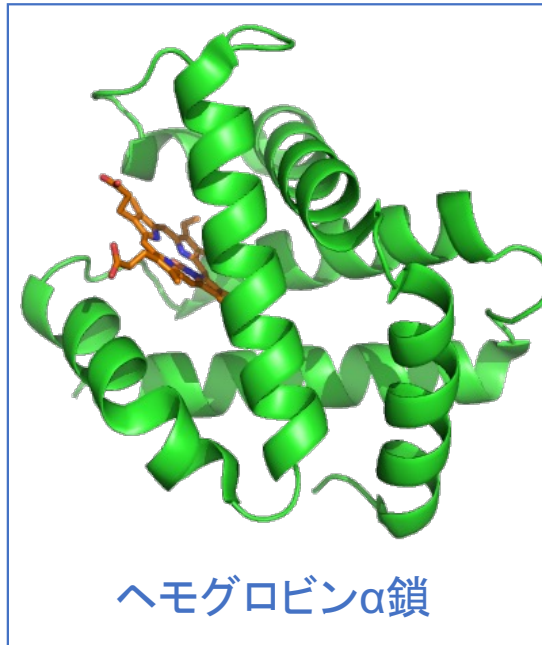


内側：

疎水性残基が集まる

外側：

親水性残基が多い



独自の安定な構造をとる

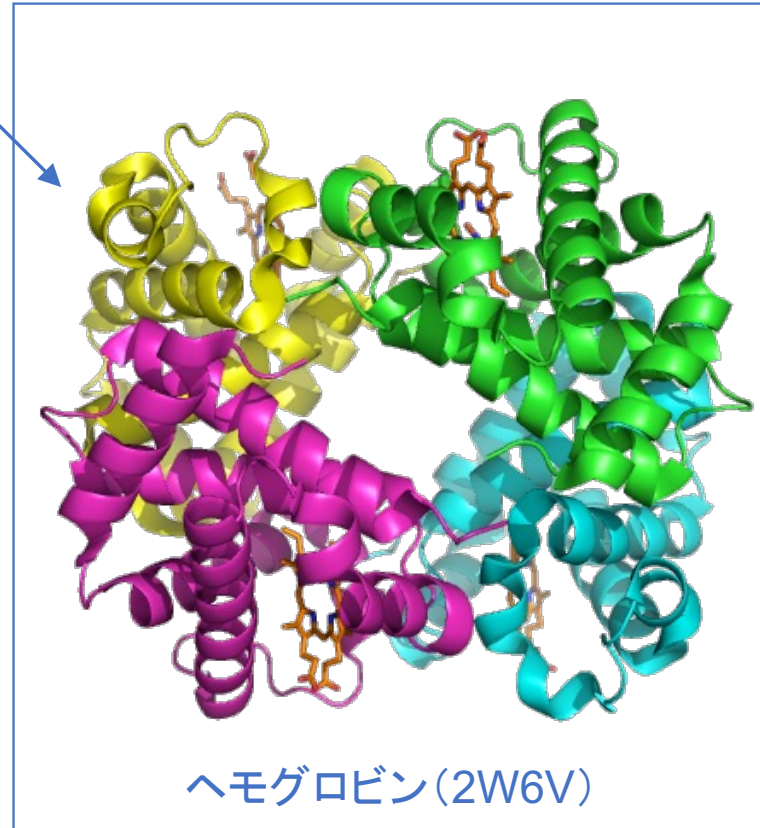
ドメイン：

- ・比較的安定な構造をとる単位
- ・三次構造より
(たいていの場合)小さな構造
- ・ドメインは、様々な三次構造
中に見られる

二次構造の組み合わせで、ある形に折りたたまれたもの。
一本のポリペプチドからなる。

四次構造

サブユニット



多量体の呼び方

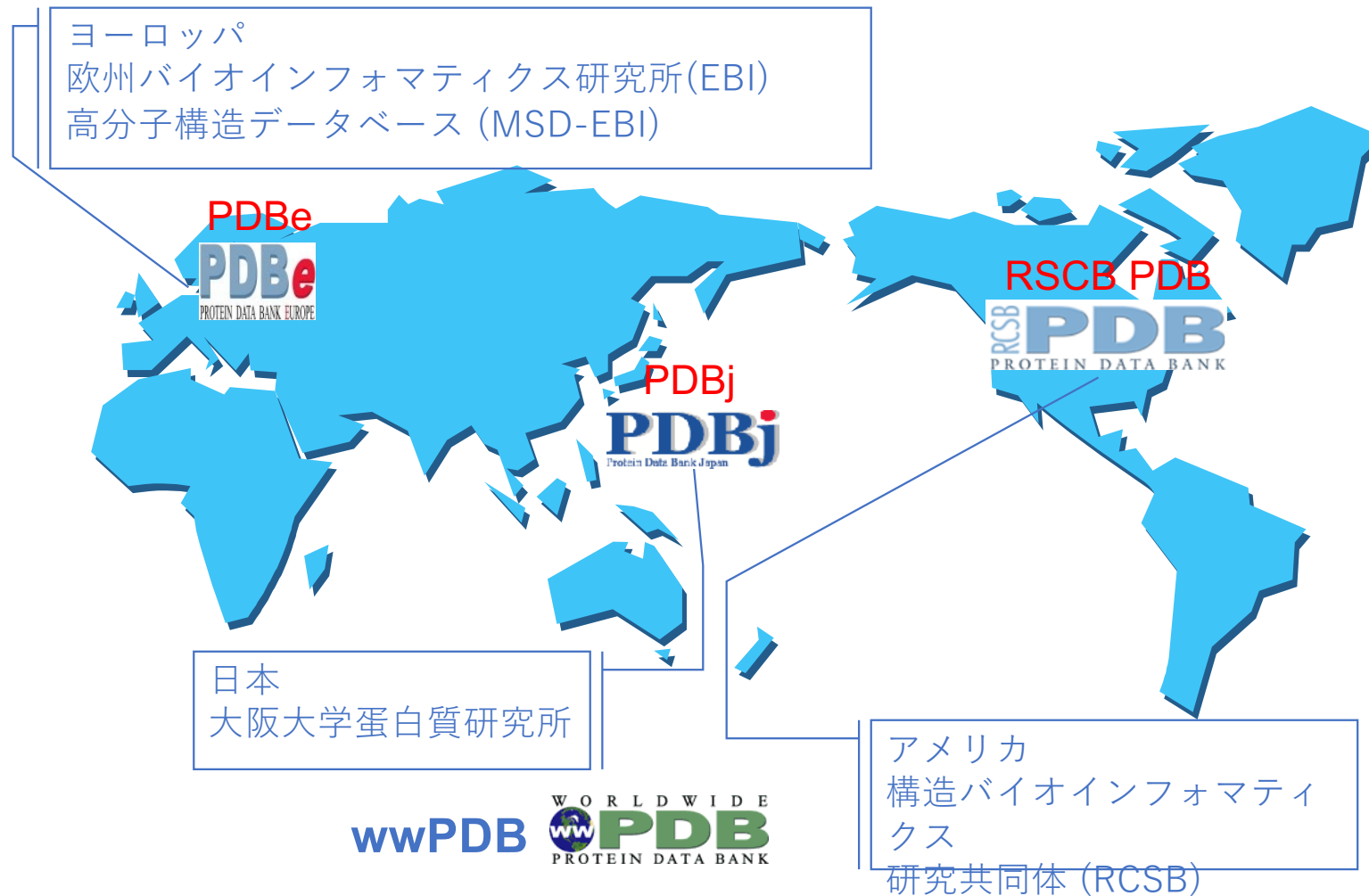
- 1本 : モノマー
- 2本 : ダイマー
- 3本 : トリマー
- 4本 : テトラマー
- 5本 : ペンタマー
- 6本 : ヘキサマー

生体内で機能しているタンパク質の多くは、四次構造を形成している

複数のポリペプチドによって形成される立体構造

生体高分子の立体構造を扱ったデータベース

- 網羅的に立体構造情報を収集しているもの



生体高分子の立体構造を扱ったデータベース

• PDBファイルの中身

HEADER	TRANSFERASE	ヘッダー	ボディ
TITLE	CRYSTAL STRUCTURE OF THE	HEADER	物質の分類名、日付、ファイル名
TITLE	2 GLEEVEC	COMPND	タンパク質名
COMPND	MOL_ID: 1;	SOURCE	物質が由来する生物名
COMPND	2 MOLECULE: TYROSINE KINASE	AUTHOR	座標を作成した著者名
COMPND	3 CHAIN: A;	REVDAT	データが登録された日
COMPND	4 SYNONYM: SPLEEN	JRNL	文献情報
COMPND	5 EC: 2.7.1.112	REMARK	実験条件などに関する情報
COMPND	6 ENGINEERED: YES	SEQRES	アミノ酸配列
SOURCE	MOL_ID: 1;	HET	アミノ酸以外の原子
SOURCE	2 ORGANISM_SCIENTIFIC	HELIX	α -ヘリックスに関する情報
SOURCE	3 ORGANISM_COMMON	SHEET	β -シートに関する情報
SOURCE	4 ORGANISM_TAXID	TURN	β -ターンに関する情報
SOURCE	5 GENE: SYK;	SSBOND	ジスルフィド結合に関する情報
SOURCE	6 EXPRESSION_SYSTEM	ATOM	個々のアミノ酸の原子の座標
SOURCE	7 EXPRESSION_SYSTEM	HETATM	アミノ酸以外の原子の座標
SOURCE	8 EXPRESSION_SYSTEM	CONTACT	原子の結合に関する情報
SOURCE	9 EXPRESSION_SYSTEM	END	ファイルの終了
SOURCE	10 EXPRESSION_SYSTEM		
KEYWDS	GLEEVEC, STI-571		
KEYWDS	2 ACTIVE CONFORMATION		
KEYWDS	3 GENOMIX, TRANSDOMAIN		
EXPDTA	X-RAY DIFFRACTION		
AUTHOR	V. L. NIENABER, S. J. DODD		
AUTHOR	2 I. K. FEIL, K. J. VAN		
AUTHOR	3 H. J. MULLER-DEBALT		
AUTHOR	4 A. RAMOS, M. RUSSELL		
(省略)			

生体高分子の立体構造を扱ったデータベース

• 座標データ

ATOM	1	N	VAL	A	363	22.741	-1.397	11.729	1.00	33.32	N
ATOM	2	CA	VAL	A	363	21.557	-0.831	11.024	1.00	32.13	C
ATOM	3	C	VAL	A	363	20.954	-1.757	9.943	1.00	31.73	C
ATOM	4	O	VAL	A	363	19.737	-1.906	9.845	1.00	30.94	O
ATOM	5	CB	VAL	A	363	21.883	0.552	10.391	1.00	33.45	C
ATOM	6	N	TYR	A	364	21.798	-2.389	9.135	1.00	29.77	N
ATOM	7	CA	TYR	A	364	21.310	-3.035	7.928	1.00	27.96	C
ATOM	8	C	TYR	A	364	20.929	-4.485	8.157	1.00	26.43	C
ATOM	9	O	TYR	A	364	21.735	-5.274	8.650	1.00	29.31	O
ATOM	10	CB	TYR	A	364	22.349	-2.901	6.808	1.00	28.13	C
ATOM	11	CG	TYR	A	364	22.619	-1.461	6.442	1.00	28.30	C
ATOM	12	CD1	TYR	A	364	21.725	-0.751	5.638	1.00	30.33	C
ATOM	13	CD2	TYR	A	364	23.772	-0.815	6.893	1.00	30.80	C
ATOM	14	CE1	TYR	A	364	21.959	0.588	5.285	1.00	31.59	C
ATOM	15	CE2	TYR	A	364	24.019	0.523	6.541	1.00	32.22	C
ATOM	16	CZ	TYR	A	364	23.101	1.208	5.733	1.00	32.48	C
ATOM	17	OH	TYR	A	364	23.330	2.524	5.389	1.00	36.29	O

1原子

1残基

1残基

原子の名前

アミノ酸の名
前

チェーンの名
前

アミノ酸残基番
号

x座標

y座標

z座標

占有率

温度因子

原子

現在のPDBのフォーマット
旧PDBフォーマットは1行80文字の
固定幅であったため、原子番号、
残基番号、チェーン番号などの
文字数に制限があり、100個以上
チェーンからなるような巨大な
タンパク質複合体の構造を取り扱え
なくなった。

data 1B21																				
# entry.id 1B21																				
#																				
_pdbx_database_status.recvd_initial_deposition_date 1998-11-05																				
#																				
_struct.title 'HEMOGLOBIN (ALPHA + MET) VARIANT'																				
_struct.pdbx_descriptor 'PROTEIN (HEMOGLOBIN) VARIANT (CHAIN A, C, ADDITIONAL NH2-TERMINAL MET)'																				
loop_																				
atom_site.group_PDB																				
atom_site.id																				
atom_site.type_symbol																				
atom_site.label_atom_id																				
atom_site.label_alt_id																				
atom_site.label_comp_id																				
atom_site.label_asym_id																				
atom_site.label_entity_id																				
atom_site.label_seq_id																				
atom_site.pdbx_PDB_ins_code																				
atom_site.Cartn_x																				
atom_site.Cartn_y																				
atom_site.Cartn_z																				
atom_site.occupancy																				
atom_site.B_iso_or_equiv																				
atom_site.pdbx_formal_charge																				
atom_site.auth_seq_id																				
atom_site.auth_comp_id																				
atom_site.auth_asym_id																				
atom_site.auth_atom_id																				
atom_site.pdbx_PDB_model_num																				
ATOM	1	N	N	.	MET	A	1	1	?	15.774	28.408	41.946	1.00	87.06	?	1	MET	A	N	1
ATOM	2	C	CA	.	MET	A	1	1	?	17.105	28.442	42.578	1.00	83.35	?	1	MET	A	CA	1
ATOM	3	C	C	.	MET	A	1	1	?	18.021	29.477	41.921	1.00	75.77	?	1	MET	A	C	1
ATOM	4	O	O	.	MET	A	1	1	?	17.695	30.170	40.943	1.00	76.91	?	1	MET	A	O	1
ATOM	5	C	CB	.	MET	A	1	1	?	17.757	27.078	42.696	1.00	100.06	?	1	MET	A	CB	1
ATOM	6	C	CG	.	MET	A	1	1	?	18.477	26.602	41.467	1.00	108.30	?	1	MET	A	CG	1
ATOM	7	S	SD	.	MET	A	1	1	?	19.741	25.379	41.972	1.00	109.12	?	1	MET	A	SD	1
ATOM	8	C	CE	.	MET	A	1	1	?	18.713	24.184	42.861	1.00	108.71	?	1	MET	A	CE	1
ATOM	9	N	N	.	VAL	A	1	2	?	19.204	29.547	42.527	1.00	63.43	?	2	VAL	A	N	1
ATOM	10	C	CA	.	VAL	A	1	2	?	20.227	30.471	42.079	1.00	50.48	?	2	VAL	A	CA	1
:																				
HETATM	4401	C	CHA	.	HEM	E	3	.	?	18.729	18.645	20.255	1.00	12.22	?	143	HEM	A	CHA	1
HETATM	4402	C	CHB	.	HEM	E	3	.	?	21.007	20.618	24.060	1.00	16.53	?	143	HEM	A	CHB	1
HETATM	4403	C	CHC	.	HEM	E	3	.	?	18.570	17.704	27.086	1.00	11.28	?	143	HEM	A	CHC	1
:	原子番号	原子名	残基名	残基番号	X座標	Y座標	Z座標	占有率	温度因子	残基番号	残基名	原子名	残基番号	残基名	原子名	残基番号	残基名	原子名	残基番号	残基名

可視化ツールが必要

- 分子グラフィクス

RasMol <http://www.umass.edu/microbio/rasmol/>

PyMol <http://pymol.sourceforge.net/>

Chime
<http://www.mdli.com/downloads/downloadable/index.jsp>

PDBj viewer http://www.pdbj.org/PDBjViewer/index_j.html

MolScript <http://www.avatar.se/molscript/>

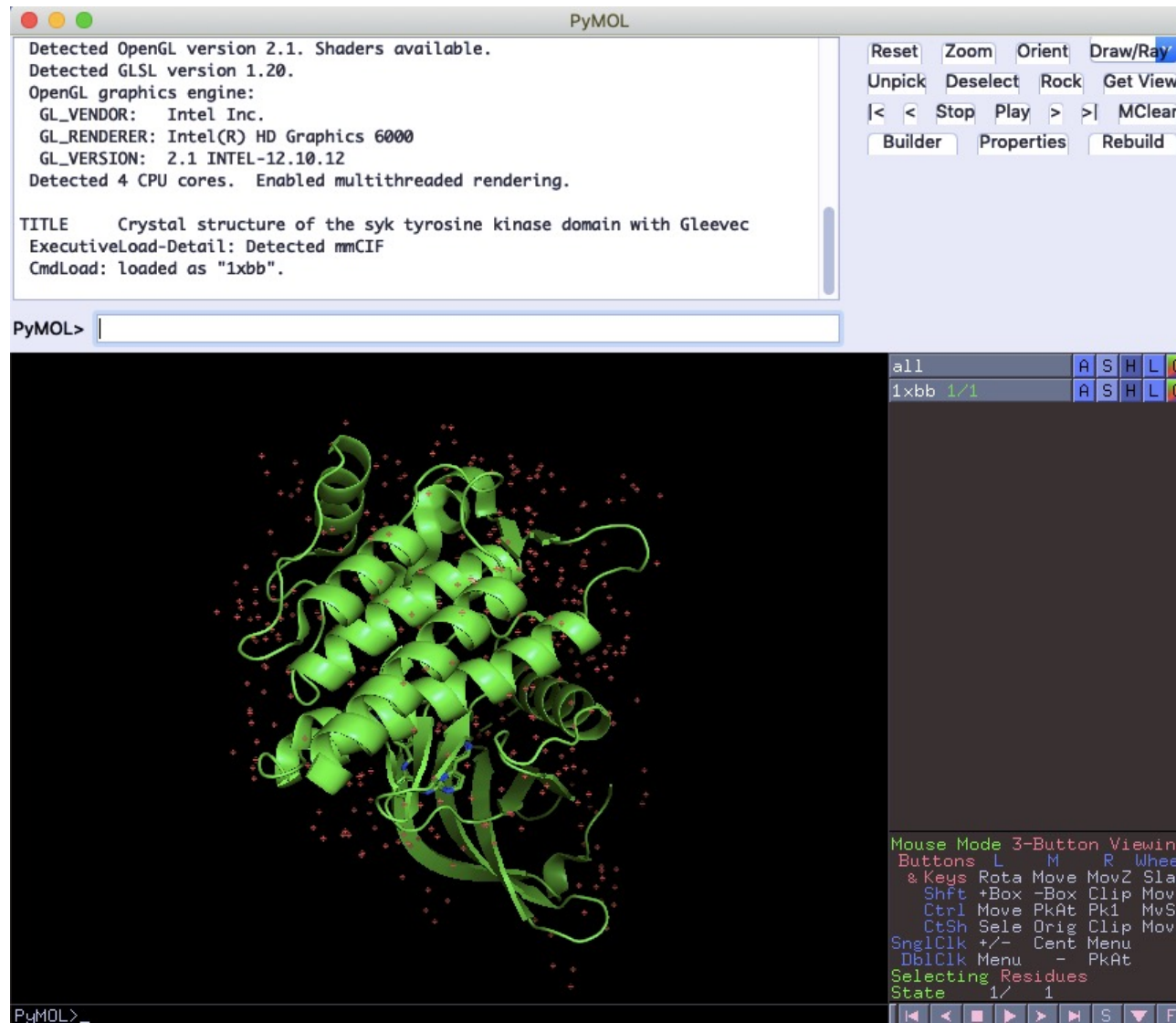
- 粗視化

コンタクトマップ

ラマチャンドランプロット

TOPS <http://www.tops.leeds.ac.uk/>

LIGPLOT <http://www.biochem.ucl.ac.uk/bsm/ligplot/ligplot.html>



チロシンキナーゼ
の構造がPyMolの
ウィンドウに表示
される

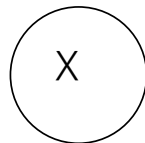
赤いドットは水

タンパク質の機能

生化学的機能

Biochemical Function

酵素活性、リガンド結合能

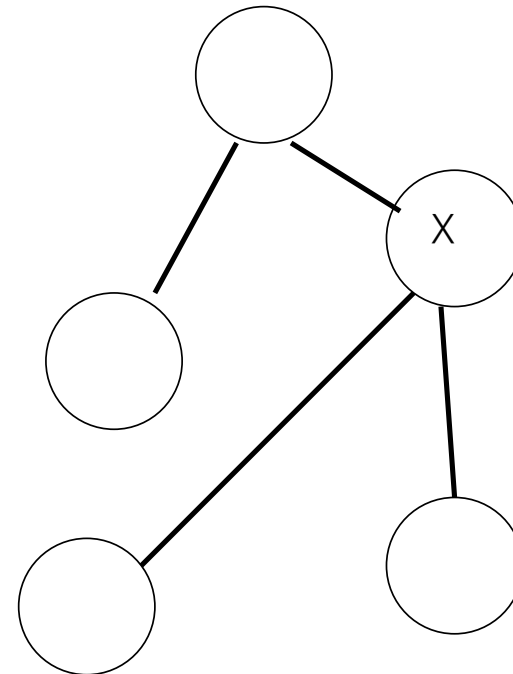


vs

生物学的機能

Biological Function

記憶、発生、行動



相互作用データベース

STRING

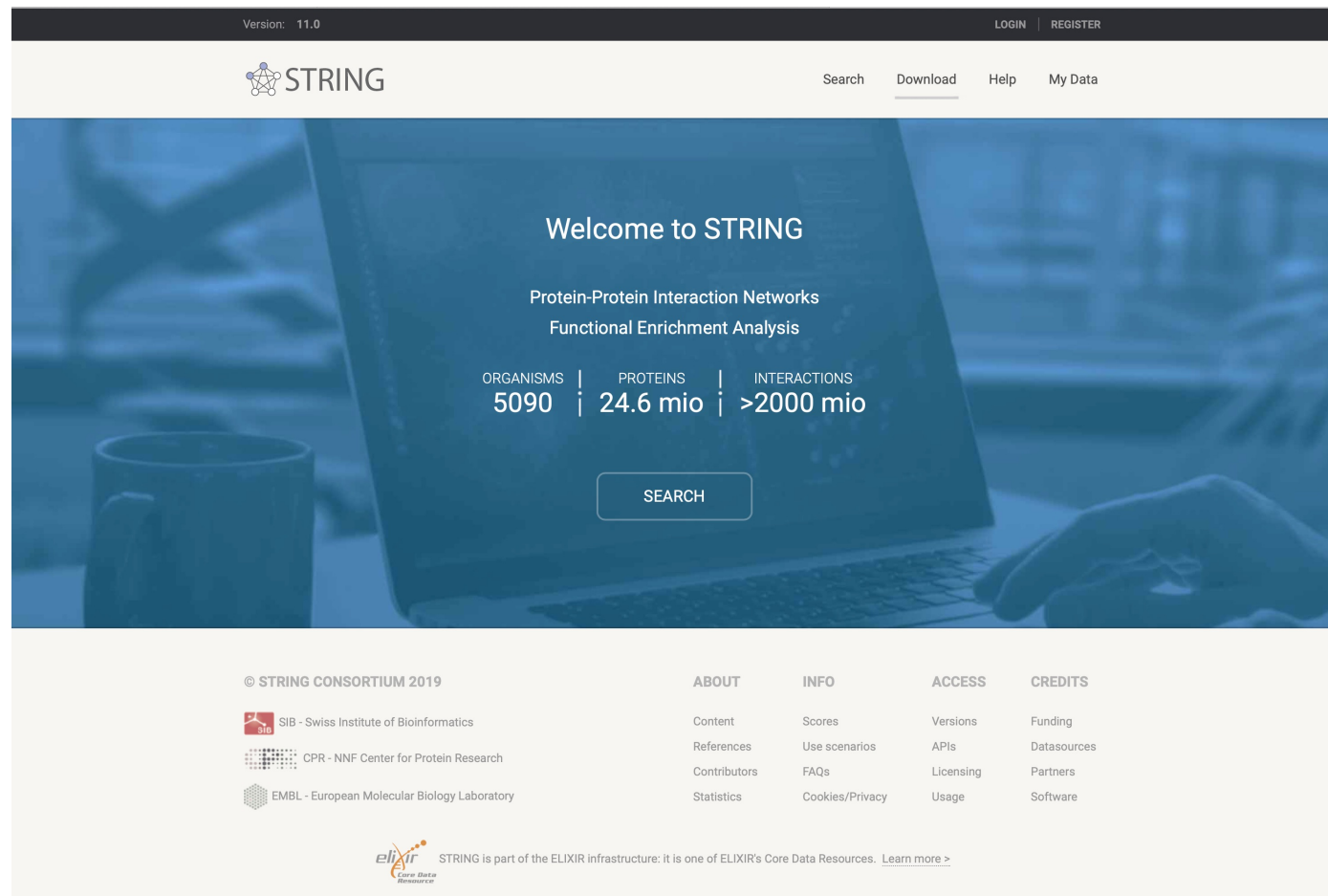
IntAct

BioGRID

STRING

<https://string-db.org>

クリック



The screenshot shows the STRING database homepage. At the top, a dark header bar contains 'Version: 11.0' on the left and 'LOGIN | REGISTER' on the right. Below this is a light beige navigation bar with the STRING logo on the left and links for 'Search', 'Download' (which is underlined), 'Help', and 'My Data' on the right. The main content area has a blue-tinted background image of a laptop and a mug. It features the text 'Welcome to STRING' followed by 'Protein-Protein Interaction Networks' and 'Functional Enrichment Analysis'. Below this, statistics are listed: 'ORGANISMS 5090 | PROTEINS 24.6 mio | INTERACTIONS >2000 mio'. A 'SEARCH' button is centered below the statistics. The footer is a light beige bar containing copyright information '© STRING CONSORTIUM 2019' and logos for SIB, CPR, and EMBL on the left. On the right, there are four columns of links: 'ABOUT' (Content, References, Contributors, Statistics), 'INFO' (Scores, Use scenarios, FAQs, Cookies/Privacy), 'ACCESS' (Versions, APIs, Licensing, Usage), and 'CREDITS' (Funding, Datasources, Partners, Software). At the bottom center, the ELIXIR logo is shown with the text 'STRING is part of the ELIXIR infrastructure: It is one of ELIXIR's Core Data Resources. Learn more >'.

Version: 11.0

LOGIN | REGISTER

STRING

Search Download Help My Data

Welcome to STRING

Protein-Protein Interaction Networks
Functional Enrichment Analysis

ORGANISMS | PROTEINS | INTERACTIONS
5090 | 24.6 mio | >2000 mio

SEARCH

© STRING CONSORTIUM 2019

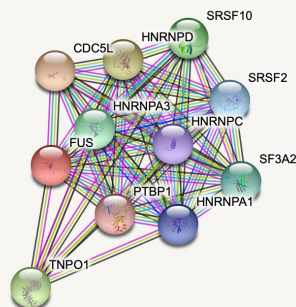
SIB - Swiss Institute of Bioinformatics
CPR - NNF Center for Protein Research
EMBL - European Molecular Biology Laboratory

ABOUT INFO ACCESS CREDITS

Content Scores Versions Funding
References Use scenarios APIs Datasources
Contributors FAQs Licensing Partners
Statistics Cookies/Privacy Usage Software

ELIXIR
Core Data Resource

STRING is part of the ELIXIR infrastructure: It is one of ELIXIR's Core Data Resources. [Learn more >](#)



[Viewers](#) ▾
 [Legend](#) >
 [Settings](#) >
 [Analysis](#) >
 [Exports](#) >
 [Clusters](#) >
 [More](#)
[Less](#)



Network currently showing
Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.



Experiments
Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.



Databases
Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.



Textmining
Automated, unsupervised textmining - searching for proteins that are frequently mentioned together.



Cooccurrence
Gene families whose occurrence patterns across genomes show similarities.



Coexpression
Proteins whose genes are observed to be correlated in expression, across a large number of experiments.



Neighborhood
Groups of genes that are frequently observed in each other's genomic neighborhood.



Fusion
Genes that are sometimes fused into single open reading frames.

STRING allows inspection of the interaction evidence for any given network. Choose any of the viewers above (disabled if not applicable in your network).

Server load: low (9%) [HD]

[Permalink](#)

© STRING CONSORTIUM 2019

 SIB - Swiss Institute of Bioinformatics
 CPR - NNF Center for Protein Research
 EMBL - European Molecular Biology Laboratory

ABOUT

[Content](#)
[References](#)
[Contributors](#)
[Statistics](#)

INFO

[Scores](#)
[Use scenarios](#)
[FAQs](#)
[Cookies/Privacy](#)

ACCESS

[Versions](#)
[APIs](#)
[Licensing](#)
[Usage](#)

CREDITS

[Funding](#)
[Datasources](#)
[Partners](#)
[Software](#)



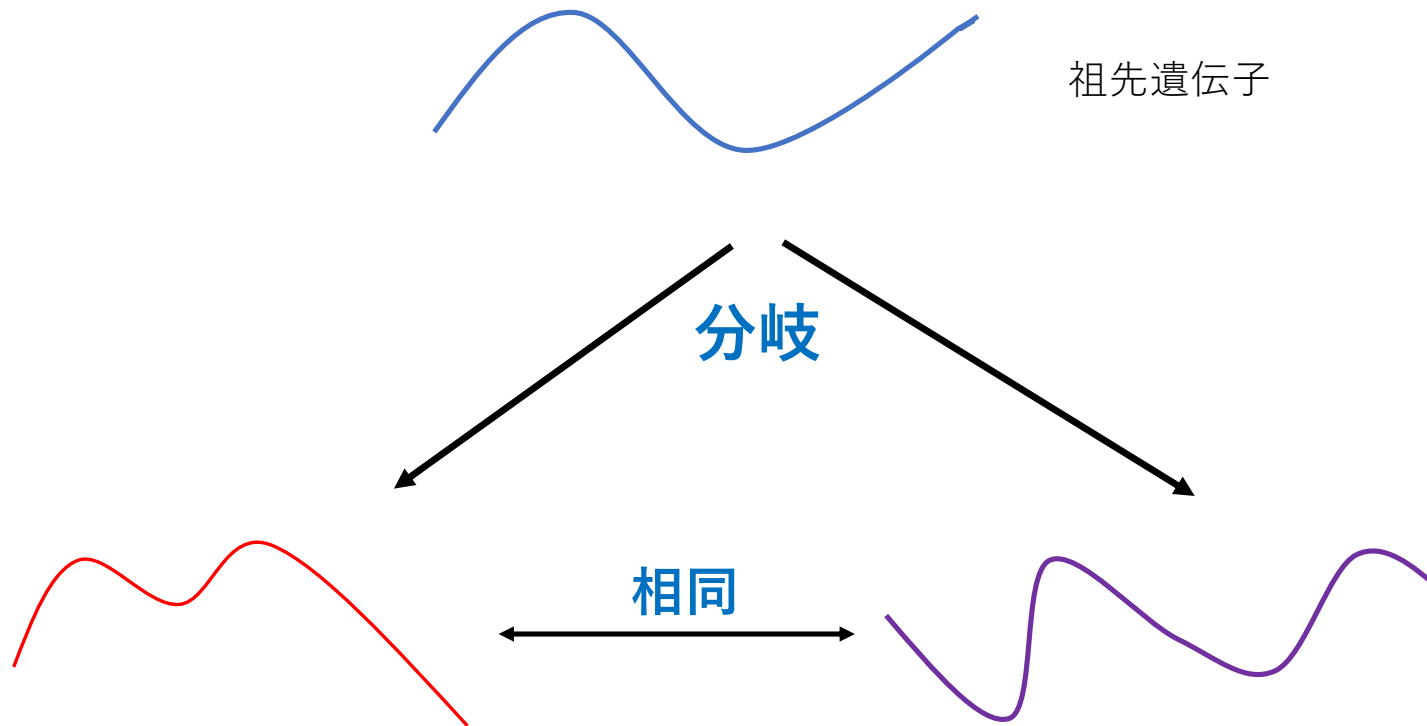
STRING is part of the ELIXIR infrastructure: it is one of ELIXIR's Core Data Resources. [Learn more >](#)

進化的視点からのタンパク質の機能解析

相同タンパク質の比較解析

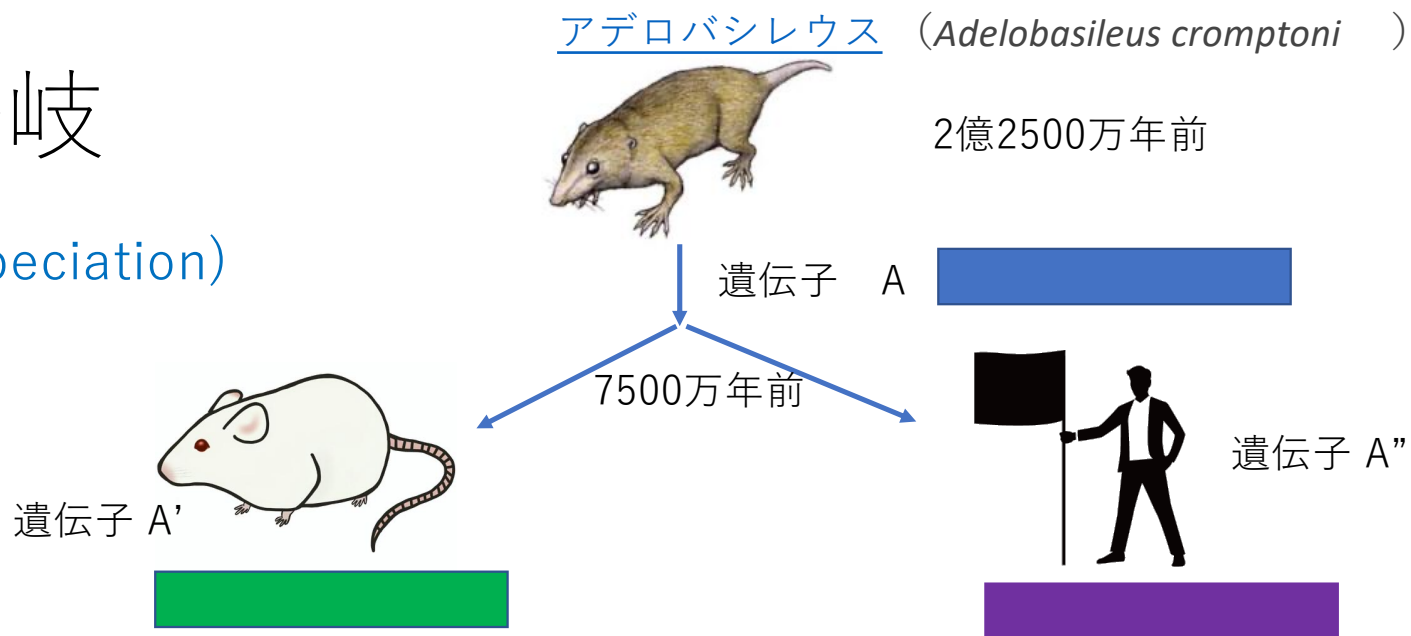
相同性 (homology)

共通の起源から分岐してきた遺伝子間の関係性



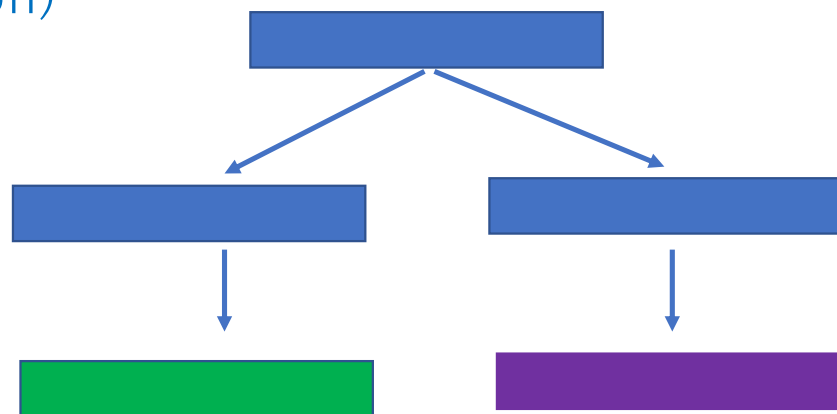
2 種類の分岐

(1) 種分化 (speciation)



(2) 遺伝子重複 (gene duplication)

重複した時には同じコピー



分子進化 (molecular evolution)

進化の過程で配列が変化すること

向上、複雑化などの意味は持たない
(分子進化だけでなく、進化自体も同じ)

分子進化の基本ステップ

点変異・挿入変異・欠失変異

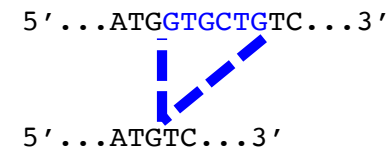
(a) 点変異 (point mutation)



(b) 挿入変異 (insertion)



(c) 欠失変異 (deletion)



種分化に伴う相同タンパク質の形成

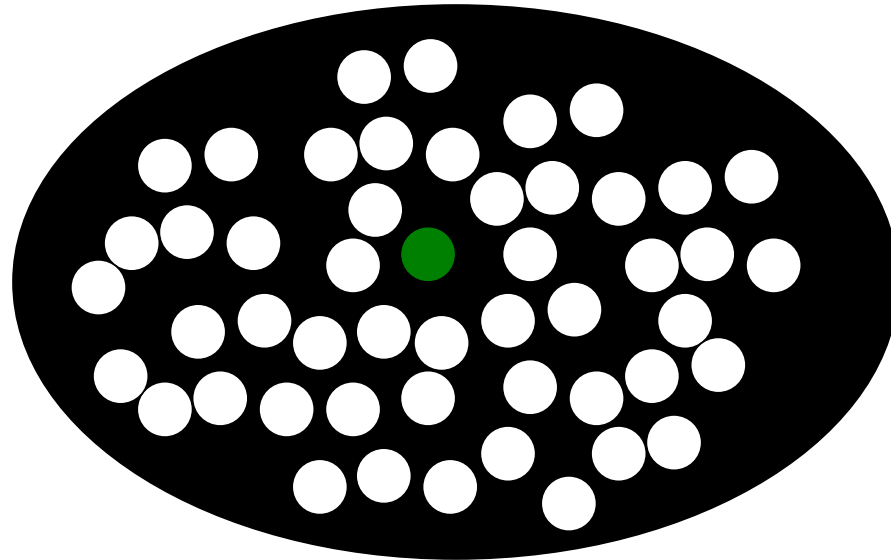
種分化

	10	20	30	40	50
human	VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH				
	::: ::				
mouse	VLSGEDKSNIKAAWGKIGHGAIEYGAEALERMFAISFPTTKTYFPHFDVSH				
	10	20	30	40	50
	60	70	80	90	100
human	GSAQVKGHGKKVADALTNVAHVDDMPNALSALSDDLHAHKLRVDPVNFKL				
	::				
mouse	GSAQVKGHGKKVADALASAAGHLDDLPGALSALSDDLHAHKLRVDPVNFKL				
	60	70	80	90	100
	110	120	130	140	
human	LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR				
	::::::::::::: ::				
mouse	LSHCLLVTLASHPADFTPAVHASLDKFLASVSTVLTSKYR				
	110	120	130	140	20/141



生物の種分化を反映したタンパク質の増加

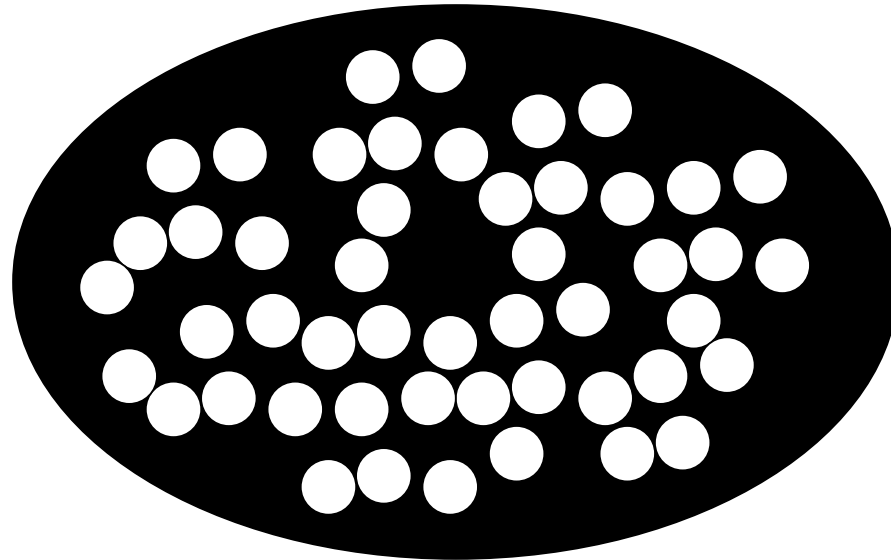
突然変異(mutation)と置換 (substitution)



突然変異は集団中の個体に生じる

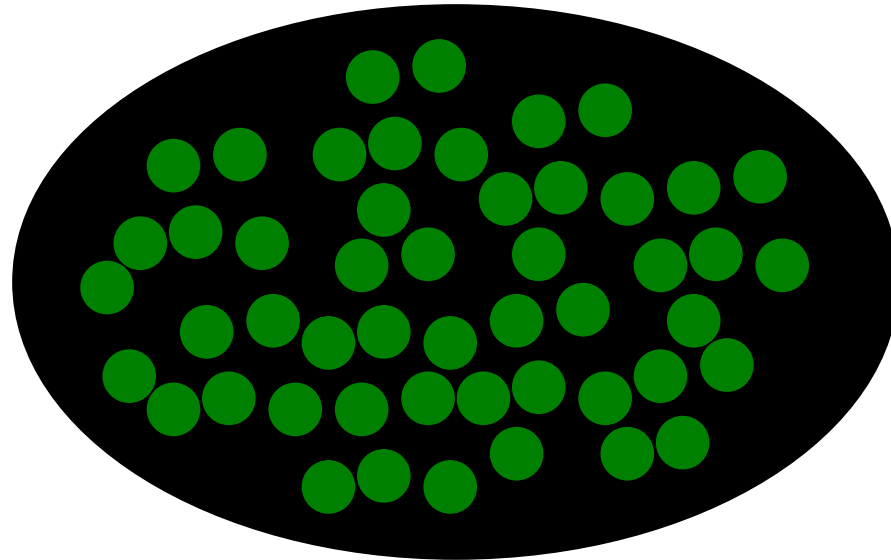
※ 進化に寄与するのは体細胞ではなく生殖系列
の細胞に生じる突然変異

突然変異(mutation)と置換 (substitution)



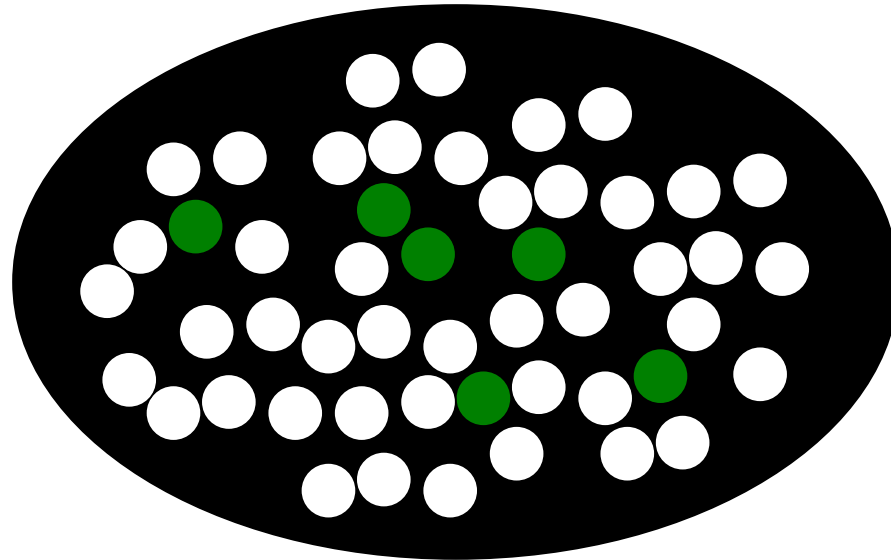
有害な突然変異は、集団から除去される
(負の選択 or 純化淘汰)

突然変異(mutation)と置換 (substitution)



有利な突然変異は急速に集団中に広まり
集団全体がその突然変異遺伝子で置き換
えられる (正の選択)

突然変異(mutation)と置換 (substitution)



中立な突然変異の頻度はランダムに浮動し
確率的に集団中に固定 (選択的に中立)

固定にいたる途中の過程 多型(polymorphism)

タンパク質をコードする遺伝子における
分子レベルの変化の大部分は中立的



中立的な変化

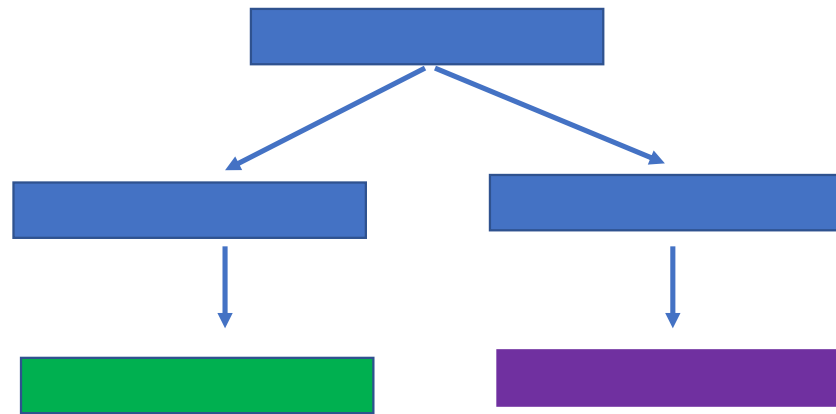
自然選択

遺伝子重複に伴う相同タンパク質の形成

遺伝子重複

古典的な遺伝子重複の描像

重複した時には同じコピー



多くの変異は、**偽遺伝子**を生じ、まれに**新機能を持った遺伝子**が生じる

タンパク質が新たな機能を獲得する進化機構

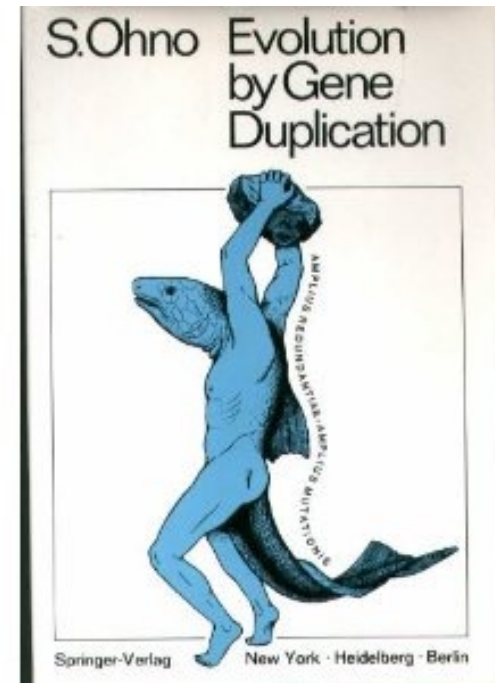
2R WGD 仮説

2 round whole genome duplication

脊椎動物のゲノムは進化の初期の段階で2回の全ゲノム重複を起こしたという仮説

大野乾により「Evolution by Gene Duplication」

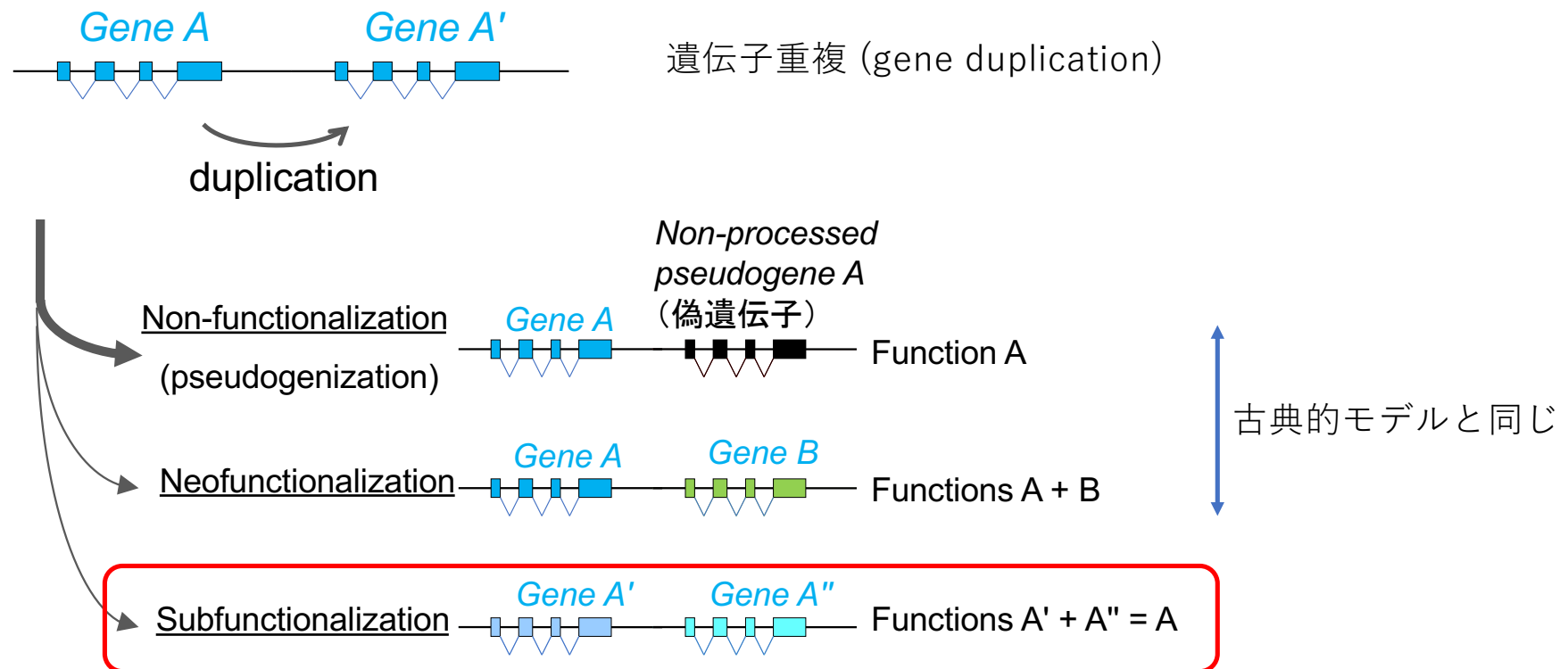
Susumu Ohno
(1928 - 2000)



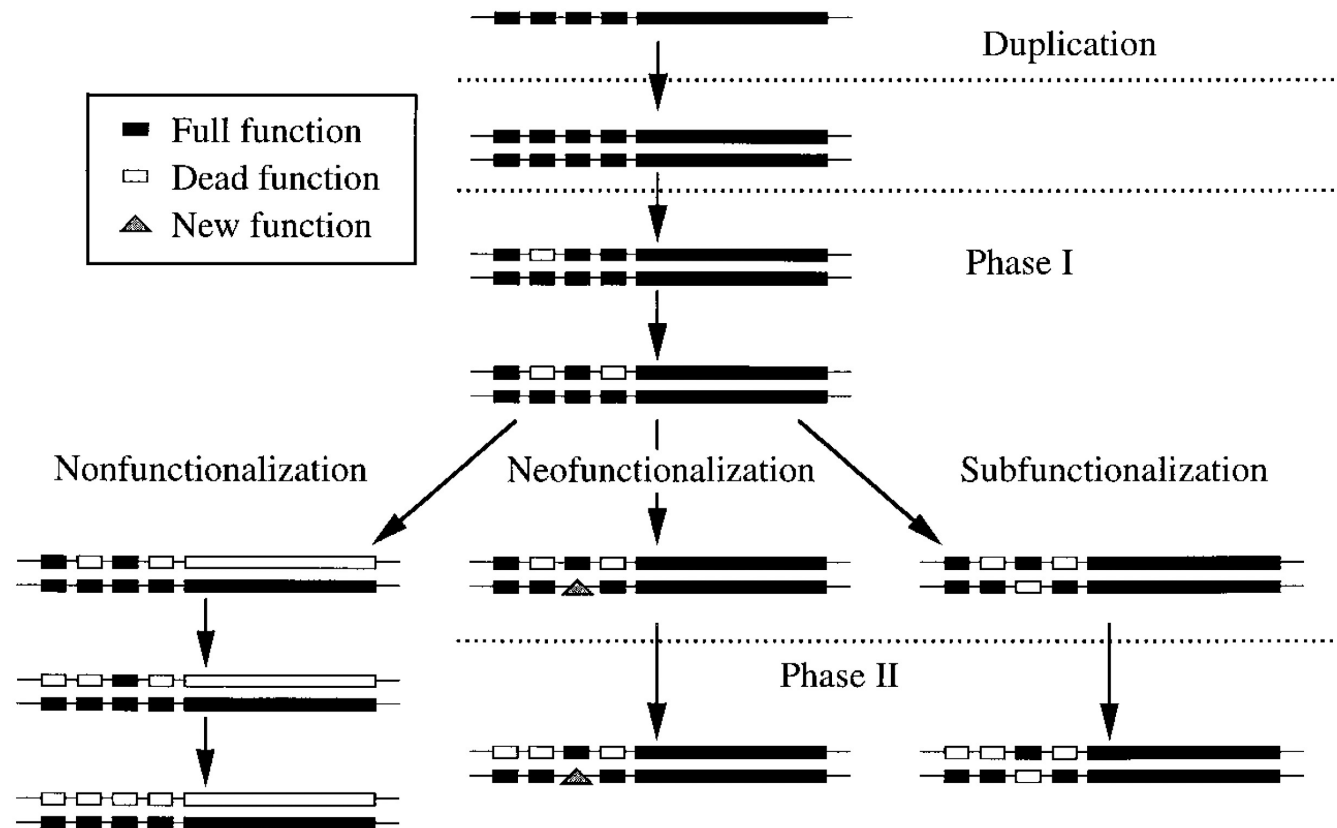
Evolution by Gene Duplication (1970)

DDC (duplication, degeneration, complementation) model

3つのシナリオよりなる

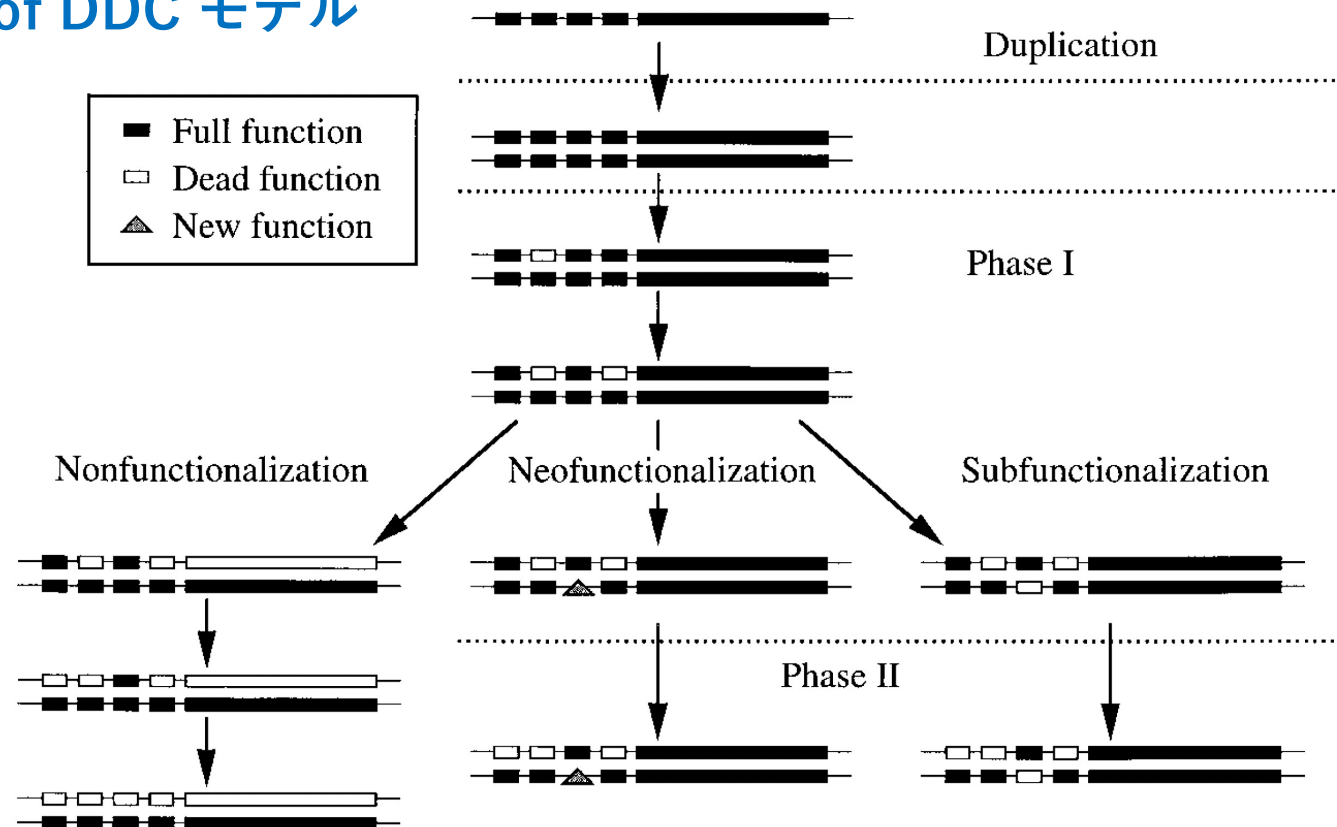


Phase I of DDC モデル



小さなボックスが発現制御領域を表す。大きなボックスはコード領域を表す。
Phase Iにおいて、subfunctionalizationでは、制御領域のいくつか壊れ、重複したコピーは相補的な発現パターンを示し、1コピーの時の機能を維持できる。
Neofunctionalizationにおいても、制御領域が変化することが知られてている。

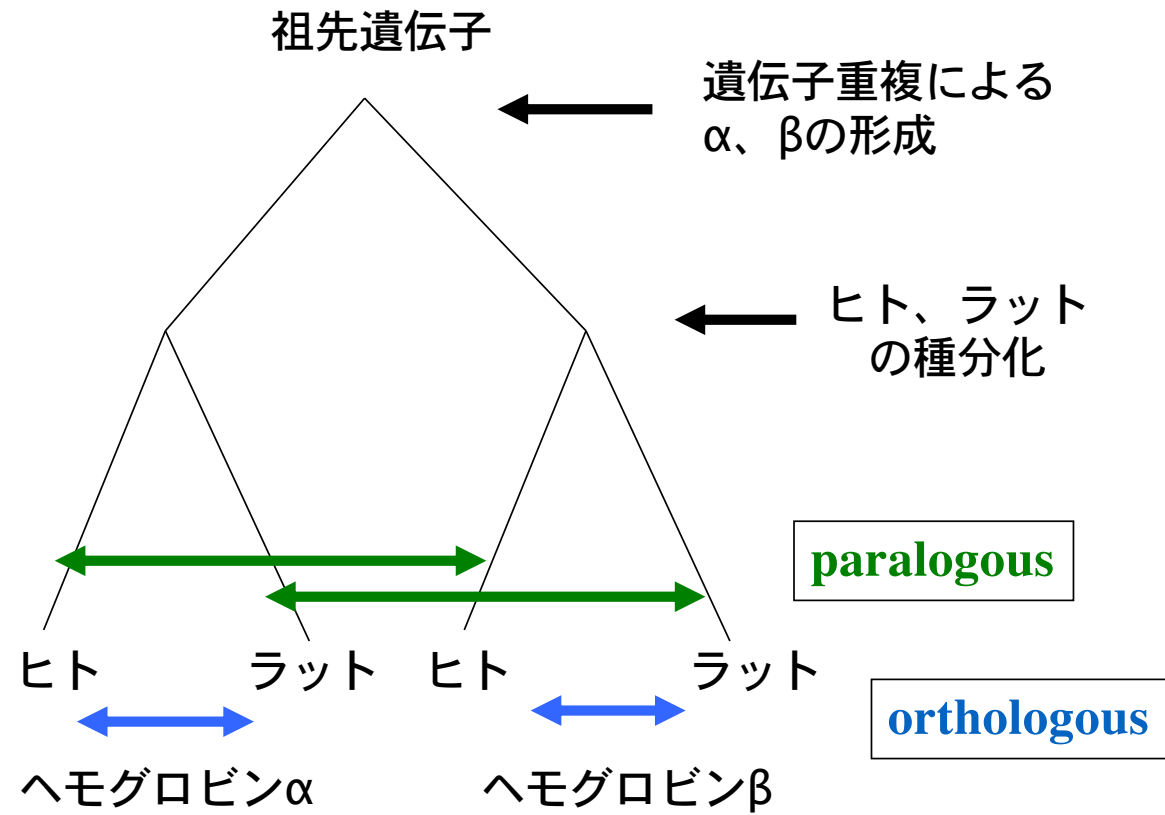
Phase II of DDC モデル



Force *et al.* (1999)

Neofunctionalization また subfunctionalization で生じたコピーの持つ制御領域の中の冗長な領域が壊れ、発現パターンの分化が進む。

Pseudogenization は、Phase II には関係しない



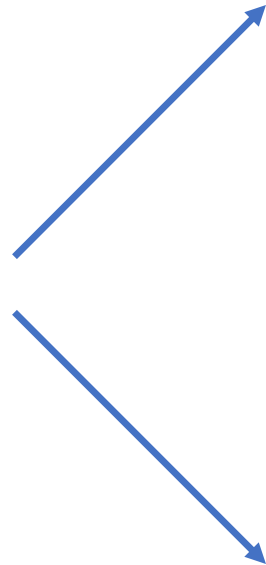
相同
(homologous)

オーソログス (orthologous)

種分化に伴い分岐した相同な関係

パラログス (paralogous)

遺伝子重複によって分岐した相同な関係



何故、相同タンパク質の解析をするのか？

“A major objective of the analysis of protein sequences is to discover which share a common ancestor. **Determining an evolutionary relationship is useful** because it generally implies **a common or similar function.**”

Hubbard,T.J.P. (1997) *Current Opinion in Str. Biol.* 7, 190-193.

相同配列の比較解析の手順

第一ステップ：相同配列の収集



第二ステップ：相同配列のマルチプルアラインメント



第三ステップ：アラインメントからの機能・構造・進化的
情報の抽出

相同配列の比較解析の手順

第一ステップ：相同配列の収集

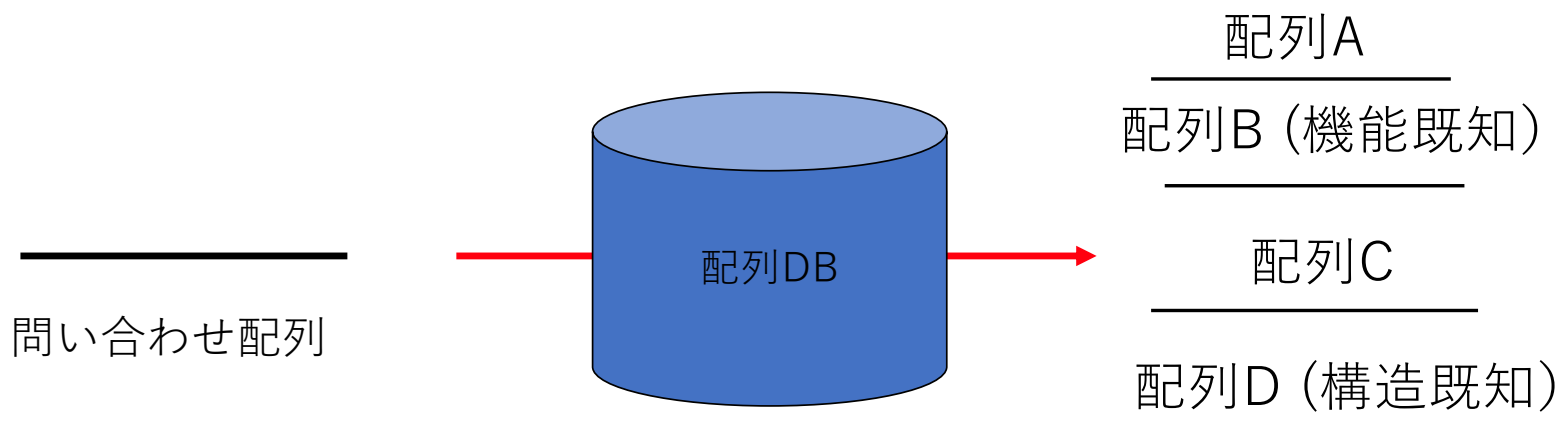


第二ステップ：相同配列のマルチプルアラインメント



**第三ステップ：アラインメントからの機能・構造・進化的
情報の抽出**

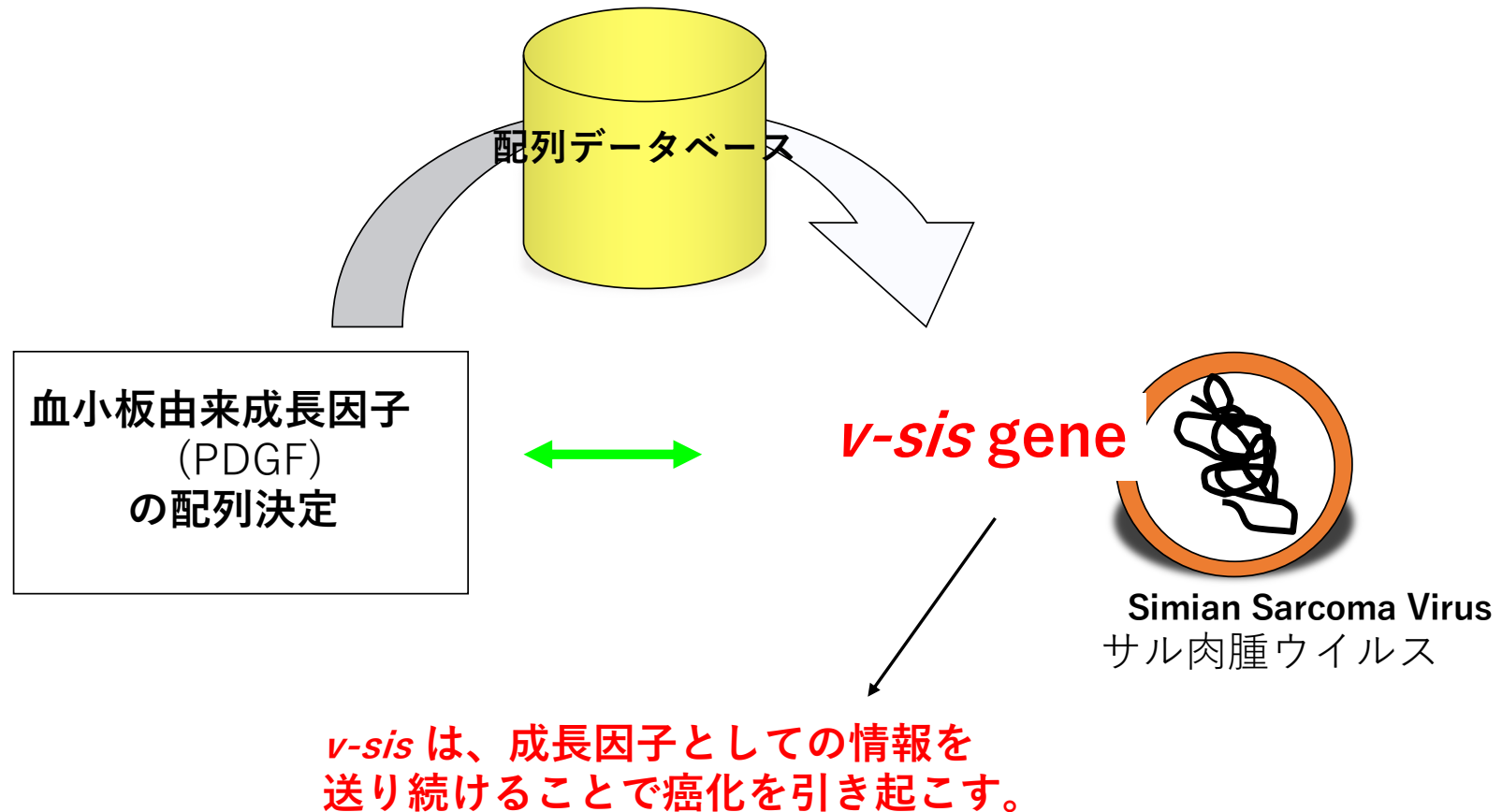
データベース検索



問い合わせ配列に類似した
配列を配列DBの中から検索

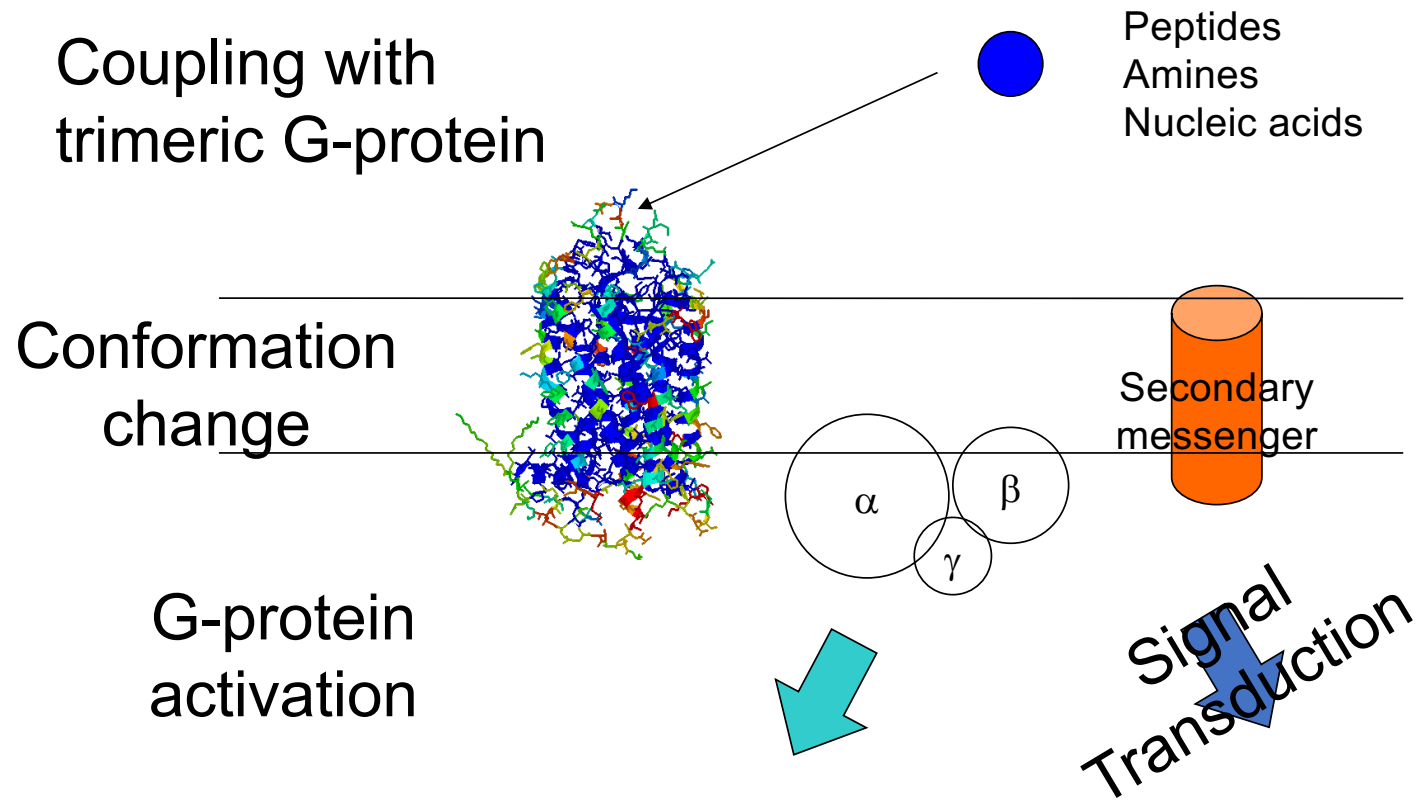
類似配列の中で、構造や機能既知の
ものがあれば、それらと同様の構造
あるいは機能を有するものと推測

発癌遺伝子 *v-sis*



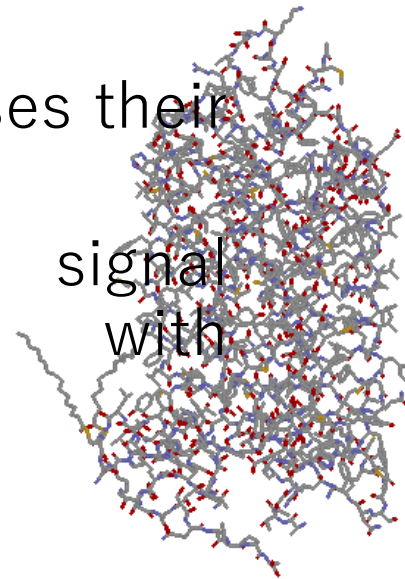
GPCRs

Monomer function



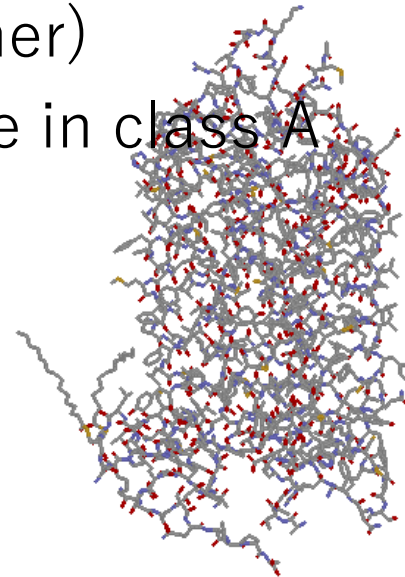
GPCRs

- Membrane proteins
- Bind neurotransmitters (physiologically active peptides, amines, nucleic acids, etc).
- Ligand binding to GPCRs causes their conformation changes.
- It leads to several transductions conjugated trimeric G-proteins.



GPCRs

- About 1000 genes in human genome
- Target for ~45% of clinically marketed drugs
- Divided into 5 classes based on sequence similarity (Class A-E, the other)
- Atomically resolved structure in class A
GPCR: Bovine Rhodopsin



既知のGPCRの配列

ゲノムにコード
されているタンパク質
との類似性を検索

ヒト・ゲノム
データベース

既知のGPCRの配列



新規のGPCRの配列

創薬ターゲット
としての検討

GPCR以外のターゲット探索の例

中外製薬(株)

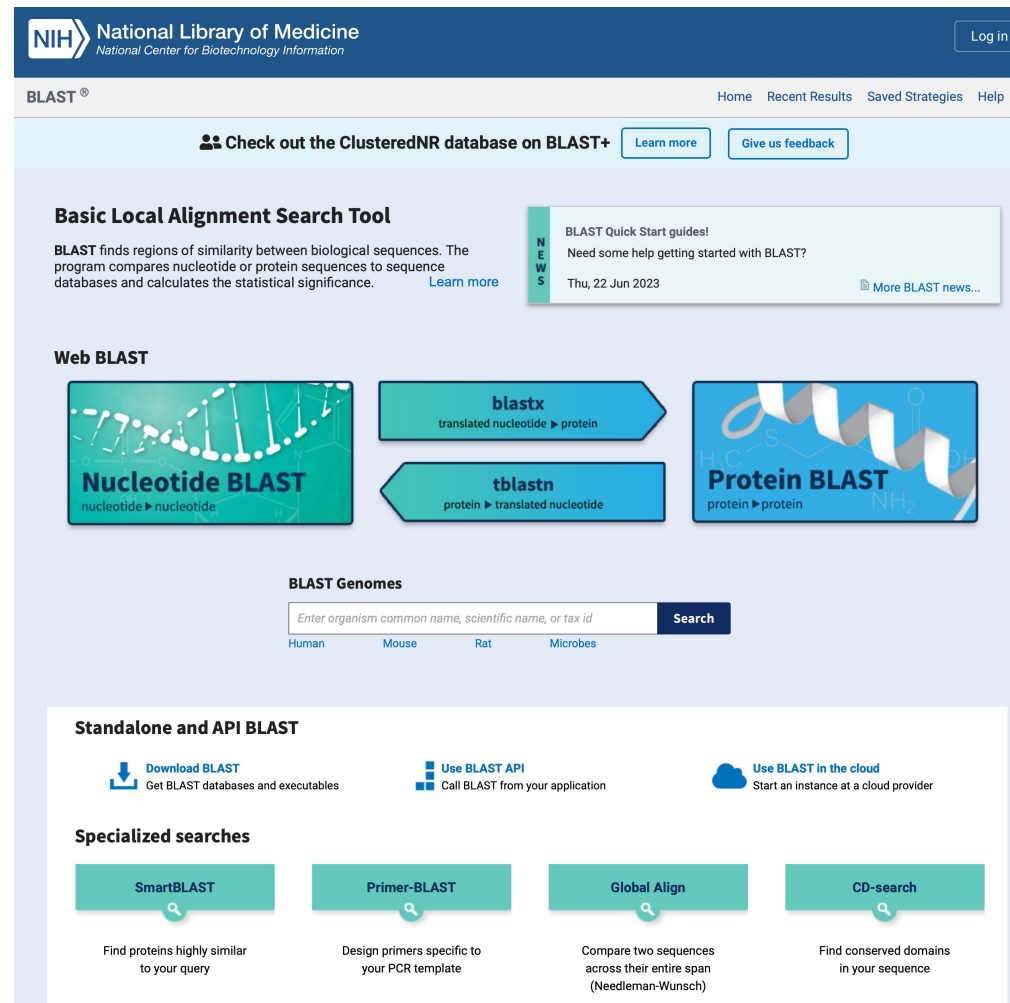
中外分子医学研究所

ドラッグデザインのターゲットとしての
新規サイトカイン及び新規サイトカイン受容体
のゲノム配列、EST配列からの探索

「ゲノム創薬」 個別化医療とゲノムデータマイニング
野村仁 著 サイエンス社 (2005)

BLAST (*Basic Local Alignment Search Tool*)

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>



The screenshot shows the NCBI BLAST homepage. At the top is the NIH logo and 'National Library of Medicine' text. Below this is a navigation bar with 'BLAST®', 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. A 'Log in' button is in the top right. A banner for 'ClusteredNR database on BLAST+' is present. The main section is titled 'Basic Local Alignment Search Tool' and describes its function. To the right is a 'NEWS' sidebar with a date and a link to 'More BLAST news...'. Below this is the 'Web BLAST' section with three main tools: 'Nucleotide BLAST' (nucleotide to nucleotide), 'blastx' (translated nucleotide to protein), and 'tblastn' (protein to translated nucleotide). There is also a 'Protein BLAST' tool (protein to protein). Below these is the 'BLAST Genomes' section with a search bar and buttons for 'Human', 'Mouse', 'Rat', and 'Microbes'. The 'Standalone and API BLAST' section offers three options: 'Download BLAST', 'Use BLAST API', and 'Use BLAST in the cloud'. The 'Specialized searches' section includes four tools: 'SmartBLAST', 'Primer-BLAST', 'Global Align', and 'CD-search', each with a brief description of its purpose.

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® Home Recent Results Saved Strategies Help

Check out the ClusteredNR database on BLAST+ [Learn more](#) [Give us feedback](#)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS
BLAST Quick Start guides!
Need some help getting started with BLAST?
Thu, 22 Jun 2023 [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

BLAST Genomes

Enter organism common name, scientific name, or tax id [Search](#)

Human Mouse Rat Microbes

Standalone and API BLAST

[Download BLAST](#)
Get BLAST databases and executables

[Use BLAST API](#)
Call BLAST from your application

[Use BLAST in the cloud](#)
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

相同配列の比較解析の手順

第一ステップ：相同配列の収集



第二ステップ：相同配列のマルチプルアラインメント



第三ステップ：アラインメントからの機能・構造・進化的
情報の抽出

2. 配列アラインメント (sequence alignment)

進化の過程でのINDELを考慮しながら、相同な配列の間の対応する塩基（あるいはアミノ酸）を同じ位置に並べる操作あるいは、その操作によってできたもの。

INDELに対応してギャップ(gap)とよばれる空記号を挿入し位置をずらして、塩基やアミノ酸を対応づける。

通常、動的計画法(dynamic programming algorithm)や、そのバリエーションを用いて、配列間の類似度が高くとなるようにアラインメントが構築される。

アラインメントの原理は次回

マルチプルアラインメント (multiple alignment)

分子系統樹を構築するには、まず複数本の相同配列についてのマルチプルアラインメントを作成する。

リゾチームのアミノ酸配列

>LYC_HUMAN ヒト Lysozyme C

MKALIVLGLVLLSVTVQGVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRST
DYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVR
QYVQGCGV

>LYC1_BOVIN ウシ Lysozyme C 1

MKALIILGFLFLSVAVQGVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESSYNTKATNYPGSEST
DYGIFQINSKWWCNDGKTPNAVDGCHVSCSELMENDIAKAVACAKQIVSEQGITAWVAWKSHCRDHDVSS
YVEGCTL

>LYC_CHICK ニワトリ Lysozyme C

MRSLLIILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTD
YGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNC AKKIVSDGNGMNAWVAWRNRCKGTDVQA
WIRGCRL

>LYC2_ONCMY マス Lysozyme C II

MRAVVVLLLVAVASAKVYDRCELARALKASGMDGYAGNSLPNWWCLSKWESSYNTQATNRNTDGSTDYGI
FQINSRYWCDDGRTPGAKNVCGIRCSQLLTADLTVAIRCAKRVVLDPNGIGAWVAWRLHCQNQDLRSYVA
GCGV

>LYC_BOMMO カイコ Lysozyme

MQKLIIFALVVLVCGSEAKTFTRCGLVHELKRKHGFEENLMRNWVCLVEHESSRDTSKTNTNRNGSKDYGL
FQINDRYWCSKGASPGKDCNVKCSDLLTDDITKA AKCAKKIYKRHRFDAWYGWKNHCQGS L PDISSC

>LYSP_DROME ハエ Lysozyme P

MKAFLVICALTLTAVATQARTMDRCSLAREMSKLGVP RDQLAKWTCIAQHES SFRTGVVGPANSNGSNDY
GIFQINNKYWCKPADGRFSYNECGLSCNALLTDDITNSVKARKIQRQQGWTAWSTWKYCSGSLPSINSC
F

マルチfasta形式

```
>gi|443546|pdb|7HVP|A Chain A, Protease Of Human Immunodeficiency Virus 1
PQITLWQRPLVTIRIGGQLKEALLDTGADDTVLEEMNLPKGWKPKMIGGIGGFIKVRQYDQIPVEIXGHK
AIGTVLVGPTPVNIIGRNLLTQIGXTLNF

> HIV2 protease (NP_663784 REGION: 523..603)
    vtayiedqpv evlldtgadd sivagielgd nytpkivggi ggfintkeyk nveikvlnkr
    vratimtgdt pinifgrnil t

> simian immunodeficiency virus SIV-mnd 2 (NP_758887 REGION: 72..166)
    slwnrpttvv eiegqkveal ldtgadtdvi kdldlkgnwk pqiiggiggs invkqffnck
    vtiagkttha svlvgptpvn ivgrnvlkkl gctl n

>gi|4389337|pdb|1BAI|A Chain A, Rous Sarcoma Virus Protease
LAMTMEHKDRPLVRVILTNTGSHPVKQRSVYITALLDTGADDTVISEEDWPTDWPVMEAAANPQIHGIGGG
IPVRKSRDMIELGVINRDGSLERPLLLFPLVAMTPVNILGRDCLQGLGLRLTNL

>gi|224443|prf||1104339A MoMuLV protease
TLDDQGGQGQEPPEPRITLKVGGQPVTFLVDTGAQHSVLTQNPGPLSDKSAWVQGATGGKRYRWTTDRK
VHLATGKVTHSFLHVPDCPYPLLGRDLLTKLKAQIHFEFGSGAQVMGPMGQPLQVL
```

1. 配列は、スペースがあいていても構わないことに注意
2. Mafftは大文字も、小文字も区別せずに配列を処理してくれるが、ソフトによっては大文字あるいは小文字しか処理できない場合があるかもしれない

MAFFTによるマルチプルアライメント

mafftは宮田研究室で開発され、加藤和貴によって継続的に開発されているマルチプルアライメントのフリーソフトウェア

海外の多くの研究機関で利用されている。

Web上でのアライメントサービスに加え、ダウンロードして自身のPC上で利用できる。Mac, Windows, Linuxなど様々なOSに対応している

ここでは、web serviceとして公開されているmafftを利用する。

<https://mafft.cbrc.jp/alignment/software/>

リゾチームのアミノ酸配列 マルチプル・アラインメント：Clustal形式

ヒト
ウシ
マス
ニワトリ
カイコ
ハエ

CLUSTAL W (1.81) multiple sequence alignment

```

LYC_HUMAN      -MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR
LYC1_BOVIN     -MKALIILGFLFLSVAVQGVFERCELARTLKRLGMDGYRGVSLANWLCLTKWESSYNTK
LYC2_ONCMY     ----MRVVVLLLVAVASAKVYDRCELARALKASGMDGYAGNSLPNWVCLSKWESSYNTQ
LYC_CHICK      -MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQ
LYC_BOMMO      -MQKLIIFALVVLVGVSEAKTFTRCGLVHELKRGHGFENL---MRNWVCLVEHESSRDTS
LYSP_DROME     MKAFLVICALTLTAVATQARTMDRCSLAREMSKLGVPDQ---LAKWTCIAQHESFRTG
               :      . .      . . .  * * * . :      * .      : : * * : ** . *

LYC_HUMAN      ATNYNAGDRSTDYGIFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVR
LYC1_BOVIN     ATNYPGSESTDYGIFQINSKWWCNDGKTPNAVDGCHVSCSELMENDIAKAVACAKQIVS
LYC2_ONCMY     ATNRN-TDGSTDYGIFQINSRYWCDDGRTPGAKNVCGIRCSQLLTADLTVAIRCAKRVVL
LYC_CHICK      ATNRN-TDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNC AKKIVS
LYC_BOMMO      KTNTN-RNGSKDYGLFQINDRYWCS--KGASPGKDCNVKCSDLLTDDITKA AKCAKKIYK
LYSP_DROME     VVG PANSNGSNDYGIFQINNKYWCKPADGRFSYNECGLSCNALLTDDITNSVKCARKIQR
               ..      . * .***::***.::***.      . . * : * . * : : : : **:::

LYC_HUMAN      DPQGIRAWVAWRNRCQNRDVRQYVQGC GV
LYC1_BOVIN     E-QGITAWVAWKSHCRDHDVSSYVEGCTL
LYC2_ONCMY     DPNGIGAWVAWRLHCQNQDLRSYVAGCGV
LYC_CHICK      DGNGMNAWVAWRNRCKGTDVQAWIRGCR L
LYC_BOMMO      R-HRFD AWYGWKNHCQGS--LPDISSC--
LYSP_DROME     Q-QGTAWSTWK-YCSGS--LPSINS CF-
               :      ** * : * .      : . *
    
```

相同配列の比較解析の手順

第一ステップ：相同配列の収集

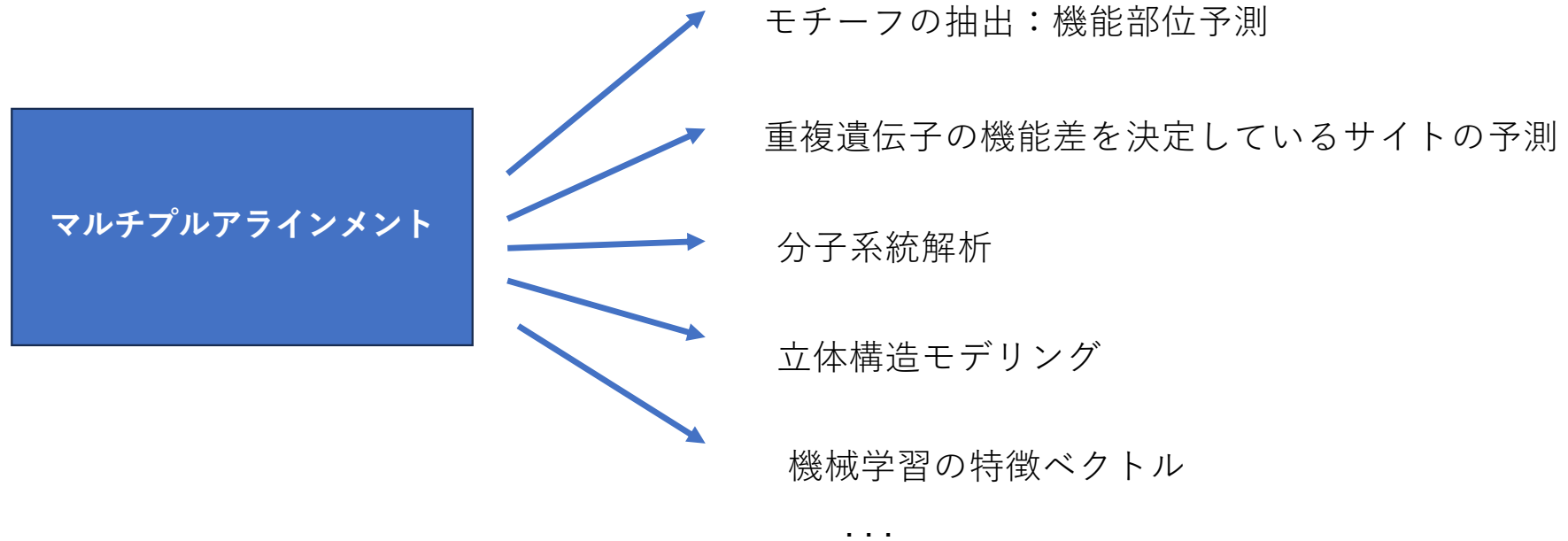


第二ステップ：相同配列のマルチプルアラインメント



第三ステップ：アラインメントからの機能・構造・進化的
情報の抽出

第三ステップ：アラインメントからの 機能・構造・進化的情報の抽出



ヒト
ウシ
マス
ニワトリ
カイコ
ハエ

LYC_HUMAN	-MKALIVLGLVLLSVTVQGVFERCELARTLKLGMGDGYRGISLANWMCLAKWESGYNTR
LYC1_BOVIN	-MKALIILGFLFLSVAVQGVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESSYNTK
LYC2_ONCMY	----MRVVVLLLLVAVASAKVYDRCELARALKASGMDGYAGNSLPNVVCLSKWESSYNTQ
LYC_CHICK	-MRSLLILVLCFLPLAALGKVFGRCELAAMKRHGLDNYRGYSLGNVWVCAAKFESNFNTQ
LYC_BOMMO	-MQKLIIFALVVLVCGSEAKTFTRCGLVHELKRGHGFENL---MRNVVCLVEHESSRDT
LYSP_DROME	MKAFLVICALTLTAVATQARTMDRCSLAREMSKLGVPDQ---LAKWTCIAQHESSEFRTG
	: . . .: ** . : * . : : * * : ** . *
LYC_HUMAN	ATNYNAGDRSTDYGFQINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVR
LYC1_BOVIN	ATNYPGSESTDYGFQINSKWWCNDGKTPNAVDGCHVSCSELMENDIAKAVACAKQIVS
LYC2_ONCMY	ATNRN-TDGSTDYGFQINSRYWCDDGRTPGAKNVCGIRCSQLLTADLTVAIRCAKRVVL
LYC_CHICK	ATNRN-TDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCACKIIVS
LYC_BOMMO	KTNTN-RNGSKDYGFLQINDRYWCS--KGASPGKDCNVKCSDLLTDDITKAAKCAKKIYK
LYSP_DROME	VVGPNANSNGSNDYGFQINNKYWCKPADGRFSYNECGLSCNALLTDDITNSVKCARKIQR
	. . . * .***: :***. : ** . . . * : * . * : : : : ** : :
LYC_HUMAN	DPQGIRAWVAWRNRCQNRDVRQYVQGC
LYC1_BOVIN	E-QGITAWVAWKSHCRDHDVSSYVEGCTL
LYC2_ONCMY	DPNGIGAWVAWRLHCQNQDLRSYVAGCGV
LYC_CHICK	DGNMGNAWVAWRNRCKGTDVQAWIRGCRL
LYC_BOMMO	R-HRFDAWYGWKNHCQGS--LPDISSC--
LYSP_DROME	Q-QGWTAWSTWK-YCSGS--LPSINSCF-
	: ** * : * . : . *

Clustal形式アラインメント下段のシンボルの意味

“*”では、完全に保存

“:”では、強い物理化学的類似性のあるグループで保存

“.”では、弱い類似性のあるグループで保存

強い弱いの基準は、PAM250 行列において、アミノ酸間のスコアが0.5より大きいか、0.5以下かで分けている

スコア・テーブル

多くのアミノ酸置換は保存的

Dayhoffのスコアテーブル

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

- Dayhoffコード
- (1) G,A,S, T, P: small hydrophilic residues
 - (2) L, I, M, V: hydrophobic residues
 - (3) D, N, E, Q: negatively charged residues and the relatives
 - (4) F, Y, W: aromatic residues
 - (5) K, R, H: positively charged residues
 - (6) C: Cys

Multiple Alignment of Lysozyme Amino Acid Sequences

Human
Bovine
Trout
Chicken
Silkworm
Fruitfly

CLUSTAL W (1.81) multiple sequence alignment

```

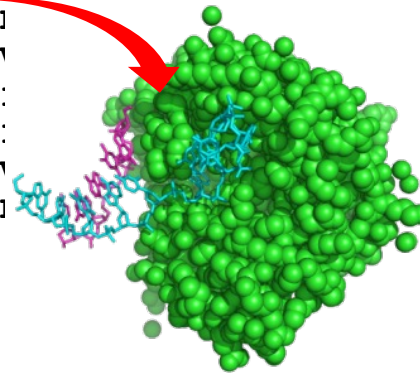
LYC_HUMAN      -MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTR
LYC1_BOVIN     -MKALIILGFLFLSVAVQGKVFERCELARTLKRLGMDGYRGISLANWLCITKWESSYNTK
LYC2_ONCMY     ----MRAVVVLLLVAVASAKVYDRCELARALKASGMDGYAGNSLPNWVCLSKWESSYNTQ
LYC_CHICK      -MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNRYGYSLGNNVCAAKFESNFNTQ
LYC_BOMMO      -MQKLIIFALVVLVGVGSEAKTFTRCGLVHELKRGHGFENL---MRNWVCLVEHESSRDTS
LYSP_DROME     MKAFLVICALTLTAVATQARTMDRCSLAREMSKLGVPDQ---LAKWTCIAQHESFRTG
               :   . .       .:.  ** *.   :   *.       : : * *   : **.  *
    
```

```

LYC_HUMAN      ATNYNAGDRSTDYGIFQINSRYWCNDGKTPCAVNACH
LYC1_BOVIN     ATNYPGSESTDYGIFQINSKWWCNDGKTPNAVDGCH
LYC2_ONCMY     ATNRN-TDGSTDYGIFQINSRYWCDDGRTPGAKNVCG
LYC_CHICK      ATNRN-TDGSTDYGILQINSRWWCNDGRTPGSRNLCH
LYC_BOMMO      KTNTN-RNGSKDYGLFQINDRYWCS--KGASPGKDCN
LYSP_DROME     VVGPA NSNGSNDYGIFQINNKYWCKPADGRFSYNECG
               ..   .  *.***::***.::*.   . .  *
    
```

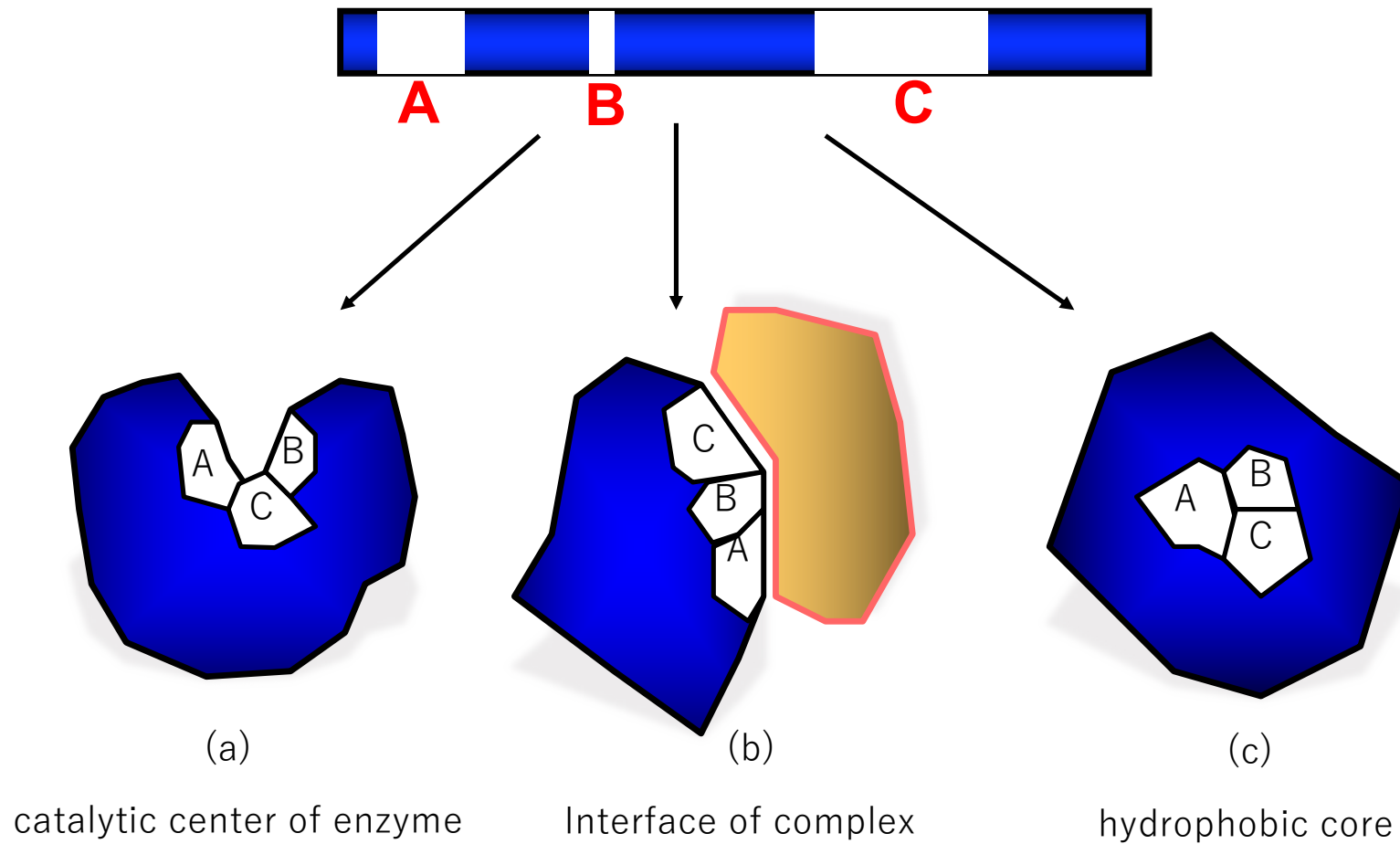
```

LYC_HUMAN      DPQGI RAWVAWRNRCQNRDVRQYVQGC GV
LYC1_BOVIN     E-QGITAWVAWKSHCRDHDVSSYVEGCTL
LYC2_ONCMY     DPNGIGAWVAWRLHCNQDLRSYVAGCGV
LYC_CHICK      DGNGMNAWVAWRNRCKGTDVQAWIRGCRL
LYC_BOMMO      R-HRFD AWYGWKNHCQGS--LPDISSC--
LYSP_DROME     Q-QGWTAWSTWK-YCSGS--LPSINSCF-
               :   **  *:  * .       :  .*
    
```



モチーフ: short stretch of conserved amino acid sequences

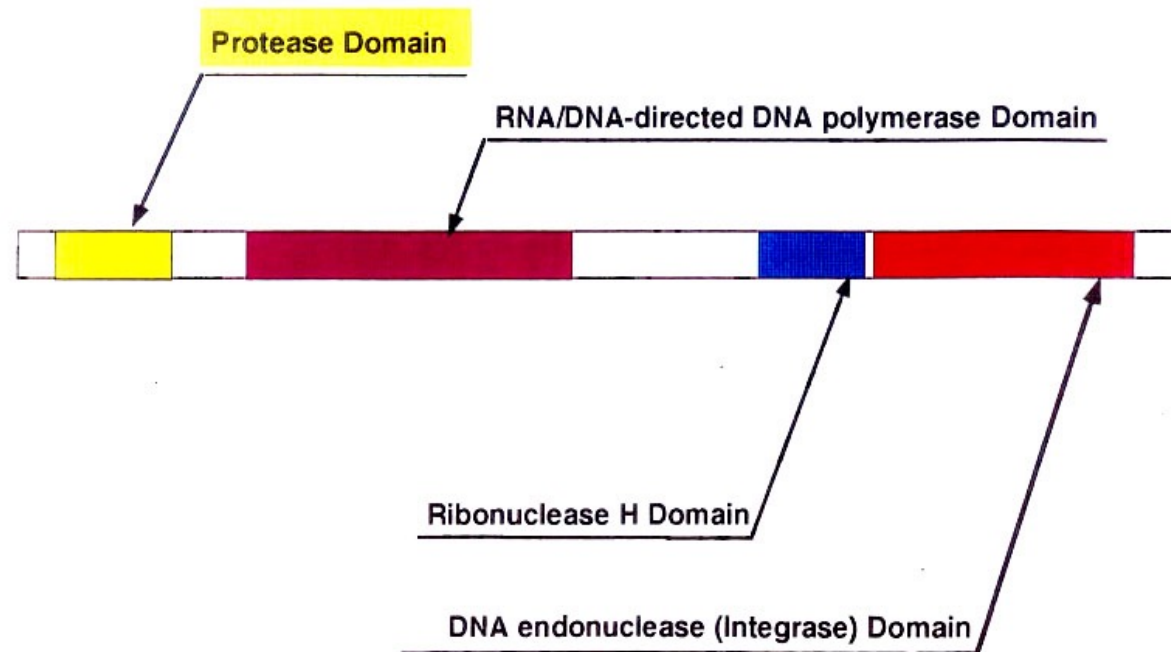
When several motifs are found by the sequence comparison, the sequences are closely located in the 3 dimensional structure.



Characterization of Retroviral Proteases

- Development of HIV Protease Inhibitor -

Retroviral Reverse Transcriptase



CLUSTAL format alignment by MAFFT L-INS-i (v7.130b)

```
gi|443546|pdb|7 PQITLW-----QRPLVTIRIGGQL-----KEALLDTGADDTVLEEMNLP  
HIV2              -----VTAYIEDQP-----VEVLLDTGADDSIVAGIELGD  
simian            ---SLW-----NRPTTVVEIEGQK-----VEALLDTGADDTVIKDLDLKG  
gi|4389337|pdb| LAMTMEHK-----DRPLVRVILTNTGSHPVKQRSVYITALLDTGADDTVISEEDWPT  
gi|224443|prf|   ---TLDDQGGQGQEPPEPRITLKVGGQP-----VTFLVDTGAQHSVLTQNPGPL
```

: .

*:****:..::

```
gi|443546|pdb|7 KW-----KPKMIGGIGGFIKVRQ---YDQIPVEIXGHKAIGTVL----VGPTPVNIIGR  
HIV2              NY-----TPKIVGGIGGFINTKE---YKNVEIKVLNKRVRATIM----TGDTPINIFGR  
simian            NW-----KPQIIGGIGGSINVKQ---FFNCKVTIAGKTTHASVL----VGPTPVNIVGR  
gi|4389337|pdb| DWPVMEAANPQ-IHGIGGGIPVRKSRDMELGVINRDGSLERPLLLFPLVAMTPVNILGR  
gi|224443|prf|   SD-----KSAWVQGATGGKRYRW---TTDRKVHLATGKVTHSFLH---VPDCPYPLLGR
```

. .. : * * : : : ..: . * :.**

```
gi|443546|pdb|7 NLLTQIGXTLN-----F  
HIV2              NILT-----  
simian            NVLKKLGCTLN-----  
gi|4389337|pdb| DCLQGLGLRLT-----NL  
gi|224443|prf|   DLLTKLKAQIHFEGSGAQVMGPMGQPLQVL
```

: *

レトロウイルス・プロテアーゼのコンセンサス配列

(Hydrophobic)-Asp-Thr-Gly-(Small Hydrophilic)

**Input: Retroviral
Protease Motif**

**Amino Acid Sequence
Database**

**Proteases with the Sequences
Similar to Retroviral Protease Motif**



Retroviral Protease

califlower mosaic virus	LHCFV-DT G ASLCIAS
copi-like element 17.6 fruit fly	LKCL I -DT G STVNMTS
Moloney murine leukaemia virus	VTFLV-DT G AQHSVLT
human T-cell leukaemia virus	IEALL-DT G ADMTVLP
Rous sarcoma virus	ITALL-DS G ADITIIS

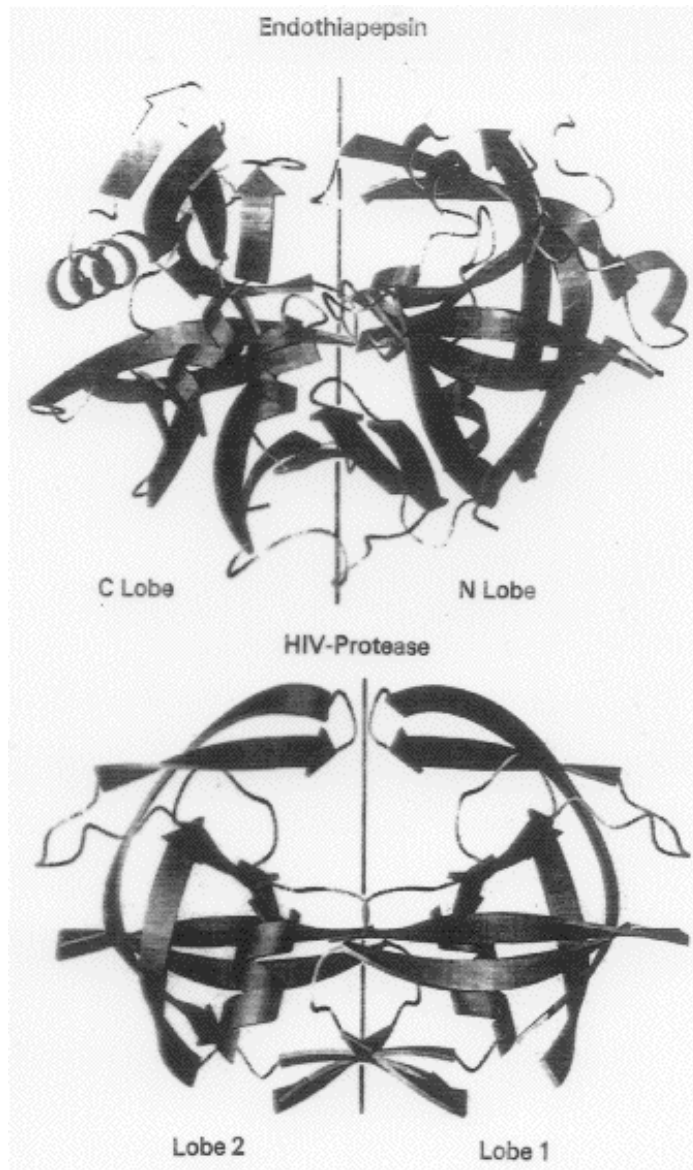
Aspartic Acid Protease (C-terminal domain)

pepsinogen (human)	CQAI V -DT G TSLLTGP
pepsinogen (pig)	CQAI V -DT G TSLLTGP
prochymosin (cow)	CQAIL-DT G TSKLVGP
penisilopepsin	FSGIA-DT G TTLLLLP
renin (mouse)	CEVV V -DT G SSFISAP
renin (human)	CLAL V -DT G ASYISGS

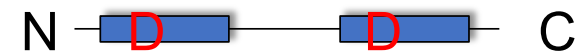
Aspartic Acid Protease (N-terminal domain)

pepsinogen (human)	DFT V FD T GS S NLWVP
pepsinogen (pig)	DFT V IFD T GS S NLWVP
prochymosin (cow)	EFT V LF D T G SS D FWVP
penusilopepsin	TLNLN F D T GS S DFWVP
renin (mouse)	TFK V MF D T G SANLWVP
renin (human)	TFK V FD T GS S NVWVP

モチーフ配列は、酸性プロテアーゼの活性中心に類似

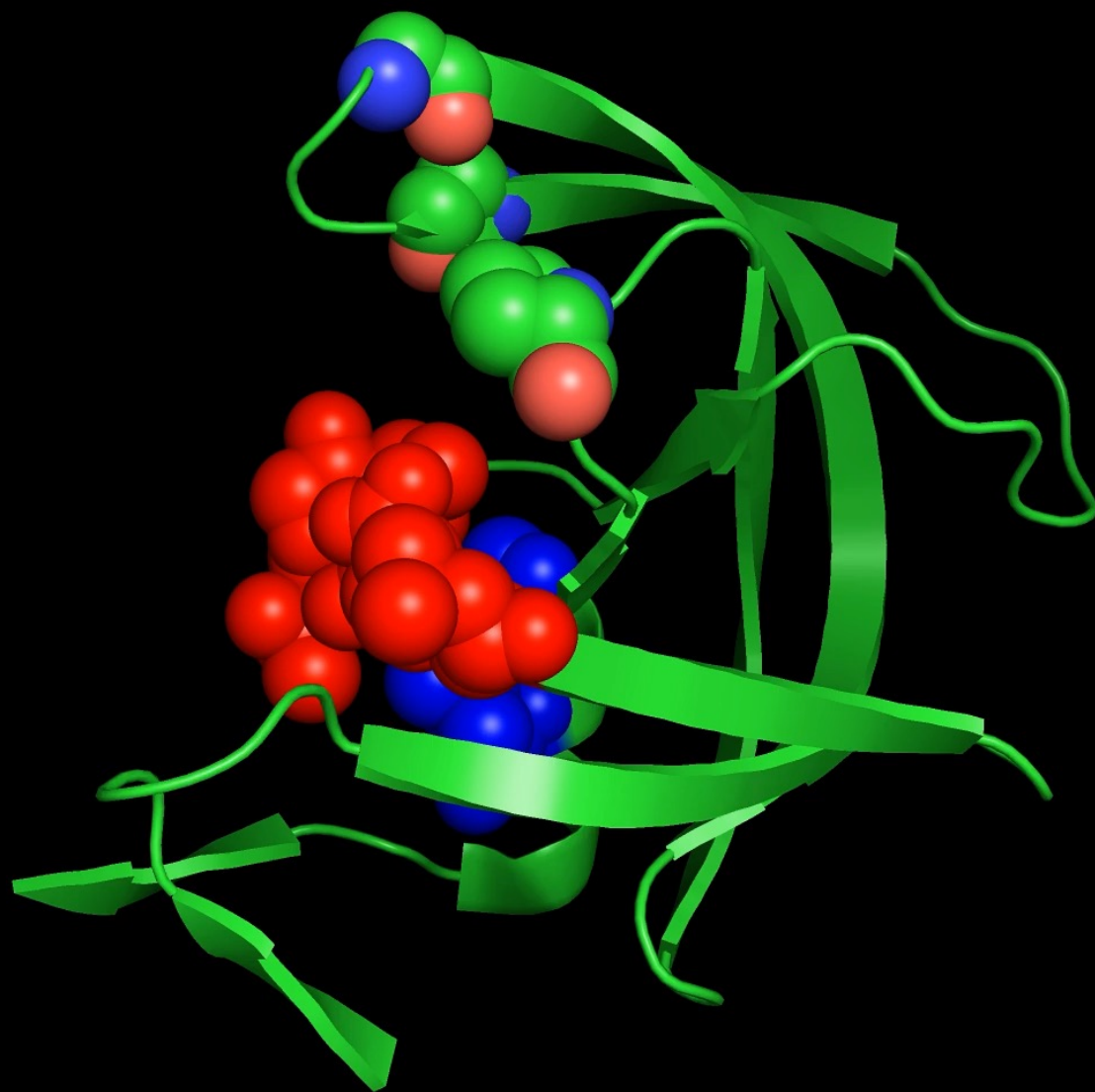


N末ドメインと
C末ドメインが
相同



ホモダイマー





予 測

レトロウイルスのプロテアーゼは酸性プロテアーゼである。

H. Toh et al. *EMBO J.* **4**, 1267 (1985)

H. Toh et al. *Nature* **315**, 691 (1985)

実 験

[1] 既知酸性プロテアーゼの立体構造を鋳型としたHIV プロテアーゼのホモロジー・モデリング
L.H. Pearl and W.R. Taylor, *Nature* **329**, 351 (1987).

[2] 酸性プロテアーゼ阻害剤（ペプスタチン）によるHIV プロテアーゼの阻害
R.F. Nutt et al. *Proc. Natl. Acad. Sci. USA* **85**, 7129 (1988).
P.L. Darke et al. *J. Biol. Chem.* **264**, 2307 (1989).

検 証

X 線結晶構造解析からHIV プロテアーゼは酸性プロテアーゼと類似の構造である。
M. Miller et al. *Science* **246**, 1149 (1989).

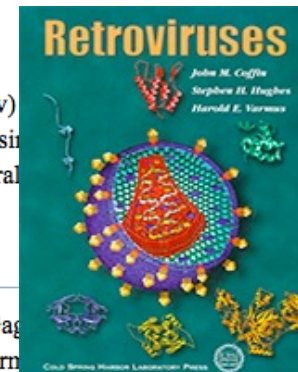
The Retroviral Protease Coffin, JM (1997)

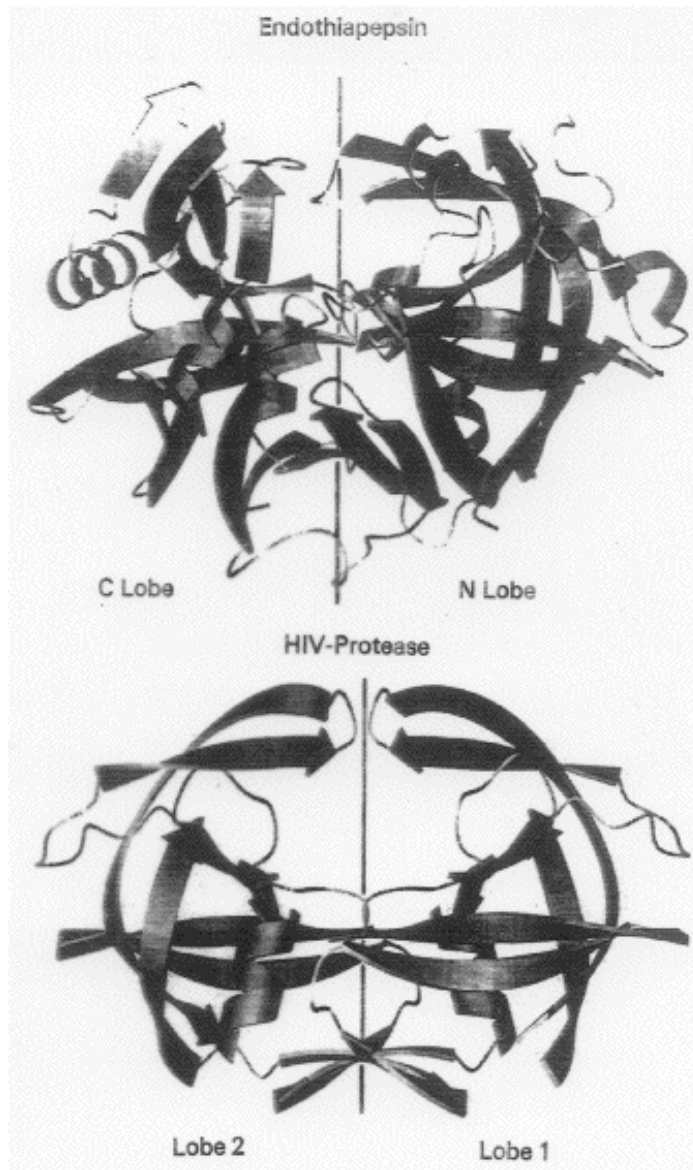
Proteolytic processing at specific sites in the Gag and Gag-Pro-Pol (and sometimes Env) is an essential step in the viral life cycle. Since PR has a central role in proteolytic processing, it is a target for the design of inhibitors of viral replication. This section describes the retroviral protease and its central role late in the viral life cycle.

Identification of the Viral Protease as an Aspartic Proteinase

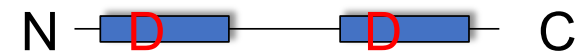
Evidence for a role for a PR in the viral life cycle came from the observation that the Gag was synthesized as a larger precursor that is subsequently processed through a series of internal cleavages to yield the protein products found in the mature virion (Vogt and Eisenman 1973; Vogt et al. 1975). Subsequently, a virion-associated PR was detected that had the specificity to produce these proteolytic cleavages (Von der Helm 1977; Yoshinaka and Luftig 1977a; Dittmar and Moelling 1978; Vogt et al. 1979; Lillehoj et al. 1988; for review, see Oroszlan and Luftig 1990). A demonstration of the critical role of PR in the viral life cycle came from genetic studies—initially with MLV and then with HIV-1—in which it was shown that mutations placed near the 5' end of the *pol* gene resulted in the production of noninfectious viral particles containing unprocessed Gag and Gag-Pro-Pol precursors (Crawford and Goff 1985; Katoh et al. 1985; Kohl et al. 1988). These experiments defined the *pro* gene downstream from *gag* and upstream of the RT-coding domain of the *pol* gene (see also Levin et al. 1984; Kramer et al. 1986; Farmerie et al. 1987). The *pro* gene appears in different reading frames relative to *gag* and *pol*, depending on the genus (see Chapter 2 and Fig. 5).

A major step was taken in our understanding of PR when sequence alignments revealed similarities with the aspartic proteinase family of enzymes (Toh et al. 1985). Proteinases are grouped into families based on the chemical entity that effects hydrolysis of the peptide bond. All aspartic proteinases share the feature that two aspartic acid residues, each placed in the highly conserved motif Asp-Thr/Ser-Gly, coordinate a water molecule used to hydrolyze the target peptide bond. Sequence alignments of retroviral genomes showed that this conserved motif is encoded in all known retroviruses.





N末ドメインと
C末ドメインが
相同



ホモダイマー



mafftによる配列アラインメント

CLUSTAL format alignment by MAFFT L-INS-i (v7.402)

```
2APRN      AGVGTVPMTDYGNDIEYYGQVTIGTPGKKFNLDFTGSSDLWIASTLCTNCGSGQTKYDP
2APRC      -----YD-----STKFKG
                :*                      .**:.

2APRN      NQSSTYQADGRT-WSISYGDGSSASGILAKDNVNLGGLLIKQTI-----ELAKREAASF
2APRC      SLTTVPIDNSRGWGITVDRA TVGTSTVAS---SFDGILDTGTTLLILPNNIAASVARAY
                . ::.      :.*  *.*: . .: .:. :*.      .:.*:.* .* *:      ::*      * :.

2APRN      ASGPN-DGLLGLGFDT-----ITTVRGV--KTPMDNLISQGLISRPIFGVYLGK-----
2APRC      GASDNGDGTYTISCDTSAFKPLVFSINGASFQVSPDSLVEEFQGGQCIAGFGYGNWGFAL
                .:. * **      :. **      : :.:.*.      :.. *.*: : : .: * *.      *:

2APRN      -----AKNGGGGEYI-----FGG
2APRC      IGD TFLKN----NYVVFNQGVPEVQIAPVAE
                **      :*:      ..
```

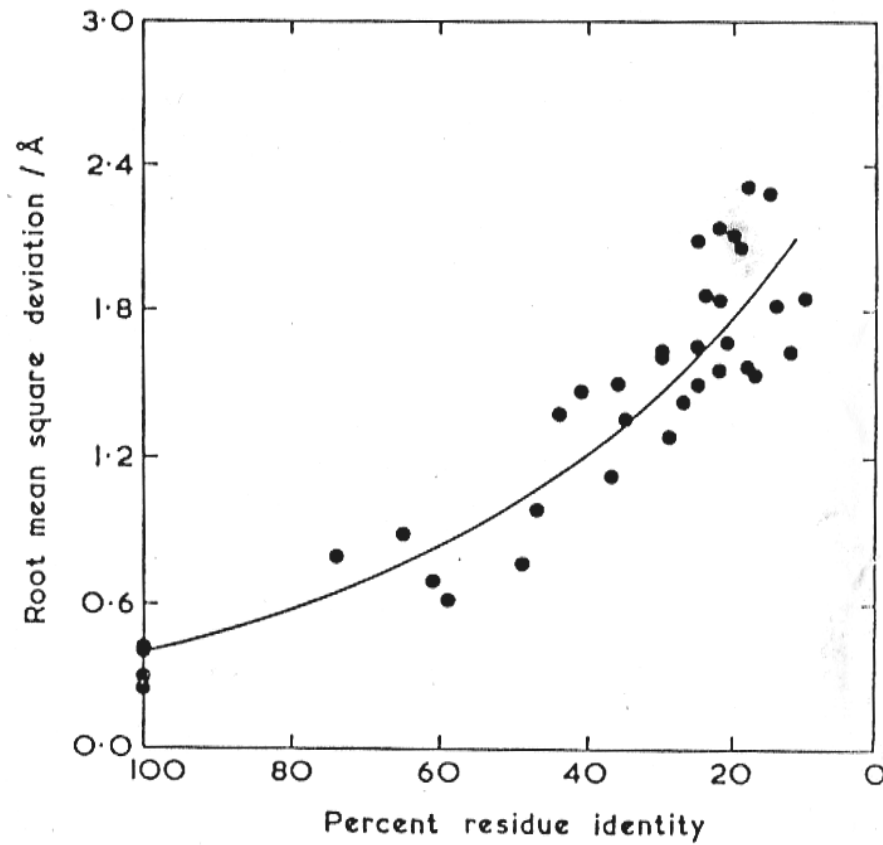
進化的関係が遠いと配列情報だけでは正しいアラインメントが作れない
入力ファイルは 2aprNC.fasta

構造を比較することの意味

アミノ酸配列に比べ、立体構造は保存しやすい



アミノ酸配列は類似していない、つまり配列レベルの検索で検出できないような類似性(遠い進化的関係)を、構造比較により検出できる。(構造データベース検索)
また、立体構造比較から、配列レベルでは検出できない機能部位の同定も可能(構造アラインメント)



$$\Delta = 0.40 \exp(1.87H)$$

Δ : common core 主鎖の重ねあわせのRMSD

H : 配列の相違度

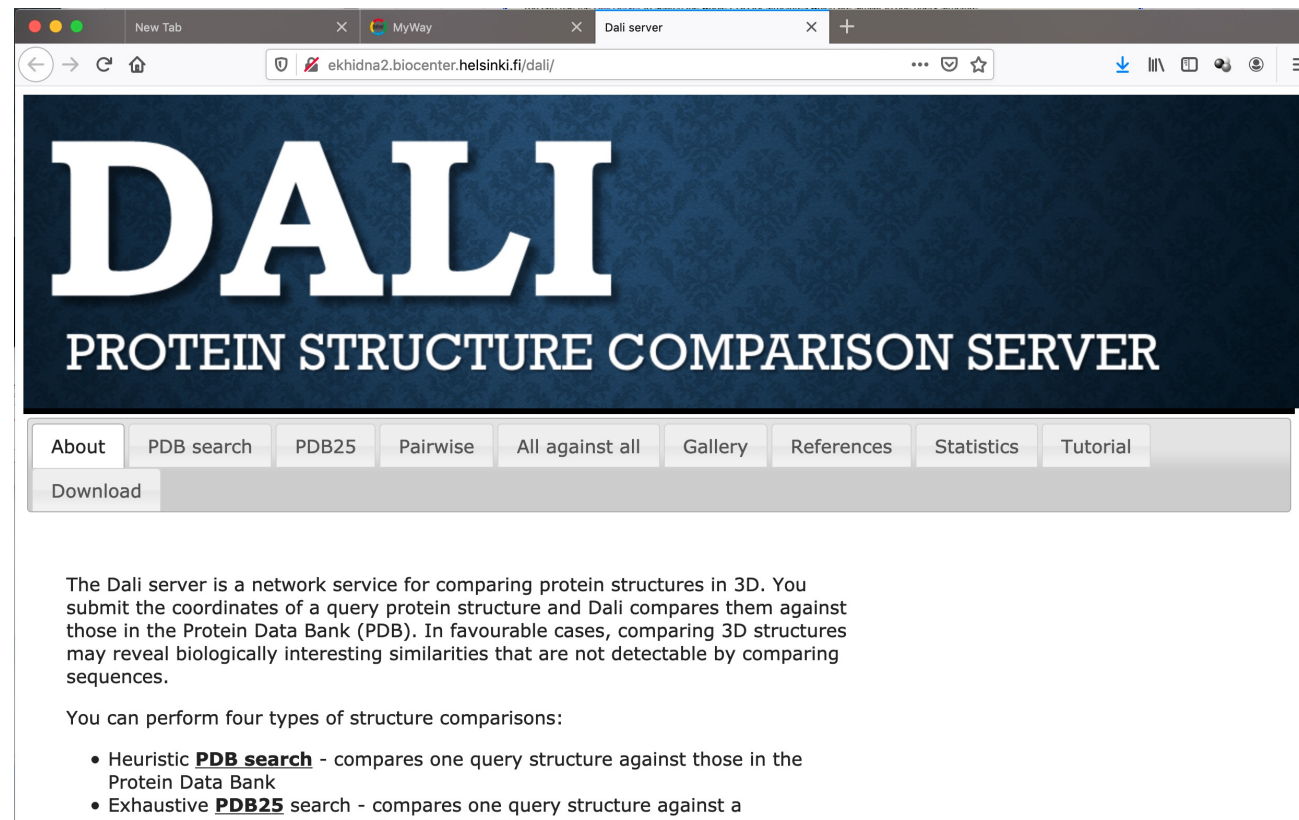
Hが小さい時 : 表面のみの変化

Hが大きい時 : 表面と内部の変化

内部の変化は表面に比べ大きな構造変化をもたらす

構造比較のサイト

<http://ekhidna2.biocenter.helsinki.fi/dali/>



[back to top](#)

DSSP	llllllllLLHhhl111EEEEEEeeee1111leeEEEE-----eeEEEEELLEEE	
Query	tncgsgqtKYDPnqsstYQADGRTwsisygdgssaSGIL-----akDNVNLGGLLI	99
ident		
Sbjct	-----IAAS--varAYGASDN-----gDGTytiscdtsafkpLVFSINGASF	96
DSSP	-----HHHH--hhhHHLLEEL-----1111Eee111hhh111EEEEELLEEE	

DSSP	E-----eeeEEEEeElhhhl	lllEEEEELl	hhhl	lllllllLHHHHHh	
Query	K-----gqtIELAKreaasfasgpn	DGLLG	Lgfdtittvrgvk	TPMDN	Li 144
ident					
Sbjct	QvspdslvfeefgqgciagFGYGN-----	wgFAIIGD-----	TFLKNN-		134
DSSP	EelhhhheeeeeeelleeeleEEEEL-----	llEEEEELH-----	HHHLL	E-	

```

DSSP  hlllllllLEEEEEELlhhhllLEEEEEEL---l
Query  sqglisrpIFGVYLGkaknggGGEYIFG---g  173
ident  |                               |
Sbjct  -----YVVFNQG-----VPEVQIApvae  152
DSSP  -----EEEEELL-----LLEEEEEEl11

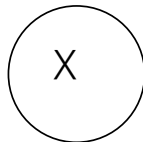
```

タンパク質の機能

生化学的機能

Biochemical Function

酵素活性、リガンド結合能

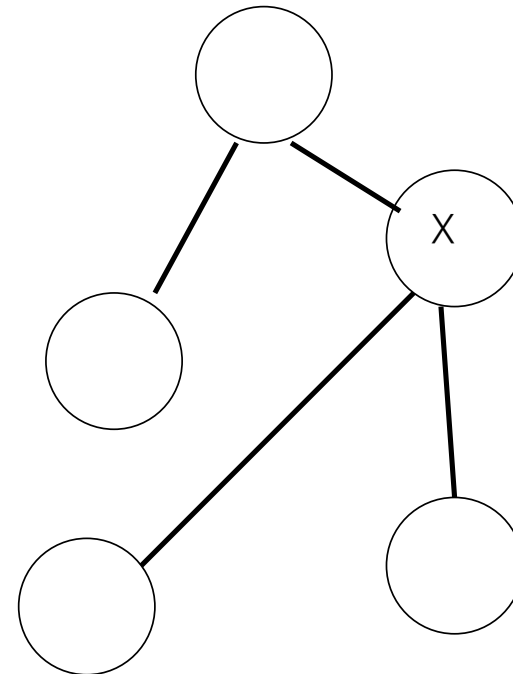


vs

生物学的機能

Biological Function

記憶、発生、行動



ネットワークの解析のツール

(1) R igraph

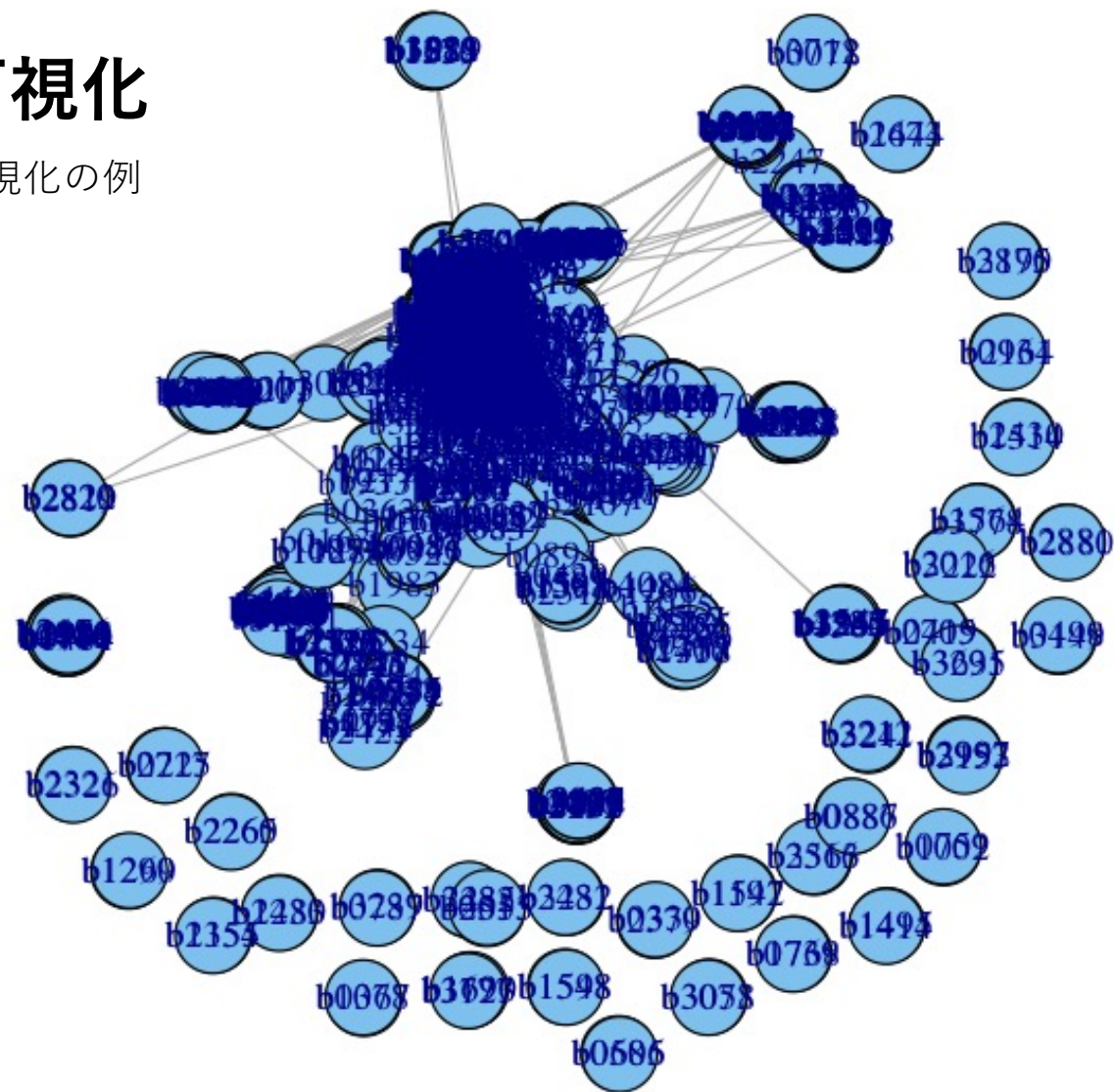
(2) Cytoscape

タンパク質間相互作用ネットワーク解析

1. ネットワーク可視化
2. 次数分布
3. 中心性解析
4. コミュニティ抽出
5. ...

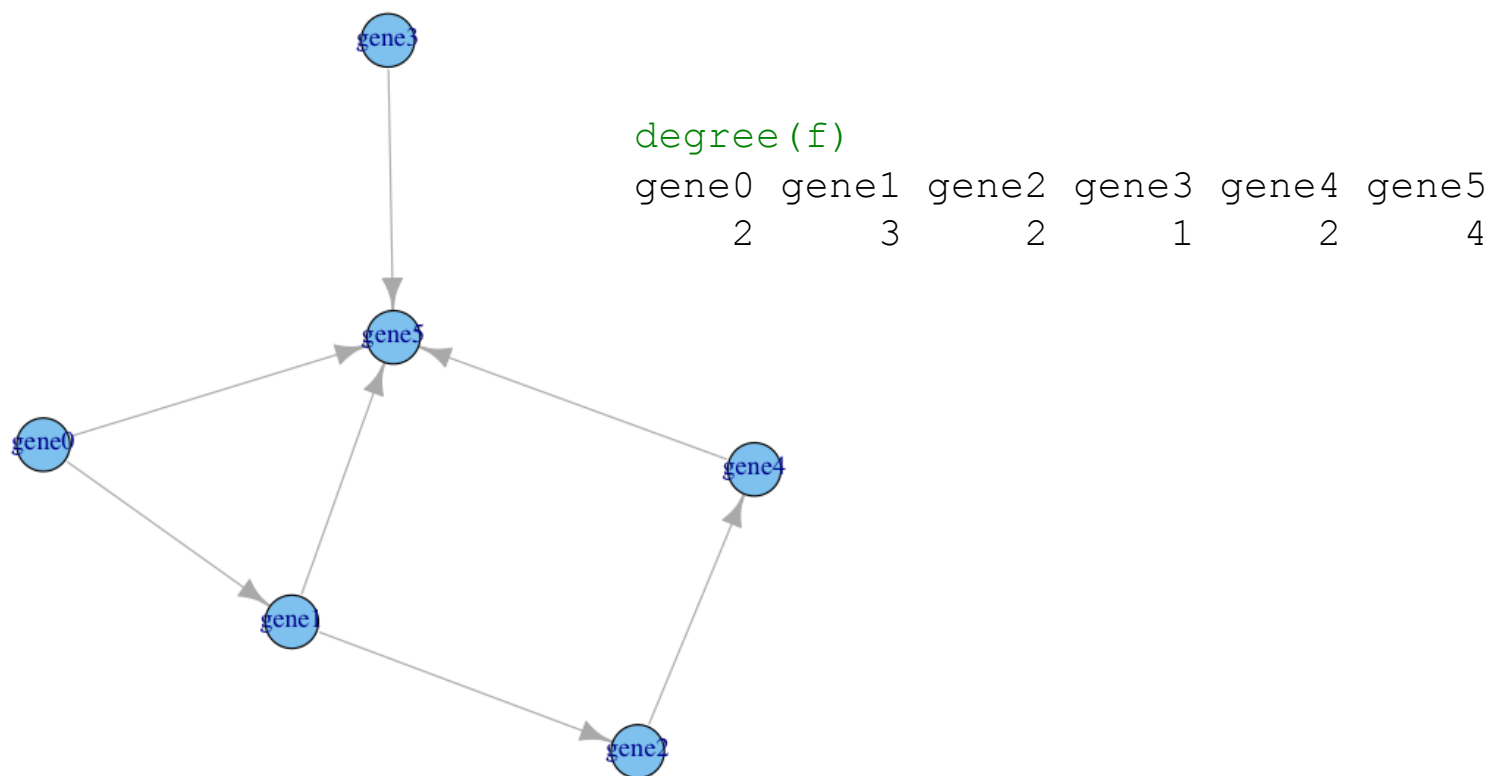
ネットワークの可視化

Igraphによる可視化の例

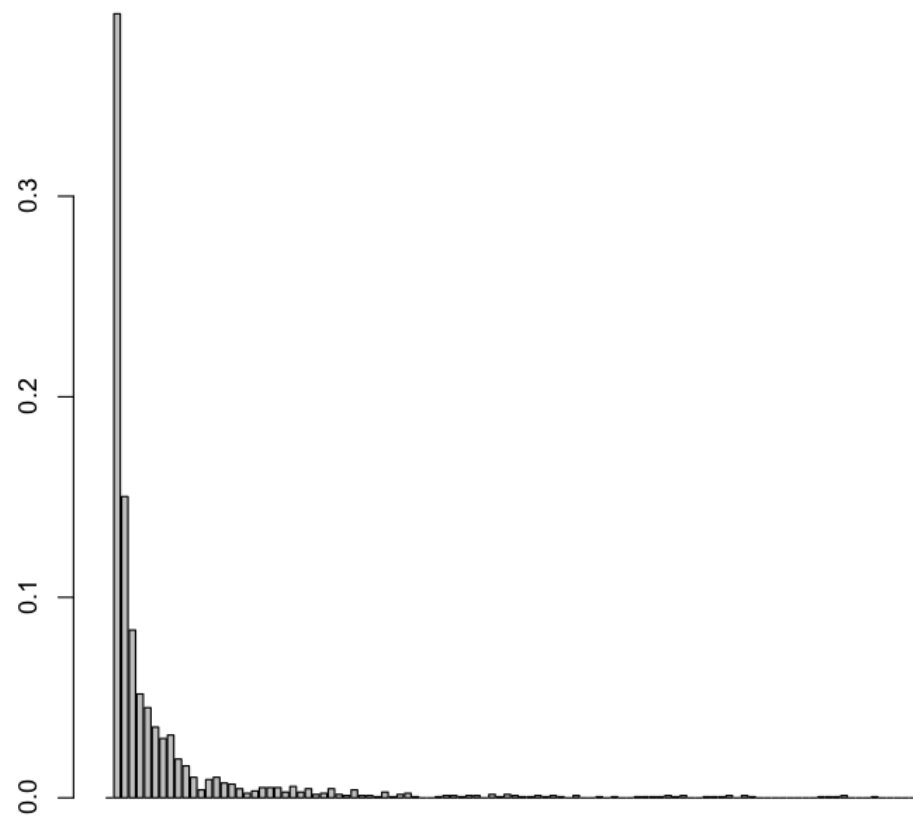


次数分布

次数：各ノードに接続する辺の数



次数分布を調べる



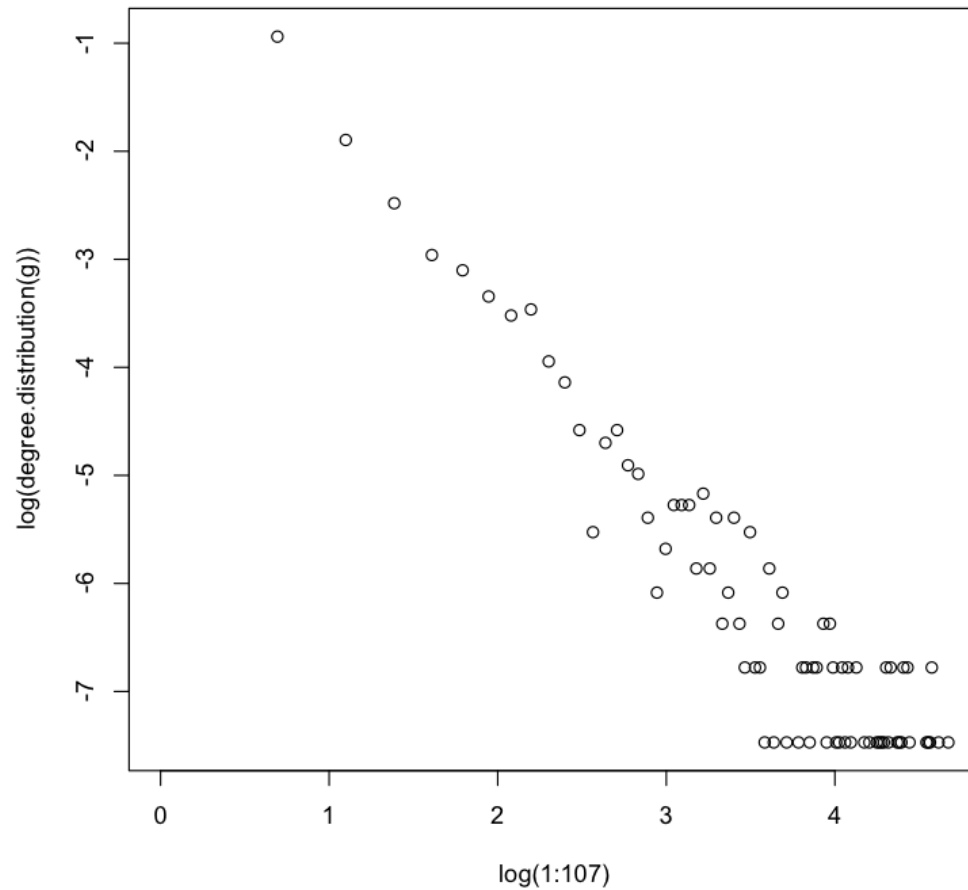
次数分布

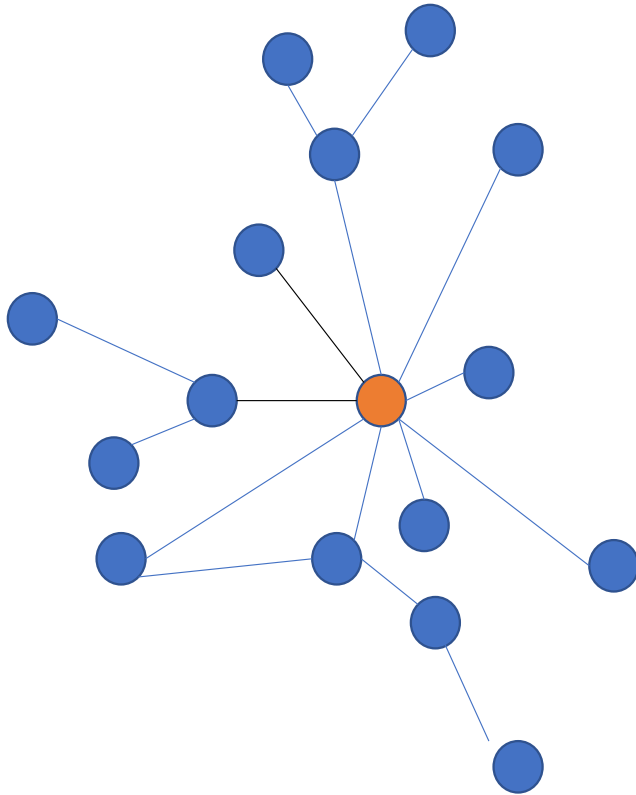
log-log plot

log-log プロットをとると
直線関係

---→ スケールフリー
であることを示している

少数のハブ(hub = 次数の
高いノード) と多数の
次数の低いノードから
構成されている





スケールフリーネットワーク

少数のハブと多数の相互作用の数の少ないノードからなる。

ハブ



多数の相互作用相手を持つ
従来の鍵-鍵穴モデルやinduced fitモデルでは説明できない

天然変性領域がハブの役割を担っていることが多い

スケールフリー性は、社会学をはじめとするこれまでの研究により、現実世界のネットワークで幅広く観察されている。

- ・ 人々の持っている知人関係の数をみると、一部の人は非常にたくさんの知人を持っているが、大多数の人々の知人の数は限られている。
- ・ WWWではごく少数の有名サイトが数百万単位のリンクを集めているが、大多数のサイトはわずかなリンク先からしかリンクされていない。
- ・ 生体内の相互作用でも、ごく一部のタンパク質が多数のタンパク質と相互作用する構造になっている。

数学的には、スケールフリー性は頂点が次数 k を持つ確率 $p(k)$ の確率分布が $p(k) \propto k^{-\gamma}$ のべき乗則になると表現される
 $\log(p(k)) \propto -\gamma \log(k)$

スケールフリーグラフが持つ注目すべき特性として、ネットワーク障害など「ランダムな故障や攻撃」に対して頑強性が高いことがあげられる。

ただし、スケールフリーなネットワークの場合、特定の重要なハブをピンポイントで狙った攻撃に対しては脆弱

<https://ja.wikipedia.org/wiki/複雑ネットワーク> より

中心性

ノードがネットワーク中でどの程度中心的であるかを評価する指標

ネットワークの中の各ノードの重要性の評価に用いられる

様々な中心性の定義がある

次数中心性 次数の高いノードが重要

近接中心性 他のノードに短い距離で繋がるノードが重要

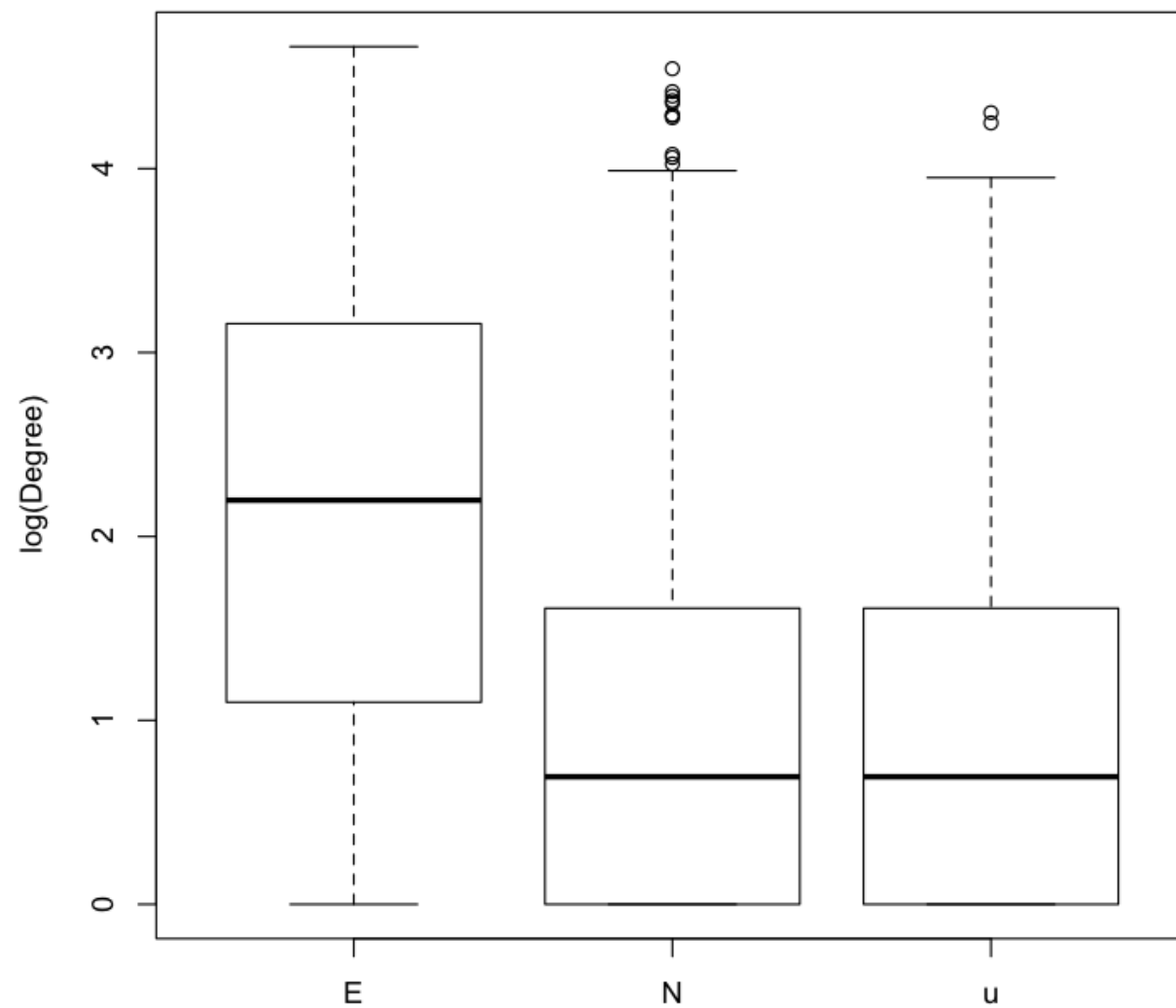
媒介中心性 ノード間最短距離を考えた時によく通過する
ノードが重要

固有ベクトル中心性 隣接するノードの中心性も考慮
次数中心性の拡張版 (Google Page rankで使用)

タンパク質間相互作用ネットワーク の中心性解析

ecoli_proteins_essentiality_Baba2006MSB.txt
Baba *et al.* (2006) *Mol Syst Biol* **2**, 0008

```
gene essential
. . .
b0023 N      non essential
b0024 u      unknown
b0025 N
b0026 E      essential
. . .
```



中心性が高いものほどessentialなものが多い

重複可能性(duplicability)とネットワークの関係

Evolutionary Rate and Duplicability in the *Arabidopsis thaliana* Protein–Protein Interaction Network

David Alvarez-Ponce^{1,2,*} and Mario A. Fares^{1,2,*}

Genome Biol. Evol. 4(12):1263–1274. (2012)

進化速度と中心性の相関は一般には低いですが、特定の機能カテゴリーについては、発現レベルと進化速度の間の相関と同程度かそれ以上の相関を示す。

ネットワーク内で相互作用するタンパク質をコードする遺伝子は比較的類似した進化速度を持つ

シロイヌナズナでは重複している遺伝子由来のタンパク質の方が、シングルトン（重複のない）遺伝子由来のタンパク質の方が高度に相互作用が多く、ヒトでも同じ傾向。一方、*E. coli*, 酵母、線虫、ハエでは重複遺伝子はネットワークの周辺に位置する傾向がある。

---→ 中心性と重複可能性(duplicability)の関係が、真核生物の進化の過程で少なくとも2回逆転した

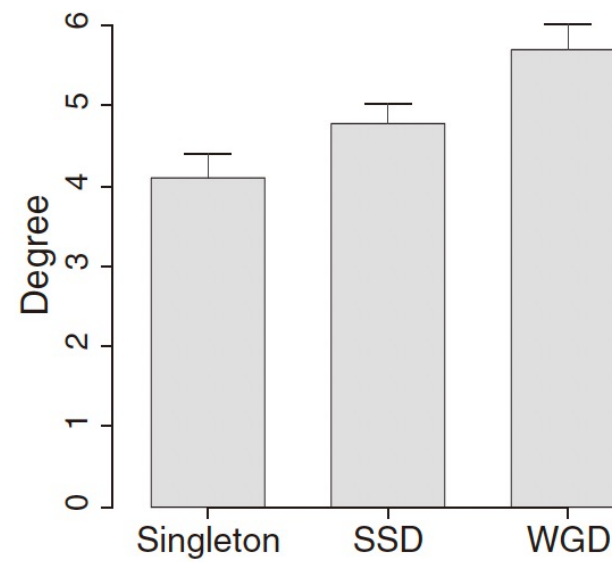


FIG. 2.—Average number of interactors for proteins encoded by singleton, whole-genome duplication (WGD), and small-scale duplication (SSD) genes. Error bars represent the standard error of the means.

Increased Genome Sampling Reveals a Dynamic Relationship between Gene Duplicability and the Structure of the Primate Protein–Protein Interaction Network

Aoife Doherty,[†] David Alvarez-Ponce,[†] and James O. McInerney*

Mol. Biol. Evol. 29(11):3563–3573 (2012)

中心性と重複可能性との関係の方向性は普遍的ではない：

Escherichia coli、酵母、ハエ、および線虫では、重複した遺伝子はネットワークの周辺に存在する傾向
ヒトでは重複しが遺伝子はネットワークの中心的な位置を占める傾向

霊長類の異なる系統の各枝で起きた重複について調査

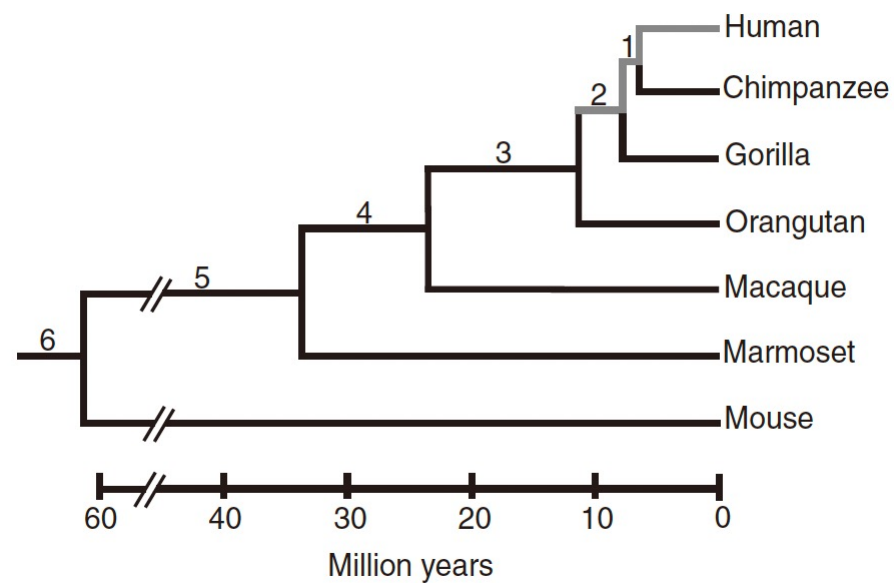
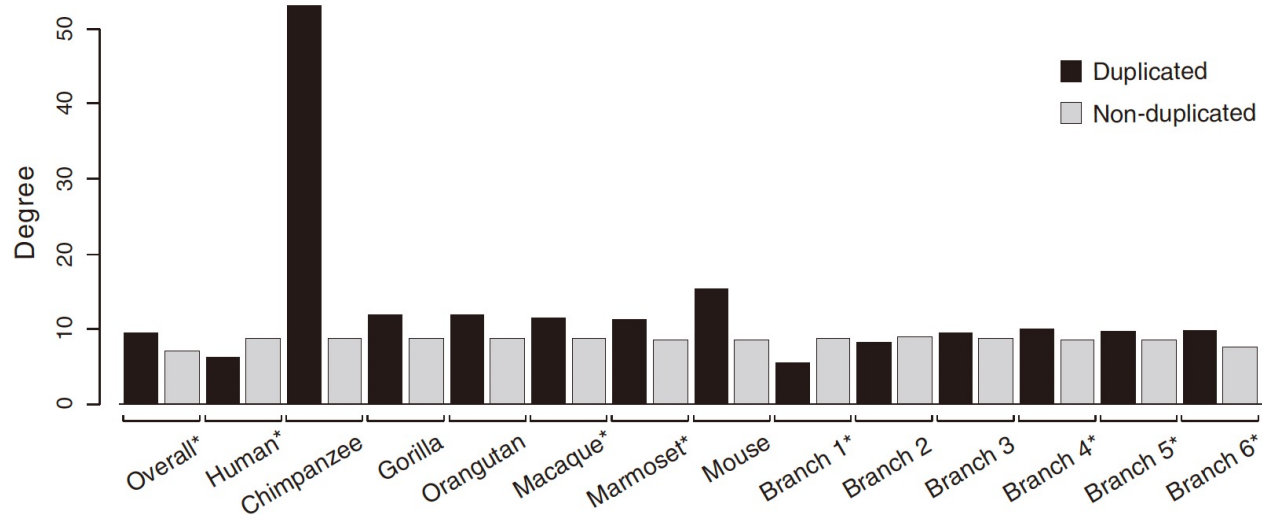
種系統樹の13の枝のうち、10個の枝で重複した遺伝子の方が、重複のない遺伝子よりも次数が高い
そのうち、5つの枝でその差は統計的に有意

3個の枝（ヒトへ至る外部枝を含む）で逆のパターン。そのうち2つ（ヒトへ至る外部枝含む）で
その差は統計的に有意

次数以外の中心性でも同様の傾向が見られる。

中心性と重複可能性には関連はあるが、その関係性は霊長類の進化の過程で変化している。

相互作用するタンパク質の遺伝子は、同じ枝で同時に重複する傾向がある。



Conclusion

生化学的機能の解析:

アミノ酸配列データ、立体構造データ

生物学的機能の解析

相互作用データ

タンパク質の情報解析には、その他にも使えるデータがある

(塩基配列、ゲノムデータ、etc)

異なるタイプのデータを組み合わせることで、個別のデータのみを用いるより豊かな情報を得ることができる