

生物情報学とは？

生物学と情報学の出会い

宮澤 三造

群馬大学大学院工学研究科

2007年 11月28日 白鷗大学にて

## 生物情報学とは:

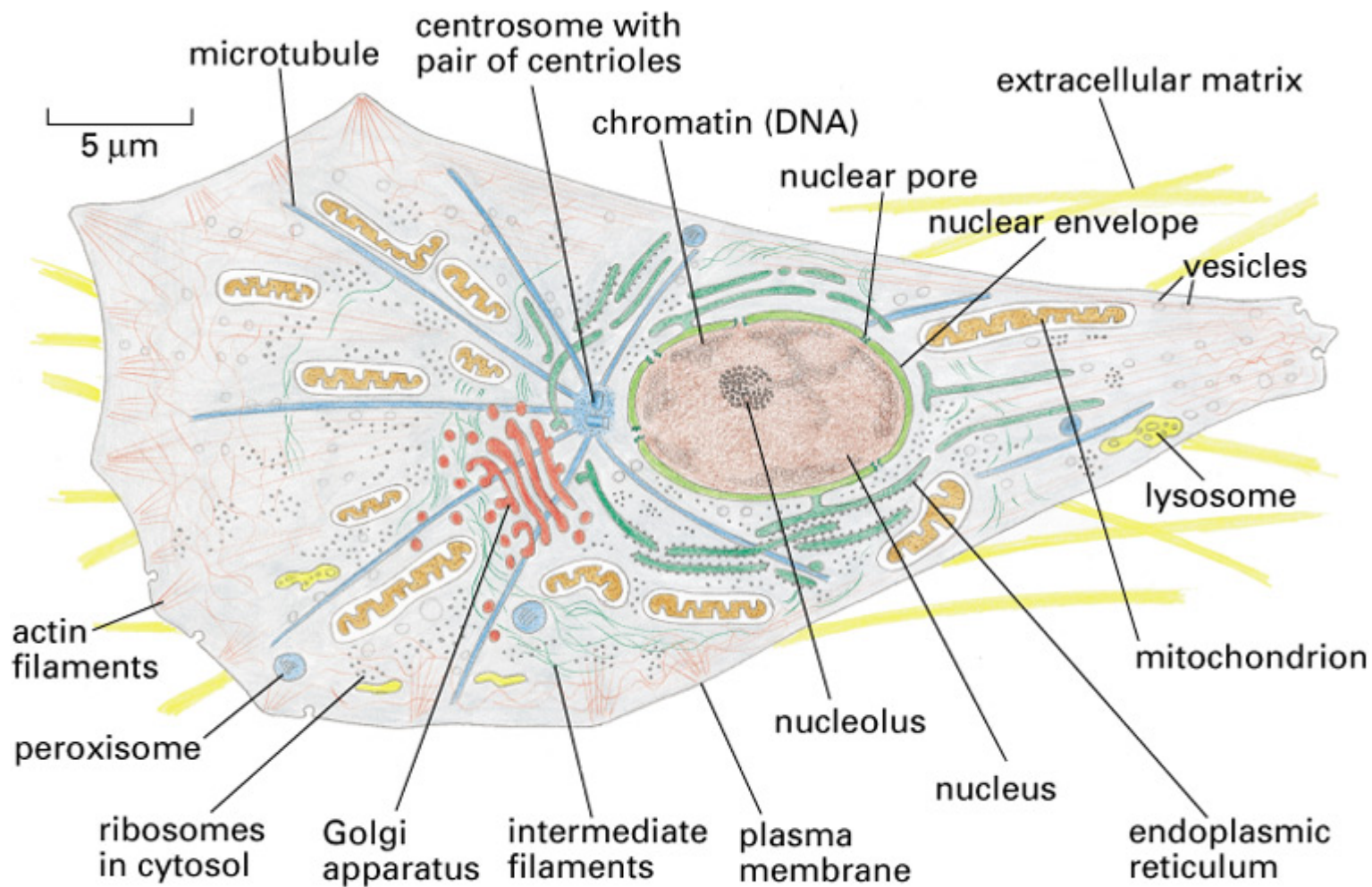
ゲノムにコードされている情報を、情報学の手法で読み解くこと。

## 広義には以下の領域を含む:

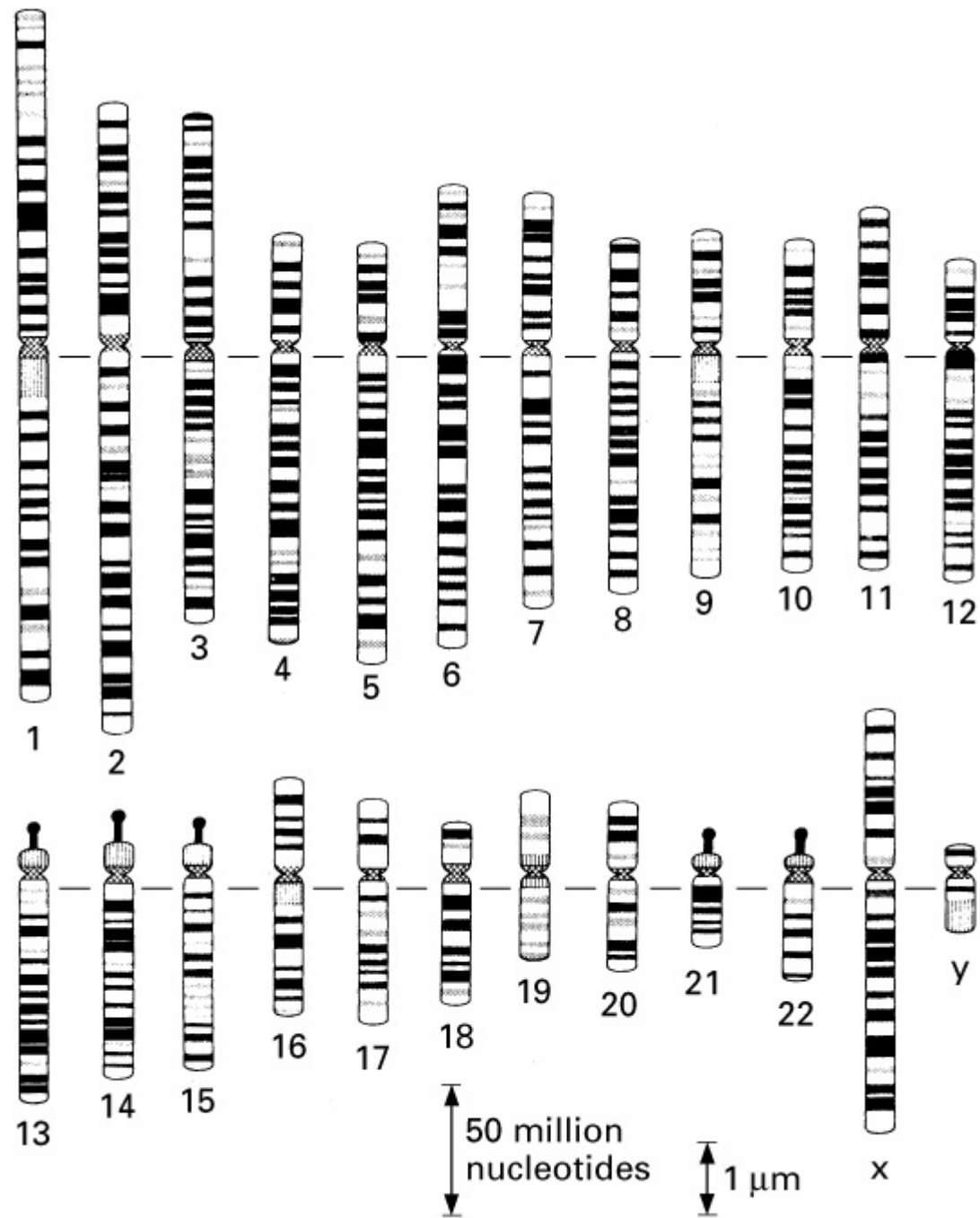
分子進化学(分子系統学)  
構造生物情報学  
システムバイオロジー

## 境界領域における研究能力:

生物学／生化学／物理学の知識 X 情報学の知識



動物細胞の模式図

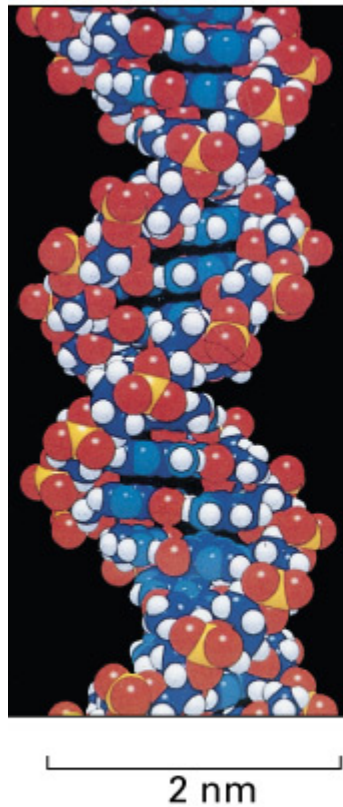


ヒトの染色体: 22対の常染色体と性染色体 XX または XY

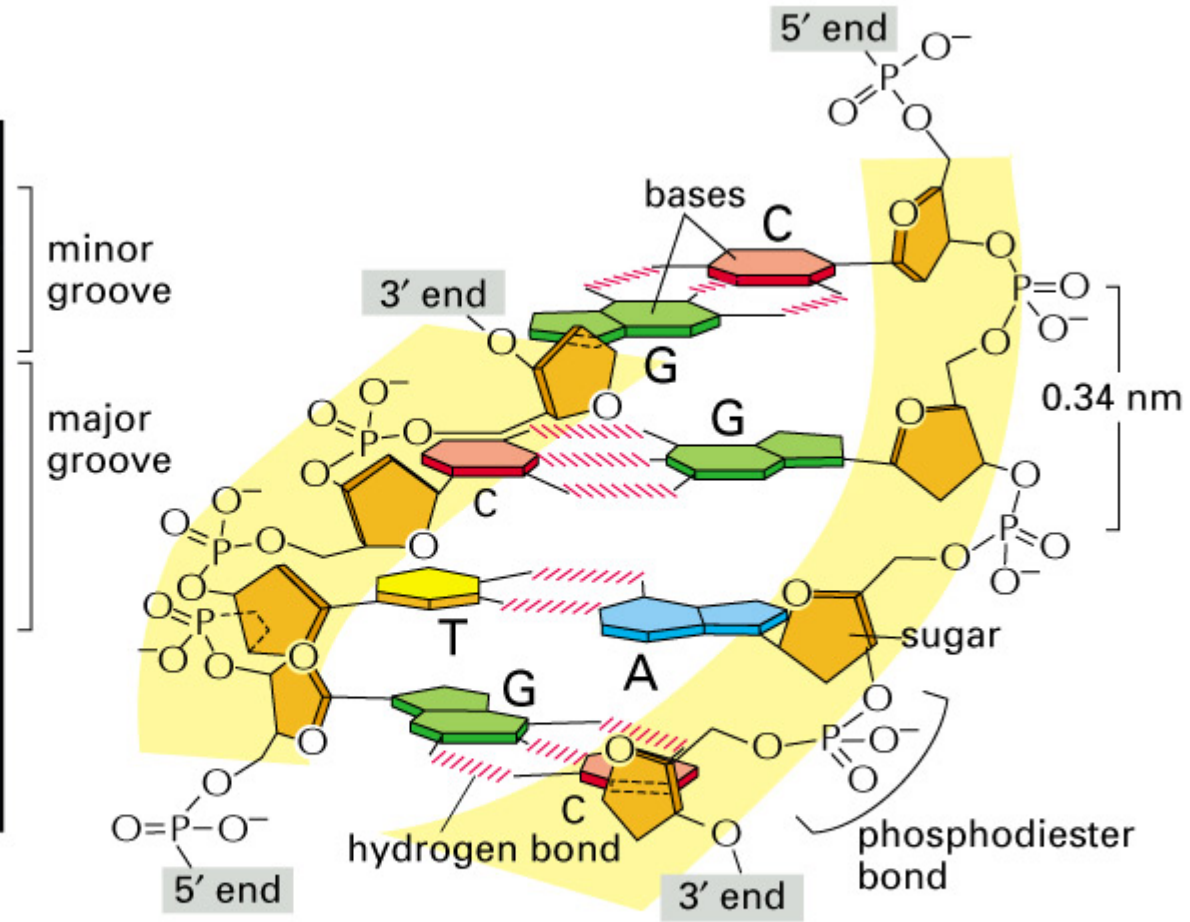
染色体: DNAと蛋白質の複合体

# DNAの分子模型

# DNA分子の模式図



(A)

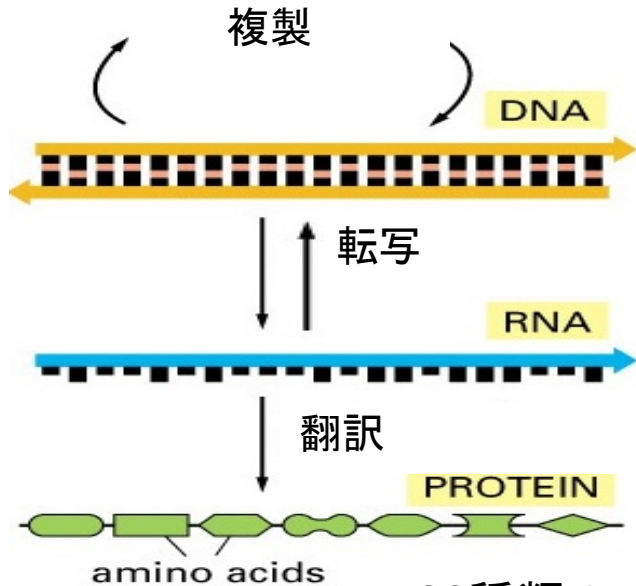


(B)

(A,T,C,G) 4種類の塩基からなる鎖状分子の2本鎖が A-T, C-G と相補的に結合し、2重らせん構造をなす。(Watson & Crick, 1953)

ヒトのDNA: 30億塩基 x 0.34 nm = 1 m

# 遺伝情報の流れ:



...TAATA...TCGGAT ...TAC...TTCCAG...CA...TC...CACATT...  
...ATTAT...

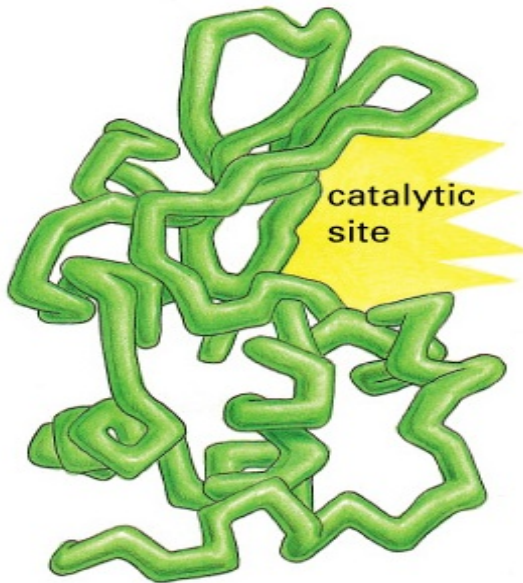
転写

AGCCUA .....AUG.....AAGGUC..... .....GUGUAA.....

翻訳

M ... K V ..... V

折り畳み  
20種類のアミノ酸が鎖状に結合した高分子



(A) lysozyme

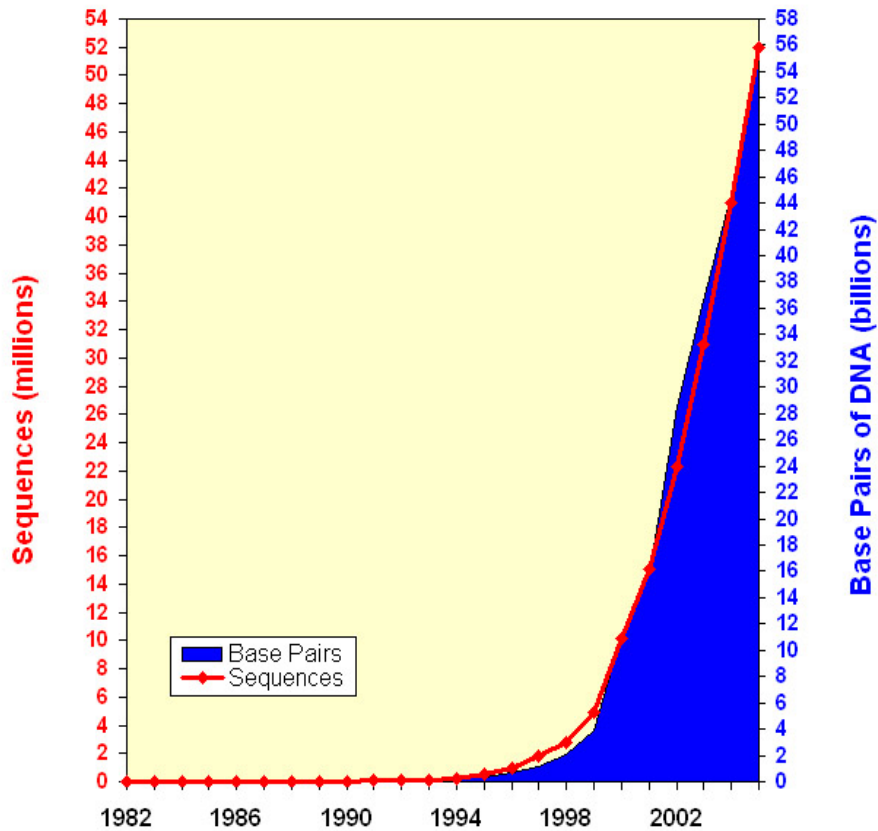
Standard Code Table  
2<sup>nd</sup> position

	U		C		A		G	
U	UUU	Phe, F	UCU	Ser, S	UAU	Tyr, Y	UGU	Cys, C
	UUC	Leu, L	UCC		UAC	UGC	UGA	Term
	UUA		UCA		UAA	UGA	Trp, W	
	UUG		UCG		UAG	UGG		
C	CUU	Leu, L	CCU	Pro, P	CAU	His, H	CGU	Arg, R
	CUC		CCC		CAC	CGC		
	CUA		CCA		CAA	CGA		
	CUG		CCG		CAG	CGG		
A	AUU	Ile, I	ACU	Thr, T	AAU	Asn, N	AGU	Ser, S
	AUC		ACC		AAC		AGC	
	AUA		ACA		AAA		AGA	
	AUG		ACG		AAG		AGG	
G	GUU	Val, V	GCU	Ala, A	GAU	Asp, D	GGU	Gly, G
	GUC		GCC		GAC		GGC	
	GUA		GCA		GAA		GGA	
	GUG		GCG		GAG		GGG	

3<sup>rd</sup>

生物情報の爆発的増加は、データベース技術だけでなく、生物情報を読み解くための手法、さらには各種情報を統合化し、システムとして理解する研究を必要としている。

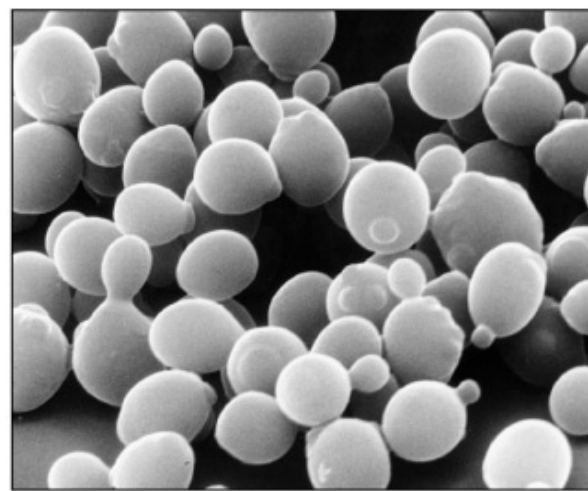
Growth of GenBank  
(1982 - 2005)



- 1975 Sanger's sequencing method
- 1977 Maxam-Gilbert's sequencing method
- 1993-2003 Human genome project

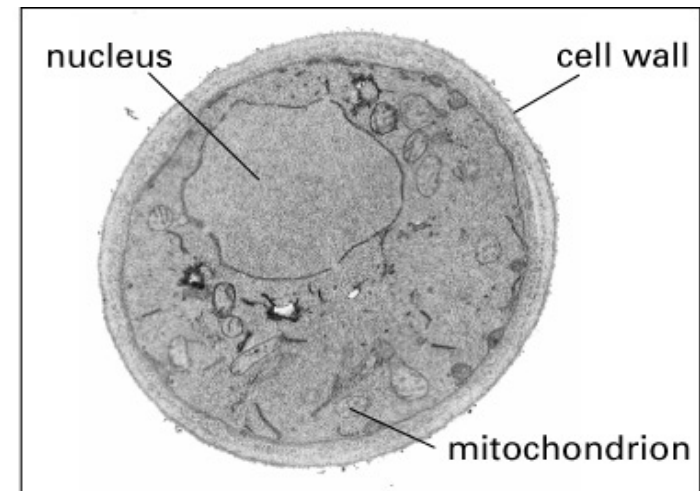
配列解析が完了している生物種

生物分類	生物種の数	例	更新日
古細菌 (Archeabacteria)	16		2002/07/09
細菌 (Bacteria)	89	大腸菌	2003/01/13
菌類 (Fungi)	2	イースト菌	2002/04/14
原生動物 (Protozoa)	1		2003/01/13
植物 (Plant)	2	シロイヌナズナ、稲	2002/04/15
動物 (Animalia)	1	線虫	2002/04/14
	1	ショウジョウバエ	2002/04/14
	2	マウス、ラット	2006/01/18
	2	にわとり、犬	2006/01/19
	2	人、チンパンジー	2006/01/18



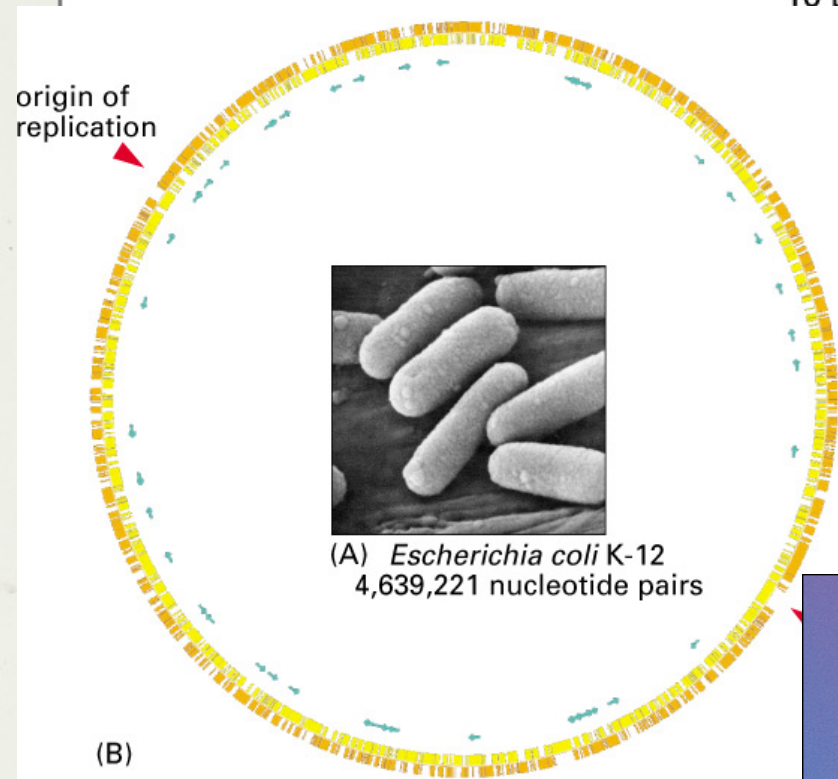
(A)

10 μm



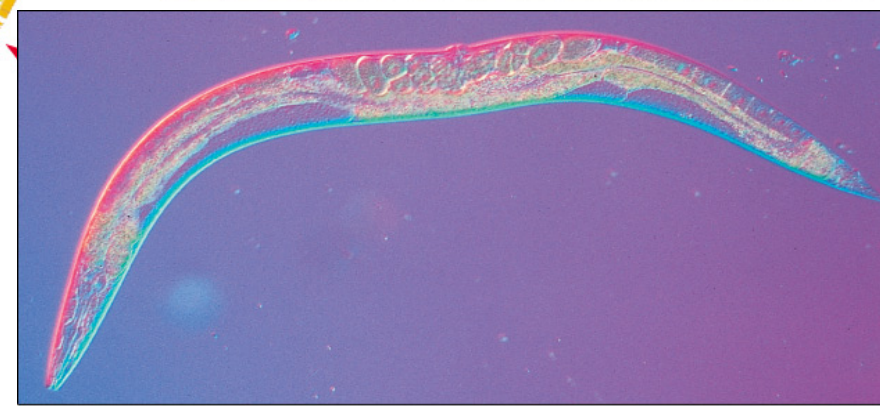
(B)

2 μm



(A) *Escherichia coli* K-12  
4,639,221 nucleotide pairs

(B)



0.2 mm

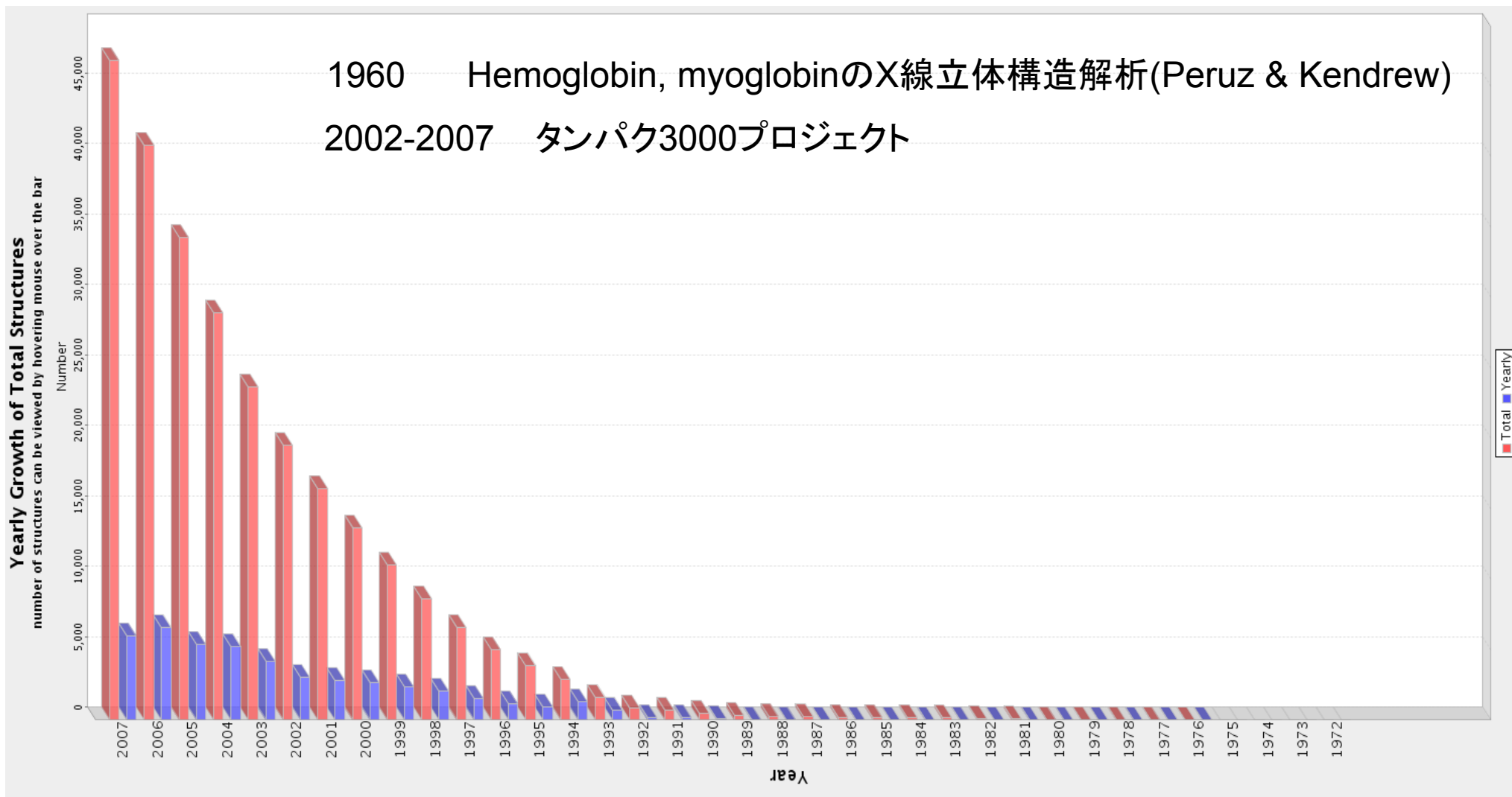
Figure 1-46. Molecular Biology of the Cell, 4th Edition.



# 蛋白質立体構造データの増大

1960 Hemoglobin, myoglobinのX線立体構造解析(Peruz & Kendrew)

2002-2007 タンパク3000プロジェクト



## 生物情報学関連のWeb上のリソース<sup>†</sup>

- データベース及びBrowser: ポータル: [NCBI](#) | [EBI](#) | [DDBJ](#) | [GenomeNet](#)
  - 塩基配列 (データ量の変遷): [GenBank](#) | [EMBL](#) | [DDBJ](#)
  - RNA配列/2次構造; RNA ファミリー ([Rfam](#))
  - 蛋白質配列 ([UniProt](#)) | 蛋白質ファミリー ([Pfam](#)) | Functional site ([PROSITE](#))
  - 蛋白質構造 (データ量の変遷) ポータル: [EBI\(DB | Analyses\)](#)  
構造 ([WW PDB](#) ([RCSB](#) | [MSD](#) | [PDBj](#) | [BMRB](#)) | [PDBsum](#)) | 分類 ([SCOP](#) | [CATH](#)) | 比較 ([Dali](#)) | 予測 | その他 ([GTOP](#))
  - ゲノム: [ゲノム計画](#) | Genome browser ([NCBI](#) | [Ensembl](#))  
配列解析が完了したゲノム ([表](#) | [リンク](#)) | [代表的な生物のゲノム比較](#) | [コドン使用頻度データベース](#)
  - ネットワーク: [Protein-protein interactions](#) | [代謝パスウェイ: KEG](#) | [遺伝子発現パスウェイ](#) | [シグナル伝達](#)
  - 遺伝子発現: [MGED](#), ポータル: [GEO](#) | [ArrayExpress](#)
  - Ontology: [Gene Ontology](#)
  - その他: 多数 ([COG](#))
- 配列解析: 各種ツール ([NCBI](#) | [EBI](#) | [DDBJ](#) | [Pasteur](#))

## 代表的な生物のゲノム比較

	大腸菌 E. coli	酵母 S. cerevisiae	ショウジョウバエ Drosophila	シロイヌナヅナ A. thaliana	マウス M. musculus	人 H. sapiens
ゲノムサイズ(Mb)	4.6	12	120+	115+	2500+	3000+
遺伝子数	4,300	6,250	13,600	25,500	30,000	50,000
1 / 平均遺伝子密度(kb)	1.1	1.9	8.8	4.5	80	60
遺伝子族数	2,500	4,500	8,000	11,000	10,000	10,000

“A Primer of Genome Science”より引用

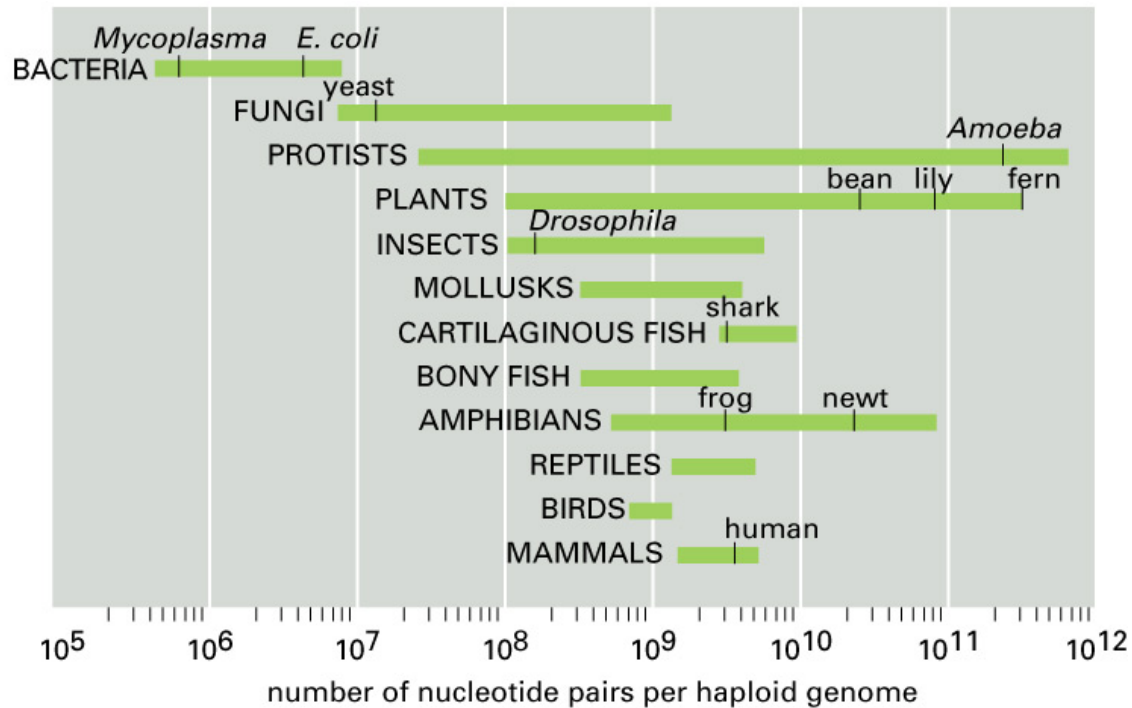
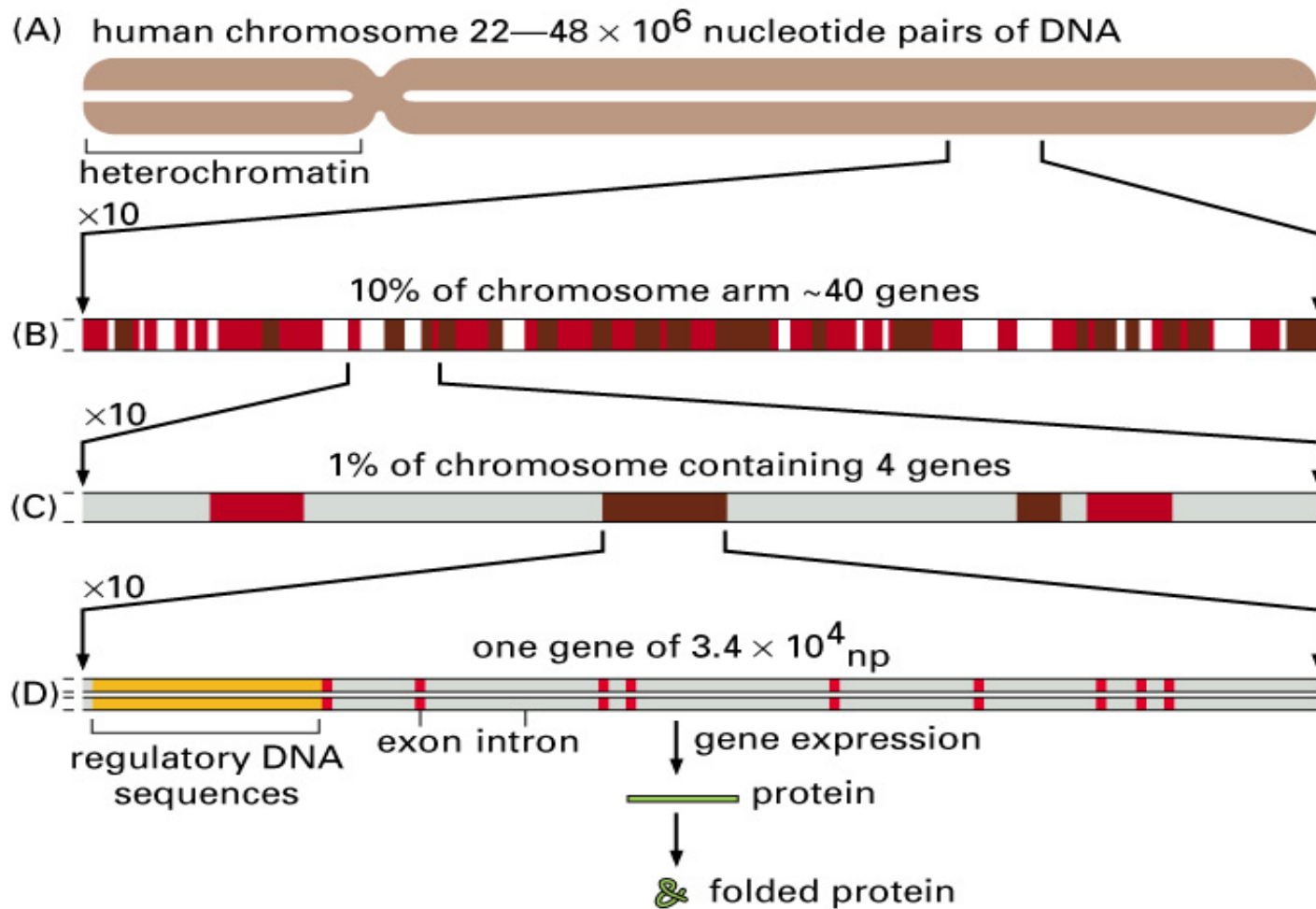


Figure 1-38. Molecular Biology of the Cell, 4th Edition.

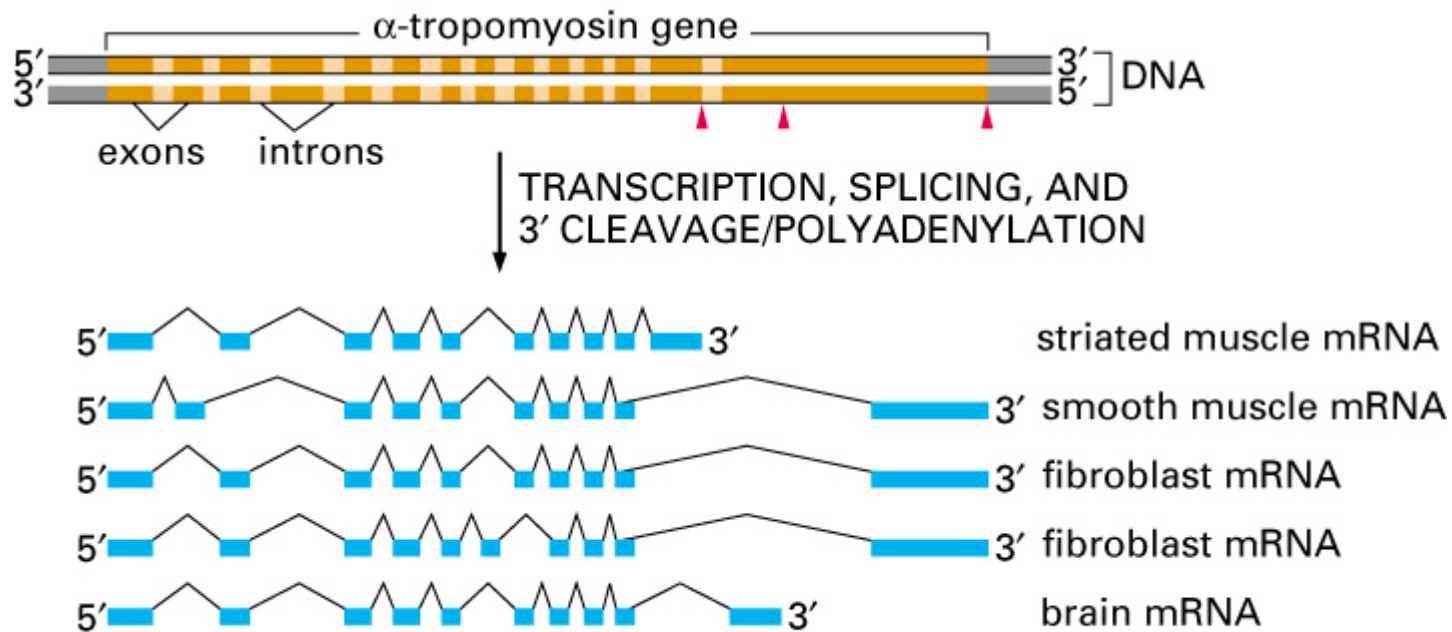


## 生物情報学における研究対象

- 遺伝子制御部位予測
- 遺伝子同定
- 配列特徴抽出
- エクソン-イントロン部位予測

## 生物情報学における研究対象

- エクソン-イントロン部位予測
- Alternating splicing予測



# ExonとIntronの部位認識

確率モデル: 位置依存の配列;  $P(s|\theta) = \prod_i \theta_i(s_i)$

Intronの5'端付近の配列プロファイル:  $\hat{\theta}_i(x)$

Frequencies (%) in 1254 donor splice sites

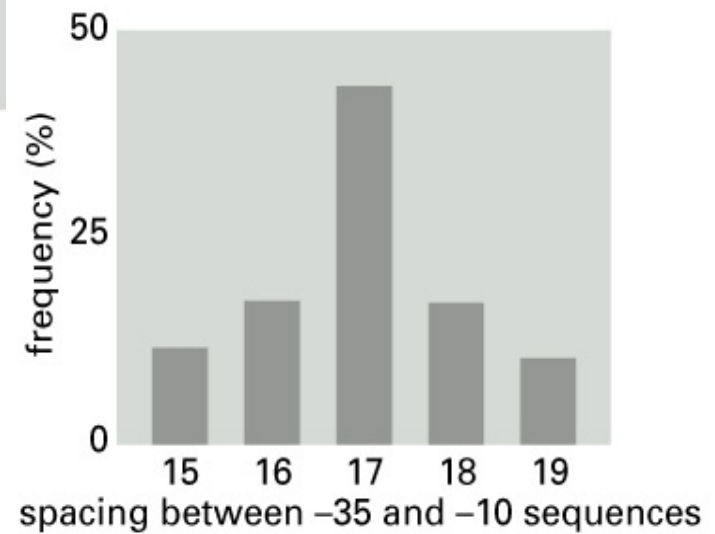
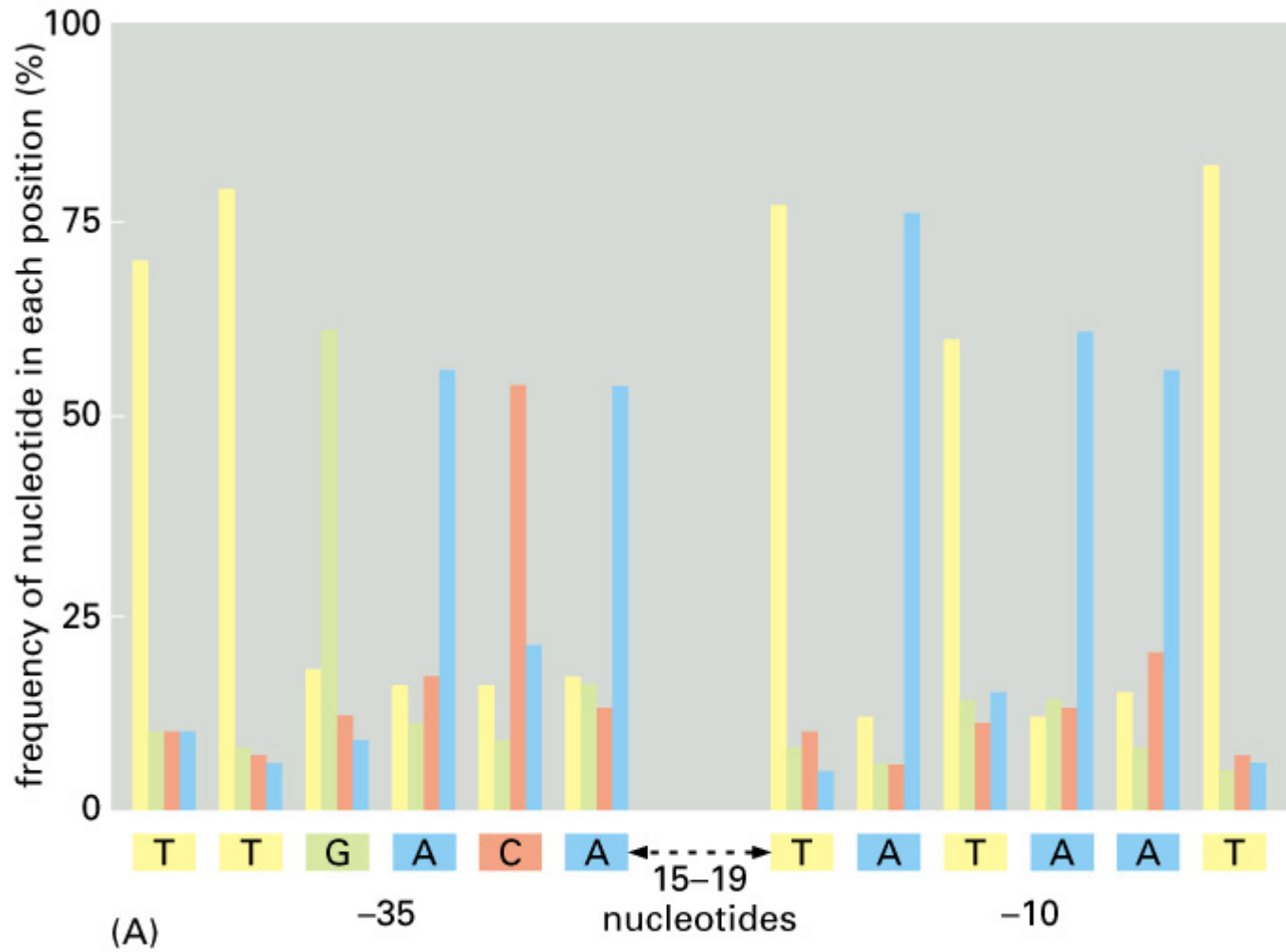
Base\Position	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	33	60	8	0	0	49	71	6	15
C	37	13	4	0	0	3	7	5	19
G	18	14	81	100	0	45	12	84	20
U/T	12	13	7	0	100	3	9	5	46

(Burge & Karlin, 1997)

DNA	C	A	G	G	T	A	A	G	T	$P(s \hat{\theta})$
	0.37	0.60	0.81	1.00	1.00	0.49	0.71	0.84	0.46	0.024

より一般的なモデルとしては隠れマルコフモデル(HMM)

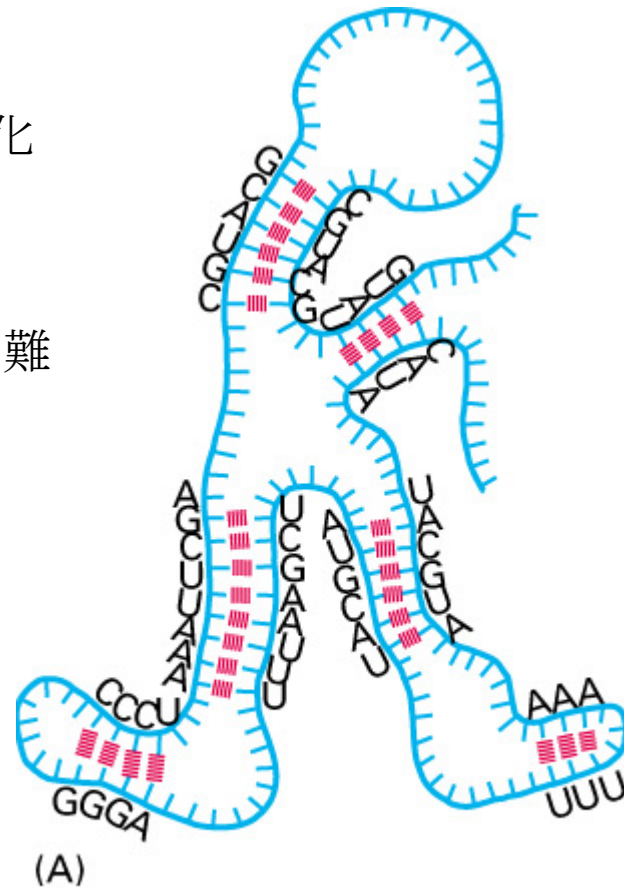
# 大腸菌DNAの転写開始点の上流 -35と -10領域における配列プロファイル



(B)

## 生物情報学における研究対象

- RNA 2次構造予測  
塩基対の数を極大化
- 機能RNA  
2次構造をもつか?
- RNA 3次構造予測は困難



RNA 2次構造

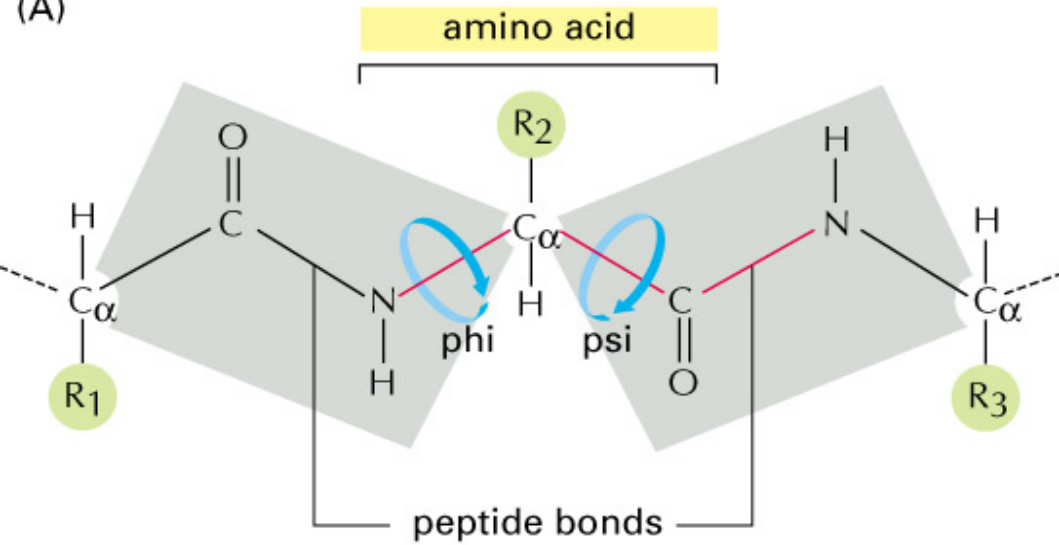


RNA 3次構造

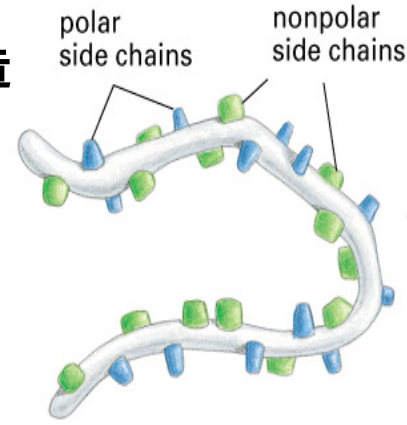


# 蛋白質の構造

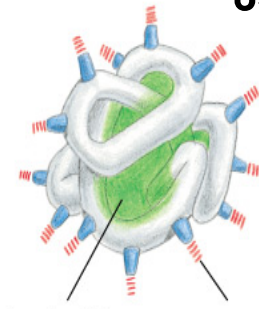
(A)



1次構造



3次構造



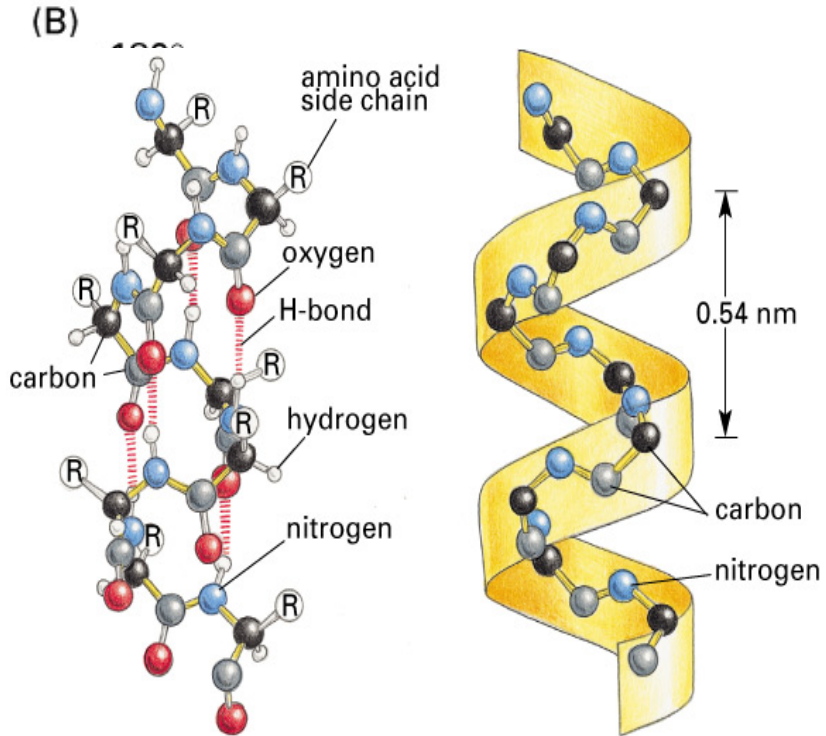
hydrophobic core region contains nonpolar side chains  
polar side chains on the outside of the molecule can form hydrogen bonds to water

unfolded polypeptide

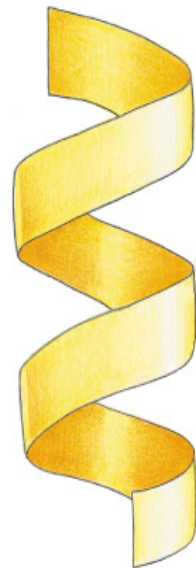
folded conformation in aqueous environment

2次構造

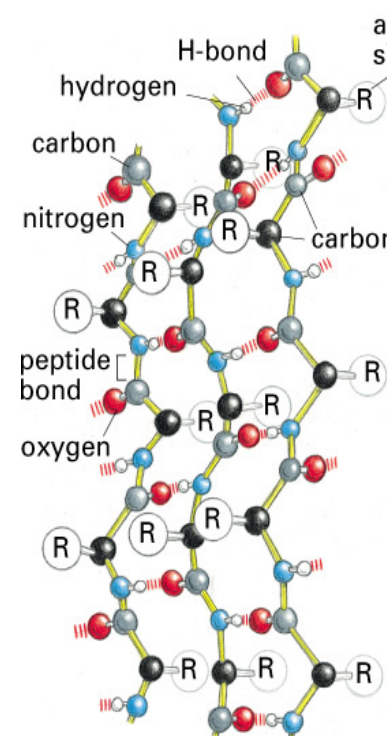
(B)



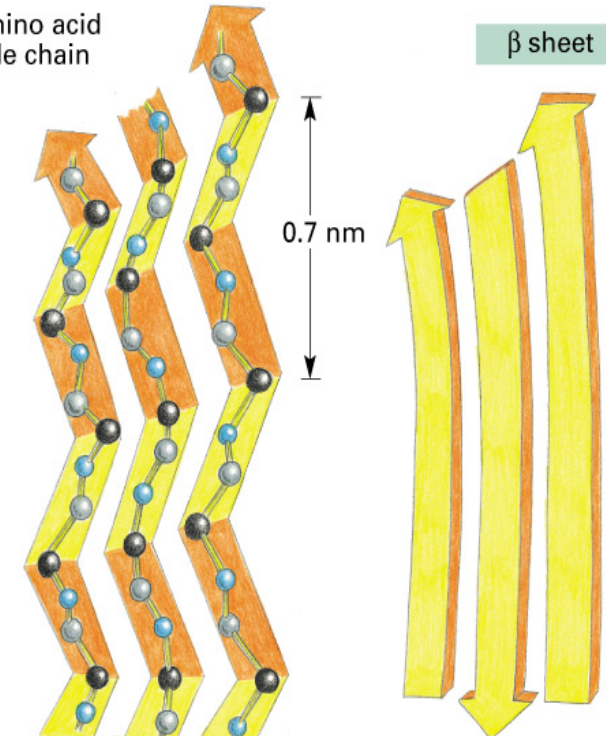
α helix



(C)



β sheet



(E)

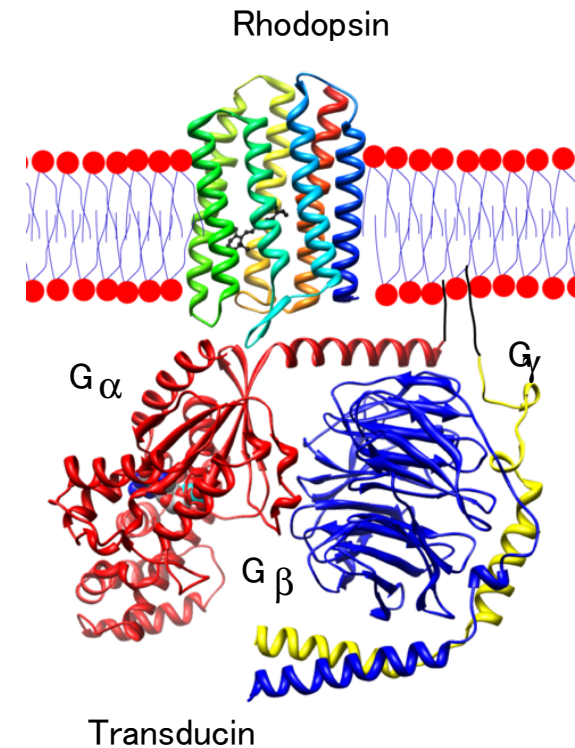
(F)

# 蛋白質立体構造ドメインの折り畳みの分類

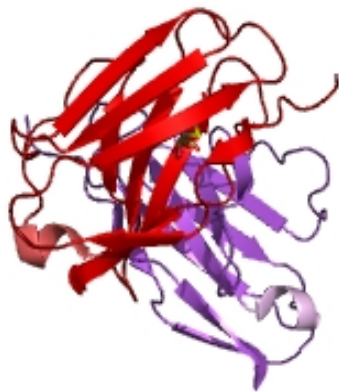
ドメイン(< 200アミノ酸)の折り畳み(fold)の種類は、たかだか数千種類と考えられている。(Chothia,1992)

Scop Classification Statistics (Release 1.71)

	#Folds	#Superfamilies	#Families
All alpha proteins	226	392	645
All beta proteins	149	300	594
a/b proteins	134	221	661
a+b proteins	286	424	753
Multi-domain	48	48	64
Membrane proteins	49	90	101
Small proteins	79	114	186
Total	971	1589	3004



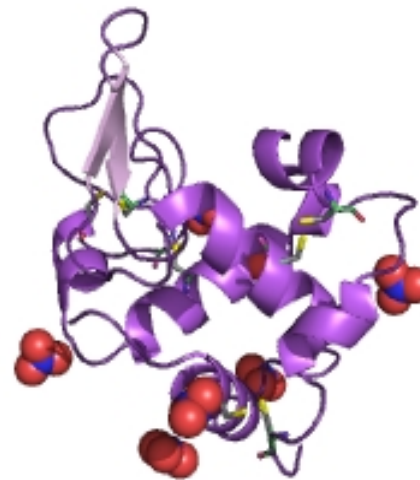
2mm1: Human myoglobin



1bww: Ig Kappa V



1hti: Triosephosphate isomerase, Human

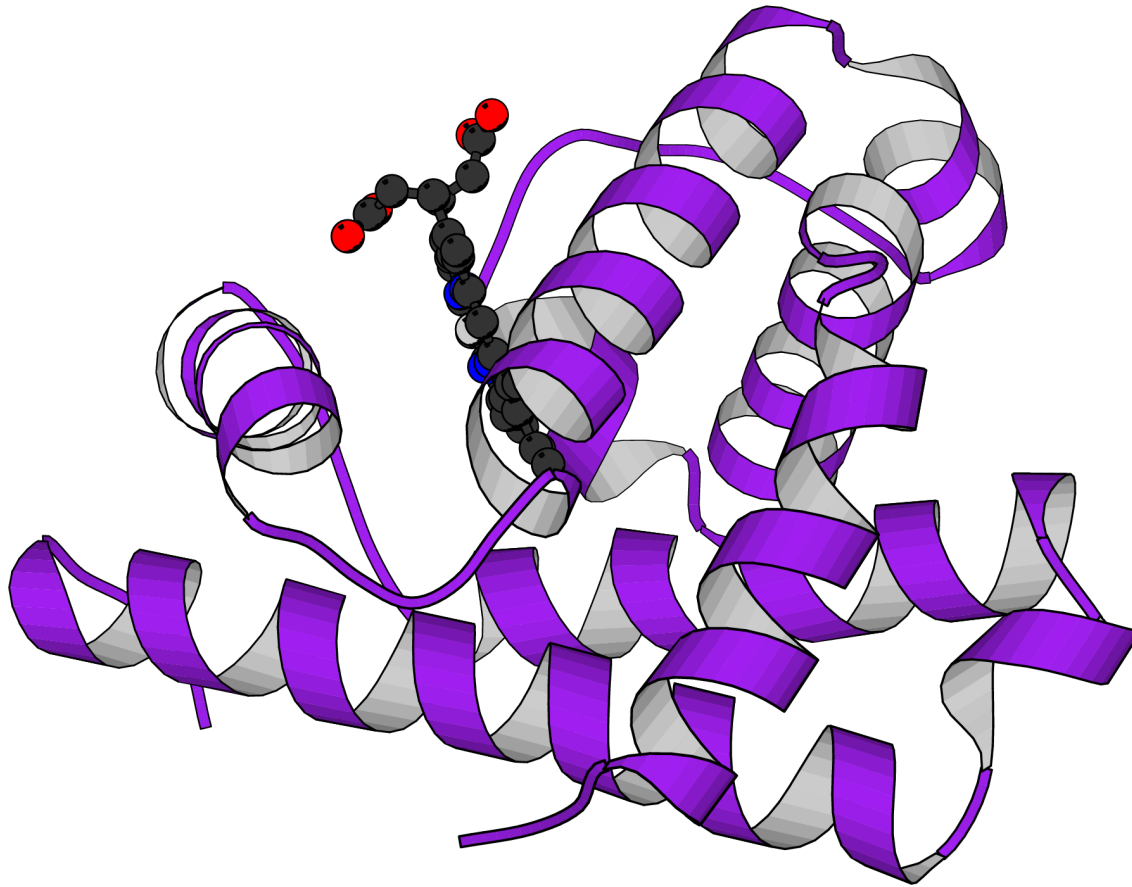


1jsf: Human lysozyme



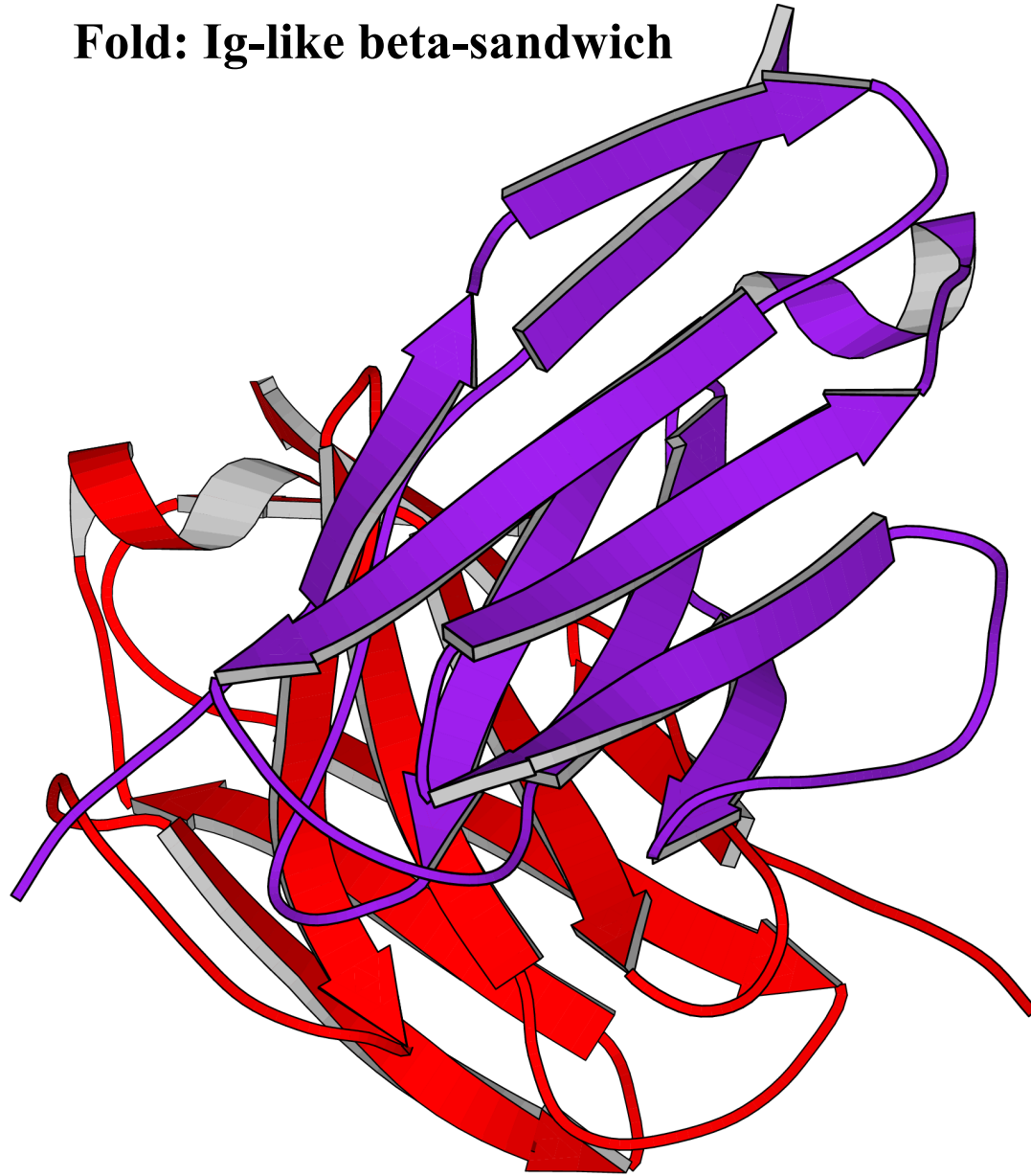
1qmn: Antichymotrypsin, Human

**Class: All alpha**  
**Fold: Globin-like**



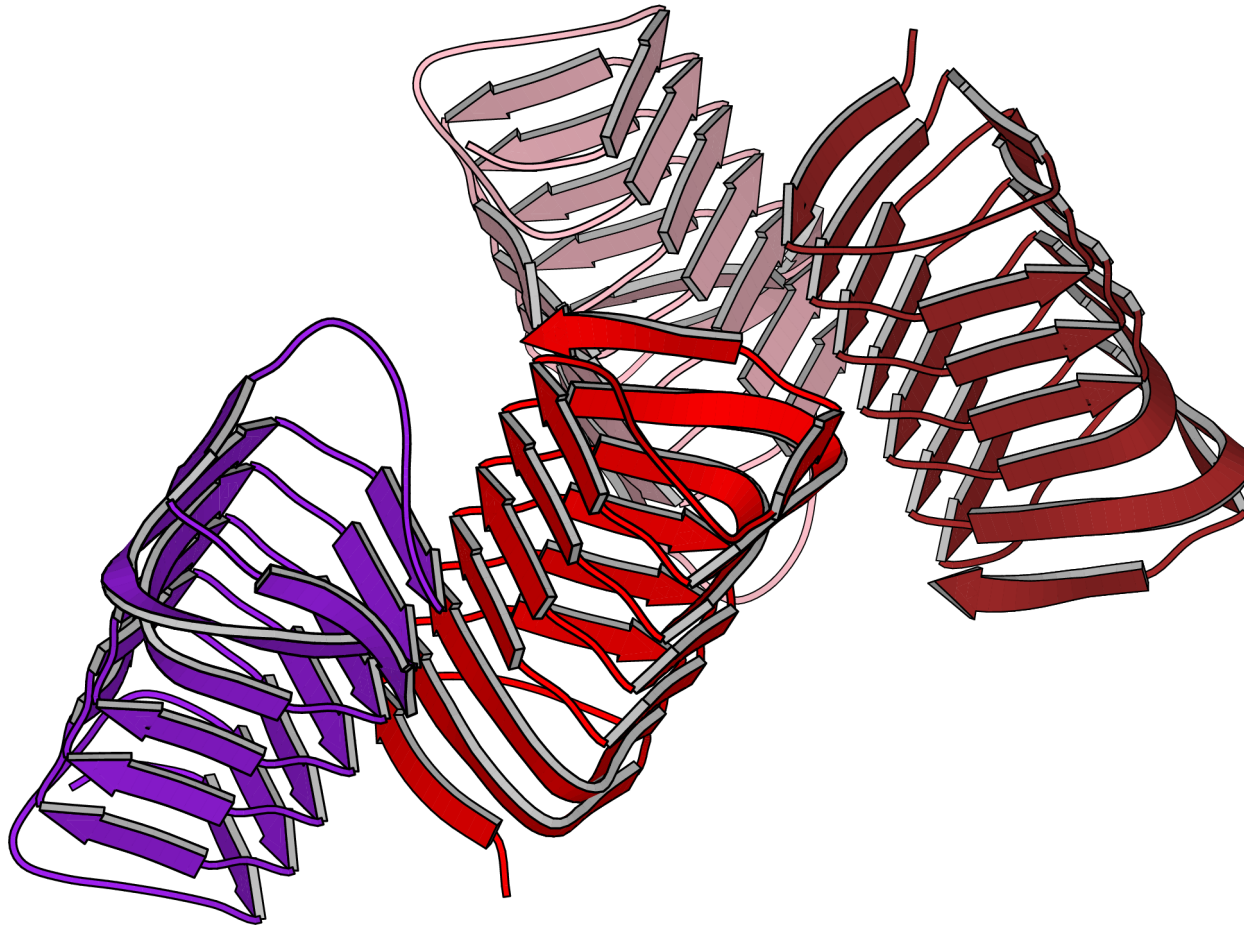
2mm1: Oxygen transport, Human myoglobin

**Fold: Ig-like beta-sandwich**



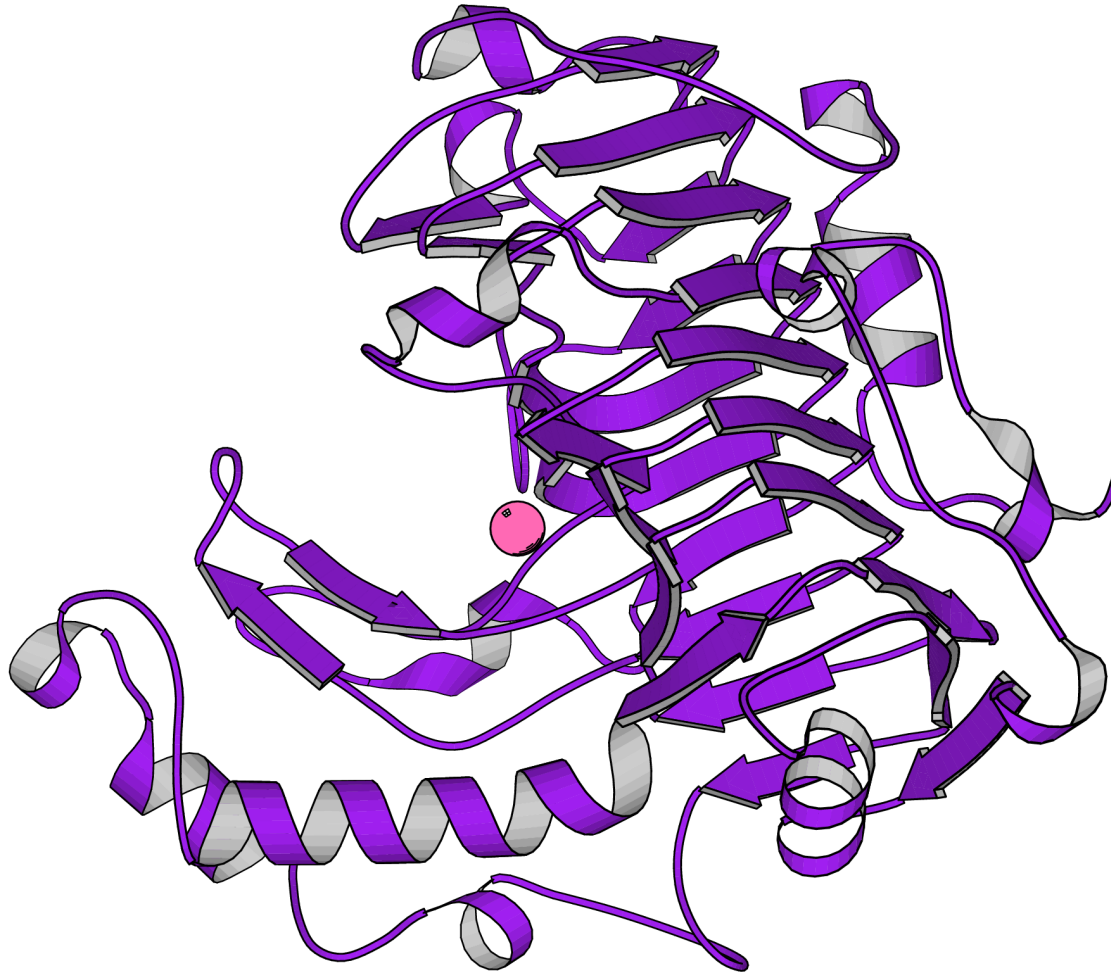
1bww: Ig Kappa V

**Fold: Single-stranded left-handed beta-helix**



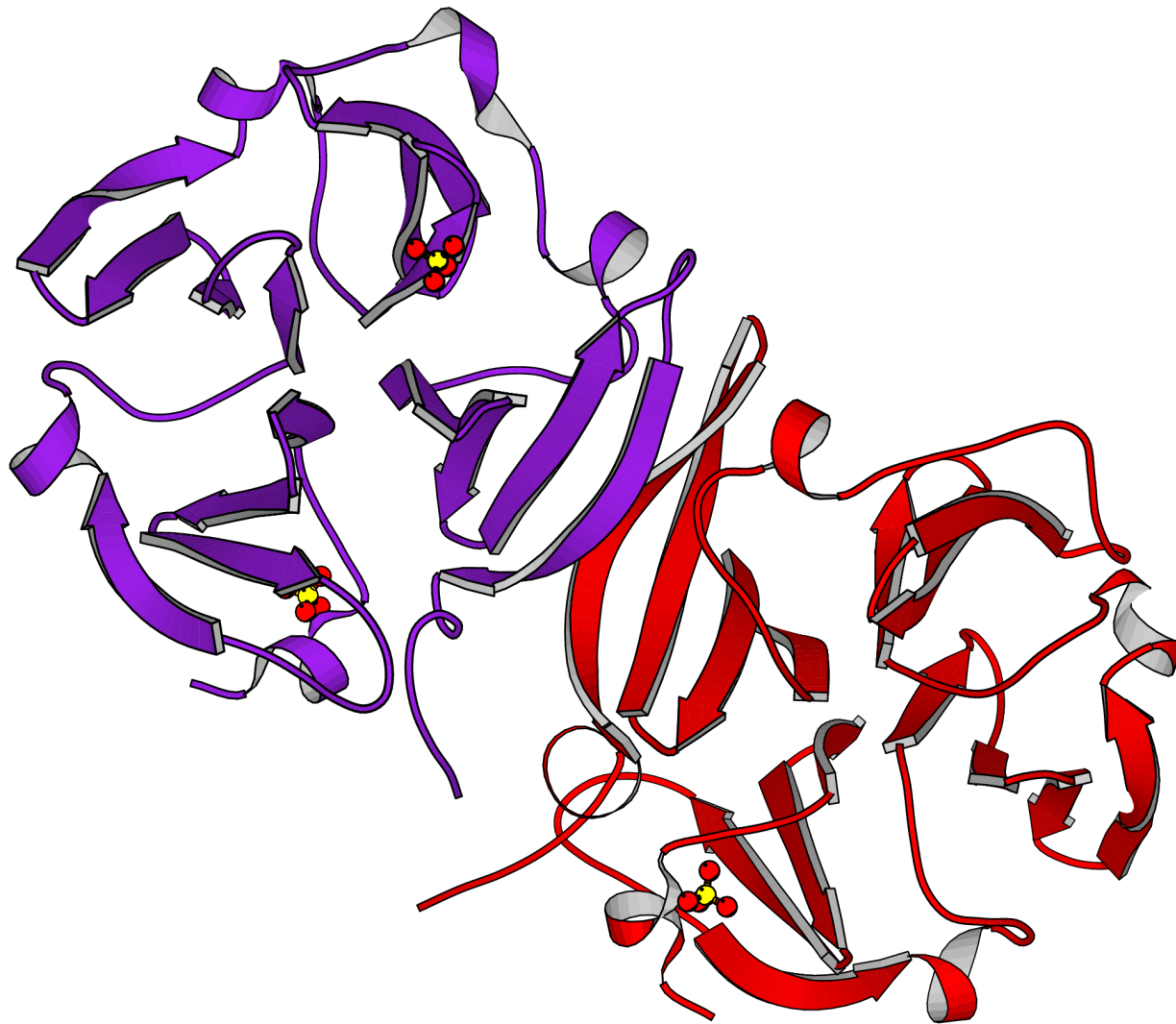
1m8n: Antifreeze protein, Spruce budworm

**Fold: Single-stranded right-handed beta-helix**



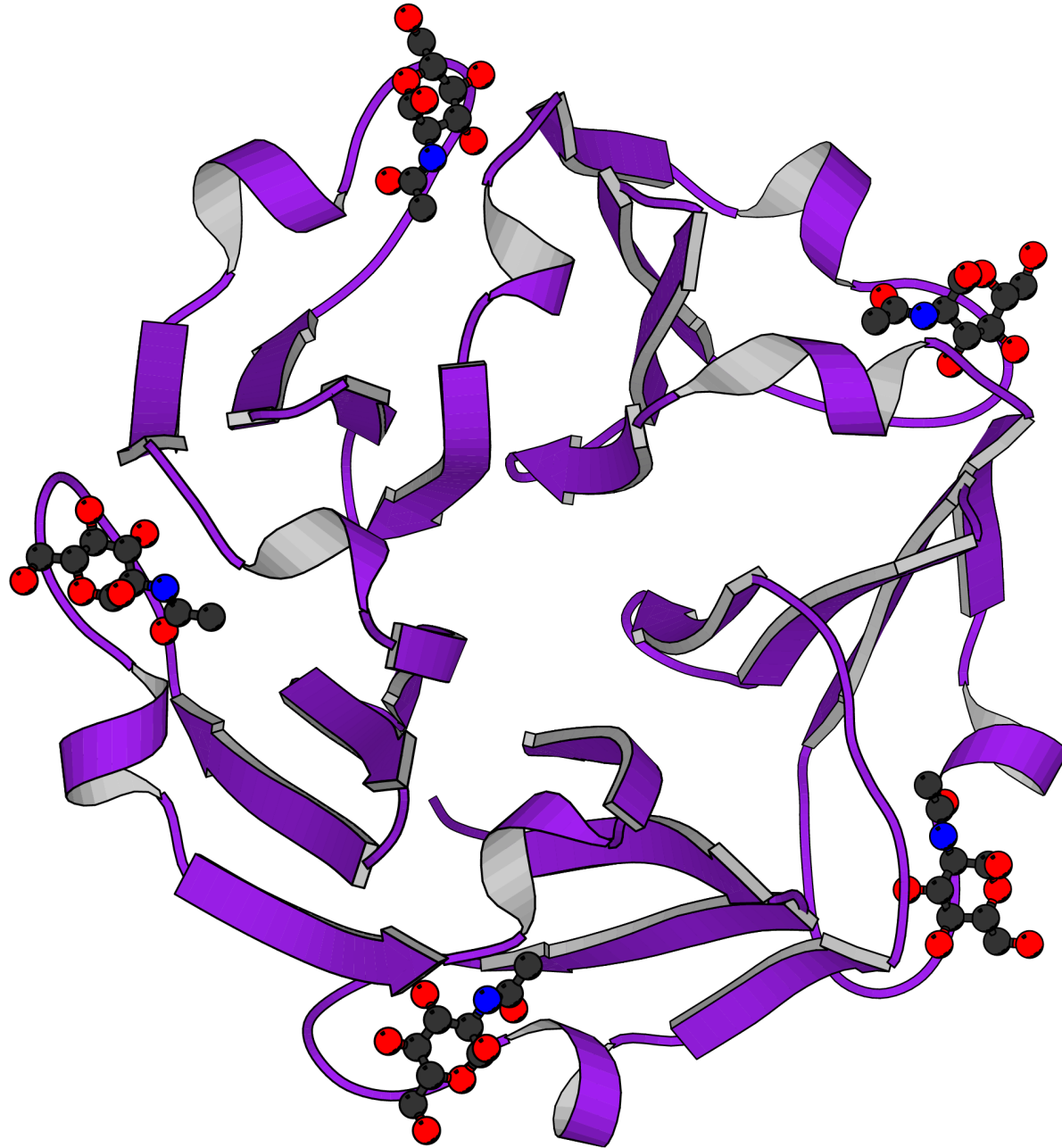
1bn8: Pectate lyase, *B. subtilis*

## Fold: 4-bladed beta-propeller



1itv: Hydrolase, Human gelatinase B (MMP-9)

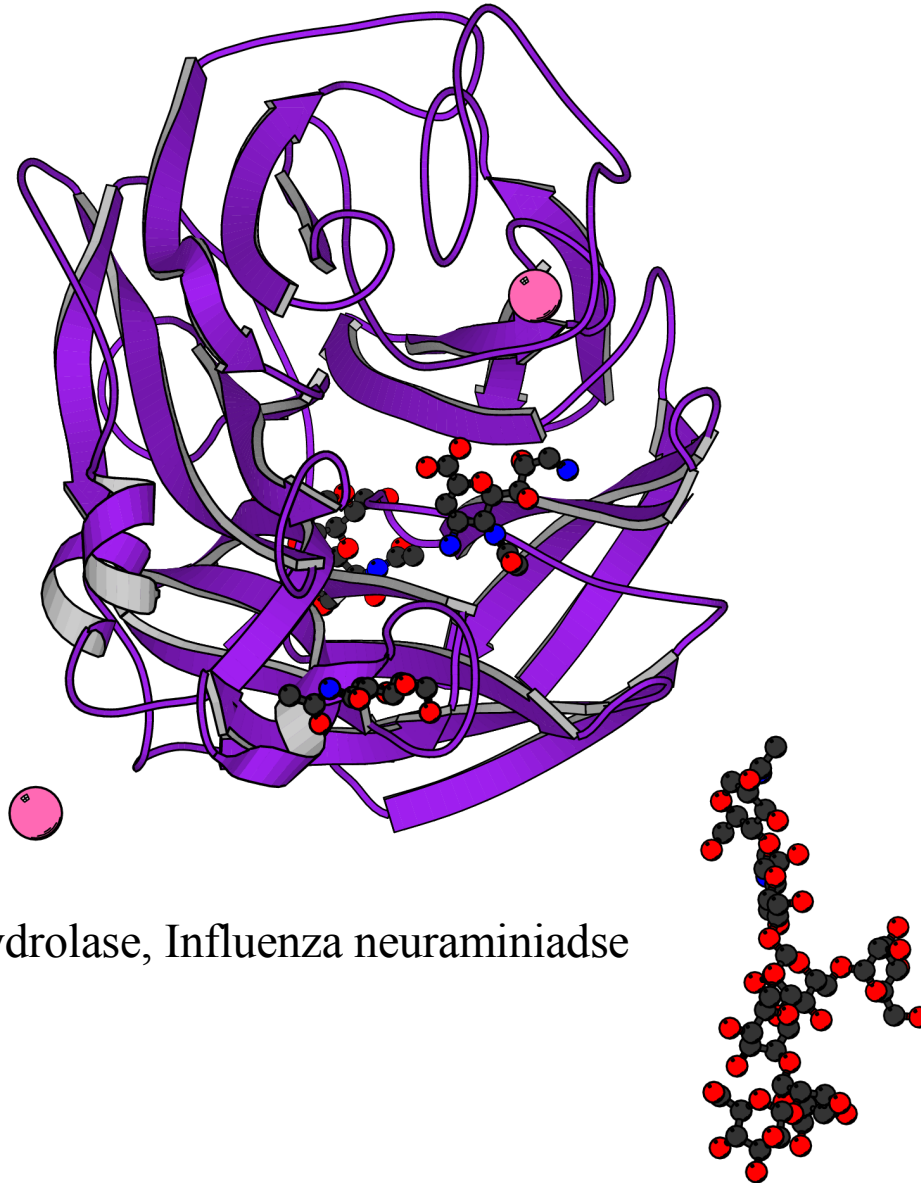
## Fold: 5-bladed beta-propeller



1tl2: Sugar binding protein, Horseshoe crab tachylectin-2

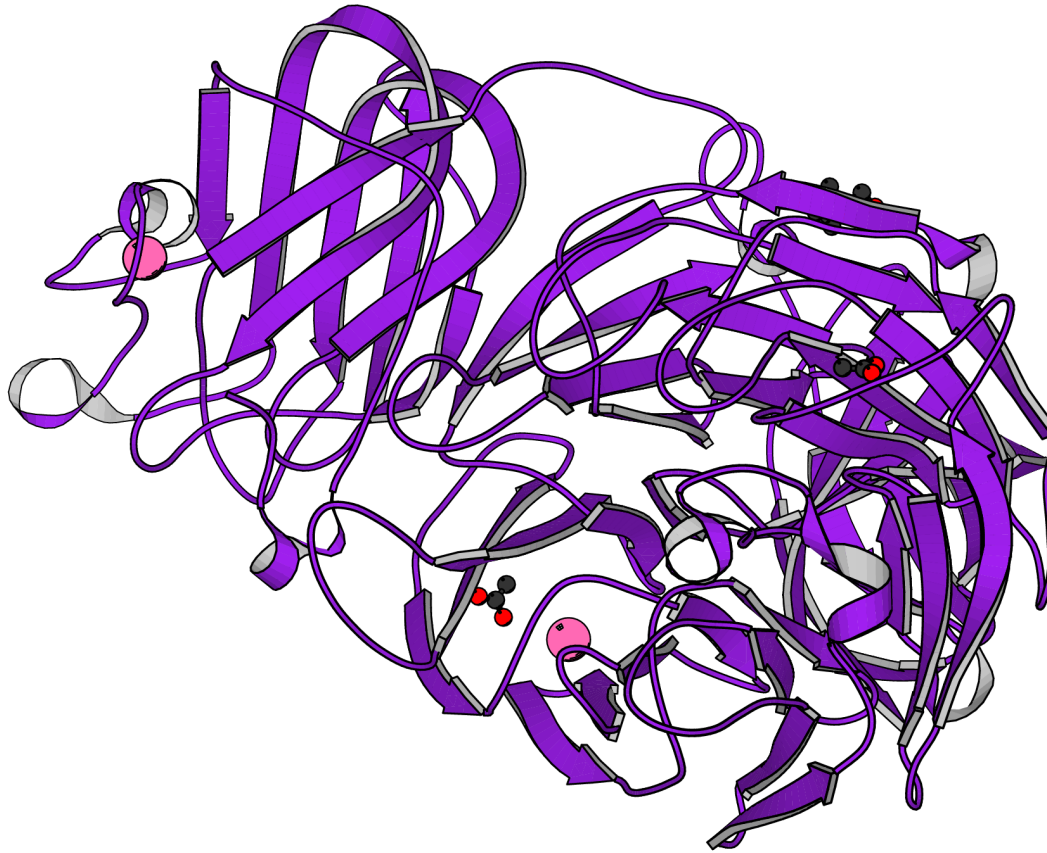


## Fold: 6-bladed beta-propeller



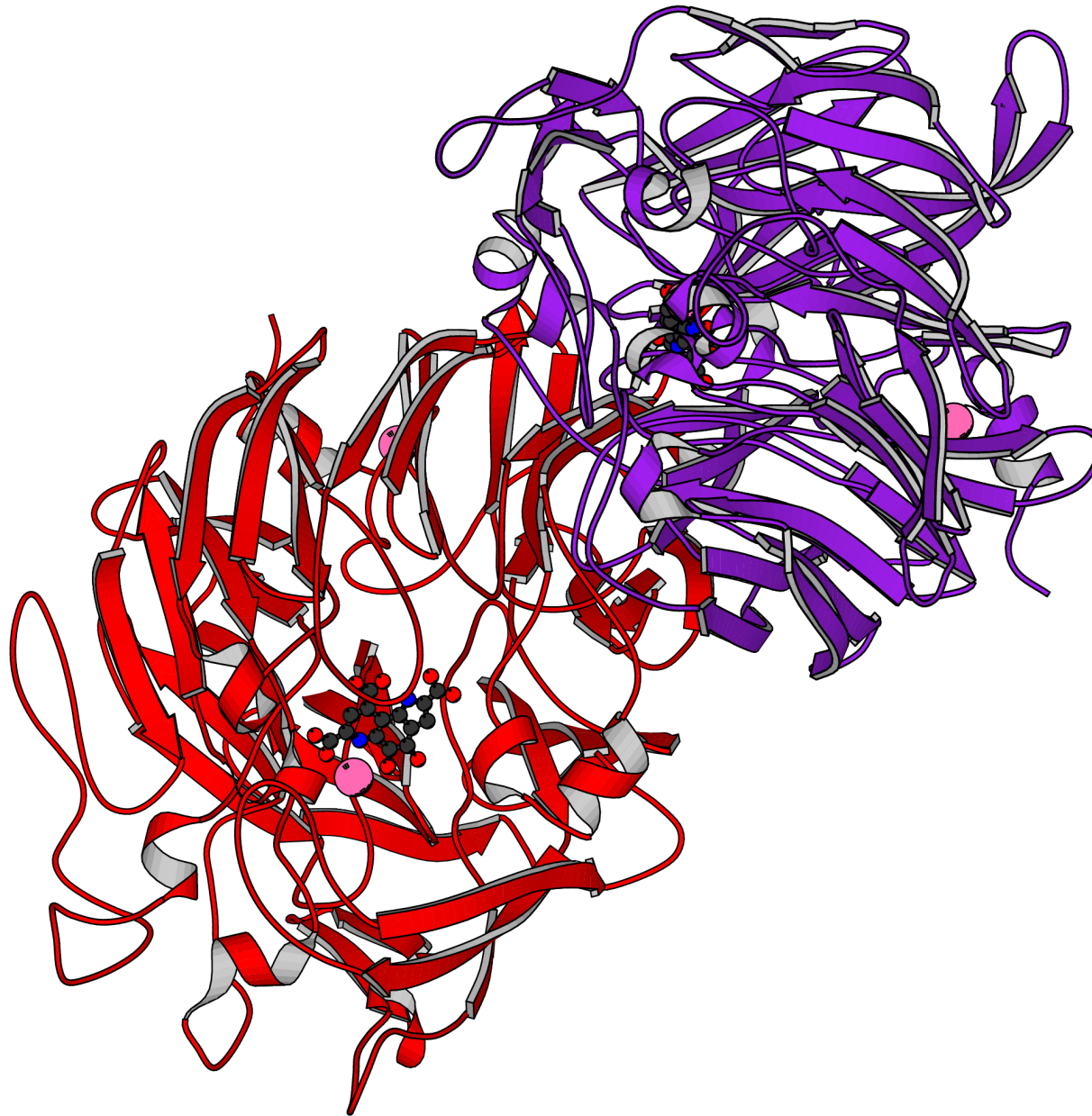
1f8e: Hydrolase, Influenza neuraminiadse

**Fold: 7-bladed beta-propeller**



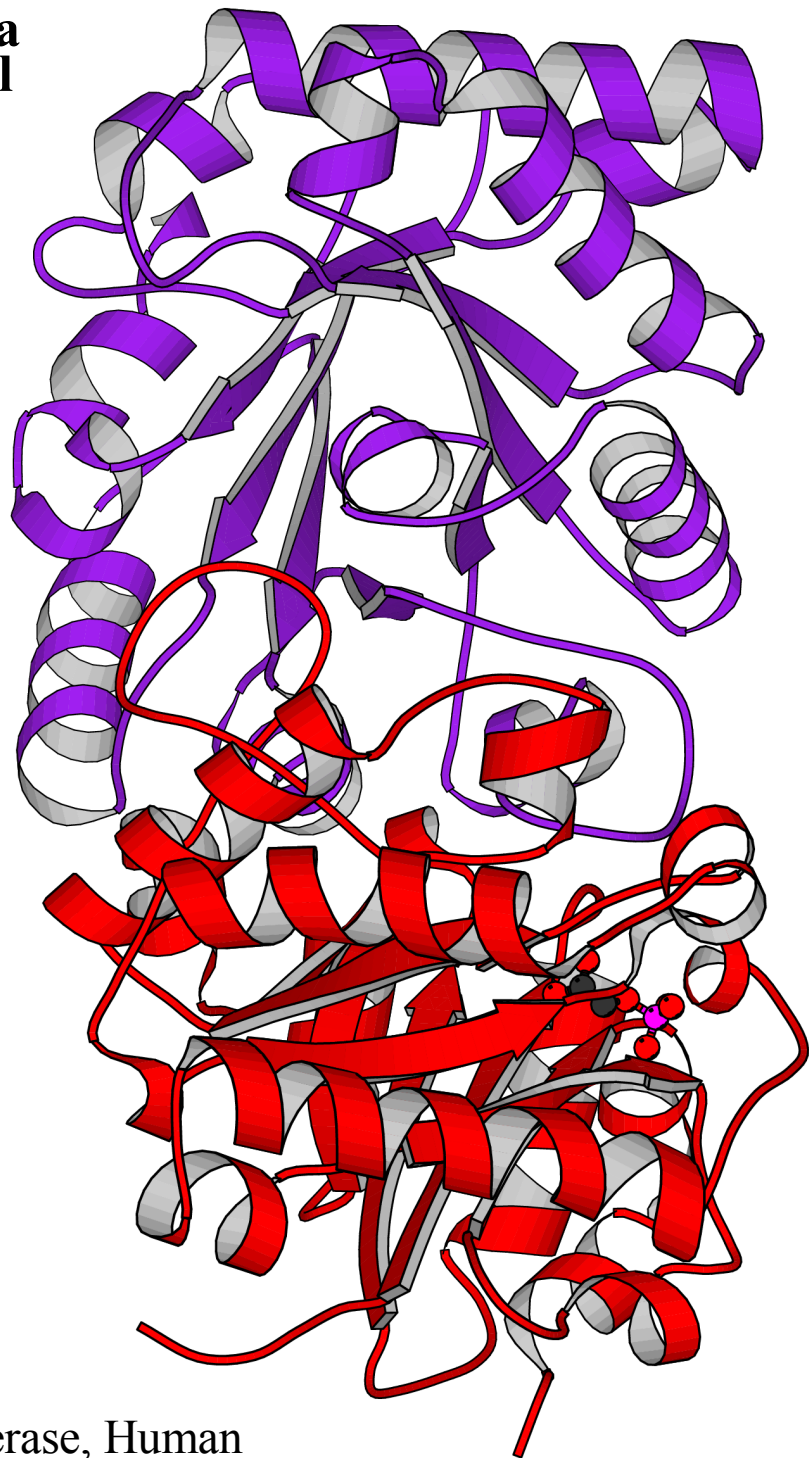
1k3i: Oxidoreductase, Galactose oxidase, Fungi

## Fold: 8-bladed beta-propeller



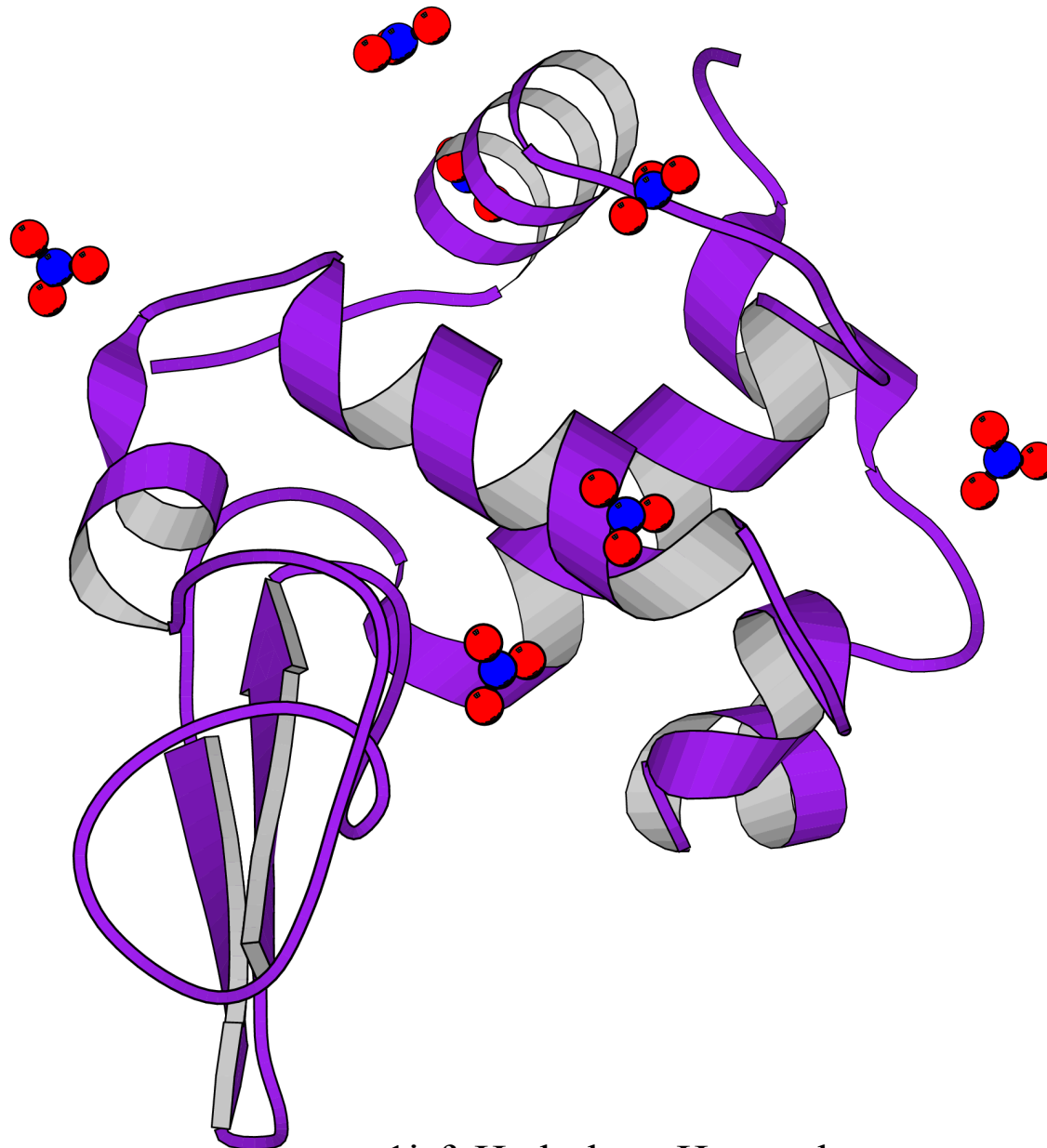
1fig: Oxidoreductase, Quinoprotein ethanol dehydrogenase, Pseudomonas

**Class: Alpha/beta**  
**Fold: TIM barrel**



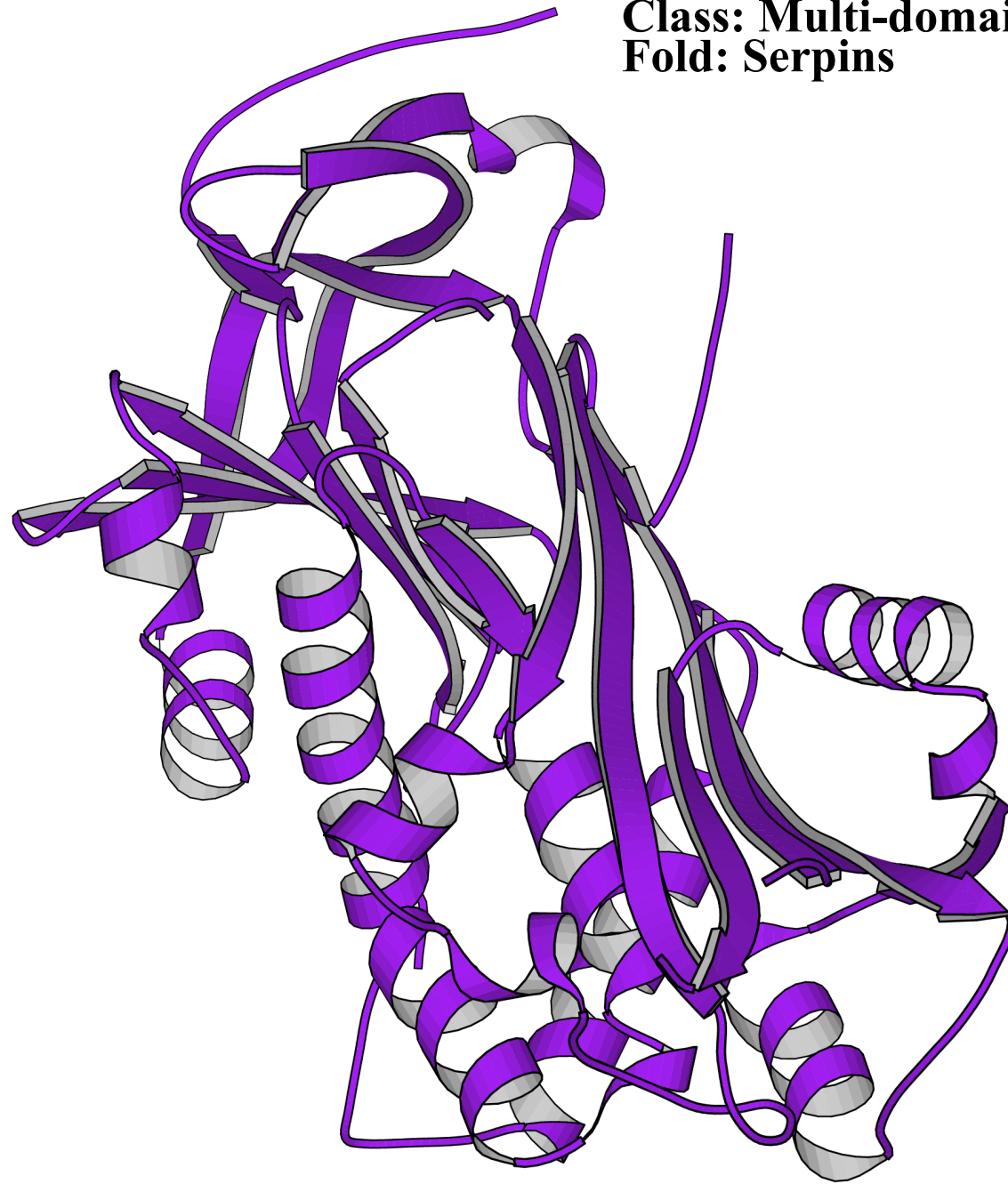
1hti: Triosephosphate isomerase, Human

**Class: Alpha+beta**  
**Fold: Lysozyme-like**



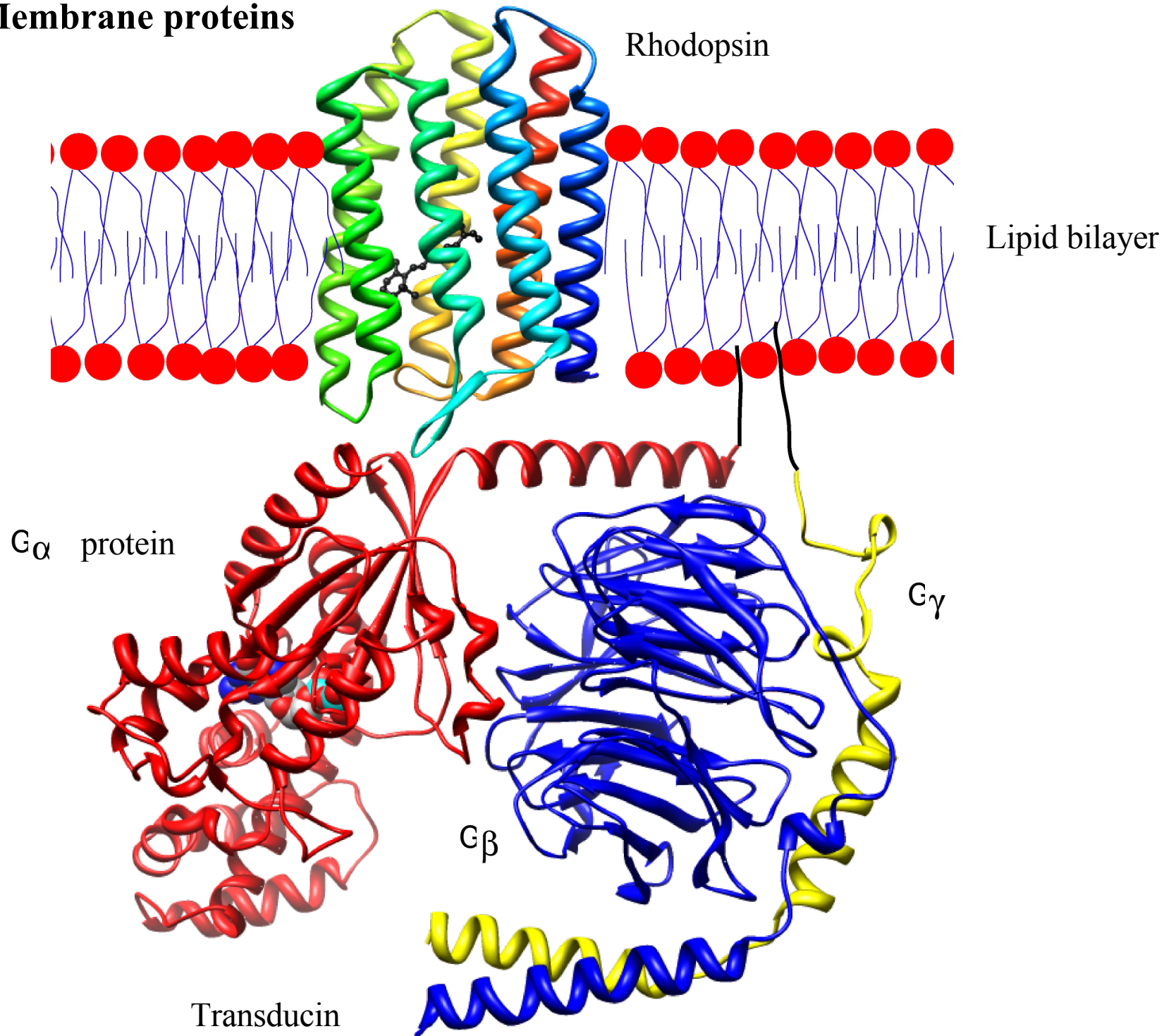
1jsf: Hydrolase, Human lysozyme

**Class: Multi-domain**  
**Fold: Serpins**

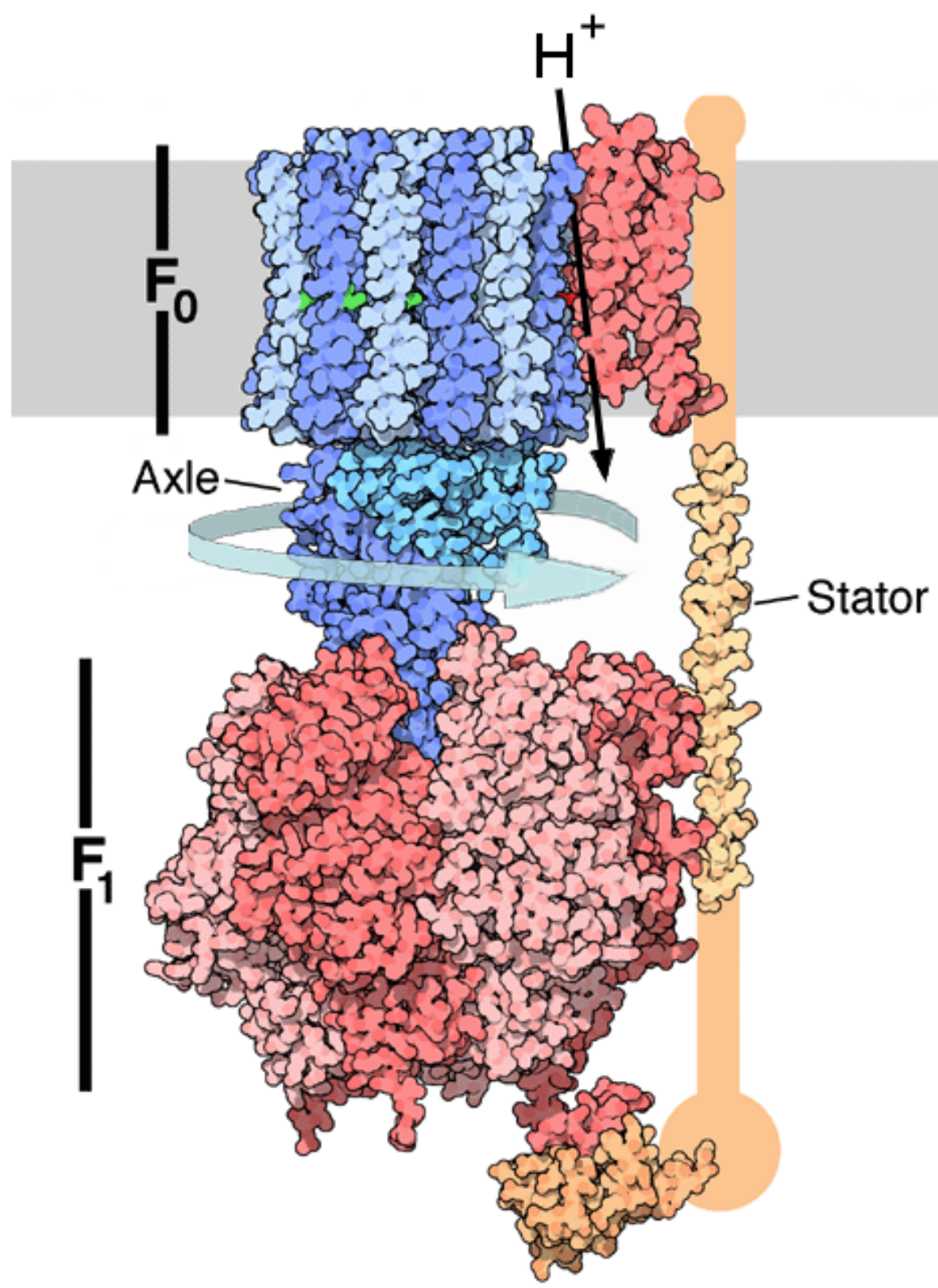


1qmn: Serpin, Human antichymotrypsin

# Class: Membrane proteins



ATP synthase





# 蛋白質のサイズの比較

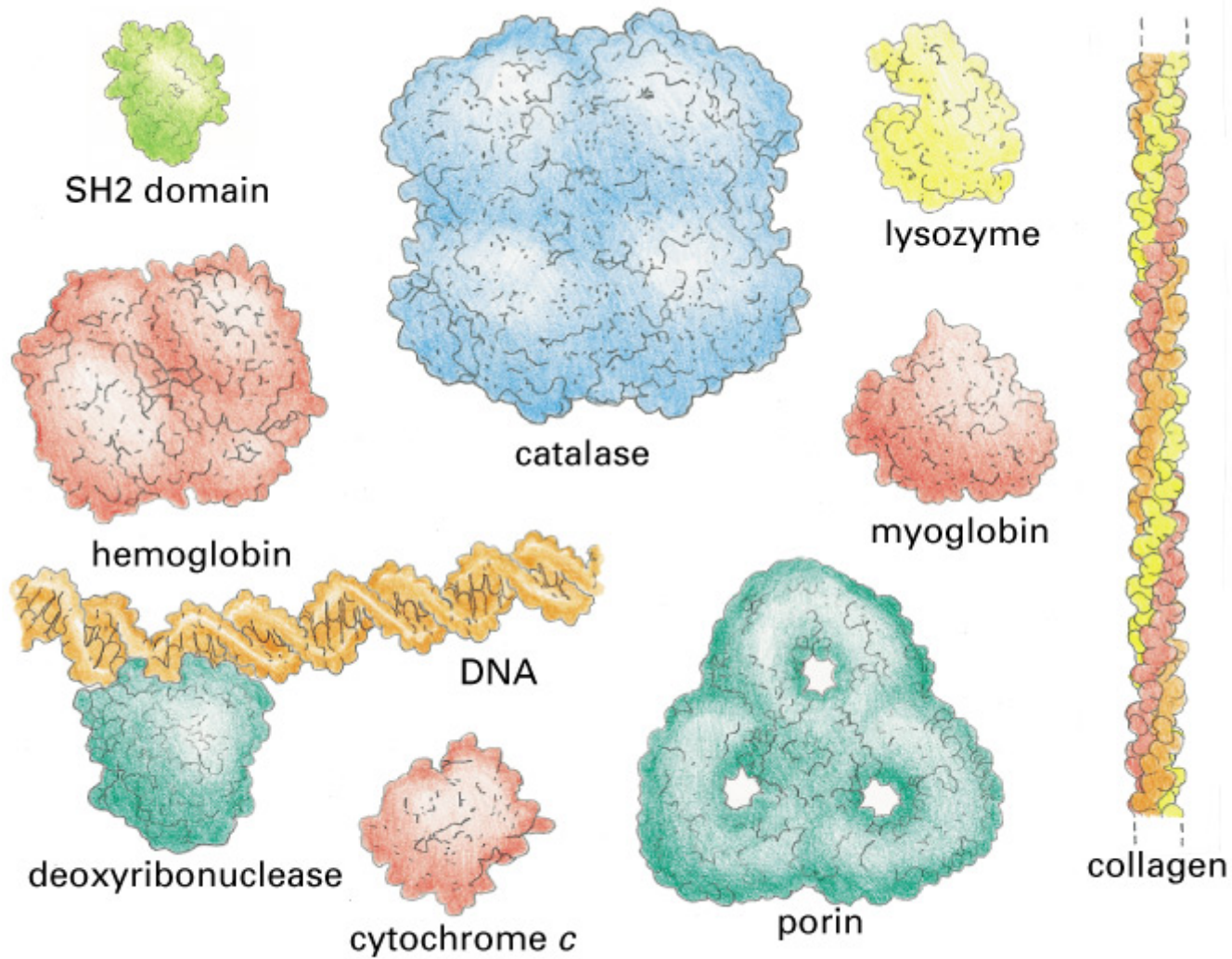


Figure 3-24 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

# 配列レベルでの類似

```

GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAAATTTGACTCAACACGGGAAACCTCACCC
|
GCCGCCTGGGGAGTACGGTGGCAAGACTGAAACTTAAAGGAATTGGCGGGGGAGCACTACAACGGGTGGAGCCTGCGGTTTAAATTTGGATTCAAACGGCGGGCATCTTACCA
|
ACCGCCTGGGGAGTACGGCAGGTTAAAACTCAAANTGAATTGACGGGGGGCCCGC.ACAAGCGGTGGAGCATGTGGTTTAAATTTGGATTCAAACGGCAAGAACTTACCT
|
GTTCCGGGGGGAGTATGGTTGCAAAGCTGAAACTTAAAGGAATTGACGGAAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAAATTTGACTCAACACGGGAAACCTCACCC

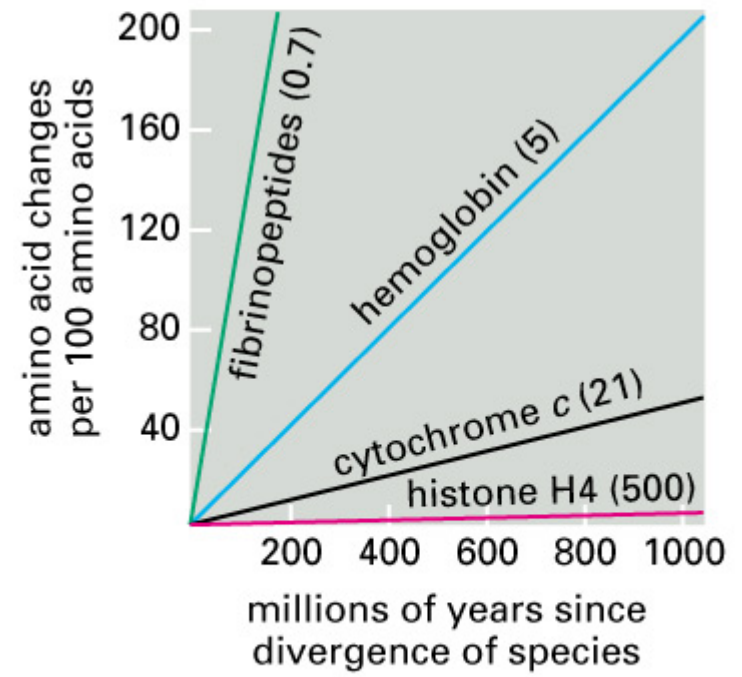
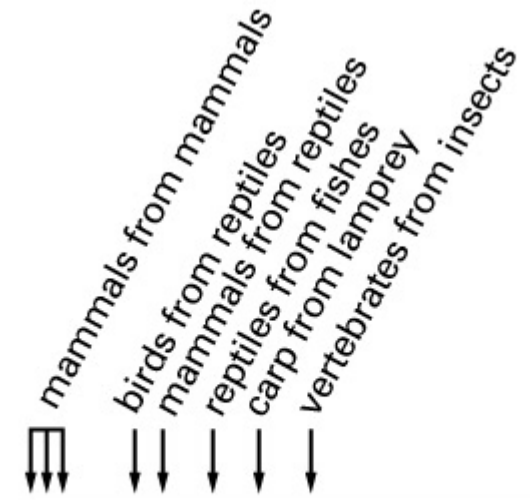
```

human  
*Methanococcus*  
*E. coli*  
human

## 分子進化学のあゆみ

— 進化は遺伝情報(DNA)の変化に刻まれる。

- 1901 抗体反応による類縁関係の推定(Nutall)
- 1962 分子進化速度一定性 (Zuckermandle & Pauling)
- 1967 Cytochrome Cの系統樹(Fitch & Margoliash)
- 1967 分子時計によるヒトの起源(Sarich & Wilson)
- 1968 分子進化の中立説 (木村資生)  
 分子レベルでの変異は、致死的か中立的。  
 中立ならば 分子進化速度 = 中立突然変異率.
- 1970 遺伝子の重複による進化(大野乾)
- 1977 rRNAの系統樹による古細菌の発見(Woese)
- ...



進化機構： 遺伝子重複／混成／変換による多様化

表 4. ウイルス遺伝子と核およびオルガネラ遺伝子の進化速度  
(同義置換速度)

ウイルス	エイズウイルス(HIV-1)	$3.2 \times 10^{-2}$
	インフルエンザウイルスA型	$1.1 \times 10^{-2}$
	〃 B型	$0.21 \times 10^{-2}$
	〃 C型	$0.14 \times 10^{-2}$
	デング熱ウイルス(デング 2)	$0.25 \times 10^{-2}$
	日本脳炎ウイルス	$< 0.28 \times 10^{-2}$
	ウシ口蹄疫ウイルスO型	$0.12 \times 10^{-2}$
	〃 C型	$0.10 \times 10^{-2}$
	パラインフルエンザ 3	$< 0.29 \times 10^{-2}$
	核遺伝子	哺乳類
齧歯類		$> 6.2 \times 10^{-9}$
棘皮動物		$5.6 \times 10^{-9}$
高等植物 <sup>a)</sup>		$7.1 \times 10^{-9}$
ミトコンドリア		
ミトコンドリア	類人猿 <sup>b)</sup>	$55.0 \times 10^{-9}$
	高等植物 <sup>a)</sup>	$0.8 \times 10^{-9}$
葉緑体	高等植物 <sup>a)</sup>	$2.6 \times 10^{-9}$

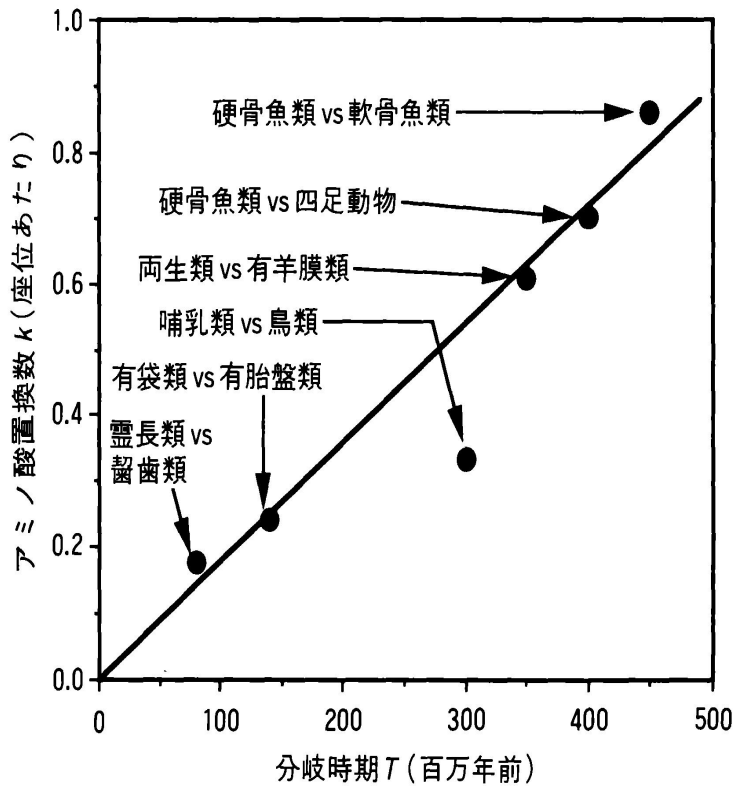


図 5. ヘモグロビンの分子時計

a) 単子葉/双子葉の分岐を 1 億年前とした。

b) ヒト/チンパンジーの分岐を 500 万年前とした。



# 配列アライメント

CTG - C - C C -  
 CTGTCTCCT

## 最良アライメント

個々の文字の対応に適切な評価値を与え、アライメントに関する総和 (アライメントスコア) が最大のアライメントと計算する。

### Needleman-Wunsch 法 (動的計画法)

評価値の例: 一致 = 1, 不一致 = -1/3, 欠失/挿入 = -1

$$s(i, j) = 1 \text{ or } -1/3$$

$$g = -1$$

$$S(i, 0) = S(0, i) = i * g$$

$$S(i, j) = \max \begin{cases} S(i-1, j-1) + s(i, j) \\ S(i-1, j) + g \\ S(i, j-1) + g \end{cases}$$

$$S(3, 2) = 1 \begin{matrix} 3 \\ G \\ T \\ 4 \end{matrix} - 1/3 = 2/3$$

$$S(3, 3) = 3 \begin{matrix} \bar{\bar{}} \\ T \\ 4 \end{matrix} - 1 = 2 \quad S(4, 3) = 2$$

$$S(4, 2) = 0 \begin{matrix} 3 \\ G \\ \bar{\bar{}} \end{matrix} - 1 = -1$$

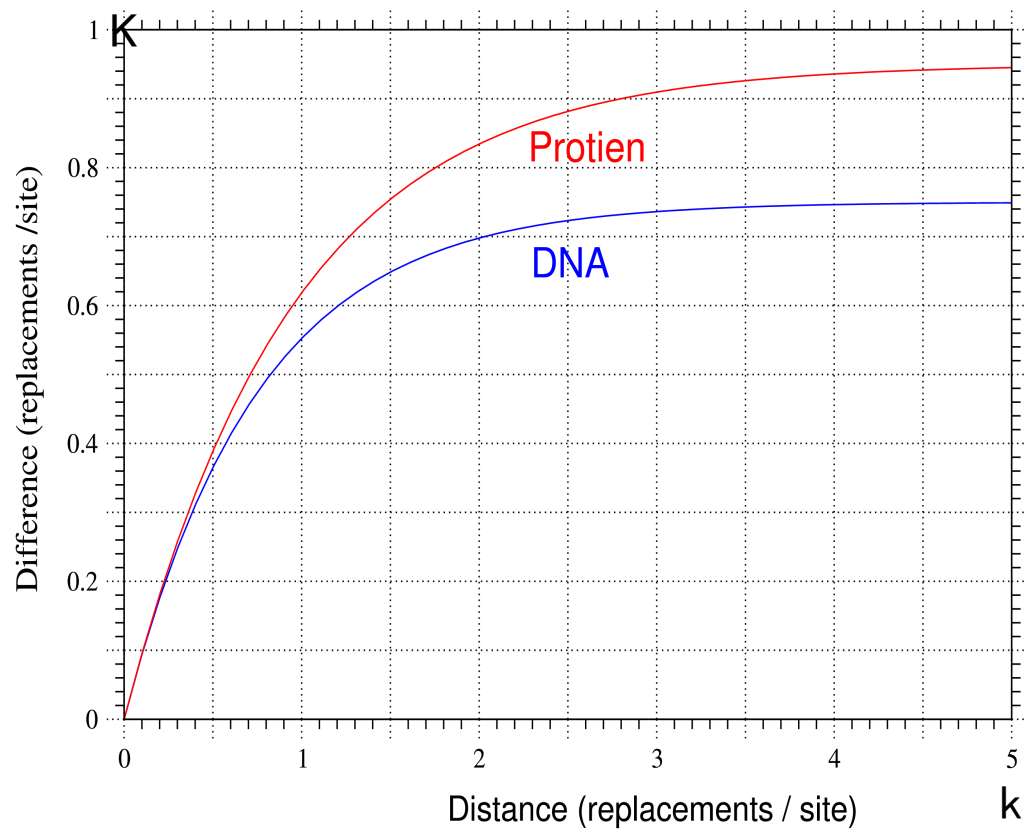
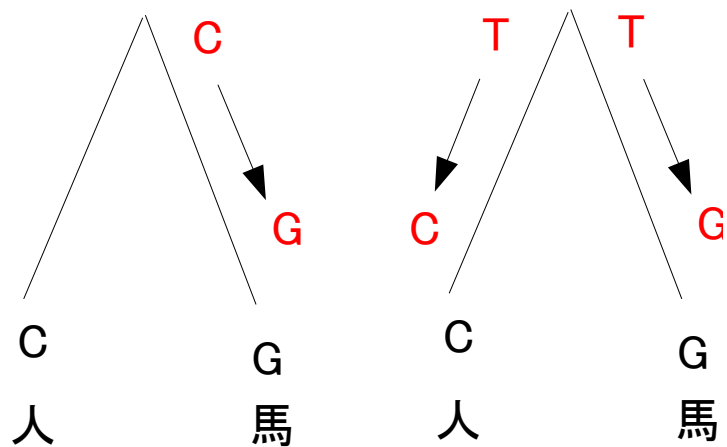
i \ j	0	1	2	3	4	5	6
配列		C	T	G	C	C	C
0	0	-1	-2	-3	-4	-5	-6
1 C	-1	+1 -1	1	0	-1	-2	-3
2 T	-2	0	2	1	0	-1	-2
3 G	-3	-1	1	3	2	1	0
4 T	-4	-2	0	-1/3 -1	2	1.7	0.7
5 C	-5	-3	-1	1	3	3.7	2.7
6 T	-6	-4	-2	0	2	2.7	3.3
7 C	-7	-5	-3	-1	1	3	3.7
8 C	-8	-6	-4	-2	0	2	4
9 T	-9	-7	-5	-3	-1	1 -1/3 -1	3

## 多重置換の補正

	V	L	S	P	A
人	GTG	CTG	TCT	CCT	GCC
馬	GTG	CTG	TCT	GCC	GCC
	V	L	S	A	A

DNA:  $K_d = (1 - \exp(-4 k_d / 3)) 3 / 4$

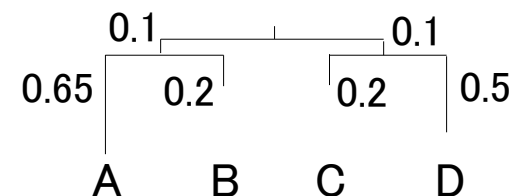
Protein:  $K_p = (1 - \exp(-20 k_p / 19)) 19 / 20$



# 系統樹推定

## 1. 距離行列法

		距離			
		A	B	C	D
差異	A		0.85	1.05	1.35
	B	0.51		0.60	0.90
	C	0.57	0.41		0.70
	D	0.63	0.52	0.46	

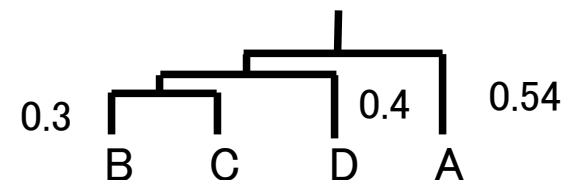
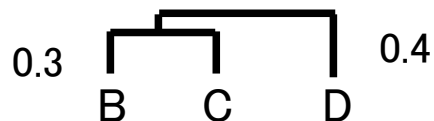


### 1.1 UPGWA法; 分岐した2つの枝の長さは等しいと仮定

	A	B	C	D
A		0.85	1.05	1.35
B	0.85		0.60	0.90
C	1.05	0.60		0.70
D	1.35	0.90	0.70	

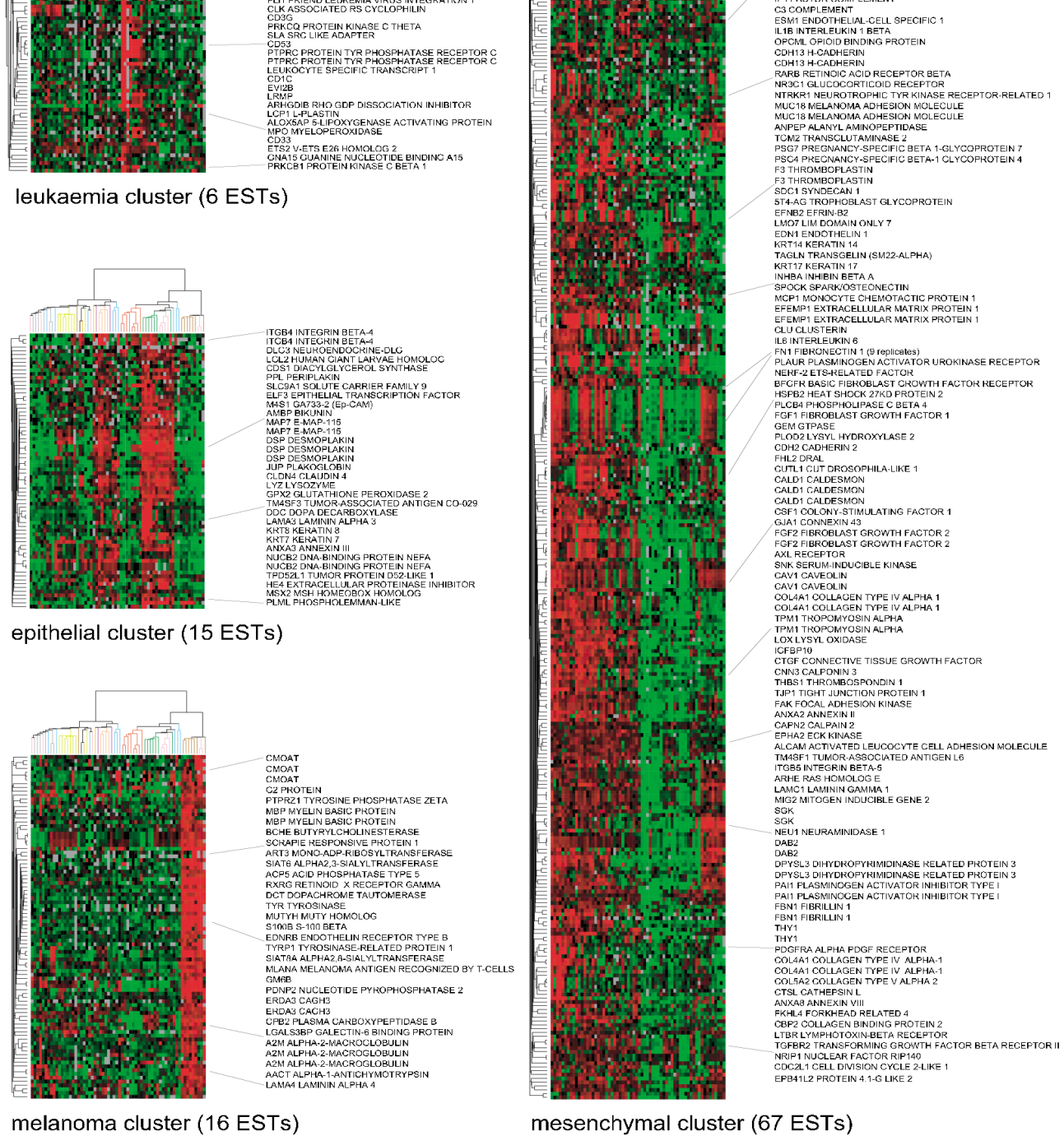
	A	BC	D
A		0.95	1.35
BC	0.95		0.80
D	1.35	0.80	

	A	BCD
A		1.08
BCD	1.08	



# 遺伝子発現プロファイル の階層的クラスタリング

© 2000 Nature America Inc. • <http://genetics.nature.com>



**Fig. 3** Gene clusters related to tissue characteristics in the cell lines. Enlargements of the regions of the cluster diagram in Fig. 1 showing gene clusters enriched for genes expressed in cell lines of ostensibly similar origins. **a**, Cluster of genes highly expressed in the leukaemia-derived cell lines. Two sub-clusters distinguish genes that were expressed in most leukaemia-derived lines from those expressed exclusively in the erythroblastoid line, K562 (note that the triplicate hybridizations cluster together). **b**, Cluster of genes highly expressed in all colon (7/7) cell lines and all breast-derived cell lines positive for the oestrogen receptor (2/2). This set of genes was also moderately expressed in most ovarian lines (5/6) and some non-small-cell-lung (4/6) lines, but was expressed at a lower level in all renal-carcinoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB-231 and MDA-MB-435) (Ross et al., Nat. Genetics, 24, 227, 2000). **c**, Cluster of genes highly expressed in all melanoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB-231 and MDA-MB-435) (Ross et al., Nat. Genetics, 24, 227, 2000). **d**, Cluster of genes highly expressed in all mesenchymal cell lines (67/67) and two related lines ostensibly derived from breast cancer (MDA-MB-231 and MDA-MB-435) (Ross et al., Nat. Genetics, 24, 227, 2000). Names are shown only for all known genes whose identities were independently re-



# 系統樹推定

## 1. 距離行列法

1.1. UPGWA法; 分岐した2つの枝の長さは等しいと仮定

1.2. 近隣結合法 (N-J法); 距離が正しいならば、正しい木、距離を与える。

2. 節約法; 必要最低限の置換総数に関し最少の木を計算

3. 最尤法; 置換確率、木に関し尤度を極大化

$$\arg \max_T \log P(D | T)$$

## 4. ベイズ法

$$\arg \max_T ( \log P(D | T) + \log P(T) )$$

$$\log P(T | D) = \log P(D | T) + \log P(T) - \log P(D)$$

T: Tree, D: Data

# グロビン蛋白質のアライメント

α  
ヘモグロビン  
β  
ミオグロビン

```

1IRD-A  -----VLSPADKTNVKAAWGKVGAGHAGEYGAELERMFLSFPTTKTYFPHF-----DLSHGSAQVKGHGKKVADALTNVAHAVDD
1IBE-A  -----VLSAADKTNVKAWSKVGGHAGEFCAELERMFLGFPTTKTYFPHF-----DLSHGSAQVKAHGKKVGDALTLAVGHIDD
1HBR-A  -----MLTAEDKKLIQQAWKAASHQEEFGAELTRMFTTYPQTKTYFPHF-----DLSPGSDQVRGHGKKVLGALGNVKNVND
1V4X-A  -----TTLSDKDKSTVKALWGIKSKSADAIADALGRMLAVYPQTKTYFSHWP-----DMSPGSGPVKAHGKKVMGVALAVSKIDD
1IRD-B  -----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN
1IBE-B  -----VQLSGEEKAAVLAALWDKVN--EEVCGEALGRLLVVYPWTQRFFESFGDLSNPQAVMGNPKVKAHGKKVLHFSFGEGVHHLDN
1HBR-B  -----VHWTAEKQLITGLWGKVN--VAECGAELARLLIVYPWTQRFFASFGNLSSTPTAILGNPMVRAHGKKVLTFSFGDAVKNLDN
1V4X-B  -----VEWTQQERSIAGIFANLN--YEDIGPKALARCLIVYPWTQRYFGAYGDLSTPDAIKGNAKIAAHGVKVLHGLDRAVKNMDN
NP_005359 -----MGLSDGEWQLVLNVWGVKVEADIPGHGQEVLIIRLFKGHPEETLEKFDKFKHLKSEDEMKASEDLKKGATVLTALGGILKKKGH
1GJN    -----GLSDGEWQQLVLNVWGVKVEADIAGHQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKKGHTVLTALGGILKKKGH
1A6M    -----VLSEGEWQVLVHVWAKVEADVAGHQDILIRLFKSHPEETLEKFDKFKHLKTEAEMKASEDLKKGHTVLTALGAILKKKGH
2LHB    PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSQVDILVKEFTSTPAAQEFKFKGLTTADELKKSDVVRWHAERIINAVDDAVASMD
1FSL    -----VAFTEKQDALVSSSEFAFKANIPQYSVVFYTSILEKAPAAKDLFSLAN-----GVDPTNPKLTGHAEKLFALVRDSAGQLKA
consensus 1.....10.....20.....30.....40.....50.....60.....70.....80.....

```

Hemoglobin a  
馬  
まわとり  
まくち

```

---MPNALSAISDLHAHKLRVDPVNFKLISHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTTSKYR-----
---LPGALSDLSNLHAHKLRVDPVNFKLISHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTTSKYR-----
---LSQAMAELSLSLHAYNLRVDPVNFKLISQCIQVVLAVHMGKDYTPVHAAFDFKFLSAVSAVLAEKYR-----
---LTTGLGDLSELHAEKMRVDPVNFKILSHCILVVAKMFPKEFTPDHAVSLDKFLASVALALAEYR-----
---LKGTFATLSELHCDKLHVDPENFRLLGNVLVLCVLAHHFGKEFTTPVQAAYQKVVAGVANALAHKYH-----

```

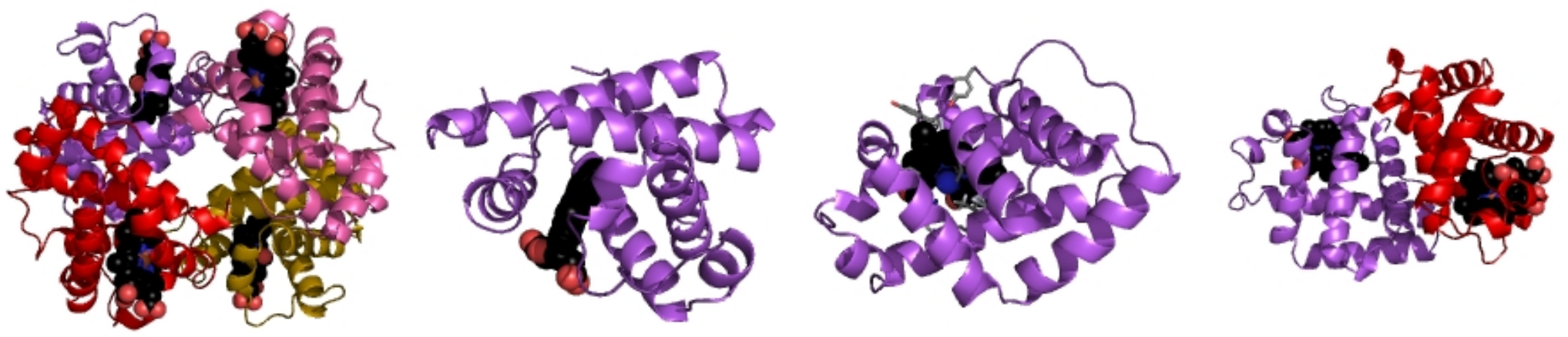
Hemoglobin β  
Myoglobin 人  
馬  
まっこうくら

```

---IKNTFSQLSELHCDKLHVDPENFRLLGDILITVLAAHFSKDFTEPCQAAWQKLVRVVAHALARKYH-----
---INEAYSELVLSHSDKLHVDPENFRILGDCLTVVIAANLGDFTVETQCAFQKFLAVVVFALGRKYH-----
---HEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
---HEAELKPLAQSHATKHKIPVKYLEFISDAIIHVLHSHKHPGDFGADAQGAMTKALELFRNDIAAKYKELGFQG
---HEAELKPLAQSHATKHKIPVKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIKAYKELGY--
TEKMSMKLRNLSGKHAKSFQVDEYFKVLA AVIADTVAG-----DAGFEKLMSMICILRSAY-----
SG-TVVADAAAGSVHAQKAVTDP-QFVVVKEALLKTIKAAVGDKWSDELSRAWEVAYDELAALIKKA-----
consensus 1.....*.....*.....

```

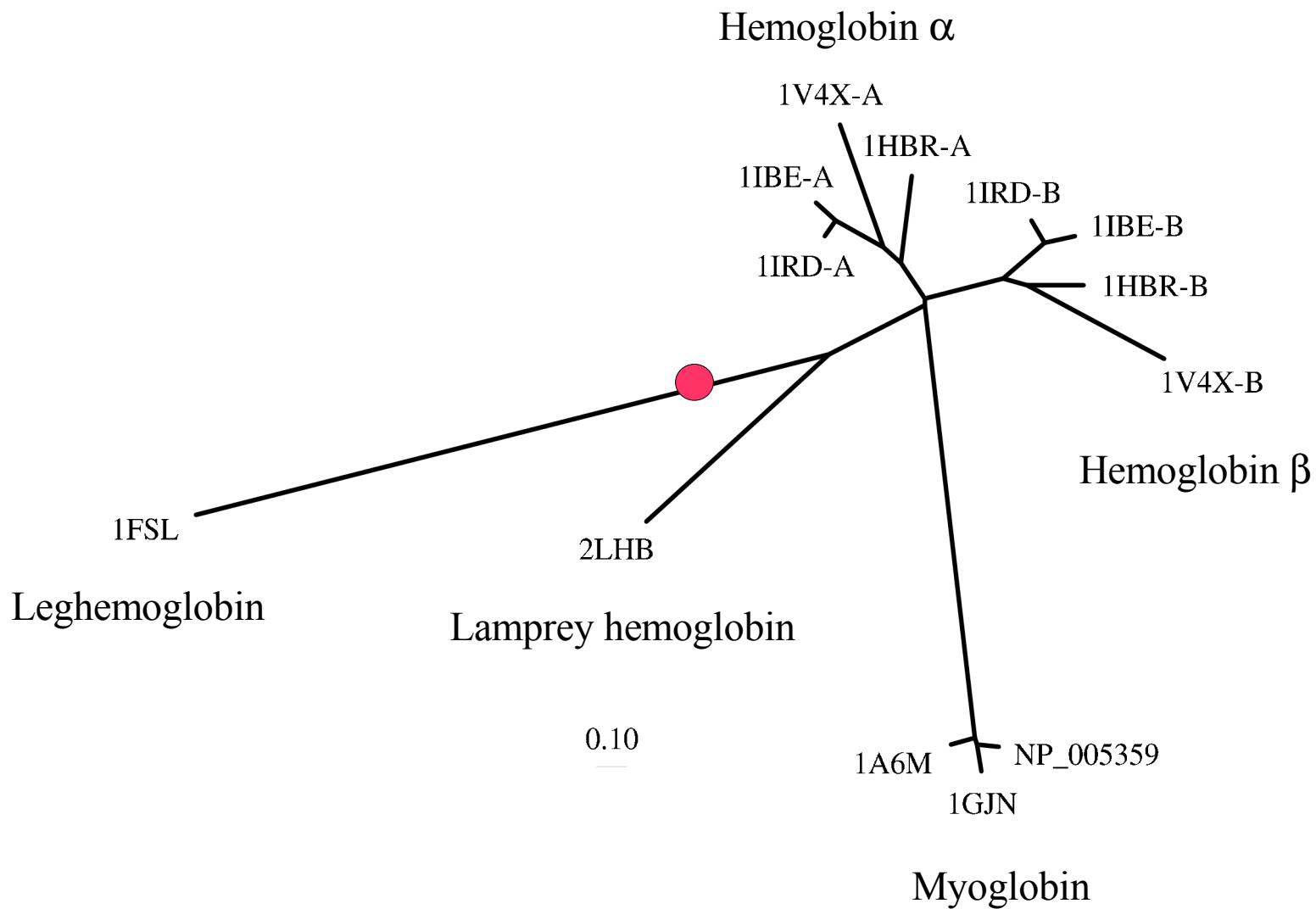
91.....100.....110.....120.....130.....140.....150.....160...



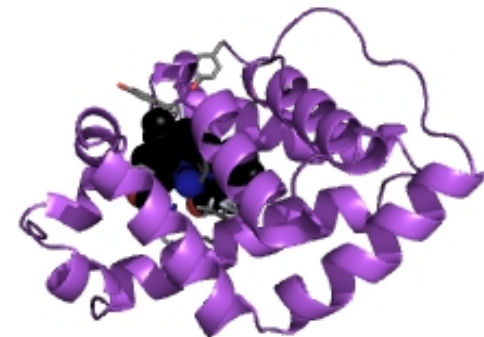
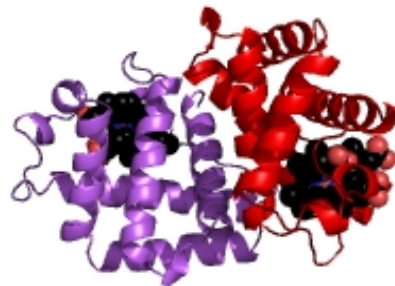
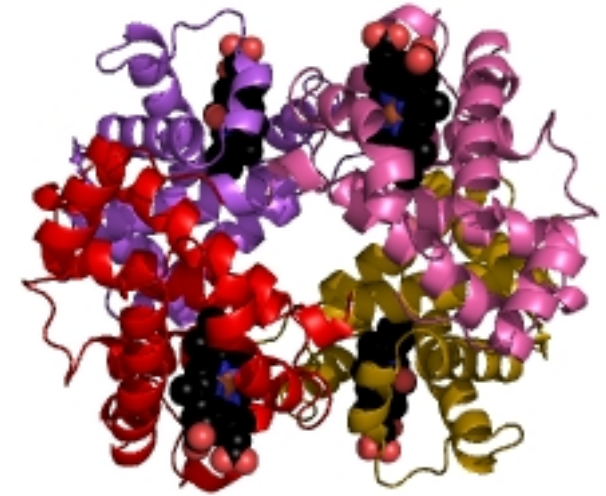
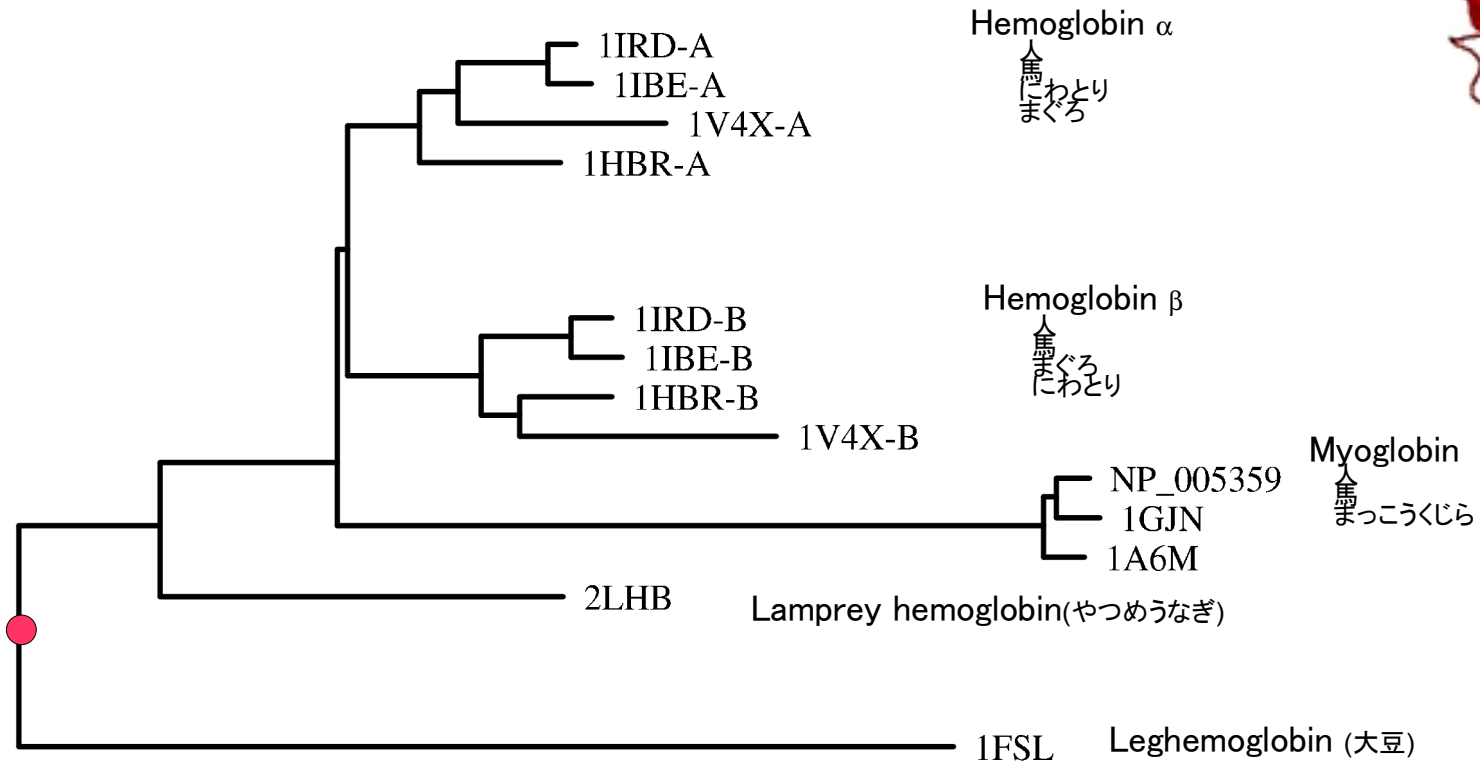
## グロビン蛋白質間の差異と距離

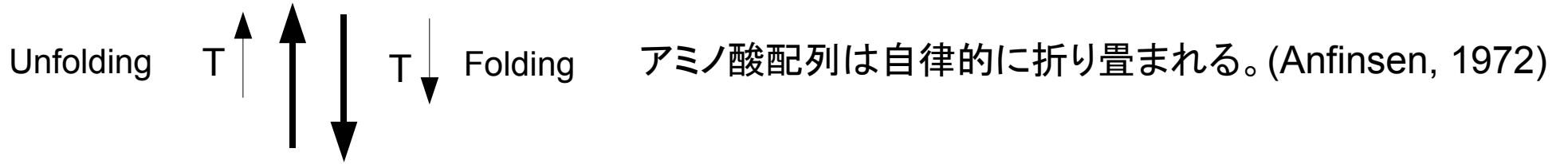
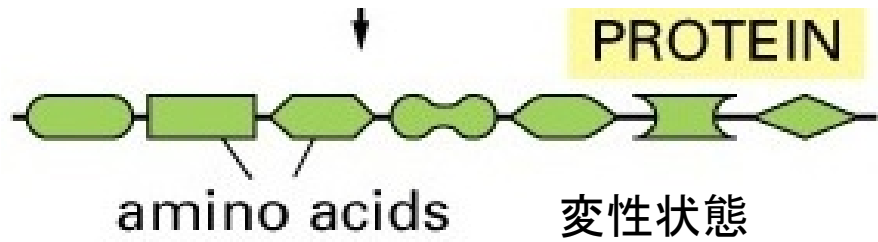
	1IRD-A	1IBE-A	1HBR-A	1V4X-A	1IRD-B	1IBE-B	1HBR-B	1V4X-B	NP_00	1GJN	1A6M	2LHB	1FSL
ヘモグロビン $\alpha$ ヒト	0.00	0.14	0.55	0.65	1.00	1.00	1.11	1.31	1.83	1.90	1.90	1.27	2.98
ヘモグロビン $\alpha$ ウマ	0.13	0.00	0.52	0.64	1.00	0.97	1.11	1.38	1.83	1.90	1.97	1.34	3.30
ヘモグロビン $\alpha$ にわとり	0.39	0.38	0.00	0.79	1.00	1.02	0.90	1.11	1.77	1.83	1.90	1.42	2.98
ヘモグロビン $\alpha$ まぐろ	0.44	0.43	0.50	0.00	1.00	1.02	1.24	1.24	2.00	2.07	2.14	1.83	2.98
ヘモグロビン $\beta$ ヒト	0.57	0.57	0.57	0.57	0.00	0.18	0.39	0.92	1.90	1.83	1.97	1.97	2.71
ヘモグロビン $\beta$ ウマ	0.57	0.56	0.58	0.58	0.16	0.00	0.45	0.92	1.97	1.83	1.97	2.00	2.71
ヘモグロビン $\beta$ にわとり	0.60	0.60	0.54	0.63	0.31	0.34	0.00	0.67	1.90	1.77	1.83	1.83	2.71
ヘモグロビン $\beta$ まぐろ	0.65	0.66	0.60	0.63	0.54	0.54	0.45	0.00	2.61	2.61	2.50	1.97	2.98
ミオグロビン ヒト	0.73	0.73	0.72	0.76	0.74	0.75	0.74	0.81	0.00	0.15	0.20	2.31	4.22
ミオグロビン ウマ	0.74	0.74	0.73	0.76	0.73	0.73	0.72	0.81	0.13	0.00	0.16	2.31	4.56
ミオグロビン マッコウクジラ	0.74	0.75	0.74	0.77	0.75	0.75	0.73	0.80	0.17	0.14	0.00	2.14	3.96
グロビン ヤツメウナギ	0.64	0.65	0.67	0.73	0.75	0.76	0.73	0.75	0.79	0.79	0.77	0.00	2.84
Leghemoglobin 大豆	0.84	0.85	0.84	0.84	0.82	0.82	0.82	0.84	0.88	0.89	0.87	0.83	0.00

Unrooted tree by neighbor-joining method

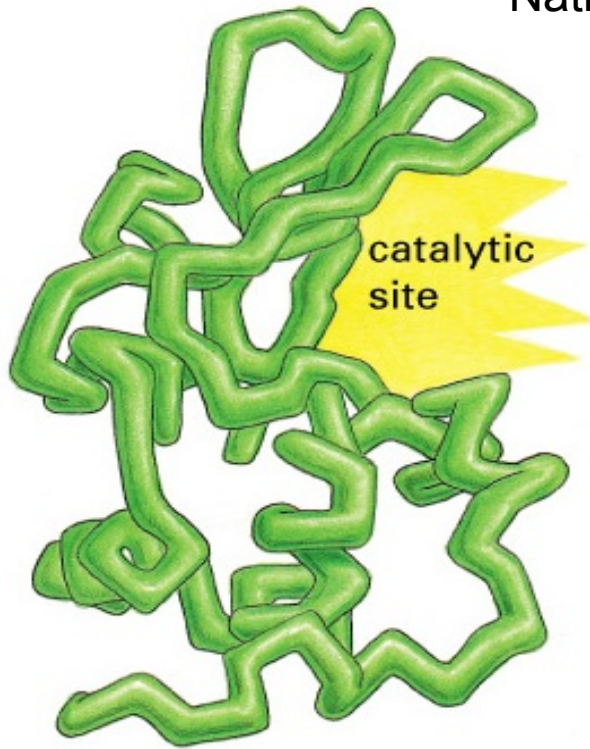


# Rooted tree by neighbor-joining method





Native状態



立体構造と機能は不可分

生物情報学における研究対象

- 蛋白質構造予測
- 蛋白質機能予測
- 蛋白質間相互作用予測

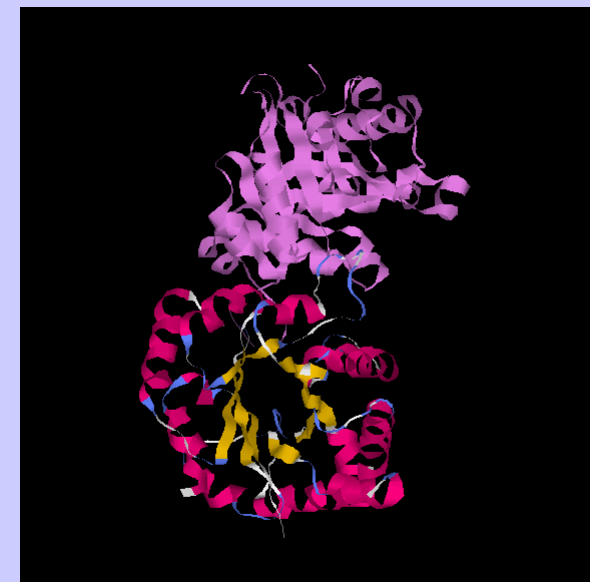
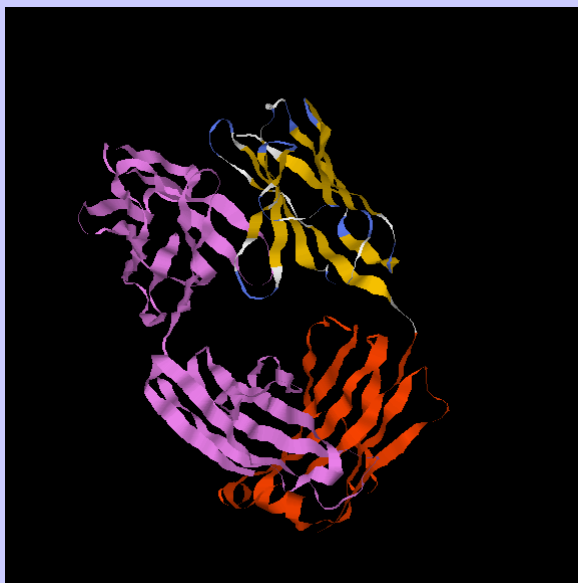
(A) lysozyme

# 配列に適合する構造の検索

YNNIP ... VTLR

Threading

適合度評価関数



# 配列—構造アライメント

```

minimum energy alignment
sequence 3GRS 364 YNNIPTVV-FSHPPIGTVGLT EDEA IHKYGIENVKTYSTS FTPMYHAVTKRKTICVM
matched to:
structure 1NPX 322 GVQGSGLAVFDYKFASTGINEVMA-QQLGK-ETKAVTVV -EDYLMDFNPDKQKAWF
probability alignment
sequence 3GRS 364 YNNIPTVV-FSHPPIGTVGLT EDEA IHKYGIENVKTYSTS FTPMYHAVTKRKTICVM
matched to:
structure 1NPX 322 GVQGSGLAVFDYKFASTGINEVMA-QQLGKKE-TKAVT-V VEDYLMDFNPDKQKAWF
7777664334334698999887541577776424333203 34444444455566666

1NPX 322 bbbbb bbbbb aaaa aaaa bbbb b bbbb bbbbb
#####
3GRS 364 bbb bbbbbbaaaaaaaa bbbbbbb b aaaaa bbb

minimum energy alignment
structure 3GRS 364 YNNIPTVVFVSH PIGTVGLTEDEA IHKYGIENVKTYSTS FTPMYHAVTKRKTICVM
matched to:
sequence 1NPX 322 GVQGSGLAVFD YKFASTGINEVMAQQLGKETKAVTVVE DYLMDF--NPDKQKAWF
probability alignment
structure 3GRS 364 --YNNIPTVVFVSH PIGTVGLT EDEA IHKYGIENVKTYSTS-FTPMYHAVTKRKTICVM
matched to:
sequence 1NPX 322 GV--QGSGLAVFDYKFASTGINE-VMAQQLGKETKAVTVVEDY---LMDFNPDKQKAWF
43223344444430345556554145667776543322222021112233566677777

minimum energy alignment
sequence 3GRS 420 KMVCA-NKEEKVVG IHMQGLGCEMLQGFVA VKMGATKADFDNT-VAIHPTSSEE L
matched to:
structure 1NPX 376 KLVYDPETTQILGALQMSKADLTANINAISLA IQAKMTIEDLAYADFFFQPAFDKPNW
probability alignment
sequence 3GRS 420 KMVCA-NKEEKVVG IHMQGLGCEMLQGFVA VKMGATKADFDNT-VAIHPTSSEE-L
matched to:
structure 1NPX 376 KLVYDPETTQILGALQMSKADLTANINAISLA IQAKMTIEDLAYADFFFQPAFDKPNW I
666666677777654045679999998888888888888888876434455433222212

1NPX 376 bbbb bbbbb aaaaaaaaa aaaaa a a
#####
3GRS 420 bbbbb b bbbbbbbb aaaaaaaaa aaaaa aaaaa

minimum energy alignment
structure 3GRS 420 KMVCA-NKEEKVVG IHMQGLGCEMLQGFVA VKMGATKADFDNT----VAIHPTSSEE L
matched to:
sequence 1NPX 376 KLVYDPETTQILGALQMSKADLTANINAISLA IQAKMTIEDLAYADFFFQPAFDKPNWII
probability alignment
structure 3GRS 420 KMVCA-NKEEKVVG IHMQGLGCEMLQGFVA VKMGATKADFDN----TVAIHPTSSEE L
matched to:
sequence 1NPX 376 KLVYDPETTQILGALQMSKADLTANINAISLA IQAKMTIEDLAYADFF-----
7766534434443034443344444455555677778888888764335622222111100

minimum energy alignment
sequence 3GRS 475 VTLR ----- min. ene. rmsd #aligned identities
matched to:
structure 1NPX 433 NIIN TAAL EAVKQER -26.4 3.9 112 0.12
probability alignment
sequence 3GRS 475 VTLR -----
matched to:
structure 1NPX 435 I--- NTAALEAVKQER 3.7 108 0.12
2011 246789999999 3.0 73

1NPX 435 a aaaaaaaaaa
3GRS 475 aa #####

minimum energy
structure 3GRS 475 VTLR -----
matched to:
sequence 1NPX 436 NTAA LEAVKQER -20.0 4.3 113 0.11
probability alignment
structure 3GRS 475 VTLR-----
matched to:
sequence 1NPX 424 ---FQPAFDKPNW IINTAALEAVKQER 3.5 92 0.12
0112533343233344344577788999 3.0 45

```





## 生物情報学における研究対象

- 相互作用ネットワークの解析
  - データベース構築
  - 遺伝子発現ネットワーク
  - 蛋白質相互作用ネットワーク
  - 代謝物パスウェイ
  - シグナリングネットワーク
  - パスウェイ比較
- 生体システムの計算機シミュレーション

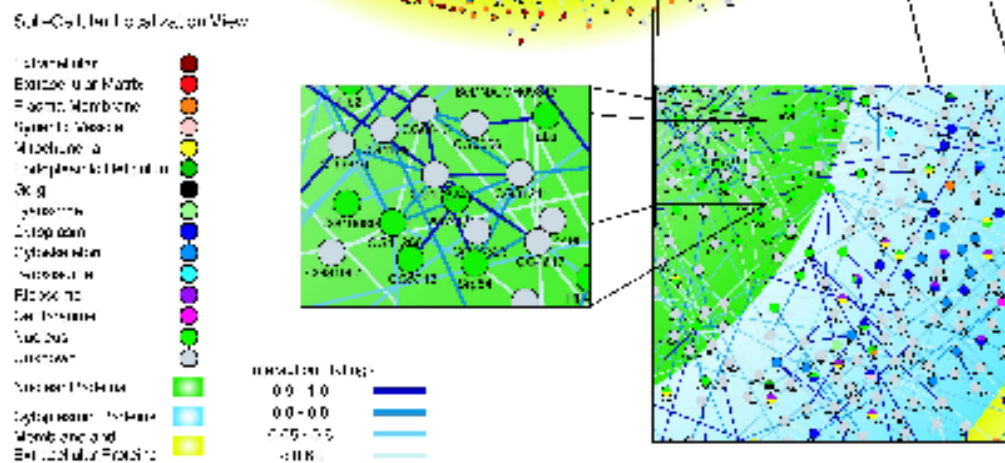
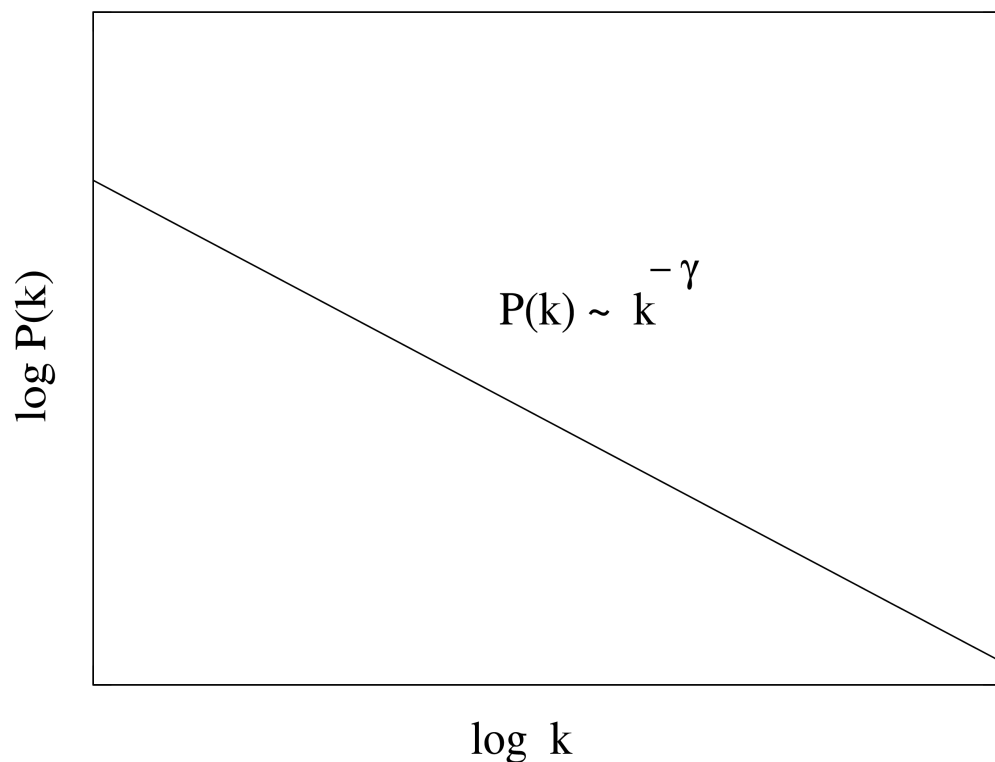


Fig. 4. Global views of the protein interaction map. (A) Protein family enrichment map showing a jagged border between human disease proteins and other proteins. Interactions were sorted according to interaction confidence score and the top 1000 interactions are shown with their corresponding 3522 proteins. This corresponds roughly to a confidence score of 0.62 and a p-value. (B) Subcellular localization

view. This view also shows the interaction map with each protein colored by its Gene Ontology Cellular Component annotation. This view has been filtered by any missing protein data was then separated into 26 interactions and with at least one Gene Ontology annotation that necessarily cellular component annotation. We show proteins for all interactions with a confidence score of 0.5 or higher. This results in a map with 2246 proteins and 2268 interactions.

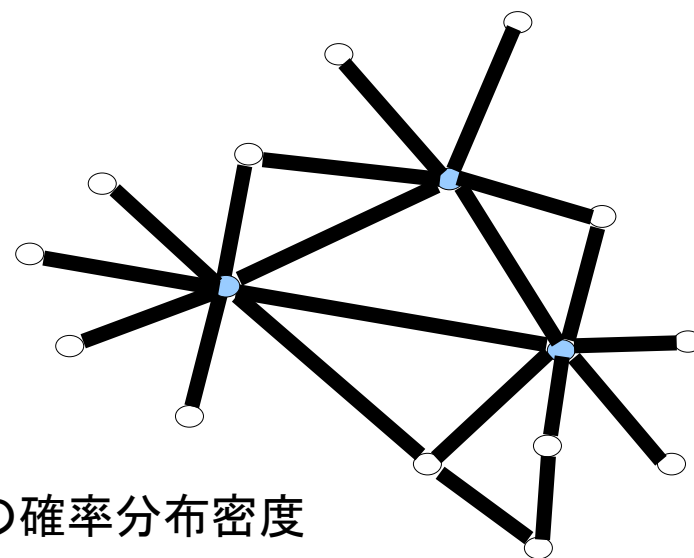
# Scale Free Network

- 生物系のネットワーク: 蛋白質相互作用ネットワーク;  $\gamma = 2.2$
- Internet: WEB ページのリンクに関するネットワーク
- 社会構造: 知人関係のネットワーク

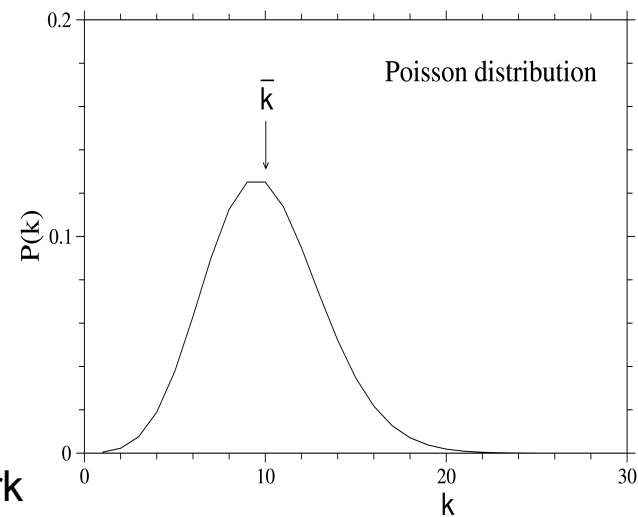


## 生成機構

Preferential attachment  
/ Rich-gets-richer



$k$ : 次数  
 $P(k)$ : 次数の確率分布密度



Random Network

## **生物情報学とは:**

**目的:** ゲノムにコードされている情報を、情報学の手法で読み解くこと。

**背景:** ゲノム解析技術の驚異的な発展による生物(遺伝)情報の爆発的増加。

## **現状:**

生物/化学/物理学的方法論に加え、多量のデータの分析を志す生物情報学が誕生した。

## **将来:**

各種情報を統合化しシステムとして生物を理解しようとする方法へ発展しつつある。

引用文献: 画像の多くは Molecular Biology of the Cell, V.4から引用  
分子グラフィクスはMolScriptを用いた。