生物情報学とは?

生物学と情報学の出会い

(DNA/蛋白質配列解析のための確率モデル)

宮澤　三造

群馬大学大学院工学研究科

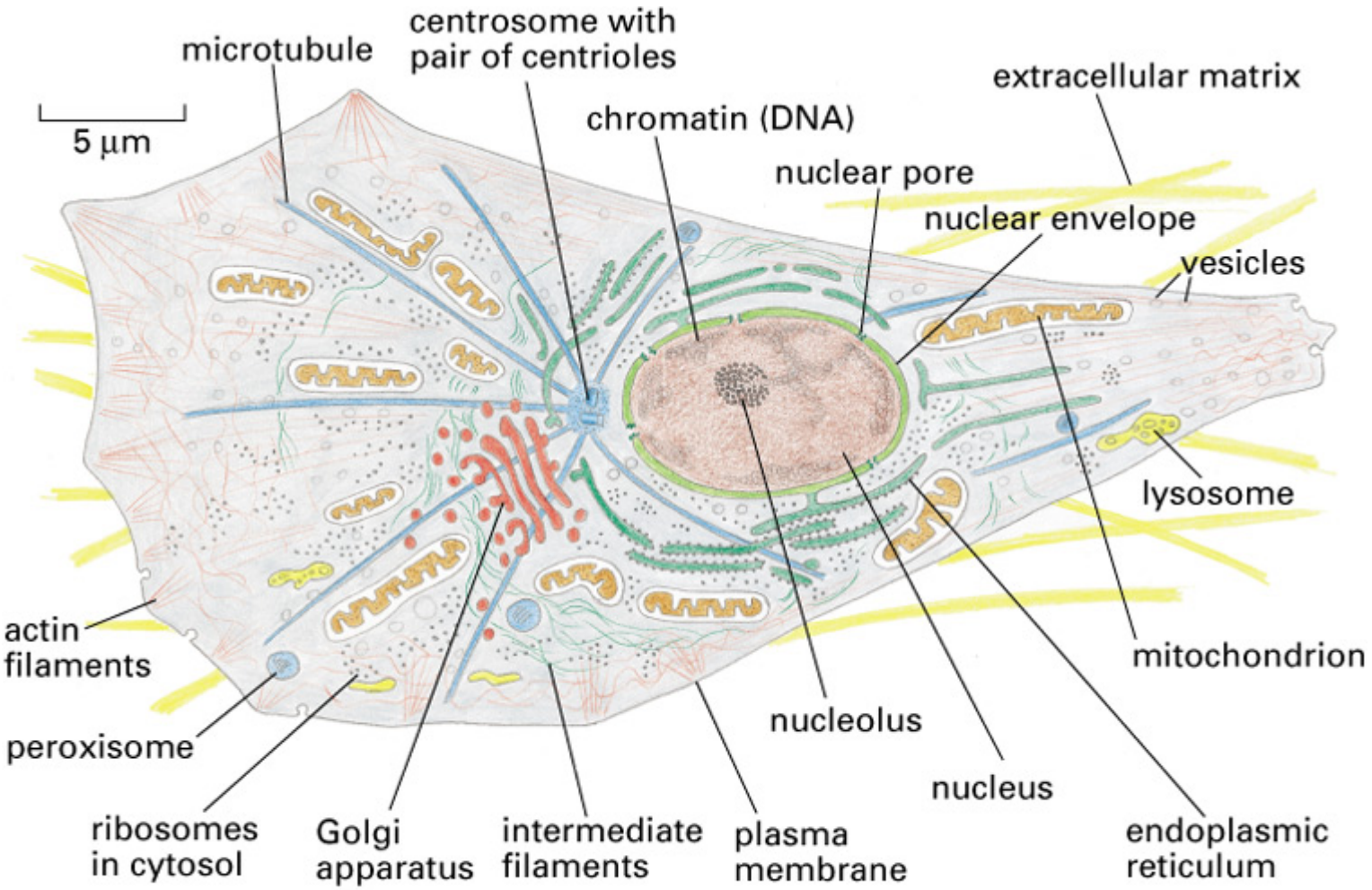
2009/06/19 群馬大学科学技術振興会セミナー

生物情報学とは:

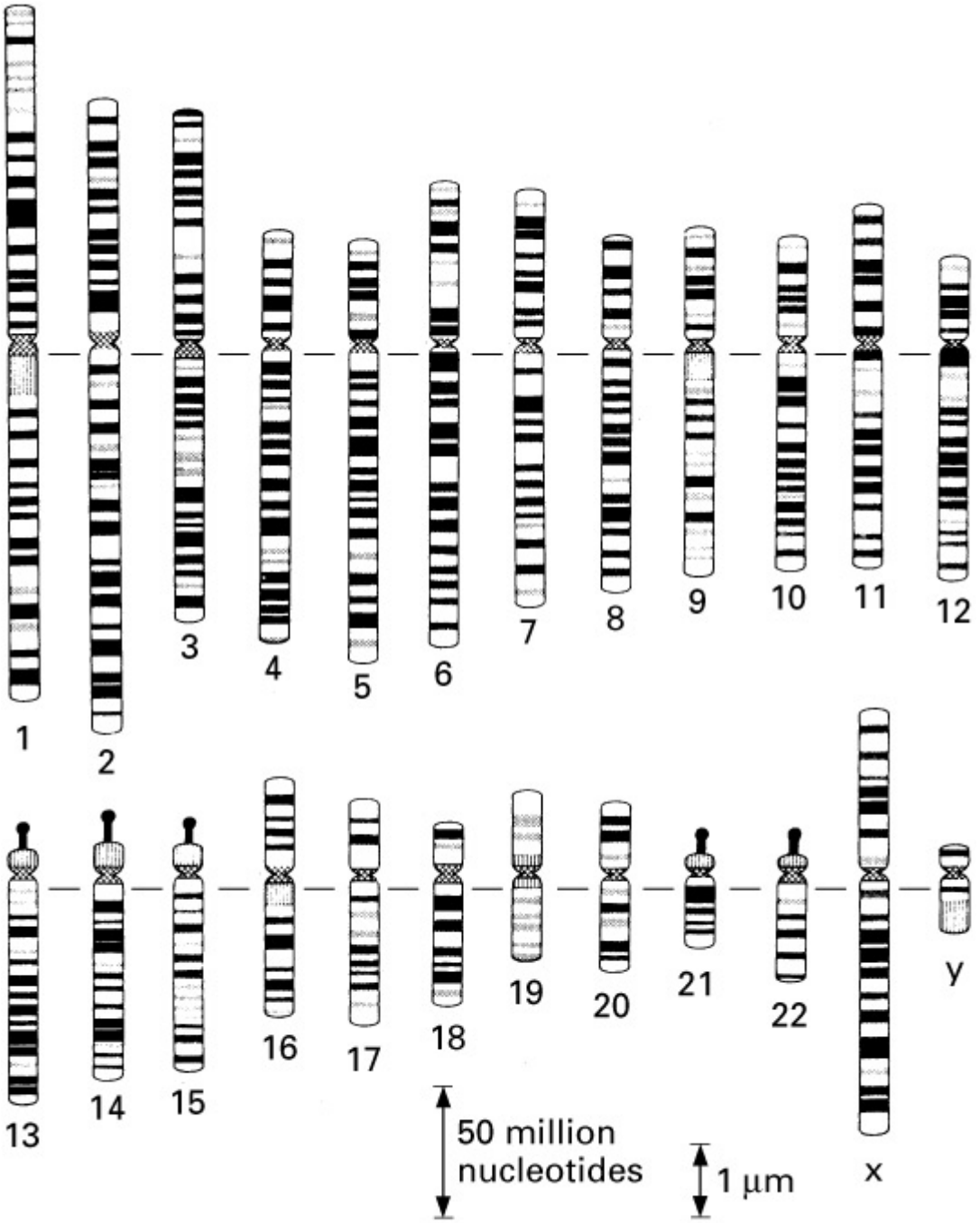
ゲノムにコードされている情報を、情報学の手法で読み解くこと。

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



動物細胞の模式図 (Molecular Biology of the Cell)



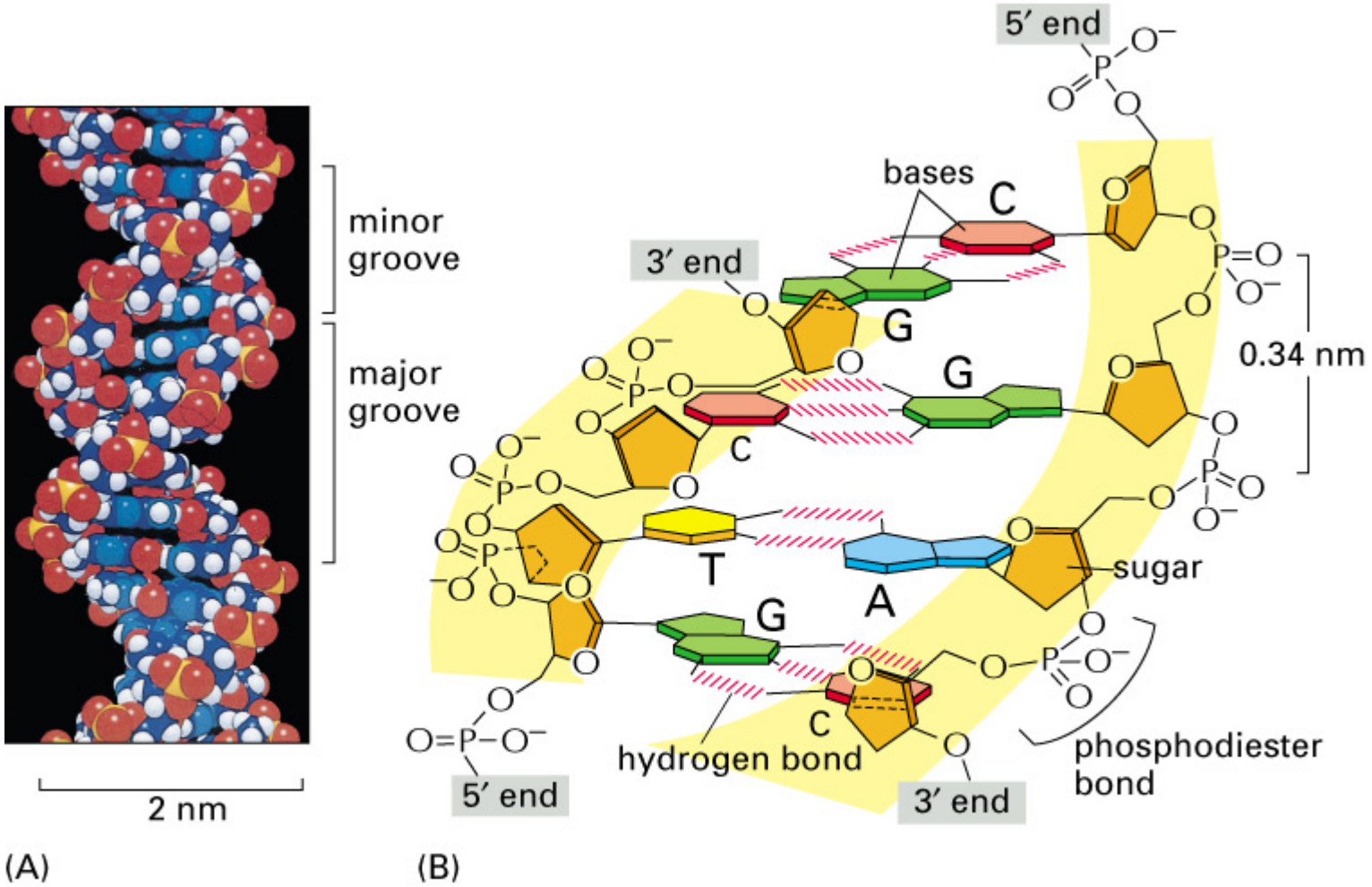
ヒトの染色体: 22対の常染色体と性染色体 XX または XY

染色体: DNAと蛋白質の複合体

(Molecular Biology of the Cell)

DNAの分子模型

DNA分子の模式図

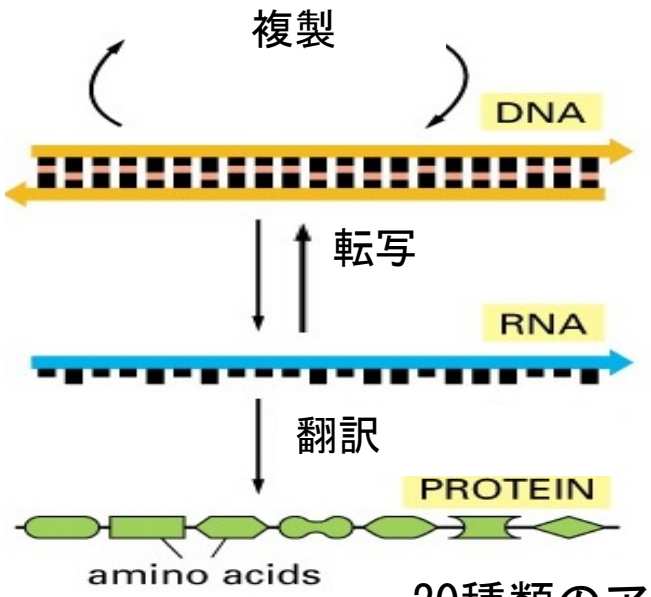


(A,T,C,G) 4種類の塩基からなる鎖状分子の2本鎖が A–T, C–G と相補的に結合し、2重らせん構造をなす。(Watson & Crick, 1953)

ヒトのDNA: 30億塩基 x 0.34 nm = 1 m

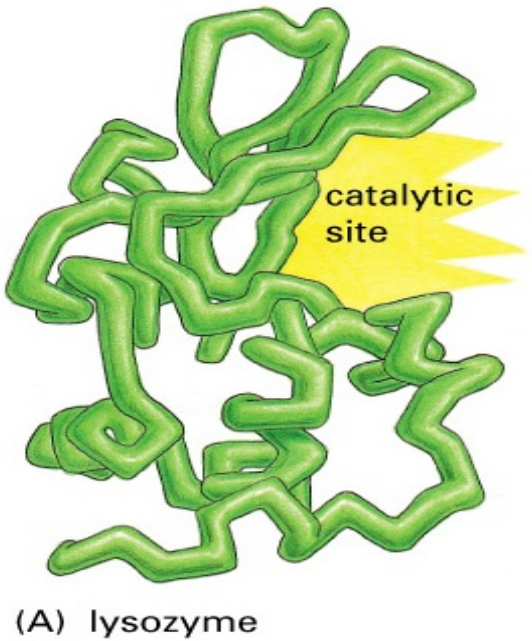
(Molecular Biology of the Cell)

遺伝情報の流れ：



折り畳み

20種類のアミノ酸が
鎖状に結合した高分子



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

転写

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

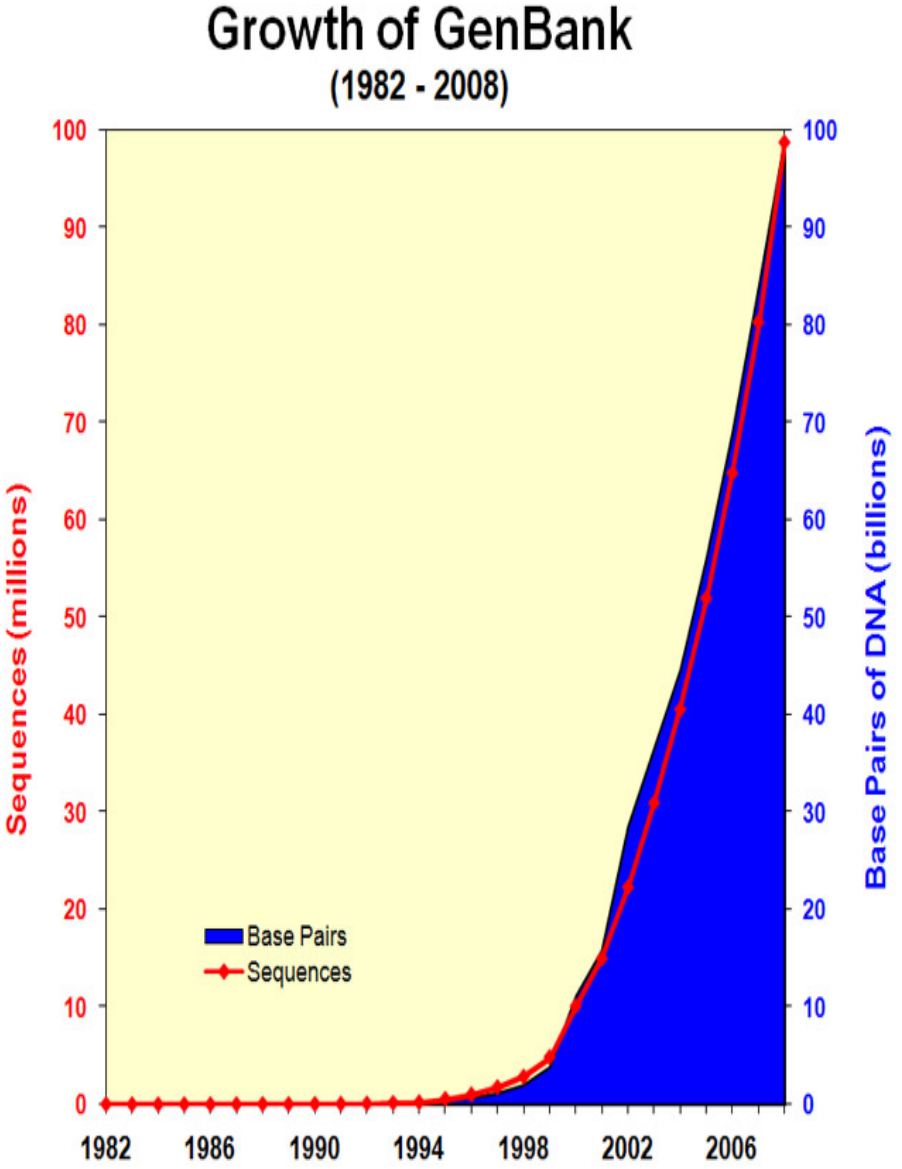
翻訳

M .... K V V

Standard Code Table

1st	2nd U	2nd C	2nd A	2nd G	3rd
U	UUU Phe, F	UCU Ser, S	UAU Tyr, Y	UGU Cys, C	U
	UUC Phe, F	UCC Ser, S	UAC Tyr, Y	UGC Cys, C	C
	UUA Leu, L	UCA Ser, S	UAA Term	UGA Term	A
	UUG Leu, L	UCG Ser, S	UAG Term	UGG Trp, W	G
C	CUU Leu, L	CCU Pro, P	CAU His, H	CGU Arg, R	U
	CUC Leu, L	CCC Pro, P	CAC His, H	CGC Arg, R	C
	CUA Leu, L	CCA Pro, P	CAA Gln, Q	CGA Arg, R	A
	CUG Leu, L	CCG Pro, P	CAG Gln, Q	CGG Arg, R	G
A	AUU Ile, I	ACU Thr, T	AAU Asn, N	AGU Ser, S	U
	AUC Ile, I	ACC Thr, T	AAC Asn, N	AGC Ser, S	C
	AUA Ile, I	ACA Thr, T	AAA Lys, K	AGA Arg, R	A
	AUG Met, M	ACG Thr, T	AAG Lys, K	AGG Arg, R	G
G	GUU Val, V	GCU Ala, A	GAU Asp, D	GGU Gly, G	U
	GUC Val, V	GCC Ala, A	GAC Asp, D	GGC Gly, G	C
	GUA Val, V	GCA Ala, A	GAA Glu, E	GGA Gly, G	A
	GUG Val, V	GCG Ala, A	GAG Glu, E	GGG Gly, G	G

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。



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

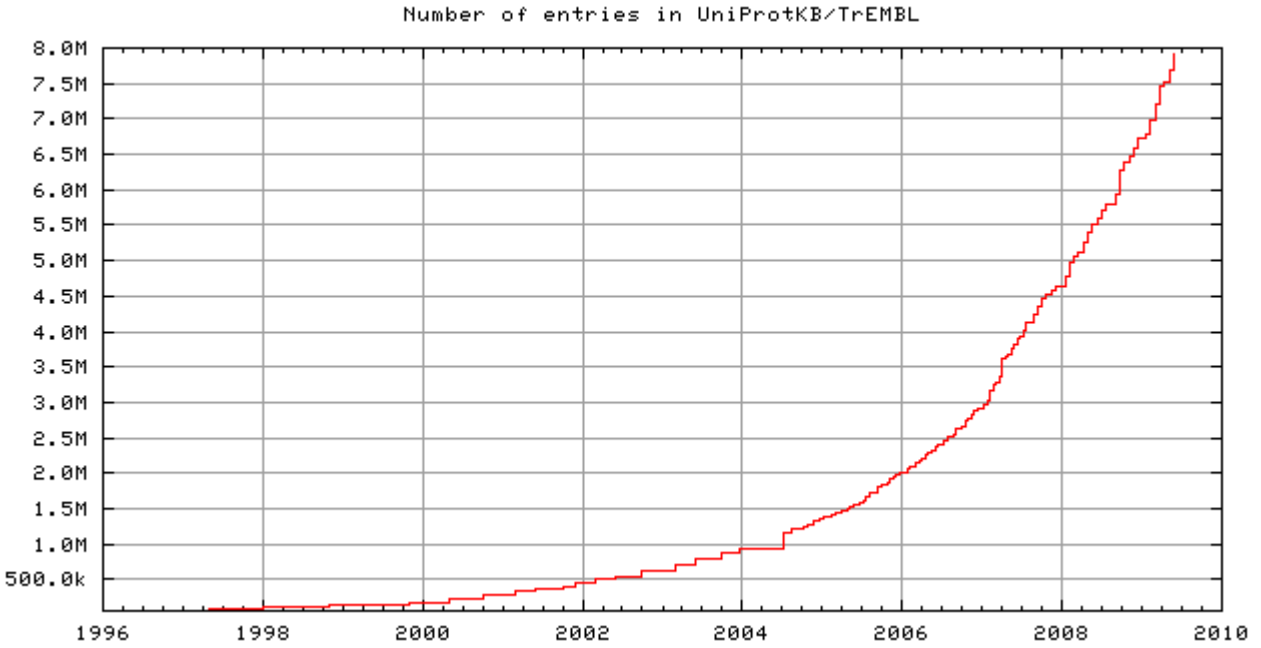
配列解析が完了している生物種

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

http://www.nslij-genetics.org/seq/ より引用

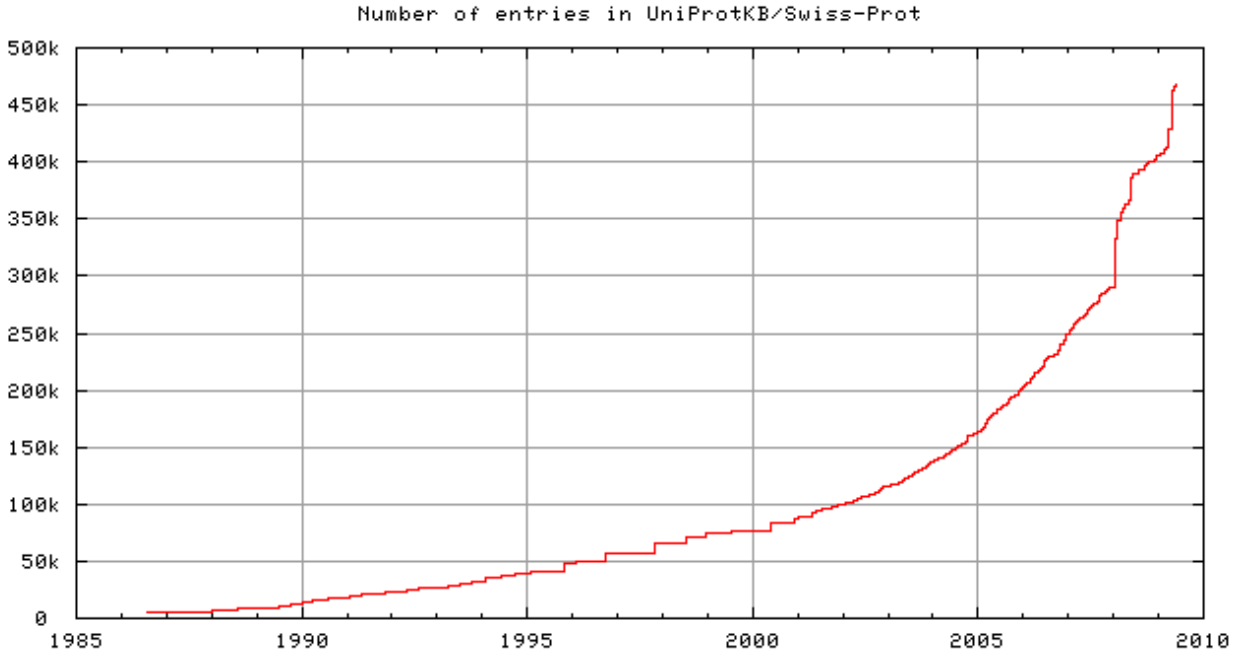
蛋白質配列データ

注釈：自動

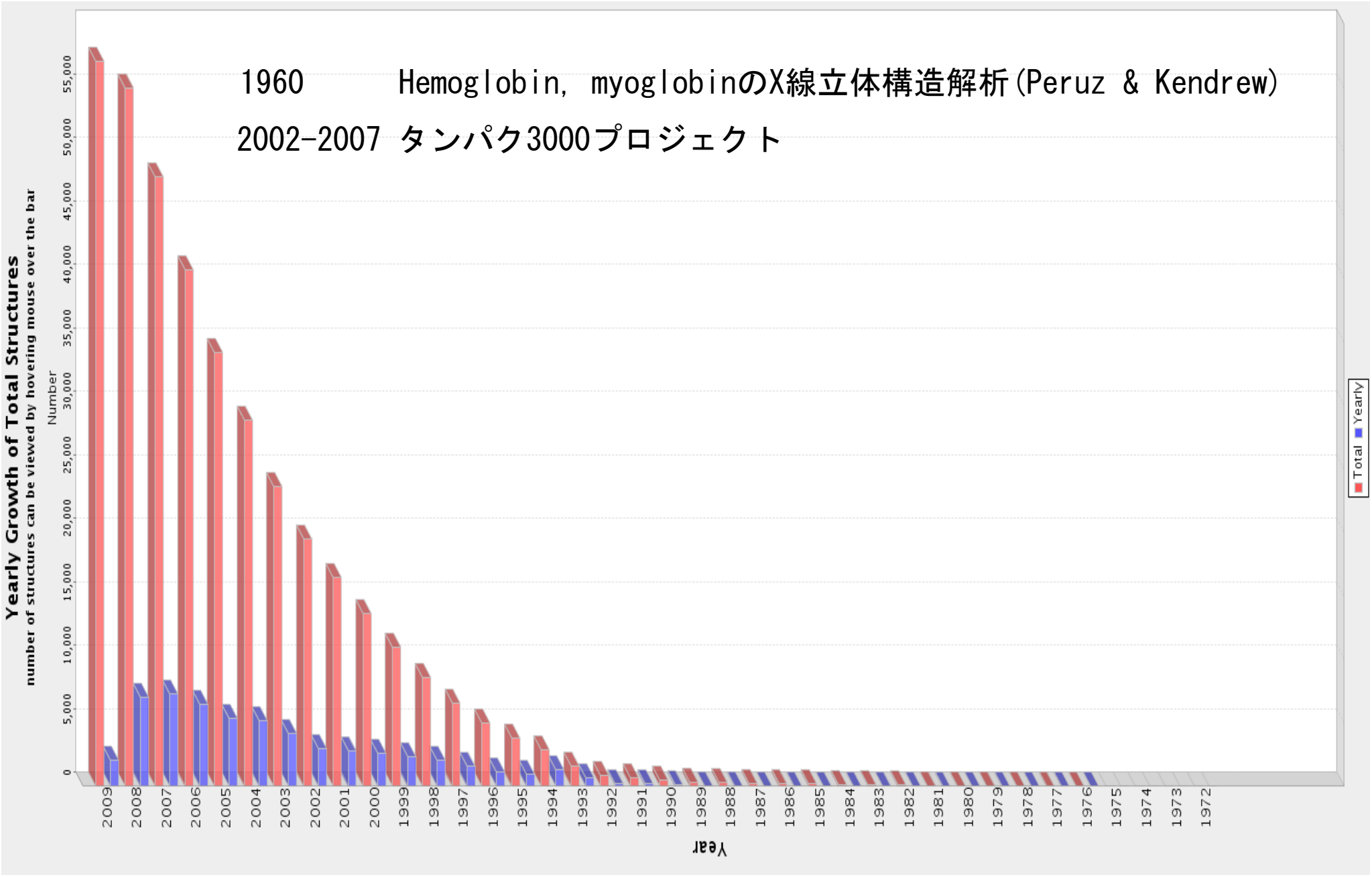


蛋白質配列データ

注釈：人手



蛋白質立体構造データの増大



生物情報学関連のWeb上のリソース †

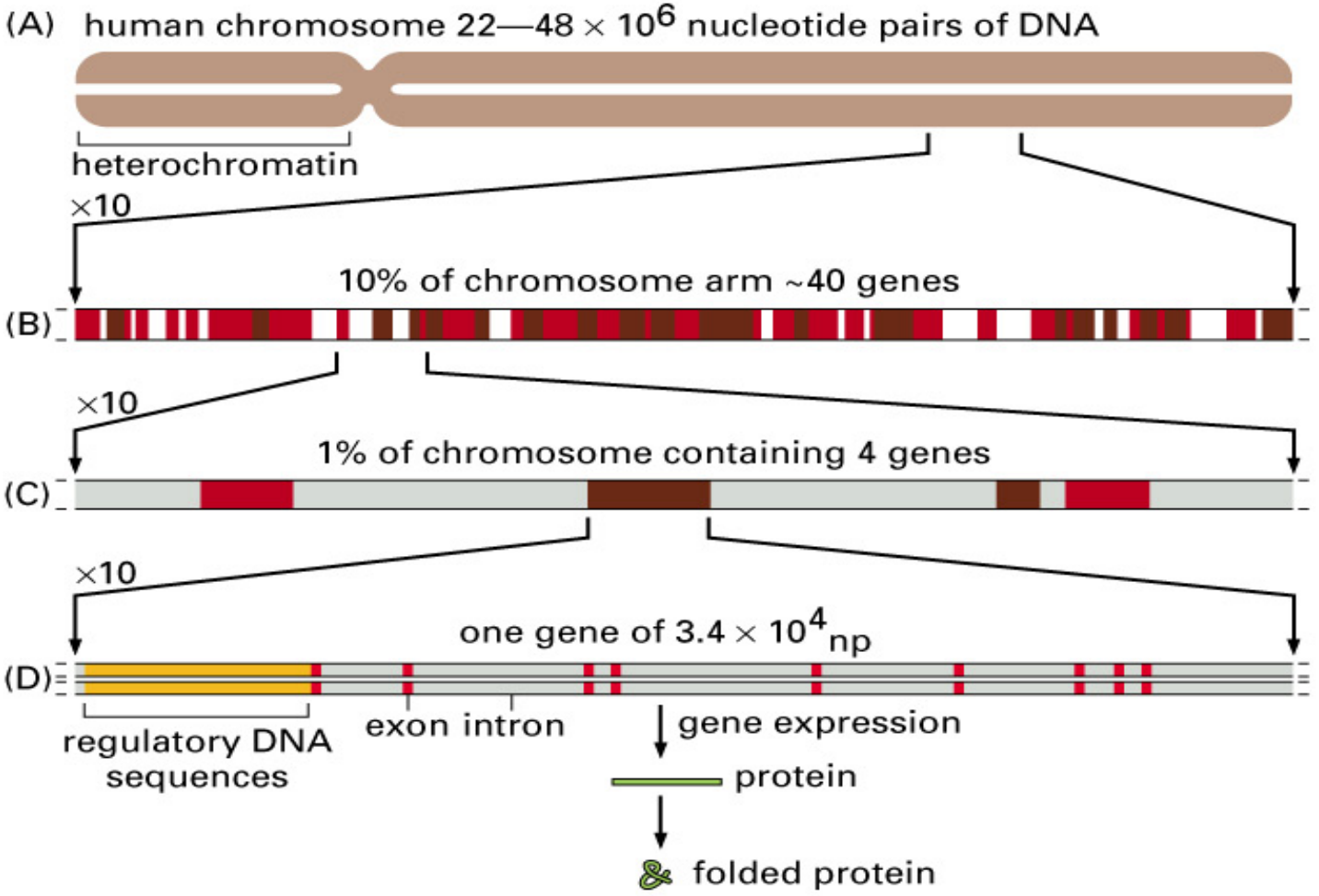
- データベース及びBrowser: ポータル: NCBI | EBI | DDBJ | GenomeNet
  - 塩基配列 (データ量の変遷): GenBank | EMBL | DDBJ
  - RNA配列/2次構造; RNA ファミリー (Rfam)
  - 蛋白質配列 (UniProt) | 蛋白質ファミリー (Pfam) | Functional site (PROSITE)
  - 蛋白質構造 (データ量の変遷) ポータル: EBI(DB | Analyses)
    構造 (WW PDB (RCSB | MSD | PDBj | BMRB) | PDBsum) | 分類 (SCOP | 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 | 代謝パスウェー: KEG | 遺伝子発現パスウェー | 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)

生物情報学における研究対象

各種データベースの構築/維持

- DNA/RNA/Proteinデータベース
- ゲノムデータベース
- 派生的なデータベース
- ネットワークデータベース
- 用語データベース

解析ソフトウエアの提供

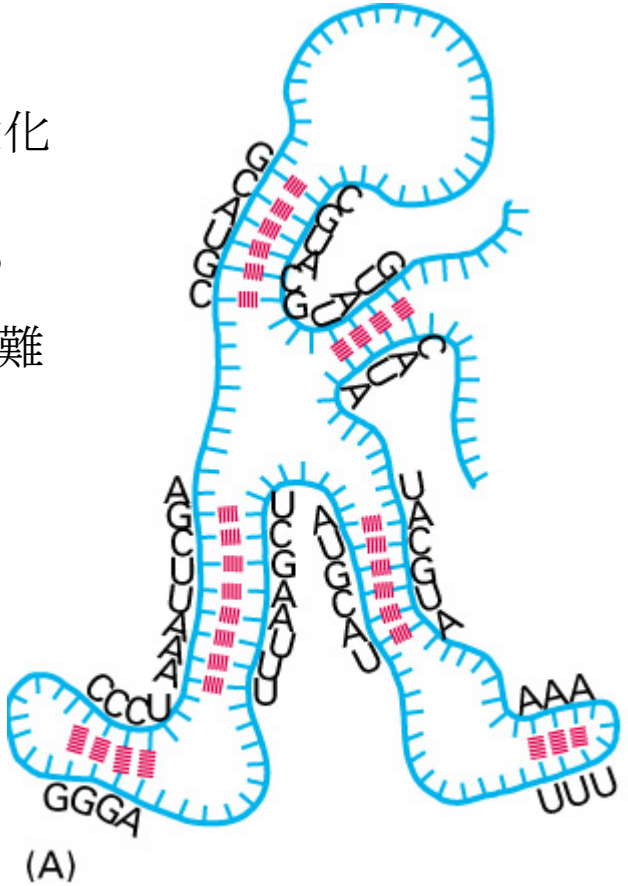


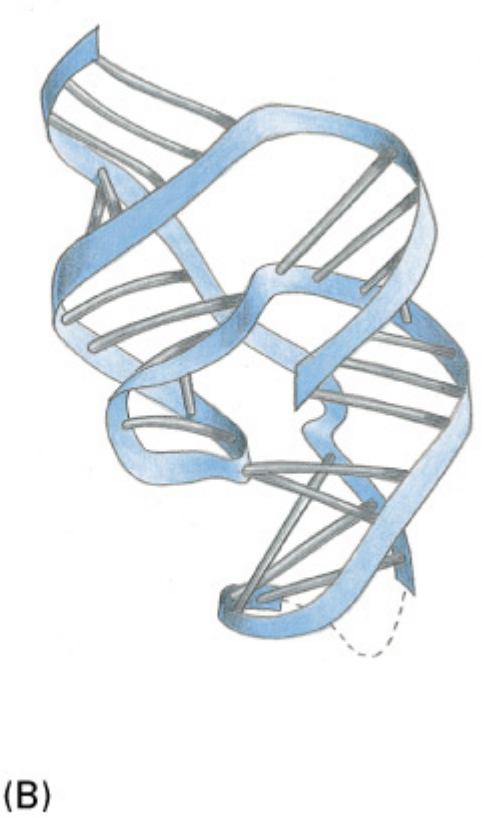
生物情報学における研究対象

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

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



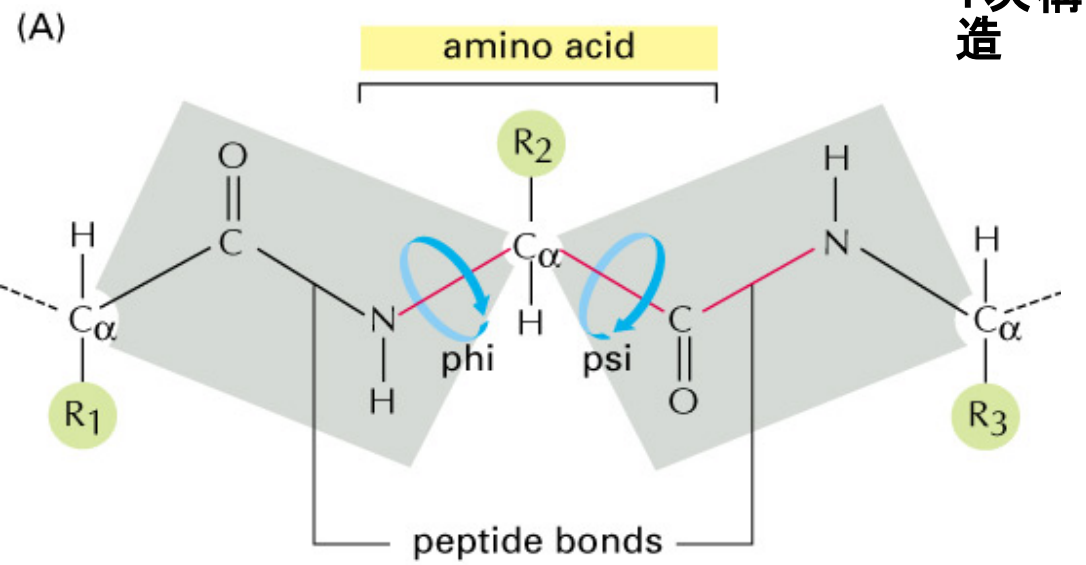


RNA 2次構造

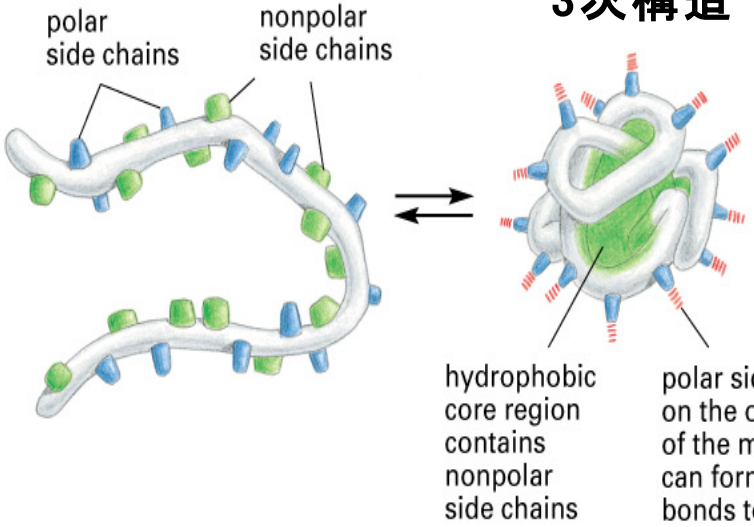
RNA 3次構造

蛋 白 質 の 構 造

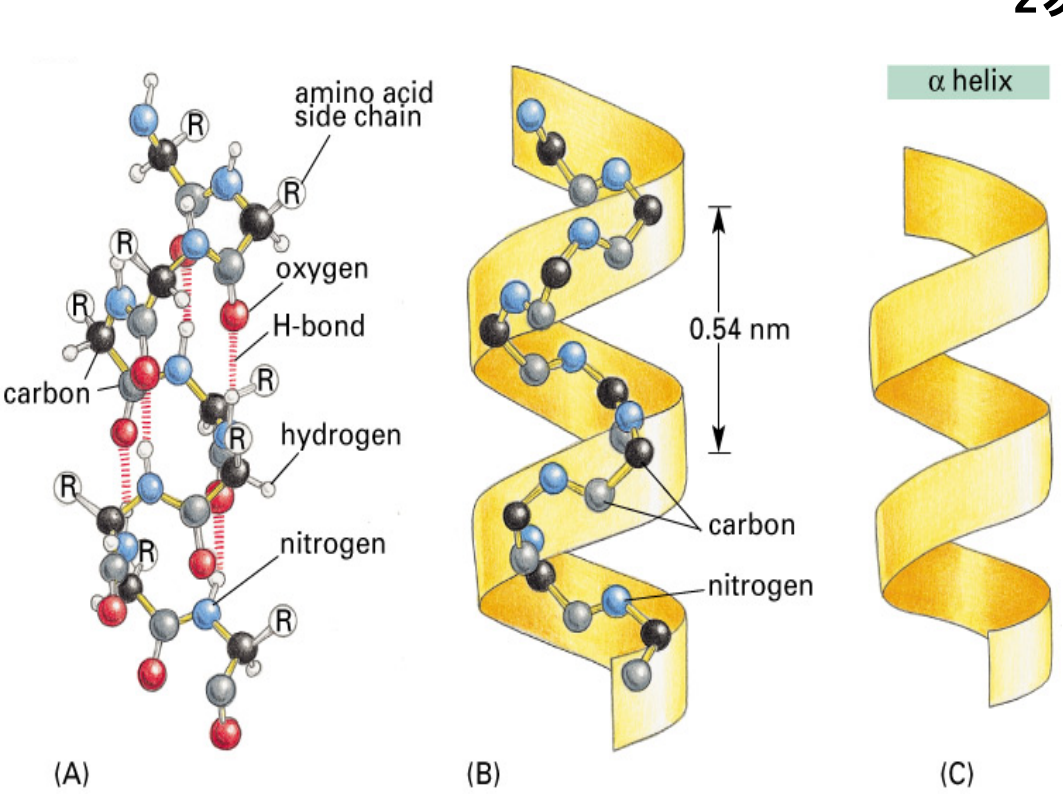
1次構造



3次構造



2次構造



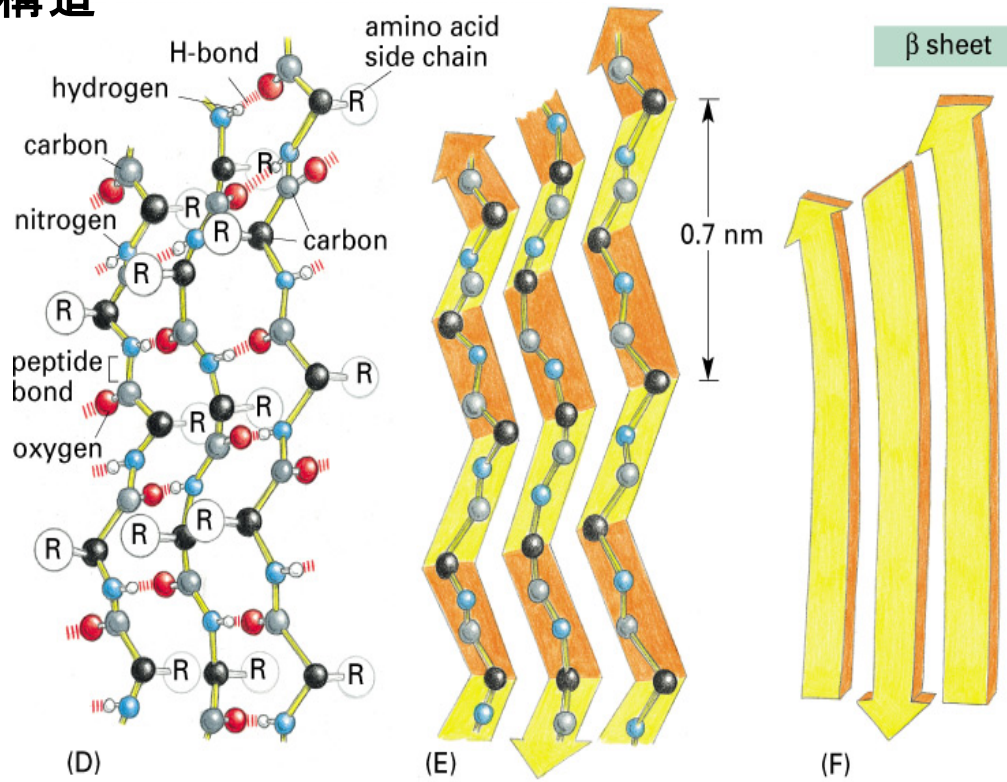
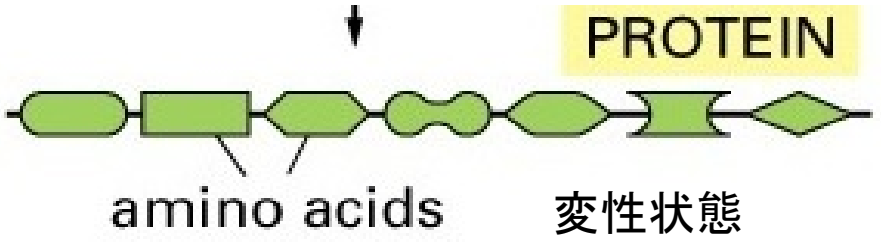


Figure 3–9 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

Figure 3–9 part 2 of 2. Molecular Biology of the Cell, 4th Edition.

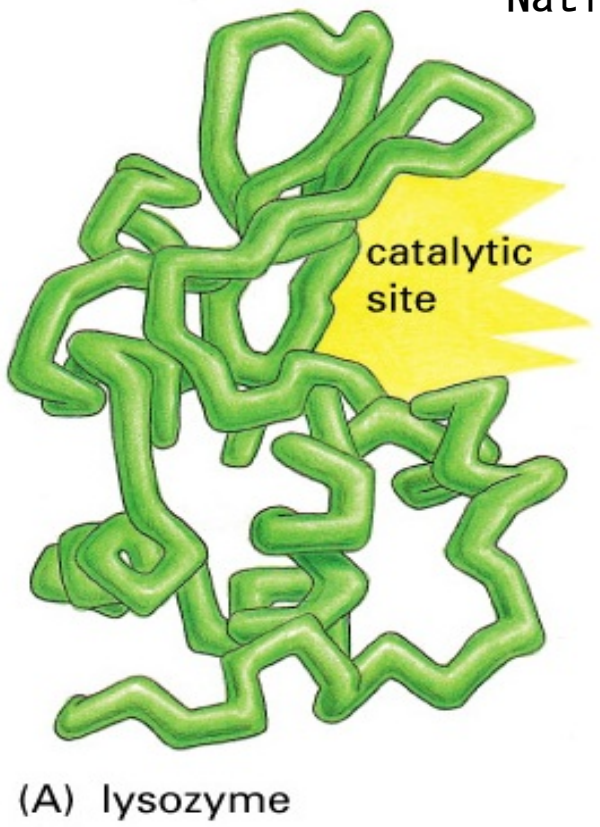


変性状態

Unfolding T ↑ ↓ T Folding

アミノ酸配列は自律的に折り畳まれる。(Anfinsen, 1972)

Native状態



(A) lysozyme

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

配列レベルでの保存と進化

生物情報学における研究対象

- 類似配列アライメント
- 類似配列検索
- ゲノム比較
- 分子系統樹作成
- 分子進化

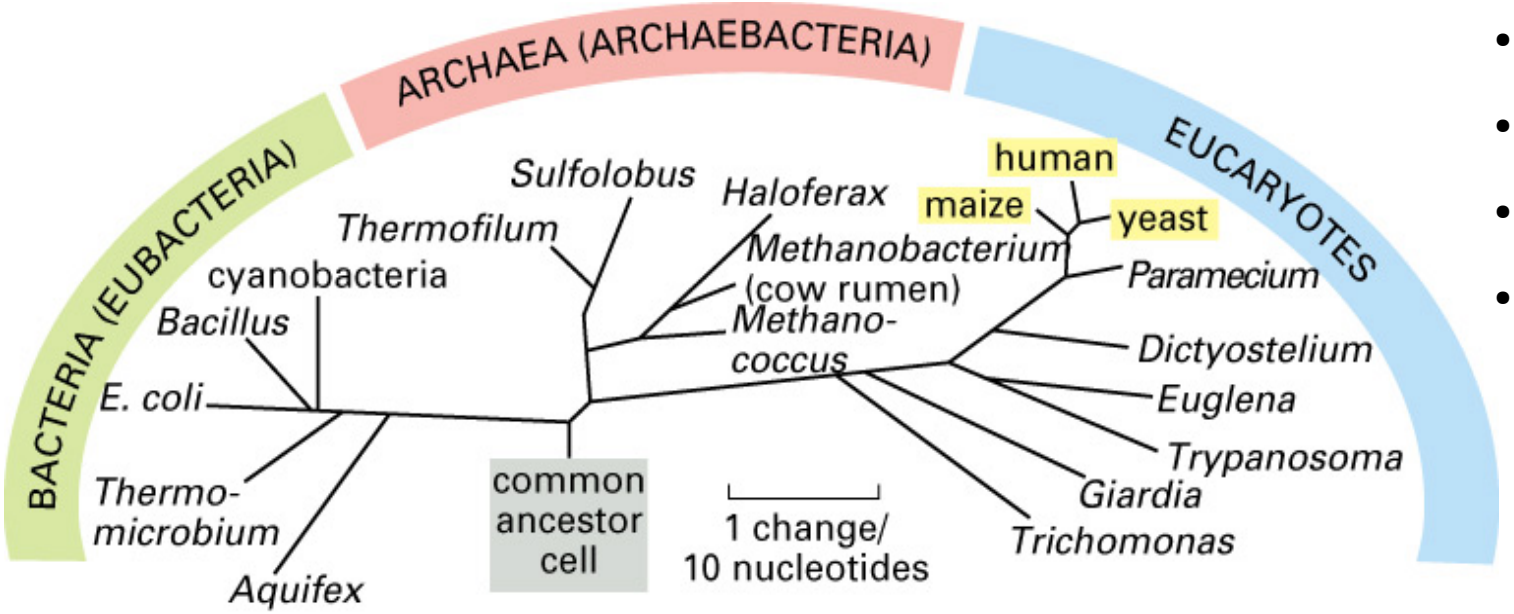


Figure 1–21. Molecular Biology of the Cell, 4th Edition.

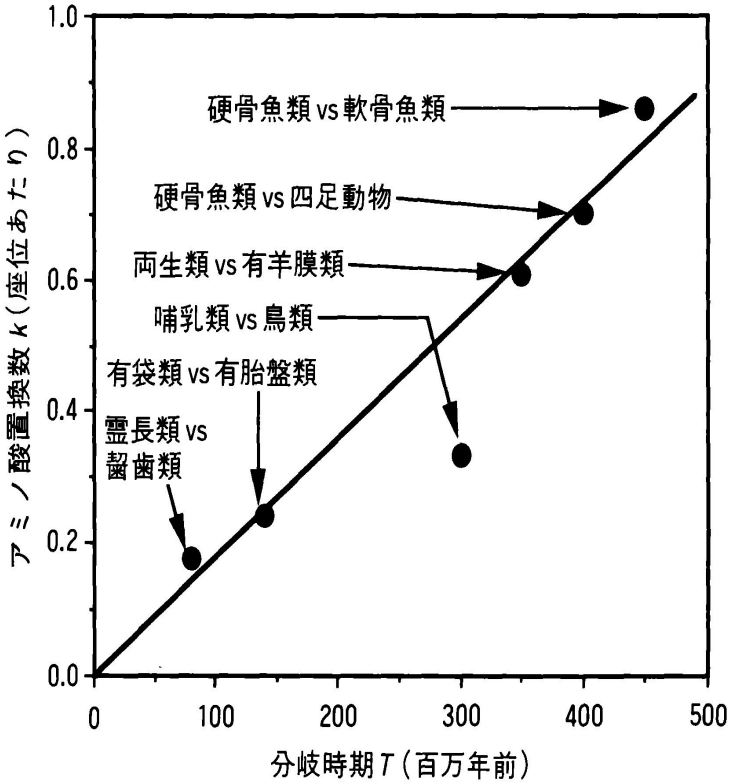


図 5. ヘモグロビンの分子時計

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)	3.2×10^{-2}
	インフルエンザウイルスA型	1.1×10^{-2}
	〃 B型	0.21×10^{-2}
	〃 C型	0.14×10^{-2}
	デング熱ウイルス(デング2)	0.25×10^{-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	0.12×10^{-2}
	〃 C型	0.10×10^{-2}
	パラインフルエンザ3	$< 0.29 \times 10^{-2}$
核遺伝子	哺乳類	2.8×10^{-9}
	齧歯類	$> 6.2 \times 10^{-9}$
	棘皮動物	5.6×10^{-9}
	高等植物[a)]	7.1×10^{-9}
ミトコンドリア	類人猿[b)]	55.0×10^{-9}
	高等植物[a)]	0.8×10^{-9}
葉緑体	高等植物[a)]	2.6×10^{-9}

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.

分子進化(宮田隆)より引用)

配列解析の確率モデル

Promoter, Exon/Intron境界, 遺伝子発見

モデル M は、確率 $P(\mathbf{a}|\theta_M)$ で配列 **a** を生成するものとする。

学習: 既知の配列データ $D(=\{\mathbf{a}\})$ よりパラメーターθ_Mを推定

$$P(\theta_M \mid D) = \frac{P(D|\theta_M)\ P(\theta_M)}{P(D)}$$

仮定: $P(\theta_M)$ = constant

最尤推定: $\hat{\theta}_M = \arg\max_\theta P(D|\theta_M)$ (Baum-Welch algorithm)

予測: $\log \{P(\mathbf{a}|\hat{\theta}_M) / P(\mathbf{a})\} > S_{threshold}$ (forward/backward algorithm)

例　Exon/Intron境界予測

確率モデル: 境界からの相対的位置のみに依存する塩基頻度を持つ特異配列

$$P(\mathbf{a}|\theta_M) = \prod_i p_i(a_i) \qquad P(\mathbf{a}) = \prod_i p(a_i)$$

学習:　$p_i(a_i=\alpha) = f_i(\alpha)$,　$p(\alpha) = f(\alpha)$,　$\alpha \in \{A, T, C, G\}$

予測: $\log P(\mathbf{a}|\hat{\theta}_M) / P(\mathbf{a}) = \Sigma_i \log p_i(a_i)/p(a_i) > S_{threshold}$

Intronの5'端付近の配列プロファイル: $p_i(a_i)$

Frequencies (%) in 1254 donor splice sites

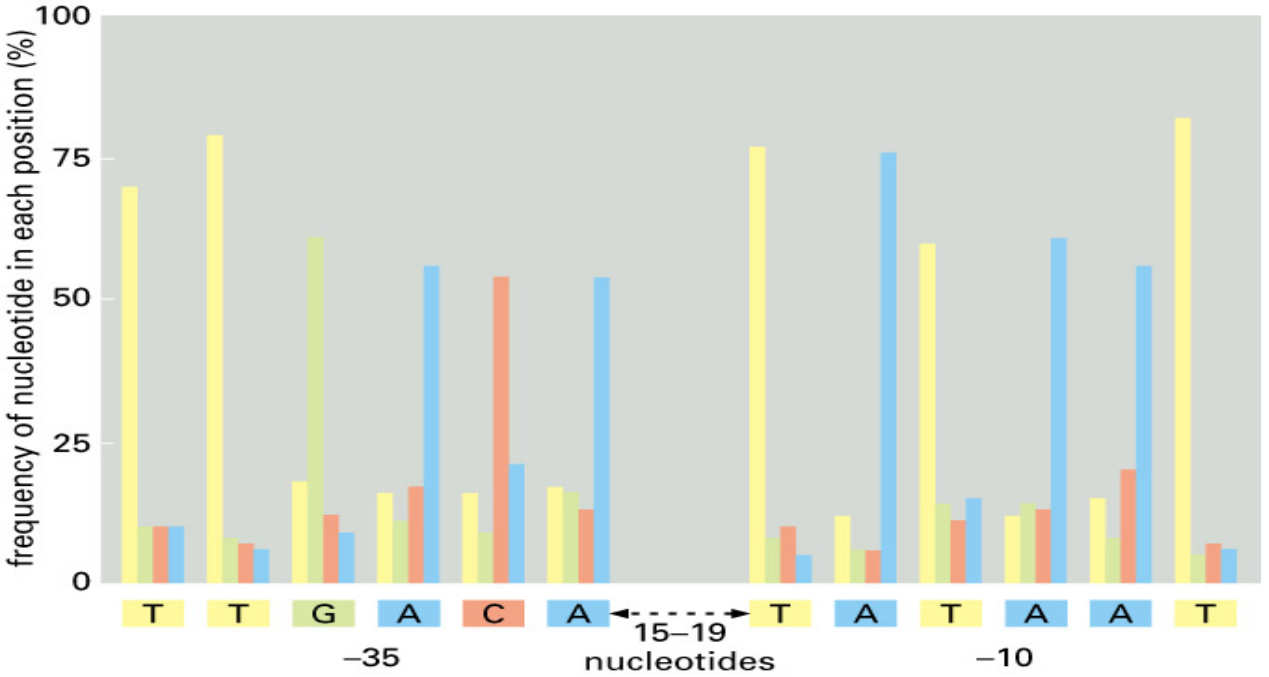
Base\Position	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	33	60	8	0	0	49	71	6	15
C	37	13	4	0	0	3	7	5	19
G	18	14	81	100	0	45	12	84	20
U/T	12	13	7	0	100	3	9	5	46

(Burge & Karlin, 1997)

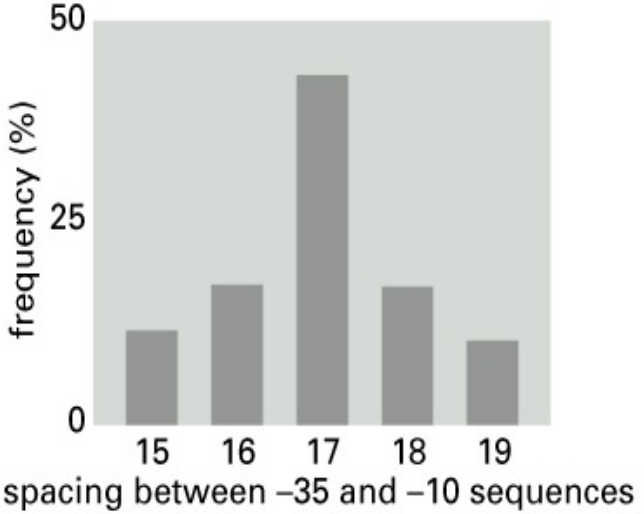
a_i	C	A	G	G	T	A	A	G	T	$\sum_i$
$p_i(a_i)$	0.37	0.60	0.81	1.00	1.00	0.49	0.71	0.84	0.46	
$p(a_i)$	0.10	0.27	0.42	0.42	0.22	0.27	0.27	0.42	0.22	
$\log(p_i(a_i)/p(a_i))$	1.33	0.80	0.66	0.87	1.51	0.60	0.97	0.69	0.74	8.17

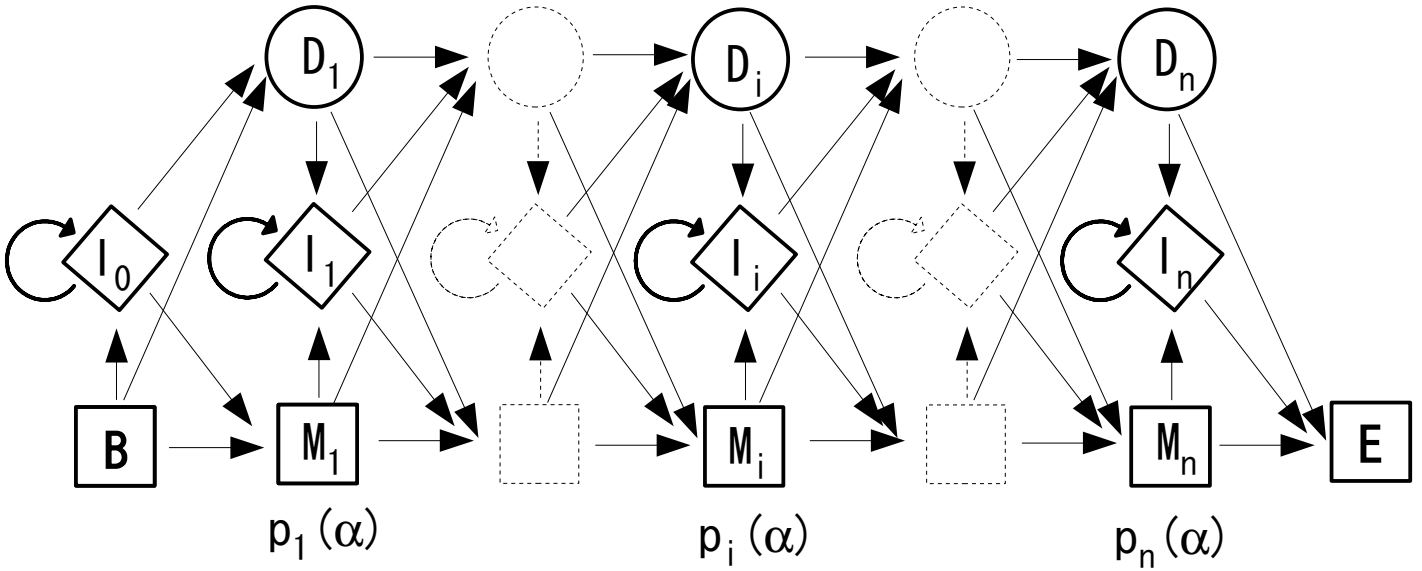
例 Promoter予測

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



(Molecular Biology of the Cell)





Profile Hidden Markov Model

Pairwise Alignment

$$A \equiv \begin{bmatrix} \cdots & a_{i-1} & a_i & - & - & a_{i+1} & \cdots \\ \cdots & - & b_j & b_{j+1} & b_{j+2} & b_{j+3} & \cdots \end{bmatrix}$$

$$P(A \mid \mathbf{a}, \mathbf{b}, \theta) = \frac{P(\mathbf{a}, \mathbf{b} \mid A, \theta) P(A \mid \theta)}{\Sigma_A P(A, \mathbf{a}, \mathbf{b} \mid \theta)}$$

$$= \frac{\exp(S(A \mid \mathbf{a}, \mathbf{b}, \theta))}{\Sigma_A \exp(S(A \mid \mathbf{a}, \mathbf{b}, \theta))}$$

最も確からしいアラインメント　(Viterbi/Needlman-Wunsch algorithm)

$$A_{max} = \arg\max_A P(A \mid \mathbf{a}, \mathbf{b}, \theta) = \arg\max_A S(A \mid \mathbf{a}, \mathbf{b}, \theta)$$

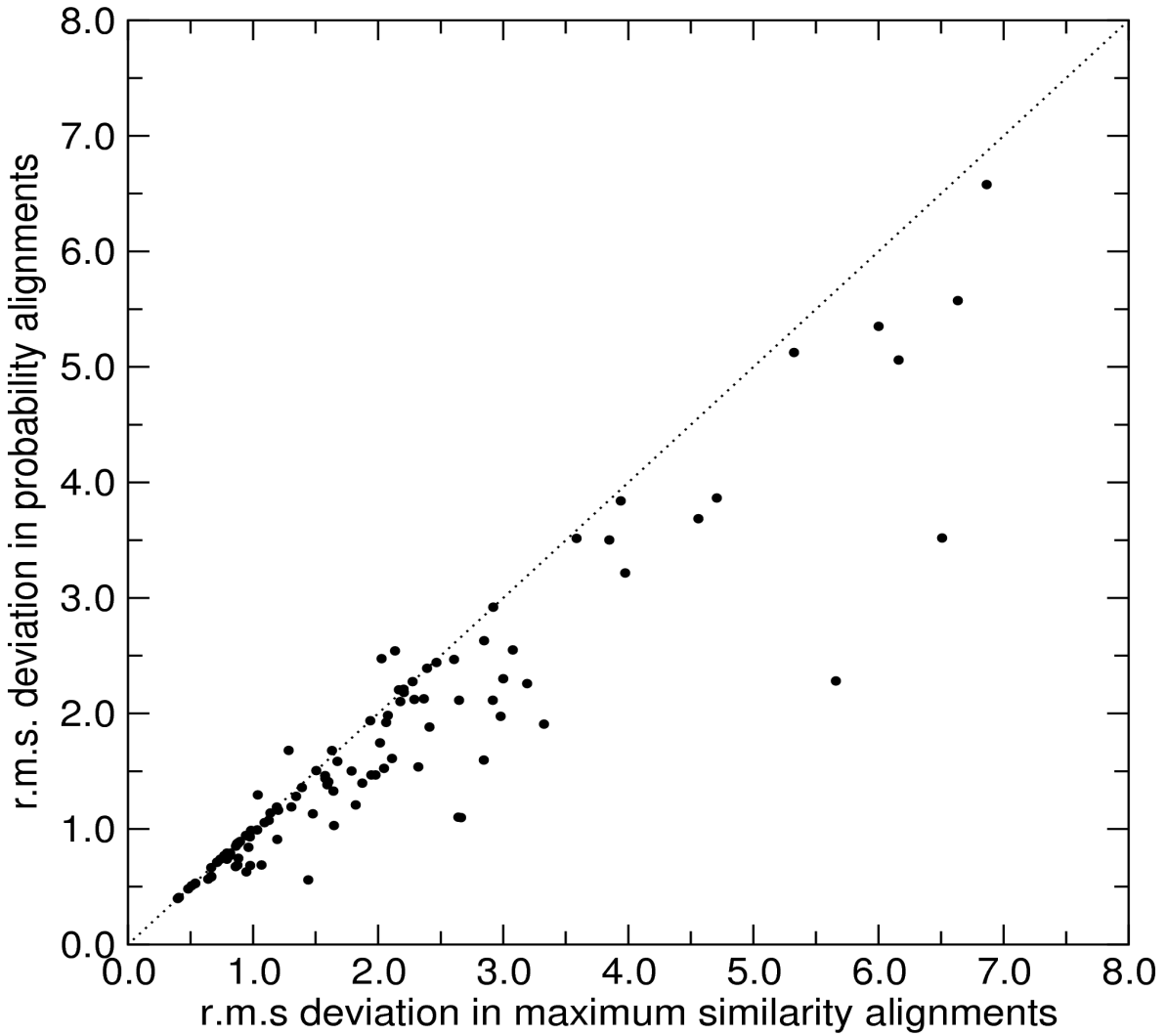
確率アライメント　(forward/backword/transfer-matrix algorithm)

$$A_{prob} = \begin{array}{ll} a_i : b_j & \text{if } p(a_i : b_j) > 0.5 \\ a_i : - & \text{if } p(a_i : -) > 0.5 \\ - : b_j & \text{if } p(- : b_j) > 0.5 \end{array}$$

$$p(a_i : b_j) = \Sigma_{A \in (.,ai:bj,.)} P(A \mid \mathbf{a}, \mathbf{b}, \theta), \quad p(a_i : -) = 1 - \Sigma_j p(a_i : b_j)$$

確率アラインメントと最も確からしいアラインメントの比較

$$r.m.s.d. \equiv \left[\min_{R,t} \left\{ \sum_{(ai,bj)\subset A} (\mathbf{r}_{ai} - R(\mathbf{r}_{bj} - \mathbf{t}))^2 \Big/ \sum_{(ai,bj)\subset A} 1 \right\} \right]^{1/2}$$



(Prot. Eng., **8**, 999, 1995)

アラインメントのための評価関数の例:

$S(A|\ \mathbf{a},\ \mathbf{b},\ \theta\ ) = \Sigma_{(ai:bj)\in A}\ \ s_{aibj} + \Sigma_{gap\in A}\ s\_(gap\text{-}length)$

s_{aibj} :　塩基/アミノ酸間($a_i \rightleftarrows b_j$)の置換の生じ易さを表す評価値

$s\_(k)$: 下に凸の単調非増加関数; 例えば $s\_(k) = s\_ + (k-1)\Delta s\_\ ,\ s\_ \le \Delta s\_ \le 0$

パラメーター推定

- 配列 **a**, **b** に対して最尤推定:

$$\hat{\theta} = \arg\max_{\theta} P(\mathbf{a},\ \mathbf{b}|\ \theta) = \arg\max_{\theta} \Sigma_A P(A,\ \mathbf{a},\ \mathbf{b}|\ \theta)$$

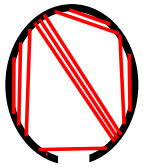
- 既知のデータ D より最尤推定:

$$\hat{\theta} = \arg\max_{\theta} P(D|\ q)$$

例:　$s_{\alpha\beta} = \log\ \{f(\alpha:\beta)\ /\ f(\alpha)f(\beta)\}$,　$s\_(k) = \log\ \{f_k\ /\ f_{no\text{-}gap}\}$

- 定常可逆マルコフ過程モデルに基づく塩基/コドン/アミノ酸置換確率行列 ($S_{\alpha\beta}$) :

$$s_{\alpha\beta} = \log\ \{f(\alpha)S_{\alpha\beta}(t\rightarrow\text{大})\ /\ f(\alpha)f(\beta)\}$$



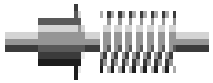
RNA 2次構造予測 （塩基対の交差を無視）

$$P(R_2 \mid \mathbf{a}, \theta) = \frac{\exp(-\beta E(R_2 \mid \mathbf{a}, \theta))}{\sum_{R2} \exp(-\beta E(R_2 \mid \mathbf{a}, \theta))}$$

最も確からしい 2次構造: $\mathbf{R}_{2,\max} = \arg\max_{R2} P(R_2 \mid \mathbf{a}, \theta)$ (CYK algorithm)

塩基対確率(> 0.5)の塩基対からなる2次構造:

$$P(i\text{-}j \mid \mathbf{a}, \theta) = \frac{\sum_{R2} \exp(-\beta E(R_2 \mid \mathbf{a}, \theta))\ \delta_{R2 \ni i\text{-}j}}{\sum_{R2} \exp(-\beta E(R_2 \mid \mathbf{a}, \theta))}$$ (inner/outer)



蛋白質 2次構造予測 (transfer-matrix/forward/backward algorithm)

$$P(\mathbf{C}_2 \mid \mathbf{a}, \theta) = \frac{\exp(-\beta E(\mathbf{C}_2 \mid \mathbf{a}, \theta))}{\sum_{C2} \exp(-\beta E(\mathbf{C}_2 \mid \mathbf{a}, \theta))}$$

最も確からしい 2次構造: $\mathbf{C}_{2,\max} = \arg\max_{R2} P(\mathbf{C}_2 \mid \mathbf{a}, \theta)$

各残基の２次構造確率に基づく2次構造: $\mathbf{C}_{2,\mathrm{prob},i} = \arg\max_{\gamma} P(C_{2i}=\boldsymbol{\gamma} \mid \mathbf{a}, \theta)$

$$P(C_{2i}=\gamma \mid \mathbf{a}, \theta) = \frac{\sum_{C2} \exp(-\beta E(\mathbf{C}_2 \mid \mathbf{a}, \theta))\ \delta_{c2i,\gamma}}{\sum_{C2} \exp(-\beta E(\mathbf{C}_2 \mid \mathbf{a}, \theta))}$$

3次元系: 3次元構造上の相互作用を含む

NP-完全のため発見的方法による

- RNA 2次/3次構造予測（塩基対交差を含む）

- 蛋白質 構造アラインメント

$$A_{\max}(d_{\text{threshold}}) = \arg\max_{A} \max_{R,\mathbf{t}} \sum_{(ai,bj)\subset A} H(d_{\text{threshold}} - |\mathbf{r}_{ai} - R(\mathbf{r}_{bj} - \mathbf{t})|)$$

- 蛋白質 配列-構造アラインメント / 配列-構造適合度評価

  配列設計(蛋白質構造に最適な配列の設計)問題に類似

- 蛋白質3次構造予測

$$\boldsymbol{C}_{\text{native}} = \arg\min_{\mathbf{C}} E(\boldsymbol{C} \mid \boldsymbol{a}, \theta)$$

蛋白質立体構造ドメインの折り畳みの分類

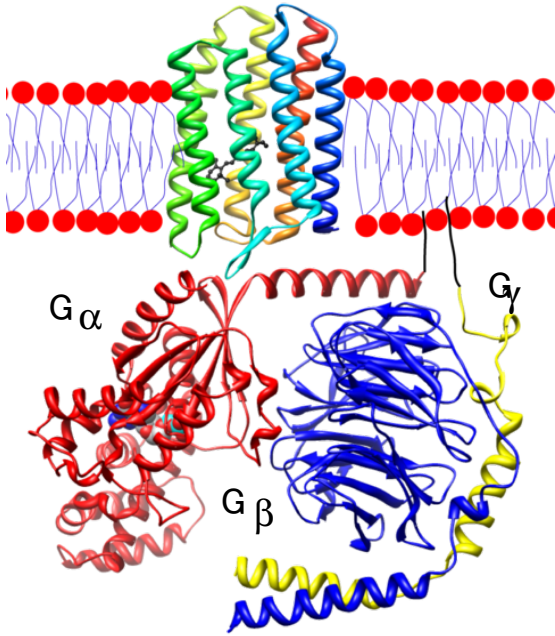
ドメイン(< 200アミノ酸)の折り畳み(fold)の種類は、たかだか数千種類と考えられている。(Chothia, 1992)

2mm1: Human myoglobin

Scop Classification Statistics (Release 1.73)

	#Folds	#Superfamilies	#Families
All alpha proteins	259	459	772
All beta proteins	165	331	679
a/b proteins	141	232	736
a+b proteins	334	488	897
Multi-domain	53	53	74
Membrane proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

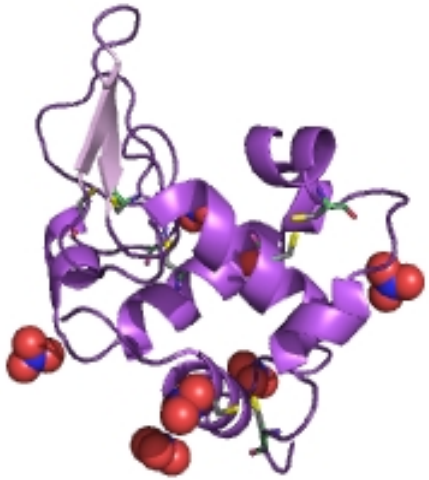
Rhodopsin



Transducin

1bww: Ig Kappa V

1hti: Triosephosphate isomerase, Human



1jsf: Human lysozyme

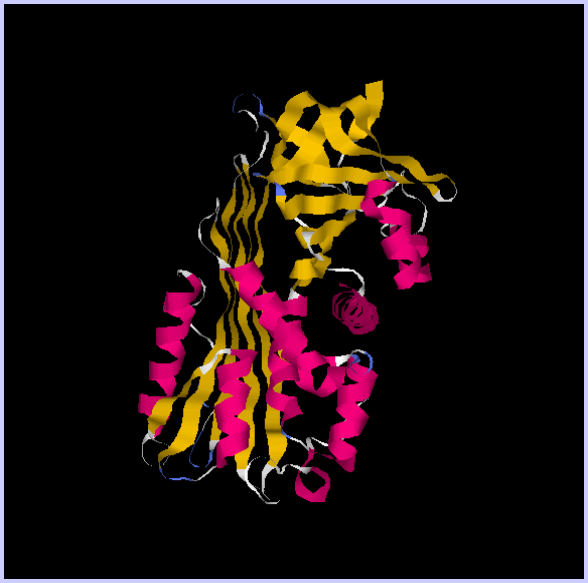
1qmn: Antichymotrypsin, Human

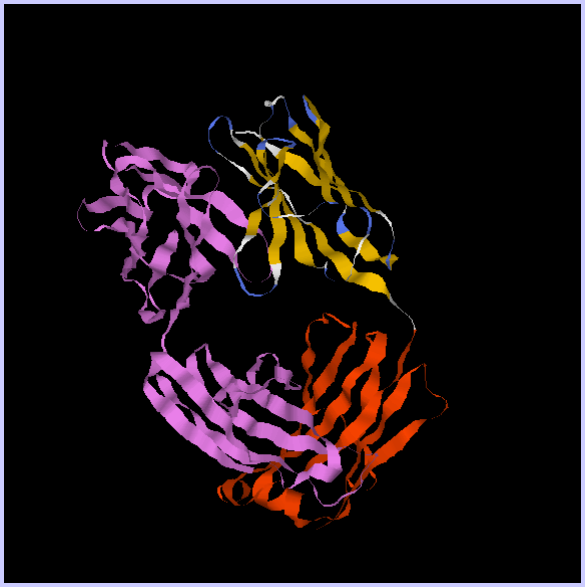
配列に適合する構造の検索

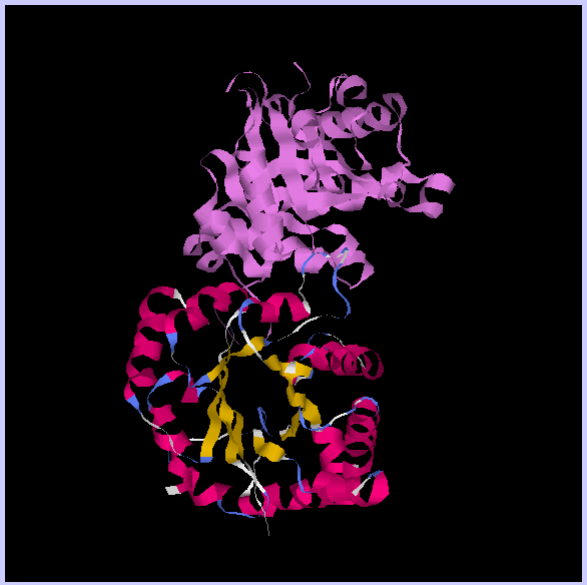
YNNIP ...
VTLR

Threading

適合度評価関数







蛋白質 配列-構造アラインメント / 配列-構造適合度評価

$$P(A \mid \mathbf{a}, \mathbf{C}, \theta) = \frac{P(\mathbf{C} \mid A, \mathbf{a}, \theta)\, P(A \mid \mathbf{a}, \theta)}{\sum_A P(A, \mathbf{C} \mid \mathbf{a}, \theta)} = \frac{\exp(-\beta\, \mathcal{E}(A \mid \mathbf{a}, \mathbf{C}, \theta))}{\sum_A \exp(-\beta\, \mathcal{E}(A \mid \mathbf{a}, \mathbf{C}, \theta))}$$

$$\log P(\mathbf{C} \mid A, \mathbf{a}, \theta) = -\beta\, E(\mathbf{C} \mid A, \mathbf{a}, \theta) - \log \sum_{\mathbf{C}} \exp(-\beta\, E(\mathbf{C} \mid A, \mathbf{a}, \theta))$$

第2項(分配関数)の評価: 高温近似

$$\mathcal{E}(A \mid \mathbf{a}, \mathbf{C}, \theta) \equiv \Delta E(\mathbf{C} \mid A, \mathbf{a}, \theta) + n_{\mathrm{aligned}}\, E_0 + \sum_{\mathrm{gap}} E(\text{gap-length})$$

$$\Delta E(\mathbf{C} \mid A, \mathbf{a}, \theta) \equiv E(\mathbf{C} \mid A, \mathbf{a}, \theta) - \langle E(\mathbf{C} \mid A, \mathbf{a}, \theta) \rangle_{\mathrm{native}}$$

配列設計: 蛋白質構造 C に最適な配列 a を設計

$$P(\mathbf{a} \mid \mathbf{C}, \theta) = \frac{P(\mathbf{C} \mid \mathbf{a}, \theta)\, P(\mathbf{a} \mid \theta)}{P(\mathbf{C} \mid \theta)}$$

$$\mathbf{a}_{\mathrm{opt}} = \arg\max_{\mathbf{a}} \log [\, P(\mathbf{C} \mid \mathbf{a}, \theta)\, P(\mathbf{a} \mid \theta)\,]$$

$$\log P(\mathbf{C} \mid \mathbf{a}, \theta) = -\beta\, E(\mathbf{C} \mid \mathbf{a}, \theta) - \log \sum_{\mathbf{C}} \exp(-\beta\, E(\mathbf{C} \mid \mathbf{a}, \theta))$$

困難: 第2項(分配関数)の評価

蛋白質 配列-構造アラインメント，蛋白質3次構造予測における相互作用ポテンシャルの条件

- $E(\mathbf{C}_M, \mathbf{C}_S \mid \mathbf{a}, \theta)$ は、水との相互作用を含む。

  水分子を陽的に考慮せず、水との相互作用を精度高く評価することは困難

- 主鎖構造評価のために

  $F(\mathbf{C}_M \mid \mathbf{a}, \theta) = -\beta \log \int \exp(-\beta E(\mathbf{C}_M, \mathbf{C}_S \mid \mathbf{a}, \theta))\, d\mathbf{C}_S$　も必要。

  高精度の評価は困難

⇒

細かい粒度の原子間相互作用ポテンシャル
粗い粒度の原子間/残基間統計ポテンシャル
は同程度の精度

水溶液中の蛋白質の主鎖構造エネルギーを的確に評価できるポテンシャルが必要とされている。

統計ポテンシャル:

残基間接触エネルギー（$e_{ab} \equiv e_{rr} + \Delta e_{ab}$）　を

既知の蛋白質構造における残基間接触数より評価

格子模型

$$E_{total} = \Sigma_i \Sigma_{j>i} e_{aibi} \Delta_{ij} + \text{constant} = \Sigma_i \Sigma_{j>i} \Delta e_{aibi} \Delta_{ij} + n_{rr} e_{rr} + \text{constant}$$

$\Delta_{ij} \equiv 1$ if in contact ($|\mathbf{r}_i - \mathbf{r}_j| < d_c$), 0 otherwise

$e_{ab} \equiv e_{rr} + \Delta e_{ab} \equiv \varepsilon_{ab} + \varepsilon_{ss} - \varepsilon_{as} - \varepsilon_{sb}$

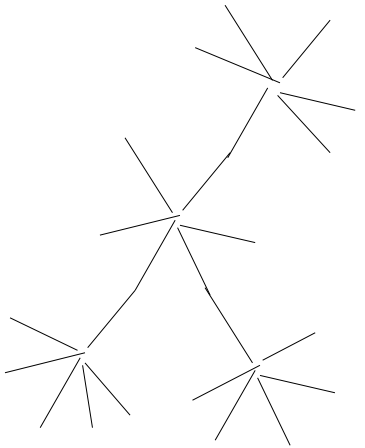
ε_{ab}, ε_{ss}, $\varepsilon_{as} = \varepsilon_{sa}$:
アミノ酸a, b間、水間、アミノ酸aと水間の相互作用エネルギー

Bethe近似

$$\Delta e_{ab} = -\log \frac{n_{ab}\, n_{rs}\, n_{rs}}{n_{rr}\, n_{as}\, n_{sb}} \qquad e_{rr} = -\log \frac{n_{rr}\, n_{ss}}{n_{rs}\, n_{sr}}$$

接触数　　$n_{ar} = \Sigma_b n_{ab}$　　$n_{rr} = \Sigma_a n_{ar}$　　$n_{rs} = \Sigma_a n_{as}$

$n_{as} + n_{ar} = q\, n_a / 2$



配列ー構造アラインメント

```
minimum energy alignment
    sequence  3GRS 364  YNNIPTVV-FSHPPIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                           |  |  |  | |    |                 | |
    structure 1NPX 322  GVQGSSGLAVFDYKFASTGINEVMA-QKLGK-ETKAVTVV -EDYLMDFNPDKQKAWF
probability alignment
    sequence  3GRS 364  YNNIPTVV-FSHPPIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                    ?      |  |  ?  | | |  |   ?             | |
    structure 1NPX 322  GVQGSSGLAVFDYKFASTGINEVM-AQKLGKE-TKAVT-V VEDYLMDFNPDKQKAWF
                        7777664334334698999887541577776424333203 344444444455566666

              1NPX 322       bbbbb  bbbbbb  aaaa aaaa     bbbb b bbbb        bbbbb
                          ######                 ######                  #########
              3GRS 364       bbb     bbbbbbbaaaaaaaaaa  bbbbbbbb b aaaaa       bbb
minimum energy alignment
    structure 3GRS 364    YNNIPTVVFSHP PIGTVGLTEDEAIHKYGIENVKTYSTS FTPMYHAVTKRKTKCVM
    matched to:                                      | | |            |       | |
    sequence  1NPX 322    GVQGSSGLAVFD YKFASTGINEVMAQKLGKETKAVTVVE DYLMDF--NPDKQKAWF
probability alignment
    structure 3GRS 364  --YNNIPTVVFSHP-PIGTVGLTEDEAIHKYGIENVKTYSTS-FTPMYHAVTKRKTKCVM
    matched to:           ??          ?     |  |?    | | |       ?? ???       | |
    sequence  1NPX 322  GV--QGSSGLAVFDYKFASTGINE-VMAQKLGKETKAVTVVEDY---LMDFNPDKQKAWF
                        4322334444444303455565554145667776543322222021112233566677777


minimum energy alignment
    sequence  3GRS 420  KMVCANKEEKVVGIHMQG-LGCDEMLQGFAVAVKMGATKADFDNT-VAIHPTSSEE   L
    matched to:         | |         |                  |     |  |         |
    structure 1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKP   W
probability alignment
    sequence  3GRS 420  KMVCANKEEKVVGIHM-QGLGCDEMLQGFAVAVKMGATKADFDNT-VAIHPTS-SEE- L
    matched to:         | |         |   ?              |     |  |    ?    |      ?
    structure 1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKPWN I
                        66666667777776540456799999888888888888888764344554332222221 2

              1NPX 376  bbbb        bbbbbb       aaaaaaaaaa  aaaaaaa             a a
                        #####              #########################
              3GRS 420  bbbbb b  bbbbbbbbbb      aaaaaaaaaa  aaaaaaa           aaaaa
minimum energy alignment
    structure 3GRS 420  KMVCA-NKEEKVVGIHMQGLGCDEMLQGFAVAVKMGATKADFDNT----VAIHPTSSEEL
    matched to:         | |             |              |     |  |
    sequence  1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFFFQPAFDKPWNII
probability alignment
    structure 3GRS 420  KMVCANKEEKVVG-IHMQGLGCDEMLQGFAVAVKMGATKADFDN----TVAIHPTSSEEL
    matched to:         | |         |?  |              |     |  |       ?     ??????
    sequence  1NPX 376  KLVYDPETTQILGAQLMSKADLTANINAISLAIQAKMTIEDLAYADFF------------
                        77665344344430344433444444555556777888888764335622222111100


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

              1NPX 435  a               aaaaaaaaaaa
                                            ########

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

(Prot. Eng., **13**, 459, 2000)

分子系統樹推定: n 配列からなる無根系統樹の可能な数: (2n − 5)!!

$$P(T_p,\ T_B \mid A,\ \{\mathbf{a}_\mu\},\ \theta) = \frac{P(A,\ \{\mathbf{a}_\mu\} \mid T_p,\ T_B,\ \theta)\ P(T_p,\ T_B \mid \theta)}{P(A,\ \{\mathbf{a}_\mu\} \mid \theta)}$$

最尤法: $(\hat{T}_p,\ \hat{T}_B) = \arg\max_{Tp,TB} P(A,\ \{\mathbf{a}_\mu\} \mid T_p,\ T_B,\ \theta)$ (NP-hard)

塩基/コドン/アミノ酸置換モデル θ: 定常可逆マルコフ過程

置換確率行列 $S(t) = \exp(R\ t)$

$\Sigma_\mu\ f_\mu\ R_{\mu\nu} = 0$ $R_{\mu\nu} = r_{\mu\nu}\ f_\nu$ $r_{\mu\nu} = r_{\nu\mu}$

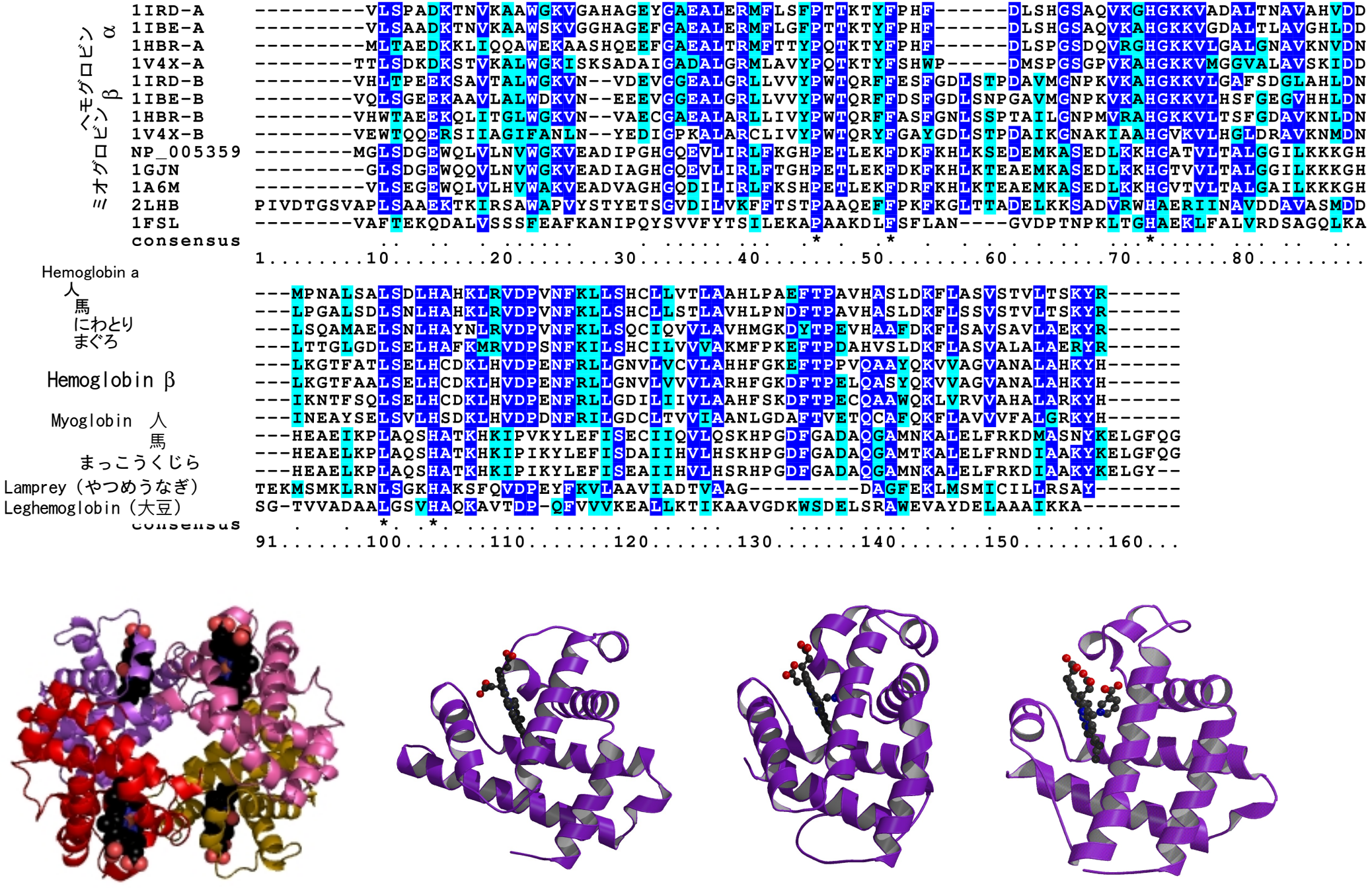
$r_{\mu\nu}$の推定:

- $(\hat{T}_p,\ \hat{T}_B,\ \hat{\mathbf{r}}) = \arg\max_{\mathbf{r}} P(A,\ \{\mathbf{a}_\mu\} \mid T_p,\ T_B,\ \mathbf{r},\ \mathbf{f})$
- $\exp(\hat{R}\hat{t}) = S^{obs}$
- コドン置換モデル: $R_{\mu\nu} = m_{\mu\nu}\ f_\nu \exp(w_{\mu\nu})$

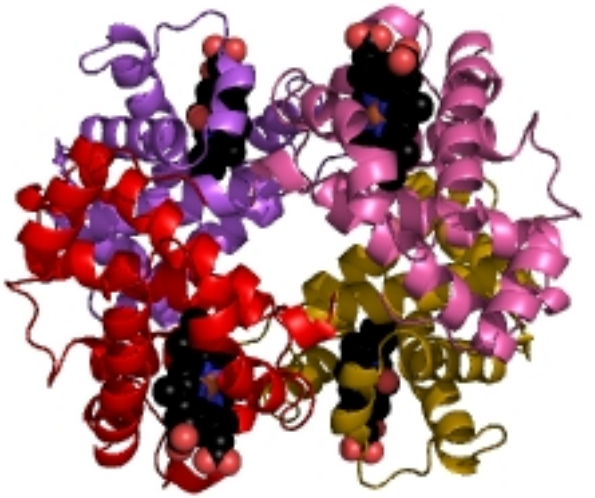
$$w_{\mu\nu} = -\beta\ \Delta\varepsilon_{\ a(\mu)b(\nu)} + w_0\ (\ 1 - \delta_{a(\mu)b(\nu)})$$

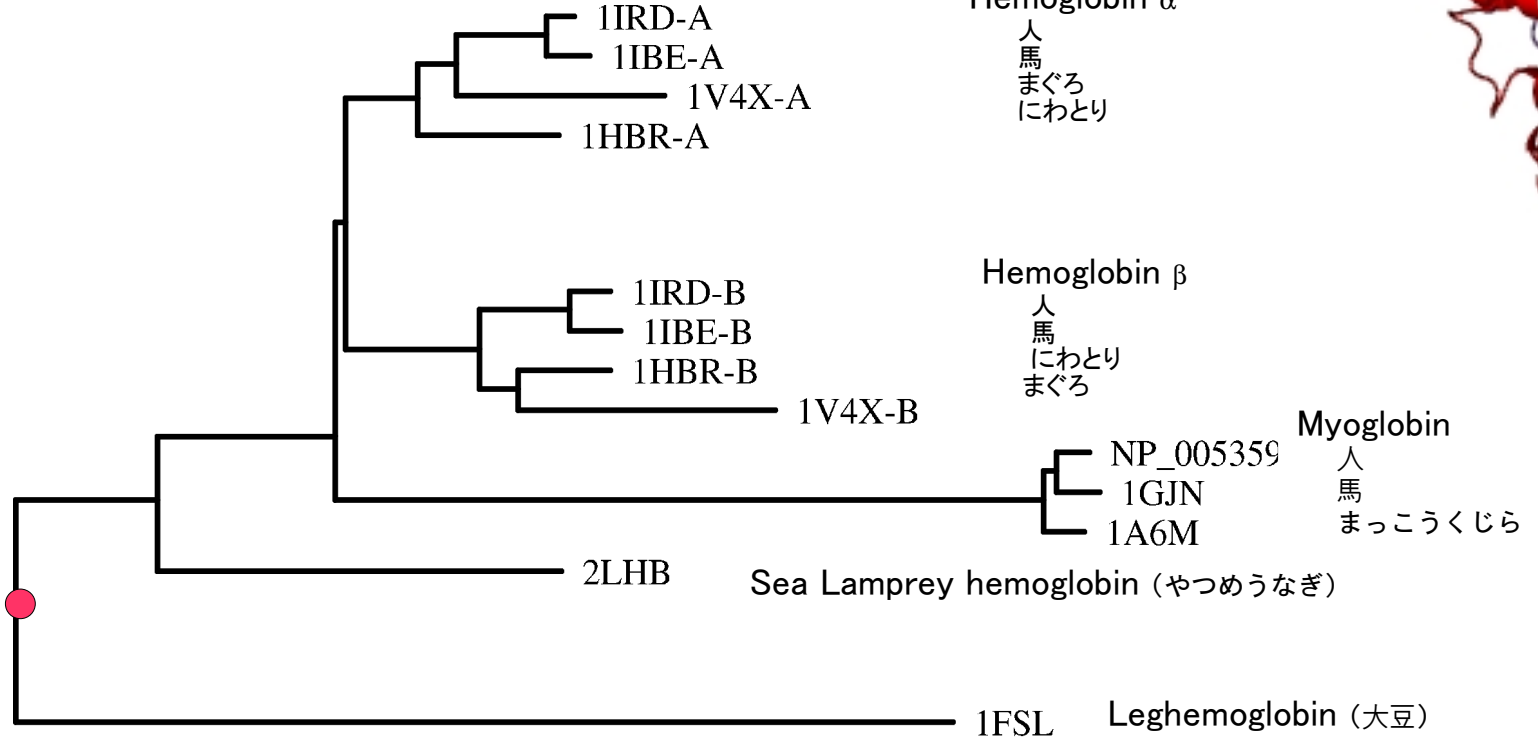
$$\Delta\varepsilon_{\ ab} = \Sigma_c\ (e_{bc} - e_{ac})\ (N_{ac}/N_a - N_{bc}/N_b) \geq 0$$

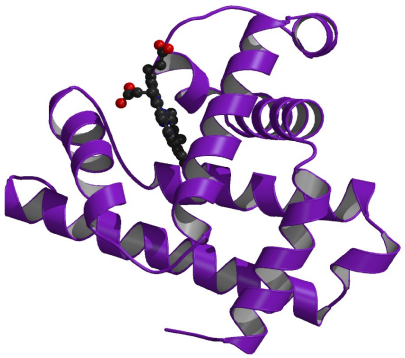
グロビン蛋白質のアライメント

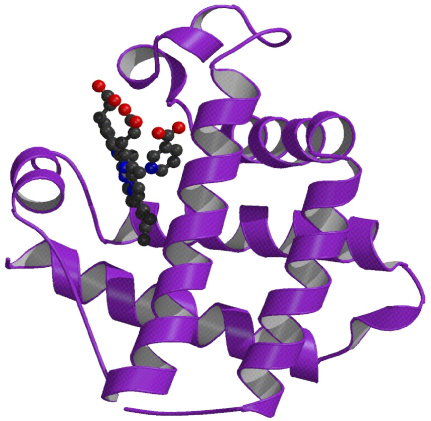


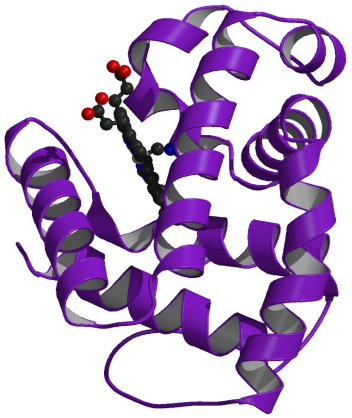
Rooted tree by neighbor-joining method



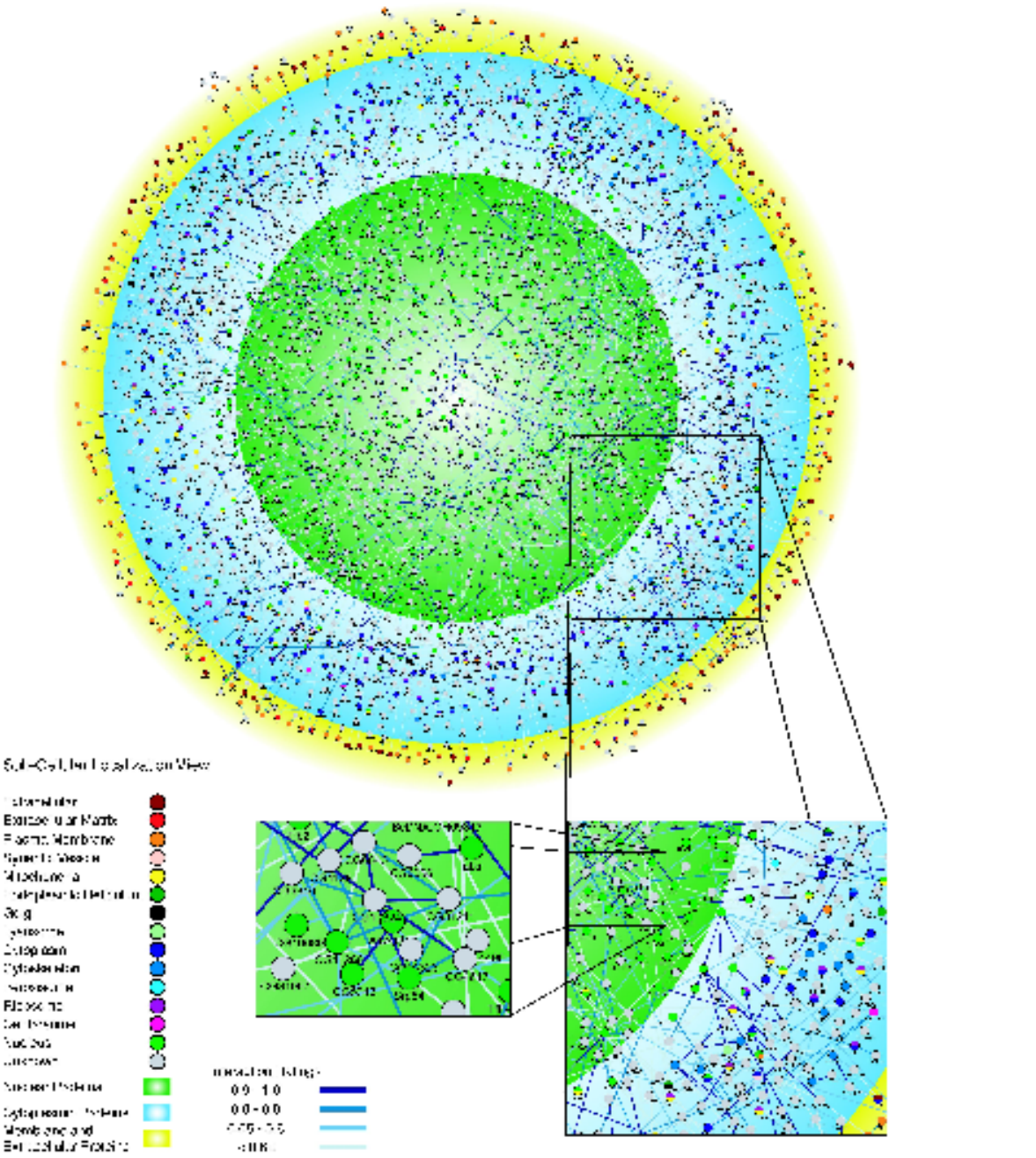








タンパク質相互作用マップ（Giot et al., Science, 203, 1727, 2003）



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
  - 代謝物パスウェー
  - シグナリングネットワーク
  - パスウェー比較
- 生体システムの計算機シュミレーション

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向へ発展しつつある。

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Rasmol/MolScript/Raster3Dを用いた.