立体構造からの情報抽出

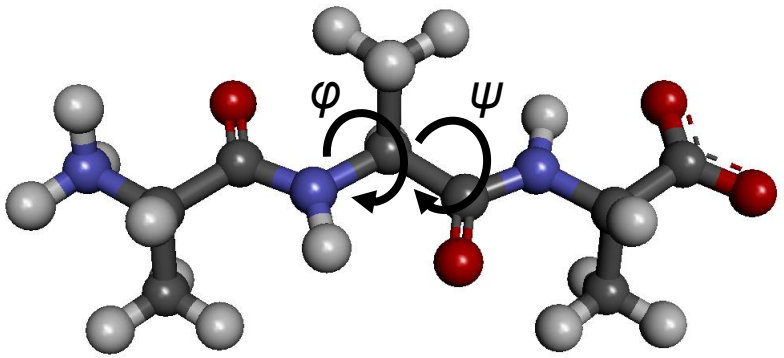
東京大学大学院農学生命科学研究科
アグリバイオインフォマティクス
教育研究ユニット
寺田　透

本日の講義内容

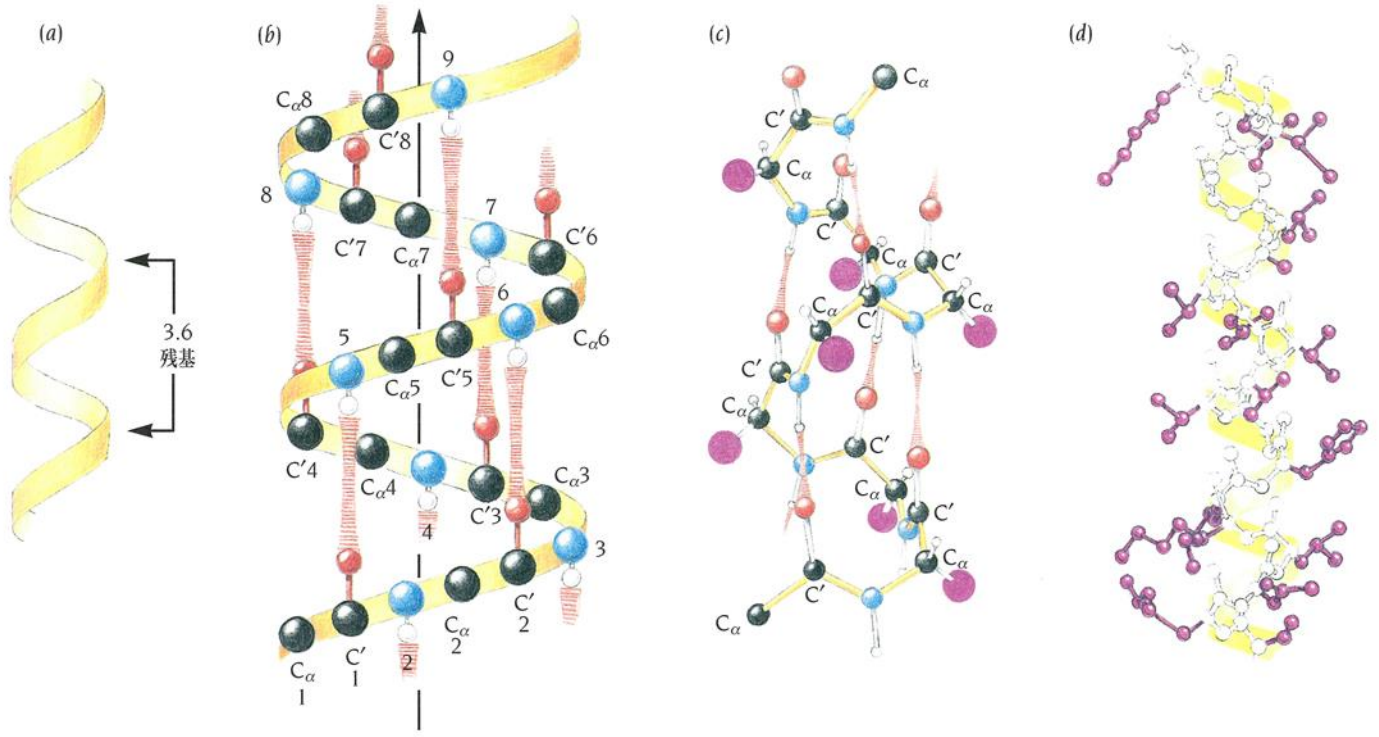
- ２次構造形成傾向
  - 実習課題１
- 3D profile法
- 立体構造比較
- 立体構造類似性と機能・進化
- 立体構造分類データベース
  - 実習課題２
- 配列類似性と立体構造類似性

2次構造

- α helix
  - 残基*i*のC=Oと残基*i*+4のN-Hが水素結合を形成
  - 主鎖の二面角(φ, ψ) ≈ (–60˚, –60˚)
- β sheet
  - 隣り合うβ strandとの間で水素結合を形成
  - 主鎖の二面角(φ, ψ) ≈ (–120˚, 120˚)
  - 平行と逆平行がある

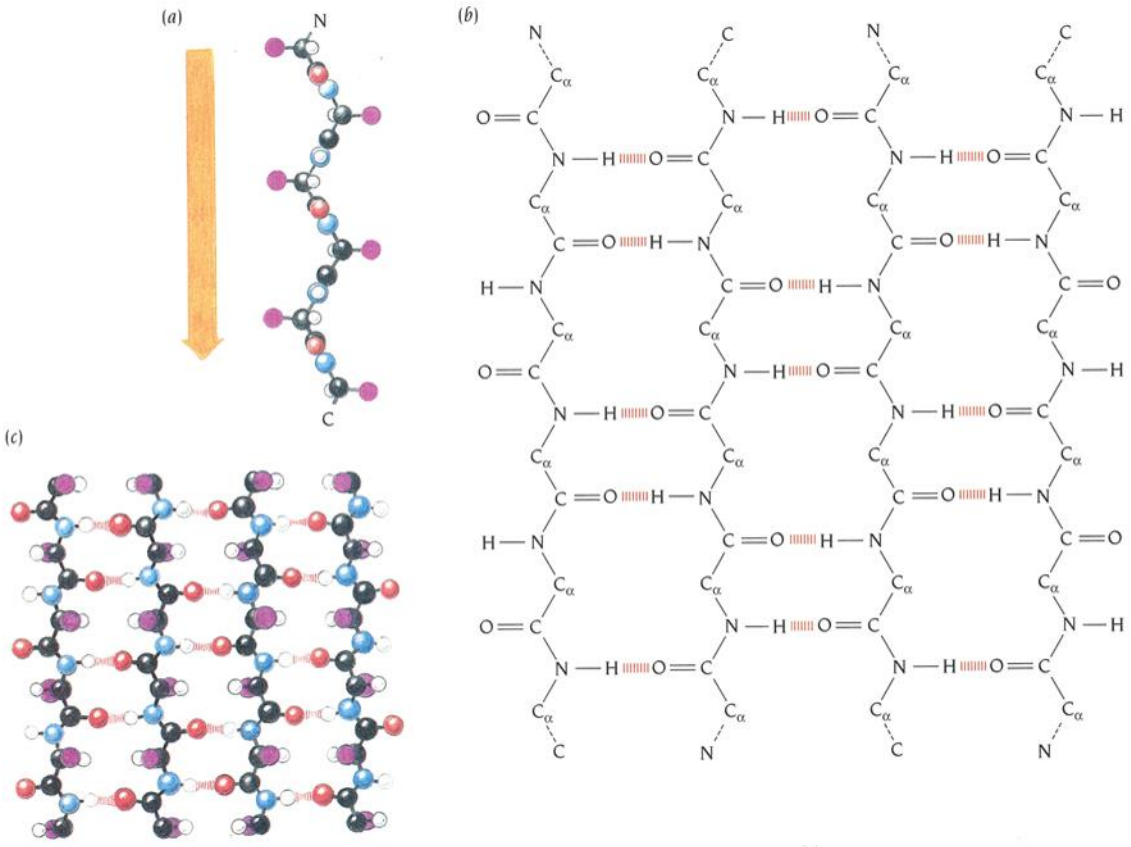


α helix



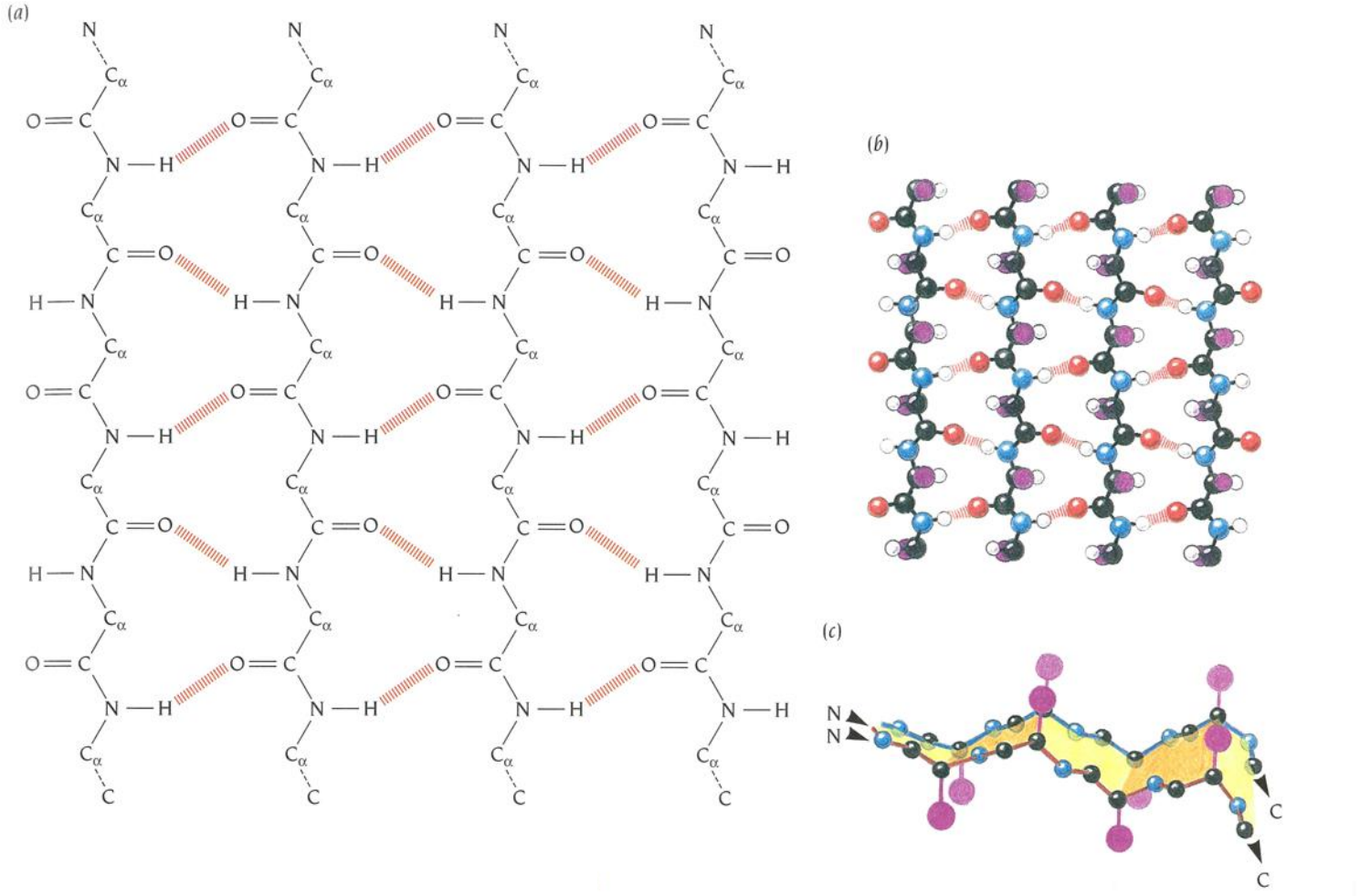
Branden & Tooze, “Introduction to Protein Structure” 2nd Ed.

逆平行β sheet



Branden & Tooze, “Introduction to Protein Structure” 2nd Ed.

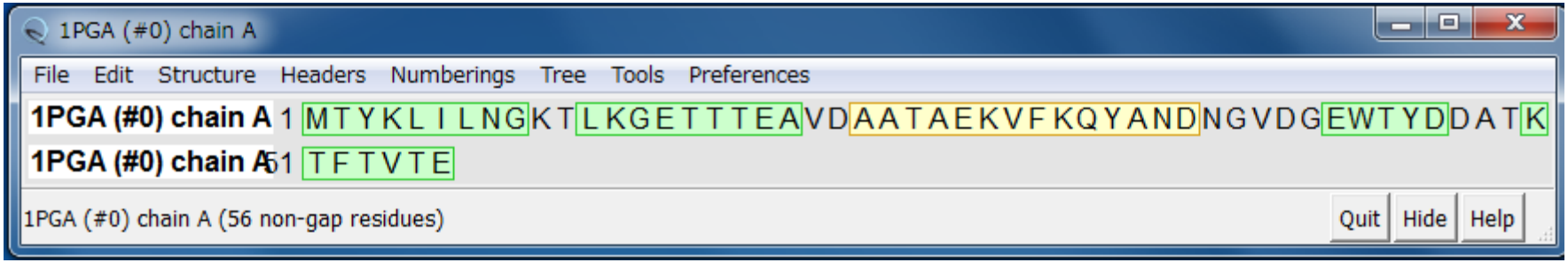
平行β sheet



Branden & Tooze, “Introduction to Protein Structure” 2nd Ed.

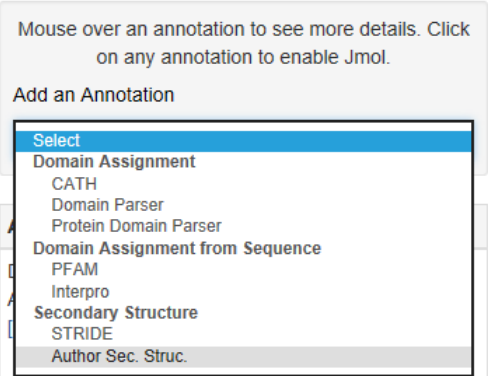
2次構造の確認（1）

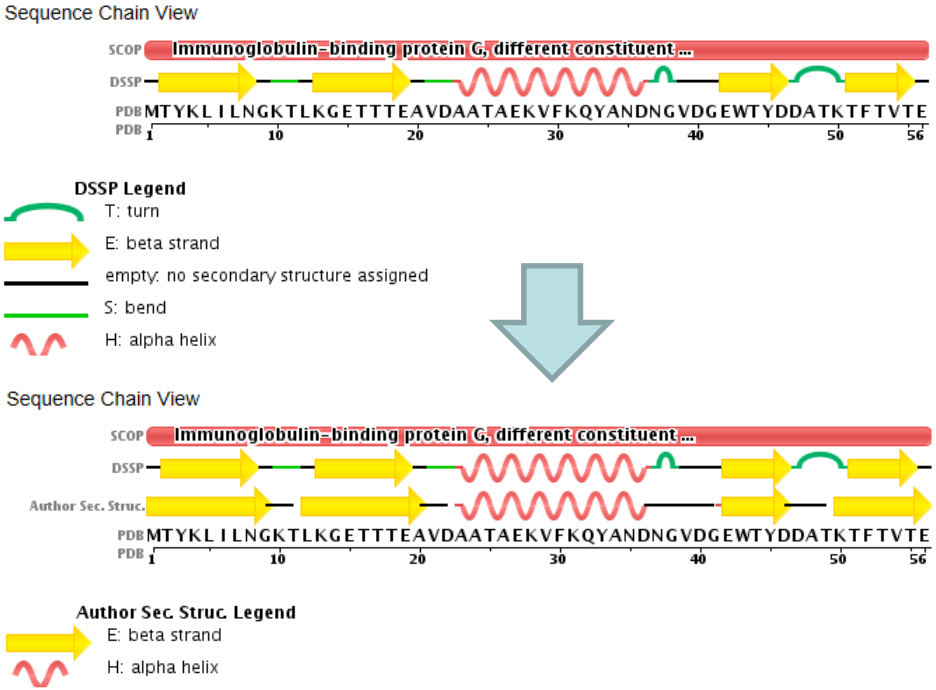
1. Chimeraで1PGAを開く
   - α helix、β sheet（並行および逆並行）を確認
2. メニューの「Tools」→「Sequence」→「Sequence」
   →α helixが黄色、β strandが緑色



2次構造の確認(2)

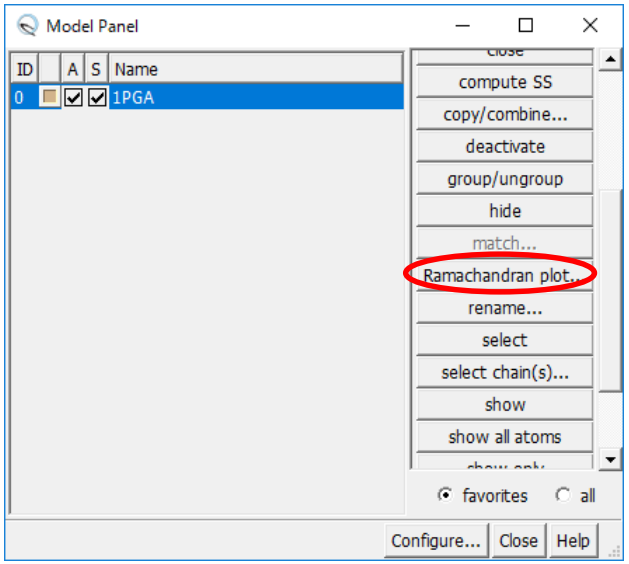
1. RCSBのサイトで1PGAを表示する
2. Sequenceタブを開く
3. 以下のようにAuthor Sec. Struc.を選択

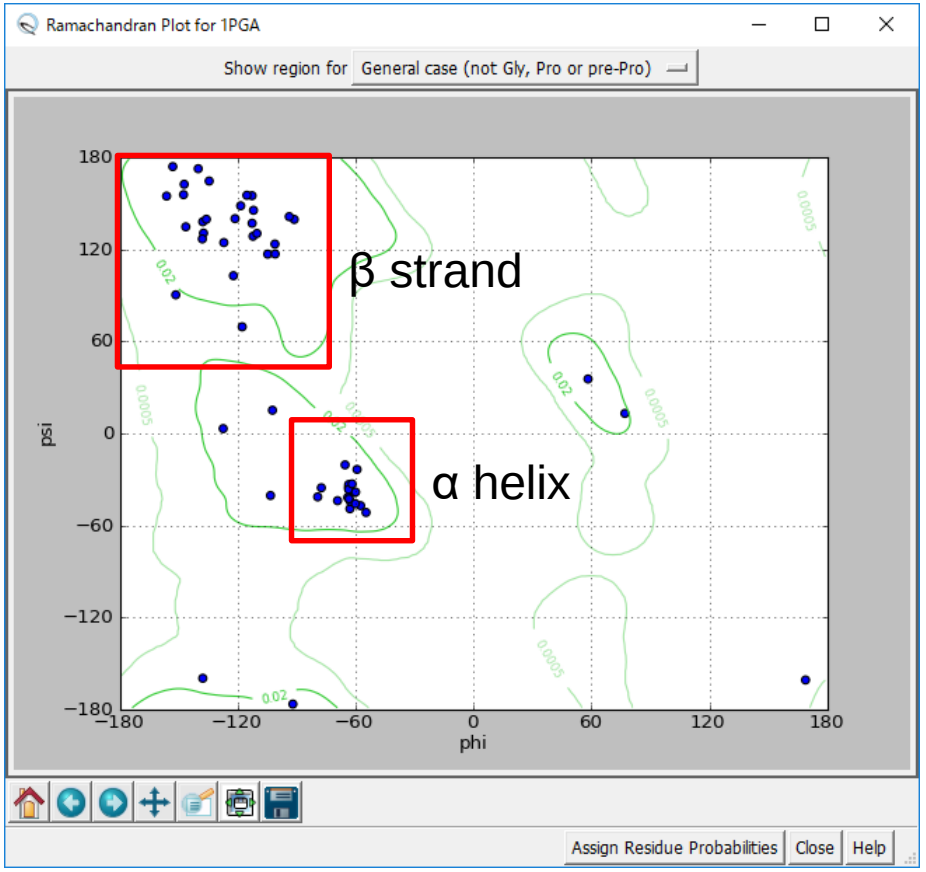




Ramachandran plot

- Chimeraでメニューから「Favorites」→「Model Panel」を選択
- 「Ramachandran plot」をクリック





２次構造データの由来

- PDBのヘッダから
  - 立体構造を決定した人が記述

```
HELIX    1   1 ALA A   23  ASP A   36  1
SHEET    1  S1 4 LEU A  12  ALA A  20  0
SHEET    2  S1 4 MET A   1  GLY A   9 -1
SHEET    3  S1 4 LYS A  50  GLU A  56  1
SHEET    4  S1 4 GLU A  42  ASP A  46 -1
```

- 立体構造から
  - 立体構造の特徴（水素結合ネットワークや主鎖の二面角）から判定する

DSSP

- 立体構造に基づいて2次構造を判定するソフトウェア*
  - 水素結合ネットワークに基づいて判定
  - DSSPの出力（H: α helix; E: β strand）

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI
1	1	A	M	0	0	116	0, 0.0	19,-2.8	0, 0.0	2,-0.5	0.000	360.0	360.0	360.0	151.3
2	2	A	T E	-A	19	0A	91	17,-0.2	2,-0.3	19,-0.1	17,-0.2	-0.900	360.0	-168.8	-101.0	123.5
3	3	A	Y E	-A	18	0A	5	15,-2.8	15,-2.8	-2,-0.5	2,-0.3	-0.774	9.0	-143.9	-112.6	155.1
4	4	A	K E	-Ab	17	51A	64	46,-1.7	48,-2.7	-2,-0.3	2,-0.5	-0.882	3.2	-156.2	-118.4	148.5
5	5	A	L E	-Ab	16	52A	0	11,-3.4	11,-1.9	-2,-0.3	2,-0.6	-0.995	1.8	-161.2	-126.9	124.5
6	6	A	I E	-Ab	15	53A	60	46,-2.8	48,-2.5	-2,-0.5	2,-0.6	-0.954	13.6	-159.1	-104.6	116.9
7	7	A	L E	+Ab	14	54A	6	7,-2.2	7,-1.8	-2,-0.6	2,-0.4	-0.907	18.6	174.1	-100.8	117.1
8	8	A	N E	+Ab	13	55A	75	46,-3.1	48,-2.4	-2,-0.6	5,-0.2	-0.695	26.5	149.4	-117.6	69.5
（中略）																
23	23	A	A H	> S+	0	0	25	-2,-0.3	4,-2.9	1,-0.2	5,-0.2	0.805	117.0	60.8	-63.5	-35.1
24	24	A	A H	> S+	0	0	49	1,-0.2	4,-1.0	2,-0.2	-1,-0.2	0.859	109.0	44.1	-63.5	-33.4
25	25	A	T H	> S+	0	0	60	-3,-0.5	4,-1.9	2,-0.2	3,-0.3	0.918	112.7	49.9	-79.2	-41.4
26	26	A	A H	X S+	0	0	0	-4,-1.9	4,-2.8	1,-0.2	5,-0.3	0.912	107.5	57.1	-63.5	-36.6
27	27	A	E H	X S+	0	0	61	-4,-2.9	4,-1.9	1,-0.2	-1,-0.2	0.855	107.8	46.7	-59.9	-38.3

*Kabsch & Sander, *Biopolymers* **22**, 2577 (1983).

２次構造形成傾向（１）

- アミノ酸によってとりやすい2次構造があるかどうか調べる
- 右図は15個のタンパク質について、α helix、β sheet、coilに含まれるアミノ酸をカウントしたもの*

	α helix	β sheet	Coil	Total
Ala	119	38	71	228
Arg	22	12	44	78
Asn	35	15	83	133
Asp	39	15	57	111
Cys	15	12	27	54
Gln	40	20	35	95
Glu	62	5	46	113
Gly	45	32	155	232
His	33	9	32	74
Ile	38	29	39	106
Leu	94	41	61	196
Lys	67	22	86	175
Met	12	8	8	28
Phe	33	18	31	82
Pro	18	9	58	85
Ser	57	25	120	202
Thr	47	32	77	156
Trp	18	9	17	44
Tyr	22	22	56	100
Val	74	51	56	181
	890	424	1159	2473

*Chou & Fasman, *Biochemistry* **13**, 211 (1974).

2次構造形成傾向(2)

- アミノ酸ごとに、2次構造の割合を求める
  - どのアミノ酸もβ sheetを作りにくい？
- 2次構造ごとに存在確率が異なる

	α helix	β sheet	Coil	Total
Ala	0.522	0.167	0.311	1.000
Arg	0.282	0.154	0.564	1.000
Asn	0.263	0.113	0.624	1.000
Asp	0.351	0.135	0.514	1.000
Cys	0.278	0.222	0.500	1.000
Gln	0.421	0.211	0.368	1.000
Glu	0.549	0.044	0.407	1.000
Gly	0.194	0.138	0.668	1.000
His	0.446	0.122	0.432	1.000
Ile	0.358	0.274	0.368	1.000
Leu	0.480	0.209	0.311	1.000
Lys	0.383	0.126	0.491	1.000
Met	0.429	0.286	0.286	1.000
Phe	0.402	0.220	0.378	1.000
Pro	0.212	0.106	0.682	1.000
Ser	0.282	0.124	0.594	1.000
Thr	0.301	0.205	0.494	1.000
Trp	0.409	0.205	0.386	1.000
Tyr	0.220	0.220	0.560	1.000
Val	0.409	0.282	0.309	1.000
	0.360	0.171	0.469	1.000

$$p(s|a) = \frac{n(s,a)}{n(a)}$$

２次構造形成傾向（２）

- アミノ酸ごとに、２次構造の割合を求める
  - どのアミノ酸もβ sheetを作りにくい？
- ２次構造ごとに存在確率が異なる

- ２次構造の存在割合で割ればよい

	α helix	β sheet	Coil	Total
Ala	1.45	0.97	0.66	1.00
Arg	0.78	0.90	1.20	1.00
Asn	0.73	0.66	1.33	1.00
Asp	0.98	0.79	1.10	1.00
Cys	0.77	1.30	1.07	1.00
Gln	1.17	1.23	0.79	1.00
Glu	1.52	0.26	0.87	1.00
Gly	0.54	0.80	1.43	1.00
His	1.24	0.71	0.92	1.00
Ile	1.00	1.60	0.79	1.00
Leu	1.33	1.22	0.66	1.00
Lys	1.06	0.73	1.05	1.00
Met	1.19	1.67	0.61	1.00
Phe	1.12	1.28	0.81	1.00
Pro	0.59	0.62	1.46	1.00
Ser	0.78	0.72	1.27	1.00
Thr	0.84	1.20	1.05	1.00
Trp	1.14	1.19	0.82	1.00
Tyr	0.61	1.28	1.19	1.00
Val	1.14	1.64	0.66	1.00
	1.00	1.00	1.00	1.00

２次構造形成傾向（３）

- アミノ酸の２次構造形成傾向

$$P_{s,a} = \frac{p(s|a)}{p(s)}$$

$p(s|a)$: アミノ酸aにおける２次構造sの出現確率
$p(s)$: ２次構造sの平均出現確率
$P_{s,a}$: アミノ酸aが２次構造sをとる確率は、平均と比べてどのくらい高いか
→アミノ酸aの２次構造s形成傾向

- 配列の場合

$$p(s_1 s_2 | a_1 a_2) = p(s_2 | s_1 a_1 a_2) p(s_1 | a_1) \approx p(s_2 | a_2) p(s_1 | a_1)$$

位置2における2次構造の出現確率は位置1の2次構造やアミノ酸によらないと仮定

$$P_{s_1 \cdots s_N, a_1 \cdots a_N} = \prod_{i=1}^{N} \frac{p(s_i | a_i)}{p(s_i)}$$

配列の2次構造形成傾向（1）

- seq_score.xlsxをダウンロード
  - 緑色のセルにアミノ酸の$P_{s,a}$の値を貼り付け
  - 橙色のセルに配列を入力
  - Productに、その配列がα helixをとった場合、β sheetをとった場合、coilをとった場合のスコアが表示される
- 1PGAの23–36残基（AATAEKVFKQYAND）の2次構造形成傾向

Sequence	A	A	T	A	E	K	V	F	K	Q	Y	A	N	D	Product
α helix	1.45	1.45	0.84	1.45	1.52	1.06	1.14	1.12	1.06	1.17	0.61	1.45	0.73	0.98	4.14
β sheet	0.97	0.97	1.20	0.97	0.26	0.73	1.64	1.28	0.73	1.23	1.28	0.97	0.66	0.79	0.25
Coil	0.66	0.66	1.05	0.66	0.87	1.05	0.66	0.81	1.05	0.79	1.19	0.66	1.33	1.10	0.14

  - スコアはα helixのものが一番大きい
  - この配列はα helixをとりやすいことがわかる

配列の2次構造形成傾向（2）

- 2次構造が与えられた時に、どのアミノ酸が適しているかを表す指標としても利用できる

$$\frac{p(s|a)}{p(s)} = \frac{p(s,a)}{p(s)p(a)} = \frac{p(a|s)}{p(a)} \quad \because\ p(s,a) = p(s|a)p(a)$$

- 1PGAの1–14残基（MTYKLILNGKTLKG）

Sequence	M	T	Y	K	L	I	L	N	G	K	T	L	K	G	Product
α helix	1.19	0.84	0.61	1.06	1.33	1.00	1.33	0.73	0.54	1.06	0.84	1.33	1.06	0.54	0.31
β sheet	1.67	1.20	1.28	0.73	1.22	1.60	1.22	0.66	0.80	0.73	1.20	1.22	0.73	0.80	1.49
Coil	0.61	1.05	1.19	1.05	0.66	0.79	0.66	1.33	1.43	1.05	1.05	0.66	1.05	1.43	0.58

  - α helixには、1–14残基（0.31）より、23–36残基（4.14）の配列が適していることがわかる

データの更新

- ChouとFasmanのデータは非常に古いので、現在のPDBを用いてデータを更新する
- 立体構造データの偏りに注意
  - 例えばhen egg lysozyme（PDB ID: 3AW6など）と100%配列が一致するエントリは631ある
- 類似した配列を除いたデータベースを使う
  - ここではPISCES*から冗長性を除いたPDBのリストを取得
- ２次構造はDSSPを用いて決定
  - 著者が決めたものより客観性がある

*http://dunbrack.fccc.edu/PISCES.php

参考:データの作成法

- dssp-2.0.4-win32.exeとcount.plをダウンロードし、デスクトップに保存
- スタートメニューから「すべてのプログラム」→「アクセサリ」→「コマンドプロンプト」を起動
- 以下を実行

```
C:¥Users¥iu>cd Desktop
C:¥Users¥iu¥Desktop>dssp-2.0.4-win32.exe 1pga.pdb > 1pga.txt
C:¥Users¥iu¥Desktop>count.pl 1pga.txt > count.csv
```

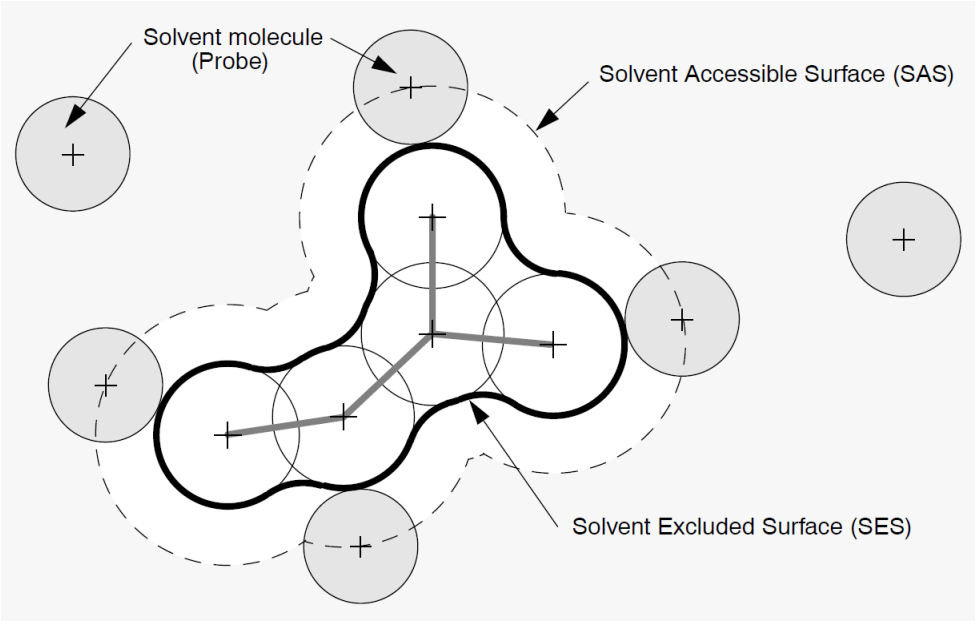
- 生成されたcount.csvをダブルクリックして開く

実習課題1

1. 講義のページからダウンロードしたnew_data.xlsを利用してアミノ酸の2次構造形成傾向$P_{s,a}$を計算せよ。
2. α helix、β sheet、coilそれぞれについて、2次構造形成をしやすくすると考えられるアミノ酸、2次構造形成をしにくくすると考えられるアミノ酸をそれぞれ2つずつ挙げよ。
3. ここで求めた$P_{s,a}$の値とseq_score.xlsxを用いて、1PGAの42–55残基（EWTYDDATKTFTVT）の2次構造形成傾向スコアを求めよ。この配列はどの2次構造を最も取りやすいか？

溶媒露出表面積

- 分子に接触するように転がしたプローブ球の中心の軌跡→溶媒露出表面
- DSSPのACCカラムの値は溶媒露出表面積に相当
- アミノ酸の大きさによって表面積が異なるので、Ala-X-Ala（またはGly-X-Gly）ペプチドにおけるそのアミノ酸の溶媒露出表面積で割った、溶媒露出度も指標として良く用いられる



Sanner *et al. Biopolymers* **38**, 305 (1993)から引用

溶媒露出度の比較

- 冗長性を除いたPDBのリストのうち、単量体で存在するものを選び、各アミノ酸の溶媒露出度を計算
- アミノ酸の種類ごとに平均溶媒露出度を計算
- 親水性アミノ酸の溶媒露出度が大きく、疎水性アミノ酸の溶媒露出度が小さいことがわかる

	溶媒露出度
Ala	0.21
Arg	0.39
Asn	0.38
Asp	0.43
Cys	0.10
Gln	0.41
Glu	0.44
Gly	0.33
His	0.28
Ile	0.11
Leu	0.13
Lys	0.48
Met	0.17
Phe	0.12
Pro	0.34
Ser	0.34
Thr	0.30
Trp	0.14
Tyr	0.17
Val	0.13

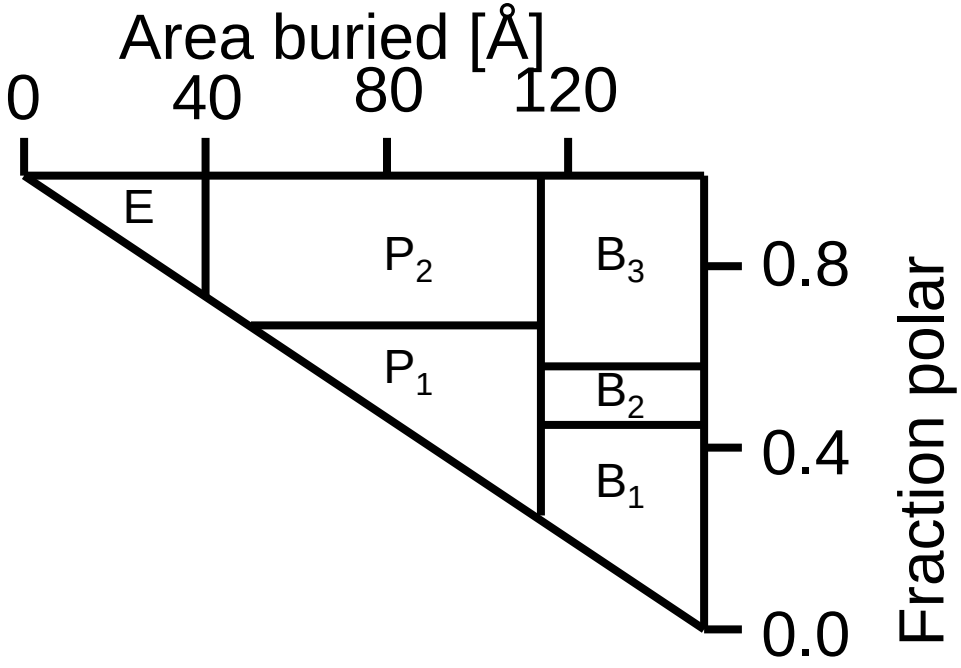
3D profile法

- アミノ酸によって、それが好む「環境」（2次構造や溶媒露出度）が異なる
- 「環境」は立体構造から求めることができる

- 立体構造から、適合するアミノ酸配列を推定する→逆フォールディング問題

Bowieらの方法（1）

- 各残基の溶媒露出表面積を計算し、Gly-X-Glyペプチドの溶媒露出表面積との差をArea buriedとする
- 各残基の表面のうち、極性原子（溶媒を含む）に覆われている割合をFraction polarとする
- これに主鎖構造（α、β、その他）を加えて18種類の環境を定義



Bowie *et al. Science* **253**, 164 (1991).

Bowieらの方法（2）

1. １次配列に沿って各ポジションの「環境」を求める
   - 例：E B_2 Eα P_2α Eα P_2α B_2α Eα P_2α B_1α
2. 「環境」が与えられた時、その環境（env）におけるアミノ酸配列（seq）の出現確率、p(seq|env)を求める
   - １次配列の各ポジションは独立と考え、各ポジションの環境におけるアミノ酸の出現確率の積で近似する
3. アミノ酸配列の平均出現確率との比を求める

$$\frac{p(\mathrm{seq}|\mathrm{env})}{p(\mathrm{seq})} \approx \frac{\prod_{i=1}^{N} p(a_i|\mathrm{env}_i)}{\prod_{i=1}^{N} p(a_i)} = \prod_{i=1}^{N} \frac{p(a_i|\mathrm{env}_i)}{p(a_i)}$$

Bowieらの方法（3）

- 立体構造データベース中のタンパク質について、アミノ酸aの環境envにおける出現確率を計算

$$3D\text{-}1D\ \text{score} = \ln\left[\frac{p(a|\text{env})}{p(a)}\right]$$

logをとることで積を和で計算する

Environment class	W	F	Y	L	I	V	M	A	G	P	C	T	S	Q	N	E	D	H	K	R
$B_1\ \alpha$	1.00	1.32	0.18	1.27	1.17	0.66	1.26	-0.66	-2.53	-1.16	-0.73	-1.29	-2.73	-1.08	-1.93	-1.74	-1.97	-0.34	-1.82	-1.67
$B_1\ \beta$	1.17	0.85	0.07	1.13	1.47	1.09	0.55	-0.79	-2.02	-0.94	-0.22	-1.12	-2.91	-1.67	-1.42	-1.93	-2.56	-1.91	-2.69	-1.16
B_1	1.05	1.45	0.17	1.10	1.11	1.02	0.98	-0.91	-1.92	0.26	-1.22	-1.53	-2.81	-1.17	-2.42	-2.52	-1.76	-1.12	-2.59	-2.16
$B_2\ \alpha$	0.50	0.90	0.85	1.01	0.63	0.68	1.12	-0.69	-1.49	-2.21	-0.10	-1.50	-1.47	-0.23	-0.61	-0.71	-1.62	0.23	-0.78	0.06
$B_2\ \beta$	0.01	1.18	1.06	0.76	1.31	1.06	0.64	-1.55	-2.26	-0.49	-0.87	-2.27	-1.77	-1.22	-2.07	-1.07	-1.41	-0.77	-1.14	-0.20
B_2	1.02	1.05	1.12	0.84	0.81	0.60	0.90	-0.66	-1.66	0.19	-0.05	-0.76	-1.17	-0.76	-0.66	-1.35	-1.28	0.46	-2.34	-0.80
$B_3\ \alpha$	0.92	-0.03	0.58	0.15	0.04	-0.02	0.89	-0.57	-1.86	-0.68	-1.56	-0.57	-0.96	0.22	-0.06	0.08	-0.50	0.73	0.43	0.96
$B_3\ \beta$	0.75	0.81	1.30	0.18	0.54	0.56	-0.57	-0.93	-1.93	-0.34	-0.54	-0.44	-0.74	0.21	-0.24	-0.14	-0.86	0.82	-0.53	0.13
B_3	1.07	0.70	1.13	0.35	-0.17	-0.03	0.23	-0.96	-0.98	-0.13	-1.20	-0.53	-0.54	0.05	0.04	-0.36	-1.05	1.01	0.10	0.66
$P_1\ \alpha$	-1.35	-0.82	-0.59	-0.52	-0.24	0.10	-0.03	0.73	-0.49	-0.25	0.95	0.31	0.34	-0.14	-0.54	-0.17	-0.25	-0.52	-0.21	-0.28
$P_1\ \beta$	0.36	-0.49	0.17	-1.03	0.20	0.46	-0.27	0.64	-0.82	-0.55	1.49	0.93	0.33	-2.27	-1.32	-0.73	-1.07	-0.42	-1.21	-0.77
P_1	-1.26	-1.20	-1.31	-0.62	-0.23	-0.01	-1.19	0.46	-0.24	0.66	1.35	0.56	0.49	-0.63	-0.13	-0.61	0.38	-1.12	-0.74	-1.29
$P_2\ \alpha$	-1.14	-1.43	-0.79	-0.35	-0.54	-0.48	-0.45	0.06	-0.50	-0.26	-0.93	-0.05	-0.18	0.55	-0.05	0.56	0.28	0.06	0.61	0.50
$P_2\ \beta$	-0.79	-0.54	-0.84	-1.30	-0.33	0.13	-0.72	-0.55	-0.98	-1.29	-0.57	0.84	0.59	-0.08	-0.16	0.32	0.19	-0.87	0.59	0.10
P_2	-0.82	-0.86	-0.51	-0.70	-1.09	-0.88	-0.89	-0.15	-0.40	0.44	-0.60	0.06	0.26	0.27	0.50	0.27	0.49	0.13	0.44	0.30
$E\ \alpha$	-1.35	-2.20	-2.10	-1.58	-2.76	-1.10	-0.72	0.46	0.68	0.04	-0.44	-0.17	0.15	0.36	0.28	0.59	0.44	-0.19	0.13	-0.34
$E\ \beta$	0.64	-0.90	0.30	-1.66	-1.47	-1.74	-0.68	0.06	1.46	-0.96	-0.24	0.14	0.65	-0.19	-0.06	-0.16	-0.78	-0.83	-0.52	-0.49
E	-2.14	-1.90	-0.94	-1.19	-1.61	-0.91	-1.67	0.12	1.13	0.20	-0.46	0.12	0.32	-0.03	0.41	0.03	0.22	-0.25	-0.14	-0.32

Bowieらの方法（4）

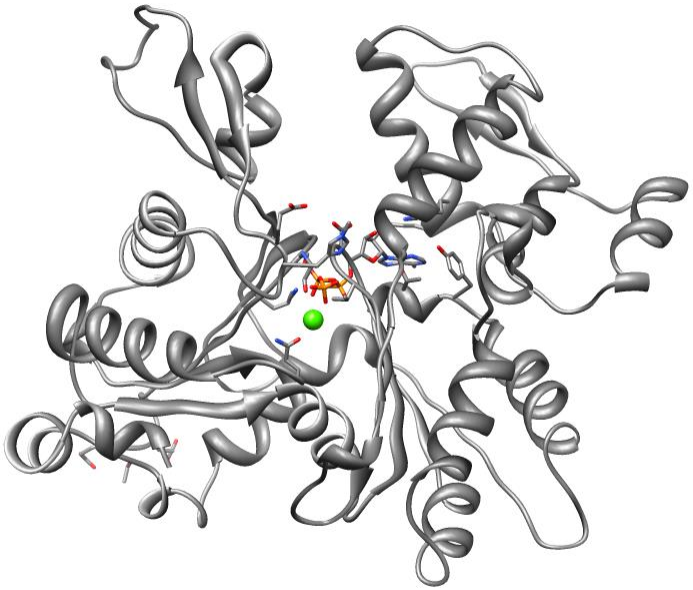
- 3D-1D scoreを用いて、与えられた配列について、その配列の出現確率を計算できる
- 3D-1D scoreを位置特異的スコア行列（3D profileと呼ぶ）とみなすことで、最も大きい出現確率を与えるアラインメントを求めることができる

		Amino acid type												Gap penalty	
Position in fold	Environment class	A	C	D	E	F	G	...	R	S	T	V	W	Y	Opn	Ext
1	E	12	-46	22	3	-190	113	...	-32	32	12	-91	-214	-94	2	0.02
2	B_2	-66	-5	-128	-135	105	-166	...	-80	-117	-76	60	102	112	2	0.02
3	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
4	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
5	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
6	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
7	B_2α	-69	-10	-162	-71	90	-149	...	6	-147	-150	68	50	85	200	200
8	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
9	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
10	B_1α	-66	-73	-197	-174	132	-253	...	-167	-273	-129	66	100	18	200	200
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.

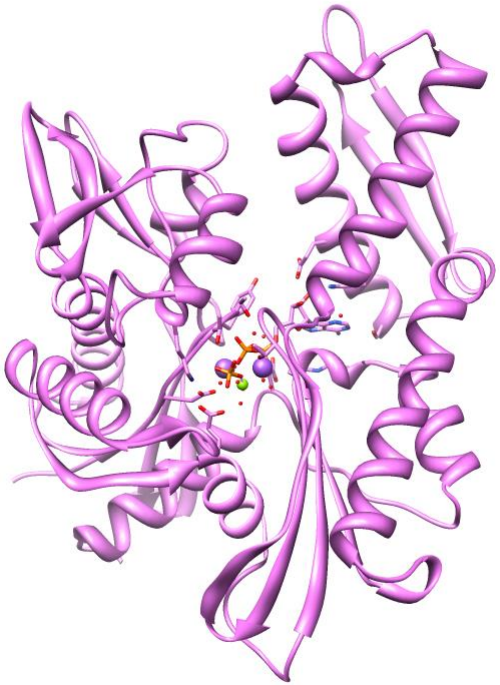
3D profile法の応用（1）

- 立体構造は配列より保存されやすいことを利用した、遠縁のホモログの検出
  - Actinとheat shock proteinは配列相同性はほとんどないが、立体構造は類似している
  - Actinの3D profileを用いると、actin自身の配列に次いでheat shock proteinが高いスコアを示す
- 様々な立体構造について3D profileを計算しておき、与えられた配列に適合する立体構造を探すこともできる→フォールド認識

3D profile法の応用（2）



Actin
PDB ID: 1ATN
（A鎖のみ）



Heat shock protein
PDB ID: 3HSC

3D profile法の応用（3）

- 予測構造の評価
  - 予測に使用した配列と、予測構造の適合度を予測構造の精度の指標とする
  - 予測構造同士の精度の比較ができるほか、各ポジションにおける3D-1D scoreを用いて局所構造の精度を評価できる
  - Verify3D（http://services.mbi.ucla.edu/Verify_3D/）

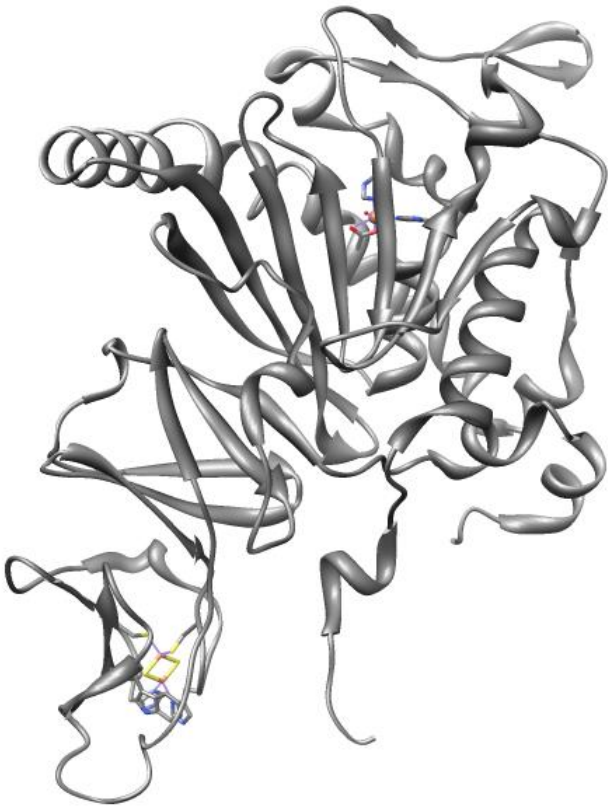
Lüthy *et al. Nature* **356**, 83 (1992).

立体構造比較（1）

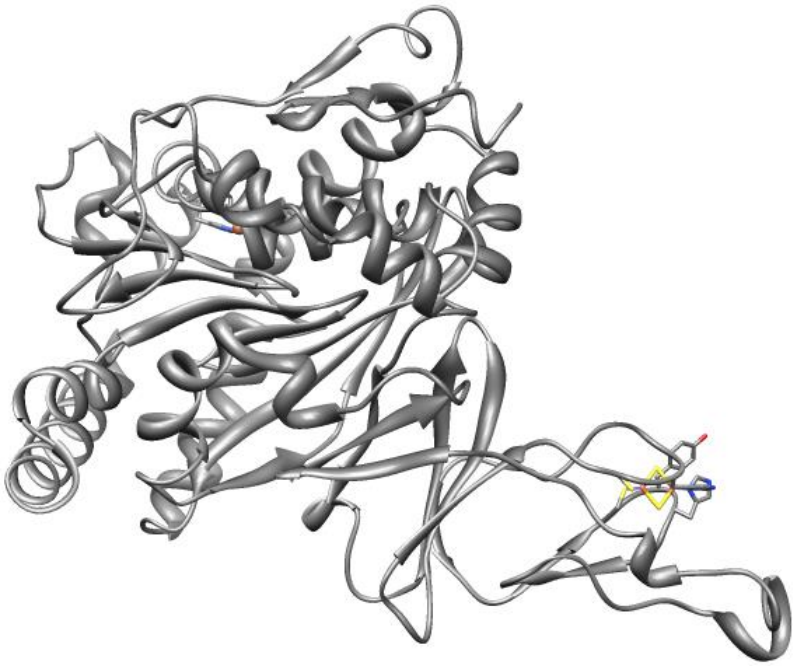
- 配列の場合
  - BLAST等で配列のアラインメント
  - 配列一致度、E-valueを指標に類似度を判断

- 立体構造の場合
  - 主鎖構造がどの程度似ているか
  - 一方の立体構造を並進・回転させ、もう一方とどの程度重なるかを調べる

立体構造比較（2）



PDB ID: 1WW9

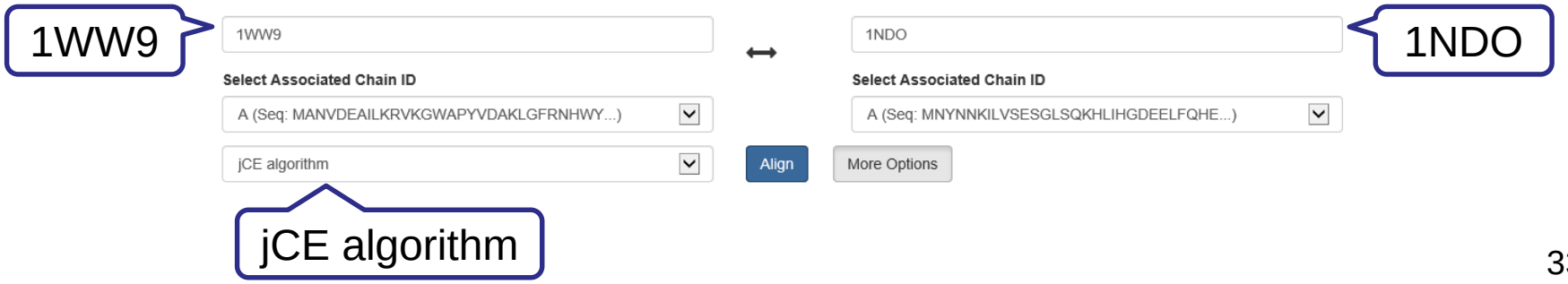


PDB ID: 1NDO
（A鎖のみ）

立体構造比較（3）

1. RCSBのホームページを開く
2. 上部のメニューから「Analyze」→「Sequence & Structure Alignment」を選択

3. 以下のように設定し、「Align」



立体構造比較（4）

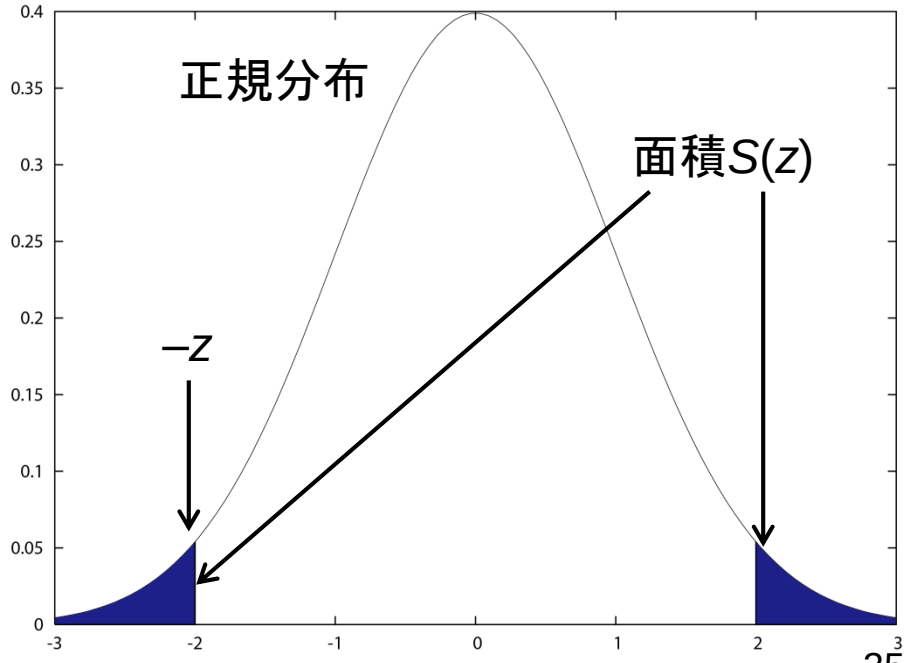
- 類似性の指標
  - Z-Score: 5.86（この値が大きいほど偶然に類似する確率が低い）

>4.5	family level similarity
4.0-4.5	superfamily level similarities, strong function related similarities or strong recurring fold
3.7-4.0	twilight zone where some similarities of biological significance can be seen
<3.7	similarities of low significance, but still some biologically important similarities can be revealed, but interpretation normally requires additional evidence.

  - RMSD: 4.34 Å（対応するCα原子間の平均距離）

Z-scoreの意味

- タンパク質の立体構造をランダムに比較したときに、偶然に同程度以上類似した立体構造のペアが見つかる確率をpとする
- $S(z) = p$となるzをZ-scoreとする
  - $S(3.7)=2.2\times10^{-4}$
  - $S(4.0)=6.3\times10^{-5}$
  - $S(4.5)=6.6\times10^{-6}$



立体構造比較（5）

- Select Comparison Methodで「blast2seq」を選択すると、配列アラインメントが表示される

blast2seq

```
1WW9: 41  GEPKTLKLLGENLLVNRI-DGKLYCLKDRCLHRGVQLSVKVECKTKSTITCWYHAWTYRW 99
          G+  T K+  + ++V+R  DG +    + C HRG  L V VE        C YH W +
1NDO: 52  GDYVTAKMGIDEVIVSRQNDGSIRAFLNVCRHRGKTL-VSVEAGNAKGFVCSYHGWGF-- 108

1WW9: 100 EDGVLCDILTNPTSAQIGRQKL--------KTYPVQEAKGCVFIY-LGDGDPPPL----- 145
            G   ++ + P    +  + L        +   V+   G  FIY   D + PPL
1NDO: 109 --GSNGELQSVPFEKDLYGESLNKKCLGLKEVARVESFHG--FIYGCFDQEAPPLMDYLG 164

1WW9: 146 --ARDTPPNFL-DDDMEILGK--NQIIKSNWRLAVEN 177
            A    P F     +E++G     +IK+NW+   EN
1NDO: 165 DAAWYLEPMFKHSGGLELVGPPGKVVIKANWKAPAEN 201
```

CE（一部）

```
1WW9:  13 KGW----------APYV---DAKLG----FRNHWYPVMFSKEINE-GEPKTLKLLGENLLVNRID-GKLY
          |||          ||||   |||||    |||||||||||||||| |||||||||||||||||| ||||
1NDO:   6 KILVSESGLSQKHLIHGDEELFQHELKTIFARNWLFLTHDSLIPAPGDYVTAKMGIDEVIVSRQNDGSIR

1WW9:  64 CLKDRCLHRGVQLSVK---VECKTKSTITCWYHAWTYRWEDGVLCDILTNP----TSAQIGRQKLKTY-P
          ||||||||||||||||   |||    |||||||||||||| ||||||||||    |||||||||||| |
1NDO:  76 AFLNVCRHRGKTLVSVEAGNAK----GFVCSYHGWGFGSN-GELQSVPFEKDLYGESLNKKCLGLKEVAR

1WW9: 126 VQEAKGCVFIYLGDGDPPPLARDT--PPNFLD---DDMEILGK---NQIIKS--NWRLAVENG-FDPSHI
          |||||||||||| ||||||||||  ||||||   ||||||||   ||||||  ||||||||| |||| |
1NDO: 141 VESFHGFIYGC-FDQEAPPLMDYLGDAAWYLEPMFKHSGGLELVGPPGKVVIKANWKAPAENFVGDAYHV
```

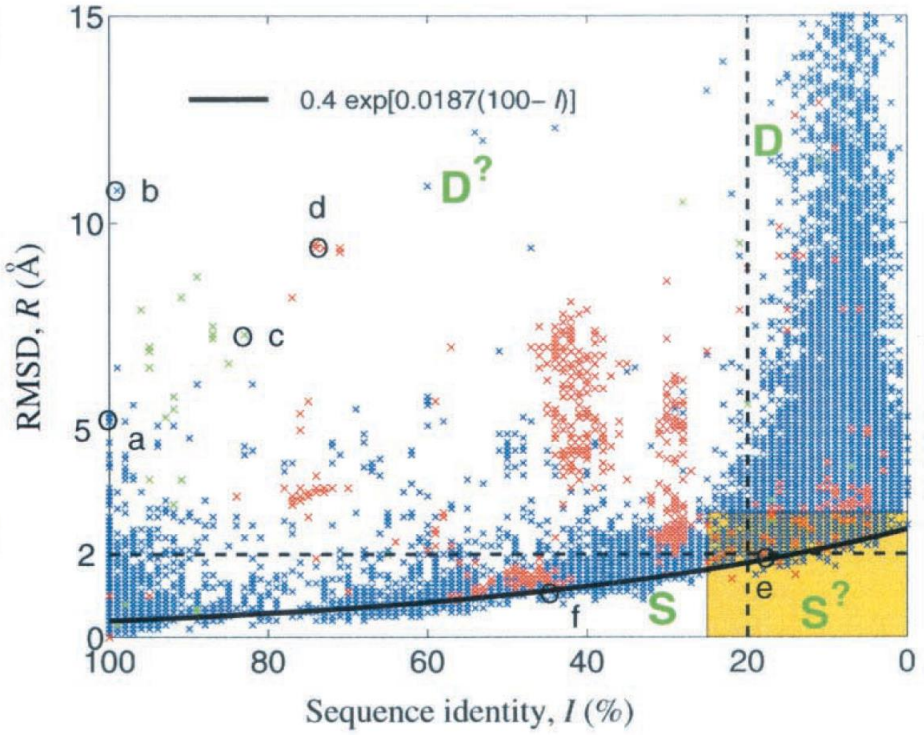
立体構造比較（6）

- 配列比較（blast2seq）
  - E-value: 3×10^{-5}、Identity: 26%
- 立体構造比較（CE）
  - Z-score: 5.86、RMSD: 4.34
- 機能
  - 1WW9: carbazole 1,9a-dioxygenase
  - 1NDO: naphthalene dioxygenase

- 進化的類縁関係が明らか→family

配列類似性と立体構造類似性

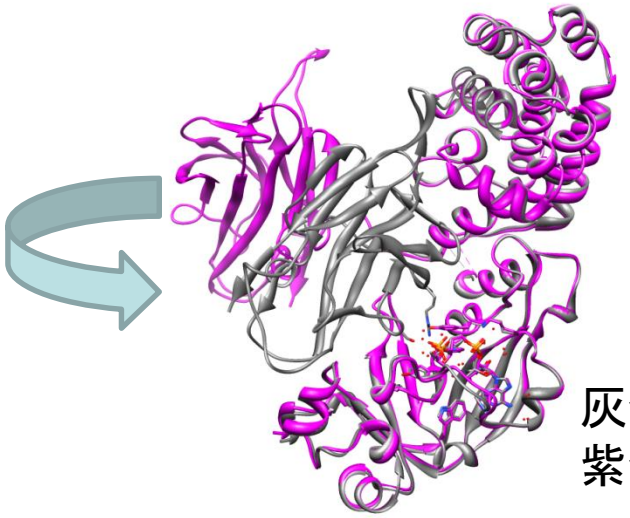
- 一般に、配列一致度が高くなると、立体構造類似性も高くなる
- 配列一致度が20%を切ると、立体構造が大きく異なるケースが増えてくる
- 配列が似ていなくても立体構造が類似しているケースがあることにも注意



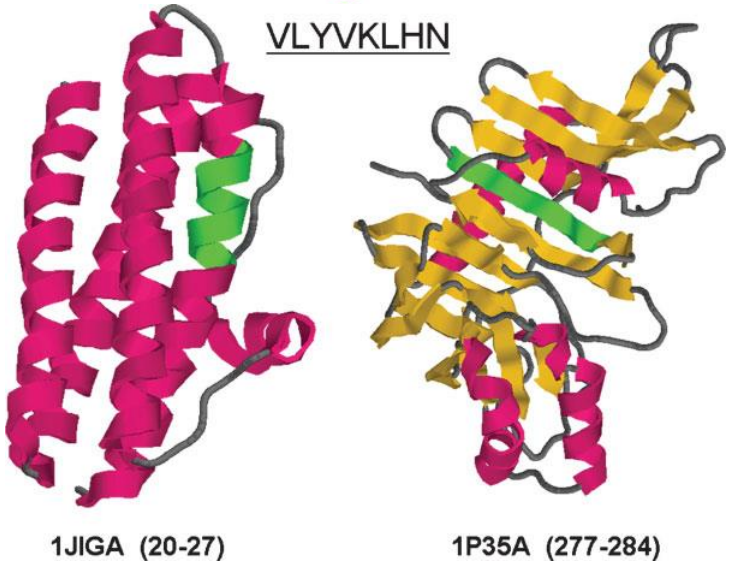
Gan *et al. Biophys J.* **83**, 2781.

配列が似ていても立体構造が異なる例

- ドメインがフレキシブルなリンカで繋がっている場合
  - ドメイン運動によってRMSDが大きくなる
- カメレオン配列を持つ場合
  - 同じ部分配列が異なるタンパク質中で異なる2次構造をとる
  - 最長でも8残基（右図）*
  - 事例は多くない

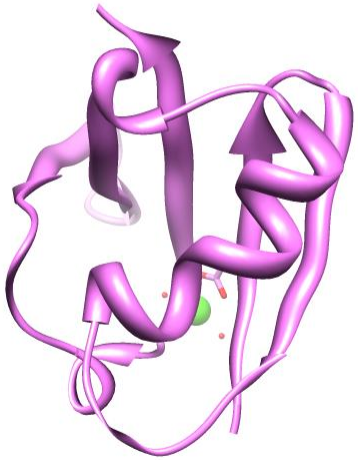


灰色：1MDT
紫色：1DDT



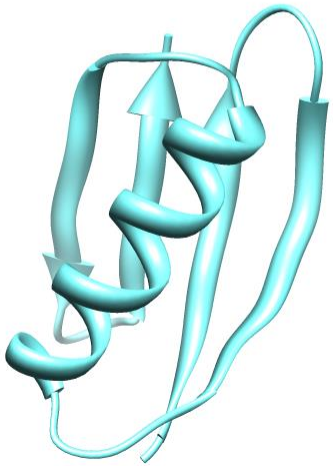
*Guo *et al. Proteins* **67**, 548 (2007).

立体構造類似性と機能・進化（１）



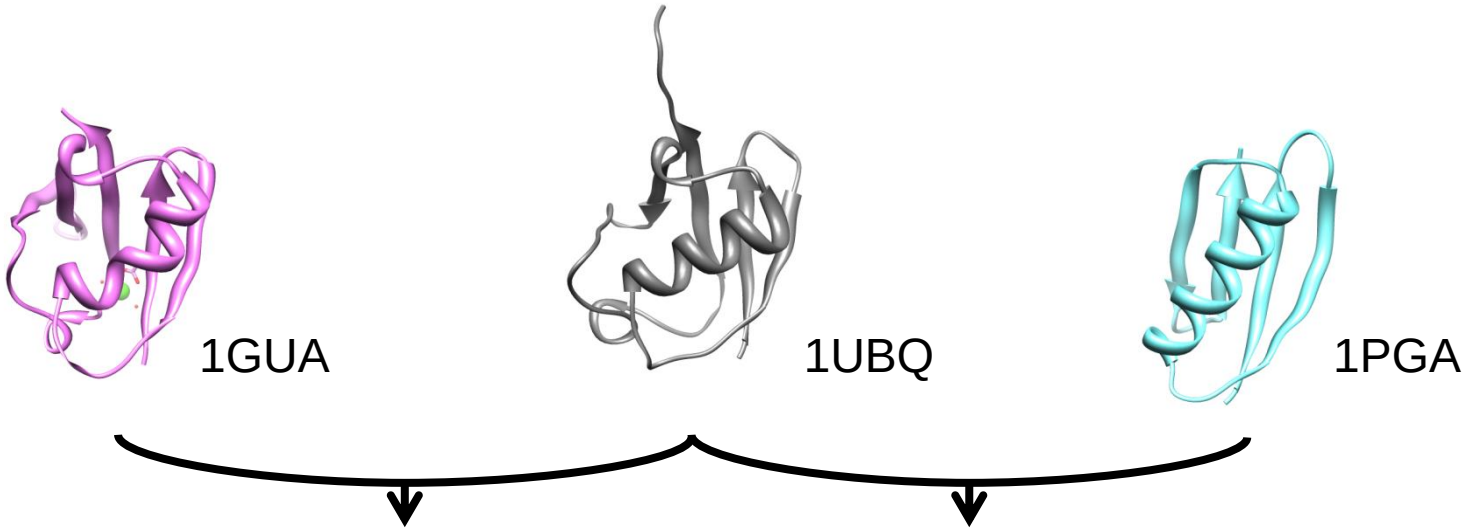
Ras-binding
domain of c-Raf-1
PDB ID: 1GUA
（B鎖のみ）

Ubiquitin
PDB ID: 1UBQ



Immunoglobulin-binding
domain of protein G
PDB ID: 1PGA

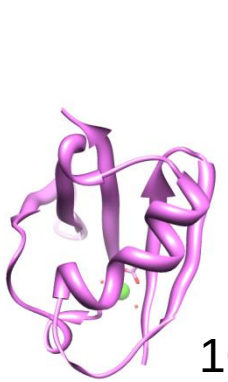
立体構造類似性と機能・進化（2）



- 配列比較
  - No hits
- 立体構造比較
  - Z-score: 4.25
  - RMSD: 2.20

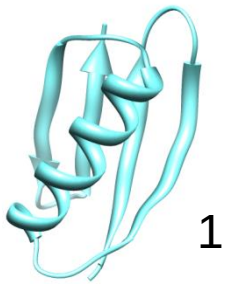
- 配列比較
  - E-value: 0.010
- 立体構造比較
  - Z-score: 3.29
  - RMSD: 3.13

立体構造類似性と機能・進化（3）



1GUA

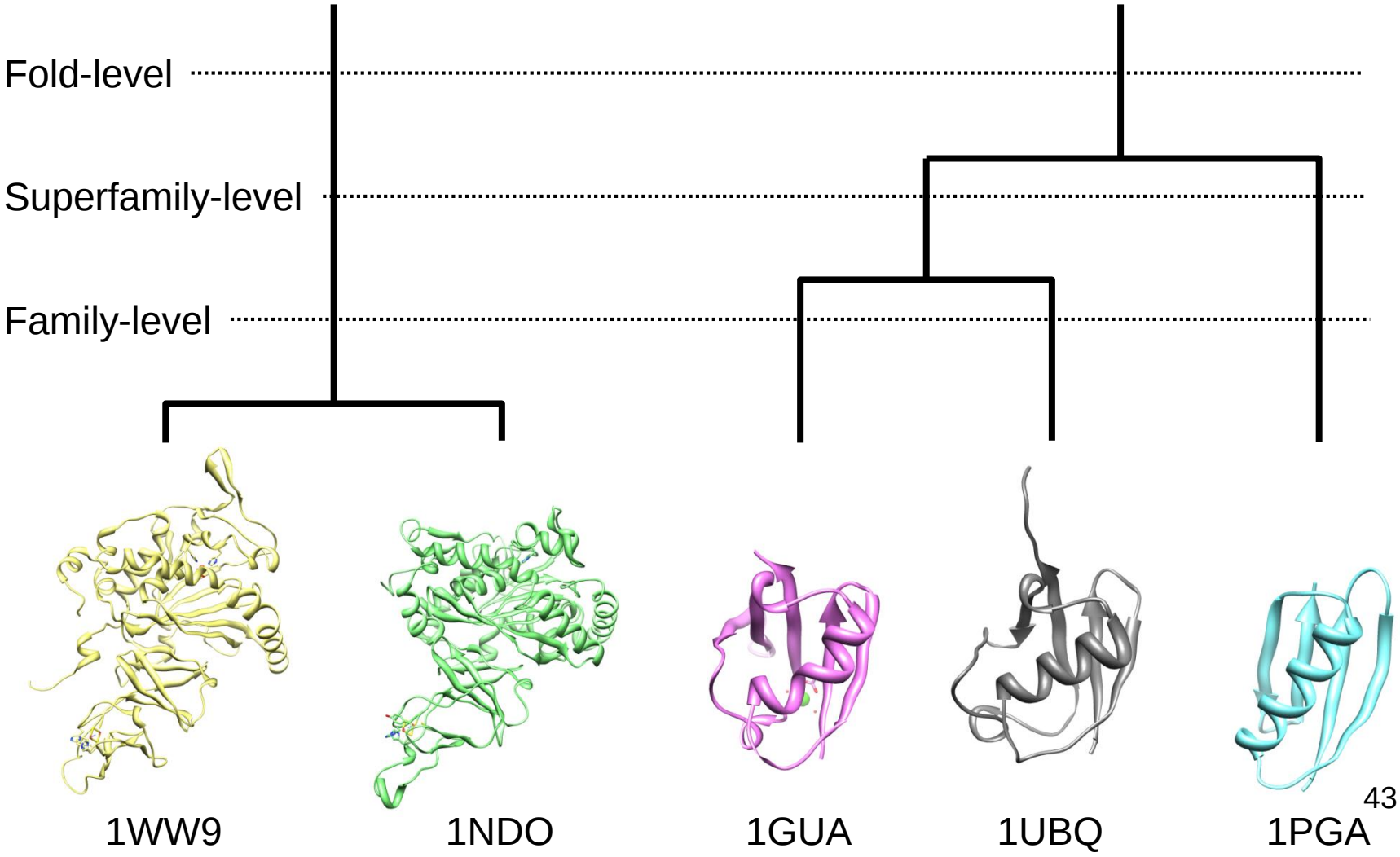
1UBQ



1PGA

- 立体構造の類似度が高い
  - 進化的類縁関係が推定される
    →Superfamily
- 配列・立体構造の類似度が低い
  - 進化的類縁関係無し
  - 単に折りたたみ様式（fold）が似ているだけ

立体構造類似性と機能・進化（４）



立体構造分類データベース

- 立体構造を立体構造類似性に基づいて階層的に分類したデータベース
  - SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/)
    主に人手による分類
    Class → Fold → Superfamily → Family
  - CATH (http://www.cathdb.info/)
    主に計算機による分類
    Class → Architecture → Topology → Homologous Superfamily

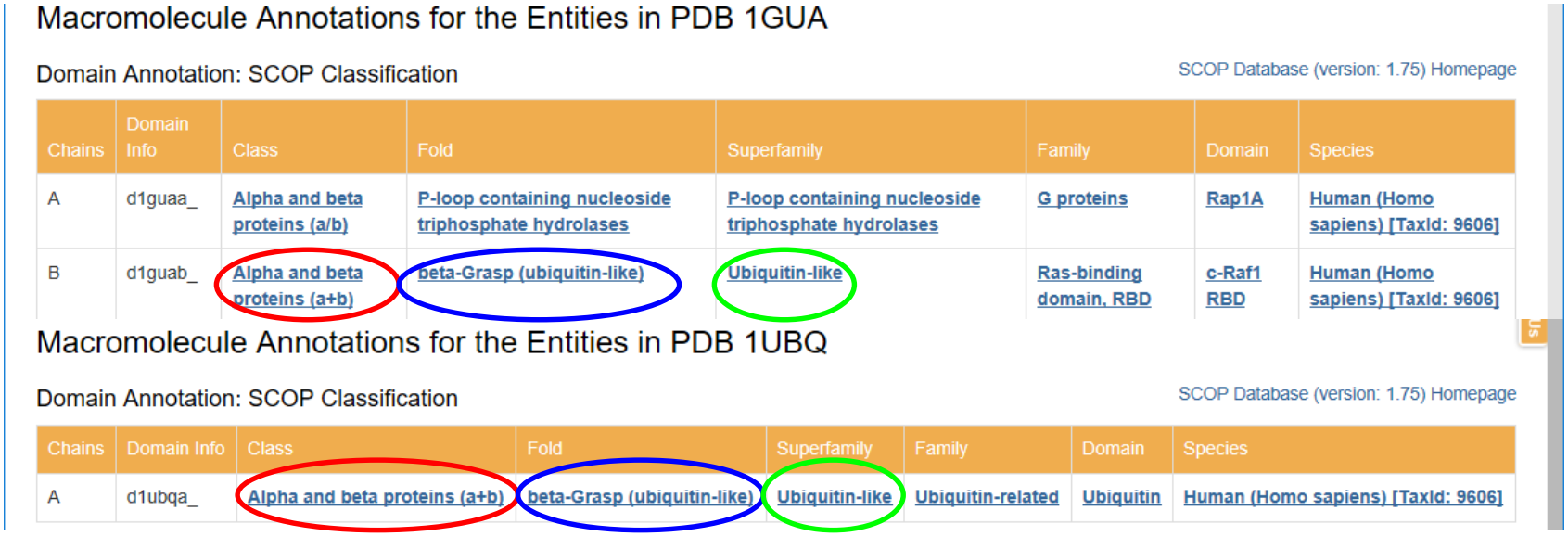
SCOPにおける階層（1）

- Family：配列一致度が概ね30%以上あり、進化的な類縁関係が明らかなもの
- Superfamily：配列一致度は低いが、立体構造と機能の類似性から、進化的な類縁関係が推定されるもの
- Fold：主な2次構造の配置とトポロジーが同じだが、進化的な類縁関係はないもの
  →構造類似性は特定のパッキングやトポロジーを好むタンパク質の物理化学的な性質に起因

http://scop.mrc-lmb.cam.ac.uk/scop/intro.html

SCOPにおける階層（2）

- RCSBのAnnotationsタブを開く



Macromolecule Annotations for the Entities in PDB 1GUA

Domain Annotation: SCOP Classification

SCOP Database (version: 1.75) Homepage

Chains	Domain Info	Class	Fold	Superfamily	Family	Domain	Species
A	d1guaa_	Alpha and beta proteins (a/b)	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases	G proteins	Rap1A	Human (Homo sapiens) [TaxId: 9606]
B	d1guab_	Alpha and beta proteins (a+b)	beta-Grasp (ubiquitin-like)	Ubiquitin-like	Ras-binding domain, RBD	c-Raf1 RBD	Human (Homo sapiens) [TaxId: 9606]

Macromolecule Annotations for the Entities in PDB 1UBQ

Domain Annotation: SCOP Classification

SCOP Database (version: 1.75) Homepage

Chains	Domain Info	Class	Fold	Superfamily	Family	Domain	Species
A	d1ubqa_	Alpha and beta proteins (a+b)	beta-Grasp (ubiquitin-like)	Ubiquitin-like	Ubiquitin-related	Ubiquitin	Human (Homo sapiens) [TaxId: 9606]

- 1GUAのB鎖と1UBQではSuperfamilyレベルまで一致

実習課題2

1. 1GUAのB鎖と1PGAのA鎖の配列比較（blast2seq）と立体構造比較（CE）を行い、E-valueとZ-score、RMSDを報告せよ
2. 1PGAが属する、SCOPのClass、Fold、Superfamily、Familyを報告せよ
3. これを、1UBQ、1GUAのc-Raf-1 RBDが属するClass、Fold、Superfamily、Familyと比較し、どの階層まで一致するか述べよ

課題の提出

- 実習課題1の項目1は、エクセルファイルをメールに添付すること
- 実習課題1の項目2、3と、実習課題2はメールの本文に記載すること
- メールは寺田宛 tterada@iu.a.u-tokyo.ac.jp に送ること
- その際、件名は「構造実習」とし、本文に氏名と学生証番号を必ず明記すること