

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人材養成プログラム
バイオインフォマティクスリテラシーI
平成19年5月28日(月)、31日(木) @農学部2号館化学第一講義室

立体構造予測 I [Web版]

フォールディング問題、構造分類、構造比較、相同性検索など

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人材養成ユニット
特任助教

古田 忠臣

講義の予定

■ 5月28日(月)、31日(木)

- 構造データベース:PDB
- 構造分類データベース:SCOP、CATH

構造類似性 □ 構造比較サーバー:CE、DALI/FSSP、VAST

配列類似性 □ 相同性検索:BLAST、PSI-BLAST、FASTA、CLUSTALW
1D検索

■ 6月4日(月)、7日(木)

2D予測

- 二次構造予測:PSIPRED、PHDsec、PREDETOR、NPS@

3D予測

- 立体構造予測 ……参考: CASP
 - 比較モデリング法
 - ホモロジーモデリング:MODELLER、SWISS-MODEL
 - フォールド認識法:meta server (3D-Jury)
 - de novo / ab initio予測法: Robettaなど

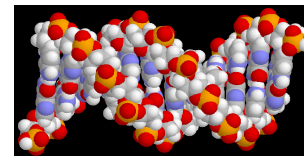
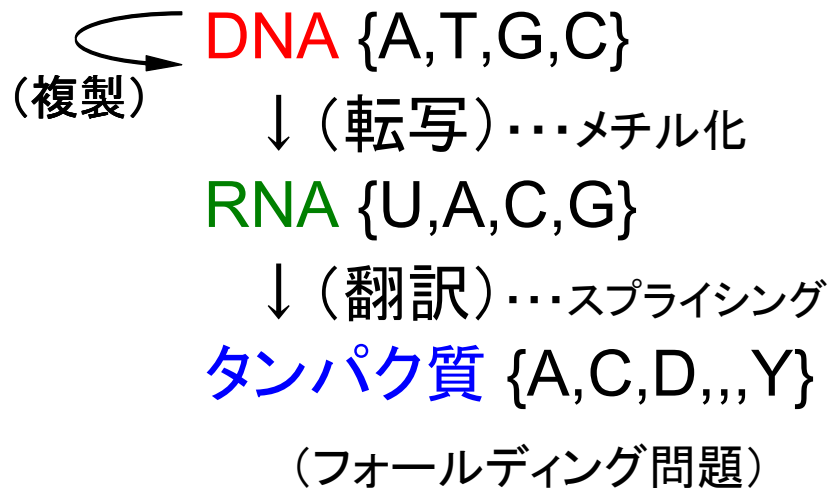
分子生物学のセントラルドグマ

Webで顔写真を
探して下さい。

■ F.H.C.Crick 1958

F. Crick

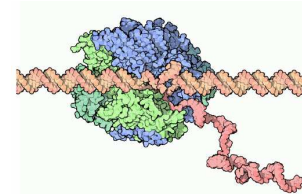
遺伝的情報は以下の様に一方向的に伝達される



DNA

Watson & Crick, 1953

(ヒトゲノム: 約30億塩基対、染色体: 23対)
(32億5,400万bp)



(RNAポリメラーゼ)



Myoglobin

PDB:1MBN

Kendrew, 1960

(ヒト遺伝子: 約3万(26,808))
(リボソーム、tRNA)

(レトロウィルスの逆転写酵素: RNA→DNA)

F.H.C. Crick, *Symp. Soc. Exp. Biol.* **12**, 138-163 (1958), "On protein synthesis"

[URL] <http://www.lif.kyoto-u.ac.jp/genomemap/>

現在300種以上のゲノムが解読されている

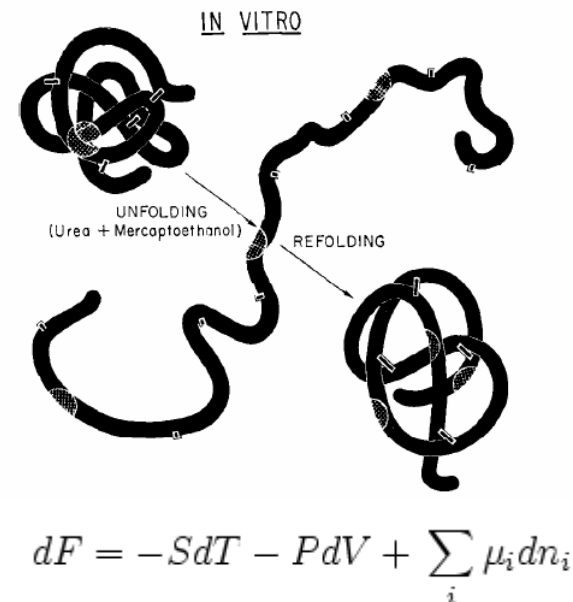
Anfinsenのドグマ 1973

- タンパク質の天然構造は熱力学的に最も安定であり、「一次構造が決まれば立体構造も決まる」

- 変性しても、元の生理的条件下に戻すと再び折り畳まる
- 現在では、巨大なタンパク質やミスfoldしたタンパク質は分子シャペロンの助けを借りて折り畳まることが知られている
 ……βアミロイド形成 → 病気

Webで顔写真を探して下さい。

C. Anfinsen



C.B. Anfinsen, *Science* **181**, 223-230 (1973), “Principles that govern the folding of protein chains”

Levinthalのパラドックス 1969

Webで顔写真を
探して下さい。

C. Levinthal

- ランダム探索では、天然構造への折り畳むまでに天文学的時間が掛かる
 - 各アミノ酸が3つのconformationsを持つとして、150残基のタンパク質の場合、可能なconformationsは $3^{150} \sim 10^{71}$ ある。一回のconformation変化に 10^{-13} s掛かるとして、全探索には $10^{71} \times 10^{-13}\text{s} = 10^{58}\text{s} \sim 10^{50}$ 年掛かる。
 - しかし、実際のタンパク質は数ms程度で折れたたまる... 10^8 回程度の探索しかしていない(パラドックス)
 - 特定のFolding pathwaysがあるのではないか。

エネルギー・ランドスケープ理論

Webで顔写真を
探して下さい。

N. Go P.G. Wolynes

■ タンパク質のエネルギー地形はファネル状の形をとる

□ GoのConsistency Principle 1984

- タンパク質は、天然状態で様々な相互作用が最適になっている(進化の所産)。

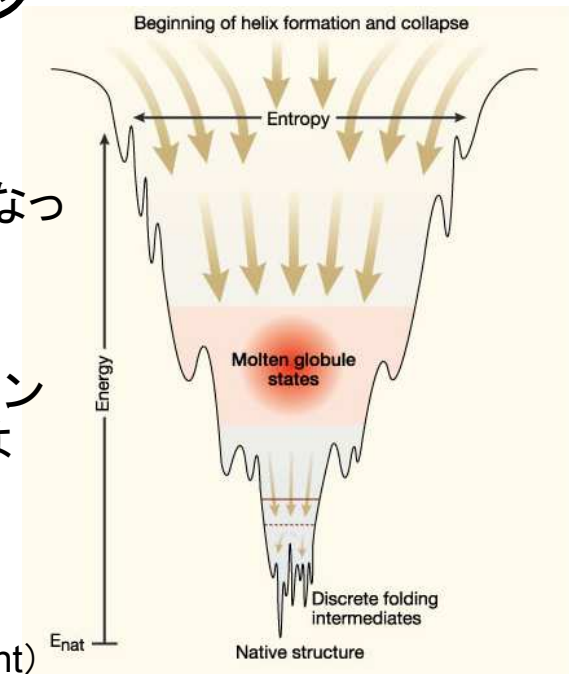
□ Wolynesらのフラストレーション最小原理 1987

- タンパク質は、天然状態でエネルギーフラストレーションを最小にする様に設計されている。エネルギー地形はファネル状の形をとる。

(フォールディング問題: モルテングロビュール状態(K.Kuwajimaら)

遷移状態(H.A.Kramers)・・・ ϕ 値解析(A.Fersht)

2状態・3状態(H.A.Sheragaら))



N. Go, *Adv. Biophys.* **18**, 149-164 (1984), "The consistency principle in protein structure and pathways of folding"
 J.D. Bryngelson, and P.G. Wolynes, *PNAS* **84**, 7524-7528 (1987), "Spin glasses and the statistical mechanics of protein folding"

分子動力学の講義 → 分子モデリングと分子シミュレーション(寺田先生)
 量子化学の講義 → 量子化学入門と分子軌道法(岩岡先生)

分子シミュレーションの簡単な歴史 (手法の改良およびQMIは除く)

1953 最初のMonte Carlo(MC)シミュレーション	Metropolisら
1957 最初のMolecular Dynamics(MD)シミュレーション	Alder & Wainwright
1969 液体(水)のMC	Barker & Watts
1971 液体(水)のMD	Rahman & Stillinger
1973 ヌクレオチドの真空中でのエネルギー最小化(GpC)	Stellmanら
(以下、タンパク質)	
1971 タンパク質のエネルギー最小化(Lysozyme)	Levitt & Lifson
1977 最初のタンパク質の真空中でのMD(BPTI, 10ps)	McCammion, Gelin & Karplus
1981 AMBER	Kollman, Caseら
1982 最初のタンパク質の溶液中でのMD(BPTI, 25ps)	van Gunsteren & Karplus
1983 CHARMM	Karplus, Brooksら
1990 GROMOS/GROMACS	van Gunsteren, Berendsenら
1996 NAMD	Schultenら
1997 世界最長時間! ?のフォールディングのMD (villin headpiece(36aa, 4,000atoms), 1 μs (2 months cpu time))	Duan & Kollman
2000 Folding@Home like SETI !?	Pandeら
2001 K ⁺ チャンネルのMD (40,000 atoms, 38ns)	BerneÅche & Roux
2001 アクアポリンのMD (101,449 atoms, 10ns)	de Groot & Grubmüller
2002 F1-ATP synthaseのMD (183,674 atoms, 7ns (~400 months cpu time))	Böckmann & Grubmüller

スペースの都合上、
文献情報がないですが
必要でしたらご連絡下さい。

小タンパク質以外のフォールディングのMDは困難

タンパク質立体構造データベース: PDB – Protein Data Bank

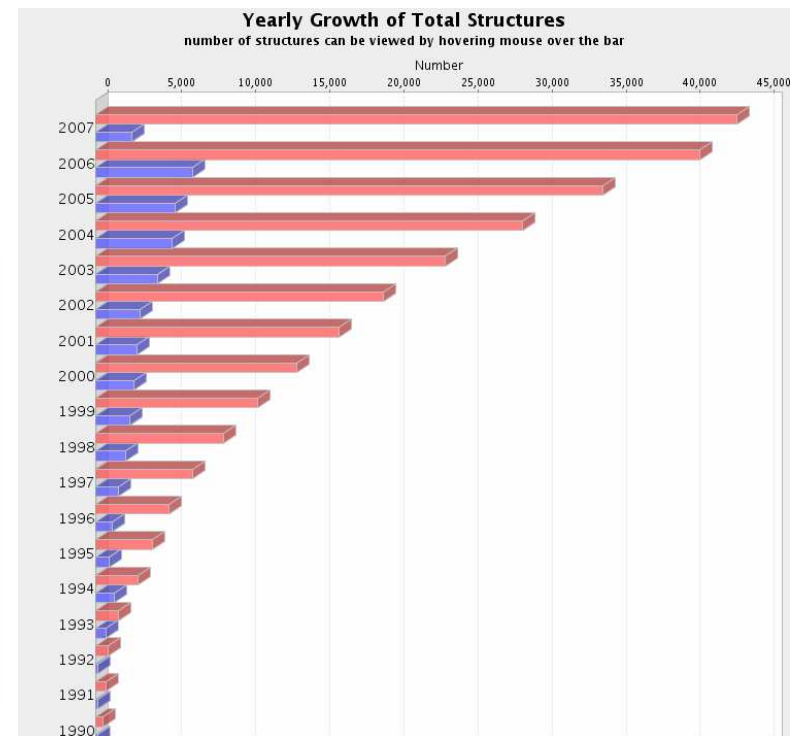
wwPDB [URL] <http://www.wwpdb.org> RCSB PDB [URL] <http://www.rcsb.org/pdb/>
MSD EBI [URL] <http://www.ebi.ac.uk/msd/>
PDBj [URL] <http://www.pdbj.org/>

- 現在 (May/08/2007)、
43,339構造が登録されている

PDB ID: 4文字の英数字 ex) 1AB1

PDB Current Holdings Breakdown

		Molecule Type				Total
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	
Exp. Method	X-ray	34259	964	1581	28	36832
	NMR	5375	760	129	7	6271
	Electron Microscopy	101	10	38	0	149
	Other	80	4	3	0	87
	Total	39815	1738	1751	35	43339



Chothiaのフォールド数の見積もり

Webで顔写真を
探して下さい。

C. Chothia

■ Chothia 1992

- タンパク質は約1000ファミリーしかない...今で言うフォールド数
 - 新規配列の1/3が既知配列と相同的
 - 既知配列の1/4が既知120ファミリーに属する
 - 統計的偏りがなければ、 $120 \times 4 \times 3 \sim 1,500$ ファミリー

- 構造(フォールド)は配列よりもよく保存されている
 - 新規フォールドの割合は収束に向かっていない、、、CASP6 (vs. CASP5)
CASP7では! ?

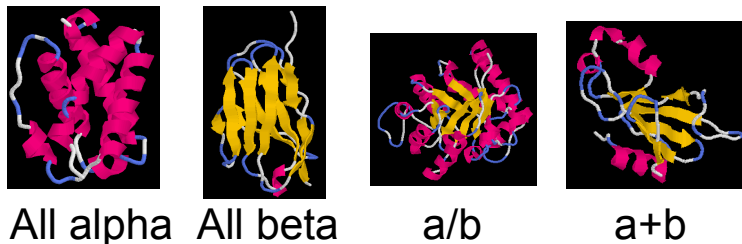
タンパク質立体構造分類データベース: SCOP – Structural Classification of Proteins

[URL] <http://scop.mrc-lmb.cam.ac.uk/scop/>

- 2005年時点での27,599PDB
が75,930ドメインに分割され、

現在、約1,000(971)
のフォールドが登録
されている(1.71)

Murzinが中心となり、
人の手・目！？で分類されて
いる



→ top of the hierarchy

スーパーファミリー: 機能・構造的特徴から恐らく共通の進化的起源
例) アクチン、ヒートショックタンパクのATPase、ヘキサキナーゼ

階層:

クラス、フォールド、スーパーファミリー、ファミリー

ex) scs: a. 1. 1. 1

All alpha proteins. Globin-like. Globin-like. Truncated hemoglobin

Scop Classification Statistics

SCOP: Structural Classification of Proteins. 1.71 release
27599 PDB Entries (18 Jan 2005). 75930 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	226	392	645
All beta proteins	149	300	594
Alpha and beta proteins (a/b)	134	221	661
Alpha and beta proteins (a+b)	286	424	753
Multi-domain proteins	48	48	64
Membrane and cell surface proteins	49	90	101
Small proteins	79	114	186
Total	<u>971</u>	1589	3004

A.G. Murzin, et al., *J. Mol. Biol.* **247**, 536-540 (1995),

“SCOP: a structural classification of proteins database for the investigation of sequences and structures”

タンパク質立体構造分類データベース:

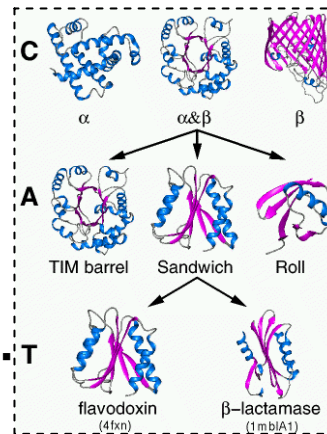
CATH – Class, Architecture, Topology, Homologous superfamily

[URL] <http://cathwww.biochem.ucl.ac.uk/latest/>

→ Browse or search the classification

- 現在、1,084 topologies (folds) が登録されている (v3.1.0)

- かなり自動的に分類されるが、最後は人手



CATH v3.1.0	
Version	3.1.0
Date	19-01-2007
Number of Domains	93885
Number of Chains	63453
Number of PDBs	30028

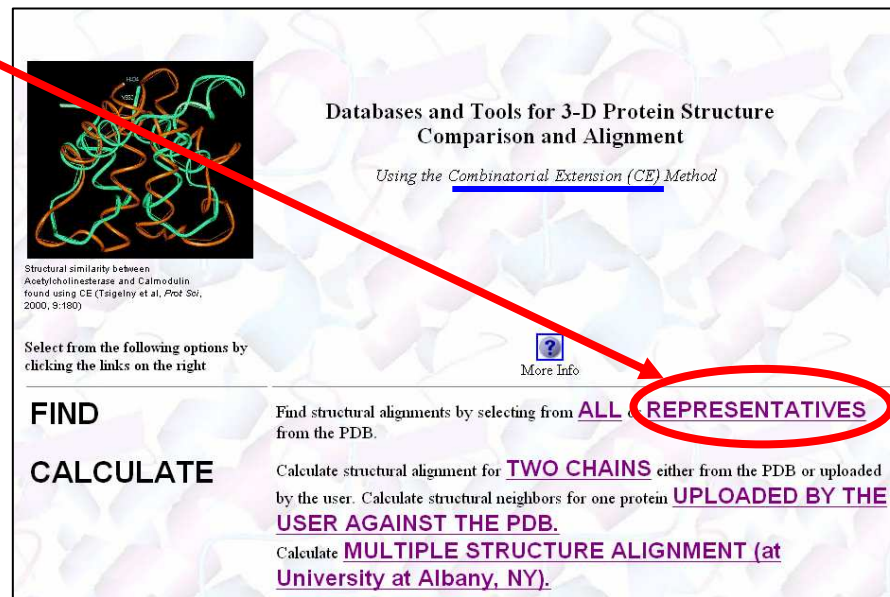
	A	T	H	S
Mainly Alpha	5	305	652	1850
Mainly Beta	20	191	415	1860
Alpha Beta	14	496	922	3922
Few Secondary Structures	1	92	102	162
Total	40	1084	2091	7794

階層 C A T H S
 クラス、アーキテクチャー、トポロジー、ホモログスーパーファミリー、シーケンスファミリー

ex) CATH code: 1. 10. 8. 10. 1
 Mainly Alpha
 Orthogonal Bundle
 Helicase,,,
 DNA helicase RuvA subunit,,
 DNA helicase Ruv subunit,,

タンパク質立体構造比較サーバー1: CE

- CEサーバーで構造比較したアラインメントを得て、DS1.7で実際の構造を見てみましょう
- CEのホームページ[URL] <http://cl.sdsc.edu/> を開く
- 「**REPRESENTATIVES**」をクリック



Databases and Tools for 3-D Protein Structure Comparison and Alignment

Using the Combinatorial Extension (CE) Method

Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al. *Prot Sci*, 2000, 9:180)

Select from the following options by clicking the links on the right

[?](#)
More info

FIND	Find structural alignments by selecting from ALL REPRESENTATIVES from the PDB.
CALCULATE	Calculate structural alignment for TWO CHAINS either from the PDB or uploaded by the user. Calculate structural neighbors for one protein UPLOADED BY THE USER AGAINST THE PDB .
	Calculate MULTIPLE STRUCTURE ALIGNMENT (at University at Albany, NY).

I.N. Shindyalov, P.E. Bourne, *Protein Engineering* **11**, 739-747 (1998), "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path"

- Specify Protein Chainに「1BFR:D」を入力
- 「Search Database」をクリック

CE FIND REPRESENTATIVES Find structural alignments by selecting from a representative part of the PDB.

Specify a polypeptide chain to use as a search against a database of [representative structures](#) and optionally specify other than the default selection criteria. Hit the **Search Database** button to initiate the search.

Specify Protein Chain:

Specify Selection Criteria:

Z-Score: ? Length Difference: ? Sequence Identity: ? Sort by: ?

RMSD: ? Gaps: ? Select:

- Seq.(%)が13.7の「1EUM:B」を**チェック**
(1NF6のチェックをはずす)
- 配列の一致度が13.7%なのに、RMSD=2.1 Å

Structure Neighbors for Chain 1BFR:D (Size=158)

MOL_ID: 1; MOLECULE: BACTERIOFERRITIN; CHAIN: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X; SYNONYM: CYTOCHROME B

Selected 288 chains

Criteria used: Z-Score>4.0 RMSD<5.0Å

Sorted by: Z-Score

Use the checkboxes to select one or more chains to align with 1BFR:D. In the ID column select the ID to get further information on that structure. In the ID column select "Neighbors" to display the structure neighbors matching that ID.

?	ID	Z-Score	RMSD (Å)	Seq. (%)	Aligned/ Size	Gap	Exp.	Name
	1BFR:D						X-Ray	MOL_ID: 1; MOLECULE: BACTERIOFERRITIN; CHAIN: A, B, C, D, E, F, G, H, I, J, K, L
	<input type="checkbox"/> 1NF6:J <small>Neighbors</small>	6.7	1.2	24.0	150 / 179	1	X-Ray	MOL_ID: 1; MOLECULE: BACTERIOFERRITIN; CHAIN: A, B, C, D, E, F, G, H, I, J, K, L
	<input type="checkbox"/> 1FHX: <small>Neighbors</small>	6.6	1.9	22.7	145 / 183	5	X-Ray	FERRITIN (H-CHAIN) MUTANT (LYS 86 REPLACED BY GLN) (K86Q)
	<input checked="" type="checkbox"/> 1EUM:B <small>Neighbors</small>	6.6	2.1	13.7	153 / 165	2	X-Ray	MOL_ID: 1; MOLECULE: FERRITIN 1; CHAIN: A, B, C, D, E, F; ENGINEERED: YES

- どれほど似ているか構造を見るために、ページ上の「**GET ALIGNMENT**」をクリックしアラインメントをしましょう

- 構造比較したアラインメントが得られます。

Structure Alignment - 1BFR:D Neighbors

Sequence alignment based on structure alignment between 1BFR:D and its neighbors. Light color indicates not-aligned residues in structural neighbors. Position numbers according to sequence (starting from 1) and according to PDB are given as SSSS:PPPP, SSSS - sequence, PPPP - PDB.

```

1BFR:D 3/4  GDIKVINYLKLLGNELVAIQYFLHARMFNWGLKRLNDVEYHESIDEMKHADRYIERI
1EUM:B 2/3  LKPEMIEKLNENLNLELYSSLYQQMSAWCSYHTFEGAAAFLRHAQEEMTHMQLFDYL

1BFR:D 63/64 LFLGLEPNIQLDGLKLNIGE-DVEEMLRSDLALELDGAKNLRERAIQYADSVHDVYSRDMMI
1EUM:B 62/63  TDIGNLFRINIVESFFAEY-SLEDLFQETVYRHEQLTQKINELARAAHNIQDYFIPFLQ

1BFR:D 122/123 EILRDEGHIDWLEFELDIQKMLQ-NYLQAQIR
1EUM:B 122/123 WYVSEQEHEEELFKSIIDKLSLAKS-EGLYFIDK
  
```

1EUM:B	2.1	13.7
	153	6.6
1BFR:D		

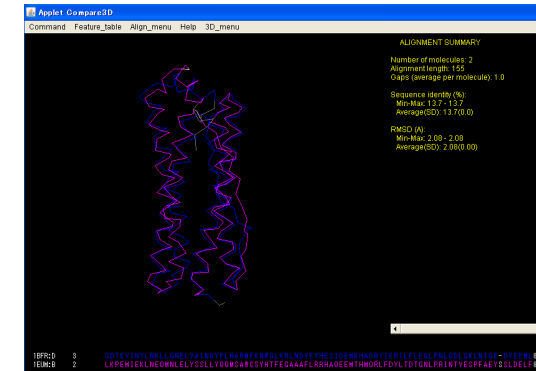
Each cell in distance matrix provides:

RMDS(A)	Sequence identity(%)
Length of alignment	Z-score

View Results

- [Download alignment as a PDB file](#)
- [Quick view of structure alignment \(using Rasmol\)](#)
- [Detailed analysis of alignment \(using Compare3D Java applet\)](#)
- [Press to Start Compare3D](#)

Notes: (i) On some platforms (InternetExplorer7) Compare3D starts from the second click (ii) Compare3D may not work with some versions of InternetExplorer or across the firewall

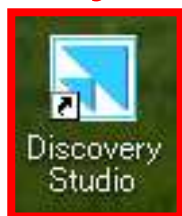


Press to Start Compare 3Dをクリックすると、Java Appletで比較した構造を見ることも出来ます。

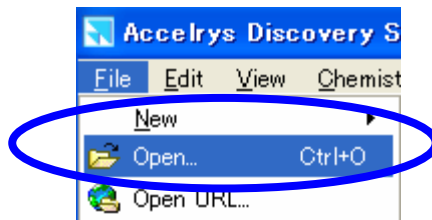
- 構造をダウンロードする

- 「**Download alignment as a PDB file**」を右クリック
- →「**対象をファイルに保存**」
- デスクトップに「リテラシー」などのフォルダを作成し、ファイル名「**1BFR-1EUM.pdb**」、ファイルの種類「**すべてのファイル**」で保存

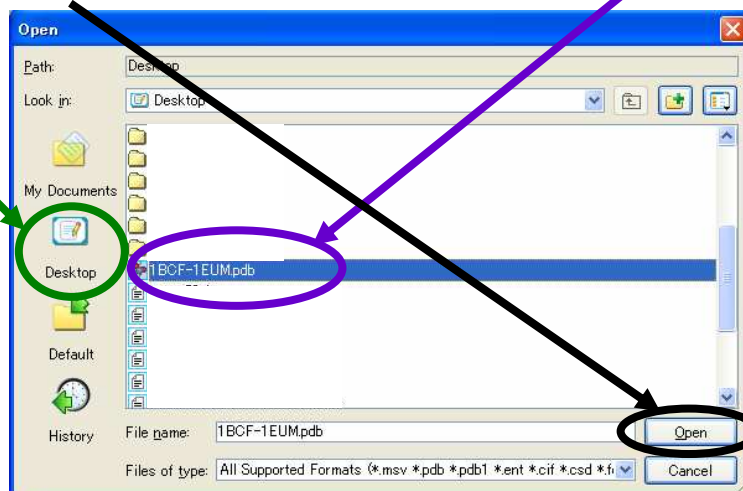
- 「Discovery Studio 1.7」を起動(ダブルクリック)



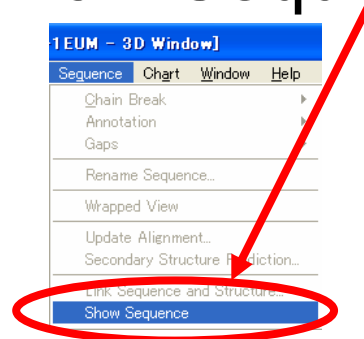
- 「File」メニュー→「Open...」を選択



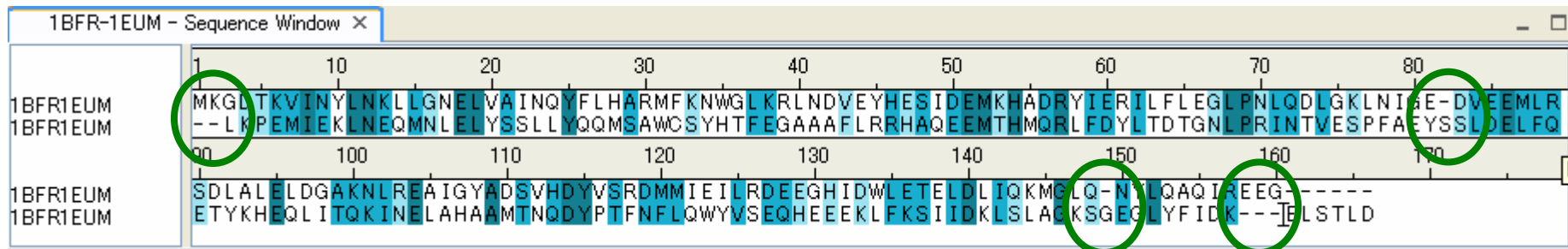
- Desktopをクリックし「リテラシー」フォルダから、ファイル「1BCF-1EUM.pdb」を選択し、「Open」をクリック



- 「Sequence」メニュー→「Show Sequence」を選択し、配列を表示



- カーソルをGapの部分にして、「Space」を入力し、先程のアラインメントに合わせてみましょう

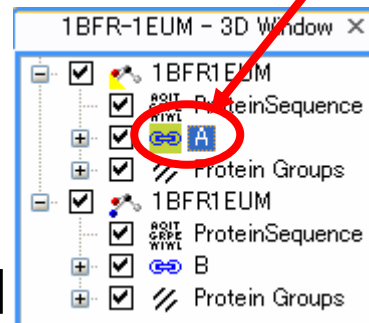


濃い緑が一致した残基を表しています。
Windowの右下に配列の一致度13.1%が表示されています

Sequence Identity: 13.1% Sequence Similarity

- 既に構造アラインメントされていますが、今編集したアラインメントを基に2つの構造を重ね合わせ (superimpose)、RMSDを計算してみましょう

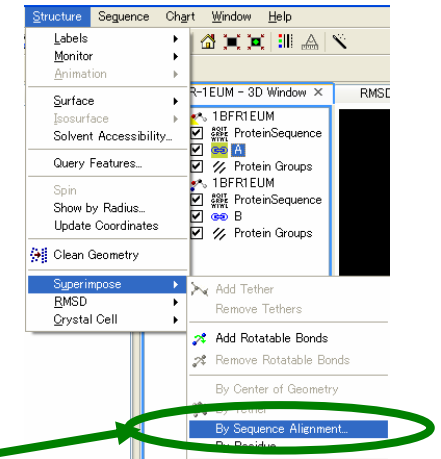
- 「3D Window」で1BFRのA chainを選択



- 「Structure」

→「Superimpose」

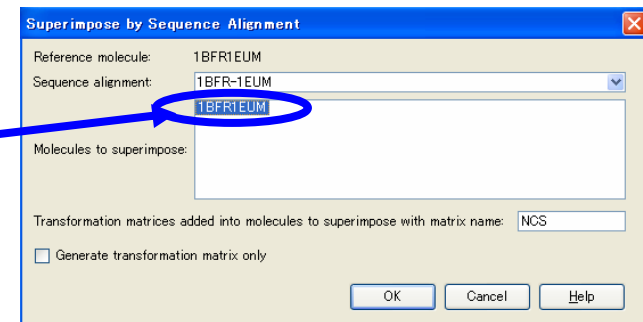
→「By Sequence Alignment」を選択



- Windowが開くので

→「1BCF1EUM」を選択し、

→「OK」をクリック



- 「Text Window」が表示され、
 - 153残基を使って、
 - RMSD=2.08と計算されました。
- 同時に「3D Window」の構造は重ねあわされています。

1BFR-1EUM - 3D Window		RMSD_Report_1 - Text Window ×
Superimpose By Sequence Alignment		
C-Alpha atom RMSD to reference protein: 1BFR1EUM based on all residues		
Protein	RMSD	Residues Used
1BFR1EUM	2.08	153

- では、どれくらい似た構造か、表示 (Display Style) を変えてみましょう。

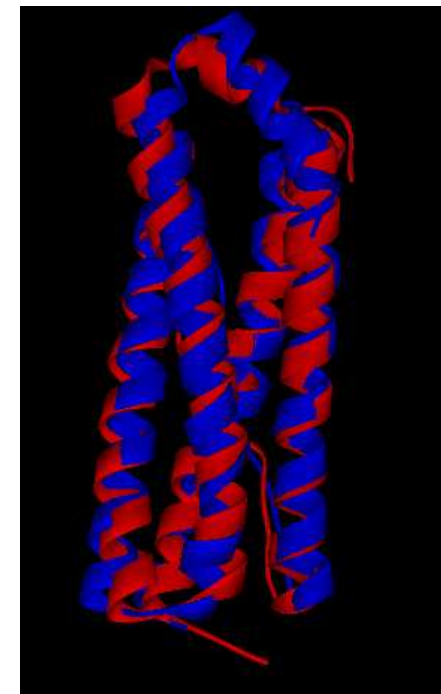
- A chain(1BFR:D)を選択し、
 - 「Ctrl」+「D」(ショートカット)・・・ボタンをクリックと同じ

- Atom: None
- Protein: Solid ribbon
 - Coloring—Custom:赤



- B chain(1EUM:B)を選択し、
 - 「Ctrl」+「D」(ショートカット)

- Atom: None
- Protein: Solid ribbon
 - Coloring—Custom:青



タンパク質立体構造比較サーバー2: DALI / FSSSP

■ [URL] <http://ekhidna.biocenter.helsinki.fi/dali/>

□ →Keyword Search: **1BFR** → (1bfrD/1-158)**browse**

The Dali Database website interface. The search box contains '1bfr'. Below the search box, there are options for 'Fold Index' and 'Fold Tree'.

Dali database query: 1BFR

Click on the Repres. links to browse the alignments and structural neighbours of the fold class.

PDB chain	Repres.	Browse	Interact	Fold	Compound
1bfrA/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN
1bcfA/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN (CY
1bcfB/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN (CY
1bfrB/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN
1bfrC/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN
1bfrD/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN
1bfrE/1-158	1bfrA_1	browse	interact	520	BACTERIOFERRITIN

1bfrA: Structural Neighbours in PDB90 and structural alignments

- PDB90 is a representative subset of PDB chains that are less than 90 % sequence identical to each other
- No: the top 20 alignments, sorted by Z-score, are shown
- Chain: PDB entry code plus chain identifier
- raw-score: the sum of weighted similarities of intramolecular distances that Dali maximizes
- Z-score: normalized score that depends on the size of the structures
- id: percentage of identical amino acids over all structurally equivalent residues
- lali: number of structurally equivalent residues
- rmsd: root-mean-square deviation of C-alpha atoms in the least-squares superimposition of the structurally equivalent C-alpha atoms
- Description: the COMPND record from the PDB entry

No	Chain	raw-score	Z-score	Xid	lali	rmsd	Description
1	1bfrA	1845.2	30.7	100	158	0.0	BACTERIOFERRITIN
2	1j5ca	1806.0	29.8	47	155	0.8	BACTERIOFERRITIN
3	1nf8F	1483.8	22.1	28	158	1.8	BACTERIOFERRITIN
4	1vlgA	1847.2	20.3	19	158	2.3	FERRITIN
5	1rfsA	1955.9	20.1	21	158	2.0	MITOCHONDRIAL FERRITIN
6	1h86A	1832.8	20.0	17	154	1.9	FERRITIN LIGHT CHAIN 1
7	1nf8A	1835.3	19.7	19	154	1.9	M FERRITIN
8	1tesA	1323.7	19.5	17	159	2.0	FERRITIN
9	1eumA	1278.9	19.3	13	158	2.2	FERRITIN 1
10	2fha	1299.8	19.1	21	154	1.9	FERRITIN
11	1kr9A	1262.5	18.9	14	157	2.2	FERRITIN
12	1j5aA	1190.4	18.7	24	155	1.8	DLP-2
13	1n1aA	1186.2	18.5	24	135	1.6	DPS PROTEIN
14	1j5A	1193.0	18.0	20	133	1.7	DLP-1
15	1db9A	1187.4	17.8	15	151	2.8	RUBREVERTIN
16	2b6cC	1157.7	17.7	15	135	2.0	NON-HEME IRON-CONTAINING FERRITIN
17	1o9rA	1165.4	17.4	17	140	2.2	AGROBACTERIUM TIMEFACTENS DPS
18	1j14A	1105.1	17.3	15	154	2.0	NEUTROPHIL-ACTIVATING PROTEIN A
19	1j5vA	1175.3	17.0	17	141	2.4	PSEUDOCALASE
20	1unnC	1114.7	16.4	18	134	2.0	DPS-LIKE PEROXIDE RESISTANCE PROTEIN

chain A ≠ B

No 9: 1eumA FERRITIN 1

Raw-score: 1278.9, Z-score: 19.3
Percent-identity: 13, Number of equivalences: 156, C-alpha rmsd: 2.15965

```

.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....
.LKPEMIEKLEQDMQLYSSLLYDONSAGVCSYTFEGAAAFRRHAEEMTHWRDFDYLDTGNLPRINTVESPFAEY
.LL.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....
.LL.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....
..
..
id
11

Domain pairs:
1. 1bfrA/1-158 (fold 520) and 1eumA/1-161 (fold 520) pairwise domain Z-score 19.3
  
```

DALIサーバー(<http://www.ebi.ac.uk/dali/>)ではPDBにない、新しく自分が決めた構造をデータベースと比較したり、2つの構造をアラインメントしたりもできます。(CEやVASTでももちろんできます)

タンパク質立体構造比較サーバー3: VAST

- [URL] <http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>
 - →try: **1BFR** → **VAST** → [D]entire chain → 1EUM Aをチェック
(Cn3D 4.1をインストール) → View 3D Alignment

The image shows a sequence of screenshots from the NCBI Structure VAST search process:

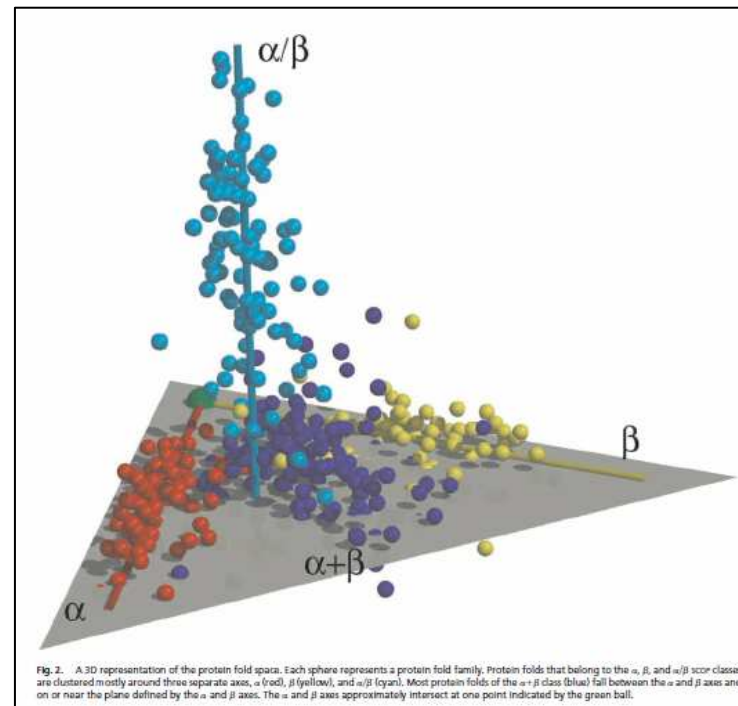
- NCBI Structure VAST Search Page:** Shows the search interface with 'try' entered in the search box. The PDB/MMDB ID '1BFR' is entered in the 'PDB/MMDB ID Code' field.
- MMDB Structure Summary Page:** Shows the search results for '1BFR', including the protein name 'Iron Storage And Electron Transport' and the 'Structure Neighbors' link.
- VAST Structure Neighbors Page:** Shows the list of structure neighbors. The 'entire chain' option is selected for chain D.
- 3D Alignment View:** Shows the 3D alignment of the query structure (1BFR) with its neighbors (1EUM A) in a Cn3D 4.1 window.

Key elements in the screenshots are circled in red and blue to highlight the search process and the selected options.

T. Madej, J.F. Gibrat, S.H. Bryant, *Proteins* **23**, 356-369 (1995), "Threading a database of protein cores"

タンパク質のフォールド空間へのマッピング

- Kimらは、**SCOP**の498フォールドを**DALI**のスコアを基に計量行列を作り、クラスター解析することにより、フォールド空間へマッピングしました。

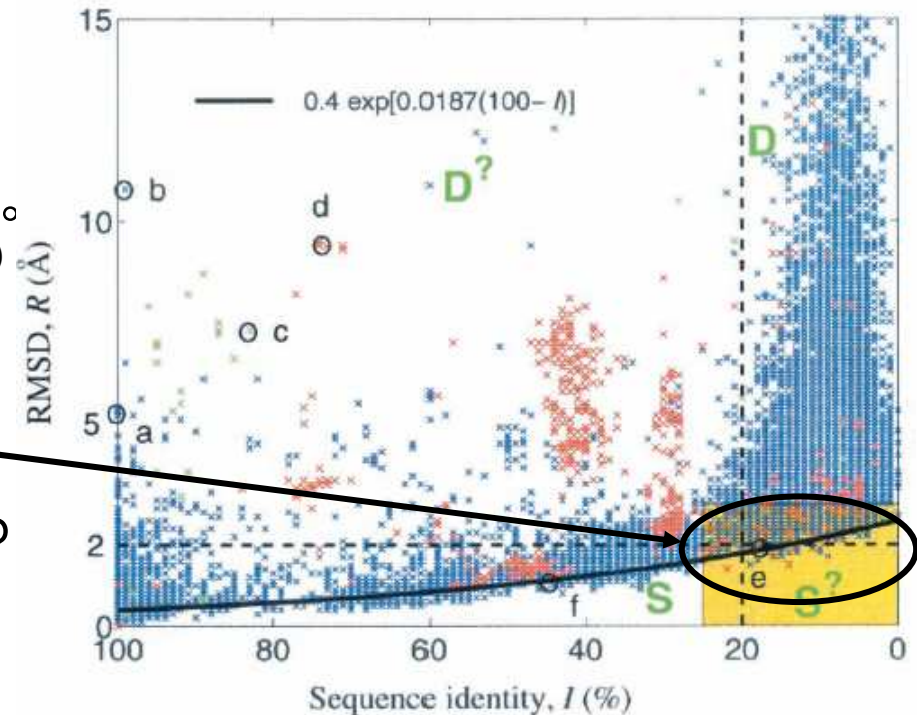


Webで顔写真を
探して下さい。

S.-H. Kim

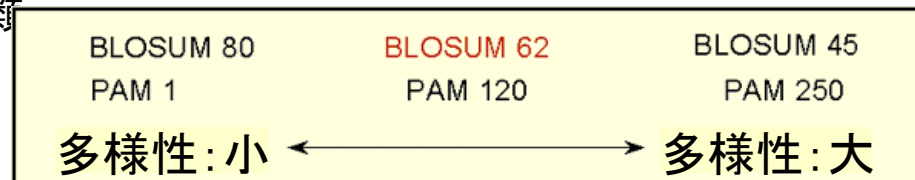
配列類似性と構造類似性の関係

- アラインした53,383タンパク質ペアの RMSD vs. 配列一致度
- 配列一致度が20%以上のものはほとんど(×)RMSD<3Åに入っています。
 - カルモジュリン(×)、イムノグロブリン(×)を除く
- 配列一致度が20%以下でもRMSD<3Åのものも多くある
→配列より構造の方が保存されている



相同性検索の簡単な歴史

- 1970 **ドットマトリックス** Gibbs-McIntyre・・・フィルタリングし一致配列を可視化
 - ダイナミックプログラミング(DP、動的計画法)・・・最適なアラインメントを検索
- 1970 **Needleman-Wunschのアルゴリズム**・・・グローバルDP
- 1978 **PAM行列** Dayhoffら・・・進化率(受け入れられた変異率)に基づくアミノ酸置換行列
 - PAM1・・・平均1%の配列上の位置が置換する時間
- 1981 **Smith-Watermanのアルゴリズム**・・・ローカルDP
- 1985 **FASTA** Lipman & Pearson
- 1990 **BLAST** Altschul et al.
- 1992 **BLOSUM行列** Henikoffら・・・ファミリーのアラインメントに基づくアミノ酸置換行列
 - BLOSUM62・・・62%類似の配列を分類
- 1997 **PSI-BLAST** Altschul et al.



A.J. Gibbs, G.A. McIntyre, *Eur. J. Biochem.* **16**, 1-11 (1970), "The diagram, a method for comparing sequences"

S.B. Needleman, C.D. Wunsch, *J. Mol. Biol.* **48**, 443-453 (1970),

"A general method applicable to the search for similarities in the amino acid sequence of two proteins"

M.O. Dayhoff et al., In *Atlas of Protein Sequence and Structure*, Chap. **22** (1978), "A model of evolutionary change in proteins"

T.F. Smith, M.S. Waterman, *J. Mol. Biol.* **147**, 195-197 (1981), "Identification of common molecular subsequences"

D.J. Lipman, W.R. Pearson, *Science* **227**, 1435-1441 (1985), "Rapid and sensitive protein similarity searches"

S.F. Altschul, et al., *J. Mol. Biol.* **215**, 403-410 (1990), "Basic local alignment search tool"

S. Henikoff, J.G. Henikoff, *PNAS* **89**, 10915-10919 (1992), "Amino acid substitution matrices from protein blocks"

S.F. Altschul, et al., *Nucleic Acids Res.* **25**, 3389-3402 (1997),

"Gapped BLAST and PSI-BLAST: a new generation of protein database search programs"

相同性検索： BLAST, PSI-BLAST

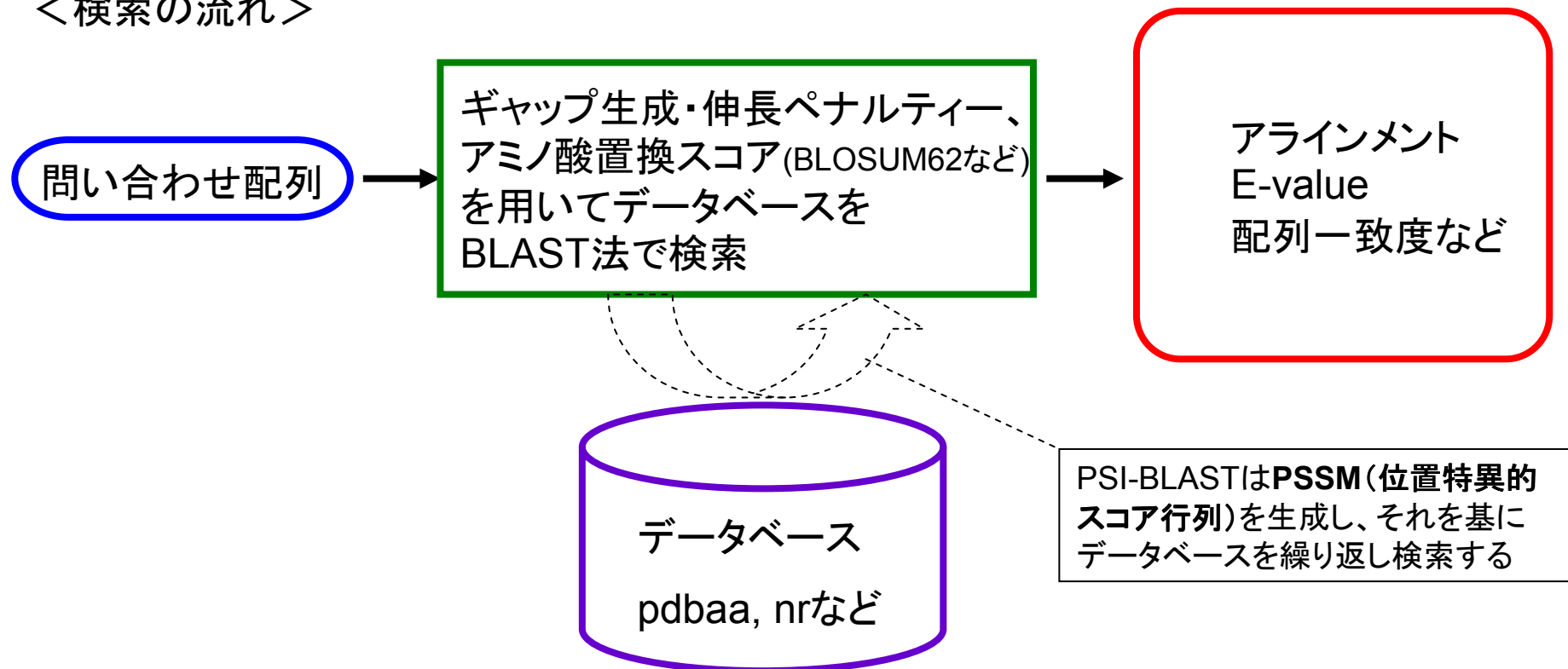
Webで顔写真を
探して下さい。

[URL] <http://www.ncbi.nlm.nih.gov/BLAST/>

S. Altschul

- 1990/1997 Altschulらは、Smith-Watermanのアルゴリズムを改良し、局所的な類似部分配列を高速検索する手法を開発した

<検索の流れ>

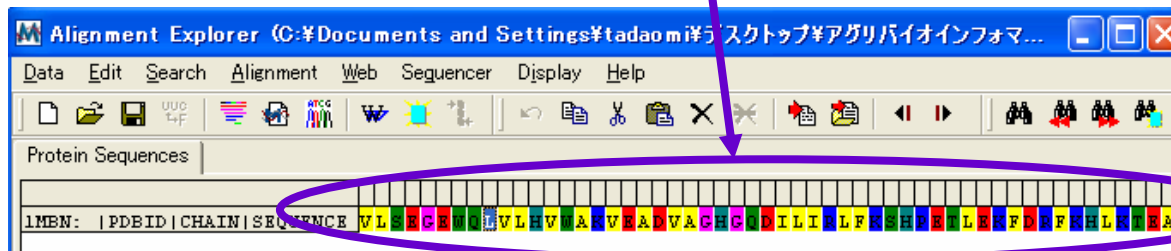


PDBサイトで配列を得る

- PDBサイト([URL] <http://www.rcsb.org>)を開き、「1MBN」を入力、検索(SEARCH)



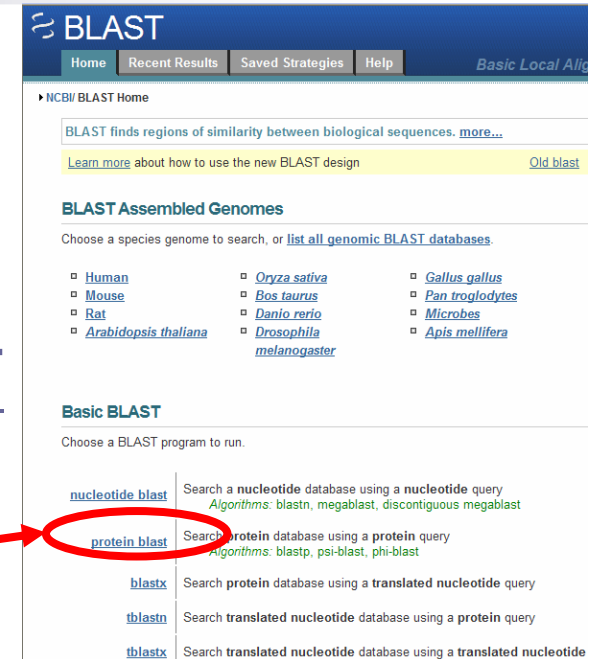
- PDBの情報が表示されるので、「FASTA Sequence」をクリックし、ファイル名「1MBN.fasta」で保存
- ファイル「1MBN.fasta」をダブルクリックで開き、Ctrl+A、Ctrl+Cでfasta形式の配列をコピー



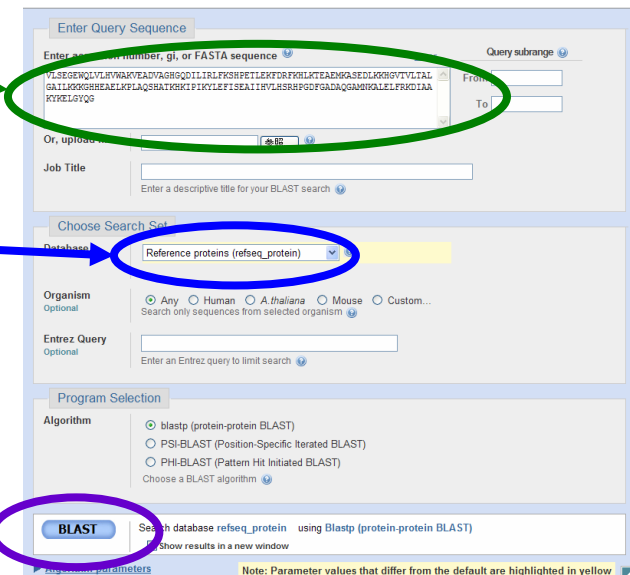
開いたMEGAの2つの
ウィンドウは閉じる

NCBIサイトでblastp検索

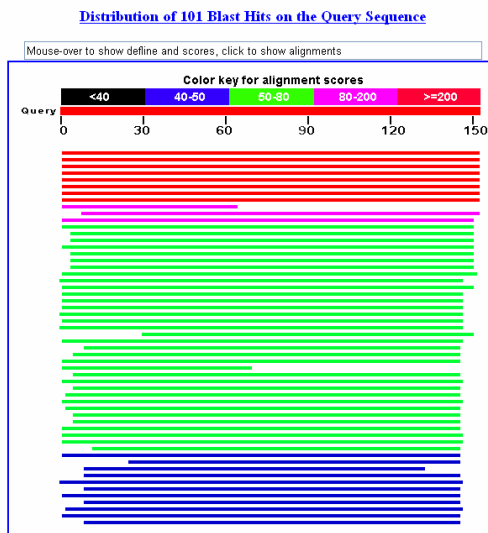
- NCBIのBLASTサイト([URL] <http://www.ncbi.nlm.nih.gov/BLAST/>)を開き、Basic BLASTの項目の中からprotein BLASTをクリック



- FASTA sequenceボックスに「1MBNの配列」を貼り付け、Databaseから「Reference proteins」を選択し、「BLAST!」をクリック



- 検索結果、スコアが**色分け**されて表示され、下の方にヒットした **gi**、**スコア**、**E-value**、さらに下の方に**アラインメント**などが表示されます。



Sequences with E-value BETTER than threshold

Sequences producing significant alignments:

	Score	E
	(Bits)	Value
gi 4885477 ref NP_005339.1 myoglobin [Homo sapiens] >gi 4495...	246	9e-65
gi 47523546 ref NP_999401.1 myoglobin [Sus scrofa]	245	2e-64
gi 21359820 ref NP_038621.2 myoglobin [Mus musculus]	233	1e-60
gi 11024650 ref NP_067599.1 myoglobin [Rattus norvegicus]	233	1e-60
gi 73969220 ref XP_862271.1 PREDICTED: similar to Myoglobin ...	231	3e-60
gi 27806939 ref NP_776306.1 myoglobin [Bos taurus]	227	6e-59
gi 73969218 ref XP_862240.1 PREDICTED: similar to Myoglobin iso...	212	3e-54
gi 50728806 ref XP_416292.1 PREDICTED: similar to myoglobin - c...	206	1e-52
gi 55661130 ref XP_515101.1 PREDICTED: similar to Myoglobin [Pa...	118	4e-26
gi 41053652 ref NP_956880.1 myoglobin [Danio rerio]	105	5e-22
gi 55742013 ref NP_001006870.1 cytoglobin [Xenopus tropicalis]	81.3	8e-15
gi 56961659 ref NP_001008786.1 hypothetical protein LOC417972	79.3	3e-14

Alignments

Get selected sequences Select all Deselect all Tree View

[gi|4885477|ref|NP_005339.1](#) myoglobin [Homo sapiens]
[gi|47523546|ref|NP_999401.1](#) myoglobin [Sus scrofa]
[gi|21359820|ref|NP_038621.2](#) myoglobin [Mus musculus]
 Length=154

Score = 246 bits (629), Expect = 9e-65, Method: Composition-based stats.
 Identifiers = 129/132 (98%), Positives = 192/152 (93%), Gaps = 0/152 (0%)

Query 2 LSEDEKQLVHWARVADVAHNSGQILSLFRHPHLEKFFDFPHLEKTRKAWASLQ 61
 LSEDEKQLVHW KTEAA QHQGSLSEKFEKLEKFFPKHLEKREKASLQ 62
 Subject 3 LSEDEKQLVHWVQVFEADIPHQSQEVLISLFRHPHLEKFFDFPHLEKREKASLQ 62

Query 42 KRKQVTVLTAIGALIKKQKHKLEKLAGHATKHKIPKLEKLEFISEAIIHQLKSRPQ 121
 KRKQ TVLTAIGALIKKQKHKLEKLAGHATKHKIPKLEKLEFISE 11 VL SHRPG
 Subject 63 KRKQVTVLTAIGALIKKQKHKLEKLAGHATKHKIPKLEKLEFISEKIIIVLQKSRPQ 122

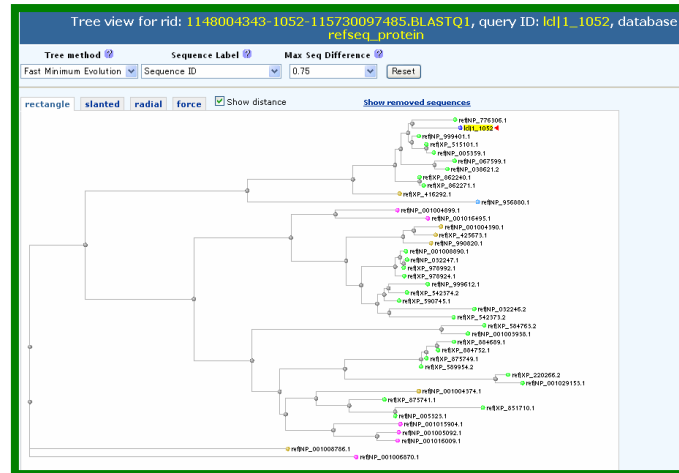
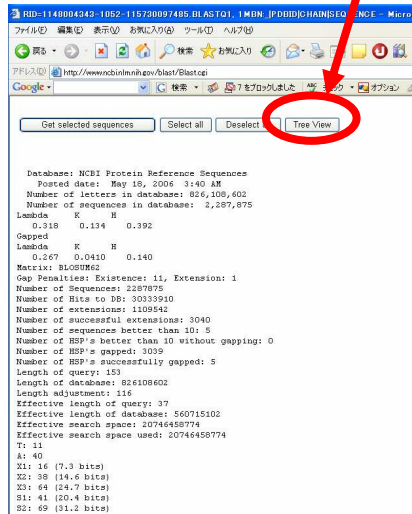
Query 122 DFOADQDAHMKLEKLEFPRDIAKRYELOYQ 153
 DFOADQDAHMKLEKLEFPRDIAKRYELOYQ 154
 Subject 123 DFOADQDAHMKLEKLEFPRDIAKRYELOYQ 154

Homo sapiens(ヒト)やSus scrofa(イノシシ), Mus musculus(ハツカネズミ)などのmyoglobin
 そしてXenopus tropicalis(アフリカツメガエル)のcytoglobinがヒットしています。

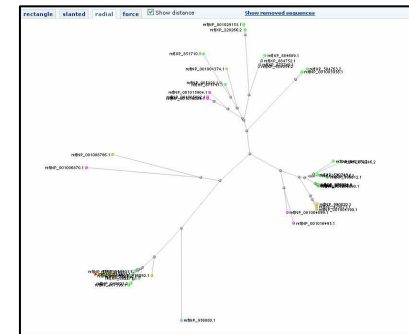
系統樹を表示

- 一番下で、「Distance tree of results」をクリックすると系統樹も描けます。

(実際は、いくつか見たい種、遺伝子などを選択(チェック)し表示すると進化的な関係が得られます)



rectangle表示

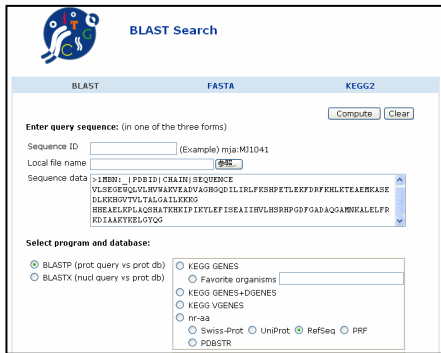


radial表示

→ バイオスタティスティクス基礎論(西田先生ら)
生物配列統計学(岸野先生ら)

同様な検索は多くのサイトで提供されています

■ 例) Genome Net [URL] <http://blast.genome.jp>

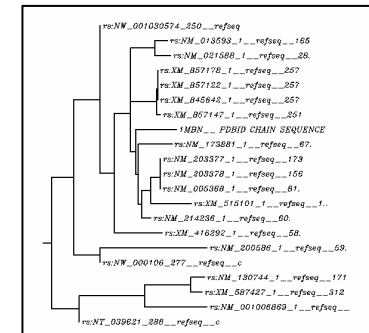
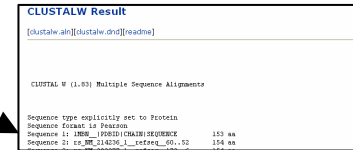


BLASTP Search Result

Database: nr-prot

Protein sequence database entries related to IMBNL_[PDBID]CHAIN|SEQUENCE - 321 hits

Entry	bits	E-val
Top 20		
<input checked="" type="checkbox"/> tc:NM_214236.1 [NM_214236] myoglobin [Sus scrofa]	274	8e-73
<input checked="" type="checkbox"/> tc:NM_203377.1 [NM_203377] myoglobin [Homo sapiens]	273	1e-72
<input checked="" type="checkbox"/> tc:NM_003376.1 [NM_003376] myoglobin [Homo sapiens]	273	1e-72
<input checked="" type="checkbox"/> tc:NM_005368.1 [NM_005368] myoglobin [Homo sapiens]	273	1e-72
<input checked="" type="checkbox"/> tc:NM_173881.1 [NM_173881] myoglobin [Bos taurus]	261	4e-69
<input checked="" type="checkbox"/> tc:NM_057179.1 [NM_057179] similar to Myoglobin isoform 4 [Canis...	261	7e-69
<input checked="" type="checkbox"/> tc:NM_057122.1 [NM_057122] similar to Myoglobin isoform 2 [Canis...	261	7e-69
<input checked="" type="checkbox"/> tc:NM_045642.1 [NM_045642] similar to myoglobin isoform 1 [Canis...	261	7e-69
<input checked="" type="checkbox"/> tc:NM_013593.1 [NM_013593] myoglobin [Mus musculus]	255	4e-67
<input checked="" type="checkbox"/> tc:NT_039621.288 [NT_039621] myoglobin [Mus musculus]	255	4e-67
<input checked="" type="checkbox"/> tc:NM_000106.277 [NM_000106] myoglobin [Mus musculus]	255	4e-67
<input checked="" type="checkbox"/> tc:NM_001030574.250 [NM_001030574] myoglobin [Mus musculus]	255	4e-67
<input checked="" type="checkbox"/> tc:NM_021508.1 [NM_021508] myoglobin [Rattus norvegicus]	255	1e-66
<input checked="" type="checkbox"/> tc:NM_057147.1 [NM_057147] similar to Myoglobin isoform 3 [Canis...	243	2e-63
<input checked="" type="checkbox"/> tc:XM_416292.1 [XM_416292] similar to myoglobin - chicken [Gallu...	229	2e-59
<input checked="" type="checkbox"/> tc:XM_515101.1 [XM_515101] similar to Myoglobin [Pan troglodytes]	118	6e-26
<input checked="" type="checkbox"/> tc:NM_000586.1 [NM_000586] myoglobin [Danio rerio]	109	4e-23
<input checked="" type="checkbox"/> tc:NM_001006869.1 [NM_001006869] cytoglobin [Xenopus tropicalis]	87	2e-16



「配列」を貼り付け、
「RefSeq」を選択し、
「Compute」をクリック
「FASTA」検索もできる

「Top 20」など選択し、
「CLUSTALW」を選択し、
「Exec」をクリック

一番下で
「N-J Tree with branch length」を選択し、
「Exec」をクリック

BLAST, FASTA

ペアワイズシークエンスアラインメント

CLUSTALW, PSI-BLAST

マルチプルシークエンスアラインメント

D. Higgins et al., *Nucleic Acids Res.* **22**, 4673-4680 (1994), "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice"

バイオインフォマティクスの有用なリンク

■ 統合サイトなど

- Entrez/NCBI [URL] <http://www.ncbi.nlm.nih.gov/gquery/>
 - GenBank, PubMed, BLAST,, Structure(VAST), CDD
- Services/EMBL-EBI [URL] <http://www.ebi.ac.uk/services/>
 - <Toolbox> BLAST, FASTA, InterProScan, CLUSTALW, DALI,,,
- GenomeNet/Kyoto Univ. [URL] <http://www.genome.jp/>
 - KEGG, KEGG2(PATHWAY,,),,, BLAST, FASTA, CLUSTALW
 - DBGET Database Links [URL] <http://www.genome.jp/dbget/dbget.links.html>
 - GenBank, EMBL, PubMed,, PDB, Prosite, Pfam, Blocks, ProDom, PRINTS
- DDBJ/NIG [URL] <http://www.ddbj.nig.ac.jp/Welcome-e.html>
 - <Search&Analysis> BLAST, FASTA, ClustalW,,,
- ExpASy [URL] <http://www.expasy.ch/>
 - PROSITE,,,
 - <Site Map> <http://www.expasy.ch/sitemap.html>
 - SWISS-MODEL,,,
- ANGIS [URL] <http://www.angis.org.au/>
 - <Links> [URL] <http://www.angis.org.au/links.shtml>
- Bio-mirror [URL] <http://bio-mirror.net/>

参考までに、
バイオインフォマティクスのWeb上での勉強に
(少し古いかも知れませんが)

■ JST Webラーニングプラザ

□ <http://weblearningplaza.jst.go.jp/>

- 分野・映像から選ぶ→ライフサイエンス

■ GenomeNet バイオインフォマティクス入門コース

□ <http://www.genome.jp/Japanese/lect/course.html>

- 昔の京大での講義

【課題1】 構造比較（構造類似性）

1ISKの(1OUNとの)構造比較をして、結果をPowerPointにまとめよ

1. CEサイト([URL] <http://cl.sdsc.edu/>)で「1ISK:A」の類似構造を検索し、「1OUN:A」との構造アライメントを得て、そのPDBをダウンロードする
2. Discovery Studio 1.7で、ダウンロードしたPDBを表示し、アライメントを合わせ、RMSDを計算し、構造を色を変えたSolid Ribbon表示にする
3. 配列のアライメント、構造の画像をPowerPointに貼り、配列一致度、RMSD、アライメントされた残基数などを記述する

1BFR-1EUM - Sequence Window	1ISK-1OUN - 3D Window	1ISK-1OUN - Sequence Window ×	RMSD_Report_2 - Text Window
	1 10 20 30 40 50 60 70 80		
1ISK1OUN	MN-----TTEHMTAVVQRYVAALNAGFLDGLVALFADDAITVEDEFGSEPRSCATAAIREFYANSLKPLAVELTQEVRAA-----NEAA		
1ISK1OUN	--GDKPIWFGIGSSFIQHYYQLFDNDRT-QISAIYIDASCLTWEG--DFQKKAATVEKLSL-PQKIQHSITAQDHPPTPDSCDLSM		
	90 100 110 120 130 140 150 160 170		
1ISK1OUN	FAFTVSEFYQGRKTYVAPIDHFRNGA---GVVSMRALFGEKNTIHAGA		
1ISK1OUN	VVGQLKARE---DRIMGFHQMLKYNINDAWVCTNDMFRLLALHFE-----		

【課題2】 相同性検索（配列類似性）

1UB4の相同性検索を行い、結果をPowerPointにまとめよ

1. PDBサイト([URL] <http://www.rcsb.org>)で「1UB4」を検索し、FASTA形式の配列を保存する
2. 保存したファイルを開き、chain Aをコピーする
3. NCBI([URL] <http://www.ncbi.nlm.nih.gov/blast/>)かGenome Net([URL] <http://blast.genome.jp/>)のBLASTサイトでその配列をrefseqデータベースを用いて相同性検索する
 1. [オプション] 系統樹を描く
4. 課題の配列はどんなファミリーに属するか、検索された相同性のあるいくつか配列に関して、どの様な生物種の何という遺伝子・タンパク質か、そしてそのE-value、アラインメントなどを含め記述する
 1. [オプション] 系統樹から進化的な関係を述べる

(参考) タカラバイオ [URL] <http://www.takara-bio.co.jp/news/2005/08/17.htm>

[URL] <http://www.takara-bio.co.jp/news/2006/05/15-5.htm>

Nature Japan [URL] <http://www.natureasia.com/japan/jobs/tokusyuu/050908-2.php>

<課題の提出>

- 上記、【課題1】、【課題2】をPowerPointで2ページの1つのファイルにまとめる
- PowerPointファイルを添付し、Subject(件名)を「課題:構造比較・相同性検索」とし、本文に学籍番号、講義用ID、氏名等記入し、E-mailで以下のメールアドレスへ送信する
 - E-mail address: tadaomi@iu.a.u-tokyo.ac.jp