

平成27年4月13日

構造バイオインフォマティクス基礎

# 立体構造データベースと その利用

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

# 講義の予定

---

1. 4月13日(月)  
担当: 寺田 透  
内容: 立体構造データベースの利用と立体構造データの可視化
2. 4月20日(月)  
担当: 永田宏次  
内容: X線結晶構造解析による立体構造決定のインフォーマティクス
3. 4月27日(月)  
担当: 寺田 透  
内容: 立体構造からの情報抽出
4. 5月11日(月)  
担当: 清水謙多郎  
内容: 立体構造のモデリング

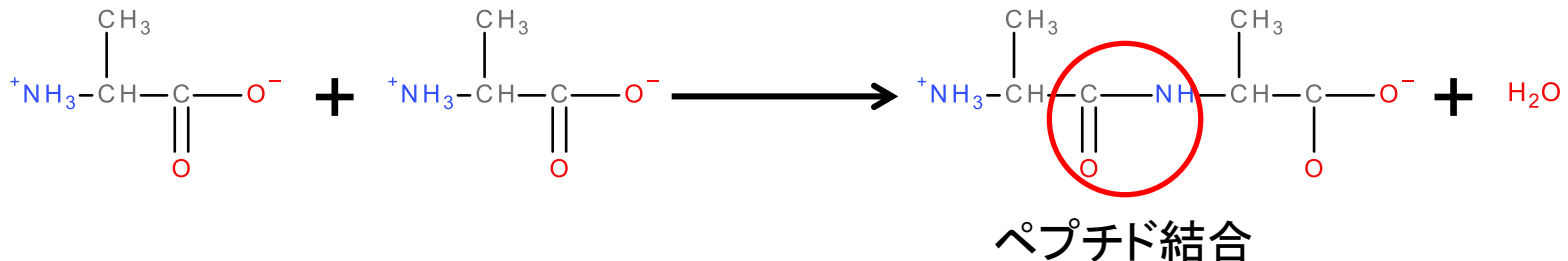
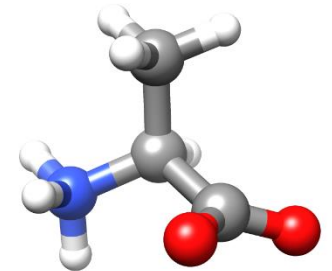
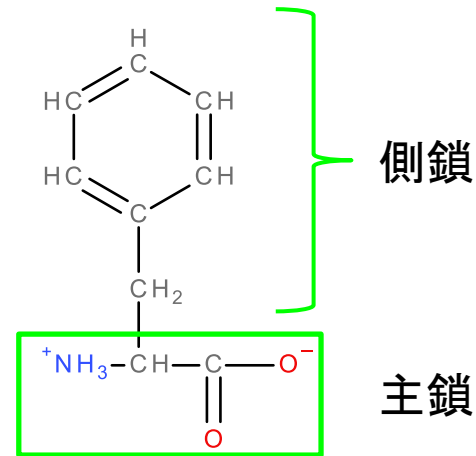
# 本日の講義内容

---

- タンパク質立体構造データベース
  - 検索
  - データのダウンロード
- 立体構造データの可視化
- 立体構造データフォーマット
- 配列データベースとの連携
- 実習

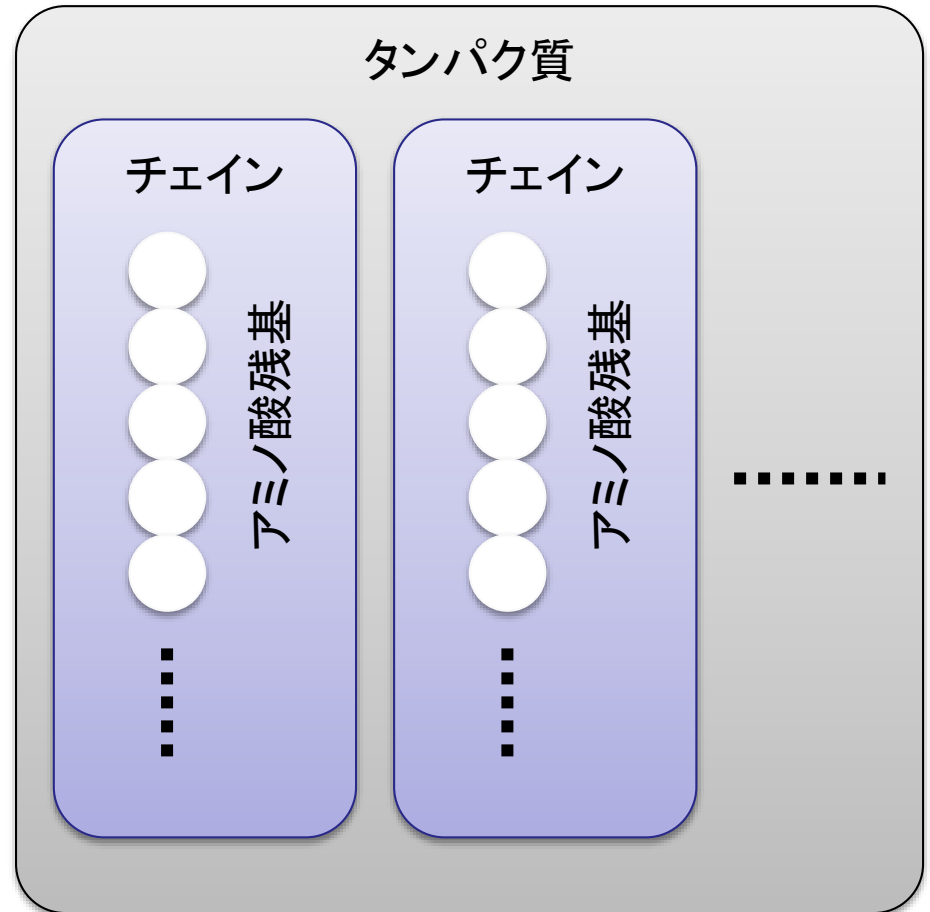
# タンパク質の構造(1)

- アミノ酸
  - 主鎖と側鎖
  - 20種類の標準アミノ酸
  - L型のみ
- ポリペプチド
  - アミノ酸がペプチド結合を介して重合したもの



# タンパク質の構造(2)

- 単量体
  - 1本のポリペプチド鎖(チェーン)からなる
- 多量体
  - 複数のポリペプチド鎖からなる
- 特定の立体構造をとる
  - 部分的に特定の立体構造をとらない領域を持つ場合もある



# 立体構造データベース

---

- Protein Data Bank (PDB)
- タンパク質、核酸などの生体高分子の立体構造を収集、公開している世界で唯一のデータベース
- 2015年4月時点でのエントリ数は約108,000
- 主なWebサイト
  - 米国 : <http://www.rcsb.org/>
  - 欧州 : <http://www.ebi.ac.uk/pdbe/>
  - 日本 : <http://www.pdbj.org/>

# データベースへのアクセス

- RCSBのサイト (<http://www.rcsb.org/>)

The screenshot shows the top navigation bar of the RCSB PDB website. It includes a dark blue header with white text for navigation: "RCSB PDB", "Deposit", "Search", "Visualize", "Analyze", "Download", "Learn", and "More". On the right side of this bar is a "MyPDB Login" button. Below the navigation bar is a light blue banner with the "RCSB PDB PROTEIN DATA BANK" logo on the left. To the right of the logo is the text "An Information Portal to 107754 Biological Macromolecular Structures". A search bar is located in the center of the banner, with the placeholder text "Search by PDB ID, author, macromolecule, sequence, or ligands" and a "Go" button. Below the search bar are links for "Advanced Search", "Browse by Annotations", "Search History (1)", and "Previous Results (107754)". At the bottom of the banner are logos for "PDB-101", "wwPDB", "EMDataBank", "NUCLEIC ACID DATABASE", and "StructuralBiology Knowledgebase". On the far right of the banner are social media icons for Facebook, Twitter, YouTube, Apple, and Android.

A vertical navigation sidebar on the left side of the page. It has a dark blue background with white text and icons. The items are: "Welcome" (with a bookmark icon), "Deposit" (with an upload icon), "Search" (with a magnifying glass icon), "Visualize" (with a camera icon), "Analyze" (with a grid icon), "Download" (with a download icon), and "Learn" (with a book icon).

## A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

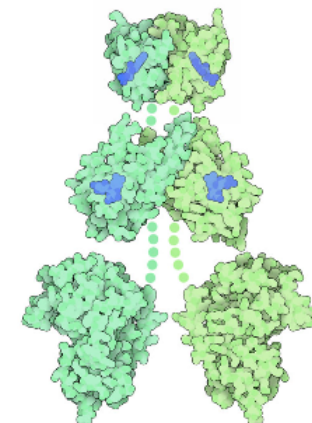
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

### Structure and Health Focus: HIV

A video thumbnail for "HIV Resources". It features the RCSB PDB logo and the text "PDB-101" and "rcsb.org". The image shows a 3D model of a virus particle.

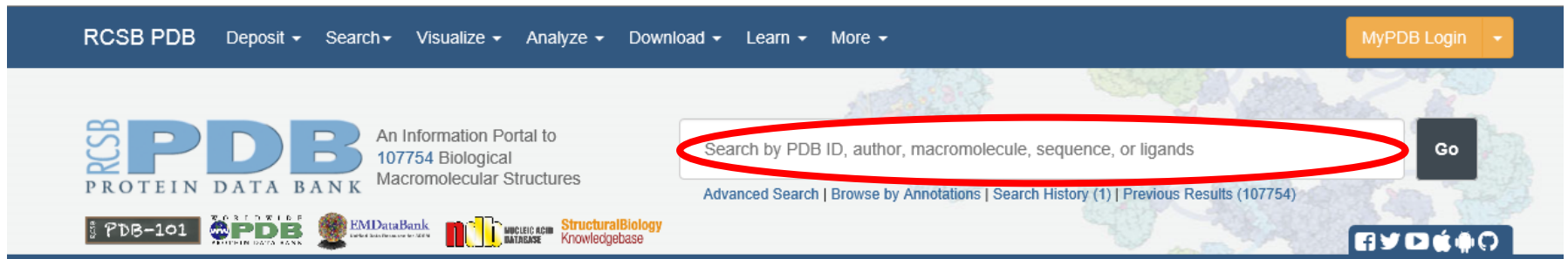
A video thumbnail for "High School Video Challenge". It shows a 3D model of a cell with a red structure and a green structure. The word "CYTOPLASM" is visible in the background.

## March Molecule of the Month



Phototropin

# 検索

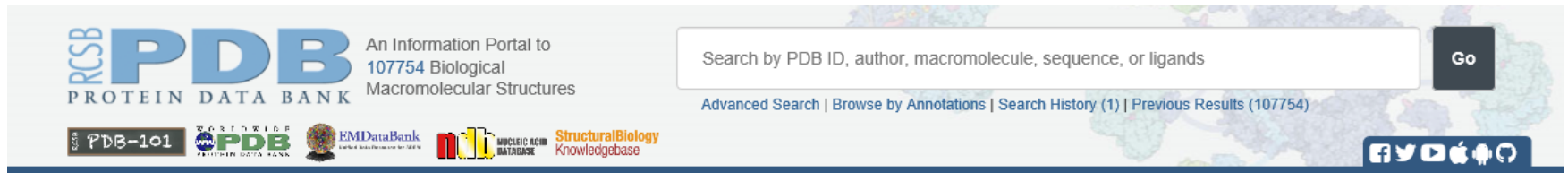


- 立体構造データに対するテキスト検索
  - 例: “HIV Protease”, “Green Fluorescent Protein”, etc.
- PDB IDを直接指定することも可能
  - PDB ID: 数字1文字と英数字3文字からなる、各立体構造データに固有のID
  - 例: 1HVR, 1J4N, etc.



# 検索結果の表示

- 上部のタブをクリックして表示を切り替える



- Summary
- 3D View
- Sequence
- Annotations
- Seq. Similarity
- 3D Similarity
- Literature
- Biol. & Chem.
- Methods
- Links


RATIONAL DESIGN OF POTENT, BIOAVAILABLE, NONPEPTIDE CYCLIC UREAS AS HIV PROTEASE INHIBITORS

1HVR [Display Files](#) [Download Files](#) [Download Citation](#)

- Summary: 文献、組成
- Sequence: アミノ酸配列、2次構造
- Annotations: 立体構造分類、ファミリー分類
- Methods: 立体構造決定法

# データのダウンロード

- 右上の「Download Files」から立体構造データをダウンロードできる

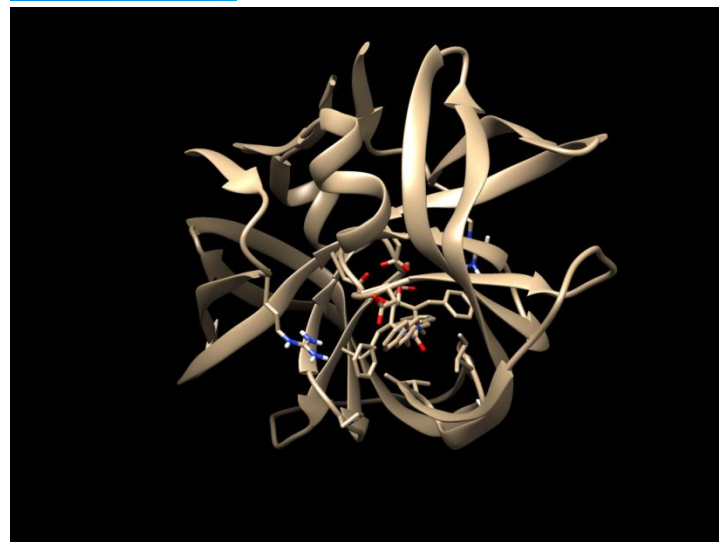
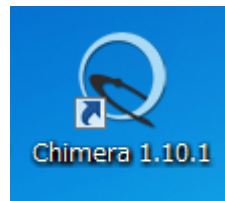


The screenshot shows the PDB website interface. At the top left is the PDB logo with the text 'RCSB PDB An Information Portal to 107754 Biological Macromolecular Structures'. To the right is a search bar with the text 'Search by PDB ID, author, macromolecule, sequence, or ligands' and a 'Go' button. Below the search bar are links for 'Advanced Search | Browse by Annotations | Search History (1) | Previous Results (107754)'. At the bottom left of the header are logos for 'PDB-101', 'PDB', 'EMDataBank', 'NUCLEIC ACID DATABASE', and 'StructuralBiology Knowledgebase'. At the bottom right of the header are social media icons for Facebook, Twitter, YouTube, Apple, and Android. Below the header is a navigation menu with buttons for 'Summary', '3D View', 'Sequence', 'Annotations', 'Seq. Similarity', '3D Similarity', 'Literature', 'Biol. & Chem.', 'Methods', and 'Links'. The main content area shows the title 'RATIONAL DESIGN OF POTENT, BIOAVAILABLE, NONPEPTIDE CYCLIC UREAS AS HIV PROTEASE INHIBITORS' and the resolution '1HVR'. To the right of the resolution are three buttons: 'Display Files', 'Download Files' (circled in red), and 'Download Citation'.

- 「PDB File (Text)」を右クリックし、「対象をファイルに保存」を選び、デスクトップに保存

# 立体構造データの可視化

1. PDB ID「1HVR」を検索し表示
2. ファイルをダウンロードし、デスクトップに保存
3. Chimera 1.10.1のアイコンをダブルクリックし起動
4. メニューの「File」→「Open」で1HVR.pdbを開く



# UCSF Chimeraの操作(1)

---

- 回転
  - マウスの左ボタンを押しながらドラッグ
- 並進
  - マウスのホイールを押しながらドラッグ
- ズーム
  - マウスの右ボタンを押しながらドラッグ
  - マウスのホイールを回転

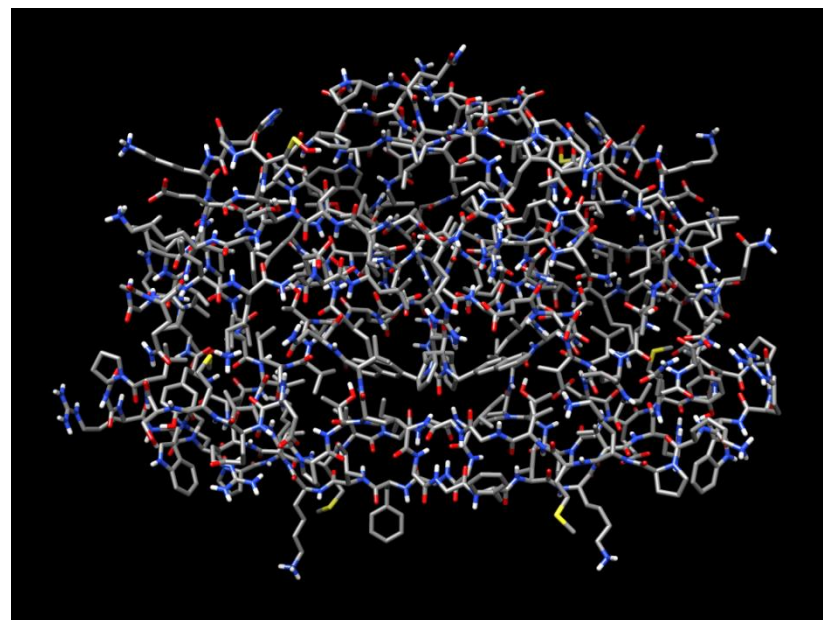
# UCSF Chimeraの操作(2)

---

- 選択(selection)
  - 「Ctrl」キーを押しながら左クリック
  - 選択を追加する時は、「Ctrl」と「Shift」キーを押しながら左クリック
  - 何もないところを「Ctrl」キーを押しながら左クリックすると解除
  - 「↑」キーで、選択範囲を原子→残基→チェーン→分子の順に拡大
- フォーカス
  - メニューの「Actions」→「Focus」で選択された原子を拡大表示する

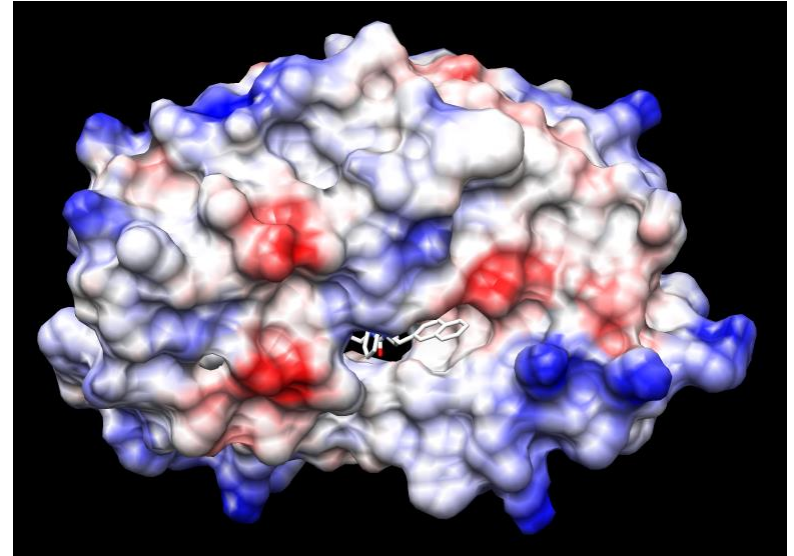
# 表示の変更(1)

- メニューの「Actions」を用いる
  1. 「Actions」→「Atoms/Bonds」  
→「show」
  2. 「Actions」→「Ribbon」  
→「hide」
  3. 「Actions」→「Color」  
→「by element」
- 選択している場合は、選択された原子の表示が変わる



# 表示の変更(2)

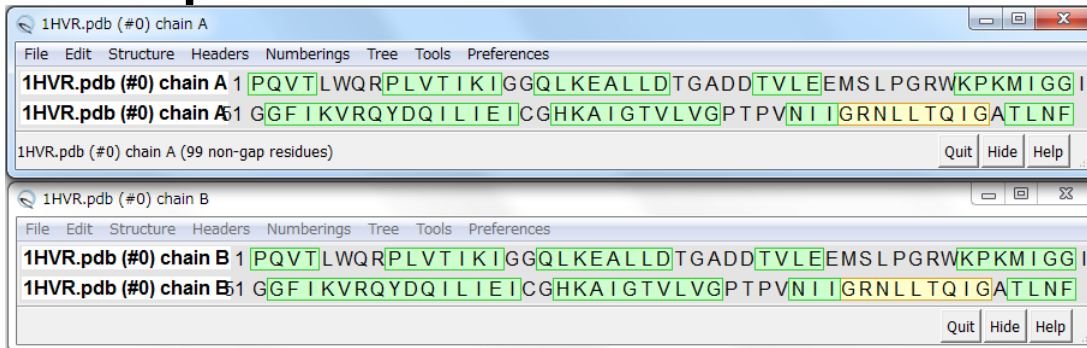
- 「Actions」→「Surface」→「show」で分子表面を表示
- 「Tools」→「Surface/Binding Analysis」→「Coulombic Surface Coloring」で静電ポテンシャルで色分け  
(少し時間がかかる)



青: 正に帯電  
赤: 負に帯電

# 配列の表示

- メニューの「Tools」→「Sequence」→「Sequence」

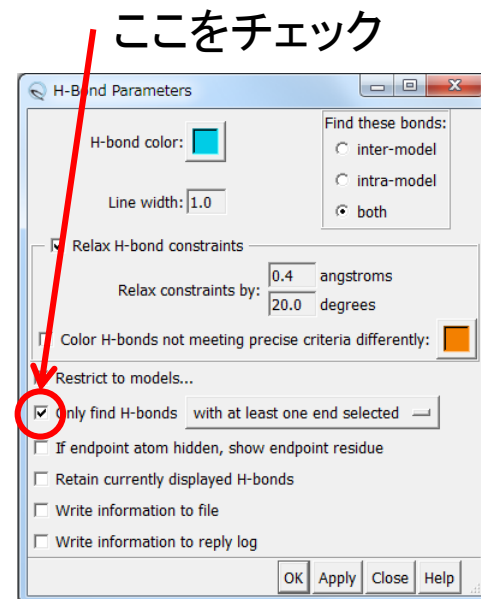


- アミノ酸を選択すると、立体構造上でも選択される
  - マウスの左ドラッグで領域を選択
  - 「Shift」キーを押しながら左ドラッグで追加



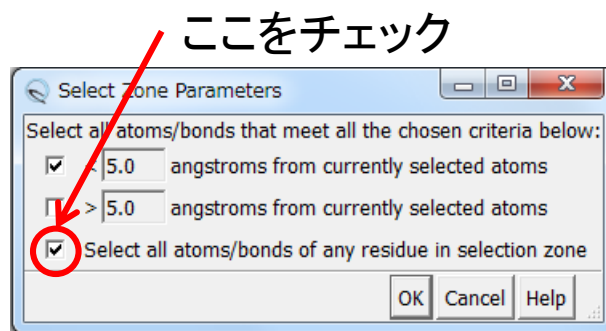
# 相互作用の検出(1)

- 水素結合の検出
  - X線結晶構造解析から得られた構造には水素原子の座標が含まれていないことが多いため、重原子間の距離で判定する
  - 水素結合を形成する重原子(窒素や酸素)間の距離は概ね  $2.8 \text{ \AA} \sim 3.5 \text{ \AA}$
- メニューの「Select」→「Residue」→「XK2」でリガンドを選択
- メニューの「Tools」→「Structure Analysis」→「FindHBond」を選択し右のように設定  
→水素結合が青色の線で表示される



# 相互作用の検出(2)

- 疎水性相互作用は原子間距離で検出
- リガンドXK2を選択
- メニューの「Select」→「Zone」を選択し、右のように設定し「OK」→リガンドから5 Åにある残基が選択される
- 「Select」→「Name Selection」で選択範囲を保存して後で呼び出すことができる



# データの保存

---

- メニューの「File」→「Save Session As」で作業状態を保存できる
- 保存した作業状態は、「File」→「Restore Session」で呼び出すことができる
- 画像は「File」→「Save Image」で保存できる
- 「File」→「Close Session」で立体構造データは閉じられ、初期状態に戻る

# PDBフォーマット(1)

The screenshot shows the PDB website interface. At the top, there is a search bar with the text "Search by PDB ID, author, macromolecule, sequence, or ligands" and a "Go" button. Below the search bar, there are navigation links: "Advanced Search | Browse by Annotations | Search History (4) | Previous Results (81)". The main content area displays the entry for "1HVR" with the title "RATIONAL DESIGN OF POTENT, BIOAVAILABLE, NONPEPTIDE CYCLIC UREAS AS HIV PROTEASE INHIBITORS". Below the title, there is a "Primary Citation" section with the text "Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors." and a list of authors. On the right side, there is a "Display Files" dropdown menu with the following options: "FASTA Sequence", "PDB File", "PDB File (Header)", "mmCIF File", "mmCIF File (Header)", "PDBML/XML File", and "PDBML/XML File (Header)". The "PDB File" option is circled in red.

- PDBファイルには、座標データを含む様々な情報が記載されている
- 「Display Files」→「PDB File」で中身を表示することができる

# PDBフォーマット(2)

- 冒頭部分には、生体高分子の名前や由来、文献等のデータが記載されている

```
HEADER      HYDROLASE (ACID PROTEINASE)                14-FEB-94   1HVR
TITLE       RATIONAL DESIGN OF POTENT, BIOAVAILABLE, NONPEPTIDE CYCLIC
TITLE       2 UREAS AS HIV PROTEASE INHIBITORS
COMPND      MOL_ID: 1;
COMPND      2 MOLECULE: HIV-1 PROTEASE;
COMPND      3 CHAIN: A, B;
COMPND      4 ENGINEERED: YES
SOURCE      MOL_ID: 1;
SOURCE      2 ORGANISM_SCIENTIFIC: HUMAN IMMUNODEFICIENCY VIRUS 1;
SOURCE      3 ORGANISM_TAXID: 11676;
SOURCE      4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE      5 EXPRESSION_SYSTEM_TAXID: 562
KEYWDS      HYDROLASE (ACID PROTEINASE)
EXPDTA      X-RAY DIFFRACTION
AUTHOR      C. -H. CHANG
```

# PDBフォーマット(3)

①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	
ATOM	1	N	PRO	A	1	-12.735	38.918	31.287	1.00	39.83	N
ATOM	2	CA	PRO	A	1	-12.709	39.097	29.830	1.00	39.29	C
ATOM	3	C	PRO	A	1	-13.575	38.051	29.162	1.00	39.78	C
ATOM	4	O	PRO	A	1	-14.097	37.126	29.753	1.00	38.67	O
ATOM	5	CB	PRO	A	1	-11.243	39.010	29.398	1.00	37.79	C
ATOM	6	CG	PRO	A	1	-10.636	38.128	30.469	1.00	38.69	C
ATOM	7	CD	PRO	A	1	-11.368	38.593	31.729	1.00	37.10	C
ATOM	8	H2	PRO	A	1	-13.142	39.756	31.758	0.00	15.00	H
ATOM	9	H3	PRO	A	1	-13.429	38.158	31.502	0.00	15.00	H
ATOM	10	N	GLN	A	2	-13.682	38.255	27.876	1.00	41.01	N

①レコード名(標準アミノ酸はATOM、非標準はHETATM)

②原子番号

③原子名(主鎖アミド窒素:N、 $\alpha$ 炭素:CA、 $\beta$ 炭素:CBなど)

④残基名(3文字表記)

⑤Chain ID

⑥残基番号(配列データベース中の番号に一致させる)

⑦⑧⑨それぞれ原子のx, y, z座標 [Å]

⑩occupancy(その原子の重み因子、通常は1.00)

⑪温度因子B [Å<sup>2</sup>](X線結晶解析で決定されている場合のみ意味がある)

# 参考：可視化ソフトウェア入手先

---

- Discovery Studio Visualizer  
(<http://accelrys.com/products/discovery-studio/visualization-download.php>)
- PyMol (<http://www.pymol.org/>)
- RasMol (<http://www.openrasmol.org/>)
- Swiss-PdbViewer (<http://spdbv.vital-it.ch/>)
- UCSF Chimera  
(<http://www.cgl.ucsf.edu/chimera/>)

# 実習課題1(1)

---

1. RCSBのサイトで「Green Fluorescent Protein」を検索する
2. たくさんヒットするが、その中からPDB IDが「1GFL」のものを選びSummaryを表示する
3. PDBファイルをダウンロードしてデスクトップに保存
4. Chimeraでこのファイルを開く
5. メニューから「Select」→「Chain」→「B」でB鎖を選択した後、「Actions」→「Atoms/Bonds」→「delete」でB鎖を削除する
6. 「Actions」→「Focus」で位置と大きさをウィンドウに合わせる



# 実習課題1(2)

このタンパク質は、紫外線を受け取って緑色の蛍光を発する。  
発色団はSer65-Tyr66-Gly67が自発的に環化してできる。

7. 「Tools」→「Sequence」→「Sequence」で配列を表示し、  
Ser65-Tyr66-Gly67を選択
8. 「Actions」→「Atoms/Bonds」→「show」でこれらの残基  
を表示
9. 「Actions」→「Ribbon」→「hide」でこれらの残基のリボン  
表示を消去(環化部分を確認せよ)
10. 選択を解除し、「File」→「Save Image」で画像を保存

# 立体構造決定法

立体構造決定法	エントリー数	割合
X線結晶構造解析	95849	89.0
核磁気共鳴(NMR)	10886	10.1
電子顕微鏡	758	0.7
その他	261	0.2
<b>合計</b>	<b>107754</b>	<b>100.0</b>

- 全エントリー中9割近くがX線結晶構造解析法により、立体構造が決定されている。
- 残りのほとんどは核磁気共鳴法
- X線結晶構造解析法については、次回解説

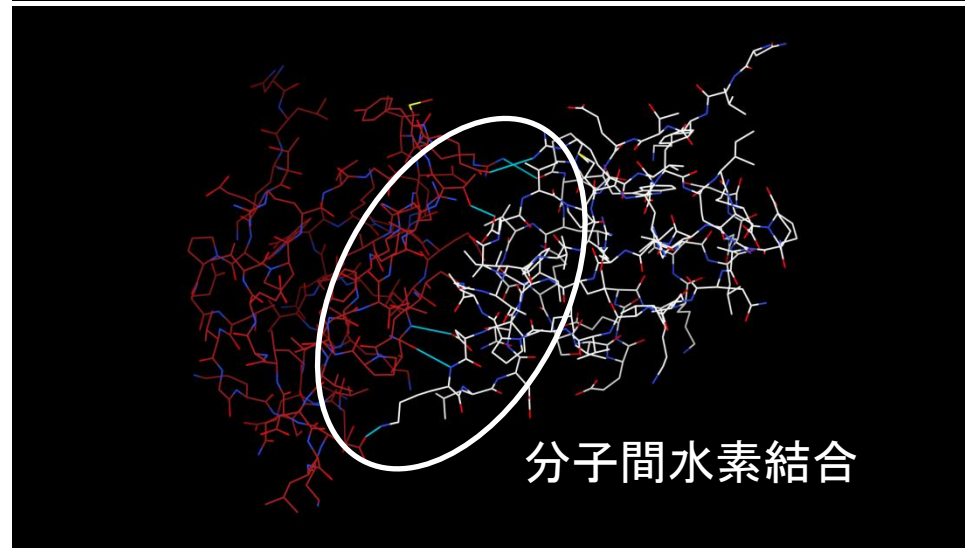
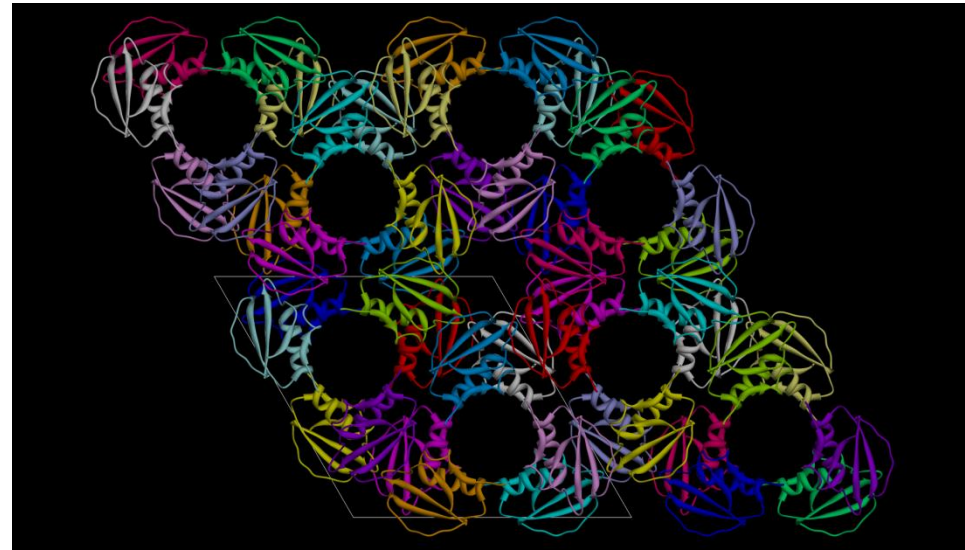
# 座標データに表れる違い

	X線結晶構造解析法	核磁気共鳴法
サンプルの状態	結晶(分子間接触あり)	溶液
分子量の上限	なし	200残基程度まで
水素原子	座標データに含まれない	座標データに含まれる
欠失原子	あり	通常なし
モデル数	通常1つ(部分的に複数)	複数
精度の指標	分解能 <sup>注</sup>	モデル構造のばらつき
原子の分布	温度因子	モデル構造のばらつき

注: X線結晶構造解析法ではどれだけ回折像を用いたかによって決まる分解能 (resolution) が全体の精度の指標。2.0~2.5 Åが普通、1.5 Å以下だと高分解能。

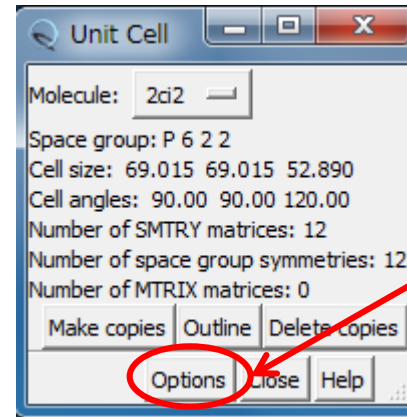
# 結晶構造の再現(1)

- 結晶中では、タンパク質分子が規則正しく並んでいる
- PDBに登録されている座標は、繰り返しの最小単位 (asymmetric unit; 非対称単位)
- 隣接したタンパク質分子間で相互作用していることがわかる

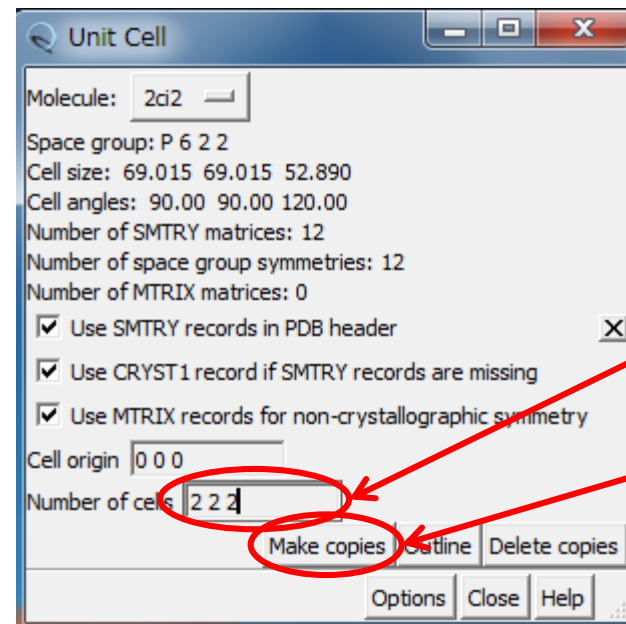


# 結晶構造の再現(2)

- メニューの「File」→「Fetch by ID」を選択し、PDB IDに2CI2を指定して「Fetch」
- 結晶構造の再現には、「Tools」→「High-Order Structure」→「Unit Cell」で右図のように指定する



ここをクリックし  
以下の表示に  
する



変更  
ここを  
クリック

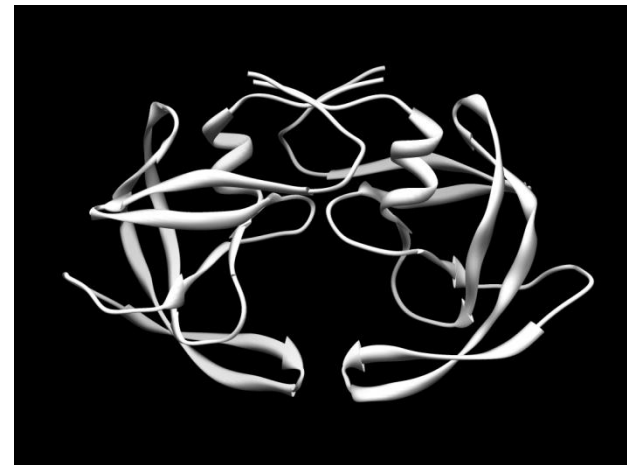
# Biological assembly (unit) (1)

- 生物学的に機能しうる最小限の分子構成をbiological assembly (unit)と呼ぶ
- 分子の対称性が、結晶の対称性と偶然一致すると、非対称単位には多量体の一部しか含まれない場合がある
- RCSBのサイトではbiological assemblyの座標がダウンロードできる



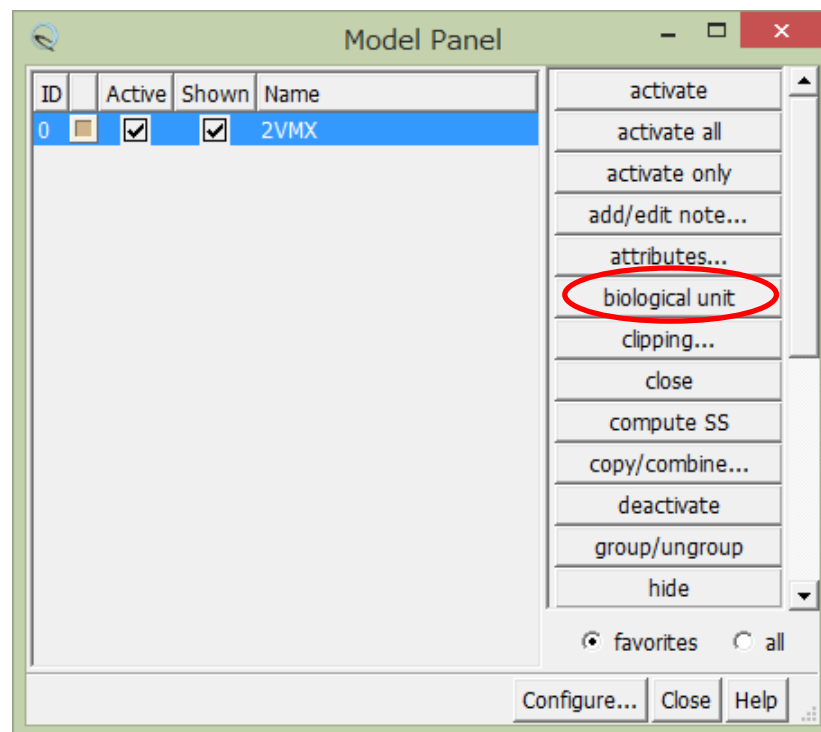
3PHVに登録されている座標

Biological assemblyの座標



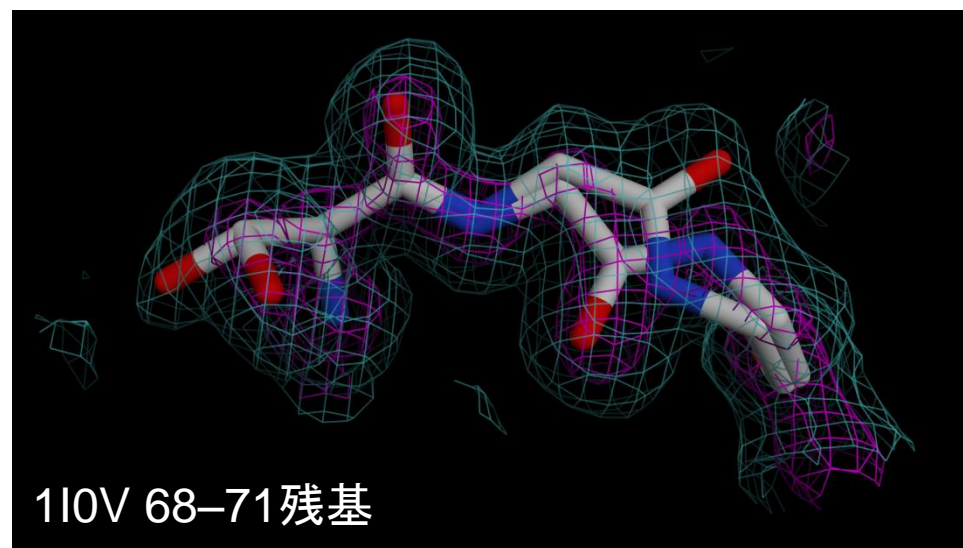
# Biological assembly (unit) (2)

- Chimeraでの操作法
  1. メニューの「File」→「Fetch by ID」を選択し、PDB IDに2VMXを指定して「Fetch」
  2. 「Favorites」→「Model Panel」を選択
  3. 「biological unit」をクリック



# Alternative conformation

- 結晶には異なるコンフォメーションを持つ複数の構造が含まれる可能性がある
- このような場合、X線回折データから得られる電子密度図では、それらの構造が存在割合に応じて複数見えることになる
- PDBファイルでは、occupancyに1より小さい重みを与え、同じ名前の原子を複数の座標で表す
- この時、原子名の残基名の間(17文字目)に、コンフォメーションを区別するIDを記入する



ATOM	548	N	AGLY	A	70	8.699	28.734	14.638	0.75	16.18
ATOM	549	N	BGLY	A	70	8.755	28.829	14.563	0.25	15.67
ATOM	550	CA	AGLY	A	70	9.857	29.561	14.390	0.75	16.18
ATOM	551	CA	BGLY	A	70	9.772	29.792	14.136	0.25	18.27
ATOM	552	C	AGLY	A	70	10.666	29.621	15.667	0.75	11.67
ATOM	553	C	BGLY	A	70	11.119	30.042	14.811	0.25	15.91
ATOM	554	O	AGLY	A	70	10.224	29.152	16.720	0.75	15.70
ATOM	555	O	BGLY	A	70	12.083	30.400	14.131	0.25	22.24



Alternate location  
indicator

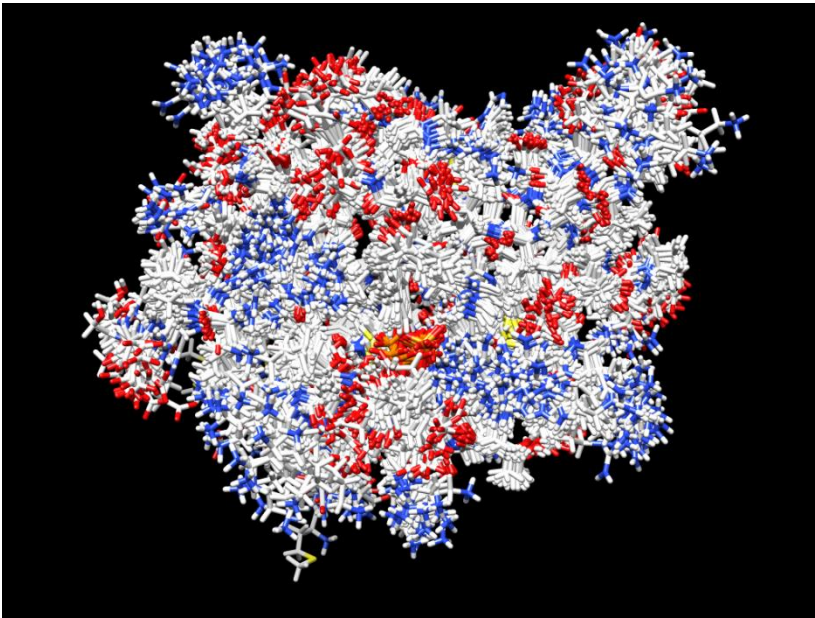


Occupancy  
32

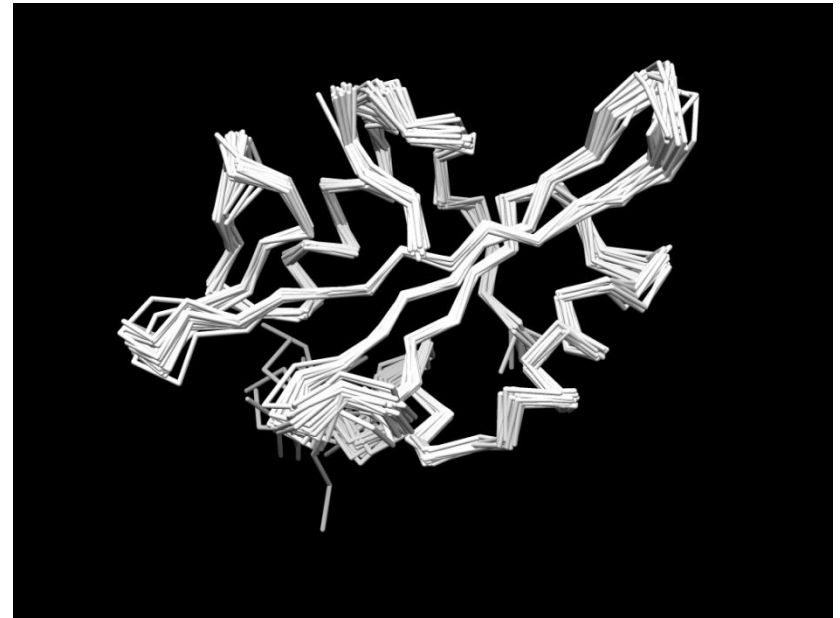


# NMR構造

- ヒトSrc SH2ドメインと基質ペプチドの複合体の立体構造、1HCTを開く



「Actions」→「Atoms/Bonds」→「show」  
「Actions」→「Ribbon」→「hide」



「Actions」→「Atoms/Bonds」→  
「backbone only」→「chain trace」

# NMR構造の特徴

- 立体構造が複数のモデルの重ね合わせで表現される
- モデル構造のばらつきを精度の指標とする
  - モデル構造の平均構造からのばらつき  
RMSD (root-mean-square deviation)

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^N |\mathbf{r}_i - \langle \mathbf{r}_i \rangle|^2}$$

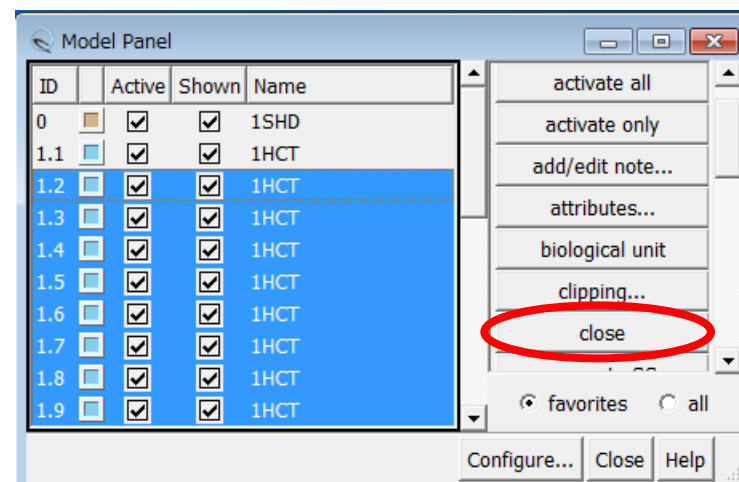
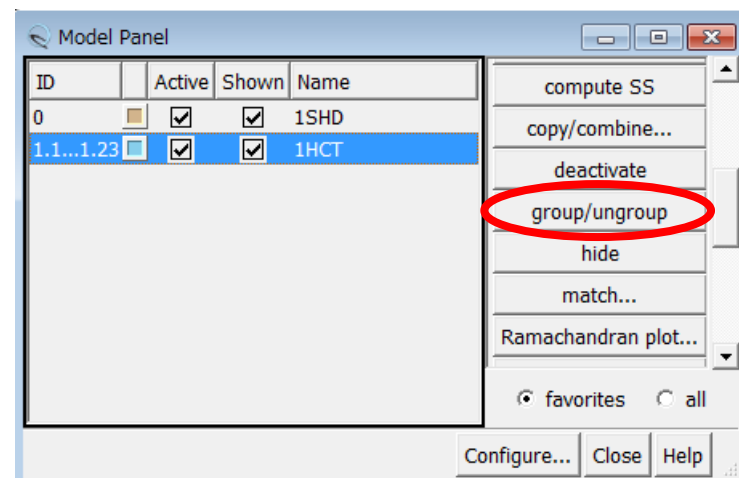
$\mathbf{r}_i$ : 原子*i*の座標

$\langle \mathbf{r}_i \rangle$ : 平均構造の原子*i*の座標

# 構造の比較(1)

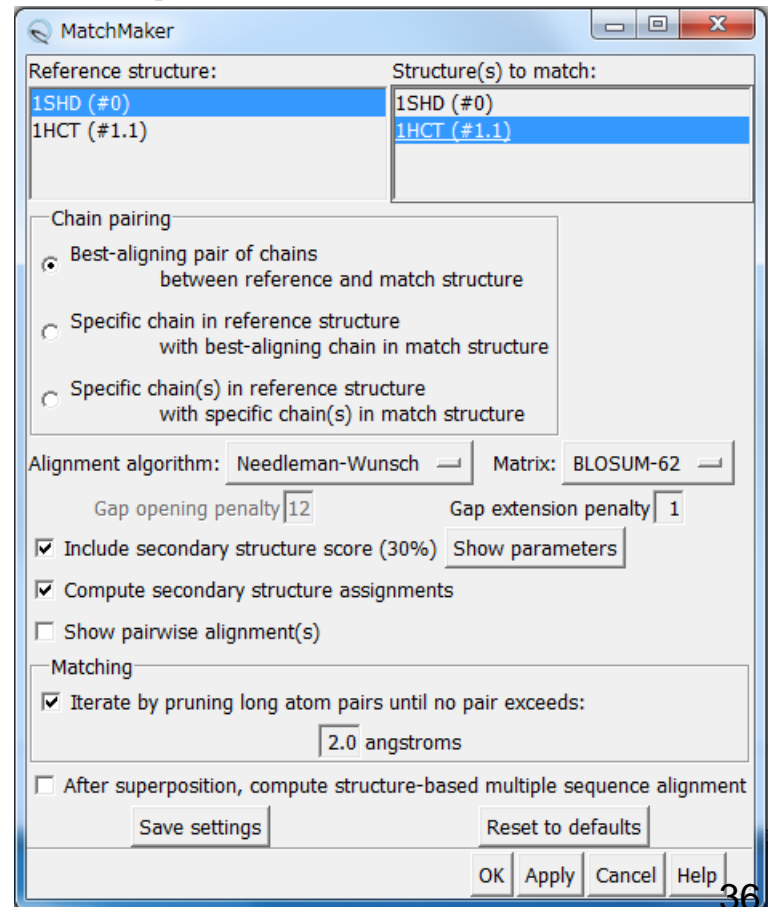
- ヒトSrc SH2ドメインと基質ペプチドの複合体についてX線構造とNMR構造を比較する

1. 「File」→「Fetch by ID」で1SHDを開く
2. 同様に1HCTを開く
3. 「Favorite」→「Model Panel」を開く
4. 右上図のように、1HCTの行をクリックして選択したのち、「group/ungroup」をクリックして展開
5. 右下図のようにID 1.2の行をクリックして選択したのち、「Shift」キーを押しながらID 1.23をクリック
6. 「close」をクリックしてこれらの構造を閉じる

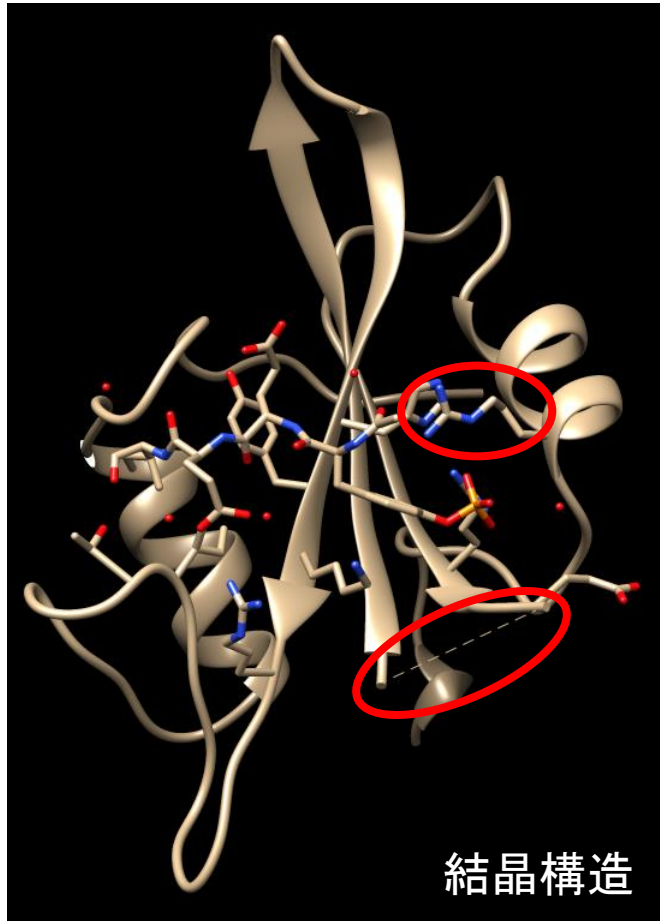


# 構造の比較(2)

7. 「Tools」→「Structure Comparison」→「MatchMaker」を選択
8. 右図のように設定し「OK」
9. ChimeraのWindowの下部に重ね合わせに使われた残基数(90残基)とRMSD(0.933 Å)が表示される

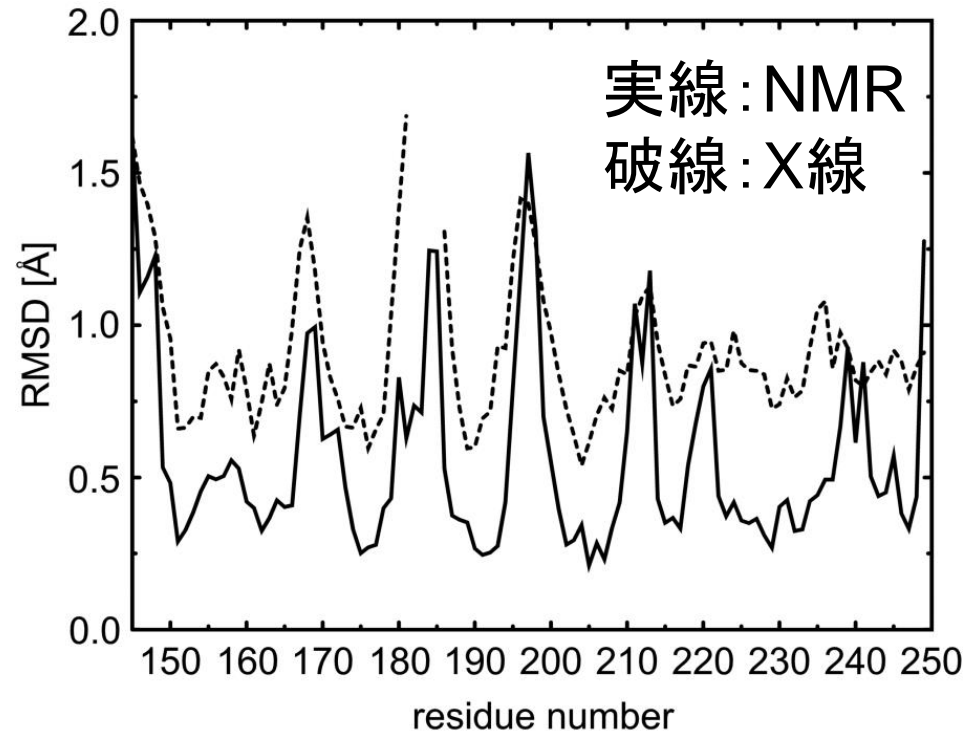


# 構造の比較(3)



# 構造の比較(4)

- NMR構造については、各モデルのC $\alpha$ 原子の平均構造からのずれの平均値(RMSD)
- X線構造では温度因子から換算
  - $B = 8\pi^2/3 (\Delta r)^2$
  - $B = 30$ で $\Delta r = 1.07 \text{ \AA}$
- 温度因子が大きい残基は、NMRでも構造のばらつきが大きい傾向



# 配列データベースとの連携

---

- 配列データベースへのリンク
  - RCSBの検索結果のSequenceタブ
- 配列データベースからのリンク
- 配列からの検索

# 配列データベースからのリンク

---

1. タンパク質配列データベースUniProt (<http://www.uniprot.org/>)を開く
2. QueryにSRC\_HUMANと入力し「Search」
3. 検索結果の下のほうに、“3D structure databases”のセクションがあり、1HCTや1SHDが現れていることを確認すること



# 配列からの検索(1)

---

1. NCBI BLASTのサイトにアクセス  
(<http://blast.ncbi.nlm.nih.gov/Blast.cgi>)
2. Basic Blastにある「protein blast」をクリック
3. 講義のページで1HCT\_B.fastaをクリック
4. 右クリックして、「すべて選択」を選んだあと、再び右クリックして、「コピー」
5. BLASTのページの「Enter accession number(s), gi(s), or FASTA sequence(s)」のテキストエリアの中で右クリックし、「貼り付け」

# 配列からの検索(2)

6. Choose Search SetのDatabaseを「Protein Data Bank proteins (pdb)」に設定
7. BLASTをクリック

The screenshot shows the NCBI BLAST Standard Protein BLAST interface. The 'Database' dropdown menu is highlighted in yellow and set to 'Protein Data Bank proteins(pdb)'. A red arrow points to this dropdown. The 'Job Title' field contains '1HCT.B|PDBID|CHAIN|SEQUENCE'. The 'Program Selection' section shows 'blastp (protein-protein BLAST)' selected. The footer indicates 'Search database Protein Data Bank proteins(pdb) using Blastp (protein-protein BLAST)'.

# 実習課題2

---

1. 講義のページで、kadai.fastaを表示し、この配列をもつタンパク質の立体構造データを検索せよ
2. 配列一致度100%のヒットのPDB IDを用いてRCSBのサイトで検索せよ
  - タンパク質名、立体構造決定の方法を確認する
3. PDBファイルをChimeraで開き表示せよ
4. 全体像をpng形式で保存せよ

# 課題の提出

---

- 課題1で保存した画像をPowerPointのスライドに貼り付け、発色団の位置を赤いマルで囲んで示せ
- 同じPowerPointファイルの別のスライドに課題2の全体像を貼り付け、PDB IDとタンパク質名、立体構造決定の方法を記入せよ
- PowerPointファイルはメールに添付して寺田宛 (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