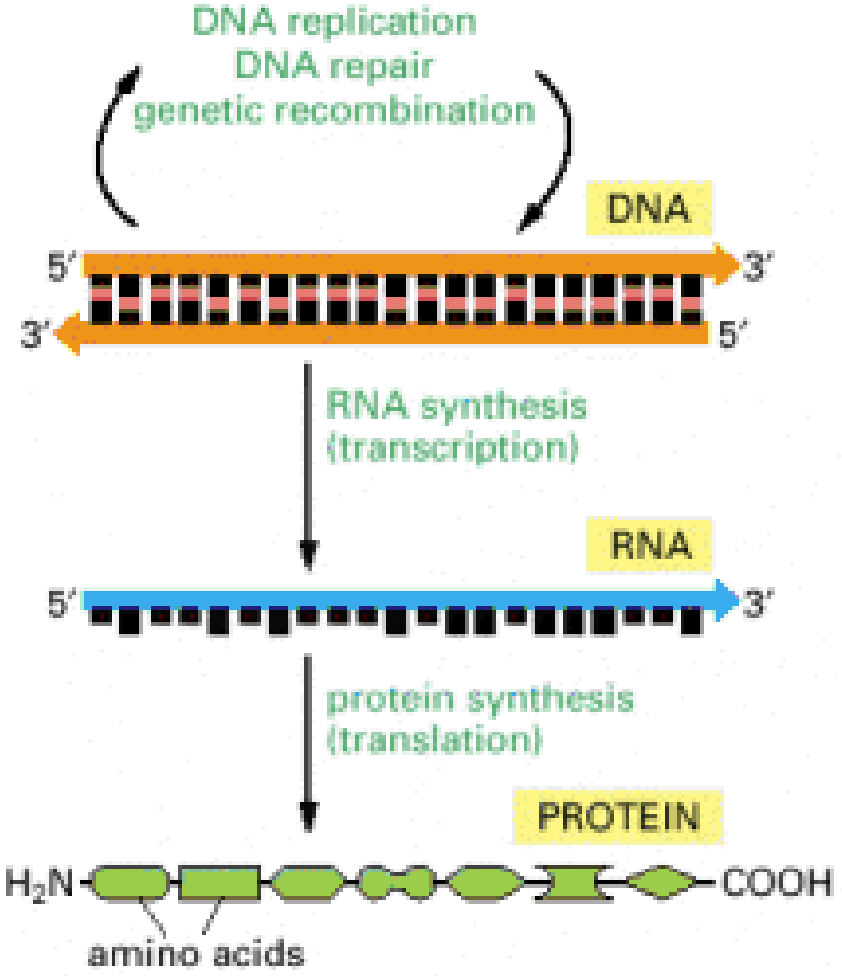
蛋白質立体構造データベースとその検索

きんじょう　あきら
金城　玲
大阪大学蛋白質研究所
プロテオミクス総合研究センター
および
情報・システム研究機構
ライフサイエンス統合データベースセンター

第1部

蛋白質とは？Protein Data Bank (PDB)とは？

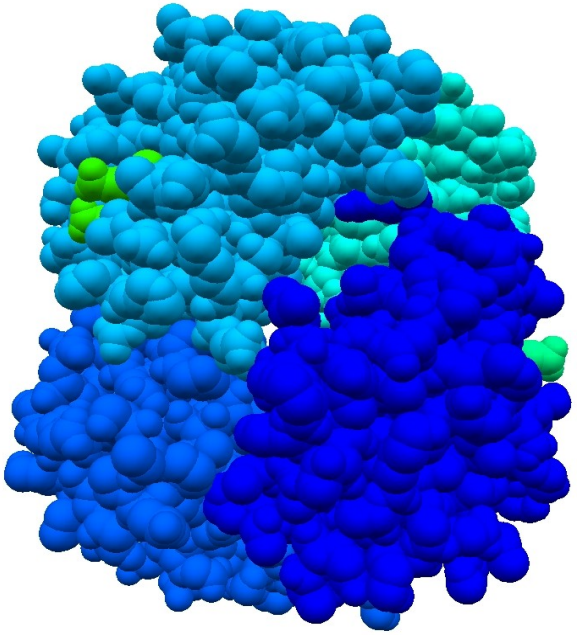
蛋白質入門

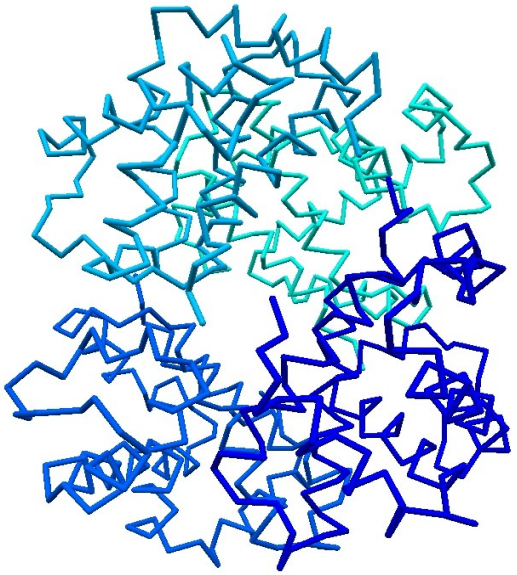


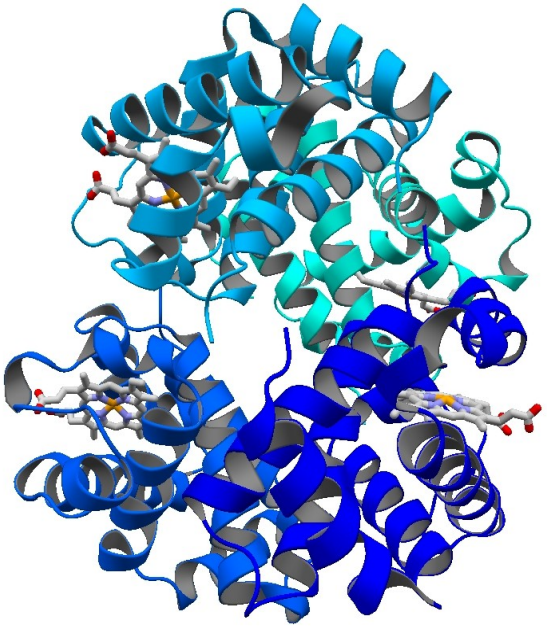
- 遺伝子のDNA配列がmRNAを介して（転写）アミノ酸配列に翻訳されたものが蛋白質。
- アミノ酸は20種類。
- アミノ酸がペプチド結合で線形につながったものがポリペプチド。
- 遺伝子配列に対応するポリペプチドが蛋白質。
- アミノ酸配列の長さは数十から数千残基。
- “Central Dogma”

Alberts *et al.*, *Molecular Biology of the Cell (4/e)*

蛋白質の構造入門

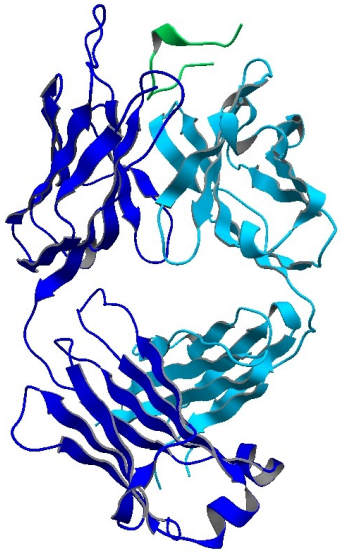


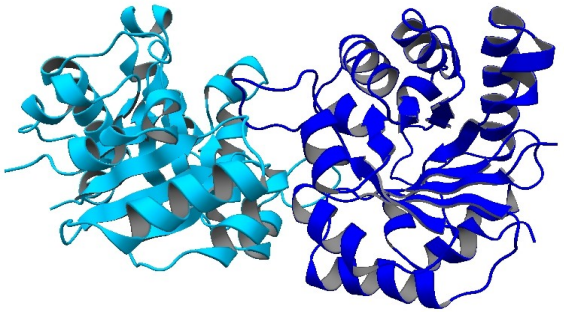


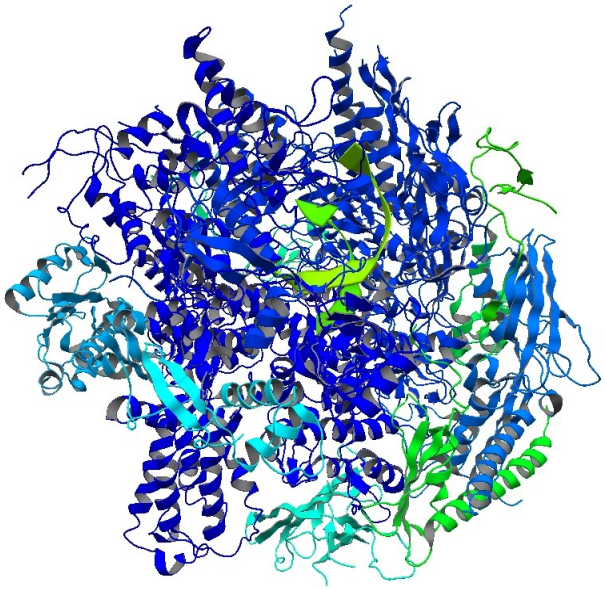


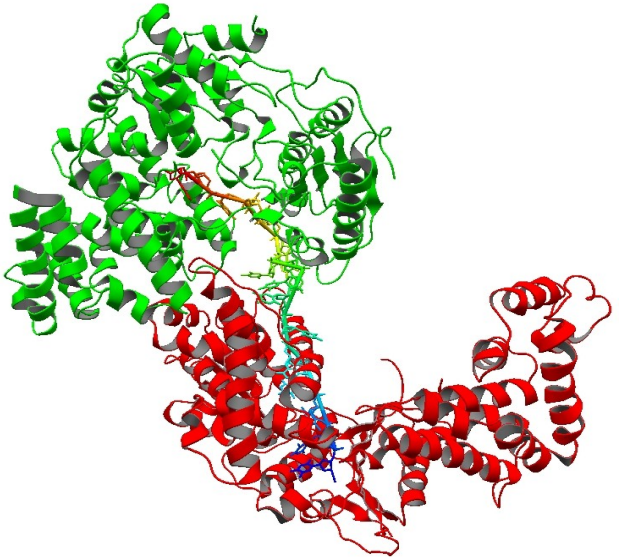
- 特定のアミノ酸配列をもつ蛋白質は生理条件下で特定の立体構造(天然構造)をとる。
- Anfinsenのドグマ:天然構造は熱力学的な平衡状態(アミノ酸配列は「境界条件」)。
- 特定の機能(酵素活性など)は特定の構造が担う。

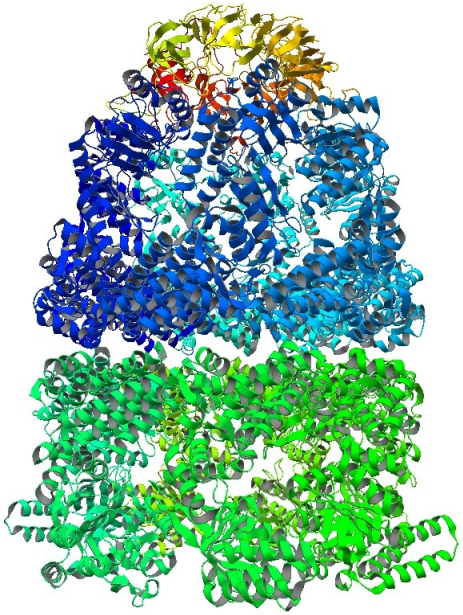
蛋白質の構造入門(2)











蛋白質の機能と構造は直結している。

Protein Data Bank

- 実験的に決定された蛋白質の立体構造情報を集積したデータバンク。
- 1970年代初頭に米国Brookhavenで活動開始。
- 2003年より world-wide PDB (wwPDB) として、米国(RCSB)、欧州(PDBe)、日本(PDBj)の三極体制で運営。
- PDBj (Protein Data Bank Japan)は大阪大学蛋白質研究所で運営されている。
- 蛋白質構造決定の論文を出版する際にはPDBへの登録が必須。
- PDBのデータはすべて無料で公開されている。

PDB入門

- 「エントリー」の概念：一つの実験で決定された構造が一つのエントリーに対応する。
- 各エントリーの情報は一つのファイルに記述される。
- ファイルフォーマットは3種類：
  - PDBフォーマット（行指向のフラットファイル）
  - mmCIF (国際結晶連合の規格 CIF を拡張したもの）
  - PDBML（mmCIFを XML形式に焼き直したもの）
- wwPDB内部では、mmCIFを基本にして他の形式に変換して公開している。

蛋白質の立体構造決定法

- 現在PDBに収められている構造は右のいずれかの手法で決定されたもの。
- 多様な実験手法に統一的に対応する必要がある。
- 基本は、アノテーションと原子座標。

1. X-ray diffraction
2. Neutron diffraction
3. Fiber diffraction
4. Electron crystallography
5. Electron microscopy
6. Solution NMR
7. Solid-state NMR
8. Solution scattering
9. Powder diffraction
10. Infrared spectroscopy

データの実際：PDBフォーマット

```
HEADER    OXIDOREDUCTASE(OXYGEN(A))                30-SEP-93   1GOF
TITLE     NOVEL THIOETHER BOND REVEALED BY A 1.7 ANGSTROMS CRYSTAL
TITLE    2 STRUCTURE OF GALACTOSE OXIDASE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: GALACTOSE OXIDASE;
COMPND   3 CHAIN: A;
COMPND   4 EC: 1.1.3.9;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HYPOMYCES ROSELLUS;
SOURCE   3 ORGANISM_TAXID: 5132
KEYWDS    OXIDOREDUCTASE(OXYGEN(A))
EXPDTA    X-RAY DIFFRACTION
AUTHOR    N.ITO,S.E.V.PHILLIPS,P.F.KNOWLES
REVDAT   3   24-FEB-09 1GOF    1       VERSN
REVDAT   2   01-APR-03 1GOF    1       JRNL
REVDAT   1   31-JAN-94 1GOF    0
JRNL        AUTH   N.ITO,S.E.PHILLIPS,C.STEVENS,Z.B.OGEL,
JRNL        AUTH 2 M.J.MCPHERSON,J.N.KEEN,K.D.YADAV,P.F.KNOWLES
JRNL        TITL   NOVEL THIOETHER BOND REVEALED BY A 1.7 A CRYSTAL
JRNL        TITL 2 STRUCTURE OF GALACTOSE OXIDASE.
JRNL        REF    NATURE                         V. 350    87 1991
JRNL        REFN                   ISSN 0028-0836
JRNL        PMID   2002850
JRNL        DOI    10.1038/350087A0
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   N.ITO,S.E.V.PHILLIPS,K.K.S.YADAV,P.F.KNOWLES
REMARK   1  TITL   THE CRYSTAL STRUCTURE OF A FREE RADICAL ENZYME,
REMARK   1  TITL 2 GALACTOSE OXIDASE
REMARK   1  REF    TO BE PUBLISHED
REMARK   1  REFN
REMARK   1 REFERENCE 2
REMARK   1  AUTH   M.J.MCPHERSON,Z.B.OGEL,C.STEVENS,K.D.S.YADAV,
REMARK   1  AUTH 2 J.M.KEEN,P.F.KNOWLES
REMARK   1  TITL   GALACTOSE OXIDASE OF DACTYLIUM DENDROIDES: GENE
REMARK   1  TITL 2 CLONING AND SEQUENCE ANALYSIS
REMARK   1  REF    J.BIOL.CHEM.                   V. 267  8146 1992
REMARK   1  REFN                   ISSN 0021-9258
REMARK   2
REMARK   2 RESOLUTION.    1.70 ANGSTROMS.
```

PDBフォーマット（つづき）

```
ATOM      1  N   ALA A   1      38.840   0.236   1.012  1.00 34.65           N
ATOM      2  CA  ALA A   1      38.356  -0.999   0.357  1.00 42.26           C
ATOM      3  C   ALA A   1      37.098  -1.547   1.056  1.00 41.25           C
ATOM      4  O   ALA A   1      36.619  -0.946   2.028  1.00 29.44           O
ATOM      5  CB  ALA A   1      39.398  -2.114   0.379  1.00 40.70           C
ATOM      6  N   SER A   2      36.610  -2.666   0.495  1.00 32.67           N
ATOM      7  CA  SER A   2      35.411  -3.244   1.202  1.00 34.90           C
ATOM      8  C   SER A   2      35.683  -4.740   1.081  1.00 38.30           C
ATOM      9  O   SER A   2      36.827  -5.147   0.747  1.00 28.59           O
ATOM     10  CB  SER A   2      34.063  -2.660   0.823  1.00 24.49           C
ATOM     11  OG  SER A   2      33.031  -3.308   1.686  1.00 20.37           O
ATOM     12  N   ALA A   3      34.660  -5.537   1.334  1.00 35.91           N
ATOM     13  CA  ALA A   3      34.815  -6.995   1.246  1.00 33.38           C
ATOM     14  C   ALA A   3      33.416  -7.594   1.259  1.00 21.71           C
ATOM     15  O   ALA A   3      32.529  -6.833   1.679  1.00 30.82           O
ATOM     16  CB  ALA A   3      35.687  -7.414   2.433  1.00 25.44           C
ATOM     17  N   PRO A   4      33.335  -8.786   0.733  1.00 36.18           N
ATOM     18  CA  PRO A   4      32.068  -9.552   0.674  1.00 40.82           C
ATOM     19  C   PRO A   4      32.067 -10.418   1.934  1.00 37.76           C
ATOM     20  O   PRO A   4      33.145 -10.698   2.466  1.00 42.84           O
ATOM     21  CB  PRO A   4      32.222 -10.479  -0.536  1.00 40.12           C
ATOM     22  CG  PRO A   4      33.729 -10.691  -0.634  1.00 48.00           C
ATOM     23  CD  PRO A   4      34.452  -9.579   0.148  1.00 34.36           C
```

mmCIF

```
data_1GOF
#
_entry.id   1GOF
#
_audit_conform.dict_name       mmcif_pdbx.dic
_audit_conform.dict_version    1.0670
_audit_conform.dict_location   http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic
#
_database_2.database_id     PDB
_database_2.database_code   1GOF
#
loop_
_database_PDB_rev.num
_database_PDB_rev.date
_database_PDB_rev.date_original
_database_PDB_rev.status
_database_PDB_rev.replaces
_database_PDB_rev.mod_type
1 1994-01-31 1993-09-30 ? 1GOF 0
2 2003-04-01 ?          ? 1GOF 1
3 2009-02-24 ?          ? 1GOF 1
#
loop_
_database_PDB_rev_record.rev_num
_database_PDB_rev_record.type
_database_PDB_rev_record.details
2 JRNL  ?
3 VERSN ?
#
_pdbx_database_status.status_code    REL
_pdbx_database_status.entry_id       1GOF
_pdbx_database_status.deposit_site   ?
_pdbx_database_status.process_site   ?
_pdbx_database_status.SG_entry       .
#
loop_
_audit_author.name
_audit_author.pdbx_ordinal
'Ito, N.'          1
'Phillips, S.E.V.' 2
'Knowles, P.F.'    3
#
```

mmCIF (つづき)

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.Cartn_x_esd
_atom_site.Cartn_y_esd
_atom_site.Cartn_z_esd
_atom_site.occupancy_esd
_atom_site.B_iso_or_equiv_esd
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . ALA A 1 1   ? 38.840  0.236   1.012   1.00 34.65 ? ? ? ? ? ? 1    ALA A N   1
ATOM   2    C  CA  . ALA A 1 1   ? 38.356  -0.999  0.357   1.00 42.26 ? ? ? ? ? ? 1    ALA A CA  1
ATOM   3    C  C   . ALA A 1 1   ? 37.098  -1.547  1.056   1.00 41.25 ? ? ? ? ? ? 1    ALA A C   1
ATOM   4    O  O   . ALA A 1 1   ? 36.619  -0.946  2.028   1.00 29.44 ? ? ? ? ? ? 1    ALA A O   1
ATOM   5    C  CB  . ALA A 1 1   ? 39.398  -2.114  0.379   1.00 40.70 ? ? ? ? ? ? 1    ALA A CB  1
ATOM   6    N  N   . SER A 1 2   ? 36.610  -2.666  0.495   1.00 32.67 ? ? ? ? ? ? 2    SER A N   1
ATOM   7    C  CA  . SER A 1 2   ? 35.411  -3.244  1.202   1.00 34.90 ? ? ? ? ? ? 2    SER A CA  1
ATOM   8    C  C   . SER A 1 2   ? 35.683  -4.740  1.081   1.00 38.30 ? ? ? ? ? ? 2    SER A C   1
ATOM   9    O  O   . SER A 1 2   ? 36.827  -5.147  0.747   1.00 28.59 ? ? ? ? ? ? 2    SER A O   1
ATOM   10   C  CB  . SER A 1 2   ? 34.063  -2.660  0.823   1.00 24.49 ? ? ? ? ? ? 2    SER A CB  1
ATOM   11   O  OG  . SER A 1 2   ? 33.031  -3.308  1.686   1.00 20.37 ? ? ? ? ? ? 2    SER A OG  1
ATOM   12   N  N   . ALA A 1 3   ? 34.660  -5.537  1.334   1.00 35.91 ? ? ? ? ? ? 3    ALA A N   1
ATOM   13   C  CA  . ALA A 1 3   ? 34.815  -6.995  1.246   1.00 33.38 ? ? ? ? ? ? 3    ALA A CA  1
ATOM   14   C  C   . ALA A 1 3   ? 33.416  -7.594  1.259   1.00 21.71 ? ? ? ? ? ? 3    ALA A C   1
ATOM   15   O  O   . ALA A 1 3   ? 32.529  -6.833  1.679   1.00 30.82 ? ? ? ? ? ? 3    ALA A O   1
ATOM   16   C  CB  . ALA A 1 3   ? 35.687  -7.414  2.433   1.00 25.44 ? ? ? ? ? ? 3    ALA A CB  1
ATOM   17   N  N   . PRO A 1 4   ? 33.335  -8.786  0.733   1.00 36.18 ? ? ? ? ? ? 4    PRO A N   1
ATOM   18   C  CA  . PRO A 1 4   ? 32.068  -9.552  0.674   1.00 40.82 ? ? ? ? ? ? 4    PRO A CA  1
ATOM   19   C  C   . PRO A 1 4   ? 32.067  -10.418 1.934   1.00 37.76 ? ? ? ? ? ? 4    PRO A C   1
ATOM   20   O  O   . PRO A 1 4   ? 33.145  -10.698 2.466   1.00 42.84 ? ? ? ? ? ? 4    PRO A O   1
ATOM   21   C  CB  . PRO A 1 4   ? 32.222  -10.479 -0.536  1.00 40.12 ? ? ? ? ? ? 4    PRO A CB  1
ATOM   22   C  CG  . PRO A 1 4   ? 33.729  -10.691 -0.634  1.00 48.00 ? ? ? ? ? ? 4    PRO A CG  1
ATOM   23   C  CD  . PRO A 1 4   ? 34.452  -9.579  0.148   1.00 34.36 ? ? ? ? ? ? 4    PRO A CD  1
```

PDBML

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock datablockName="1GOF-noatom"
  xmlns:PDBx="http://pdbml.pdb.org/schema/pdbx-v32.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://pdbml.pdb.org/schema/pdbx-v32.xsd pdbx-v32.xsd">
  <PDBx:atom_sitesCategory>
    <PDBx:atom_sites entry_id="1GOF">
      <PDBx:Cartn_transform_axes xsi:nil="true" />
      <PDBx:fract_transf_matrix11>0.011535</PDBx:fract_transf_matrix11>
      <PDBx:fract_transf_matrix12>0.000000</PDBx:fract_transf_matrix12>
      <PDBx:fract_transf_matrix13>0.000000</PDBx:fract_transf_matrix13>
      <PDBx:fract_transf_matrix21>0.000000</PDBx:fract_transf_matrix21>
      <PDBx:fract_transf_matrix22>0.011186</PDBx:fract_transf_matrix22>
      <PDBx:fract_transf_matrix23>0.000000</PDBx:fract_transf_matrix23>
      <PDBx:fract_transf_matrix31>0.006081</PDBx:fract_transf_matrix31>
      <PDBx:fract_transf_matrix32>0.000000</PDBx:fract_transf_matrix32>
      <PDBx:fract_transf_matrix33>0.011534</PDBx:fract_transf_matrix33>
      <PDBx:fract_transf_vector1>0.00000</PDBx:fract_transf_vector1>
      <PDBx:fract_transf_vector2>0.00000</PDBx:fract_transf_vector2>
      <PDBx:fract_transf_vector3>0.00000</PDBx:fract_transf_vector3>
    </PDBx:atom_sites>
  </PDBx:atom_sitesCategory>
  <PDBx:atom_sites_footnoteCategory>
    <PDBx:atom_sites_footnote id="1">
      <PDBx:text>CIS PROLINE - PRO      52</PDBx:text>
    </PDBx:atom_sites_footnote>
    <PDBx:atom_sites_footnote id="2">
      <PDBx:text>CIS PROLINE - PRO     163</PDBx:text>
    </PDBx:atom_sites_footnote>
    <PDBx:atom_sites_footnote id="3">
      <PDBx:text>CIS PROLINE - PRO     350</PDBx:text>
    </PDBx:atom_sites_footnote>
  </PDBx:atom_sites_footnoteCategory>
```

PDBML (つづき)

```
<PDBx:atom_siteCategory>
  <PDBx:atom_site id="1">
    <PDBx:B_iso_or_equiv>34.65</PDBx:B_iso_or_equiv>
    <PDBx:B_iso_or_equiv_esd xsi:nil="true" />
    <PDBx:Cartn_x>38.840</PDBx:Cartn_x>
    <PDBx:Cartn_x_esd xsi:nil="true" />
    <PDBx:Cartn_y>0.236</PDBx:Cartn_y>
    <PDBx:Cartn_y_esd xsi:nil="true" />
    <PDBx:Cartn_z>1.012</PDBx:Cartn_z>
    <PDBx:Cartn_z_esd xsi:nil="true" />
    <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
    <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
    <PDBx:auth_comp_id>ALA</PDBx:auth_comp_id>
    <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
    <PDBx:group_PDB>ATOM</PDBx:group_PDB>
    <PDBx:label_alt_id></PDBx:label_alt_id>
    <PDBx:label_asym_id>A</PDBx:label_asym_id>
    <PDBx:label_atom_id>N</PDBx:label_atom_id>
    <PDBx:label_comp_id>ALA</PDBx:label_comp_id>
    <PDBx:label_entity_id>1</PDBx:label_entity_id>
    <PDBx:label_seq_id>1</PDBx:label_seq_id>
    <PDBx:occupancy>1.00</PDBx:occupancy>
    <PDBx:occupancy_esd xsi:nil="true" />
    <PDBx:pdbx_PDB_ins_code xsi:nil="true" />
    <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
    <PDBx:pdbx_formal_charge xsi:nil="true" />
    <PDBx:type_symbol>N</PDBx:type_symbol>
  </PDBx:atom_site>
```

```
  <PDBx:atom_site id="2">
    <PDBx:B_iso_or_equiv>42.26</PDBx:B_iso_or_equiv>
    <PDBx:B_iso_or_equiv_esd xsi:nil="true" />
    <PDBx:Cartn_x>38.356</PDBx:Cartn_x>
    <PDBx:Cartn_x_esd xsi:nil="true" />
    <PDBx:Cartn_y>-0.999</PDBx:Cartn_y>
    <PDBx:Cartn_y_esd xsi:nil="true" />
    <PDBx:Cartn_z>0.357</PDBx:Cartn_z>
    <PDBx:Cartn_z_esd xsi:nil="true" />
    <PDBx:auth_asym_id>A</PDBx:auth_asym_id>
    <PDBx:auth_atom_id>CA</PDBx:auth_atom_id>
    <PDBx:auth_comp_id>ALA</PDBx:auth_comp_id>
    <PDBx:auth_seq_id>1</PDBx:auth_seq_id>
    <PDBx:group_PDB>ATOM</PDBx:group_PDB>
    <PDBx:label_alt_id></PDBx:label_alt_id>
    <PDBx:label_asym_id>A</PDBx:label_asym_id>
    <PDBx:label_atom_id>CA</PDBx:label_atom_id>
    <PDBx:label_comp_id>ALA</PDBx:label_comp_id>
    <PDBx:label_entity_id>1</PDBx:label_entity_id>
    <PDBx:label_seq_id>1</PDBx:label_seq_id>
    <PDBx:occupancy>1.00</PDBx:occupancy>
    <PDBx:occupancy_esd xsi:nil="true" />
    <PDBx:pdbx_PDB_ins_code xsi:nil="true" />
    <PDBx:pdbx_PDB_model_num>1</PDBx:pdbx_PDB_model_num>
    <PDBx:pdbx_formal_charge xsi:nil="true" />
    <PDBx:type_symbol>C</PDBx:type_symbol>
  </PDBx:atom_site>
```

第2部

PDBjのバックエンドデータベースの開発

PDBj@阪大蛋白研

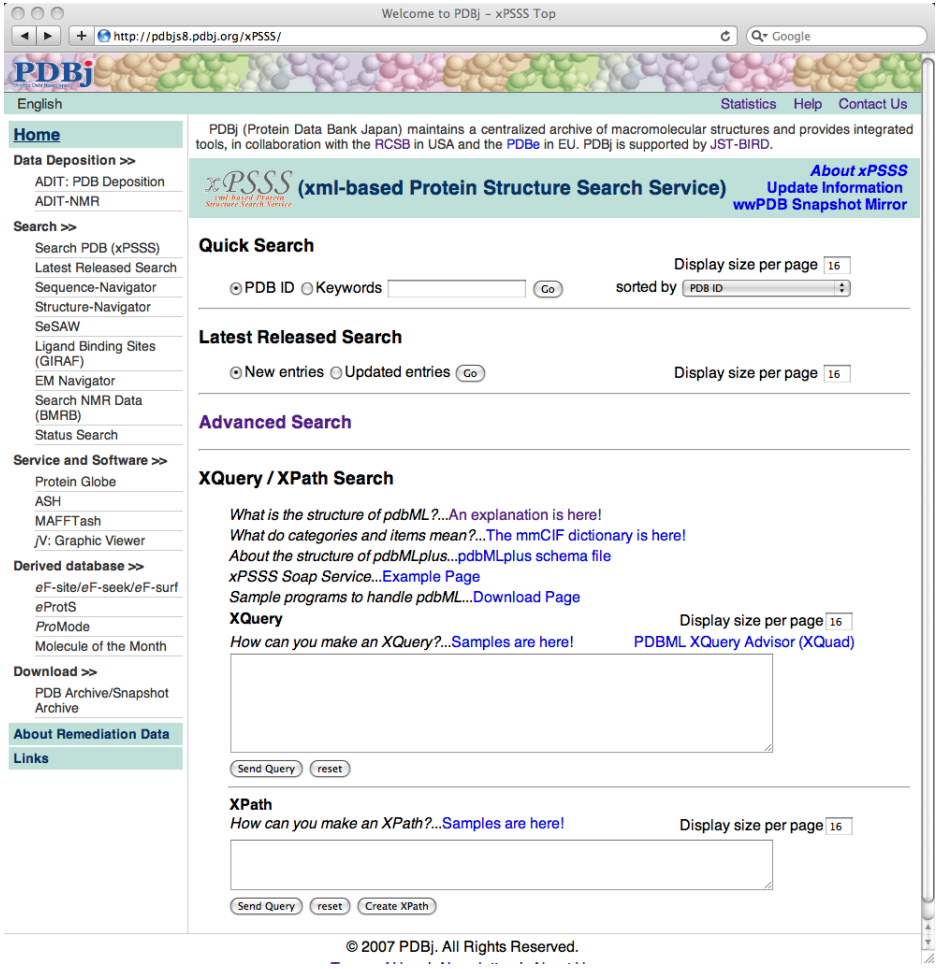
http://www.pdbj.org/

- データの登録作業
  - 実験家からの受付
  - データの査定
  - 公開
- 新規サービスの開発
  - 検索サービス
  - 二次データベース
  - アノテーション付加
- JST-BIRD事業の支援
- 統括責任者：中村春木教授
- 総勢約30名

基本の検索：キーワード、条件指定

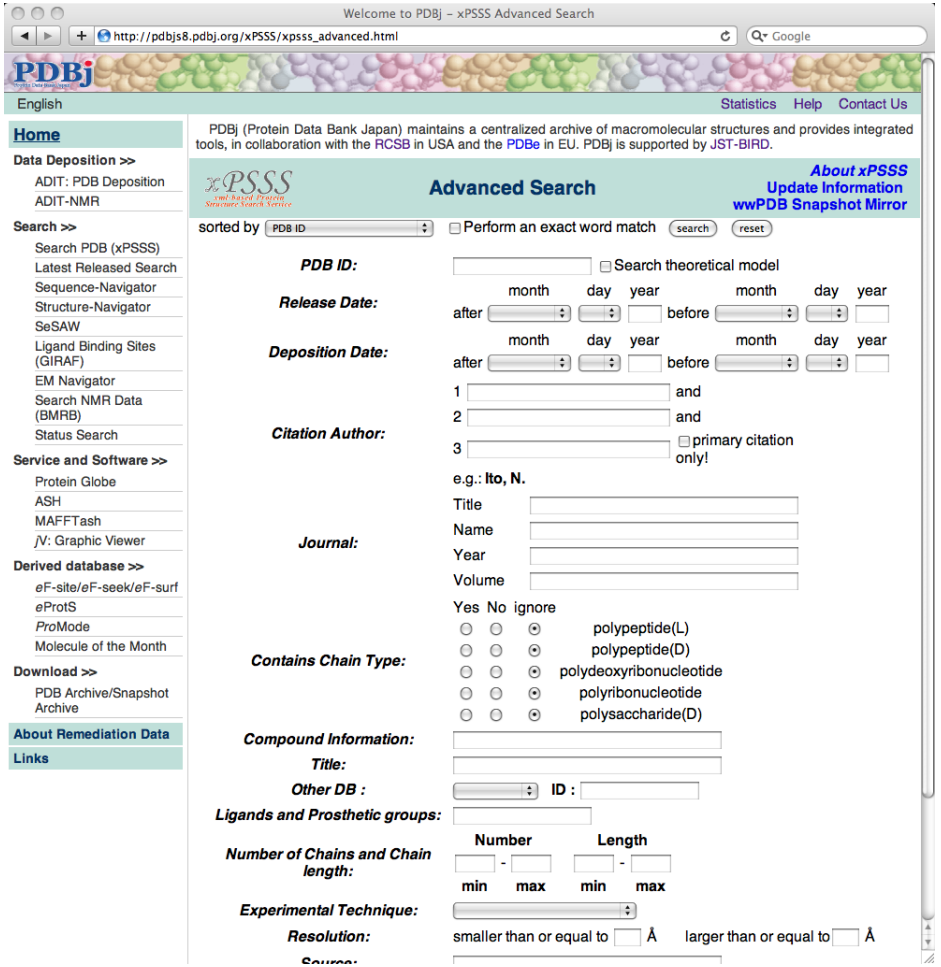
- PDBの公式なデータ配布形式は各エントリーのファイルのみ（2009年8月現在約6万）。
- しかしこれでは、検索には向かない。
  - 「ヒトのヘモグロビンα鎖はどのエントリー？」
  - 「蛋白質とDNAの複合体はどのエントリー？」
  - 「解像度1.5Å以上のX線結晶構造がほしい」
  - などなど

XML-based Protein Structure Search Service (xPSSS)



- PDBML (PDBのXML形式)をそのまま native XML DB に格納。
- 「ふつう」の検索はとくに問題ない。
- 「ややこしい」検索には時間がかかる。
- XQuery/XPathによる検索も直接できる。
  - が、誰も使わない(使えない)。

Advanced Search: 条件指定検索



- 「よく使う」項目のみに特化した検索。
  - 引用文献
  - ポリマーのタイプ
  - リガンド分子名
  - 実験手法
  - 公開日
  - ……
- キーワード検索

Native XML DBの問題点

- （商用）ライセンス料が高い。
  - PDBjの運営予算は限られている。
  - ミラーサイトが構築できない。
- （OSS)未成熟。PDBjの規模には使えない。
- （商用＆OSS）技術が枯れてない（？）
  - XQuery 1.0 は2007年1月にようやくW3C recommendationになったばかり。
  - 複雑な検索をするととたんに遅くなる。

PDBMLをRDBに格納する。

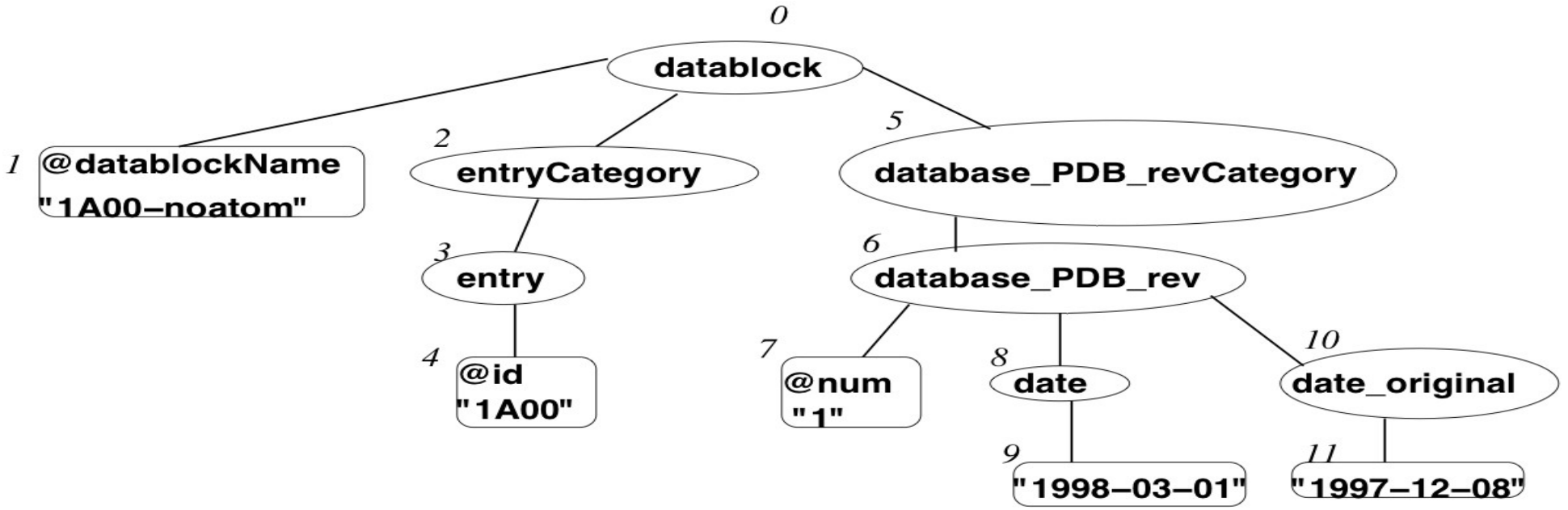
- PDBML Schema → RDB table
  - (XML→SQLコンパイラ)
- Xpath による階層構造の分解。
- 検索のためのサマリーテーブル。

しょうがないので自前で開発することにした。

なぜRDBか？

- PDBのデータ構造は複雑なので、単なる全文検索では意味をなさない。
- RDBは良質のOSSが複数存在する。
  - 今回はPostgreSQLを選択した。

XML文書の分解



(*c.f.* M. Yoshikawa *et al*. (2001) *ACM Trans. IT*, **1:**110–141)

- XML Schemaに基づいて、可能なXPathをすべて列挙する。
- Xpath ⇄テーブル名
- 各ドキュメントノードを深さ優先で番号（ポインタ）付けする。

テーブルの構造

```
CREATE TABLE “E:/datablock/entryCategory/entry” (
  Docid INT REFERENCES xmldoc(docid),  /* エントリーID */
  Pstart INT, /* エレメントの開始点（括弧ひらく）*/
  Pend INT,   /* エレメントの終了点（括弧とじる）*/
  PRIMARY KEY(docid,pstart,pend));

CREATE TABLE “/datablock/entryCategory/entry/@id” (
  Docid INT REFERENCES xmldoc(docid),
  Pos INT,  /* エレメントの位置（PCDataも属性も同様に扱う）*/
  Val TEXT, /* エレメントの値（PCDataまたは属性） */
  PRIMARY KEY(docid, pos));

 CREATE TABLE
  “/datablock/database_PDB_revCategory/database_PDB_rev/date” (
   Docid INT REFERENCES xmldoc(docid),
   Pos INT,
   Val DATE,
   PRIMARY KEY(docid,pos));
```

- このようなテーブルが約8，000個定義される。
- Advanced Search用にmaterialized view を定義する。
- 現在バックエンドはほぼ出来上がり、β版公開に向けてウェブIFを構築中。

检索例

```
<datablock datablockName=”101M”>
 ......
  <struct_refCategory>
    <struct_ref>
      <db_name>UNP</db_name>
      <db_code>P02185</db_code>
    </struct_ref>
  </struct_refCategory>
  ......
</datablock>
```

```
SELECT docid, p2.val
FROM "E:/datablock/struct_refCategory/struct_ref" e
  JOIN "/datablock/struct_refCategory/struct_ref/db_name" p1
        ON (p1.docid = e.docid AND p1.pos BETWEEN e.pstart AND e.pend)
  JOIN "/datablock/struct_refCategory/struct_ref/db_code" p2
        ON (p2.docid = e.docid AND p2.pos BETWEEN e.pstart AND e.pend)
  WHERE p1.val = 'UNP'
```

第3部

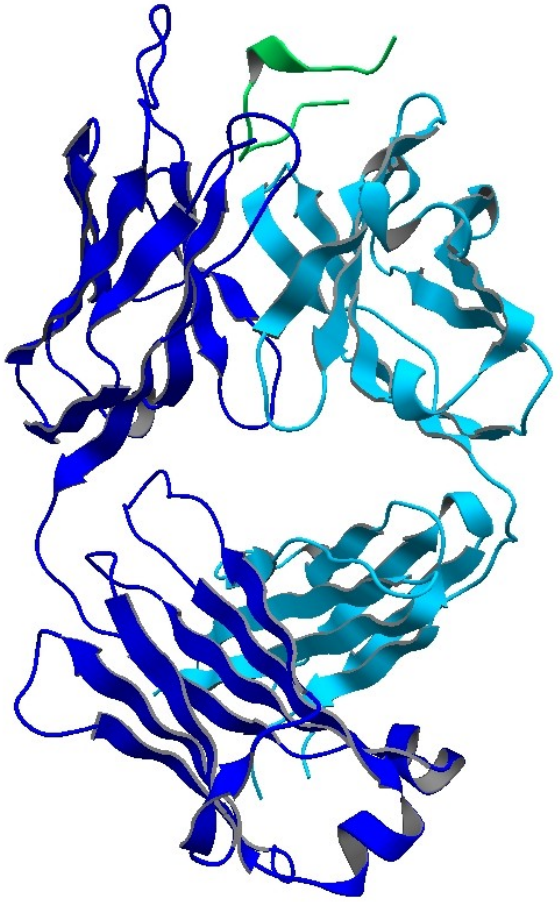
立体構造のかたちの検索

PDBの最重要情報は「構造そのもの」

- 構造決定は分子機能のメカニズムを詳細に理解するために行われる。
- PDBのエントリに付加されたアノテーションは「おまけ」にすぎない（？）
- 「構造そのもの」から機能を理解するのが本来の目的。
- では、「構造そのもの」とは何か？

→原子座標の組

原子座標からわかること



- かたち
- 分子間相互作用

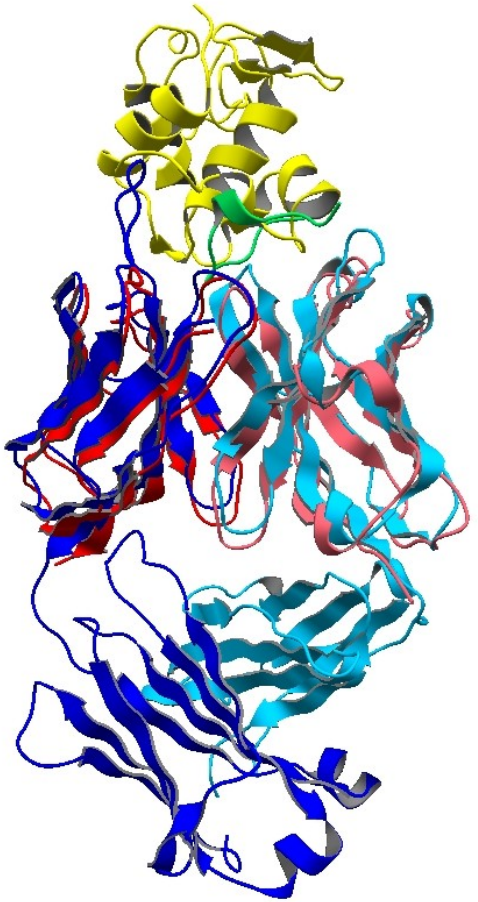
これらを介して機能のメカニズムを推測する。

- しかし一つの構造をみるだけでは複雑すぎてどこを見たらよいのかわからないことも多い。

→

構造の比較・パターン分類

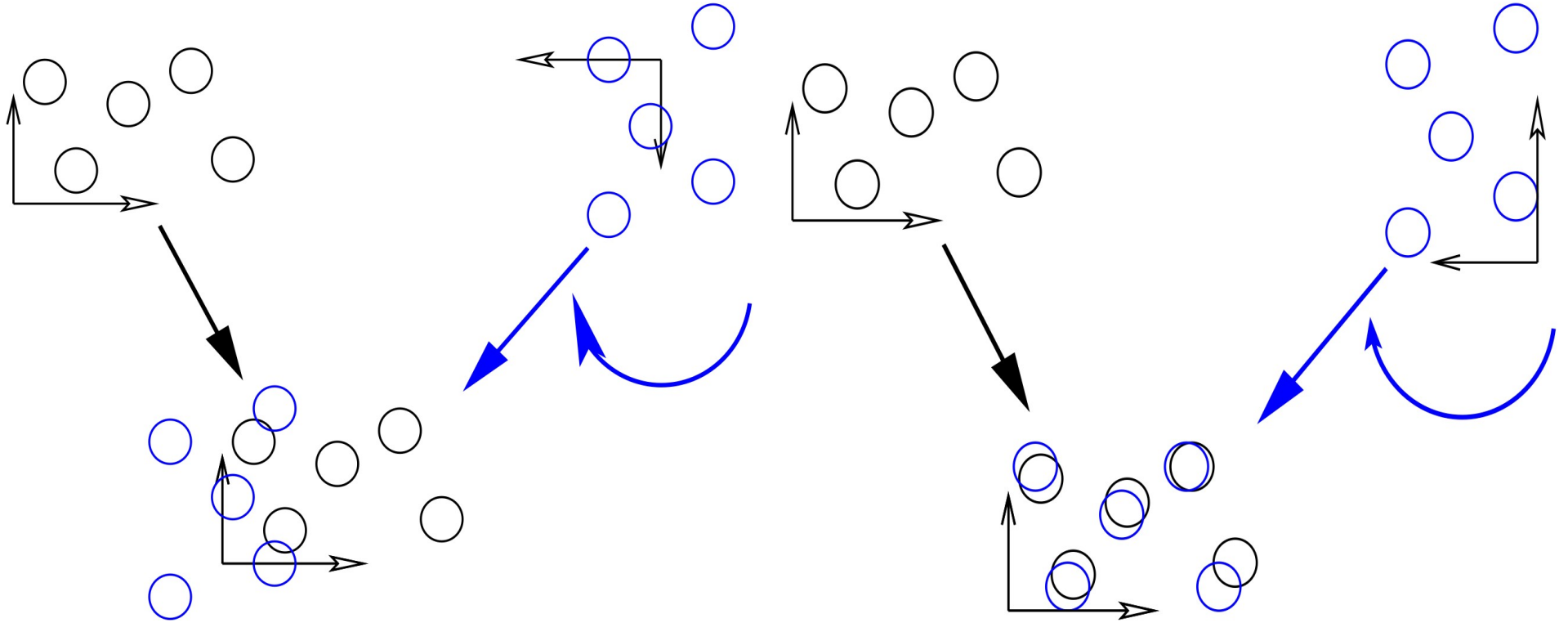
立体構造比較



- 似たような構造同士を並べて重ね合わせることで、機能部位の普遍性と多様性が見えてくる。
- そのためには、
  - 似た構造を探し出し、
  - 原子間の対応付ける

という作業が必要になる。

構造類似性の検出法（の一つ）

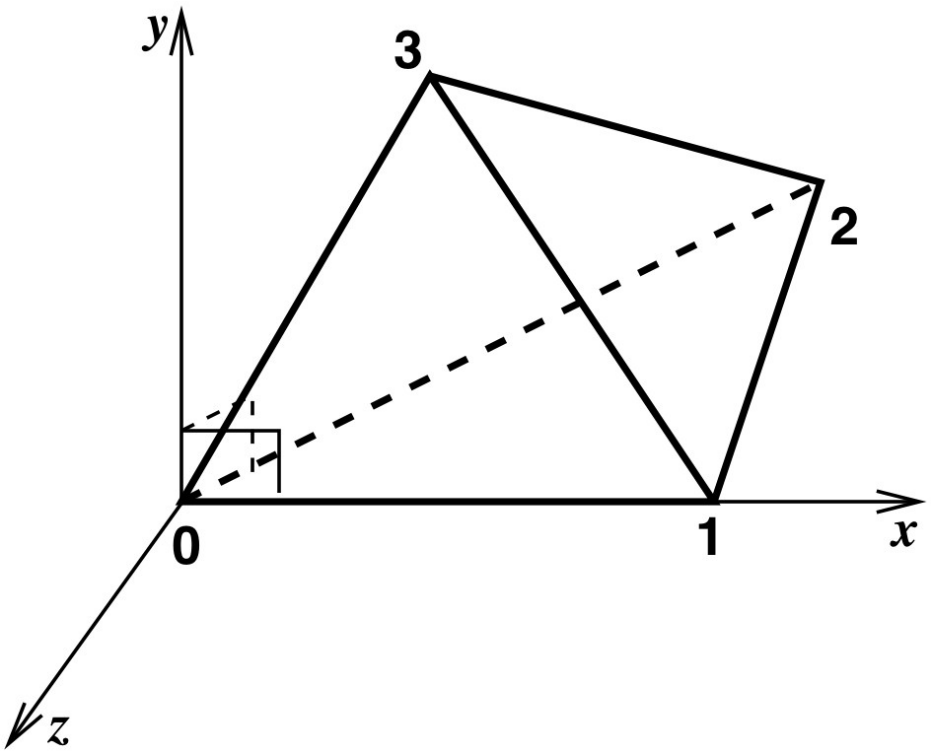


- （いろいろな）局所座標系を定義する。(i.e., 並進・回転の自由度を取り除く)
- それぞれの局所座標系で重なる原子を数える。
- 重なる原子が多ければ「似ている」と判定できる。

構造比較の計算複雑性

- 3つの原子の座標で局所座標系が定義できる。
- 蛋白質A(M原子) vs. 蛋白質B(N原子)
  - 局所座標系A: $O(M^3)$
  - 局所座標系B: $O(N^3)$
- これらの座標系の総当たり比較をするので、結局 $O(M^3 N^3)$ の手間が必要になる。
  - 実際は発見的手法で $O(M^2 N^2)$ 程度に減るが……
- 一対の構造比較の時間は馬鹿にならない！
- 十万件以上の構造データに対して検索はできない！

幾何学的索引付け



- 原子座標のDelaunay分割で決まる四面体。
- 四面体で局所座標系を定義。
- 四面体の属性（体積、面積、辺の長さ、各面の周囲の原子組成など）で特徴付け。
- 四面体の属性と変換後の原子座標を丸ごとRDBに保存。
- 原理的にはSQLの１クエリで類似性検索が可能。
  - 実際は効率化のために、外部プログラムのハッシュテーブルも併用する。

RDBへの格納と検索

```
CREATE TABLE refsetdb (
  lbsml_id INTEGER,
  irs INTEGER,
  PRIMARY KEY (lbsml_id, irs)
  tetra TEXT,
  tvol DOUBLE PRECISION,
  td01 DOUBLE PRECISION,
  td02 DOUBLE PRECISION,
  td03 DOUBLE PRECISION,
  td12 DOUBLE PRECISION,
  td23 DOUBLE PRECISION,
  td31 DOUBLE PRECISION,
  atype_id INTEGER [ ],
  xco DOUBLE PRECISION [ ],
  yco DOUBLE PRECISION [ ],
  zco DOUBLE PRECISION [ ]
);
```

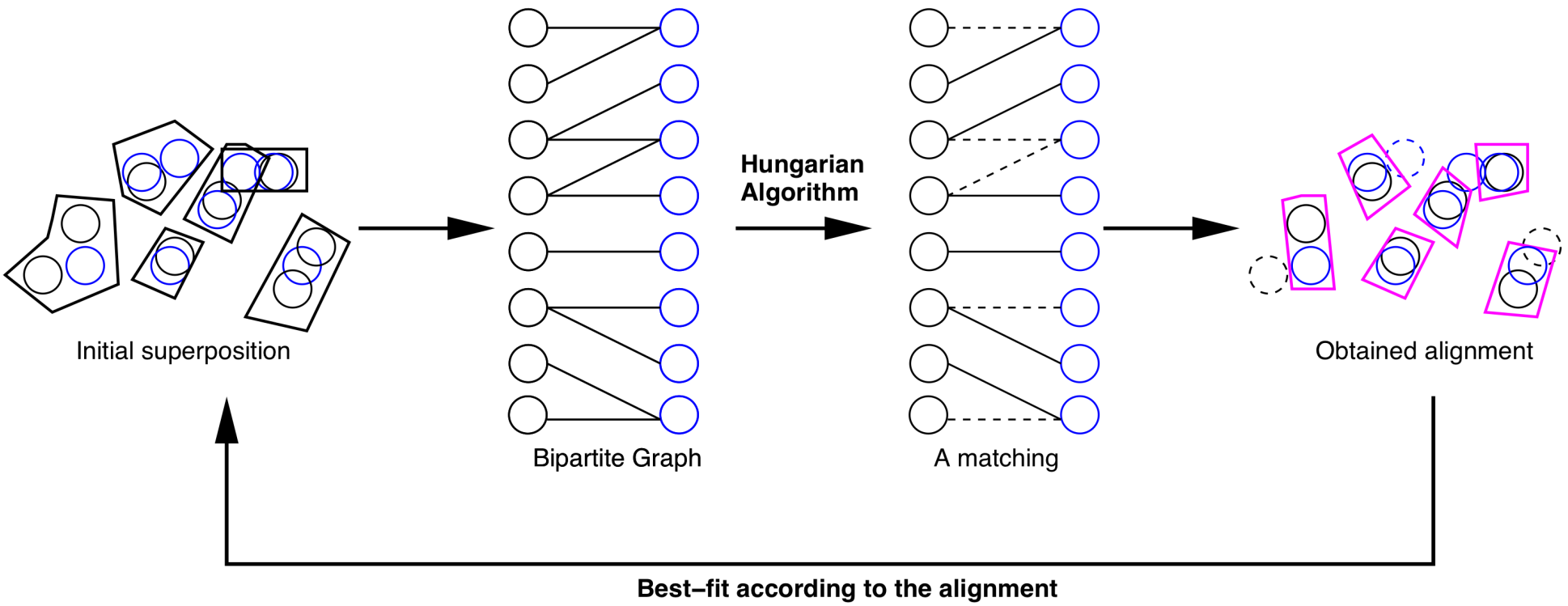
（実際は四面体には40の属性がある）

実際はまとめてバイナリデータのコラムにする

```
SELECT atype, xco, yco, zco, lbsml_id, irs
FROM refsetdb
WHERE tetra = 'tq'
  AND tvo1 BETWEEN vq −Δv AND vq +Δv
  AND td01 BETWEEN d01−Δd AND d01+Δd
  AND td02 BETWEEN d02−Δd AND d02+Δd
  AND td03 BETWEEN d03−Δd AND d03+Δd
  AND td12 BETWEEN d12−Δd AND d12+Δd
  AND td23 BETWEEN d23−Δd AND d23+Δd
  AND td31 BETWEEN d31−Δd AND d31+Δd
  AND ...
```

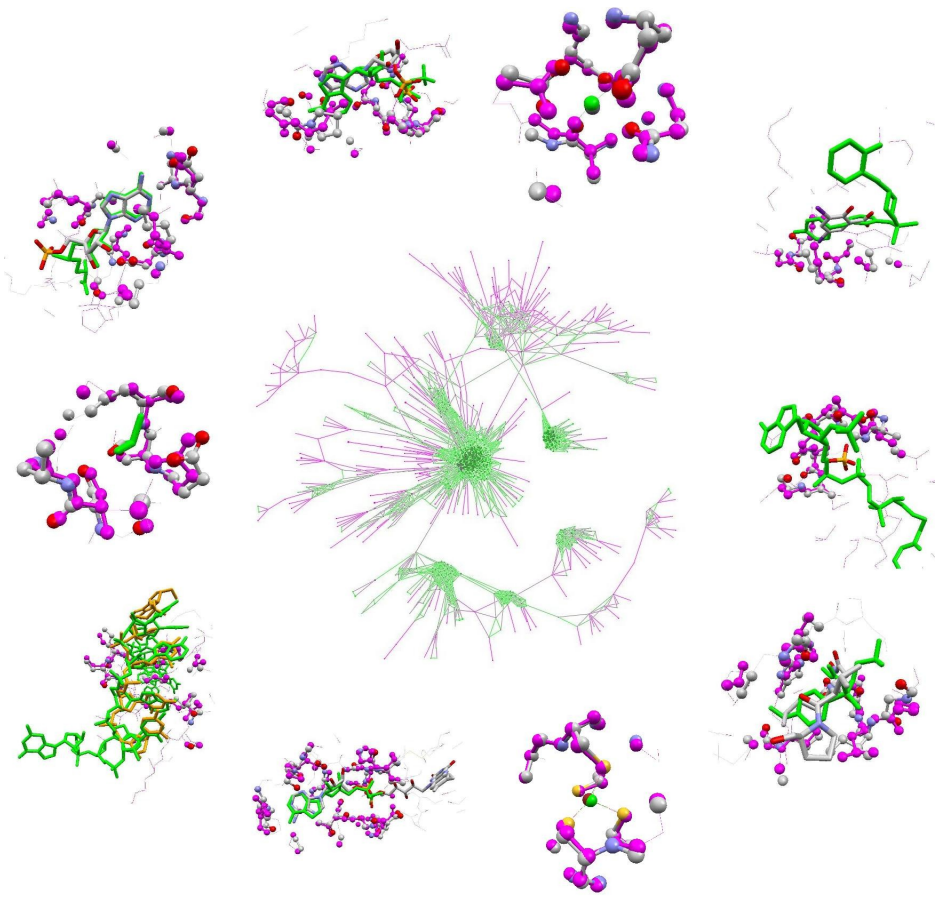
A.R. Kinjo & H. Nakamura (2007) *BIOPHYSICS* **3**:75-84

「似ている」からアラインメントへ



- 近くにある原子対を辺（重み付き）で結ぶ。
- 二部グラフのマッチングで原子の１対１対応が得られる。
- その対応に基づいて、最小二乗法で構造重ね合わせ。
- もう一度二部グラフを作り直して繰り返し計算。

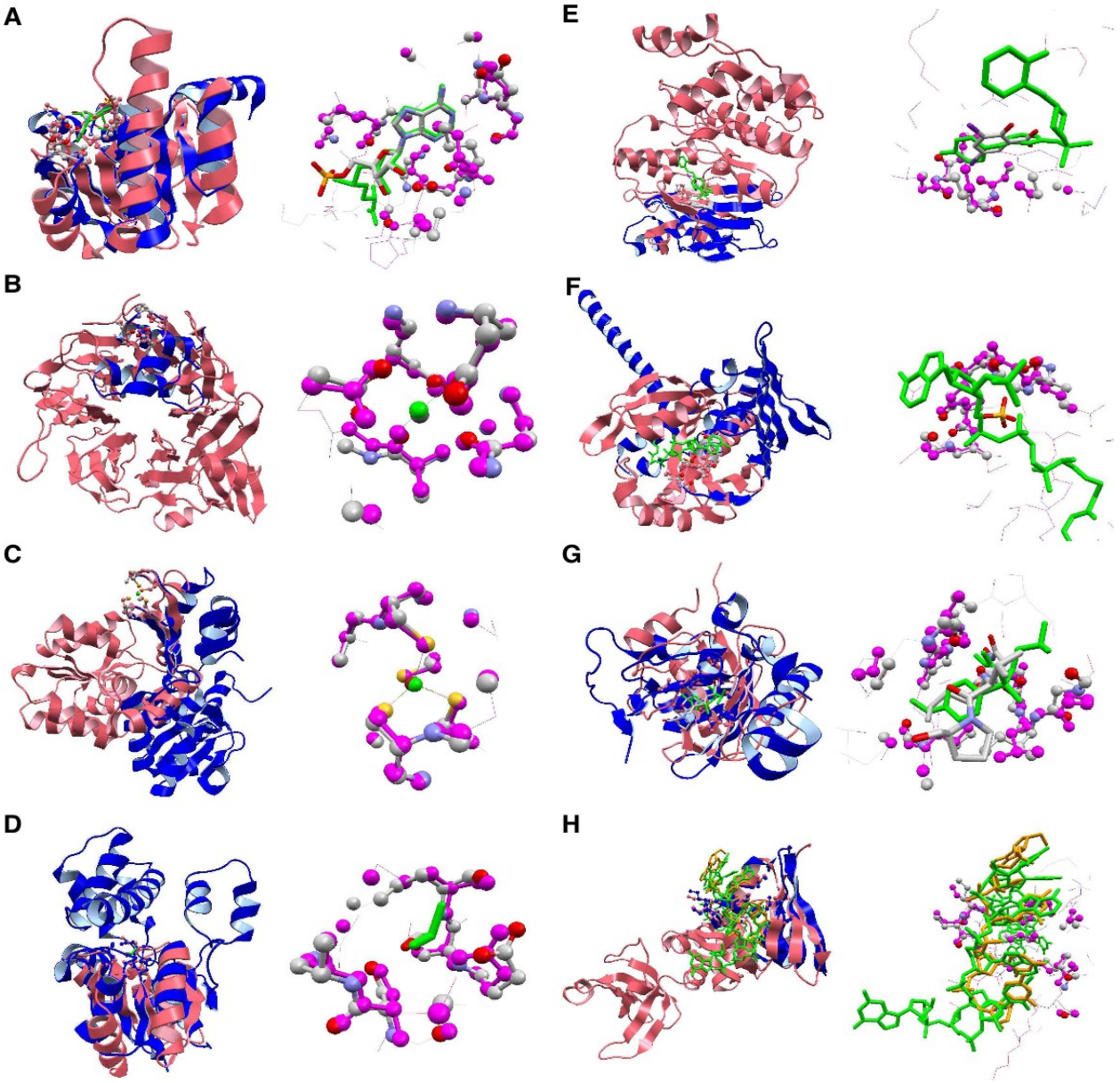
リガンド結合部位の総当たり比較



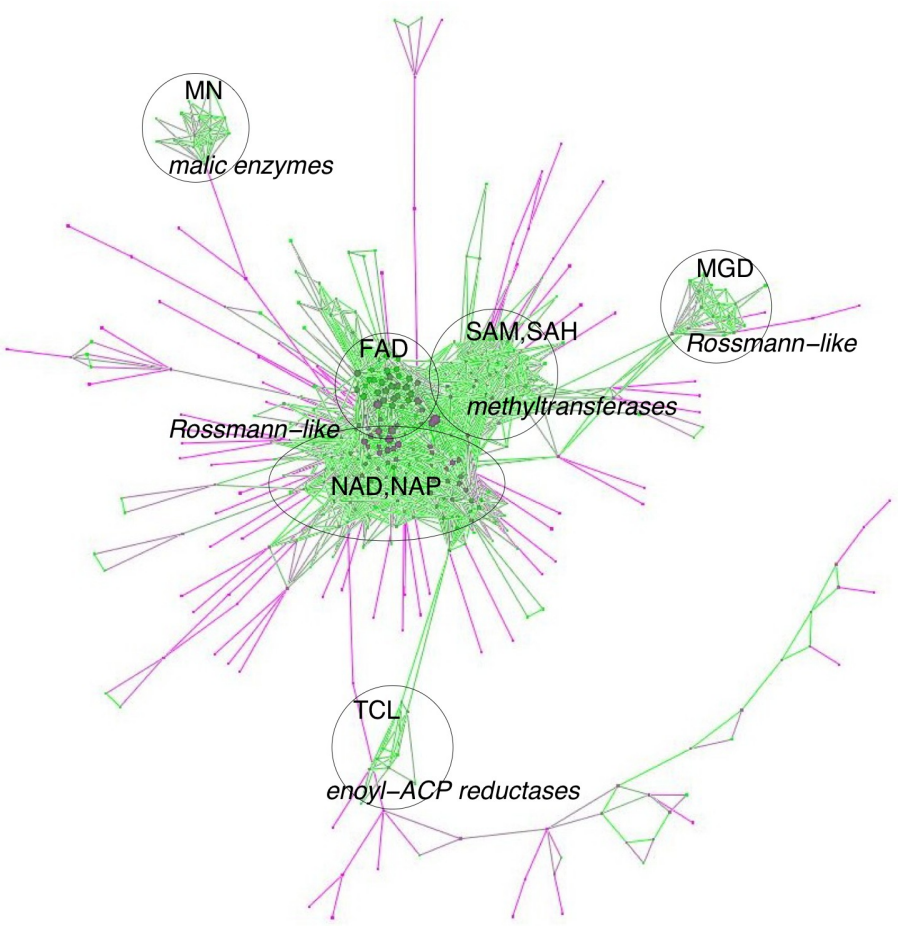
- １８万件以上の蛋白質のリガンド結合部位の総当たり比較。
  - 3 days/160 CPU cores
- 3000程度の「頻出パターン」を同定。
- 4000以上のパターンが全体構造の類似性とは無関係に出現していることが明らかになった。

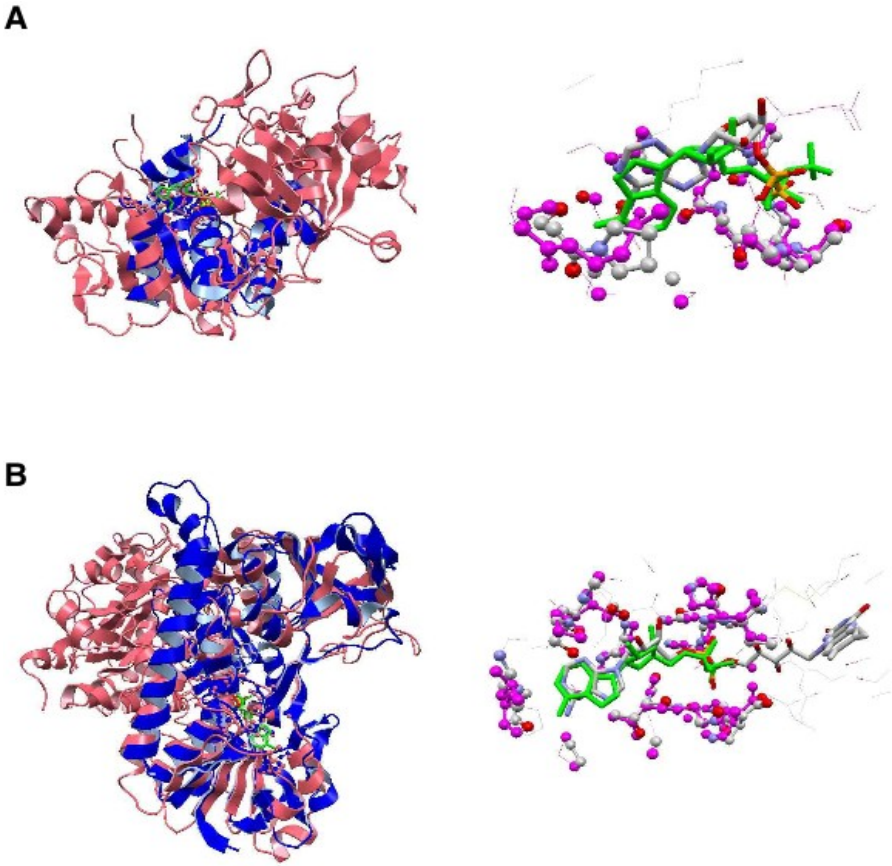
A.R. Kinjo & H. Nakamura (2009) *Structure* **17**:234-246

構造類似性の例



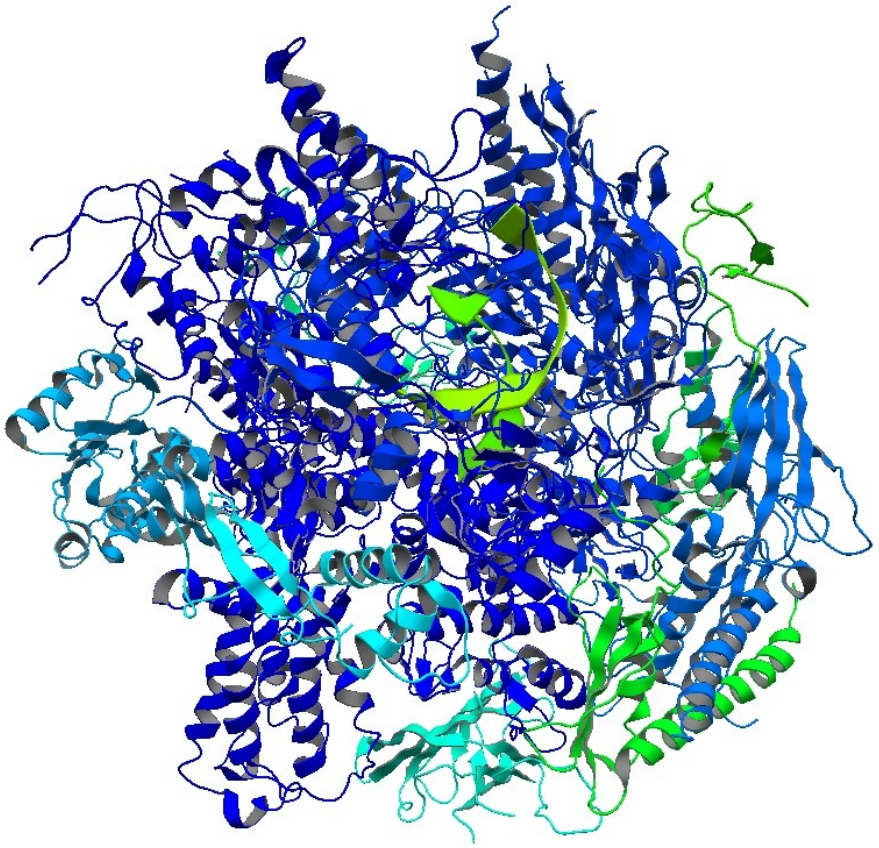
構造類似性のネットワークの例





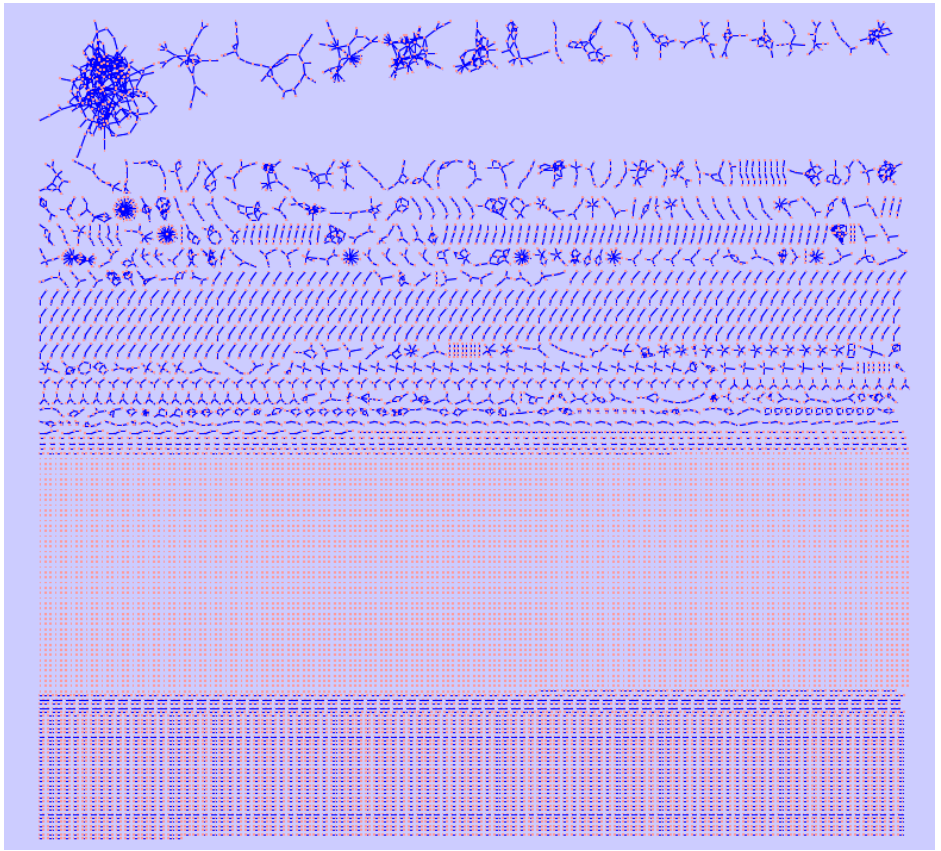
蛋白質と（低分子）化合物の相互作用のパターンが分類できた。

蛋白質-蛋白質相互作用への応用



- 蛋白質-低分子化合物相互作用は蛋白質の「生化学的機能」（分子機能）
- 蛋白質-蛋白質相互作用は蛋白質の「生物学的機能」（細胞機能）
  - 分子・細胞生物学の教科書を見よ。
  - あるいは藤博幸（著）『タンパク質機能解析のためのバイオインフォマティクス』（講談社サイエンティフィク）

蛋白質結合モチーフのネットワークはシンプル



- リガンド結合部位の比較と同様に蛋白質結合部位（約18万件）の総当たり比較を行った。
- その結果得られる相互作用ネットワークは「小さい」(いわゆる “small world”という意味ではない！）
- 蛋白質間相互作用は精密に維持されていることが推測される。

まとめ

- タンパク質立体構造データベースは二つの要素からなる。
  - アノテーション：実験情報、モノに関する情報（由来、配列、機能）
    - 通常のRDBなどの通常の運用で検索可。
  - 原子座標
    - 現実的な検索をするためには特別な工夫が必要。
    - うまく前処理できればRDBが活用できる。
- アノテーションと原子座標の両方をRDBで扱えれば、ちょっとした「統合DB」が可能になる（？）
  - これまではアノテーションと構造は全く別物扱いだった。

PDBjのメンバー

- 統括責任者
  - 中村 春木 (大阪大学蛋白質研究所・教授)
- PDBjデータベース管理運営グループ
  - 中川 敦史 (大阪大学蛋白質研究所・教授)
  - 松浦 孝範 (大阪大学蛋白質研究所・特任研究員(客員准教授))
  - 五十嵐 令子 (科学技術振興機構・研究補助員)
  - 見学 有美子 (科学技術振興機構・研究補助員)
  - 松浦 かんな (科学技術振興機構・研究補助員)
  - 井上 真由美 (大阪大学蛋白質研究所・特任研究員)
  - 陳 旻瑜 (大阪大学蛋白質研究所・特任研究員)
- PDBj国際的な運営高度化グループ
  - 金城 玲 (大阪大学蛋白質研究所・准教授)
  - 岩崎 憲治 (大阪大学蛋白質研究所・准教授)
  - 鈴木 博文 (大阪大学蛋白質研究所・特任研究員)
  - 山下 鈴子 (科学技術振興機構・技術員)
  - 鎌田 知佐 (科学技術振興機構・研究補助員)
  - 清水 有希子 (科学技術振興機構・研究補助員)
  - 工藤 高裕 (大阪大学蛋白質研究所・特任研究員)
- BMRBデータベース管理運営グループ
  - 藤原 敏道 (大阪大学蛋白質研究所・教授)
  - 阿久津 秀雄 (大阪大学蛋白質研究所・客員教授)
  - 小林 直弘 (大阪大学蛋白質研究所・特任研究員)
  - 中谷 英一 (科学技術振興機構・技術員)
  - 原野 陽子 (大阪大学蛋白質研究所・特任研究員)
- 九州大学生体防御医学研究所グループ
  - 藤 博幸 (九州大学生体防御医学研究所・教授) (for ASH)
  - 加藤和貴 (九州大学デジタルメディシンイニシャティブ・准教授)
  - 大津美希 (九州大学生体防御医学研究所・研究補助員)
- 研究協力者
  - 輪湖 博 (早稲田大学社会科学部・教授)
  - 伊藤 暢聡 (東京医科歯科大学大学院・教授)
  - 木下 賢吾 (東京大学医科学研究所・准教授)
  - Standley, M.Daron (大阪大学免疫学フロンティア研究センター・准教授)