# タンパク質立体構造の配列および原子間距離による分類と非冗長化されたPDB代表タンパク質チェインデータベース (PDB-REPRDB) の作成

野口 保[†] 秋山 泰[†]
鬼塚 健太郎[†] 安藤 誠[†]

タンパク質立体構造データベース (PDB) は、近年のX線結晶回折やNMRによる構造解析技術の進歩により、その内容は現在７５００エントリー (3.4Gbytes) を越え、今後もさらに増え続けると予想されている。しかしながら、冗長性やデータの不完全性のためにPDB全てのエントリーがタンパク質の立体構造の解析に適しているとは言えず、何らかの基準でタンパク質立体構造を分類し、代表タンパク質を決定する必要がある。タンパク質立体構造データの分類は、立体構造の取扱いの困難さとそれに基づく分類に膨大な計算が必要なため、近似的に配列の相同性（ID%）を指標にして行われてきた。我々は、従来のID%による分類に、タンパク質分子を重ね合わせた時の原子間距離の最大値（Dmax）を分類の指標として加え、より正確な分類を可能にした。

本論文では、本手法をMPIライブラリを用いて並列化し、新しい分類指標の追加に伴う計算量増加の問題を解決した。本研究で実装した並列版では、従来の約110倍の高速化を実現し、およそ1週間を必要としていた代表タンパク質決定処理を約1.5時間で実行できるようになった。我々は、本手法を用い、様々なID%とDmaxの値の組合せでPDBのチェインを分類し、代表を決定した“PDB代表タンパク質チェインデータベース（PDB-REPRDB）”を作成した。本データベースをWWWで公開し、既に世界から2100回以上アクセスされている。

# The classification of protein structures based on the sequential and structural similarity, and the database of representative protein chains (PDB-REPRDB)

TAMOTSU NOGUCHI,[†] YUTAKA AKIYAMA,[†] KENTARO ONIZUKA[†]
and MAKOTO ANDO [†]

The Protein Data Bank (PDB) is a rich library of atomic-coordinate data of biological macromolecules. The PDB entries has been increasing rapidly by the improvement of X-ray crystallography and NMR experimental techniques, and the number of current entries is more than 7,500 (3.4Gbytes), though not all entries are competent for the purpose of computational protein structure analysis. A lot of entries have insufficiently-refined coordinate data, or have some or many similar entries in terms of structural or sequential similarity. Thus the need for a classification procedure of protein sturcures has become quit obvious. We have proposed a representative chain database PDB-REPRDB, which startegy of selection is based on the sequential and structural similarity.

In this paper, we have developed a representative chain database PDB-REPRDB, and we report the MPI- parallelization of our automatic construction system for PDB-REPRDB. Now that a calculation of a representative set can be done within 1.5 hours rather than 1 week, with 110-folds speed-up achieved in this study. We have opened a WWW service for the PDB-REPRDB, which have been accessed more than 2100 times.

## 1. はじめに

PDB(Protein Data Bank)[1] は、米国のブルックヘブン国立研究所が提供しているタンパク質立体構造データベースで、X線やNMRなどの構造解析により明らかにされた生体高分子（タンパク質, DNA, RNA など）の立体構造が、その解析結果ごとに1ファイル（エントリー）の形式で登録されている。

近年のX線結晶回折やNMRによる構造解析技術の進歩により、そのデータ量は１９９１年ごろから急激に増加し、１９９８年４月版で７５００エントリー (3.4Gbytes) を越え、さらに増え続けている（図1）。

しかしながら、そのエントリーの多くは配列と立体構造がともに類似している“近縁”のタンパク質である。近縁タンパク質の基準として、たとえば、

† 技術研究組合 新情報処理開発機構
Real World Computing Partnership

- **配列の相同性基準**：ID%（配列を重ね合わせた際の同一アミノ酸残基の比率）≥ 75%、かつ、
- **立体構造の類似性基準**：Dmax（構造を重ね合わせた際の原子間距離の最大値）≤ 10.0 Å、

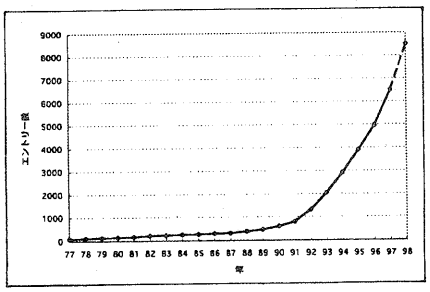


**図1** タンパク質立体構造データベース (PDB) のエントリー数の推移

を採用すると実に全エントリーの85%は他のタンパク質と近縁関係にある。またPDBデータは、実験方法の差異、分解能やリファインメント☆の度合いなどによってデータの質（信頼度）が様々である。PDBデータを利用する場合、類似のデータがあれば、より質の良いデータを利用した方が、解析誤差を低く抑えられる。

このような需要のもとにHobohmら[2),3)]は、**配列間の相同性**のみを考慮して、PDBの代表タンパク質チェインを決定する方法を提案し、"PDB_SELECT"[4)]として公開している。現在この代表は**配列の相同性**：ID% < 25%の基準で用意されており、タンパク質立体構造の研究者の間で広く用いられている。

また、Holmら[5)]は、**配列の相同性**で代表タンパク質チェインを決定し、その代表タンパク質チェインをPDBの全チェインに対して構造アライメントして、立体構造の類似したチェインを検索したデータベース(FSSP[5)])を作成し、公開している。

しかし一方で、たとえ**配列の相同性**が高いタンパク質であっても、立体構造を重ね合わせた時に、**部分構造**が大きく異なることがある。このような局所的構造のバラエティを残して、研究用のデータセットを作成したい場合には、従来からの**配列の相同性**だけを基準とする方法では不十分である。そこで我々は、**配列の相同性**が高いチェイン同士を比較し、部分的に立体構造が異なるチェインは別の代表点とする"PDB-REPRDB" V.1.0を作成した。ただし、この時点での選定作業には、多くの手作業が残されていた。その後、PDBのエントリー数の急激な伸びに対応するため、PDBの代表タンパク質決定システムの自動化（逐次版）を行った[8)]。

本論文では、この自動決定システムを説明するとともに、さらに処理を高速にするため、システムの並列版を作成したので報告する。

☆ リファインメント(refinement): 実験データをもとに立体構造を構築していく段階で、実験データと矛盾なく、かつエネルギー的により安定な構造を力学計算により決める処理。

## 2. タンパク質立体構造の分類

タンパク質立体構造をチェインごとに分類したデータベースとしては、**配列の相同性**だけを考慮して分類した"PDB_SELECT"[4)]や"FSSP"[5)]の他に、配列と立体構造のトポロジーを解析して分類した"SCOP"[7)]や"CATH"[6)]がある。SCOPとCATHは、ともにタンパク質の**全体構造**を分類したデータベースで、部分的な構造の違いは考慮されていない。また、各グループの代表構造と言ったものは特に決めていない。

したがって、現在までに配列と立体構造を同時に比較しながら、タンパク質立体構造を分類し、代表タンパク質チェインを決定しているデータベースは、PDB-REPRDBだけである。

我々は、**配列の相同性**と**立体構造の類似性**を同時に考慮しながら、代表タンパク質チェインを決定するためのシステム（PDB代表タンパク質決定システム）を作成した[8)]。

分類の基準は、図2のように、

- **配列の相同性基準**：ID%（配列を重ね合わせた際の同一アミノ酸残基の比率）、
- **立体構造の類似性基準**：Dmax（構造を重ね合わせた際の原子間距離の最大値）、

の両方を用いた。**立体構造の類似性基準**に関しては、全体構造を重ね合わせた時のr.m.s.d値を基準にするのが一般的であるが、部分構造の違いを検出するのには適さないので、我々はDmaxを採用した。

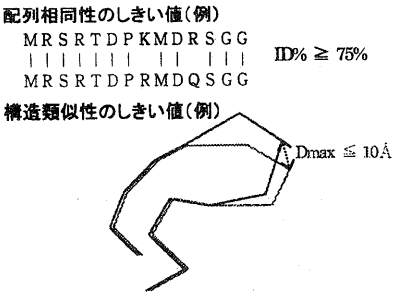


**図2** 近縁タンパク質の基準

## 3. PDBの代表タンパク質決定システム

PDB-REPRDBは、PDBをもとに、以下の手順で作成する（図3）。

### 3.1 不適切なデータの除外

PDBのエントリーをまずチェイン☆☆単位に分離した

☆☆ タンパク質が単数の／複数のポリペプチド鎖で構成されるときの

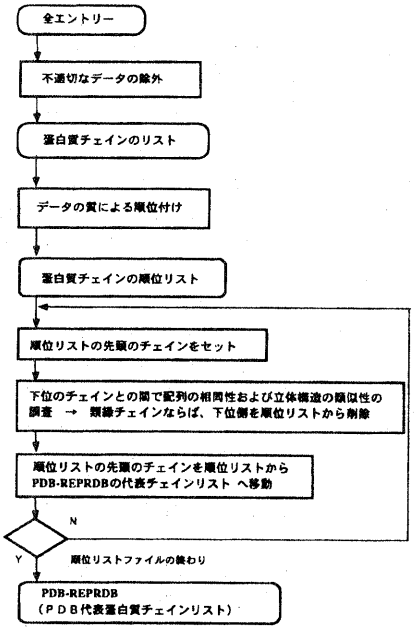


図3 PDBの代表タンパク質決定システムの流れ

のち、下記に該当するデータを取り除く。

a) DNA と RNA データ
b) 理論計算だけで求められたモデルデータ
c) チェインの長さが短いデータ ($l < 40$ 残基)
d) 全ての残基において主鎖座標が欠落したデータ
e) 全ての残基において側鎖座標が欠落したデータ

### 3.2 データの質による順位付け

PDBデータのチェインごとに、下記の優先度で並び替えを行い、順位リストを作成する。

最初にX線結晶回折によって構造解析されたデータを、まず分解能、次にRファクターの小さい順に並び替え、順位リストを作成する。分解能、Rファクターがともに等しい場合は、さらに下記の項目を順に調べて順位付けを行なう。NMRのデータは、分解能やRファクターに相当するパラメータがないので、下記の項目だけを順に調べて順位付けを行い、X線結晶回折の順位リストの下位に置いた。

(1) チェイン・ブレイク☆の数 (少ないほど上位)
(2) 標準的なアミノ酸残基種以外の残基の数 (少ないほど上位)
(3) 主鎖原子の座標を欠く残基の数 (少ないほど上位)
(4) 側鎖原子の座標を欠く残基の数 (少ないほど上位)
(5) チェイン名のアルファベット順

各鎖

☆ チェイン・ブレイク (chain break):PDBの座標において、チェインの途中で座標を決定できなかった原子が存在したためチェインが切れたように見える状態。または、リファインメントが不十分なため、主鎖の原子間距離が異常に離れた状態。

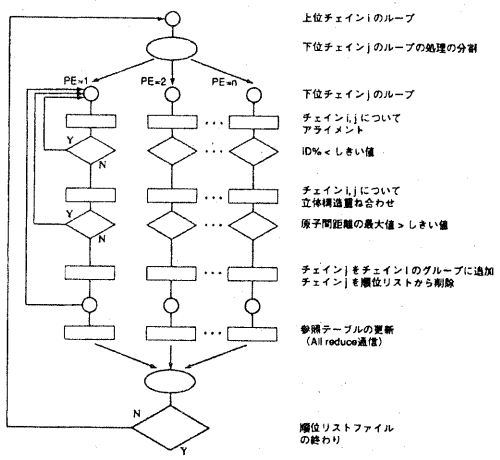


図4 並列PDB代表タンパク質決定システムにおける計算の流れ

(例: 1MCD < 1MCE, 5AT1A < 5AT1C)

### 3.3 類似タンパク質チェインの検索および代表タンパク質チェインの決定

上記の処理により作られた順位リストの上位のものを優先しながら、互いに近縁関係がないような代表チェインを選び出し、選択されなかったチェインについては、どの代表に近いかでグループ分けを行なう。

具体的には、まず上位のチェインのアミノ酸配列をキーにして、それ以下のチェインの配列相同性をDP(動的計画法)を用いたペアワイズアライメントの手法[11]で調べる。その相同性がしきい値以上であれば、下位側は消去しても良いと考えられる。しかし、我々の手法では、ここでさらに構造類似性のチェックを $C_\alpha$ 原子に注目して行なう。チェイン全体の骨格の重ね合わせを行い、全ての原子間のズレがしきい値以下であり、立体構造の差異もないと認められる時に初めて下位側をリストから削除し、近縁タンパク質チェインとして、代表点(上位側)と同じグループのリストに加える。(図3)

この処理を順にリストの最後まで行うことにより、近縁グループおよびその代表タンパク質チェインを決定する。立体構造の比較にはKabschによる最小2乗フィット法[12]を採用し、重ね合わせた原子間距離の最大値(Dmax)を求める。

## 4. 代表タンパク質決定システムの並列化実装

PDBのデータの急激な増加に対応し、かつ、様々な基準でのPDB代表タンパク質チェインを決定するために、PDB代表タンパク質決定システムの処理を高速化する必要があり、我々は、MPIライブラリを利用して、PDB代表タンパク質決定システムの並列化を行なった。

逐次版システムにおいて、処理時間の90%以上を要

していた“類似タンパク質チェインの検索および代表タンパク質チェインの決定”の部分（図3におけるループ）の内部において、上位側チェイン$i$が与えられたとき下位側チェイン$j$との比較処理は各$j$について同時に行なえることから、これをいわゆるSPMD(Single Program Multiple Data)方式で並列化した（図4）。

順位リストの各チェインが近縁タンパク質として削除された状態か、未削除かを記録する参照テーブルを用意する。この参照テーブルをもとに比較を行なうべきチェインが決められ、以下の処理が並列実行される。

並列に処理されるのは、配列間アライメント、立体構造重ね合わせ、および参照テーブルの更新である。タンパク質チェインのリストと全配列データは、$n$台のプロセッサの全てに配布しておく。

上位側チェイン$i$と比較すべき下位側チェイン$j$の各プロセッサへの分担法は計算の当初から静的に決められており、チェイン番号にしたがいブロックサイクリック的に対応づけられる。すなわち$m$本のチェイン$c_0$から$c_{m-1}$があるとき、第$i$番目のチェイン$c_i$を担当すべきプロセッサの番号$p$ $(1 \leq p \leq n)$は、

$$p = (\left\lfloor \frac{i}{k} \right\rfloor \bmod n) + 1 \qquad (1)$$

で決定される。ただし$k$はブロック幅（今回は1）、$n$は使用するプロセサ台数とする。

配列間アライメントで用いる各チェインの配列データは各プロセッサのメモリ上に保持し、立体構造重ね合わせで用いる原子座標データは、必要に応じてPDBファイルから読み込むことにした。立体構造重ね合わせが行なわれるのは、アライメントの結果、相同性が高かった時のみであり、その実行の割合はアライメント約500回に対し1回程度である。また原子座標のデータ量は、$C_\alpha$原子部分だけでも大きい（約120Mbytes）ため、各プロセッサのメモリには配列データ（約10Mbytes）のみを置いた。

必要となる通信は、最初に参照テーブルと全配列データを各プロセッサにブロードキャストし、以降は図4の上位側チェイン$i$のループが終了するごとに、各プロセッサで削除したチェイン名を収集して、参照テーブルの内容を更新して再びブロードキャストすることである。プロセッサ間通信については、MPIライブラリを用いて実装した。

ブロックサイクリック化により、ある程度の負荷分散が期待されるが、配列長のバラツキなどのため、必ずしも充分には均一化されていない。

## 5. 並列化の性能評価

PDB代表タンパク質決定システムを並列化することにより、どれだけ処理時間が短縮できるかを実測により調べた。速度性能は、日立製のSR2201/256を用いて評価した。表1は実験に用いたSR2201の仕様である。

表1 SR２２０１の仕様

| 機種 | プロセッサチップ | プロセッサ数 | 主記憶 |
|---|---|---|---|
| Hitachi SR2201 | PA-RISC1.1+PVP-SW 150MHz | 256 | 64GB 分散メモリ |

使用したPDBは、リリース#78(エントリー数: 4873, 全チェイン数: 8870, 順位リストに残るチェイン数：6127 )である。チェイン数による性能の違いを評価するため、順位リストの上位1000本のチェインだけとったサブセットと、全6127本のチェインからなるフルセットとを作り、性能評価に利用した。

図5にSR2201での処理時間と速度向上比を示す。順位リストのチェイン数が1000本の場合と6127本の場合とで、ほぼ同様の性質を示している。両者とも計算粒度は十分に大きく、通信コストは隠蔽されていると言える。順位リストのチェイン数が6127本の場合、256プロセッサ利用時で、約110倍の台数効果を得た。このとき、約1.5時間で6127本の順位リストのチェインを分類することができた。

今後、PDBエントリー数の増加とともに、順位リストのチェイン数も増加し、計算粒度もさらに大きくなるので、台数効果はさらに向上すると予想される。

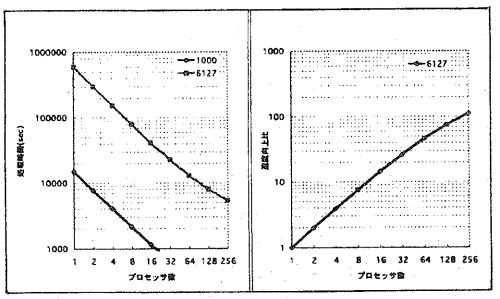


図5 SR2201上での処理時間と速度向上比

## 6. 結　　果

本論文で分類実験に使用したPDBは、1998年4月版のリリース#84である。総エントリー数は、75578で、“不適切なデータの除外”や“データの質による順位付け”の結果、実際に分類を行なった順位リストのチェインの数は、11062本であった。

この順位リストのチェインに対し、代表タンパク質決定システムを用いて、**配列の相同性**：ID% < 25% ～ 95%まで10%刻みの8通りと、**構造の類似性**：Dmax >10 Å ～ 50 Åまで10 Å刻みと∞ Åを加えた6通りの基準を組み合わせて、合計8 × 6 = 48通りの代表タンパク質を決定した。（表2、図6）

図6を見てまず気がつくのは、Dmax > 10Åの基準で決定した代表チェイン数と、他のDmaxの基準で決定した代表タンパク質チェイン数の差が、ID%のしき

表2 決定した代表タンパク質チェインの数

| ID% | Dmax(Å) | | | | | |
|---|---|---|---|---|---|---|
| | > 10 | > 20 | > 30 | > 40 | > 50 | ∞ |
| < 25 | 1689 | 1176 | 994 | 898 | 882 | 874 |
| < 35 | 1792 | 1455 | 1399 | 1381 | 1378 | 1377 |
| < 45 | 1900 | 1620 | 1579 | 1567 | 1564 | 1562 |
| < 55 | 2019 | 1784 | 1749 | 1734 | 1732 | 1729 |
| < 65 | 2152 | 1934 | 1900 | 1886 | 1884 | 1882 |
| < 75 | 2267 | 2064 | 2033 | 2020 | 2019 | 2018 |
| < 85 | 2449 | 2272 | 2239 | 2228 | 2227 | 2225 |
| < 95 | 2812 | 2672 | 2645 | 2638 | 2637 | 2637 |

い値によらず常に大きいことである。ID% < 95% でも、175 (=2812-2637) 本のチェインを別のチェインとして分類している (表2)。このことから、ID% ≥ 95% の配列相同性があっても、配列の置換や挿入・欠損によって、Dmax > 10Å 部分構造が変化していることがわかる。

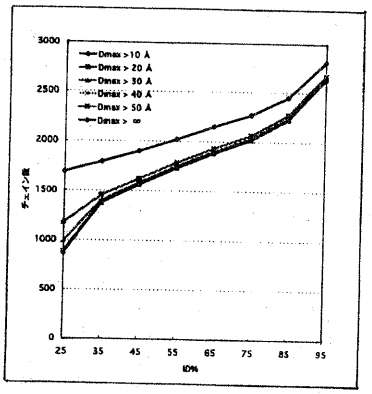


図6 類似基準 (ID% と Dmax) と代表タンパク質チェイン数の関係

図6を見て次に気づくのは、ID% < 25% と ID% < 35% の間で、代表タンパク質チェイン数が急激に変化していることである。表2の Dmax > 20 Å と ∞ の列を比較すると、ID% < 35% では、75 (=1455-1377) 本の代表タンパク質チェインしか増加していないが、ID% < 25% では、302 (=1176-874) 本も代表タンパク質チェインが増えている。このことは、ID% < 25% になると、配列の相同性だけを考慮した分類だと、本来分けるべき、構造が異なる他のグループを数多く吸収してしまっていることを示している。

ID% < 85% でありながら、Dmax > 50 Å の基準で別のチェインと分類された例を図7に示す。図7は、抗トロンビンのLチェインとIチェインを重ね合わせた図である。ID% は 94.0% あるがLチェインのC末端にあるβシート構造が、Iチェインでは解けてしまっていている。このためC末端の部分構造がLチェインとIチェインでは、大きく異なっており、Dmax の値が 62.4 Å と非常に大きくなっている。r.m.s.d 値も 21.5 Å と比較的大きな値になっているが、この結果からも、Dmax 値の方が、部分構造の違いを調べるには、適していると言える。

このように本システムでは、配列の相同性だけの分類では見落としてしまう立体構造の違いを見逃さず分類することができた。また、上記の結果から、タンパク質の立体構造は、配列相同性だけで単純に分類できるものではなく、ID% のしきい値によっては、構造の類似性も考慮する必要があることを示すことができた。


図7 抗トロンビン (PDB エントリー名:2ANT) のLチェイン (薄いリボン) とIチェイン (濃いリボン) の重ね合わせ図

## 7. PDB 代表タンパク質チェインの公開

本システムの処理結果である PDB 代表タンパク質チェインは、PDB-REPRDB[8] として、図8のように WWWで公開している。WWW ページは、我々の研究室で公開している PAPIA システム[9] とゲノムネットの WWW サーバー[10] とリンクし、多くの利用者に使われている。

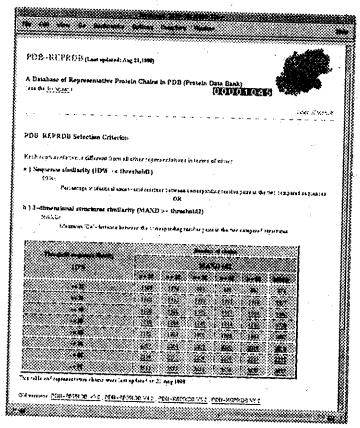


図8 WWW 上の PDB-REPRDB

ホームページ (図8) では、様々な基準で決定した代表タンパク質チェインの数が記された表が表示され、あ

る基準での代表タンパク質チェインが知りたい場合、その基準のマス目の数字をクリックすると、図9のように、その基準での代表タンパク質チェインのリストが表形式で表示される。エントリー名とチェインIDの部分は、分類された類似タンパク質チェインのリストとホットリンクしており、クリックするとその類似タンパク質チェインの詳細を知ることができる。類似タンパク質チェインリストのエントリー名とチェインIDの部分は、PDBとホットリンクしており、クリックすると該当するPDBエントリーの内容が表示される。

また、エントリー名とチェインIDと残基数の間の"*"印をクリックするとRasMolプログラムを用いた立体構造のグラフィック表示もできる。

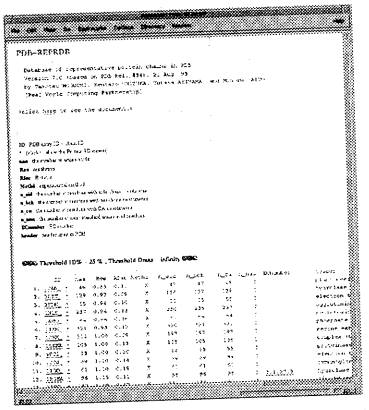


図9 WWW上のPDB代表タンパク質チェインリスト（例）

## 8. まとめ

配列の相同性（ID%）だけでなく、構造の類似性にも着目し、タンパク質分子を重ね合わせた時の原子間距離の最大値（Dmax）を分類の指標にした新たなタンパク質立体構造の分類手法を提案し、その手法を用いたPDB代表タンパク質決定システムを作成した。

また、PDB代表タンパク質決定システムをMPIライブラリを用いて並列化し、処理の高速化を実現した。この並列化により、SR2201の256プロセッサ利用時で、約110倍の台数効果を得て、順位リストのチェイン6127本を約1.5時間で分類することができた。

提案した分類手法を用いた結果、配列の相同性だけでは分類できなかった、配列は相同性だが、立体構造の異なるチェインを、別々のグループとして分類することが可能になった。

本手法により決定されたPDB代表タンパク質チェインは、非冗長化されたPDB代表タンパク質チェインデータベース(PDB-REPRDB)としてWWWで公開され、既に世界から2100回以上アクセスされている。

今後は、様々なタンパク質立体構造解析の研究者の要求にきめこまかく対応できるように、順位リストのチェイン間の配列相同性(ID%)や構造相同性(Dmax)のテーブルだけをあらかじめ用意しておき、オンデマンドで様々な基準の代表タンパク質チェインを決定し、提供できるようなシステムを構築していく予定である。

## 参考文献


1) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; and Tasumi, M.: The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *Journal of Molecular Biology* **112**, pp.535-542 (1977). http://www.pdb.bnl.gov/

2) Hobohm, U.; Scharf, M.; Schneider, R.; and Sander, C.: Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Science* **1**, pp.409-417 (1992).

3) Hobohm, U.; and Sander, C.: Enlarged representative set of protein structures. *Protein Science* **3**, pp.522 (1994).

4) Hobohm, U.; and Sander, C.: PDB_SELECT: Representative list of PDB chain identifiers. http://www.sander.embl-heidelberg.de/whatif/select

5) Holm, L.; and Sander, C.: Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, pp.316-319 (1998). http://www2.ebi.ac.uk/dali/fssp/fssp.html

6) Orengo, C. A.; Michie, A. D.; Jones, S.; Swindells, M. B.; Jones, D. T.; and Thornton, J. M.: CATH: Protein Structure Classification, version 1.0. http://www.biochem.ucl.ac.uk/bsm/cath

7) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C.: scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **247:** 536-540 (1995). http://scop.mrc-lmb.cam.ac.uk/scop/

8) Noguchi, T.; Onizuka, K.; Akiyama, Y.; and Saito, M.: PDB-REPRDB:A Database of Representative Protein Chains in PDB (Protein Data Bank). *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology.*, pp.214-217 (1997). http://pdap1.trc.rwcp.or.jp/papia/papia.html

9) Akiyama, Y.; Onizuka, K,; Noguchi, T.; and Ando, M.: Parallel Protein Information Analysis (PAPIA) system. *Proc. 1998 RWC Symposium*, RWC TR-98001, pp.123-128 (1998). http://pdap1.trc.rwcp.or.jp/papia/papia.html

10) "GenomeNet WWW Server" http://www.genome.ad.jp/dbget/dbget.links.html

11) Needleman, S. B. and Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology*, **48**, pp.443-453 (1970).

12) Kabsch, W.: A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **A34**, pp.827-828 (1978).