タンパク質立体構造データベースにおける代表タンパク質チェイン決定システムに関する研究

2001年1月

野口　保

内容梗概

タンパク質立体構造データベース (PDB) は，近年のＸ線結晶解析やＮＭＲによる構造解析技術の進歩により急激に増加し，その内容は 2000 年 11 月の時点で 13,600 エントリーを越えている．今後も，各種生物種のゲノムプロジェクトの後を受けて開始した“構造ゲノミックスプロジェクト (ゲノムの中に含まれるタンパク質の立体構造をすべて決める．)” によって，さらにその増加は加速す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 の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ない場合（例えば，タンパク質立体構造予測の研究）があり，何らかの基準でタンパク質立体構造を分類して，その中から代表タンパク質を決定する必要がある．代表の決め方としては，分類されたグループ内の中心を求めて，そのタンパク質を代表にする方法があるが，そうした場合，その代表タンパク質データの質が悪く（分解能が悪く，チェインが途中で切れているなど），タンパク質立体構造予測の基礎データとしては，相応しくないことがある．そのような観点では，各エントリーの内容を調べ，解析に適さない質の悪いデータを除去し，各エントリーに対して他のエントリーが配列および立体構造上，類似のタンパク質かどうかを調べあげ，分類した上で代表を決定する必要がある．PDB のエントリー内には，構造上，一つにつながったタンパク質チェインが複数本含まれる場合があるため，一般に，そのチェイン同士を比較・分類し，その中から任意の優先度で選ばれたタンパク質チェインを代表にしている．タンパク質チェイン同士の比較・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，近似的に配列の類似性 (ID%) を指標にして行なわれてきた．

本研究では，まず従来の 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，部分構造の違いも考慮した，より正確な立体構造分類を行って，非冗長な PDB 代表タンパク質チェインデータベース (PDB-REPRDB) を作成することを可能にした，PDB 代表タンパク質チェイン決定システムの初期バージョンの作成を行った．次に，自動化が不十分であった初期バージョンの自動化を進め，高速に代表タンパク質チェインを得られるように，PDB 代表タンパク質チェイン決定システムの改良を行った．新しい分類指標 (Dmax) の追加による計算量増加の問題は，MPI ライブラリを用いてプログラムを並列化すること

によって処理の高速化を行い解決した．また，あらかじめ決めた基準 (配列の相同性：ID% $\geq$ 25% $\sim$ 95% で 10% 刻みの 8 通りと，構造の相同性：Dmax $\leq$ 10 Å $\sim$ 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 $8 \times 6 = 48$ 通り) の代表タンパク質チェインを決定し，それらの結果を PDB-REPRDB として WWW 上に公開した．

しかしながら，研究の内容や研究者によって，代表を選ぶ基準は様々で，とてもすべての要求に応じきれない．そこで，最新バージョンでは，WWW によるインターフェースを作成し，研究者が得たい基準での代表セットを自身で指定することができるように改良し，代表タンパク質チェインをオンデマンドで提供することを可能にした．

本論文では，PDB 代表タンパク質チェイン決定システムの初期バージョン，並列版および最新の会話形式バージョンにおけるシステム構成及び特徴について述べる．

最後に，本システムで作成した代表タンパク質チェインの利用例として，タンパク質二次構造予測の基礎データとなる構造ライブラリのセットや，並列タンパク質情報解 (PAPIA) システムの検索対象となる立体構造データベースを作成し，実際のタンパク質立体構造予測の研究に役立てているので，それについても述べる．

本研究の成果は，1997 年 8 月に初期バージョンを WWW で公開以来，既に世界中から 4,000 回以上アクセスされ，多くの研究者に利用されている．

関連発表論文

1. 学術論文誌掲載論文

(1) Ken Nishikawa and Tamotsu Noguchi, "Predicting Protein Secondary Structure Based on Amino Acid Sequence", Methods in Enzymology, Vol.202, pp.31-44 (1991).

(2) 野口 保, 秋山 泰, 鬼塚 健太郎, 安藤 誠: "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", 情報処理学会論文誌, Vol.40, No.SIG2 (TOM1), pp.117-128 (1999).

(3) Tamotsu Noguchi, Kentaro Onizuka, Makoto Ando, Hideo Matsuda and Yutaka Akiyama: "Quick Selection of Representative Protein Chain Sets Based on Customizable Requirements", Bioinformatics, Vol.16, No.6, pp.520-526 (2000).

(4) Tamotsu Noguchi, Hideo Matsuda and Yutaka Akiyama: "PDB-REPRDB: A Database of Representative Protein Chains from PDB (Protein Data Bank)", Nucleic Acids Res., Vol.29, No.1, pp.219-220 (2001).

(5) Tamotsu Noguchi, Masahiro Ito, Hideo Matsuda, Yutaka Akiyama and Ken Nishikawa: "Prediction of Protein Secondary Structure Using the Threading Algorithm and Local Sequence Similarity", Research Communications in Biochemistry, Cell and Molecular Biology (掲載予定).

2. 国際会議会議録掲載論文

(1) Tamotsu Noguchi, Kentaro Onizuka, Yutaka Akiyama and Minoru Saito: "PDB-REPRDB: A Database of Representative Protein Chains in PDB (Protein Data Bank)",Proc. The fifth Int'l Conf. on Intelligent Systems for Molecular Biology, pp.214-217, AAAI Press (1997).

3．国内研究会，全国大会発表論文

(1) 野口 保，西川 建：" タンパク質の二次構造予測法の開発 (新ジョイント法)" 第 27 回生物物理学会年会　(1989).

(2) 野口 保, 秋山 泰, 鬼塚 健太郎, 斎藤 稔, 安藤 誠, 志澤 由久: "蛋白質立体構造データベース (PDB) の代表蛋白質決定システムの並列化", 情報処理学会研究報告 97-HPC-67-6, pp.31-36 (1997).

(3) 野口 保, 鬼塚 健太郎, 秋山 泰, 斎藤 稔: " 配列の相同性と立体構造の類似性を考慮した PDB 代表蛋白質データベース (PDB-REPRDB)" , 第 4 回「タンパク質立体構造の構築原理」ワークショップ予稿集, pp.52 (1997).

(4) Yutaka Akiyama, Kentaro Onizuka, Tamotsu Noguchi and Makoto Ando: "Parallel Protein Information Analysis (PAPIA) System Running on a 64-node PC Cluster"(64 ノード PC クラスタ上で動作する並列タンパク質情報解析 (PAPIA) システム), Proc. the 9th Genome Informatics Workshop, pp.131-140 (1998).

(5) 秋山 泰, 鬼塚 健太郎, 野口 保, 安藤 誠, 斎藤 稔: "並列タンパク質情報解析 (PAPIA) システムの PC クラスタ上での実現", 情報処理学会研究報告 97-HPC-70-6, pp.31-36 (1998).

(6) Yutaka Akiyama, Kentaro Onizuka, Tamotsu Noguchi, and Makoto Ando: "Parallel Protein Information Analysis (PAPIA) system." (並列タンパク質情報解析 (PAPIA) システム), Proc. 1998 RWC Symposium (RWC TR-98001), Real World Computing Partnership, pp.123-128 (1998).

(7) 野口 保, 秋山 泰, 鬼塚 健太郎, 安藤 誠: "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", 情報処理学会研究報告 98-MPS-21-6, pp.31-36 (1998).

(8) Yutaka Akiyama, Kentaro Onizuka, Tamotsu Noguchi, and Makoto Ando: "Development of Biological- and Chemical-Applications on a 64-node PC Cluster", Int'l Workshop on Innovative Architectures for Future Generation High-Performance Processors and Systems (IWIA'98), pp.27-34 (1998).

(9) 野口 保, 伊藤 將弘, 秋山 泰, 西川 建: " 3D-1D 法を用いたタンパク質二次構造予測法の改良" , 第5回「タンパク質の立体構造の構築原理」ワークショップ予稿集, pp.48 (1998).

(10) Yutaka Akiyama, Kentaro Onizuka, Tamotsu Noguchi, and Makoto Ando: "Biological- and Chemical- Parallel Applications on a PC Cluster" (PCクラスタ上での生物学と化学の並列応用), Proc. of International Symposium on High Performance Computing (ISHPC'99) (Lecture Notes on Computer Sciences. Springer-Verlag), pp.220-233 (1999).

(11) 秋山 泰, 鬼塚 健太郎, 野口 保, 安藤 誠: "大規模ＰＣクラスタを用いたインターネット上の公開計算サービス～並列タンパク質情報解析 (PAPIA) システムの構築と利用実績～", 情報処理学会研究報告 99-OS-81-11, pp.59-64 (1999).

(12) Tamotsu Noguchi, Kentaro Onizuka and Yutaka Akiyama: "PDB-REPRDB: An Interactive Database of Representative Protein Chains from the Protein Data Bank (PDB)", The Seventh International Conference on Intelligent Systems for Molecular Biology (1999).

(13) 秋山 泰, 鬼塚 健太郎, 野口 保, 安藤 誠: "大規模 PC クラスタ上での並列タンパク質情報解析 (PAPIA) システムの構築", 「タンパク質立体構造の分類・予測・デザイン」 研究会予稿集 (1999).

(14) 秋山 泰, 鬼塚 健太郎, 野口 保, 安藤 誠: "大規模 PC クラスタを用いたタンパク質情報解析 (PAPIA) システムの構築", 第 180 回 CBI 研究会 (1999).

(15) 秋山 泰, 鬼塚 健太郎, 野口 保, ポール ホートン, 安藤 誠: "大規模並列処理による構造生物学への挑戦", 第 7 回 SIF 講演会 (1999).

(16) Yutaka Akiyama, Tamotsu Noguchi, Kentaro Onizuka and Makoto Ando: "PAPIA (Parallel Protein Information Analysis) System and MolTreC Parallel Molecular Dynamics Simulator Running on a Compact 8-node Linux PC Cluster", The 10th Genome Informatics Workshop, pp.202-203 (1999).

(17) Yutaka Akiyama, Tamotsu Noguchi, Kentaro Onizuka and Makoto Ando: "A Compact 8-node Linux PC Cluster for Protein Information Analysis" (タンパク質情報解析用小型 8 ノード Linux PC クラスタ), Proc. of the fifth Int. Symp. on Artificial Life and Robotics (AROB 5th '00), pp.729-732 (2000).

目次

図一覧

表一覧

第 1 章

序論

1.1 研究の背景

タンパク質は生物細胞の主要な構成物質で，細胞内の構造形成に関与したり，酵素として様々な生体反応を触媒するなどの機能を果たしている生体高分子である．タンパク質はDNA上にコードされた遺伝子配列がRNA配列に転写され，そのRNA配列に従いリボソーム上でアミノ酸に翻訳され，一次元の鎖状のアミノ酸配列に生合成されたもので．このアミノ酸配列が，折れたたまり，固有の立体構造を形成し，特異的な機能を発現している．この一つながりのアミノ酸配列の鎖はチェインと呼ばれる．タンパク質は，このチェインが単独あるいは複数縮合して存在している．

このようなタンパク質の立体構造を決定している主な要因は，20種類のアミノ酸の配列順序で，「タンパク質の立体構造は，アミノ酸配列より決定される．」と言うアンフィンセンの仮説 [1] に基づき，立体構造が既知であるタンパク質データを用いて，アミノ酸配列から立体構造を予測する研究が盛んに行われている．また，アミノ酸配列を基にした分子進化の研究などから，タンパク質のアミノ酸配列上に保存部位(モチーフ)があり，その部位が機能に関わっていることが確認され，部分的な配列及びその構造と機能の関係が議論されている．

一方，生物の生体情報を蓄えたデータベースは，現在100を越えて，今後もさらに増える傾向にある．主なデータベースとしては，GenBankやSWISS-PROTに代表されるDNAやアミノ酸の配列データベースと，PDBに代表される立体構造のデータベースがある．前者は，近年の配列解析技術の発達と各種ゲノムプロジェクトの展開により，飛躍的にその情報量を増やしているが，後者は，近年のX線結晶解析やNMRによる構造解析技術の進歩，さらに新たに電子顕微鏡による解析結果が加わり急激に増加しているが，立体構造解析の困難さゆえに，その情報量は飛躍的に増加たが，GenBankの配列数と比べるとまだ700分の1程度，SWISS-PROTのエントリー数に比べても7分の1程度のデータ量しかない．各種生物種のゲノムプロジェクトの後を受けて開始した“構造ゲノミックスプロジェクト(ゲノムの中に含まれるタンパク質の立体構造をすべて決める．)”によって，立体構造データの増加は，今後さらに加速すると予想されているが，実験の性質上，配列データのような増加は考えられない．そのため，現在利用できる立体構造データから必要な情報を抽出しなければならない．

現在利用できる立体構造データベースの代表であるPDB(Protein Data Bank)[2, 3]は，米国のブルックヘブン国立研究所が公開を始め，現在は米国のノンプロフィットのコンソーシアムであるRCSB(Research Collaboratory for Structural Bioinformat-

ics) が引き継いで維持しているタンパク質立体構造データベースで，X 線結晶回折や 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 1 エントリーの形式で登録されている．現在世界中にデータの配布と登録を受け持つセンターが存在し，大阪大学 蛋白質研究所が日本におけるセンターの役割を負っている．

図 1.1に PDB のエントリー例を示す．最初の 6 カラムは，その行の内容を示すヘッダで，タイトル部，一次構造部，二次構造部，座標部等に分類され，現在 50 種類ある．それぞれのヘッダーに対応した情報が，9-10 カラム目の継続情報を挟んで，11 カラム目からの記述されている．図 1.1の例では，スペースの都合で主なヘッダーの情報のみを示した．PDB のエントリーは，タンパク質の種類と公開日および PDB のコードが記述されている “HEADER” 行で始まり，“END” 行で終わる．その間に，原子座標が，生物種，実験方法，文献情報，分解能や配列などの付加情報とともに記述されている．

近年，PDB 登録システムによるデータの登録が確立して以来，その記述が統一されて来ているが，それ以前のデータは，個々の解析の登録者が必要に応じて自由に記述して登録していたものが多く見かけられ，エントリーごとに記述にばらつきがある．また，分解能が不十分のため解析が困難で，一部の原子座標が決められず，チェインブレイク[1]が存在したり，Cα のみ，あるいは主鎖原子のみの座標だけしか登録されていないエントリーが存在する．このため，一次構造部の “SEQRES” の残基数と座標部の “ATOM” の残基数が一致しない場合が生じるので，この点にも注意が必要である．

一方，前述のように，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， PDB のデータ量は 1991 年ごろから急激に増加し，その内容は 2000 年 11 月の時点で 13,600 エントリーを越え，さらに増え続けている (図 1.2).

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，

- 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) $\geq$ 70% ，かつ，

[1]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．

```
HEADER    HYDROLASE                                25-SEP-99   1D2M
TITLE     UVRB PROTEIN OF THERMUS THERMOPHILUS HB8; A NUCLEOTIDE
TITLE    2 EXCISION REPAIR ENZYME
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: EXCINUCLEASE ABC SUBUNIT B;
COMPND   3 CHAIN: A;
COMPND   4 SYNONYM: UVRB;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: THERMUS THERMOPHILUS;
SOURCE   3 ORGANISM_COMMON: BACTERIA;
SOURCE   4 STRAIN: HB8;
SOURCE   5 EXPRESSION_SYSTEM_COMMON: BACTERIA;
SOURCE   6 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE   7 EXPRESSION_SYSTEM_PLASMID: PET11A
KEYWDS    MULTIDOMAIN PROTEIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    N.NAKAGAWA,M.SUGAHARA,R.MASUI,R.KATO,K.FUKUYAMA,S.KURAMITSU
REVDAT   1   22-MAR-00 1D2M    0
JRNL        AUTH   N.NAKAGAWA,M.SUGAHARA,R.MASUI,R.KATO,K.FUKUYAMA,
JRNL        AUTH 2 S.KURAMITSU
JRNL        TITL   CRYSTAL STRUCTURE OF THERMUS THERMOPHILUS HB8 UVRB
JRNL        TITL 2 PROTEIN, A KEY ENZYME OF NUCLEOTIDE EXCISION REPAIR
JRNL        REF    J.BIOCHEM.(TOKYO)             V. 126 986 1999
JRNL        REFN   ASTM JOBIAO JA ISSN 0021-924X
.
.
REMARK   2 RESOLUTION. 1.90 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3  PROGRAM      : CNS
REMARK   3  AUTHORS      : BRUNGER,ADAMS,CLORE,DELANO,GROS,GROSSE-
REMARK   3               :  KUNSTLEVE,JIANG,KUSZEWSKI,NILGES, PANNU,
REMARK   3               :  READ,RICE,SIMONSON,WARREN
.
.
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : NULL
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE            (WORKING SET) : 0.234
REMARK   3   FREE R VALUE                     : 0.253
REMARK   3   FREE R VALUE TEST SET SIZE   (%) :  NULL
REMARK   3   FREE R VALUE TEST SET COUNT      : 8293
REMARK   3   ESTIMATED ERROR OF FREE R VALUE  : NULL
REMARK   3
.
.
SEQRES   1 A  665  MET THR PHE ARG TYR ARG GLY PRO SER PRO LYS GLY ASP
SEQRES   2 A  665  GLN PRO LYS ALA ILE ALA GLY LEU VAL GLU ALA LEU ARG
SEQRES   3 A  665  ASP GLY GLU ARG PHE VAL THR LEU LEU GLY ALA THR GLY
.
.
HELIX    1   1 ASP A   13  ILE A   18  1                                   6
HELIX    2   2 GLY A   20  ASP A   27  1                                   8
.
.
HELIX   25  25 SER A  552  GLY A  577  1                                  26
SHEET    1   A 7 ALA A  82  TYR A  85  0
SHEET    2   A 7 VAL A 133  SER A 138  1  O  ILE A 134   N  GLU A  84
.
.
SHEET    6   F 6 ALA A 467  LEU A 470  1  O  ARG A 468   N  VAL A 495
.
.
ATOM      1  N   THR A   2      42.615  18.433  30.439  1.00 38.52           N
ATOM      2  CA  THR A   2      43.827  17.779  29.866  1.00 38.11           C
ATOM      3  C   THR A   2      43.773  17.788  28.345  1.00 37.22           C
ATOM      4  O   THR A   2      42.793  17.339  27.749  1.00 37.29           O
ATOM      5  CB  THR A   2      43.949  16.312  30.331  1.00 38.53           C
ATOM      6  OG1 THR A   2      43.964  16.263  31.763  1.00 39.71           O
ATOM      7  CG2 THR A   2      45.234  15.688  29.795  1.00 38.76           C
ATOM      8  N   PHE A   3      44.832  18.296  27.722  1.00 35.62           N
ATOM      9  CA  PHE A   3      44.902  18.360  26.270  1.00 35.24           C
.
.
ATOM   4459  CG2 VAL A 583      60.790  24.365   1.539  1.00 40.91           C
TER    4460      VAL A 583
.
.
END
```

図 1.1: PDB のエントリー例

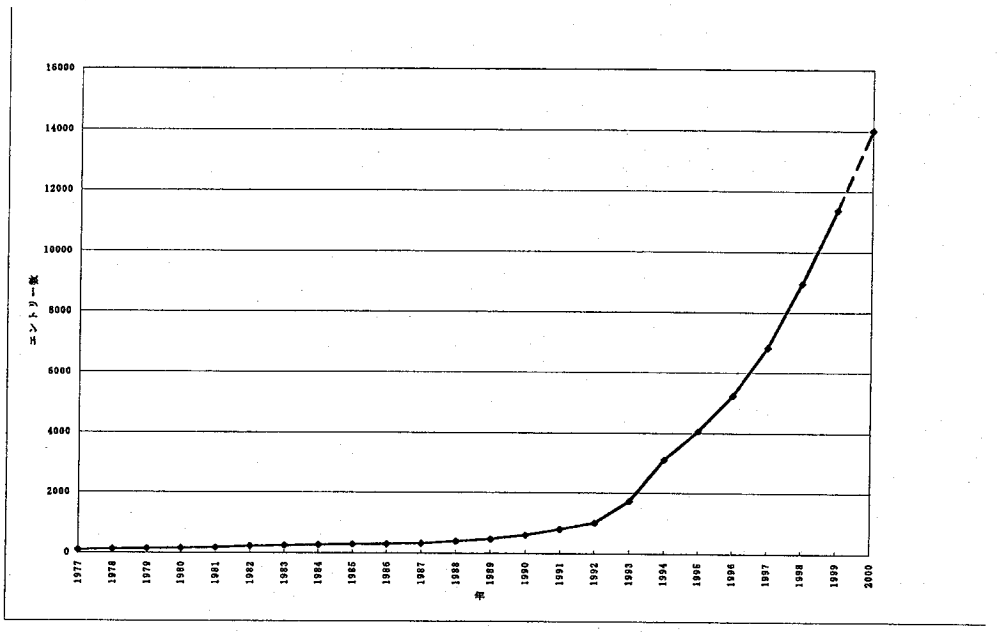

図 1.2: PDB のエントリー数の推移

- 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) $\leq$ 10.0 Å,

を採用すると実に全エントリーの約 80% 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[2]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．

立体構造が既に明らかなタンパク質の配列と立体構造の関係を調べ，未知の立体構造を予測する経験的立体構造予測法の研究では，前述の近縁タンパク質を無視して統計をとると情報の偏りを生じてしまい，誤った予測をする可能性が高い．そのため，一定の基準 (たとえば，配列の相同性：ID% $<$ 30%) で近縁タンパク質の代表を選んでおくことが，この種の研究を進める上できわめて重要である．このような用途での "代表点" は，比較的遠い関係のタンパク質もカバーする "半径の大きな" ものとなる．

他方，よく類似した配列の立体構造をもとにタンパク質の未知立体構造をモデリングしたい場合には，別の基準 (たとえば，配列の相同性：ID% $<$ 95%) で近縁タンパ

[2]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．

ク質の代表点を決めておくことが有益である．この場合の “代表点” は “半径の小さな” ものとなり，よく類似したタンパク質の中で良質の構造が選ばれる．これにより，近接した良質な立体構造を選んでモデリングを始めることができる．

このような需要のもとに Hobohm らは，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た．この代表タンパク質チェインは，“PDB_SELECT” [4, 5] として公開され，現在では，配列の相同性：ID% $< 25\%$ と $< 95\%$ の基準のリストが用意されており，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 の立体構造の類似したチェインに対して構造アライメントして登録したデータベース (FSSP[6]) を作成し，公開している．また，水口らも PDB の X 線結晶回折による解析データの中で，配列相同性のあるタンパク質を構造アライメントしたデータベース “HOMSTRAD”[7] を作成している．

また，Sander らは，代表チェインは決めていないが，PDB と SWISS-PROT の相同配列をアライメントした配列データベース (HSSP[8]) 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．

“PDB_SELECT” や “HSSP”，“FSSP” 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データベースとして，“SCOP”[9] や “CATH”[10] がある．SCOP は，all-α，all-β，α/β，$\alpha+\beta$ などの構造クラスに分類した後，それらをさらに折れ畳みのタイプ別に分類して，そこから配列の相同性を調べ，ファミリー分類を行なっている．CATH は，タンパク質のドメイン[3] 構造を分類したデータベースで，ドメインを SCOP のように構造分類している．SCOP と CATH は，ともに立体構造の全体構造の分類を行なったデータベースで，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，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存在しなかった．

[3] ドメイン: チェインを構成する部分構造で，タンパク質の折れ畳みの単位と考えられている．

1.2 研究の目的と効果

従来のタンパク質立体構造分類の方法には，大きく分けて以下の2種類があった．

1) 配列相同性による分類

2) 配列相同性＋立体構造全体の類似性

前者は，配列相同性によって，ファミリー分類を行うことにより，近似的に立体構造を分類する方法である．そのため，部分構造の違いを正確に考慮して分類することは困難である．後者は，タンパク質のファミリー分類やフォールタイプの分類を目的としたもので，立体構造全体の類似性は見ているが，部分構造の違いを考慮せず分類しているため，この方法でも，部分構造の違いを正確に考慮して分類することは困難である．

本研究の主な目的は，PDB内のタンパク質チェインの立体構造を分類し，代表タンパク質チェインを決めることにより，非冗長なPDB代表タンパク質チェインデータベース (PDB-REPRDB) を作成することである．その際，従来法では考慮されていなかった，部分構造の違いを検出し，部分的に立体構造が異なるチェインを別の代表点とすることにより，より正確な構造分類を可能する．また，タンパク質立体構造予測などの研究に利用することを考慮して，グループの中心構造を代表にするのではなく，分解能が良く，チェインブレイクが少ないなど，できるだけ信頼度の高いデータを優先して代表に選ぶことにする．本研究では，PDBデータの質にばらつきがあることを考慮し，データの質に応じてPDBのタンパク質チェインを順位付けし，その上位にあるチェインを代表に選ぶ方法を採用した．

本研究の初期段階では，あらかじめデータの質に関するデータ項目を決め，それらに代表を選ぶ際の優先度を与えることで，PDBデータの順位付けを行い，代表タンパク質チェインを決定していた．しかしながら，研究内容によって，それらの優先度は変わるので，最終的には利用者が，優先度を変更できるようにした．また，本研究の初期段階では，あらかじめ決めた分類基準で作成したPDB-REPRDBを複数用意して公開しているだけであったが，これも研究内容によって，それらの基準が変わるので，最終的には利用者が，それも変更して，利用者が得たい基準でPDB-REPRDBを作成できるようにした．

これにより，配列相同性と立体構造の類似性を実際に比較しながら，タンパク質チェインを分類することが可能となり，経験的立体構造予測法の研究の基となる立体構造

のデータセットを正確に決めることが可能になる．また，部分構造が異なるチェインを別の代表とすることが可能になり，特徴的な部分構造を効率よく網羅的に調査する場合のデータベースを容易に作成することが可能になる．

また，WWW を用いた会話型 PDB-REPRDB 作成システムは，研究内容によって，研究者が欲しい代表チェインセットを数分で提供することを可能とし，タンパク質立体構造予測法の研究やタンパク質立体構造の特徴抽出などの研究に貢献するものと期待している．

1.3 本論文の構成

本論文は以下の章によって構成される.

第1章の序論では，本研究の背景と目的を述べた後，特に本研究の重要性や従来法の問題点について述べた.

第2章では，本研究の基礎となった配列相同性と構造類似性を考慮したPDB代表タンパク質チェインデータベース (PDB-REPRDB) を作成するPDB代表タンパク質チェイン決定システムについて述べる.

第3章では，第2章で述べたPDB代表タンパク質チェイン決定システムを改良し，並列化することによって，処理の高速化を実現したので，その並列化手法とその結果について述べる.

第4章では，WWWと並列化技術を利用することにより，利用者が得たい基準のPDB代表タンパク質チェインデータベース (PDB-REPRDB) を，会話形式で作成できるシステムを構築したので，そのシステムについて述べる.

第5章では，本研究で得られたPDB代表タンパク質チェインデータベース (PDB-REPRDB) がどのように利用されるか具体例を示して紹介する.

第6章の結論では，本研究のまとめと今後の課題について述べる.

第 2 章

PDB 代表タンパク質チェイン決定システム

2.1 はじめに

経験的な手法を用いたタンパク質立体構造予測の研究を行う際に，常に先立つ問題として起るのが，その基本となるタンパク質立体構造をどうやって選ぶかである．序章で述べたように，配列相同性を用いて近似的に立体構造を分類し，代表を決めたデータベースは既に存在していたが，タンパク質立体構造予測に用いるデータとしては，近似を用いていると言う意味で不十分であった．本来は，配列相同性があっても，立体構造が大幅に異なるタンパク質であれば，別の代表にして，そこから情報を得るべきである．

また，部分構造の特徴抽出を行う際にも，従来法では，部分構造の違いを考慮せずに代表が決められているため，それらを見逃してしまう可能性が高い．

本章では，分類の指標として配列相同性に構造の類似性を加えた新たなPDB代表タンパク質チェイン決定システムを構築したので，そのシステムおよびシステムで得られたPDB代表タンパク質チェインデータベース (PDB-REPRDB) について述べる．

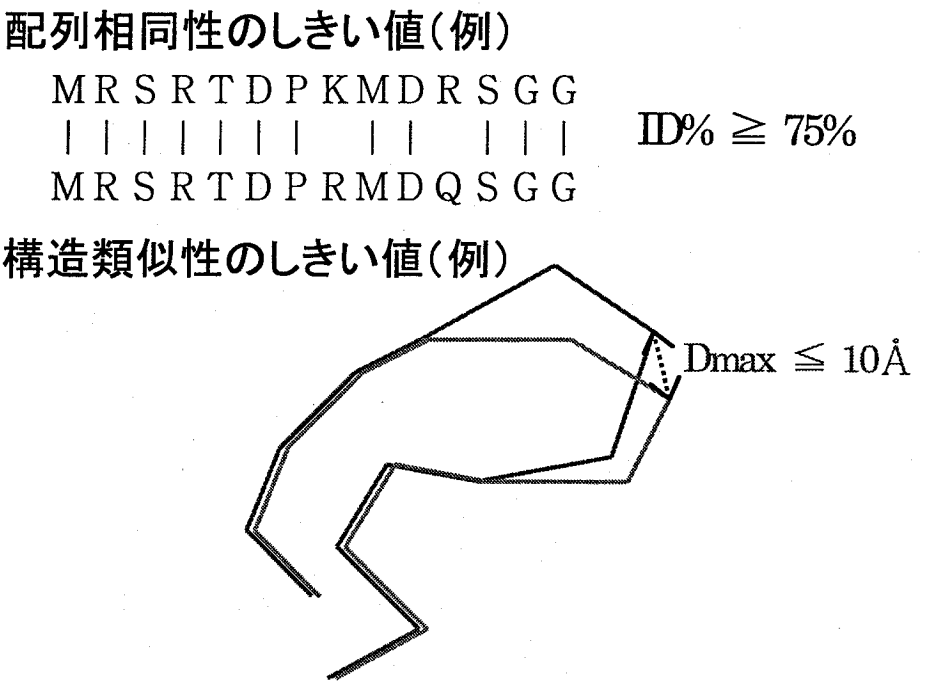

図 2.1: 近縁タンパク質の基準

2.2 PDB 代表タンパク質チェイン決定システム

2.2.1 タンパク質立体構造の分類

本研究における分類の基準は，二次構造や活性部位などの部分構造を対象とした研究に利用するために，より多くの特徴ある部分構造を含んだ代表タンパク質チェインを決定する必要があったので，

- 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 (図 2.1).

タンパク質の全体構造を比較する場合には，全体構造を重ね合わせた時の RMSD を基準にするのが一般的であるが，全体構造を重ね合わせた時の RMSD は，構造が類似している部分の原子間距離が小さいため， Dmax の値より小さな値になる．このように RMSD は，構造全体の類似性を示す基準としては適しているが，構造の一部

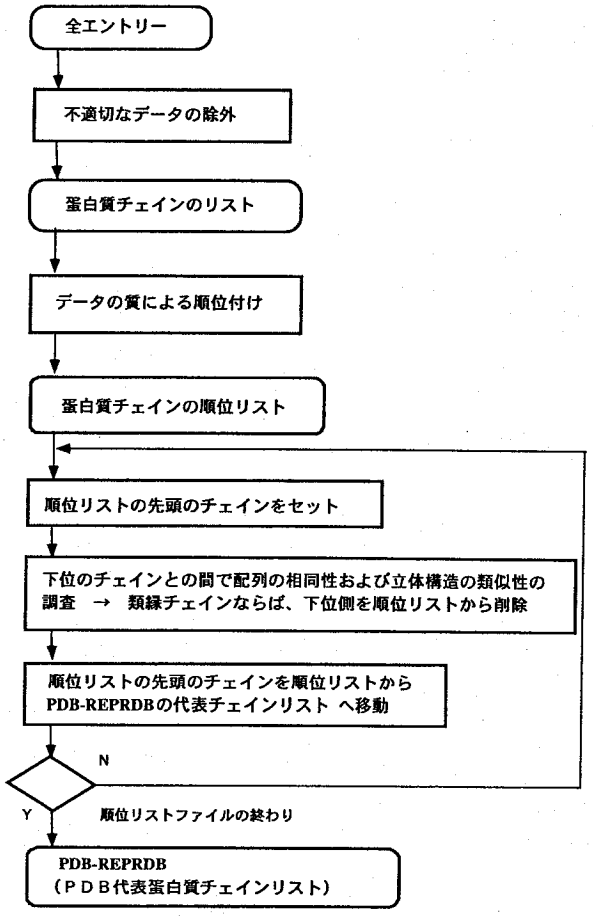

図 2.2: PDB 代表タンパク質決定システムの流れ

だけが異るタンパク質同士を比較し，その違いを検出しようとすると，その部分以外の類似している構造の影響を受けるため，部分構造の違いを正確に検出する基準としては適していない．また，全体構造を重ね合わせた時の RMSD は，重ね合わせた原子数に依存するので，分類の指標となるしきい値を決める時に，重ね合わせた原子数を考慮しなければならない．実際に重ね合わせる原子数は，同じタンパク質であっても，その相手によって異なるため，RMSD のしきい値はその相手ごとに異なる値にする必要が生じてしまう．以上のように，部分構造の違いの検出に適している点と，RMSD を用いると決められたしきい値で分類することが困難になるため，Dmax を分類の指標とした．

PDB-REPRDB は，PDB を基に，図 2.2に示した手順で作成する．

2.2.2 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) NMR で解析されたデータ

c) 理論計算だけで求められたモデルデータ

d) チェインの長さが短いデータ ($l < 40$ 残基)

e) 全ての残基において主鎖座標が欠落したデータ

f) 全ての残基において側鎖座標が欠落したデータ

f) リファインメントされていないデータ

NMR で解析されたデータは，現状では，X 線結晶回折によって解析された立体構造と比較することの妥当性に疑問があり，また，構造比較をするモデル選びの方法が決められないので，あらかじめ分類対象から削除した．

2.2.3 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ない，順位リストを作成する．始めに準備として，X 線結晶回折によって構造解析されたデータを，分解能が 3.0 Å 以下かつ R ファクターが 0.3 以下の質の高いチェインと，それ以外のチェインに分類し，前者をクラス A，後者をクラス B とする．データの優先度は，クラス A ＞ クラス B とする．

クラス A とクラス B のチェインは，それぞれのクラス内で，まず分解能，次に R ファクターの小さい順に並び替えられ，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．

(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) 変異型と野生型 (野生型が上位)

(6) 単量体と複合体 (単量体が上位)

(7) チェイン名のアルファベット順 (若いほど上位)
(例： 1MCD > 1MCE, 5AT1A > 5AT1C)

2.2.4 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，各クラスごとにデータの良質度でソートされたリストが得られるので，クラス A ～ B の 2 クラスを合わせて，1 つの順位リストを作成する．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った．

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，相同配列検索プログラム (FASTA)[11, 12] で，それ以下のチェインの配列相同性を調べる．あらかじめ決めた相同性しきい値 (95%，85%，75% など) 以上であれば，さらに，構造類似性のチェックを行なう．FASTA によるペアワイズアライメントの結果，同じ残基種で並置された (例えば，図 2.1の “配列相同性しきい値 (例)” で線で結ばれた) 残基ペアの C_α 原子同士を，Kabsch による最小 2 乗フィット法 [13] により重ね合わせ，重ね合わせた原子間距離の最大値 (Dmax) を求める．この Dmax があらかじめ決めたしきい値 (10 Å) 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．

この処理を順にリストの最後まで行なうことにより，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．

表 2.1: 配列相同性 (ID%): 75% による PDB_SELECT と PDB-REPRDB の数の比較

	Number of chains	
	PDB_SELECT (11 Nov. 1996)	PDB-REPRDB ver.2.1 (1997)
Total	1255	1064
X-Ray Data ($l \geq 40$)	1025	1064
X-Ray Data ($l < 40$)	38	0
NMR Data	191	0
Other Data	1	0

注) PDB-REPRDB は，PDB Release 78 (Oct. 1996) で作成. "l" は，主鎖原子の座標データを持つ残基の数.

2.3 PDB_SELECT との比較

本研究の PDB-REPRDB と Hobohm らの PDB_SELECT との違いを明らかにするために，両者の比較を行った．表 2.1に，ID% のしきい値が 75% の代表チェイン数を実験方法ごとに示す．PDB-REPRDB は，X 線結晶回折のデータしか利用していないので，全体の代表チェイン数では少ないが，X 線結晶回折のデータだけで比較すると，ほぼ同数の代表が選ばれた．選ばれたタンパク質の "ID"(PDB エントリ ID + チェイン ID) を比較した結果では，890 チェインが同一であった．この結果，配列相同性の基準では，代表の選び方に違いはあるが，ほぼ同様の代表チェインを選んでいることが確認された.

配列相同性 (ID%) のしきい値を変化させた時に，代表チェイン数がどのように変わるか調べた結果を，表 2.2に示す．配列相同性 (ID%) のみを考慮した場合 (すなわち，Dmax: ∞ Å)，PDB_SELECT の代表チェイン選出の方法が，最も多くのチェインの代表チェインを選ぶアルゴリズムを採用している関係で，ID% のしきい値が 25% の時に，PDB_SELECT の代表チェイン数が PDB-REPRDB の代表チェイン数より小さくなるが，それ以外では，PDB-REPRDB には NMR で解析されたデータが含まれないため，PDB_SELECT の代表チェイン数が多くなっている．ただし，ID% のしきい値が 95% の時にその差が縮まっている．これは，相同性の範囲が狭まったため，

表 2.2: 配列相同性 (ID%) と構造類似性 (Dmax) による PDB_SELECT と PDB-REPRDB の数の比較

Threshold sequence identity (ID%)	Number of chains		
	PDB_SELECT (6 May 1997) (Dmax: ∞ Å)	PDB-REPRDB ver.3.0 (1997) (Dmax: ∞ Å)	PDB-REPRDB ver.3.0 (1997) (Dmax: 10 Å)
25	635	645	1102
35	998	897	1147
45	1140	1008	1208
55	1255	1113	1280
65	1368	1209	1346
75	1450	1293	1437
85	1590	1436	1553
95	1765	1695	1786

注) PDB-REPRDB は，PDB Release 80 (April 1996) で作成.

上記のような PDB_SELECT のアルゴリズムで代表を決める方法と単純にデータの質の良いものを代表チェインに決める方法で，相同性の中心になる代表チェインの配列に差がなくなったためで，Dmax のしきい値が，10 Å の時に，その数が逆転しているのもそれが主な原因である.

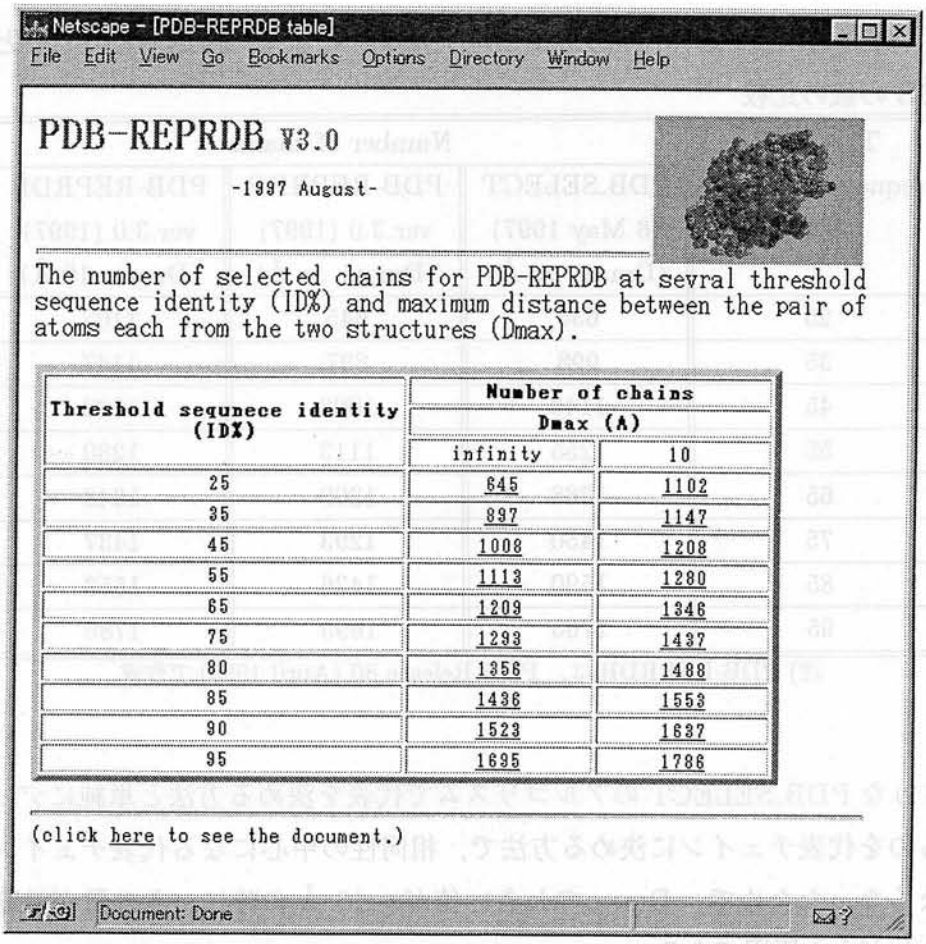

図 2.3: WWW 上の PDB-REPRDB

2.4 PDB 代表タンパク質チェインデータベース (PDB-REPRDB) の公開

本システムにより，あらかじめ決めた配列相同性 (ID%) と構造類似性 (Dmax) のしきい値で，PDB 代表タンパク質チェインデータベース (PDB-REPRDB) を作成し，WWW 上で公開した (図 2.3).

本ホームページは，新情報開発機構 つくば研究センタ 並列応用つくば研究室で公開している PAPIA システム (URL: http://www.rwcp.or.jp/papia/) [14] のサーバー上に置かれ，ゲノムネットの WWW サーバー (URL: http://www.genome.ad.jp/dbget) とリンクしていた.

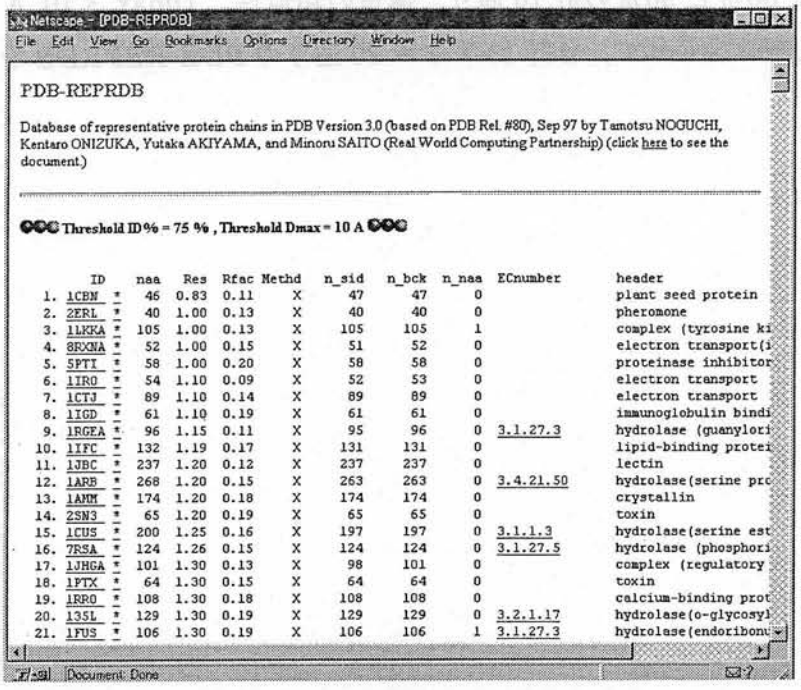

図 2.4: WWW 上の PDB 代表タンパク質チェインリストの例

ホームページ (図 2.3) では，あらかじめ決められた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ており，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 2.4のような，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．リストには，選ばれた代表タンパク質の ID 名 (エントリー名＋チェイン ID)，残基数，分解能，R ファクター，実験方法，側鎖原子の座標を持つ残基数，主鎖原子の座標を持つ残基数，非標準アミノ酸の残基数，EC(酵素) 番号，タンパク質の分類名 (HEADER) が記されている．また，ID 名の部分は，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，クリックすると類似タンパク質のリストを見ることができる．また，代表タンパク質チェインのリストの ID 名と残基数の間に表示されている “*” 印をクリックすると RasMol プログラムを用いた立体構造がグラフィック表示される．類似タンパク質チェインリストの ID 名は，さらに 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．

本システムでは，自動化が不完全であったため，多くの基準で代表タンパク質チェインセットを作成できなかったが，配列の相同性：ID% $\geq$ 25% $\sim$ 95% まで 10% 刻

みの 8 通りと 80% と 90% の計 10 通り，構造の類似性：Dmax $\leq$10 Å と ∞ Å の 2 通りの基準を組み合わせて，合計 $10 \times 2 = 20$ 通りの PDB-REPRDB を作成し，公開した．

2.5 まとめ

本章では，従来の配列相同性に基づくタンパク質立体構造の分類手法およびその代表タンパク質を決定する方法に，新たに構造類似性にも着目し，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分類手法を提案し，その手法を用いたPDB代表タンパク質チェイン決定システムを作成した．本システムの検証は，本システムで作成した代表タンパク質チェインと，従来法であるPDB_SELECTの代表タンパク質チェインと比較して行い，結果の妥当性を確認した．

本システムにより，従来法では配列相同性を用いて近似的に立体構造を分類し，代表を決めていたため，タンパク質立体構造予測に用いるデータとしては，不十分であった代表タンパク質チェインデータを，直接立体構造を比較し，分類することによって，近似によらない分類を可能にし，正確な代表タンパク質チェインデータが得られるようになった．また，従来法では，見逃してしまう可能性が高かった特徴ある部分構造も，本システムで代表タンパク質チェインを選ぶことによって，見逃すことなく効率的に調査できるデータベースを作成できるようになった．

本システムは，自動化が不十分で，かつ処理に膨大な時間を要したため，研究者の要求に応えられるような，様々な分類基準で作成した代表タンパク質チェインセットを用意することができなかった．また，本システムにおいては，X線結晶回折によって解析された立体構造と比較することの妥当性に疑問があり，また，構造比較をするモデル選びの方法が決められず，NMRによって解析された立体構造をあらかじめ分類対象から削除していた．しかしながら，NMRデータが無視できないほどPDBに登録されて来ており，今後も増える傾向にあるので，今後は，NMRによって解析された構造も分類に含めることにする．

第 3 章

PDB 代表タンパク質チェイン決定システムの並列化

3.1 はじめに

第2章では，配列相同性と立体構造の類似性を考慮したPDB代表タンパク質チェイン決定するシステムを作成した．しかしながら，まだ自動化が不十分でかつ計算に膨大な時間を要するため，研究者の要求に応えられるような，様々な分類基準で作成した代表タンパク質チェインデータベース (PDB-REPRDB) を用意することができなかった．また，X線結晶回折によって解析された立体構造と比較することの妥当性に疑問があり，また，構造比較をするモデル選びの方法が決められなかったため，NMRによって解析された立体構造データはあらかじめ分類対象から削除していた．しかしながら，現在NMRのデータは無視できないほどPDBに登録されて来ており，今後も増える傾向にあるため，今後は，NMRによって解析された構造も分類に含める必要がある．PDBデータの増加に加え，NMRのデータを加えることにより，分類する立体構造はさらに増加し，PDB-REPRDBを作成するために要する処理時間は，膨大になると予想された．

本章では，これらの問題を解決するために，PDB代表タンパク質チェイン決定システムのさらなる自動化を進めるとともに，処理の高速化を目指して，システムの並列化を行ったので，並列化されたPDB代表タンパク質チェイン決定システムについて述べる．また，このシステムにより，配列相同性と構造類似性の組合わせで，合計 $8 \times 6 = 48$ 通りの基準で分類し，それぞれの基準での代表タンパク質チェインのセットを決定したので，その結果についても述べる．

3.2 PDBの代表タンパク質決定システム

本研究のPDBの代表タンパク質決定システムの処理の流れは，第2章で述べたシステムとほぼ同じであるが，システムの自動化を進め，並列化を考慮したシステム作りを行うために，鬼塚らによるPAPIAライブラリ[15, 16]と呼ぶオブジェクト指向の共通プログラムライブラリを導入することにした．PAPIAライブラリでは，タンパク質の構造データや配列データが，C++言語上のクラス構造として明確に定義されている，例えば，proteinクラスのオブジェクトは，タンパク質立体構造を表し，複数のchainオブジェクトを持っている．各chainオブジェクトはアミノ酸残基を表すresidueのリストで構成され，各residueは，atomオブジェクトから構成されている．タンパク質の階層的構造が，このようなオブジェクト指向言語に適していると言える．また，タンパク質立体構造解析で良く用いられる立体構造の回転や移動などの操作が，ライブラリに用意されており，立体構造同士の重ね合わせと言った複雑な操作もそのライブラリを使うことによって簡単に行える．また，タンパク質配列のペアワイズアライメントもライブラリに用意されていて，本システムの作成の効率化に大きく貢献した．

3.2.1 不適切なデータの除外

本システムから，“NMRで解析されたデータ”を分類に含めることにしたため，第2章で除外していた“NMRで解析されたデータ”と“リファインメントされていないデータ”を本処理から除いた．したがって，PDBのエントリーをまずチェイン単位に分離したのち，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

3.2.2 データの質による順位付け

本システムから，“NMR で解析されたデータ” を分類に含めることにしたため，下記のように新たにクラス C を設け，処理を行った．

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ない，順位リストを作成する．始めに準備として，X 線結晶回折によって構造解析されたデータを，分解能が 3.0 Å 以下かつ R ファクターが 0.3 以下の質の高いチェインと，それ以外のチェインに分類し，前者をクラス A，後者をクラス B とする．また，X 線結晶回折以外の構造解析技術 (NMR など) で構造解析されたデータを，クラス C とする．データの優先度は，クラス A > クラス B > クラス C とする．

クラス A とクラス B のチェインは，それぞれのクラス内で，まず分解能，次に R ファクターの小さい順に並び替えられ，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．クラス C に関しては，NMR のデータだけを抽出し，(NMR には，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) 同様に下記の項目を順に調べて順位付けを行なう．

(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) 変異型と野生型 (野生型が上位)

(6) 単量体と複合体 (単量体が上位)

(7) チェイン名のアルファベット順 (若いほど上位)
(例： 1MCD > 1MCE, 5AT1A > 5AT1C)

3.2.3 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，各クラスごとにデータの良質度でソートされたリストが得られるので，クラス A ～ C の 3 クラスを合わせて，1 つの順位リストを作成する．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[17]で調べる．第2章のシステムでは，ここの処理はFASTAを使用していたが，本システムでは，PAPIAライブラリ内に並列化されたペアワイズアライメント法のプログラムが用意されているので，それを利用した．

ペアワイズアライメントの結果，その相同性がしきい値以上であれば，さらに構造類似性のチェックを行なう．ペアワイズアライメントの結果において，同じ残基種で並置された残基ペアのC_α原子同士を，Kabschによる最小2乗フィット法[13]により重ね合わせ，重ね合わせた原子間距離の最大値(Dmax)を求める．このDmax値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．

この処理を順にリストの最後まで行なうことにより，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．

上記の処理の流れは，第2章で述べたシステム(図2.2)と同じである．

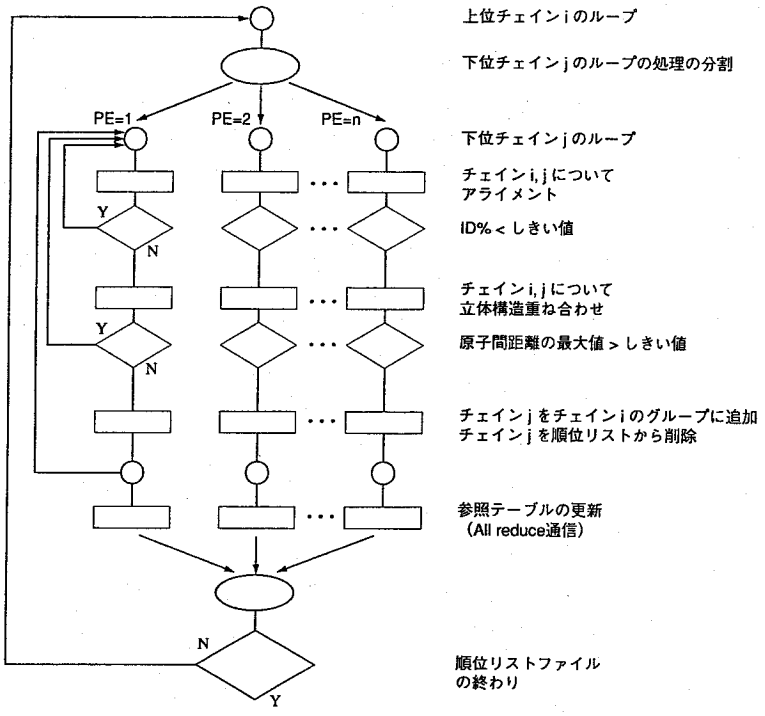

図 3.1: 並列版 PDB 代表タンパク質決定システムの流れ

3.3 代表タンパク質決定システムの並列化実装

PDB 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は，PDB 代表タンパク質決定システムの処理をさらに高速化する必要がある．そこで我々は，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．

Sun SPARC center2000E (Super SPARC II 85MHz, メモリ 5GB) で動作していた逐次版システムにおいて，1,000 チェインの分類で，約 8 時間半かかっていた処理時間のうち，“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 2.2におけるループ) の内部の処理に約 8 時間を要していた（それ以外の処理で主なものは，ハードディスクへの I/O である）．その部分において，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 3.1).

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，

$$p = (\lfloor \frac{i}{k} \rfloor \bmod n) + 1 \tag{3.1}$$

で決定される．ただし k はブロック幅 (今回は 1)，n は使用するプロセッサ台数とする．

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_α 原子部分だけでも大きい (約 120 Mbytes) 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することと，以降は図 3.1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．

上記の処理の計算量については，配列間アライメントが配列長の二乗のオーダ，重ね合わせが一乗のオーダであるが，それぞれのタンパク質の配列長のバラツキが大きいため，計算時間は配列ペアごとに大きく変わる．システム全体では，チェイン数 m に対して二乗のオーダとなる．配列および構造の類似性のしきい値，およびデータベースの内容によって，削除されるチェイン数が異なり，計算量が変動する．

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，静的割り当てをしているため，必ずしも充分には均一化されていない．

表 3.1: 並列 PDB 代表タンパク質決定システムの性能評価を行なった SR2201 の仕様

プロセッサ数	256 計算ノード +16 I/O ノード
プロセッサ	HARP-1E 150 MHz 0.3 GFLOPS
ネットワーク	3 次元クロスバネットワーク 300MB/sec
総メモリー容量	256MB × 256=64GB
ローカルディスク	なし

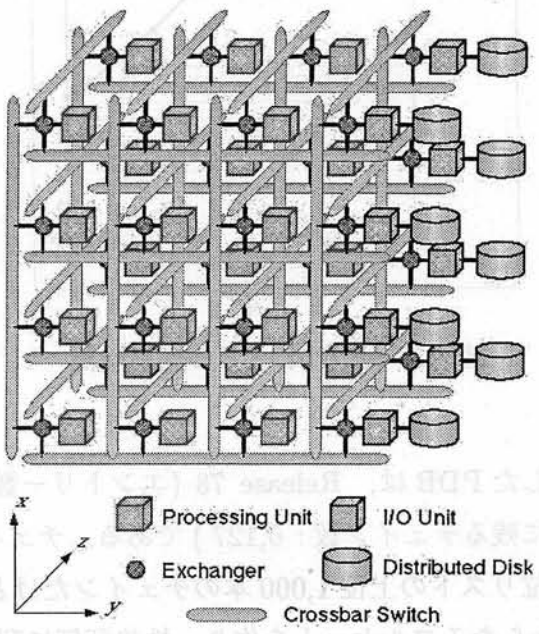

図 3.2: 3 次元クロスバネットワーク概念図

3.4 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 3.1は実験に用いた SR2201 の仕様である． SR2201 の特徴は，各ノード間の通信が効率よく行えるように設計された 3 次元クロスバネットワークにあり，これにより高い転送スループット性能を実現している (図 3.2).

性能評価テストは，システム作成時と PDB エントリー数の増加による影響を調べるためにチェイン数が約 2 倍になった時期とで，計 2 回行った．

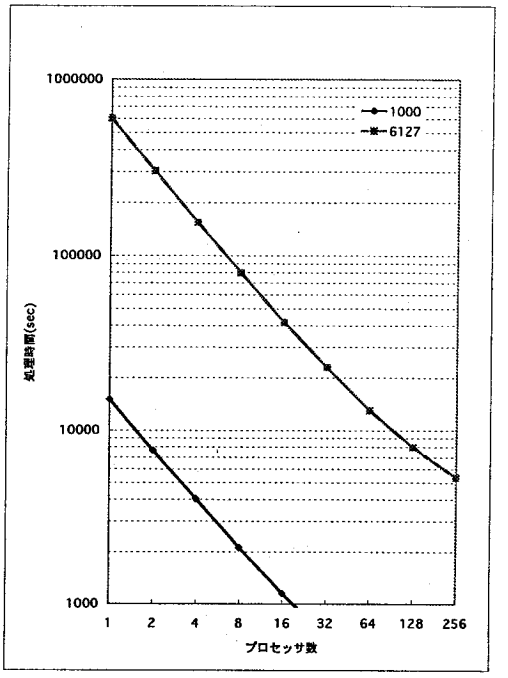

図 3.3: SR2201 上での処理時間

最初のテストで使用した PDB は， Release 78 (エントリー数: 4,873，全チェイン数: 8,870，順位リストに残るチェイン数 : 6,127) 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 1,000 本のチェインだけとったサブセットと，全 6,127 本のチェインからなるフルセットを作り，性能評価に利用した．

図 3.3に SR2201 での処理時間，図 3.4に速度向上比を示す．順位リストのチェイン数が 1,000 本の場合と 6,127 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．

順位リストのチェイン数が 6,127 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 時間で順位リストの 6,127 本のチェインを分類することができた．この時点では，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，各プロセッサが担当しなければならない計算の粒度は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た．

2 回目のテストで使用した PDB は， Release 84 (エントリー数: 7,578，全チェイン数: 11,257，順位リストに残るチェイン数 : 11,062) である．今回は全 11,062 チェインのフルセットのみを作り，性能評価に利用した．

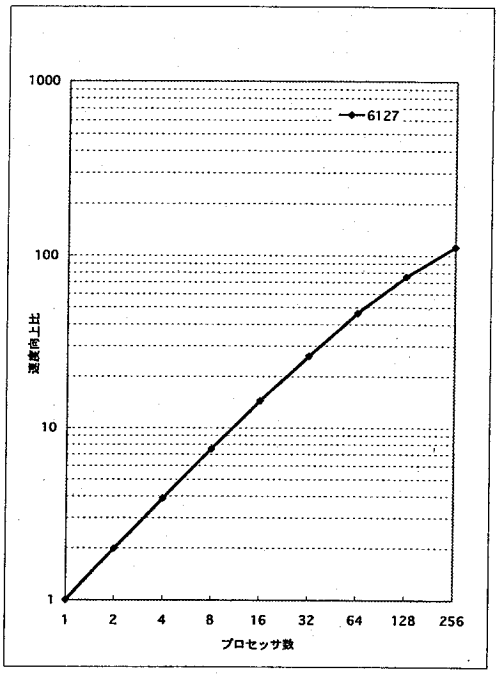

図 3.4: SR2201 上での速度向上比

前回の性能評価テストの結果を含む，処理時間と速度向上比を図 3.5と図 3.6に示す．16 プロセッサ以下では，全ての場合でほぼ同様の性質を示しているが，そのプロセッサ数を越えると，順位リストのチェイン数が 1,000 本の場合と 11,062 本の場合で並列効率が悪化している．その理由は，6,127 本の場合では，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が，1,000 本の場合，プロセッサ数が増えると各プロセッサが担当する計算量が少なくなり，計算粒度が小さくなるためと考えられる．また，11,062 本の場合では，各プロセッサが担当する計算量は多くなるが，参照テーブルの更新時に行われる通信量およびその回数が増えることによって，計算粒度が小さくなるためと考えられる．

前回の性能評価テストでは，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，各プロセッサが担当しなければならない計算の粒度は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していたが，今回のテストで，参照テーブルの更新時に行われる通信量を下げるための改良などを行い，並列効率を高める必要があることが明らかになった．

現状（11,062 本の処理）で，PDB-REPRDB を 1 セット作成するのに約 7 時間を要

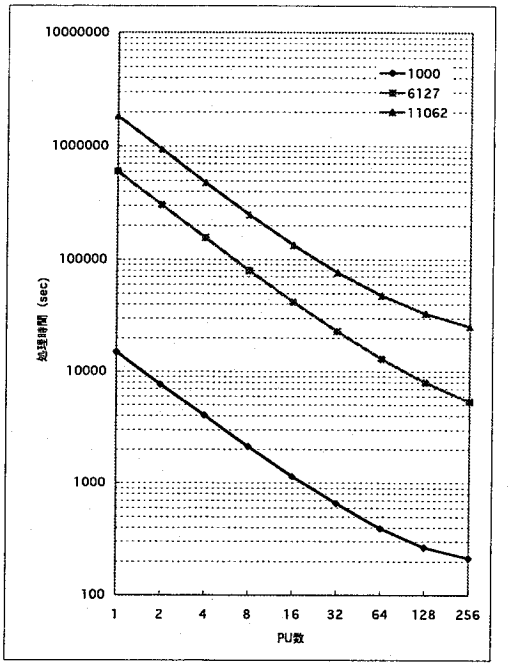

図 3.5: SR2201 上での処理時間 (2)

することから，それを 48 通り作成するのに要する時間は，約 7 時間× 48 = 約 336 時間（約 2 週間）である．現状の速度向上比 (約 73) を，6,127 本の処理での速度向上比（約 110）にすることができれば，約 4 時間半で PDB-REPRDB を 1 セット作成することができ，全体の処理時間を約 214 時間 (9 日) に短縮することが可能である．

しかしながら，オンラインによる PDB の更新が可能になり，PDB を毎日更新できる環境を考えると，PDB-REPRDB の更新を 1 週間以内で行うシステムが理想であり，今後の PDB の増加も考慮すると，抜本的なシステムの改良が必要である．

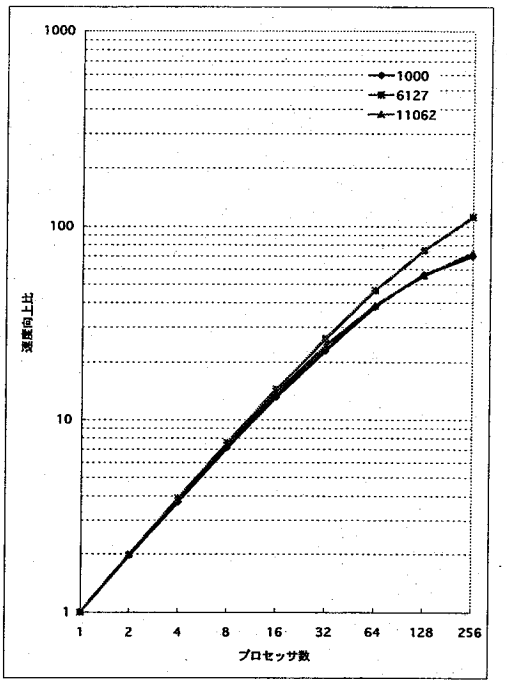

図 3.6: SR2201 上での速度向上比 (2)

3.5 結果

本研究で分類実験に使用した PDB は，1998 年 4 月版の Release 84 (表 3.2, 表 3.3) である．

“不適切なデータの除外”(3.1 節) や “データの質による順位付け”(3.2 節) 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，11,062 本であった (表 3.4)．

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% $\sim$ 95% まで 10% 刻みの 8 通りと，構造の類似性：Dmax $>$10 Å $\sim$ 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 $8 \times 6 = 48$ 通りの代表タンパク質を決定した (表 3.5，図 3.7)．

図 3.7を見てまず気がつくことは，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% のしきい値によらず常に大きいことである．最も差が縮まる ID% $<$ 95% の場合でも，175 (=2,812-2,637) 本のチェインを別のチェインとして分類している (表 3.5)．このことから，ID% $\geq$ 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，

表 3.2: PDB Release 84 の内容 (分子タイプによる分類)

分子タイプ	数
タンパク質，ペプチド，ウイルス	6,723
タンパク質，核酸の複合体	298
核酸	545
炭水化物	12
計	7,578

表 3.3: PDB Release 84 の内容 (解析法による分類)

実験法	数
理論モデル	183
NMR	1,191
X 線結晶回折	6,204
計	7,578

Dmax $>$ 10Å をこえる部分構造の変化があることがわかる．また，ID% $<$ 25% では，ほぼ半数の 815 (=1,689-874) 本のチェインを構造の差異によって別のチェインとして分類している (表 3.5)．この結果を見ると，タンパク質の部分構造の解析を行なう場合，ID% $<$ 25% の配列相同性だけを考慮した代表タンパク質チェインを用いたのでは，他の多くの有用な構造データを使わずに解析していたことがわかる．

図 3.7を見て次に気づくのは，ID% $<$ 25% と ID% $<$ 35% の間で，選ばれた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 3.5の Dmax $>$ 20 Å と ∞ の列を比較すると，ID% $<$ 35% では，75 (=1,455-1,377) 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 (=1,176-874) 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．

最後に，構造の差異を見ることが重要であることを実例をもって示す．図 3.8は，

表 3.4: データの質によるクラス分け

	数
総チェイン	11,257
Class A	9,105
Class B	1,110
Class C	1,042
順位リストのチェイン	11,062
Class A	9,105
Class B	1,110
NMR	847

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ている例である．図 3.8 [18] 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 95.0% あるが，L チェインの C 末端にある β シート構造が I チェインではほどけてしまっている．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ずれており，Dmax の値が 61.6 Å と非常に大きな値になっている．RMSD 値も 9.1 Å と比較的大きな値であるが，ずれていない (対応する原子間距離の小さい) 部分の影響を受け，Dmax の値より小さい値になっている．このことは，部分構造の違いを検出する指標として，Dmax 値の方が RMSD 値より適していることを示している．

上記のように，タンパク質の立体構造は，ID% のしきい値が高くても，部分構造が異なるタンパク質が数多く存在する．したがって，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，厳密に分類するためには，構造の類似性を考慮することが重要である．

表 3.5: 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	1,689	1,176	994	898	882	874
< 35	1,792	1,455	1,399	1,381	1,378	1,377
< 45	1,900	1,620	1,579	1,567	1,564	1,562
< 55	2,019	1,784	1,749	1,734	1,732	1,729
< 65	2,152	1,934	1,900	1,886	1,884	1,882
< 75	2,267	2,064	2,033	2,020	2,019	2,018
< 85	2,449	2,272	2,239	2,228	2,227	2,225
< 95	2,812	2,672	2,645	2,638	2,637	2,637

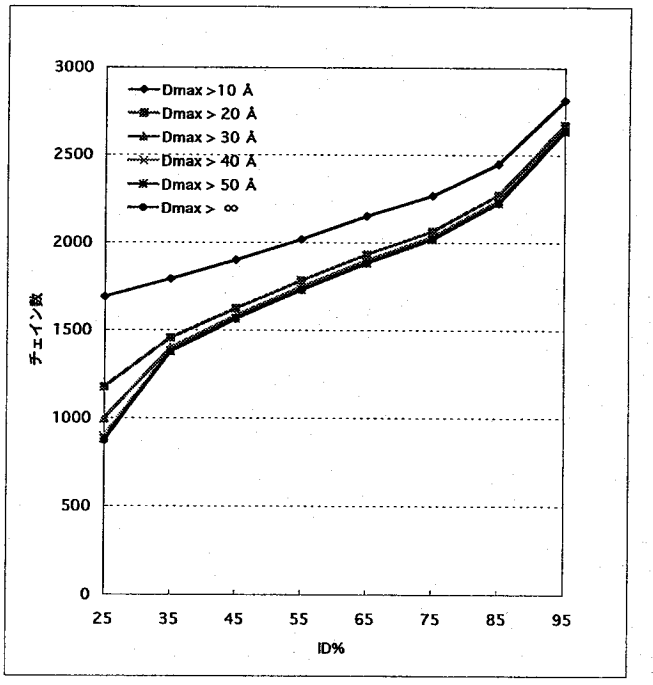

図 3.7: 類似基準 (ID% と Dmax) と代表タンパク質チェイン数の関係

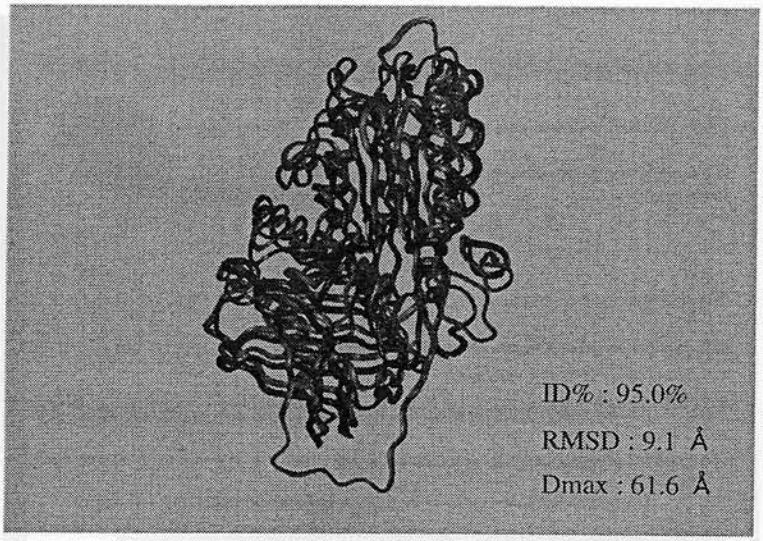

図 3.8: 抗トロンビン (PDB エントリー名:2ANT) の L チェイン (薄いリボン) と I チェイン (濃いリボン) の重ね合わせ図

3.6 PDB 代表タンパク質チェインの公開

本システムは，第 2 章の PDB 代表タンパク質チェイン決定システムを引き継ぎ，本システムにより，WWW 上に公開した全ての PDB-REPRDB の作成が行われた．PDB-REPRDB の更新は，PDB の更新（最新版の CD-ROM が届いたタイミング) に合わせて行ったため，年 4 回であった．作成した PDB-REPRDB は，第 2 章のシステムと同様に図 3.9のように WWW で公開した (URL: http://www.rwcp.or.jp/papia/)．

ホームページ (図 3.9) 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 3.10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．システムの改良及び並列化による処理の高速化により，2 倍以上の PDB-REPRDB を作成することができるようになった．様々な研究用途に対応するため，配列の相同性：ID% $\geq$ 25% $\sim$ 95% まで 10% 刻みの 8 通りと，構造の類似性：Dmax $\leq$10 Å $\sim$ 50Å 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 $8 \times 6 = 48$ セットを公開した．

WWW 上のシステムは，代表タンパク質チェインのリストに表示するデータ項目に，Cα 原子の座標を持った残基数が追加された点と，タンパク質分類名 (HEADER) が

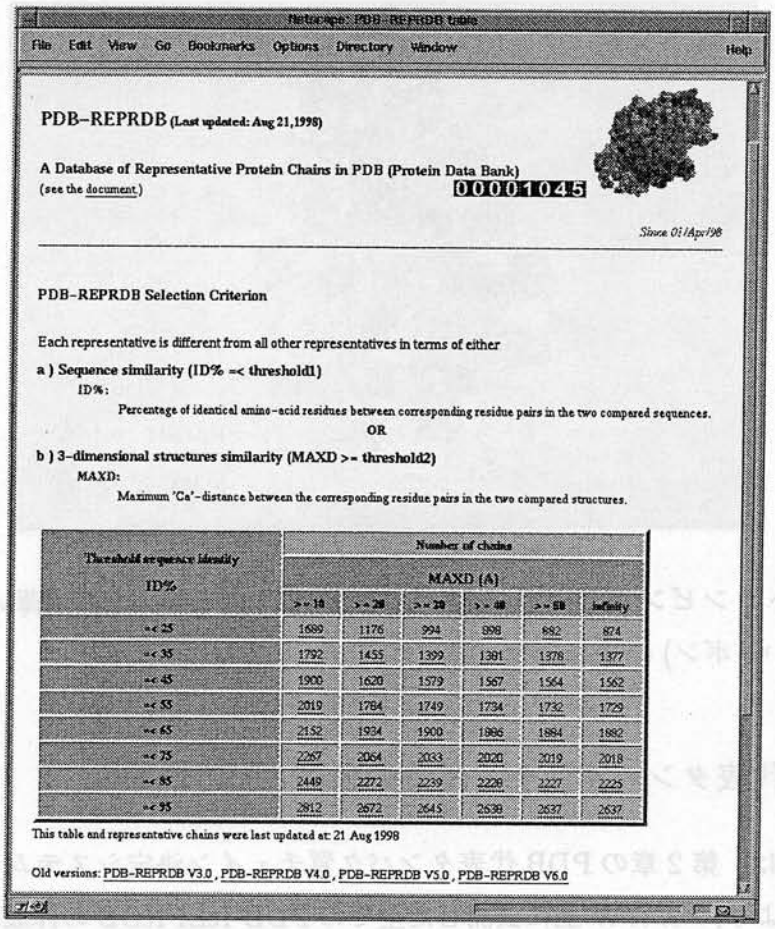

図 3.9: WWW 上の PDB-REPRDB (並列版)

タンパク質名 (COMPND) に変更された点を除くと，第 2 章のシステムと同じである．代表タンパク質チェインのリストは表形式で示され，選ばれた代表タンパク質の ID 名 (エントリー名＋チェイン ID)，および残基数，分解能，R ファクター，実験方法，側鎖原子の座標がそろっている残基数，主鎖原子の座標がそろっている残基数，Cα 原子の座標を持った残基数，非標準アミノ酸の残基数，EC(酵素) 番号，タンパク質名が記されている．また，ID 名の部分は，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，クリックすると類似タンパク質のリストを見ることができる．また，代表タンパク質チェインのリストの ID 名と残基数の間に表示されている “*” 印をクリックすると RasMol プログラムを用いて立体構造がグラフィック表示される．類似タンパク質チェインリストの ID 名は，さらに PDB とホットリンクし

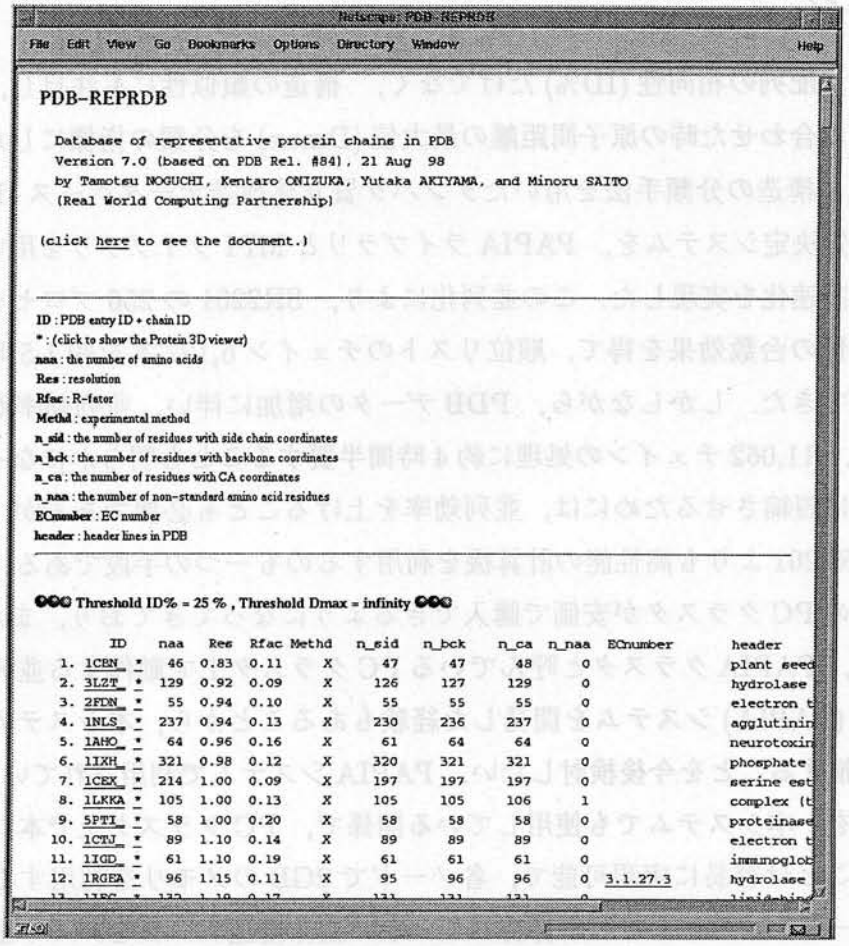

図 3.10: WWW 上の PDB 代表タンパク質チェインリストの例 (並列版)

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．

システム改良以降，本システムで作成した PDB-REPRDB は，世界から 2,100 回以上アクセスされた．

3.7 まとめ

本章では，配列の相同性 (ID%) だけでなく， 構造の類似性にも注目し，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用いたタンパク質立体構造データベース (PDB) の代表タンパク質決定システムを， PAPIA ライブラリと 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 6,127 本を約 1.5 時間で分類することができた．しかしながら， PDB データの増加に伴い，並列効率が悪化する問題が生じ， 11,062 チェインの処理に約 4 時間半要することも明らかになった．処理時間を大幅に短縮させるためには，並列効率を上げることも必要であるが，現在利用している SR2201 よりも高性能の計算機を利用するのも一つの手段である．最近，大容量メモリの PC クラスタが安価で購入できるようになってきており，また， 5.4 節で述べるが， PAPIA クラスタと呼んでいる PC クラスタ上で動作する並列タンパク質情報解析 (PAPIA) システムを開発した経験もあることから，本システムを PC クラスタに移植することを今後検討したい． PAPIA システムで利用されている PAPIA ライブラリを，本システムでも使用している関係で， PC クラスタ上で本システムを動作させることは容易に実現可能で，各ノードで 2GB のメモリを利用することにより，現在ハードディスク上でしか保持していない立体構造データをメモリ上に持つことが可能になり，処理速度はさらに向上すると予想される．

第 2 章のシステムでは，自動化が不十分でかつ，処理時間が膨大だったたため，研究者の要求に応えられるような，様々な分類基準で作成した代表タンパク質チェインセットを用意することができなかった．また，第 2 章のシステムでは， X 線結晶回折によって解析された立体構造と比較することの妥当性に疑問があり，また，構造比較をするモデル選びの方法が決められずに， NMR で解析されたデータをあらかじめ分類対象から削除していたが，本研究では， NMR データも分類対象に加え，代表チェインセットの数を大幅に増やすことができた．

本システムにより決定された PDB 代表タンパク質チェインは， 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，第 2 章のシステムから合計すると世界から 2,500 回以上アクセスされている．

本研究により，高速に PDB-REPRDB を作成できるようになり， PDB の更新のたびに 48 セットの PDB-REPRDB を作成し， WWW 上に公開できるようになった．

しかしながら，これだけのセットでは，研究者全員の要望に応えることはできない．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，順位リストの全チェイン間の配列相同性 (ID%) や構造類似性 (Dmax) の計算結果をテーブルとしてあらかじめ用意しておき，オンデマンドで様々な基準 (良質の基準や各配列および構造の類似性のしきい値など) で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する必要がある．

第 4 章

会話形式による PDB 代表タンパク質チェイン決定システム

4.1 はじめに

PDB代表タンパク質チェイン決定システムの並列化による処理の高速化によって，それ以前より多くのPDB代表タンパク質チェイン(PDB-REPRDB)のセットを用意できるようになったので，できるだけ多くの研究用途に対応するため，配列の相同性：ID% $\geq$ 25% $\sim$ 95% まで10% 刻みの8通りと，構造の類似性：Dmax $\leq$10 Å $\sim$ 50Å 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 $8 \times 6 = 48$ 通りのPDB-REPRDBを作成してWWWで公開した．しかしながら，この程度の数の代表チェインセットを用意しても，様々な研究手法でタンパク質を研究している研究者の要望を全て満たすことはできない．また，今後のPDBデータの増加を考慮すると，PDBの更新のたびに多数のPDB-REPRDBを作成するのでは，計算にかかる時間も膨大になると予想され，処理の効率化を何らかの方法で行わなければならない．

本章では，従来のPDB代表タンパク質チェイン決定システムを発展させ，研究者の要望に応じた代表セットを，会話形式ですばやく提供するPDB代表タンパク質チェイン決定システムを構築した．これにより，上記の問題を一度に解決することがきたので，そのシステムについて詳しく述べる．

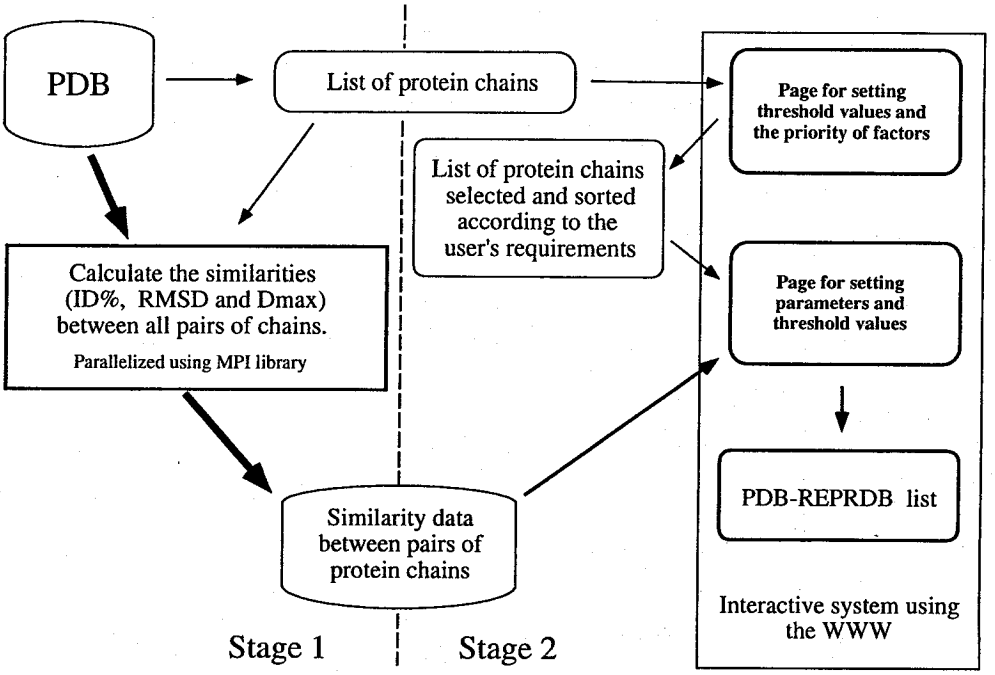

図 4.1: 会話形式による PDB 代表タンパク質チェイン決定システムの概略図

4.2 方法

新しいシステムは，以下のような従来の PDB-REPRDB 作成ポリシーに従って構築した．

a) 原子座標データの質の良いデータを代表とする

b) 他の代表と配列が似ていないチェインを代表とする

c) 他の代表と (部分的であっても) 立体構造が似ていないチェインを代表とする

本研究の代表の決め方の特徴は，最も原子座標データの質の良いデータを選ぶ点にある．PDB のような立体構造データベースから代表を決める場合，その質の違いによって，微妙に構造が違う場合があり得る．立体構造から二次構造の位置を定義する DSSP [19] のような，構造に敏感なアルゴリズムを用いると，この微妙な違いによって，結果に違いが生じる可能性があるため，代表はそのグループ内で，最もデータの質の良いチェインを用いるべきである．

従来のPDB代表タンパク質チェイン決定システムは，あらかじめ決めたデータ項目の順序で，このデータの質を決めていたが，これも研究者によって，優先するデータ項目が異なる場合がある．我々は，この点も考慮して，研究者がデータ項目に優先順位を付けられるようにするとともに，データ項目ごとにしきい値を設定できるようにし，不要なデータを事前に取り除けるようにした．さらに，研究者によっては，全体構造の違いで分類し，代表を決めたい場合もあり得るので，構造類似性の基準としてRMSDも加えた．

新しいPDB代表タンパク質チェイン決定システムは，以下の2つの処理部に分割した(図 4.1).

- 計算部(全てのチェインペアの類似性を計算する．)
- 分類部(利用者によって指定された優先度等に応じて，チェインを分類し，代表を選ぶ．)

4.2.1 計算部

計算部の流れ図を，図 4.2に示す．この処理は，PDBを更新するたびに，1度だけ行われる．

計算部では，全てのタンパク質チェインペアの類似性(ID%，Dmax，RMSD)を計算するが，以下のデータをあらかじめ削除する．

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) 短いチェイン ($l < 40$ 残基)

d) 標準残基が1残基もないデータ

これにより，PDB内に含まれるタンパク質以外(DNA，RNA，ペプチド等)のデータや実験(X線結晶回折やNMR等)以外の方法で求めた立体構造を削除する，従来のPDB代表タンパク質チェイン決定システムでは，以下のデータも事前に削除していた．

(1) 全ての残基に主鎖の座標がない(C α のみ)チェイン

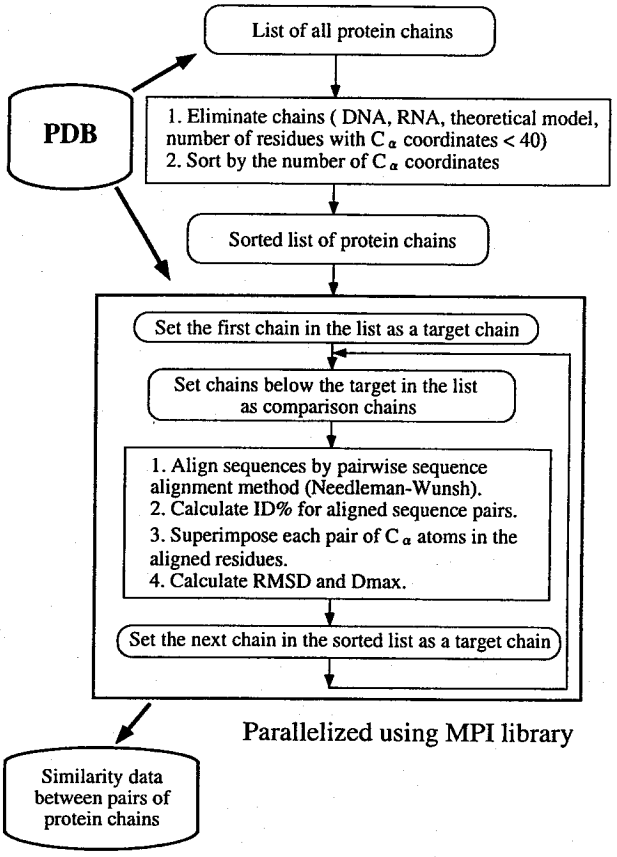

図 4.2: 計算部の流れ

(2) 全ての残基に側鎖の座標がない (主鎖のみ) チェイン

(3) リファインメントされていないチェイン

新しいシステムでは，分類部で利用者の判断によって，以下のデータ項目のしきい値を指定して削除できるようにした．

(1)' Ｃ α 原子だけの残基の比率

(2)' 主鎖原子だけの残基の比率

(3)' Ｒファクター値

次に，残ったチェインをその残基長の長いものから順にソートし，そのソートリストの先頭から順に類似度の計算を行う．類似度の計算は，最初にダイナミックプログ

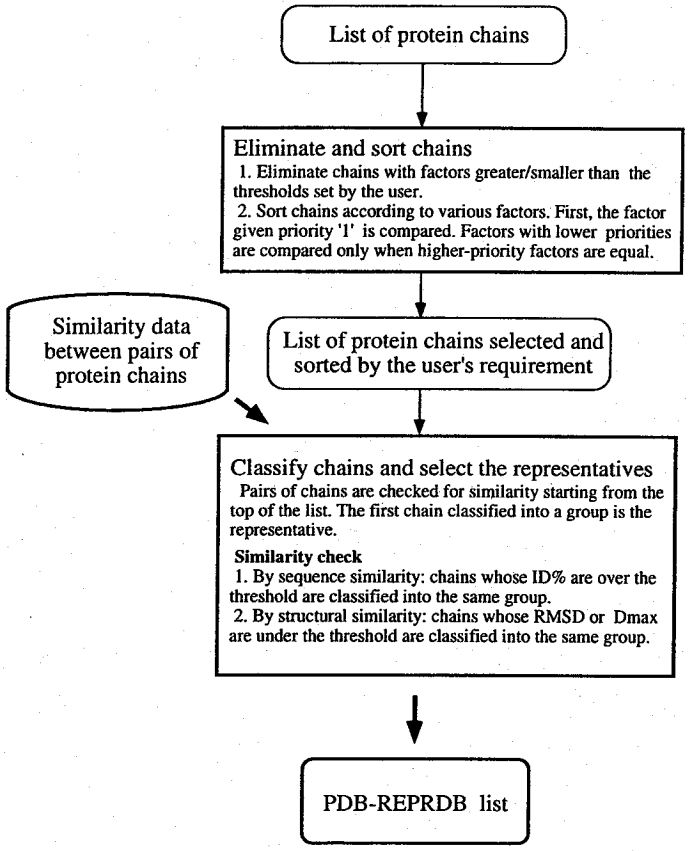

図 4.3: 分類部の流れ

ラミング法を用いたペアワイズアライメント [17] を行い，そのアライメント結果から ID% を求める．次に，そのアライメントされた残基の C α 原子同士を Kabsch による最小 2 乗フィット法 [13] により重ね合わせ， RMSD と Dmax を求める．

この処理は PAPIA ライブラリと MPI ライブラリを用いて並列化され，大量のチェインペアの類似度計算を高速に行うことができる．

4.2.2 分類部

分類部の流れ図を，図 4.3に示す. 分類部では， WWW のインターフェイスを利用し， PDB-REPRDB を作成する．

チェインの分類と代表チェインの決定方法は，従来の PDB 代表タンパク質チェイン決定システムと同じままであるが，新しいシステムでは，分類対象となるチェイン

表 4.1: データ項目による除外と優先度のデフォルト値

Factors for elimination	Default priority
分解能	1
R ファクター	2
チェインブレイク数 (少ないほど上位)	3
標準的なアミノ酸残基種以外の残基の比率 (小さいほど上位)	4
主鎖原子の座標を欠く残基 (Cα のみ) の比率 (小さいほど上位)	5
側鎖原子の座標を欠く残基 (主鎖の原子のみ) の比率 (小さいほど上位)	6
配列長 (長いほど良い)	7
変異型を含む (野生型の方が変異型より良い)	8
複合体を含む (非複合体が良い)	9
NMR データを含む	-

を，様々なデータ項目にしきい値を設定することによって，制限を加えることができ，さらに，それらデータ項目のしきい値や優先度を変えたり，類似度のパラメータを選択し，しきい値を設定することによって，様々な基準で PDB-REPRDB を作成することができる．さらに，全チェインペアの配列相同性と構造類似性計算が，すでに済んでいるため，分類部の処理は短時間で行える．

チェインを削除するためのデータ項目と優先度のデフォルト値を表 4.1に示す．配列長以外は，計算部で類似度を計算したチェインから，各データ項目しきい値より大きいチェインを削除する．また，変異体や複合体，それに NMR によって解析されたチェインを分類対象から外したい場合，削除することができる．

まず，利用者によって最初のページで設定されたこれらデータ項目のしきい値と優先度に応じて，計算部で類似度を計算したチェインから分類するチェインを抽出し，それらをソートする．

もし，全てのデータ項目の値が同じ場合，チェイン名のアルファベット順 (若いほど上位) (例： 1MCD > 1MCE, 5AT1A > 5AT1C) にソートする．この優先度の順番は，先のバージョンの優先度と同じである．次に，チェインの分類がソートされた順位リストの先頭から順に行われる．まず，順位リストの先頭のチェインが最初の代表チェインとして選ばれ，WWW の 2 ページ目で，利用者によって選択，設定された類似度のパラメータとそのしきい値を近縁の基準とし，それと近縁であった場合 (ID% はしきい値より大きい，RMSD と Dmax はしきい値より小さい)，その代表のグループに含め順位リストから削除する．この処理が順位リストの最後まで終了したら，順位リストの次のチェインを代表チェインにして，同様の処理を行う．上記の処

理を繰り返し，順位リストのチェインが全てなくなったら，処理が終了する．

分類部では，計算部で計算された類似度 (ID%, RMSD, Dmax) の値を，全てオンメモリで持って処理を行っている．これにより，膨大な計算部での計算結果の I/O 時間を削減している．

4.3 WWWによるPDB-REPRDBの利用

新しいPDB-REPRDB決定システムは，WWWユーザインターフェースを介して，代表チェインセットを作成するように設計した．まず，利用者は最初のページ(図4.4)で不必要なチェイン(例えば，分解能が悪い，チェインブレイクがあるなど)をそれぞれのデータ項目ごとにしきい値を設定することによって，削除することができる．また，代表チェインを選ぶ際のデータ項目の優先度を変更することができ，これらの設定によって，利用者が得たい代表チェインを指定することが可能となる．

このページの“apply constraints”は，そのデータ項目のしきい値(“threshold”)を適用するかどうかを決める．“No”が指定された場合は，そのしきい値は適応されず，そのデータ項目の値によってチェインが削除されることはない．“Yes”が指定された場合は，そのしきい値が適応され，条件に当てはまるチェインは削除される．

優先度(“priority”)の値は，1から9までの整数でなければならず，同じ値が複数あってはならない．優先度は，1が一番高く順に低くなる．

全ての設定を完了し，“Make List”をクリックすると，代表チェインを選ぶための順位リストが作成され，類似度パラメータ設定ページ(図4.5)が表示される．利用者はこのページで分類の基準(例えば，ID% $\geq$ 30% かつ RMSD $\leq$ 15 Å，ID% $\geq$ 90% かつ Dmax $\leq$ 5 Å)を指定することができる．分類の基準は，配列相同性の基準: 残基一致率(ID%)の1種類と構造類似性の基準: 平均原子間距離(RMSD)と最大原子間距離(Dmax)の2種類を，単独もしくは組み合わせて選ぶことができ，それぞれのしきい値を指定して，“Submit”をクリックすることによって，分類処理が開始し，代表チェインが選ばれる(図4.6上のページ)．

図4.6は，分類基準: ID% $\geq$ 30% and Dmax $\leq$ 10 Å の代表チェインリスト(図上部)と分類データ(図下部)のページの例である．代表タンパク質チェインのリストの“1GCI”をクリックすると矢印で示した分類データのページの類似チェインを表示する．代表チェインリストのページには，代表チェインの“ID”の他に以下の情報を加えて表示している．

- ID: PDB エントリ ID + チェイン ID
- *: (クリックで“RasMol”にリンクし，分子図を表示)
- naa: 残基数(PDB の SEQRES 行による)

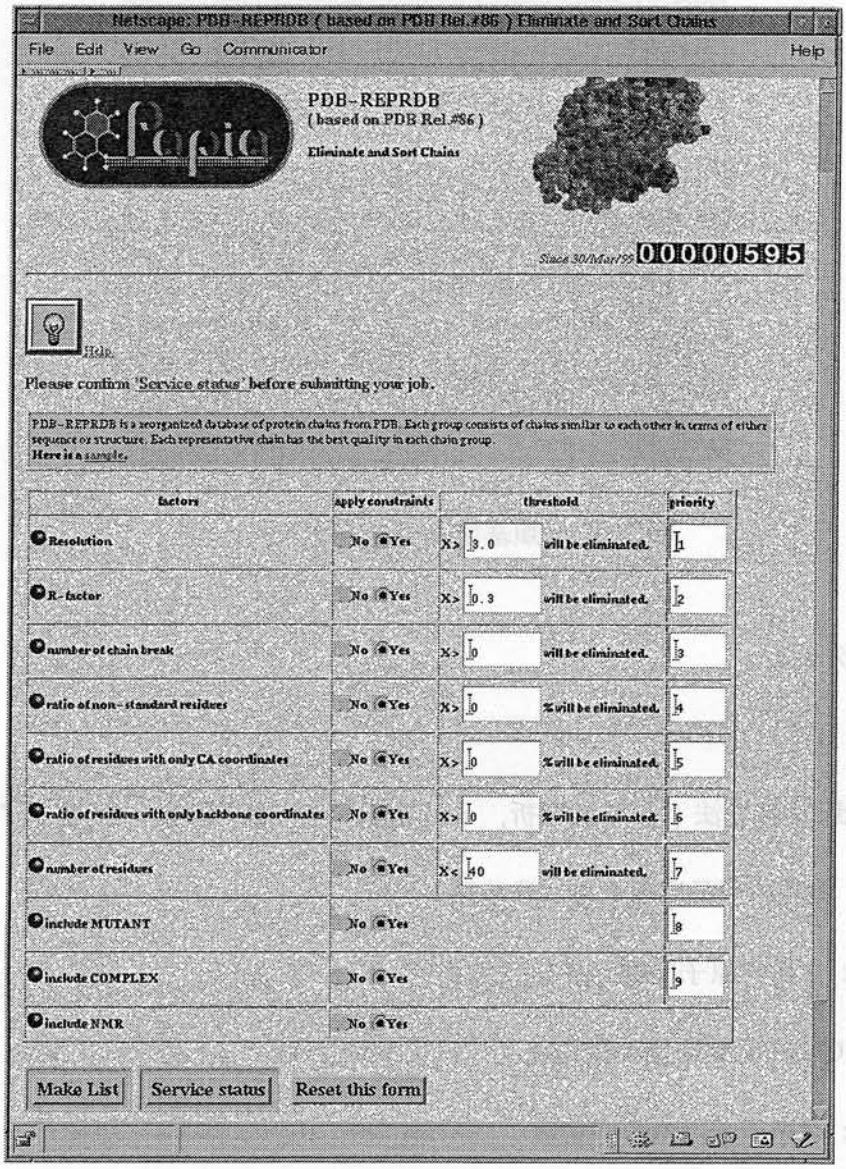

図 4.4: PDB 代表タンパク質チェイン決定システムのトップページ

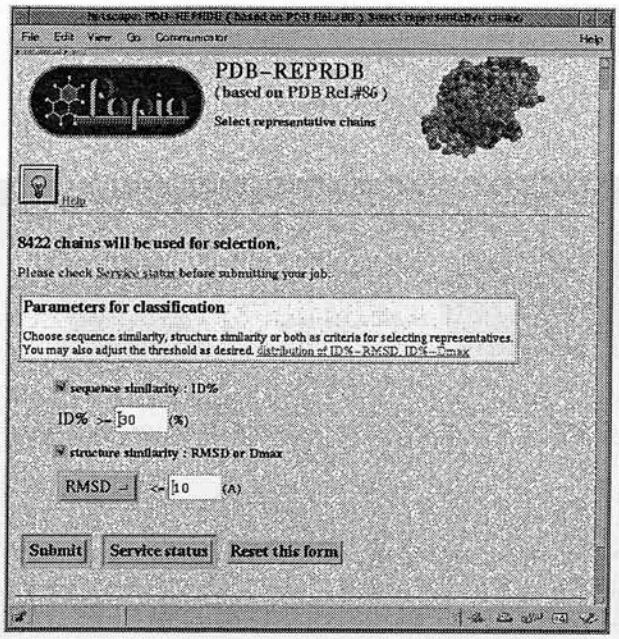

図 4.5: 分類基準をセットするページ

- Res: 分解能
- Rfac: R ファクター
- Methd: 実験方法 (X: X 線回折， n: NMR， E: 電子回折， F: ファイバー回折)
- n_sid: 側鎖の原子座標を持つ残基数
- n_bck: 主鎖の原子座標を持つ残基数
- n_ca: Cα の座標を持つ残基数
- n_naa: 標準残基以外の残基数
- brk: チェインブレイク数
- mutant: 変異型と野生型の区別 (m: mutant， W: wild)
- complex: 単量体と複合体の区別 (c: complex， N: non-complex)
- ECnumber: 酵素番号

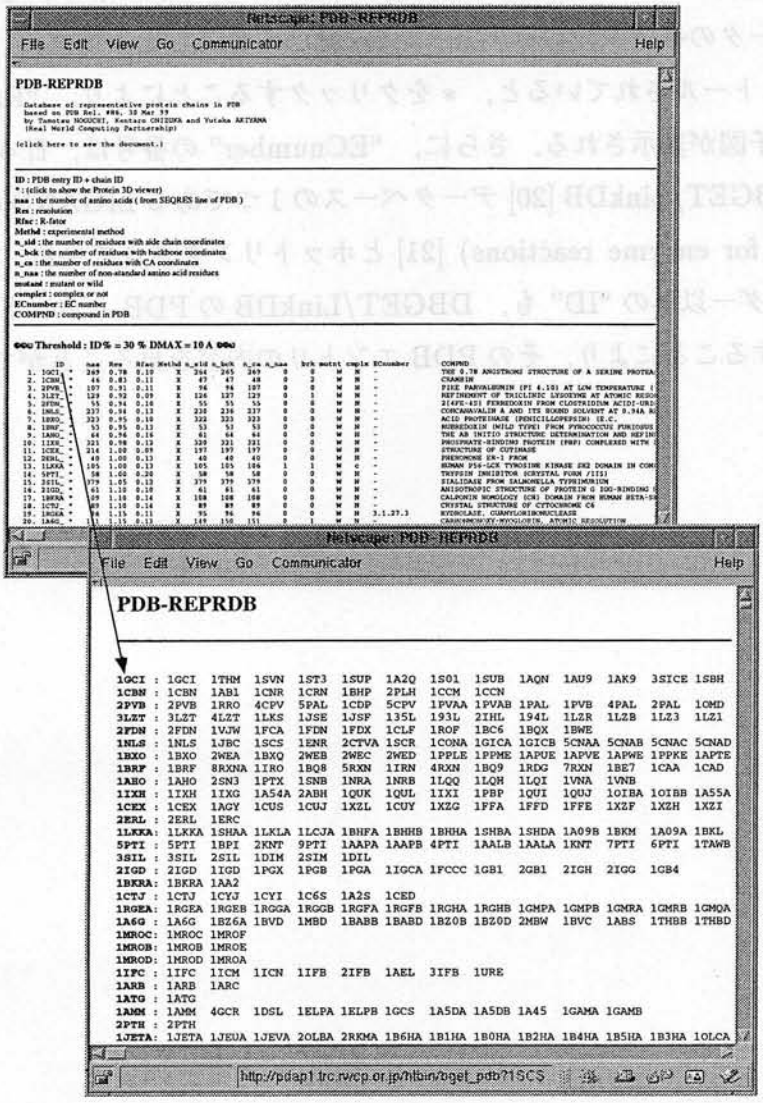

図 4.6: PDB-REPRDB の代表タンパク質チェインリストと分類データリスト

- COMPND: タンパク質名

分類データのページは，代表チェインの “ID” をヘッダーにして，その代表チェインと類似しているチェインが一行に記述されている．代表チェインリストの “ID” の文字と，その代表チェインの分類データはホットリンクされており，それをクリックすると分類データのページの類似チェインの “ID” を見ることができる．また，“RasMol” がインストールされていると，* をクリックすることにより，“RasMol” が起動され，その分子図が表示される．さらに，“ECnumber” の番号は，日本におけるゲノムネットの DBGET/LinkDB [20] データベースの 1 つである LIGAND (Ligand chemical database for enzyme reactions) [21] とホットリンクされている．分類データのページの ヘッダー以外の “ID” も，DBGET/LinkDB の PDB とホットリンクしており，クリックすることにより，その PDB エントリの内容を見ることができる．

4.4 まとめ

本章では，利用者が会話話形式で配列と構造の違いに基づいた選択基準を指定して，PDB 代表タンパク質チェインデータベース (PDB-REPRDB) を作成する新しいシステムについて述べた．本システムによって，我々が指向している，配列の相同性は高いが，残基の挿入，欠損や置換による部分的な構造の違い，それに複合体を形成したことによって生じる微妙な構造変化などを見落とさずに選ぶ代表チェインをオンデマンドで選ぶことが可能となった．本システムは，1999 年 4 月から PAPIA WWW サーバーで利用可能となり，2000 年 11 月時点で約 1,300 件利用されている．

現在，PDB-REPRDB 用のデータベースの更新は，現環境で PDB の更新が毎週行えるようになっので，従来の 3ヶ月に 1 度のペースから 1 から 2ヶ月に 1 度のペースで行っている．

第 5 章

PDB 代表タンパク質チェイン決定システムの利用

5.1 はじめに

本章では，まず本研究を行うきっかけとなった，蛋白工学研究所 西川建氏（現 国立遺伝学研究所 生命情報研究センター教授）との共同研究「アミノ酸配列に基づくタンパク質二次構造予測」について述べる．次に本研究の結果得られたPDB代表タンパク質チェインデータベース (PDB-REPRDB) の利用例として，前述の西川建教授，国立遺伝学研究所生物遺伝資源情報総合センターの伊藤將弘氏および新情報処理開発機構つくば研究センター 並列応用つくば研究室の秋山泰室長（現 工業技術院 電子技術総合研究所主任研究官）との共同研究である「3D-1D 法と部分配列相同性を用いたタンパク質二次構造予測」と新情報処理開発機構 つくば研究センター 並列応用つくば研究室で開発された「並列タンパク質情報解析 (PAPIA) システム」について述べる．また，本システムの利用状況に関しても簡単に紹介する．

5.2 アミノ酸配列に基づくタンパク質二次構造予測

タンパク質二次構造予測は，各アミノ酸残基が形成することができる二次構造を，α-ヘリックス，β-ストランド，コイルのいずれかに限定できると言う点で，デジタルな問題である．また，β-シートを形成するβ-ストランドは，本来相手も予測しないといけないが，現状では，配列に沿ったローカルな二次構造として予測しているので，線形問題として扱うことができる．したがって，二次構造予測は，配列と二次構造間の一次元のデジタルな関係を見つけることが最終目的である．それは，単純で理想的な予測システムがあれば，簡単に解けると思われていた．二次構造予測の歴史の初期においては，研究者は単純な予測方法で，満足できる予測結果を得られたので，それを簡単な問題として見なしていた．しかし，それらの方法が後に明らかになったタンパク質立体構造に対しては無力で，予測精度が落ちたことによって，その困難さが明らかになった．しかしながら，現在，二次構造予測は，タンパク質立体構造予測における主要技術のうちの一つと見なされている．

この困難さは，タンパク質のフォールディング過程において，二次構造が部分的相互作用だけでなく，配列上離れた残基同士の相互作用によるグローバルな効果に依存しているためである．このグローバルな効果を表現する試みは，幾つかのグループで行われているが，一般に一次元の配列情報から予測を行なっているため，グローバルな効果を考慮しているとは言えない．したがって，現状の部分配列を基にした予測法では，配列上近傍の残基同士の相互作用が支配的なままなので限界がある．

最近発表された予測法は，新規または既存方法の改良であるが，全てローカル配列に基づく方法である．新しい方法の一つに，パターン認識の方法として発展したコンピュータ学習アルゴリズムであるニューラルネットを用いた方法がある．その他の新規方法では，ホモロジーを基にしている．立体構造既知のタンパク質の中に，弱い配列類似性がある部分構造の二次構造を割り当てて，予測する残基において，割り当てられた二次構造タイプで多いものを予測としている．この方法は，立体構造が明らかになったタンパク質の数が増えれば，最も可能性がある方法である．類似配列を用いた別の方法は，ターゲット配列と相同ないくつもの配列を用いている．同じ進化上のファミリーからなるタンパク質は，一般に同じ三次元構造を取るので，二次構造も同様に考えることができる．配列データベースから利用できる相同な配列データが予測精度を改善している．

上記の幾つかの予測法を含め，８つの代表的予測法を比較した．そして，その中で

表 5.1: 8 種類の二次構造予測精度の比較（テストセット A）

PDB code	Chain length	Prediction scores (%)							
		CF	GOR	Li	QS	NO	Na	PF	GGR
1CTF	68	41	43	68	47	47	50	52	56
1LH1	153	50	58	48	62	61	64	61	74
2CDV	107	52	54	51	72	73	68	68	71
2CTS	437	54	59	62	61	67	64	73	74
2WRP	104	56	59	53	60	62	71	67	71
1ACX	108	47	57	63	65	69	69	64	69
1HMG(A)	328	57	58	52	66	58	57	59	63
1HMG(B)	175	50	59	56	55	50	53	67	59
1FC1	207	57	48	54	53	60	62	44	60
1NXB	62	71	58	61	69	63	68	58	89
1PSG	365	65	60	52	62	64	67	61	63
2ALP	198	53	58	49	59	56	52	54	72
4RHV	255	53	55	58	64	50	61	61	63
1ABP	306	49	53	47	57	49	56	53	53
1WSY	385	60	67	65	64	71	64	62	66
3PFK	319	56	56	54	62	61	64	61	63
3GAP	208	45	50	64	60	51	53	67	51
1UBG	76	71	74	63	58	75	59	75	51
2CI2	65	52	62	39	62	31	45	60	62
2CPP	405	58	56	68	62	66	67	56	58
2OVO	56	54	55	63	64	55	52	57	61
6API	374	51	46	50	48	63	55	60	49
Average		54.9	56.2	56.5	60.3	60.6	61.0	61.0	62.4

CF: Chou-Fasman, GOR: Garnier-Osguthorpe-Robson, QS: Qian-Sejnowski, GGR: Gibrat-Garnier-Robson, Na: Nagano, NO: Nishikawa-Ooi, Li: Lim, PF: Ptitsyn-Finkelstein

5 つの高精度の方法を組み合わせてジョイント法を作成し，個々の方法との比較を行なった．最後に PDB に公表されていない X 線結晶回折によって決められたタンパク質の構造からテストセットを作りテストした．

Chou-Fasman(CF) 法 [22] は，いくつかの経験的なルールを組み合わせ，単独残基のパラメータを用いたシングレットタイプに含まれる．Garnier-Osguthorpe-Robson の GOR 法 [23] は，求める残基とそれを中心にした ± m 残基の周辺アミノ酸の残基ペアを考慮したが，アミノ酸残基の特定を，中心残基ではせず，周辺残基にだけした疑似ダブレットタイプである．ニューラルネット法は，Qian- Sejnowski(QS)[24] によって最初に二次構造予測に応用された．Gibrat-Garnier-Robson による GGR 法 [25] は，GOR 法における疑似ダブレットを特定のアミノ酸ペアを考慮するダブレットに置き換えた方法で，ダブレットの代表法である．長野法（Na） [26] は，トリプレットタイプで，一度に 3 残基を考慮している．ただし，独立なパラメータを減らすために，

```
PREDICTION CODE AND NAME: 1UBQ         UBIQUITIN
NUMBER OF RESIDUE = 76
        1         +         +         +         +         +         +         +      76
SEQUENCE MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
X-RAY    CBBBBBBCCCCCBBBBBCCCCCCAAAAAAAAAAAACCCCCCBBBBBCCBBCCCCCCCCCCCCCCCCCBBBBBBCCCCC
           **  ********** **** ********    ******  *** *  ****************        ***
PREDICT  AABBAABCCCCCBBBBBBCCCCAAAAAAAAACCCCCCCCCCCCABBBACCCCCCCCCCCCCCCCCCCCAAAAAAAACCC
RATE     34333333443345554555533555453445555555554403330544445554355555343334444333555
GGR      AABBAABACCCCCBBBBCCCCAAAAAAACACCCCCCCCCCAAAAAAACAAAACCCCACCCCCAAAAAAAAABBBCCC
PF       BBBBBBBCCCCBBBBBBCCCCAAAAAAAAACCCCCCCCCCCCBBBCCCCCCCCCCCCCCCCCCCCAAAAAAAACCC
NA       BAAAAABBBBBBBBBBCCCCBBAAAAAAACCCCCCCCCCCCBBBCCCCCCCCCBBCCCCCCCCBBBBBBBBCCC
NO       AAAAAAACCCBBBBBBCCCCCAAAAAAAAACCCCCCCCCCCBBBBBCCCCCCCCCCCCCCCCCCCBBBBBBCCC
QS       AABBBBCCCCCCBBBBBCCCCCCAAACACCCCCCCCCCCCCCAAAAACCCCCCCCCCCCCCCCACAAAAAAAAAACCC
```

図 5.1: ジョイント予測の例

アミノ酸残基を 7 つのグループに分類している．既知構造のタンパク質と弱いホモロジーを利用した西川 - 大井のホモロジー法 [27] は，一度に約 10 残基の部分配列を考慮し，同一アミノ酸を見つける代わりに類似残基を探すので，疑似マルチプレットタイプと見なせる．Lim 法（Li）[28] は，既知の X 線結晶回折によるタンパク質立体構造から得られた数値パラメータとは独立な，経験的なルールだけを用いている．1970 年代前半に発表されたこの方法は，今日用いられているエキスパートシステムのパイオニア的試みであった．Ptitsyn-Finkelstein（PF）法は [29]，アイシング理論の統計力学を基にした幾つかの方法の中に含まれる．経験的なパラメータを用いる，α- ヘリックス形成の基本法則がこの理論でできているので，このタイプの予測は有利である．

各予測法を調べるために，下記の 22 タンパク質を PDB から選らんだ．それらの幾つかは，上記の予測法のテストに用いられていたが，それらのパラメータセットやルールを決める際には用いられていなかったことを確認した．PDB の座標データから二次構造を定義する方法としては，Kabsch-Sander 法 [19] を用いた．予測精度は，各アミノ酸残基において α- ヘリックス，β- ストランド，コイルの 3 状態を正しく予測した割合によって計った．

予測結果を表 5.1に示す．22 タンパク質の平均予測精度で判断すると，GGR 法が最も高い予測精度であり，Chou-Fasman 法が最も低かった．上位 5 手法の予測精度は，60 ～ 63% でそんなに差は大きくなかった．個々のタンパク質の予測精度にばらつきはあるが，異なる方法論によることを考えるとこれは驚くべきことである．予測精度が 60% 以下の 3 つの方法（CF, GOR, Li）は，1970 年代に開発された方法であった．唯一の例外は長野法で，60% 以上の予測精度で，1980 年代の残りの方法と同じ

表 5.2: ジョイント予測による予測精度 (テストセット A)

PDB	Prediction scores (%)			
Code	Jo	J-5	J-54	J-543
1CTF	50	47 (0.25)	55 (0.69)	50 (1.00)
1LH1	68	82 (0.43)	75 (0.67)	69 (0.97)
2CDV	77	84 (0.52)	80 (0.88)	76 (0.94)
2CTS	74	91 (0.34)	84 (0.66)	75 (0.97)
2WRP	74	81 (0.40)	79 (0.64)	75 (0.95)
1ACX	70	75 (0.57)	72 (0.83)	70 (1.00)
1HMG(A)	65	76 (0.49)	73 (0.73)	68 (0.94)
1HMG(B)	57	78 (0.21)	64 (0.49)	56 (0.93)
1FC1	62	72 (0.44)	70 (0.69)	65 (0.94)
1NXB	66	91 (0.57)	72 (0.76)	66 (1.00)
1PSG	67	81 (0.38)	76 (0.66)	69 (0.95)
2ALP	63	87 (0.31)	75 (0.58)	62 (0.94)
4RHV	61	80 (0.41)	69 (0.71)	63 (0.96)
1ABP	57	67 (0.38)	61 (0.64)	59 (0.94)
1WSY	70	86 (0.40)	77 (0.67)	73 (0.93)
3PFK	66	83 (0.35)	78 (0.67)	69 (0.93)
3GAP	56	81 (0.36)	64 (0.66)	58 (0.94)
1UBG	70	90 (0.41)	75 (0.67)	71 (0.96)
2CI2	54	82 (0.17)	58 (0.53)	54 (0.94)
2CPP	69	79 (0.35)	77 (0.62)	72 (0.92)
2OVO	59	69 (0.64)	62 (0.80)	59 (0.96)
6API	55	83 (0.28)	65 (0.63)	57 (0.95)
Average	64.8	80.3 (0.38)	72.8 (0.66)	66.5 (0.95)

精度であった．60% ぐらいの予測精度では，新しいタンパク質に用いる場合，十分高いとは言えないので，個々の方法を組み合わせたジョイント法を開発した．

タンパク質二次構造におけるジョイント予測法は，並行に複数の異なる予測法を用いる方法である．この方法の効果に影響する最も重要な要因の一つは，組み合わせる個々の予測法をどのように選ぶかである．我々は，方法論の異なる 8 つの代表的方法を調べた後，予測精度の良い上位 5 つの方法を選んだ．

ジョイント法のアルゴリズムは簡単で，各残基における個々の二次構造予測法の結果の多数決を取る．図 5.1に結果の例を示す．各残基における二次構造は，各々 A （α-ヘリックス），B （β-ストランド），C(コイル) で示される．例えば，タンパク質の最初の残基では，5 つの個々の方法は， A, B, B, A, A で，ジョイント法の予測は A である．最初の残基の予測では，5 つのうち 3 つが一致したので，予測レイトは 3 で示した．予測レイトは， 5,4,3,0 のいずれかである．レイト 5 は，個々の予測結果が全て一致したことを示し，レイト 0 は，個々の予測結果が 2:2:1 （例えば， A, A, B, B,

表 5.3: テストセット B のタンパク質リスト

Abbreviation	Protein	Source	Fold Type
PLC	Phospholipase C	Bacillus cereus	α
GH	Growth hormone	Pig	α
IL-2	Interleukin 2	Human	α
HAP	Aspartyl protease	HIV-1 virus	β
MADH	Methylamine dehydrogenase, L subunit	Thiobacillus versutus	β
IL-1B	Interleukin 1β	Human	β
BLG	β-Lactoglobulin	Bovine	β
YE	Enolase	Yeast	α/β
XYI	Xylose isomerase	Streptomyces olivochromogenes	α/β
DLH	Dienelactone hydrolase	Pseudomonas sp.	α/β
TS	Thymidylate synthase	Lactobacillus casei	α/β
MI	Muconolactone isomerase	Pseudomonas putida	$\alpha+\beta$
PCD	Protocatechuate 3,4-dioxygenase, α subunit	Pseudomonas aeruginosa	$\alpha+\beta$
HLA1	Histocompatibility antigen HLA-A2, heavy chain	Human	$\alpha+\beta$
HLA2	Histocompatibility antigen HLA-A2, light chain	Human	β

C) に分散したことを示している．この場合，予測精度の比較で一番予測精度が高かった GGR 法の結果を用いることにした．

ジョイント法を個々の予測法のテストで用いたのと同じ 22 タンパク質に適応した結果を表 5.2に示す．平均予測精度は，64.8% であった．ジョイント法によって，個々の二次構造予測法よりも 2 ～ 5% 高い予測精度で，予測できるようになった．これは，用いた予測法の中で一番予測精度の高い方法と同じレベルの予測精度しか得られなかった従来のジョイント予測とは，異なる新しい発見である．今回の場合，注意深く個々の方法を選んだことが，予測精度の改善を導き，注意深く適用したとき，ジョイント予測は，1 つの方法で予測するより一般に良い精度が得られることを示唆している．ジョイント予測において，全ての予測法で同じ予測をした残基では，他の予測に比べて予測の信頼性があることが知られている．表 5.2は，全残基に対する全一致で予測できた割合を示すカバレージでは 0.38 と低いが，全一致の予測精度が，80.3% と高精度であることを示している．同様に，4 と 5 方法で一致している予測の J-54 の平均予測精度は，0.66 のカバレージで 72.8% であった．これらの結果はかなり注目すべきなので，論文でのみ構造が明らかにされ，PDB にまだ登録されていない 15 個のタンパク質 (表 5.3) をテストセット B として，さらに詳細なテストを行なった．

表 5.4に，ジョイント法と個々の二次構造予測法で予測した結果を示す．テストセット A の結果と比較すると，ジョイント法も個々の二次構造予測法同様に，4 から 7%

表 5.4: 8 種類の二次構造予測精度の比較 (テストセット B)

PDB	Chain	Prediction scores (%)						
code	length	QS	NO	Na	PF	GGR	Jo	J-5
PLC	245	48	57	52	40	58	59	52 (0.28)
GH	191	70	71	70	81	69	81	93 (0.40)
IL-2	133	45	59	61	39	50	50	59 (0.37)
HAP	94	54	40	47	57	53	59	81 (0.28)
MADH	121	55	57	53	43	55	56	56 (0.43)
IL-1B	153	32	58	50	59	41	48	64 (0.34)
BLG	162	30	44	40	59	29	38	62 (0.28)
YE	436	67	68	71	68	72	76	89 (0.39)
XYI	388	60	46	52	56	52	59	53 (0.47)
DLH	236	56	63	57	51	55	61	78 (0.31)
TS	316	64	59	60	59	60	64	82 (0.36)
MI	96	49	43	50	65	54	50	84 (0.26)
PCD	200	46	49	48	49	38	49	63 (0.29)
HLA1	270	52	56	51	47	56	56	64 (0.42)
HLA2	97	60	55	59	65	57	62	75 (0.49)
Average		54.8	56.5	56.2	56.4	55.2	60.1	70.3 (0.37)

表 5.5: テストセット A と B における平均予測精度と標準偏差

	For test set A		For test set B	
Method	%	Coverage	%	Coverage
Jo5	64.8 ± 7.1	1.00	60.1 ± 10.4	1.00
J-5	80.3 ± 9.4	0.38	70.3 ± 13.1	0.37
J-54	72.8 ± 7.6	0.66	65.9 ± 12.3	0.64
J543	66.5 ± 7.3	0.95	61.8 ± 10.9	0.94
GGR	62.4 ± 9.2		55.2 ± 10.6	
PF	61.0 ± 9.0		56.4 ± 11.0	
Na	61.0 ± 7.1		56.2 ± 8.1	
NO	60.6 ± 9.9		56.5 ± 8.7	
QS	60.3 ± 6.0		54.8 ± 11.1	
Li	56.5 ± 7.5		54.8 ± 8.9	
GOR	56.2 ± 6.5		53.4 ± 11.6	
CF	62.4 ± 9.2		50.3 ± 8.5	

平均予測精度が落ちている．テストセットAとBを直接比較で，この差をもっとはっきりさせた(表5.5)．ジョイント法の予測精度が悪化したのは，個々の二次構造予測法の予測精度が落ちたことによる．テストセットBの予測では，平均予測精度が悪化しただけでなく，標準偏差も大きくなっている．予測精度が平均から大きく振れる予測法は，一般的に信頼性に欠ける．テストセットAでは，X線結晶回折の座標からKabsch-Sanderの二次構造定義法で，テストセットBでは，各実験者が二次構造を定義していることが，この予測精度の振れの理由である可能性がある．この可能性を確かめるために，テストセットAでも実験者による定義を用いて精度を調べたところ，ジョイント法で61.1%に下がった．これは，実験者による定義の二次構造を用いると数パーセント程度予測精度が変化することを示している．テストセットAでこの二つの定義による二次構造の一致率を調べたところ，たった80%であった．

もう一つ悪化の原因として考えられるのは，サンプルとして選んだタンパク質の違いである．現状の予測法による二次構造予測では，一般にα-ヘリックスの方がβ-ストランドより予測精度が高い．テストセットAとテストセットBのα-ヘリックスとβ-ストランドの含有率は，それぞれ31%，22%と31%，26%であった．テストセットBのβ-ストランド含有量が，テストセットAより高かったことが，精度を落とした原因と考えられる．また，個々のサンプルタンパク質の予測精度を見ると，ジョイント法では，予測精度が良い場合と(GH，YE)と極端に悪い場合（BLG，IL-1B，PCD，IL-2，MI）があり，個々の二次構造予測法でも同様な傾向を示している(表5.5)．GHとIL-2は，ともにall-α型のタンパク質であるが，予測精度はジョイント法で81%と50%であった．α+β型のタンパク質で予測精度がすべて65%以下だった他は，このように全てのフォールドタイプで予測精度が良いタンパク質とそうでないタンパク質が混在している．したがって，予測の成功は，フォールドタイプには依存していない．

興味深い例として，BLG（β-lactoglobulin）を取り上げる(図5.2)．BLGは9個所のβ-ストランドと1個所のα-ヘリックスからなるall-βタイプのタンパク質であるが，図5.2に示すようにほとんど全てのβ-ストランドの位置をα-ヘリックスと予測してしまっているため，40%以下の予測精度になってしまった．一方，桑島らは，BLGに関して，興味深い現象を観測していた．変性剤によるリフォールド実験で，ネイティブ状態に向かうフォールドの初期段階で，β-ストランドに加えてα-ヘリックスが相当な量形成されていることを発見したのである．これは，BLGの配列がα-ヘリックスを形成する傾向性を持っていたことを示していると言え，ジョイント法の結

```
PROTEIN CODE: BLG (162 AA)
 1        +        +        +        +        +        +        +        +       90
  LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNEN
  CCCCCCCCCCCCCCCCBBBBBBBBBBBBCCCCCCCCCCCCBBBBBBCCBBBBBBBBBBBBCCCBBBBBBBBBBBBBBCCCBBBBBCCCCBB
      *******                    *****  **  **                           ***
  BBBBCCCCCCCAAAACCCAAAAAAAACAABBAACCCCCAABBAACCCCCCCCAAAAAAAAACCCCCAAAAAAAACCCCCAAAAAAAAAAC
  343333355543333333333444443003300355550033333345555535555554345543444444333444044443344333
91        +        +        +        +        +        +       162
  KVLVLDTDYKKYLLFCMENSAEPEQSLVCQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI
  BBBBBBBCCCCBBBBBBBBCCCCCBBBBBBBBBBCCCCCAAAAAAAAAAACCCCBBBBBBCCCCCCCCCCCC
    ***  * *  ****   *****        * ****************   * **********      *
  AABBBACCACAABBBBACCCCCCCCCAACCCCBCCCCCCAAAAAAAAAAAAAAACABBBBBCCCCCAAAAAAC
  333330343333343305555544433300033455554455555555555543033443355554444433
```

図 5.2: BLG (β-lactogloburin) のジョイント予測結果

果が必ずしも悪いとは言えないことを示唆している．しかしながら，タンパク質の最終構造は α- ヘリックスよりもむしろ β- シートによって安定化した全体のエネルギーバランスで決まっている．球状タンパク質は，配列上に沿ったショートレンジの相互作用が，常に配列上離れているが空間上近くにあるアミノ酸残基間のロングレンジの相互作用より先に起こる．これは，現状の二次構造予測法が，ある程度成功していることからも明らかである．しかしながら，これ以上の予測精度を得ようとしたとき，このロングレンジの相互作用を考慮しないと発展は困難であることをこの実験は示唆している．

表 5.6: New SSThread 構造ライブラリ用 PDB 代表タンパク質チェインの基準

Factors for elimination	Threshold	Default priority
チェインブレイク数 (少ないほど上位)	>0	1
標準的なアミノ酸残基種以外の残基の比率 (小さいほど上位)	>0	2
主鎖原子の座標を欠く残基 (Cα のみ) の比率 (小さいほど上位)	>0	3
分解能	なし	4
R ファクター	なし	5
側鎖原子の座標を欠く残基 (主鎖の原子のみ) の比率 (小さいほど上位)	なし	6
配列長 (長いほど良い)	なし	7
変異型を含む	-	8
複合体を含む	-	9
NMR データを含む	-	-

5.3 3D-1D 法と部分配列類似性を用いたタンパク質二次構造予測

序論でも簡単に述べたが，タンパク質二次構造予測法の研究は，経験的タンパク質立体構造予測の一種で，立体構造既知のタンパク質の配列と二次構造の関係を調べて，立体構造未知のタンパク質の配列からその二次構造を予測する研究である．“立体構造既知のタンパク質” がその基礎となるため，その基礎データとしてどの立体構造を選ぶかが予測にとって非常に重要になってくる．「3D-1D 法を用いたタンパク質二次構造予測法 (SSThread)」は，構造と配列の適合性評価法 (3D-1D 法)[30] によって，予測するタンパク質の配列と適合する立体構造を，構造ライブラリと呼ばれる PDB から一定の基準を満たしたタンパク質チェインをタンパク質立体構造の代表（テンプレート）として登録したデータベースから抽出し，その立体構造情報を用いて二次構造予測する方法である [31]．したがって，この構造ライブラリに登録する代表タンパク質立体構造によって，予測精度が変わってくる．このように「3D-1D 法を用いたタンパク質二次構造予測法」は，配列と二次構造間の関係から予測をする一般のタンパク質二次構造予測法と異なるが，“立体構造既知のタンパク質” がその基礎となると言う意味では，その重要度は全く同じである．

本研究に伴い，構造ライブラリの更新を行った．構造ライブラリの作成は，PDB 代表タンパク質チェイン決定システムを用い，表 5.6の基準でデータの質を決め，配列相同性 (ID%) のしきい値を 30% 以上として行った．構造ライブラリの残基長による数の分布で，更新前の構造ライブラリと比較したグラフを (図 5.3) に示す．選出の基となる PDB が Release 74 から Release 84 と新しくなったため，抽出された構造ラ

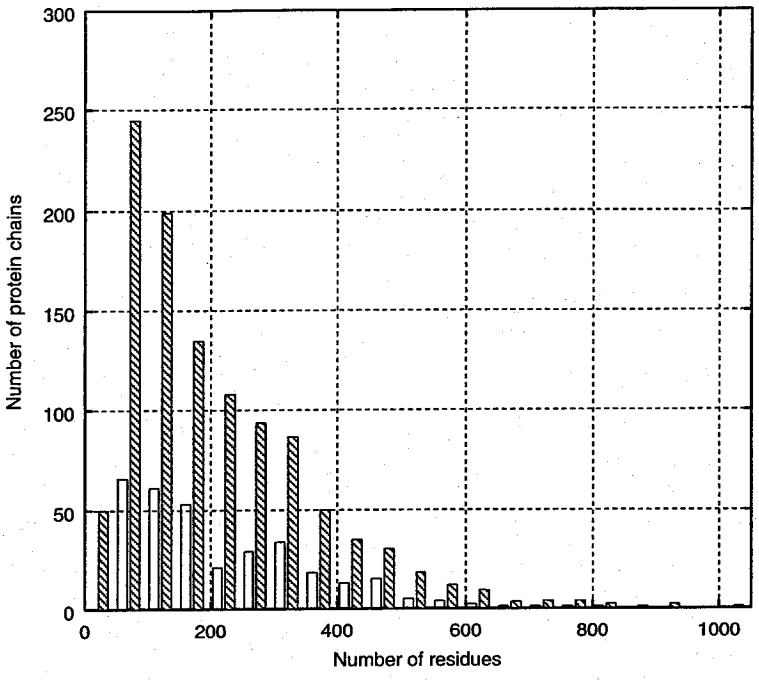

図 5.3: 旧構造ライブラリと新構造ライブラリに含まれるタンパク質の大きさの分布
白抜きと斜線の棒グラフは，それぞれ旧構造ライブラリと新構造ライブラリに含まれる 50 残基長ごとにまとめられたタンパク質の数である．全体の数は，それぞれ 325 と 1,089 である．

イブラリの数は，325 個から 1,089 個に増加し，400 残基以上のチェインも増えている．また，本研究の新しいテストセットのタンパク質は，構造ライブラリを作成した PDB の Release より更に新しい Release から PDB 代表タンパク質チェイン決定システムを用い選んだ．テストセット作成の基準は，構造ライブラリ作成時と同じとし，構造ライブラリと同一，もしくは類似 (ID% $>$ 30 %) の代表チェインを除いたものをテストセットとした．

本研究では，SSThread に，部分配列類似性と 3D-1D 法における適合性スコアの値でそれぞれ予測の重みづけを行う改良を行い New SSThread 法を作成した．また，SSThread の研究で，大きなタンパク質（$>$ 400 残基）の予測精度が悪化する問題が指摘されており，その研究ではドメインで区切って予測を行うことによって，予測精度を改善していた．しかしながら，本来，立体構造がわからないタンパク質のアミノ酸配列を，正確にドメインに分けることは，現状では不可能である．そこで，大きなタンパク質では，配列を二等分し，それぞれに 100 残基の重なり領域を付加して New SSThread で二次構造予測を行い，重なり部分でそれぞれの予測結果の前後 50 残基分

づつを用いてつなぎ合わせて全体の予測とする方法を用いた．新たに選んだテストセット (> 400 残基含む) における平均予測精度は，New SSThread が，SSThread より 3% 高く，71.3% であった．

5.3.1 方法

New SSThread の概略図を図 5.4に示す．まず，3D-1D 法により，予測するタンパク質配列と構造ライブラリ全体のテンプレートタンパク質との適合性スコアを，全体構造の適合性スコア（S_{tot}）と残基ごとの適合性スコア（S_{tot}^{res}）の 2 種類計算し，S_{tot} でソートしたテーブルを作成する．次に予測するタンパク質配列と S_{tot} と S_{tot}^{res} でソートしたそれぞれのリスト上の上位 50 個のタンパク質配列をペアで構造アライメントする．ここまでは，SSThread の処理と同じである．

SSThread における次の処理に改良が加えられた．SSThread では，S_{tot} の構造アライメントにおいて，各残基ごとに構造既知のタンパク質の二次構造を，α- ヘリックス，β- ストランド，コイル別にカウントして，最も多い数値の二次構造を予測としていた．この際，アライメント上のギャップにより，二次構造が特定されない場合があり得る．このような場合，その部分の予測精度が極端に下がるため，その数が 40 より大きくなった時，S_{tot} と同様の方法で求めた S_{tot}^{res} での各二次構造のカウント値が加算された．

今回の改良では，タンパク質全体の適合性スコア (S_{tot}) の 3D-1D アライメント結果に対し，SSThread では単に二次構造タイプ別に多数決をとって予測としていたのに対し，2 種類の重み関数を導入して，各二次構造タイプ別に加算されたスコア値が高いものを予測とした．1 つ目の重み関数 ($f(C_r)$) は，前後 5 残基 (計 11 残基) の部分配列の相関係数 (C_r) [32] の値に応じてその中心残基の二次構造に重みを与えるものである．西川と大井は，この相関係数を用いて部分的ホモロジー配列を定義して，ホモロジー法と言う二次構造予測法を開発した [27]．もう一つの重み関数 ($g(S_{tot})$) は，二次構造全体に対しての適合性スコア (S_{tot}) の値に応じて，その配列全体の二次構造に重みを与えるものである．

これらの関数の導入によって計算される α- ヘリックス，β- ストランド，コイルのスコア値：$V_{\alpha,i}$, $V_{\beta,i}$, $V_{c,i}$ は，それぞれ以下の式で計算され，その値が最大の二次構造タイプが予測結果となる．

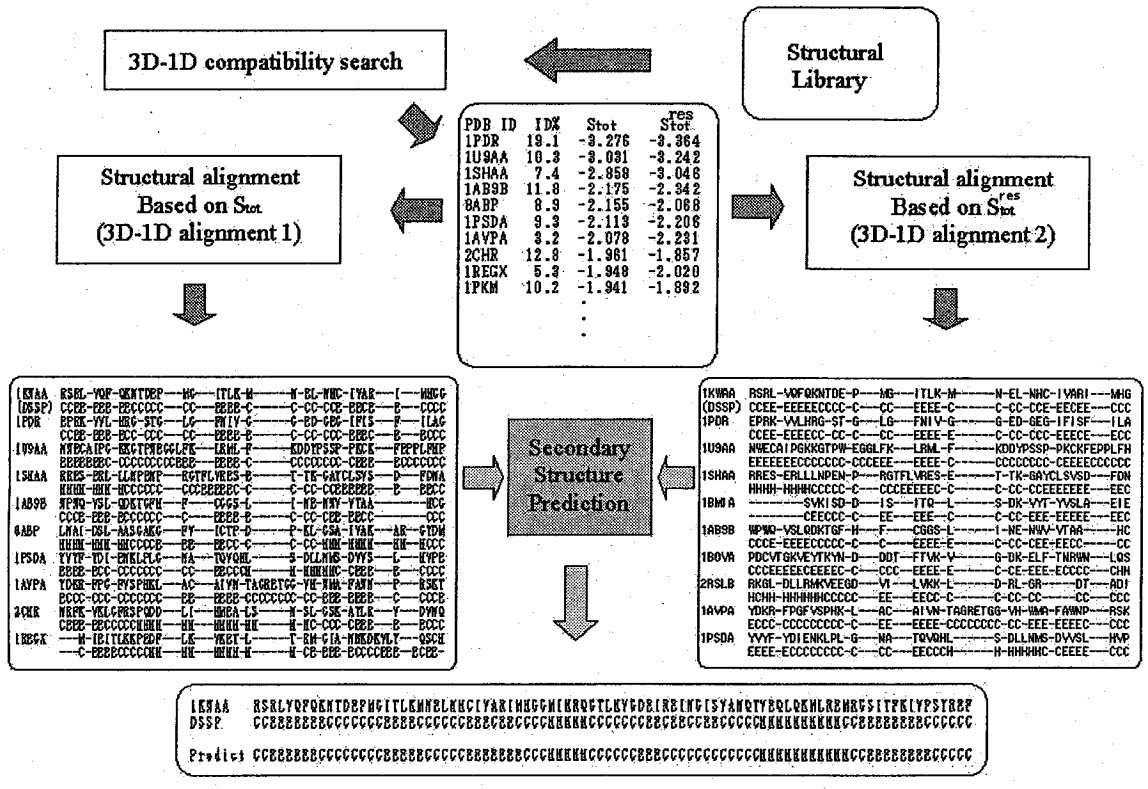

図 5.4: New SSThread の概念図

$$V_{\alpha,i} = \sum_{j=1}^{50} W_\alpha(f_\alpha(C_r(i,j)) + g_\alpha(S_{tot}(j)))\delta_\alpha(i,j) + \sum_{j=1}^{50} W_{\alpha,res}\delta_{\alpha,res}(i,j), \qquad (5.1)$$

$$V_{\beta,i} = \sum_{j=1}^{50} W_\beta(f_\beta(C_r(i,j)) + g_\beta(S_{tot}(j)))\delta_\beta(i,j) + \sum_{j=1}^{50} W_{\beta,res}\delta_{\beta,res}(i,j), \qquad (5.2)$$

and

$$V_{c,i} = \sum_{j=1}^{50}(f_c(C_r(i,j)) + g_c(S_{tot}(j)))\delta_c(i,j)) + N_{c,res}(i) \qquad (5.3)$$

ここで，iは，3D-1D アライメント上の残基位置，jは，ソートリスト上の順位を表し，W_α と W_β は，3D-1D アライメント 1 における α- ヘリックスと β- ストランドと重み係数，f_α, f_β と f_c は，α- ヘリックス，β- ストランドとコイルの C_r の重み関数，g_α, g_β と g_c は，それぞれ α- ヘリックス，β- ストランドとコイルの S_{tot} の重み関数，$\delta_k(i,j)$ は，0 (kでない) または 1 (kである) ($k = \alpha$- ヘリックス，β- ストランドとコイル) である．

また，$W_{\alpha,res}$ と $W_{\beta,res}$ は，3D-1D アライメント 2 における α- ヘリックスと β- ストランドと重み係数，$\delta_{k,res}(i,j)$ は，0 (kでない) または 1 (kである) ($k = \alpha$- ヘリックス，β- ストランドとコイル)，$N_{c,res}(i)$: 3D-1D アライメント 2 におけるコイルの数である．

5.3.2 結果

本研究において，各種の重み関数やパラメータを決めるためのデータセット（学習データセット）として，SSThread の研究で用いられたデータで残基長が 400 以下のタンパク質を用いた．

最初に，部分配列の相関係数 C_r の重み関数を求めた．学習データセットを用い，次式でそれぞれの重み因子 ($\mathrm{E}_{C,k}(C_r)$) を計算し，C_rと$E_{C,k}(C_r)$の関係を求めた．

$$E_{C,k}(C_r) = (n_{c,k}(C_r)/n_{u,k}(C_r)) \times (N_{u,k}/N_{c,k}) \qquad (5.4)$$

($k = \alpha$- ヘリックス, β- ストランド，コイル)

ここで，$n_{c,k}$ と $n_{u,k}$ は，C_r の相関係数値で，それぞれの二次構造タイプで予測が正しかった数と間違った数，$N_{c,k}$ と $N_{u,k}$ は，それぞれの二次構造タイプで予測が正

表 5.7: 部分配列の相関係数 C_r に対する重み因子の値

C_r	-0.55	-0.45	-0.35	-0.25	-0.15	-0.05	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
α	0.00	0.45	0.77	0.79	0.85	0.94	0.97	1.03	1.15	1.18	1.24	1.41	1.68	2.02	4.85	2.24
β	0.00	1.59	1.05	1.00	1.01	0.98	0.99	1.01	0.99	1.01	1.06	1.03	1.15	1.25	3.08	5.75
coil	0.77	1.13	1.01	1.03	0.98	0.98	0.99	0.99	1.02	1.03	1.07	1.05	1.16	1.29	1.32	4.60

表 5.8: 3D-1D 適合性スコア S_{tot} に対する重み因子の値

S_{tot}	-3.05	-2.95	-2.85	-2.75	-2.65	-2.55	-2.45	-2.35	-2.25	-2.15	-2.05	-1.95	-1.85	-1.75	-1.65	-1.55	-1.45
α	2.82	1.34	2.05	0.85	1.59	1.24	0.93	1.03	1.30	1.07	1.00	0.95	1.00	0.97	0.93	0.93	0.92
β	3.38	2.17	2.01	1.68	1.85	1.21	1.27	1.38	1.00	1.13	1.07	0.95	0.98	0.91	0.98	0.88	0.94
coil	1.83	1.22	1.46	1.14	1.22	1.02	1.07	1.16	1.03	0.99	1.03	1.02	0.97	0.96	0.98	0.98	0.93

しかった数と間違った数の合計である．結果を表5.7に示す．この結果を表す最適関数を各二次構造タイプ別（f_α, f_β, f_c）に求めた．

次に，3D-1D 適合性スコア (S_{tot}) の重み関数を求めた．求め方は，部分配列の相関係数 (S_{tot}) の重み関数と同じで，次式でそれぞれの重み因子 ($E_{S,k}(S_{tot})$) を計算し (表5.8)，最適関数を各二次構造タイプ別 (g_α, g_β, g_c) に求めた．

$$E_{S,k}(S_{tot}) = (n_{c,k}(S_{tot})/n_{u,k}(S_{tot})) \times (N_{u,k}/N_{c,k}) \tag{5.5}$$

$$(k = \alpha\text{- ヘリックス}, \beta\text{- ストランド}, \text{コイル})$$

ここで，$n_{c,k}$ と $n_{u,k}$ は，S_{tot} の相関係数値で，それぞれの二次構造タイプで予測した結果が正しかった数と間違った数，$N_{c,k}$ と $N_{u,k}$ は，それぞれの二次構造タイプで予測が正しかった数と間違った数の合計である．

その他の重みパラメータ (W_α, W_β, $W_{\alpha,res}$, $W_{\beta,res}$) は，SSThread と同じ方法によって，W_α=1.25, W_β=1.2., $W_{\alpha,res}$=1.74, $W_{\beta,res}$=1.21 を得た．

次に，先の SSThread の実験で指摘された大きなタンパク質の予測精度が悪化する問題を確認するテストを行った．400 残基より長い配列を含めた学習セット全てのタンパク質の二次構造予測を行い，タンパク質を 100 残基ごとにグループ分けしてその平均予測精度を求めた (図 5.5)．400 残基を越えると 10% 以上予測精度が下がっている．

この問題に対応するために，先の研究では，ドメインごとに配列を分割して予測をしていたが，本来，立体構造が明らかにならないとドメインの切れ目はわからない．

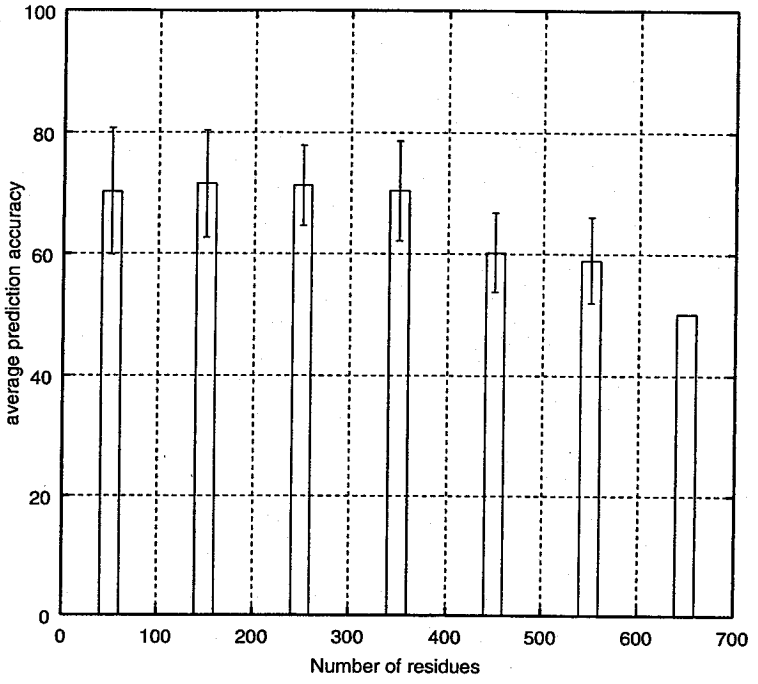

図 5.5: New SSThread における残基数の違いによる予測精度の分布

そこで本研究では，予測する配列を自動的に二分割し，重なる部分を 100 残基づつ加えた配列で二次構造予測を行い，その結果をつなぎ合わせて全体の予測とした．つなぎ目の 100 残基の予測は，前半と後半の 50 残基を，それぞれ前と後の配列の予測結果を採用した．この方法の評価テストを，学習セット内の 400 残基より長いタンパク質 9 個を使って行った (表 5.9)．この分割法を用いることによって，平均予測精度が 58.1% から 64.7% まで改善された．

上記のテストで確立された New SSThhread （400 残基より大きいタンパク質は分割法を使う）を用い，テストセットの 62 タンパク質の二次構造予測を行った (表 5.10)．比較のために，SSThread でも同じ条件（新しい構造ライブラリを用い，大きなタンパク質では分割法を使用）で予測を行った．その結果，平均予測精度が 3.4% 改善され，71.3% の値を得た．

5.3.3 改良の効果

部分配列類似性と 3D-1D 法における適合性スコアの値の重み因子を SSThread に導入する改良を行い，New SSThread を開発した．その際，構造ライブラリの更新と大きなタンパク質の予測に対応するための自動分割の手法も開発した．本研究の成果で

表 5.9: 学習セット内で 400 残基より大きなタンパク質の予測精度 (Q_3)

Code	Length	Class	new SSThread (%)	
			whole	two segments
1CHMA	401	α / β, $\alpha + \beta$	63.3	61.8
1HPLA	449	α / β, β	54.8	62.1
1SRP	471	$\alpha + \beta$, β	62.8	65.8
1DDT	535	$\alpha + \beta$, α, β	61.3	62.1
1CTN	540	α / β, β	59.3	63.1
1GTR	547	β, α / β, $\alpha + \beta$	48.1	57.2
1AOZ	552	β, β, β	65.2	69.7
1DLC	584	α, β, β	60.8	70.0
1TRKA	678	α / β, α / β, α / β	50.1	68.1
average	4807		58.1	64.7

ある New SSThread の性能評価は，表 5.10に示したが，各改良点の寄与を調べるために，学習セットの 400 残基以下のタンパク質を用いて，詳細なテストを行った．各改良ごとの平均予測精度を表 5.11に示す．構造ライブラリの更新により，その数は 325 から 1,089 にまで増え，その効果で全体の予測精度が 0.8% 改善された．個々の二次構造タイプの平均予測精度では，β- ストランドが目立て良くなった．これは，構造ライブラリの拡大が，配列上離れた残基同士の相互作用によって形成される，β- ストランドの予測に効果があったことを示しており，十分なテンプレート構造が β- ストランド予測には必要であることを示唆している．部分配列類似性の重み因子の効果は，全体平均で 0.4% しか予測精度が上がらず，効果が少ないように見えるが，α- ヘリックスの予測精度を確実に上げている．α- ヘリックスが，配列上近傍の残基同士の相互作用によって形成されるので，これはリーズナブルな結果と言える．3D-1D 法における適合性スコアの値の重み因子の導入は，効果的であった．α- ヘリックスの予測精度を約 8% 上げ，全体の予測精度でも 1.1% 改善している．ただ，この効果は，3D-1D 法の適合性スコアの高いテンプレートが見つからないと効果を発揮しないため，効果にばらつきがあると考えられる．両方の重み因子を用いた効果により，全体の予測精度で 1.7% の改善が得られた．

表 5.10: テストセット内のタンパク質の予測精度 (Q_3)

Code	Length	New SSThread Q_3 (%)	SSThread Q_3 (%)
1bazA	49	85.7	81.6
1aojB	60	48.3	45.0
1kigI	60	73.3	76.7
1tuc	61	55.7	67.2
1sknP	74	70.3	74.3
1am9B	75	77.3	64.0
1bb9	83	53.0	72.3
1a32	85	88.2	83.5
1kwaA	88	85.2	60.2
1f36A	89	80.9	79.8
1a4pA	92	75.0	81.5
1ecmB	95	91.6	90.5
1mb1	98	56.1	48.0
1g31A	107	61.7	66.4
1sfp	111	76.6	76.6
1bnkA	120	70.8	71.7
1buoA	121	66.9	62.8
1byl	122	64.8	54.9
1bdyB	123	69.9	52.0
1dfx	125	53.6	60.0
2eifA	133	63.2	66.9
1akr	147	88.4	67.3
1a95B	150	68.7	63.3
1amx	150	66.7	62.0
1bd8	156	90.4	67.9
1bfrA	158	82.9	82.3
1a73A	162	56.8	60.5
1au1A	166	71.7	68.1
1alvA	173	85.0	83.2
1cv8	173	56.6	51.4
1bnlA	178	61.8	69.7
1tyfA	183	69.4	57.9
1behA	184	79.3	72.3
1np4	184	62.0	64.7
1nkr	195	63.6	67.7
1oakA	196	76.5	73.0
2xat	208	81.3	73.1
1bquB	215	66.5	69.8
1a2zA	220	70.0	71.4
1rypF	233	86.3	57.9
1d2nA	246	77.6	80.1
1rypB	250	77.2	64.0
1a8p	257	61.1	63.4
1jfrA	260	68.1	67.3
1a81A	266	72.6	68.0
1eny	268	81.0	68.3
1a02N	280	74.3	73.6

(Continue)

Code	Length	new SSThread Q_3 (%)	SSThread Q_3 (%)
1fts	295	73.2	71.2
1uox	295	52.5	54.6
2ptd	296	71.3	64.2
1fsz	334	80.8	73.7
3thi	362	69.6	74.6
1bhe	376	61.7	72.1
1jdbF	384	71.9	73.7
2qwc	388	83.0	67.5
1bag	425	77.9	67.3
1a0cA	437	73.2	70.7
1a6cA	513	65.5	65.9
1bfd	523	75.0	71.5
1akn	547	66.0	62.9
1b0l	691	65.4	57.2
1bf2	750	71.3	73.1
average	13845	71.3	67.9

表 5.11: 各改良ごとの平均予測精度 (Q_3)

	α helix %	β strand %	Coil %	Total %
SSThread (old structural lib.)	58.3	54.8	79.3	68.5
SSThread (new structural lib.)	59.8	61.9	77.1	69.3
New SSThread (new structural lib., weight of C_τ)	66.9	55.7	76.7	69.7
New SSThread (new structural lib., weight of S_{tot})	67.9	60.7	75.9	70.4
New SSThread (new structural lib., weights of $C_\tau and S_{tot}$)	69.0	57.0	78.2	71.0

5.4 並列タンパク質情報解析 (PAPIA) システム

並列タンパク質情報解析 (PAPIA:PArallel Protein Information Analysis) システム [14] は，タンパク質配列や立体構造に関する情報解析を，分散メモリ型の並列計算機で高速に並列実行するもので，主な計算機能として，1) タンパク質類似構造検索，2) タンパク質相同配列検索，3) マルチプルアライメント等があり，現在，WWW 上 (URL: http://www.rwcp.or.jp/papia/) で公開されている (図 5.6).

PAPIA システムは，現在，図 5.7に示す PAPIA クラスタ上で動作している．PAPIA クラスタは，新情報処理開発機構 つくば研究センタ 並列分散ソフトウェアつくば研究室で開発された PC クラスタと NetBSD を基本 OS にして動作する SCore クラスタソフトウェア [33] の技術を導入して，PAPIA システムを動作させるために作成

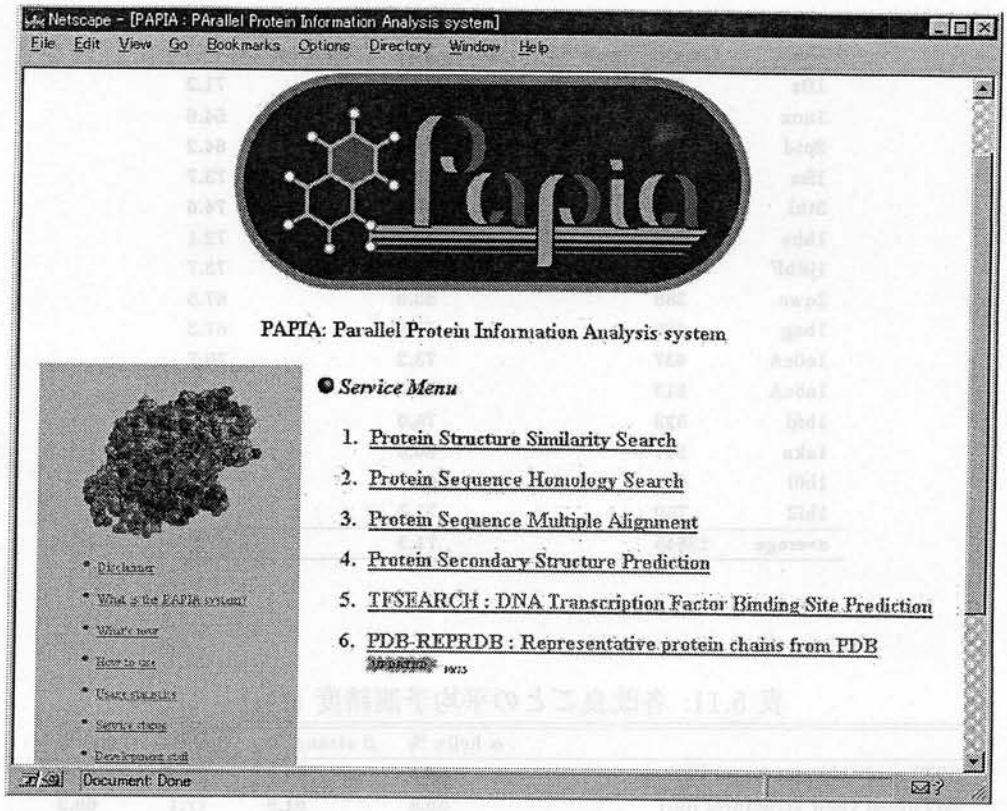

図 5.6: PAPIA ホームページ

したクラスタである (図 5.7, 表 5.12).

PAPIA システムの計算機能の大部分は，鬼塚らによる PAPIA ライブラリ [15, 16] と呼ぶオブジェクト指向の共通プログラムライブラリで書かれてる．PAPIA ライブラリは，PDB 代表タンパク質チェイン決定システム並列版より導入しているライブラリで，我々のシステム開発の効率化に大きく貢献している．

PAPIA システムが検索対象としているデータベースは，立体構造検索では PDB の 1 種類，配列検索では PDB と SWISS-PROT の 2 種類である．そのうち PDB は，全てのエントリを立体構造検索の対象とした場合，WWW 上で対話型の検索サービスを行うには，計算時間がかかりすぎるため，冗長性をなくしたデータベースにする必要があった．構造を検索対象としているため，配列相同性 (ID%) だけを指標に代表を選ぶと部分構造の違う構造を対象から落としてしまう危険性が高いので，PDB 代表タンパク質チェイン決定システムを用い，表 5.13の基準でデータの質を決め，配列相同性 (ID%) と構造類似性 (Dmax) のしきい値を，それぞれ 95% 以上と 10 Å 以下とした非冗長化した PDB 代表セットを作成し，立体構造データベースとした．

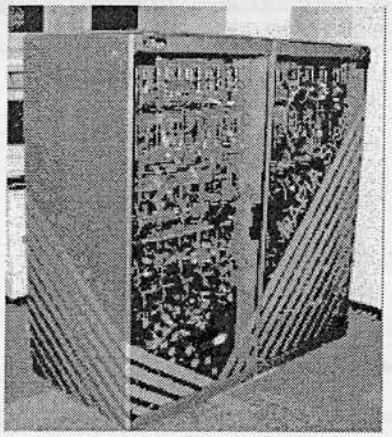

図 5.7: PAPIA クラスタ

表 5.12: PAPIA クラスタの仕様

プロセッサ数	64 計算ノード + 2 モニタノード
プロセッサ	Pentium Pro, 200MHz
メモリ	256MB ノード
ローカルディスク	4.1GB ノード
ネットワーク	Myrinet 1.28 Gbit 秒 + 100-BaseT
OS	NetBSD 1.2.1 + SCore-D
寸法	H1600 × W1600 × D800 mm
製作年導入年	1998 年 2 月 製作

表 5.13: PAPIA システム用 PDB 代表タンパク質チェインの基準

Factors for elimination	Threshold	Default priority
分解能	なし	1
R ファクター	なし	2
チェインブレイク数 (少ないほど上位)	なし	3
標準的なアミノ酸残基種以外の残基の比率 (小さいほど上位)	なし	4
主鎖原子の座標を欠く残基 (Cα のみ) の比率 (小さいほど上位)	なし	5
側鎖原子の座標を欠く残基 (主鎖の原子のみ) の比率 (小さいほど上位)	なし	6
配列長 (長いほど良い)	なし	7
変異型を含む	-	8
複合体を含む	-	9
NMR データを含む	-	-

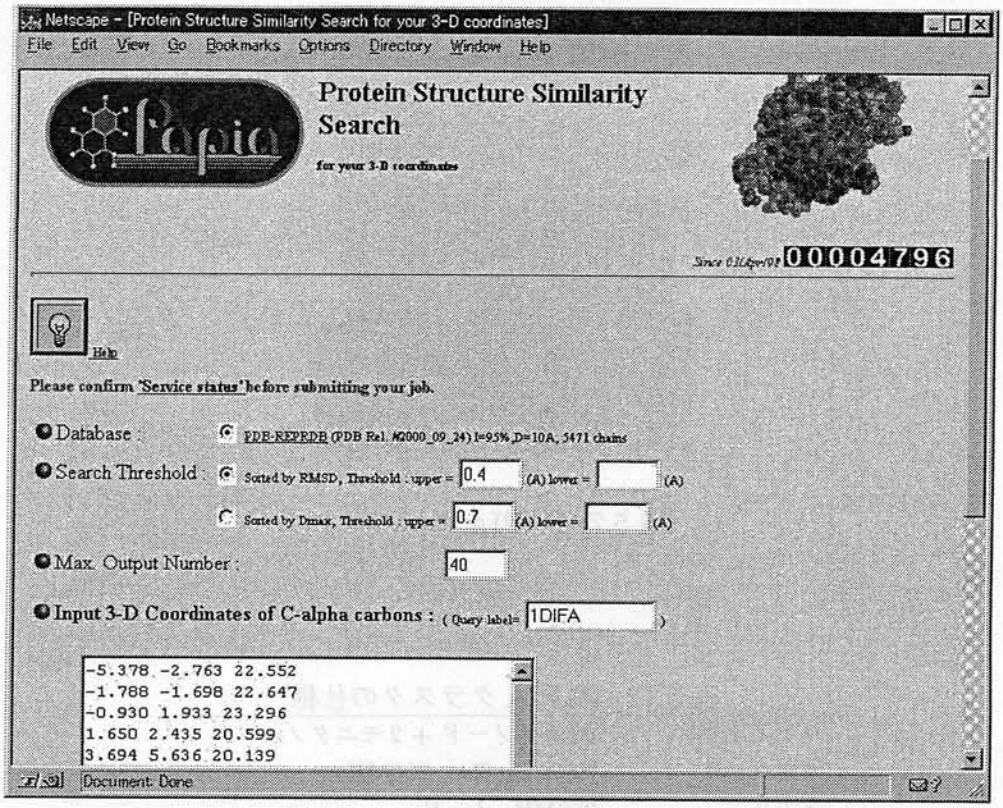

図 5.8: PAPIA 立体構造検索ページ

PAPIA システムでは，全ての計算メニューを融合した単一のプログラム（PAPIA 計算サーバ）が，起動時に PC クラスタ上の全ノードで立ち上がる．PC クラスタ上の１ノードが入出力用のマスタ役となり，残りのノードが計算担当のスレーブとなる．マスタおよびスレーブのプロセスは，デーモンとしてシステム上に常駐し，外部からの計算リクエストの到着を待つ．リクエストはソケット通信で，“プログラム名＋入力データ” の形式でマスタに到着し，マスタは必要に応じて，仕事をスレーブに対して割り当てる．スレーブ上での計算あるいは検索結果をまとめて整形し，返信するのもマスタの仕事である．

立体構造データベースと SWISS-PROT は，検索要求に迅速にこたえるため，起動時にスレーブ上に読み込まれ，内容をパージングして，メモリ上にオブジェクトとして展開してある．メモリ容量の制限のため，スレーブごとに担当するエントリの部分集合を定めてある．

ここでは “タンパク質類似構造検索” の操作のみ紹介する．図 5.8に WWW 上の立体構造検索のパラメータ設定画面の例を示す．この例では，検索したい構造を X,Y,Z

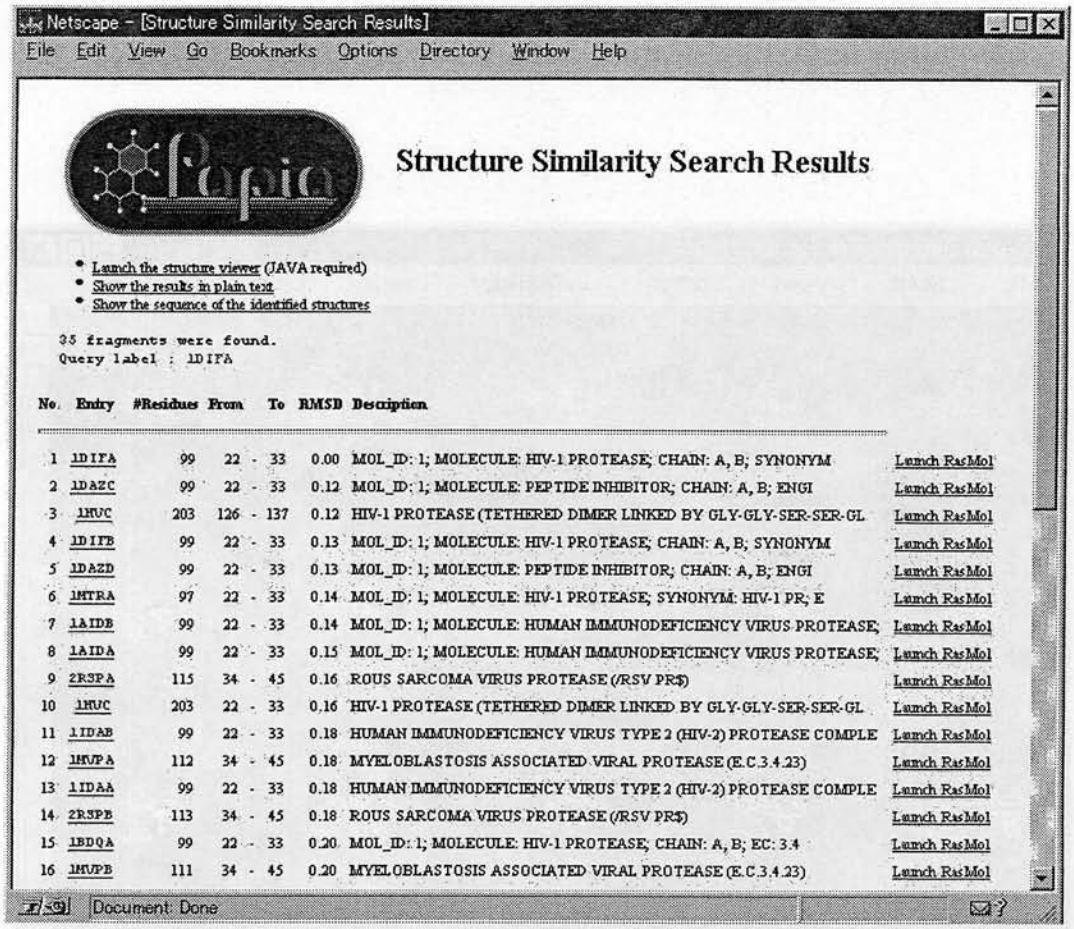

図 5.9: PAPIA 立体構造検索

の座標値を入力しているが，その他の PDB の “ID” と残基番号で構造を指定して検索することもできる．検索対象の立体構造データは，上記のように PDB-REPRDB で決められた代表タンパク質チェインの立体構造である．検索のしきい値として RMSD か Dmax のどちらかを指定し，実行することによって，指定した立体構造と類似した構造をしきい値範囲内で検索することができる．

PDB エントリ全体の立体構造の重ね合わせには，膨大な時間が必要となるため，このようなタンパク質立体構造検索サービスは世界的にもほとんど例がない．発見された類似構造は，図 5.9のようにリスト表示される．それらの構造が見たい場合は，JAVA による立体構造ビューワー (図 5.10) や RasMol プログラムで即座に確認することができる．

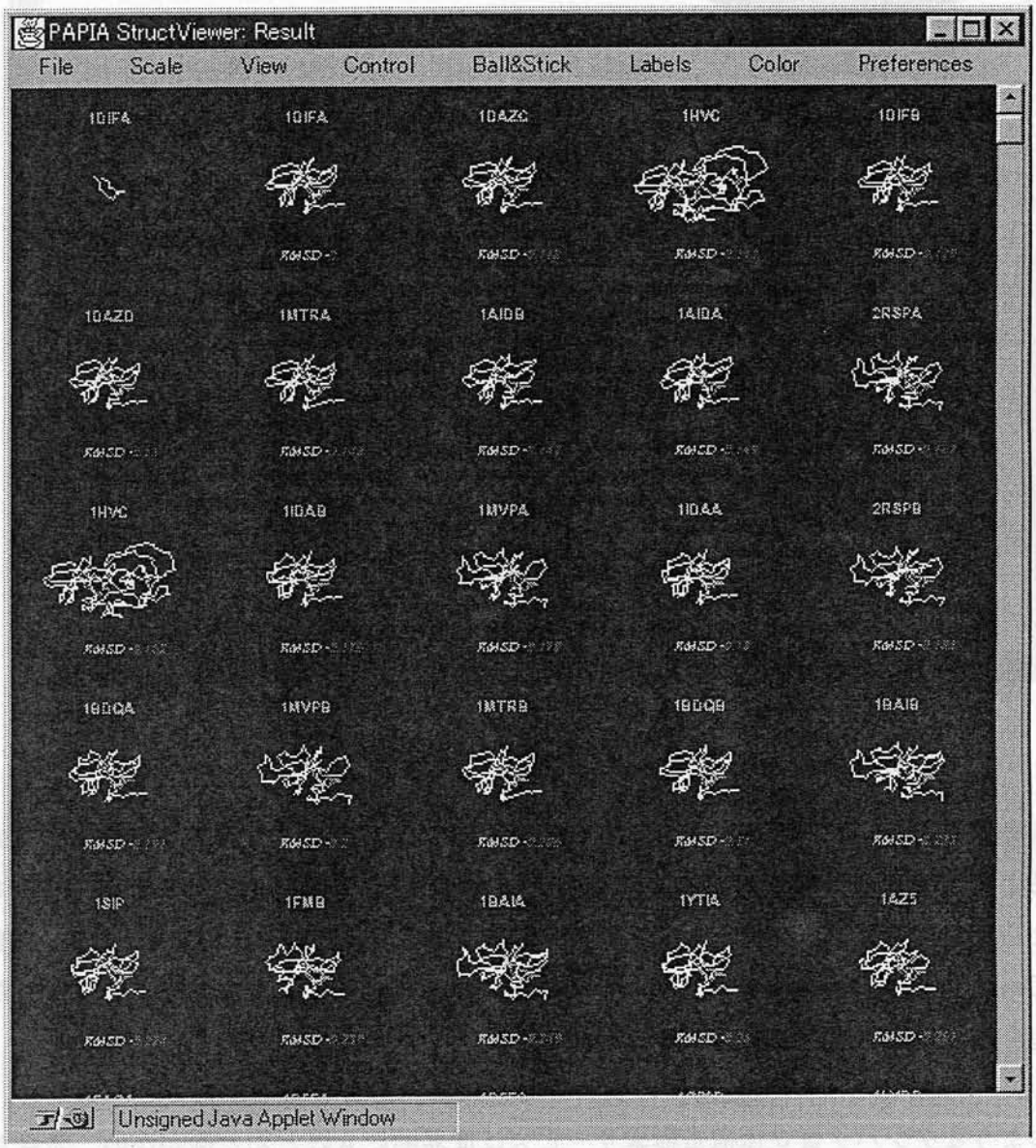

図 5.10: JAVA による PAPIA 立体構造検索結果の表示

5.5 まとめ

本章では，本研究を始める動機となった「アミノ酸配列に基づくタンパク質二次構造予測」の研究と，本研究の利用例として，「3D-1D 法と部分配列類似性を用いたタンパク質二次構造予測」と「並列タンパク質情報解析 (PAPIA) システム」の研究について述べた．

「アミノ酸配列に基づくタンパク質二次構造予測」では，アミノ酸配列に基づいく二次構造予測法の中から，方法論が異なる 8 種類の方法を取り上げ，それぞれの予測精度の比較を行った．その結果から，平均予測精度の良かった上位 5 手法を組み合わせたジョイント予測法を作成し，平均予測精度で個々の予測法より 2-5% 向上させ，64.8% を得た．経験的（統計的）二次構造予測の基となるタンパク質立体構造の選出の重要性を認識させられた研究で，良質のタンパク質立体構造選出法である本研究に着手するきっかけとなった．

「3D-1D 法と部分配列類似性を用いたタンパク質二次構造予測」では，経験的立体構造予測の基礎データとして，一般に利用されている基準 (ID% $>$ 30 %) で PDB-REPRDB を基に作成し，構造ライブラリを充実させたことによって，予測精度を約 1% 向上させることができた．今後も定期的に構造ライブラリを更新することよって，予測精度の改善を行っていきたい．また，配列の長い，大きなタンパク質に対応するために，配列を分割して予測する方法を導入したが，部分構造の違いをもっと厳密に調べ，チェインより小さい構造単位で基礎データを作成する必要があると思う（例えば，ドメインレベルでの予測法の開発など）．そのためには，現在のチェインの代表決定システムからドメインの決定システムへの移行が必要になってくる．この点は，今後の研究課題としたい．

「並列タンパク質情報解析 (PAPIA) システム」への応用は，まさに立体構造の類似性を分類の基準に入れた効果が出た良い例だと思う．本システムにより，配列相同性による分類では見落としていた，部分構造の異なるデータを含んだ代表セットを作成し，検索対象のデータベースにすることによって，特徴的な部分構造を見落とす可能性はほとんどなくなり，かつ効率的な検索が可能となった．今後，部分構造の研究を行う上で，本基準の効果がさらに生かせると期待している．

本システムは，初期バージョンが PAPIA WWW サーバーで公開されてた 1997 年 8 月から，2000 年 11 月の間に 4,000 件以上利用されており，ここに上げた例は，我々の研究への利用例であるが，同様な利用法を含め，様々な研究で本システムが利用さ

れると期待している.

第 6 章

結論

6.1 はじめに

本論文では，冗長でデータの質にもばらつきがあるタンパク質立体構造データベース (PDB) を配列相同性と立体構造類似性を考慮して分類し，非冗長な PDB 代表タンパク質チェインを決定する新たな手法を提案した．この手法により，従来法では考慮されなかったタンパク質立体構造の部分的な違いを検出して，従来法では見落としていた部分構造が異なるチェインを別の代表チェインとして選ぶことが可能になった．

本章では，各章における達成成果と，研究の発展性についてまとめる．

6.2 研究成果

本研究で提案したPDB代表タンパク質チェイン決定法の特徴は，PDBデータの質とタンパク質立体構造の部分構造に着目した点にある．

二次構造予測法やスレッディング法などの経験的タンパク質立体構造予測法の基礎データとなるトレーニング用タンパク質は，データの質によって構造が微妙に異なる場合があるので，できるだけ質の良いデータを利用するべきである．また，構造の中にチェインブレイクがあったり，残基種がわからない残基が存在するデータだと，統計が取りづらいばかりか，経験的パラメータに悪影響を及ぼす危険があるので，できるだけそのようなデータは使用するべきではない．PDBデータの質を優先して代表チェインを決める本手法は，選ばれた代表の個々のデータの質をチェックする必要なく，利用できる点で優れている．

また，タンパク質の配列は，挿入，欠損，置換を繰り返し進化を遂げているが，その立体構造は比較的保存されていると言われている．そのため，タンパク質の立体構造を配列の相同性のみで分類する方法が一般的であったが，部分構造に着目すると配列の相同性が高くても，構造が変化している場合がある．このようなケースを無視して，経験的タンパク質立体構造予測の基礎データにどちらかを利用した場合，これもまた経験的パラメータに悪影響を及ぼす危険がある．このようなケースを抽出するのには，本システムは有効である．

また，部分構造が異なるチェインを別の代表チェインにすることが可能になり，特徴的な部分構造を網羅的に調査する研究や，タンパク質配列における，挿入，欠損，置換による部分構造の変化を調べる研究などに用いるチェインセットを容易に作成することが可能になった．これにより，研究対象となるチェインセット作成に多大な労力を費やす必要がなくなり，部分構造の研究に専念できるようになり，部分構造の研究にも貢献できたと思う．

第2章では，従来の配列相同性に基づくタンパク質立体構造の分類手法およびその代表タンパク質を決定する方法に，新たに構造類似性にも着目し，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分類手法を提案し，その手法を用いたPDB代表タンパク質チェイン決定システムを作成した．本システムにより，従来法では配列相同性を用いて近似的に立体構造を分類し，代表を決めていたため，タンパク質立体構造予測に用いるデータとしては，不十分であった代表タンパク質チェインデータを，直接立体構造を比較し，分

類することによって，近似によらない分類が可能になり，正確な代表タンパク質チェインデータを得られるようになった．また，従来法では，見逃してしまう可能性が高かった，特徴ある部分構造も本システムによって，見逃す可能性をなくすことができるようになった．

第3章では，第2章で開発した配列の相同性 (ID%) だけでなく，構造の類似性にも注目し，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PDBの代表タンパク質決定システムを，PAPIAライブラリと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6,127本を約1.5時間で分類することができた．

第2章のシステムでは，自動化が不十分でかつ，処理時間が膨大だったたため，研究者の要求に応えられるような，様々な分類基準で作成した代表タンパク質チェインセットを用意することができなかった．また，第2章のシステムでは，X線結晶回折によって解析された立体構造と比較することの妥当性に疑問があり，また，構造比較をするモデル選びの方法が決められず，あらかじめ分類対象から削除していたNMRで解析されたデータを分類に加え，代表チェインの数を大幅に増やすことができた．

本システムにより決定されたPDB代表タンパク質チェインは，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，世界から2,500回以上アクセスされた．

第4章では，利用者が会話話形式で配列と構造の違いに基づいた選択基準を指定しで，PDB代表タンパク質チェインデータベース (PDB-REPRDB) を作成する新しいシステムについて述べた．本システムは，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，順位リストの全チェイン間の配列相同性 (ID%) や構造類似性 (Dmax) の計算結果をテーブルとしてあらかじめ用意しておき，オンデマンドで様々な基準 (良質の基準や各配列および構造の類似性のしきい値など) で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である．本システムによって，我々が指向している，配列の相同性は高いが，残基の挿入，欠損や置換による部分的な構造の違い，それに複合体を形成したことによって生じる微妙な構造変化などを考慮して代表チェインを，すばやく得ることが可能となったので，様々な基準でその変化を検出できるようになった．

本システムは，1999年4月からPAPIA WWWサーバーで利用可能となり，2000年11月時点で約1,300件利用されている．現在，PDB-REPRDB用のデータベース

の更新は，現環境でPDBの更新が毎週行えるようになったので，従来の3ヶ月に1度のペースから1から2ヶ月に1度のペースで行っている．

第5章では，PDB代表タンパク質チェイン決定システムを作成する動機となった「アミノ酸配列に基づくタンパク質二次構造予測」の研究について述べた後，PDB代表タンパク質チェイン決定システムを用いて作成したPDB代表タンパク質チェインデータベース(PDB-REPRDB)が，どのように利用できるかを例をあげて述べた．「3D-1D法と部分配列類似性を用いたタンパク質二次構造予測」では，一般に利用されている基準(ID% $>$ 30 %)でPDB-REPRDBを作成して，経験的立体構造予測の基礎データとしたが，部分構造の違いをもっと厳密に調べ，チェインより小さい構造単位で基礎データを作成する必要があると思う．この点は，今後の研究課題としたい．PAPIAシステムへの応用例は，配列相同性による分類では見落としていた，部分構造の異なるデータを含んだ代表セットを作成し，検索対象のデータベースにすることによって，特徴的な部分構造を見落とす可能性はほとんどなくなり，かつ効率的な検索が可能となったと言う点で，まさに立体構造の類似性を分類の基準に入れた効果が出た良い例だと思う．今後，部分構造の研究を行う上で，本システムの効果がさらに生かせると期待している．

本システムは，初期バージョンがPAPIA WWWサーバーで公開されてた1997年8月から，2000年11月の間に4,000件以上利用されており，ここに上げた例は，我々の研究への利用例であるが，同様な利用法を含め，様々な研究で本システムが利用されると期待している．

6.3 今後の課題

PDBエントリーの増加のペースは急激で，第4章の最新システムであらかじめ行っておく全ペアに対する配列相同性と構造類似度の計算量は，今後さらに増加し，処理時間も膨大になり，近い将来，データの更新が現在のペースで行えない事態になる．この問題を解決するために，PDBの更新に伴うこの“類似データベース”（チェイン同士のID%，RMSDとDmaxの値を保存)の作成処理を，現在の全ての計算を更新の度に行う仕様から，更新されたPDBデータのチェインと類似データベース内の既存チェイン同士で類似データを計算するだけで，類似データベースを更新する仕様に変更する予定である．これによって，類似データベース更新の効率化が可能となり，PDBの更新に即応したシステムの提供が可能となる．本システム改良は，早期に実現したいと考えている．

また，本研究を利用した研究として，特徴的な部分構造の抽出を行い，本研究の有効性についても実証していく予定である．

また，本システムは，PDB内のチェインを配列の相同性（ID%）と立体構造の類似性（RMSDまたはDmax）を基に分類し，データの質の良いチェインを代表に決めるシステムで，主にタンパク質立体構造予測の学習データの選択やユニークな部分構造を含んだチェインをもれなく代表にすることを目的にしたシステムなので，タンパク質立体構造のファミリー代表とは，必ずしも一致しない．PDB_SELECTの代表との比較やSCOPやCATHなどドメインレベルでの全体構造の類似性で分類されたデータと比較することによって，本システムで作成されるPDB-REPRDBとそれらの関係を把握したいと考えている．

謝辞

本研究を行うにあたり，御指導，御教授と格別なるご配慮を賜りました大阪大学大学院基礎工学研究科情報数理系専攻 橋本昭洋教授に深く感謝致します．

また，本論文をまとめるにあたり，貴重な時間を割いて頂き，懇切なる御指導と有益な御助言を賜りました大阪大学蛋白質研究所生体分子解析研究センター蛋白質立体構造データ解析研究系 中村春木教授，ならびに，大阪大学大学院基礎工学研究科情報数理系専攻 萩原兼一教授，柏原敏伸教授に心より感謝致します．

また，本研究を行うにあたり，終始直接御指導してくださった，大阪大学大学院基礎工学研究科情報数理系専攻 松田秀雄助教授に深く感謝致します．

工業技術院電子技術総合研究所知能情報部生命情報科学ラボ 秋山泰主任研究官には，本研究全般の御指導とともに，情報生物学における並列計算機の有用性およびその効率的利用法を御教授頂きました．ここに深く感謝致します．

また，国立遺伝学研究所生命情報センター 西川建教授には，永年にわたり情報生物学全般の御指導，御教授頂きました．ここに心より感謝致します．

本研究を進めるにあたり貴重な御意見と御助言を頂いた，京都大学化学研究所の五斗進助手と金久實教授，生物分子工学研究所情報解析研究部門 藤博幸部門長に深く感謝致します．

また，「3D-1D 法と部分配列類似性を用いたタンパク質二次構造予測」の共同研究者である国立遺伝学研究所生物遺伝資源情報総合センターの伊藤將弘氏，3D-1D 法を開発した理化学研究所ゲノム科学総合研究センター 松尾洋氏には，本研究を進めるにあたり貴重な御意見と御助言を頂きました．ここに深く感謝致します．

また，工業技術院電子技術総合研究所ならびに技術研究組合新情報処理開発機構つくばセンターの皆様には，有意義なご討論と研究の場を提供して頂きました．ここに深く感謝致します．

最後に，本研究を行うにあたり，ともに研究を進め御討論，御協力を頂いた松下技研株式会社 鬼塚健太郎氏，日本鋼管株式会社 安藤誠氏，株式会社情報数理研究所志澤由久氏，株式会社インテック ゲノム・インフォマティクス・センター 辻宇俊氏ならびに工業技術院電子技術総合研究所知能情報部生命情報科学ラボの皆様に深く感謝致します．

参考文献

[1] Anfinsen, C.B.: Principle that govern the folding of protein chains, *Science*, Vol.181, pp.223 (1973).

[2] 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, *J. Mol. Biol.*, Vol.112, pp.535-542 (1977).

[3] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E.: The Protein Data Bank, *Nucleic Acids Res.*, Vol.28, pp.235-242 (2000).

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

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

[6] Holm, L. and Sander, C.: The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.*, Vol.22, pp.3600-3609 (1994).

[7] Mizuguchi, K., Deane, C.M., Blundell, T.L. and Overington, J.P.: HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science*, Vol.7, pp.2469-2471 (1998).

[8] Sander, C. and Schneider, R.: Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, Vol.9, pp.56-68 (1991).

[9] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C.: scop: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, Vol.247, pp.536-540 (1995).

[10] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M.: CATH - a hierarchic classification of protein domain structures. *Structure*, Vol.5, pp.1093-1108 (1997).

[11] Pearson, W.R. and Lipman, D.J.: Improved Tools for Biological Sequence Analysis. *Proc. Natl. Acad. Sci.*, Vol.85, pp.2444-2448 (1988).

[12] Pearson, W.R.: Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Methods in Enzymology*, Vol.183, pp.63-98. (1990).

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

[14] Akiyama, Y., Onizuka, K., Noguchi, T. and Ando, M.: Parallel Protein Information Analysis (PAPIA) system running on a 64-node PC cluster. *Proc. of the Ninth Workshop on Genome Informatics*, Universal Academy Press, pp.131-140 (1998).

[15] Onizuka, K., Noguchi, T. and Akiyama, Y.: Parallel PDB Data Retriever 'PDB Driving Booster'., *in Lecture Notes in Computer Science*, Vol.1336, pp.389-396 (1997).

[16] 鬼塚, 野口, 斎藤, 秋山: タンパク質立体構造研究支援のための並列統合解析システムの構築, 情処研報 97-HPC-68-8, pp.45-50 (1997).

[17] Needleman, S.B. and Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, Vol.48, pp.443-453 (1970).

[18] Skinner, R., Abrahams, J.P., Whisstock, J.C., Lesk, A.M., Carrell, R.W. and Wardell, M.R.: The 2.6 Å structure of antithrombin indicates a conformational change at the heparin binding site. *J. Mol. Biol.*, Vol.266, pp.601-609 (1997).

[19] Kabsch, W. and Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymer*, Vol.22, pp.2577-2637 (1983).

[20] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y. and Kanehisa, M.: DBGET/LinkDB: An integrated database retrieval system. *Pac. Symp. Biocomput. 1998*, pp.683-694 (1998).

[21] Goto, S., Nishioka, T. and Kanehisa, M.: LIGAND database for enzymes, compounds and reactions. *Nucleic Acids Res.*, Vol.27, pp.377-379 (1999).

[22] Chou, P.Y. and Fasman, G.D.: Prediction of the secondary structure of protein from their amino acid sequence. *Adv. Enzymol.*, Vol.47, pp.45-148 (1978).

[23] Garnier, J., Osguthorpe, D.J. and Robson, B.: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, Vol.120, pp.97-120 (1978).

[24] Qian, N. and Sejnowski, T.J.: Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, Vol.202, pp.865-884 (1988).

[25] Gibrat, J.F., Garnier, J. and Robson, B.: Further development of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, Vol.198, pp.425-443 (1987).

[26] Nagano,K.: Triplet information in helix prediction applied to the analysis of super-secondary structures. *J. Mol. Biol.*, Vol.109, pp.251-274 (1977).

[27] Nishikawa, K. and Ooi, T.: Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim. Biophys. Acta*, Vol.871, pp.45-54 (1986).

[28] Lim, V.I.: Algorithms for prediction of alpha-helicies and beta-structural regions in globular proteins. *J. Mol. Biol.*, Vol.88, pp.873-894 (1974).

[29] Ptitsyn, O.B. and Finkelstein, A.V.: Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, Vol.22, pp.15-25 (1983).

[30] Matsuo, Y. and Nishikawa, K.: Assessment of a protein fold recognition method that takes into account four physicochemical properties: side-chain packing,

solvation, hydrogen-bonding, and local conformation. *Proteins*, Vol.23, pp.370-375 (1995).

[31] Ito, M., Matsuo, Y. and Nishikawa, K.: Prediction of protein secondary structure using the 3D-1D compatibility algorithm. *Comput. Appl. Biosci.*, Vol.13, pp.415-423 (1997).

[32] Kubota, Y., Nishikawa, K., Takahashi, S. and Ooi, T.: Correspondence of homologies in amino acid sequence and tertiary structure of protein molecules. *Biochim. Biophys. Acta*, Vol.701, pp.242-252 (1982).

[33] Ishikawa, Y., Tezuka, H., Hori, A., Sumimoto, S., Takahashi, T., O'Carrol, F. and Harada, H.: RWCP PC Cluster II and SCore Cluster System Software - High Performance Linux Cluster., *in Proc. 5th Annual Linux Expo*, pp.55-62 (1999).