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，その内容は現在約7,500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ず，何らかの基準で代表タンパク質を決定する必要がある．この代表を決定するには，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たり，各エントリーに対して他のエントリーが配列的および立体構造的に類似のタンパク質かどうかを調べあげ，分類を行なわなければならない．

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，近似的に配列の類似性（ID%）だけを指標にして行なわれてきた．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に対応する原子間距離の最大値（Dmax）を分類の指標として加え，従来よりも正確な分類を可能にしたPDBの代表タンパク質決定システムを開発した．

本論文では，本システムを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PDBのリリースごとに作成している．本データベースをWWWで公開し，既に世界から2,200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 construction of 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 have 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 structures has become quite obvious. We have proposed a representative chain database PDB-REPRDB, whose strategy of selection is based on the sequential and structural similarity.

We have developed a representative chain database PDB-REPRDB, and in this paper 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 already been accessed more than 2,200 times.

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

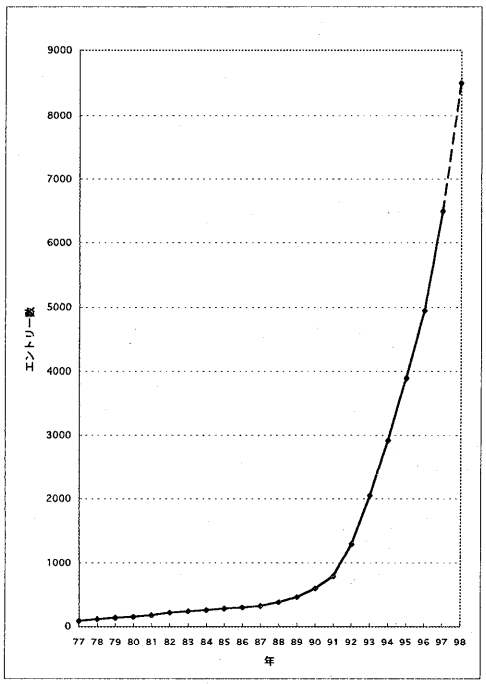


図1 タンパク質立体構造データベース (PDB) のエントリー数の推移
Fig. 1 Number of entries in Protein Data Bank (PDB)

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ファイルに1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7,500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，

- **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）$\geq$ 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）$\leq$ 10.0 Å,

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．

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統計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[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た．この代表タンパク質チェインは，“PDB_SELECT”[4)]として公開され，現在は，**配列の相同性**：ID% < 25%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[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

☆ 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．

☆☆ チェイン: 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各鎖．

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）を行なった[6]．

本論文では，この自動決定システムについて述べるとともに，さらに処理の高速化を目指して，システムの並列版を作成したので報告する．また，タンパク質立体構造を配列の相同性と構造の類似性の組み合わせで，合計8×6＝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集合セットを決定したので，その結果についても報告する．

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”[8]がある．

SCOPでは，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☆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，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なく，本論文で述べるPDB-REPRDBが最初である．

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[6]．

本論文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，特徴ある部分構造を漏れなく含むように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対応する各原子間距離の最大値），

の両方を用いた(図2)．

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r.m.s.d(root mean square deviation)値を基準にするのが一般的であるが，部分構造だけが異なるタンパク質を比較する場合，全体構造を重ね合わせた時のr.m.s.d値は，構造が類似している部分の原子間距離が小さいため，その影響を受け，部分構造の違いを的確に検出することができない．また，全体構造を重ね合わせた時のr.m.s.d値は，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，r.m.s.d値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，r.m.s.d値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，本研究では，Dmaxを分類の指標とした．

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

☆ 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．

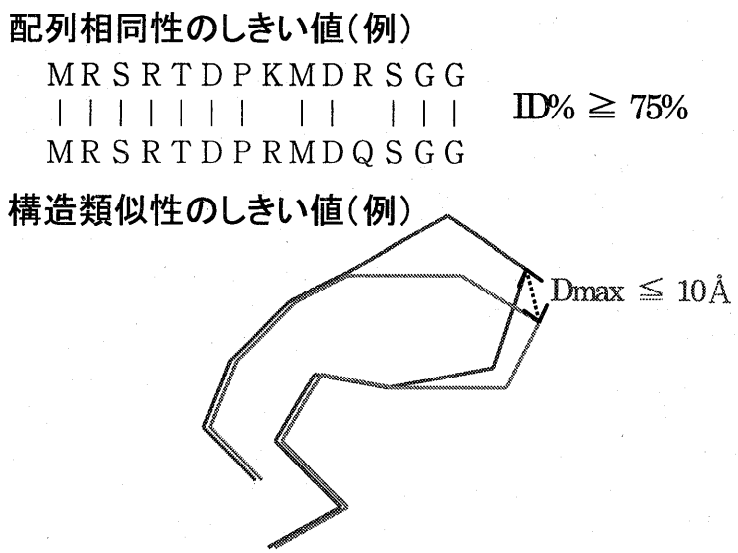


図2 タンパク質の分類基準（例）
Fig. 2 An example criteria for the classification of protein structures

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) 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)

3.3 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[9]で調べる．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，同じ残基種で並置された (例えば，図2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[10]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（図3)．

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．

4. 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．

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

☆ 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．

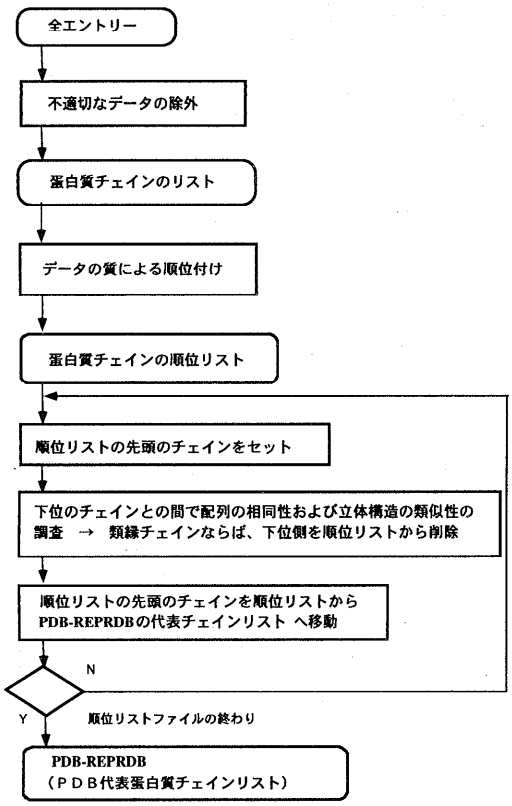


図3 PDB代表タンパク質決定システムの流れ
Fig. 3 Flow of the PDB-REPRDB system

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 は使用するプロセッサ台数とする．

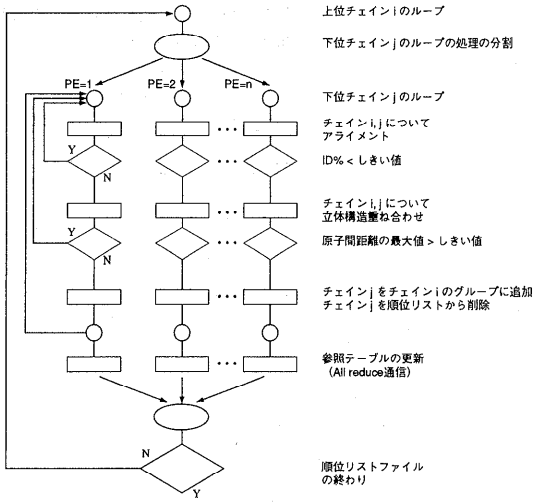


図4 並列版PDB代表タンパク質決定システムにおける計算の流れ（模式図）
Fig. 4 Diagram of the parallelized PDB-REPRDB system

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することと，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．

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
Table 1 SR2201 specification used for the performance test of the parallelized PDB-REPRDB system

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	256MB × 256=64GB 分散メモリ

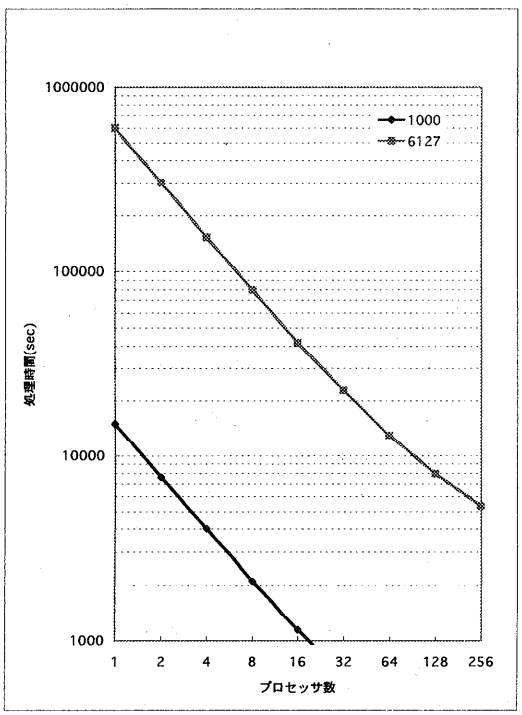


図5 SR2201上での処理時間
Fig. 5 Execution time on SR2201

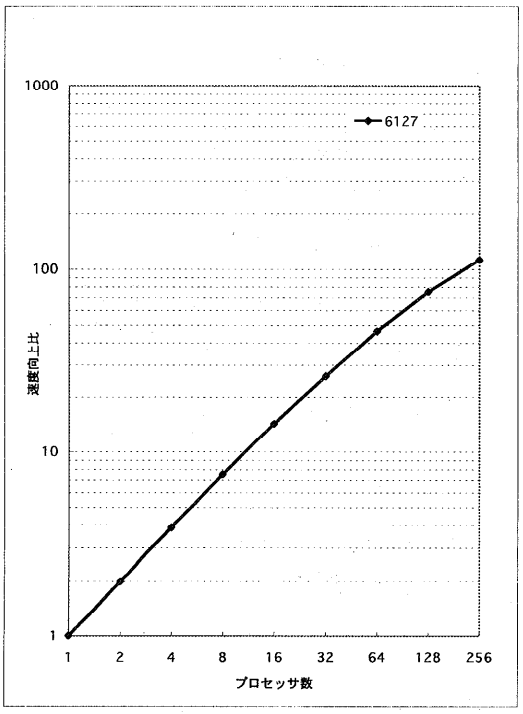


図6 SR2201上での速度向上比
Fig. 6 Speedup ratio on SR2201

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．

使用したPDBは，リリース#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とを作り，性能評価に利用した．

図5にSR2201での処理時間，**図6**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と言える．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．

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る．

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 (**表2**, **表3**）である．

総エントリー数は7,578で，“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数は，11,062本であった．(**表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% ～ 95%まで10%刻みの8通りと，**構造の類似性**：Dmax

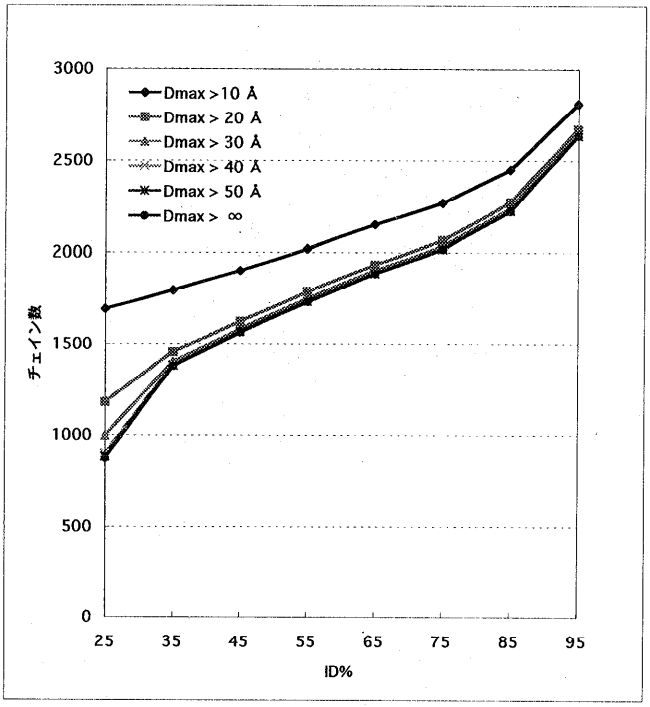


図7 類似基準 (ID% と Dmax) と代表タンパク質チェイン数の関係
Fig. 7 Relationship between the number of representative protein chains and the several threshold of the sequential similarity (ID%) and the structural similarity (Dmax)

>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．

上記の基準で決定した代表タンパク質チェインの数を表5と図7に示す．図7の横軸は配列相同性のしきい値 (ID%)，縦軸は代表タンパク質チェイン数で，構造の類似性のしきい値 (Dmax) の基準ごとにその数をプロットし折れ線で示している．

図7を見てまず気がつくことは，Dmax の基準が10Å の時の代表チェイン数と，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% のしきい値 95% の場合（右端）でも，Dmax が ∞ の代表チェイン数と比較すると，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5)．このことから，ID% ≥ 95% の配列相同性があっても，配

表2 PDB リリース #84 の内容（分子タイプによる分類）
Table 2 Number of atomic coordinate entries on PDB Release #84 (Molecule Type)

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 PDB リリース #84 の内容（解析法による分類）
Table 3 Number of atomic coordinate entries on PDB Release #84 (Experimental Technique)

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

表4 タンパク質チェインリスト
Table 4 Number of protein chains classified according to the priority

	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

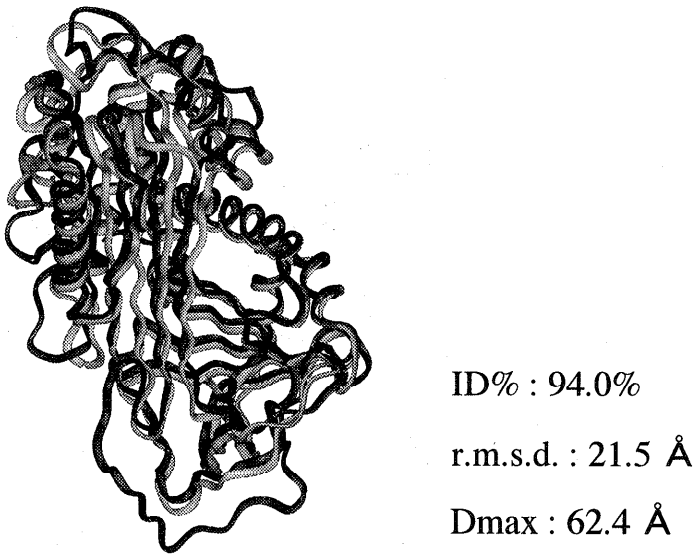


図8 抗トロンビン (PDB エントリー名:2ANT) のLチェイン（薄いリボン）とIチェイン (濃いリボン) の重ね合わせ図
Fig. 8 View of superposition of L (white ribon) chian and I (black ribon) chain of Antithrombin (PDB ID:2ant)

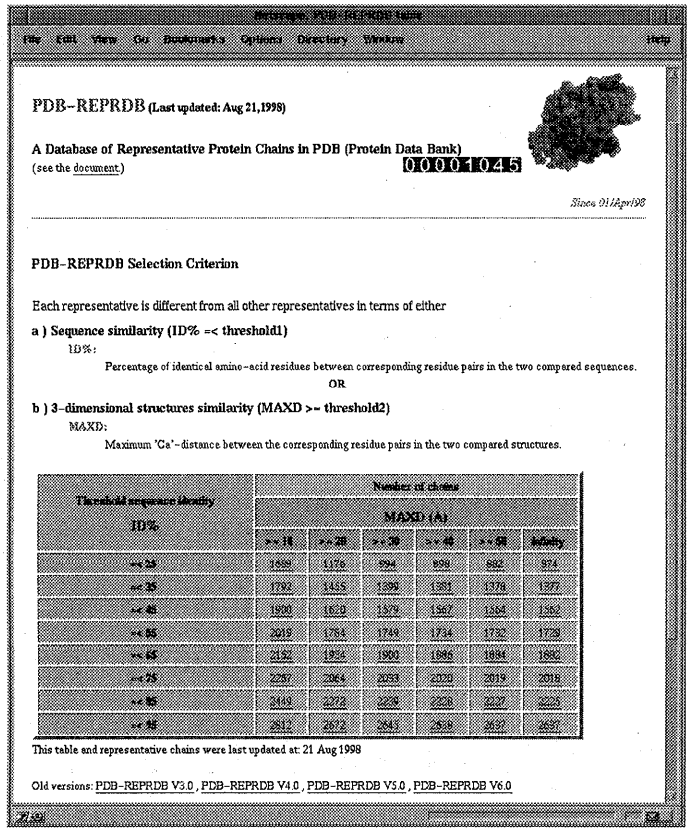
列の置換や挿入・欠損によって，Dmax が 10Å をこえる部分構造の変化があることがわかる．また，ID% のしきい値 25% で，同様に Dmax が ∞ の時の代表チェイン数と比較すると，ほぼ半数の 815 (=1,689-874) 本のチェインを構造の差異によって別のチェインとして分類している (表 5)．この結果を見ると，タンパク質の部分構造の解析を行なう場合，ID% が 25% の配列相同性だけを考慮した代表タンパク質チェインを用いたのでは，多くの有用な構造データを使わずに解析していたことになる．

図 7 を見て次に気づくのは，ID% のしきい値 25% と 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 5 の Dmax が 20 Å と ∞ の列を比較すると，ID% が 35% の行では，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% が 25% の行では，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% のしきい値が 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．

表5 決定した代表タンパク質チェインの数
Table 5 Number of representative protein chains at the several threshold of the sequential similarity (ID%) and the structural similarity (Dmax)

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

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 8 の例では，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．図 8 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 62.4 Å と非常に大きな値になっている．r.m.s.d 値も 21.5 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，あまり突出した値にはなっていない．このことは，部分構造の違いを検出する指標として，Dmax 値



PDB-REPRDB (Last updated: Aug 21,1998)

A Database of Representative Protein Chains in PDB (Protein Data Bank)
(see the document)

00001045

Since 01Apr98

PDB-REPRDB Selection Criterion

Each representative is different from all other representatives in terms of either

a) Sequence similarity (ID% =< threshold1)

ID%:

Percentage of identical amino-acid residues between corresponding residue pairs in the two compared sequences.

OR

b) 3-dimensional structures similarity (MAXD >= threshold2)

MAXD:

Maximum 'Ca'-distance between the corresponding residue pairs in the two compared structures.

Threshold sequence identity ID%	Number of chains					
	MAXD (A)					
	=< 10	=< 20	=< 30	=< 40	=< 50	Infinity
=< 25	1589	1176	994	898	882	874
=< 30	1791	1455	1309	1301	1278	1277
=< 40	1920	1659	1579	1567	1564	1562
=< 50	2015	1784	1748	1734	1732	1729
=< 65	2182	1924	1900	1886	1884	1882
=< 75	2257	2064	2031	2020	2019	2018
=< 85	2349	2272	2228	2220	2217	2215
=< 95	2412	2372	2343	2326	2297	2287

This table and representative chains were last updated at: 21 Aug 1998

Old versions: PDB-REPRDB V3.0 , PDB-REPRDB V4.0 , PDB-REPRDB V5.0 , PDB-REPRDB V6.0

図 9 WWW 上の PDB-REPRDB
Fig. 9 PDB-REPRDB WWW page

の方が r.m.s.d 値より適していることを示している．

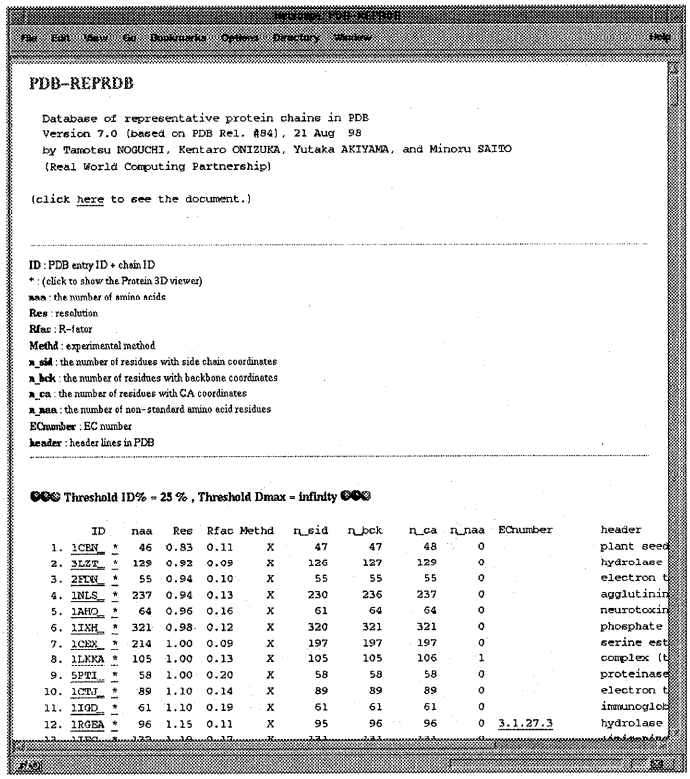
上記のように，タンパク質の立体構造は，ID% のしきい値が高くても，部分構造が異なるタンパク質が数多く存在する．したがって，タンパク質の立体構造は，配列相同性だけで単純に分類できるものではなく，厳密に分類するためには，構造の類似性を考慮することが重要である．

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[6] として，WWWで公開している．この WWW ページは，我々の研究室で公開している PAPIA システム[11] とゲノムネットの WWW サーバー[12] とリンクされており，既に世界から 2,200 回以上アクセスされている．

ホームページ (図 **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を知りたい場合，その基準のマス目の数字をクリックすると，図 **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．現在は，様々な研究用途に対応するため，配列の相同性：ID% $\geq$ 25% ~ 95% まで 10% 刻みの 8 通りと，構造の類似性：Dmax $\leq$10 Å ~ 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 $8 \times 6 = 48$ 通りの代表タンパク質のリストを作成している．

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 (エントリー名とチェイン ID)，および残基数，分解能，R ファクター，実験方法，主鎖原子の座標がそろっている残基数，側鎖原子の座標がそろっている残基数，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．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．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そのタンパク質分子の



PDB-REPRDB

Database of representative protein chains in PDB
Version 7.0 (based on PDB Rel. #84), 21 Aug 98
by Tamotsu NOGUCHI, Kentaro ONIZUKA, Yutaka AKIYAMA, and Minoru SAITO
(Real World Computing Partnership)

(click here to see the document.)

ID : PDB entry ID + chain ID
* : (click to show the Protein 3D viewer)
naa : the number of amino acids
Res : resolution
Rfac : R-fator
Methd : experimental method
n_sid : the number of residues with side chain coordinates
n_bck : the number of residues with backbone coordinates
n_ca : the number of residues with CA coordinates
n_naa : the number of non-standerd amino acid residues
ECnumber : EC number
header : header lines in PDB

Threshold ID% = 25 % , Threshold Dmax = infinity

	ID		naa	Res	Rfac	Methd	n_sid	n_bck	n_ca	n_naa	ECnumber	header
1.	1CBN_	*	46	0.83	0.11	X	47	47	48	0		plant seed
2.	3LZT_	*	129	0.92	0.09	X	126	127	129	0		hydrolase
3.	2FDN_	*	55	0.94	0.10	X	55	55	55	0		electron t
4.	1NLS_	*	237	0.94	0.13	X	230	236	237	0		agglutinin
5.	1AHO_	*	64	0.96	0.16	X	61	64	64	0		neurotoxin
6.	1IXH_	*	321	0.98	0.12	X	320	321	321	0		phosphate
7.	1CEX_	*	214	1.00	0.09	X	197	197	197	0		serine est
8.	1LMKA	*	105	1.00	0.13	X	105	105	106	1		complex (t
9.	5PTI_	*	58	1.00	0.20	X	58	58	58	0		proteinase
10.	1CTJ_	*	89	1.10	0.14	X	89	89	89	0		electron t
11.	1IGD_	*	61	1.10	0.19	X	61	61	61	0		immunoglob
12.	1RGEA	*	96	1.15	0.11	X	95	96	96	0	3.1.27.3	hydrolase

図10 WWW上のPDB代表タンパク質チェインリスト（例）
Fig. 10 An example of the chain list on PDB-REPRDB

立体構造がグラフィック表示される．

8. まとめ

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提案し，その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作成した．

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．

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6,127本を約1.5時間で分類することができた．

従来は，代表チェインの選出は手仕事で行なわれており，様々に条件を変えて代表リストを作成したり，研究者の注文に合わせて即時に計算することなどは全く不可能であった．本研究での並列化をはじめとする高速化，自動化により，現在ではそれが可能になった．

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2,200回以上アクセスされている．

また，我々の研究室で公開しているPAPIAシステム[11]のPDBデータや，我々の研究室で行なっているタンパク質立体構造予測の研究で利用するPDBデータは，このPDB-REPRDBの代表リストを利用して決めており，PDB-REPRDBは我々の研究室において重要な位置を占めている．

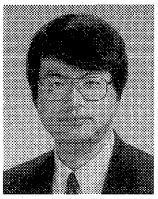
今後は，様々なタンパク質立体構造解析の研究者の要求にきめこまかく対応できるように，順位リストの全チェイン間の配列相同性（ID%）や構造相同性（Dmax）の計算結果をテーブルとしてあらかじめ用意しておき，オンデマンドで様々な基準（良質の基準や各配列および構造の相同性のしきい値など）での代表タンパク質チェインを決定し，即時に提供できるようなシステムを構築していく予定である．

謝辞 本研究を始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 實教授に深謝致します．

参考文献

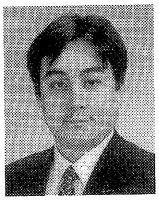
1) Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M.: The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures, *J. Mol. Biol.*, Vol. 112, pp. 535-542 (1977). http://www.pdb.bnl.gov/
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*, Vol. 1, pp. 409-417 (1992).
3) Hobohm, U. and Sander, C.: Enlarged representative set of protein structures, *Protein Science*, Vol. 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.*, Vol. 26, pp. 316-319 (1998). http://www2.ebi.ac.uk/dali/fssp/fssp.html
6) 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/pdbreprdb/
7) 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). http://scop.mrc-lmb.cam.ac.uk/scop/
8) 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
9) Needleman, S. B. and Wunsch, C. D.: A general method applicable to the search for similarities in the amino acid sequences of two proteins, *J. Mol. Biol.*, Vol. 48, pp. 443-453 (1970).
10) 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).
11) 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 (GIW'98)*, Universal Academy Press (1998), (to appear). http://www.rwcp.or.jp/papia/
12) "GenomeNet WWW Server" http://www.genome.ad.jp/dbget/

(平成10年9月10日受付)
(平成10年10月29日再受付)
(平成10年11月12日採録)



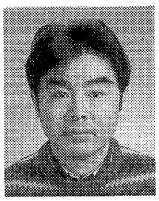
野口 保（正会員）

昭和32年生．昭和58年東京農工大学大学院工学研究科応用物理学専攻修士課程修了．昭和59年富士通（株）入社．同年富士ファコム制御（株）に出向．コンピュータケミストリ分野のシステム開発に従事．平成元年蛋白工学研究所（PERI）に出向．タンパク質立体構造予測の研究に従事．平成3年より富士通（株）にてバイオ分野のシステム開発に従事．平成6年から平成7年にかけて九州工業大学情報工学部受託研究員．平成8年技術研究組合新情報処理開発機構主任研究員．並列応用つくば研究室にて，並列計算機を用いたタンパク質立体構造予測の研究に従事．日本物理学会会員．



秋山 泰（正会員）

昭和36年生．平成2年慶應義塾大学大学院理工学研究科電気工学専攻博士課程修了．工学博士．同年通産省電子技術総合研究所研究官．平成4年京都大学化学研究所助教授．平成8年技術研究組合新情報処理開発機構並列応用つくば研究室室長．現在に至る．並列計算機を用いたタンパク質立体構造および遺伝子配列情報解析等の研究に従事．電子情報通信学会，日本生物物理学会，分子生物学会，神経回路学会，IEEE各会員．



鬼塚健太郎

昭和38年生．平成2年東京都立大学理学研究科物理学専攻修士課程修了．同年松下電器産業（株）入社．（財）新世代コンピュータ技術開発機構（ICOT）に出向，同機構の開発する並列論理型コンピュータの遺伝子情報処理分野への応用に関する研究に従事．平成6年より松下技研（株）において音声認識の研究，平成7年より松下電子工業（株）において画像処理プロセッサの開発に従事，平成8年技術研究組合新情報処理開発機構に出向，並列応用つくば研究室にて並列計算機上でのタンパク質立体構造解析に関する研究に従事．日本生物物理学会会員．

安藤　誠（正会員）

昭和42年生．平成4年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．同年日本鋼管（株）入社．データベース（関係型，オブジェクト指向など）のデータモデルに関する研究に従事．平成6年同社より米国コンベックスコンピュータ社（現ヒューレットパッカード社）に派遣．並列計算機 Exemplar シリーズの OS 開発グループに所属．平成8年技術研究組合新情報処理開発機構に出向．並列応用つくば研究室にて，並列計算機上でのタンパク質立体構造解析等の研究に従事．