蛋白質立体構造の進化的解析のための Ninf 版並列 MGG とその性能評価

小野功[1)]　今出広明[1)]　中田秀基[2,3)]　小野典彦[1)]　松岡聡[3,4)]　関口智嗣[2)]　楯真一[5)]

核磁気共鳴法(NMR)は，ポストシーケンスにおける最重要課題の一つである蛋白質立体構造解析の有望な手段である．しかし，専門家でさえ，一つの蛋白質のデータ解析に数ヶ月程度の試行錯誤が必要なことが深刻な問題となっている．これに対し，小野らは遺伝的アルゴリズム(GA)に基づくデータ解析の自動化手法を提案し，小規模な問題において比較的良好な性能を得られたと報告している．本報告では，本手法を高速化するために，産総研が提案しているミドルウェア Ninf を用いて，GA の世代交代モデル Minimal Generation Gap の並列化実装を行い，評価実験によりその動作検証および性能評価を行う．

A Parallel Minimal Generation Gap Model Using Ninf for Evolutionary Analysis of Protein Structures and Its Performance Evaluation

Isao Ono[1)], Hiroaki Imade[1)], Hidemoto Nakada[2,3)], Norihiko Ono[1)], Satoshi Matsuoka[3,4)], Satoshi Sekiguchi[2)] and Shin-ichi Tate[5)]

Nuclear Magnetic Resonance (NMR) spectroscopy is a promising method for the three-dimensional structure determination of proteins that is one of the most important problems in post-sequence era. This method has a serious problem that it takes several months for experts to analyze the data of only one protein. In order to remedy the problem, Ono et al. have proposed an automatic method based on a genetic algorithm (GA) for analyzing the data and determining the three-dimensional structures of proteins and reported that they had good results on relatively small-size problems. In this report, to speed up the GA, we propose an parallel implementation of the generation alternation model, Minimal Generation Gap (MGG), which is employed in the GA. In the implementation, we employ Ninf proposed by National Institute of Advanced Industrial Science and Technology (AIST) as a middleware. In order to examine the performance, we perform some experiments.

１．はじめに

ポストシーケンスにおいて蛋白質立体構造解析は最重要課題の一つであり，核磁気共鳴法(Nuclear Magnetic Resonance; NMR)は有望な構造解析技術の一つである．NMR による構造解析の過程は，1) 連鎖帰属，2) NOE (Nuclear Overhauser Effect)帰属および立体構造決定，の２つのフェーズに大別されるが，後者が律速となる．現状の NOE 帰属および立体構造決定のフェーズでは，高度な専門知識と豊富な経験に基づき，試行錯誤的に NOE シグナルを ^{1}H ペアに帰属し，徐々に立体構造を構築していく作業が行われているため，経験豊富な専門家でも１つの蛋白質の立体構造決定に数ヶ月を要する．この人的および時間的コストの問題を解決するために，自動化および高速化技術の開発が強く望まれている[3)]．

小野らは，観測された NOE シグナルを満たす立体構造を遺伝的アルゴリズム(Genetic Algorithm; GA)により探索することにより，NOE 帰属および立体構造決定の自動化を行う手法を提案している[1)]．本手法を蛋白質における典型的な部分構造である α-helix に適用したところ，専門家の領域知識によらず，専門家とほぼ同じ立体構造を求めることに成功した．しかし，現状の実装では，アミノ酸残基数 13 の α-helix でも半日程度，アミノ酸残基数 27 の α-helix であると数日程度の計算時間がかかる．そのため，これ以上大きな構造に適用することは現実的に困難であり，高速化が緊急の課題であるといえる．

GA は，複数の解候補のサンプリングを膨大な回数繰り返しながら探索を進めていく．ここで，各解候補の評価値計算は全く独立に行えるため，本質的に並列化が可能である．小野らの手法[1)]において，計算時間の大半を占めているのは，探索中に新しくサンプリングされた立体構造の評価計算の部分であることから，この部分を並列化することにより，大幅な高速化が期待できる．そこで，本稿では，小野らの手法において採用されている世代交代モデル Minimal Generation Gap (MGG)[5)]のマスター・ワ

1) 徳島大学　The University of Tokushima
2) 産業技術総合研究所 National Institute of Advanced Industrial Science and Technology (AIST)
3) 東京工業大学　Tokyo Institute of Technology
4) 国立情報学研究所　National Institute of Information
5) 生物分子研究所　Biomolecular Engineering Research Institute

ーカー・モデルによる並列化実装を提案する．本実装は，長時間にわたる計算処理を考慮して，1) 一部のワーカーが計算途中で落ちた場合にその分の処理速度は低下するが全体の計算は止まらない，2) マスターが計算途中で落ちた場合にも定期的に保存された内部状態を用いて再スタートできる，3) 計算途中で，全体の計算を中断することなく，ワーカーの追加・削除ができる，などの特徴をもつ．また，グリッド・アプリケーションへの拡張を考慮して，ミドルウェアとして GridRPC を実装した Ninf [2),7)]を用いている．本稿では，15 残基α-helix 構造決定問題への適用を通じて，提案した MGG の並列化実装の動作検証および簡単な性能評価実験を行う．

２．問題設定と解法の構成

２．１　NMR 蛋白質立体構造決定問題

蛋白質は複数のアミノ酸がペプチド結合により鎖状に結合したものである．図１に示すように，通常，結合周りで回転することにより複雑に折りたたまれエネルギー的に安定な立体構造をとっている．結合周りの回転各を二面角と呼び，二面角ω, ϕ, ψが定義されている鎖を主鎖，二面角χが定義されている鎖を側鎖と呼ぶ．

ある立体構造をもつ蛋白質を溶液に溶かし，NMR の装置にかけると，蛋白質中に多数含まれる水素原子核(^{1}H)に由来する化学シフトと呼ばれるシグナルデータを得ることが出来る．化学シフトは，各 ^{1}H を取り巻く電子密度に依存することから各 ^{1}H に固有の値である．いくつかの条件で測定されたシグナルデータ(HNCA や HCACO など)を用いることにより，観測された化学シフトを元の ^{1}H に帰属していくことが可能である．この作業は，連鎖帰属とよばれ，現状で 1〜2 週間程度の作業である．

連鎖帰属に用いたものとは別のある特別な測定条件において，NOE (Nuclear Overhauzer Effect) シグナルと呼ばれる化学シフト．ペアを数多く測定することができる．この NOE シグナルは，お互いの距離が十分に近い(5Å程度) 2 個の ^{1}H の化学シフトが観測されたものであり，これらの NOE シグナルを対応する ^{1}H ペアに帰属することにより，蛋白質中

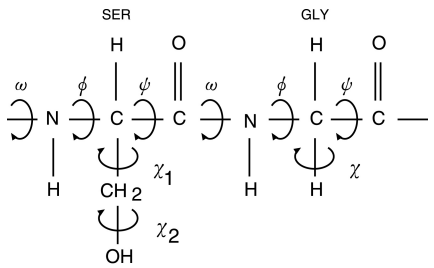

図 1　蛋白質と二面角

の複数の ^{1}H ペア間の距離制約を得ることができる．この作業は，NOE 帰属とよばれ，連鎖帰属で帰属されている各 ^{1}H の化学シフトを手がかりに，NOE シグナルを対応する ^{1}H ペアに帰属していくことが可能である．NOE 帰属により得られた膨大な数の距離制約を用いることにより，蛋白質の立体構造を決定できる．しかし，実際には，測定装置の分解能や誤差の問題から NOE シグナルを対応する ^{1}H ペアに一意に割り付けることは困難であり，専門家でもかなりの試行錯誤を要する上，数ヶ月程度の時間がかかることから，自動化および高速化が望まれている．

小野らは，上述の NOE 帰属の過程を，観測された NOE シグナルをなるべくよく説明するような立体構造を求める構造最適化問題として定式化することを提案している [1)]．すなわち，NOE 帰属は，以下の評価関数を最小化するように，蛋白質立体構造のシステムパラメータである二面角を最適化する問題として設定される：

（評価値）＝$-w_1$・(観測 NOE と予測 NOE の一致度)
＋ w_2・(原子の重なり具合)

ここで，予測 NOE は評価対象の立体構造において原子間距離が 8Å以下の ^{1}H ペアの化学シフト・ペアを列挙したものであり，観測 NOE を全てカバーしているとき，第 1 項は最小値$-w_1$・(観測 NOE の数) をとる．また，第 2 項目は，評価対象の立体構造において原子同士の衝突がないとき，最小値 0 をとる．

２．２　遺伝的アルゴリズムに基づく解法 [1)]

２．２．１　遺伝的アルゴリズム (GA)

遺伝的アルゴリズム(Genetic Algorithm; GA)は，自然界の生物の進化過程を模倣した最適化の枠組みである．GA は，目的関数の微分情報を必要としない直接探索法であることから，実数値変数最適化に限らず，組合せ最適化，構造最適化も取り扱うことができる．また，複数の解候補が探索空間内において競争的に探索を行う確率的多点探索法であることから，膨大な数の局所解が存在する多峰性の探索空間においても効率よく大域的に良好な解を探索可能である．このような特徴から，GA はさまざまな応用分野において有望な最適化手法として注目を集めている．GA の一般的なアルゴリズムは以下のとおりである：

(1) 初期集団の生成
　ランダムに複数の解候補（個体）を生成し，それらを初期集団とする．
(2) 複製選択
　集団から，新しい解候補（子個体）を生成させる

個体のペア（両親）を複数組生成する.

(3)子の生成

ステップ 2 で生成されたそれぞれの個体のペアに交叉および突然変異オペレータを複数回生成し，複数個の子個体を生成する.

(4)生存選択

ステップ 3 で生成された子集団と元の集団の中から，各個体の評価値を参照しながら，次世代へ残す個体を選択し，残りの個体は淘汰する.

(5) 終了条件がみたされるまで，ステップ 2 から 4 を繰り返す.

GA の設計項目は，個体の計算機上での表現方法と子の生成方法を設計するコード化/交叉・突然変異設計，複製/生存選択方法を設計する世代交代モデル設計に大別される．探索効率の観点から，コード化/交叉/・突然変異設計においては形質遺伝が重要であり，得られる解の質の観点から，世代交代モデル設計においては集団内の形質の多様性維持が重要となる [4].

２．２．２　解法の構成 [1]

コード化としては，アミノ酸残基の二面角 ω, ϕ, ψ, χ 角を要素とする実数値ベクトルを採用し，交叉としては，一様交叉(Uniform Crossover; UX)を採用している．UX は，図 2 に示すように， 2 つの親個体が与えられたとき，50%の確率で染色体の各変数を入れ換える． 突然変異オペレータは，各変数について 1%の確率で，[-1.0°, +1.0°]の範囲の一様乱数を加えるものとする.

世代交代モデルとして，多様性維持能力に優れ，多峰性の探索空間において優れた性能を示す Minimal Generation Gap (MGG)[5]に基づくモデルを採用している．図 3 にその概要を図示する．この世代交代モデルでは，集団の多様性維持の観点から，両親は集団からランダムに選ばれ，世代交代は局所的に行われる．両親の存在する領域のランドスケープを正しく見積もるため，交叉を複数回適用して 1 ペアの親から複数個の子を生成する．また，局所的な選択圧を高めるために，生存選択において両親と子個体群を合わせた家族の中から最良 2 個体を選択している．アルゴリズムの詳細を以下に示す：

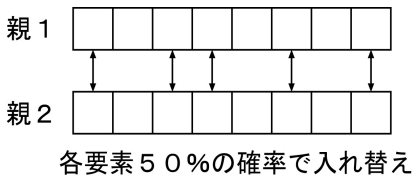

図 2　一様交叉

(1) 初期集団の生成

ランダムに複数個の個体を生成，評価値を計算し，それらを初期集団とする.

(2) 複製選択

集団からランダムに交叉のための 2 つの親を選択する.

(3) 子の生成

ステップ 2 で選択された両親に対し，交叉 UX を n_c 回適用し，子を $2n_c$ 個生成する．また，生成された子個体に対し，突然変異を適用する.

(4) 子個体の評価

ステップ 3 で生成された全ての子個体の評価値を計算する.

(4) 生存選択

両親と生成された全ての子を合わせた個体集合から最良 2 個体を選択し，集団中の両親と置き換える.

(5) 停止条件が満たされるまで，ステップ 2 からステップ 4 を繰り返す.

３．世代交代モデル MGG の並列化

３．１　並列化実装の要件

利用者の観点およびグリッド環境への拡張の観点から，MGG モデルの並列化実装は以下の要件を満たしていることが望ましいと考えられる.

(1) 拡張性

投入した計算資源に比例して，計算時間が短縮されることが望ましい.

(2) 頑健性

計算ノードやネットワークに障害が起こった場合でも，全体の計算は動作しつづけることが望ましい．長時間の計算時間を要することから，最悪の場合でも計算途中から再開できるようになっているべきである.

(3) 柔軟性

計算途中で，全体の計算を止めることなく，計算ノードの追加および削除が出来るようになっていることが望ましい．これは，グリッド環境など，

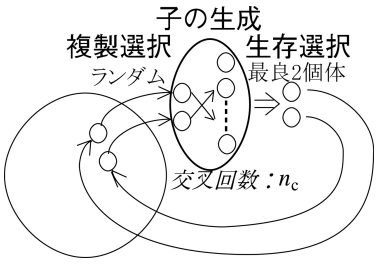

図 3　Minimal Generation Gap (MGG)

他のユーザーと計算資源を共有している場合に便利であると考えられる.

(4) 移植性
多くのプラットフォーム上で動作することが望ましい.

(5) 価格性
専用の PC クラスタを購入しなくても，現有の計算機環境で実行できるように,性能の異なる計算ノードが含まれていても,性能を発揮できることが望ましい. 特に，グリッド環境では，このようなヘテロな環境への対応は重要であると考えられる.

３．２　MGG の並列化実装

３．２．１　ミドルウェアと開発言語

本並列化実装では，2.2.2 節において示した MGG のアルゴリズムにおける個体の評価値計算処理を複数のワーカー・ノードで実行し，その他の処理をマスター・ノードで実行するマスター・ワーカー・モデルを採用する.

ミドルウェアとしては,グリッド環境への拡張を考慮して,Ninf Ver.1[2),7)]を用いる.現在,Ninf Ver.1 は，SunOS 4.X，5.X，Cray J90/C90，Alpha OSF1，FreeBSD，Linux，IRIX，SP-2 (AIX)，SR2201 といったさまざまなプラットフォームで動作することから,移植性の要件をよく満たしていると考えられる.

開発言語としては,個体の評価値計算処理モジュールが C++を用いて開発されていたことからワーカー側の開発に C++を採用し，マスター側の開発に JAVA を採用した. マスター側に JAVA を採用した理由は，移植性に優れているほか，強力なスレッド機構，メモリ管理機構および例外処理機構を利用できることから頑健性を満たす実装を容易に構築できると考えられるためである.

３．２．２　ワーカーの設計

NMR 蛋白質立体構造決定のための評価器[1)]では，個体が持っている二面角情報から蛋白質の立体構造を構築し，評価値を計算するためには，蛋白質中に含まれる全ての種類のアミノ酸の立体構造情報，全ての種類の原子半径，全ての原子核の化学シフト情報，全ての NOE シグナル情報が必要となる. これらのデータサイズはかなり大きいが，全ての個体評価において使いまわせる情報である. そこで，ワーカーでの処理を，初期化，個体評価，終了化の３つのメソッドに分割し，初期化処理メソッド呼び出し時にアミノ酸の立体構造情報などの共通に用いるデータを全て評価器に読み込ませるようにしている. また，個体評価処理メソッドにおいては，各ワーカーの処理能力に応じて計算量を変更できるように，マスター側で指定された数の個体を一度に評価するようになっている. Ninf Ver.1 は基本的には RPC であるが，メモリ上のデータを共有する異なる複数のメソッドの呼び出しができるように工夫を施してある.

３．２．３　マスターの設計

マスターの概念図を図 4 に示す. マスターにおいては，図 4 に示すように，メイン・スレッド，世代交代スレッド，クライアント・スレッドの 3 種類のスレッドが協調しながら動作している.

メインスレッドは，他のスレッドの初期化，世代交代スレッドとクライアント・スレッドの通信用キューなどデータ構造の初期化，初期集団の生成などの処理を行う. また，ユーザーからのワーカーの追加・削除の要求を処理する.

世代交代スレッドは，複製選択，子の生成，生存選択を行うスレッドである. 世代交代スレッドの主な処理の流れを以下に示す：

(1) 集団から親個体を選択する.
(2) 選択された親個体に交叉および突然変異を適用して複数個の子個体を生成する.
(3) 評価値計算のため，生成した子個体をクライアント・スレッドとの通信用のキューに登録する.
(4) 全ての子個体の評価値計算が正常に終了するか，ユーザーが定めた時間を経過するまで待機する.
(5) 評価値計算が正常に終了した子個体と両親の中から最良 2 個体を選択して，集団中の両親と入れ替え，世代数カウンタを増す.
(6) 打ち切り世代数に達した場合，世代交代スレッドに打ち切り世代数に達したことを通知して終了する. 打ち切り世代数に達していない場合はステップ１へ.

キューを通じてクライアント・スレッドに個体を絶え間なく供給するために，図 4 に示すように，複数の世代交代スレッドが動作している. そのため，集団中の親個体は複数の世代交代スレッドに選択されないようにロック機構を有している. また，世代交代スレッドは全ての生成個体の評価が終了するまで新たにキューに個体を供給できないため，効率性の観点から，同時に複数のスレッドがステップ１～３を実行できないようになっている.

クライアント・スレッドは，ワーカーに一対一に対応して動作するスレッドである. クライアント・スレッドの主な処理の流れを以下に示す：

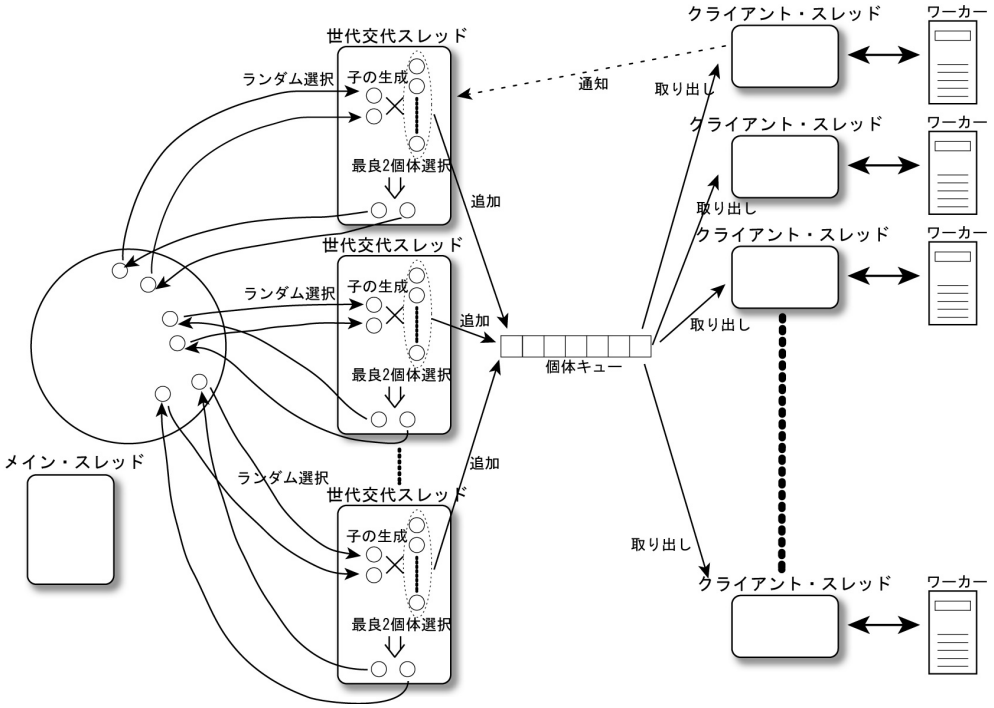

図 4　MGG の並列実装

(1) ユーザーが指定した数の個体をキューから取り出す. 指定された数の個体がキューに供給されていない場合は, ある一定時間, 個体の供給を待つ.
(2) ステップ 1 において, キューから取り出せた個体の数が 0 であり, かつ打ち切り世代数に達している場合, 終了する.
(3) Ninf GridRPC API を用いて, キューから取り出された個体群の評価値計算をワーカー側で行う.
(4) 世代交代スレッドに個体の評価値計算が終了したことを通知する.
(5) ステップ 1 へ.

ステップ 3 が異常終了した場合, メインスレッドにその旨を通知して終了することにより, 対応するワーカーに異常が起きても, 全体の計算は止まらないようになっている.

3．3　提案実装モデルの特徴

提案実装モデルの特徴を以下に述べる：

(1) 拡張性への対応
複数の世代交代スレッドを同時に動作させ, 評価値計算の必要な個体をキューに絶え間なく供給しつづけることにより, ワーカーの遊び時間を極力抑えるような工夫をしている.

(2) 頑健性への対応
ワーカーに異常が生じた場合, 対応するクライアント・スレッドのみを終了させることにより, 全体の計算は止まらないような工夫をしている. また, 定期的に集団の状態をディスクに保存することにより, マスターに異常が生じても, 途中から計算を再開できる.

(3) 柔軟性への対応
クライアント・スレッドを削除／追加することにより, 計算途中でワーカーを切り離したり, 追加したりすることができる. これにより, 異常が生じて落ちてしまったワーカーを再び追加することも可能となっている.

(4) 移植性への対応
ワーカーは 3.2.1 節で述べた幅広いプラットフォーム上で動作し, マスターは JAVA ランタイム環境の動作するプラットフォームで動作する.

(5) 価格性への対応
ワーカーの処理能力に応じて, ワーカーへ投げる評価個体数を変更できるようになっている. この変更は計算途中でも可能であり, ユーザーが様子を見ながらチューニングできるようになっている. ワーカーへ投げる評価個体数の自動チューニングについては, 現在, 研究が進行中であり, 別の機会に発表する予定である.

4．実験

4．1　実験設定

実験を行った計算機構成は，以下のとおりである：

- マスター・ノード：Athlon MP 1.2GHz × 2
- ワーカー・ノード：Pentium III 800MHz × 10
- ネットワーク：100BaseTX スイッチングハブ

対象問題は，15 残基α-helix 構造決定問題である．GA のシステムパラメータである集団サイズは 500，交叉回数は 100 と設定した．また，提案した MGG の並列実装のシステムパラメータである世代交代スレッドの数は 3，キューの長さは 1,000，各ワーカーへ投げる個体数は 1 と設定した．

４．２　頑健性，柔軟性の検証

頑健性に関する工夫の動作検証を行うため以下の実験を行い，動作を確認した：

- 計算中に，ワーカーのプロセスを強制終了する．
- 計算中に，ワーカー・ノードを再起動する．
- 計算中に，ワーカー・ノードのネットワークケーブルを抜く．
- 計算中に，マスター・ノードを再起動する．

また，計算中にワーカー・ノードの削除と追加を行い，柔軟性に関する工夫の動作を確認した．

４．３　拡張性の検証

ワーカーの数を１から１０まで変化させて，50 世代分の計算が終了するまでの時間を 5 試行ずつ計測した．5 試行の平均の結果を図 5 に示す．また，性能の向上率を図 6 に示す．これより，ワーカー数 10 まではほぼリニアに性能が向上していることがわかる．

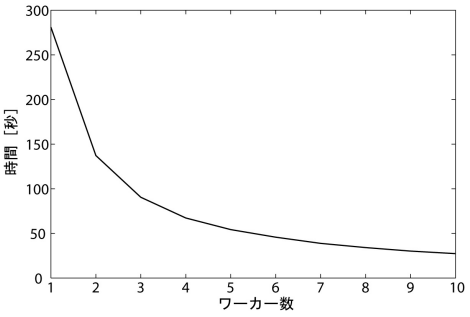

図 5　ワーカー数と計算時間の関係

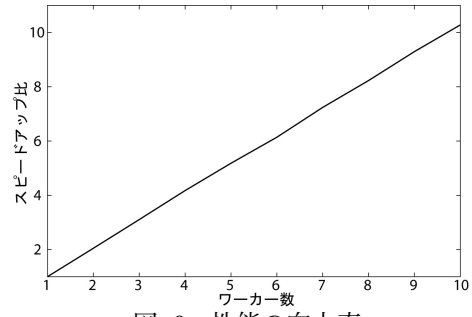

図 6　性能の向上率

５．おわりに

本稿では，GA に基づく NMR 蛋白質立体構造決定法を高速化するために，Ninf GridRPC を用いた世代交代モデル MGG の並列化実装を提案し，15 残基α-helix 構造決定問題への適用を通じてその有効性を検証した．

今後は，さらに多くのワーカー・ノードを用いての拡張性の検証，性能の異なるワーカー群を用いての実験およびその上での評価個体数の自動チューニング，Ninf-G[2),6)]を用いてのグリッド環境における実験などを行っていきたいと考えている．

謝辞

本研究を進めるにあたってご協力いただいた平成 13 年度修士課程修了生の菅正和氏，現在，学部 4 年生の井上和豊君，修士 1 年生の中島直敏君に感謝の意を表します．

参考文献

1) Ono, I. Fujiki, H., Ootsuka, M., Nakashima, N., Ono, N. and Tate, S.: Global Optimization of Protein 3-Dimensional Structures in NMR by A Genetic Algorithm, Proc. 2002 Congress on Evolutionary Computation, pp.303-308 (2002).
2) http://ninf.apgrid.org/
3) 伊藤隆 : 効率的で迅速な NMR タンパク質構造解析法の模索，実験医学，Vol.19，No.8，pp.954-957 (2001).
4) 喜多一，山村雅幸 : 機能分担仮説に基づく GA の設計指針，計測と制御，Vol.38,No.10,pp.612-617 (1999).
5) 佐藤浩，小野功，小林重信 : 遺伝的アルゴリズムにおける世代交代モデルの提案と評価，人工知能学会誌，Vol. 12，No.5，pp.734-744 (1997).
6) 田中良夫，中田秀基，平野基孝，佐藤三久，関口智嗣 : Globus による Grid RPC システムの実装と評価，情報処理学会ハイパフォーマンスコンピューティング研究会，Vol.2001，No.77，pp.165-170 (2001).
7) 中田秀基，高木浩光，松岡 聡，長嶋雲兵，佐藤 三久，関口 智嗣 : Ninf による広域分散並列計算，情報処理学会論文誌 Vol.39，No.6，pp. 1818-1826 (1998).