Grid 環境におけるタンパク質のエネルギー最小化による立体構造予測

青井 桂子

1　はじめに

自然に存在するタンパク質の立体構造は系の自由エネルギーの最小状態に対応している．このため，アミノ酸配列情報からコンピュータシミュレーションにより最小化問題としてタンパク質分子の立体構造が予測可能であると考えられる．タンパク質のエネルギー関数は局所的に無数の，大域的にも複数の極小値を持つ．このため，我々は SA と GA のハイブリッドアルゴリズムである遺伝的交叉を用いた並列シミュレーテッドアニーリング (PSA/GAc) による立体構造予測を行っている．

一方で，ネットワーク上につながれた広い地域に配置された計算資源を結びつけて，広域的に分散／並列処理を行う Grid と呼ばれる新しい計算モデルが研究されるようになった．

本研究では，ネットワークを介して遠隔地にある計算資源を利用するための Grid ミドルウェアである NetSolve を用いて Grid 環境を構築し，Grid 環境に適した PSA/GAc のモデルを実装する．

2　これまでの研究

Grid 環境におけるタンパク質のエネルギー最小化による立体構造予測の研究において，これまでに検討したことは以下の 4 点である．

1. NetSolve 上に PSA/GAc の計算モデルを構築
2. NetSolve Farming 機能の適用
3. 本システムの導入による資源の供給量の測定
4. オーバーヘッド，Server での計算時間の測定
5. 3D 表示のアプリケーションを作成

3　NetSolve の PSA/GAc 実装モデル

NetSolve は，Tennessee 大学の Jack Dongarra らによって開発された Grid RPC System である．NetSolve システムはネットワーク上にあるハードウェアとソフトウェアの両方の計算資源にリモートアクセスを可能にする．GA と SA のハイブリッドである PSA/GAc は，複数の逐次 SA を並列に実行し，一定間隔で遺伝的交叉を行う．遺伝的交叉の処理では，もとの親と生成した子との 4 個体のうち評価値の高い 2 個体を選択して，選択された 2 個体から次の探索を行う．

タンパク質のエネルギー計算は，その 1 つ 1 つの計算に時間がかかるのではなく，エネルギー計算回数が膨大になり，計算時間がかかることがわかっている．このため，Server 側に逐次の SA を実行させ，交叉周期になると Client 側に個体を返すモデルを考案し，構築した．

4　NetSolve Farming 機能の適用

NetSolve Farming 機能を用いて，Grid 環境における PSA/GAc マスタースレーブモデルを作成した．これまで NetSolve の API として Client 側で用いてきた関数では，一つの実行要求を行った場合に Server での実行が終了して Client に値を返すまで次の実行要求を行うことができなかった．このため，並列処理を Server で実行させる場合，Fig. 1 のように Client 側で複数のプロセスを立ち上げて各々のプロセスが NetSolve の実行要求を行う必要があった．NetSolveFarming 機能は類似処理を一括して実行できる．このため，Fig. 2 のように，Client 側で複数のプロセスを立ち上がらせることなく，複数の実行要求を行うことができる．

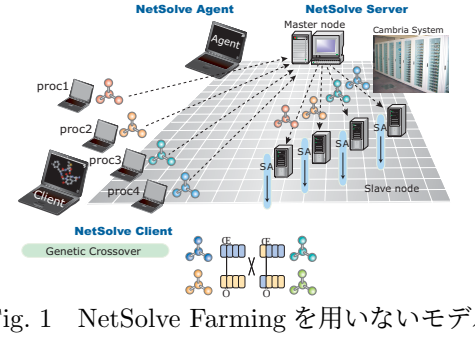

Fig. 1　NetSolve Farming を用いないモデル

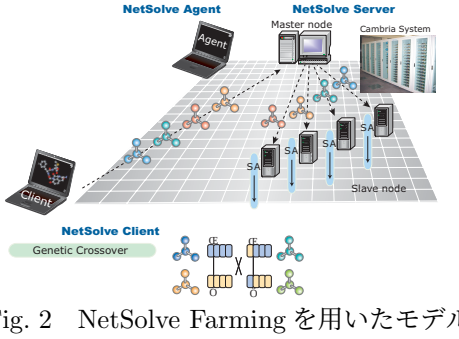

Fig. 2　NetSolve Farming を用いたモデル

5　システムにおける単位時間あたりの計算量

個体数を増やした場合の，Grid 環境の PSA/GAc と通常の PSA/GAc の単位時間あたりの評価計算回数の比較を行う．

Grid 環境の PSA/GAc では，Grid RPC システムの一つである NetSolve を用いる．また，ユーザは 1 台の PC しか持たないものとし，NetSolve Server としては，CambriaSystem の 100 ノードを利用可能な資源とする．

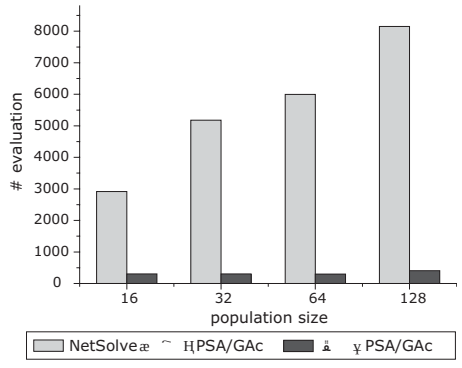

Fig. 3　単位時間 (1hour) あたりの評価計算回数の比較

Fig. 3 に単位時間あたりの評価計算回数の比較を示す．実験結果より，Grid 環境の PSA/GAc では個体数が増えると逐次 SA の並列数が増えるため，外部 Server で利用する計算資源が増える．ユーザーが PC 1 台しか持たない場合でも，16 個体で 10 倍近く，128 個体の時には 20 倍もの計算資源を手に入れられた．

6　システムにおける時間の測定

実装したシステムでは，Server での計算時間の他に通信時間，NetSolve システムの NetSolve の Farming 機能を用いており，Farming のさいにはシステムの待機時間が生じる．

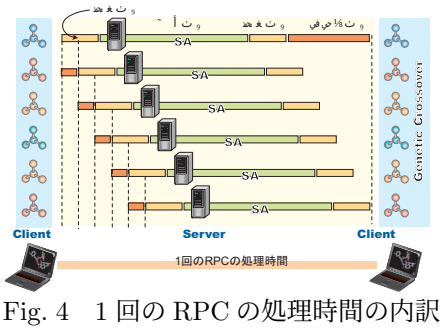

Fig. 4　1 回の RPC の処理時間の内訳

対象タンパク質は 19 個の設計変数 (二面角) を持つ Met-enkephalin と，30 個の二面角を持つ $(Ala)_{10}$ である．

本実験では，各対象問題に対して，Server での SA のステップ数を 32MCsweep, 64MCsweep, 128MCsweep, 256MCsweep にしたパターンのもので実験を行う．また，PSA/GAc における個体数 (並列数) を 2, 4, 8, 16, 32, 64, 128 個体として実験を行った．各 MCsweep 数，個体数で比較を行う．実験では Cambria Cluster を用い，NetSolve Server として 224 ノードを割り当てた．

6.1　Server での計算時間の測定

Fig. 5 と Fig. 6 に Met-enkephalin と $(Ala)_{10}$ における各 MCsweep の時の Server での計算時間の比較を示す．Fig. 5，Fig. 6 より MCsweep 数の増加に応じて計算時間も長くなることが示された．

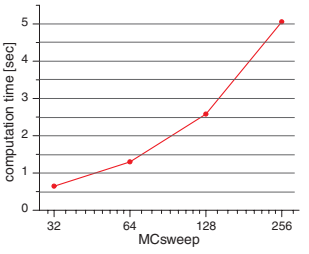

Fig. 5　Met-enkephalin 計算時間

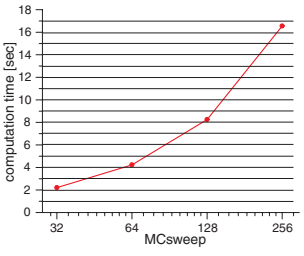

Fig. 6　$(Ala)_{10}$ 計算時間

6.2　システムの累積待機時間の測定

"isdl_netsl_farm" を用いた場合の各個体数に要する累積待機時間 (待機時間) を Fig. 7 と Fig. 8 に示す．1 回の RPC に要する RPC は約 0.5sec であり，個体数が増えるほど累積待機時間は増加する．対象問題が大規模で，Server での計算時間が待機時間に比べて極端に長い時間を要するものでなければ，非同期にすることなどを考える必要がある．

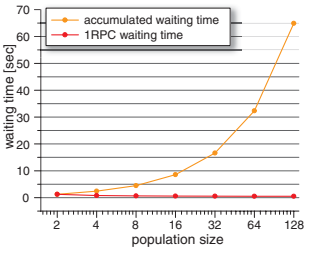

Fig. 7　Met-enkephalin 待機時間

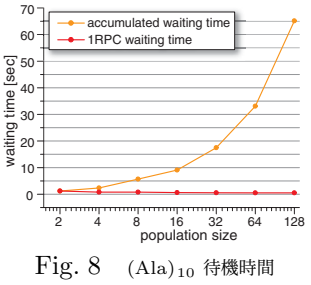

Fig. 8　$(Ala)_{10}$ 待機時間

7　3D 表示のアプリケーションを作成

SC2002 で行うデモ用の，タンパク質立体構造予測の探索過程を 3D 表示するアプリケーションを作成した．

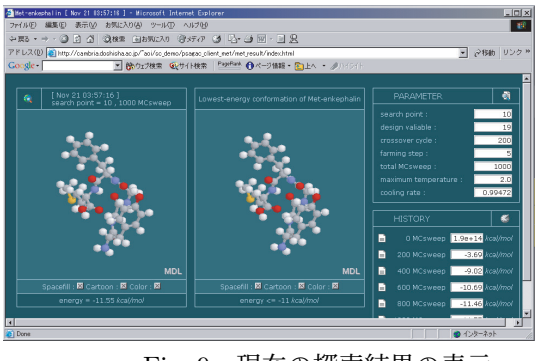

Fig. 9　現在の探索結果の表示

Fig. 9 では，現在計算中の個体における最優良個体の構造と，これまで経験的に最適解とされる構造を並べて表示している．最優良個体は，タンパク質の立体構造予測においては，得られたエネルギーが最も低かった個体を示している．Fig. 9 の右側の表にはこのとき実験で用いたパラメータと探索過程で得られた履歴 (MCsweep とエネルギー値) が表示される．

8　今後の研究課題

NetSolve を用いた PSA/GAc の非同期モデルの実装