タンパク質立体構造決定におけるグリッドコンピューティングによる分子動力学計算の精度向上

高島 浩幸*, 井本　祐司*, 三村 典夫*

インターネット技術の進歩により、グリッドコンピューティングは実用化の段階に入りつつある。その大きな特徴である分散性とハイパフォーマンスを活用するために、我々は、タンパク質の立体構造決定計算に適用した。そして、従来の計算では解釈が曖昧だった初期構造依存性の問題が高分散処理によって解決できること、さらに、それによって構造計算の精度が飛躍的に向上することを示す。これは、従来の科学技術計算へのグリッドコンピューティングの有用性を示す実例となる。

High Resolution Protein Structure Determinations by Molecular Modeling Calculations and GRID Computing Implementations

Hiroyuki Takashima*, Yuji Imoto* and Norio Mimura*

This is an implementation of a GRID computing. We utilized the distributed high performance computing for protein structure determinations by NMR and molecular modeling calculations. We indicate that an initial structure dependency problem in conformational sampling can be solved by the distributed computing. The results demonstrate the efficiency of the GRID computing with striking improvements of structural resolutions for proteins.

1 はじめに

生体高分子の分子動力学計算は、動的構造のシミュレーションや立体構造決定などに広く用いられている。タンパク質の構造のシミュレーションには、国際的構造データベースである Protein Data Bank (PDB, url: http://www.rcsb.org/pdb/)に登録されている結晶構造や Nuclear Magnetic Resonance (NMR) による溶液構造が計算の初期構造として用いられることが多い。ここで問題になるのが初期構造依存性である。計算の出発点となる構造の選び方でシミュレーションの結果に影響が出る可能性が極めて高い。分子動力学計算では、この問題を解決するために Simulated Annealing (SA)法など様々な手法が取り入れられてきた。しかし、問題は、タンパク質のような生体高分子の立体構造計算における莫大な conformational space（構造自由度の大きさ）に起因するため、簡単には解決できないでいる。これは、NMR などの実験データをもとにした一次構造（アミノ酸配列）からの立体構造決定計算ではさらに深刻な問題となる。そのため、NMR による立体構造決定で

*ノバルティスファーマ筑波研究所
(Novartis Institutes for BioMedical Research)

は、できるだけ多数の初期構造を用いて計算を行う必要がある。どれだけの初期構造を用い SA 法などの分子動力学計算をどれだけ長い時間行えば初期構造依存性を除去することができるのかは、実は、検証されていない。従来は計算機の能力の限界から、100 個程度の初期構造が使用されてきた。そして、それが技術的限界として黙認されてきた。

我々は、この問題を検証するために、グリッドコンピューティングを使用し、十分に多数の初期構造からの計算を試みた。NMR による立体構造決定では、各初期構造の計算は互いに独立であるため、分散処理に適している。我々は、グリッドコンピューティングによって、単に処理速度を上げるだけではなく、従来見過ごされてきた問題が解決できること、そして、それによってタンパク質の立体構造の精度を劇的に向上させることができることを示す。

2　NMR による立体構造決定法

まず、NMR 測定から得られた実験データをもとにタンパク質の立体構造を決定する方法についてごく簡単に概説する。タンパク質の NMR 測定によって、水素原子間の空間的な距離を大まかに推測することができる。例えば、100 残基のタンパク質であれば、800 個程度の水素原子を持ち、計算上 800×800 個程度の距離情報を得ることができる。ただし、この距離情報は 5Å以上離れた原子間では観測できない。また、NMR シグナルのオーバーラップやアミノ酸側鎖の運動性の問題などにより、情報の数は大幅に減少する。それでも、1000 個から 2000 個程度の情報が得られる。数は多いものの網羅性に乏しく、かつ、測定誤差をもった情報の集合である。それを一様に満足（目的関数の最小化）させる構造を conformational space 内で検索する。この計算は乱数によって作成した初期構造から開始するのが一般的であり、一様に構造の不確かさを含むので、複数の計算結果を重ね合わせた構造のアンサンブルが最終産物となる（例えば、距離情報を満足させた度合いの高い構造 20 個を選び出したもの, 図４参照）。得られたアンサンブル内で均一な構造を持っていれば構造の精度が高いことになる。逆に分散していれば、なんらかの原因で構造が決まらないということになるが、その原因は、分子の運動性に由来するものと実験・計算誤差がある。問題は、運動性と計算誤差を区別することは困難な点にある。

* 生理活性物質の構造を模し、受容体に拮抗的に結合することで活性の抑制を行う薬剤分子

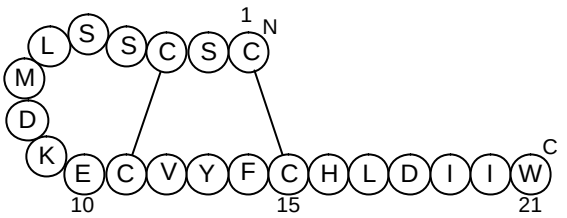

図１エンドセリン-1 の一次構造
アミノ酸配列を1文字コードで示す。直線は、SS 結合の位置。数字は、残基番号。N 末端および C 末端をそれぞれ N とC の添え字で示す。

3　エンドセリン-1 の立体構造解析の問題点

構造決定計算に使用する初期構造の数が最終産物の構造精度におよぼす影響を検証するために、我々は、比較的分子量の小さな 21 残基の生理活性ペプチドであるエンドセリン-1（ET-1）の立体構造計算を行った (1)。

ET-1（図１）は、ヒト血管内皮細胞から分泌される血管収縮作用を持つペプチドで、高血圧、心臓疾患、ガンなどへ関与すると考えられている。その生理活性は C 末端 21 番のトリプトファン残基（Trp21）に強く依存するため、ET-1 受容体の拮抗阻害剤* の設計ではトリプトファンの側鎖を模した構造が中心となっている。しかし、過去に発表されている ET-1 の結晶構造（PDB ID: 1EDN）と

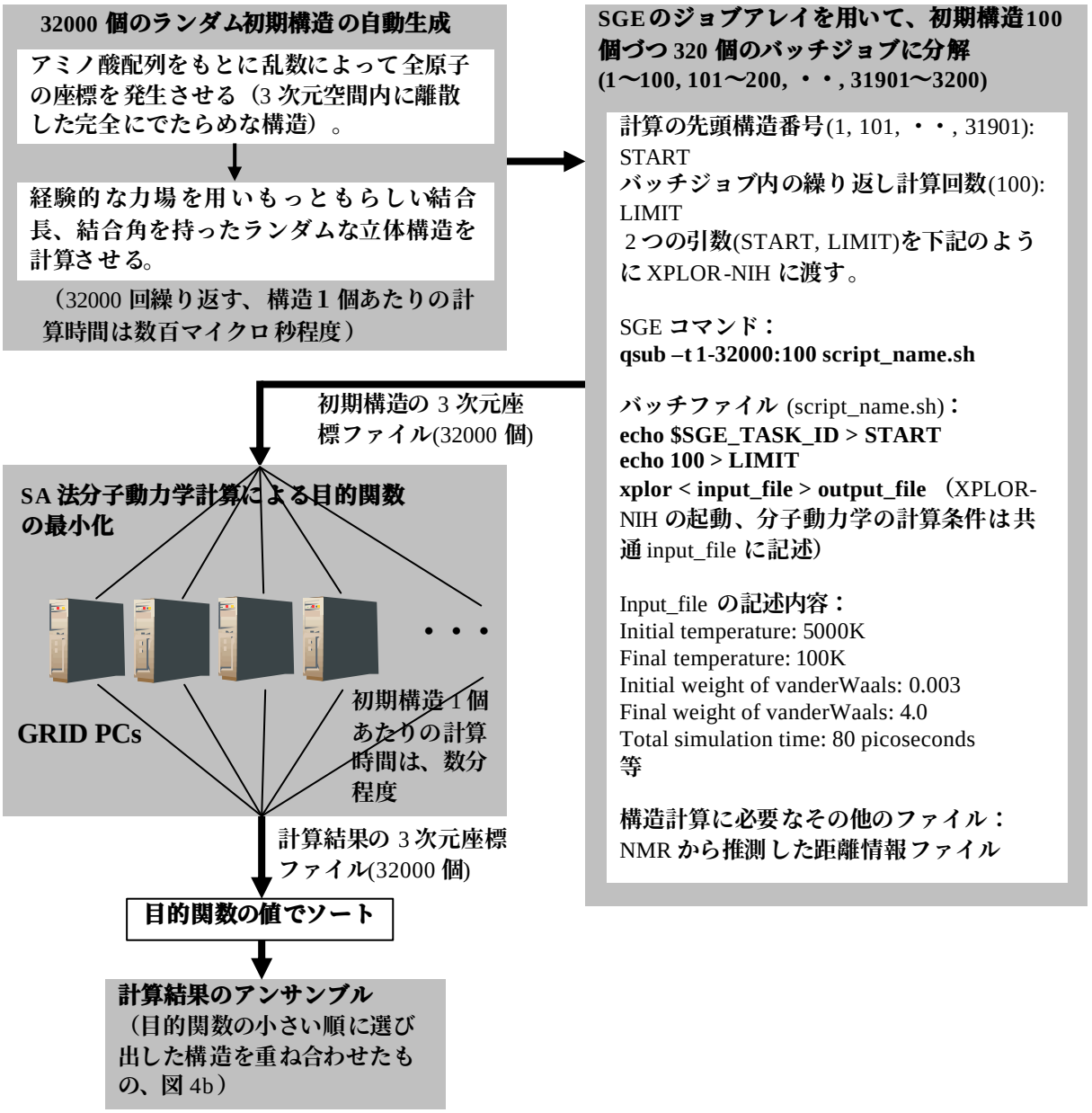

図 2：グリッドコンピューティングを使用した構造計算の手順

分子動力学の計算は、Simulated Annealing (SA)法のアルゴリズムを使用している。これは、高温から低温に擬似的な焼きなましを行って目的関数の最小化を図るものである。NMR から得られた距離情報はハーモニックポテンシャルとして力場（共有結合の経験値等）に加え、SA 法の計算を行っている（restrained molecular dynamics calculations）。3 次元座標ファイルの大きさはアミノ酸残基数×11K バイト程度（ET-1 の場合、230K バイト）。NMR から推測した原子間距離情報は不確かさを含むので、距離の上限と下限の二つで与える（数 K バイト程度のテキストデータ）。計算のインプットもアウトプットも小さなファイルのみであり、かつ、初期構造ごとの計算が全て独立であるためグリッドコンピューティングに適しており、今回、コンピュータノードの数に比例した処理速度が得られることを確認した。ランダムな初期構造の数は、sampling scale issue として古くから知られた問題であるが計算パワーの制限からこれまで検証されることはなく、計算結果の構造精度との関連はこの十数年の間ほとんど議論されなかった。PC 単体の高速化とグリッドコンピューティングによるスケーラブルなパフォーマンスアップにより初めて検証が可能になった。

NMR 溶液構造では、このトリプトファン周辺の C 末端の立体構造に大きな食い違いがある。従来の研究において NMR 溶液構造では、C 末端は分散して構造が決まっていなかった。そのため、大きな自由度つまり構造のゆらぎを持つと解釈されてきた。一方、結晶構造では C 末端の構造は決まっているため、従来、この構造が薬剤開発におけるスタンダードと考えられてきた。ところが、過去に開発された拮抗阻害剤の分子構造は結果的にどれもこの結晶構造とは合致せず、結晶構造が水溶液中の活性構造を再現しているという仮定に疑問の声が上がっていた。そのため、溶液中における ET-1 の構造を C 末端を含めて決定する必要がある。

4　グリッドコンピューティングの構造計算への実装

我々は、グリッドコンピューティングを用いて、初期構造の数を 100 個から 32000 個まで増やした（図２）。これは、従来法の 300 倍の計算量となる。

計算には、17 台の PC を SUN GRID Engine version 5.3p2 (SGE)でコントロールして行った。PC は、HP 社製 Evo 500, Intel PentiumⅣ 1.6 GHz 1CPU で、OS は、RedHat Linux 8.0を使用した。これらの PC を NIS クライアントとして、SGE administrator アカウントと、計算を実行する一般ユーザアカウントを定義した。一般ユーザアカウントのホームディレクトリを NFS サーバー上に置くことで、全コンピュータノードで計算データとジョブファイルを共有した。コンピュータノード PC、SGE master PC、NIS サーバー、NFS サーバーは、コンピュータルームに設置し、オフィスの SGE ジョブコントローラー PC との間を汎用 Virtual Private Network (VPN)ルーターを用いて構築した簡便な VPN ネットワーク（L2TP トンネルを使用）で接続した。これによって、既存の社内 LAN 内で、高いセキュリティと可用性を両立させることができた。

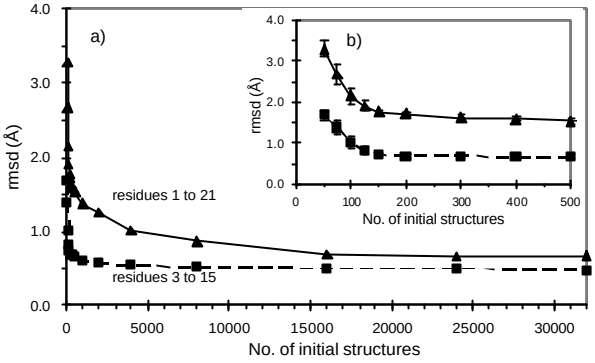

図３:初期構造の数（横軸）に対する構造収束（縦軸）のプロット
a) 構造の収束を、root mean square deviations (rmsd) で示す。構造計算の結果の中から実験値をより満足させた 20 個の構造を選び出し、そのアンサンブルの中での構造類似度を数値化したものであり、値が小さいほど構造収束が良い。実線は、N 末端から C 末端まで全体で rmsdを計算したもの。破線は、N 末端側の 13 残基のみで rmsdを計算したもの。b) 初期構造の数 500 個までを拡大表示したもの。

分子動力学の計算は、XPLOR-NIH v2.0.6(2)を使用した。このプログラムは従来から結晶構造解析、NMR による構造決定計算に用いられてきた一般的なもので、クラスターやグリッドへのソースレベルでの対応は行われていない。我々は、バッチジョブを SGE に投入するスクリプトと SGE のジョブ Array を用いて 32000 個の初期構造ごとに計算を分散処理させた（図２）。

5　エンドセリン-1 の C 末端側立体構造の決定

グリッドコンピューティングを用いて、初期構造の数を増やし、初期構造の数に対して構造の収束をプロットしたところ（図３）、従来の予想を覆すような結果を得た。従来、100 個程度で十分計算できると考えられていた ET-1 の構造が、その 100 倍以上の 16000 個の初期構造を用いることでようやく収束したのだ（図４）。そして、問題になっていた

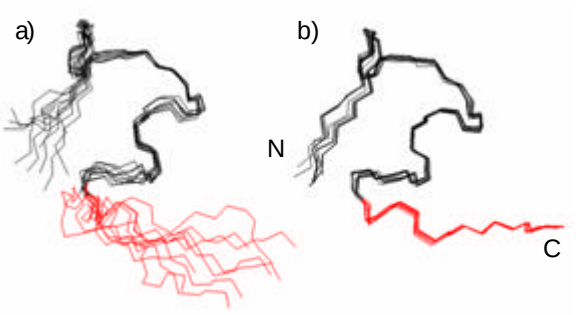

図４:エンドセリン-1 の NMR 溶液構造アンサンブルの主鎖の重ね合わせ
a) 従来法と同様、100 個の初期構造を用いた結果。b) グリッドコンピューティングを用い、32000 個の初期構造から計算させた結果。C 末端側 16 番から 21 番までの残基を赤で示す。

C 末端側の構造も決めることができた（PDB ID: 1V6R）。それは、13 番のチロシン（Tyr13）の側鎖を中心とした疎水性のコアである（図５）。

ET-1 の N 末端側は、2 本の SS 結合（図１）で安定化されたβストランドとαヘリックスを持つ比較的強固な構造を持っている（図５）。これは、蛇、サソリ、ヒトなど様々な生物種で見つかっている特徴的な構造モチーフ(3)で、ET-1 でも同じ構造が決定されていた（図４a 参照）。図３b の構造収束を見ると、初期構造 100 個程度の計算ですでにこの N 末端側 13 残基は構造が決まっていることがわかる。これは、2 本の SS 結合により構造の自由度が部分的に減少しているためである。従来の初期構造の少ない計算では、この部分的収束と全体の収束を区別する方法は無く、見誤りを犯していたことになる。つまり、従来考えられてきた NMR 溶液構造における ET-1 の C 末端側のゆらぎは計算誤差だったことがわかった。

今回行った計算は、条件設定、検算を含めて、3 週間ほどを要した。単一の CPU で計算させた場合、1 年以上を要する計算となり、事実上不可能だった。（ちなみに、10 年前のスーパーコンピュータでは、初期構造 100 個程度の計算で大学の研究室の CPU 時間の割り当てを使い果たしてしまうほどだった。）

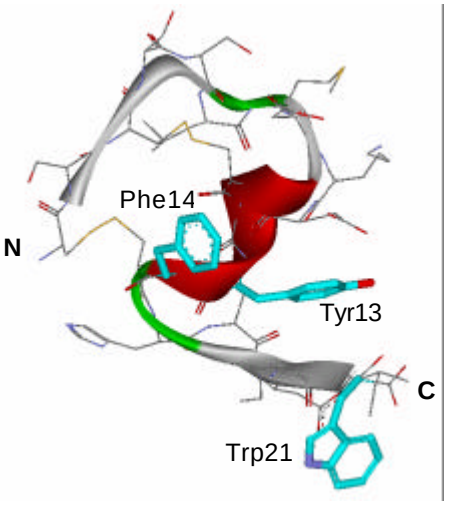

図５:エンドセリン-1 の NMR 溶液構造
図 3bのアンサンブル中の最小エネルギー構造を示す。主鎖はリボンモデル。赤はα ヘリックス、緑は、β ターン。Tyr13, Phe14, Trp21 のアミノ酸側鎖をスティックモデルで示す。

一方、グリッドコンピューティングを使えば、CPU 数を 10 倍に増やすことは簡単にできる。今回の計算を 10 倍の CPU で行えば数日で結果が出ることになり、さらに 10 倍の CPU で行えば数時間で構造決定が完了することになる。現在、製薬企業では常時数千台から数万台の PC が LAN 上に接続されているため、数千 CPU 程度のグリッドコンピューティングは構築可能である。筆者らの所属している Novartis では、スイスを中心に 2002 年からネットワーク内のグリッドコンピューティング導入に取り組んでおり、現在は、United Devices の GRID Engine を用いて数万 CPU 規模の環境を構築中である。

6　Tyr13 を中心とした疎水性コアの実証

今回得られた Tyr13 の側鎖を中心とした C 末端側残基の疎水性コア（図５）は、これまで開発されてきた ET-1 受容体の拮抗阻害剤の構造とも良い一致を示す（トリプトファン、フェニルアラニン、チロシンおよび電荷を持

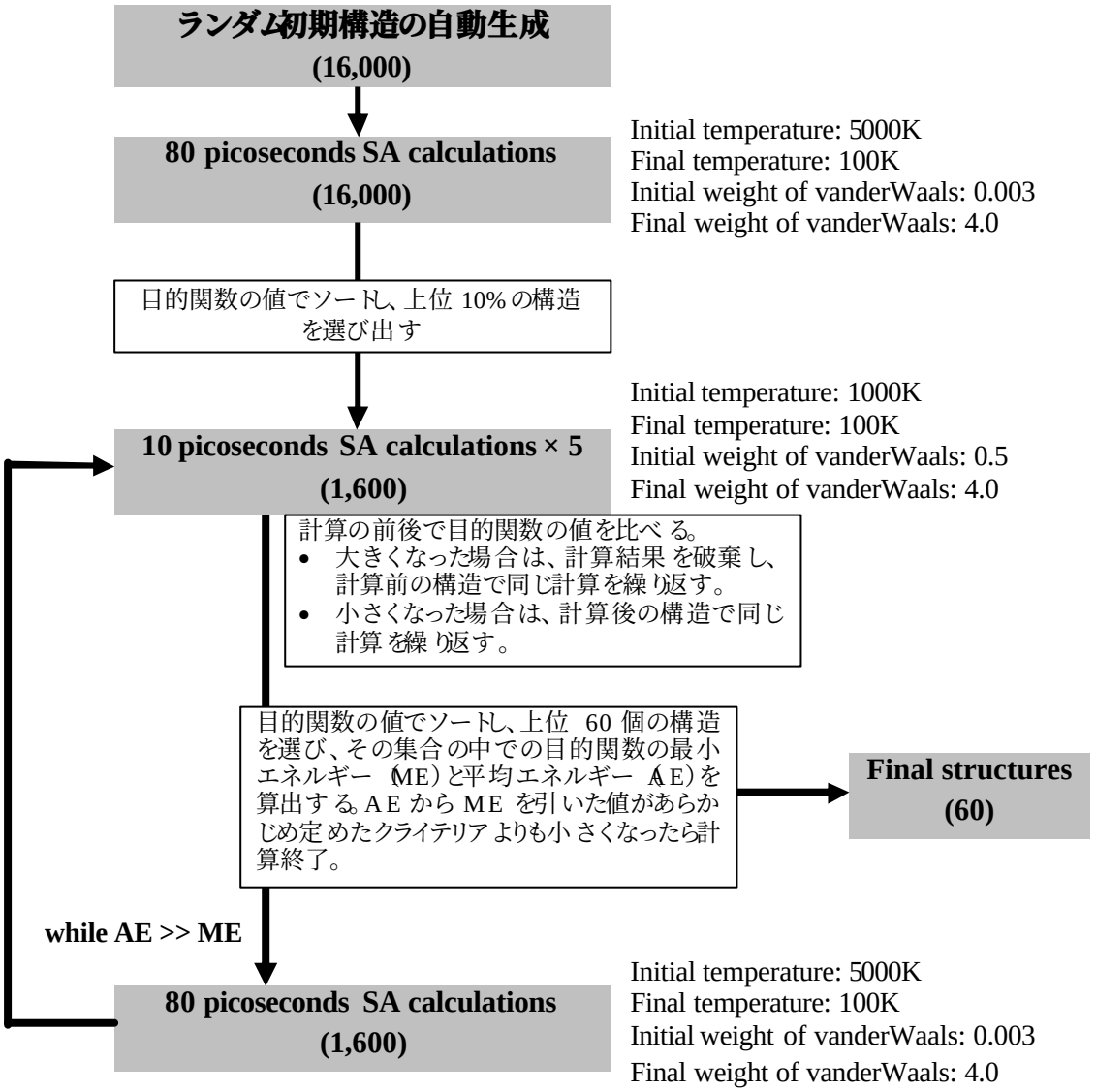

図 6：ネオカルチノスタチンの構造決定に用いた分散化計算手順

ランダム初期構造の自動生成とそれに続く SA 計算までは、図 2の手順と同じ。全ての SA 計算は、NMR から得られた距離情報で制限をかけて実行している。括弧内の数字はそれぞれのステップで計算する構造の総数を示す。SA 法の計算条件を右横に示す。計算時間は、80 ピコ秒のシミュレーションで数十分程度（Kolo-NCS の場合）。タンパク質の計算において、ランダムな初期構造から出発すると、間違ったフォールディングによってエネルギーの最小化ができなくなる場合がある。間違ったフォールディングを取り除くため、最初の SA 計算の後、目的関数の値の大きな 90% の構造を捨てた。その後、初期温度 1000K の SA による比較的細かい構造検索と初期温度 5000K の SA による粗く広い構造検索を交互に繰り返した。初期温度 1000K の SA では、vanderWaals **半径の重みづけの初期値を大きく設定し、より狭い範囲の構造検索を効率的に行うようにした**。繰り返し計算の終了を判定するために新たに導入した AE-ME のクライテリアには、構造計算の経験上もっともらしい値（アミノ酸残基数×0.2 Kcal/mol）を用いた。繰り返し計算は、1600 個の構造で実行されるので、それぞれのステップで構造 20 個づつのバッチジョブに分けて SGE に投入した。今回は、ET-1 の場合と同じグリッドコンピューティング環境（17CPU）を用いて、全体の計算で 3 ヶ月ほどを要した。なお、繰り返し計算中のソートのステップに統計解析を組合わせれば、個々の構造が計算途中であっても繰り返し終了の判定の予測が可能になる。

つ側鎖の位置関係を重ね合わせた場合）。そのため、今後の拮抗阻害剤設計に有用と考えられる。

ET-1 の C 末端側の構造を検証するために、我々は、他の測定手法である Photochemically Induced Dynamic Nuclear Polarization NMR と Matrix-assisted laser desorption ionization Time-of-Flight Mass Spectrometry を用いた実験を行い、Tyr13 の側鎖が分子内疎水性コアの中に入り込んでいることを確認した(4)。そして、ET-1 の結晶構造では、Tyr13 とグルタミン酸側鎖が分子間で水素結合を持っているために NMR 溶液構造と異なった構造を持っていることがわかった(4)。

7　より大きなタンパク質の構造決定へのグリッドコンピューティングの適用

Conformational space は、残基数の大きさに依存して急激に増大するため、より大きなタンパク質ではより大量の計算が必要になる。しかし、一般的なタンパク質で数万個以上の初期構造を用いると、現状ではディスクスペースの面でも問題が発生する。一方、初期構造の数を増やすことと、シミュレーション時間を長くすることは Conformational space 内の構造検索で同等の効果を持つと期待される。それを確認するため、ET-1 の計算で 1600 個の初期構造（10 分の 1）から出発し、800 ピコ秒（10 倍）の SA 計算を実行した。シミュレーション時間を長くすればそれに比例して計算に必要な実時間も増えるので、総計算時間は同じである。そして、図４b と同じ構造精度を得た。つまり、シミュレーション時間を長くすることと初期構造の数を増やすことでは、最終構造の精度に等価の効果を持つことが確認できた。

グリッドコンピューティングの適用性をタンパク質で検証するため、我々は、114 残基の抗腫瘍性タンパク質ホロ体ネオカルチノスタチン（holo-NCS）の構造決定計算を行った(5)。ET-1 の結果（図３）から、80 ピコ秒の SA 計算 16000 個では sampling scale が小さすぎることが容易に予測されるため、シミュレーション時間を長くする必要がある。しかし、どれだけ長くすれば良いかは計算をかけてみないとわからない（現状でも膨大な CPU 時間が必要になる）。さらに、単発の長い計算をかけるだけでは、目的関数の最小化を判断するのは不可能に近いという問題もある。SA 計算の初期温度も構造検索の速度に影響を与えるが、より高い温度ではより粗い検索しかできない。また、グリッドコンピューティングは一続きの長い計算には不向きである。これらの問題を克服するため、計算を断片化し、さらに、全体計算の終了を一意的に判定できるように、異なった初期温度での短い SA 計算の繰り返しによる図６の計算手順を考案した。

一方、シミュレーションの長い計算を実行する目的で Rocks (http://rocks.npaci.edu/Rocks) を用いた PC クラスターも検討した。現在、クラスターに対応している分子動力学計算のソフトウエアには Amber (http://amber.scripps.edu), GROMACS (http://www.gromacs.org) などが知られており、NMR の距離情報と組合わせることで構造決定計算も可能である。しかし、現状では、計算のオーバヘッドが大きすぎるため、16 台の PC を使用して 8 倍程度の計算パフォーマンスの向上にとどまり、図６の分散化手法によるパフォーマンスをずっと下回ってしまった。したがって、今回、PC クラスターは使用しなかった。

8　ネオカルチノスタチンの構造解析

Holo-NCS は、βシートに富む球状タンパク質で、バルク DNA の選択的な切断活性を持つ NCS クロモフォアを疎水性の結合部位の中に包み込んでいる。NCS クロモフォアは、2 本のアセチレン結合とエポキシ結合を持つ非常に反応性の高い分子でありながら、タンパク質との複合体で極めて高い安定性を持っ

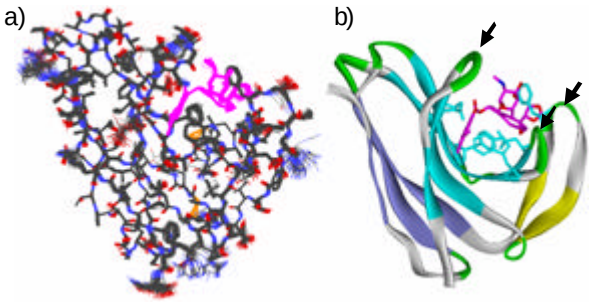

図 7：ネオカルチノスタチンの NMR 溶液構造
a) グリッドコンピューティング用い計算させた結果の中から 60 個を重ね合わせたものを示す。構造収束を示す rmsd の値は、0.32Å。図には、重原子のみを表示している。NCS クロモフォアは、マジェンダで示す。b) a の重ね合わせの中の最小エネルギー構造を主鎖リボンモデルで示す。NCS クロモフォアとその結合に関与するアミノ酸側鎖をスティックモデルで示す。β シート β ターンの部分構造を図 6 と同じ色分けのリボンで示す。

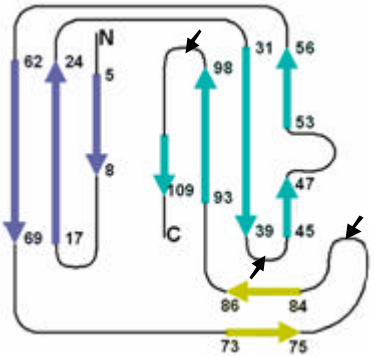

図 8：ネオカルチノスタチンのβシー 構造
色つきの矢印でβ シー トの位置を示す。数字は、残基番号を示す。黒い矢印は、図 5と同じ NCS クロモフォアの結合に関与するループ構造を示す。

ている。つまり、分子版トロイの木馬のようなもので、表面上無害のように見えて、標的に達した時に強力な刺客を放出するのだ。この強い結合と選択的な解離を両立させる分子機構が興味深く、ドラッグデリバリーシステムへの応用も検討されている。

過去に発表されている holo-NCS の NMR 溶液構造では、NCS クロモフォアを風呂敷の紐のように包み込んでいるループ部分の構造の精度が非常に悪かった。そのため、ループ部分の運動性と holo-NCS の活性との関連が示唆されてきた。しかし、ループ部分が他のβシートに比べて有意に高い運動性あるいは構造のゆらぎを持っているという実験的証拠は無い。我々は、この構造精度の悪さも ET-1 の C 末端同様計算誤差に由来すると考えて、グリッドコンピューティングによる精度向上を図った。

計算結果を図 7 に示す。グリッドコンピューティングと図 6 の計算手順の適用により、ループ部分（図 7b 矢印）についても原子レベルの非常に高い構造精度を得ることができ、その結果は、NMR の緩和時間測定の結果(5)とも良い一致を示した。全体構造の精度も従来法の計算に比べて劇的に向上した（rmsd 1Å 程度だったものが 0.32Å に向上）（図 7 a, PDB ID: 1O5P）。

Holo-NCS は、7 本鎖の短くいり込んだβシート構造を持っている（図 7、図 8）。βシートは、並びあったペプチド鎖間に多数の水素結合を持っている。この水素結合による構造安定化と短いβシートのパッキングが ET-1 の構造モチーフと同様の影響を持ち、従来の不十分な計算量において全体構造の収束を妨げていたと考えられる。実際、βシートは、NMR 測定から得られる距離情報も多く、技術的に“決め易い”構造であり、逆に溶媒に露出しているループ部分は、情報が少なく“決め難い”構造である。こういった情報精度の部分構造特異性は従来の計算法では見逃されることが多く、間違った結果の解釈を生んできたと考えられる。

9　まとめ

グリッドコンピューティングによる計算量の増大は、単に速く計算できるという利点に留まらず、従来の技術的限界のもとに見逃されてきた問題を掘り起こす点でも有用性が高いことが実証できた。

従来考えられてきた溶液中における ET-1 の C 末端構造のゆらぎと holo-NCS のループ部分の運動性は、他の部分構造における SS 結合や水素結合による構造安定化とそれによる情報精度の差から生じた計算誤差だという結論を得た。グリッドコンピューティングを用い計算量を増やすことで、初めて初期構造依存性の問題が未解決であったことがわかった。そして、計算誤差を取り除くことでタンパク質の立体構造決定の精度が飛躍的に向上することがわかった。

構造検索の網羅性が個々の構造に依存するのではなく全体としての構造の分散性によって実現されていることは特筆に価する。図６の計算手順で幾つかの構造が失われたとしても、それがランダムに起こり、初期構造の数に比べて無視できるほど少ない限り最終結果に影響を及ぼさない。これは、インターネット上でグリッドコンピューティングを行う場合に大きな利点になる。異なるネットワーク環境にある PC 上で無作為な計算のドロップアウトもしくは遅延が起こっても問題が無いということが保障されるからだ。

今回は、比較的簡単に構築できるグリッドコンピューティング環境を既存の計算手法に適用しただけであるが、それでも、従来の知見を覆す結果が得られたことも注目に値する。既存の技術の組み合わせだけでこの結果を得たのは、それが空白の境界領域に位置していたためであろう。現在、情報伝達の速度に革命的な変化が起こっているにもかかわらず情報から抽出される知識がその担体である人間から人間（あるいはその集合）へと伝達される速度には大きな進歩は無い。むしろ、学問の細分化と境界の峻険さによってその速度が妨げられているようにも見える。したがって、今回の報告のような境界領域は手付かずのまま大量に残されていると考えられる。今後も日進月歩を続けるコンピュータとネットワーク環境の中、様々な研究分野の境界を越えることで、個々の情報の精度が見直され、知識・知見の質的向上が図られるものと期待される。

10　謝辞

SUN GRID Engine のサポートをいただいたサン・マイクロシステムズ（株）荒木万里子、林憲一両氏に謝意を表します。NMR 測定などの実験は大阪大学薬学部小林祐次教授によって行われたものであり（文献 1, 5）、計算では、その原子間距離情報データ（PDB ID: 1V6R および 1O5P）を使用させていただきました。

参考文献

1. Takashima, H., Mimura, N., Ohkubo, T., Yoshida, T., Tamaoki, H., and Kobayashi, Y. (2004) Distributed computing and NMR constraint-based high-resolution structure determination: applied for bioactive peptide endothelin-1 to determine C-terminal folding. *J. Am. Chem. Soc.* **126**, 4504-4505.

2. Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Res.* **160**, 65-73.

3. Tamaoki, H., Miura, R., Kusunoki, M., Kyogoku, Y., Kobayashi, Y., and Moroder, L. (1998) Folding motifs induced and stabilized by distinct cystine frameworks. *Protein Eng.* **11**, 649-659.

4. Takashima, H., Tamaoki, H., Teno, N., Nishi, Y., Uchiyama, S., Fukui, K., and Kobayashi, Y. (2004) Hydrophobic core around tyrosine for human endothelin-1 investigated by photo-CIDNP NMR and MALDI-TOF-MS. *Biochemistry* **43**, 13932-13936.

5. Takashima, H., Yoshida, T., Ishino, T., Hasuda, K., Ohkubo, T., and Kobayashi, Y. (2005) Solution NMR Structure Investigation for Releasing Mechanism of Neocarzinostatin Chromophore from the Holoprotein. *J. Biol. Chem.* **280**, 11340-11346.