グリッドコンピュータを利用した蛋白質立体構造の予測に関する研究

１．2002 年度の具体的な研究計画

神戸大学理学部で開発中の蛋白質立体構造予測プログラムをグリッド上に適した形に整備し、グリッド上にインストールして、開発研究を進める。とくに今年度は、

- 蛋白質立体構造データベースと連携した形での、モデルパラメータ最適化の並列計算
- 構造サンプリングにおける大量並列計算のグリッド環境での試行

をおこなう。

２．2002 年度の進捗状況と研究成果

蛋白質立体構造データベースによるモデルパラメータの学習

モデルのパラメータ最適化は、構造予測の成否を分ける最重要ステップの一つであり、ネットワーク上に位置する既知構造データベースとの連携を計りながら、大規模分散処理をおこなうプロセスである。

私たちは、物理化学的考察に基づいて、蛋白質内の原子間相互作用の経験的エネルギー関数を構築した。

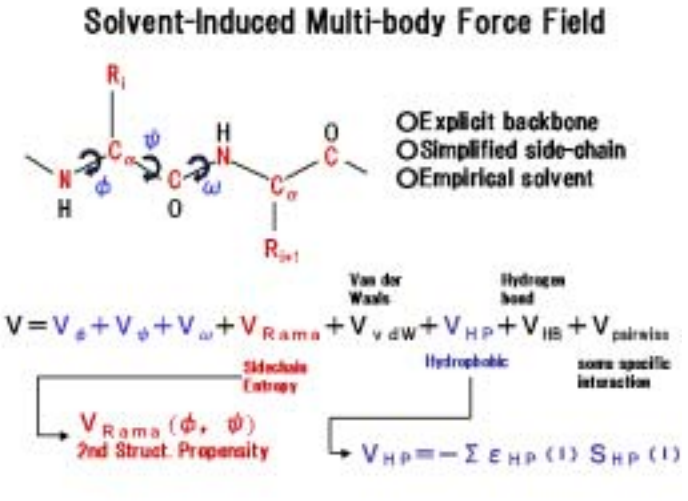

図１．私達が独自に構築した蛋白質相互作用エネルギー関数の模式図

これは、粗視化された表現形をとりながら、出来る限り高精度に相互作用を記述するためにいろいろな工夫が施されている。結果として、エネルギー関数にはかなりの数のパラメータ（今回は 80 個のパラメータ）が含まれている。これらを精度よく決めることが、構造予測成功の鍵を握る。これらパラメータの多くは、その値を実験によって直接観測できるものではない。そこで、なんらかの方法で実験で得られる情報からモデルに含まれるこれらパラメータを高精度に推定することが問題となる。

私たちは、拡張し続ける立体構造データを利用して、これらパラメータを最適化する方法を提案する。そこで、一種の学習アルゴリズムを利用した。すなわち、蛋白質は天然構

造で自由エネルギー最小であるから、「天然構造のエネルギーE_Dが無数にある非天然構造のエネルギーの平均$<E>_D$より十分低くなる」ようなパラメータを探すのである。より詳細には、

$$Z = \frac{E_N - \langle E \rangle_D}{\Delta E_D}$$

を考える。ここで、右辺分母は変性状態アンサンブルにおけるエネルギーの標準偏差である。エネルギーEがパラメータの関数なので、Zもそうである。Zが負でその絶対値が大きくなるようにパラメータを最適化する。上記 Z はある一つの蛋白質についてのスコアであるが、モデルパラメータを一つの構造既知蛋白質に対して最適化しても意味がない。私たちは、これを 40 個のトレーニング蛋白質セットについて平均したものZ_{AV}を目的関数として最適化を行った。こうして得られるパラメータは、トレーニング蛋白質以外に対してもうまく機能するものと考えるのである。

さて、Z_{AV}の最適化は、非常に大規模ではあるが、分散処理可能な過程である。パラメータを変数とした 80 次元のモンテカルロ法によって、最適化を行った。このなかでZ_{AV}を数万回程度繰り返し計算する。Z_{AV}は、40 個のトレーニング蛋白質についての Z の平均であり、各蛋白質の Z の計算は互いに独立している。Z の計算は、沢山の（今回は約千個）の非天然構造のエネルギー計算を含むが、この計算も互いに独立している。従って、このプロセスは、非常に高い効率で並列計算が可能である。

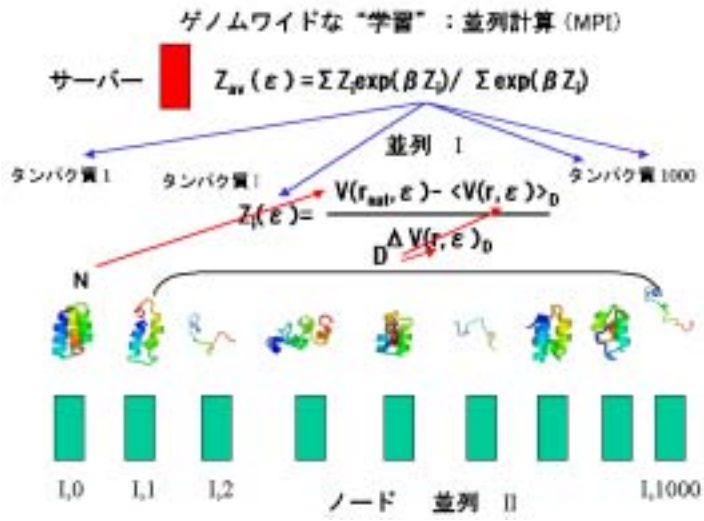

図２．分散処理による、モデルパラメータの最適化の模式図

今年度は、均質な計算能力を持つ PC クラスタによる並列処理を行い、この計算を高速化することに成功した。グリッド環境でこれを動かすことが出来れば、（現在 40 蛋白質に対してトレーニングしているところを）1 万以上ある現在既知の全蛋白質をトレーニングセットにしたパラメータチューニングも視野に入ってくる。これは、来年度以降の課題である。

蛋白質の構造サンプリング

蛋白質の立体構造予測解決へのもう一つの鍵は、高速に、広大な蛋白質の構造空間を探

索するシステムを構築することであり、このプロセスは元来分散処理に最適であり、グリッドコンピューティングが威力を発揮する問題である。

私たちは、Baker らによって展開されてきたフラグメント組合せ法の拡張版の開発を行った。フラグメント組合せ法について簡単に説明しておく。構造を予測しようとするターゲット配列の各 3 アミノ酸並びについて、それと類似性の高い配列を構造データベースのなかから検索する。検索スコアがトップ 20 程度の配列について、その部分の構造をとりだして、それをターゲットの対応する部分の構造候補とする。こうして各 3 残基について、平均で 20 から 30 程度の構造候補を作成、保存しておく。問題は、ターゲットの全体構造を組上げる段階である。目標は、各アミノ酸部位について約 20 程度の候補のなかから一つずつを選び、全体としてエネルギー最小の構造を探すのである。フラグメント組合せ法は、各アミノ酸部位のとる構造を 20 個程度の有限個に抑え、それでいて全体としてかなりの高精度で構造を組上げることが出来る点で優れている。しかし、100 程度のアミノ酸長をもつ比較的小規模の蛋白質でさえ、この全体構造の場合の数は、有限ではあるが、莫大である。とてもすべての場合を計算し尽くすことは不可能である。そこで、シミュレーテッドアニーリングモンテカルロ（SAMC）法によって構造サンプリングをおこなうのである。SAMC 法は、全探索よりははるかに効率よいサンプリングを行うことが出来るが、残念ながら 100 残基程度より大きなもの、あるいは複雑なトポロジーをもつ蛋白質では、その構造サンプリング能は、十分でないことが分かっている。

SAMC 法でサンプリングが不十分な場合、物理などの計算科学の分野では、よく知られた拡張アンサンブル法を使うことになる。拡張アンサンブル法には、レプリカ交換法やマルチカノニカル法などが存在し、それらはそれぞれ、SAMC 法より高い効率でサンプリングが行えることが分かっているのである。ところがここで問題がある。すなわち、拡張アンサンブル法を適用するためには、いわゆる詳細つりあいを満たしたアルゴリズムでなければならないのだが、実はフラグメント組合せ法はこれを満たしていない。そこで、私たちは、フラグメント組み立て法を拡張して、詳細釣り合いを満たし、したがって熱力学的平衡状態のアンサンブルを実現することが出来る新しいアルゴリズムを開発した。さらにそれをマルチカノニカル法と組み合わせることで、従来よりはるかに高効率な構造サンプリングを可能にした。

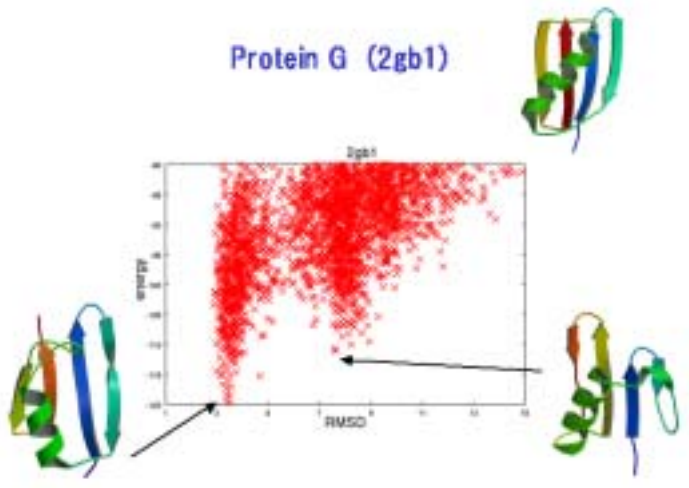

図３．分散処理によるテスト蛋白質の構造サンプリングの例。縦軸にエネルギー、横軸に天然構造からの構造のずれ（RMSD）をつかって、サンプルされた構造を点でプロットした。右上の図が蛋白質の天然構造、左下が計算によって得られた最低エネルギーをもつ構造であり、正しいフォールドを持っていることがわかる。

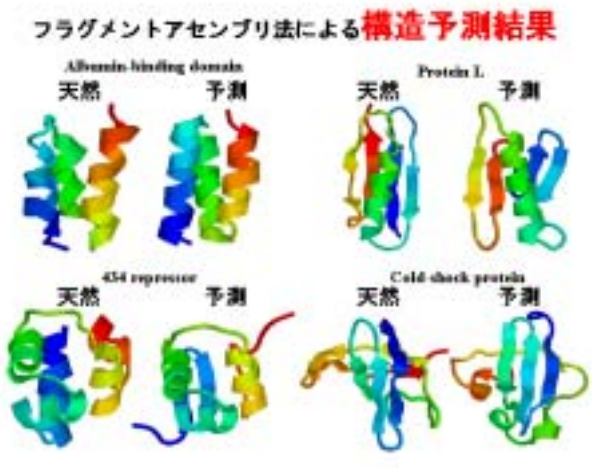

図４．フラグメント組合せ法による構造予測のベンチマークテスト。各蛋白質について左が天然構造、右が予測構造をあらわす。簡単な蛋白質では、その全体フォールドを予測することができる。一方、難しいもの（右下）では、一部分の構造が正しく予測されるに留まっている。

マルチカノニカルモンテカルロ法による構造サンプリングにおいて、非常に単純なやり方ではあるが、分散処理を実現した。今年度は、グリッドおよび PC クラスタ様の環境で、自動的に無数のジョブを投入、結果集積するツールを作成、テストした。多数のノードにおいて、全く独立なシミュレーションを無数（たとえば 1 万、あるいは 10 万回程度）に行うことにより、全構造空間に渡る探索を高速化した。大量の同時計算を行う上で、計算結果のデータサイズなど多くの問題点が見つかった．2 年目には、グリッド環境で効率よくシミュレーションを行うことが出来るよう、これら問題点を解決するツールの整備を進める。

国際構造予測コンテスト（CASP5）

蛋白質の立体構造予測に関する､国際ブラインドコンテストが 2 年に一度行われている。構造があらかじめ知られていない蛋白質のアミノ酸配列がインターネット上で提出され、参加者が各自の手法でその立体構造を予測する。年末にその蛋白質について、実験により得られた構造と予測構造を、第 3 者が客観的に比較し、採点される。2002 年行われた第 5 回コンテスト CASP5 に、私たちも参加した。コンテストの予測期間は 2002 年の 5 月ごろはじまり 8 月末頃に終わったので、グリッド環境を使うには整備が間に合わなかった。したがって、もっぱら単純な PC クラスタによる計算を行った。

アミノ酸配列から、自由に立体構造を組み立てていく NewFold 部門が私たちのターゲットとなる。5 本のヘリックスよりなる新奇フォールドをとっていた T0170 において、私たちは正しいフォールドを予測することが出来た。また、T0181 などで部分構造をほぼ正解し、相対的に高い評価を得た。

今回私たちの方法は、比較的小型（110 残基程度以下）で、とくにヘリックスが多いターゲットについては、ある程度機能した。一方それより長い蛋白質、あるいは複雑なトポロジーをもつものについては、構造サンプリングの効率はまだ不十分であり、結果としてよい予測構造を出すことが出来なかった。CASP5 において、私たちの弱点は明確になった。すなわち、ある程度以上長い蛋白質の構造をいかにサンプリングするか、に焦点は絞られる。グリッドコンピューティングの適用が、構造サンプリング問題の解決に向けて必須であると考えられる。