計算科学フロンティアの展開　その1
ー CASP ノススメー

佐々木　　尚

生命現象は，その担い手であるタンパク質によって支えられている。多くの場合タンパク質は，生体内でアミノ酸のならび（アミノ酸配列）で決定される一つの決まった立体構造を保持して機能を発現しており，タンパク質の機能を理解するためには，まずその立体構造構築原理を理解する必要がある。また，ヒトゲノムの解読が完了し，他の生物種のゲノムデータの蓄積も進んでおり，その DNA 情報から創薬や食品などの分野への応用の期待が高まっているが，これらにおいても，その配列がコードしているタンパク質の立体構造情報が必要となってくる。現在では，アミノ酸配列の情報から立体構造を決定する手法の開発が世界中で精力的に進められている。これらの手法の優劣を決める場として，CASP（Critical Assessment of Techniques for Protein Structure Prediction）と呼ばれるタンパク質立体構造予測コンテストが 1994 年から 2 年に一度開催されるようになった [1,2]。CASP では，世界中の研究グループが 2 つの部門に分かれて立体構造予測法の優劣を競い合う。似通ったアミノ酸配列をもつタンパク質がタンパク質立体構造データベース中にすでに存在する問題配列の立体構造予測を行う部門と，似通ったアミノ酸配列をもつタンパク質が，タンパク質立体構造データベース中に見つからない問題配列の立体構造予測を行う部門が存在し，数ヶ月以内に構造決定される予定のタンパク質のアミノ酸配列が問題として出題される。前者では，既存のタンパク質の構造テンプレートを用いることが可能であり，構造を予測することは比較的容易である。しかし，後者では，既存のタンパク質の構造テンプレートの転用が難しいタンパク質のアミノ酸配列が問題として出題される。このようなタンパク質の立体構造予測は非常に困難であり，これまでさまざまな手法が考案されてきた。中でも，Baker らのフラグメントアセンブリ法 [3] は，この分野で成功を収めてきた手法の一つである。この手法の特徴は，配列プロファイルと呼ばれるタンパク質の進化的な情報を用いて，構造未知のタンパク質のフラグメント構造候補を絞り込む点にある。この手法によって，比較的小さなタンパク質の立体構造予測が可能になりつつあり，今現在，構造予測を行う研究者の間でもっとも広く使われている手法である。しかしながら，このフラグメントアセンブリ法をもってしても，大きなタンパク質やトポロジーの複雑なタンパク質になると途端に予測が困難になる。立体構造予測の分野がさらに進歩するためには，これからさらなる改良や全く新しい発想の手法が求められるだろう。

そこで我々は，これまでにない新しい構造予測法の開発を行ってきた [4,5]。本手法では，アミノ酸残基を一つの粒子で記述し，それを仮想的なバネでつないでタンパク質の粗視化モデルを構築した。また，タンパク質立体構造データベースから経験的ポテンシャルを構築し，ランジュ

バン動力学計算に用いた。構造サンプリング法にはシミュレーテッドアニーリング法を採用した。本手法を用いてタンパク質の折り畳みシミュレーションによる立体構造予測のデモンストレーションを行った結果が図 1 である。左図から右図に向かって，伸びた構造から徐々にコンパクトな構造になり，ネイティブ構造を再現している（PDBid:1R69 を参照 [6]）。

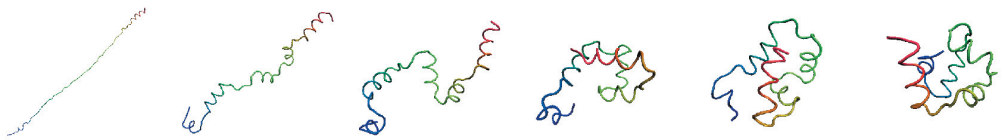


図 1

我々は，本手法の予測性能を知るため CASP7 に参加した。チーム名は“KORO”とした。CASP ではウェブページ上にこれから構造決定される予定のタンパク質の配列が公開され，参加者はそれらの立体構造を予測することとなる。CASP7 においては 5 月末から 8 月上旬にかけて，およそ 1 週間に 10 個程度の問題が順次出題された。それぞれの問題配列には期限が設けられており，それまでに構造を提出しなければならない。ただし，必ずしも全問解く必要はなく，自分たちが参加する部門に合わせて解くべき問題配列を自分たちで選別すればよいのである。我々は，自分たちの参加する部門，開発した手法，持っている計算機クラスターの規模などに応じて解く問題配列を決めた。幸いにも，研究室の計算機クラスターだけでなく，情報連携基盤センターの協力の下に HPC2500 も利用できたため，我々が解くべき問題配列のほぼすべてに関して構造を提出することができた。特に，500 残基を超えるような大きなタンパク質の構造予測には，この HPC2500 が大きく貢献してくれた。

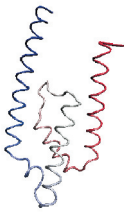
図 2 は，実際に我々が提出した予測構造と，その答えであるネイティブ構造を示したものである。予測された構造とネイティブ構造が互いによく似ていることが見て分かる。ここで指標“GDT_TS”は以下の式で定義されている。

$GDT\text{-}P_x$：ネイティブ構造に対して，誤差 x（Å）以下で重ねられる残基の割合

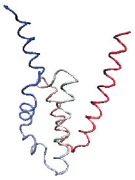
$$GDT\text{-}TS = \frac{(GDT\text{-}P_1 + GDT\text{-}P_2 + GDT\text{-}P_4 + GDT\text{-}P_8)}{4} \times 100(\%)$$

GDT_TS は予測構造とネイティブ構造との類似度を表現し，特に部分的に似ているときにも値が高くなる指標である。近年の CASP ではこの指標がよく用いられている。また，RMSD（root mean square deviation の略）も同様に予測構造とネイティブ構造との類似度を表現する指標であるが，構造全体に渡って互いの残基を重ね合わせられないと低い値にはならない厳しい指標である。GDT_TS の値は 100％に近づくほどよく，また RMSD の値は 0（Å）に近づくほどよいということになる（CASP においては，RMSD による採点は行われていない）。図 2 の RMSD の隣の括弧内の数字は RMSD を計算する際に用いた残基の番号を記したものである。我々の参

問題配列：T0283（112 残基）



予測構造
GDT_TS=57.143
RMSD=4.877(Å)



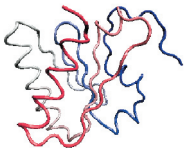
ネイティブ構造

問題配列：T0348（61 残基）

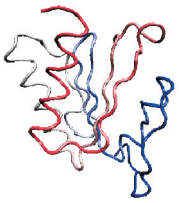
予測構造
GDT_TS=52.869
RMSD(2-62)=11.246(Å)
RMSD(2-42)=5.029(Å)

ネイティブ構造

問題配列：T0354（120 残基）



予測構造
GDT_TS=51.875
RMSD(1-120)=9.869(Å)
RMSD(1-100)=3.657(Å)



ネイティブ構造

図２

加した部門では，GDT_TS が 50%を超えるような結果は稀であり，チーム KORO は非常によく健闘していると言える。実際の評価はこれを執筆している今から 1 週間後に開催される CASP7 ミーティングで発表されるため，現時点ではあきらかではないが，本文が皆さんに読まれている頃には結果が出ている。是非 CASP7 のホームページ [2] から我々の結果を覗いてみて欲しい。そして何よりも皆さんに興味を持っていただき，これからどんどん参加していただきたいと思う。

【参考文献等】

[1] http://predictioncenter.org/

[2] http://predictioncenter.org/casp7/

[3] K.T. Simons, C. Kooperberg, E. Huang, and D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol*. **268**（1997）209.

[4] T.N. Sasaki, and M. Sasai, A coarse-grained langevin molecular dynamics approach to protein structure reproduction. *Chem. Phys. Lett*. **402**（2005）102-106.

[5] T.N. Sasaki, and M. Sasai, Development of a Technique to Dynamically Predict Protein Structures for the Reproduction of Protein Folding Process., EABS&BSJ2006, Okinawa, Nov. 13-14,（2006）, poster presentation.

[6] http://www.rcsb.org/pdb/home/home.do

（ささき　たけし：名古屋大学大学院工学研究科計算理工学専攻 21 世紀 COE プログラム「計算科学フロンティア」COE 研究員）