【技術資料】

CASP5 における FAMS の成績

平野 敏行[*], 土屋 俊夫[*], 岩館 満雄[†], 竹田-志鷹 真由子[†], 梅山 秀明[†]

2003 年 8 月 27 日

1　はじめに

CASP(Critical Assessment of techniques for protein Structure Prediction) とは California にある Protein Structure Prediction Center が主催している国際的なタンパク質立体構造予測コンテストです。CASP は既に 4 回 (1994, 1996, 1998, 2000)、隔年で行われており、今回 (2002) で 5 回目になります。これらの結果の詳細は CASP web サイト (http://PredictionCenter.llnl.gov/) を参照してください。CASP5 は 2002 年 5 月に始まり、8 月まで行われ、187 のチームが、67 のターゲットとなるアミノ酸配列に対し、タンパク質の立体構造の予測を行いました。

本報告は、北里大学梅山研究室の研究チーム FAMS, FAMSD に注目し、CASP5 における成果を、客観的に評価したものです。**FAMS はホモロジーモデリングの手法を使って、全自動でタンパク質の立体構造予測を行うソフトウェアです。**CASP コンテストでは FAMS のようなソフトウェアの他に、市販のソフトウェアや、独自開発のモデリングソフトウェアなどを用いて、最終的には経験者が判断してモデリングを行うチーム (手動モデリングチーム) や、全自動であるものの、FAMS のようなソフトウェアの結果を利用する、メタサーバと呼ばれるソフトウェアも参加しています。一般的に、手動モデリングチームやメタサーバの結果は、FAMS のような全自動のソフトウェアに比べ、精度が良くなります。これまで FAMS はこうした CASP の条件下においても、好成績を残しています。ここでは、FAMS は CASP5 においてどのような結果を残したのか、CASP5 web 上で得られる数値データを基に客観的に判断しました。

2　調査方法

データは以下の出題されたターゲットとなるアミノ酸配列ついて調査しました。ターゲットの難易度はホモロジー検索で使用されるツールによって分類されます。本調査で使用したターゲットは表 1 の通りです。

[*] 日本 SGI(株)
[†] 北里大

表 1: 本調査で使用したターゲット

ホモロジー検索システム	ターゲット
BLAST	T0137, T0140, T0142, T0143, T0150, T0151, T0153, T0154, T0155, T0160 T0167, T0177, T0178, T0179, T0182, T0184, T0185, T0188, T0190, T0191
PSI-BLAST	T0133, T0141, T0149, T0152, T0165, T0169, T0172, T0176, T0184 T0185, T0186, T0189, T0192, T0195
Transitive PSI-BLAST	T0130, T0132, T0136, T0159, T0168, T0193

データはチーム毎にこれらターゲットの平均値を用いています。ただし、それぞれの難易度 (BLAST, PSI-BLAST, Transitive PSI-BLAST) 毎に予測したターゲットの数が半分以下のチームについては除外しています。また明らかに予測が不完全なものも除外しました。

3 結果

CASP5 での原子位置の精度に対する評価は、実験データと予測モデル間の対応する原子間の距離の差が、ある範囲内に収まった (ある範囲外を cutoff) した予測確率として評価されています。すなわち cutoff が 1Å であれば、その確率はそのモデルのある原子がターゲットの相当する原子の位置から 1Å 以内に存在した割合になります。cutoff の値が大きくなるにしたがって、その確率も 100%に近くなります。cutoff が小さい段階で予測率の高いものは精度が良いということになります。

図 1 と図 2 に BLAST でホモロジーが検出できるレベルにおける確率について示しました。図 1 は C_α について評価したものであり、図 2 は全原子について評価したものになります。緑色は手動モデリングチームを示しており、青色は全自動のソフトウェア、灰青色はメタサーバを示しています。

この 2 つの図から、C_α の予測、全原子の予測ともに、FAMS は他のソフトウェアと比べてかなり良い精度を持っていることがわかります。特筆すべき点は、C_α のみならず全原子においても精度が高いことであります。C_α の位置はホモロジーが高ければ比較的容易に決定できますが、全原子が対象となると側鎖の原子の精度が効いてきます。実際、大多数のメタサーバを含む全自動のソフトウェアは、全原子の結果は悪くなっています。その点、**FAMS は全原子の結果においてもトップクラスの精度を保持しており、側鎖の原子も精度が良く予測していることが示唆されています。**

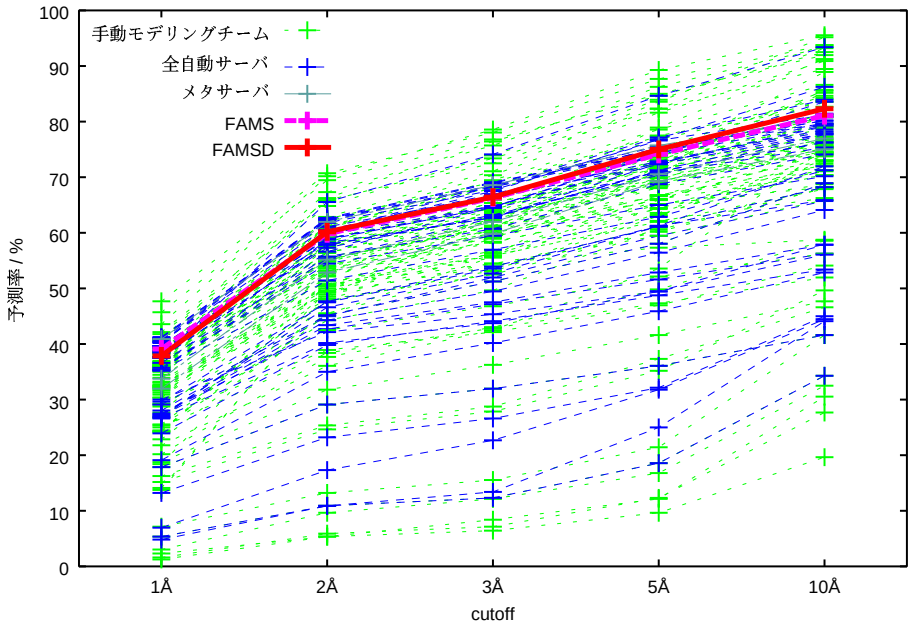

図 1: C_α における予測精度

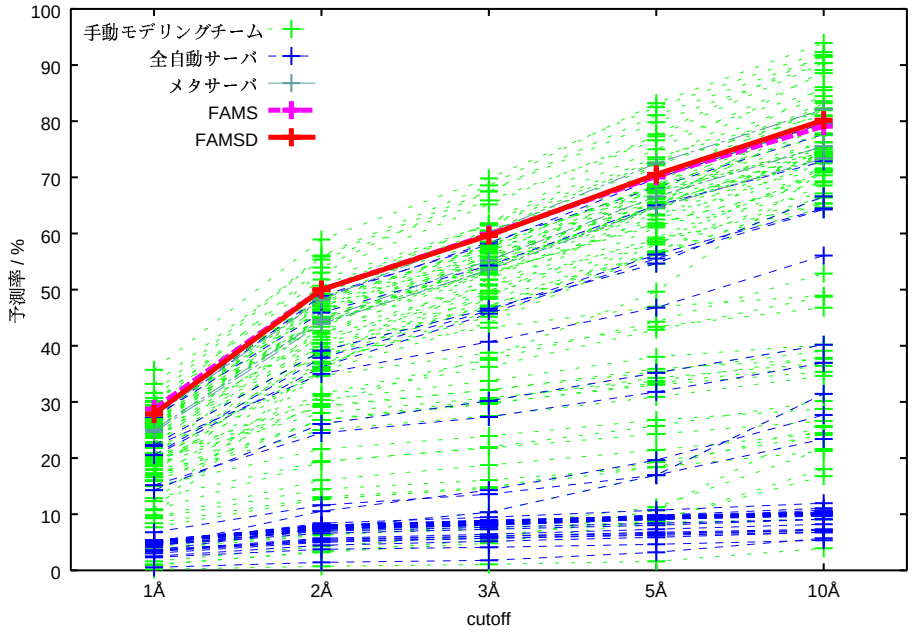

図 2: 全原子における予測精度

ターゲットの難易度を上げた場合の精度を評価してみます。図 3 の横軸はターゲットの難易度 (どのツールでホモロジーが検出できるか) を示し、縦軸は GDT_TS というパラメータを示しています。GDT_TS(Global Distance Test, Total Score) は、CASP において精度を総合的に評価するパラメータです。GDT_TS は、実験データと予測モデルの対応する原子位置 (C_α のみ) の差が、1Å, 2Å, 4Å, 8Å 以内にある割合をそれぞれ求め、その 4 つの割合の平均として求められます。予測精度の良いチームは、GDT_TS が大きくなります。

全体的にどのチームも難易度が上るにつれ精度も落ちているものの、FAMS は世界トップクラスの精度を保持していることがわかります。GDT_TS のパラメータが C_α の原子位置を基に求められており、各難易度における全原子から求められた結果が公表されていないのが残念ですが、前述の結果 (図 1、2) を踏まえれば、FAMS は難易度を上げても世界トップクラスの精度を維持していることが示唆されます。

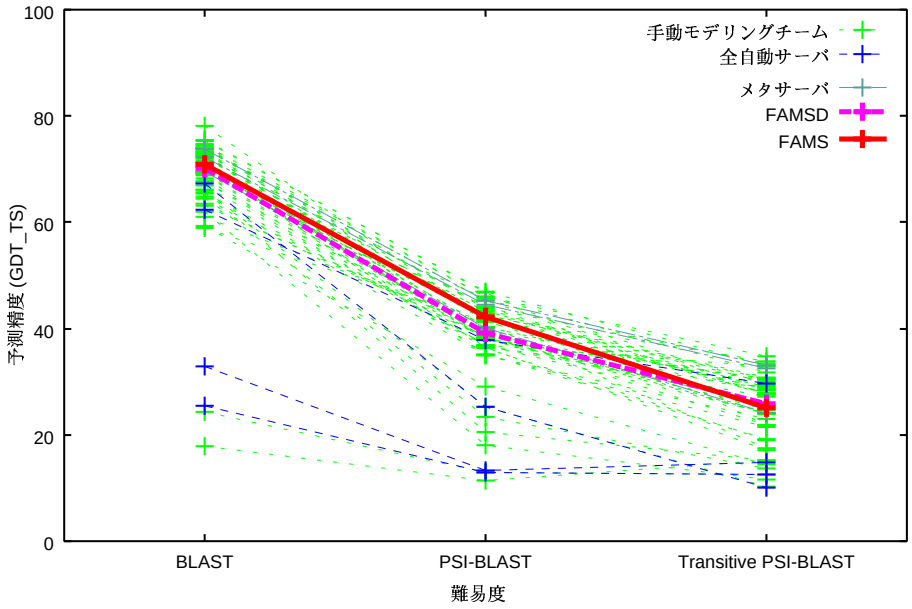

図 3: ホモロジーの違いと予測精度の関係

4　まとめ

FAMS は全自動でタンパク質の立体構造予測を行うソフトウェアです。FAMS は完全全自動でありながら、世界でトップクラスのモデリング熟練者に匹敵する程の予測精度を有しています。手動によるモデリングでは、解析途中にて人の判断を加えるため、そのオペレータの塾練度がモデリングの精度に直接反映されます。一方、FAMS のような全自動の予測ソフトウェアでは、アミノ酸配列を入力するだけで、オペレータの熟練度に関係なくタンパク質の立体構造を予測することができます。FAMS は、精度の高い全自動の立体構造予測ツールとして、非常に有用であると言えます。