解　説

タンパク質立体構造予測の現状と未来

Present and Future Prospects of Protein Structure Prediction

朴 聖俊 Sung-Joon Park　神戸大学理学部化学科 Department of Chemistry, Faculty of Science, Kobe University park@proteinsilico.org, http://www.proteinsilico.org/

千見寺 浄慈 George Chikenji　（同　上） chikenji@theory.chem.sci.kobe-u.ac.jp, http://theory.chem.sci.kobe-u.ac.jp/~chikenji/

広川 貴次 Takatsugu Hirokawa　産業技術総合研究所生命情報科学研究センター Computational Biology Research Center, Advanced Industrial Science and Technology t-hirokawa@aist.go.jp, http://www.cbrc.jp/

富井 健太郎 Kentaro Tomii　（同　上） k-tomii@aist.go.jp, http://www.cbrc.jp/

高田 彰二 Shoji Takada　神戸大学理学部化学科 Department of Chemistry, Faculty of Science, Kobe University stakada@kobe-u.ac.jp, http://theory.chem.sci.kobe-u.ac.jp/

keywords: protein structure prediction, comparative modeling, fold recognition, ab initio, CASP

1. はじめに

タンパク質の立体構造予測問題とは，与えられたアミノ酸配列をもつタンパク質が生体内でとる三次元立体構造を理論的に予測する，という至ってシンプルな設定の問題である．構造予測研究は 40 年にもわたる長い歴史をもち，多くの挑戦と失敗とを繰り返しながら，徐々に理論と実践の両面から成熟してきた．

現在の構造予測研究の最先端では，物理化学とバイオインフォマティクスを融合させており，これまでに蓄積してきた物理化学の側面と情報学の側面を巧みに活用している．そして，場合によっては X 線構造と同レベルの予測が可能になり実用化されるとともに，タンパク質の構築原理を理解するための基礎研究の一つとなっている．

この分野の現状は，1994 年以降 2 年に一度行われている構造予測技術評価会議 CASP（Critical Assessment of Techniques for Protein Structure Prediction）の結果から非常によく概観できる．

本解説は，タンパク質立体構造予測における最新の研究成果と動向について，2004 年に行われた CASP6 を一望することによって理論と実践の両面から解説し，構造ゲノム科学における応用と期待について述べる．

2. タンパク質立体構造予測問題の背景

2·1 タンパク質の階層構造

タンパク質の構成要素であるアミノ酸は中心炭素原子（C_α）に水素原子（H），アミノ基（NH_2），およびカルボキシル基（COOH）が結合する共通構造をもち，さらにアミノ酸の化学的性質を決める側鎖も C_α に結合している．隣接するアミノ酸間で，アミノ基とカルボキシル基との脱水重合によって 1 本のポリペプチド鎖＝タンパク質となる．タンパク質は，鎖に沿った化学結合周りの回転自由度によってねじれ曲がった立体構造（三次構造）をとるが，その個性は鎖上に突き出た側鎖の並び順（アミノ酸配列，一次構造とも呼ぶ）に起因する．鎖には方向性があり，鎖の端が NH_2，および COOH である末端をそれぞれ，N 末端と C 末端と呼ぶ．

三次構造は一つあるいは複数のドメインからなる．一つのドメインは複数のモチーフから構成されている．モチーフは二次構造と呼ばれる α ヘリックスと β ストランド，それ以外のコイルがつながった構造であり，三次構造の密な球状形成に寄与する．三次構造は集まって多量体の四次構造を形成する場合がある．

2·2 進化的関係に基づく立体構造予測

配列空間はほとんど無限である．20 種類のアミノ酸を 200 残基つなげて作るタンパク質の配列数は 20^{200} という天文学的な数となる．しかし，実在するタンパク質の三次構造の大まかな形（フォールド）のバリエーションには限りがあり，フォールドはアミノ酸配列の変異に対して保守的であると考えられている [Liu 04].

進化的に類縁関係にある二つのタンパク質は相同タンパク質と呼ばれ，類似する立体構造と生化学的機能を保っている．遠縁タンパク質の配列類似度は，しばしば 30%以

下の同一残基率であり，フォールドは保存されているが，局所構造の変異によって異なる機能を発現する場合が多い [朴 05b]．一般に，構造未知のタンパク質に対して，それと有意な配列類似性をもつ既知構造タンパク質が見つけられれば，未知タンパク質についての構造予測はその既知構造タンパク質を鋳型として進めることができる．

2·3 第一原理的立体構造予測の難しさ

第一原理に基づく立体構造予測は，依然として未解決の難問題である．予測を困難にする要因として，探索すべき構造空間が広大であることと，適切なエネルギー関数の設計が非常に難しいことがあげられる．

たとえば，各アミノ酸の各回転自由度がとりうる配位をそれぞれ三つと考えても，アミノ酸 100 残基の小型タンパク質の構造数は $3^{200}(\approx 10^{95})$ という天文学的な数字となる．その網羅的な探索は不可能であり，非常に効率的な構造探索法が必要である．さらに，構造探索時の最適化関数＝エネルギー関数が，ファンデルワールス力，静電相互作用，水和エネルギーなどの膨大かつ複雑な相互作用からなっており，それら相互作用エネルギーを精密かつ高速に評価することは極め難しい．

2·4 立体構造予測への世界的な挑戦

§1 構造生物学とバイオインフォマティクスの登場

1990 年代以降，構造生物学とバイオインフォマティクスの台頭は，それまでの純粋な物理化学に基づく予測方法に格段の変化と改善をもたらした．

構造データの蓄積・整備と類似配列性検索方法の開発・成熟は，構造既知タンパク質を鋳型構造とする予測を可能にした [Tramontano 03]．これは，予測対象配列と近縁・遠縁関係にあるタンパク質が構造的・機能的類似性をもっている，という経験則を利用した帰納的予測である．

また，大規模構造データの解析によって，タンパク質の局所構造の分布に偏りが存在することがわかった [Kolodny 02]．これにより，予測対象配列の局所部位を既知構造データベースに存在する部分立体構造で表現するアイディアが生まれ，局所部位のエネルギー相互作用を粗視化して立体構造予測を行う方法が出現した [Simons 99]．

両アプローチにおけるスコア関数とエネルギー関数は，二次構造予測や残基間コンタクト予測などの情報を利用して改良することができ，予測構造を構築していくうえで重要な方針を与える．このような「立体構造予測のための予測」を行うことは，次に説明する CASP に成功している手法の基本的な共通点である．

§2 構造予測技術評価会議 CASP

1994 年から始まった CASP は，2 年に一度行われる世界的規模の立体構造予測のブラインドテストである．CASP のスキームは非常に単純である．CASP 開催年の 5 月から 9 月頃まで立体構造未公開の予測対象配列（ターゲット）を実験者から集めて，次々と予測者へ出題する．手動予測チームは数週間以内に，全自動予測サーバーは 48 時間以内に 1 問に付き 5 個までの予測構造を提出する．12 月の会議で，第三者である評価者による採点，評価と成績優秀者の講演が行われ，翌年のジャーナル “Proteins: Structure, Function, and Bioinformatics” に特集を組む．

表 1 CASP における予測チームのエントリ数（Predictor），予測対象配列数（Target），予測構造数（Prediction）の推移

	Predictor	Target*	Prediction
CASP1	35	33	100
CASP2	152	42	947
CASP3	120	43	3807
CASP4	198	43	11136
CASP5	259	67	28728
CASP6	266	87	41283

* 取り消された配列を含む

構造予測分野における CASP の役割は重要性を増してきている．参加者数とターゲット数は年々増加しており，2004 年の CASP6 では，200 以上のグループが 87 問の立体構造予測に膨大な計算資源を投入するなど（表 1），世界的に熾烈な競争が繰り広げられている．

3. 立体構造予測の現状

3·1 鋳型構造による予測

§1 比較モデリング

立体構造が未知であるタンパク質のアミノ酸配列が与えられたとする．最も単純なアプローチは，立体構造既知のタンパク質に，与えられたアミノ酸配列と類似した配列が存在するかどうか検索を行うことであろう．現在では，PSI-BLAST[Altschul 97] をはじめとする高度な配列類似性検索技術が発達してきている．これらの技術は，与えられた予測対象配列と構造既知であるアミノ酸配列とのアラインメントを与える．これにより，予測対象配列中の個々のアミノ酸が，立体構造においてどの位置に対応するか推定することができる．こうしたアラインメントに基づき，立体構造モデルを構築することを比較モデリング（Comparative Modeling，CM）という．

CM では，上記のように鋳型構造（テンプレート）とのアラインメントが与えられれば，最終的な予測対象タンパク質の原子座標は，一般に以下のようなプロトコルで得ることができる．

(1) テンプレートの座標をアラインメントに従って参照しながら予測タンパク質の実際のアミノ酸残基を割り当てる
(2) 原子間接触などの立体障害を分子力学計算等で回避する
(3) タンパク質構造としての適正度合いを評価関数で

チェックする

(4) 3) の過程で不適切な部位が見つかれば，アラインメントの修正などを行い，再度同じ手順を繰り返しながら最終構造を決定する

ただし，このプロトコルも予測の難易度によって留意点や必要な処方箋が加わってくる．

CM において，相同タンパク質をテンプレートとする予測は「ホモロジーモデリング」と呼ばれるが，これは予測精度が比較的安定しているため，実用レベルで用いられることが多い．ホモロジーモデリングでは，アラインメント情報を利用できないループ部分のモデリングや側鎖構造の配向予測が重要なポイントとなる．これらはタンパク質構造に基づくドラッグデザインでも鍵となるため，実用化に向けての課題とされている．

§2 フォールド認識

数多くのタンパク質の立体構造が明らかになるにつれ，必ずしも配列が類似していないタンパク質であっても，類似したフォールドをとり得ることが明らかになってきた [Chothia 92]．

フォールド認識（Fold Recognition，FR）は与えられたアミノ酸配列が，有限個であると考えられるフォールドのうち，いずれのものをとるかという問題設定に基づくアプローチの総称である．配列類似性の有無に関わらず，予測対象配列がとるフォールドを予測（認識）するための手法は，フォールド（しばしば日本語では構造）認識法と呼ばれる．近年では，後述するプロファイル比較法が FR の有力な手法であると考えられている．

FR 法で検出されたテンプレートを用いる予測問題では，ループ構造や側鎖構造予測はもちろんのこと，全体構造を構築する過程で工夫が必要となる．ループ構造のモデリングでは，残基の一致度や周辺環境との距離制限を考慮するデータベース検索や，二面角の経験的ポテンシャル関数によるモデリングなどがある [Fiser 00]．複数のテンプレートを用いるリコンビナントモデリングでは，大局的な構造を保ちながら部分領域ごとに参照する鋳型を変えて全体構造を構築する方法 [Kosinski 03] などがある．

§3 鋳型検索技術

配列類似性検索技術は，開発当初，ペアワイズのアミノ酸配列比較技術に依拠していた．ペアワイズのアミノ酸配列比較は，アミノ酸種相互の類似度を定義するアミノ酸スコア行列と，ギャップペナルティー，比較アルゴリズムにより達成される．問い合わせ（予測対象）配列を配列データベースに格納されている配列と逐一比較し，アラインメントスコアを計算する．それらのスコアを比較したデータベース中の配列の長さによる補正を行い，各々の統計的有意性を見積もるのである．

次世代においては，PSI-BLAST に代表されるように，問い合わせ配列の類似配列をまず集めておき，それらを用いてマルチプルアラインメントを構築し，そこからプロファイルを作成する．プロファイルとは，構築されたマルチプルアラインメントの任意の残基位置において，観測される出現頻度に応じたスコアをアミノ酸種毎に割り振った行列であり，位置特異的スコア行列とも呼ばれる．PSI-BLAST では，このプロファイルを用いてデータベース検索が行われる．これとは逆に，データベース中の配列各々について作成されたプロファイルデータベースを問い合わせ配列検索する IMPALA[Schaffer 99] と呼ばれる手法も存在する．

配列相互の比較，プロファイルと配列の比較，あるいは配列とプロファイルの比較の歴史の後に，更なる発展として登場し，近年非常に洗練されつつある手法が，プロファイル比較法である．この手法では，問い合わせ配列，データベース中の配列のそれぞれにおいてプロファイルが作成され，相互の比較が行われる．プロファイル比較法を用いた類似性検索の枠組みは，ペアワイズのアミノ酸配列比較を利用したデータベース検索と同様のものが利用可能である．異なる点は，ペアワイズのアミノ酸配列比較に用いられたアミノ酸スコア行列により与えられたアミノ酸種相互の類似度が，プロファイル比較では，任意の残基位置におけるプロファイルスコア相互の類似度に置き換わることである．プロファイルスコア相互の類似性尺度として，内積や相関係数をはじめ，様々な手法が提案されている [Wang 04]．

鋳型構造を用いる予測は，極論すれば構造未知タンパク質のアミノ酸配列と構造既知タンパク質とのできるだけ正確なアラインメントを得ることにつきるといえよう．こうしたアプローチの最大の欠点は，その性質上，どうしても新規構造予測には向かない点にある．

3·2 鋳型構造によらない予測

§1 *ab initio* 法

予測したいタンパク質の立体構造が既知構造データベースになかった場合，シミュレーションによって新しい構造を作り出し，その構造の妥当性を評価し，もっともらしい構造を予測構造とする，という作業をしなければならない．このようなアプローチを *ab initio* 法という．

本来，*ab initio* という言葉の意味は「第一原理から」ということであり，物理や化学の業界では量子力学の基礎方程式を近似なしに解く，というふうに連想される．しかし，タンパク質立体構造予測では程度の差こそあれ経験的な要素を含まざるをえないので，この分野では単に「新規構造を予測できる能力を持つ方法」を *ab initio* 法と呼んでいる．なお，最近では “*de novo*” 法や，CM との対比として “free modeling” などと呼ばれることがある．

さて，*ab initio* 法の指導原理は「与えられたアミノ酸配列のとりうる立体構造の中で最も安定な構造を探す」ということである．これを実現する為には，なるべく精密に，しかし（計算時間の都合上）計算可能なレベルに

まで粗視化したタンパク質の相互作用エネルギーを設計し，最も安定な構造を探索する必要がある．このような *ab initio* 法のアプローチのはじまりは，1960 年代にまでさかのぼるが，その歴史はまさに失敗の連続であったというべきだろう．しかし，1990 年代の終わり頃から状況は一変した．具体的には，1998 年に行われた CASP3 にて，それまで全く実用とは程遠かった *ab initio* 法の成功例が幾つか報告されたのである [Simons 99].

§ 2 フラグメントアセンブリ法

特筆すべきものは，CASP3 で注目をあびた Baker らによるフラグメントアセンブリ（Fragment Assembly, FA）[Simons 99] であろう．FA 法の基本的な考え方は，「新規フォールドであっても，その断片構造は既知構造データベースに存在するはずなので，それを組み合わせることにより新しい構造を作ることができる」というものである．FA 法のプロトコルは次の通りである．まず，予測したいタンパク質のフラグメント配列 (例えば 9 残基長) と似た配列をプロファイル比較法などを用いて既知構造データベースから探し，そのフラグメントがとりうる構造の候補を幾つか (例えば 20 個) 選ぶ．次に，そのフラグメント候補をつなぎ合わせることによって全体構造を構築し，その構造の妥当性を疎水性相互作用などのエネルギーによって評価する．Baker らはこの方法を用いて CASP3，4，5 の新規フォールド部門でダントツの成績を残した [Simons 99, Bonneau 01, Bradley 03].

現在では多数のグループが FA 法を用いており，*ab initio* 法のスタンダードになりつつあるが，フラグメント検索法とエネルギー関数の問題は依然として残されているため，改良の試みが続けられている [Takada 01, Chikenji 03, Fujitsuka 04, 朴 05a]．他にも，Skolnick や Kolinski らによる構造認識法と物理化学の手法を組み合わせた方法 [Zhang 03] や，Scheraga らによる物理化学的な方法 [Liwo 99] も近年注目を浴びている方法である．

4. CASP6

4·1 CASP6 の概要

§ 1 ターゲットの分類

CASP6 の最終的な評価は 64 個のターゲットを 90 個のドメインにわけて行われた．各ドメインは，進化的に関連する構造既知タンパク質がどのレベルで発見可能かによって分類される（表 2）.

類似配列検索やプロファイル比較によって有意なテンプレートの発見が可能なターゲットは，CM 法などの鋳型構造による予測手法が利用できる．この種の予測対象配列は，CM ターゲットあるいは FR/H ターゲットに分類される．一方，テンプレートの発見が困難な場合と存在しない場合はそれぞれ，FR/A ターゲットと NF ターゲットに区別され，*ab initio* 法が用いられる．

ターゲットのドメイン分割と難易度分類は評価者の裁量によるもので，特に FR/H と FR/A ターゲットに関しては激しい議論が交わされる．なぜなら，予測精度の評価がターゲット分類に従う部門ごとに行われるため，予測者にとっては成績に関わる重要な問題だからである．

表 2 CASP6 におけるターゲットの分類と数

BLAST*	●				
PSI-BLAST§	●	●			
プロファイル比較，または機能・立体構造類似性	●	●	●		
立体構造類似性	●	●	●	●	
新規フォールド					●
ターゲット分類	CM		FR		NF
	easy	hard	H	A	
	25	18	22	15	10

CM/easy: Comparative Modeling (easy)
CM/hard: Comparative Modeling (hard)
FR/H: Fold Recognition (Homologous)
FR/A: Fold Recognition (Analogous)
NF: New Fold
*§E-value< 0.01，§5 iterations
● はターゲットの鋳型が見つかるレベルを示している

§ 2 予測構造の評価

ターゲットの分類と並んで，予測構造の精度評価についての論争は絶えない．現状として，CASP に用いられる評価方法は，下式で定義される GDT_TS（Global Distance Test Total Score）である [Zemla 03].

$$\mathrm{GDT\_TS} = \frac{\mathrm{GDT\_P}_1 + \mathrm{GDT\_P}_2 + \mathrm{GDT\_P}_4 + \mathrm{GDT\_P}_8}{4}, \quad (1)$$

ここで $\mathrm{GDT\_P}_x$ は主鎖の C_α 原子が誤差 xÅ 以下で天然構造へ重なる予測配列の割合である

図 1 は，Model 1*[1]の GDT_TS における上位 10 チームのスコアから計算した平均 GDT_TS をターゲットごとにプロットしている．また，1 位の予測構造と天然構造との最小二乗距離（Root Mean Squared Distance, RMSD）を第 2Y 軸に表示している．全体的にターゲットの難易度と上位チームの予測精度は相関している．各部門には，相対的に難易度の高いターゲットが必ず存在し，NF ターゲットの予測精度は十分なレベルから程遠い．CASP の精度評価が部門ごとに行われている理由は，このような現状を反映している．

結果的に，“usual suspects”[Cozzetto 05] と呼ばれる常連は，やはり今回も上位 10 チームに名前を連ねている．特記すべき点は，CBRC-3D（産業技術総合研究所），Chimera（北里大学），Rokko（神戸大学），Rokky（神戸大学, `http://www.proteinsilico.org/rokky/`）がそ

*1 5 個までの予測構造のうち，最も予測者の確信が高い構造

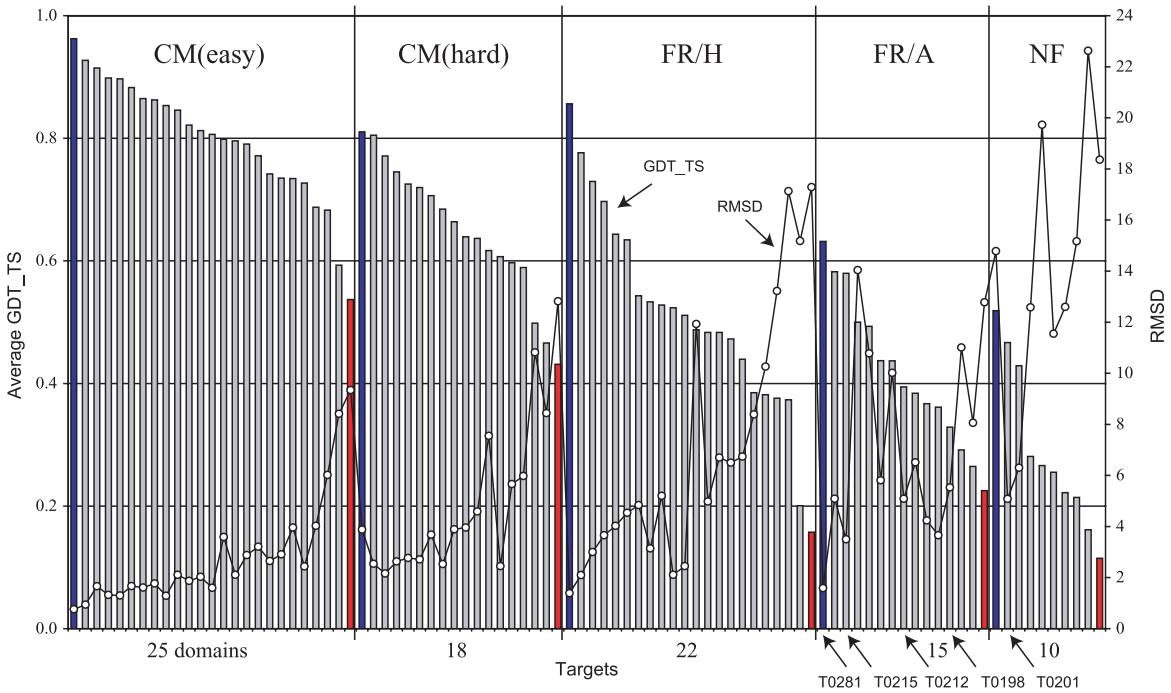

図 1 Model 1 における GDT_TS の上位 10 チームの平均 GDT_TS および 1 位の予測構造と天然構造との RMSD

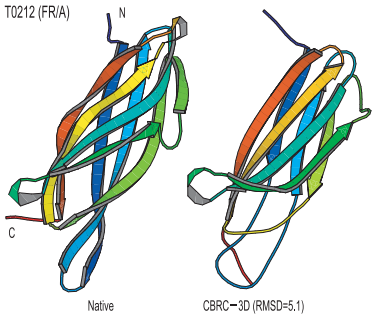

図 2 T0212 の天然構造と予測構造

れぞれの部門で好成績を収めたことである．また，日本からの参加チームは米国に次いで 2 番目に多く，構造ゲノムプロジェクト（理化学研究所）が 21 個のターゲットを提供するなど，CASP6 における日本の貢献度と成績は例年以上であった．

4·2 CM・FR/H 部門の動向

これらの部門のターゲットは，論理的には全て，鋳型構造を利用した立体構造予測が可能である．そのために，これまで様々な遠縁タンパク質検出法やフォールド認識法が提案されている．CASP5 から登場したメタサーバーが，一般に最も強力な鋳型発見のための手法の一つとなってきている．メタサーバーは，独立した複数のサーバーの予測から，コンセンサス予測を行うものである．コンセンサス予測には，ニューラルネットワークを用いる手法 Pcons[Lundstrom 01] や，協調アルゴリズムを用いた 3D-SHOTGUN[Fischer 03]，予測モデルの中で，できる限り重心構造に近いモデルを選択しようとする 3D-Jury[Ginalski 03] などがある．

CASP6 のこの部門において，Ginalski は群を抜いた成績を残した．彼は自身が構築した 3D-Jury システム以外に，更に二次構造予測の結果も利用するメタサーバー Meta-BASIC[Ginalski 04] も予測に用いていた．問題によっては，ターゲットの相同タンパク質についてもこうしたシステムを用いて予測を行うことで，有用な情報の蓄積をはかっていた．ただ，こうした予測システムの充実だけではなく，彼自身のタンパク質に対する深い造詣が，特にモデリングの段階での予測結果の向上に大きく寄与した部分もあろう．

FR/H 部門では，2 位の GeneSilico も独自のメタサーバーを予測に用いていた．しかし 3 位の CBRC-3D は，コンセンサス予測ではなく，独自に開発したプロファイル比較法 FORTE[Tomii 04] を用いて構築されたターゲットとテンプレートとのアラインメントに基づく網羅的モデル構築とそれらの評価を行うことで，メタサーバーの結果を利用した他の多くのグループよりも良好な結果を得ることができた．その典型例は，T0212（図 2）である．主要なメタサーバーは，依拠するサーバーの多くが正確なフォールド認識に失敗していたために，信頼度の低い誤った予測結果を与えていたのに対し，CBRC-3D のアプローチでは，正解である免疫グロブリンフォールドを認識することに成功し，このターゲットに対してもテンプレートに基づく立体構造予測を行うことができた．

4·3 FR/A・NF 部門の動向

CASP6 にみられる *ab initio* 法の潮流を，四つのターゲットに着目して見てみよう．

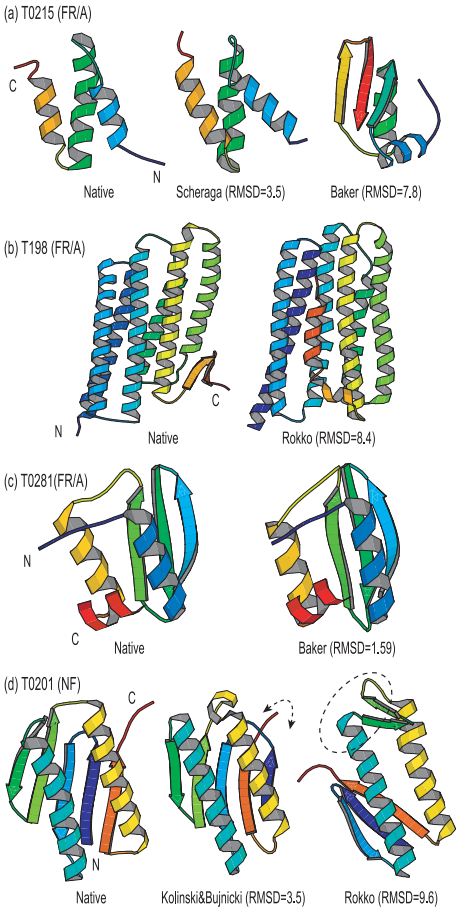

図 3 CASP6 における代表的な FR/A，NF ターゲットの天然構造と予測構造

§1 T0215（FR/A）

T0215 の天然構造（PDB[*2]コード 1X9B，図 3a）は，53 残基の小型 All α 膜タンパク質である．T0215 の予測に成功しているチームは，全原子ポテンシャルの分子動力学を用いた Scheraga である．Samudrala-AB も同様に，PDB 構造の統計情報を活用してモンテカルロ SA（Simulated Annealing）を行う “energy-based prediction” を用いて，天然構造との RMSD 5.0Å の構造を予測した．

一方，FA 法を用いる Baker と Rokko は二次構造レベルで大きく外れている．これは，ヘリックス部位のフラグメントライブラリがストランドへ偏っているためであり，二次構造予測の失敗がその原因である．

T0215 は膜と関係するタンパク質である．また，水溶性タンパク質用の二次構造予測ツールが得意とするヘリックス予測に失敗していることは興味深く，1X9B の論文出版が待ち遠しい．

§2 T0198（FR/A）

T0198 の天然構造（PDB コード 1SUM ，図 3b）は，225 残基の中型 All α タンパク質である．

T0198 の予測に成功してチームは FA 法による Baker と Rokko である．このことは，FA 法の技術的進歩を反映しており，簡単な形の中型タンパク質の予測が FA 法によって精度よく行えることを証明した．

§3 T0281（FR/A）

CASP6 の *ab initio* 部門において，最も驚くべき予測成果は Baker の T0281 である．

T0281 の天然構造（PDB コード 1WHZ，図 3c）は，分解能 1.52Å の X 線解析による 70 残基の小型 α/β タンパク質である．Baker は，多数の遠縁タンパク質についての大規模 FA シミュレーションと全自動予測サーバーの予測構造を分析して残基間コンタクト情報を取り出した．このような情報を T0281 予測の制約条件として用いて，天然構造との RMSD 2.2Å の高精度予測構造を組み立てた．その後，側鎖パッキングなどの全原子レベルの精密化を行って，RMSD 1.59Å のスーパークオリティの予測モデルを構築したのである．

§4 T0201（NF）

T0201 の天然構造（PDB コード 1S12，図 3d）は 94 残基の小型 α/β タンパク質である．

Kolinski&Bujnicki は，“FRankenstein’s monster” とよばれるメタアプローチ [Kosinski 03] と *ab initio* 予測サーバーの予測結果から残基間距離情報を解析した．その情報を拘束条件として CABS 力場 [Zhang 03] と高分解能（0.61Å）の格子モデルを用いて，RMSD 3.5Å の予測構造を構築した．この構造は，一つの β ペアリングだけが天然構造と異なっている（図中，矢印の部位）．

一方，FA 法は天然構造のような β シートを作るための良いループ型フラグメントが用意できず，予測に失敗している（図中，点線で示された部位）．

§5 *ab initio* 法の新展開

100 残基以下の小型タンパク質では幾つかのグループが *ab initio* 予測に成功している例が多い．中型 All α タンパク質には，FA 法による粗視化 *ab initio* 法が効果的である．複雑なトポロジーの予測は困難であり，バイオインフォマティクスを駆使して様々観点から予測配列を解析することが重要である．また，*ab initio* 予測の粗い構造は主鎖のずれが生じるため，側鎖構造の最適化と全原子ポテンシャルによる予測構造の精密化が必要である．

現状，All β タンパク質のような非局所的相互作用が多く含まれている立体構造の予測は非常に困難である．この問題への挑戦として， Baker は CASP6 で “Broken-chain FA” を提案した（図 4）．つまり，β ペアリングやコンタクト残基などの予測による拘束条件は，二面角を動かす FA 法では簡単に崩れる．そこで，拘束されている残基の周りにループ切断を入れて，条件を満たすようにしたのである．結果的に，T0212 の予測に成功（RMSD 6.2Å）するなど，幾つかのターゲットにおいてその有効性が証明されたが，うまく機能しなかったケースも多く，さらなる改善が望まれる．

*2 Protein Data Bank，`http://www.rcsb.org/pdb/`

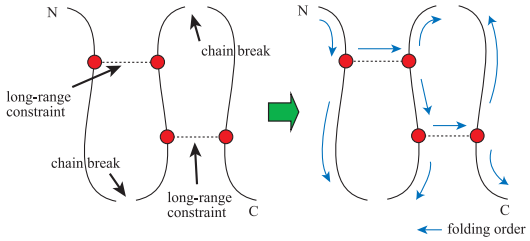

図 4 Broken-chain によるフラグメントアセンブリ法の概念図

5. 立体構造予測の課題

5·1 鋳型構造を用いた高精度モデリング

配列類似度が 60%以上のテンプレートが存在する場合，主鎖レベルで誤差 1Å のモデリングが可能となっている [Baker 01]．これは，X 線と NMR 解析による構造に匹敵するものである．類似度 30%–60%のテンプレートに基づくモデリングの誤差は，およそ 1Å–2Å である．30%以下のテンプレートの場合，精度は急激に悪くなり，テンプレートそのものが間違っている場合もある．

精密な構造モデリングには，側鎖構造の配向や配列アラインメント，ループモデリングの改善はもちろん，信頼性の高い立体構造評価法の開発・改善が欠かせない．NIH（National Institutes of Health）は，2005 年から X 線結晶構造に相当する高精度立体構造モデリングへむけて “High-accuracy Protein Structure Modeling” プロジェクトを 3 年計画（年間 75 万ドル規模）で進めており，このような状況の打開を図っている．

5·2 *ab initio* 法による基礎科学

ab initio 法による予測は，本質的に難しい問題であり，X 線結晶構造レベルの高精度を期待するのは，やや時期尚早である．それでも，現時点で *ab initio* 法が基礎科学として役に立つ方向性が幾つか存在する．

まず，配列類似性検索によって鋳型構造を容易に決定できない場合に，*ab initio* 法によるシミュレーションから既知構造トポロジーが予測される場合がしばしばある．これは CASP の FR/A に属する場合であり，ここではシミュレーションによって，鋳型構造を認識できることになる．実際，FR/A の多くのターゲットに対して，原理的にはフォールド認識によって PDB 構造から鋳型が選べるはずだが，実際にはそれは難しく，むしろ *ab initio* 法によって鋳型構造に近い予測構造が得られる．フォールドが決まれば，その鋳型タンパク質の機能と類似機能をもつと推定できるので，未知タンパク質の機能アノテーションに役立つことになる [Bonneau 02]．

より基礎的に，*ab initio* 法は，実験的に未知な構造を調べる唯一の方法である．例えば，新規フォールドをもつ人工タンパク質設計のためには，配列設計のアルゴリズムと *ab initio* 構造予測法の融合が不可欠であった [Kuhlman 03]．また，*ab initio* 法は，実験的には知ることが難しいタンパク質のアンフォールドした構造集団を作り出すことができ，それは折りたたみ機構の解析 [Chikenji 04] および，天然状態でアンフォールドしたタンパク質の構造集団を推測するのに貴重である．さらに，鋳型構造が極めて少ない膜貫通型タンパク質の構造モデリングにも応用が期待される．

6. おわりに

利用可能な技術を全て活用することが立体構造予測の真髄である．原子レベルから粗視化レベルまで入手可能なツールと方法を駆使して，知見と経験を生かして，あり得る可能性を全て試してみることが肝心である．そして，成功と失敗を学んで次の挑戦へつなげることが重要であり，それが CASP 精神である．CASP に成功している研究者は過去 10 年間，この精神に徹した．その結果，解決すべき問題ととるべき接近法が明らかになり，予測問題の核心へ徐々に迫っている．

とるべき接近法の一つとして，CASP6 で見られるように「立体構造予測のための予測」を行うことがあげられる．つまり，最初に二次構造や残基間コンタクトなどを予測し，その結果を充分に利用して全体構造の予測を行うことによって，予測精度の向上が期待できるということである．そのうえ，コンタクト予測情報などは比較モデリングの評価関数と *ab initio* 法のエネルギー関数の改良に密接に関係し，構造空間の効率的な探索方法設計に有益な指針を与える．

タンパク質立体構造予測は科学的重要性と実用性を増してきている．鋳型構造による構造予測は（良い鋳型が見つかる場合には），実験的構造に匹敵する構造モデルを作ることができ，構造ゲノム科学において欠かせない技術になっている．また，粗視化 *ab initio* 法は，まだ確度・精度ともに未熟ではあるが，タンパク質構造構築原理の理解にむけて重要な技術として，さらに発展するものと期待する．

◇ 参 考 文 献 ◇

[Altschul 97] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Miller, W., and Lipman, D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402 (1997)

[Baker 01] Baker, D. and Sali, A.: Protein structure prediction and structural genomics, *Science*, Vol. 294, pp. 93–96 (2001)

[Bonneau 01] Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M., and Baker, D.: Rosetta in CASP4: Progress in ab initio protein structure prediction, *Proteins*, Vol. 45, pp. 119–126 (2001)

[Bonneau 02] Bonneau, R., Rohl, C. E. S. C. A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., and Baker, D.: De novo prediction of three-dimensional structures for major protein families, *J. Mol. Biol.*, Vol. 322, pp. 65–78 (2002)

[Bradley 03] Bradley, P., Chivian, D., Meiler, J.,

Misura, K. M., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C. E., and Baker, D.: Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation, *Proteins*, Vol. 53, pp. 457–468 (2003)

[Chikenji 03] Chikenji, G., Fujitsuka, Y., and Takada, S.: A reversible fragment assembly method for de novo protein structure prediction, *J. Chem. Phys.*, Vol. 119, pp. 6895–6903 (2003)

[Chikenji 04] Chikenji, G., Fujitsuka, Y., and Takada, S.: Protein folding mechanisms and energy landscape of src SH3 domain studied by a structure prediction toolbox, *Chemical Phys.*, Vol. 307, pp. 99–109 (2004)

[Chothia 92] Chothia, C.: One thousand families for the molecular biologist, *Nature*, Vol. 357, pp. 543–544 (1992)

[Cozzetto 05] Cozzetto, D., Matteo, A. D., and Tramontano, A.: Ten years of predictions ... and counting, *FEBS Journal*, Vol. 272, pp. 881–882 (2005)

[Fischer 03] Fischer, D.: 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor, *Proteins*, Vol. 51, pp. 434–441 (2003)

[Fiser 00] Fiser, A., Do, R. K., and Sali, A.: Modeling of loops in protein structures, *Protein Sci.*, Vol. 9, pp. 1753–1773 (2000)

[Fujitsuka 04] Fujitsuka, Y., Takada, S., Luthey-Schulten, Z. A., and Wolynes, P. G.: Optimizing physical energy functions for protein folding, *Proteins*, Vol. 54, pp. 88–103 (2004)

[Ginalski 03] Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L.: 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics*, Vol. 19, pp. 1015–1018 (2003)

[Ginalski 04] Ginalski, K., Grotthuss, von M., Grishin, N. V., and Rychlewski, L.: Detecting distant homology with Meta-BASIC, *Nucleic Acids Res.*, Vol. 32, pp. W576–581 (2004)

[Kolodny 02] Kolodny, R., Koehl, R., Guibas, L., and Levitt, M.: Small libraries of protein fragments model native protein structures accurately, *J. Mol. Biol.*, Vol. 323, pp. 297–307 (2002)

[Kosinski 03] Kosinski, J., Cymerman, I. A., Feder, M., Kurowski, M. A., Sasin, J. M., and Bujnicki, J. M.: A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation, *Proteins*, Vol. 53, pp. 369–379 (2003)

[Kuhlman 03] Kuhlman, B., Dantas, B., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D.: Design of a novel globular protein fold with atomic-level accuracy, *Science*, Vol. 302, (2003)

[Liu 04] Liu, X. and Wang, W.: The number of protein folds and their distribution over families in nature, *Proteins*, Vol. 54, pp. 491–499 (2004)

[Liwo 99] Liwo, A., Lee, J. S., Ripoll, D. R., Pillardy, J., and Scheraga, H. A.: Protein structure prediction by global optimization of a potential energy function, *Proc. Natl. Acad. Sci. USA.* (1999)

[Lundstrom 01] Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A.: Pcons: a neural-network-based consensus predictor that improves fold recognition, *Protein Sci.*, Vol. 10, pp. 2354–2362 (2001)

[Schaffer 99] Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., and Altschul, S. F.: IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices, *Bioinformatics*, Vol. 15, pp. 1000–1011 (1999)

[Simons 99] Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D.: Ab initio protein structure prediction of CASP III targets using ROSETTA, *Proteins*, Vol. 37, pp. 171–176 (1999)

[Takada 01] Takada, S.: Protein folding simulation with solvent-induced force field: folding pathway ensemble of three-helix-bundle proteins, *Proteins*, Vol. 42, pp. 85–98 (2001)

[Tomii 04] Tomii, K. and Akiyama, Y.: FORTE: a profile-profile comparison tool for protein fold recognition, *Bioinformatics*, Vol. 20, pp. 594–595 (2004)

[Tramontano 03] Tramontano, A. and Morea, V.: Assessment of homology-based predictions in CASP5, *Proteins*, Vol. 53, pp. 352–368 (2003)

[Wang 04] Wang, G. and Dunbrack, R. L.: Scoring profile-to-profile sequence alignments, *Protein Sci.*, Vol. 13, pp. 1612–1626 (2004)

[Zemla 03] Zemla, A.: LGA: a method for finding 3D similarities in protein structures, *Nucleic Acids Res.*, Vol. 31, pp. 3370–3374 (2003)

[Zhang 03] Zhang, Y., Kolinski, A., and Skolnick, J.: TOUCHSTONE II: a new approach to ab initio protein structure prediction, *Biophysics J.*, Vol. 85, pp. 1145–1164 (2003)

[朴 05a] 朴 聖俊, 高田 彰二：アミノ酸配列情報から蛋白質立体構造予測へ向けて：配列プロファイル比較に基づく部分立体構造ライブラリの構築, SICE 第 32 回知能システムシンポジウム, pp. 289–294 (2005)

[朴 05b] 朴 聖俊, 高田 彰二, 山村 雅幸：実数値 GA によるタンパク質立体構造の 2 層比較, 情報処理学会論文誌, Vol. 43, pp. 898–910 (2005)

著 者 紹 介

朴　聖俊（正会員）

1998 年専修大学経営学部卒業，2005 年東京工業大学大学院総合理工学研究科知能システム科学専攻修了．博士（工学）．2005 年 4 月より神戸大学理学部学術研究員，現在に至る．進化型計算，バイオインフォマティクス，立体構造予測の研究に従事．情報処理学会，日本バイオインフォマティクス学会，日本分子生物学会，各会員．

千見寺　浄慈

1997 年東京都立大学理学部卒業，2002 年大阪大学大学院理学研究科物理学専攻修了．理学博士．2002 年日本学術振興会特別研究員，2005 年 4 月より神戸大学理学部学術研究員，現在に至る．タンパク質のフォールディング，立体構造予測の研究に従事．日本生物物理学会，日本蛋白質科学会，各会員．

広川　貴次

1998 年東京農工大学大学院工学研究科修了．工学博士．（株）菱化システム科学技術計算部での勤務を経て 2001 年に産業技術総合研究所生命情報科学研究センターに入所．2003 年より同センター分子設計チーム長．2003 年 4 月より東京医科歯科大学疾患生命科学研究部客員助教授を併任．現在，受容体タンパク質を対象としたドラッグデザインへの応用研究に従事．

富井　健太郎

1998 年京都大学大学院理学研究科生物物理学専攻修了．理学博士．1998 年生物分子工学研究所ポスドク，2000 年 University of California, Berkeley ポスドク，2001 年産業技術総合研究所生命情報科学研究センター研究員，現在に至る．タンパク質立体構造予測の研究に従事．日本生物物理学会会員

高田　彰二

1988 年京都大学理学部卒業．1990 年同大学院理学研究科化学専攻修士修了．1991 年岡崎国立共同研究機構技官．理学博士．1995 年日本学術振興会研究員，1998 年神戸大学理学部講師を経て，2001 年より同大学助教授，現在に至る．生物物理，理論的タンパク質構造と機能解析，タンパク質立体構造予測の研究に従事．