蛋白質構造のデノボ予測のための
フラグメント整合スコアの開発

CETIN HIKMET

蛋白質構造のデノボ予測のための
フラグメント整合スコアの開発

CETIN HIKMET

博士論文

2012 年 1 月

名古屋大学大学院工学研究科

計算理工学専攻

謝辞

この博士論文の謝辞を書くに当たって、第一の言葉として残しておきたいのが、「全人類と共に私をも一滴の汚い水から創り上げ、この世の中に送ってくださった在りて在るもの・全知全能の慈悲遍く万有の王・唯一成るアッラーに、彼からの預け物にしかならない命と共に感謝の気持ちを捧げる」という言葉です。

そして、アッラーのみつかい様、アダム・アブラハム・モーゼ・イェス・モハーマド預言者様たち（彼らに平安あれ）にも、信仰の光をこの世に齎し私が生きる支えをつくって下さったことを深く感謝いたします。

また、私を生みそして育てたお母さんや私のはるか遠い日本での留学の希望を快くサポートしてくれたお父さん、そして現在トルコにいる家族の皆にも、私を常にあたたかい気持ちにさせてくれたことを感謝します。

そして、何よりも私を研究室に迎えて、大変面倒をみて、いつも指導してくれた笹井理生先生に心から感謝いたします。笹井先生のお陰で、研究テーマを決められ、研究計画をつくり、学術論文を投稿でき、博士論文を書く事ができました。副査の長岡正隆先生と美宅成樹先生にも私のプレゼンテーションを聞いて、博士論文を直していただいた事に対して心から謝辞しております。

そして、この研究の応用に使ったデーターを提供してくださって、つまったときにいつもたすけてくれた佐々木尚先生に特に心から感謝します。また私のぼろぼろのプレゼンテーションを修正し読んでおくべき論文を教えてくれた寺田智樹講師、Ｃ言語についてなど多くを教えてくれた西村信一郎先輩、コンピューターの様々な設定やネットワークやインターネットに関して指導してくれた長尾知生子先輩、ソフトの使い方やネットの検索など細かいところで助けてくれた堀田剛史先輩、事務的なことや生物物理の基礎的な学習について色々相談にのってくれた岡部ゆりえさんをはじめ、笹井研究室の皆様に心から感謝いたします。

また、工学研究科で製薬の研究をしている友達　ＭＵＨＡＭＭＥＴ　ＵＹＡＮＩＫ助教や国際開発研究科を修了した友達ＥＭＲＥ　ＭＥＲＣＡＮさんに、夕食などに付き合って食事をおごってくれたり、いつも一緒にいてくれたことを深く感謝します。

目次

第Ⅳ章　予測の難しい TBM 構造への応用

序章

現在、多くの生物のゲノム配列解読が完了しており、技術革新によって、ヒトゲノムの塩基配列を短時間で読み取ることも可能になった。また、こうして得られたデーターベースを使って、塩基配列をアミノ酸配列に翻訳すれば、ヒトが生産しうるすべての蛋白質のアミノ酸配列を推定することができるようになるであろう。

しかしこれだけでは、ヒトが持つすべての蛋白質の機能を理解するという目的を達成することはできない。なぜならば、自然界に現存する蛋白質は、密にパックされ２次構造などの局所構造を含んである特定の３次元構造をとってはじめて機能を発現していることが知られているため、蛋白質の立体構造を解明することが、その機能を理解する上で不可欠だからである。この目的を達成するための立体構造解明の研究を、主に「実験的な手法」と「理論的な手法」の 2 つに分けることが出来る。実験的な手法による構造ゲノム研究は盛んに行われており [1,2]、特に、ポストゲノムプロジェクトでは「すべての蛋白質の構造を NMR によって、又は、精製・結晶化し X 線結晶解析によって解こう」という目標に向かって努力が行われている。しかし、蛋白質一個の構造の全解析には一年以上の歳月がかかるケースが多く、生体内には多種多様な蛋白質が存在していて、膜蛋白質など結晶化が非常に難しいものもある。従って、実験的に全蛋白質の構造を解析するには、かかる時間を初めとして、多くの問題点がある [3]。

理論的な手法は、所謂、アミノ酸配列を基に計算機を使った蛋白質の構造予測をする方法であり、この方法もまた更に「分子動力学（Molecular Dynamics, MD）などによる物理的な方法」と「構造比較などによるバイオインフォマティックな方法」という 2 つのグループに分けられる。しかしどちらにせよ、1 次元（1D）情報であるアミノ酸配列から直接、全原子の 3 次元（3D）の立体構造を予測することは非常に困難である。分子量の小さな蛋白質においても、可能な立体構造のバリエーションが膨大に存在することがその理由である。

従って、アミノ酸残基から蛋白質の構造を予測する事は理論的な生命科学研究における主たる挑戦である。ターゲット蛋白質の配列が既知構造の蛋白質の配列に類似する時は、それらの既知構造をテンプレートとしてターゲット蛋白質の未知構造をモデリングするために使用できる。テンプレートを基準にしたモデリング（Template Based Modeling, TBM）の方法は大幅に進歩しているが [4, 5]、的確なテンプレートを有しないターゲット蛋白質に関しては、一貫性がある体系的な方法は未だに確立されていない [6]。後者の問題はデノボ予測、又はフリー・モデリング（Free Modeling, FM）と呼ばれ、TBM より困難な問題であるが、蛋白質構造構築の規則を理解する [7, 8]という構造生物

学のより大きな問題の応用にとって、重要な問題である。

まず、過去の研究の流れを振り返って、本研究における新しい方法の位置づけについて述べておきたい。上述したように、これまで「MD 法などによる物理的な計算」と「バイオインフォマティックな構造評価法」に基づいた様々な研究がなされたが、その全てにおいて、1D のアミノ酸配列から 3D の蛋白質立体構造にたどり着く最終的なゴールには、およそ到達できていなかった。Chikenji ら[9] のような物理的な方法では、構造の多数の候補を作成することができるものの、その中からより良い構造を選び出す指標を物理的な方法と伴に用いれば、さらに構造予測の能力を高めることができると思われる。実際、候補構造と正解構造の近さと相関を持った指標を開発する Sasaki ら[10]のような研究が進んでいる。また、Chistopher ら[11]のような研究では、成果は活性部位のみのシミュレーションが成功しているに留まっている。一方の「バイオインフォマティックな構造評価法」では、既知のネイティブ構造から経験則を抽出し、これを適用して未知の構造を予測をする。こうした研究は下記のようにグループ分けができる

A）1D アミノ酸配列から 2 次構造を予測する。このタイプの研究は一番進んでおり、Clark ら[12]のように高い予測率を得ることが出来るが、2 次構造予測は 3D 立体構造予測の途中の目標に過ぎない。また、Clark ら[12]において 2 次構造予測のために、残基ごとの「構造密度」（詳しくは第 1 章で述べる）などの、3D 立体構造の特徴が利用されていることに注意が必要である。

B）2D-3D 法では、2 次構造から 3D 立体構造や 3D 立体構造の「構造密度」などの特徴を予測する。このタイプの研究としては、2 次構造からニューラル・ネットワークを用いて「構造密度」の実測値を予測する研究 [13] や、Clark ら[12]のような 2 次構造予測の研究を踏み台にし、粗視化された「構造密度」を用いて更に 3D 立体構造により近い概念である「フォールド」を予測する Hargbo ら[14]のような研究が良い例である。

C）3D-1D 法では、3D 立体構造を 1D の情報に置換し、それをアミノ酸配列と比較して評価する。Bowie ら[15]の研究では、ネイティブ構造のフォールド状況を調べ、アミノ酸配列との間の相関を求めている。Etchebest ら[16]の研究では、アミノ酸残基の一個一個が取り得る立体構造のクラスを定義し、与えられたアミノ酸配列の残基それぞれが入るクラスを予測した。Rice ら[17]の研究では 2 次構造や「構造密度」やフォールドも使って更に複雑な「クラス分け」を行っている。

それ以外にも、蛋白質構造予測に直接関係はないが、3D 構造同士の類似性を調べたり [18, 19]、未知フォールドを使用して 1D アミノ酸配列と 3D 立体構造の関係を表す簡単なモデルを作る研究[20]がなされている。

Gallicchio ら[21]の研究において、MD 法と「構造密度」の計算の併用が用いられていることを参考にすると、MD 法によって生成された候補構造をバイオインフォマティックな比較法によって選

ぶことができると考え、本研究では、候補構造からよりよい構造を選び出すモデル品質評価（Model Quality Assessment, MQA 又は QA）法の開発を試みた。Sasaki ら[9]の MD 法で求められた候補構造の品質を評価するために、Etchebest,および Rice ら[16,17]のように、残基一個一個を「クラス分け」する手法を新たに導入した。そして、そのために、Adamczak ら[12]のように先に 2 次構造を定義し、Garg ら[13]のように「構造密度」を 計算し、Hargbo ら[14]のような粗視化を導入して、Rice ら[17]のように各残基を 3D 立体構造クラスに分けたが、更に進んで、クラス分けから「クラスに基づく構造配列の作成」を新たに定義した。また、今まで蛋白質全体で定義されていた「コンタクト・オーダー」[22]という概念をアミノ酸残基ごとに新たに「LCO」として定義し、クラスに基づく構造配列をより実用的にすることを目指した改良を行った。

第 I 章では、デノボ予測手法を進展させるために、上記のクラス構造配列の考え方をもとに、さらにフラグメントの方法を導入したフラグメント整合スコア（Fragment-based Consistency Score, FCS）法を提案し、テストする。本研究では、主にデノボ予測問題に集中するために、既知構造配列とターゲット配列とのホモロジー関係性に頼らないスコア関数を開発した。しかし、ターゲットと他の構造既知蛋白質の間のホモロジーが見出せない FM 問題でも、多数の蛋白質の中の短いフラグメントの間に見いだされるローカルな配列－構造関係が、構造を推測する上で役立つはすである。FM ターゲットのデノボ予測に於いて、この点は、フラグメントアセンブリ法 [8, 23-26]、あるいは様々な長さのローカルな構造パーツを集める方法 [27-29]の成功の実例によって、最も明確に示された。ここで最も重要な特徴は、ローカルな 1D-3D 関係を推測するために、残基 1 個を使う代わりに、有限な長さのフラグメントの集合を用いる事である。これらのデノボ予測技法と同様に、MQA 法も、各残基周辺の有限な長さのフラグメントを比較する事によって改善されるはずである。MQA のためのスコア関数を導くために、フラグメント比較を有効に使った研究が既にあるが [30]、本論文の中では 1D-3D 関係を導くために、フラグメント構造のみならずフラグメント周辺のローカルな構造環境も使用した。ここでは、フラグメント整合スコア関数を計算するために Cα座標のみを使用しており、各残基の側鎖の原子配置に関する情報は使われていない事に注目してもらいたい。

この方法は、VERIFY3D [31]と関連する手法 [53,62]による 1D-3D 法の直接的な拡張版とみなすこともできる。モデルの各残基に関して、ローカルな構造環境を構造指標で表したクラスに分類し、本論文で構造配列とよぶ構造指標の配列によって、蛋白質構造を表した。各残基の周りで、Position Specific Iterative (PSI)-BLAST [32]によって作られた配列プロファイルを考慮する事によって、重複しない（Non-Redundant, NR）蛋白質構造のライブラリーから 9 残基の長さのフラグメントを選び、フラグメントの集まりの中における構造配列の出現数を数える事によって、モデル構造フラグメント整合スコア（FCS）を評価した。第 I 章はこの方法論を解説する。「構造密度」や「構造配列」

や「LCO」を定義し、更にそれらの概念を使って実際に計算する際の、具体的な計算法における相違点を述べる。

第 II 章では、FCS 法とランジュバン MD 法を併用すれば、ランジュバン MD 法が作り出す構造候補を有効に選別できることを説明する。デノボ予測のひとつの戦略は、フォールディング過程をシミュレートするモンテカルロ法 [8,33-34] 又は、ランジュバン MD 法 [35-37,9]を使用する戦略である。物理的フォールディング過程をシミュレートすることによる利点は、幾つか存在する。1 つ目には、予測問題のために開発された方法がフォールディング過程に対する洞察を与えるという利点を挙げることができる。2 つ目には、予測問題以外でも、開発した方法を蛋白質の機能に於ける大規模な構造変化に応用できる可能性が期待できる。更には、重要な事であるが、自然の中に存在する構造の生成過程を真似する事は、複雑な構造の決定の合理的な手法であろうと予想されることを強調しておきたい。そこで第 II 章では 、ランジュバン MD による構造候補の作成[9]という物理的な方法と、既知構造の情報をデーターベースとして使用し、これらの構造候補の中で実際の構造に近いものを選択的に見分けるバイオインフォマティックな FCS 法を合併させた。

第 III 章では、第 8 回目の Critical Assesment of techniques for protein Structure Prediction（CASP8）に提出された、10 個の FM ターゲット蛋白質ドメインに対する全てのサーバーモデル構造を評価するために、FCS 法を応用した結果を説明する。

デノボ予測技法を進展させるために役立つ手法の一つが、MQA 法 [38]である。MQA はターゲットの正しい解答構造を知る前に、予測されたモデル構造と正しいターゲット構造の類似の程度を見積るという問題である。モデル構造が多数提案されて利用可能な時には、それらの中からもっと良い候補を選別するのに適切な MQA が役立つはすであり、それが予測の能力を向上させるはずである。今日、MQA の重要性は広く認識される様になり、第 7 回目の Critical Assesment of techniques for protein Structure Prediction(CASP7) [39] 以来、多数の MQA 法が提案され、比較された [40]。CASP における QA カテゴリーに提出された MQA 法の解析を通して、多数の異なる方法で予測されたモデル構造の間の共通の特徴を抽出し、モデル構造がそうした共通の特徴を多く持つほどよい構造として評価する、コンセンサスに基づく方法 [41-49,30] がコンセンサスに基づかない方法より大幅に良い結果を出す事が示された [40,50-51]。これは、おそらく、違う予測法によって生成されたモデルのエラーがノイズとして処理され、多数のモデルの比較を通して除かれたからである。しかしながら、コンセンサスに基づく方法はそれ自身の限度を持っている。予測法の多数派が失敗していくつか例外的な方法が良い予測を行った時に、コンセンサスに基づく MQA は、提案されたモデルの集合から良いモデルを選別することに失敗する。この状況は特に、FM 問題でよく遭遇する状況である、特に、多数の予測法がターゲットとテンプレートの間の配列アラインメントに頼っているが、

正しいターゲット構造がどのテンプレートとも類似しない時に、コンセンサスに基づく MQA 法は正しい選別ができない。更に、第Ⅱ章で述べた粗視化されたランジュバン分子動力学法の様な、一つ又は少数の予測法のみ使える時は、同じ方法を通して生成されて多数のモデルが、同じく偏ったエラーを持つはずである。このとき、コンセンサスに基づいた方法は多数の違う予測サーバーのモデルが使える CASP でのそれらの活躍の様に、よく機能しないかもしれない。これらの理由から、コンセンサス解析に頼らない MQA 法を開発する価値がある。こうして開発されたコンセンサスに基づかない方法と他のコンセンサスに基づく方法の併用 [44] によって、デノボ予測手法を進展させる更に良い MQA 法を提供できる事が期待される。コンセンサスに基づかない方法は、文献の中で頻繁にシングル・モデル法と呼ばれるが [40,46-48,52]、我々は各モデル構造の品質を評価する方法の性能を試しているのではなく、主に一つの予測法によって生成された多数の構造から良いモデルを選別するのにどうやってコンセンサスに基づかない方法が機能するのかに注目するので、我々はここでその呼び方を避ける。これまで、1D-3D 相関 [31,53] についての既知の性質から導かれたスコア関数や、残基分布や残基ペアのコンタクト形成 [54-57] や、配列アラインメントによる蛋白質間のホモロジーの活用[46,58]などを用いる、多種多様なコンセンサスに基づかない MQA 法が開発された。第 1 章で提案され、テストされた FCS 法は、コンセンサスに基づかない新しい MQA 法である。第 III 章において、FCS 法を CASP8 に参加した他のコンセンサスに基づかない MQA 法の成績と比較した結果、FCS 法はそのうち有力な MQA 法とほぼ同等の能力を持っていることを示すことができた。

第 IV 章では、FCS 法の応用範囲を FM ターゲット蛋白質ドメイン以外にも拡げる。CASP8 に出題されたターゲット蛋白質のうち、とくに興味深いのは、アミノ酸配列にはわずかの違いしか持たないが、構造が大きく違う２つの人工的に設計された蛋白質である T0498 と T0499 である。これらの蛋白質とホモロジーの関係にある既知構造が存在するため、CASP8 ではこれらの蛋白質は TBM ターゲットとして扱われた。しかし、既知構造とのホモロジーを使った予測法は、配列の似ている T0498 と T0499 に対して同じ構造を予測するため、２つの大きな構造の違いを予測できない。従って、大多数のサーバーモデル構造が誤ったために、コンセンサス法はうまく働かなかった。コンセンサスに基づかない FCS 法は、T0498 と T0499 の構造を識別できる。

又、TBM ターゲットの構造予測における最も難しい問題は、正規な 2 次構造に分類できない不規則なループやコイル構造をとる部分の予測であり、フォールディングを行うチーム [59]でもテンプレートを使うチーム [60,61] でも、こうした部分については予測を不得意としている。FCS 法はこの問題においても、ある程度の予測の性能を示した。

引用文献

1. Zhou Y. F, Mi W, Li L: Protein preparation, crystallization and preliminary X-ray crystallographic analysis of Smu. 1475c from caries pathogen Streptococcus mutants, *Biochim Biophys Acta*. 1764:324-326 (2005).
2. Flemming H. D, Gorelsky S. I, Sarangi R: Reinvestigation of the method used to map the electronic structure of blue copper proteins by NMR relaxation, *Biol Inorg Chem*. 11:277-285 (2006).
3. Schlenker O, Hendricks A, Sinning I, Wild K: The Structure of the Mammalian Signal Recognition Particle(SRP) Receptor as Prototype for the Interaction of Small GTPases with Longin Domains, *Biol Chem*. 281:8898-8906 (2006).
4. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: Improving physical realism, stereochemistry and side-chain accuracy in homology modeling Four approaches that performed well in CASP8, *Proteins: Struct. Funct. Bioinform*. 77:114-122 (2009).
5. Peng J, Xu J: A multiple-template approach to protein threading, *Proteins: Struct. Funct. Bioinform*. 79:1930–1939 (2011).
6. Jauch R, Yeo H. C, Kolatkar P. R, Clarke N. D: Assessment of CASP7 structure predictions for template free targets, *Proteins*. 69:57-67 (2007).
7. Ben-David M, Noivirt-Birk O, Paz A, Prilusky J, Sussman J L, Levy Y: Assessment of CASP8 structure predictions for template free targets, *Proteins: Struct. Funct. Bioinform*. 77:50-65 (2009).
8. Chikenji G, Fujitsuka Y, Takada S: Shaping up the protein folding funnel by local interaction lesson from a structure prediction study, *Proc. Natl. Acad. Sci. USA*. 103:3141-3146 (2006).
9. Sasaki T. N, Sasai M: A coarse-grained Langevin molecular dynamics approach to protein structure reproduction, *Chemical Physics Letters*. 402:102−106 (2005).
10. Chistopher M S, Michael L, William F. D: An Atomic environment Potential for use in Protein Structure Prediction, *J.Mol.Biol*. 352:986-1001 (2005).
11. Clark M, Guarnieri F, Shkurko I, Wiseman J: Grand Canonical Monte Carlo Simulation of Ligand-Protein Binding, *Chem Inf Model*. 46:231-242 (2006).
12. Adamczak R, Porollo A, Meller J: Combining Prediction of Secondary and Solvent Accessibility in Proteins, *PROTEINS:Structure, Functions and Bioinformatics*. 59:467-475 (2005).
13. Garg A, Kaur H, Raghava G.P.S: Real Value Prediction of Solvent Accessibility in Proteins Using Multiple Sequence Alignment and Secondary Structure, *PROTEINS:Structure, Functions and Bioinformatics*. 61:318-324 (2005).
14. Hargbo J, Elofsson A: Hidden Markov Models That Use Predicted Secondary Structures For Fold Recognition, *PROTEINS:Structure, Functions and Genetics*. 36:68-76 (1999).
15. Bowie JU, Luthy R, Eisenberg D: A method to identify protein sequences that fold into a known three-dimensional structure, *Science*. 253:164-170 (1991).
16. Etchebest C, Benros C, Hazout S, De Breven A.G.: A Structural alphabet for Local Protein Structures Improved Prediction Methods, *PROTEINS:Structure, Functions and Bioinformatics*. 59:810-827 (2005).
17. Rice D.W, Eisenberg D: A 3D-1D substitution Matrix for Protein Fold Recognition that includes Predicted Secondary structure of the Sequence, *J.Mol.Biol*. 267:1026-1038 (1997).
18. Burbaum JJ, Starzyk RM, Schimmel P: Understanding structural relationships in proteins of unsolved three-dimensional structure, *Proteins*.7:99-111 (1990).
19. Sippl MJ, Weitckus S: Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations, *Proteins*. 13:258-271 (1992).
20. Fischer D: Modelling three-dimensional protein structures for amino acid sequences of the CASP3 experiment using sequence-derived predictions, *Proteins*. 3:61-65 (1999).
21. Gallicchio E, Levy R.M: AGBNP an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling, *Comput Chem*. 25:479-499 (2004).
22. Kinjo A.R, Horimoto K, Nishikawa K: AGBNP an analytic implicit solvent model suitable for

molecular dynamics simulations and high-resolution modeling, *PROTEINS:Structure,Functions and Bioinformatics*. 58:158-165 (2005).
23. Rohl C.A, Strauss C.E.M, Misura K.M.S, Baker D: Protein structure prediction using Rosetta, *Methods Enzymol*. 383:66–93 (2004).
24. Bradley P, Misura K.M.S, Baker D: Toward high-resolution de novo structure prediction for small proteins, *Science*. 309:1868–1871 (2005).
25. Lee J, Kim S.Y, Lee J: Protein structure prediction based on fragment assembly and parameter optimization, *Biophys. Chem*. 115:209–214 (2005).
26. Fujitsuka Y, Chikenji G, Takada S: SimFold energy function for de novo protein structure prediction: consensus with Rosetta, *Proteins: Struct. Funct. Bioinform*. 62:381–398 (2006).
27. Zhang Y, Arakaki A.K, Skolnick J: TASSER: an automated method for the prediction of protein tertiary structures in CASP6, *Proteins: Struct. Funct. Bioinform*. 61:91–98 (2005).
28. Zhou H, Skolnick J: Ab initio protein structure prediction using Chunk-TASSER, *Biophys J*. 93:1510-1518 (2007).
29. Wu S, Skolnick J, Zhang Y: Ab initio modeling of small proteins by iterative TASSER simulations, *BMC Biol*. 5:page number not for citation purpose (2007).
30. Zhou H, Skolnick J: Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential, *Proteins: Struct. Funct. Bioinform*. 71:1211–1218 (2008).
31. Eisenberg D, Lothy R, Bowie J: VERIFY3D: Assessment of protein models with three-dimensional profiles, *Methods Enzymol*. 277:396-404 (1997).
32. Altschul S.F, Madden T.L, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman D.J: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*. 25:3389-3402 (1997).
33. Kazmierkiewicz R, Liwo A, Scheraga H.A.: Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte-Carlo method, *J. Comput. Chem*. 23:715-723 (2002).
34. Nanias M, Chinchio M, Oldziej S, Czaplewski C, Sheraga H.A: Protein structure prediction with the UNRES force-field using replica-exchange monte carlo-with-minimization; comparison with MCM, CSA and CFMC, *J. Comput. Chem*. 26:1472-1486 (2005).
35. Liwo A, Khalili M, Scheraga H.A: Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains, *Proc. Natl. Acad. Sci. USA*. 102:2362-2367 (2005).
36. Papoian G..A, Ulander J, Eastwood M.P, Z. Luthey-Schulten, P. G. Wolynes: Water in protein structure prediction, *Proc. Natl. Acad. Sci. USA*. 101:3352-3357 (2004).
37. C. Hardin, M. P. Eastwood, Luthey-Schulten Z, Wolynes P.G.: Associative memory hamiltonians for structure prediction without homology: alpha-helical proteins, *Proc. Natl. Acad. Sci. USA*. 97:14235-14240 (2000).
38. Bartlett GJ, Taylor WR: Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction, *Proteins: Struct., Funct. Bioinform*. 71:950–959 (2008).
39. Available from http://predictioncenter.gc.ucdavis.edu/ .
40. Cozzetto D, Kryshtafovych A, Tramontano A: Evaluation of CASP8 Model Quality Predictions, *Proteins: Struct., Funct. Bioinform*. 77:157-166 (2009).
41. Ginalski K, Elofsson A, Fischer D, Rychlewski L: 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics*. 19:1015–1018 (2003).
42. Wallner B, Fang H, Elofsson A: Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller, *Proteins: Struct. Funct. Genet*. 53:534-541 (2003).
43. Wallner B, Elofsson A: Prediction of global and local model quality in CASP7 using Pcons and ProQ, *Proteins: Struct. Funct. Bioinform*. 69:184-193 (2007).
44. Larsson P, Skwark M.J, Wallner B, Elofsson A: Assessment of global and local model quality in CASP8 using Pcons and ProQ, *Proteins: Struct. Funct. Bioinform*. 77:167-172 (2009).
45. DeRonne K.W, Karypis G: Improved estimation of structure predictor quality, *BMC Structural Biology*. 9: page number not for citation purpose (2009).
46. Archie J.G, Paluszewski M, Karplus K: Applying Undertaker to Quality Assessment, *Proteins: Struct.*

Funct. Bioinform. 77:191-195 (2009).
47. Benkert P, Tosatto S.C.E, Schwede T: Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust, *Proteins: Struct. Funct. Bioinform*. 77:173-180 (2009).
48. Cheng J, Wang Z, Tegge A.N, Eickholt J: Prediction of global and local quality of CASP8 models by MULTICOM series, *Proteins: Struct. Funct. Bioinform*. 77:181-184 (2009).
49. McGuffin L.J: Prediction of global and local model quality in CASP8 using the ModFOLD server, *Proteins: Struct. Funct. Bioinform*. 77:185-190 (2009).
50. Kihara D, Chen H, Yang Y.D: Quality Assessment of Protein Structure Models, *Current Protein and Peptide Science*. 10: 216-228 (2009).
51. Kryshtafovych A, Fidelis K: Protein structure prediction and model quality assessment, *Drug Discovery Today*. 14:386-393 (2009).
52. Wang Z, Tegge A.N, Cheng J: Evaluating the absolute quality of a single protein model using structural features and support vector machines, *Proteins: Struct. Funct. Bioinform*. 75:638–647 (2009).
53. Zhou H, Zhou Y: Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins: Struct. Funct. Bioinform*. 55:1005–1013 (2004).
54. Sippl M.J: Knowledge-based potentials for proteins, *Curr. Opin. Struct. Biol*. 5:229–235 (1995).
55. Panchenko A.R, Marchler-Bauer A, Bryant S.H: Combination of threading potentials and sequence profiles improves fold recognition, *J. Mol. Biol*. 296:1319–1331 (2000).
56. Wallner B, Elofsson A: Can correct protein models be identified?, *Protein Sci*. 12:1073–1086 (2003).
57. Paluszewski1 M, Karplus K: Model quality assessment using distance constraints from alignments, *Proteins: Struct. Funct. Bioinform*. 75:540–549 (2009).
58. Archie J, Karplus K: Applying undertaker cost functions to model quality assessment, *Proteins: Struct. Funct. Bioinform*. 75:550–555 (2009).
59. Anfinsen C.B: Principles that govern the folding of protein chains, *Science*. 181:223-230 (1973).
60. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: Improving physical realism, stereochemistry and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8, *Proteins: Struct. Funct. Bioinform*. 77:114-122 (2009).
61. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: Evaluation of template-based models in CASP8 with standard measures, *Proteins: Struct. Funct. Bioinform*. 77:18-28 (2009).
62. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H: fams-ace: A combined method to select the best model after remodeling all server models, *Proteins: Struct. Funct. Bioinform*. 69:98–107 (2007).

第Ｉ章
フラグメント整合スコアの方法

Ｉ.1 序論

アミノ酸配列から蛋白質の構造を予測する問題は、構造生物学の主たる挑戦の一つである。蛋白質の構造予測の分野で 3D-1D 法が広く用いられているので [1,2,3]、3D-1D 法における新しい指標を用いて、各構造候補とアミノ酸配列の整合性を評価することは興味深い。しかし、3D 情報である立体構造と 1D 情報のアミノ配列の整合性はそのままでは測定できないので、各構造候補の立体構造が持つ 3D 情報をなんらかの 1 次元の配列になおして、その 1D 構造情報と別の 1D 情報であるアミノ酸配列を相互比較する方法が考えられてきた[4]。本論文では、この方法によって構造候補の善し悪しを評価する MQA 法を開発する。

本論文では、構造候補の善し悪しを Global Distance Test Total Score（GDTTS）によって評価する 。そこで、構造候補の中で GDTTS 値が高い構造を有効に選択する方法を開発することが本論文の目標である。そのため、PDB の既知構造のライブラリーを参考にして、ネイティブ蛋白質の立体構造を「クラスに基づく構造配列」という新しい１D 情報に変換し、既知構造のライブラリーを利用してアミノ酸配列と構造配列の整合性を調べ、それを基に「FCS スコア」という新しい指標を提案する。本章では、FCS スコアの方法を説明し、アミノ酸配列とコンシステントな構造、すなわち標的構造により近い（GDTTS 値が高い）構造が上位のスコアを得るように、スコアの算出法を改善した結果を報告する。

ターゲット蛋白質の配列が既知構造の蛋白質の配列と類似している時は、それらの既知構造をテンプレートとして未知のターゲット構造のモデリングのために使用できる。テンプレートに基づくモデリング（TBM）技法 [5,6] は、比較的に前進しているのだが、適切なテンプレートを持たないターゲット蛋白質を予測する方法については、一貫性のある体系的な方法はまだ開発されていない。後者の問題はデノボ予測、又はテンプレートに基づかないフリーモデリング（FM）とよばれ、TBM より困難な問題であるが、構造生物学の広範な問題への応用のために重要であり、そして、蛋白質構造構築の原則 [7-11] の理解のためにも重要である。本論文においては主にデノボ予測問題に集中するため、本章では、配列のホモロジー関係に頼らない知識データベースに基づくスコア関数として、FCS スコア法を開発する。

FM問題では、ターゲットと他の構造が解かれた蛋白質の間のホモロジー関係が確立されていな

いが、こうしたFM問題でも、多数の蛋白質の間のローカルな配列－構造（1D－3D）関係は構造を推論する上で助けとなる。この点は、FMターゲットのデノボ予測の中で、フラグメントアセンブリ法 [11,13-16] と様々な長さの局所部分を集める方法 [16-18] が成功したことによって一番明確に実証されている。ここで、これらの方法の重要な特徴は、ひとつの残基の代わりに有限の長さのフラグメントの集合を用いて、ローカルな1D－3D関係を推論する事である。このデノボ予測技法と同様のやり方をすれば、各残基の周辺の有限の長さのフラグメントの比較によって、構造評価のMQA方法を改良することができるはずである。既にこうした発想に基づいて、スコア関数を導き出すためにフラグメントの比較が有効に用いられているが [19]、この章では、1D－3D関係を導き出すために、フラグメントの構造だけではなく、フラグメント周辺のローカルな構造環境も使用されているところが新しい点である。ここで、この論文の中のFCS関数の計算のために既知構造の蛋白質や構造候補のpdb形式の出力のCαの座標情報のみ用いており、側鎖の原子位置に関する情報は用いられていない事に注目してもらいたい。

FCS法による解析は、VERIFY3D [19] および関連する方法 [21-22] による1D－3D方法の直接的な拡張とみなすこともできる。モデルの各残基に関して、ローカルな構造環境が構造的な指標に準じて分類されて、蛋白質構造が構造的な指標の配列として表現されている。重複しない蛋白質構造のライブラリーから、各残基周辺の９残基の長さのフラグメントが選ばれて、Position Specific Iterative (PSI)-BLAST [23]を考慮した配列プロファイル表が生成された。そして、選ばれたフラグメントの中における構造配列の出現の数を数えて、モデル構造のフラグメント整合スコアが評価されている。実際の計算の実施では、フラグメント整合スコアを評価するために、様々な計算法をとることが可能である。第Ⅱ章では、ランジュバンMD法で生成された構造候補の評価を行い、第Ⅲ章では、CASP7とCASP8のFMターゲットに対して参加したサーバーチームが提出した構造候補を対象としてFCS法を用いるが、その前に、本章では可能な計算法のうちどれを採用すべきかという方法を決めるためのテストをする。これらの複数の計算法をCASP７のFMターゲットの18ドメインを参照として用いてテストすることにより、本章では、ローカルスコア評価のための的確な平均化が、FCSのベストな性能を得るために基本的に必要な方法であることを示す。第III章では、本章のテストの結果選ばれた計算法が、CASP8の12個のFMターゲットに応用される。

図１は、FCSスコア関数の使い方の概略図とFCS法の有用性を表している。1Dアミノ酸配列からスタートして、付録Bで解説されるランジュバンMDにより、多数の3D立体構造候補（モデル構造）が生成される。また、こうして得られた構造候補は、PDBに収録された既知の構造の間に見いだされる経験則を利用して、1D構造配列に翻訳される。この1D構造配列と1Dアミノ酸配列の相互比較をすることにより、FCSスコアが導かれる。第II章では、このFCSスコアとランジュバンMDで

計算されたエネルギーを比較して、構造候補の選択が行われる。第III章では、FCSスコアによる選択と、CASPで用いられた他のMQA法による選択の結果が比較される。

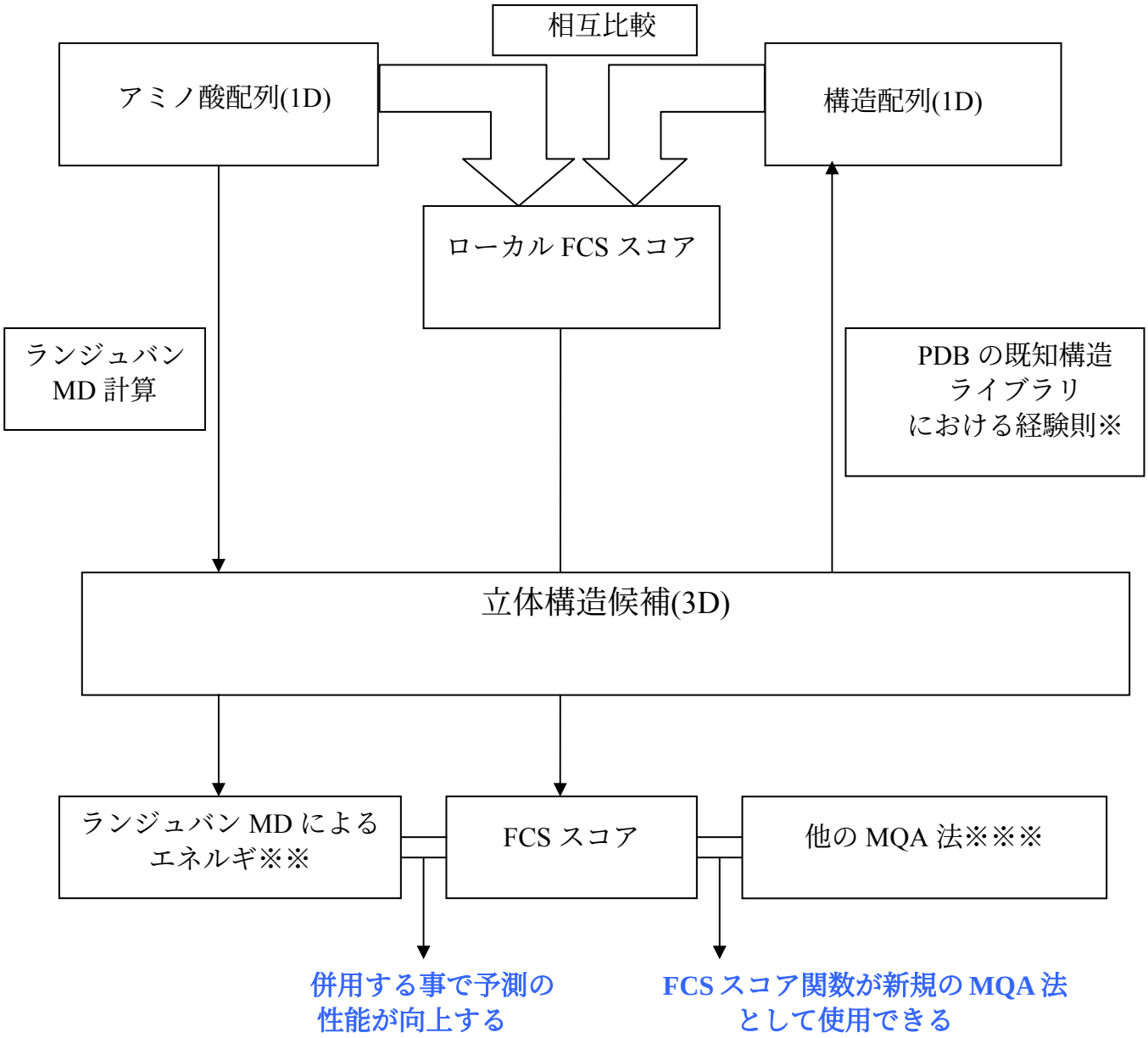


図 1　3D-1D 法による FCS スコアの使い方の概略図と FCS 法の有用性。
※の印をつけた項目は、第Ⅰ章で詳しく述べている。※※の項目は第Ⅱ章で、※※※の項目は第Ⅲ章の中で詳しく述べる。

Ｉ.2 方法

FCS法の最初のステップは、構造ライブラリーの中からフラグメントを選ぶステップである。ライブラリーから抽出することのできる全てのフラグメントの残基プロファイルと、ターゲット蛋白質のフラグメントの残基プロファイルの類似性を計算し、この類似性の高いライブラリーフラグメントの集合をその後の計算に用いる。次のステップは、ライブラリー蛋白質から選んだフラグメントの構造特徴とモデルのフラグメントの構造特徴を比較して、フラグメント整合スコアを導くことである。この方式全般を通して、実際の評価計算の中には、フラグメント整合スコアの様々な計算方法が存在している。それらの各方式をこのセクションで説明して、それらの性能の違いはI.3.1のセクションの中で議論される。

I.2.1　配列プロファイル

ターゲット蛋白質と構造ライブラリーの蛋白質の配列プロファイルは、それらの配列を、重複しない（non-redundant , NR）配列データベース [24] の配列と比較することによって得られる。この比較は、E-valueを0.001に設定してPSI-BLAST [23]を繰り返し適用することによって実行され、その結果、配列プロファイルが得られる。

I.2.2 フラグメントの収集

まず、PISCES server [25-26] をパラメータをデフォルト値に設定した状態で使用することにより、Protein Data Bank（PDB）のリストからnon-redundant な（NR）蛋白質構造のライブラリーが選ばれた。我々の方法と以前のCASPに参加したMQA法との公平な比較を行うために、CASP７に提出されたモデル構造の品質を評価する際には、CASP7が実施される以前に発表されたPDBデーターの中から3624個のライブラリー構造を選び、CASP8に提出されたモデル構造の品質を評価する際には、CASP8以前に発表されたPDBデーターの中から5164個のライブラリー構造を選んだ。

ターゲット蛋白質のi番目の残基を中心にした９残基領域を$W(i)$と書く。ただし、Nをターゲットの総残基数として$i = 4, ..., N–4$である。各 $W(i)$ について、構造ライブラリーの中から、ライブラリー蛋白質の中の９残基領域の配列プロファイルと$W(i)$の配列プロファイルの間の相関係数を調べ、最も相関係数が大きいN^{fr}個の領域が$W(i)$ のフラグメントとして選ばれた。この選ばれたフラグメントを$F_k(i)$と書く。ただし、$k = 1, ..., N^{\mathrm{fr}}$ である。我々は、N^{fr}を決定する方式として、以下の２つの

方式を比較した。

(A) *Collecting Correlated Fragments* （高い相関を持つフラグメントの収集、*CF*法）：決まったしきい値より高い相関係数を有するフラグメントが収集された。我々は、CASP7の18FMターゲットのモデルの品質を評価する能力を最大にする様に、そのしきい値を、0.6と定めた。ただし、選ばれたフラグメントの数が380以下になるように制限された。0.6以上の相関係数を有するフラグメントの数が80以下だったら最も高い相関係数を持つ80個のフラグメントが選ばれる。従って、N^{fr}は i に依存しており、$80 \le N^{\mathrm{fr}}(i) \le 380$である。

(B) *Collecting Relatively Correlated Fragments* （相対的に高い相関を持つフラグメントの収集、*RCF*法）: $W(i)$の配列プロファイルとライブラリーの可能な全ての９残基領域の配列プロファイルの間の最高の相関係数が$C^{\mathrm{Top}}(i)$だとして、ライブラリーの中から、相関係数$C_k(i)$が$\eta C^{\mathrm{Top}}(i) \le C_k(i) \le C^{\mathrm{Top}}(i)$の条件を満たすフラグメントが収集された。$\eta$のパラメーターは、CASP7の18FMターゲットのモデルの品質を評価する能力を最大にする様に、0.75と定められた。フラグメントの数は、上記に説明された*CF*法と同様に、$80 \le N^{\mathrm{fr}}(i) \le 380$に設定された。

I.2.3 構造配列

I.2.3.1　3D-1D 変換による構造配列の作成

構造配列の作成は、構造ライブラリーから得られるフラグメントの集まり、及び第Ⅱ章と第Ⅲ章で述べる様なターゲット蛋白質に対する各モデルの構造候補の両方について行うが、ランジュバン動力学では Cα しか扱わないので、側鎖原子の位置が求められている場合に関しても、Cαだけの座標情報に絞り込むことによって、構造配列を作成した。蛋白質の鎖をα炭素のビーズの繋がりで表現し、その座標を$\{\mathbf{r}_i\}$で示す。立体構造を 1D アミノ酸配列と比較するために、立体構造を１Dで表現せねばならない。そのために、アミノ酸配列の種類によって偏らないバランスのいい分布を示す立体構造の特徴を、指標として使用すべきである。このため、ローカルな構造環境を３つの指標で表した。この３つの指標の重要なポイントを、以下の I.2.3.2 と I.2.3.3 と I.2.3.4 に簡潔に説明しておく。さらに詳細な点は、付録 A に説明する。

I.2.3.2　２次構造判定

蛋白質の立体構造を特徴付けるものとして、すぐに思いつくのは 2 次構造である。アミノ酸の種類によって、どの 2 次構造をとりやすいかという傾向があるのはいうまでもなくよく知られていることなので、有用な指

標と言えよう。また蛋白質はフォールディングする段階において実際に 2 次構造を形成しながら更に複雑に折りたたむのであるし [27]、3D 同士の構造比較に於いても [28]、3D-1D の変換 [2]やスレッディング法に於いても、2 次構造を参照して行われるケースが多い。それで、3D の立体構造を 1D で表現するにはまず、各残基に関して 2 次構造を判定せねばならない。

図 2 のように、注目する残基 *i* の周りの残基のつくる2つ三角形を含む面を考え、2つの面の間の角度のコサイン関数を計算し [29]、その値を基に、残基 *i* の周囲が、α ヘリックスと β シートの 2 次構造、またはその他の構造、合わせて 3 つの構造の中のどれをとるか判定する。

但し、この場合、連続した2次構造の (特に α ヘリックスの) 末端部分の 2 つの残基をその他の構造と評価する傾向があるという問題点が残るが、それに関して、ペナルティーを与えれば、データーベースの情報の個数が減少する。別の方法として、連続して同じ 2 次構造をとらないとその 2 次構造として評価しないようにすれば、末端部分のエラーが増加する。よって、この問題点に関して微細な修正をすることは全体の成績を上げることにつながらない。Adamczak ら [30] でも見られるように、I.2.3.3 の N_{10} の利用により、2 次構造判定の精度が向上するというデーターがあるので、この問題点は構造配列作成に大きな影響を及ぼさないものと考えた。まとめると以下の通りである。

我々は、2 面角 θ_i を $\mathbf{r}_{i-1}, \mathbf{r}_i,$ と $\mathbf{r}_{i+1}$ で描かれた面と $\mathbf{r}_i, \mathbf{r}_{i+1}$ と $\mathbf{r}_{i+2}$ で描かれた面の間の角度として定義した。図 2 で参照できるように、θ_i が $\alpha_0 < \cos\theta_i < \alpha_1$ を満たす時に、*i* 番目の残基周辺のローカル構造 S_i を $S_i = \alpha$ と分類し、$\beta_0 < \cos\theta_i < \beta_1$ の時に $S_i = \beta$、それ以外の時に $S_i = C$ と分類した。しきい値 α_0 と β_0 の決め方は、付録 A「2 次構造判定のしきい値決定」に詳しく述べるが、ライブラリー蛋白質の 2 次構造の割り当てを最適に再現できる様に、$\alpha_0 = 0.35$, $\alpha_1 = 0.82$ および $\beta_0 = -1$, $\beta_1 = -0.78$ と定められた。

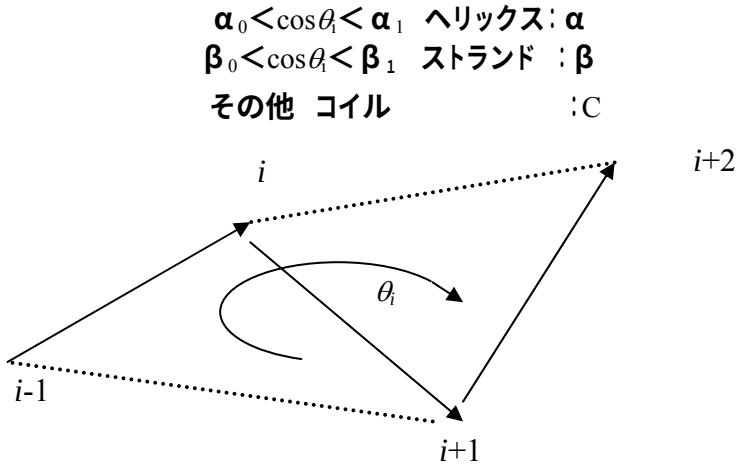


図 2 残基 *i* の周囲の 2 次構造判定

I.2.3.3　　構造密度 N_{10}

蛋白質の各アミノ酸残基の立体構造において、もう１つの重要な特徴として、各残基が蛋白質の中心に埋もれているか蛋白質の表面に出ているかという状況を区別する量である構造密度を挙げることができる。この指標は、その残基が疎水性か親水性かという区別に関係しており、アミノ酸の種類によって異なる分布を示している。

そこで図 3 のように、10Å以内の周辺残基数 N_{10} を各アミノ酸残基の Cα から 10Å以内にある C_α の数と定義して、構造密度を表す指標とした。アミノ酸残基が中に埋もれているほどこの値が大きくなるはずであり、逆に、アミノ酸残基が露出しているほど、この値が小さくなるはずである。また、N_{10} は 3D-1D や構造比較においてよく用いられる溶媒接触率(SAR) [27]とも関係しており、SAR が大きくなれば小さくなる量のはずである。溶媒接触率（SAR）とは、1 つのアミノ酸残基の全体表面の中で水と接触している表面の割合を意味する。アミノ酸残基は中に埋もれているほどこの値が小さくなるし、逆に、アミノ酸残基が露出しているほど、この値が大きくなるはずであり、最大で 1 最小で 0 になる。

本研究の中では Garg ら[27]と違って、N_{10} を 2 段階ではなく、4 段階にわけて表現した。すなわち、0 個～6 個、7 個～12 個、13 個～18 個、19 個以上の 4 段階である。まとめると、以下のようになる。

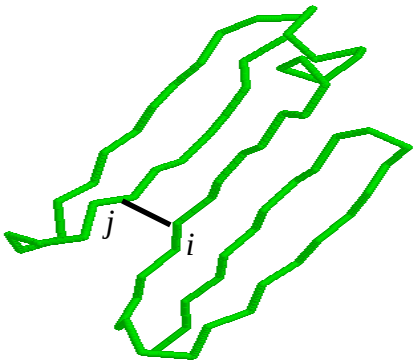
N_{10} は $\mathbf{r}_i$ の位置を中心とした半径 10Å の球内の α 炭素の数で定義される。構造密度の指標 D_i は、$N_{10} \le 6$ の時に $D_i = 1$、$7 \le N_{10} \le 12$ の時に $D_i = 2$、$13 \le N_{10} \le 18$ の時に $D_i = 3$、$19 \le N_{10}$ の時に $D_i = 4$ と定義する。

図 3　N_{10} の概念の導入

I.2.3.4　　ローカル･コンタクト･オーダー (LCO)

最終的な結果をよくするためには、構造を捉える効果的な指標が必要である。そのため、Kinjoら[31]の研究で蛋白質全体に対して定義されたコンタクト・オーダーを図４のように残基毎に定義して、LCO（ローカル・コンタクト・オーダー）と名づけた。フラグメントの中心残基が i 番目であるとき、$\mathbf{r}_i$ を中心とした半径 10Å の球内に位置する j 番目との配列に沿っての距離$|i-j|$と書く。LCO は、$\mathbf{r}_i$ を中心とした半径 10Å の球内に位置する残基すべての残基について、この配列に沿っての距離を平均した量として定義される。

指標L_iを次のように定義する。 $\mathrm{LCO} \le L_0$ の時に$L_i = 1$、$\mathrm{LCO} > L_0$の時に$L_i = 2$と書く。我々は、構造ライブラリーにおいて、$L_i = 1$の総残基数と$L_i = 2$の総残基数が大体同じになる様に、CASP7より以前に構造が公表された3624個の重複しない（NonRedundant NR）ライブラリー蛋白質と、CASP8より以前に構造が公表された5164種のNRライブラリー蛋白質でそれぞれテストした結果、CASP7では$L_0 = 27.79$ 、CASP8では$L_0 = 28.861$に設定すべきであることが判明したため、計算対象に合わせてそれぞれ使用した。



i 番目の残基のCα周辺の、半径10Åの球の中に入っている Cα がj 番目の残基のCαであったとき、残基番号の差 $|i-j|$ を半径10Åの球の中の残基すべてについて平均した値

図 4　LCO の概念の導入

I.2.3.5　　クラスに基づく構造配列

上述した通りに、各アミノ酸残基を2次構造について3クラスに分け、さらにN_{10}の構造密度について4クラスに分け、LCOについて2クラスに分け、３Ｘ４Ｘ２の２４クラスを作成した。このように、i番目の残基周辺のローカル構造を$x_i = (S_i, D_i, L_i)$.で表現する。この定義をイラストで表したのが、図5である。アミノ酸の種類によってどのクラスに入るかという傾向があるものの、必ず同じ種類のアミノ酸がいつも同じクラスに入るわけではないということに注意していただきたい。例えば、同じアラニンでも、周辺の状況によって、２次構造や構造密度が変化すれば、入るクラスも当然異なり、同じ蛋白質の違う箇所においても変化している。

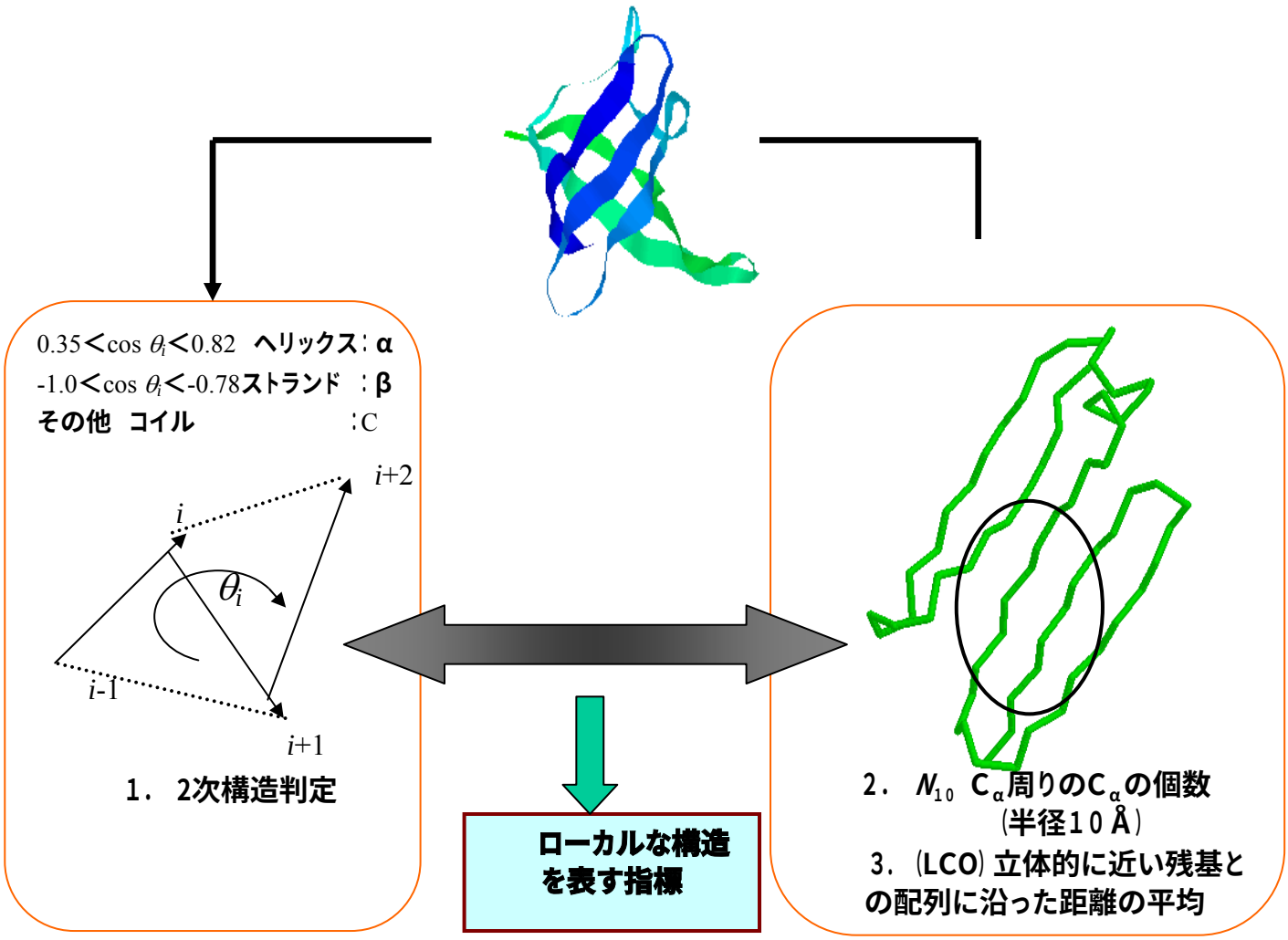


図５ ローカルな構造を表す指標の定義

こうして定義された三つ組の指標を用いると、蛋白質構造は$\{x_i\}$の配列で表現される。我々はそれを構造配列と呼ぶ。これによって、ライブラリー蛋白質の3D立体構造も、第II・第III章で述べるターゲットを予測するためにつくられた各モデル構造候補も、1Dの構造配列に変換できるようになり、1Dの構造配列と1Dのアミノ酸配列が相互に並べられるのである。

I.2.4 フラグメント整合スコア

モデル構造のフラグメントの構造とライブラリー蛋白質から抽出したフラグメントの構造を比較することによって、ローカルなフラグメント整合スコアが得られる。残基 i を中心に持つモデル構造のフラグメント $W(i)$ に属する残基 j のローカル構造を $x_j = (S_j, D_j, L_j)$ と書き、ライブラリー蛋白質から抽出したフラグメント$F_k(i)$ に属する n 番目の残基ローカル構造を $x_n = (S_n, D_n, L_n)$ と書くとき、x_j とx_n の類似度を表す指標、$p_k(i)\{x_j: x_n\}$を以下の２つの方法で定義し、後にこの２つの方法による違いを評価する。

(A) *All or None Comparison* (全か無かの指標比較法、*ANC*法)：３つの指標がすべて一致する$x_j = x_n$の時に$p_k(i)\{x_j: x_n\} = 1$、それ以外の$x_j \neq x_n$の時に$p_k(i)\{x_j: x_n\} = 0$だと定める。

(B) *Fuzzy Comparison* (ファジー指標比較法、*FC*法)：構造がどれくらい異なるかという度合いが計算に取り入れられる。 $x_j = x_n$ の時に $p_k(i)\{x_j: x_n\} = 1$、x_j と x_n の3つの指標の内2つが同一の場合 $p_k(i)\{x_j: x_n\} = 0.5$、x_j と x_n の3つの指標の内1つが同一の場合 $p_k(i)\{x_j: x_n\} = 0$、x_j と x_n の3つの指標 $_j$ の全てが違う場合に $p_k(i)\{x_j: x_n\} = -0.5$と定める。例えば、$S_j = S_n, D_j = D_n,$ で $L_j \neq L_n$ の時に $p_k(i)\{x_j: x_n\} = 0.5$ になり、$S_j \neq S_n, D_j = D_n,$で $L_j \neq L_n.$ の時に $p_k(i)\{x_j: x_n\} = 0$ になる。

我々は、フラグメント$F_k(i)$の中心残基をk_0と書き、k_0周辺の連続した5つの残基を $k_0 - 2, k_0 - 1, k_0, k_0 + 1, k_0 + 2$ と書く。$p_k(i)\{x_j: x_n\}$ からローカルな構造比較指標として $q_k(i)$ を以下の２つの方法で導いた。後にこの２つの方法による違いを評価する。

(A) *Center-to-Center Matching* (中心残基比較法、*CCM*法)：モデル蛋白質の配列領域の中心残基とフラグメントの中心のみを比較する。すなわち、$q_k(i) = p_k(i)\{x_i: x_{k0}\}$とする。

(B) *Finite Width Matching* (有限幅比較法、*FWM*法)：$W(i)$の構造環境と$F_k(i)$ の構造環境が、以下の様に比較される。$q_k(i) = p_k(i)\{x_i: x_{k0}\} + 0.5p_k(i)\{x_{i+1}: x_{k0+1}\} + 0.5p_k(i)\{x_{i-1}: x_{k0-1}\} + 0.25p_k(i)\{x_{i+2}: x_{k0+2}\} + 0.25p_k(i)\{x_{i-2}: x_{k0-2}\}$である。

$q_k(i)$ から、$W(i)$ のローカル・フラグメント整合スコア $LFCS_i$, が導かれる。ターゲットの中の i 番目の残基周辺の 9 残基領域に対して、構造ライブラリから選ばれたフラグメントの数を $N^{fra}(i)$ と定義する。$N^{fra}(i)$ 個のフラグメントを相関係数の大きさ順に並べたとき、k 番目のフラグメントの順番を$O_k(i)$ と書いて、k 番目のフラグメントの規格化の重み数 $f_k(i)$を $f_k(i) = \left(N^{fra}(i) - O_k(i) + 1\right)/N^{fra}(i)$と定義した。

こうして、ローカル・フラグメント整合スコア$LFCS_i$ が以下の様に計算される。

$$LFCS_i = \sum_{k=1}^{N^{fr}(i)} q_k(i) f_k(i) \qquad (1)$$

ローカル・フラグメント整合スコアをモデル構造全体にわたって足すことにより、モデル構造のフラグメント整合スコアが以下の様に得られる

$$FCS = \sum_{i=4}^{N-4} LFCS_i \qquad (2)$$

上に説明したように、実際の計算においては、フラグメント整合スコアの計算のために、*CF* 又は *RCF, ANC* 又は *FC,* そして *CCM* 又は *FWM* という具合に、様々な方法が可能である。この後のセクションにおいて、それらの$2^3 = 8$の方法の性能が比較される。図6に構造配列の作成とFCSスコアの計算の概略図をまとめた。

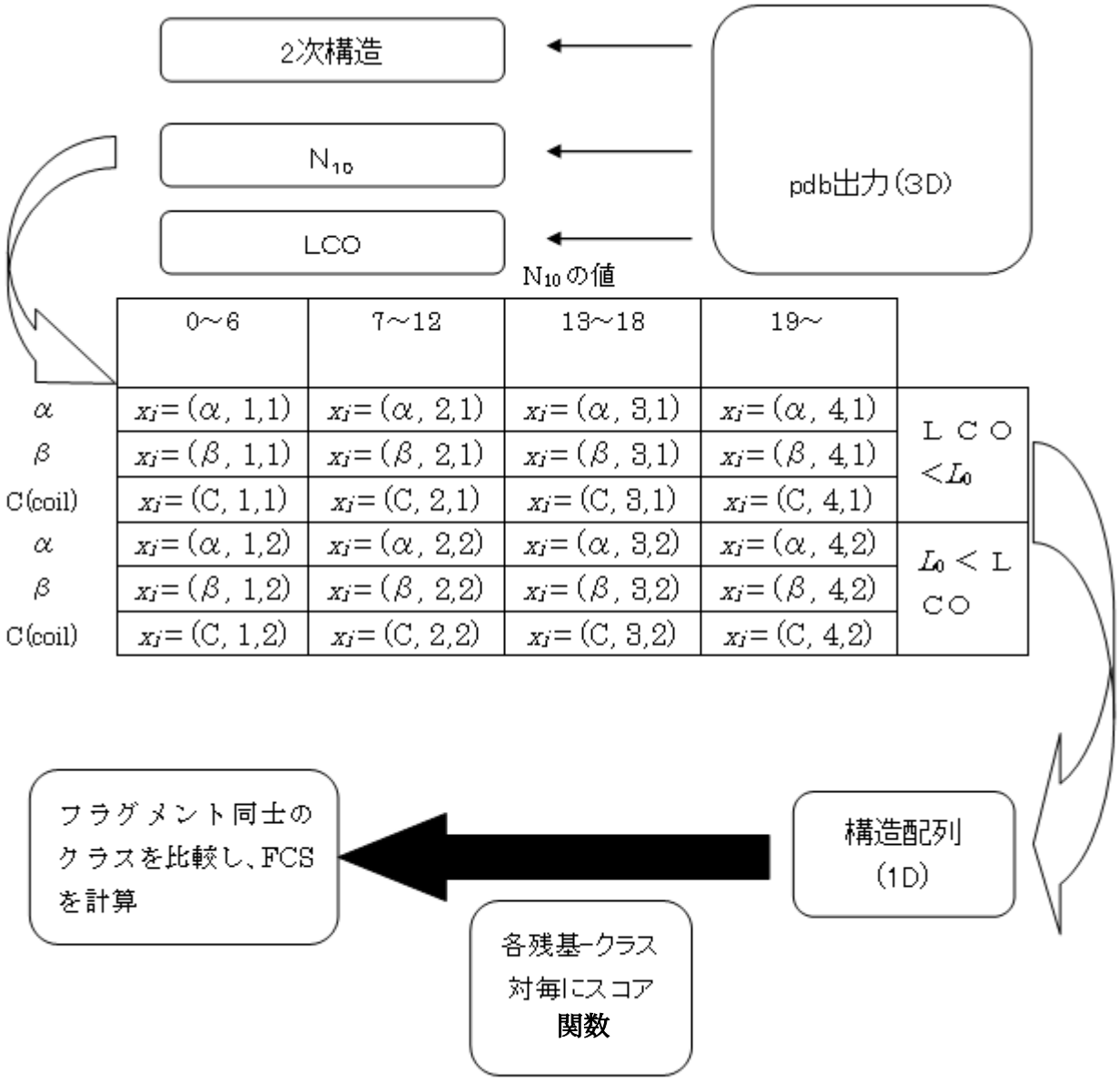


図6　FCSスコア関数の計算の流れ

Ⅰ.3　FCS の計算方式の選別

FCS計算を実行するために、前節では８つの異なる方法が定義された。この節では、これらの８つの方法が、CASP7に出題された18個のFMドメインを使って相互に比較された。それらの18個のターゲット・ドメインは付録Dの表D1に説明されている。各ターゲットに関して、CASP７に参加したサーバーチームによってモデル構造が予測された。各サーバーチームは自動計算の方法を用いる事によって、モデル構造をCASPの問題への回答として作成しており、提出されたモデル構造はCASPのウェブページで公開されている[32]。これらのモデル構造が本論文のFCS法によって評価された。ターゲットによって、提出されたサーバー・モデル構造の数は異なるが、利用可能なサーバー・モデル構造数は、１ターゲット当たり640から940個の程度である[32]。我々は方法の違いをμで表す。すなわち、μ = *CF-ANC-CCM, RCF-ANC-FWM,* 等々である。方法μで計算された、*n*番目のターゲット蛋白質の a 番目のモデル構造のフラグメント整合スコアを$FCS(a, n, \mu)$と書く。n = 1-18である。

モデル構造と正しい解答構造がどの程度類似した構造であるかを表す指標として、我々はGlobal Distance Test Total Score (GDTTS) [33]を使用する。CASP大会に参加した各予測チームは、CASP主催者に5つのモデル構造を提出できる。この論文の中でも同じ基準を利用して5つの構造を選ぶ。すなわち、与えられた n とμに対して、全サーバーモデルの中で5つの最も高いFCSを持つモデル構造を選ぶ。この5つのモデル構造のうちで、最も高いGDTTSを持つモデル構造を$a^{\text{best}}(n, \mu)$と表す。

*n*番目のターゲットドメインの a 番目のモデル構造のGDTTSを$GDTTS(a, n)$と書いた。図7にプロットしたのは、異なるμに対する相対GDTTS、すなわち$\Delta GDTTS$である。

$$\Delta GDTTS(n,\mu) = \frac{GDTTS(a^{\text{best}}(n,\mu),n) - GDTTS(\max)}{GDTTS(\max)} \qquad (3)$$

ここで、$GDTTS(\max)$ は図7の各パネルの中のμで区別された異なる方式の間における $GDTTS(a^{\text{best}}(n, \mu), n)$ の値の最高値である。$\Delta GDTTS(n, \mu)$ の曲線が*n*の関数として、$\Delta GDTTS(n, \mu)$ の高い値から低い値に向かってプロットされている。

図 7 では、$\Delta GDTTS$ が高い曲線を示す方法が良い方法である。図 7a と 7c を見ると、*FC* が常に *ANC* よりいい結果をもたらすことが分かる。図 7a と 7b を見ると、*FWM* が一般的に *CCM* よりいい結果を与え、図 7b と 7c からは、*RCF* が *CF* よりいい結果を与えることが分かる。*FC* と *FWM* を優先させる事が、$LFCS_i$ の計算の最適な実行に於いて、良い性能を得るために基本的に重要である。

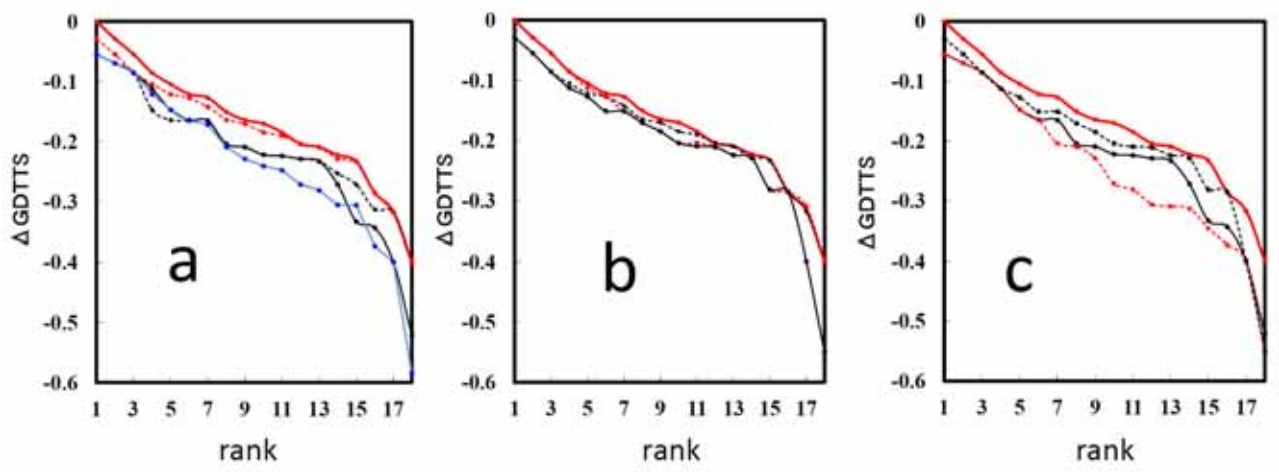


図7　様々なFCS計算法の比較

FCS計算法の8種を、CASP7の18個のFMドメインのΔGDTTSを基準として比較する。各パネルの中で、ΔGDTTSが下る順番に18個のターゲットを並べてプロットした。FCSは以下の方法で計算された。(**a**) *RCF-FC-FWM* (赤の実線), *RCF-ANC-FWM* (黒の実線), *RCF-FC-CCM* (赤の点線), *RCF-ANC-CCM* (黒の点線),そして*CF-ANC-CCM* (青の実線), (**b**) *RCF-FC-FWM* (赤の実線), *CF-FC-FWM* (黒の実線), *RCF-FC-CCM* (赤の点線),そして*CF-FC-CCM* (黒の点線), (**c**) *RCF-FC-FWM* (赤の実線), *CF-FC-FWM* (黒の点線), *RCF-ANC-FWM* (黒の実線), そして *CF-ANC-FWM* (赤の点線)。

Ⅰ.4　まとめ

CASP7 の FM ターゲットを用いて調べたところ、8 通り考案された FCS 法の方式の中で、他の方式より明確に良いものが見つかり、特に、有限幅比較（*FWM*）法とファジー指標比較（*FC*）法を使った予測の成績が、FCS 法の他の方式より優れていた。

付録 A では、2 次構造判定のための面角度のしきい値決定について説明しておく。2 次構造判定法をテストした結果、β構造の判定率がやや高いことがわかるが、コイルやループの構造を持つ環境もβとして間違って判定する傾向があり、本格的なβシート判定のためには二面角以外の指標も必要と考えられる。

引用文献

1. Bowie JU, Luthy R, Eisenberg D: A method to identify protein sequences that fold into a known three-dimensional structure, *Science*. 253:164-170 (1991).
2. Etchebest C, Benros C, Hazout S, De Breven A.G.: A Structural alphabet for Local Protein Structures Improved Prediction Methods, *PROTEINS:Structure, Functions and Bioinformatics*. 59:810-827 (2005).
3. Rice D.W, Eisenberg D: A 3D-1D substitution Matrix for Protein Fold Recognition that includes Predicted Secondary structure of the Sequence, *J.Mol.Biol*. 267:1026-1038 (1997).
4. Berglund A, Head R.D, Welsh E.A, Marshal G.R: ProVal A Protein-Scorong Function for the Selection of Native and Near-Native Folds, *PROTEINS:Structure, Functions, and Bioinformatics*. 54:289-302 (2004).
5. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: Improving physical realism, stereochemistry and side-chain accuracy in homology modeling Four approaches that performed well in CASP8, *Proteins: Struct. Funct. Bioinform*. 77:114-122 (2009).
6. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: Evaluation of template-based models in CASP8 with standard measures, *Proteins: Struct. Funct. Bioinform*. 77:18-28 (2009).
7. Ben-David M, Noivirt-Birk O, Paz A, Prilusky J, Sussman J L, Levy Y: Assessment of CASP8 structure predictions for template free targets, *Proteins: Struct. Funct. Bioinform*. 77:50-65 (2009).
8. DeBartolo J, Colubri A, Jha A.K, Fitzgerald J.E, Freed K.F, Sosnick T.R: Mimicking the folding pathway to improve homology-free protein structure prediction, *Proc. Natl. Acad. Sci. USA*. 106:13748-13753 (2009).
9. Voelz V.A, Shell M.S, Dill K.A: Predicting peptide structures in native proteins from physical simulations of fragments, *PLoS Comput Biol*. 5:1-12 (2009).
10. Papoian G..A, Ulander J, Eastwood M.P, Z. Luthey-Schulten, P. G. Wolynes: Water in protein structure prediction, *Proc. Natl. Acad. Sci. USA*. 101:3352-3357 (2004).
11. Chikenji G, Fujitsuka Y, Takada S: Shaping up the protein folding funnel by local interaction lesson from a structure prediction study, *Proc. Natl. Acad. Sci. USA*. 103:3141-3146 (2006).
12. Rohl C.A, Strauss C.E.M, Misura K.M.S, Baker D: Protein structure prediction using Rosetta, *Methods Enzymol*. 383:66–93 (2004).
13. Bradley P, Misura K.M.S, Baker D: Toward high-resolution de novo structure prediction for small proteins, *Science*. 309:1868–1871 (2005).
14. Lee J, Kim S.Y, Lee J: Protein structure prediction based on fragment assembly and parameter optimization, *Biophys. Chem*. 115:209–214 (2005).
15. Fujitsuka Y, Chikenji G, Takada S: SimFold energy function for de novo protein structure prediction: consensus with Rosetta, *Proteins: Struct. Funct. Bioinform*. 62:381–398 (2006).
16. Zhang Y, Arakaki A.K, Skolnick J: TASSER: an automated method for the prediction of protein tertiary structures in CASP6, *Proteins: Struct. Funct. Bioinform*. 61:91–98 (2005).
17. Zhou H, Skolnick J: Ab initio protein structure prediction using Chunk-TASSER, *Biophys J*. 93:1510-1518 (2007).
18. Wu S, Skolnick J, Zhang Y: Ab initio modeling of small proteins by iterative TASSER simulations, *BMC Biol*. 5:page number not for citation purpose (2007).
19. Zhou H, Skolnick J: Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential, *Proteins: Struct. Funct. Bioinform*. 71:1211–1218 (2008).
20. Eisenberg D, Lothy R, Bowie J: VERIFY3D: Assessment of protein models with three-dimensional profiles, *Methods Enzymol*. 277:396-404 (1997).
21. Zhou H, Zhou Y: Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins: Struct. Funct. Bioinform*. 55:1005–1013 (2004).
22. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H: fams-ace: A combined method to select the best model after remodeling all server models, *Proteins: Struct. Funct. Bioinform*. 69:98–107 (2007).

23. Altschul S.F, Madden T.L, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman D.J: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*. 25:3389-3402 (1997).
24. Available from http://www.ncbi.nlm.nih.gov/ .
25. Wang G., Dunbrack Jr R.L: PISCES a protein sequence culling server, *Bioinformatics*. 19:1589-1591 (2003).
26. Available from http://dunbrack.fccc.edu/PISCES.php .
27. Garg A, Kaur H, Raghava G.P.S: Real Value Prediction of Solvent Accessibility in Proteins Using Multiple Sequence Alignment and Secondary Structure, *PROTEINS:Structure, Functions and Bioinformatics*. 61:318-324 (2005).
28. Sippl MJ, Weitckus S: Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations, *Proteins*. 13:258-271 (1992).
29. 斉藤　静司、博士論文　「モデルタンパク質の折れ畳みの動力学とそのエネルギー面の特徴」
30. Adamczak R, Porollo A, Meller J: Combining Prediction of Secondary and Solvent Accessibility in Proteins, *PROTEINS:Structure, Functions and Bioinformatics*. 59:467-475 (2005).
31. Kinjo A.R, Horimoto K, Nishikawa K: AGBNP an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling, *PROTEINS:Structure,Functions and Bioinformatics*. 58:158-165 (2005).
32. Available from http://www.predictioncenter.org/casp7/index.cgi .
33. Zemla A: LGA: A method for finding 3D similarities in protein structures, *Nucleic Acids Res*. 31:3370-3374 (2003).

第Ⅱ章
粗視化したランジュバン分子動力学と
フラグメント整合スコア法の併用

Ⅱ.1 序論

蛋白質構造予測のために、多くの計算法が提案されてきた。Anfinsenの熱力学仮説 [1] に従えば、ネイティブ構造は低い自由エネルギーを持つはずである。デノボ予測のために、多くの研究グループが、様々なタイプの有効なエネルギー関数の応用によって、そのような低い自由エネルギーの構造を見つけるサンプリング技法を開発した。それらの中で比較的、優秀な方法は、9残基の長さのフラグメントを集めるフラグメントアセンブリ法 [2-6] や、それより長い鎖の立体配置を集める戦略をとるTASSER法 [7-10]である。これらの方法では、まずターゲットとデーターベース蛋白質の間のローカルな配列類似性を活用してローカルな構造候補が集められる。そして、低い有効エネルギーを持つ全体構造が形成されるように、ローカル構造候補の整合的な組み合わせを見つける事によって、全体の鎖の構造が予測される。これらの方法の成功は、ローカル構造と全体構造の間に整合性があること [11] 、あるいはフラストレーションが最小化されていること [12] が蛋白質の構造形成の指針である事を意味している。

デノボ予測の別の戦略は、モンテカルロ [13-15]、又はランジュバン分子動力学（MD）法 [16-19] を使用する、フォールディング過程のシミュレーションによる方法である。物理的フォールディングの過程をシミュレートする利点は幾つか存在する。1つ目は、予測問題のために開発された方法がフォールディング過程に対する洞察を与えるという点である。2つ目には、予測問題以外でも蛋白質の機能に於ける大規模な構造変化に方法が応用できるという可能性を拡げる点である。3つ目には、重要な事として、自然の中に存在する構造の生成過程を真似する事は、複雑な構造の決定の合理的な手法として期待できるという点である。

デノボ予測を改善する方法は、予測法によって提案された多数の構造候補の品質を評価し、よりよい構造候補を選び出すモデル品質評価（Model Quality Assesment, MQA 又は QA）法を適用することである。最近の CASP では、多くの MQA 法の能力が比較され、その結果、多数の構造候補に共通な性質をより多く備えている構造候補をよい候補とするコンセンサス評価法がよい成績を収めることが判明した[25-27]。しかし、一つあるいは少数の予測法だけしか用いる事ができない時は、同じ方法を通して生成される多数のモデルが同じ間違った見方のエラーを持っていることが予想されるため、コンセンサスに基づいた方法は、CASP の様に多数の違う予測サーバーのモデルが利

用可能な機会と同じようにうまく機能するとは限らない。しかし、コンセンサスに基づかない方法である FCS 法は、このような時にも利用可能であると考えられる。

この章では、我々は、上記の２つ戦略の両方の合併である新規に開発されたデノボ予測法を議論する。すなわち、第II章で解説した、コンセンサスに基づかないMQA法であるFCS法と、有効エネルギー関数をベースにした予測法であるランジュバンMD法を組み合わせて用いる。

ランジュバン MD 法は、何種類かのポテンシャルの合計からなる、粗視化したエネルギー関数を利用する [20,21]。それらのポテンシャルの幾つかは、ターゲット蛋白質に適用するフラグメントがどのような構造の傾向を持つかを表しており、他の多残基ポテンシャルは、フラグメントが疎水性相互作用と水素結合を通してどうやって集合するか表す。この手法によって、合計エネルギー関数が充分に低くなった時に、ローカルな構造予測のための条件とローカル構造間のフラストレーションを最小にする条件の両方を同時に満たすように、フラグメントを集めることが実現されるはずである。この様に定義されたエネルギー関数を使用して、低いエネルギーの構造を探索するために、ランジュバン MD シミュレーションが行われた。フラグメントを利用した粗視化されたランジュバン MD の方法は、最近、連想記憶ハミルトニアン法の文脈の中でも利用されている [22]。

この章では、CASP8のFMカテゴリーに属する10個のターゲット蛋白質、又はターゲットドメインを対象として、この方法のベンチマーク試験を行った結果を説明する。ランジュバンMDにより、各ターゲットに対して90-270個のモデル構造を生成し、これらの生成された構造をフラグメント整合スコアによって分類して、低い有効エネルギーと高いフラグメント整合スコアを同時に持つモデル構造を選ぶ。こうして選ばれたモデル構造をCASP8のサーバーとメタサーバーグループのモデル構造と比較し、FCSで選ばれたいくつかのモデルが、サーバー又はメタサーバーモデルより良い品質を持っている事を示す。

Ⅱ.2 方法

Ⅱ.2.1　フラグメント整合スコア関数

第 I 章では、FCS 計算を実行する際には複数の方法が可能であること、そしてそれらの方法のうち、最も良い方法は *RCF-FC-FWM* 法であることが示された。本章では、この結果に従い、*RCF-FC-FWM* 法を用いて FCS 計算を行う。

ランジュバン MD によって、CASP8 [22]の FM カテゴリーの対象となった 10 個のターゲット蛋白質ドメインのそれぞれに対して、約 90-270 個のモデル構造が生成された。我々はランジュバン MD を用いて CASP8 に参加したのであるが、本章で使うモデル構造は、この CASP 参加時に作成されたモデル構造である。CASP の参加者はターゲット蛋白質の配列が出題されてから、短い期間の間に予測構造を CASP 組織者に提出しなければならない。この締切までの時間に可能なランジュバン MD 計算の回数は、蛋白質の大きさ、その他の条件によって異なるため、上記のようにターゲットごとに対応するモデル構造の数にバラツキがあることは避けられなかった。CASP で蛋白質の全体の長い配列がターゲットとして出題された際には、その予測構造の CASP での評価は、長い鎖をいくつかのドメインに分割してドメインごとに行われることが多い。しかし、どこがドメインかは、ターゲットが出題された時点では明らかでないため、ここでのランジュバン MD 計算は、出題されたターゲット配列全長を用いた計算である。

Ⅱ.2.2　ランジュバン MD

以下の過減衰ランジュバン方程式に従う変化を数値的に追跡することによって、ペプチド鎖のフォールディングをシミュレートする。

$$d\mathbf{r}_i / dt = -\partial V_{\text{total}} / \partial \mathbf{r}_i + \xi_i(t) \qquad (2)$$

上式の中で、$\xi_i\ (t)$ は $\left\langle \xi_i(t)\xi_j(t') \right\rangle = 2T\delta_{ij}\delta(t-t')$ を満たすガウシアンホワイトノイズであり、T は構造ゆらぎの振幅を制御する温度に類似した意味を持つパラメーターである。V_{total} が有効エネルギーであり、結果と考察のセクションにおいて、構造選択のために使われる。V_{total} は$\{\mathbf{r}_i\}$によってあらわに微分することができる関数形を持つ多残基間のポテンシャルであり、以下の様に分割して表現できる。

$$V_{\text{total}} = V_{\text{fragment}} + V_{\text{assemble}} \quad (3)$$

上式の中で、V_{fragment} はローカルなフラグメント構造を形成する相互作用を表しており、$V_{\text{fragment}} = w_1 V_{\text{fragment}}^{\text{pair}} + w_2 V_{\text{fragment}}^{\text{angle}}$ と書くことができる。V_{assemble} はフラグメントを集合する相互作用を表しており、$V_{\text{assemble}} = w_3 V_{\text{nn}} + w_4 V_{\beta}$ と書ける。重み係数 w_1, w_2, w_3, と w_4 は、構造候補の中から正しいターゲット構造を識別できるように、決定されるべきであるが、その詳細は付録 B で説明する。

$V_{\text{fragment}}^{\text{pair}}$ と $V_{\text{fragment}}^{\text{angle}}$ と V_{nn} は、上述した様に、構造ライブラリーから選ばれたフラグメントの構造の統計的な傾向を見積もる事によって構築される。$V_{\text{fragment}}^{\text{pair}}$ はフラグメントの中の残基-残基ペア間距離の分布を再現するように決定され、$V_{\text{fragment}}^{\text{angle}}$ はフラグメント内の鎖がつくる面角度の分布を再現するように決定される。V_{nn} は $\mathbf{r}_i$ の位置の近くに位置する α 炭素の数で決定される。V_{β}は β シートを形成するために β ストランドの集合を促進する構造制約を与える。ポテンシャルについての詳しい説明は、付録 B に記述した。

Ⅱ.2.3　冷却スケジュール

ランジュバン MD 計算は大きい T の高温から開始し、徐々に低温にして有効エネルギーが低い構造を探す。この冷却スケジュールは以下のように設定した。MD の i 番目の MD ステップに於ける T を $T(i)$として、x より小さい最大の整数を Int(x)と書く。$T_0 = 0.3$ から開始して、$T(i) = T_0(1\text{-Int}(i/N_{\text{ann}})/(N_{\text{step}}/N_{\text{ann}}))$ によって、ランジュバン分子動力学（MD）が行われた。$N_{\text{step}} = N^{\text{res}}$ X10^6 は MD の全ステップの数であり、N_{ann} は 10 にセットされている。

Ⅱ.3 結果および考察

CASP8 の FM カテゴリーに属する 10 個のターゲットドメインを対象として、粗視化されたランジュバン MD によって各ターゲットに対して約 90-270 個のモデル構造が生成された。ターゲットと生成された構造の数は表 1 にまとめられている。こうして生成されたモデル構造から 2 つの異なる方法によって構造候補が選別された。選別方法の 1 つは、有効エネルギーが最も小さい 5 つの構造を選ぶ方法であり、選別方法のもう 1 つは、有効エネルギーとフラグメント整合スコアの併用によって 5 つの構造を選ぶ方法である。有効エネルギーとフラグメント基準スコアの併用の場合、生成された構造の全体から、まずエネルギーが最も低い 20%の構造が選ばれ、そしてそれらの中から

フラグメント整合スコアが最も高い 5 つの構造が選ばれた。我々はここで、モデル構造と正しいターゲット構造の類似の程度を GDTTS によって評価する。表 1 では、有効エネルギーとフラグメント整合スコアの併用によって得られた 5 つの構造のうち、最も良い構造の GDTTS が、ランジュバン MD で生成されたすべての構造のうち、最もよい構造の GDTTS と比較されている。この比較より、有効エネルギーとフラグメント整合スコアの併用は可能な最もよい選別に近い成績を残していることがわかる。

表 1　ターゲット蛋白質ドメインと、ランジュバン MD および FCS による選別の結果

*Target ID in CASP	PDB code	Number of structures generated with Langevin MD	GDTTS of best structures generated by Langevin MD	**GDTTS of structures selected by Scr/Ene/LangevinMD
T0397-D1	3d4r	259	29.88	29.27
T0405-D1	-	132	60.07	55.9
T0405-D2	-	132	22.48	22.48
T0443-D1	3dee	91	47.35	44.32
T0443-D2	3dee	91	40.42	37.08
T0465-D1	3dfd	214	36.72	34.11
T0476-D1	2k5c	250	47.99	33.05
T0482-D1	2k4v	270	52.61	51.87
T0496	3do9	178	29.43	29.43
T0496-D1	3do9	178	37.29	32.71

*例えば T0397-D1 は CASP8 で使われたターゲット蛋白質 T0397 のドメイン 1 を表す。
**ランジュバン MD によって生成された構造の中から有効エネルギーとフラグメント整合スコアの併用によって選ばれた 5 つの構造の GDTTS のうちの最大値。

２つの選別方法、すなわち有効エネルギーのみによる選別と、有効エネルギーとフラグメント整合スコアの併用による選別の能力を比較するため、図 1 に２つの選別の結果を比較した。図 1 の縦軸は、CASP8 の全サーバーモデル構造の持つ最も高い GDTTS であり、各ターゲットドメインの構造予測の容易さ/難しさの基準となっている。ランジュバン MD はサーバーチームとして CASP に参加しておらず、縦軸の値の導出には関係していない。試された 10 ターゲット中 7 ターゲットに対して、有効エネルギーとフラグメント整合スコアの併用が有効エネルギーのみで選別する方法より高い GDTTS を出した。試された 10 ターゲット中 1 ターゲットに対して、有効エネルギーのみでの選別が有効エネルギーとフラグメント整合スコアの併用より高い GDTTS を出したが、残りの 9 ターゲットについては、有効エネルギーとフラグメント整合スコアの併用が高い GDTTS を出した。図 1 に示したように、エネルギーとフラグメント整合スコアの併用によって選ばれた 2 構造が、CASP8 の全サーバーのベストモデルよりも高い GDTTS を出した。この様に、ランジュバン MD に

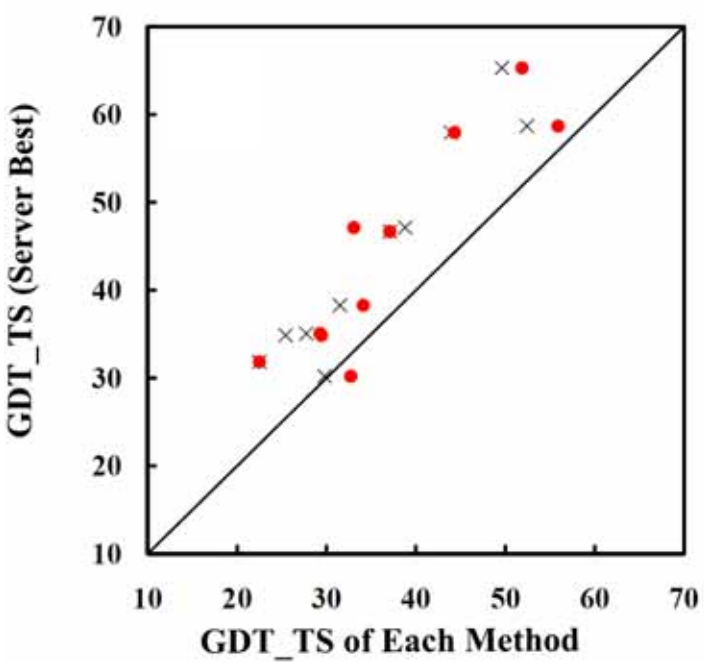


図 1　２つの選別方法、すなわち有効エネルギーのみによる選別と、有効エネルギーとフラグメント整合スコアの併用による選別の能力比較。横軸は、表 1 に示された FM ターゲットに対して、各々の方法で選別された 5 つのモデル構造の最大の GDTTS の値。縦軸は CASP8 のサーバー予測によって生成された全モデル構造の最大の GDTTS。有効エネルギーのみによって選ばれた構造の GDTTS (×印)と有効エネルギーとフラグメント整合スコアの併用によって選ばれた構造の GDTTS (赤丸)。

よる有効エネルギーとフラグメント整合スコアの併用が、デノボ予測の能力向上のために効果的である。

図 2 では、有効エネルギーとフラグメント整合スコアの併用による選別と CASP8 に参加した他の MQA 法の能力を比較するために、有効エネルギーとフラグメント整合スコアの併用によって選ばれた 5 つのモデルのうちの最大の GDTTS が、CASP8 でトップランクの成績を収めたメタサーバーグループによって選ばれたモデル [23]の GDTTS と比較されている。メタサーバーグループは、各グループの開発した MQA 法を用いる事によって CASP8 のサーバーの提出した構造候補からモデル構造を選別して、さらに、各グループ独自のシミュレーションにより、選んだ構造を改良している。例えば Circle と呼ばれる方法[24,23] は、極性側鎖が溶媒に露出している分率と、他の原子によって埋もれている領域の分率、および 2 次構造形成傾向を指標として、1D-3D 関係によって導かれたコンセンサスに基づかない MQA を使用した。GeneSlico [23] や Pcons [25-27] などのメタサーバーグループは、他の MQA 法 [28]と同様なコンセンサスに基づく方法を使用している。FCS 法はコンセンサスに基づかず、本章では、一つの予測法によって生成された構造群のみから選別しているのにも関わらず、有効エネルギーとフラグメント整合スコアの併用による選別で得られた構造

は、Circle (コンセンサスに基づかないメタサーバー方法)の構造より高い GDTTS を持っていて、GeneSilico (コンセンサスに基づくメタサーバー方法)の構造と類似するか、少し高い GDTTS を持っている。

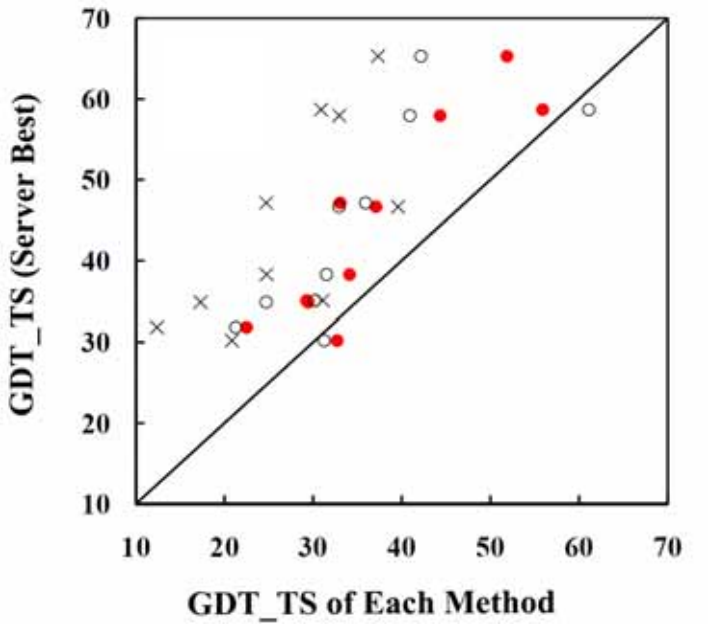


図 2　ランジュバン MD によって生成された構造の中から、有効エネルギーとフラグメント整合スコアの併用によって選ばれた構造の GDTTS (赤丸) が CASP8 に参加した他のメタサーバーによる構造と比較されている。Circle (×印)、GeneSilico (塗りつぶされていない丸)。

図 3 に、いくつかの FM ターゲットに対して、有効エネルギーとフラグメント整合スコアの併用によって予測された構造と CASP8 の全サーバーの予測した構造のうち、最も GDTTS の高い構造が比較されている。有効エネルギーとフラグメント整合スコアの併用法が難しい FM ターゲットの予測を大幅に進展させているのだが、図 3 から、予測された構造の GDTTS はそれでもまだ低くて、FM ターゲットについての一貫性のある体系的な予測法が未だに開発されていないことが理解できる。表 1 で見られた様に、粗視化したランジュバン MD 法によって生成された構造の中に、FCS 法によって選ばれた構造よりも、良い構造がまだ存在していることがわかるが、更に改善された MQA 法をランジュバン MD 法と組み合わせることが、FM ターゲットのための体系的な予測法を開発するのに役立つはずである。

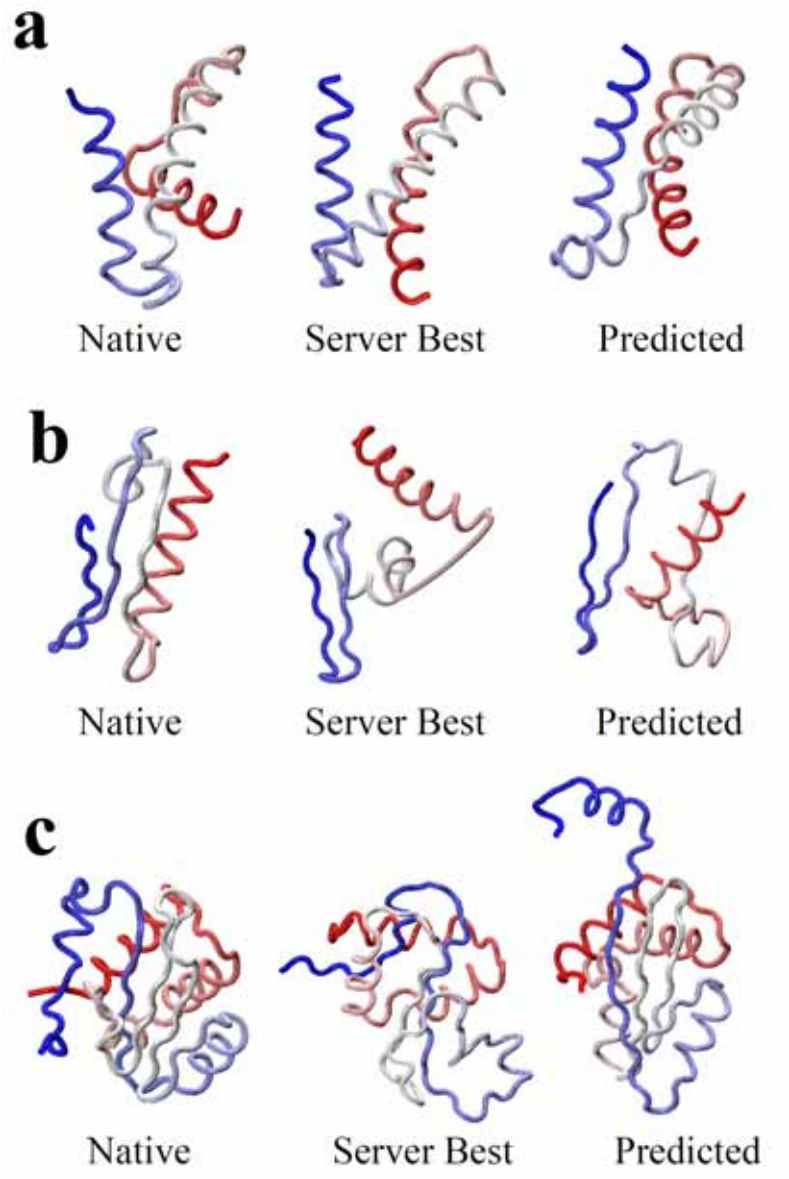


図 3　FM ターゲットに対する予測された構造と、実験的に観察された構造。CASP8 に出題された FM ターゲットの例として、ターゲット T0405 のドメイン 1 (T0405-D1)(**a**)、T0443 のドメイン 2(T0443-D2)(**b**)と T0496 のドメイン 1 (T0496-D1)(**c**)が比較されている。各ターゲットに対する対して、三つの構造が示されている。実験的に観察されたネイティブ構造(左)、CASP8 の全サーバーモデルのうちの最大 GDTTS を持つ構造(中)とランジュバン MD によって生成された構造の中から、有効エネルギーとフラグメント整合スコアの併用によって選ばれた構造(右)。T0405-D1(**a**) に対して、ベストサーバーモデル(中)の GDTTS が 58.68 で平均２乗平方変位（root mean square deviation, RMSD）が 4.33 Å、有効エネルギーとフラグメント整合スコアの併用によって選ばれた構造(右)の GDTTS が 55.90 で RMSD が 4.58 Å。T0443-D2(**b**)に対して、ベストサーバーモデル(中)の GDTTS が 36.73 で RMSD が 9.28 Å、有効エネルギーとフラグメント整合スコアの併用によって選ばれた構造(右)の GDTTS が 37.08 で RMSD が 10.29 Å。T0496-D1(**c**)に対して、ベストサーバーモデル(中)の GDTTS が 30.21 で RMSD が 11.49 Å、有効エネルギーとフラグメント整合スコアの併用によって選ばれた構造の GDTTS が 32.71 で RMSD が 10.81 Å。

第 II 章の主な目的は、新しく考案した MQA 法が実際にデノボ予測を進展させる事を示す事であったが、もっと一般的に FCS 法の能力を議論すべきであろう。CASP8 の FM と TBM カテゴリーの両方を含む合計 194 個のターゲットに対して、CASP ではサーバーチムがモデル構造を提出している。図 4 では、これらの全てのモデル構造からフラグメント整合スコア法により最大の FCS を持つ 5 つの構造を選別し、そのうち最大の GDTTS を CASP8 に参加したメタサーバーグループ、Fams-ace2 と Circle [24,23]によって予測されたモデルの GDTTS と比較している。ランジュバン MD はサーバーチームとして CASP に参加していない、つまり、図 1―3 とは異なり、図 4 はランジュバン MD の結果を含んでいないことに注意が必要である。Circle はコンセンサスに基づいていないが、Fams-ace2 は Circle と同じグループによって実行され、全サーバーモデルの中からトップ 10%のモデルを選ぶために、コンセンサスに基づく選択をし、選ばれた構造を改良・再構築して、その後に 5 つのモデルを選ぶために Circle を使用している[24,23]。

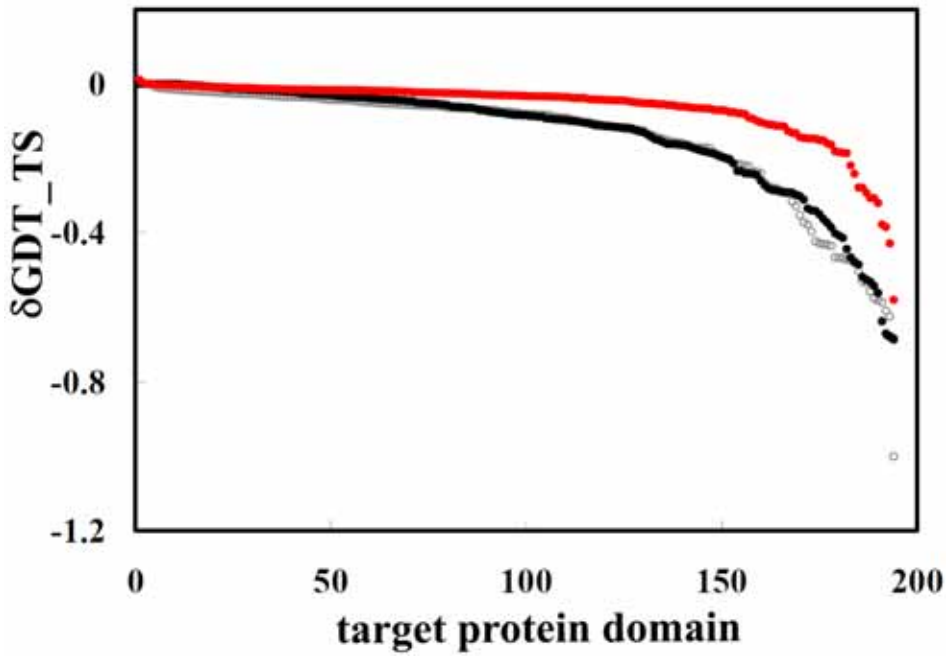


図 4　CASP8 の全サーバーモデルの中から選ばれたモデル構造の GDTTS の比較。CASP8 の 184 個の TBM ターゲットと 10 個の FM ターゲットに対して、フラグメント整合スコアによって選ばれた構造の GDTTS（黒丸）が CASP8 のメタサーバーによって選ばれた構造の GDTTS と比較されている。Circle（塗りつぶされていない丸）、Fams-ace2（赤丸）。全サーバーモデルの最大 GDTTS を GDTTS(ベスト)と書き、各方法によって予測された 5 つのモデルの最大 GDTTS を GDT_TS(予測)と書いて、194 ターゲットの δGDTTS ＝｛ GDT_TS(ベスト) – GDT_TS(予測) ｝/ GDT_TS(ベスト)を δGDTTS の値の順番にプロットしている。

図 4 が示すように、FCS 法の能力はコンセンサスに基づかない Circle と同じか少し高いが、コンセンサスに基づく Fams-ace2 の能力が、FCS 法と Circle に比べて高いことがわかる。これは、コンセンサスに基づく方法がコンセンサスに基づかない方法に勝るという一般的な傾向 [29]と一致する結果である。Fams-ace2 は、Circle にコンセンサスによる評価を加えた方法であることから推測すると、FCS 法とコンセンサスに基づく方法の併用が、TBM 問題に対する役立つ MQA を与えると予想される。

Ⅱ.4 結論

第 II 章の中で、我々は複数の尺度を組み合わせて用いると、提案された予測モデル構造の品質を評価する能力が高くなり、実際に蛋白質構造のデノボ予測の改良につながることを示すことができた。複数の尺度として、我々は粗視化したランジュバン MD によって計算された有効エネルギーとフラグメント整合スコアを使用した。フラグメント整合スコアは、多数の違う予測法の間のコンセンサスに基づかないので、CASP 以外の多数の違う予測法が利用できない状況で有力な方法となり得るし、ターゲットのテンプレートの存在を推定しないため、デノボ予測に適した方法であると言うことができる。フラグメントのローカル構造環境の多残基に及ぶ特徴をうまくとらえることが、適切なスコア法を開発する基礎であったことを考えると、良いフラグメントを使う事がフラグメント整合スコア法を改善するはずなので、フラグメントの収集法の改良がスコア法の改良をもたらすかどうかをみる事が興味深い。

引用文献

1. Anfinsen C.B: Principles that govern the folding of protein chains, *Science*. 181:223-230 (1973).
2. Rohl C.A, Strauss C.E.M, Misura K.M.S, Baker D: Protein structure prediction using Rosetta, *Methods Enzymol*. 383:66–93 (2004).
3. Bradley P, Misura K.M.S, Baker D: Toward high-resolution de novo structure prediction for small proteins, *Science*. 309:1868–1871 (2005).
4. Lee J, Kim S.Y, Lee J: Protein structure prediction based on fragment assembly and parameter optimization, *Biophys. Chem*. 115:209–214 (2005).
5. Fujitsuka Y, Chikenji G, Takada S: SimFold energy function for de novo protein structure prediction: consensus with Rosetta, *Proteins: Struct. Funct. Bioinform*. 62:381–398 (2006).
6. Ishida T, Nishimura T, Nozaki M, Terada T, Nakamura S, Shimizu K: Development of an ab initio protein structure prediction system ABLE, *Genome Inform*. 14: 228-237 (2003).
7. Zhang Y, Arakaki A.K, Skolnick J: TASSER: an automated method for the prediction of protein tertiary structures in CASP6, *Proteins: Struct. Funct. Bioinform*. 61:91–98 (2005).
8. Zhou H, Skolnick J: Ab initio protein structure prediction using Chunk-TASSER, *Biophys J*. 93:1510-1518 (2007).
9. Wu S, Skolnick J, Zhang Y: Ab initio modeling of small proteins by iterative TASSER simulations, *BMC Biol*. 5:page number not for citation purpose (2007).
10. Zhou H, Pandit S.B, Lee S.Y, Borreguero J, Chen H, Wroblewska L, Skolnick J: Analysis of TASSER-based CASP7 protein structure prediction results, *Proteins*. 69: 90-97 (2007).
11. Go N: Theoretical studied of protein folding, *Ann. Rev. Biophys. Bioeng*. 12:183-210 (1983).
12. Onuchic J.N, Luthey-Schulten Z, Wolynes P.G.: Theory of protein folding: the energy landscape perspective, *Ann. Rev. Phys. Chem*. 48:545-600 (1997).
13. Chikenji G, Fujitsuka Y, Takada S: Shaping up the protein folding funnel by local interaction lesson from a structure prediction study, *Proc. Natl. Acad. Sci. USA*. 103:3141-3146 (2006).
14. Kazmierkiewicz R, Liwo A, Scheraga H.A.: Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a Monte-Carlo method, *J. Comput. Chem*. 23:715-723 (2002).
15. Nanias M, Chinchio M, Oldziej S, Czaplewski C, Sheraga H.A: Protein structure prediction with the UNRES force-field using replica-exchange monte carlo-with-minimization; comparison with MCM, CSA and CFMC, *J. Comput. Chem*. 26:1472-1486 (2005).
16. Liwo A, Khalili M, Scheraga H.A: Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains, *Proc. Natl. Acad. Sci. USA*. 102:2362-2367 (2005).
17. Papoian G..A, Ulander J, Eastwood M.P, Z. Luthey-Schulten, P. G. Wolynes: Water in protein structure prediction, *Proc. Natl. Acad. Sci. USA*. 101:3352-3357 (2004).
18. C. Hardin, M. P. Eastwood, Luthey-Schulten Z, Wolynes P.G.: Associative memory hamiltonians for structure prediction without homology: alpha-helical proteins, *Proc. Natl. Acad. Sci. USA*. 97:14235-14240 (2000).
19. Sasaki T. N, Sasai M: A coarse-grained Langevin molecular dynamics approach to protein structure reproduction, *Chemical Physics Letters*. 402:102−106 (2005).
20. Sasaki T.N, Cetin H, Sasai M: A coarse-grained Langevin molecular dynamics approach to de novo protein structure prediction, *Biochem. Biophys. Res. Comm*. 369:500-506 (2008).
21. He Y, Chen Y, Alexander P, Bryan P.N, Orban J: NMR structures of two designed proteins with high sequence identity but different fold and function, *Proc. Natl. Acad. Sci. USA*. 105:14412-14417 (2008).
22. Hegler J.A, Lätzera J, Shehuc A, Clementi C, Wolynes P.G: Restriction versus guidance in protein structure prediction, *Proc. Natl. Acad. Sci. USA*. 106:15302-15307 (2009).
23. Available from http://www.predictioncenter.org/casp8/index.cgi .
24. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H: fams-ace: A combined method to select the best model after remodeling all server models, *Proteins:*

Struct. Funct. Bioinform. 69:98–107 (2007).
25. Wallner B, Fang H, Elofsson A: Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller, *Proteins: Struct. Funct. Genet*. 53:534-541 (2003).
26. Wallner B, Elofsson A: Prediction of global and local model quality in CASP7 using Pcons and ProQ, *Proteins: Struct. Funct. Bioinform*. 69:184-193 (2007).
27. Larsson P, Skwark M.J, Wallner B, Elofsson A: Assessment of global and local model quality in CASP8 using Pcons and ProQ, *Proteins: Struct. Funct. Bioinform*. 77:167-172 (2009).
28. Pawlowski M, Gajda M.J, Matlak R, Bujnicki J.M: MetaMQAP: a meta-server for the quality asessment of protein models, *BMC Bioinformatics*. 9:page number not for citation purpose (2008).
29. Cozzetto D, Kryshtafovych A, Tramontano A: Evaluation of CASP8 Model Quality Predictions, *Proteins: Struct., Funct. Bioinform*. 77:157-166 (2009).

第Ⅲ章
フラグメント整合スコア法と他の MQA 法との比較

Ⅲ.1　序論

第Ⅱ章では、フラグメント整合スコアをランジュバンMD法と併用することによって、デノボ構造予測において良い効果を生むことが示された。MQA法としてのフラグメント整合スコアは、ランジュバンMD法に限らず、他の多くの構造予測法と併用できるはずである。本章では、CASPに提出されたサーバーチームの予測構造モデルの品質を、フラグメント整合スコアによって評価し、その評価の能力を、CASPに参加した他MQAチームの結果と比較することによって検討する。

MQAは、予測されたモデル構造と正しいターゲット構造の類似の程度を、ターゲット蛋白質構造の正しい解答を知る前に、見積もるという問題である。提案されたモデル構造をたくさん利用することが可能な場合、的確なMQAはそれらのモデル構造の中からよい候補を選別する助けとなるはずであり、予測能力の向上につながるはずである。例えば、Bartlett と Taylor [1]は、進化的に関連した配列間の、残基ペアの相関に基づいたMQA法を開発して、MQAがデノボ予測法で生成されたモデル構造の内、間違ったモデル構造を区別する事に役立つということを示した。MQAの重要性は、第7回CASP（CASP7）から広く認知されるようになり[2] 、多数のMQA法がCASPのQAカテゴリー部門で比較されてきた[3]。CASPのQAカテゴリーの解析を通じて、品質評価の対象となる予測モデル構造の多くに共通の特徴を抽出する、コンセンサスに基づいた方法 [4-13] が、コンセンサスに基づいていない方法より大幅によい結果を与える事が示された[4-13]。それはおそらく、多数のモデルの比較を通して、違う予測法から生成されたモデルのエラーがノイズとして処理され除去されるからであろうと考えられる。しかしながら、コンセンサスに基づいた方法にもそれ自身の制限がある：予測法の大部分が失敗して、例外的に幾つかの方法だけが良い予測をした時に、コンセンサスに基づいたMQAは、提案されたモデルの集合体から良いモデルを選別することに失敗する。この問題は、特に、多数の予測方法がターゲットとテンプレートの間の配列相同性に頼るが、正しいターゲット構造がどのテンプレートにも類似しない時に、発生する。これはデノボ問題を扱うときに頻繁に遭遇する状況である。さらに、第Ⅱ章で説明したようなランジュバンMD法の様に、一つ、あるいは少数の予測法だけを用いる事ができる時は、同じ方法を通して生成される多数のモデルが同じ間違った見方のエラーを持っているため、コンセンサスに基づいた方法は、CASPの様

な多数の違う予測サーバーのモデルが利用可能な場合と同じようには、うまく機能しない可能性がある。こうしたことを考えると、コンセンサス解析に頼らないMQA法を開発する価値があるはずである。そのようにして改良されたコンセンサスに基づいていない方法と、他のコンセンサスに基づいた方法 [7]を併用することにより、デノボ予測のための新たなMQA法を開発することができると期待される。コンセンサスに基づいていない方法は、文献の中でシングル・モデル方法と呼ばれることが多いが[3, 9, 11, 12, 14]、本論文においては、各モデル構造の品質を評価する方法の能力を試みるわけではなく、主に、コンセンサスに基づいていない方法が多数の構造の中から良いモデルをどう選別するかに集中するので、我々はここでは、その呼び方を避ける。

これまで、コンセンサスに基づいていない多種の MQA 法が開発された。残基ペアやコンタクトの形成確率からスコア関数を導く方法[15-18]や、配列アラインメントを通して蛋白質間のホモロジーを活用する方法[9,19,20]、配列と構造の対応から導かれた経験的なスコア関数を用いる方法[21-23]など、多くの例がある。第Ⅰ章では、デノボ予測法を進展させるために、コンセンサスに基づいていない新たな MQA 法として、フラグメント整合スコア（FCS）法を提案してテストし、その実行方法を吟味した。本章では、FCS 法の能力をその他のコンセンサスに基づいた方法、基づかない方法と比較し、検討する。

Ⅲ.2　方法

第Ⅰ章に説明された、FCS 計算方法の選択によって、最も良い方法は *RCF-FC-FWM* 法であることが示された。第Ⅱ章の CASP8 のターゲットについての計算と同様に、本章でも、その方法を用いた。

Ⅲ.3　結果・考察

Ⅲ.3.1　ローカルフラグメント整合スコア

第I章の式(2)で説明された様に、FCSはローカルフラグメント整合スコア、LFCSの合計であるので、FCSと他のMQA法の比較の前に、LFCSの能力を試みる事が有意義だと思われる。我々は、CASP8に出題された12個のFMターゲットの中から、CASP8のサーバーモデルのうち最もよいモデル構造が比較的高いGDTTSを示した4つのターゲットを例として選んだ。表1には、それらの4例のターゲ

ットに対して、全てのサーバーモデルの中から5つの最も高いFCSを持つモデルを選び、テストした結果が示されている。選ばれた各モデルに対して全てのフラグメント構造を抽出し、抽出した*i*番目のフラグメント構造と、正しい解答構造の*i*番目のフラグメント構造の間のRMSDを$RMSD_i$と書くことにする。モデル構造の残基数をNとして、$i = 4, ..., N–4$である。$LFCS_i$ と$RMSD_i$ のPearson相関係数をP_{model} と定義する。表１からは、GDTTSが比較的高いモデルが高いP_{model}を持つことがわかる。すなわち、モデルが良いフラグメントを含む場合に、LFCSがローカル構造を区別する能力があることが想像できる。第IV章では、モデル構造のコイルやループ領域の品質を評価するために、FCSの能力を試めす事によって、ローカル構造評価に関する能力がさらにテストされる。

表1　各モデル構造におけるLFCSの性能

CASP8の4つのFMターゲットの例について、全てのサーバーモデル構造から、最も高いFCSを持つ5つのモデルが選ばれた。各モデルの全フラグメントのLFCSとRMSDを比較して、相関係数P_{model}**が計算された。**

ターゲット	モデル	P_{model}	モデルのGDTTS
T0405_D1	FALCON2	0.484	55.21
	PSI4	0.484	55.21
	PSI1	0.418	57.99
	Zhang-Server2	0.383	41.67
	Zhang-Server5	0.348	43.40
T0416_D2	Zhang-Server5	0.603	66.23
	SAM-T08-server2	0.473	36.84
	fais-server5	0.353	52.19
	Zhang-Server1	0.342	32.46
	Zhang-Server3	0.257	34.65
T0443_D1	MUFOLD-Server1	0.432	31.82
	MUFOLD-Server4	0.431	31.82
	Zhang-Server2	0.349	38.64
	Zhang-Server4	0.295	53.41
	Zhang-Server3	0.273	41.29
T0513_D2	Zhang-Server2	0.710	51.45
	FEIG3	0.621	44.20
	Zhang-Server3	0.555	44.56
	BAKER-ROBETTA4	0.537	43.84
	MULTICOM-RANK5	0.264	40.94

Ⅲ.3.2　他のMQA法との比較

FCS法と他のMQAの方法の性能を比較するために、FCS法が、CAS7とCASP8のサーバーモデルに適用された。使用されたターゲットは、CASP7の18個のFMターゲットドメインとCASP8の12個のFMターゲットドメインである。これらのターゲットについては、付録Dの表D1と表D2を参照していただきたい。CASP7の18個のターゲットは、第I章において、FCSの方法の調整のために使われたターゲットと同じである。つまり、CASP7のターゲットはトレーニング・セットで、CASP8のターゲットはテスト・セットとみなすことができる。

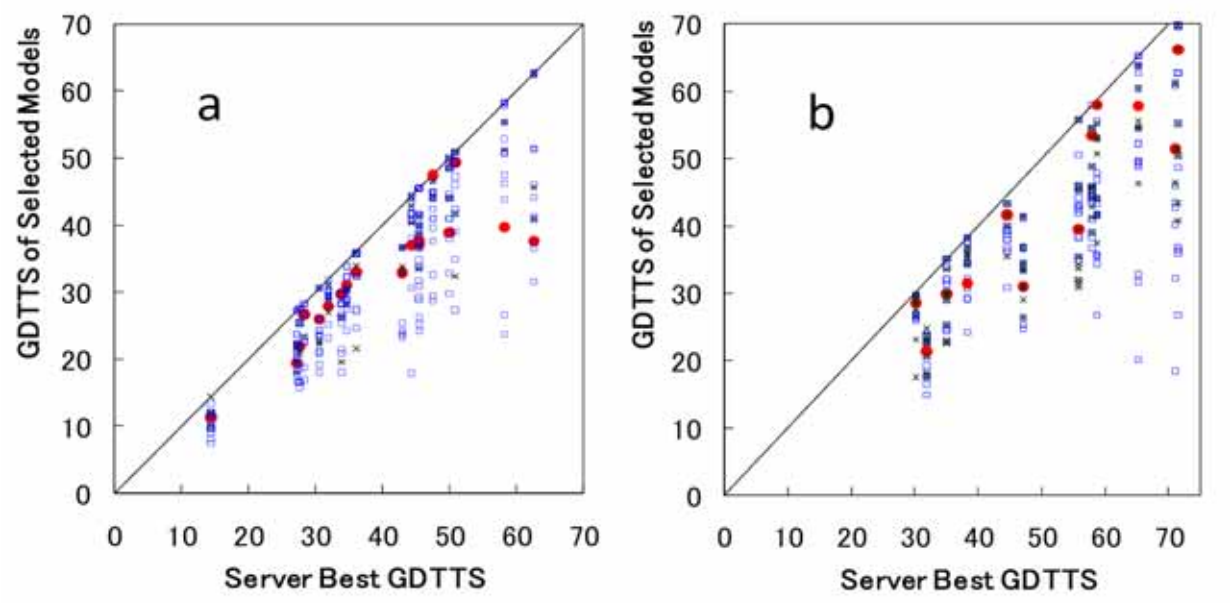


図2　GDDTSに基づくFCS法と他のMQAの方法の比較。各MQA法によって選ばれたモデルの最大のGDTTSによって、FCS法と他のMQAの方法を比較した。横軸は各ターゲットの全サーバーモデルのうちの最大のGDTTSの値である。各ターゲットに対する、コンセンサスに基づくMQA法によって得られたGDTTS (X)、FCS 法によって得られたGDTTS（赤丸）、そしてコンセンサスに基づかないMQAの方法によって得られたGDTTS（□）。FCS法はCASP7に参加したMQAチーム(**a**) とCASP8に参加したMQAチーム(**b**)の方法と比較されている。

図2では、CASPのQAカテゴリーに参加したすべてのMQAチームによる評価結果が、FCS法による評価結果と比較されている。それぞれのMQA法で選ばれた5つのモデル構造のうち、最大のGDTTSを持つモデル構造を$a^{best}(n, \mu)$ と書く。ここで n はターゲットを区別する指標であり、μ はFCS法やその他のMQA法を区別する指標である。図2a には、CASP７の22のMQAチームによる $a^{best}(n, \mu)$ のGDTTS がCASP7の18個のFMターゲットドメインについてプロットされており、図2b

にはCASP8の41のMQAチームによる $a^{best}(n, \mu)$ のGDTTS がCASP8の12個のターゲットドメインについてプロットされている。図2の横軸は、各ターゲット蛋白質に対するサーバーモデル構造のうちの最大のGDTTSを示す。これを *GDTTS*(*n*, best_server) と書く。

FCS法がうまく機能しない幾つかの例があるが、図2から、CASP7においてもCASP8においても、FCS法は他の能力が高いMQA法に匹敵する成績を示すことが分かる。CASP7においては、*GDTTS*(*n*, best_server) > 55であるターゲット(T0348とT0350)では、テンプレートのホモロジー探索と併用するQAの方法はテンプレートを使わないFCS法より良い結果が得られる。しかし、*GDTTS*(*n*, best_server) < 55である難しいターゲットについては、FCS法は他の方法と同程度、あるいはそれ以上の成績を示す。

MQA法の能力を相関係数によって比較することができる。ターゲット n のモデルa を方法 μ によって計算したMQAスコアを*Score*(a, n, μ)と書き、ターゲット n のモデルa のGDTTSを*GDTTS*(a, n)と書いたとき、すべての a 、つまり各ターゲットの可能な全サーバー・モデルの集合にわたって計算した*Score*(a, n, μ)と*GDTTS*(a, n)の間のPearson相関係数を*CC*(n, μ)と書く。図3には、

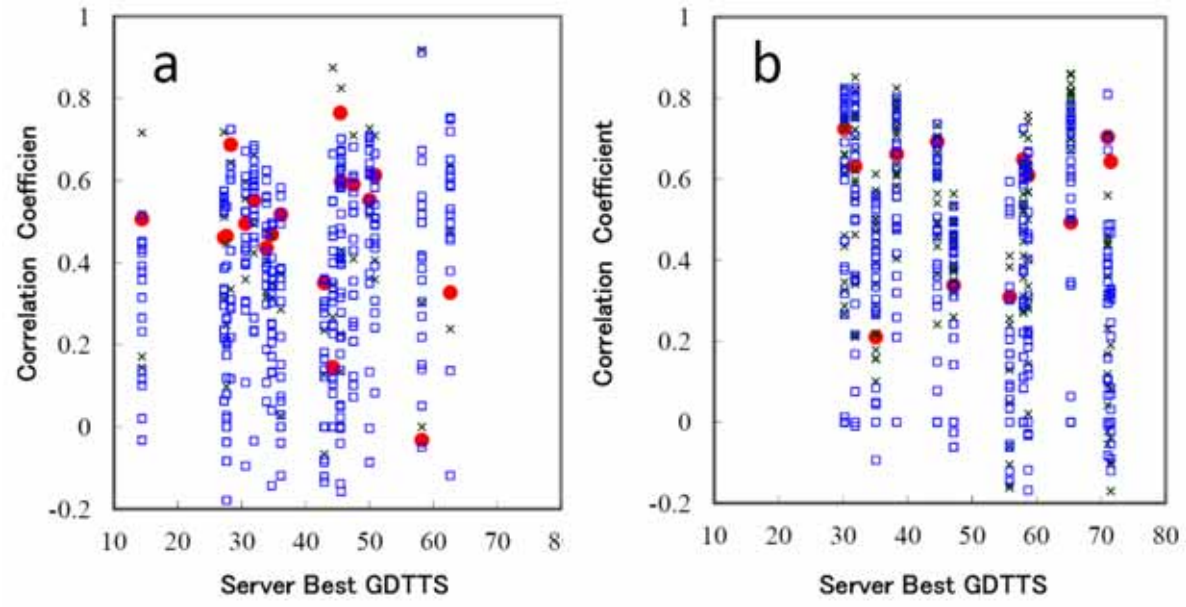


図3　相関係数に基づくFCS法と他のMQAの方法の比較

MQAスコアとGDTTSの間のPearson相関係数に基づき、FCS法とCASP7、CASP8に参加した他のMQA法が比較されている。横軸は各ターゲットの全サーバーモデルのうちの最大のGDTTSの値である。各ターゲットに対する、コンセンサスに基づくMQA法によって得られたGDTTS (X)、FCS 法によって得られたGDTTS（赤丸）、そしてコンセンサスに基づかないMQAの方法によって得られたGDTTS（□）。FCS法はCASP7に参加したMQAチーム(**a**) とCASP8に参加したMQAチーム(**b**)の方法と比較されている。

CASP7とCASP8のFMターゲットについての$CC(n, \mu)$がプロットされている。CASPのヴェブ・ページでは、MQA法の相関係数は蛋白質全体のMQAスコアとモデルの部分的なドメインのGDTTSを比較して計算されている。それで、我々もここで、ドメイン・ターゲットの$CC(n, \mu)$を得るために同じ評価をした。図3は図2と同じ様な結果を示している。

図2と図3の結果が表2と表3に要約されている。表2から、多くのターゲットに対して、$CC(n, FCS)$は、CASPに参加したMQA法の平均より良い事がわかる。図3と表2でわかる様に、FCS法は、特に、CASP7(T0321_D2は例外) の$GDTTS(n, \text{best_server}) < 55$のターゲットについて、あるいは、T0397_D1, T0476_D1とT T0482_D1を除いた他のCASP8のターゲットについて、能力の高い他のMQA法と同程度の成績を示している。

表3はFCS法と、CASP7とCASP8で使用された幾つかのトップクラスのMQA法との比較である。表3の中で、特に、FCSと似た方法を用いるTASSER（CASP8でID# 057）[13]のよい性能を注目すべきである。TASSERでは、モデルのフラグメントは、我々のモデルの構造指標の代わりにRMSDを用いて評価されている。さらにTASSERでは、フラグメントの評価に加えて、多くのモデルの間のコンタクトのパターンのコンセンサス・マッチングもスコアの計算に用いられている。表3から、TASSERがFCSより常にいい結果を与える事がわかる。TASSERで使われたコンセンサス技法が、そうしたよい成績の理由かもしれない、FCS法とTASSERのさらなる比較は、どうやって現在のFCS法とコンセンサス・データーを併用するか、という方法についての手掛かりとなるかもしれない。

表2　図2と図3に示されたデーターの要約

各ターゲットについて、FCSによって選ばれた構造のGDTTSとFCSとの相関係数が、他のMQA法による同様な相関係数の全MQA法にわたる平均値、および全MQA法の中の最高値と比較されている。括弧の中には、各ターゲットに使用された全MQA法の中でのFCS法の順位も示されている。太文字で示された数字はMQA法の平均値より高い。

Target		GDTTS FCS (Rank/[# of MQA teams])	Averaged GDTTS of MQA methods	Best GDTTS of MQA methods	Pearson FCS (Rank/[# of MQA teams])	Averaged Pearson of MQA methods	Best Pearson of MQA methods
CASP7	T0287	**26.71** (4/14)	24.31	28.26	**0.688** (2/14)	0.447	0.725
	T0296	**11.36** (12/22)	10.93	14.37	**0.506** (3/22)	0.272	0.717
	T0300	**47.48** (1/21)	41.53	46.63	**0.590** (5/21)	0.375	0.710
	T0304	37.13 (16/22)	37.85	45.55	**0.598** (10/22)	0.507	0.825
	T0307	27.85 (19/23)	29.13	31.91	**0.552** (12/23)	0.497	0.685
	T0309	**31.05** (6/22)	29.95	33.87	**0.469** (3/22)	0.280	0.497
	T0314	**25.97** (5/23)	24.40	30.58	**0.495** (10/23)	0.427	0.672
	T0316_D2	**32.92** (11/18)	32.35	36.33	**0.351** (6/24)	0.110	0.360
	T0319	19.45 (19/23)	21.64	27.22	**0.461** (13/23)	0.434	0.718
	T0321_D2	36.99 (19/21)	40.17	44.26	0.145 (20/24)	0.328	0.875
	T0347_D2	37.68 (12/22)	39.07	45.42	**0.765** (3/24)	0.280	0.626
	T0348	39.75 (20/22)	51.23	58.19	-0.032 (22/23)	0.397	0.919
	T0350	37.64 (20/23)	52.05	62.64	0.327 (20/23)	0.508	0.753
	T0353	38.86 (17/21)	44.19	50.00	**0.552** (11/22)	0.504	0.727
	T0356_D1	21.98 (12/22)	22.48	27.62	0.465 (3/24)	0.111	0.537
	T0361	**29.75** (1/23)	26.50	29.75	**0.436** (10/23)	0.407	0.625
	T0382	**49.37** (9/22)	45.57	50.84	**0.613** (3/22)	0.459	0.709
	T0386_D2	33.02 (14/22)	33.08	35.80	**0.517** (6/24)	0.263	0.583
CASP8	T0397_D1	29.88 (17/35)	30.13	35.06	0.210 (34/42)	0.350	0.613
	T0405_D1	**57.99** (1/41)	44.46	57.99	**0.610** (12/42)	0.370	0.758
	T0405_D2	21.39 (27/39)	21.78	24.76	**0.632** (20/42)	0.552	0.852
	T0416_D2	**66.23** (13/36)	57.12	69.74	**0.643** (2/42)	-0.039	0.489
	T0443_D1	**53.41** (8/40)	46.52	57.95	**0.649** (5/42)	0.460	0.725
	T0443_D2	**41.67** (8/40)	40.72	43.33	**0.693** (9/42)	0.522	0.737
	T0465_D1	31.51 (34/41)	35.32	38.28	**0.662** (23/42)	0.626	0.824
	T0476_D1	31.03 (31/39)	34.94	41.38	0.337 (34/42)	0.384	0.564
	T0482_D1	**57.84** (23/40)	56.30	65.30	0.493 (38/42)	0.674	0.861
	T0496_D1	**28.54** (24/38)	27.96	29.79	**0.724** (20/42)	0.597	0.826
	T0510_D3	39.53 (20/40)	40.99	55.81	**0.309** (9/42)	0.049	0.594
	T0513_D2	51.45 (31/40)	62.03	71.01	**0.704** (4/42)	0.315	0.810

表3. 代表的なMQA法との比較

CASP7とCASP8で評価されたFMターゲットすべてにわたっての、各MQA法によって得られたPearson相関係数とGDTTSの平均値が比較されている。

MQA team in CASP7	ID # of the MQA team in CASP7	Averaged Pearson	Averaged GDTTS	# of targets assessed	MQA team in CASP8	ID # of the MQA team in CASP8	Averaged Pearson	Averaged GDTTS	# of targets assessed
Circle-QA	713	0.569	36.24	18	Circle	396	0.635	45.71	12
Pcons	634	0.553	33.24	18	ModFOLD	199	0.607	44.22	12
ProQ	633	0.508	35.26	18	TASSER	057	0.575	45.42	12
GeneSilico	38	0.490	37.51	17	Pcons_ ProQ	469	0.567	41.79	12
Ma-OPUS	91	0.484	34.91	18	**FCS**	**present work**	0.555	42.54	12
FCS	**present work**	0.472	32.50	18	Bilab-UT	325	0.548	45.17	12
ABIpro-h	699	0.469	37.72	18	FAMSD	140	0.523	42.55	12
QA-MODCHECK	703	0.455	32.78	16	SAM-T08-MQAU	365	0.515	43.69	12
ProQlocal	692	0.453	34.15	18	Fiser-QA-COMB	177	0.413	42.89	12
CaspIta-FRST	717	0.390	32.61	17	MULTICOM	453	0.327	43.40	12
Bliab	178	0.359	35.54	18	DISTILLF	117	0.307	40.80	12
QA-ModFOLD	704	0.350	36.12	18	MODCHECK-Jury	052	0.266	36.38	12
LEE	556	0.333	29.59	18	LEE	407	0.214	39.83	12
Jones-UCL	13	0.310	32.47	16	MODCHECK-HD	094	0.063	33.49	12

Ⅲ.4 結論

この論文で、我々は、個別の残基の特徴によってのみならず、有限の長さのフラグメントのローカルな構造環境を考慮に入れて予測モデル構造の品質を評価する、FCS 法という新しい MQA 法を開発した。現段階の結果から、評価のスコアを計算する際に、フラグメントのローカル構造環境の多残基の特徴を適切に平均すれば、FCS 法はデノボ予測を改善するために有効であることが示された。実際、第 II 章に示したように、FCS 法はランジュバン MD 法と併用することで効果を発揮する。もっと良いフラグメントの使用は FCS 法を更に進展させるはずであり、フラグメント構造を物理的な方法によって精密化すれば FCS 法の改善がもたらされるかどうかを調べる事は興味深い問題である。FCS 法と他のコンセンサスに基づく方法の適切な併用が、モデルの品質評価技法の更なる改善に向けて有用であると予想される。

引用文献

1. Bartlett GJ, Taylor WR: Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction, *Proteins: Struct., Funct. Bioinform.* 71:950–959 (2008).
2. Available from http://predictioncenter.gc.ucdavis.edu/ .
3. Cozzetto D, Kryshtafovych A, Tramontano A: Evaluation of CASP8 Model Quality Predictions, *Proteins: Struct., Funct. Bioinform.* 77:157-166 (2009).
4. Ginalski K, Elofsson A, Fischer D, Rychlewski L: 3D-Jury: a simple approach to improve protein structure predictions, *Bioinformatics*. 19:1015–1018 (2003).
5. Wallner B, Fang H, Elofsson A: Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller, *Proteins: Struct. Funct. Genet*. 53:534-541 (2003).
6. Wallner B, Elofsson A: Prediction of global and local model quality in CASP7 using Pcons and ProQ, *Proteins: Struct. Funct. Bioinform*. 69:184-193 (2007).
7. Larsson P, Skwark M.J, Wallner B, Elofsson A: Assessment of global and local model quality in CASP8 using Pcons and ProQ, *Proteins: Struct. Funct. Bioinform*. 77:167-172 (2009).
8. DeRonne K.W, Karypis G: Improved estimation of structure predictor quality, *BMC Structural Biology*. 9: page number not for citation purpose (2009).
9. Archie J.G, Paluszewski M, Karplus K: Applying Undertaker to Quality Assessment, *Proteins: Struct. Funct. Bioinform*. 77:191-195 (2009).
10. Benkert P, Künzli M, Schwede T: QMEAN server for protein model quality estimation, *Nucleic Acids Res*. 37:510-514 (2009).
11. Cheng J, Wang Z, Tegge A.N, Eickholt J: Prediction of global and local quality of CASP8 models by MULTICOM series, *Proteins: Struct. Funct. Bioinform*. 77:181-184 (2009).
12. McGuffin L.J: Prediction of global and local model quality in CASP8 using the ModFOLD server, *Proteins: Struct. Funct. Bioinform*. 77:185-190 (2009).
13. Zhou H, Skolnick J: Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential, *Proteins: Struct. Funct. Bioinform*. 71:1211–1218 (2008).
14. Wang Z, Tegge A.N, Cheng J: Evaluating the absolute quality of a single protein model using structural features and support vector machines, *Proteins: Struct. Funct. Bioinform*. 75:638–647 (2009).
15. Sippl M.J: Knowledge-based potentials for proteins, *Curr. Opin. Struct. Biol*. 5:229–235 (1995).
16. Panchenko A.R, Marchler-Bauer A, Bryant S.H: Combination of threading potentials and sequence profiles improves fold recognition, *J. Mol. Biol*. 296:1319–1331 (2000).
17. Wallner B, Elofsson A: Can correct protein models be identified?, *Protein Sci*. 12:1073–1086 (2003).
18. Paluszewski1 M, Karplus K: Model quality assessment using distance constraints from alignments, *Proteins: Struct. Funct. Bioinform*. 75:540–549 (2009).
19. Archie J, Karplus K: Applying undertaker cost functions to model quality assessment, *Proteins: Struct. Funct. Bioinform*. 75:550–555 (2009).
20. Kalman M, Ben-Tal N: Quality assessment of protein model-structures using evolutionary conservation, *Bioinformatics*. 26:1299-1307 (2010).
21. Eisenberg D, Lothy R, Bowie J: VERIFY3D: Assessment of protein models with three-dimensional profiles, *Methods Enzymol*. 277:396-404 (1997).
22. Zhou H, Zhou Y: Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition, *Proteins: Struct. Funct. Bioinform*. 55:1005–1013 (2004).
23. Shirota M, Ishida T, Kinoshita K: Absolute quality evaluation of protein model structures using statistical potentials with respect to the native and reference states, *Proteins: Struct. Funct. Bioinform*. 79:1550–1563 (2011).

第Ⅳ章

予測の難しい TBM 構造への応用

Ⅳ.1 序論

FCS 法には第Ⅱ章・第Ⅲ章で示した FM ターゲットの予測以外にも、応用できる余地がある。この章で述べる 2 つの例は、ループやコイルといった不規則な構造の予測、および配列が似ているが構造が大きく異なる人工的に設計された２つの蛋白質の構造の識別である。CASP においては、これらはいずれも TBM ターゲットのカテゴリーに属する構造であるが、FM ターゲットと同様の難しさを持ち、新しい MQA 法の応用が期待される。

Ⅳ.1 .1 ループおよびコイルの構造予測

ループやコイルは、典型的な 2 次構造に分類されないため、蛋白質全体についてホモロジーが見つかっても、蛋白質に含まれるループやコイル部分に相当するテンプレートを発見することが難しい。従って、TBM カテゴリーに分類されるターゲット蛋白質においても、ターゲットに含まれるループやコイルの予測構造を得る事は困難であり、コンセンサスに基づく MQA 法は正確な評価に失敗する事が多い。そこで本章では、CASP8 に提出された TBM ターゲット蛋白質の中で、α ヘリックスや β シートなどの決まった 2 次構造を持たない 10 残基以上の長さの領域に注目し、そのうち、多くのサーバーチームによる予測構造の GDTTS 値が低かったループやコイルの領域について、FCS 法がサーバーチームの予測構造の集まりの中から、よりよい構造を選別する能力をもつかどうかを試した。

Ⅳ.1 .2 似た配列が異なる構造をつくる例における構造評価

CASP8 の中で注目すべきターゲットは、T0498 (PDB code: 2kdl)と T0499(PDB code: 2kdm) [1] の、2 つの人工的に設計された蛋白質である。この 2 つのアミノ酸残基配列には 3 箇所でしか違いがなく、残りは同じ残基を持っているが、実験で観測されたこの 2 つのターゲットのネイティブ構造は、大きく異なっていた。この 2 つターゲットに関しては、配列が近くホモロジー関係にあるとみなせる蛋白質が既知構造の中に存在するため、このホモロジーを利用すると、TBM 問題として扱うことができるが、その場合、２つのターゲットのどちらにも、ほぼ同じ構造を予測することになる。

つまり、２つのうち片方の構造について、誤った予測をすることになる。多くのサーバーチームが間違った予測をするため、コンセンサスに基づく MQA 法も間違った評価をしてしまう。このような難しいケースについて、配列のホモロジーに依存せず、コンセンサスに基づかない MQA 法を適用することは興味深い。本章では、そうした MQA 法の例として、FCS 法の適用を試みる。

Ⅳ.2　　方法

第Ⅰ章に説明された、FCS 計算方法の選択によって、最も良い方法は *RCF-FC-FWM* 法であることが示された。本章でも、*RCF-FC-FWM* 法を用いた。

Ⅳ.3　　結果および考察

Ⅳ.3 .1　ループおよびコイルの構造予測

FMターゲットの予測されたモデルを選別するために提案されたMQA法が、TBMターゲットにも応用できるかどうかを試みる事は有意義である。テンプレートの2次構造の形式と並び方がTBMターゲットの正しい解答構造のそれらと、大体の場合、類似しているのにも関わらず、2次構造を繋ぐ長いループや、末端部位に近いコイル領域の様な、不規則な領域の正しい構造が、それらのホモロジー・テンプレートと大きく違う事が頻繁にある。TBM問題の最も困難な部分は、それ故に、これらの不規則な領域を予測する事である[2-4]。我々はここで、そういった領域のモデル構造の選別にFCS法を応用した。

CASP8の全ターゲットの正しい解答構造の座標を調べたところ、通常の2次構造を示さない10残基より長い不規則な領域が175個見つかった。それらの175個の領域の中で70個について、サーバー・モデルの70%以上がGDTTSが70以下の値を持ち、予測困難な領域であることが示された。残りの105個については、TBMの通常の方法、またはFMの方法で予測可能な領域であったことがわかる。

我々は、ξ 番目の領域のフラグメント整合スコアを以下の様に計算した。

$$FCS(\xi) = \sum_{i \in \mathrm{region}(\xi)} LFCS_i \qquad (4)$$

この$FCS(\xi)$ を使用して、上記の70領域を含むターゲットを予測した全サーバーモデルの該当領域

の構造を評価した。*FCS*(ξ) の値が最も大きい、5つのモデル領域構造を選ぶと、70領域中の13領域について*GDTTS* が70以上の値を持つモデル領域構造を選別することができた。さらに不規則領域全体という制限にこだわらず、領域の１部分について、対象とする長さを変えながら構造評価能力を検討したところ、52領域について、*GDTTS* の値が70以上となる領域構造を全サーバーモデルから選別することができた。図４と図５に、FCS法を用いて、全サーバーモデル構造のうち最もGDTTSの大きい構造か、それに近い構造を選別することに成功した例を幾つか示す。

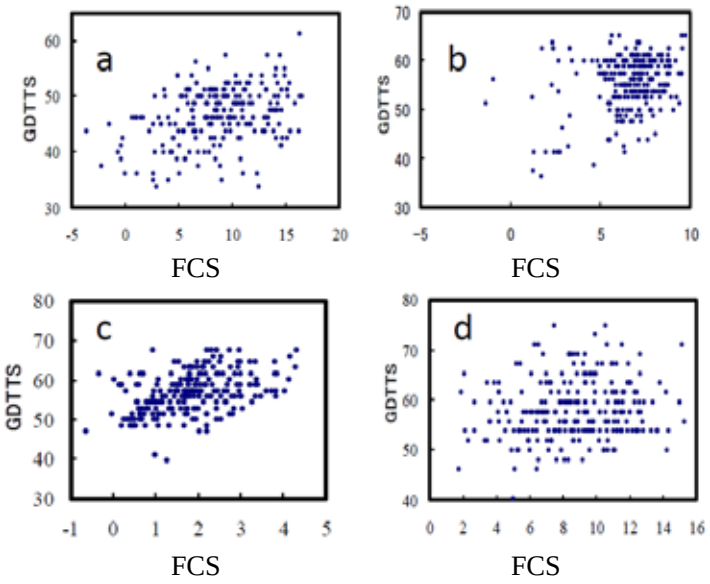


図4. FCS法による不規則な領域のモデル構造の品質評価。CASP8のTBMターゲットの中のループやコイル領域を予測したサーバー・モデル構造の品質をFCS法で評価した。その結果得られた、領域モデル構造のFCSとGDTTSの相関のプロット。(**a**) CASP8のターゲット番号T0413 (PDB code: 3d0k)の蛋白質の174-193番残基、(**b**) T0427 (PDB code: 3d3y)の219-238番残基、(**c**) T0464 (PDB code: 2k5r)の1-17番残基、そして(**d**) T0468 (PDB code: 2k5w)の73-85番残基。各領域の相関係数は0.36(**a**)、0.23(**b**)、0.42(**c**)、0.15(**d**)である。

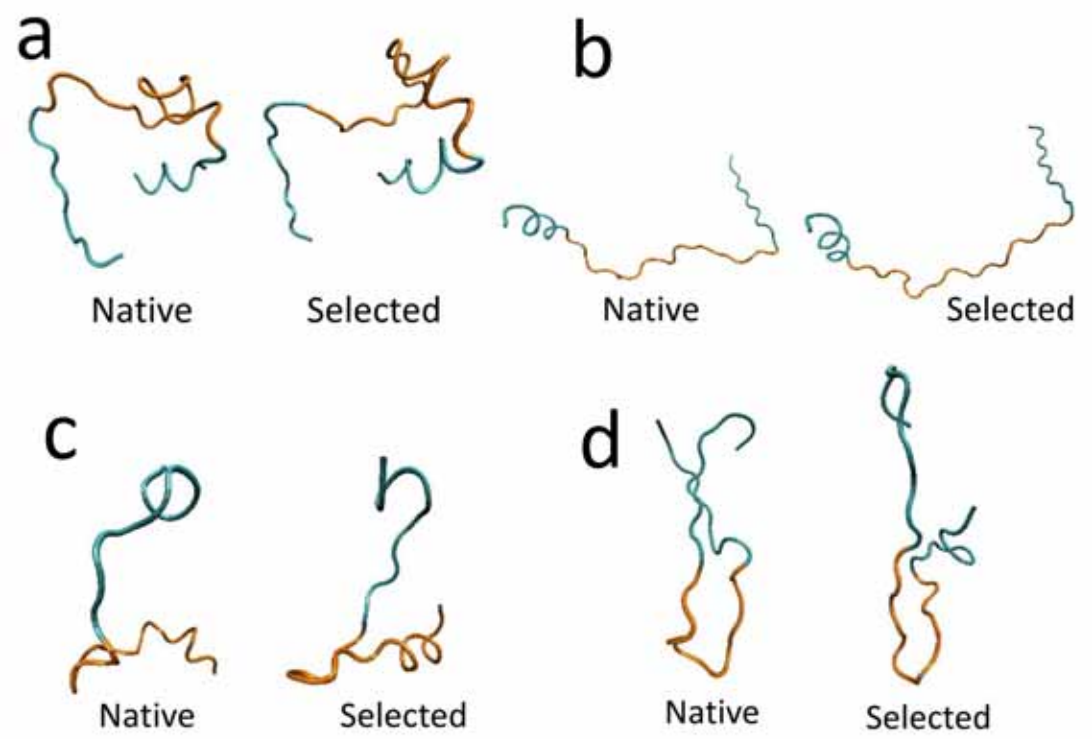


図5. FCS法によって選ばれた不規則な領域のモデル構造。FCS法によって選ばれたループとコイルについての５つのサーバーモデルの中のベストの構造がネイティブ構造と比較されている。オレンジ色に色付けされた領域の品質がFCS法によって評価された。この図a-dのターゲットが図４のa-dと対応している。

図4a-4dは、不規則な領域の例におけるFCSとGDTTSの相関を示している。図4aと図4cでは、相関係数が高くて、図4bと図4dでは比較的低い。図4a、図4b、図4cの例では、FCS法は、図5a-5cで見られる様な最もGDTTSが大きいモデルを選別でき、また、図4dの例については、可能な全部のサーバー・モデルの中で4番目にGDTTSが大きいモデルを選別できた。図5に見るように、それらの選別されたモデルは、ネイティブ構造と詳細には一致していなくて、品質がまだ足りない事がわかるが、FCS法によってテストされたモデルの内では、相対的に良い構造を選別する能力があることが示されたので、この方法はTBM問題の中のループやコイル構造の予測を進展させる効果的な手段として期待できるであろう。

Ⅳ.3 .2 似た配列が異なる構造をつくる例における構造評価

CASP8でTBMターゲットに分類された蛋白質の中で、特に興味深いのは、ある蛋白質の組、T0498とT0499である。T0498とT0499はお互いに3つの残基だけ違うが、構造が大きく違う[1]。T0498は構

造が既に解かれたProtein Gおよびその類縁蛋白質に配列が近いので、T0498はCASPに際してはTBMターゲットに分類されたが、その構造トポロジーは予想されたテンプレートと違い、そのため実際上、デノボ予測が必要な蛋白質である。図6aと図6bに示すのは、それぞれT0498とT0499に対する全サーバー・モデルのFCSとGDTTSである。図6aと図6bの両方で、サーバー・モデルの結果は、高いGDTTSのグループと低いGDTTSのグループの２つのグループに分類される。図6aに示したT0498に関しては、低いGDTTSグループの個数が高いGDTTSグループの個数よりずっと多い事がわかる。そのため、コンセンサスを基準にしたMQA法は、高いGDTTSグループからモデル構造を選別する事に完全に失敗している。他方、図6bに示したT0499に関しては、高いGDTTSグループの個数が低いGDTTSグループの個数より多く、そのため、コンセンサスを基準にしたMQA法がT0499に対して成功している。T0498に関しては、高いGDTTSグループの構造がヘリックス構造を持ち、低いGDTTSグループの構造がシート構造をより多く含んでいる。T0499に関しては、低いGDTTSグループの構造が高いGDTTSグループの構造より、より多くヘリックスを含んでいる。結果がその様に、２つのグループに分裂する可能な理由の一つは、これらの各蛋白質が、ヘリックスを多く含む構造と、Protein G や関連した蛋白質の様にシートを含む構造の、２つの低い自由エネルギーを持つ構造を取り得るためではないかと想像される。それ故に、少数の残基の変異が、これらの２つの構造の自由エネルギーの相対的な深みを変えて、それが構造の分布のバランスの変化を誘導し、結果としてターゲット構造の様に違う解答が得られると予想する事ができるかもしれない。CASP8のサーバー予測のほとんどがターゲットとテンプレートの間の配列アラインメントに頼ったため、Protein Gに類似した構造を持つテンプレートがT0498とT0499の両方に関して選別され、自由エネルギーの深みの変化がそれらの予測方法によって検出されなかった。

T0498に対してFCSは、多くのサーバー・モデルの中から高いGDTTSを持つ少数の構造を良く識別でき、高いGDTTSグループの構造に関してFCSとGDTTSの間に明確な相関が見られる。これらの特徴は、FCS法はT0498に対してモデル品質を評価する事に成功している事を示している。T0499に対してFCS法は良いモデルを選別せず、幾つかの悪いモデルを区別できるのみである。これは、おそらく、正しいβシート構造がFCSでは適切に評価されていないことに起因しており、FCS法の更なる改良の方向を示唆している。第Ｉ章でも述べたように、現在用いているFCS法における２次構造判定率は、βシートに関しては高くないため、この欠点がβシートの構造を含むT0499に関して、悪いモデルを除く能力を示すににとどまった原因と考えられる。FCS法における２次構造判定率を上げれば、この問題の解決につながると期待される。

図6cと6dに示したのは、FCS法で選別された構造である。T0498に対して、FCS 法は図6cの様に良いモデルを選別した。T0499に対しては、図6dに示した様に、FCS法は低いGDTTSグループと高

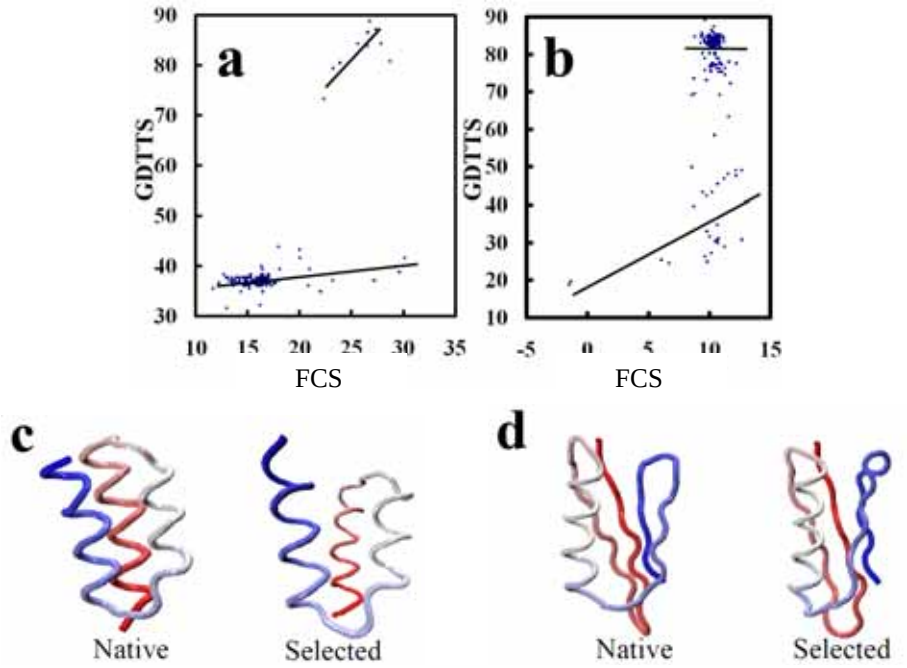


図6. ３つの残基のみ違うように設計された蛋白質の構造トポロジーの違いの、FCS法による識別。GDTTSとFCSが、CASP8の中のサーバーによって提出されるT0498とT0499の全てのモデル構造に対して2次元のプロットで示されている。ターゲットT0498 (**a**)に対して298個、ターゲットT0499(**b**)に対して268個の構造のGDTTSとFCSをプロットしている。**a**と**b**における線の傾きが、GDTTSとFCSの相関係数を意味する。T0498に対して、高いGDTTSを持つグループの相関係数が0.69であり、低いGDTTSを持つグループの相関係数が0.37である。T0499に対しては、高いGDTTSを持つのグループの相関係数が0.04であり、低いGDTTSを持つグループの相関係数が0.55である。T0498（**c**）とT0499（**d**）に対して、実験的に観察されたネイティブ構造(左)とFCS法によって全サーバー・モデルの中から選ばれた5つの構造のうちの最もGDTTSの大きい構造(右)が示されている。T0498 に対して選別された構造のGDTTSは88.44、RMSDは1.49Åである。T0499 に対して、選別された構造のGDTTSは77.68、RMSDでは2.97である。

いGDTTSグループの構造の両方を選別しており、FCSの５つのトップ構造は、高いGDTTSグループの構造を含む。従って、FCS法はまだ不十分ではあるが、FCS法の更なる改良が、コンセンサスを基づいた手法では到達できないここで紹介したような問題についても、より良いMQA法の開発への道を開くと期待される。

Ⅳ.3 結論

FCS法は長いループやコイルなどの不規則な構造について予測されたモデル構造の選別をする能力を持つ。またFCS法は、互いに配列上少しだけ違うが構造が大きく違う人工的に設計された蛋白

質を識別を助ける事もできる。これらの問題は、コンセンサスに基づく方法では扱いが難しく、本論文で提案されたFCS法と他のコンセンサスに基づく方法の併用がモデル品質評価を更に改善するはずである。

引用文献

1. Alexander P.A., He Y, Chen Y, Orban J, Bryan P.N: A minimal sequence code for switching protein structure and function, *Proc. Natl. Acad. Sci. USA*. 106:21149–21154 (2009).
2. Arnautova Y.A, R. Abagyan A, Totrov M: Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling, *Proteins: Struct. Funct. Bioinform*. 79:477–498 (2011).
3. Soto C.S, Fasnacht M, Zhu J, Forrest L, Honig B: Loop modeling Sampling, filtering, and scoring, *Proteins: Struct. Funct. Bioinform*. 70:834–843 (2008).
4. Lee J, Lee D, Park H, E. Coutsias A, Seok C: Protein loop modeling by using fragment assembly and analytical loop closure, *Proteins: Struct. Funct. Bioinform*. 78:3428–3436 (2010).

終章

この論文では、蛋白質のデノボ構造予測という目標に向かって、予測技術を進展させると期待ができる「フラグメント整合スコア法（FCS 法）」という新しい手法を開発し、その能力を評価した。蛋白質の各アミノ酸残基の周りのローカルな構造環境を上手く評価することができれば、FCS 法は、ある程度の範囲で有用な手法である事が、各章を使って示された。

残基の周りのローカルな構造環境の評価法は、アミノ酸配列を有限幅で切ったフラグメントと呼ばれる領域が、蛋白質の中でどのような構造をとっているかを評価するために使われた。フラグメント整合スコアの計算では、ターゲット蛋白質から切り出した 9 残基領域に対応するフラグメントが、ライブラリーに記録されている多数の蛋白質から配列プロファイルの相関によって選ばれているが、こうして選ばれたフラグメントとターゲット蛋白質の 9 残基領域の構造が、ローカルな構造環境の指標を用いて比較され、さらに適切に評価されることにより、ローカルなフラグメント整合スコアが導出される。このローカル・フラグメント整合スコアを蛋白質全体にわたって合計したものが、フラグメント整合スコアである。

従来の構造予測の手法のうち、テンプレートを使う手法 [1, 2] は、ターゲット蛋白質の構造全体が、配列プロファイル比較で見つかるターゲットに類似した蛋白質の既知構造の全体（テンプレート）と似ていることを前提にした手法であるが、FCS 法ではそのような前提を用いていない。テンプレートを使わない構造予測手法のうち、最も広く使われる手法のひとつは、フラグメントアセンブリ法 [3, 4-7] である。フラグメントを蛋白質構造ライブラリーから集める方法においては、FCS 法は多くのフラグメントアセンブリ法とほぼ同じ方法を用いているが、フラグメントアセンブリ法はフラグメントを組み立ててモデル構造を生成する方法であるのに対し、FCS 法は、集めたフラグメントをモデル構造の評価（Model Quality Assessment, MQA）のために用いている。フラグメントをこうした目的のために使うことは、これまでのフラグメントアセンブリ法には無かった、独創的な方法である。

第 I 章では、FCS 法を計算する際のフラグメントの収集方法、構造環境の指標の比較法などにおいて、可能な組み合わせとして考えられる計算方式の中からベストな計算方式を、CASP7 の FM ターゲットを基準にしたテストによって見つける事ができたことが示されている。どの方式がベストか、という判定の結果は、使われたターゲットすべてについて一貫して成り立つ判定であり、また、基準とするターゲットの数を入れ替えてもベストな成績を残すため、こうして選ばれた計算方式は、ターゲット毎に用いるべき計算方式が変わるような一貫性のない手法ではないということが示唆

された。また、2 次構造を判定する際に用いた二面角のコサイン値のしきい値を、3624 個の重複しないライブラリー蛋白質の既知構造を元に計算して決めた。その際、しきい値を最もうまく選んでも、β シートではないコイルやループの領域が β シートとして判定される傾向が残ることが示された。これは、FCS 法が β シート領域を含む構造を評価することが苦手な原因であると考えられ、現在の FCS 法の計算手法の難点とも言うことができる。2 次構造の判定を二面角以外の指標も用いて行えば、β シートを含む構造に対する予測の能力が向上するであろうと期待される。

第 II 章では、FCS 法というバイオインフォマティックな方法とランジュバン分子動力学計算における有効エネルギーという物理的な計算手法を併用することにより、ランジュバン分子動力学計算によって生成されたモデル構造の品質を評価する高い能力が得られる事が示された。このことは、「単独の計算手法によって得られたモデルの品質の評価」というコンセンサスに基づく MQA 手法にとって苦手な問題を扱うことができるという利点を、我々の手法が持っている事を示している。フラグメント整合スコアは、多数の違う予測法によってつくられたモデル構造間のコンセンサスに基づかないので、CASP 以外の多数の違う予測法が利用できない状況でも有力な方法となり得るし、ターゲットのテンプレートの存在を推定しないため、デノボ予測に適した方法であると言うことができる。フラグメントのローカル構造環境の多残基に及ぶ特徴をうまくとらえることが、適切なスコア法を開発する基礎であったことを考えると、良いフラグメントを使う事がフラグメント整合スコア法を改善するはずなので、フラグメントの収集法の改良がスコア法の改良をもたらすかどうかを調べる事が、興味深い問題として残されている。

第 III 章では、CASP7 と CASP8 の FM ターゲットドメインに対して提出された多数のサーバーモデル構造を用い、これらのモデル構造の品質を評価する能力について、CASP7 と CASP8 に参加した多くのチームによる MQA 法を用いた成績と、FCS 法による成績を比較した。比較対象となった多くの MQA 法には、コンセンサスに基づくもの、基づかないものの両方が含まれているが、比較の結果、FCS 法は CASP で良い成績を収めた他のコンセンサスに基づかない MQA と同じ程度の評価能力を持つことが示された。

更に、上述の多数の MQA 法の中から、Circle と Fams-ace2 という二つの手法を例として選び、詳細な検討を行った。これらは、いずれも北里大学のチームによって開発された方法であるが、Circle はコンセンサスを使わないのに対し、Fams-ace2 は Circle を土台にしてさらにコンセンサスによる評価を加えた方法である[8] 。FM と TBM を合わせた CASP8 の全ターゲットドメインに対して、FCS と Circle と Fams-ace2 の予測能力を比較したところ、FCS は Circle よりやや品質評価能力が高いものの、Fams-ace2 には全体的に劣っている。これは「TBM におけるモデル品質評価についてはコンセンサスに基づくチームが断然と有利である」という一般的に見られる傾向に一致した結

果である。Fams-ace2 は Circle にコンセンサスによる評価を加えた方法であることを考えれば、Fams-ace2 の評価能力が大幅に向上したことは興味深い。同様に、FCS 法と他のコンセンサスに基づく方法の適切な併用が、モデルの品質評価技法の更なる改善に向けて有用であると予想される。

第 IV 章では、TBM に分類されているが、予測の難しい構造に対して FCS 法を適用した結果を議論した。CASP8 に出題されたターゲット蛋白質のうち、とくに興味深いのは、アミノ酸配列にわずかの違いしか持たないが、構造が大きく違う 2 つの人工的に設計された蛋白質である T0498 と T0499 である。既知構造とのホモロジーを使った予測法は、配列の似ている T0498 と T0499 に対して同じ構造を予測するため、2 つの大きな構造の違いを予測できない。従って、大多数のサーバーモデル構造が誤ったために、コンセンサス法はうまく働かなかった。コンセンサスに基づかない FCS 法は、T0498 と T0499 の構造を識別するために役立つことが示された。

又、TBM ターゲットの構造予測における最も難しい問題は、正規な 2 次構造に分類できない不規則なループやコイル構造をとる部分の予測である。決まった 2 次構造を持たない不規則な構造であるが故に、品質評価の際にテンプレートを使う事も難しいし、また、多くのサーバーチームがテンプレートを使うので、それらの中のコンセンサスに頼る MQA によるサーバー構造の品質評価は困難である。この困難に着目して、コンセンサスを用いない MQA 法による不規則部分のモデル構造品質評価を行ったのは、本研究の新しい試みである。FCS 法はこの問題においても、ある程度の予測の性能を示した。

以上の通り、フラグメント整合スコア（FCS）法という新規の方法を開発して、その能力を検討した結果、FCS 法はデノボ予測に於いて、コンセンサスに基づかない MQA 法としての機能を持つと共に、単独の計算手法によってのみ生成される構造候補の集まりから良い品質の構造を選ぶ問題や、予測の難しい TBM ターゲットのモデル構造品質評価にも応用できる事が示された。本研究で開発された FCS 法は、蛋白質の構造予測という生物物理学の主たる目標に向かって、もう一歩の前進をするために役立つ、有意義な新しい方法である。

引用文献

1. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: Improving physical realism, stereochemistry and side-chain accuracy in homology modeling Four approaches that performed well in CASP8, *Proteins: Struct. Funct. Bioinform*. 77:114-122 (2009).
2. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A: Evaluation of template-based models in CASP8 with standard measures, *Proteins: Struct. Funct. Bioinform*. 77:18-28 (2009).
3. Chikenji G, Fujitsuka Y, Takada S: Shaping up the protein folding funnel by local interaction lesson from a structure prediction study, *Proc. Natl. Acad. Sci. USA*. 103:3141-3146 (2006).
4. Rohl C.A, Strauss C.E.M, Misura K.M.S, Baker D: Protein structure prediction using Rosetta, *Methods Enzymol*. 383:66–93 (2004).
5. Bradley P, Misura K.M.S, Baker D: Toward high-resolution de novo structure prediction for small proteins, *Science*. 309:1868–1871 (2005).
6. Lee J, Kim S.Y, Lee J: Protein structure prediction based on fragment assembly and parameter optimization, *Biophys. Chem*. 115:209–214 (2005).
7. Fujitsuka Y, Chikenji G, Takada S: SimFold energy function for de novo protein structure prediction: consensus with Rosetta, *Proteins: Struct. Funct. Bioinform*. 62:381–398 (2006).
8. Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H: fams-ace: A combined method to select the best model after remodeling all server models, *Proteins: Struct. Funct. Bioinform*. 69:98–107 (2007).

付録A　２次構造判定のしきい値決定

全ライブラリー蛋白質の全残基について、pdb ファイルに指定された実際の 2 次構造と下記のようなしきい値による判定とを比較した。

2 次構造をαと判定する際のしきい値決定

しきい値α_0、α_1を、$-1<\alpha_0<1$、$-1<\alpha_1<1$の範囲で 0.05 の間隔ごとに選び、成功率 *C* を計算した。成功率 *C* は、

$$C = (A-B)/A \times 100\ (\%)$$

で定義されているが、*A* は、全ライブラリー蛋白質の全残基の中で、実際の 2 次構造がαヘリックスで、しきい値による判定でもαとして判定ができた残基数であり、*B* は、実際に 2 次構造がαではないが、しきい値による判定でαとして判定した残基数である。

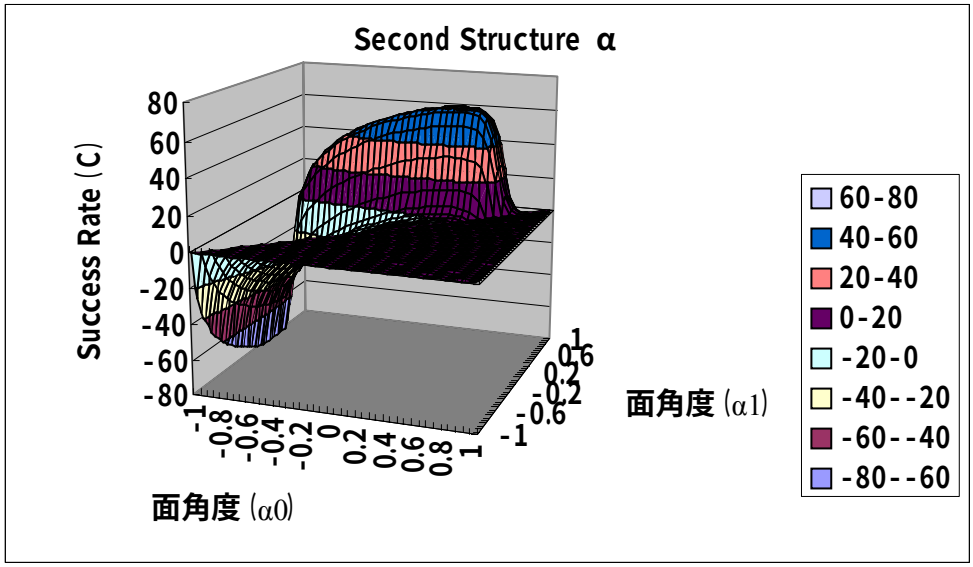


図 A1　しきい値α_0, α_1の大まかな決定。

図 A1 に示す結果のように、*C* が最大となる領域は、$-0.33 < \alpha_0 < 0.37$，$0.81 < \alpha_1 < 0.83$ であったので、その領域に対して、更に、しきい値を 0.01 の間隔で変えて *C* を計算した。その結果が図 A2、および表 A1 に示されている。これらの結果から、しきい値を$\alpha_0=0.35$，$\alpha_1=0.82$と決めるのが最適であることがわかった。

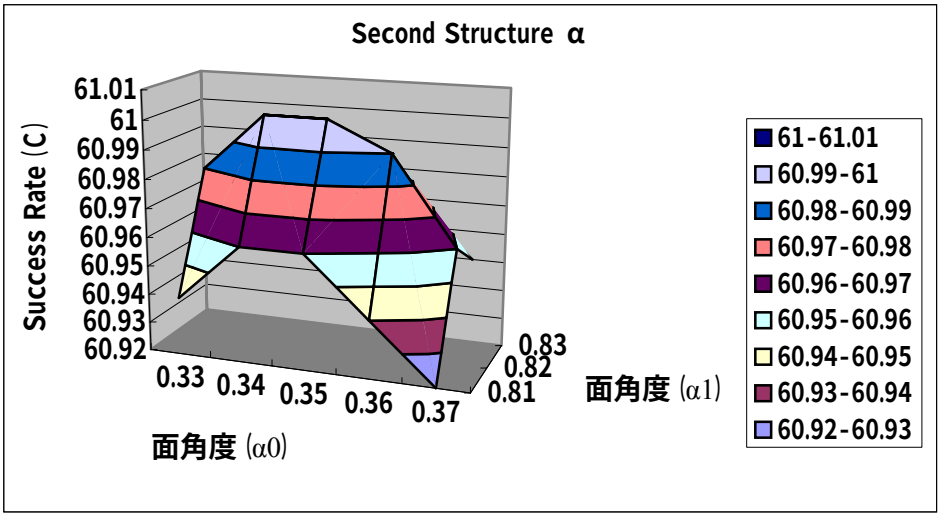


図 A2　しきい値 α_0, α_1 の細かい決定。

表 A1　図 A2 の結果を数字で表した表。

面角度 (α0)	面角度 (α1)	実際にαの残基数	実際にαの残基をαとして判定した残基数 (A)	実際にαではない残基をαとして判定した残基数 (B)	A－B	A－BのAに対する割合 (C)
0.33	0.81	375776	285652	56649	229003	60.94
0.33	0.82	375776	287146	57985	229161	60.98
0.34	0.81	375776	284826	55765	229061	60.96
0.34	0.82	375776	286320	57101	229219	61
0.35	0.81	375776	283927	54850	229077	60.96
0.35	**0.82**	**375776**	**285421**	**56186**	**229235**	**61**
0.35	0.83	375776	286740	57555	229185	60.99
0.36	0.82	375776	284467	55298	229169	60.99
0.36	0.83	375776	285786	56667	229119	60.97
0.37	0.83	375776	284806	55773	229033	60.95

2 次構造をβと判定する際のしきい値決定

α_0 と α_1 を決めたのと同様な方法で β_0 と β_1 を決めた。ただし、全ライブラリー蛋白質の全残基の中で、実際の 2 次構造がβストランドである残基数に対する、実際の 2 次構造がβでしきい値による判定でもβと判定ができた残基数（*A*）の割合を%で表した値を *D* とし、実際の 2 次構造がβストランドではない残基数に対する、実際の 2 次構造がβではないのにしきい値による判定でβとして判定してしまった残基数（*B*）の割合を%で表した値を *E* として、成功率を $D-E$ で定義した。

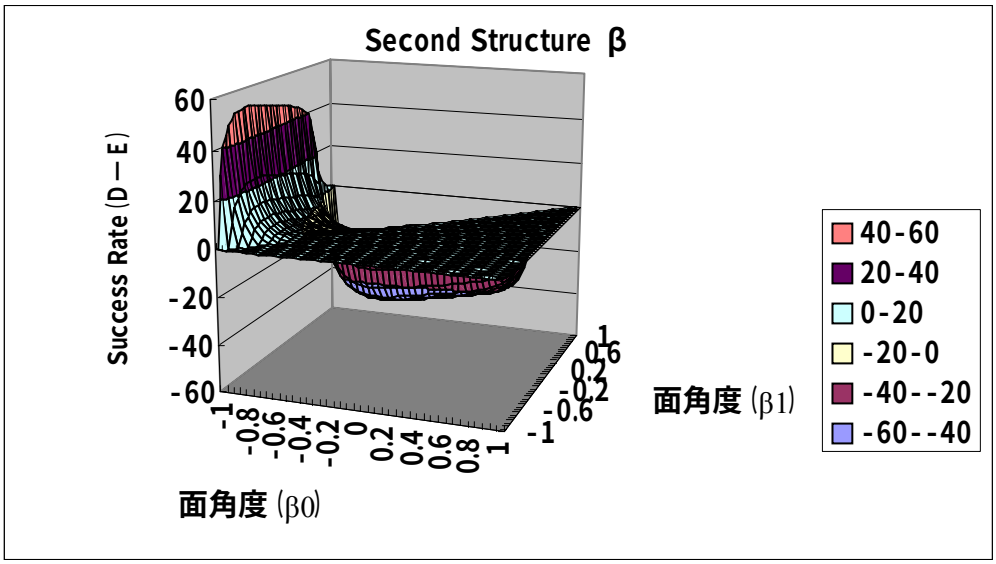


図 A3　しきい値 β_0, β_1 の大まかな決定。

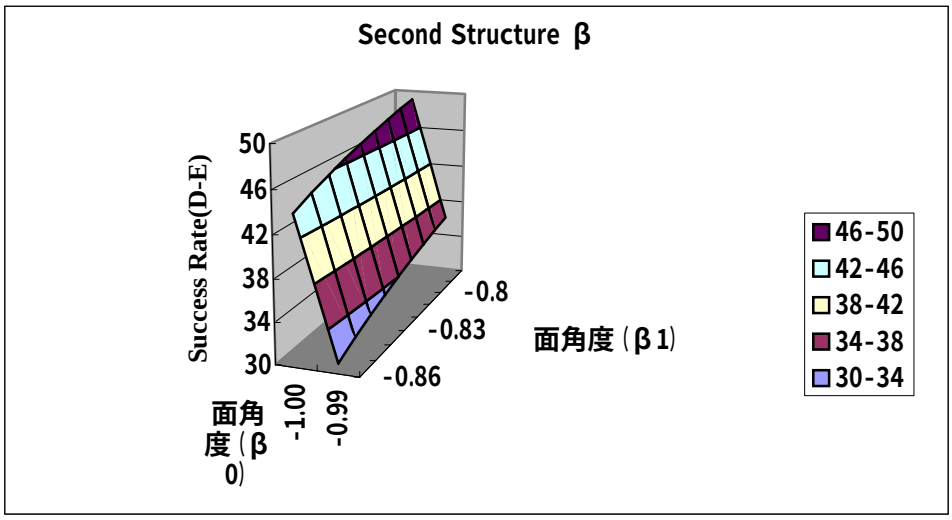


図 A4　しきい値 β_0, β_1 の細かい決定。

図 A3 に示すように、成功率（$D-E$）が最大となる領域は $-1 < \beta_0 < -0.99$，$-0.86 < \alpha_1 < 0.78$ 付近であったので、その領域に対して更に 0.01 の間隔でしきい値を調べた。図 A4、表 A2 に示された結果から、β_0=-1，β_1=-0.78 に決めるのが最適であることがわかった。

表 A2　図 A4 の結果を数字で表した表。

面角度（β0）	面角度（β1）	実際にβの残基数	実際にβの残基をβとして判定した残基数（A）	実際にβではない残基をβとして判定した残基数（B）	実際にβの残基をβとして判定した割合%（D）	実際にβではない残基をβとして判定した割合%（E）	D－E
-1	-0.86	75258	45462	128564	60.41	16.34	44.07
-1	-0.85	75258	46519	132546	61.81	16.85	44.96
-1	-0.84	75258	47461	136322	63.06	17.33	45.73
-1	-0.83	75258	48350	140047	64.25	17.8	46.45
-1	-0.82	75258	49171	143625	65.34	18.26	47.08
-1	-0.81	75258	49915	147035	66.33	18.69	47.64
-1	-0.8	75258	50645	150432	67.3	19.12	48.18
-1	-0.79	75258	51342	153859	68.22	19.56	48.66
-1	**-0.78**	**75258**	**51959**	**157001**	**69.04**	**19.96**	**49.08**
-0.99	-0.86	75258	32210	92845	42.8	11.8	31
-0.99	-0.85	75258	33267	96827	44.2	12.31	31.89
-0.99	-0.84	75258	34209	100603	45.46	12.79	32.67
-0.99	-0.83	75258	35098	104328	46.64	13.26	33.38
-0.99	-0.82	75258	35919	107906	47.73	13.72	34.01
-0.99	-0.81	75258	36663	111316	48.72	14.15	34.57
-0.99	-0.8	75258	37393	114713	49.69	14.58	35.11
-0.99	-0.79	75258	38090	118140	50.61	15.02	35.59
-0.99	-0.78	75258	38707	121282	51.43	15.42	36.01

表 A2 から読みとれることであるが、実際にはβストランドの構造をとっていない残基をβと判定してしまう事が良くある。このことは第Ⅳ章で議論するように、残基上の違いが小さいが構造が大きく違う T0498 と T0499 の識別の際に、FCS 法はヘリックスを多く含む T0498 をかなり良く選別できたのに、シートを多く含む T0499 に関しては悪いモデル構造を除けたものの、良い選別が得られたとはいえない理由であると考えられる。この問題点の解決のためには、ターゲットモデルの 2 次構造を判別する指標として二面角以外の指標も導入して、FCS を計算すべきだと思われる。

付録 B

デノボ構造予測のための粗視化されたランジュバン分子動力学手法

ペプチド鎖がα炭素の繋がれたビーズで表されていて、各ビーズの座標が$\{\mathbf{r}_i\}$で示されているとする。以下のような過減衰するランジュバン方程式を数値的に解くことによって、ペプチド鎖のフォールディングがシミュレートされた。

$$d\mathbf{r}_i/dt = -\partial V_{\text{total}}/\partial \mathbf{r}_i + \xi_i(t) \tag{4}$$

$\xi_i(t)$ は $\left\langle \xi_i(t)\xi_j(t') \right\rangle = 2T\delta_{ij}\delta(t-t')$ の条件を満たすガウシンアン乱数であり、T は乱数の重みを制御する温度に似たパラメーターである。V_{total} は、$\{\mathbf{r}_i\}$によってあらわに微分可能な関数である多体ポテンシャルである。

$$V_{\text{total}} = V_{\text{fragment}} + V_{\text{assemble}} \tag{5}$$

V_{fragment} はローカルなフラグメント構造を形成するための相互作用を表し、V_{assemble} はフラグメントを集める相互作用を表す。w_1, w_2, w_3, w_4をそれぞれ重み係数として、$V_{\text{fragment}} = w_1 V_{\text{fragment}}^{\text{pair}} + w_2 V_{\text{fragment}}^{\text{angle}}$ そして、$V_{\text{assemble}} = w_3 V_{\text{nn}} + w_4 V_{\beta}$ である。ここで、$V_{\text{fragment}}^{\text{pair}}$、$V_{\text{fragment}}^{\text{angle}}$ と V_{nn} は、重複のない蛋白質構造のライブラリーから選ばれたフラグメント候補構造の統計的な傾向を見積もる事によって構成される。

B.1 フラグメント収集

PISCESサーバー[1,2]を既定のパラメーターのまま使用して、CASP7 以前に公開されたPDBリストから選ばれた3624の重複しない蛋白質構造のライブラリーを作成し、ターゲット蛋白質の配列の各9残基領域のフラグメントと比較した。配列プロファイルは、ターゲット蛋白質と構造ライブラ

リーの各蛋白質について、しきい値をE-value 0.001とし、PSI-BLAST [3]の三回の繰り返しによって、重複しない（NR）配列データーベース [4]から導かれた。ターゲット蛋白質の配列の中の９残基領域すべてに対して、構造ライブラリーの中から最も配列プロファイル相関係数の大きいフラグメントを、20ないし40個選ぶ。*j*番目残基の周りの９残基領域に対して、最も相関があるフラグメントを $F_i(j)$と書く。$i = 1, \ldots, N^{\text{fragment}}$ として $N^{\text{fragment}} = 20$ 又は40である。

B.2 フラグメントに基づく二体ポテンシャル

j 番目の９残基領域が *j* 番目の残基から *j*+8 番目の残基までの長さのものである時に、$V_{\text{fragment}}^{\text{pair}}$ が $\{F_i(j)\}$から構成される。$k = 1, 2, 3, 4$ として、領域の中で *j*+4 番目の残基と $j + 4 \pm k$ 番目の残基の間の距離を $r_{j+4,\, j+4\pm k} = \left|\mathbf{r}_{j+4} - \mathbf{r}_{j+4\pm k}\right|$ と書く。 $r_{j+4,\, j+4\pm k}$ が取る値の統計的な傾向を表すために、$V^{j,\pm k}(r_{j+4,\, j+4\pm k})$ が以下の様に定義された。

$$V^{j,\pm k}(r_{j+4,\, j+4\pm k}) = \frac{-1}{\sum_{i=1}^{N^{\text{fragment}}} C(i,j)} \sum_{i=1}^{N^{\text{fragment}}} C(i,j) \times \exp\left[-\frac{1}{2c_k}\left(r_{j+4,\, j+4\pm k} - r_{5,\,5\pm k}^{i}\right)^2\right] \tag{6}$$

$c_k = 0.5k^{2/3}$ で $N^{\text{fragment}} = 20$ とする。$C(i, j)$ は、ターゲットの *j* 番目の９残基領域の配列プロファイルと $F_i(j)$.の配列プロファイルの間の相関係数である。$r_{5,\,5\pm k}^{i}$ は、フラグメントの N 末端から 5 番目の残基と $F_i(j)$の中の $5 \pm k$.番目の残基との距離である。このように足し算されたガウシアン関数が、ターゲット構造の中の９残基フラグメントの距離に対する統計的な制約として働く。$V^{j,\pm k}(r_{j+4,\, j+4\pm k})$ を説明するイラストとして、図 B1 を参照していただきたい。N^{res} がターゲット蛋白質の残基数だとして、次式のように、$V_{\text{fragment}}^{\text{pair}}$ は $V^{j,\pm k}(r_{j+4,\, j+4\pm k})$ を k と全９残基領域に対して合計したものである。

$$V_{\text{fragment}}^{\text{pair}} = \sum_{j=1}^{N^{\text{res}}-8}\left(\sum_{k=1}^{4} V^{j,-k}(r_{j+4,\, j+4-k}) + \sum_{k=1}^{4} V^{j,k}(r_{j+4,\, j+4+k})\right), \tag{7}$$

B.3　フラグメントに基づく二面角ポテンシャル

$V_{\mathrm{fragment}}^{\mathrm{angle}}$ は{F$_i$(j)}から構成される。$\mathbf{r}_{j,\,j+1}=\mathbf{r}_{j+1}-\mathbf{r}_j$ として、2 つの疑似二面角、θ_j とϕ_j,が $\mathbf{r}_j$、$\mathbf{r}_{j+1}$ と $\mathbf{r}_{j+2}$ の３つの点を含む面と $\mathbf{r}_{j+1}$、$\mathbf{r}_{j+2}$ と $\mathbf{r}_{j+3}$ を含む面によってつくられる角度として、以下の様に定義される。

$$\cos\theta_j=\frac{\mathbf{g}_j\cdot\mathbf{g}_{j+1}}{\left|\mathbf{g}_j\right|\left|\mathbf{g}_{j+1}\right|},\quad \cos\phi_j=\frac{\mathbf{g}_j\cdot\mathbf{r}_{j+2,\,j+3}}{\left|\mathbf{g}_j\right|\left|\mathbf{r}_{j+2,\,j+3}\right|}\ , \tag{8}$$

ただし、ここで $\mathbf{g}_j=\mathbf{r}_{j,\,j+1}\times\mathbf{r}_{j+1,\,j+2}$ である。$C_{\mathrm{angle}}=0.05$ で $N^{\mathrm{fragment}}=20$ とし、$\theta_3^{\,i},\theta_4^{\,i},\phi_3^{\,i}$,および $\phi_4^{\,i}$ が F$_i$(j).を中心にした周辺の５つの継続する残基によって定義される二面角だとして、ターゲットの j 番目の９残基領域を中心にした周辺における、二面角の値の統計的な傾向が、以下の様なポテンシャルで表される。

$$\begin{aligned}V_\theta^{\,j}(\cos\theta_{j+2},\cos\theta_{j+3})&=\frac{-1}{\sum_{i=1}^{N^{\mathrm{fragment}}}C(i,j)}\\ &\times\sum_{i=1}^{N^{\mathrm{fragment}}}C(i,j)\times\exp\left[-\frac{1}{2C_{\mathrm{angle}}}\left\{\left(\cos\theta_{j+2}-\cos\theta_3^{\,i}\right)^2+\left(\cos\theta_{j+3}-\cos\theta_4^{\,i}\right)^2\right\}\right],\end{aligned} \tag{9}$$

$$\begin{aligned}V_\phi^{\,j}(\cos\phi_{j+2},\cos\phi_{j+3})&=\frac{-1}{\sum_{i=1}^{N^{\mathrm{fragment}}}C(i,j)}\\ &\times\sum_{i=1}^{N^{\mathrm{fragment}}}C(i,j)\times\exp\left[-\frac{1}{2C_{\mathrm{angle}}}\left\{\left(\cos\phi_{j+2}-\cos\phi_3^{\,i}\right)^2+\left(\cos\phi_{j+3}-\cos\phi_4^{\,i}\right)^2\right\}\right],\end{aligned} \tag{10}$$

図 B1 のイラストで示した様に、これらのポテンシャルによって、ターゲット蛋白質の j 番目の領域の中の $j+2$ から $j+6$ までの５つの継続する残基に対する、幾何学的な制約が表現される。$V_{\mathrm{fragment}}^{\mathrm{angle}}$ は、ターゲットの中の全９残基領域に対する $V_\theta^{\,j}(\cos\theta_{j+2},\cos\theta_{j+3})$ と $V_\phi^{\,j}(\cos\phi_{j+2},\cos\phi_{j+3})$ の合計である。

$$V_{\text{fragment}}^{\text{angle}} = \sum_{j=1}^{N^{res}-8} \left(V_{\theta}^{j} + V_{\phi}^{j} \right) \quad . \tag{11}$$

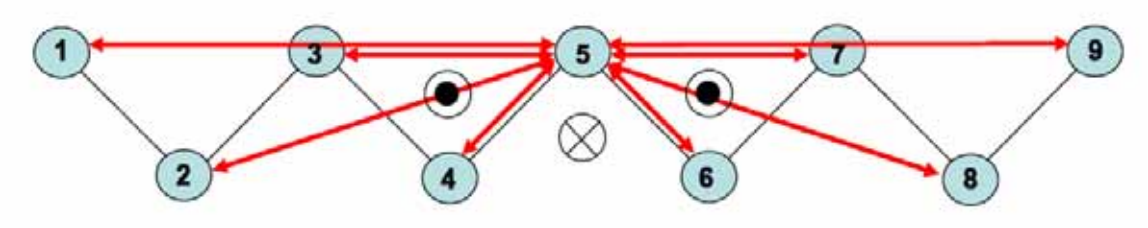


図 B1　$V_{\text{fragment}}^{\text{pair}}$ と $V_{\text{fragment}}^{\text{angle}}$ の構築を説明するイラスト。$V_{\text{fragment}}^{\text{pair}}$ は、フラグメントの中の赤い矢印によって示されるペア間距離に対する制約を表す。$V_{\text{fragment}}^{\text{angle}}$ は、フラグメントの中心周辺に位置する三つの連続する面の間の二面角に対する制約を与える。

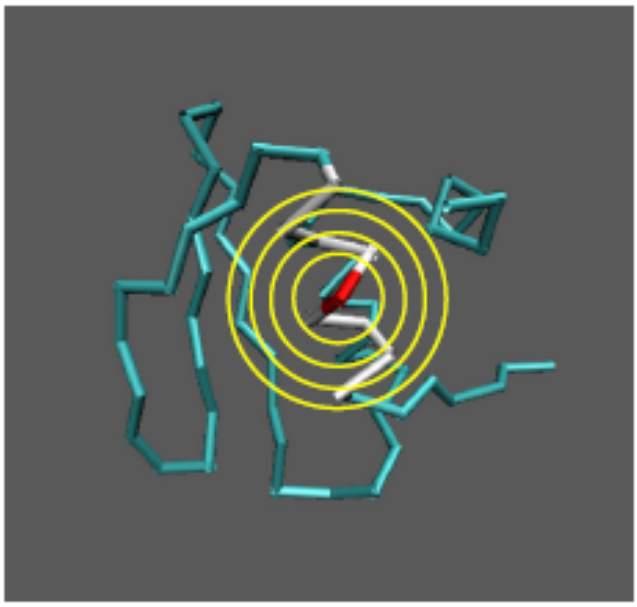


図 B2　V_{nn} の構築を説明するイラスト。V_{nn} は、各残基周辺の球の間の層に存在する残基数に対する制約を与える。

B.4　隣接数ポテンシャル

V_{nn} は疎水性相互作用による引力と排除体積による反発を表す。例えば、ターゲット蛋白質の j 番目の領域の中心である j+4 番目の残基と、その残基を中心とする球が図 B2 の様に定義される。ただし、$k = 1, 2, 3, 4$ として半径が $r(k) = 2.0 + 2.0 \times k$ (Å)となるように、4 層に重なった球が定義される。中心残基周辺の隣接する残基数を、次式のように、なめらかで微分可能な関数を定義する

事によって数えることができる。

$$u_k(r)=\begin{cases}1 & (r<r(k)-\delta r)\\ \dfrac{1}{2}\left\{1+\cos\left(\dfrac{r-\left(r(k)-\delta r\right)}{2\delta r}\pi\right)\right\} & (r(k)-\delta r\le r\le r(k)+\delta r)\\ 0 & (r>r(k)+\delta r)\end{cases}, \tag{12}$$

ただし、$\delta r=0.25$ Å である。$r<r(k-1)$ に対して $\Delta_k(r)=1$ で、$r\ge r(k-1)$ に対して $\Delta_k(r)=0$ として、半径 $r(k-1)$ の球と半径 $r(k)$ の球の間の殻の中の残基数を N_k^{j+4} と書けば、N_k^{j+4} は以下の様に計算される。

$$\begin{aligned} N_k^{j+4} &= \sum_{n\neq j+4}^{N^{\mathrm{res}}} u_k(r_{j+4,n}) - \sum_{m\neq j+4}^{N^{\mathrm{res}}} \Delta_k(r_{j+4,m}) \qquad (2\le k\le 4)\\ N_k^{j+4} &= \sum_{n\neq j+4}^{N^{\mathrm{res}}} u_k(r_{j+4,n}) \qquad (k=1) \end{aligned} \tag{13}$$

N_k^{j+4} に対する制約が $\mathrm{F}_i(j)$ の中心残基の周辺の隣接残基数である $N_k^{5,\,i}$ のサンプリングによって見積もられ、そして、$C_{\mathrm{nei}}=0.5$ で、$N^{\mathrm{fragment}}=40$ として、以下の様なポテンシャルによって表される。

$$V(N_k^{j+4})=\frac{-1}{\sum_{i=1}^{N^{\mathrm{fragmenti}}} C(i,j)}\sum_{i=1}^{N^{\mathrm{fragment}}} C(i,j)\times\exp\left[\frac{-\left(N_k^{j+4}-N_k^{5,\,i}\right)^2}{2C_{\mathrm{nei}}}\right] \tag{14}$$

V_{nn} はターゲット蛋白質の中で、k と全９残基領域に対する $V(N_k^{j+4})$ の合計によって定義される。

$$V_{\mathrm{nn}}=\sum_{j=1}^{N^{\mathrm{res}}-8}\sum_{k=1}^{4} V(N_k^{j+4}). \tag{15}$$

B.5　βシートポテンシャル

V_β は、βストランド間の擬似水素結合の形成によるβシートの安定化を表す。まずは、鎖に沿ってベクター $\mathbf{a}_n = \mathbf{r}_{n+1,n+2}/r_{n+1,n+2}$ を定義して、$l_0 = \cos(\pi/3)$、 $\Delta l = 0.3$、$\theta_0 = (1+10/180)\pi$、$\Delta c = 0.25$ を使用する事によって、βストランドの形成が次の関数で表される。

$$V_{\text{strand}}(n) = \exp\left(-\frac{\left(\cos(\theta_n - \theta_0) - 1\right)^2}{2\Delta c^2}\right) \times \exp\left(-\frac{1}{2\Delta l^2}\left[\left(\vec{a}_{n-1}\cdot\vec{a}_n - l_0\right)^2 + \left(\vec{a}_n\cdot\vec{a}_{n+1} - l_0\right)^2\right]\right). \tag{16}$$

$\mathbf{b}_{n,m} = \mathbf{r}_{n,m}/r_{n,m}$ は、水素結合に並行に配向するベクトルであり、$\Delta s = 0.3$ で $\Delta l_{hb} = 1.0$ Å で $l_{hb} = 5.0$ Å として、擬似水素結合の形成に対する幾何学的な制約が次の関数で表される。

$$V_{\text{bond}}{}^{k,l}(n,m) = \exp\left(-\frac{\left(r_{n+k,m+l} - l_{hb}\right)^2}{2\Delta l_{hb}^2}\right)\exp\left(-\frac{1}{2\Delta s^2}\left[\left(\mathbf{a}_n\cdot\mathbf{b}_{n+k,m+l}\right)^2 + \left(\mathbf{a}_m\cdot\mathbf{b}_{n+k,m+l}\right)^2\right]\right) \tag{17}$$

n 番目の残基周辺のβストランドと、m 番目の残基周辺のβストランドの間の平行シート形成と逆平行シート形成を、それぞれ次の関数で表す。

$$V_{\text{parallel}}(n,m) = w_{n+1,m+1}V_{\text{bond}}{}^{1,1}(n,m) + w_{n+2,m+2}V_{\text{bond}}{}^{2,2}(n,m) + w_0 V_{\text{bond}}{}^{1,1}(n,m)V_{\text{bond}}{}^{2,2}(n,m) \tag{18}$$

$$V_{\text{anti-parallel}}(n,m) = w_{n+1,m+2}V_{\text{bond}}{}^{1,2}(n,m) + w_{n+2,m+1}V_{\text{bond}}{}^{2,1}(n,m) + w_0 V_{\text{bond}}{}^{1,2}(n,m)V_{\text{bond}}{}^{2,1}(n,m) \tag{19}$$

ただし、$w_0 = 0.5$ である。係数 $w_{n,m}$ は、βシートの形成に対するニューラル・ネットワークアルゴリズムを用いた見積もり法である BETApro [5]を使って決定した。すなわち、$w_{bp}(n,m)$ を BETApro によって計算される、n 番目の残基を含むストランドと m 番目の残基を含むストランド間のβシート形成の確率を表す擬似エネルギーとして、$w_{n,m}$ は以下の様に定義される。

$$w_{n,m} = w_{bp}(n,m) + 0.5 \tag{20}$$

もし、n 番目の残基又は m 番目の残基が BETApro によって予測されたストランドの中に含まれなければ、$w_{\mathrm{bp}}(n, m)$は 0 にセットされる。式(16)-(20)を使って、V_β は以下の様に定義される。

$$V_\beta = -\sum_{n=1}^{N^{\mathrm{res}}-7}\sum_{m=n+4}^{N^{\mathrm{res}}-3} V_{\mathrm{strand}}(n)V_{\mathrm{strand}}(m)\times\left(V_{\mathrm{parallel}}(n,m)+V_{\mathrm{anti\text{-}parallel}}(n,m)\right) \tag{21}$$

B.6　ランジュバン MD シミュレーション

方程式(7),(11),(15)と(21)を使って、方程式(5)のV_{total}を$\{\mathbf{r}_i\}$によって解析的に微分する事が可能である。重み係数w_1, w_2, w_3とw_4 は、ターゲット構造と、ターゲット構造とは相関しないコンパクト構造の間に、充分なエネルギー・ギャップを与えるように選ばれる。そのためここでは、w_1= 1.5, w_2=0.75, w_3= 0.79 と w_4=1.0 を使用した。引き伸ばされた鎖の構造形態から開始して、低いエネルギー構造を探求するために、方程式(4)のランジュバン MD が第 II.2.2 項で説明された冷却スケジュールに従って実行された。

付録 C
ランジュバン動力学と FCS 併用による方法を、CASP7 のターゲットに対して適用した計算の結果

第Ⅱ章で述べた、ランジュバン動力学と FCS の併用による方法を、CASP8 のターゲットのみならず、CASP7 のターゲットに対しても適用した。ベンチマーク試験として、中程度に難しいターゲットと難しいターゲットが、CASP7 の TBM, FM, TBM/FM (TBM と FM の間の境界カテゴリー)のカテゴリーから選ばれた。計算方式は第Ⅰ章で見いだされた最も能力の高い *RCF-FC-FWM* 法ではなく、*CF-ANC-CCM* 法で行った。*CF* 法でフラグメント集合の時に使う相関係数のしきい値は 0.6 とした。

付録 B の中で詳しく説明された方法のランジュバン動力学を用いることによって、CASP7 に“KORO”というチーム名で参加した。この付録で紹介するのは、CASP7 の後に、各ターゲットに対して違う乱数を用いて、ランジュバン MD 計算を $N_{traj} = 400$ 回繰り返した結果である。CASP7 参加時には、予測構造提出締切までの時間が限られているため、計算できる構造数がターゲット毎に異なってしまったが、よりよく計算結果を評価するために、CASP7 参加後に計算した構造数が N_{traj} 回にそろうように、計算し直したためである。N_{traj} トラジェクトリーの最終段階で得られた N_{traj} 個の構造のうち、ランジュバン MD における有効エネルギーが 1 番目と 2 番目に低い構造を、1 番目と 2 番目のモデルとして選んだ。そして、N_{traj} 構造のクラスター分析を行って、1 番目と 2 番目と 3 番目の大きなクラスターの中心構造を 3 番目と 4 番目と 5 番目のモデルとして選び、これらの各ターゲットに対して 5 つの構造を KORO-1 の結果と呼ぶことにする。

KORO-1 の結果は、いくつかのターゲットに対してかなり高い GDTTS を示している。T0283 と T0354 がそのようなターゲットの例である。それらのターゲットに対して、$N_{traj} = 400$ 構造のエネルギーとそれらの GDTTS が図 C1 にプロットされている。1 番目のモデルの構造と実験的に観察された構造は、図 C2 に比較されている。有効エネルギーと GDTTS の相関係数の絶対値 $|C_{\text{Energy-GDTTS}}|$ は、T0283 に対して 0.34（p-value＜10^{-6}）、T0354 に対して 0.556（p-value＜10^{-6}）である。こうした $|C_{\text{Energy-GDTTS}}|$の高い値は、これらのターゲットに対するエネルギー表面がファネル型のランドスケープをとっている事を示唆している。図 C1 と図 C2 におけるもう一つの例は、T0300 に対する構造とプロットである。T0300 に対しては、$|C_{\text{Energy-GDTTS}}|$が 0.096（p-value 0.056）と低い。T0300 に対する小さい相関係数は、現在の有効エネルギー関数が、この蛋白質のエネルギーランドスケープの、いくつかの重要な特徴を表現する事に失敗している事を示す。T0300 に対する一番低い有効エネルギー構造の GDTTS は、表 C1 に示す様に比較的小さい。しかしながら、N_{traj} 構造の中で、我々はある程度高い GDTTS を持つ構造をも見つけられるので、そうした品質の高い構造を救う一つの方法は、

ランジュバン MD によって生成された構造の中から、よい候補を識別できる MQA スコア関数を用いる事である。この目的のために、我々は FCS をスコア関数として使う。

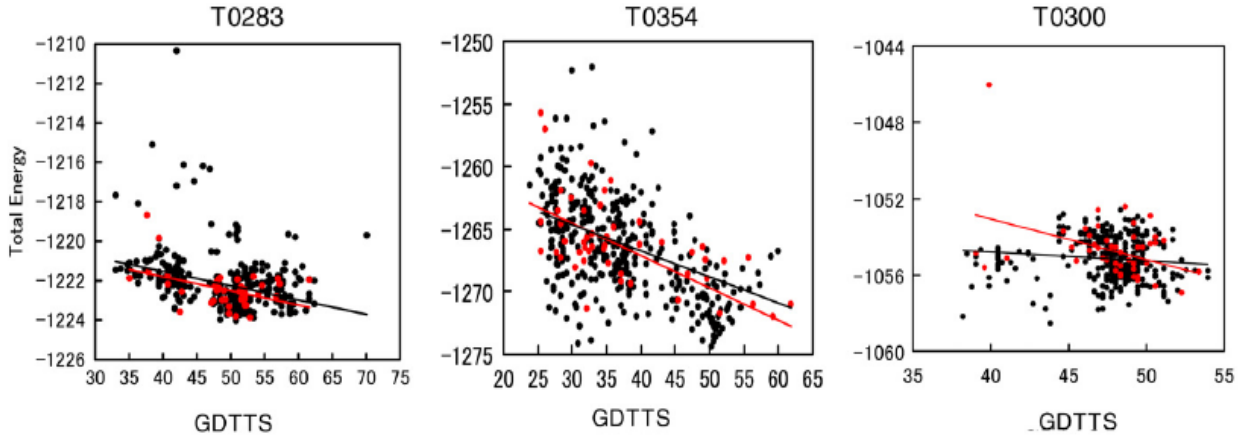


図 C1　有効エネルギーの GDTTS のプロット。赤点が最も高い N_{score}＝50 個の構造であり、黒点は他の $N_{traj} - N_{score}$＝350 個の構造。赤直線と黒直線の傾きは、それぞれ赤い N_{score} 個の点と全ての N_{traj} 個の点の相関係数を表す。左が T0283、中が T0354、そして右が T0300。

まずは、N_{traj} = 400 個の構造に FCS 法を応用して、そして、N_{traj} 構造の中で最も高い FCS を持つ N_{score} 個の構造を選ぶ。18 個ターゲット中 14 個に対して、構造候補を 400 個から N_{score} = 50 に制限する事によって、$|C_{Energy\text{-}GDTTS}|$ の値は明確に高くなった。（図 C3 を参照）。それ故に、エネルギー関数と FCS を同時に使うというクロス・チェックによって、もっと良い構造を選ぶことができるであろうと期待される。図 C1 に、N_{score} = 50 個のデータをプロットした。N_{score} = 50 構造を選ぶ事によって、T0283・T0354・T0300 に対する $|C_{Energy\text{-}GDTTS}|$ は、それぞれ 0.434（p-value 0.0016）、0.591（p-value 6.3×10^{-6}）、0.384（p-value 0.006）と上昇する。

N_{score} = 50 個の構造の中のエネルギーが 1 番低い構造を 1 番目のモデル、N_{traj} 構造の中で 1 番エネルギーが低い構造を 2 番目のモデル、N_{traj} 構造の中で 1 番目と 2 番目と 3 番目の大きなクラスターを 3 番目と 4 番目と 5 番目のモデルとして選ぶ。こうして得られた結果を、KORO-2 と呼ぶことにする。図 C2 に示すように、KORO-2 を用いることによって、T0300 に対する予測構造の品質は、相当な品質の向上を見せた。このように、T0300 に対する $|C_{Energy\text{-}GDTTS}|$ が低いという問題は、FCS の利用によって解決された。

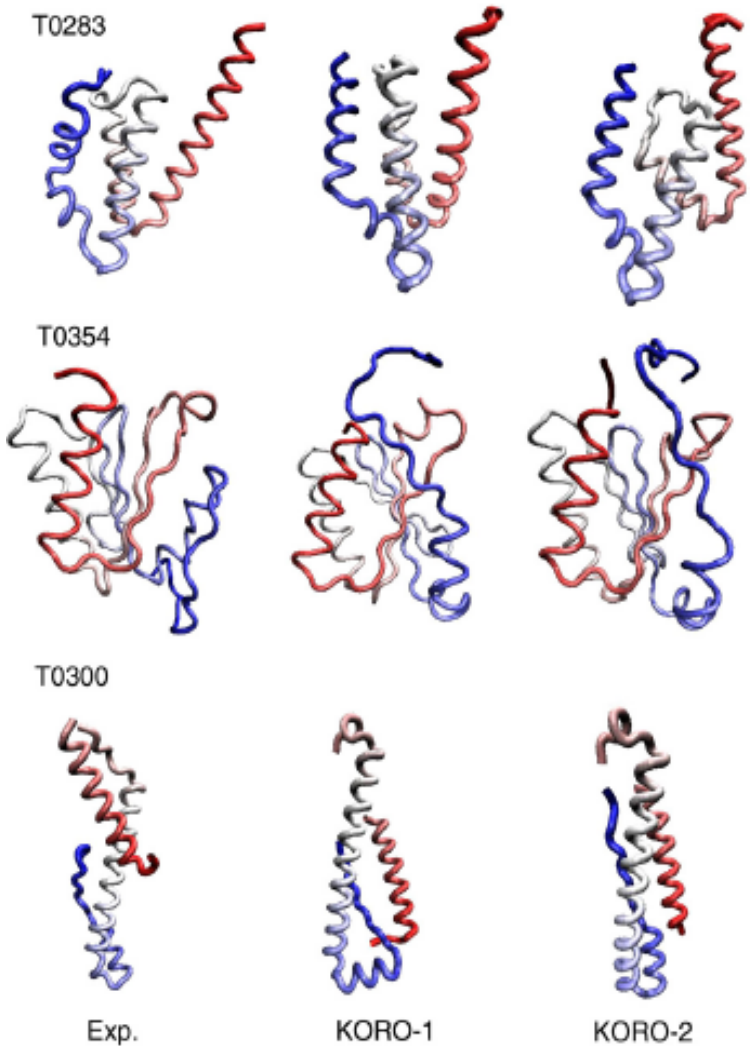


図 C2

構造の比較。鎖は赤(N 末端)から青(C 末端)に色付けされている。**上段**：T0283 に対する構造。(左) 構造の実験的に観察された構造、CASP7 の評価に使われたのは PDB コード 2HH6 に登録された 1-97 残基である。(中) KORO-1 の 1 番目のモデル、GDT_TS＝50.77 で RMSD(root mean square deviation)＝5.19Åである。(右) KORO-2 の 1 番目のモデル、GDT_TS＝52.84 で RMSD＝5.71Åである。**中段**：T0354 に対する構造。 (左) 実験的に観察された構造、PDB コード 2ID1。(中) KORO-1 の 1 番目のモデル、GDT_TS＝50.21 で RMSD＝3.23Åである。(右) KORO-2 の 1 番目のモデル、GDT_TS＝59.17 で RMSD＝4.30Åである。ここでは、1-100 残基を使用する(101-200 残基を無視する)事によって RMSD が計算されている。**下段**：T0300 に対する構造。(左) 実験的に観察された構造、PDB コード 2H3R に登録された内容では、32-38 残基の構造は、実験的に決定されていない。 (中) KORO-1 の 1 番目のモデル、GDT_TS＝43.82 で RMSD＝12.28Åである。(右) KORO-2 の 1 番目のモデル、GDT_TS＝52.25 で RMSD＝10.77Åである。

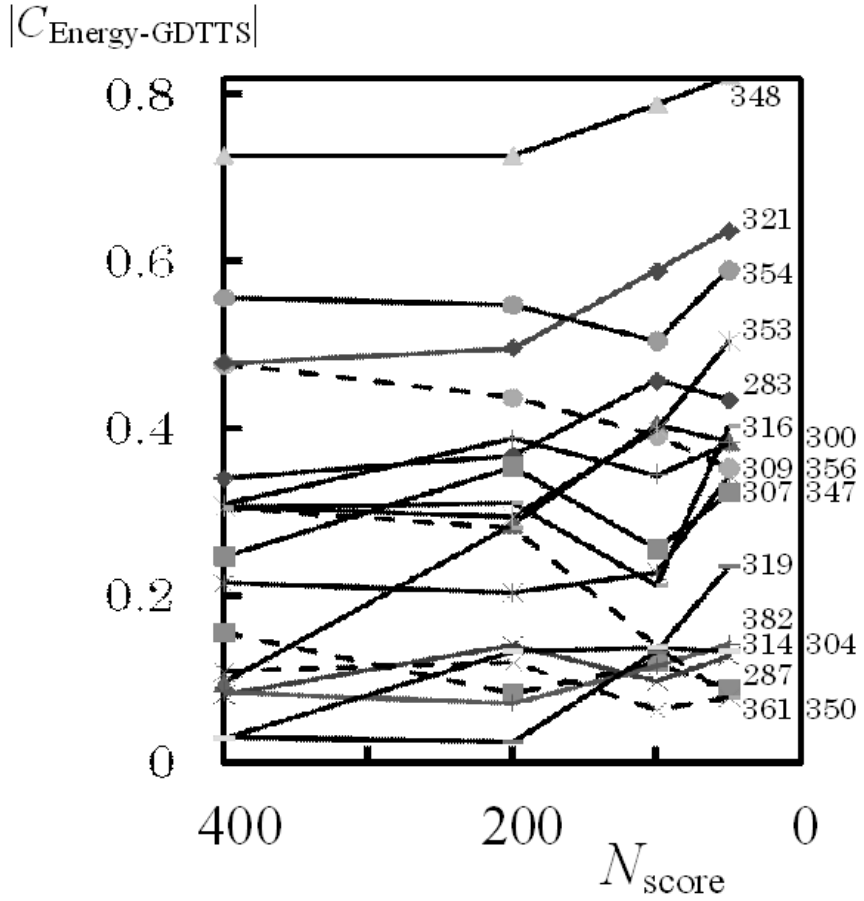


図 C3

GDTTS と有効エネルギーの間の相関係数の絶対値、$|C_{\text{Energy-GDTTS}}|$、が FCS を使用する事によって制限された構造の数、N_{score}、に対してどのように変化するかを示す。N_{score}＝50 に制限したときの$|C_{\text{Energy-GDTTS}}|$は、$N_{\text{score}}$＝400 のときの$|C_{\text{Energy-GDTTS}}|$より、4 個のターゲット蛋白質に対しては減少しているが(破線)、14 個のターゲット蛋白質に対しては大きくなっている(実線)。

18 個のターゲットについてのテストの結果が、図 C4 に要約されている。図 C4 に示されているように、18 個のうち 10 個のターゲットについて、KORO-2 の 1 番目のモデルは KORO-1 の 1 番目のモデルに比べて、より良い結果を与え、3 個のターゲットに対しては KORO-1 より変化なし、そして、5 個のターゲットに対しては、KORO-1 より低い GDTTS を持つようになった。ターゲットごとの 5 つのモデルの最もよい構造を比較すると、KORO-2 は 4 ターゲットに対して KORO-1 より改良されて、11 ターゲットに対して変化なし、そして 3 ターゲットに対してより低い GDTTS を持つようになった。従って、KORO-2 の結果は KORO-1 より、全体的に改良されたことが理解できる。

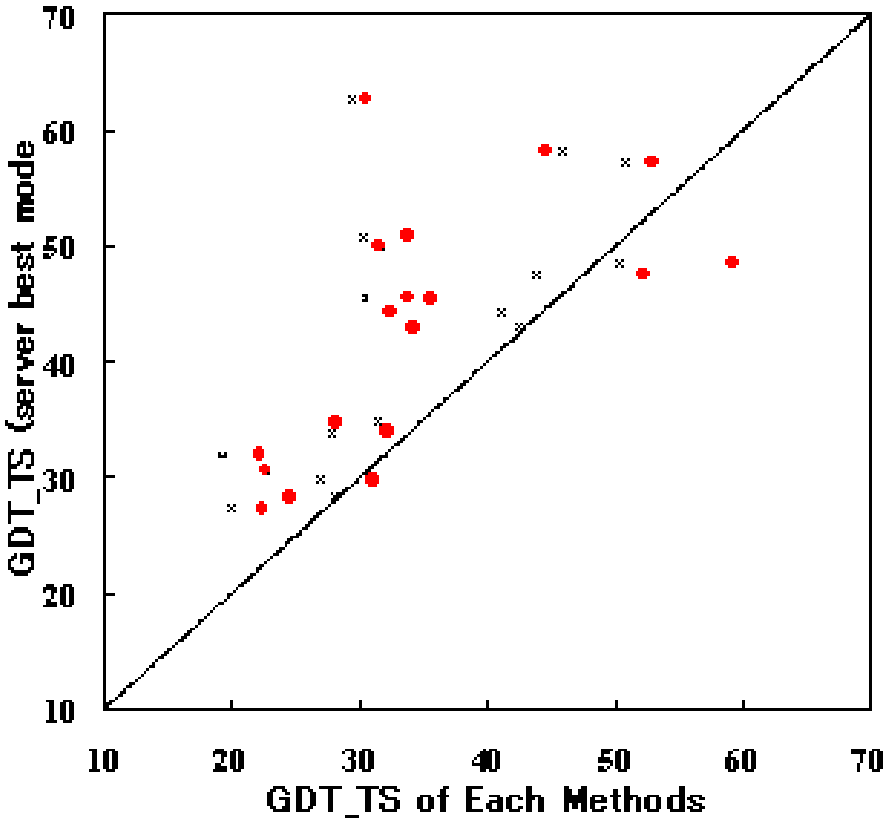


図 C4　横軸は、CASP7 の各ターゲットに対して、それぞれ KORO-2（赤丸）と KORO-1（黒×）の方法で選別された 1 番目のモデル構造の GDTTS の値。縦軸はサーバーモデル構造の GDTTS の最大値。

表 C1 と C2 では、KORO-2 の構造の GDTTS が、CASP7 に参加した他のグループ ROKKO・TASSER・Baker・Zhang の結果と比較されている。1 番目のモデルの GDTTS（表 C1）と 5 つのモデルのベストの GDTTS（表 C2）の両方で、KORO-2 が他の手法と比較できる結果を達成できる事を示している。表 C1 と C2 における 5 つの違う手法の中で、1 番高い GDTTS を示すターゲットの数を数えてみよう。KORO-2 が 1 番高い GDTTS を示すターゲットの数は、1 番目のモデルを用いたときは 6 個、5 つのモデルのベスト構造を用いたときは 4 個である。KORO-2 のかわりに KORO-1 を他の 4 つの手法と比較すれば、KORO-1 が 1 番高い GDTTS を示すターゲットの数は、1 番目のモデルを用いたときは 2 個、5 つのモデルのベスト構造を用いたときは 3 個であった。このことから、エネルギーと FCS の併用が結果を改良していることがわかる。

表 C1
5 つの違う手法によって得られた 1 番目のモデルの GDT_TS の比較

Target	KORO-2	ROKKO	TASSER	Baker	Zhang
T0283	**62.89	45.36	60.05	*75.00	40.47
T0287	*28.11	16.30	**22.36	19.72	22.05
T0300	*52.25	*no data*	38.20	**44.10	40.45
T0304	35.40	22.52	28.96	*45.05	**44.55
T0307	25.81	**28.05	24.59	*30.08	27.24
T0309	33.47	*36.69	31.05	**34.27	30.24
T0314	22.57	21.12	**23.79	23.06	*24.52
T0316_2	*42.50	20.00	22.91	**31.67	31.25
T0319	**22.41	17.78	18.33	*25.74	19.81
T0321_2	41.05	35.98	**45.77	42.91	*45.95
T0347_2	*38.03	*36.27	33.45	32.40	**35.56
T0348	45.08	30.33	**47.54	45.90	*50.00
T0350	31.04	32.42	35.44	**57.42	*57.97
T0353	31.63	40.36	37.05	**40.97	*51.51
T0354	*59.17	42.21	44.67	**56.76	43.03
T0356_3	**31.67	18.75	**31.04	22.09	*33.12
T0361	*32.12	18.04	27.85	**29.43	19.62
T0382	33.82	**47.69	46.22	38.23	*50.21

* 5 つの手法の最も高い GDTTS
** 5 つの手法の 2 番目に高い GDTTS

表 C2
5 つの違う手法によって得られたそれぞれ 5 つのモデルのベスト GDT_TS の比較

Target	KORO-2	ROKKO	TASSER	Baker	Zhang
T0283	**62.89	45.36	60.05	*82.47	58.76
T0287	*28.11	21.59	24.53	23.75	*28.11
T0300	**52.25	*no data*	38.20	*53.37	46.07
T0304	35.40	42.08	31.19	*45.05	**44.55
T0307	25.81	**34.96	24.59	*39.02	30.69
T0309	33.47	*36.69	33.06	**34.27	32.26
T0314	22.57	**25.00	24.27	23.06	*29.13
T0316_2	*42.50	20.84	25.00	34.58	**34.59
T0319	22.41	20.00	22.59	*30.92	**25.37
T0321_2	41.05	37.67	**45.77	45.44	*45.95
T0347_2	38.03	**41.90	37.67	*52.82	39.08
T0348	50.82	48.77	49.18	**53.28	*54.51
T0350	31.04	43.41	35.44	*60.16	**57.97
T0353	31.63	**46.99	44.28	44.58	*51.51
T0354	*59.13	42.21	47.13	**57.79	52.46
T0356_3	31.04	30.00	*33.96	27.71	**33.12
T0361	*32.12	18.67	27.85	**31.01	28.64
T0382	33.82	47.69	**53.36	51.05	*59.66

* 5 つの手法の最も高い GDTTS
** 5 つの手法の 2 番目に高い GDTTS

表 C1 と C2 に示された様に、既に開発されたどの手法も、ターゲットに応じてときにはよい予測構造を生成することができても、多くの FM ターゲットに対して常に、充分に高い GDTTS を持つ予測構造をつくり出せる、一貫した能力を持っていない。そのため、もっと一貫性のある予測を可能にする、新しいアイディアを試す事が強く望まれている。この論文の中で、有効エネルギーと FCS の２つの基準で構造をクロス・チェックする事によって、結果が改良された事を示したが、これは一貫性のある方法を開発するための、ひとつの可能性を示唆している。

付録 D

CASP7 と CASP8 における FM ターゲット蛋白質

表D1　CASP7におけるFMターゲット蛋白質

CASPにおけるターゲット識別番号	PDBコード	タイプ	大きさ (残基数)	**最良のサーバーモデル構造におけるGDTTS	**,# 対象を扱ったMQAチーム数
T0287	2g3v	α	161	28.26	166
T0296	2ha9	α/β	414	14.32	128
T0300	2h3r	α	89	47.48	138
T0304	2h28	α+β	101	45.55	184
T0307	2h5n	β	123	31.91	152
T0309	2h4o	α+β	62	34.68	188
T0314	2hg6	α+β	103	30.58	161
*T0316_D2	2hmA	β	90	42.92	157
T0319	2j6a	α+β	135	27.22	141
T0321_D2	2h1q	α/β	156	44.26	145
T0347_D2	2hwj	α	71	45.42	164
T0348	2hf1	β	61	58.19	168
T0350	2hc5	α+β	91	62.64	162
T0353	2hfq	α+β	83	50.00	171
T0356_D1	2idb	α+β	120	27.62	147
T0361	2hkt	α	158	33.86	165
T0382	2i9c	α	119	50.84	175
T0386_D2	2jk8	α+β	81	36.11	139

* 例えば T0316_D2 は CASP7 に使用された T0316 のターゲット蛋白質のドメイン２を表す。
**構造を局所的にのみ示すサーバーモデルは排除されている。
各サーバーチームは予測に対して 5 つのモデル構造を提出した。

表D2　CASP８におけるFMターゲット蛋白質

CASPにおけるターゲット識別番号	PDBコード	タイプ	大きさ (残基数)	**最良のサーバーモデル構造におけるGDTTS	**[#] 対象を扱ったMQAチーム数
*T0397_D1	3d4r	β	82	35.06	200
T0405_D1	-	α	72	58.68	212
T0405_D2	-	α+β	208	31.85	195
T0416_D2	3d3q	α	57	71.49	222
T0443_D1	3dee	α	66	57.95	216
T0443_D2	3dee	α+β	60	44.58	203
T0465_D1	3dfd	α	96	38.28	223
T0476_D1	2k5c	α+β	87	47.13	208
T0482_D1	2k4v	α+β	67	65.30	214
T0496_D1	3do9	α+β	120	30.21	220
T0510_D3	3doa	α+β	43	55.81	204
T0513_D2	3dup	α+β	69	71.01	170

* 例えば T0397_D1 は CASP8 に使用された T0397 のターゲット蛋白質のドメイン１を表す。
**構造を局所的にのみ示すサーバーモデルは排除されている。
[#] 各サーバーチームは予測に対して 5 つのモデル構造を提出した。

引用文献

1. Wang G., Dunbrack Jr R.L: PISCES a protein sequence culling server, *Bioinformatics*. 19:1589-1591 (2003).
2. Available from http://dunbrack.fccc.edu/PISCES.php .
3. Altschul S.F, Madden T.L, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman D.J: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*. 25:3389-3402 (1997).
4. Available from http://www.ncbi.nlm.nih.gov/ .
5. Cheng J, Baldi P: Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms, *Bioinformatics*. 21: 75-84 (2005).