論文の内容の要旨

応用生命工学　専攻
平成１５年度博士課程　進学
氏　　名　石田　貴士
指導教員名　清水謙多郎

論文題目　　機械学習を用いたタンパク質の立体構造予測・評価手法の開発

１．はじめに

タンパク質の立体構造はその機能との間に深い相関があることが知られており、タンパク質の機能解析やタンパク質をターゲットとした創薬にはその立体構造は欠かせないものとなっている。近年、X線結晶解析やNMR(Nuclear Magnetic Resonance)分光解析などの実験的な手法により数多くのタンパク質の立体構造が明らかにされているが、依然として実験的なタンパク質の立体構造決定はコストと労力を要する。計算によりアミノ酸配列情報から立体構造の決定を試みる立体構造予測は、これらの実験的な手法を補助するにとどまらず、実験的手法の限界に捕らわれないことから従来では決定が難しかったタンパク質の構造を明らかにすることも期待されており、また、新規の人工的なタンパク質のデザインにおいても力を発揮すると考えられている。

一方、近年の構造ゲノミクスの進展は数多くの新規フォールドのタンパク質の立体構造を明らかにしたが、その一方で機能を有するにもかかわらず天然状態で一定のフォールドを形成しない、もしくは数十残基にわたって一定の構造をとらないdisorderと呼ばれる領域を持つタンパク質が数多く存在することが明らかになってきた。さらに、これらのdisorder領域の中には結合サイトを有し、DNA結合などの分子認識に関与するものがある事が知られており、disorder領域を特定することは機能の推定においても重要であると考えられる。

そこで、本研究では構造未知のタンパク質のアミノ酸配列を入力とし、まずそのタンパク質のdisorder領域を予測することで不安定な構造をとりうる領域を同定し、また、安定な構造をとりうる領域に対して、その立体構造を予測する手法の開発を行った。

２．タンパク質のdisorder領域予測手法の開発

タンパク質のdisorder領域のアミノ酸構成には単純な配列の繰り返しや、A, R, G, Q, S, P, E, K といった特定のアミノ酸がその他の領域に比べ多く見られるなどの特徴的なパターンが存在している。そのため、その配列的な特徴を利用して、アミノ酸配列情報を入力としたdisorder領域の予測が行なわれてきた。それらの多くは2次構造予測手法に類似した、局所的な配列情報を入力として機械学習を用いる予測手法を利用しているが、その予測精度はQ2(2状態予測)で70～80%程度にとどまっており、更なる予測精度の向上が望まれている。

アミノ酸配列が相同なタンパク質が類似した立体構造を持つことはよく知られており、僅かな配列相同性しか存在しないタンパク質間であっても立体構造がよく保存されている例が数多く知られている。そのためdisorder領域の保存性について調査を行ったところ、disorder領域も立体構造と同様にアミノ酸配列が大きく変異した配列相同性の低いタンパク質間であっても保存されている事がわかった。特に、60%を超える高い配列相同性を示したホモログにおいては70%近い割合でdisorder領域が保存されていた。

そこで、本研究では従来型の局所配列情報からの予測としてSupport Vector Machine(SVM)と呼ばれる2クラスの分類を行う機械学習アルゴリズムを利用した予測を行い、同時に構造既知で配列相同なタンパク質との配列アライメントを利用した大域的な配列情報からの予測を併用することで予測精度の向上を図った新たなdisorder 領域予測手法を開発した。

開発された予測手法は、5-foldのクロスバリデーションテストによって性能が評価された。局所配列情報のみに基づく予測手法では4%の疑陽性(false positive)を許容した場合57.7%の感度(sensitivity)が得られたのに対して、新たに開発された手法では同じ4%の疑陽性を許容した場合に60.3%の感度を示し、3%近い感度の向上が得られた。また、構造既知のホモログを持つタンパク質チェインのみからなるテストセットに対する性能評価では同じく4%の疑陽性を許容した場合において63.1%の感度を示し5%を越える感度の向上が得られ、Q2の予測精度では81.9%と非常に高い予測精度が得られた。

３．タンパク質立体構造予測手法の開発

タンパク質立体構造予測手法は既知構造テンプレートを利用する手法と利用しない方法との大きく2つに分類できる。比較モデリング等の既知構造テンプレートを利用する手法は高い予測精度が得られる一方、新規のフォールドを予測することができず、配列アライメントの得られなかった領域についてはモデリングの精度が非常に低下するという欠点を有している。既知構造テンプレートを利用しない*de novo* (もしくは*ab initio*)と呼ばれる予測手法は、Anfinsenの仮説に基づき、アミノ酸配列に対して自由エネルギーが最小となるような構造を探索することで立体構造を予測する方法であり、多くの計算コストを要するが、未知のフォールドについても予測が可能な手法である。本

研究では現在既知構造テンプレートが存在しない際に用いられる手法として最も成功しているフラグメントアセンブリ法に基づいた予測手法の開発と改良を行った。

現在のde novo立体構造予測手法の問題点は大きく2つに分けられる。一つ目の問題は構造サンプリングの問題である。フラグメントアセンブリ法は予測対象のアミノ酸配列を9残基程度の断片に分割し、個々の配列断片に対してその配列相同性によって構造データベースから用意された候補となる局所構造をエネルギーが低くなるように組み合わせることで、探索する構造空間を効果的に制限し、効率的な構造サンプリングを行う手法である。しかし、予測対象タンパク質が100残基前後の小さなタンパク質であれば十分に良い構造サンプリングが可能であるが、数百残基といった大きなタンパク質に対してはフラグメントの長さが9残基程度では構造空間の制限が不十分で効果的なサンプリングができない。また、ループのように構造バリエーションの多い領域では9残基にわたり天然構造に一致する局所構造がデータベース中に存在しない場合、逆に構造空間を制限しすぎてしまい、サンプルされる構造空間中に天然に近い構造が存在しないという状況が起こりうる。そこで、本研究ではより効率的なサンプリングを行い、同時にサンプルする構造空間に天然構造が含まれなくなってしまうのを防ぐために、フラグメント長を既知構造に対する2次構造と配列の類似性から最適化するアルゴリズムを開発した。この手法を含めた本研究室で開発された立体構造予測システムABLEの性能を検証するため、既知構造テンプレートの存在しないターゲットについての予測を試みたところ、200残基を越えるタンパク質についても正しいフォールドの予測に成功した(図1)。

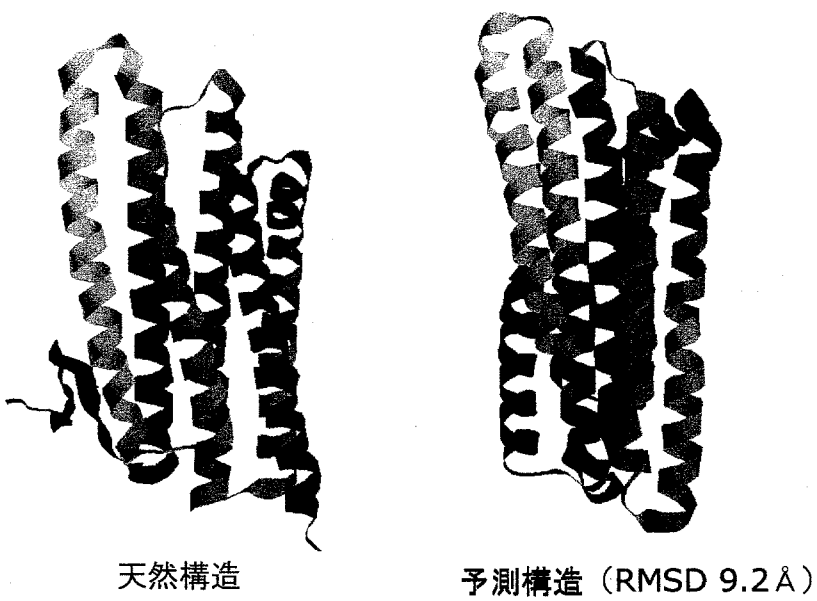

図1　予測構造と天然構造の比較（Phosphate transport system regulator PhoU, putative, T. maritime, pdbid: 1SUM, 235残基）

現在のde novo立体構造予測におけるもう一つの問題はサンプリングに用いるポテンシャルに関するものである。計算量の問題から構造のエネルギー計算には側鎖を一つの代表点で近似した粗視化モデルを利用し、そのモデルに対して残基レベルの解像度で設計されたデータベース由来の統計ポテンシャルが用いられている。埋もれ度ポテンシャルはそのような残基レベルのポテンシャルの一つであり、タンパク質のフォールディングにおいて重要な力の一つである疎水性相互作用を表すものである。ある残基の周辺に存在する残基の数で表される残基の埋もれ度はその残基のアミノ酸種毎に特徴的な分布を示すため、従来は、ある埋もれ度における各アミノ酸種の存在確率から統計ポテンシャルとして埋もれ度ポテンシャルが計算されてきた。しかし、従来のポテンシャルは計算コストが低い一方で精度に問題があり、天然構造が必ずしも有意に低いエネルギーを持たないなどの問題があった。近年では埋もれ度の予測に関する研究の進展により、配列上連続する残基のアミノ酸種情報や、アミノ酸配列プロファイル情報を利用することで高い精度で埋もれ度を予測することが可能となっている。そこで本研究では機械学習アルゴリズムの一種であるSupport Vector Regression (SVR)によって埋もれ度を予測し、その値を利用することで、精度を向上させた埋もれポテンシャルを開発した。

構造サンプリングの問題と切り離して新たに開発したポテンシャルを評価するため、天然のタンパク質様の構造からなる幾つかのデコイセットを利用したテストにより性能評価を行った。ポテンシャルの評価には、エネルギーによって天然構造がその他の構造から有意に識別可能であるかを判定するために、『天然構造のエネルギーのデコイ構造群のエネルギーに対するz-score』と、サンプリングの効率化に寄与する構造類似度とエネルギーとの間の相関をみるために『天然構造に対するrmsd値とエネルギー値とのピアソン相関係数』の2つを用いた。新たなポテンシャルはz-scoreで-1.38と良い値を示し、z-scoreが-0.08であった従来のポテンシャルに対して大幅な性能向上が見られた。また、rmsd値とエネルギー値との相関も0.4と高い値を示し、全原子を利用した従来のスコア関数であるVERIFY3Dの0.16に対して非常に高い性能を示した。

4．まとめ

本研究において、アミノ酸配列から一定の構造をとらないdisorder領域を予測し、一定の構造をとる領域に対して立体構造の予測を行う一連の手法の開発を行い、disorder領域予測、立体構造予測共に従来の手法に対して精度の改善が得られた。しかし、立体構造予測の精度は機能解析や新規タンパク質のデザインといった用途にはまだまだ不十分なものである。今後全原子のモデルを用いたリファインメントなどの更なる改良が必要であると考えられる。一方disorder予測の精度は十分に高いものであり、今後同定されたdisorder領域がどのような機能を持つのかという、タンパク質機能の推定へと繋げることができるものと期待される。