様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 １１ 日現在

機関番号：３２６０７
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１２７８０

研究課題名（和文）水和構造評価も可能なタンパク質立体構造評価のための新規評価指標の開発

研究課題名（英文）A new scoring function for protein structure assment based on the hydration structure information.

研究代表者
清田　泰臣（Kiyota, Yasuomi）

北里大学・薬学部・助教

研究者番号：５０６３１６４４

交付決定額（研究期間全体）：（直接経費）　　1,200,000円

研究成果の概要（和文）：タンパク質複合体界面において、水和構造の理解は益々重要な課題となっている。水和構造を正しく評価できると、タンパク質間相互作用を安定化する要因への理解にもつながり、人工抗体などの立体構造予測にも応用が可能となる。本研究では、このようなタンパク質精密立体構造予測において重要となる、予測構造の構造評価指標について、同じく重要視されるようになった「タンパク質の水和情報」に基づき新規に開発することを目的とした。厳密な統計力学理論である3D-RISM理論から得られた精確な溶媒和エネルギーに、機械学習などの情報学的な処理を加えることで、従来の構造スコアよりも、高速で高精度な構造スコアの開発を目指した。

研究成果の学術的意義や社会的意義
この研究は、水が強く関与するようなタンパク質機能において、その分子機構の解明にもつながるような技術である。本研究で開発される構造スコアは、従来法とは異なり、どのようなタンパク質表面が水和により安定化または不安定化するのか、という情報を含むことになる。そのため、タンパク質の相互作用部位予測やホット・スポット予測、さらにはアミノ酸の変異による親和性の変化など、薬学の幅広い分野に応用が可能になると考える。今後、小分子化合物群などにもデータベースが広がれば、今まで発見が不可能であった2次的な薬剤結合部位の予測などにも応用することが可能になるだろう。

研究成果の概要（英文）：In the field of protein complex study, to understand the hydration structure is becoming an increasingly important issue. If hydration structure can be correctly evaluated, this knowledge reveals the factor for stabilizing the protein-protein interaction and reaches accurate three-dimensional structure prediction technique. In this study, we aim to develop a new structure evaluated score for the predicted structure, which is based on the "protein hydration structure information". Machine learning was performed on the solvation energy obtained from the strict statistical mechanics theory, 3D-RISM theory.
The structure evaluated score is faster and more accurate than the conventional structure score.

研究分野：構造バイオインフォマティクス

キーワード：構造バイオインフォマティクス　タンパク質　構造評価指標　3D-RISM理論　機械学習

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

(1) 水和構造を考慮した構造予測スコアの必要性

創薬分野、特にタンパク質複合体の相互作用界面を対象とした創薬において、「水和構造」の理解は益々重要な課題となっている。2015 年には、タンパク質間相互作用解析の国際的なコンテスト“ CAPRI（Critical Assessment of PRediction of Interactions）”においても、タンパク質複合体間の水の位置を予測する課題が出題された（Round34, http://www.ebi.ac.uk/msd-srv/capri/round34/）。このコンテストに申請者の研究室も参加し、良い成績を残すことができたが、「予測精度の高い構造を構築できていた」にも関わらず、「そのような構造を選び取る（＝判定する）ことができなかった」という課題を残した。

(2) 当時の既存構造予測スコアとその問題点

従来の多くの構造スコアは、二つの要素から構成されていることが多い。一つは「統計学的スコア」であり、PDB（Protein Data Bank）などのタンパク質構造データベースに登録された、膨大な既知構造情報に基づいたスコアである。もう一つは、「物理化学的スコア」であり、物理化学ポテンシャル（分子力場など）を活用したスコアである。このスコアは、ループ構造やアミノ酸の側鎖など細やかな差異の判別を得意とする。構造スコアとして有名な ProQ2 (Ray et al.,2012)、RosettaScore(Kuhlman et al.,2000)、PRESCO(Kim et al.,2015)などは、これらの要素を何種類ものパラメータにより制御することで、タンパク質立体構造予測の国際的なコンテスト“ CASP（Critical Assessment of protein Structure Prediction）”でも好成績を残している。

これら従来の構造スコアでは、水和の効果を表面積に依存するスコアとして取り扱う場合が多い。これは、溶媒和エネルギーがタンパク質の表面積に依存するためである。このようなスコアは簡便ではあるが、表面の環境を画一的に扱ってしまうことにより、水和構造も含めた高精度な構造判定法にはなっていないと考えた。

２．研究の目的

そこで本研究では、統計力学理論であり、精確な溶媒和エネルギーを計算することのできる 3D-RISM 理論に対して、周辺残基の影響を考慮するために機械学習などの情報学・統計学的な処理を組み合わせることで、従来の構造スコアよりも高速で高精度な構造評価指標の開発に望んだ。より具体的には、

ｉ）3D-RISM 理論から得られた溶媒和エネルギーを基にデータベースを作成し、情報処理技術を組み合わせ、新規構造スコアを開発する。

ii）新規構造スコアを用いて、タンパク質立体構造予測コンテストの予測構造に対する構造評価を行う。問題点がある場合は、ｉ)へとフィードバックし、スコアの改良を行う。

ことを目的とした。

３．研究の方法

新規スコアの開発に向け、以下の順で研究を遂行した。

(1) 溶媒和エネルギー計算に用いるタンパク質立体構造情報の取得

本研究ではまず、多様なタンパク質フォールドを含んだタンパク質構造情報のデータセットを用意する必要がある。そこで、タンパク質をフォールドレベルで管理したデータベースである SCOP2（Structural classification of proteins）データベース（Andreeva, et al.,2014）に登録された、1011 鎖のタンパク質構造情報を用いた。これらの構造に対して、構造情報の欠損などをチェックした後に、網羅的に 3D-RISM 理論を適応し、水和エネルギーをアミノ酸残基単位で計算した。

(2) アミノ酸残基単位で管理された水和エネルギーデータベースの構築

アミノ酸残基毎の表面積に対する水和エネルギーデータを解析した。相関傾向はアミノ酸の種類により大幅に変化するが、全てのアミノ酸で相関が得られかを確認した。相関から外れたデータ群は、周辺残基の影響によって外れた可能性があり、次にその影響を機械学習により補うことを考えた。

(3) 水和エネルギーデータベースに対する構造スコア開発のための機械学習

構築した水和エネルギーデータベースは、「相関に帰属できるか」という情報の他に、周辺残基環境や二次構造情報、アミノ酸の存在位置（タンパク質表面、界面、内部など）の情報を識別子として持つことになる。このデータベースに対して、Python の Keras ライブラリを用いたディープラーニングによる分析を行った。ディープラーニングのバックエンドとして Theano という機械学習装置を用いることで、相関から外れる要因を特定した。

(4) 新規構造スコアによる立体構造予測コンテストの予測構造に対する構造評価

開発した新規構造スコアの構造評価能を検証するため、2018 年開催予定のタンパク質立体構造予測の国際コンテストである CASP12 で出題されたタンパク質群を対象として、ブラインドテストを行った。CASP で出題されるタンパク質を対象とする理由は、100 題近くのタンパク質に対して、各チームの最新の構造スコアにより、多くの比較・検討がなされるためである。これにより、開発したスコアの有用性を、公正にかつ客観的に確認することができる。CASP においては、各チームが予測した構造をサーバに登録する。それらの登録された予測構造に対し

て評価を行うことで、新規スコアの構造評価能を検証した。
(5) 立体構造予測コンテストでの予測性能を基にした構造スコアへのフィードバック
CASP に出題されるタンパク質には、構造データベースには存在しない、「新規フォールド」や「新規相互作用界面」が含まれることがある。これらのタンパク質は”Hard Target”と呼ばれ、構造予測が非常に困難であるとされている。このようなタンパク質のデータも新たにデータベースへ取り込むことで、構造スコアの改良を行った。また、既存の高精度タンパク質構造評価スコアである ProQ3 とを併用することで、モデリングされたタンパク質立体構造の精度予測の性能が向上するかを検証した。

４．研究成果

(1) アミノ酸残基単位で管理された水和エネルギーデータベースの構築
例えばメチオニンのような疎水性残基において、二乗誤差刈り込み平均最小化による回帰分析により、70%近くのデータが表面積に対する高い相関傾向（r＝-0.80）を持つことが確認できた（図１）。相関傾向はアミノ酸の種類により大幅に変化するが、全てのアミノ酸で相関が得られることが確認できた。

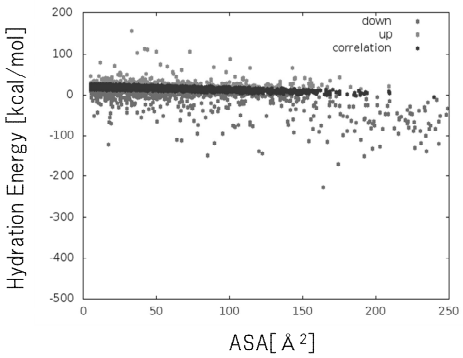

図１. メチオニン残基における表面積と水和エネルギーの相関傾向. 70%近くのデータが一つの高い相関（r＝-0.80）にのる。

(2) 水和エネルギーデータベースに対する構造スコア開発のための機械学習
構築したタンパク質表面情報－水和エネルギーを集積したデータベースは、SCOP2 に登録された様々なタンパク質を基にしており、注目するアミノ酸残基の表面積の情報、周辺残基の情報、二次構造情報、そして 3D-RISM 理論により計算されたその残基の水和エネルギーの情報を、アミノ酸残基ごとに参照することのできるデータベースとした。このデータベースをデータセットとして、Python の Keras ライブラリを用いたディープラーニングによる分析を行った。ディープラーニングのバックエンドとして Theano を用い、「相関にのらない表面」での補正値に対する回帰式を、前述したデータセットから学習させた。これにより、3D-RISM 理論により求まる、アミノ酸残基毎の「表面積－水和エネルギー」の関係を厳密に再現することに成功した。
(3) 新規構造スコアによる立体構造予測コンテストの予測構造に対する構造評価
作成した新規構造スコアは、”Hard Target”と呼ばれるような新規相互作用界面に弱く、十分な精度が得られなかったため、既存の高精度タンパク質構造評価スコアである ProQ3 とを併用することを考えた。対象にしたタンパク質群は 2016 年に開催されたタンパク質立体構造予測の国際コンテストである CASP12 において提出された、様々な研究グループにより予測されたタンパク質立体構造モデル群である。検証の結果、実際の構造精度を示すスコアと予測した構造精度を示すスコアの相関が良くなり、性能向上が認められた。これにより、3D-RISM 理論による水和エネルギーを考慮した新規構造スコアの開発に成功したと考える。機械学習による構造スコアの精密化が済み、精度検証も行ったため、現在、論文として発表する準備を行っている。

５．主な発表論文等

〔学会発表〕（計 3 件）
①Yasuomi Kiyota, Yudai Yamamoto, Katsuya Naito, and Mayuko Takeda-Shitaka
Structure prediction and evaluation for protein complex in CASP12/CAPRI Round 37
第 45 回構造活性相関シンポジウム(2017)

②清田泰臣、山本裕大、内藤克哉、志鷹真由子
タンパク質複合体立体構造予測の国際コンテストにみる予測精度の評価と現状
日本薬学会第 138 年会(2018)

③清田泰臣、竹田－志鷹真由子
水和構造評価も可能なタンパク質立体構造評価のための新規構造スコアの開発
日本薬学会第 139 年会(2019)

６．研究組織

(1)研究分担者
研究分担者氏名：
ローマ字氏名：
所属研究機関名：
部局名：
職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。