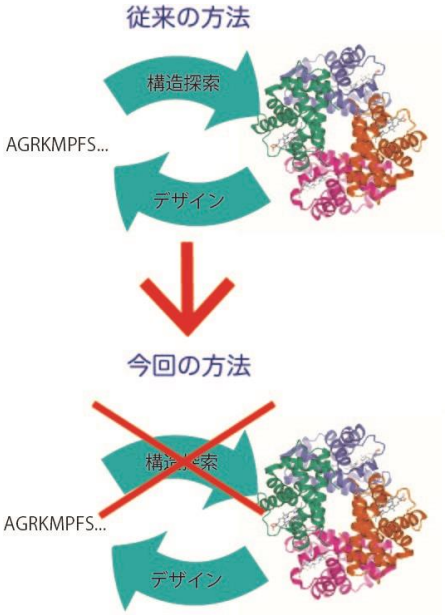

AI を用いた新薬の開発に役立つ物理学と ベイズ統計に基づいた新理論の提案

国立大学法人東海国立大学機構 名古屋大学大学院情報学研究科の高橋 智栄 博士後期課程学生、時田 恵一郎 教授、名古屋大学大学院工学研究科の千見寺 浄慈 助教は、タンパク質を効率的に設計できる新たな理論的手法を提案しました。

タンパク質は、生物の体の中で固有の複雑な立体構造を形成することで機能を発現しています。その立体構造の情報をもとに、アミノ酸配列を予測することで必要な機能を持つタンパク質を設計することを「タンパク質デザイン」と言います。「タンパク質デザイン」は創薬分野への応用が期待されている非常に重要な研究で、実用化に成功すれば安価で迅速な医薬品開発が実現するため、世界的に注目されています。

「タンパク質デザイン」は、本来は配列と構造の両方を同時に最適化せねばならず計算量的に非常に難しい問題であるため、物理学に基づいた理論的な研究は近年ほとんど進展していませんでした。また、最近ではそのような二重の最適化をすることなく現実的なタンパク質のデザインを一定の精度で実行する計算ソフトが開発されていますが、その理論的な裏付けに関する研究もほとんど進んでいません。

本研究では、この二つの方策のギャップを物理学とベイズ統計に基づく機械学習であるベイズ学習[注 1)]の方法によって埋め、「なぜタンパク質はうまくデザインできるのか？」という疑問に答えることで、さらなる応用のための基礎となるような理論を提案することに成功しました。

本研究で得られた成果は、実際のタンパク質デザインで使われている計算ソフトの理論的基礎を明らかにし、より有効な手法の実現への足掛かりとなることが期待されます。また、より現実に近いタンパク質への理論の拡張も容易です。さらに、近年、分子生物学や細胞生物学の分野で注目されている相分離生物学の研究成果を、物理学及び情報学的に裏付ける理論の一つとみなすことも可能です。

本研究成果は、2021 年 7 月 8 日付アメリカ物理学会が刊行する雑誌『Physical Review E』に掲載されました。

【ポイント】
・タンパク質デザインはアミノ酸配列と立体構造の両方の最適化を行う必要があるため膨大な計算量を要する。
・アミノ酸配列に関する新しい仮説を提案し、それをベイズ学習における事前分布[注2)]に反映させることで、立体構造部分の最適化を省略する新しいデザイン手法を考案した。
・結果的に、現在タンパク質デザインの分野で広く用いられている計算ソフトの手法を統計力学的に再現する理論を作ることができた。

【研究背景と内容】
タンパク質は、生体内においてそれぞれの機能に対して固有の複雑な立体構造を形成します。その複雑な立体構造は、遺伝情報から生成される、一次構造であるアミノ酸の配列から決まることが知られています。アミノ酸配列から立体構造を予測することをタンパク質立体構造予測と言います。

これに対し、今回の研究は「タンパク質デザイン」と呼ばれるものであり、立体構造予測とは逆に、与えられた立体構造を実現するアミノ酸配列を予測すること、すなわちタンパク質を「設計する」という問題です。

与えられた立体構造をデザインすることは、必要な機能を持つタンパク質をデザインすることにつながるため、タンパク質デザインは、高分子医薬品開発、人口酵素開発、機能性高分子材料の開発などの重要な工学的応用分野を持ちます。例えばタンパク質デザインの技術が高分子医薬品のデザインにおいてうまく実用化されれば、膨大な時間と開発費を要する医薬品開発が安価で効率的なものとなり、社会的インパクトの大きさは計り知れません。

タンパク質が固有の立体構造を形成している状態を天然状態と言います。アンフィンセンのドグマ[注3)]より、天然状態は、アミノ酸配列と適切な生理条件で決まる熱平衡状態であると考えられています。つまり、ある与えられたタンパク質の立体構造（ターゲット構造）をデザインするためには、その立体構造を低温においてただ一つの基底状態とする配列を求める必要です。そのため、デザインの過程で、ある配列を得るたびにその配列が取り得る立体構造を実験によって求めるか、計算機シミュレーションによって探索し（立体構造予測）、ターゲット構造を唯一の基底状態とする配列であるのかを確かめなければなりません。つまり、タンパク質デザインは本来、配列と立体構造の二重の最適化計算を含む、立体構造予測以上に計算量的に困難な問題なのです。実際、平衡統計力学の方法を用いた過去の理論的な研究では、このような二重のループをどのように効率的に実行するかが重要な論点でしたが、アミノ酸残基数が数百から数千にもなる現実的なサイズのタンパク質の構造に対しては計算量が非現実的に膨大となり、この二重ループに立脚した研究の進展は実質的にストップしていました。

一方で、ターゲット構造のエネルギーを最小化するというシンプルな方法でデザインを実行する計算ソフトの開発も進められていますが、この方法ではターゲット構造以外の別の構造のエネルギーを最小化することになりかねず、なぜデザインがうまく

いくのかは自明ではないという問題がありました。そこで、本研究では、この二つの状況のギャップを埋める理論を提示するということに主眼をおきました。

本研究におけるデザイン手法は、情報学における最新理論の一つであるベイズ学習と、理論物理学の一分野である統計力学を組み合わせるというものです。

ベイズ学習においては、観察に基づくデータの出現確率を表す尤度関数と、研究者の直観や事前に得られている知見・情報、あるいは計算のしやすさによって設定される、尤度関数の中のパラメータの出現確率である事前分布 の積によって、パラメータの事後分布を求め、さらに事後分布を基にパラメータの推定やデータの予測などを行います。本研究においては、データをタンパク質のターゲット構造、パラメータをアミノ酸配列とし、タンパク質デザインの問題をベイズ学習のパラメータ推定問題に帰着させました。

尤度関数には、統計力学において使われる確率分布であるグランドカノニカル分布[注4)]を用いました。これにより、タンパク質のエネルギーとして、アミノ酸残期間の疎水性相互作用だけでなく、タンパク質の表面の親水性残基に結合する水分子の寄与も計算に入れることができ、水の化学ポテンシャルによってタンパク質表面の親水性残基数をコントロールすることが可能になりました。

本研究の理論の骨子は、アミノ酸配列の事前分布に対する新しい仮説です。あるターゲット構造のみを最適化する配列はあらゆる配列パターンの中で極めて珍しいはずであり、そのような配列は、その与えられたターゲット構造の自由エネルギーを最小にするように進化してきたはずです。そこで本研究においては、「タンパク質の自由エネルギーを最小にする配列の出現確率が最も高くなる」という仮説を事前分布に反映させました。

この事前分布は配列に依存する大分配関数に比例するため、その部分が尤度関数の分母とキャンセルし、結果的に、立体構造探索に当たる計算ルーチンなしの配列の事後分布が得られました。すなわち、現実のタンパク質デザインに成功している方法であるエネルギー最小化をベイズの定理から導いた形になっています。事後分布を与えるアミノ酸配列のデザインには、ギブスサンプリングというマルコフ連鎖モンテカルロ法の一つを用いました。

結果は、2次元の格子HPモデルに対しては6割から8割という高いデザイン精度が得られました。一方、3次元では3割程度という低い結果となりましたが，この3次元構造は、全構造パターンが枚挙可能な小さな構造であるため、タンパク質のモデルとしては非現実的であり、より現実的なタンパク質モデルに対するデザイン精度を明らかにすることが今後の課題です。しかし3次元の構造でも、先行研究において一つのベンチマークとなっているアミノ酸残基数（N=3×3×3=27）の立方体型の構造の中で最もデザインの解の個数が多くデザインが容易な構造に対してはデザインが成功しました（図1）。また、2次元の場合は先行研究においてデザインに成功しているN=50という比較的大きなサイズの構造を先行研究よりも短時間でデザインすることに成功しました（図2）。

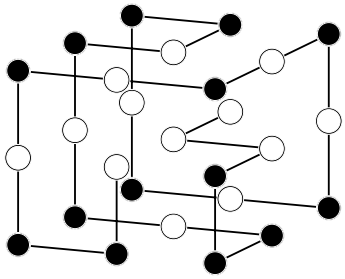

図 1：3×3×3の全てのコンパクトな構造の中で最もデザインの解の個数が大きい構造のデザイン結果。白丸が疎水性アミノ酸残基を表し黒丸が親水性アミノ酸残基を表す。この配列は正しいデザイン結果であり、つまりあらゆる可能な3×3×3の構造の中でこの構造のみを唯一の基底状態とする配列。

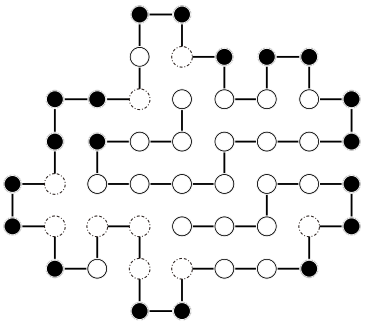

図 2：A. Irbäck らによる先行研究(A. Irbäck et al.、 Structure 1999)にて上で述べた2重ループに基づく手法でデザインに成功した構造を我々の手法でデザインした結果。この配列は Irbäck らが求めた配列と完全に一致する。点線で囲んだ残基は表面残基数を数えるときに表面残基に含めない残基として明示的に示したもので、全て疎水性アミノ酸残基である。

また、2次元モデルのデザインにおいては、コンパクトな長方形の構造よりも、一部分がほどけた部分のある完全にはコンパクトでない構造の方が、デザインの成功率が高いことも分かりました。この事実は、水に露出している部分が通常の球状タンパク質よりも多く、一部分がほどけた状態で存在してその機能を発現している天然変性タンパク質のデザインへの可能性を開く可能性を示唆しています。天然変性タンパク質のデザインはタンパク質工学における未解決の課題の一つです。

また、内側に構造探索を含む先行研究の手法と性能比較をしたところ、先行研究と比較してデザイン精度は低いが、計算時間は1/100と大幅に短縮することができることが分かりました。また、コンパクトな構造よりも、主鎖がほどけた、表面積の大きな構造の方がデザイン精度は高くなることも分かりました。また、デザイン精度を最大化する化学ポテンシャルとターゲット構造の表面残基数の関係を調べたところ、おおよそ比例関係にあることが確認されました（図3）。

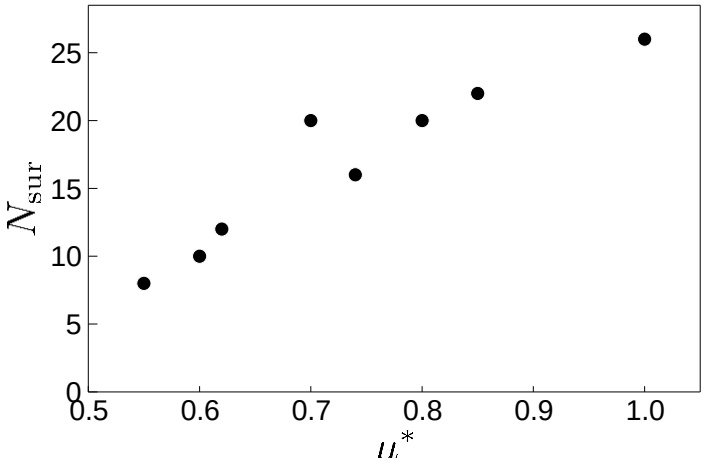

図 3：デザイン精度を最も高める化学ポテンシャル μ^* とデザインした構造の表面残基数 N_{sur} の関係。おおよそ比例関係にあることが分かる。

これらの結果が示唆することは、化学ポテンシャルによってタンパク質表面の水和効果をコントロールする要素を取り入れると、全てではないが、エネルギー最小化によってもタンパク質を正しくデザインできる場合が多くあることです。またこれは、タンパク質の周りを取り囲む水の分布の仕方が、タンパク質がある特異的な天然構造へと折れたたむ確率と密接に関係していることを統計力学的にシンプルな形で示すことができています。

【成果の意義】

- 統計力学的には、内側に構造探索を含めないと原理的には正しくデザインできないと考えられていたが、水との結合を考えると、構造探索という時間のかかるステップを踏まなくても正しくデザインできる構造も多いことを示しました。
- 構造探索は現実の、より大きなサイズのタンパク質（残基数数百～数千）においてはスーパーコンピュータを用いても現実的な時間内には不可能なため、このように計算時間が短縮できることは実用上非常に役に立ちます。
- Rosetta という現実のタンパク質を合理的にデザインするために標準的に用いられ成功を収めている計算ソフトでは、表面残基に水と強く相互作用する親水性アミノ酸残基を配置し、内側には疎水性アミノ酸によるコアを作るような束縛条件を人間が指定します。この方法でデザインできる理由は明らかではありませんでしたが、本研究ではその理由を説明する理論を提案しています。

・近年、タンパク質をはじめとする細胞内の器官が互いに液滴を形成して生命現象において本質的な機能を果たしていることが分かってきています。本研究は水との相互作用を取り入れることによってデザイン精度を高めているという点において、相分離生物学[注5)]との相性もよく、今後の展開が期待されます。
・タンパク質デザインの問題をベイズ学習によるパラメータ推定というシンプルかつ一般的な形に定式化することができました。情報学視点からは、二重のループの方法は最尤法として捉えることができ、新たなパラメータを導入してもそれに対応する事前分布を考えることでベイズ学習のモデルを拡張することが可能です。ベイズ学習の理論は数学的に精緻で一般的な形で整備されているため、そのように拡張されたモデル同士を、その理論の枠組みの中で統一的に比較検討することができます。このように、複雑な生命現象をベイズ学習という最新の情報学的手法で定式化・解析したことには、情報学的な意義も大きく、2017 年に新設された名古屋大学大学院情報学研究科が目指す教育・研究の成果の一つであると考えられます。

【用語説明】

注1)　ベイズ学習：

ベイズ統計に基づいたシンプルかつ効率的な機械学習手法。ベイズ統計学は、古典的な頻度論主義の統計学と異なり、統計モデルのパラメータ（母数）までも何らかの確率分布（事前分布）に従うとする。そして、観測されたデータを固定した上で、パラメータのデータによる条件付き確率（事後分布）を求め、事後分布をもとに推定や予測を行う。マルコフ連鎖モンテカルロ法というコンピュータによる効率的なサンプリング手法の発達により、事後分布からの推定や予測は容易になり、ベイズ学習はその実用性を格段に増している。

注 2)　事前分布：

統計モデルのパラメータが従うとする確率分布のこと。データを観察して比較的客観的に仮定できる統計モデル（尤度関数）に対し、データが観測される以前に利用可能な情報をもとに仮定する必要がある。ドメイン知識からの知見やそれに基づく分析者の直観、あるいは事後分布の計算のしやすさなどから決定する場合が多い。あるいは考えている状況や経験事実から明らかな場合などもある。

注 3)　アンフィンセンのドグマ：

温度や溶媒の濃度などの環境条件が適切であれば、タンパク質の天然構造はアミノ酸配列のみで決まる熱平衡状態であるとする分子生物学上の熱力学的仮説。1950 年代から始まる、クリスチャン B. アンフィンセンらによる、リボヌクレアーゼを一度変性させたあともう一度環境を戻すことで自発的に天然構造を取り戻すことを確かめた実験による。この業績により、アンフィンセンはノーベル化学賞（1972 年）を受賞している。

注4)　グランドカノニカル分布：

温度と化学ポテンシャルをパラメータにとった平衡状態における系の力学変数が従う確率モデル。注目する系が熱浴とエネルギーだけでなく粒子数もやりとりす

ると考える。体積が大きい極限で、カノニカル分布とミクロカノニカル分布と熱力学的に等価になる。注目する物質の界面に分子が吸着する系の平衡状態などの解析に使われる。

注5)　相分離生物学：

タンパク質をはじめとする細胞内小器官が、互いに液滴を形成する（液-液相分離という）ことで様々な生命現象を担っているという近年の研究成果に基づいた新しい見方による細胞生物学・分子生物学の潮流。近年多くの重要な成果を残しており、大きな注目を集めている。

【論文情報】

雑誌名：Physical Review E
論文タイトル：Lattice protein design using Bayesian learning
著者：Tomoei Takahashi、 George Chikenji、 and Kei Tokita
DOI：10.1103/PhysRevE.1-4.014404
HP ：https://journals.aps.org/pre/abstract/10.1103/PhysRevE.104.014404