学位論文内容要旨

北里大学 薬学部
氏名：小林 慎平

【題目】
三次元分布関数を用いたタンパク質周辺の水分子のサンプリング手法の開発

【論文目録】
Kobayashi S., Kiyota Y., Takeda-Shitaka M., Bulletin of the Chemical Society of Japan, 97(6), uoae063 (2024).

【背景・目的】
タンパク質が他のタンパク質やリガンドと相互作用する上で、水分子の存在はきわめて重要である。タンパク質周辺の水分子の位置を知ることは、タンパク質の機能や相互作用の理解、創薬研究などにつながる。

タンパク質周辺の水分子の位置を予測する手法には、様々なアプローチがある。その 1 つが、3D-RISM 理論に代表される、水の三次元分布関数を求める方法である。この方法は、標的タンパク質の周辺の水分子の存在しやすさを三次元分布関数、即ち水の確率密度として計算することができる。近年では AI を用いて三次元分布関数を計算する手法も開発されており、このアプローチは今後さらに興味深いものとなる可能性がある。

3D-RISM 計算等で得られる三次元分布関数は、コンピュータ上の三次元グリッド空間上において、標的タンパク質を格納したボックスを計算領域として計算される。水の確率密度は計算領域内の全てのグリッド点について計算される。三次元分布関数が計算されたならば、それを基に、明示的な水分子への変換、即ち水分子の位置予測を行うことができる。三次元分布関数を基に、明示的な水分子の位置を予測する手法は既にいくつか開発されており、代表的な手法としては確率密度が最大の位置への水分子の配置を繰り返す手法である「Placevent」が挙げられる。本研究においては、三次元分布関数を基に明示的な水分子の位置を予測した 1 つの水和構造を、「サンプル」と定義した。

本研究では、重み付きモンテカルロ法を水分子のサンプリングに適用した生物分子設計学教室の先行研究を応用し、水の酸素原子の三次元分布関数を基に、水の位置を予測する手法である「DroPred」を開発した。本研究では、「水の位置」として、水の酸素原子の座標を予測した。DroPred は、重み付きモンテカルロ法をベースに、三次元分布関数から得られる確率密度を調整した重みを使用し、水の確率密度が高いグリッド点を優先しながら水分子を配置するサンプリング手法であり、1 つの三次元分布関数から複数のサンプルを得るこ

とができる。

【方法】

三次元分布関数に基づく調整重みの導入

重み付きモンテカルロ法とは、モンテカルロ・サンプリングの手法の一つで、確率密度関数に従って重み付けを行い、重みの大きい事象を優先的に発生させる方法である。水分子のサンプリングについて考えた場合、一般的なモンテカルロ・サンプリングであれば、発生させた乱数と、計算に使用するボックスのグリッド点を対応させることで、水分子のサンプリングを行うことができる。これに対し重み付きモンテカルロ法では、各グリッド点を、三次元分布関数から得られる確率密度に基づいて重み付けすることで、水分子を確率密度の高いグリッド点に優先的に配置することが可能となる。

式(1)に示す累積確率密度関数を定義することにより、0 から 1 の値域で発生させた乱数を三次元座標と対応させることができる。

$$P(x',y',z') = \frac{1}{N}\int_{-\infty}^{x'}\int_{-\infty}^{y'}\int_{-\infty}^{z'} \rho_0 g_0(\mathbf{r})|_{g_o(\mathbf{r})\geq c}\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{1}$$

ここで、ρ_0は水の酸素原子の数密度であり、$g_0(\mathbf{r})$は分布関数から得られる確率密度の値、cは確率密度のクライテリアである。本研究では、確率密度のクライテリアを$g_o(\mathbf{r}) \geq 2$とした。また、Nは累積確率密度関数 Pの値域を 0 から 1 にするための規格化定数である。本研究では、確率密度の大きさによる予測水の配置されやすさを調整するため、指数 w を導入した。これにより、確率密度の高い場所への予測水の配置確率を高め、確率密度の低い場所への予測水の配置確率を抑えることができる。(式(2)) 本論文においては、重みを調整するための指数 wを「調整指数」、調整された重みを「調整重み」と定義した。

$$P(x',y',z') = \frac{1}{N_w}\int_{-\infty}^{x'}\int_{-\infty}^{y'}\int_{-\infty}^{z'} \rho_0 \left(g_0(\mathbf{r})|_{g_o(\mathbf{r})\geq c}\right)^w \mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{2}$$

3D-RISM 計算における三次元分布関数は、計算領域内の各グリッド点(x_i, y_j, z_k)に確率密度$g_o(x_i, y_j, z_k)$として与えられる。本研究では、3D-RISM 計算によって与えられた確率密度を基に、確率密度$G_o(x_i, y_j, z_k)$を再定義した。

$$G_o(x_i, y_j, z_k) = \frac{1}{8}\sum_{l=0}^{1}\sum_{m=0}^{1}\sum_{n=0}^{1}\left(g_o(x_{i+l}, y_{j+m}, z_{k+n})|_{g_o(x_{i+l},y_{j+m},z_{k+n})\geq c}\right)^w \tag{3}$$

これにより、式(2)は以下のように表される。

$$P(X',Y',Z') = \frac{1}{N_w}\sum_{i=0}^{X'}\sum_{j=0}^{Y'}\sum_{k=0}^{Z'} \rho_0 V G_o\left(x_i, y_j, z_k\right) \tag{4}$$

式(4)に基づいて水分子のサンプリングを行った。

DroPred における水分子の配置アルゴリズム

DroPred における水分子の位置予測アルゴリズムについて説明する(図 1)。

はじめに乱数を発生させ、式(4)に従ってグリッド点の座標へと変換する。次に、そのグリッド点と、既に配置した全ての予測水との距離を計算し、周囲 2.0 Å 以内に水分子が存在しない場合、水分子を挿入する。この操作を、予測水が十分に配置されるまで繰り返す。

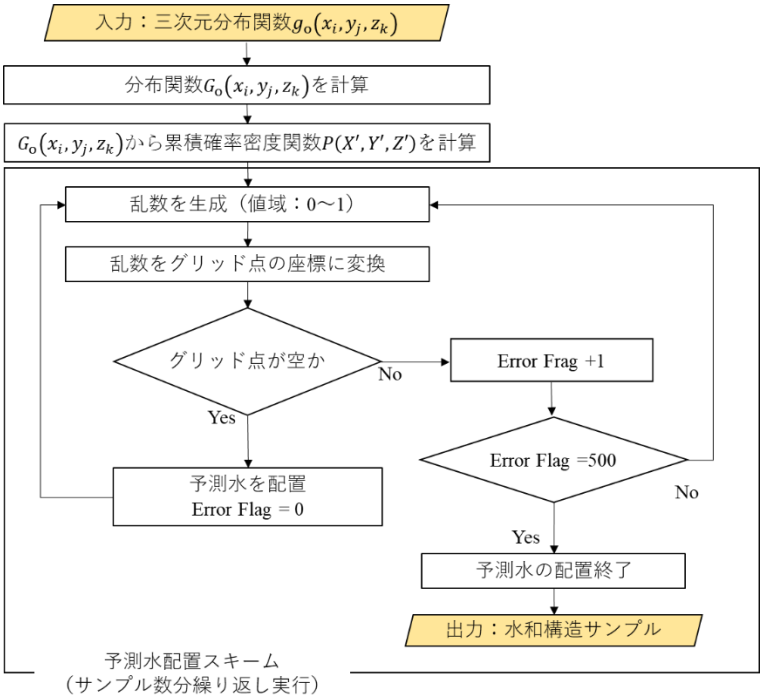

図 1. DroPred による水分子の配置アルゴリズム

データセットの準備

精度検証のため、タンパク質-タンパク質複合体のテストセット(151 ターゲット)を準備した。ターゲットの PDB ファイルに含まれる実験構造中の水分子(結晶水)のうち、タンパク質-タンパク質界面に存在し、B-factor が 40 未満であるものを評価の対象とした。

3D-RISM 計算

DroPred および Placevent の入力となる三次元分布関数は、3D-RISM 計算によって取得した。3D-RISM 計算は、AmberTools の「sander」プログラムを使用して行った。3D-RISM 計算の計算コストは計算領域の体積に依存するため、ボックスの大きさが最適となるようにターゲットの向きと位置を調整した。三次元グリッド空間上のグリッド点同士の間隔は 0.25Å とした。

DroPred 及び Placevent の実行

三次元分布関数を入力に DroPred を実行し、タンパク質界面の水分子の位置を予測した。調整指数には 0 から 10 の整数値を使用した（DP_w0～DP_w10）。 ここで DP_w0 は、一般的なモンテカルロ法によるランダムな配置、DP_w1 は三次元分布関数をそのまま重みとして使用した重み付きモンテカルロ法によるサンプリングである。DroPred は任意の数のサンプルを出力することができるが、本研究では各ターゲットについてそれぞれ 1000 サンプルを出力した。性能比較のため、同一の三次元分布関数を使用し、Placevent による予測も行った。Placevent からは、各ターゲットに対して 1 サンプルのみが出力される。

【結果】

評価基準

DroPredの性能は、タンパク質複合体界面に存在する結晶水の再現性によって評価した。本研究においては、結晶水の再現は、その結晶水が存在する位置の周囲 1 Å 以内に予測した水分子（予測水）を配置できた場合と定義した。各標的タンパク質は、複数の結晶水を持つため、サンプルの精度を、予測水が再現した結晶水の割合（coverage)で評価した。

Coverage による比較の場合、予測する水の数は重要な要素である。基本的に、予測する水分子の数が多くなるほど、coverage は上昇する。DroPred によって配置される水分子の数はサンプルごとに異なるが、性能評価においては、配置された順に Placevent と同数までの予測水を使用した。

精度評価

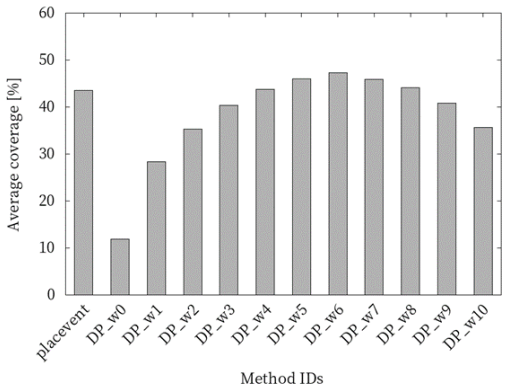

図 2. DroPred による水分子の予測精度

各標的タンパク質について、出力した 1000 個のサンプルの coverage を計算し、中央値を算出した。調整指数ごとの coverage の中央値の平均値(Average coverage)を示す。Placevent においては、標的タンパク質につき 1 個のサンプルが出力されるため、それらの Average coverage となる。

重み付きモンテカルロ法、及び調整指数による重みの調整が、水分子の予測精度に与える影響について検証した。

一般的なモンテカルロ法による予測(DP_w0)と比較して、重み付きモンテカルロ法を用いた予測(DP_w1)では、Average coverage が大きく上昇したが、Placevent と比較すると低く、結晶水を精度よく再現出来ているとは言えない。一方、調整指数によって重みを調整したDP_w2 以降では coverage が更に上昇し、DP_w6 で Average coverage＝47.4%と、coverage が最大となった。この結果は、調整指数による重みの調整が、予測精度を更に上昇させることを示唆しており、DP_w6 では Placevent（coverage=43.5%）と同等以上の精度を期待できる。一方で、DP_w8 以降では coverage が低下したことから、過剰に大きな調整指数は精度を悪化させる可能性がある。

分布関数の再現度の評価

次に、調整重みを使用した重み付きモンテカルロ法による複数サンプルの生成が、基となった三次元分布関数を表現しているかを検証した。

評価対象とした全ての結晶水(5448 個)が存在する地点で、結晶水が存在する位置の確率密度と、1000 個のサンプル中の予測水の出現率の相関を検証した(図 3) 。予測水の出現率は、1000 個のサンプルの中で、その結晶水が再現されたサンプルの割合として定義した。

図 3 に示すように、一般的なモンテカルロ法による配置(DP_w0)では、結晶水が存在する位置の確率密度と、その結晶水の出現率に相関は見られなかった。しかし、重み付きモンテカルロ法による予測(DP_w1)から相関が現れはじめ、調整指数によって重みを調整した場合(DP_w2～DP_w10)には強い相関が見られた。DP_w3 のとき、相関係数が最大となった(R＝0.778)。この結果は、重み付きモンテカルロ法によって配置された予測水が、基となった分布関数をよく表現していることを示唆していると考えられる。

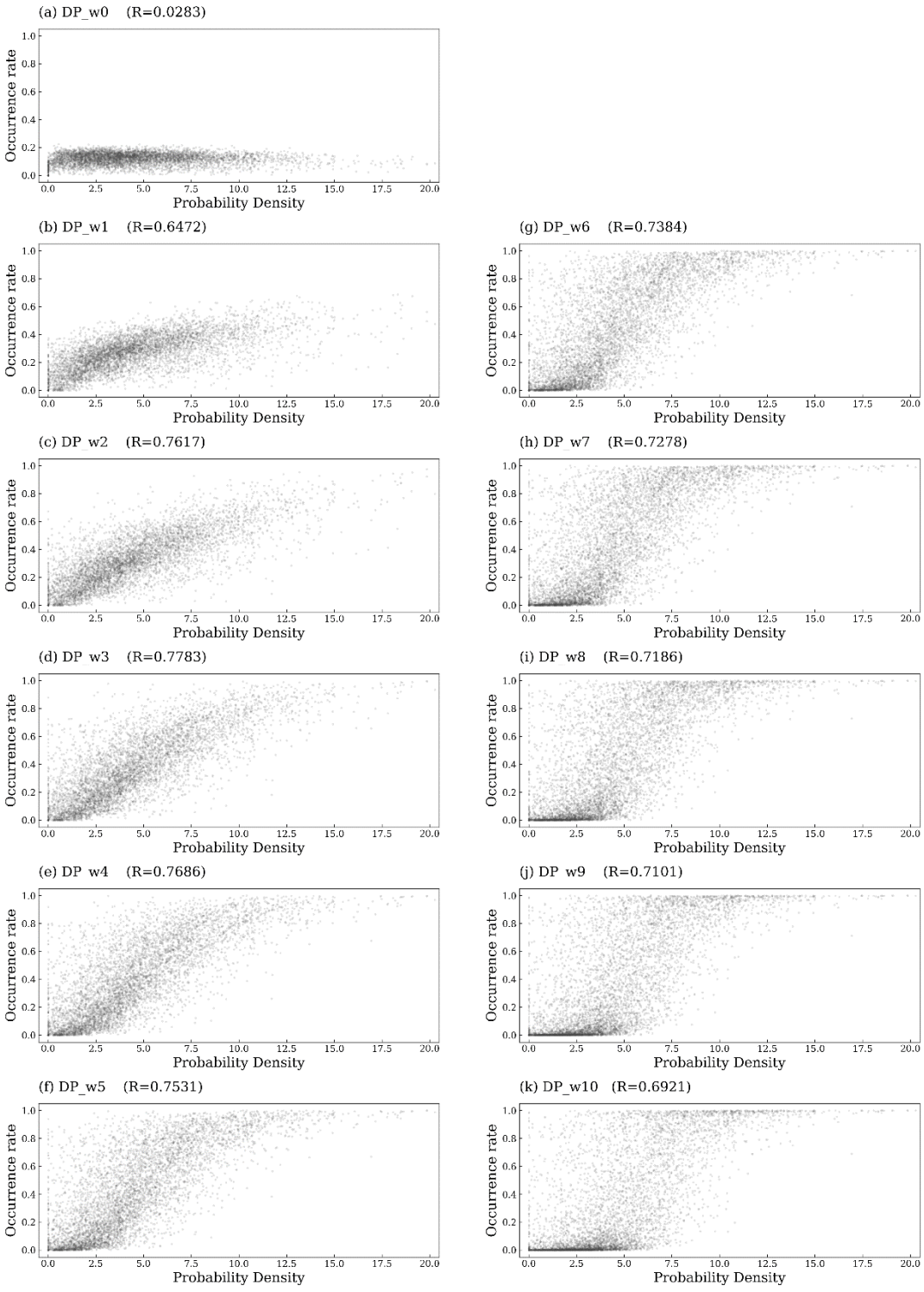

図 3. 結晶水が存在する位置での確率密度と出現率の相関

DroPred による予測例

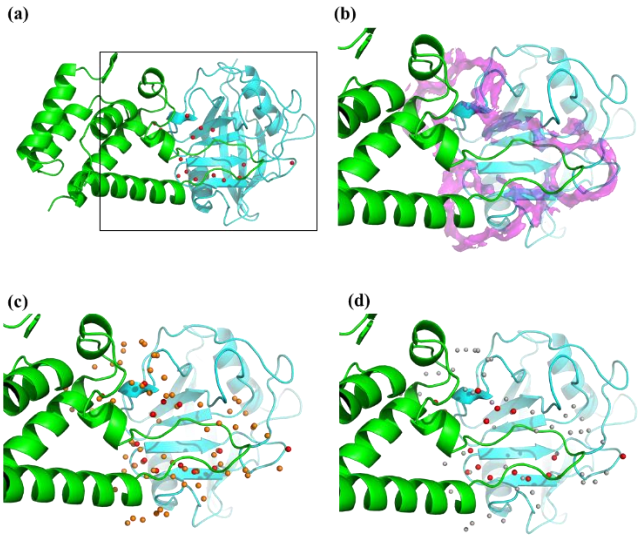

図 4. DroPred における水分子の予測が良好な結果を示した例（PDB ID: 2XGY）

(a)全体構造　赤球：結晶水　四角の枠：相互作用界面

(b) DroPred および placevent の計算で共通して用いた三次元分布関数(紫の透過面)

(c) DroPred（DP_w6）の中で最も coverage の高かったサンプル

赤球：結晶水　オレンジ球：DroPred の予測水

(d) Placevent の予測によって得られたサンプル

赤玉：結晶水　白球：Placevent の予測水

DroPred の予測結果が良好であった例を示す(図 4)。
このターゲット(PDB ID: 2XGY)には、結晶水が 15 個存在するが、DroPred(DP_w6)によって生成された 1000 個のサンプルのうち最も coverage の高かったサンプルでは、coverage は 93.3%であった(図 4(c))。Placevent によって予測されたサンプルの coverage は 60.0%であった(図 4(d))。

【考察】

本研究では、重み付きモンテカルロ法をベースにして、水の三次元分布関数から得られる確率密度を調整した重みを使用し、確率密度の高い場所を優先しながら水分子のサンプリングを行う手法を開発した。予測精度の検証の結果、調整重みの導入が有効であることが確認できた。また、生成された複数のサンプルは、基となった分布関数をよく表現していること、サンプルの中には、Placevent の精度を大きく上回るサンプルも含まれている可能性があることも示唆された。本研究の成果は、水分子を含んだタンパク質の機能や相互作用の解析などに貢献できるものであると期待される。

以上