論 文 の 内 容 の 要 旨

論文題目　A Fast Protein-Protein Docking Algorithm
Using Orthonormal Basis Functions Based on Spherical Harmonics
(球面調和関数に基づく正規直交基底関数を用いた
タンパク質間相互作用様式の高速計算法)

氏　　名　角越　和也

本論文では、球面調和関数を用いて新規に設計した正規直交基底関数による級数展開を用いた、高分子-高分子、特にタンパク質-タンパク質複合体の結合配座予測(ドッキング予測)の高速計算手法を提案する。

生体内には多種の高分子が存在し、それらが相互作用することにより生命活動に必要な各種の複雑な機能が実現されている。中でも特にタンパク質間の相互作用はそれらの機能の大部分を担っており、その機序の解明は多くの分子生物学の研究者の焦点となっている。基本的に、タンパク質間の相互作用は特異的、つまり、あるタンパク質が、ある特定のタンパク質と、ある特定の部位において結合することにより実現される。したがって、そのようにして形成された複合体がどのような配座をとっているかを知ることは、その機能機序の解明において、非常に重要な意味を持つ。複合体構造を知る方法には、X 線結晶構造解析や NMR を用いた解析などにより、実験的に求める方法がある。しかし、複合体によっては、技術的に困難または不可能であったり、多大なコスト、時間を要したりする場合が多い。そのため、計算機を用いてその配座を予測する手法が注目されている。

典型的には、計算機によるドッキング予測のプロセスは、二つの段階から成る。第一段階は、配座空間を全探索し、配座の候補リストを生成するプロセスである。計算機による

ドッキング予測が困難である大きな理由の一つは、この配座空間の広さにある。この段階の手法のほとんどのものは、分子の表現に剛体モデルを使い自由度を減らしているが、それでもまだ6自由度あり、その空間の探索は近年のプロセッサをもってしても依然として多大な時間を要する。次に、第二段階は、第一段階で生成された候補集合の精緻化を行うプロセスである。この段階の手法は、より計算コストのかかる精度の高い計算により、得られた候補リストの再順位付けを行ったり、得られた候補構造そのものの精緻化を行ったりする。本研究では、ドッキング予測で必ず必要となる第一段階のプロセスに主眼を置く。

従来提案されている主な第一段階の手法は、大きく次の三つに分けられる。(1) タンパク質表面の特徴的な点のマッチングによるもの、(2) 分散した複数の初期配座からエネルギーの最小化を行うもの、(3) Fast Fourier Transform (FFT)による高速相関計算を用いた最小エネルギー配座探索を行うものである。近年では、適用可能なエネルギー関数（スコア関数）に柔軟性があり、精度、計算時間の両面である程度バランスのとれた、上記(3)に属するもの、つまり、FFT による 6 次元配座空間の探索に基づく手法が多く提案されている。しかし、ある程度の精度で、ある程度大きなタンパク質のドッキング予測を行うには、近代的なプロセッサ 1 つを用いて、数時間から数日を要するほどの計算量があるため、より高速な手法の開発が求められている。

そこで本研究では、FFT を用いた従来手法より高速でかつ、最小化したいスコア関数の設計に柔軟性がある手法を開発することを目的とした。これらの特徴をもつアルゴリズムの実現により、ドッキング予測の際により詳細で複雑なスコア関数が適用可能になり、また複数のタンパク質間の相互作用の網羅的解析が可能になると考えられる。本論文では、上記目的の達成のため、球面調和関数を用いて新規に設計した正規直交基底関数での級数展開による高速内積計算を使ったドッキング予測アルゴリズムを提案する。本提案手法では、スコア関数を、各分子から定義されるスカラー場の内積の線形和として表す。各配座においてこのスコア関数を計算し、その値が低いものから順に順位付けし、上位のものを候補配座とする。各スカラー場は、その内積が、表現したいエネルギーもしくは性質を反映するように任意に定義できる。このスコア関数の枠組みは非常に柔軟で、分子形状の相補性や各種ペアポテンシャル、静電相互作用などを表現することが可能である。また、本手法では、スカラー場を上記正規直交基底関数で展開することにより、スコア関数の計算に必要な内積計算を高速に行う。また、配座空間の探索に必要な座標変換操作も高速に行えることを示した。球面調和関数を用いたドッキング予測手法は過去にもいくつか提案されており、中でも動径方向の基底関数をとりいれた Ritchie らの手法は、有望な高速計算手法のひとつに挙げられる。しかし、彼らの論文にも述べられている通り、彼らの用いている動径方向の基底関数が、中心からの距離 r が増えるに従い、指数的に減衰することに起因して、展開係数によるスカラー場の表現能力が、r の増加に従って劇的に劣化するという問題点がある。それにより、球形から離れたいびつな形状の分子や巨大な分子への適用が困難となっている。そこで、本手法は上記の問題点がなく、さらによい性能を発揮するため

に、球面調和関数と、中心からの距離 r による減衰のない、修正 Legendre 多項式を組み合わせた基底関数を使い、また、タンパク質の結合エネルギーをよく反映すると報告されている原子レベルの統計ポテンシャルである Atomic Contact Energy (ACE)を、本手法の枠組みで効率的に計算できる形に変換し、スコア関数に取り込んだ。

本手法の性能評価として、実際のタンパク質の構造データを用いて、計算時間と精度の測定を行った。

一組のタンパク質(PDB ID: 1AKZ, 1UGI(A))のドッキング予測を行うのに要した計算時間は、1CPU (Pentium 4, 2.4-GHz)を用いて約 40 秒であった。ターゲットとするタンパク質によって要する時間は若干変化したが、どの場合についても数秒程度の差であった。FFT を用いる手法の代表的なものの一つである FTDock で同一のタンパク質について同一の計算機環境で計算したところ、彼らのデフォルトのパラメータを用いての実行には、約 18 時間を要し、使用パラメータを本手法のデフォルトのパラメータに近いものにした場合、約 100 分の計算時間を要した。空間探索の方法が根本的に違うため、計算精度のパラメータを完全に等価なものにすることはできないことから、正確な比較は困難であるが、上記の得られた計算時間を比較すると、本手法は FTDock と比較して、160 倍から 1700 倍の高速計算が可能となった。FFT を用いる手法と比較して、以上の高速化が図れた主な要因としては、以下の三点が挙げられる。第一の点は、探索に必要なスカラー場の座標変換操作、特に回転操作が高速にできる点である。これは、座標変換されたスカラー場の展開係数が、元のスカラー場の係数から高速に直接計算できる、つまり、座標変換後のスカラー場を求めて、それを展開するという操作を経ることが不必要であるからである。第二の点は、各分子に対する一つのスカラー場の表現に必要なメモリ量が少ない点である。典型的なパラメータを用いた場合、FFT を用いた手法に比べ、本手法では約 1000 分の 1 の量のメモリしか要さない。そのため、非常に多くの数の中間結果をキャッシュしておくことができ、それにより計算の高速化が図れている。第三の点は、FFT を用いた手法では、三次元の平行移動変位の探索空間を一度にまとめて計算してしまうのに対し、本手法は、条件として与えられた一部の空間のみを重点的に探索するのに適していることである。そのため、FFT を用いた手法で計算してしまっていた、明らかに正解とはなり得ない配座、例えば、分子の中心点間の距離が近すぎて分子の衝突を起こしている、または、離れすぎていて相互作用していないような配座の計算を容易に除くことができ、計算の効率化が図れている。

精度の評価は、各候補構造のinterface r.m.s.d.、つまり、正解の複合体構造において相互作用をしている部位の構造がどれだけ予測構造のものと近いかを表す指標、を計算することによって行った。この指標は、従来のドッキング予測の研究で候補構造の評価尺度としてよく使われており、正解に近い構造（near-native構造）とそうでない構造を分ける閾値として 2.5Åから 3.0Å程度の値がよく使われる。６つの組のタンパク質について、「最初にnear-native構造が現れる順位」、「上位からある数の候補をとった際に一番低かったinterface r.m.s.d.」、「上位候補中のnear-native構造の個数」を調べた。その値をFTDockの

ものと比較したところ、ほぼ同じような精度が得られていることが分かった。また、候補構造の上位1000個をとってくると、調べた６つの組のタンパク質のすべての場合において、少なくとも１つのnear-native構造（閾値：3.0Å）が得られていることが分かった。生成された候補構造の個数が約10^7個のオーダであることを考えると、十分な絞込みができたと考えられる。