論文の内容の要旨

応用生命工学専攻
平成 19 年度博士課程進学
氏名　　角田　将典
指導教員名　清水　謙多郎

論文題目
カーネル法を用いたタンパク質の構造および機能の予測

はじめに

カーネル法とは、カーネル関数を用いた機械学習法を指し、カーネル関数はデータをある空間に写像した後のデータ間の内積を計算する関数のことである。カーネル法は 2 つの利点を有する。1 つは、カーネル関数を用いることで、既存の線形手法（主成分分析や、判別分析、正準相関分析など）を非線形化できることである。線形手法を非線形なデータに適用しても良好な結果は得られないが、適切なカーネル関数を用いることで非線形なデータにも対応できるようになる。カーネル関数で写像先の内積を計算し、この値をもとに線形手法を適用することで、元の空間における非線形な結果を得る。カーネル関数で非線形化された手法としては、サポートベクターマシン（SVM）やカーネル主成分分析、カーネル判別分析、カーネル正準相関分析などがある。もう一つの利点は、カーネル関数を用いることで、通常のデータ解析で扱われるベクトルデータだけでなく、文字列や木、グラフなどの構造を有するデータにまで適用範囲が広がることである。構造を有するデータに通常の主成分分析などを適用することはできない。しかし、カーネル関数によって写像先の内積が得られるようになり、その値を利用することで、主成分分析などのもともとはベクトルデータに対する手法を構造を有するデータにも適用可能になる。

本論文では、カーネル法を用いたタンパク質の機能予測と自由エネルギー曲面の解析手法について報告する。一つは、サポートベクターマシンを用いたタンパク質間相互作用予測、もう一つは、Isomap を用いた非線形次元削減による、ポリペプチドの自由エネルギー解析である。

タンパク質ペアに対するカーネル関数を用いたタンパク質間相互作用予測法の開発

タンパク質間相互作用は、多くの生物学的機能に関与している。そのため、細胞内でどのようなタンパク質間相互作用が存在するかを知ることは重要である。しかし、取りうるタンパク質の組合せは膨大であるため、全ての組合せを実験によって確かめるのには多大な労力を必要とする。そこで、本研究ではタンパク質間相互作用の予測手法を開発した。この手法は、2 クラスの分類を行う機械学習アルゴリズムのひとつである SVM を用い、与えられたタンパク質ペアの配列をもとに、そのペアが相互作用するか否かを予測する。

SVM はデータ間の関係（類似度）をカーネル関数によって計算する。SVM はこのカーネル関数によって計算された値を用いて識別関数を構築するため、カーネル関数は予測精度に大きな影響を与える。したがって、精度の良い予測器を得るためには、適切に類似度を計算できるカーネル関数を用いる必要がある。そのため、タンパク質配列のペアに対するカーネル関数の設計を行った。

カーネル関数設計は、構造を有するデータに対するカーネル関数の設計法として知られる Haussler の畳み込みカーネルに基づいて行った。これは、構造を有するデータのカーネル関数を、部分構造のカーネル関数の組み合わせによって定義して再帰的に計算を行うものである。タンパク質ペアのカーネルを、ペアの構成要素であるアミノ酸配列のカーネルの組み合わせで定義し、このアミノ酸配列のカーネルには文字列に対する畳み込みカーネルの一つである spectrum カーネルを用いた。spectrum カーネルを用いたのは、他の文字列に対するカーネルと比べ高速に計算できるためである。このようにして、タンパク質ペアに対するカーネルとして、アミノ酸配列のカーネルの組み合わせ方の異なる 10 種のカーネル関数を作成した。

作成したカーネル関数を評価するため、ヒト及び H. pylori のタンパク質間相互作用データを用い、5 分割の交差検定を行った。その結果、もっとも良い性能のカーネル関数では、予測の評価指標の一つである AUC （area under the ROC curve） の値が、それぞれ、0.88、0.93 であった。また、既存のタンパク質ペアを入力とする予測手法と比較し、予測精度が向上していることを確認した。また、ペア（非順序対）を適切に扱えるように設計したカーネル関数の予測精度がもっとも高いことが確かめられた。

次に、より良いカーネル関数を得るため、タンパク質のドメインを考慮したカーネル関数を設計した。タンパク質は構造的、機能的に比較的独立したドメインから構成される。相同なドメインでは構造や性質は保存していることが期待されるため、相互作用する相手も類似のタンパク質（ドメイン）であると考えられる。また、タンパク質ドメインのデータベースである Pfam を利用し調査したところ、前述のヒトの相互作用データに含まれるタンパク質のうち、85%が少なくとも 1 つのドメインを、43%が複数のドメインを有していた。そこで、アミノ酸配列に対するカーネルとして用いていた spectrum カーネルを、アミノ酸配列に加えてドメインの情報も考慮するものに変更することで、さらなる精度向

上を試みた。

ドメイン情報の利用方法の異なる 3 つのカーネル関数を作成し、ヒトのタンパク質間相互作用データを用い、5 分割の交差検定によってカーネル関数の評価を行った。いずれのカーネル関数でも改善は見られなかった。この原因を明らかにするため、SVM が出力する値の相関を調べた。その結果、もととなったアミノ酸配列のカーネル、3 つのアミノ酸配列とドメイン情報を用いるカーネルのどの予測器の組み合わせにおいても、相関係数は >0.85 と高く、ドメインを考慮に入れても、またドメイン情報の利用の仕方を変えても計算される類似度に大きな変化がないことが分かった。

Isomap を用いた非線形次元削減の自由エネルギー解析への適用

分子動力学シミュレーションやモンテカルロシミュレーションなどで生成された分子構造群から安定状態や準安定状態、遷移状態の解析を行う際に、しばしば、次元削減が行われる。これは、高次元のデータのままで解析が可能なほどの大量の構造を、シミュレーションでサンプリングするのが困難なためであるのと、データの解析を容易にするためである。構造を低次元で表現するために、分子の構造変化が大きい場合は、天然構造における原子間コンタクトの再現率や回転半径などが良く用いられる。また、構造群の分布をうまく表現できる原子間距離や二面角などが既知の場合は、それらが用いられる。しかし、分子の構造変化が複雑で、適当な座標を選択するのが困難なこともある。そのような場合、サンプリングした構造群の分布に基づいて系統的に次元を削減する手法である主成分分析や多次元尺度構成法などが用いられる。これらは、構造間の距離関係をできるだけ保つような直交座標に次元削減を行う。ここで問題となるのは、主成分分析や多次元尺度構成法で保存するのは構造間の直線距離である点である。分子は結合長、結合角、二面角、衝突、静電相互作用などから制約を受け、どのような構造でもとれるわけではないので、サンプリングされた構造の周辺を分子が取り得る構造と仮定し、取り得る構造のみを経る道なりの距離関係をできるだけ保つような次元削減が適当な場合もあると考えられる。この様な非線形の次元削減を行うのが Isomap である。

Isomap は距離の近いデータ間を辺でつないだ重み付きの無向グラフを作成し、データ間の距離をこの無向グラフ上での最短距離で定義し、最後に、得られた距離行列に多次元尺度構成法を適用して次元削減を行う手法である。Isomap のシミュレーションで生成された分子構造への適用は報告されているが、主成分分析や多次元尺度構成法との比較は、どちらが無向グラフ上での最短距離と相関が強いか、という観点からのみ行われている。しかし、自由エネルギー解析では、安定状態や準安定状態の数、その相対配置なども重要である。本研究では、Isomap と多次元尺度構成法を、ペンタアラニンとデカアラニンの 100ns（10 万構造）の定温分子動力学シミュレーションの結果に適用し、両手法から求まる空間での自由エネルギー曲面を比較した。その結果、図 1 に示すように多次元尺度構成法による次元削減では他の安定状態と重なってしまっている準安定状態が、Isomap によ

る次元削減では分離されていることを確認した。

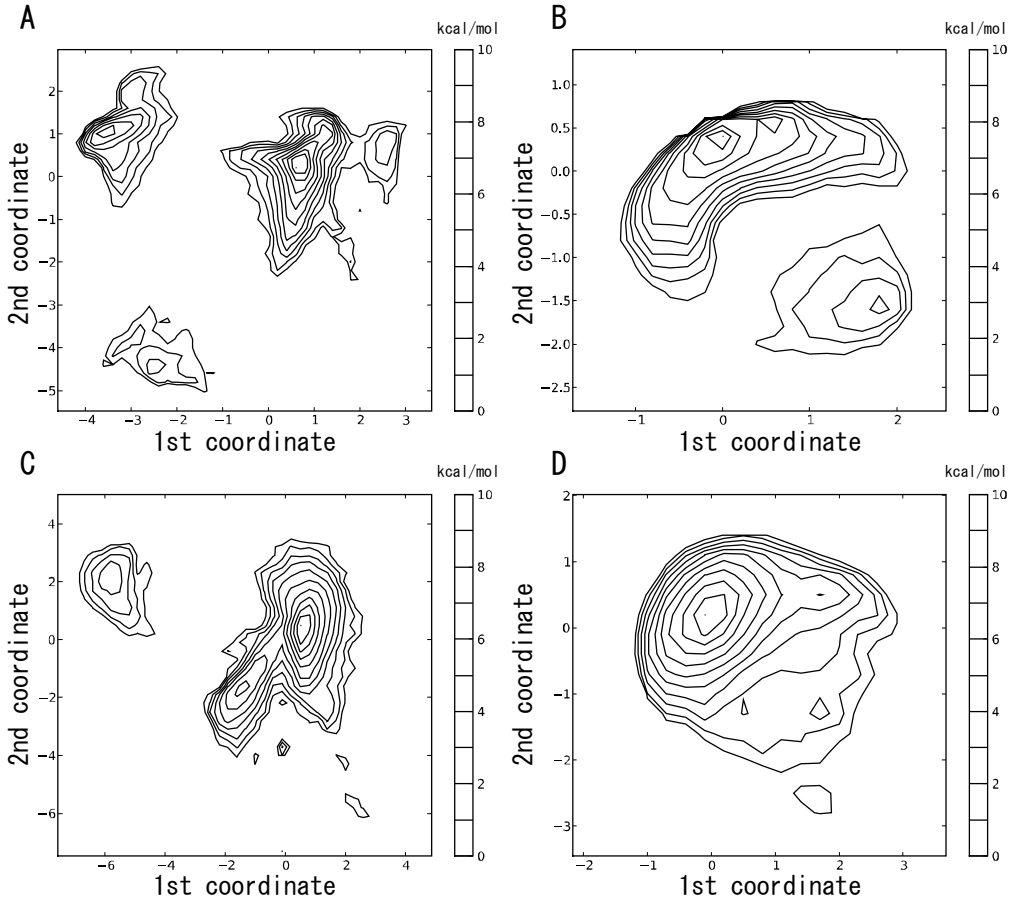

図 1 ペンタアラニンのシミュレーション結果に Isomap (A)と多次元尺度構成法(B)を、デカアラニンのシミュレーション結果に Isomap (C)と多次元尺度構成法(D)を適用して得られた自由エネルギー曲面。

まとめ

カーネル法をタンパク質間相互作用予測および分子シミュレーションの自由エネルギー解析に適用した。相互作用予測では、タンパク質ペアに対するカーネル関数を開発し、これを用いた予測法が、従来法の予測精度を改善することを確認した。また、予測精度の改善は見られなかったが、タンパク質のドメインを考慮に入れたカーネル関数の開発も行った。分子シミュレーションの自由エネルギー解析では、Isomap による非線形次元削減の有用性が確認できた。Isomap には、グラフ構築のために行う全データ間の距離計算に時間を要すると言う問題があるため、より大規模なシミュレーション結果に適用するためには、計算時間を削減する手法の開発が必要であると考えられる。