隠れ変数を用いたタンパク質の分類

三次元的な構造を取り入れた分類が可能に

すべての生命現象の根幹を成すタンパク質については、既に多くのものが知られており、ある程度データベース化されている。例えばSCOPデータベース[1]には、数千個のタンパク質が、階層的に分類されている。このような分類は、人の手によって行われてきたが、これを自動的に行おうというのが、この研究の動機である。

タンパク質は20種類のアミノ酸がつながった文字列として表される。これを分類する時にもっとも単純な方法は、各々のアミノ酸の数を数えて20次元のベクトルとし、これを既存の多変量解析の手法に与えることである（図１上）。この図では、書き表しやすいようにアミノ酸の種類を４つにしてある。多変量解析の方法としては、サポートベクターマシン（図２）などを用いることができる。

しかし、タンパク質の性質は、アミノ酸の頻度だけで決まるわけではない。特に、アミノ酸鎖は、空間上でまっすぐな形をしているのではなく、複雑に曲がり、折りたたまれていて、このような三次元的な構造が性質に大きな影響を与えている。ここで構造情報が「隠れ変数」h の系列として与えられていると仮定する（図１下）。この図ではh=1は、曲がっている部分を示し、h=2はまっすぐの部分を示す。次に、構造情報を考慮して分類を行うため、アミノ酸と隠れ変数を組として数えることを考える（図１下）。こうすることによって、アミノ酸の数は、まっすぐなところと、曲がっているところが区別されて数えられることになり、より高次元のベクトルが得られる。

しかし実は、このような構造情報が得られることは稀であるので、アミノ酸列から統計的に推定しなければならない。統計的な推定法では、h=1 であるか、h=2 であるかということは確定できず、例えば、h=1の確率が6割で、h=2が４割となる。

一方、我々が提案した Marginalized Kernelという方法では、不確定な推定結果を元にベクトルを作る[2]。例えば、アミノ酸がAの所で、h=1 である確率が６割の場合、（A,1）の数には 0.6 を加える。このような考え方に基づいて、実際のデータベースを用いて分類実験を行ったところ、従来の方法に比べて誤りの少ない優れた結果が得られた。

今後はこのような統計的な隠れ変数推定を用いる分類法を、他の対象（DNA など）にも適用していきたい。

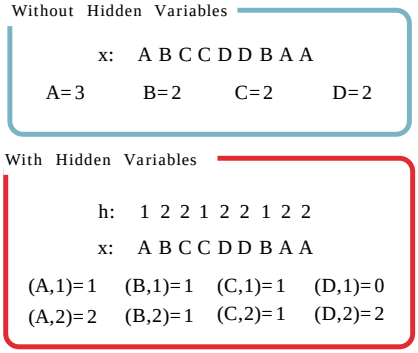

図１　隠れ変数のない場合(上）と、ある場合（下）の特徴抽出

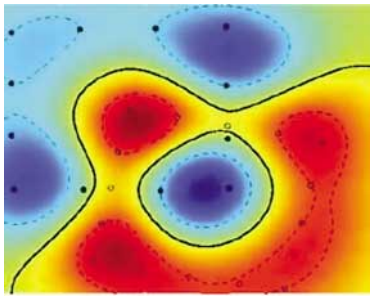

図２　サポートベクターマシンによる空間の分割例

つだ こうじ
津田宏治
koji.tsuda@aist.go.jp
生命情報科学研究センター

関連情報

● 1) http://scop.berkeley.edu
● 2) K. Tsuda, T. Kin and K. Asai: "Marginalized Kernels for Biological Sequences",Bioinformatics, 2002, in press.