3P138

サポートベクターマシン(SVM)による蛋白質水和水の物性解析

(北陸先端大院・知識) ○杉山 歩, 水上 卓, Dam Hieu Chi, Ho Tu Bao

【序】計算機と計算技術の発展は同時に膨大な情報量の蓄積・処理に対する技術的発展をもたらした。この膨大なデータには直感的には想像し難い相関性や莫大なデータ集合体間に発生する新たな知識など有用な知識を内包している。データマイニングとはこの様な大量のデータに埋もれた有用なデータを発見するデータ処理技法の総称で、統計学やパターン認識の技術を駆使し、特にバイオインフォマティクス、金融、経済活動等の分野でその有用性が認められ、急速に発展・普及している。サポートベクターマシン(Support Vector Machine (SVM))はカーネルトリックを利用した非線形解析として注目されているパターン認識手法である。SVM は極めて高速な処理のため大規模問題への適応が事や分類回帰問題の難問である局所解の問題が無いことから化学・薬学の分類問題・予測問題に適用されている。　本研究では上述 SVM を計算機シミュレーションから生成された膨大な情報量を含む蛋白質水和水分子のダイナミクスに適用し、蛋白質水和水の物性解析を試みた。

【計算方法】学習用のデータを$(\mathbf{x}^{(i)}, y_i)$ (i=1,・・l)とする。ここで、$\mathbf{x}^{(i)}\in\mathbf{R}^p$, $y_i\in\{-1, 1\}$である。この時、写像$\Phi : \mathbf{R}^p\rightarrow\mathbf{R}^{p'}$によって入力データの高次元空間への写像を考える。判別関数 $f(\mathbf{x}) = \mathrm{sgn}((\mathbf{w}\cdot\Phi(\mathbf{x}^{(i)}))+b)$ $(\mathbf{w}\in\mathbf{R}^{p'}, b\in\mathbf{R})$とする時、学習データ集合に対するマージンを最大にする超平面$(\mathbf{w}, b)$を求める問題は、

$$\min \quad \frac{1}{2}\|\tilde{w}\|^2 + C\sum_{i=1}^{l}\xi_i \tag{1}$$

$$\text{s.t.} \quad y_i\left(\left(\tilde{w}\cdot\Phi\left(x^{(i)}\right)\right)+b\right) \geq 1-\xi_i \qquad (i=1,\cdots,l)$$

$$\xi_i \geq 0 \qquad (i=1,\cdots,l) \tag{2}$$

と定式化される。この時、式(1),(2)に対する双対問題は、

$$\max \quad \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j\beta_i\beta_j \; K(x^{(i)}, x^{(j)}) \tag{3}$$

$$\text{s.t.} \qquad \sum_{i=1}^{l}\alpha_i y_i = 0 \qquad 0\leq\alpha_i\leq C \quad (i=1,\cdots,l) \tag{4}$$

と書き表せる。 ここで $K(x^{(i)},x^{(j)})$ はカーネル関数と呼ばれる。本研究では線形カーネル、Radial basis function (RBF)カーネル、d 次多項式カーネルを採用し、解析を行った。線形カーネル、Radial basis function (RBF)カーネル、d 次多項式カーネルはそれぞれ、

$$K(x,x^{(j)}) = \exp\left(-\frac{\left\|x - x^{(j)}\right\|^2}{2\sigma^2}\right)$$

で表現される。本研究では LIBSVM[1]を利用し、上述(3),(4)式から超平面からのマージンが最大となる水和水のダイナミクスを bulk water と hydration water への分類問題とみなし、考察を行った。

【計算モデル】本計算は Amber10(amber force field03)を使用し、合成ペプチド(PDBID: 1PSV[2]), CPI-9 蛋白質(PDBID:1J2M[3])の水和システム及び bulk water に対し分子動力学計算を行い SVM 用学習データ並びにターゲットデータを作成した。学習データ並びにターゲットデータは任意に選択した 100 個の水分子の 25ps 間の Root Mean Square Deviation (RMSD)を利用した。

【結果】SVM の学習データは hydration water には 1PSV 周辺に 1172 個配した水分子 100 個から生成した Water Unit (WU)のデータ(計 11WU)を、bulk water は 12071 個の水分子から生成した 120 個の WU のデータを利用した。生成された SVM を利用し 1PSV 水和水及び bulk water を判別した所、水和水と識別された水の割合はそれぞれ 92%, 0.001%となり、良好に識別されていると考えられる。本 SVM を利用し、1J2M 周辺水分子の識別を行った。SVM で識別された水和水には蛋白質間距離の依存性がみられる（図 1）。これは水和水のダイナミクスは蛋白質周辺からより広範囲にわたる事を示唆している。詳細は SVM と Kernel 関数の依存性と共に当日報告する。

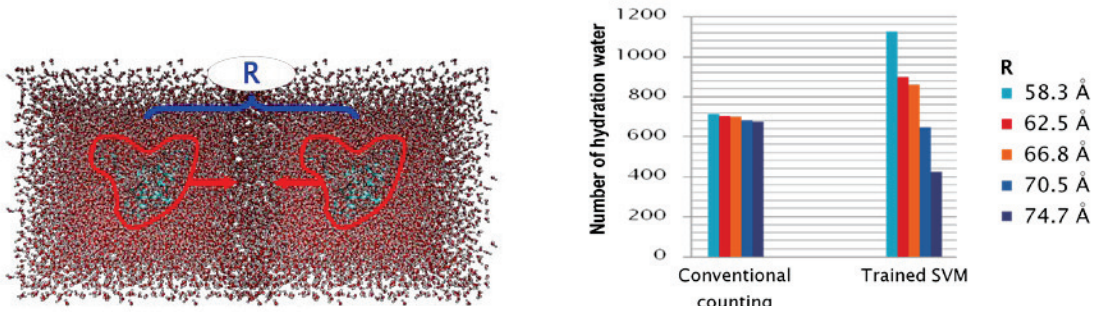

図 1: 2 蛋白間距離と水和水の数の比較

References

[1] C. C. Chang, C. J. Lin, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2] B. I. Dahiyat et. al. J.Mol.Biol. **273** 789-796, (1997)

[3] S. Ohki et. al. J.Mol.Biol. **326:** 1539-1547, (2003)