4P111

線形回帰法を用いたタンパク質原子電荷に関する研究

（東大院工*、東大生研**）〇金泰煥*、平野敏行**、佐藤文俊**

Studies of atomic charges for proteins using the linear regression method

(School of Engineering, the University of Tokyo*,
Institute of Industrial Science, the University of Tokyo**)
〇Kim Taehwan*, Hirano Toshiyuki**, Sato Fumitoshi**

【序】 ESP 電荷[1]は分子周辺の静電ポテンシャル（ESP）を再現するよう、フィッティングした原子電荷である。評価点iにおいて、カノニカル分子軌道（CMO）計算結果から得られた ESP をV_i、原子電荷$\boldsymbol{q}$から算出した ESP を$\hat{V}_i$とする。ESP 電荷は最小二乗法より残差$(V_i - \hat{V}_i)$の 2 乗和であるl_2損失$\boldsymbol{f}_{LS}(\boldsymbol{q})$が最小となる原子電荷である（式（1））。

$$\boldsymbol{f_{LS}}(\boldsymbol{q}) = \sum_{i=1}^{Ngrid} \left(V_i - \hat{V}_i\right)^2 \qquad (1)$$

ESP 電荷の評価点iは全て vdW 半径の外側に分布するため、分子内部は特異な電荷が得られる場合がある。このような最小二乗法の欠点を防ぐために、様々な線形回帰法が開発されている。本研究では、線形回帰法を用いてタンパク質の原子電荷を算出し、特性を調べた。

【理論】 l_2制約付き最小二乗学習法である Ridge 回帰はl_2損失に、l_2ノルムの正則化項$\|\boldsymbol{q}\|^2$を加え、過適合を防いだ回帰法である（式（2））。Ridge 回帰のl_2ノルムの代わりにl_1ノルムの正則化項$\|\boldsymbol{q}\|_1$を用いたl_1制約付き最小二乗学習法（Lasso 回帰法）は多くのパラメータ$\boldsymbol{q}$がスパースと推定される学習法である。パラメータ$\boldsymbol{q}$の次元が大きいとき、効率良く特徴選択ができる。式（2）、（3）のλ_R、λ_Lはそれぞれ Ridge パラメータ、Lasso パラメータとよび、損失項と正則化項のバランスを表す。

$$\boldsymbol{f_R}(\boldsymbol{q}) = \boldsymbol{f_{LS}}(\boldsymbol{q}) + \lambda_{\mathrm{R}}\|\boldsymbol{q}\|^2 = \sum_{i=1}^{Ngrid} \left(V_i - \hat{V}_i\right)^2 + \lambda_{\mathrm{R}} \sum_{A=1}^{Natom} q_A^2 \qquad (2)$$

$$\boldsymbol{f_L}(\boldsymbol{q}) = \boldsymbol{f_{LS}}(\boldsymbol{q}) + \lambda_{\mathrm{L}}\|\boldsymbol{q}\|_1 = \sum_{i=1}^{Ngrid} \left(V_i - \hat{V}_i\right)^2 + \lambda_{\mathrm{L}} \sum_{A=1}^{Natom} |q_A| \qquad (3)$$

Huber 損失$\boldsymbol{\rho_H}(r_i)$を用いたロバスト回帰は、残差$r_i = (V_i - \hat{V}_i)$の絶対値が閾値ηの以下であればl_2損失を、大きければl_1損失を使う Huber 損失$\boldsymbol{\rho_H}(r_i)$の総和が最小とする回帰法である（式（4））。l_1損失を混ぜ合わせたので高いロバスト性を持ち、閾値ηを設定することで、モデルの再現性とロバスト性のバランスを調整することができる。

$$\boldsymbol{f_H}(\boldsymbol{q}) = \sum_{i=1}^{Ngrid} \boldsymbol{\rho_H}(r_i) \qquad \boldsymbol{\rho_H}(r_i) = \begin{cases} \dfrac{r_i^2}{2} & if\ |r_i| \le \eta \\ \eta|r_i| - \dfrac{\eta^2}{2} & if\ |r_i| > \eta \end{cases} \qquad (4)$$

【方法】 本研究はオキシトシン（原子数：134 個、評価点数:5,414 個）とインスリン（原子数：786 個、評価点数：23,787 個）を計算対象にした。CMO 計算および ESP 計算は ProteinDF[2]を使用した。Ridge

回帰、Lasso 回帰、Huber 損失$\boldsymbol{\rho_H}(r_i)$を用いたロバスト回帰から見積もられた電荷を便宜上 Ridge 電荷、Lasso 電荷、Huber 電荷とする。Ridge 電荷は制限関数を調和型、参照電荷を 0 にした RESP 電荷[3]に相当する。全ての原子電荷計算は原子電荷の総和が系全体の電荷Q_{tot}に等しい条件下$\boldsymbol{g}(\boldsymbol{q})$（式(5)）、ラグランジュ未定乗数方程式から原子電荷$\boldsymbol{q}$を求めた。

$$\boldsymbol{g}(\boldsymbol{q}) = \sum_{A=1}^{Natom} q_A - Q_{tot} = 0 \qquad (5)$$

線形回帰法から求められた原子電荷を Mulliken 電荷および Amber 力場電荷 ff03 電荷と比較ならびに解析を行った。

【結果】 インスリンの Mulliken 電荷、ff03、Ridge 電荷($\lambda_R = 5 \times 10^{-4}$)の計算結果を図 1 に示した。ここでは主鎖の窒素(N)と酸素(O)のみ示した。Mulliken 電荷において N と O の原子電荷は同程度であり、分散が小さいのに対し、Ridge 電荷は隣接する原子が作る ESP によって電荷が見積もられるため、多様な原子電荷を取ることができる。51 残基すべての主鎖の N と O の平均値を調べると N が O より正に寄っていることが確認された。このことは、N と O の電気陰性度から考えると妥当な結果である。

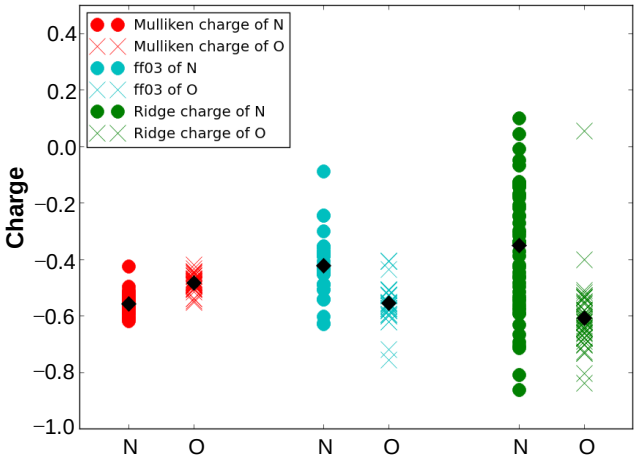

図 1 インスリンの主鎖の窒素と酸素の電荷

（黒いダイヤモンドは電荷の平均値）

(左)Mulliken 電荷, (中)ff03, (右)Ridge 電荷

次に様々な Lasso パラメータλ_Lを使って Lasso 電荷がスパースに推定されたパラメータ数（0 要素数）と ESP 再現性の指標となる RRMS の結果を表 1 に示した。Lasso パラメータ$\lambda_L = 10^{-1}$のとき、全体の 75%に相当する 596 個の原子がスパースに推定されたにも関わらず、ff03 と同程度の ESP 再現性を見せた。即ち、Lasso 電荷は少ない変数（原子電荷）で ff03 と同程度の ESP が再現できるといえる。

Huber 電荷の特徴については当日報告する。

表 1 様々な Lasso パラメータ$\boldsymbol{\lambda_L}$を使って

Lasso 電荷がスパースに推定されたパラメータ数（0 要素数）と ff03 の RRMS

原子電荷	**Lasso 電荷**					**ff03**
	$\lambda_L = 0$	$\lambda_L = 10^{-4}$	$\lambda_L = 10^{-3}$	$\lambda_L = 10^{-2}$	$\lambda_L = 10^{-1}$	
0 要素数	0	41	202	372	596	0
RRMS	0.0284	0.0287	0.0340	0.0520	0.1168	0.1210

$$\mathrm{RRMS} = \sqrt{\frac{\sum_{i=1}^{Ngrid}\left(V_i - \hat{V}_i\right)^2}{Ngrid}} \qquad (6)$$

【参考文献】 [1] U. C. Singh. et. al., *J. Comput. Chem.*, 5 (1984) 129. [2] http://proteindf.github.io/ . [3] C. Bayly. et. al., *J. Phys. Chem.*, 97 (1993) 10269.