Vol.2015-MPS-103 No.24
Vol.2015-BIO-42 No.24
2015/6/23

条件付確率場を用いたタンパク質残基間コンタクト予測

鎌田 真由美[1,a)]　林田 守広[2]

概要：生体内で重要な役割を担うタンパク質の機能や他分子との相互作用は，その立体構造に依存する．タンパク質の立体構造はアミノ酸残基間の相互作用に基づき形成される為，残基間の関係を知ることは立体構造や機能を知る上で重要な情報となる．その重要性から，残基間相互作用に関してこれまで多くの研究がなされているが，アミノ酸配列情報のみからその相互作用を予測することは，未だ難しい問題である．一方，生物学上の進化の過程において，機能的もしくは構造上重要な残基群は対応する残基と共進化し，また，それらの情報は各々のアミノ酸配列に保存されていると考えられている．故に共進化の関係は，相同タンパク質の多重配列アライメントに対する相互情報量などの統計的指標を用いて推測することが出来，これまでに活性部位やタンパク質間相互作用部位の予測などに用いられている．本研究では，残基の共進化のアイデアに従い，条件付き確率場モデルに基づくタンパク質の構造内残基間コンタクト予測手法を提案する．提案手法では，残基ペアのコンタクト予測に周辺残基の状態が有効であるとし，周辺残基の情報を考慮するモデルを考える．提案モデルの予測精度と隣接残基情報の有用性を検証する為，既存手法との比較を行い，共進化関係の推定法についても考察を行う．

1. はじめに

タンパク質は構成アミノ酸残基の相互関係に基づき，エネルギー的に安定な構造へと折りたたまれ，その機能を発現する．これまでに多くのタンパク質で立体構造解明が実験的に行われているものの，未だほとんどのタンパク質が構造未知である．故に，一次構造であるアミノ酸配列から立体構造を予測することが出来れば，新規タンパク質の機能を理解する上で大きなヒントとなる．その為には，配列情報から構造内残基間の相互作用 (コンタクト) を予測することが必要となる．タンパク質において構造の形成や安定性に重要な残基ペアは，その関係を保つ為，進化の過程において常に共進化しており，それらの関係性は配列レベルで保存されていると考えられている．この保存された関係性は，相同配列の多重配列アライメント (MSA) に対する統計的指標の適用によって推測できるとされており，相互情報量 (Mutual Information, MI) を用いた手法の幾つかで，その有用性が示されている [1]．配列情報のみを用いた残基間コンタクト予測手法としては，support vector machine や deep neural network などの機械学習を用いるものが数多く提案されている [2], [3]．しかしながらその予測精度は低く，未だ多くの課題が残る難しい問題である．

アミノ酸残基間の総当たりの関係を 2 次元で表現すると，一枚の画像として捉えることが出来る．そこで本研究では，画像処理や自然言語処理の分野で広く用いられている，条件付き確率場 (Conditional Random Filed, CRF) に基づく残基間コンタクト予測モデルを提案する．更に，コンタクト予測には，隣接残基との相互関係の情報が有用であると考えられることから，この情報を取り入れるモデルを考える．

[1] 京都大学大学院 医学研究科 臨床システム腫瘍学
[2] 京都大学 化学研究所 バイオインフォマティクスセンター
[a)] mkamada@kuhp.kyoto-u.ac.jp

2. アミノ酸残基間の相互情報量

タンパク質のアミノ酸配列 A とその相同配列を集め, MSA を計算する．ここで, $\mathcal{A}$ を 20 の必須アミノ酸と未決定のアミノ酸を表す文字の集合とし, $p_i(a)$ を配列上のポジション i におけるアミノ酸 $a \in \mathcal{A}$ の頻度, $p_{ij}(a,b)$ をポジション i と j におけるアミノ酸 $a,b \in \mathcal{A}$ の同時出現頻度とする．この時，ポジション (i,j) 間の MI は $MI(i,j) = H_i + H_j - H_{ij}$ で表される．ここで, H_i と H_j は各々ポジション i と j の周辺エントロピーであり，$H_i = -\sum_{a\in\mathcal{A}} p_i(a)\log p_i(a)$ として定義される．$H_{ij} = -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{A}} p_{ij}(a,b)\log p_{ij}(a,b)$ は結合エントロピーである．大きな $MI(i,j)$ の値は，ポジション間の相互依存が大きいことを表している．

上記で計算される MI には MSA で生じる系統発生や分子進化におけるノイズが含まれる可能性がある．その為，それらを MI から取り除く手法がこれまでに幾つか提案されている．本研究では，それらの手法の一つである MIp[4] も入力として用い，MI との比較を行う．

Vol.2015-MPS-103 No.24
Vol.2015-BIO-42 No.24
2015/6/23

3. 提案手法: 2D-CRF モデル

頂点の集合 V と辺の集合 E から成るグラフ $G(V,E)$ を考え，各頂点 s が確率変数 x_s と観測変数 y_s と関係を持つとする．グラフ G に従う確率変数 x_s が条件 y_s の元でマルコフ性を持つ場合，つまり，$P(x_s|\boldsymbol{x}_{\{t\in V|t\neq s\}},\boldsymbol{y}) = P(x_s|\boldsymbol{x}_{\mathcal{N}_s},\boldsymbol{y})$ ($\mathcal{N}_s$ は G において s に隣接する頂点の集合) を満たす時，$(\boldsymbol{x},\boldsymbol{y})$ は CRF である．CRF は，全ての $\boldsymbol{x}$ に対して $P(\boldsymbol{x}|\boldsymbol{y}) > 0$ であり，以下の式で表される．

$$P(x_s|\boldsymbol{x}_{\mathcal{N}_s},\boldsymbol{y}) = \frac{1}{Z_s}\exp\{-U_s(\boldsymbol{x},\boldsymbol{y})\} \tag{1}$$

ここで，$U_s(\boldsymbol{x},\boldsymbol{y})$ は頂点 s に関するポテンシャル関数，Z_s は規格化定数である．本研究では，ポテンシャル関数を以下の様に定義する．

$$U_{ij}(\boldsymbol{r},\boldsymbol{m}) = \boldsymbol{w}^T\boldsymbol{f}_{ij}(\boldsymbol{r},\boldsymbol{m}) + \boldsymbol{v}^T\sum_{(k,l)\in\mathcal{N}_{ij}}\boldsymbol{g}_{ijkl}(\boldsymbol{r},\boldsymbol{m}) \tag{2}$$

ここで，$r_{ij}\in\{0,1\}$ はポジション i と j 番目の残基がコンタクトするかどうかを表す確率変数であり，m_{ij} は残基ペア (i,j) の MI, MIp の値である．残基ペア (i,j) の隣接残基ペアの集合 $\mathcal{N}_{ij}$ は $\{(i,j-1),(i,j+1),(i-1,j),(i+1,j)\}$ と定義する．また，式 (2) におけるベクトル値関数 $\boldsymbol{f}_{ij}$, $\boldsymbol{g}_{ijkl}$ を以下で定義する．

$$\boldsymbol{f}_{ij} = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ m_{ij} \end{pmatrix}, \tag{3}$$

$$\boldsymbol{g}_{ijkl} = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \bar{r}_{kl} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ m_{kl} \end{pmatrix} \tag{4}$$

ここで $\bar{r}$ は r の否定 (i.e. $\bar{r} = 1-r$)，$\otimes$ はクロネッカー積である．式 (2) の右辺第 2 項で隣接残基の相互作用を考慮している．そこで，隣接残基ペア情報の有用性確認の為，ポテンシャル関数 U' を下記のように定義する．

$$U'_{ij}(\boldsymbol{r},\boldsymbol{m}) = \boldsymbol{w}^T\boldsymbol{f}_{ij}(\boldsymbol{r},\boldsymbol{m}) \tag{5}$$

2D-CRF モデルの持つパラメータ $\theta = (\boldsymbol{w},\boldsymbol{v})$ を擬似尤度最大化によって推定する．N 個のタンパク質とその配列，そして相互作用する残基 $\boldsymbol{r}^{(n)}(n=1,\ldots,N)$ が与えられたとし，相互情報量 $\boldsymbol{m}^{(n)}$ を各々のタンパク質に対して計算する．この時，擬似尤度関数は以下のように与えられる．

$$L(\theta) = \log\prod_{n=1}^{N}\prod_{i}\prod_{j}P(r_{ij}^{(n)}|\boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)},\boldsymbol{m}^{(n)},\theta) \tag{6}$$

本研究では，Broyden-Fletcher-Goldfarb-Shanno (BFGS) 法によって $L(\theta)$ の最大化を行う．

次に，推定したパラメーターが与えられた時，条件付き確率が最大になるように $\boldsymbol{r}$ を求め，コンタクトの予測を行う．推定にはエネルギー最小化手法の一つである，逐次的再重み付けツリーによるメッセージ伝搬法 (TRW-S)[5] を用いる．

表 1 提案手法と PSICOV の予測精度比較 (AUC)

		U_{ij}		U'_{ij}	
	PSICOV	+MI	+MIp	+MI	+MIp
Local	0.768	0.519	0.671	0.614	0.679
Non-local	0.719	0.477	0.528	0.581	0.541

4. 結果と考察

本稿では，提案手法と既存手法 PSICOV[6] との精度比較を行った．比較には，PSICOV の論文で提供されているデータセットを用いて 10-fold 交差検定を行い，AUC(Area Under ROC Curve) の平均値を計算した．PSICOV データセットには 150 のタンパク質が含まれているが，構造データと整合性のあった 149 のタンパク質を対象とした．構造内コンタクトの定義は，$C\beta$(Glycines は $C\alpha$) 原子間の距離が 8Å 以下の 2 残基をコンタクトペアとする，the Critical Assessment of Techniques for Protein Structure Prediction (CASP) での定義を用いた．また，アミノ酸配列上の距離が問題の難度に関わることから，配列上 6〜24 残基内の距離にある残基間のコンタクトを Local contact，24 残基以上離れている場合を Non-local contact とした．予測結果の精度を表 1 に示す．PSICOV の精度を上回ることは出来なかったが，提案手法はシンプルなモデルながらも改善の余地のある結果を示した．また，隣接残基ペア情報を取り入れたモデルよりも，考慮しないモデル (U'_{ij}) の方が精度が高かった．これは MI と MIp の結果の比較からも，入力にノイズが多い場合に周辺のノイズも加えて影響を受ける為だと考えられる．今後，入力としている残基の共進化指標の改良や，アミノ酸の物理化学特性に関する特徴量の導入等によって，精度向上を目指したい．

参考文献

[1] Marks, D. S., Hopf, T. A. and Sander, C.: Protein structure prediction from sequence variation, *Nature Biotechnology*, No. 11, pp. 1072–1080 (2012).
[2] Cheng, J. and Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set, *BMC Bioinformatics*, Vol. 8, p. 113 (2007).
[3] Di Lena, P., Nagata, K. and Baldi, P.: Deep architectures for protein contact map prediction, *Bioinformatics*, Vol. 28, No. 19, pp. 2449–2457 (2012).
[4] Dunn, S. D., Wahl, L. M. and Gloor, G. B.: Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, *Bioinformatics*, Vol. 24, No. 3, pp. 333–340 (2008).
[5] Kolmogorov, V.: Convergent Tree-Reweighted Message Passing for Energy Minimization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, pp. 1568–1583 (2006).
[6] Jones, D. T., Buchan, D. W. A., Cozzetto, D. and Massimiliano, P.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequenxe algnments, *Bioinformatics*, Vol. 28, No. 2, pp. 184–190 (2012).