条件付き確率場を用いたタンパク質 RNA 間残基塩基相互作用予測

Predicting protein-RNA residue-base contacts using two-dimensional conditional random field

化学研究所バイオインフォマティクスセンター 数理生物情報 林田守広

背景と目的

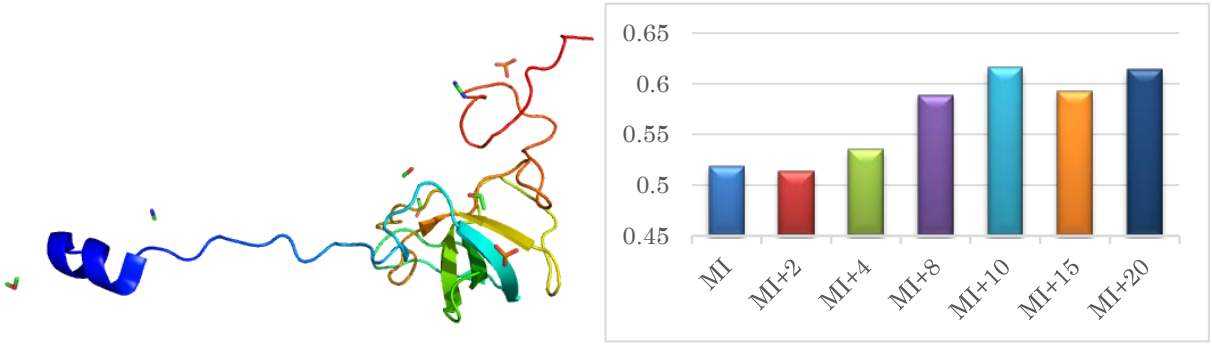
細胞内分子のネットワークと機能を解明する上で、スプライシングや転写後調節などの制御メカニズムに関わる、タンパク質と RNA の間の相互作用を解析することは重要である。タンパク質と RNA から形成される複合体の三次元構造について、どのようにしてタンパク質は選択的に特定の RNA あるいはその塩基と相互作用するのかという観点から多くの研究がなされてきた。例えば U1A タンパク質はヘアピンループまたは内部ループ構造にある塩基配列 AUUGCAC を認識して結合することが知られている[2]。タンパク質残基間相互作用予測に対するこれまでの研究では、相互作用する残基間には共進化が観測される、つまり互いに相互作用する残基の片方が突然変異により他のアミノ酸に置換されればもう片方も相互作用を維持するために置換される、という考えに基づき二残基間の依存関係を表す量として相互情報量を用いた。また縦横に相互情報量を対応する残基の位置に配置することで二次元画像とみなせ、画像に対して開発された識別手法である識別確率場[3]を用いて残基間相互作用をモデル化した。本研究では、タンパク質 RNA 間においても同様に残基塩基間の共進化が観測されると考えられるので、残基塩基間の相互情報量を用いた予測手法を開発する。

検討内容

これまでの研究で相互情報量に識別確率場を用いた場合に、残基間相互作用予測に対してであるが、必ずしも良好な結果は得られなかった。原因として Kumar らの識別確率場は自然画像の特徴である、隣り合う画素が似た値をもつことを仮定していることが考えられる。そこで次のような条件付き確率場モデルを導入する。タンパク質のi番目の残基位置と RNA のj番目の塩基位置との間の確率変数をr_{ij}とし、$r_{ij} = 1$のとき相互作用する、$r_{ij} = 0$のとき相互作用しないことを表し、また(i, j)と隣接する頂点の集合を$N_{ij} = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$とする。このとき条件付き確率を$P\left(r_{ij} \middle| r_{N_{ij}}\right) = \frac{1}{Z}\exp\{w_f f_{ij}(r) + w_g \sum_{(k,l)\in N_{ij}} g_{ijkl}(r)\}$によって与える[1]。ここで$w_f, w_g$はパラメータベクトルで、$f_{ij}, g_{ijkl}$は確率変数と相互情報量、残基または塩基位置でのラベルを入力とし実数ベクトルを返す関数である。

結果

予測手法を検証するため、タンパク質 RNA 複合体の立体構造を PDB コード 1yl4, 2hgu, 3kcr の中から 7 つのタンパク質 RNA ペアを選択し、最近接原子間距離が 3Å以内の残基塩基ペアを相互作用するとした。下図左は 1yl4 のタンパク質 RS12_THET8 とこのタンパク質から 3Å以内の RNA M26923 の原子を表す。各位置間での相互情報量を得るためにタンパク質、RNA それぞれについて Pfam、Rfam データベースの多重配列アラインメントを使用した。



またアミノ酸を物理化学的な性質に基づいて、2, 4, 8, 10, 15, 20 のグループに分類し交差検定による計算機実験を行った。上図右は各グループ分けにおけるテストデータに対する平均の AUC を表す。この結果からアミノ酸の細かいグループ分けの方が予測精度が高い傾向が見られた。

考察

タンパク質残基 RNA 塩基間相互作用を、多重配列アラインメントから得られる相互情報量と、アミノ酸と塩基のラベルから予測する条件付き確率場モデルを提案した。交差検定による計算機実験の結果はラベルを使わない場合または粗いグループ分けの場合よりも細かいグループ分けを用いた場合の方が予測精度が高くなる傾向を示した。今後の課題として精度のさらなる向上が望まれるが、例えば確率場のパラメータ推定において正則化項を加えた学習を行うなどの方法が考えられる。

参考論文

[1] M. Hayashida, M. Kamada, J. Song, T. Akutsu, Predicting protein-RNA residue-base contacts using two-dimensional conditional random field, *Proc of 2012 IEEE Conference on Systems Biology*, pp.152-157, 2012

[2] D. Scherly et al., Identification of the RNA binding segment of human U1A protein and definition of its binding site on U1 snRNA, *EMBO J*, vol. 8, pp.4163-4170, 1989.

[3] S. Kumar, M. Hebert, Discriminative random fields, *Int. J. Comput. Vision*, vol. 68, pp.179-201, 2006.