分配関数を用いたタンパク質配列の比較法

小池亮太郎（横浜市立大学大学院国際総合科学研究科）

分子生物学において、進化的類縁関係は基盤となる情報の 1 つである。進化的類縁関係の検出において配列を比較しその類似性を明らかにする方法は重要な役割を果たす[1-3]。これらの方法はアライメントを一意的に決定する。アライメントとは 2 つの配列を類似の残基が対応付けられるように並べたものである。配列の類似性はそのアライメントから求められ、その類似性で進化的類縁関係を判定する。しかし、このようにして得られるアライメントは評価関数や最適化法に左右される。この問題は配列の類似性が低くなるほど顕著となり、出力されるアライメントと等価なアライメントが多数存在することが知られている[4]。このようなアライメントを準最適アライメントという。この問題を軽減するための方法として、準最適アライメントを検出する方法が開発されてきた[5-12]。ここで、我々はこの問題に対する別の解決法として、全てのアライメントを確率的に記述する方法を提案する。

確率的アライメントは標準的な配列比較法である動的計画法を有限温度に拡張したものと捉えられる[13-16]。この方法では、全てのアライメントは温度 T での分配関数として扱われる。分配関数は次の式で与えられる。

$$Z=\sum_{\alpha}\exp\left[-E(\alpha)/T\right] \qquad (1)$$

ここで、状態αは 2 つのタンパク質を比較したときのアライメントを、ポテンシャル $E(\alpha)$ はアライメントのスコアを表す。アライメントのスコアはたいてい次の式で与えられる。

$$E(\alpha)=\sum_{i=1}^{N}\sigma\left(s_i,t_i\right)+\gamma\left(s_{i-1},t_{i-1},s_i,t_i\right) \qquad (2)$$

1 体のポテンシャルσは 2 つの残基 s_i と t_i の配列の類似度を、隣接間相互作用γは挿入や欠損を反映するギャップペナルティーを表す。このようにアライメントを統計力学的枠組みで表現することで、比較の対象となる 2 つのタンパク質の各残基 p_j と q_k, の類似度を、従来の方法が用いてきたスコア行列 $\mathbf{M}_\sigma=\left\{\sigma\left(p_j,q_k\right)\right\}$ ではなく、新たに導出される確率行列 $\mathbf{M}_\mathrm{P}=\left\{P\left(p_j,q_k\right)\right\}$ で評価する。2 つの残基 p_j と q_k, を対応させたときの確率 $P\left(p_j,q_k\right)$ は分配関数を用いて下記の式で表される。

$$P(p_j, q_k) = \frac{1}{Z} \sum_{\alpha}^{(p_j, q_k)} \exp\left[-E(\alpha)/T\right] \qquad (3)$$

ここで、p_j と q_k を対応付ける全てのアライメントαに関しての和をとる。$T \to 0$ の極限では、確率行列$\mathbf{M}_{\mathrm{P}}$はスコア行列$\mathbf{M}_{\sigma}$から導かれるアライメントと同じアライメントを与える。すなわち、$T \to 0$での0以外の確率をもつアライメントは動的計画法によるアライメントと一致する。

しかし、有限温度への拡張はある問題を引き起こす。最も単純な例として、各残基間の類似度が一定、すなわち、$\sigma(p_j, q_k) = \text{const.}$の場合を考える。このとき、各残基間の対応確率も同様に一定となることが期待される。しかし、実際には確率行列上の対角線上の対応、すなわち、ギャップの少ない、多くの残基を対応付けるアライメントに由来する対応が過剰に高い確率をもつ(Fig. 1)。

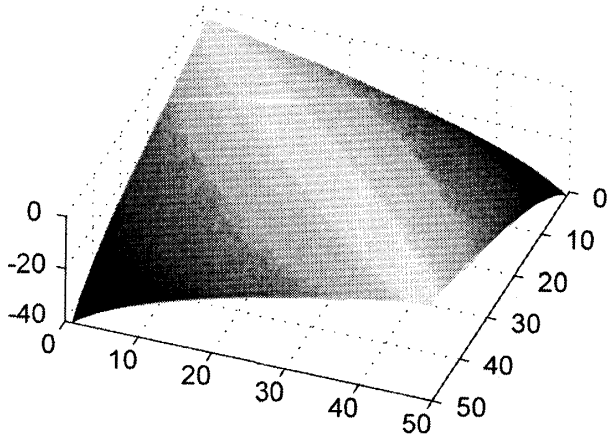

Figure 1　$\sigma(p_j, q_k) = \text{const.}$における確率行列$\mathbf{M}_{\mathrm{P}}$

確率行列の各要素は log スケールで表記。

この対応確率のバイアスは有限温度に拡張したことで生じるエントロピーに起因する。我々はこの問題を解決するためにアライメント構築の際に周期境界条件を導入した。これにより、エントロピーのバイアスを解消することができた。

この精錬した配列比較における確率的アライメント法を用い進化的類縁関係の検出を行った。ここでは、その検出能を評価するために進化的類縁関係の検索に広く使われる PSI-BLAST と比較する。PSI-BLAST はプロファイル法と呼ばれる方法を採用しており、配列間の類似性を検出する際に、着目している 2 つの配列の中間配列をデータベース中からとりだし、その情報を利用する。その結果、確率的アライメント法は PSI-BLAST と同等の高い検出能を示

した(Fig. 2)。

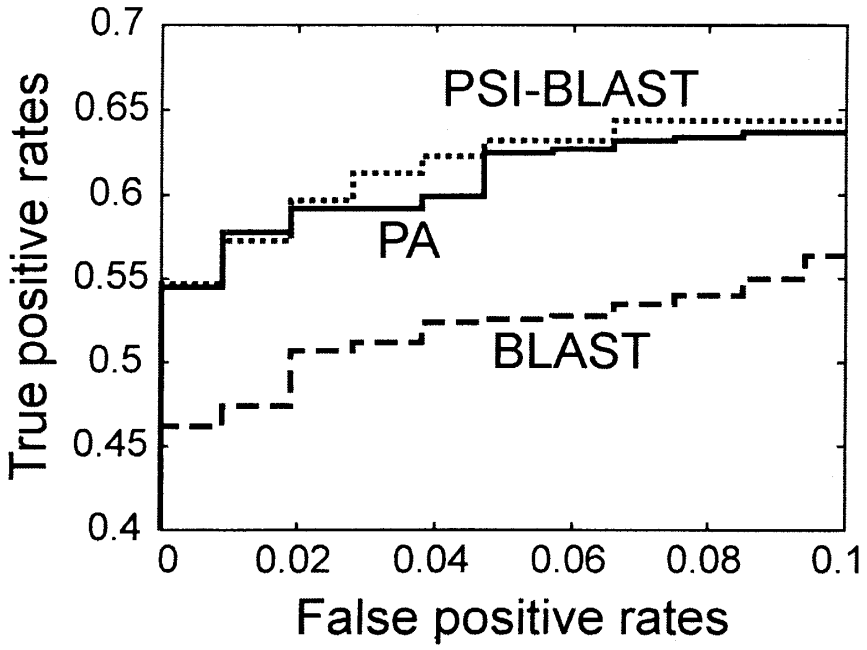

Figure 2　ROC を用いた確率的アライメントの検出能評価

実線は確率的アライメント（PA)、点線は PSI-BLAST、破線は BLAST の ROC カーブを表す。ここでは正解例（True）としては SCOP データベースで同じスーパーファミリーに分類されるペア、不正解例（False）としては同じフォールドに分類されるが異なるスーパーファミリーに分類されるペアを採用している。

この結果は確率的アライメント法が PSI-BLAST のように中間配列の情報を利用していないことを考えると、驚くべき結果である。この確率的アライメント法の検出能の向上は有限温度への拡張によるエントロピーの効果であると考えられる。

謝辞

この研究は横浜市立大学大学院国際総合科学研究科の木寺詔紀教授と東京大学医科学研究所の木下賢吾助教授との共同研究です。ここに深謝致します。

参考文献

1. Waterman MS. “Introduction to Computational Biology: Maps, Sequences and Genomes”, Chapman & Hall, 1995.

2. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-3402.
3. Durbin R, Eddy S, Krogh A, Mitchison G. "Biological sequence analysis: probabilistic models of proteins and nucleic acids", Cambridge University Press, 1998.
4. Vingron M. Near-optimal sequence alignment. Curr Opin Struct Biol. 1996;6:346-352.
5. Warerman MS. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. Proc Natl Acad Sci. USA 1983;80:3123-3124.
6. Waterman MS, Eggert M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. J Mol Biol. 1987;197:723-728.
7. Vingron M, Argos P. Determination of reliable regions in protein sequence alignments. Protein Eng 1990;3:565-569.
8. Saqi MA, Sternberg MJ. A simple method to generate non-trivial alternate alignments of protein sequences. J Mol Biol 1991 20;219:727-732.
9. Zuker M. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. J Mol Biol. 1991;221:403-420.
10. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325-1338.
11. Feng ZK, Sippl MJ. Optimum superimposition of protein structures: ambiguities and implications. Folding Design 1996;1:123–132.
12. Lackner P, Koppensteiner WA, Sippl MJ, Domingues FS. ProSup: a refined tool for protein structure alignment. Protein Eng 2000;13:745-752.
13. Miyazawa S. A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng 1994;8;999-1009.
14. Zhang MQ, Marr TG. Alignment of molecular sequences seen as random path analysis. J Theor Biol. 1995 21;174:119-129.
15. Kschischo M, Lassig M. Finite-temperature sequence alignment. Pac Symp Biocomput. 2000:624-635.
16. Muckstein U, Hofacker IL, Stadler PF. Stochastic pairwise alignments. Bioinformatics. 2002;18 Suppl2:S153-160.