統計的アラインメント作成法 statistical methods for alignments

大域的アラインメント (global alignment) 作成の統計的手法として，**最尤アラインメント (maximum likelihood alignment)** [1,2] と**確率的アラインメント (probabilistic alignment)** [3]) について述べる．タンパク質/DNA の配列対配列，もしくはタンパク質の構造対配列，構造対構造など，長さ m, n からなる **a** と **b** の**ペアワイズアラインメント (pairwise alignment)** において，特有のアラインメント A_ν は**統計的重み (statistical weight)** $\exp S(A_\nu)$ で確からしさを有するものとする．$S(A_\nu)$ は A_ν のアラインメントスコアである．$S(A) = \max_\nu S(A_\nu)$ なる最大のスコアをもったアラインメント A が**最良アラインメント (maximum score alignment)** である．特定のアラインメントの確からしさ，すなわちアラインメントにおける置換/挿入/欠失の確からしさは，配列が分子進化の産物である以上，その遺伝子の進化過程に基づいて評価されねばならない．また，塩基/アミノ酸の置換率，挿入/欠失頻度は比較する配列対に固有である．座位依存の塩基/アミノ酸置換スコアや，比較する配列対に適合する置換率に対応する置換スコアは類似配列のデータベース検索手法においても使用されている．

一方，分子進化学からのアプローチの一つでは，通常のアラインメントスコアは尤度比の対数 (log-odds) として定義されるが，尤度 ($l(A_\nu) \equiv \exp S(A_\nu)$) そのものの対数をスコアとする．このアプローチでは，塩基配列進化の確率モデルに基づき可能な全アラインメントの尤度 $L(\equiv \sum_\nu l(A_\nu))$，すなわち祖先配列が分岐し配列 **a** と **b** にいたる確率を最大化するように塩基置換率，一塩基の挿入/欠失頻度などのモデルパラメータを推定する最尤アラインメント法が提案された [1]．つまり，進化モデルを仮定し最尤法の意味で最適化されたモデルパラメータを用いた最良アラインメントを最尤アラインメントという．この方法は，多塩基の挿入/欠失，置換率が異なる 2 種類の配列領域を許す，より精緻な進化モデルへと改良された [2]．この方法はアラインメントと**系統樹 (phylogenetic tree)** の**最尤法 (maximum likelihood method)** による同時最適化を志向するものであろう．ちなみに**最大節約法 (maximum parsimony method)** による同時最適化はすでに試みられている．

ともかくも，統計的重み $\exp S(A_\nu)$ は適切に計算されたとしよう．最良アラインメントは，統計的重み $\exp S(A_\nu)$ をもつアラインメント A_ν 全体からなる**統計集団 (statistical ensemble)** における最も確からしいアラインメントである．しかし遠縁の配列比較においては，多数のアラインメントが類似のスコア値をもつため最良アラインメントだけを考慮していては不十分である．たとえば，最良アラインメントにおいて配列 **a** の i 番目の座位 a_i と配列 **b** の j 番目の座位 b_j がマッチ $(a_i : b_j)$ するとしても，座位 a_i にとって b_j とのマッチが確率最大であるとは限らない．可能なアラインメント集団における A_ν の任意の関数 $B(A_\nu)$ の平均値 $< B >$ は，統計的重みの総和，統計物理学でいうところの**状態和 (sum over states)** あるいは**分配関数 (partition function)** $Z(\equiv \sum_\nu \exp S(A_\nu))$ を用いて，$< B >= \sum_\nu B(A_\nu) \exp S(A_\nu)/Z$ として計算できる．統計的重みが尤度である場合は $Z = L$ である．たとえば B として，A_ν において $a_i : b_j$ のとき 1，それ以外では 0 をとる関数を選べば，$< B >$ は全アラインメント集団における $a_i : b_j$ の確率を与える．最も確からしいマッチに基づくアラインメント (確率的アラインメント [3]) を考えてみよう．たとえば，座位対/挿入/欠失確率 > 0.5 なる座位対/挿入/欠失を選ぶことにより各座位ごとの最大対確率に基づくアラインメントを作成できる．座位対のみに興味があるとき，また確からしい座位対に興味がある場合に有用であろう．

参考文献

1) Bishop,M. J. & Thompson, E. A. (1986) *J. Mol. Biol.*, **190**, 159
2) Thorne, J. L. et al. (1992) *J. Mol. Evol.*, **34**, 3
3) Miyazawa, S. (1995) *Protein Engineering*, **8**, 999

⇒統計：統計的推測
⇒情報：文字列検索アルゴリズム
⇒物理：統計分布
⇒配列：ペアワイズアラインメント
⇒配列：ベイジアンアラインメント
⇒ゲ遺：系統樹の種類と作成

⇒立体：構造アラインメント