日本オペレーションズ・リサーチ学会
2004年秋季研究発表会

分子系統樹構築アルゴリズムへの メタヒューリスティックアルゴリズムの応用を考える

中村政隆 (No. 01402860)
東京大学総合文化研究科広域システム科学系
153-8902 目黒区駒場３－８－１　nakamura@klee.c.u-tokyo.ac.jp

複数の種 (taxon) の遺伝子の塩基列もしくはアミノ酸列をデータとして、そこから系統樹 (phylogenetic tree) を構築するアルゴリズムは、距離行列法と、与えられたデータのもとでの各二分木のコストを定義してコスト最小の二分木を求める、という２つのタイプに大別できる。

距離行列法 (Pairwise Distance Method) としては、UPGMA / WPGMA 法 (実用にはあまり用いられてない)、及び近隣結合法 (Neighbour Joining Method, N J 法) などが知られている。

解空間の中で局所探索を繰り返すタイプの組み合わせ最適化のアルゴリズムは、一般に次の要素からなる。

A. 解空間と各解に対するコスト関数が与えられているとする。

B. 解の近傍を定義する。

C. 初期値を定める。

D. 探索ルールと打ち切り条件を定める。

生物学の文献を読むと、分子系統樹の構築アルゴリズムについて、上の局所探索アルゴリズムの各要素の取り方として、次のようなものが提案されている。

A. コスト関数 f(T) の定義 (正しくはデータを x として $f(x;T)$) : 現在提案されている或いは利用されているもの

(i) 最小二乗法 (Least Squares Method)— 辺の長さも自動的に求まる

(ii) 最小進化法 (Minimum Evolution Method) ME 法— 木 T に対してその辺の長さの推定値 b_i の総和 $f(T) = \sum_i b_i$ を最小にする木Tを求める。

(iii) 最大節約法 (Parsimony) — 辺の長さは別途求める

(iv) 最尤法 (Maximum Likelihood Method) —尤度は、木 T とその辺の長さ r_T の関数 $F(T, r_T)$ であり、その値の最小化から最適木と木の辺の長さが決まる。

$$\min_T\{f(T)\} = \min_T\{min\{F(T, r_T) : r_T\}\}$$

実際の生物学の研究者は、最大節約法か最尤法を多く用いるようである。

B. 探索のための解の近傍の定め方としては

(i) Nearest Neighbor Interchange (NII) – 辺の数で距離が３の葉の対のラベルを交換する。

(ii) Subtree Pruning and Regrafting (SPR) – 任意の辺 $e = xy$ をとる。y が次数が 3 の点であるとする。T から辺 $e = xy$ を削除し、点 y を除いてその両端点をもとのように結ぶ辺を加える。そして、y を含む成分のうちの任意の辺をとってその途中に新しい点 z を加えて、辺 xz を最後に加える。という一連の操作のこと。

(iii) Tree Bisection and Reconstruction (TBR) – 任意にひとつの内辺をひとつ削除し、得られた 2 つの連結成分でそれぞれで任意に選んだ辺に新しい点 u, u' を挿入して、最後に辺 uu' を加える。

などが提案されている。

C. 局所探索を利用するときの初期値の取り方：つまり初期解の二分木を与える方法として、NJ 法の結果を初期値にとる、あるいは Stepwise Addition (SA) 法 (greedy algorithm) で構成したものを初期値とする方法が取られている。

D. 探索ルールと停止条件については、 既存の生物学の文献の中にはまったく記述が見あたらないように見える。実際に生物学者が利用しているソフトでは、データの数が小さければ全数数え上げによって解を求め、データ数が増えて解空間が非常に大きくなる場合は、初期解の近傍にあるものをなるべくたくさん数え上げて調べるという方法で解の探索が行われているようである。

組み合わせ最適化の分野でメタヒューリスティックアルゴリズムの名の下に知られている

(1) 遺伝的アルゴリズム (Genetic Algorithm)

(2) タブー探索 (Tabu Search)

(3) 進化的アルゴリズム (Evolutionary Algorithm)

などの戦略を、系統樹アルゴリズムの樹形探索の部分に活用して実際に効果的であるかどうかを、これからの研究課題として考えてみたい。

References

[1] P. Clote and R. Backofen: Computational Molecular Biology, Wiley, 2000. (数式が多用されているが、見かけほど「数学的」な部分はむずかしくない。)

[2] R. Durbin, S. Eddy, A. Krogh and G. Mitchison: Biological Sequence Analysis, Cambridge Univ. Press, 1998. (生物学的な内容と数学的取り扱いのバランスがとれている。数学的な部分の説明が具体的で分かりやすい。例えば NJ 法がうまくいくことの証明も載っている。)

[3] M. Nei and S. Kumar: Molecular Evolution and Phylogenetics, Oxford Univ. Press, 2000. (分子進化の標準的な教科書のひとつとして良く知られている。アルゴリズムの正確な記述よりも各種アルゴリズムの比較に論点が置かれている。)

[4] R.D.M. Page and E.C. Holmes: Molecular Evolution–A Phylogenetic Approach, Blackwell, 1998. (明晰で品の良いテキストである。ネットワークの話しもきちんと視野に入れている。)

[5] C. Sempel and M. Steel: Phylogenetics, Oxford University Press, 2003. (形式化のためだけの数学的定式化による本。数学的形式化の悪い見本例のように思える。)

[6] M.S. Waterman, Intoroduction to Computational Biology, Chapman & Hall, 1995. (Reprint is published by CRC Press 2000.) (数学的記述が明晰でかつわかりやすい形で書かれている良いテキスト。)