Reliable Evolutionary Tree Reconstruction: The Generalized Neighbor Joining Method

William Pearson
Gabriel Robins
Tongtong Zhang

Appeared in the Journal of Molecular Biology and Evolution

www.cs.virginia.edu/robins

Background

Biological Sequences
DNA (gene) – RNA – Protein

cgtaacagac... MAEKPKLR...

Evolution

time

Evolutionary Tree Reconstruction

Optimality Criteria:
- Least-Squares
- Minimum-Evolution
- Maximum-Parsimony
- Maximum-Likelihood

NP-complete [Foulds & Graham 1982, Day 1987]

Previous Approaches
- Fitch • Margoliash [1967]
- Neighbor-Joining [1987]
- PHYLIP [1993]
- Split-Decomposition [1995]
- Quartet-Puzzling [1997]
- PAUP [1998]

All use greed & target best solution

Challenges

Topologically distant low-cost solutions exist

Random starting trees + heuristics?
- e.g. [Maddison 1991, Penny 1995, Swoford 1997]
Problem: getting trapped in local minima

Goal: given N taxa (sequences), find diverse low-cost evolutionary trees

Our Solution: Generalized Neighbor-Joining

Traditional Neighbor-Joining

Generalized Neighbor-Joining

Parameters controlling the solution space sampling:
- \( K \): max # partial solutions maintained
- \( Q \) (Quality): # candidates selected for low cost
- \( D \) (Diversity): # candidates selected for variety

Tradeoff quality & topological diversity: \( K = Q + D \)

Time complexity: \( O(K \cdot N^3 \cdot (\lg K + \lg N)) \)

Generalized Neighbor Joining: Experimental Results

Performance (N=8 leaves)

Solution Cost (N=16 leaves)

Solution Diversity (N=16 leaves)