第5回「計算科学による新たな知の発見・統合・創出」シンポジウム
@ May 14th, 2009

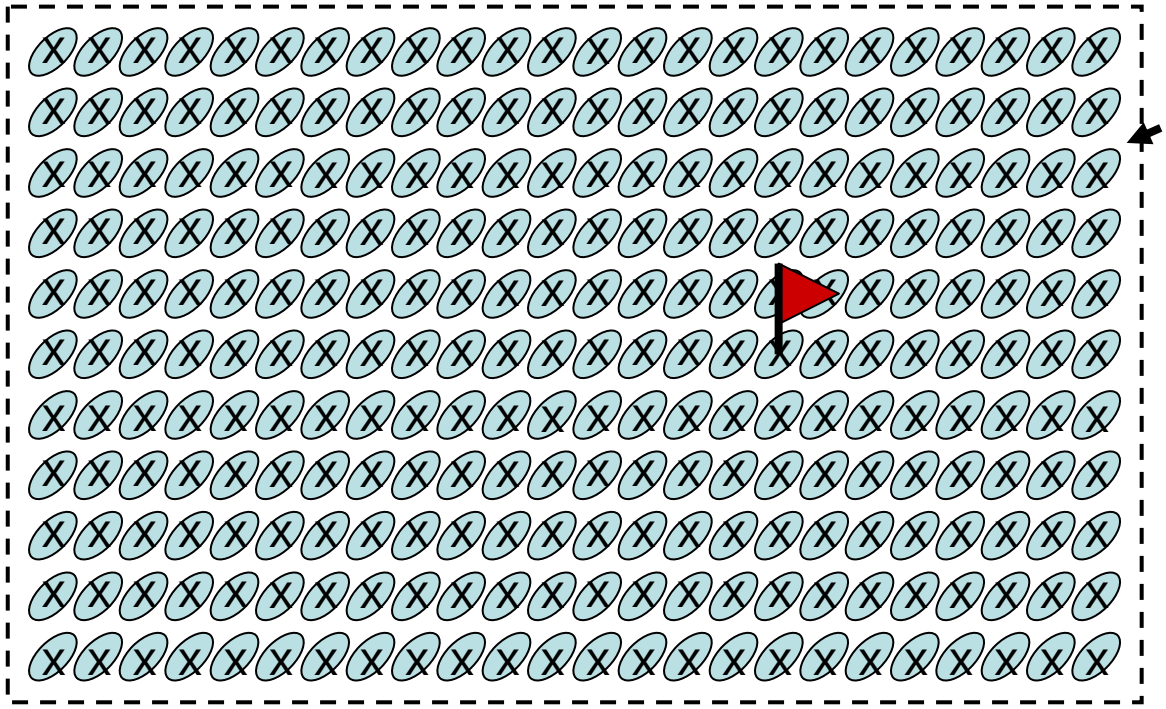
MOL EVOL MICROBES
UNIV.TSUKUBA

最尤法分子系統解析における heuristicな樹形探索の検討

稲垣祐司
（筑波大学大学院生命環境科学研究科）

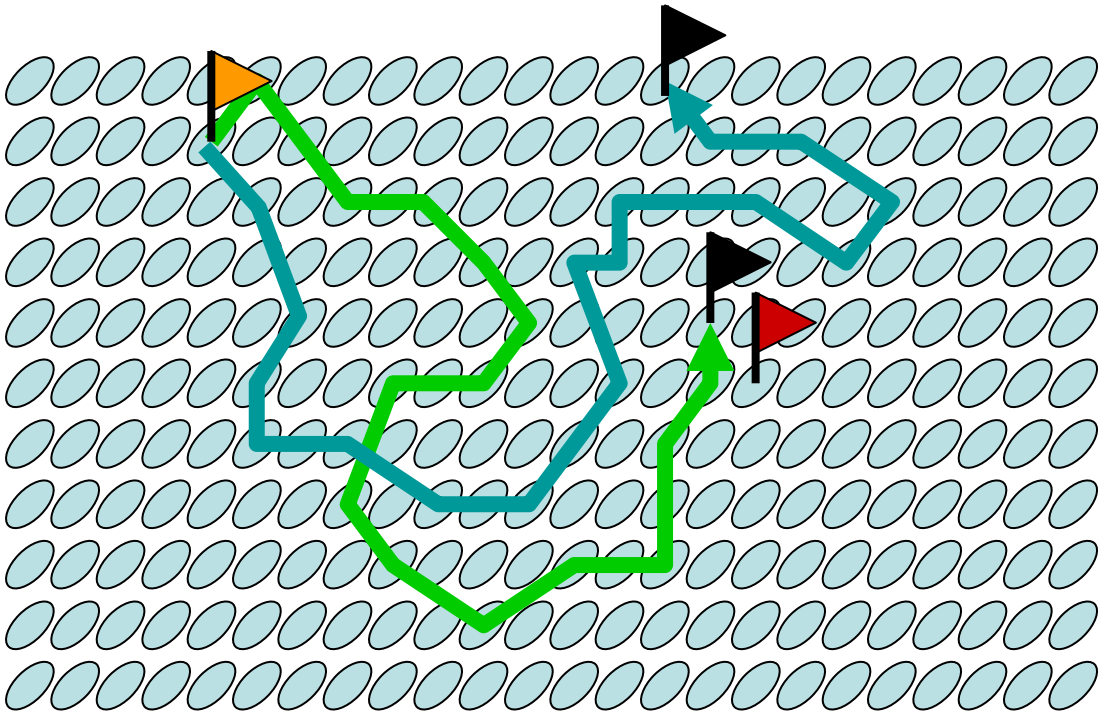
重点課題推進プログラム（08b-22）
網羅的樹形探索によるブートストラップ解析法の検討
(Study on bootstrap analyses by exhaustive tree search)
代表：橋本哲男（筑波大学大学院生命環境科学研究科）

Exhaustive tree search



- Exhaustive tree search (ETS) scores *all* possible trees
  - Never miss the ML tree
  - Computationally very intensive

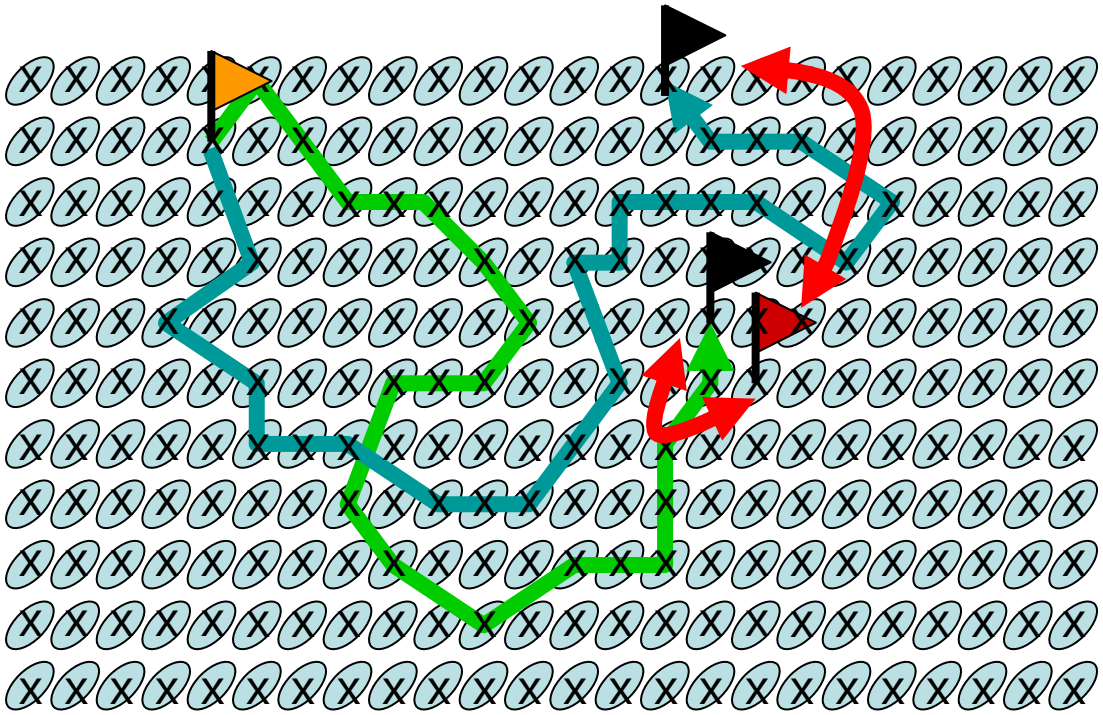
Heuristic tree search



: The output trees from HTS

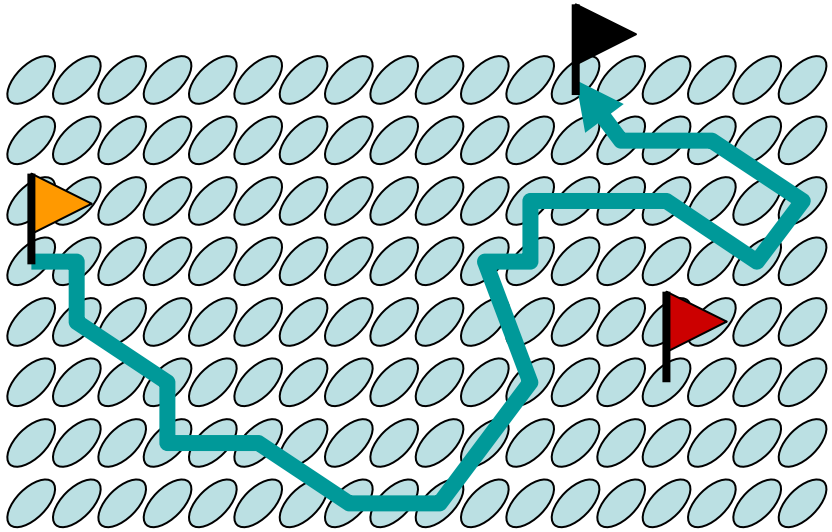
- HTS may *miss* the ML tree
  - Computationally *feasible*
  - The resultant tree could be an *artifact*
    - We usually don't know the ML tree

Motivation



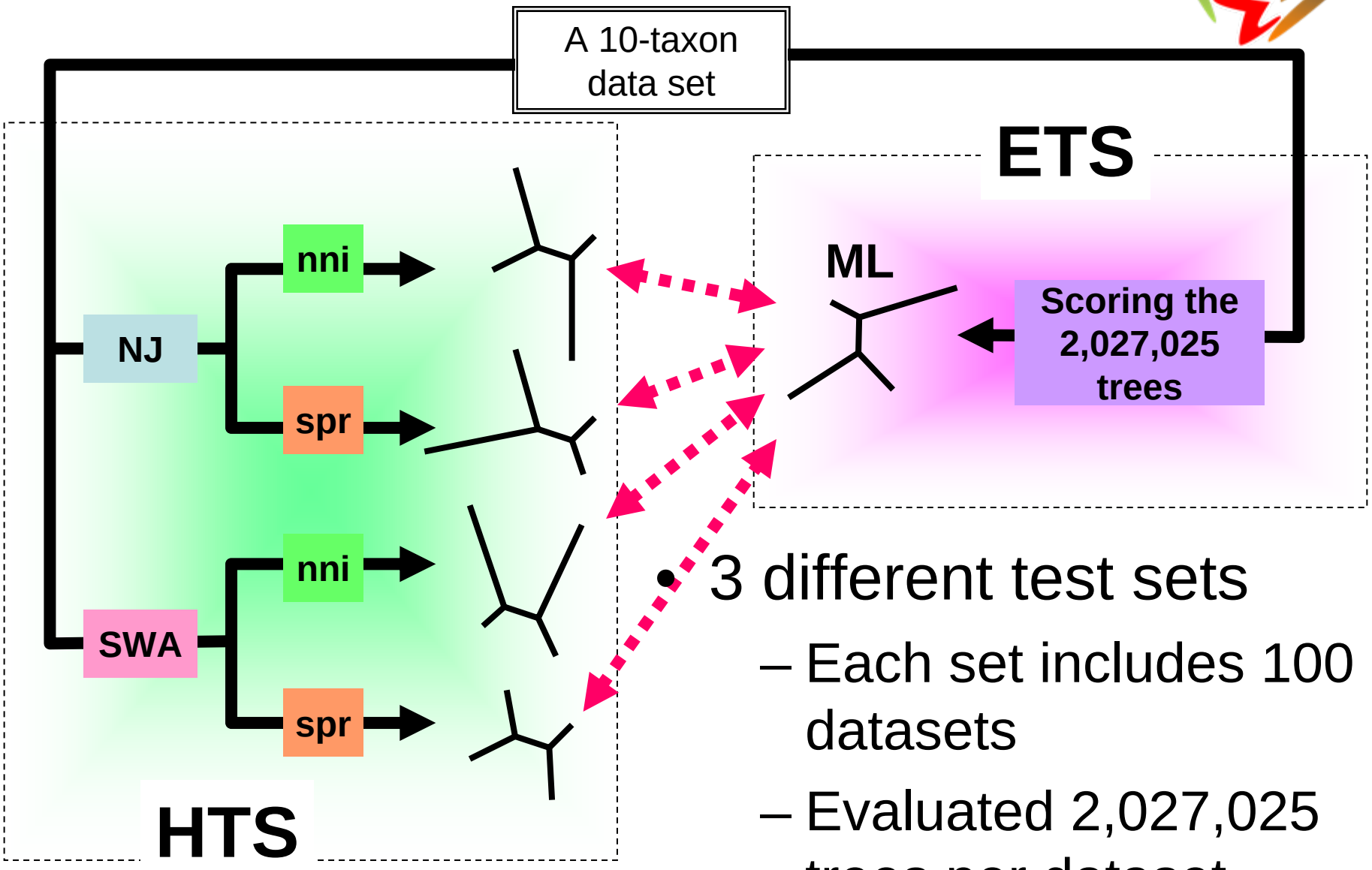
- Examine the efficiencies of HTS methods
  - Estimate the trees by HTS methods
  - Find the ML tree by the ETS method
  - Compare the ML tree and the “HTS” trees

Two parts in a HTS method



- Start tree
  - Neighbor-Joining (NJ)
  - Step-wise addition (SWA)
- Tree search
  - Nearest neighbor interchange (NNI)
  - Subtree Pruning & Regrafting (SPR)

Experimental design



- 3 different test sets
  - Each set includes 100 datasets
  - Evaluated 2,027,025 trees per dataset

Datasets

- “Mic” – including two long branches

The two LB are *not always* united in the ML tree

- “Arc” – including two long branch

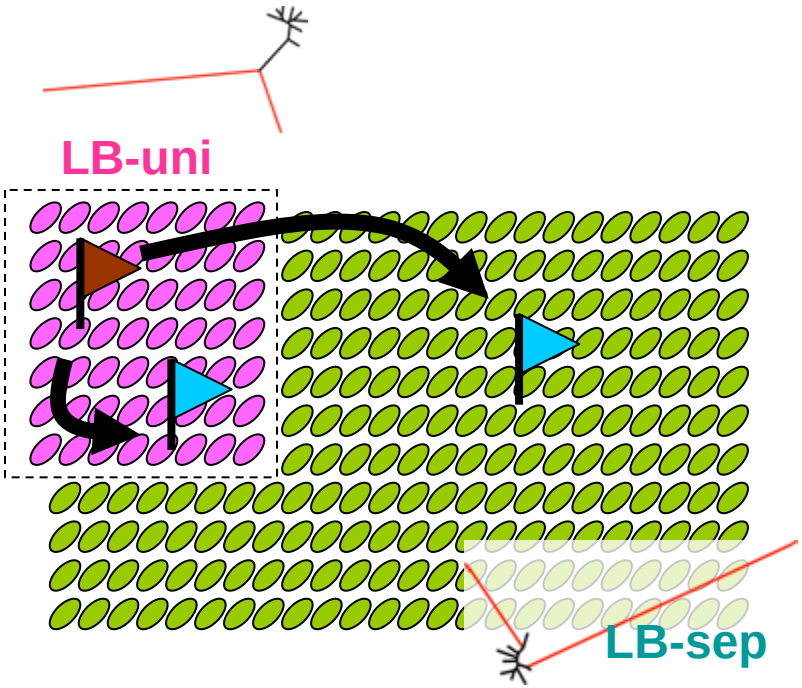
The two LB are *always* united in the ML tree

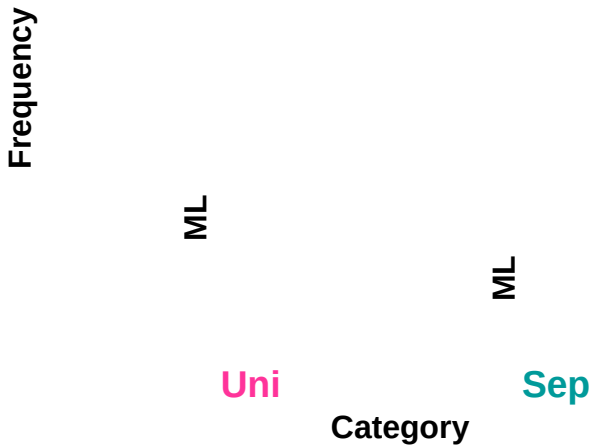
- “Sty” – including a single long branch

“Mic” – low efficiency

- Efficiencies are *unexpectedly* low
  - Start trees had a significant impact on the efficiency
    - NJ>SWA

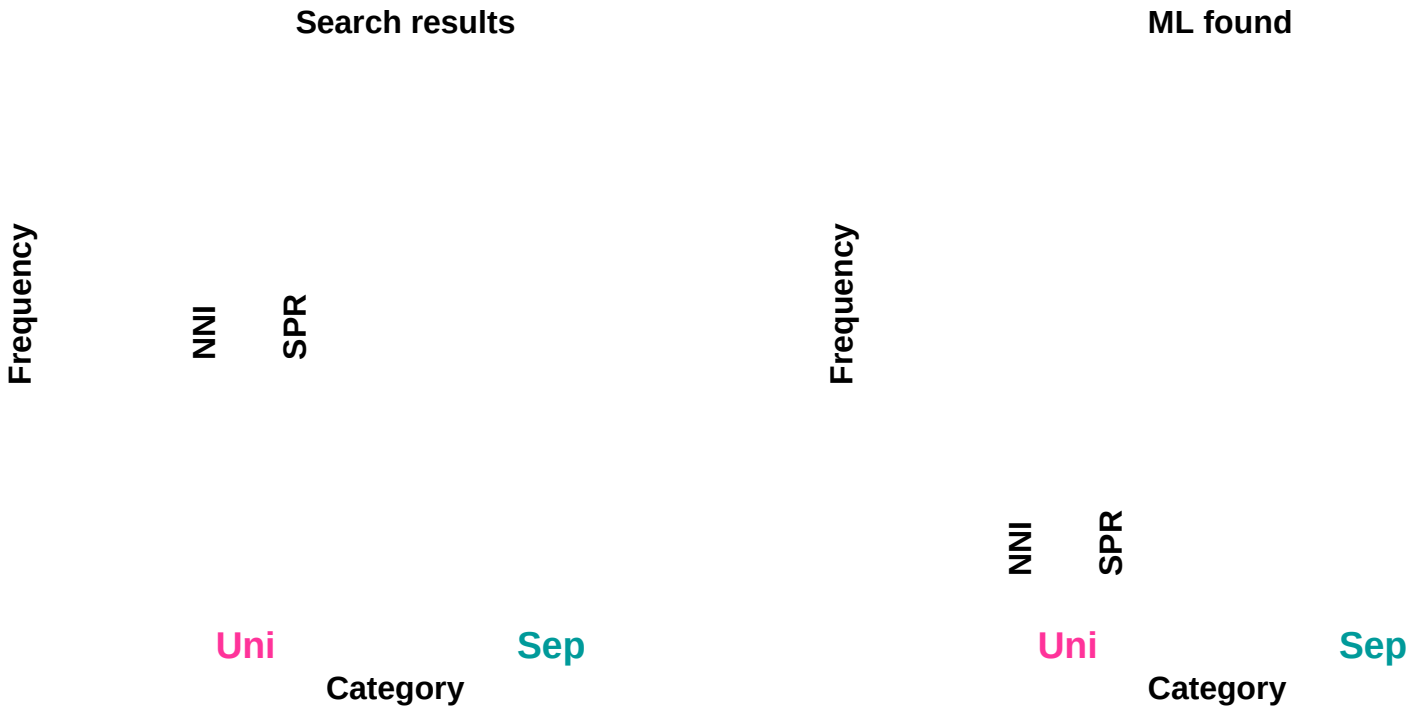
Impact of start trees – NJ





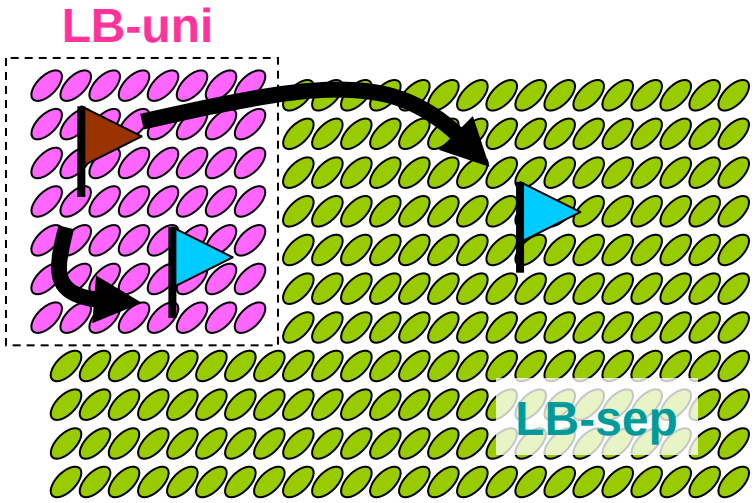
- Start trees are *all* LB-uni
- Search patterns required…
  - “Uni→Sep”; 32
  - “Uni→Uni”; 68

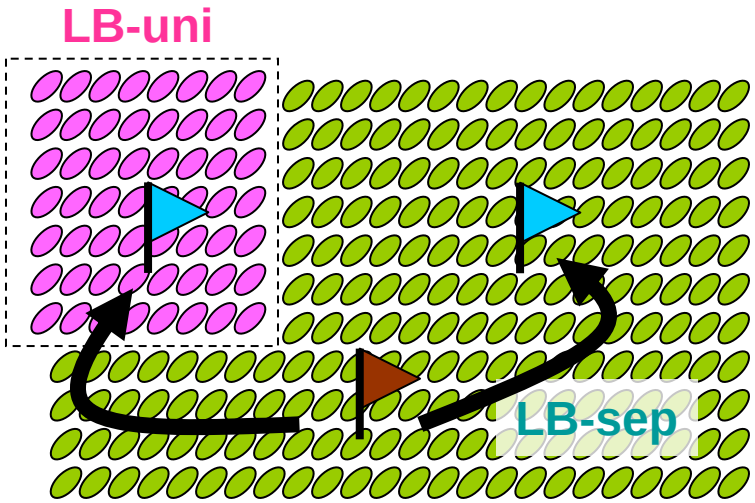
Impact of start trees – NJ

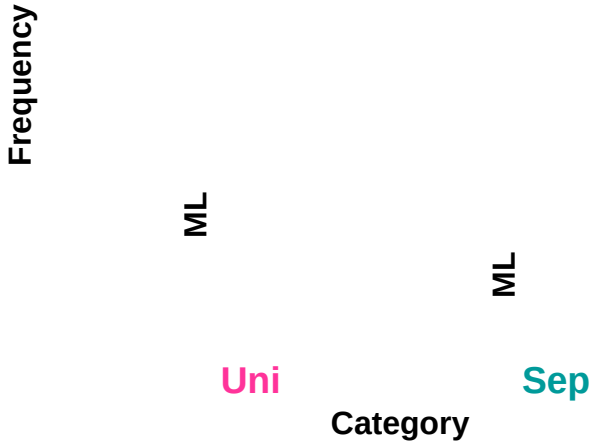


- NNI/SPR trees are *all* LB-uni
- No “Uni→Sep” search happened
  - NNI: “Uni→Uni” succeed only 27 times
  - SPR: “Uni→Uni” succeed only 33 times

Impact of start trees – SWA

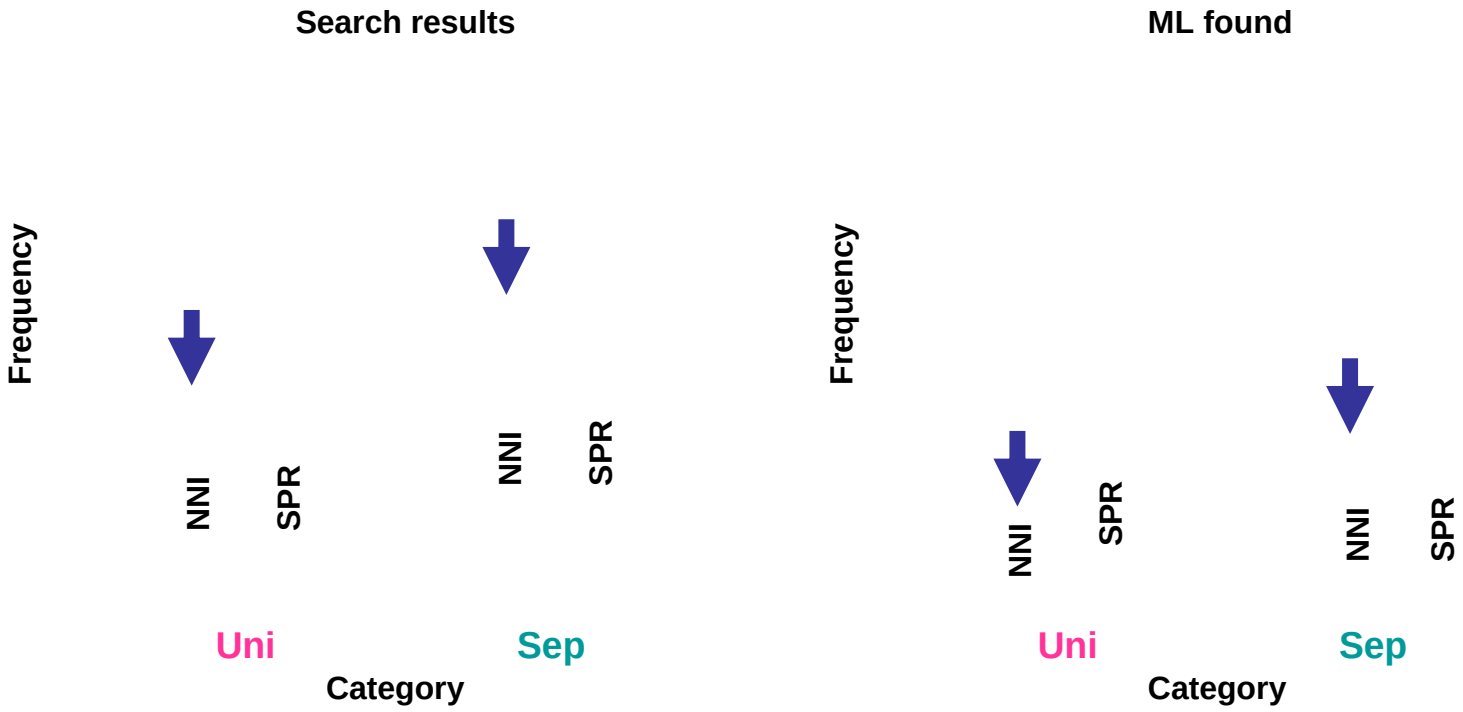






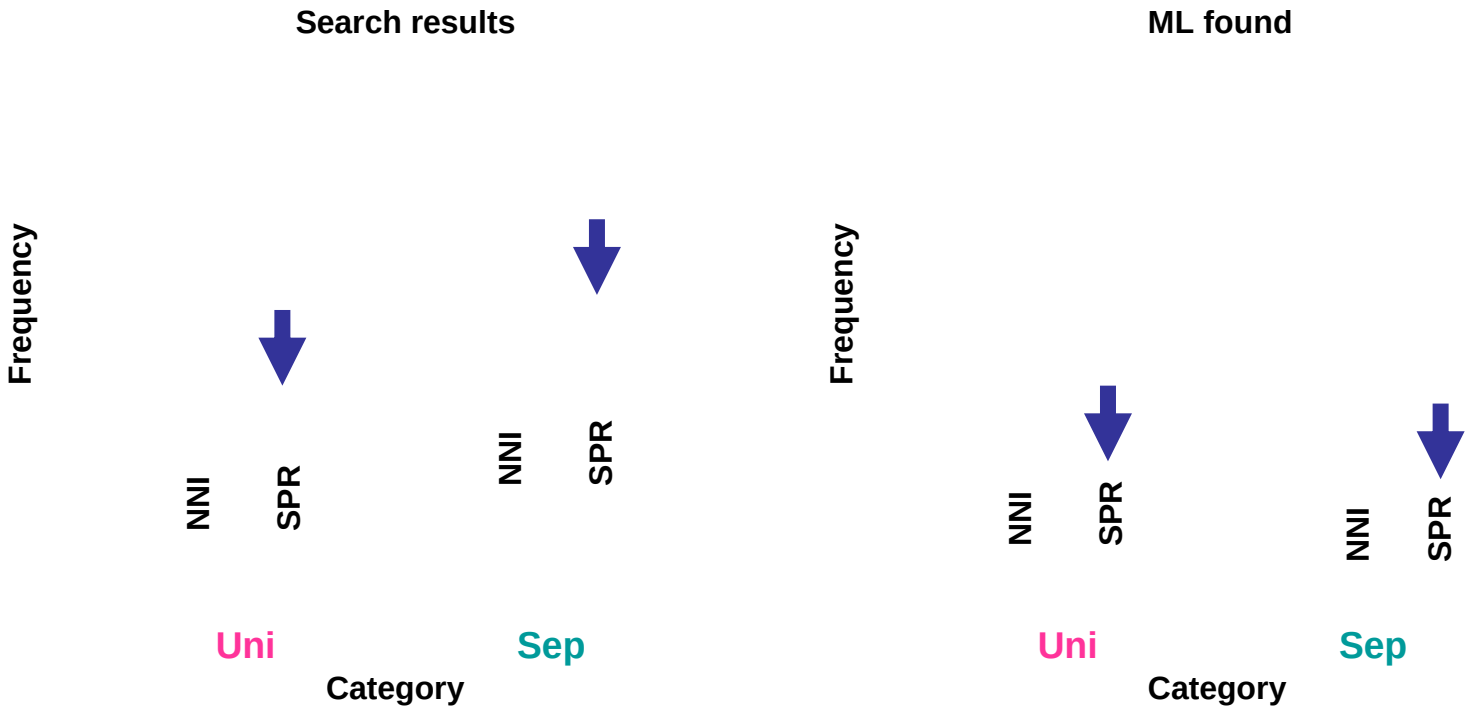
- Searches required…
  - "Uni→Sep"; 8
  - "Uni→Uni"; 24
  - "Sep→Sep"; 24
  - "Sep→Uni"; 44

Impact of start trees – SWA_nni



- “Uni↔Sep” search barely happened
  - “Uni→Uni” & “Sep→Sep”; 32 (3) & 60 (6)
  - “Uni→Sep”; 0 (0)
  - “Sep→Uni”; 8 (2)

Impact of start trees – SWA_spr



- “Uni↔Sep” search barely happened
  - “Uni→Uni” & “Sep→Sep”; 32 (6) & 58 (5)
  - “Uni→Sep”; 0 (0)
  - “Sep → Uni”; 10 (3)

Neither NNI or SPR are efficient

- NNI or SPR searches *limited* tree space
  - Uni→Sep or Sep→Uni barely observed
  - Start trees are important
    - To hit the ML^{uni} tree, you have to start from a LB-uni tree
    - To hit the ML^{sep} tree, you have to start from a LB-sep tree
- So, do we have to live with this?
  - No, increase start points
  - Improve search methods