Phylogeny
Quadtree Sampling

Rolf Backofen

Lehrstuhl für Bioinformatik
Institut für Informatik
Albert-Ludwigs-Universität Freiburg

Course Bioinformatics I

built on January 29, 2019
• **Finally:** Problem of find correct topology for unrooted tree
  • very complex problem ⇒ heuristics
  • best so far: quadtree sampling

• Quadtree \( \{a, b, c, d\} = \)

\[
\begin{array}{c}
a \\
\downarrow \\
b \\
\downarrow \\
c \\
\downarrow \\
d
\end{array}
\]

⇒ expresses neighbourship between taxa:

\( a-b \) are more closely related than \( a-c \) or \( a-d \) etc.
Example: 5 taxa

Consider given but unknown tree for 5 taxa

1. How many quadtrees? \( \Rightarrow 1 \) for every 4-subset of 5 taxa \( \Rightarrow \binom{5}{4} \) many

2. Given correct topology, what are the “correct” quadtrees?

\[
\{a, b, c, d\} \rightarrow \hspace{1cm} \{a, c, d, e\} \rightarrow \\
\{b, c, d, e\} \rightarrow \hspace{1cm} \{a, b, c, d\} \rightarrow \\
\{a, b, d, e\} \rightarrow \hspace{1cm} \{a, e\mid b, d\}
\]

Alternative tree encodings:
\[N_1(a, b|c, d), \ N_2(a, e|c, d), \ N_3(b, e|c, d), \ N_4(a, e|b, c), \ N_5(a, e|b, d)\]
Why are we interested in the \( \binom{n}{4} \) many quadtrees for \( n \) taxa?

**Theorem**

*Given the \( \binom{n}{4} \) many quadtrees associated with an arbitrary topology for \( n \) taxa. Then the set of all quadtrees uniquely determines the final topology.*

⇒ Idea of quadtree sampling

**Step 1** for each of the \( \binom{n}{4} \) subsets of \( n \) taxa, determine the optimal quadtree using Felsenstein.

3 possibilities for each subset of 4 taxa to check

**Step 2** combine the found quadtrees into final topology

- *Step 2 “simple” if all quadtrees are correct*
- *however, this is usually not the case ⇒ heuristic combination*
Idea Quadtree Sampling

- Idea of heuristic in **Step 2**: minimize error!
  - ⇒ iterative procedure, extend tree by one taxa in every step.
  - given reconstructed tree (topology) for \( k \) taxa \( \{ t_1, \ldots, t_k \} \)
  - consider every possible edge to add \( t_{k+1} \)

  **for each edge** \( e \):
  - consider all quadtrees (=neighborships) for \( \{ t_1, \ldots, t_{k+1} \} \) that contains \( t_{k+1} \)
  - count the number of neighborships not satisfied by the edge \( e \)
  - choose edge \( e \) with minimal number of unsatisfied neighborships

- Example: correct tree:

![Diagram](image.png)

- assume (**correct !**) result from step 1: associated quadtrees
  
  \[
  N_1(a, b | c, d), \; N_2(a, e | c, d), \; N_3(b, e | c, d), \; N_4(a, e | b, c), \; N_5(a, e | b, d),
  \]
Example: Effect of Addition of Edge

- add edge $e$ to the following topology for $\{a, b, c, d\}$

- now add node $e$ (incorrectly) to the $c$ edge:

- check satisfied/unsatisfied neighbourships:

  \[
  N_1(a, b|c, d) \checkmark \quad N_2(a, e|c, d) \xmark \quad N_3(b, e|c, d) \xmark \\
  N_4(a, e|b, c) \xmark \quad N_5(a, e|b, d) \xmark 
  \]

- edge will be marked with 4 (errors)
Example: Quadtree Sampling

\[ N_1(a,b|c,d) \]

\[ N_2(a,e|c,d) \]

\[ N_3(b,e|c,d) \]

\[ N_4(a,e|b,c) \]

\[ N_5(a,e|b,d) \]

CC BY-NC-SA 4.0 - R. Backofen - Bioinformatics - University Freiburg - Germany
Final Outcome:

- **Note:** there might be several edges with minimal error

- **Solutions:**
  - randomize succession order of taxa
  - select minimal edge at random
  - run several times
  - build final **majority tree** (see literature)