Phylogenetic Trees

- Representation of the evolutionary relationship between species
From Organisms to Trees

DNA Sequencing

A Set of Taxa

Ape: ACCGTAGCTT
Bear: ATAGTAACCT
Dog: CCGTATTT
Emu: CGCATAAGC
Frog: CCTAAC
Goat: GTAATAGAAC

Unaligned Sequences

Aligned Sequences

Maximum Parsimony Search

Multiple Sequence Alignment

Set of Optimal Trees

Consensus Analysis

Consensus Tree

Phylogenetic Trees in ACL2 – p.3/14
Lots and lots of trees

- Number of possible trees grows exponentially with the number of leaves in the tree

- Two main methods used to determine the correct tree
  - A heuristic search through tree space
  - A Bayesian estimation of phylogeny using Markov chain Monte Carlo

- Both of these methods may produce hundreds, or thousands of trees which are then the input to further processing
Lots and lots of trees

- Number of possible trees grows exponentially with the number of leaves in the tree
- Two main methods used to determine the correct tree
  - A heuristic search through tree space
  - A Bayesian estimation of phylogeny using Markov chain Monte Carlo
- Both of these methods may produce hundreds, or thousands of trees which are then the input to further processing

Need a system to store these trees efficiently, and perform post-tree analysis.
Why Use ACL2?

- Standard answer: Accuracy
  - Explicit specification of input and output for all functions together with proof that the specification is met within the code (guards)
  - Two representations of trees, with proof that we can accurately move from one representation to the other and back
- Additional answers: Storage space and performance speed
  - Hash-consing gives greatly reduced storage space
  - Memoization gives improved performance speed
- Overall: Medical systems of the future
TASPI High-Level Representation:

(( (A B) C) ( (D E) (F G) ))
(( (A B) C) ( (D E) F G))
(( (A B) C) (D (E (F G))))
((A (B C)) ( (D E F) G))
((A (B C)) ( (D E) (F G) ))
TASPI Low-Level Representation:

((#1=((A B) C) #5=(#6=(D E) #9=(F G)))
(#1#(#6# F G))
(#1#(D (E #9#)))
(#12=(A (B C)) ((D E F) G))
(#12##5#))
Reduced Storage Space

![Graph showing reduced storage space for different data set numbers and sizes. Two lines represent Newick and TASPI.bhz formats, indicating the size in bytes for each data set number.]
Bipartition Representation
Bipartition Representation

Parenthetical Notation:

\((A \ B \ (C \ ((D \ E) \ F)))\)  \((A \ (B \ ((D \ E) \ F)) \ C)\)  \((A \ B \ ((C \ (D \ E)) \ F))\)
Bipartition Representation

Parenthetical Notation:
(A B (C ((D E) F)))  (A (B ((D E) F)) C)  (A B ((C (D E)) F))

Bipartition Representation:

- AB | CDEF
- ABC | DEF
- ABCF | DE
- AC | BDEF
- ABC | DEF
- ABCF | DE
- ABF | CDE
- ABCF | DE
Parenthetical Notation:
\((A\ B\ (C\ ((D\ E)\ F)))\)  \((A\ (B\ ((D\ E)\ F))\ C)\)  \((A\ B\ ((C\ (D\ E))\ F))\)

Bipartition Representation:

<table>
<thead>
<tr>
<th>AB</th>
<th>CDEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>DEF</td>
</tr>
<tr>
<td>ABCF</td>
<td>DE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AC</th>
<th>BDEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>DEF</td>
</tr>
<tr>
<td>ABCF</td>
<td>DE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AB</th>
<th>CDEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF</td>
<td>CDE</td>
</tr>
<tr>
<td>ABCF</td>
<td>DE</td>
</tr>
</tbody>
</table>

Our Bipartitions:

- \((A\ B\ C\ D\ E\ F)\)
- \((C\ D\ E\ F)\)
- \((D\ E\ F)\)
- \((D\ E)\)
(defthm paren-partition-paren
  (implies (and <properties of input tree>
               <properties of ordering>
               <properties of tree and ordering>)
           (equal (tree-from-fringes (get-fringes tree ordering)
                           ordering)
                  tree)))
Strict and Majority Consensus

- **Strict consensus**: Any branch that appears in every input tree is in the consensus tree.

- **Majority consensus**: Any branch that appears in more than half of the input trees is in the consensus tree.
Example

Phylogenetic Trees in ACL2 – p.11/14
Example

Phylogenetic Trees in ACL2 – p.11/14
Improved Consensus Performance

Phylogenetic Trees in ACL2 – p.12/14
Conclusion and Future Work

- TASPI provides accuracy guarantees, while providing state of the art performance in terms of size and speed.

- TASPI is being extended to perform further post-tree analyses, as well as database operations.
Questions?