Combining Data Sets

Matt Mauldin

Definitions

- Character Independence: changes in character states are independent of others
- Character Correlation: changes in character states occur together
- Character Congruence: Different characters agree or disagree
  - Congruence among Data Sets
Outline

- History
- Philosophical Outlook
- Methodology
- Which Approach Works Best?

History

- 20 year debate still going strong
- Kluge 1989
  - Numerous papers written in response
- Independent Analysis
  - Miyamoto & Fitch (at one time)
- Combined Analysis
  - Kluge
- Conditional Combination
Philosophy

- TE: Total Evidence approach (Carnap 1950)
  
  - Requires that all evidence (old and new) be considered for proper inductive inference. It was first introduced into systematics by Kluge in his 1989 paper, in an argument directed against differential character weighting, data set partitioning, and the dismissal of any evidence before analysis.

  - “Assume nothing—it will all come out in a wash”

---

Philosophy

- Segregation Approach
  
  - Not all data is the same
  
  - Old data and new data
  
  - Evidence must be separated into distinct classes, and analyzed individually

  - Do data sets agree or disagree?
Philosophical vs. Methodological

- These two arguments have been developing in parallel for the past 20 years
- Focus on Methodology
- In the future These debates may converge influencing this debate
  - Unsure how

Why is there an issue?

- Data is data
- If all data is equal, then there’s no reason not to combine data sets
Data is Data

- Is all data the same?
- If not what do you need to determine the difference between sets or classes of data?
- How different is different?

"To qualify as distinct class of evidence, characters in a data set must, in a statistical sense, be more similar to each other than they are to characters in other data sets with respect to some property that affects phylogeny estimation by the given method." (Bull et al. 1993)

- Do separate classes of data exist? (Kluge & Wolfe 1993)
  - Continuous Versus Distinct Characters
  - Molecular evidence is undeniable (Miyamoto & Fitch 1995)
    - Introns, Exons, Pseudogenes, Mitochondrial Genes
    - Differences in Overall rate of evolution, gene recombination, gene duplication, and gene conversion
    - Differences in Character Independence, Correlation, and Congruence both within and between data sets
• “When faced with biologically defensible divisions, the nature of these partitions should be taken into account in the choice of analytical methods.”

-de Querioz 1995

Variation in Data

• Since classes of data do exist, decisions dealing with methodology must be made
  • 56 Major Models of Evolution GTR+G
    • GTR+I+G
    • HKY+G
    • HKY+I+G
  • How to analyze the data
    • Separate models make specific assumptions
    • These assumptions drive analyses in distinct ways
      • Different models produce different trees
Methodology

- Separate Analysis
  - Separate Trees
  - Consensus Trees
- Combined Analysis
  - Combined Data Set Tree

Separate Analysis

- Individual trees based on single data sets
- Compare trees, and shared supported clades
  - Bull et al. 1993
    - Rapidly and slowly evolving genes analyzed separately and through consensus. Individual tree was closer to the true tree.
- Problems
  - Two trees are completely different, which is best? Are either close to the true tree?
  - Uninformative characters or data set(s)
  - support values differ on between trees
Separate Analysis

Consensus trees
- Data sets analyzed separately
- Trees are combined
- Different types of consensus trees
  - Consensus methods
  - Consensus index

Issues with Consensus Trees
- Data is lost during consensus
  - Parts of some trees are ignored, others emphasized
- Different methods recover different trees
  - Different consensus methods often produce very different trees, producing a problem with repeatability
- Choices to be made
  - Partitioning
  - Consensus Methods & Indices
    - Strict consensus (Sokal & Rohlf 1981) Cluster Based
    - Majority consensus (Margush & McMorris 1981) Cluster Based
    - Adams consensus (Adams 1972) Intersection Based
      - Best tree that represents all the information shared among a set of trees
### Overview of Separate Analysis

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• No tainting “good” data with “bad” data</td>
<td>• May contain clades not found in any original trees</td>
</tr>
<tr>
<td>• Allows for separate models of evolution</td>
<td>• Many decisions and assumptions to be made</td>
</tr>
</tbody>
</table>

### Combined Analysis

- Arguments in favor
  - Total Evidence
  - Objections to arbitrariness in consensus methods
  - Difficulty of choosing a scheme for partitioning
    - No partitioning necessary
  - Greater descriptive and explanatory power of phylogenetic hypotheses generated from combined data
  - Greater ability to uncover “real” phylogenetic groups
Uncovering Real Phylogenetic Groups

- It is possible for combined analysis to show novel groups not present in either of the data sets
  - Olmstead & Sweare 1994
    - 3 trees AB, AC, & BC
    - Each tree had unique features
  - Frequently Increased bootstrap support

Decisions

- Weighting
  - Unweighted
    - Swamping out of small data sets
    - Could produce poor tree if big data set is “bad”
  - Weighted
    - Whole Data Sets
      - Weight smaller data sets so swamping doesn’t occur
    - Characters within data sets
      - Transitions vs. Transversions
      - Gaining vs. Losing Morphological features
      - Down weight the third position
Barrett et al. 1991 argued that weighting decisions should be explicitly defended, rather than being arbitrarily decided by the outcome of the method of analysis.

Overview of Combined Analysis

Pros
- No data lost during consensus
- Increases number of characters
- No decision regarding partitioning during analysis (fewer assumptions)

Cons
- Can be misleading if heterogeneity in data exists
- May obscure significant patterns of congruence or conflict among characters
- Combining "good" and "bad" data
What To Do?

- Run Tests
  - Incongruence Length Difference (ILD) test (Ferris et al. 1994)
    - Quantifies conflicts that occur between data sets
    - Separate Debate over ILD
      - Considered very useful when first developed
      - Recent evidence argued against the usefulness of ILD test
  - Templeton Test (Templeton 1983)
  - Topological Incongruence Test (Rodrigo 1993)

What Type of Analysis is Best

- Run tests
  - Issues and debates
- Each type has both Pros and Cons
  - Weigh Pros against Cons for each analysis
- No Simple Answer
- Perform both types
- Possibly nontraditional methods that use advantages of both
  - Cao et al 1994 (Maximum likelihood approach)
Focus

- Development and validation of specific tests for heterogeneity among data sets.
- Test for presence and cause of heterogeneity

At Present

- The debate goes on, until some definitive test is derived which is able to make clear-cut decisions about combining data is developed.