# Supertrees and SuperNets

for Agnostics

David Bryant*

At a time when huge multi-gene alignments can be assembled in a few thousand mouse-clicks, supertree methods are coming into their own. It is difficult to deny their appeal: according to papers and advocates, supertree methods will allow us to analyse hundreds of genes and thousands of taxa. They can incorporate almost any type of data. The are fast and they have a catchy name.

My own interest in supertrees started over 10 years ago, in my Ph.D work with Mike Steel. But, in all honesty, their main attraction was the neat mathematical and algorithmic problems that they generated. Later, I learnt statistics and became extremely sceptical of supertree approaches, and especially of any novel conclusions that were supported only by supertree methods. At the moment I fluctuate between advocate and sceptic, changing, perhaps, weekly.

As such, I am reluctant to give a supertree recipe talk, because I'm sure I'll regret it in the following week. Instead, what I want to do is give some of the theory background you'll need to use, and interpret, supertree type methods properly.

I will spend relatively little time describing the methods themselves - these are well covered in a recent book on Supertrees edited by Olaf Bininda Emonds [**?**], much of which can be found online using google searches...

The main sections of the talk are:

- *What is it that makes phylogenies incorrect*? This is background material - but its essential if we are going to talk about pros, cons, and dangers of supertree analysis. I'll talk about different kinds of error, and the techniques we have available for dealing with them.
- *Overview of the supertree approach.* The differences between separate and combined analysis. A quick survey of consensus and supertree methods.
- *Statistics of supertrees.* The guts of the talk. I talk about potential hazards when making inferences about multi-gene analyses, determining homogeneity, and so on.

*McGill Centre for Bioinformatics, http://www.mcb.mcgill.ca/~bryant

# 1 What is it the makes phylogenies incorrect?

- Software bugs and software misuse.
- Random error.
  - What random and sampling error is.
  - Example: coin toss.
  - Message: we can cope with random error if we have enough data and if our models are good.
  - Bootstrap as an indication of the extent of sampling error (but not necessarily of support).
  - Confidence intervals in statistics and phylogenetics (Cartoon guide)
- Bias
  - What is bias in normal statistics.
  - Bias in tree methods... even with correct models
  - Cartoon of long branch problem.
- Systematic error
  - Modelling error - why we should (and can) proceed with an assumption of incorrect models.
  - This is the most difficult error to deal with.
  - Model refinement.

# 2 Overview of the supertree approach

- The basic question: to combine or not to combine.
- Combined analysis (for likelihood and Bayesian).... how its done and how it could be done.
- Supertree and Consensus analysis.
  - Recoding methods
    - MRP analysis.
    - Average consensus trees
    - Min Flip
    - Paint-by-numbers supertree method.
  - Combinatorial trickery
    - Consensus supertrees
    - Aho et al's algorithm and a million variations.
    - The 6 million dollar supertree.
    - YAPTP
  - Bayesian approaches
  - Supernet methods

## 3 Statistics of Supertrees

- Low risk interpretation: graphical representation.
- Significance of getting the same tree.
- Significance of not getting the same tree.
- Real information, or poly-filla?