Version dated: January 13, 2016

Project thoughts

What are inapplicable characters (in a phylogenetic sense)? If we consider them as true NAs (in a mathsy way) they can just help solving regions of the tree where there are indeed applicable. For example, in an artiodactyles tree, a first character (say antlers presence/absence) will be useful as a synapomorphy for cervidae and will group them together (antlers present) in opposition to the other artidactyles (antlers absent). Then, any extra characters based on antlers will be inapplicable among non-cervidae and will only be useful in a phylogenetic sense in cervidae.

Molecular characters have any of the five states (ACGT-) and can be coded as "?" (as a token for any of these five states). In page 7 of the draft, I think the example is not correct: in my mind there is no real parallel between morphological inapplicable characters and molecular *indels* since:

1. Genes are orthologuous so *Indels* as, the four other molecular characters are based on gene's orthology (which is not always the case for morphological characters - e.g. tail color can be homologous). Molecularists seems to have fixed the problem of inapplicability from the start by not allowing it (e.g. the big problems when reconstructing the tree of life: the only attempts to date have used the rusty supertree approach).

2. *Indels* do bring some phylogenetic information (i.e. the gain/lose of a nucleotide at site X) that can be translate to some biological information (e.g. a shift in the DNA traduction frame resulting in new proteins being produced or no proteins at all).

In other words: one indel *is* a phylogenetic change as in an inapplicable character is not necessary a change. For example in a big vertebrate phylogeny, the character number of digit is not a change in fishes (but is it in snakes?)...

1 Parsimony and inapplicables

1.1 homology/homoplasy

When the presence/absence character is not synapormorphic (i.e. the tail example with presence of the tail in two independent clades), how should any inapplicable character be treated in both clades (i.e. the color of the tail)? In other words, if the presence of tail appears two times independently in the phylogeny, is the color of the the tail an homologuous character? Or should it be coded something like "tail color in clade 1" and "tail color in clade 2"? Also, a wild question: is homoplasy a biological thing or a parsimony thing?

1.2 Algorithms

One idea would be to implement a "pseudo-super-matrix" approach:

1. Separate the matrix into multiple matrices:
    - the one containing characters applicable among all taxa
    - multiple blocks containing characters only applicable to blocks of taxa (i.e. cervidae block etc...)
2. treat only applicable characters and get the most parsimonious trees (all of them)

3. on all these equally parsimonious trees, add the other matrices blocks and get the most parsimonious trees from there

Side note: I think computational problems have to be optimised (because that's the cool way to code) but it doesn't need to be the fastest. For example, if the Ramirez implementation tested all the configurations (and is justified) better run them all (and use computer clusters to speed up the yoke).

1.3 Testing the algorithms

Generate trees to match the expectations? For real datasets pick up dataset with a "third party" answer (molecular?). I thing there's potentially good candidates among mammals (see Biology Letters paper).

1.4 Characters non-independance

I think that's not a big deal... Characters are totally not independent in DNA. Especially in exons were the structure per triplets *has* to be conserved.

1.5 Other idea

Maybe go away from most parsimonious and get a "best treeS" approach? I.e. if the minimum step is x and the maximum (theoretical) is y get all the trees with the minimum steps +/- a small percentage (5%?).

2 Probabilistic approaches

Current model (Mk) suffers of the same over-simplifying assumptions than the parsimony approach. What is needed is a more relaxed model with:

- A rate heterogeneity for each characters (i.e. characters can have different rates) – This is already fairly well implemented in MrBayes/BEAST
- A relaxed Mk evolutionary model (i.e. the rules of change between states for each character) – April Wright is kind of working on that
- A more efficient way to deal with inapplicable characters?
- Get a reversibility parameter? ($0 \rightarrow 1 = 1 \rightarrow 0$?)

One idea for dealing with inapplicables would be the same one as for parsimony: just treat them independently in the concerned tree regions?

2.1 Using the CAT model?

CAT model allows to split the characters into discrete categories with their own rate and states frequencies. This allows the MCMC to "choose" the best category upon which the substitution model (Mk) can be applied (Lartillot and Philippe, 2004). The idea might be to look at the distribution of characters states *a priori* and make the algorithm decide which substitution model to use.

2.2 Combining morphological matrices

Maybe we also need a more efficient way to combine pseudo-independant morphological matrices. One idea would be to compare matrices and group characters with the same information. The user can then check if characters with the same info are (1) independent (keep them both) or (2) synonym (keep one).

3 Side projects

3.1 mulTree

Using Bayesian posterior tree distribution in comparative methods (i.e. not just the consensus tree but all the *best* trees).

3.2 dispRity

Using multidimensional approaches for solving ordination problems (i.e. using all PCA axis rather than 2...).

4 Cool idea

One good thing about solving the inapplicable data riddle is that it can have big implication in developing a "super-matrix" approach for morphological characters. One analogy is that each exon/intron/UTR in the molecular super matrix could be the equivalent to one character. In the molecular super matrix way, some regions of the matrices are just marked as "?" because the gene has not been collected (or is inapplicable!) for such and such taxa. Transposing that to morphological data, we could combine

data in a clever way and mark the data as true missing data as well ("?"). If inapplicability is well implemented, then some characters states should be estimated as "-" rather than classically any of the available states. This will allow to efficiently combine the matrices and run big morphological trees!

5 Misc

5.1 Conferences

- BES Macro (June - Oxford)
- Evolution (17-21 June - Austin)
- SVP? (26-29 Oct - Salt Lake City)
- EAVP? (6-9 July - Haarlem)

5.2 Holidays

Finland sometimes in March and Croatia sometimes in April?

References

Lartillot, N. and H. Philippe. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21:1095–1109.