Allan Wilson Centre, Biomathematics Research Centre, University of Canterbury
Christchurch, New Zealand, 29th January - 29th April 2015

The search for the most informative character

Ina Deutschmann
Project Supervisors: Mike Steel, Charles Semple and Magnus Bordewich
Joint work with Elisa Kasbohm

1 Introducing $I(f)$

A central question in evolutionary biology is how to reconstruct an evolutionary tree, a ‘phylogeny’, from present-day data. We often have data in form of discrete characters, which partition the taxa into disjoint subsets according to the states each taxon takes. Morphological characters, e.g. ‘wings’ vs. ‘no wings’, is an example of a 2-state character, while the amino acid at an aligned protein sequence site is a 20-state character. For each taxon it describes which of the 20 amino acids is present at that site.
We assume that each character is homoplasy-free, i.e. its evolution could have been described on some tree without assuming reversals or convergent evolution. Then each character constrains the set of phylogenetic trees that are possible for those taxa. Indeed it has been shown in [1], [2] that for any binary phylogenetic tree on any number of species, there exist just four multistate characters that ‘capture’ that tree, in the sense that no other tree allows the characters to be homoplasy-free.
We first investigated the question of the extent to which just a single multi-state character confines the set of possible phylogenetic trees under the homoplasy-free assumption. For a character f on X, consider the following measure $I(f) = -\ln(P(f))$, where $P(f)$ is the proportion of binary phylogenetic trees on which f is homoplasy free, which was introduced in [3]. When one has just a single character, $I(f)$ is determined just by the sizes of the subset of the taxa in the partition of X induced by f.
Our first result was to show that the character f that maximized $I(f)$ for a given size n of X and number r of states are the ones that break the set X into roughly equally sized subsets. That means, when r devides n, then all subsets will have size $\frac{n}{r}$, but when r does not divide n, then k sets will have size $\lfloor \frac{n}{r} \rfloor$, i.e. $\frac{n}{r}$ rounded down to the closest integer, and $r - k$ sets will have the size $\lceil \frac{n}{r} \rceil$, i.e. $\frac{n}{r}$ rounded up to the next integer. Here $k = r\lceil \frac{n}{r} \rceil - n$.
We then wanted to know which value of r allows the largest possible value of $I(f)$ and therefore gives us the most information. This is a more subtle question and led to some surprising findings.

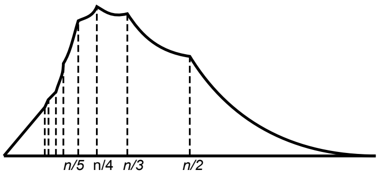

Figure 1: A scheme of the concatenation of waves.

2 Maximizing $I(f)$

Jumps: r_{max} does not grow steadily with n. From time to time r_{max} jumps from a bigger to a lower value. The bigger n the bigger is the jump back to a lower value for r_{max}. The sequence of n where a jump occurs does not seem to follow a special pattern: 9, 30, 104, 345, 1109, 3485, 10753, 32704, 98349, 293028, 866357, 2544799. We can approximate $I(f)$ and determine r_{max}^{approx} which has the advantage that no jumps occur.
Waves: Let n be fixed. The graph for $I(f)$ against r is a concatenation of curves. Rather than being smooth, the graph of $I(f)$ against r for fixed n is a concatenation of curves, *waves*. The boundaries of the waves are $n, \frac{n}{2}, \frac{n}{3}, \frac{n}{4}, \ldots$ Assume there is a jump between the r_{max} of n and $n+1$. A closer look at the peak of the two curves reveals that it seems to be a *flat* zone. However, zooming in on this zone shoes it's actual form: it is concave. For n is the maximum on the right peak, but for $n+1$ is it the left peak.
Width: When you do have a jump, you can quite nicely estimate the size of it. Due to the concatenation of *waves*, along the curve of $I(f)$, we sometimes find on the top of the curve the phenomenon that we don't have a peak but rather two mountains. From time to time r_{max} jumps back in its value. This occurs when the hight of the left mountain peak is rising above the peak of the right moutain peak from n to $n+1$. Let $n+1$ be the position of a jump. Therefore $r_{max}^{n+1} < r_{max}^{n}$. We can estimate the distance $r_{max}^{n} - r_{max}^{n+1}$ by $\frac{r_{max}^{n+1} \cdot r_{max}^{n}}{n}$. Our data indicated that this does not provide exact values. However it is a nice estimation of the size of the jump.
Alternation: The blocksizes grow and can also fall back by one. The blocksizes do not grow steadily and once a new blocksize occurs, it will not be one of the main two blocksizes right from the beginning. This also happens at the last occurrences.
The growth of r_{max} with n. We were able to show that $r_{max} = o(n)$. So $\frac{n}{r}$ goes against infinity as n goes against infinity. Also as n goes against infinity $\frac{r_{max} \ln(r_{max})}{n}$ goes against 1.

3 Open Questions

We did quite a few calculations and accumulated a lot of data. The last section revealed that there are still open questions. Two of the leading questions are: When does a jump occur? When do the blocksizes change?

4 Acknowledgements

I thank the Allan Wilson Centre for providing project funding and giving me the opportunity to gain relevant research experiences in my field of study. I would also like to thank Mike Steel, Charles Semple and Magnus Bordewich for supervising this project. I really enjoyed the fruitful joint work together with them and Elisa Kasbohm.

5 References

[1] Bordewich, M., Semple, C. and Steel, M. (2006). Identifying X-trees with few characters. *Electronic Journal of Combinatorics* 13 #R83

[2] K. Huber, V. Moulton and M. Steel. (2005). Four characters suffice to convexly define a phylogenetic tree. *SIAM Journal on Discrete Mathematics* 18(4): 835-843.

[3] M. Steel and D. Penny. (2005). Maximum parsimony and the phylogenetic information in multi-state characters. Pp. 163-178 *In Parsimony, phylogeny and genomics* (ed. V. Albert), Oxford University Press.