Alignment-Ambiguous Regions of Genes

Introduction and Background

Thinking about alignment of multiple DNA sequences, what would you do if you find that there are some regions in the sequences which do not look like they could be aligned as there may be some ambiguity in the alignment? In most studies, these alignment-ambiguous regions are simply removed before analysis is carried out. In the article in the opinion section of TRENDS in Ecology & Evolution, December 2001 issue, the author, Michael S.Y. Lee, has raised and restressed the importance of proper study and analysis of these alignment-ambiguous regions (e.g. the repidly evolving regions of genes) in Molecular Phylogenetic and Evolutionary studies. Three promising methods have been suggested for analysis of such regions with examples. I think this topic is very interesting and could draw attention from physics and bioinformatics communities to develop the methods further and provide deeper insight in the fields.

Simple Case Study

Consider a simple example which will be used throughout the discussion, from Figure 1(a) there are six protein-encoding DNA sequences from six taxa (Outgroup, A, B,C, D and E). The first four sequences (Outgroup, A, B, C) have the same length. The last two sequences (D, E) are three base pairs shorter and there are two possible ways (Figures 1(b) and 1(c)) to align them with the first four sequences. So the positions 4-9 (Region Y) comprise "alignment-ambiguous region" of the genes which cannot be aligned unequivocally and therefore would be ignored from the common analysis.

Alignment is very simple when A, B and C (Group 1) are aligned separately from D and E (Group2) since each group has the same length. In this case, the alignment-ambiguous region contains "strong phylogenetic information" for both groups. By comparing with the Outgroup and making a Phylogenetic tree (Figure 1(d)), we can see that B and C are separated from A by "four substitutions" (Positions 4, 6, 8 and 9) in Group 1 whereas in Group 2 (D and E) are excluded from Group 1 by a "three-base pairs deletion" in the alignment-ambiguous region. So by ignoring Region Y from Phylogenetic analysis, the information about "Four Substitutions and Three-base pairs Deletion" will be lost.

To include the phylogenetic and evolution information related to the alignment-ambiguous region, Region Y has to be taken into account. The following three methods have been suggested i.e. Multiple Analysis Method, Elision Method and Fragment-Level Alignment.

1. Multiple Analysis Method

In Multiple Analysis Method, Multiple Sequence Alignments are performed for many different sensible values of relevant alignment parameters e.g. gap opening, gap extension, transitions and transversions. Then the corresponding Phylogenetic trees are drawn. The majority of the trees is selected as the final result. The downside of this method is that it is very time consuming and there could be a problem if there are more than one possible final tree.

2. Elison Method

In Elison method, a range of possible alignments of codons in the alignment-ambiguous region is generated. Then these possible alignments are combined into a single matrix with a large number of columns (as the possible alignments are connected one after another). So in this case the possible alignment of the alignment-ambiguous region will be directly taken into account.

In the example above, there are two possible alignments (Figure 1(b) and 1(c)). So the resultant matrix will look like what is shown in Figure 1(e). Then this matrix is analyzed as a single multiple alignment to generate the resultant phylogenetic tree.

The disadvantage of this method is that there may be too many possible alignments in the alignment-ambiguous region then they will be downweighted too much that they are effectively deleted. This also make alignment to be computationally very expensive.

3. Fragment-Level Alignment

Fragment-Level Alignment (or Fixed Character State or INtegration of Ambiguous Aligned SEquences (INAASE)) is method based on assumption that the alignment-ambiguous region can be treated as a single multistate character (called Y in the example above) with each distinct sequence variant treated as a separate character state (Figure 1(f)) e.g. state Y=1 for Outgroup and A; Y=2 for B and C; Y=3 for D and E. An associate step-matrix (Figure 1(g) can be constructed to show how each state is related based on the changes (substitutions and gaps) to convert one sequence variant to the other.

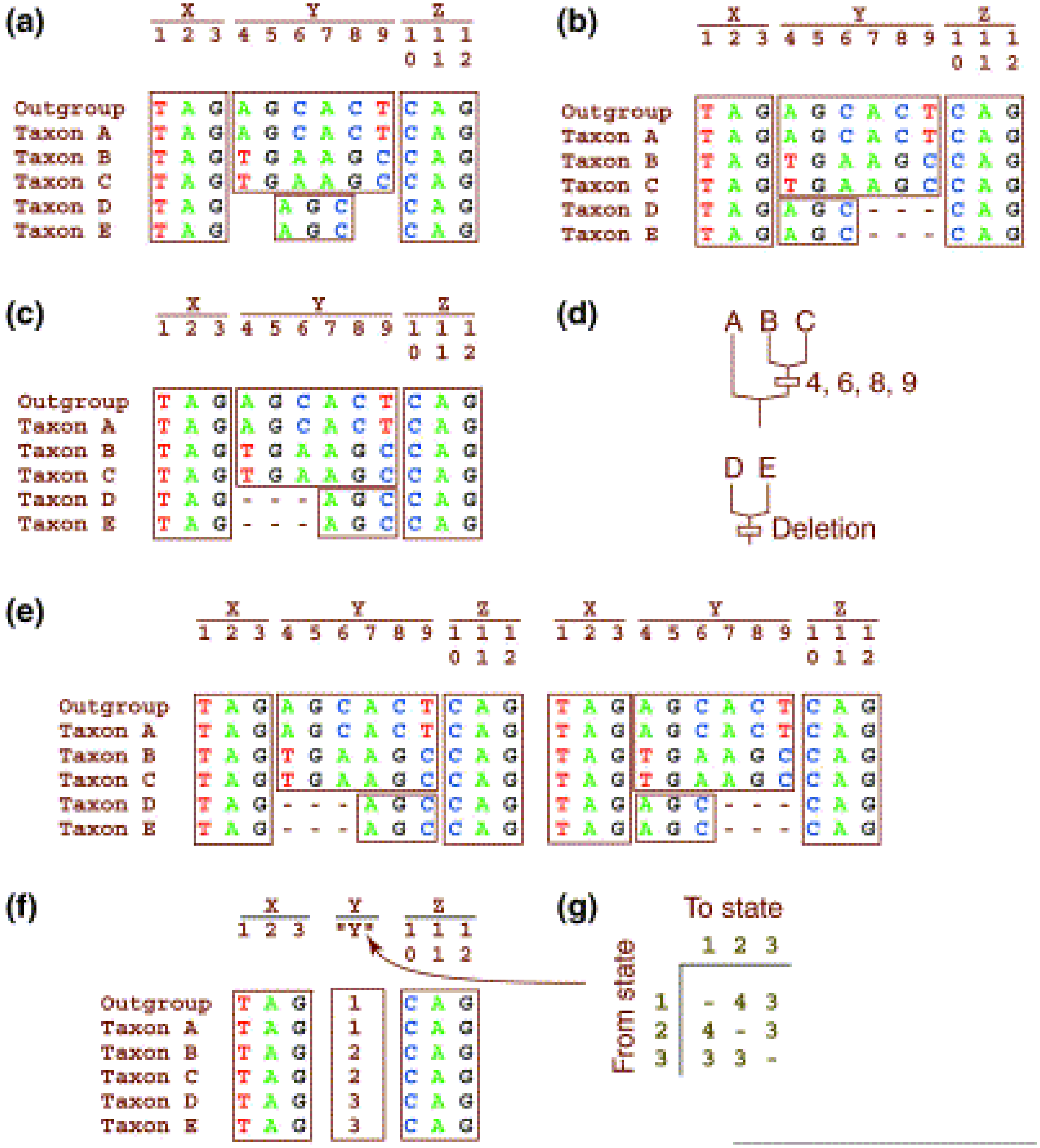
The methods allows transformation to occur and it chooses the one that is most consistent with the globally optimal one. The alignment is amplified as there are more alignment-ambiguous regions. On the other hand if there are too many states in resultant multistate character then some current phylogenetic programs may not be able to handle the problem. So it seems to be more sensible to only apply this method to alignment-ambiguous regions and analyze well-aligned regions by standard methods rather than applying it to the whole sequence.

Summary

Multiple Analysis is currently the most feasible way to take care of the ambiguous regions. Even though all methods are computationally very expensive but considering the phylogenetic and evolution significance and the effort required to obtain the sequence data from experiments it is worth considering analysis of such ambiguous regions.

Bibliography

Lee, M. S. Y. (2001) **Unalignable Sequences and Molecular Evolution**. *TRENDS in Ecology & Evolution* **16**, 681-685



TRENDS in Ecology & Evolution

Figure 1