Phylogenies based on gene content

Matthew Spencer,
Department of Mathematics and Statistics and
Department of Biochemistry and Molecular Biology,
Dalhousie University, Halifax, Nova Scotia, B3H 3J5, Canada.
matts@mathstat.dal.ca

Outline

1. Why is it difficult to infer deep phylogenetic relationships?
    - Saturation: nucleotide and amino acid sequences may have undergone so many changes that they are no longer informative about evolutionary relationships
    - Paralogy: genes may exist in multiple copies. If different copies are deleted in different lineages, the relationships among the surviving copies may not reflect the species relationships
    - Lateral gene transfer: Different genes may have different evolutionary histories
2. Gene content may provide useful phylogenetic information in these situations
    - Deletion, duplication and transfer mean that species have different sets of genes
    - Closely-related species tend to have similar gene content, while distantly-related species show large differences
    - We could use differences in gene content to measure evolutionary distance
3. What are the problems?
    - Unobservable data. If a gene family is absent from all species in our database, we may not know that it exists. This means we are likely to underestimate the number of gene families absent in both members of a pair of species.
    - Multiple changes. One difference in presence/absence state or in the number of members of a gene family may represent multiple events. We will tend to underestimate evolutionary distances.

4. What methods have been used?

    - Parsimony
    - Naive distances
    - SHOT
    - paralinear distances
    - Huson and Steel
    - Gu and Zhang
    - models with multi-gene events

5. Recommendations

    - Don't use parsimony: known to be unreliable, doesn't deal with multiple changes. We don't know how to weight gene gain and loss events
    - Don't use naive distances: they aren't appropriate measures of evolutionary distance
    - The distance measure used by SHOT is acceptable for quick analyses
    - Paralinear distances have the potential to deal with variation in expected genome size, which is not well modelled by other current methods. But the properties of paralinear distances for gene content data are not yet well understood.
    - Model-based distances (e.g. Huson and Steel, Gu and Zhang, multi-gene events) can tell us more about the biological processes affecting gene content.

6. Areas for future research

    - Better models for gene content, e.g. rate variation across the tree and among gene families
    - Full maximum likelihood methods?
    - Network methods to deal with lateral gene transfer