**Algorithms and Complexity in Phylogenetics**

**Magnus Bordewich**

**Origins**

Understanding and determining the evolutionary relationships between species has been a central goal of biology since Darwin first developed the Theory of Evolution.

This page from his 1837 notebook shows one of the first pictures of a ‘phylogenetic tree’.

The text reads:

> "I think <<diagram>>

Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction. Thus genera would be formed. –

**Comparing trees**

Analysis using different methods, or different data sets, can lead to contradicting trees on one set of species. We need appropriate methods of comparing trees and measuring the discrepancies. Subtree prune and Regraft (SPR) distance is a biologically motivated measure of tree discrepancy.

**Complexity**

Unfortunately, it transpires that determining the SPR distance is NP-complete [2]. This suggests that it is computationally intractable, i.e. no algorithm can efficiently determine the distance between an arbitrary pair of trees, regardless of the techniques applied.

**Fixed Parameter Tractability**

Despite this negative result, further analysis of the combinatorial structures underlying SPR distance has allowed the development of fixed parameter algorithms [2,3]. Such an algorithm runs efficiently when a given parameter is known to be relatively small, even when the overall input size may be large. In this case we can efficiently determine the SPR distance between two trees which are known to differ in a limited number of places.

**Approximation Algorithm**

An alternative approach to dealing with NP-completeness is to try to find approximate solutions. Due to the understanding of the common structures between trees that has enabled us to develop an approximation algorithm [3]. This approach runs quickly on an arbitrary pair of input trees, and guarantees that the result is close to the correct answer. Our 3-approximation gives the best guarantee currently known.


**Using the information**

Beyond establishing the ancestral relationships between species, phylogenetic tree also enable us to learn more about the nature of evolution and make quantitative evaluations of biodiversity.

**Hybridization**

A reticulation event, e.g. a hybridization or a horizontal gene transfer (HGT), causes the DNA sequences from transferred genes to have a different evolutionary history (i.e. phylogenetic tree) to the rest of the genome (see right). Conversely, comparison of the trees associated with different genes can reveal how much hybridization has occurred in the past, and between which ancestral species.

As with SPR distance, it is NP-complete to determine the minimum amount of hybridization required to explain the differences between two trees [4]. Again, it is possible to develop fixed parameter algorithms [5], and these have been implemented and tested on a database of grasses [6].

**Nature Reserve Selection**

There is great public and political interest in conserving the world’s ‘biodiversity’. Although the intuitive idea of biodiversity is easy to grasp, it can be difficult to formalize and quantify the biodiversity of a region, or a set of species. Furthermore, it is not clear how to best select a set of species or regions to be the focus of resources and efforts.

Phylogenetic trees can provide a concrete and logical way to quantify biodiversity [7]. This allows us to examine the complexity of the selection problem, and develop algorithms to solve or approximate them in various circumstances [8,9].

**Finding good trees**

DNA and protein sequencing has provided an immense amount of data to work with. The algorithmic challenge is to efficiently turn this data into useful information. In Phylogenetics we are trying to find an evolutionary tree that best fits the data.

One technique is to compute a pairwise distance matrix for the set of species being examined, from the number of discrepancies in the sequences. This matrix can be used for a measure of fitness of a proposed tree and the data.

‘A hill climbing’ algorithm then adjusts the tree to continually improve the fit. But when does this process lead to the best possible tree?

By analyzing the mathematics of the fitness measure and the combinatorics of the trees we can prove bounds on the data accuracy required to lead to the correct tree (1).


