Molecular Biology

DNA: Double stranded helix, A,C,G,T alphabet, A-T, C-G, has many different features, e.g. coding regions and junk DNA, mtDNA, stores genetic blueprint

RNA: Single stranded, A,C,T,U alphabet, copied from DNA, used in building proteins and enzymes, (rRNA, mRNA, tRNA)

Amino acids: Triplets of consecutive residues in RNA/DNA (called codons) define one of 20 different amino acids, using the "genetic code"

Proteins: Strings of amino acids, basic building blocks of cells

Gene: A contiguous stretch of DNA that encodes one protein

Chromosomes: Long strands of DNA

Genome: Complete set of all chromosomes inside a cell

Tree Reconstruction Problem

Given DNA sequences for a set of different species, infer the evolutionary tree along which the species evolved.

A Simple Stochastic Model of Evolution

- Given a *model tree* T, with edge weights $P(e)$

- Interpret $P(e)$ as the prob. of change at any given position in a sequence along the edge e

- Fix a sequence length k. Choose a root r and a random start sequence S at r

- Evolve sequences along the tree T, *Markov tree*.

Phylogenetic Tree Reconstruction

- The sequences at the leaves of the model tree T represent existing species

- Sequences at internal nodes are (lost) ancestors

- Assuming something is known about the underlying model of evolution, can we recover the tree T and/or the edge weights?

- The above model is unrealistic, e.g.: *rates across sites*, *transition matrices depending on e*, *sites that turn "on" and "off"* and more

Theorem (Steel) For certain simple models of evolution, T can be uniquely recovered (given infinite sequences).

Problem *Given DNA sequences for a set of different species, infer the evolutionary tree along which the species evolved.*

Daniel Huson, 1998

Some Tree Reconstruction Methods

- Maximum Parsimony
  - Popular sequence-based method
  - Solve the Hamming distance Steiner tree problem to obtain the *most parsimonious tree*. (NP-*hard*)

```
a:000010
b:100110
c:000011
d:110000
```

- Neighbor-Joining
  - Popular distance-based method
  - Successively "join" close pairs of species to infer tree. (*fast*)
- Buneman Tree
  - Distance-based method with nice mathematical properties (*low resolution*)

Maximum Parsimony

Find tree that explains data using a minimal number of *mutations*.

- For a given tree, find an optimal labeling (easy, using *Fitch's algorithm*)
- Look at *all possible* trees on given sequences (NP-hard)
- Use heuristics such as *branch-swapping*

Distance Based Methods

- Given a set of sequences, one can define a distance matrix using *Hamming distances*:

  $H(i,j)$:= # sites that sequence i and j differ

- $H(i,j)$ under-estimates true distances, due to *multiple hits* or *back mutations*

- Depending on the model of evolution, this can be corrected using a so-called *distance transformation*.

Additive Distances

A distance matrix D is called *additive*, if it comes from a tree.

$$\begin{pmatrix} 0 & 5 & 11 & 10 & 18 & 18 & 15 \\ . & 0 & 8 & 7 & 15 & 15 & 12 \\ . & . & 0 & 3 & 15 & 15 & 12 \\ . & . & . & 0 & 14 & 14 & 11 \\ . & . & . & . & 0 & 4 & 9 \\ . & . & . & . & . & 0 & 9 \\ . & . & . & . & . & . & 0 \end{pmatrix}$$

The Four Point Condition

Problem: Given a distance matrix, is it additive?

Let D be a distance matrix. The *four point condition* holds, if for every four species a, b, c, d: the two larger of $D(a,b) + D(c,d)$, $D(a,c) + D(b,d)$ and $D(a,d) + D(b,c)$ are *equal*.

Theorem (Buneman 1971) A distance matrix D is additive, iff

- the triangle inequalities hold, and
- the four point condition holds.

Accuracy and Consistency

- Accuracy: How well does a method recover the correct tree?
  - *False Positives*: wrongly inferred edges.
  - *False Negatives*: missed edges.
- Consistency: Given longer and longer sequences, does the result of the method converge to the correct tree?
  - Distance methods: yes
  - Maximum Parsimony: no

The Felsenstein Zone

Generalizing Tree Reconstruction

Tree reconstruction: approximate distance matrix D by an additive distance matrix D_{add}, i.e. a tree.

Idea: Use more general graphs to approximate D.

One such method: *Split decomposition* (Bandelt and Dress 1992)

Split graphs generalizes tree:

Split Decomposition

There are three different resolved topologies on any four species $a, b, c, d \subset X$:

A *tree* induces at most *one* such topology on each quartet a, b, c, d, e.g.:

A *splits graph* G is allowed to induce upto *two* such topologies, e.g.:

Applications

Split decomposition is useful for e.g. noisy data, hybrid data, population data, or more general distance data:

(Note: Planar split graphs are zonotopes!)

Mathematical Questions

- If the threshold w is large enough, and if the base method returns the true tree on each sub-problem, then the algorithm returns the true tree on the whole problem.

- For $\epsilon > 0$, determine an upper bound for the sequence length k required to recover every edge of length $\leq \epsilon$.