Phylogeny

Martin J Bishop

Cambridge, UK Milan, Italy

Giles Bishop was born in Horsley, Gloucester, England in1634

- Genealogy – 1000 yrs
- Phylogeny – 1000 million yrs

- Photograph of Sidney Bishop family c. 1912

Evolutionary inferences

ASSUMPTIONS

- nature of evolutionary processes; functional constraints
- evolutionary relationships; functional constraints
- evolutionary relationships; nature of evolutionary processes

INFERENCES

- evolutionary relationships of taxa
- nature of evolutionary processes
- functional constraints

Divergence time

- Primate relationships
  - recent 5-30 MY
- Orders of mammals
  - radiation in short period 70 MY ago
- Living organisms
  - very ancient divergences 1000 MY

Phylogenetic inference

- Tree model of relationships
- Observable genetic sequences
- Processes of genetic change
  - some are random
  - usually constraints e.g. CpG in mammals
  - some changes impossible (selection)

The tree model

- Genealogies describe the detail
- Inbreeding is assumed to be negligible
- Pathways of genetic transmission approximate to a tree
- There are huge numbers of possible pathways

Pre-sequencing era

- Taste polymorphism
- Cell surface polymorphisms
- Protein polymorphisms
- Immunology
- Chromosome banding
- DNA hybridisation

Taste polymorphism

20/27 chimpanzees found phenyl thiocarbamide (PTC) distasteful.

This dimorphism also exists in man.

Blood groups

- Humans have the ABO system
- Chimps have AO (O rare in other primates)
- Gorilla is B only

Major histocompatibility antigen

MHC region contains 1000 linked loci, many of which have up to 20 alleles

- Man/chimp considerable sharing
- Man/gorilla less shared

Chromosome banding

Staining techniques can visualise 1000 chromosome bands in man

- Man/chimp 13 identical chromosome pairs
- Man/gorilla 9
- Man/orang 8

Blocks of similar banding have been shuffled

Protein polymorphism

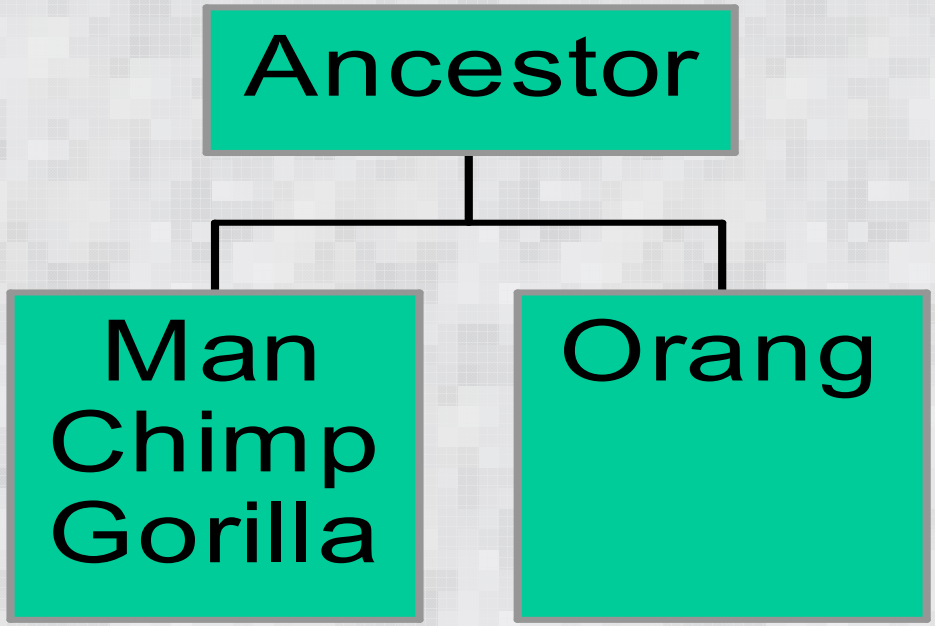
Detection of variants of proteins is possible by gel electrophoresis.

A study of 23 enzymes in man, chimpanzee and gorilla revealed very little difference.

The study was uninformative about phylogeny.

Immunology

Alan Wilson (1967) studied albumin microcomplement fixation



DNA hybridisation

Sibley & Ahlquist (1984, 1987)

This works by warming DNA in solution to separate the two strands and then measuring the rate of reannealing.

To measure similarity of DNA from two species use a mixture.

(((Man Chimp) Gorilla) Orang)

Evolutionary models for phylogenetic estimation

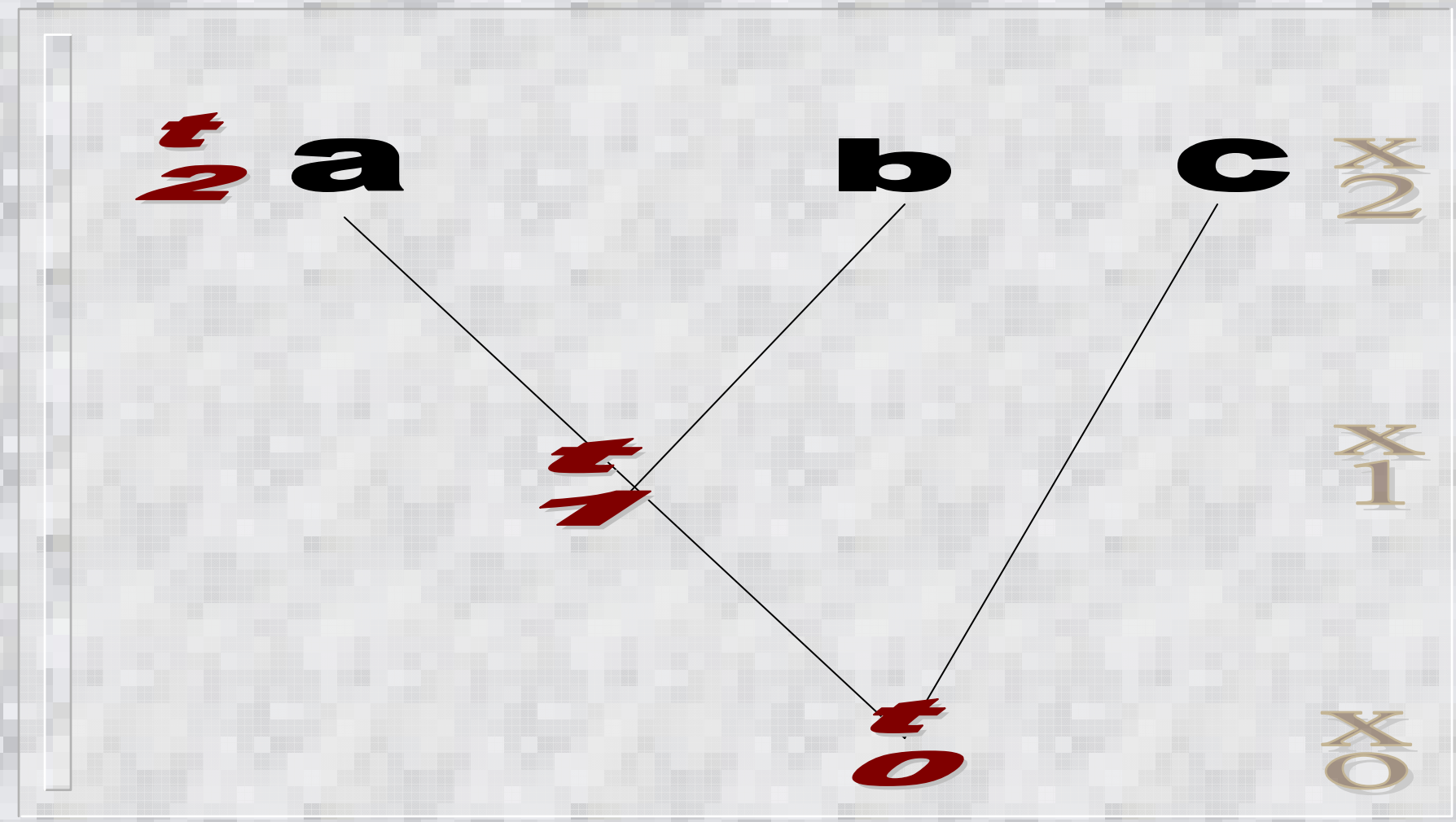
Components of the model

- Evolutionary tree model
  - rooted tree if “molecular clock” is assumed
  - unrooted tree if variable rate model
- Evolutionary process model
  - DNA sequence evolution
  - Protein sequence evolution
  - Chromosome segment evolution
  - Gene order evolution

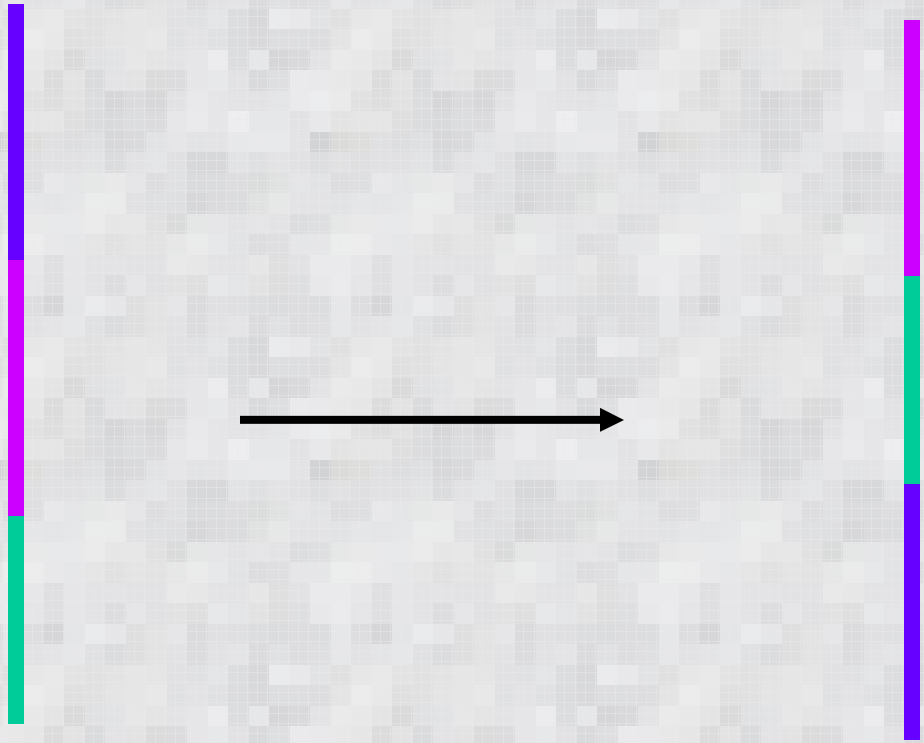
Evolutionary tree model

- Tips of the tree (leaf nodes) are the extant individuals that can be sampled
- Internal nodes are population divergences
- Arcs of the tree are the pathways of transmission of the genetic material
- Rooted tree is ordered in time
- Unrooted tree is scaled in character change space

An evolutionary tree



An evolutionary event



Probability

- System
- Conditions
  - Coin tossed by fair player
  - Coin tossed in ballistic machine
- Enumerate all know events
  - Coin lands heads
  - Coin lands tails
  - Coin lands on its side
  - Coin rolls down drain and outcome is indeterminate

Evolutionary inference

- Data
  - the observables
- Model
  - the parts of the system that are taken as given
- Hypothesis
  - one of all the possible outcomes

Likelihood inference

- Likelihood is the probability of an hypothesis given the data and the model
- Hypotheses may be compared on the basis of their likelihoods as a ratio
- More conveniently, hypotheses may be compared by the difference in their log likelihoods

$$L(H1|D) = k.P(D|H1)$$
$$L(H2|D) = k.P(D|H2)$$

$$L1/L2 = P(D|H1)/P(D|H2)$$

$$\ln L1 - \ln L2 = \ln P(D|H1) - \ln P(D|H2)$$

Maximum likelihood

- The hypothesis that has the greatest likelihood is called the maximum likelihood estimate
- For simple systems a maximum likelihood estimator can be determined analytically
  - e.g. the arithmetic mean is the maximum likelihood estimate under a normal probabilty distribution

$$\max L = \max P(D|HX)$$

Complex systems

- The maximum likelihood estimate can be obtained by iteration using a computer program as long as the model is properly constrained
- Models that are not properly constrained have no maximum likelihood estimate

Phylip package DNAMLK

- Evolutionary tree model is rooted
- Poisson process exponential failure model of DNA sequence change
- Stochastically constant rate of sequence evolution
- Independent evolution at each nucleotide site

Data do not fit the model?

- Allow stochastically variable rate of change on unrooted tree - DNAML
- Poisson process inappropriate? Replace with negative binomial model
- Nucleotide sites may change at different rates

What is wrong with parsimony?

- Model is rudimentary or absent
- No evolutionary time dependent model
- Cannot determine if the data fit the model or not
- Only considers one possible outcome rather than the sum of all possible outcomes

DNA sequence data

- Extremely heterogeneous
- Need to select regions that may be under the same selective constraints
- Only practical to compare small regions of the genome
- It is impractical to include all possible alignments in the ML estimate which a full model would require

Protein sequence data

- Very strong selective constraints
- Evolutionary processes conserve structure rather than sequence except for crucial functional sites
- Models of structural evolution have been rather unsatisfactory

Chromosome segment data

- Polytene chromosome banding patterns of Drosophila offer a rich source of data
- Models of chromosome segment evolution have been proposed
- Satisfactory estimates of phylogeny have been made using ML

Gene order data

- A huge amount of gene order data is being generated
- 1000 DNA RFLPs have been mapped in cereals
- Conserved blocks have been identified
- The phylogeny has been derived by ML (((Sorghum Millet)Sugar Cane)Rice)

An example: vertebrate phylogeny

- Study of vertebrate gene sequences and proteins has given a very confused picture
- E.g. birds cluster with mammals not within their place in the reptiles
- Gene conversion has edited sequences in gene familes
- Gene order data offer an alternative approach and selective pressures are very different