Phylogenetic Trees

Gabriella Trucco

Email: gabriella.trucco@unimi.it

Introduction

- Where do we come from?
- Systematics : detect and classify the diversity in the biological world
- Phylogeny of a species: a history of its own evolutionary development.
- Phylogenetic Systematics: organisms are classified into groups by their phylogeny

Introduction

- Carl Linné (1707-1778):
  - revolutionized the way in which species were classified and named. He proposed to group them by shared similarities into higher taxa, being: genera, orders, classes and kingdoms.
  - He also invented the ‘binomial system’ for naming species
  - Linné believed in invariant species.
  - Later on, he admitted that a certain variation was possible.
  - His most important works are Systema naturae (1735) and Genera plantarum (1737).
- Chevalier de Lamarck (1744-1829):
  - started as taxonomist in botany
  - applied Linné’s ideas also to animals: Philosophie zoologique (1809).
  - Lamarck was one of the first to believe in some kind of evolution.
  - his way of explaining the observed changes was not accurate

Introduction

- Georges Cuvier (1769-1832):
  - comparison of organisms, characterizing their differences and similarities.
  - studied fossils and observed changes in comparison with contemporary organisms.
- Charles Darwin (1809-1882): Here is an excerpt from the famous book On the origin of species by means of natural selection by Charles Darwin, 1859:

  Whatever the cause may be of each slight difference in the offspring from their parents—and a cause for each must exist—it is the steady accumulation, through natural selection, of such differences, when beneficial to the individual, that gives rise to all the more important modifications of structure, by which the innumerable beings on the face of this earth are enabled to struggle with each other, and the best adapted to survive.

Introduction

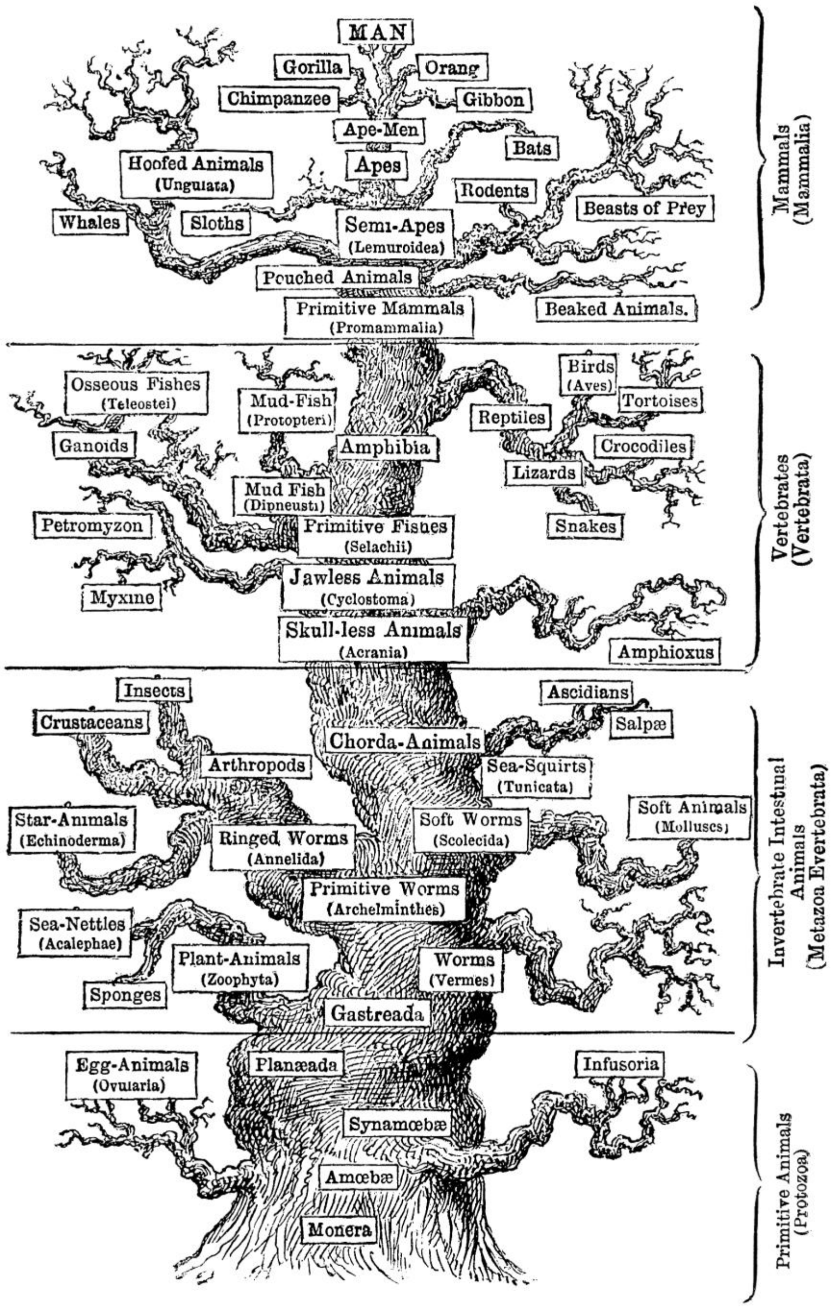
- Darwin was not the first who believed in some sort of evolution.
- Darwin was the one who was able to explain how this evolution could have occurred.
- Concept of evolution by natural selection:
  - The individual organisms in a population vary.
  - They overproduce (if the available resources allow).
  - Natural selection favors the reproduction of those individuals that are best adapted to the environment.
  - Some of the variations are inherited to the offspring.
  - Therefore organisms evolve.

Introduction

- Alfred Russel Wallace (1823-1913).
    - He and Darwin had similar ideas about natural selection
    - Wallace became a co-discoverer in the shadow of the more famous Darwin.
    - He explained the emergence of striking coloration of animals as a warning sign for adversaries and he devised the Wallace effect: Natural selection could inhibit hybridization and thus encourage speciation.
- Ernst Haeckel (1834-1919) did a lot of field work and is known for his “genealogical tree”

PLATE XV.

PEDIGREE OF MAN.



Introduction

- Emil Hans Willi Hennig (1913-1976) was specialized in dipterans (ordinary flies and mosquitoes).
  - morphological similarity of species does not imply close relationship.
  - phylogeny based systematic, stated corresponding problems, developed first, formal methods, and introduced an essential terminology
- Emil Zuckerkandl (1922) and Linus Pauling (1901-1994) were among the first to use biomolecular data for phylogenetic considerations.
  - In 1962, they found that the number of amino acid differences in hemoglobin directly corresponds to time of divergence.
  - molecular clock hypothesis in 1965

Examples in Epidemiology

- Understand the development of pandemics, patterns of disease transmission, and development of antimicrobial resistance or pathogenicity:
- Basler, C.F., et al. 2001. Sequence of the 1918 pandemic influenza virus nonstructural gene (NS) segment and characterization of recombinant viruses bearing the 1918 NS genes. *PNAS*, 98(5):2746-2751.
- Ou, C.-Y., et al. 1992. Molecular epidemiology of HIV transmission in a dental practice. *Science* 256(5060):1165-1171.
- Pradeep Kumar, N., et al. 2002. Genetic variability of the human filarial parasite, *Wuchereria bancrofti* in South India. *Acta Trop* 82(1):67-76.

Examples in Conservation Biology

- Determine which populations are in greatest need of protection, answer other questions of population structure:
- Trepanier, T.L., and R.W. Murphy. 2001. The Coachella Valley fringe-toed lizard (*Uma inornata*): genetic diversity and phylogenetic relationships of an endangered species. *Mol Phylogenet Evol* 18(3):327-334.
- Alves, M.J., et al. 2001. Mitochondrial DNA variation in the highly endangered cyprinid fish *Anaecypris hispanica*: importance for conservation. *Heredity* 87(Pt 4):463-473.

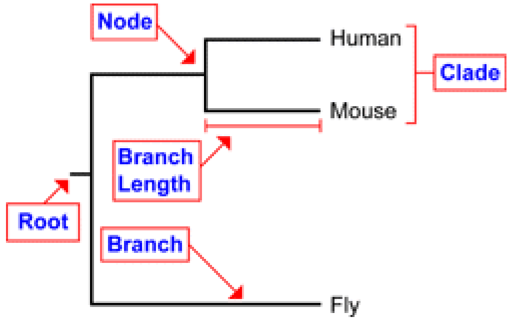
Examples in Pharmaceutical Research

- Determine which species are most closely related to other medicinal species, sharing their medicinal qualities:
- Komatsu, K., et al. 2001. Phylogenetic analysis based on 18S rRNA gene and matK gene sequences of *Panax vietnamensis* and five related species. *Planta Med* 67:461-465.

Examples in Forensic Science

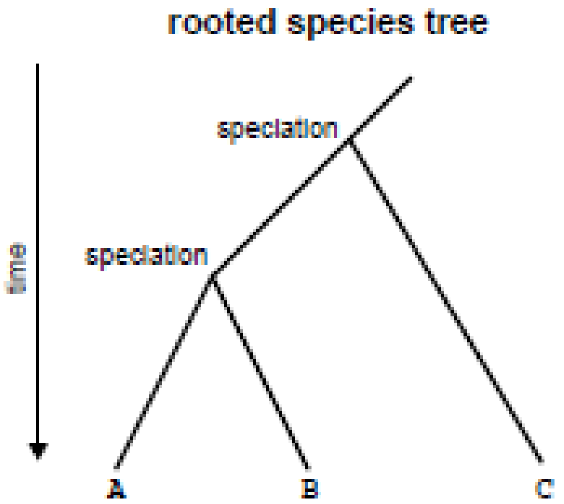
- Solve crimes, test purity of products, determine whether species have been smuggled or mislabeled:
- Vogel, G. 1998. HIV strain analysis debuts in murder trial. *Science* 282(5390): 851-853.
- Lau, D. T.-W., et al. 2001. Authentication of medicinal *Dendrobium species* by the internal transcribed spacer of ribosomal DNA. *Planta Med* 67:456-460.
- Metzker, M. L., et al. 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci USA* 99(22):14292-14297.

Terminology



- Phylogeny: "tree", which estimates the "historical" connections between species or genes that they carry.
- Parts of a phylogenetic tree include:
  - The "tips" of the tree branches represent the taxa in the study.
  - Taxa: orders, species, populations, etc.
  - OTUs, or Operational Taxonomic Units.
  - The lines within the tree are called the "branches".
  - The points at which branches connect, or the tips of branches, are both called nodes. Internal nodes connect branches; external nodes represent taxa.
  - Some trees will have a basal node, known as the "root".
  - A grouping of an ancestor and all of its descendants is known as a clade.

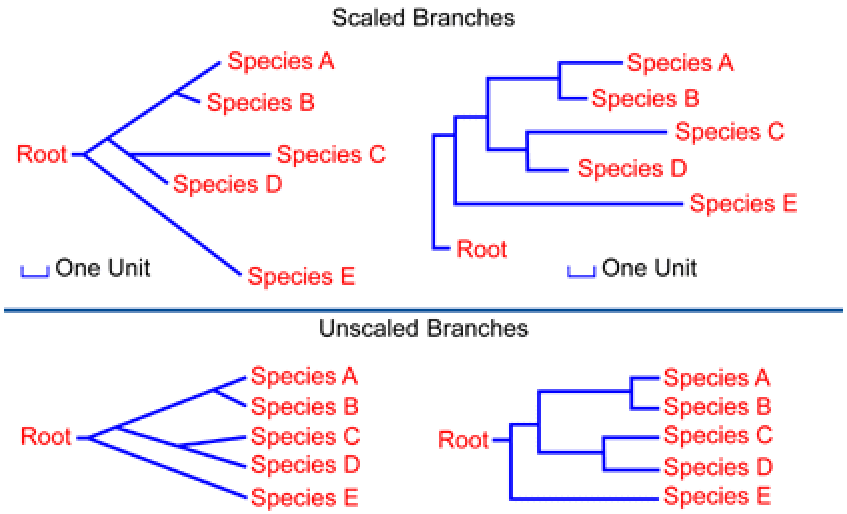
Phylogenetic tree



- Phylogenetic (also: evolutionary) trees display the evolutionary relationships among a set of objects.
- Usually, those objects are species
- contemporary species: represented by the leaves
- Internal nodes: represent the last common ancestor before a speciation event took place.
- The species at the inner nodes are usually extinct
- tree mostly based on the data of contemporary species.
- It models their evolution, showing how they are related

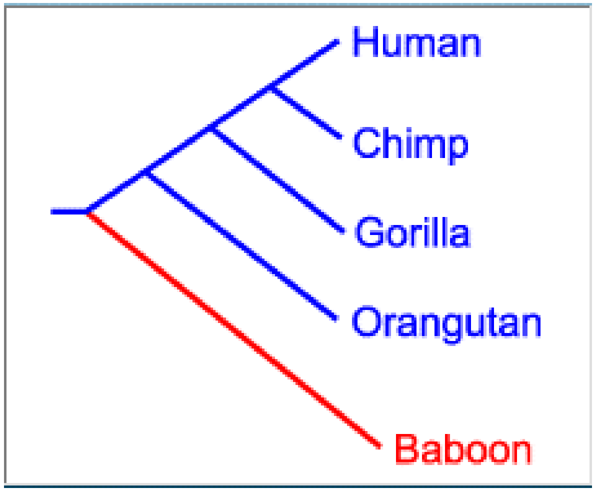
Scaled vs unscaled braches

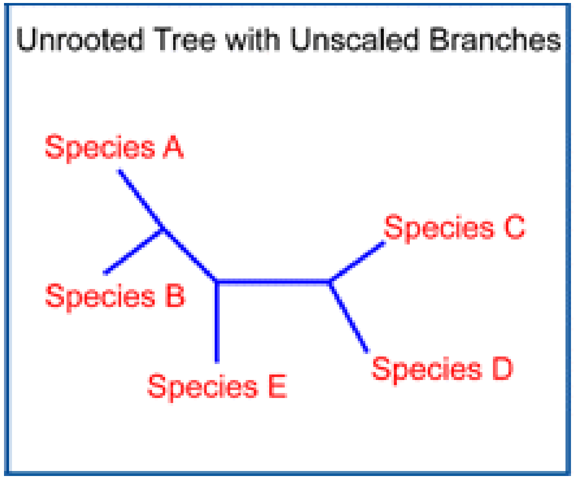
- Scaled branches - branches will be different lengths based on the number of evolutionary changes or distance.
- Unscaled branches - all branches in the tree are the same length.



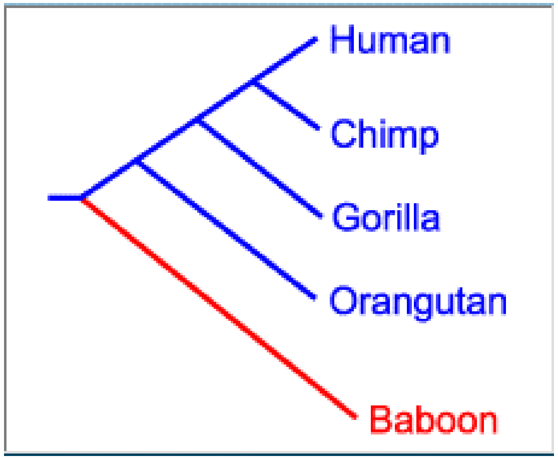
Rooted versus Unrooted Trees

- Rooted trees reflect the most basal ancestor of the tree in question.
- Unrooted trees do not imply a known ancestral root.





Speciation



- Speciation: the origin of a new species.
- Speciation event always linked to a population of organisms, not to an individual.
- a group of individuals emerges that is able to live in a new way
- After the separation of the two populations, both will diverge from each other during the course of time.
- The last common ancestor of the two will usually be extinct today
- "Species" trees recover the genealogy of taxa, individuals of a population, etc.
- Internal nodes represent speciation or other taxonomic events

Characters and States

- Given a group of species and no information about their evolution, how can we find out the evolutionary relationships among them?
- Find certain properties of these species, where the following must hold:
  - Decide if a species has this property or not.
  - Measure the quality or quantity of the property (e.g., size, number, color).
- These properties are called characters. The actual quality or quantity of a character is called its state.

Definition. A *character* is a pair C = $(\lambda; S)$ consisting of a property name λ and an arbitrary set S, where the elements of S are called character states.

Examples

- The existence of a nervous system is a binary character.
- The number of extremities (arms, legs,...) is a numerical character. Character states are elements of N.
- Here is an alignment of DNA sequences:

  Seq1: A C C G G T A

  Seq2: A G C G T T A

  Seq3: A C T G G T C

  Seq4: T G C G G A C

A nucleotide in a position of the alignment is a character. The character states are elements of {A,C,G,T}.

- The definition given for characters and states is not restricted to species. Any object can be defined by its characters.
  - vector of characters

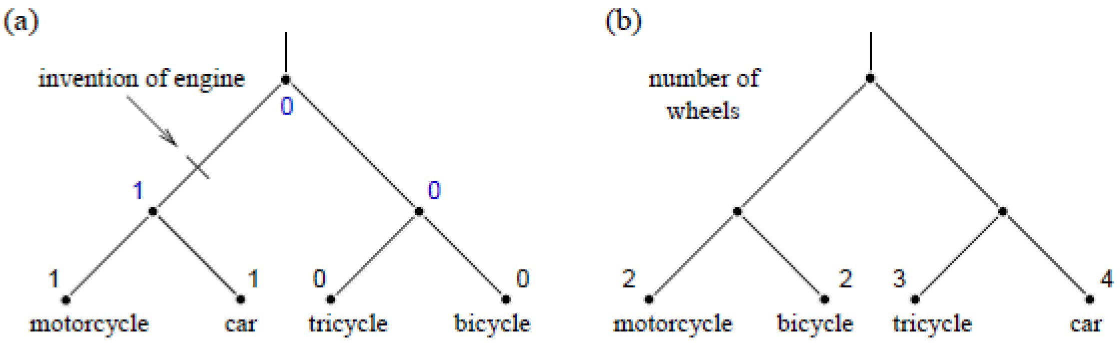
Examples

Bicycle, motorcycle, tricycle and car are objects. The number of wheels and the existence of an engine are characters of these objects. The following table holds the character states:

	# wheels	existence of engine
bicycle	2	0
motorcycle	2	1
car	4	1
tricycle	3	0

Compatibility

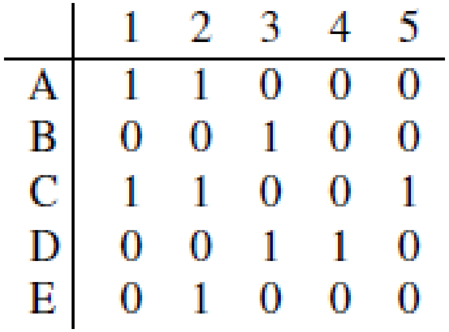
- Goal: find correct phylogenetic trees for the species under consideration.
- **Definition:** A character is *compatible* with a tree if all nodes of the tree can be labeled such that each character state induces one connected subtree.
- **Example:** Given a phylogenetic tree on a set of objects and a binary character c = {0; 1}, if the tree can be divided into two subtrees, where all nodes on one side have state 0, and 1 on the other, we count only one change of state.



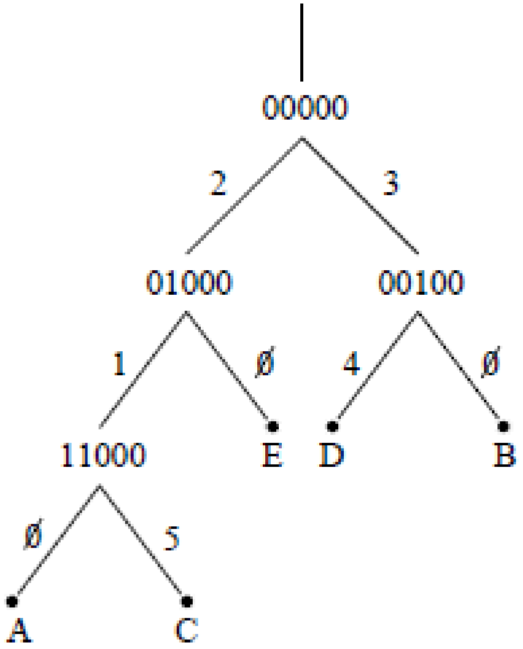
Perfect Phylogenies

- Let a set *C* of characters and a set *S* of objects be given.
- **Definition:** A tree T is called a *perfect phylogeny* (PP) for *C* if all characters in *C* are compatible with T.

- **Example:** The objects {A;B;C;D;E} share five binary (two-state) characters. The matrix M holds their binary states:



	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0



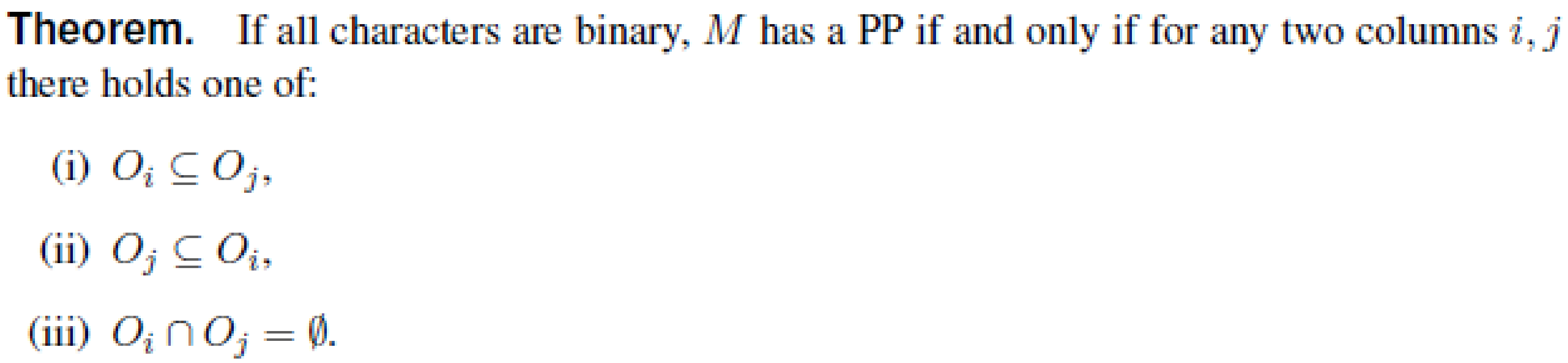
Perfect Phylogenies

- Object set for each character: character i corresponds to the set of objects O_i where the character i is “on”. Then each column of the matrix M corresponds to a set of objects: O_1={A;C}; O_2={A;C;E}; O_3={B;D}; O_4={D}; O_5={C}.
- The Perfect Phylogeny Problem (PPP) addresses the question if for a given matrix M there exists a tree and, given its existence, how to construct it.
- An alternative formulation → Character Compatibility Problem: Given a finite set S and a set of splits (binary characters) {A_i;B_i} such that

$$A_i \cap B_i = \emptyset \text{ and } A_i \cup B_i = S$$

is there a tree that realizes these splits?

Perfect Phylogenies



Theorem. If all characters are binary, M has a PP if and only if for any two columns i, j there holds one of:

(i) $O_i \subseteq O_j$,

(ii) $O_j \subseteq O_i$,

(iii) $O_i \cap O_j = \emptyset$.

- The condition in the above theorem describes the compatibility of two characters. Two characters are compatible if they allow for a PP.
- Gusfield’s theorem becomes: “A set of binary characters has a PP if and only if all pairs of characters are compatible.”

Perfect Phylogenies

- The algorithm to check if a PP exists is to test all pairs of columns for the above conditions which has time complexity O(nm^2), where n is the number of rows, and m is the number of columns in M.
- We will present here a more sophisticated method for recognition and construction of a PP with a time complexity of O(nm).
- 1. Check the inclusion/disjointness property of M:

  a) Sort the columns of M by their number of ones. (Note that this can also be done in O(nm) time.)

  b) Add a column number 0 containing only ones.

  c) For each '1' in in the sorted matrix, draw a pointer to the previous '1' in the same row.
- **Observation.** M has a PP if and only if in each column, all pointers end in the same preceding column.

If the condition is fulfilled, we can proceed constructing a PP.

Perfect Phylogenies

- 2. Build the phylogenetic tree:

  a) Create a graph with a node for each character.

  b) Add an edge (i; j) between direct successors in the partial order defined by the O_i set inclusion as indicated by the pointers. If the pointers of column j end in column i < j, draw an edge from i to j. Because of the set inclusion/disjointness conditions, the obtained graph will form a tree.

- 3. Refine and annotate the tree:

  a) Annotate all leaf nodes labeled j with the object set Oj , mark these objects as ‘used’, and annotate the parent edge with j.

  b) In a bottom-up fashion, for each inner node which is labeled with character j and whose children are all re-labeled, do

    i. Re-label the node with all objects of Oj which are not yet marked as ‘used’,

    ii. mark these objects as ‘used’, and

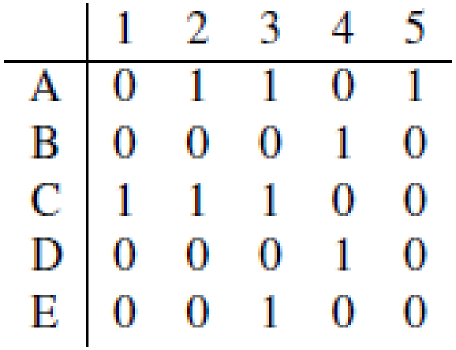
    iii. annotate the parent edge with j (not for the root node).

  c) For each node u and each object o it is labeled with, if u is not a leaf labeled only with o, remove o from the labeling and append a new leaf v labeled with o.

Perfect Phylogenies

- Finally, we obtain a PP where the edges are annotated with the invented characters and the leaves are labeled with the objects.

- Example:



	1	2	3	4	5
A	0	1	1	0	1
B	0	0	0	1	0
C	1	1	1	0	0
D	0	0	0	1	0
E	0	0	1	0	0