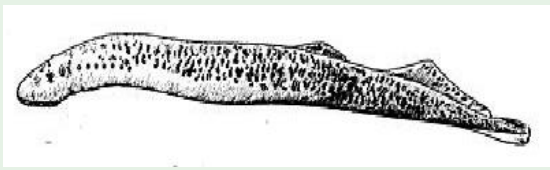
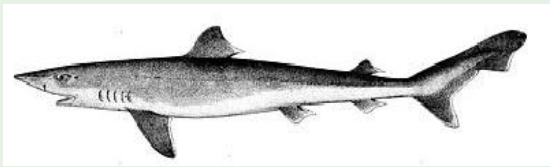
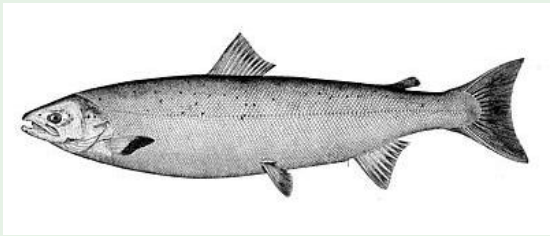
Definition

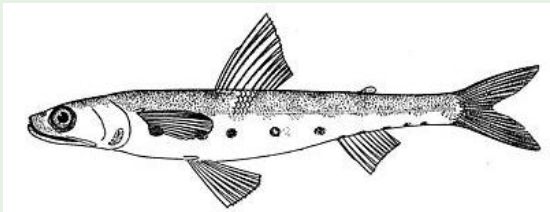
Let M be a boolean matrix with n rows and m columns that represents n objects in terms of m characters describing them, meaning that $M(i,j) = 1$ if and only if object i has character j.

Example









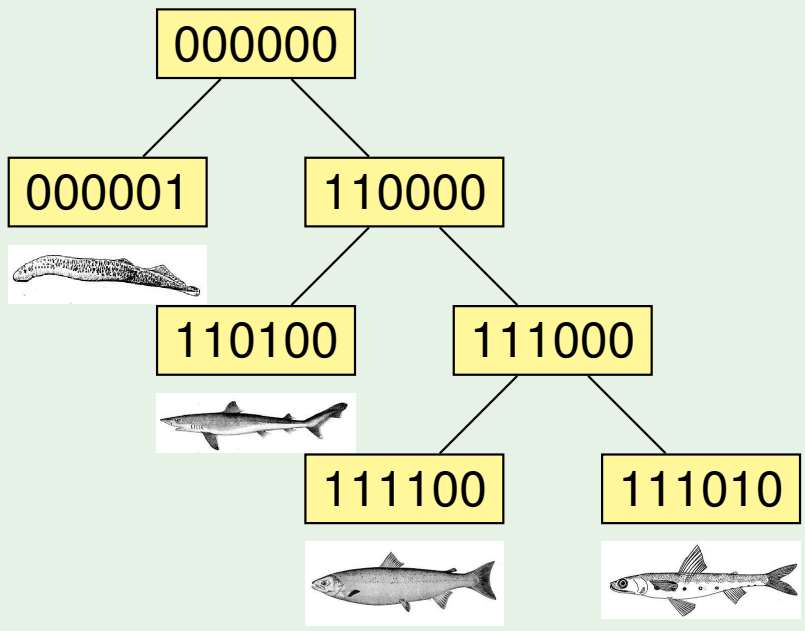
	paired fins	jaws	large dermal bones	fin rays	lungs	rasping tongue
lamprey	0	0	0	0	0	1
shark	1	1	0	1	0	0
salmon	1	1	1	1	0	0
lizard	1	1	1	0	1	0

Definition

A phylogenetic tree for M is a rooted tree T with n leaves (one for each object) and m of the branches associated with each of the characters, such that for any leaf x, the characters associated with the branches along the unique path from the root to x correspond to the character vector for object x.

Example

	paired fins	jaws	large dermal bones	fin rays	lungs	rasping tongue
lamprey	0	0	0	0	0	1
shark	1	1	0	1	0	0
salmon	1	1	1	1	0	0
lizard	1	1	1	0	1	0



Remark

A phylogenetic tree gives an estimate of the evolutionary history (in terms of branching) of the objects, based on the biological assumption that the state of each character is 0 in the ancestral object, and that there are no back mutations (each of the characters changes from the 0 state to the 1 state exactly once).

Remark (Biological interpretation in cladistics)

- n taxa
- m cladistic characters
- two states, 0 (absent) and 1 (present)
- unordered

Remark (Biological interpretation in genomics)

- n sequences
- m sites, possibly SNP sites
- two states, 0 (base) and 1 (mutant)
- ordered (on the chromosome)

Example

The alcohol dehydrogenase enzyme is one of the most abundant proteins in Drosophila melanogaster, and it is encoded by a single gene.

The alcohol dehydrogenase gene was studied on a sample of eleven species from five natural populations of Drosophila melanogaster, and the sampled sequences contain 44 polymorphic (segregating) sites.

```
1  CCGCAATAATGGCGCTACTCTCACAATAACCCACTAGACAGCCT
2  CCCCAATATGGGCGCTACTTTCACAATAACCCACTAGACAGCCT
3  CCGCAATATGGGCGCTACCCCCCGGAATCTCCACTAAACAGTCA
4  CCGCAATATGGGCGCTGTCCCCCGGAATCTCCACTAAACTACCT
5  CCGAGATAAGTCCGAGGTCCCCCGGAATCTCCACTAGCCAGCCT
6  CCCCAATATGGGCGCGACCCCCCGGAATCTCTATTCACCAGCTT
7  CCCCAATATGGGCGCGACCCCCCGGAATCTGTCTCCGCCAGCCT
8  TGCAGATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
9  TGCAGATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
10 TGCAGATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
11 TGCAGGGGAGGGCTCGACCCCACGGGATCTGTCTCCGCCAGCCT
```

Example

Under the infinite sites assumption, by which mutations are rare enough to discard the possibility of more than one mutation to occur at the same site in a sample of sequences, no site of a sample can contain more than two different nucleotides.

The most frequent nucleotide along a site is often taken as the base, with the least frequent nucleotide being taken as the mutant.

The base and mutant nucleotides for each site of the previous sequences are as follows.

```
CCCCAATAAGGGCGCGACCCCCCGGAATCTCTATTCGCCAGCCT
TGGAGGGGTTTCGTATGTTTTAACAGTAACGCCCCAAAGTATTA
```

Example

This allows for a binary representation of a sample of sequences, where the base nucleotide in each site is encoded as 0 and the mutant nucleotide is encoded as 1.

```
1  00100000010000010010101111011110101010100000
2  00000000010000001001110111011111010101010000
3  00100000100000010000000000000001010111000101
4  00100000100000011100000000000001010111011000
5  00111000001100101100000000000001010100000000
6  00000000100000000000000000000000000010000010
7  00000000100000000000000000000001010100000000
8  11011000001110000000000000000001010100010000
9  11011000001110000000000000000001010100010000
10 11011000001110000000000000000001010100010000
11 11011111000001000000010001000010101000000000
```

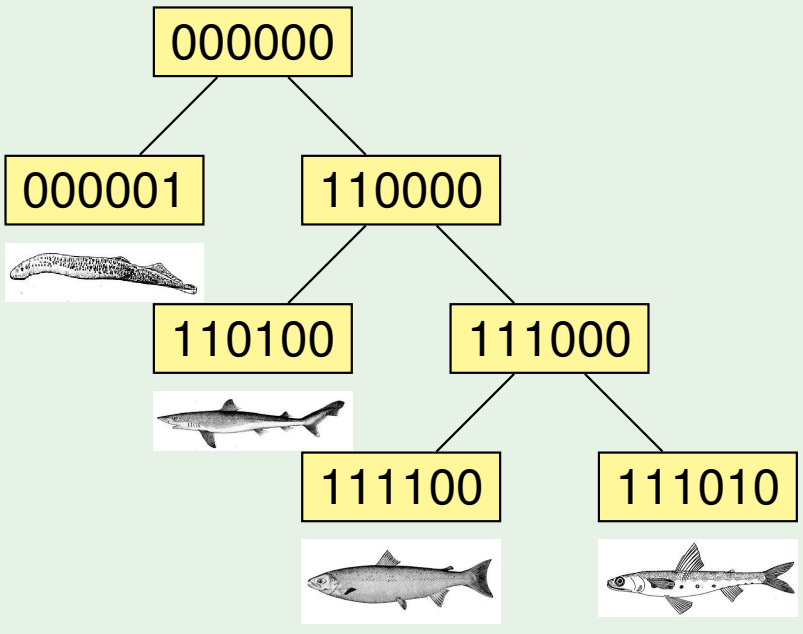
Definition

The set of objects which have character k is denoted by O_k. In a matrix M, O_k is the set of objects with a 1 in column k.

Example

lamprey	0	0	0	0	0	1
shark	1	1	0	1	0	0
salmon	1	1	1	1	0	0
lizard	1	1	1	0	1	0

$O_1 = \{shark, salmon, lizard\}$
$O_2 = \{shark, salmon, lizard\}$
$O_3 = \{salmon, lizard\}$
$O_4 = \{shark, salmon\}$
$O_5 = \{lizard\}$
$O_6 = \{lamprey\}$

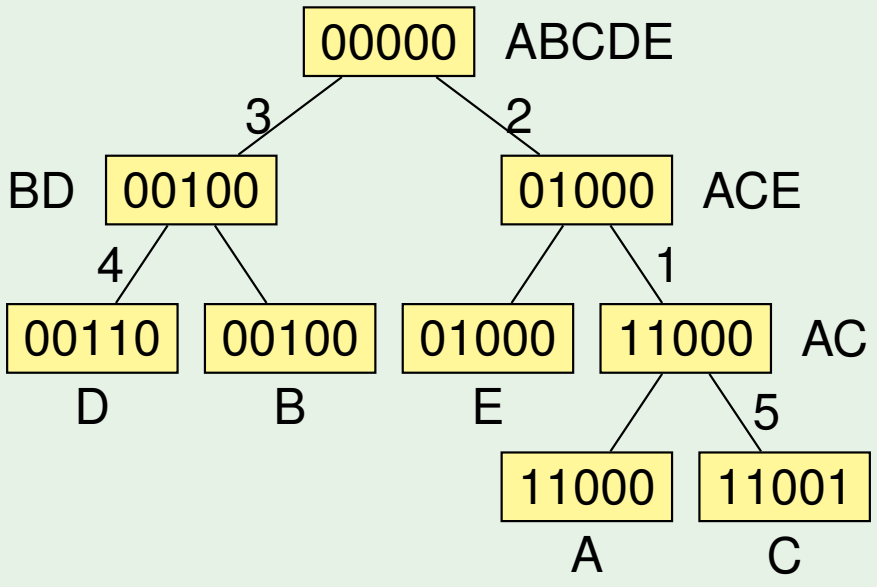


Example

A matrix M that has a phylogenetic tree.

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$O_1 = \{A, C\}$
$O_2 = \{A, C, E\}$
$O_3 = \{B, D\}$
$O_4 = \{D\}$
$O_5 = \{C\}$

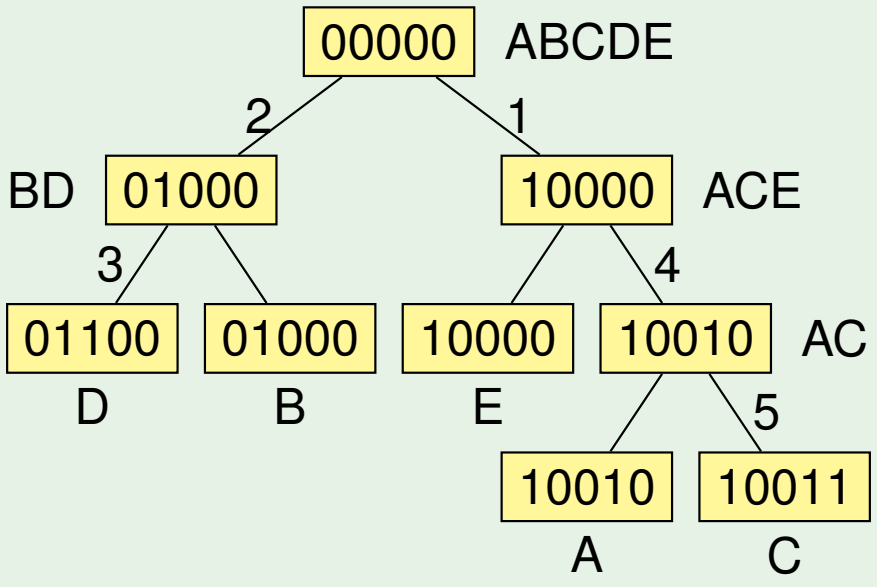


Example

Another matrix M that has the same phylogenetic tree.

$$M = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$O_1 = \{A, C, E\}$
$O_2 = \{B, D\}$
$O_3 = \{D\}$
$O_4 = \{A, C\}$
$O_5 = \{C\}$

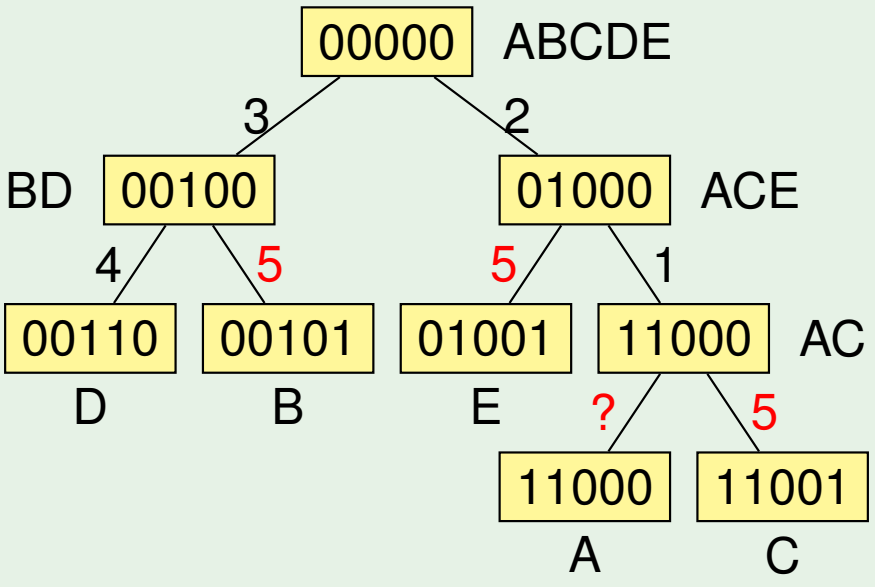


Example

A matrix M that has no phylogenetic tree.

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

$O_1 = \{A, C\}$
$O_2 = \{A, C, E\}$
$O_3 = \{B, D\}$
$O_4 = \{D\}$
$O_5 = \{B, C, E\}$



Lemma

M has a phylogenetic tree if and only if for every two columns i and j, either O_i and O_j are disjoint or one contains the other.

Proof.

Let M be a matrix that has a phylogenetic tree T, and let i and j be two columns such that $O_i \cap O_j \neq \emptyset$. Every object of $O_i \cap O_j$ must be a descendent in T of every object of O_i and also of every object of O_j, because of the biological assumptions. Thus, each object of $O_i \cap O_j$ would have two parents in T unless $O_i \subseteq O_j$ or $O_j \subseteq O_i$.

Conversely, let M be a matrix such that for every two columns i and j, either O_i and O_j are disjoint or one contains the other. Then, M defines a unique phylogenetic tree T with leaves labeled by each of the objects in M and one internal node for each non-trivial O_i. □

Definition

Two columns i and j in a matrix M are compatible if the set of objects O_i with a 1 in column i and the set of objects O_j with a 1 in column j are disjoint or one contains the other.

Corollary

M has a phylogenetic tree if and only if every two columns i and j are compatible.