

MA3259 Lecture 7

**Molecular Phylogenetic Analysis:
Part 1**

LX Zhang

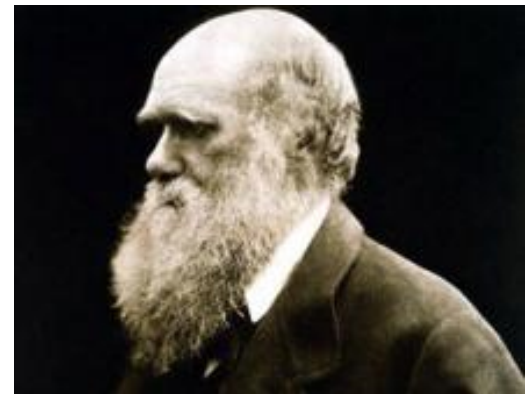
Department of Mathematics
National University of Singapore

matzlx@nus.edu.sg

Little Background

- Evolution is a basic subject in the modern biology, unifying together different fields such as genetics, microbiology.
- In 1859, Darwin published “On the origin of Species”

- Natural Selection
- Tree of life



The Tree of Life

- All living species are the modified descendants of the earlier species.
 - At the base is the common ancestor in the distant past, and out of it grows a trunk, which split again and again to create a large bifurcating tree. Each branches represents a single species; branching points are where one species becomes two. Most branches eventually come to a dead end as species go extinct. But some reach right to the top – these are living species.
 - Ever since Darwin, the tree concept has been the main principle for understanding of all living things.
- Then, how a new species is formed?
-- speciation events.



The “tree of life”
in Darwin’s notebook.

Nature Selection

- Evolution is driven by a process of natural selection.
 - All individuals struggle to survive, but some with small beneficial traits will have greater chance of survival (ecological selection) or reproducing (sexual selection)
 - Over ages, process of slow evolutionary change causes one species to evolve into another.
 - But a species splits into two in a speciation event.

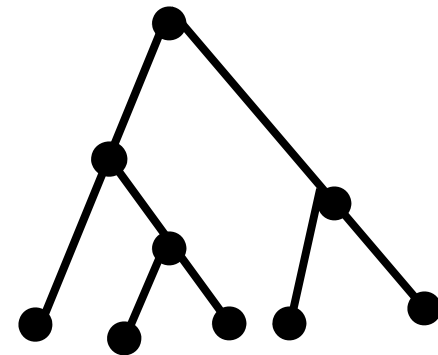
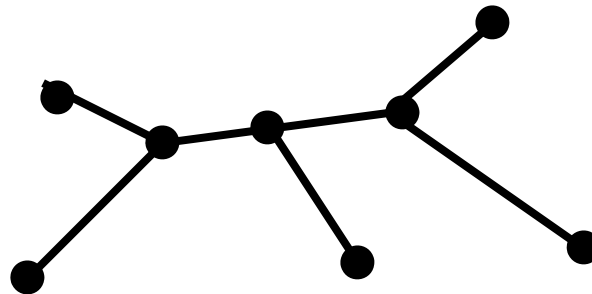
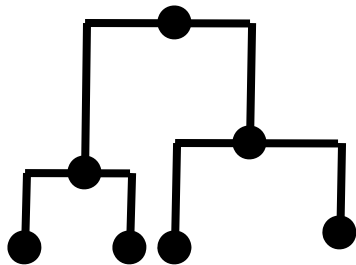


Molecular Evolution

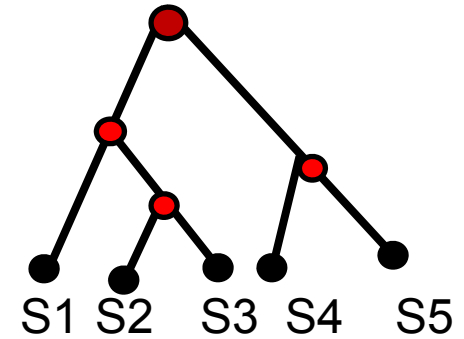
- Darwinian evolution and genetics produce the modern evolutionary biology.
- The mechanisms of inheritance began to be revealed in the start of 20th century.
- The structure of DNA was uncovered in 1953.
- As a blueprint of life, DNA must be the stuff in which the history of life was written.
 - From a common ancestral sequence, two DNA sequences are diverged. And each of the two sequences start to accumulate nucleotide substitution.
 - The more closely related two species are, the more similar their DNA ought to be.
- The molecular phylogenetic analysis is to prove a powerful tool. For instance, it has been used to
 - Study of human evolution.
 - Classification of giant panda
 - Tracing HIV infection

(Molecular) Phylogenetic Tree: Basic model of evolution

- It summarizes the evolutionary relationships (or differences) among a set of sequences
- A tree structure is composed of nodes and edges (or branches). Nodes represent the taxonomic units (genes, sequences).



Some concepts



Topology of a phylogenetic tree is the branching pattern of the tree, which is a tree in graph theory

- The degree of a node is the number of branches incident to it. It can be one, two, or three.
- Degree-1 nodes are **leaves**, usually having a label.
- Non-leaf nodes are **internal nodes**
- There is at most one node of degree 2, called the **root** if any.

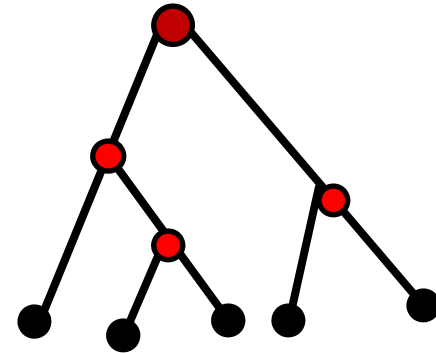
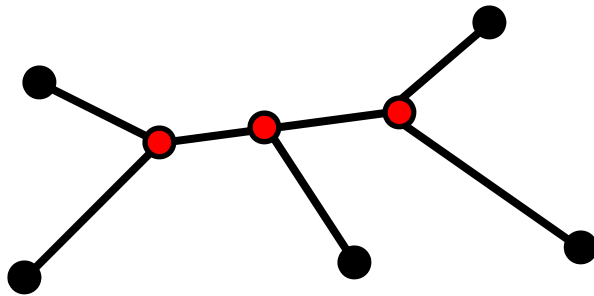
Leaves represent the sequences under comparison, called **operational taxonomic units** (OUTs) or tips

Internal nodes represent inferred ancestral sequences, called **hypothetical sequences**.

Unrooted and rooted phylogenetic trees

Unrooted phylogenetic tree, also called **phenogram**, is a tree in which

- all nodes represents related descendants,
- but there is not enough information on their common ancestor.



Rooted phylogenetic tree, also called **Cladogram**, is a tree in which

- there is a root, representing the common ancestor of the objects represented by leaves.
- The path from the root to a leaf represents an evolutionary path

Branch length and molecular clock

A branch from a node to its child often is assigned a **length** or **weight**, representing the number of mutations occurring in the corresponding course of evolution.

- Mutational events under consideration varies from study to study,
- The length of a branch may be converted into the evolutionary time with a molecular clock, i.e, mutational rate..

Molecular clock hypothesis

- All mutations occurs at the same rates in all the tree branches.
- The rate of the mutations is the same for all positions along a sequence.

Basic Mathematical Properties

Theorem 1: (a) Each unrooted phylogenetic tree of n leaves has $2n - 2$ nodes and $2n - 3$ edges for $n \geq 3$.

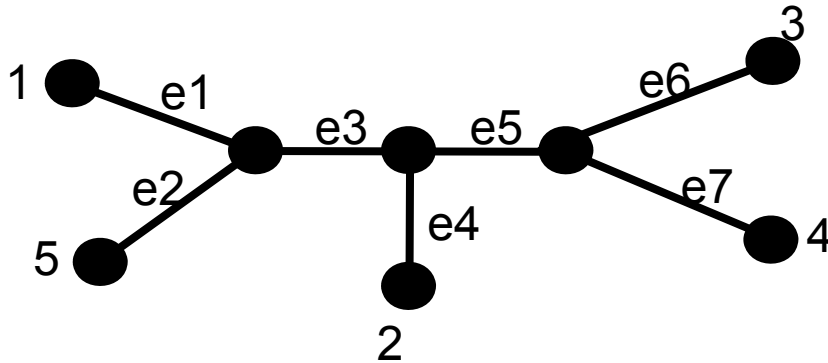
(b) Each rooted phylogenetic tree of n leaves has $2n - 1$ nodes and $2n - 2$ edges for $n \geq 3$.

Proof. (a) It is proved by induction on n .

(b) It is derived from the following facts:

Appending a leaf to the root of a rooted tree gives a
a unrooted tree with $n + 1$ leaves.

- Each branch in a unrooted phylogenetic tree defines a partition of the set of leaf labels



$$e1 \rightarrow \{1\}, \{2, 3, 4, 5\}$$

$$e2 \rightarrow \{5\}, \{1, 2, 3, 4\}$$

$$e3 \rightarrow \{1, 5\}, \{2, 3, 4\}$$

$$e4 \rightarrow \{2\}, \{1, 3, 4, 5\}$$

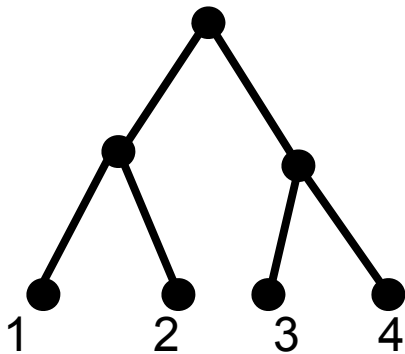
$$e5 \rightarrow \{1, 2, 5\}, \{3, 4\}$$

$$e6 \rightarrow \{1, 2, 4, 5\}, \{3\}$$

$$e7 \rightarrow \{1, 2, 3, 5\}, \{4\}$$

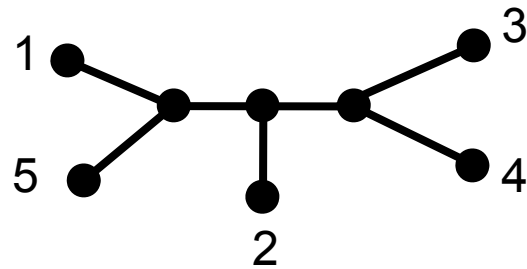
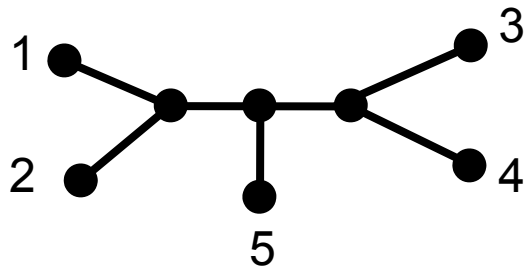
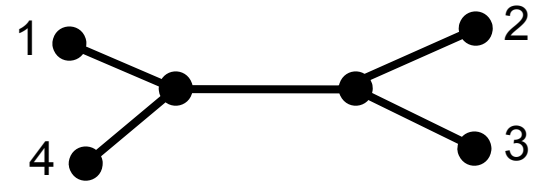
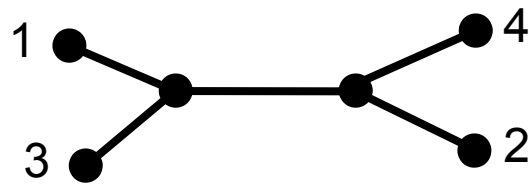
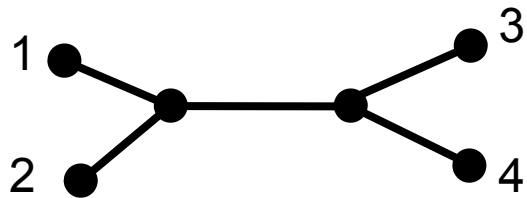
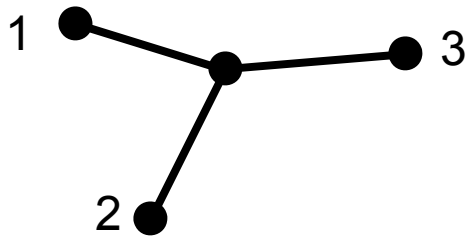
Two unrooted trees are identical if
Their edges induce same label partitions.

- Each node in a rooted phylogenetic tree defines a subset of leaf labels, composed of leaves below it. **So, rooted tree is also used in taxonomy.**



Two rooted trees are identical if
Their internal nodes induce same label subsets.

Unrooted or rooted phylogenetic trees have labeled leaves and unlabeled internal nodes



and 13 mores for
n=5

Theorem 2: (1) There are

$$(2n - 5)!! = 1 \times 3 \times 5 \times \cdots \times (2n - 5)$$

unrooted phylogenetic trees with unit-length branches, unlabeled internal nodes and n labeled leaves.

(2) There are

$$(2n - 3)!! = 1 \times 3 \times 5 \times \cdots \times (2n - 5) \times (2n - 3)$$

rooted phylogenetic trees with unit-length branches, unlabeled internal nodes and n labeled leaves.

There are $(2n-3)$ times more rooted trees than unrooted trees over n leaves

Number of all possible trees

| n | Unrooted | Rooted |
|----|------------|-------------|
| 3 | 1 | 3 |
| 4 | 3 | 15 |
| 5 | 15 | 105 |
| 6 | 105 | 945 |
| 7 | 945 | 10,395 |
| 8 | 10,395 | 135,135 |
| 9 | 135,135 | 2,027,025 |
| 10 | 2,027,025 | 34,459,425 |
| 11 | 34,459,425 | 654,729,075 |

Tree model is rich.

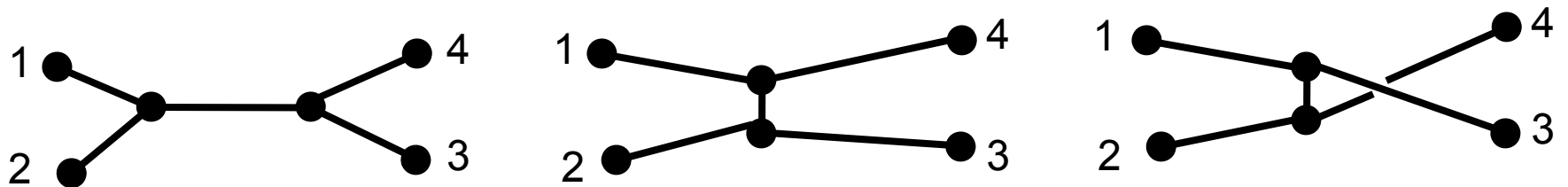
edges = branches

Sketch of Proof. (1) There is only one unrooted binary tree with 3 labeled leaves. Since such a tree has 3 edges, there are three possible ways by which we can add a new leaf. Thus, there are 3 unrooted binary trees with 4 labeled leaves. For each of these trees, there are 5 possible edges where we can add the next leaf. Continuing this procedure, we can see that the fact is correct.

(2). There are $(2n - 5)!!$ unrooted binary trees with n labeled leaves. Each such a tree can be rooted in $2n - 3$ different ways, each corresponding to an edge.

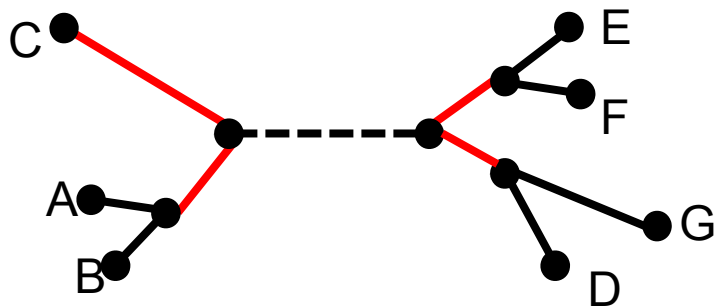
Enumerate and search tree space

There are 3 unrooted trees formed by grouping differently 4 leaves.



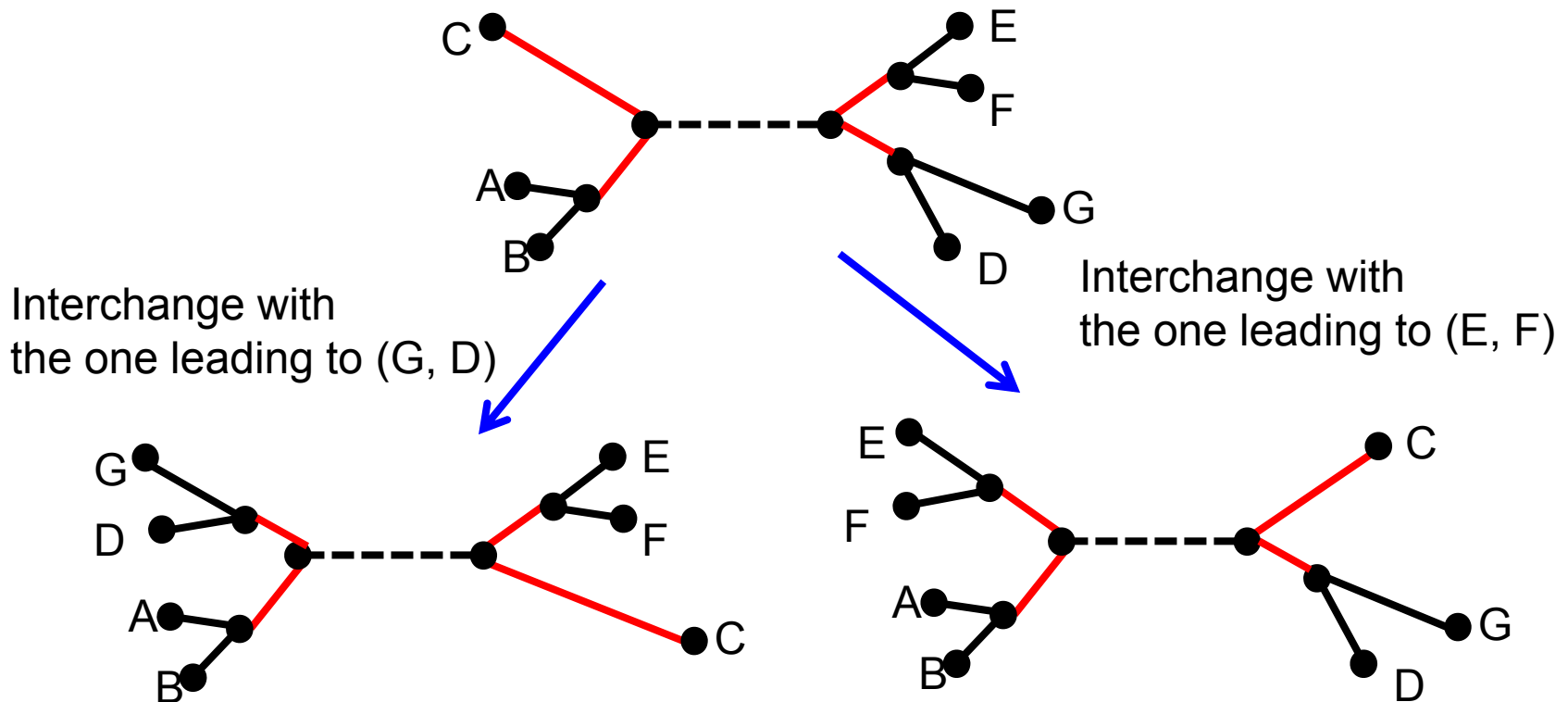
In general, each internal branch that connects two internal nodes is adjacent to 4 other branches (red in the example).

These 4 branches are called the **nearest neighbors** of each other.

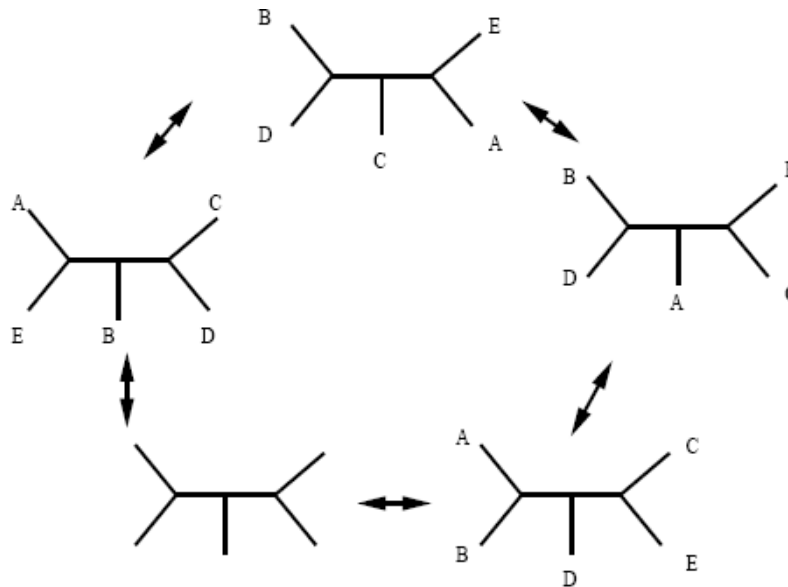


Nearest Neighbor Interchange(NNI)

Consider the branch leading to C. It is originally grouped with the nearest neighboring branch leading to (A, B). By interchange it with one of the other two nearest neighbors, we obtain different trees.



A tree can be transformed into any other in a series of NNI operations.



The space of phylogenetic trees with 5 leaves connected by NNI transformation

The **nni distance** $d(S, T)$ between two unrooted trees over the same set of leaves is the minimum number of NNI operations needed from transformation from one to another.

Theorem: For any two unrooted trees S and T of n leaves, their NNI distance is bounded as

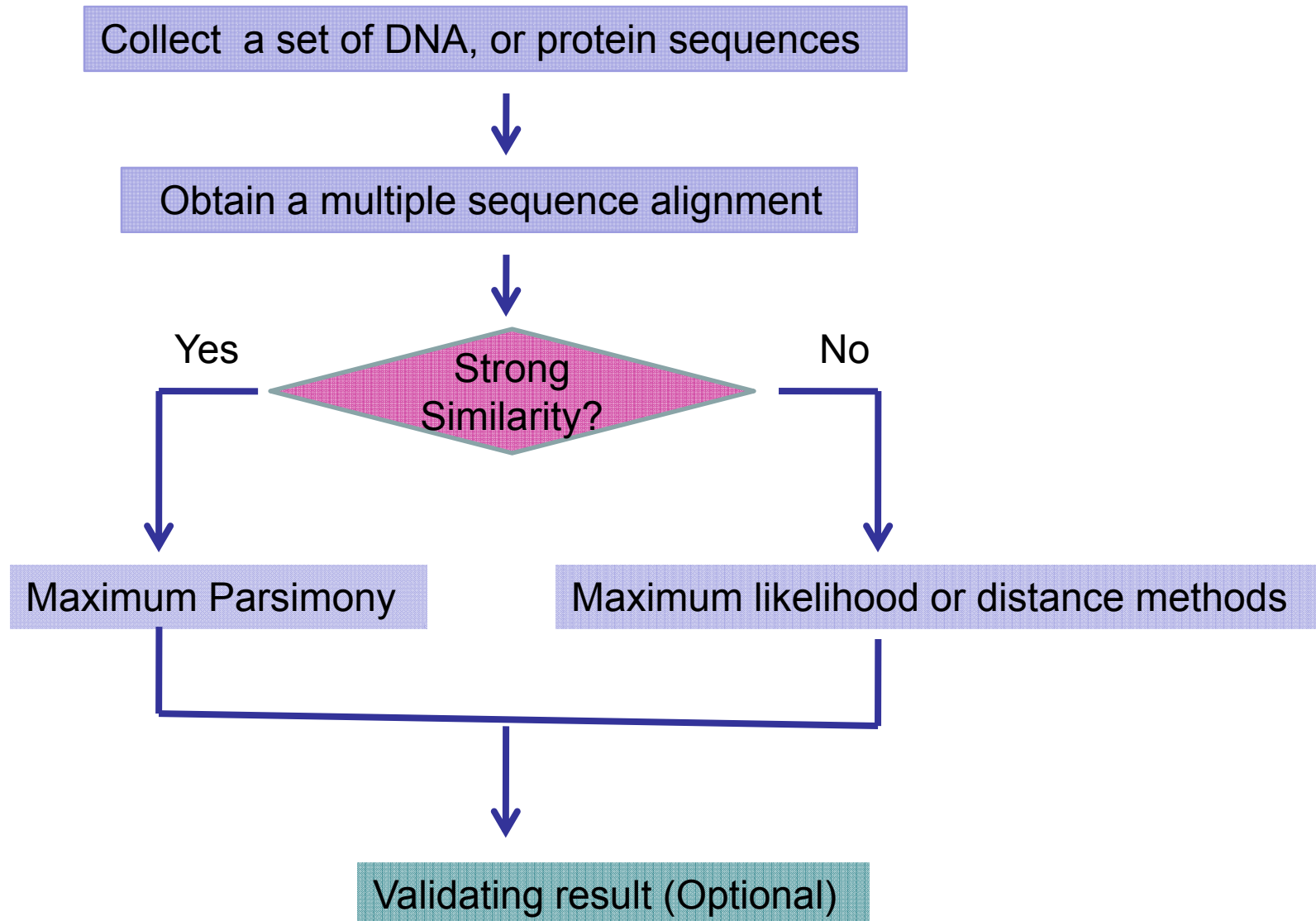
$$d(S, T) \leq n \log_2 n + 2n$$

Procedure of Building (or Reconstructing) Molecular Phylogenetic Trees

Methods:

- Character based methods
 - Maximum parsimony
 - Maximum likelihood
- Distance based methods
 - Neighbor joining
 - UPGMA method

How to choose a phylogenetic method?



Step 1: Selecting sequences

What type of sequences to use, non-coding, coding or protein sequences?

- Non-coding regions have high substitution rate than coding regions
- Proteins are much more conserved than DNA since they need to conserve their function.
- Hence, protein or gene sequences are usually used.
- Recently, even whole genome sequences are used

Step 2: Aligning multiple sequences

- CLUSTALW is often used.
- Important sites could be misaligned by the selected alignment method. Hence, alignment is crucial to phylogenetic tree building.
- Gaps are treated differently by different alignment methods and so columns containing gaps are often removed for tree building.
- Low complexity regions create random bias for alignment and should be also removed from building the tree.

Step 3: Select a phylogenetic tree method

- All existing program for tree building are based on heuristic methods. They usually produce near-optimal trees rather than optimal one. Two popular program packages are PHYLIP and PAUP.
- Different methods could output very different trees. Therefore, it is often to check whether the trees output from different methods are consistent or not.
- The researchers should caution against systematic error occurring at this stage when biological conclusion is drawn.

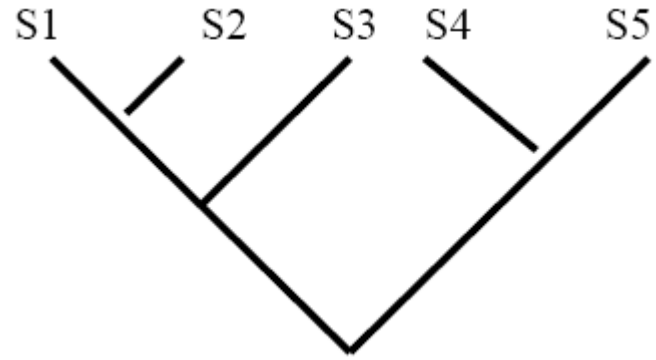
Maximum Parsimony Method

- It assumes that substitution at a position in sequence is independent from those occurring at neighbor positions.
- It outputs a tree that requires the minimum number of changes to explain the differences observed in the alignment data.
 - Ockham's razor principle:

The explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis.

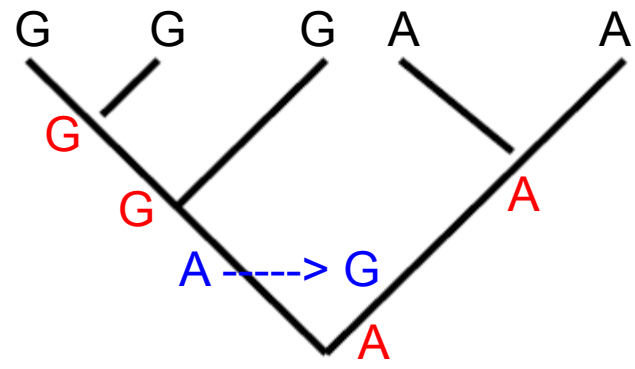
Consider the following DNA data and a proposed tree

| | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------|---|---|---|---|---|---|
| S1 | G | T | C | G | T | A |
| S2 | G | T | C | A | C | T |
| S3 | G | C | G | G | T | A |
| S4 | A | C | G | A | C | A |
| S5 | A | C | G | G | T | A |

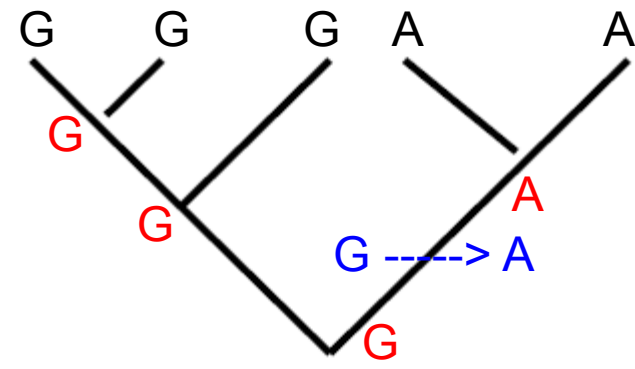


By examining all sites individually, we show that 8 substitutions are needed to explain the data.

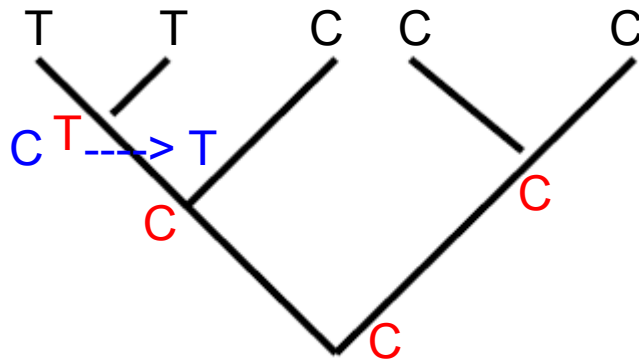
At site 1:



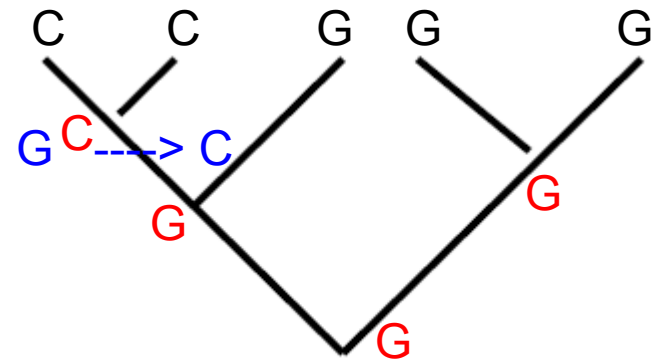
Or



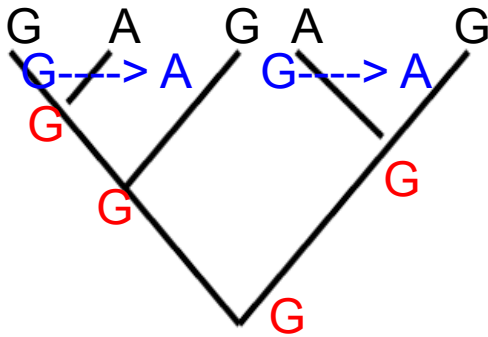
At site 2:



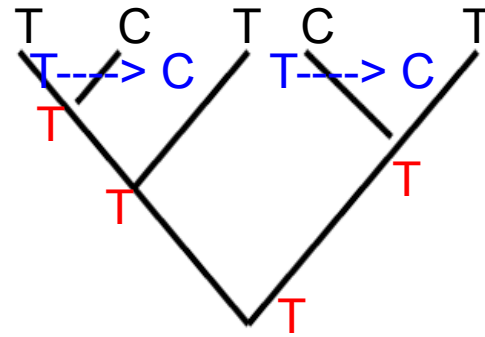
At site 3:



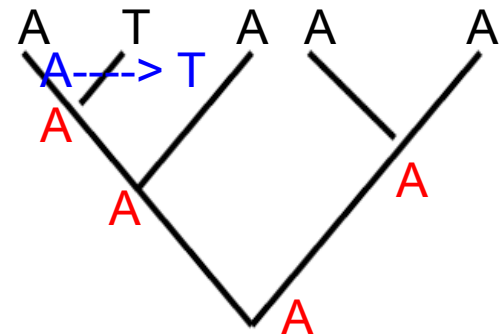
At site 4:



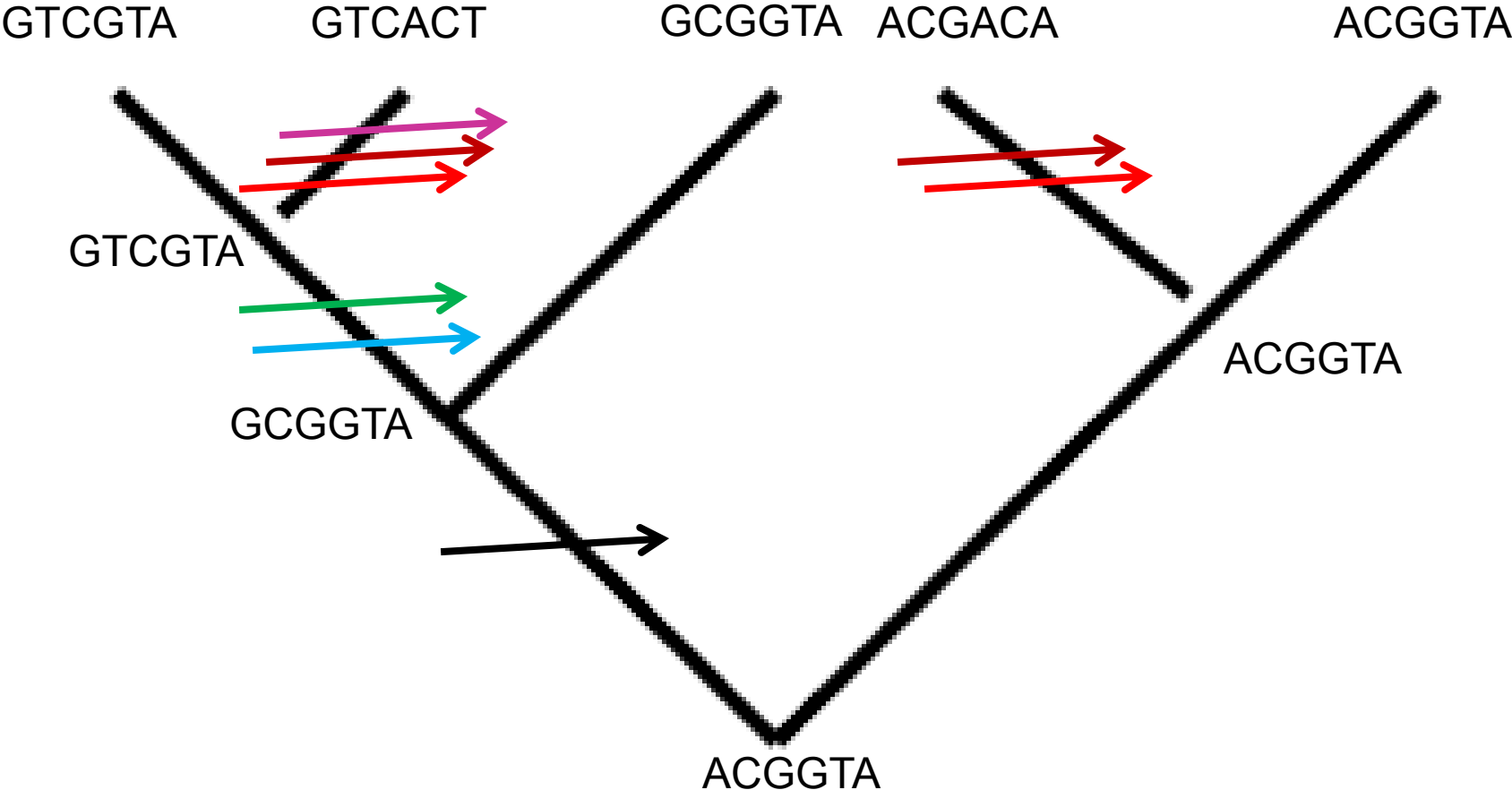
At site 5:



At site 6:



In summary,



Changing the tree will alter the number of substitutions required to explain the data. Hence, the following problem arises.

Small Parsimony Problem:

Input: A rooted phylogenetic tree with leaves labeled with letters.

Question: Label the internal nodes to minimize the number substitutions in all branches.

Assume that u is connected to v by a branch in a rooted tree.

The node u is called a **child** of v if v is closer to the root than u .

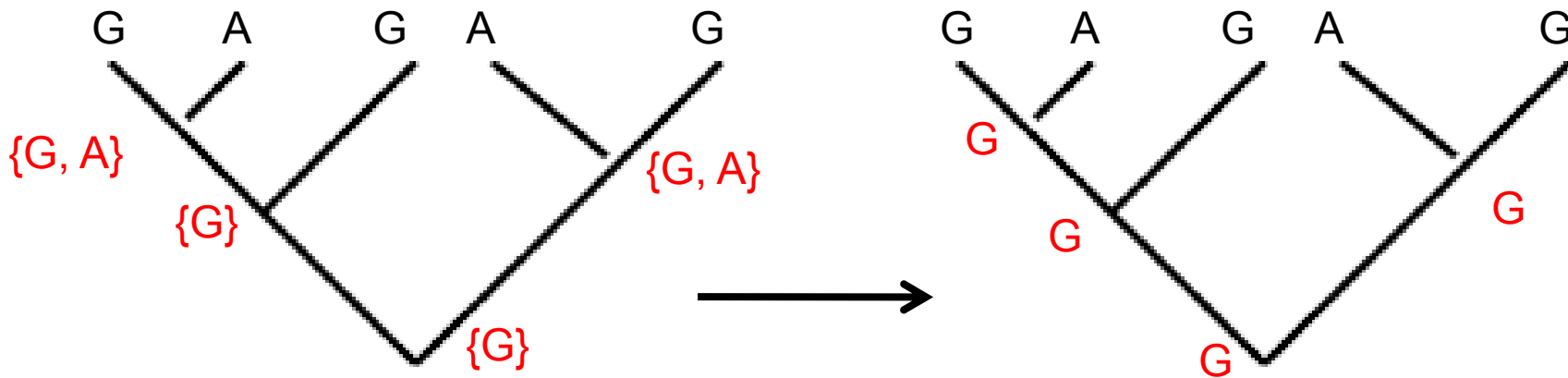
Obviously, each internal node has exactly two children.

Fitch Algorithm

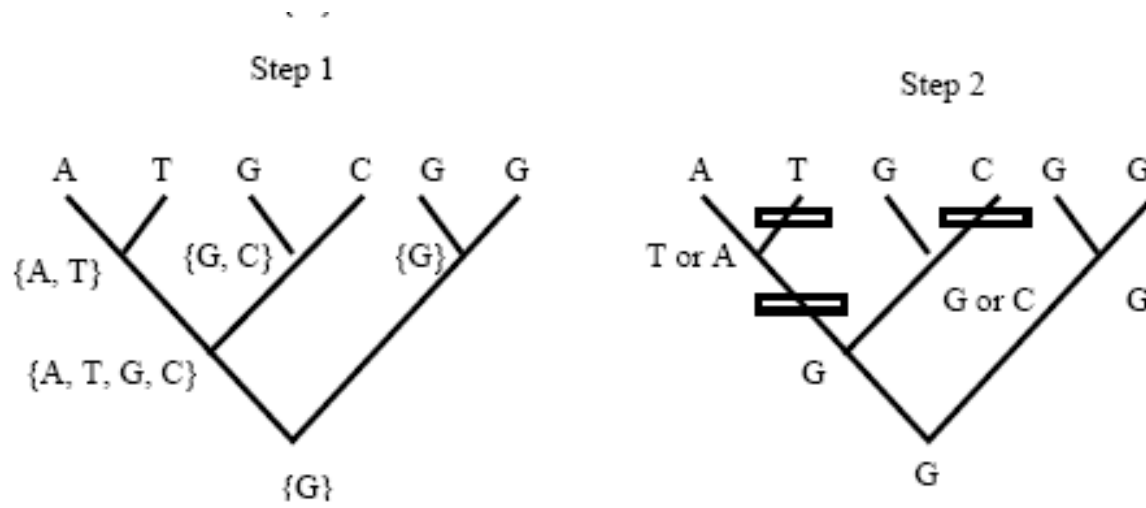
Step 2: Select a letter l_x from S_x to label a node x . This is done from the root to leaves.

- (1) Select an arbitrary letter from S_r to be l_r for the root r .
- (2) Assume u is a child of v and l_v is determined. If l_v belongs to S_u , then $l_u = l_v$. Otherwise, select an arbitrary letter from S_u as l_u .

Example:



Another Example:



Remarks 1. The number of changes in the resulting labeled tree is equal to the number of union operations taken in Step 1.

2. The above algorithm works with a single columns. It has complexity $O(nk)$. To obtain a solution for the entire data with m columns, we simply apply the algorithm to each column and then merge the solution together. This leads to an overall complexity of $O(mnk)$ for the Small Parsimony problem.

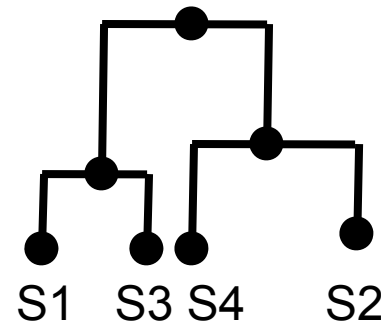
k is the number of letters appearing in the leaf labels.

Maximum Parsimony Problem:

Input: A multiple sequence alignment;

Solution: A most parsimonious tree that explains the sequence data with the minimum number of substitutions.

S1: GCTTTATTCTT
S2: GCTTCATTGAG
S3: GATTCAGTGTG
S4: GCTGTAATGTG



Heuristic Methods to Parsimony

Our final goal is to find an optimal phylogenetic tree. Since there is a huge number of phylogenetic trees, solving the Small Parsimony problem is not enough. Actually, the Parsimony problem is NP-hard. Therefore it is unlikely to find a polynomial time algorithm for it. Like other NP-hard optimization problems, we attack this problem using local search methods.

The basic idea of local search is to define a neighborhood relation for the phylogenetic tree space, and given such relation, to traverse all the trees from one tree to its neighbors, stopping at a local optimum, which will hopefully be the optimum. Typical search techniques include greedy algorithm, simulated annealing, etc.

The Nearest Neighbor Interchange is often used to define neighborhood among trees. We say that two unrooted trees are neighbors if one can be transformed into another in one NNI operation.

It takes about 2 hours.

Parsimony takes 1 hour. So, consider remove how to choose methods.
And put it in application part.

