Lecture 10

Maximum likelihood

Botany 563 — Spring 2021

- **Previous class check-up:**
  - We studied the different models of evolution
- **Learning Objectives:** At the end of today's session, you will be able to
  - Explain how the likelihood of a tree is computed
  - Explain the steps in maximum likelihood phylogenetic inference
- **Pre-class work**
  - Read HAL 1.2

Phylogenetic inference

Step 1: Choose the criterion to use: distances, parsimony, likelihood

Step 2: Search the space of trees until you find the optimum

Phylogenetic inference

Step 1: Choose the criterion to use:
~~distances, parsimony~~, likelihood

Step 2: Search the space of trees
until you find the optimum

You know how to calculate the likelihood for a given tree

Maximum likelihood

1. Choose a substitution model

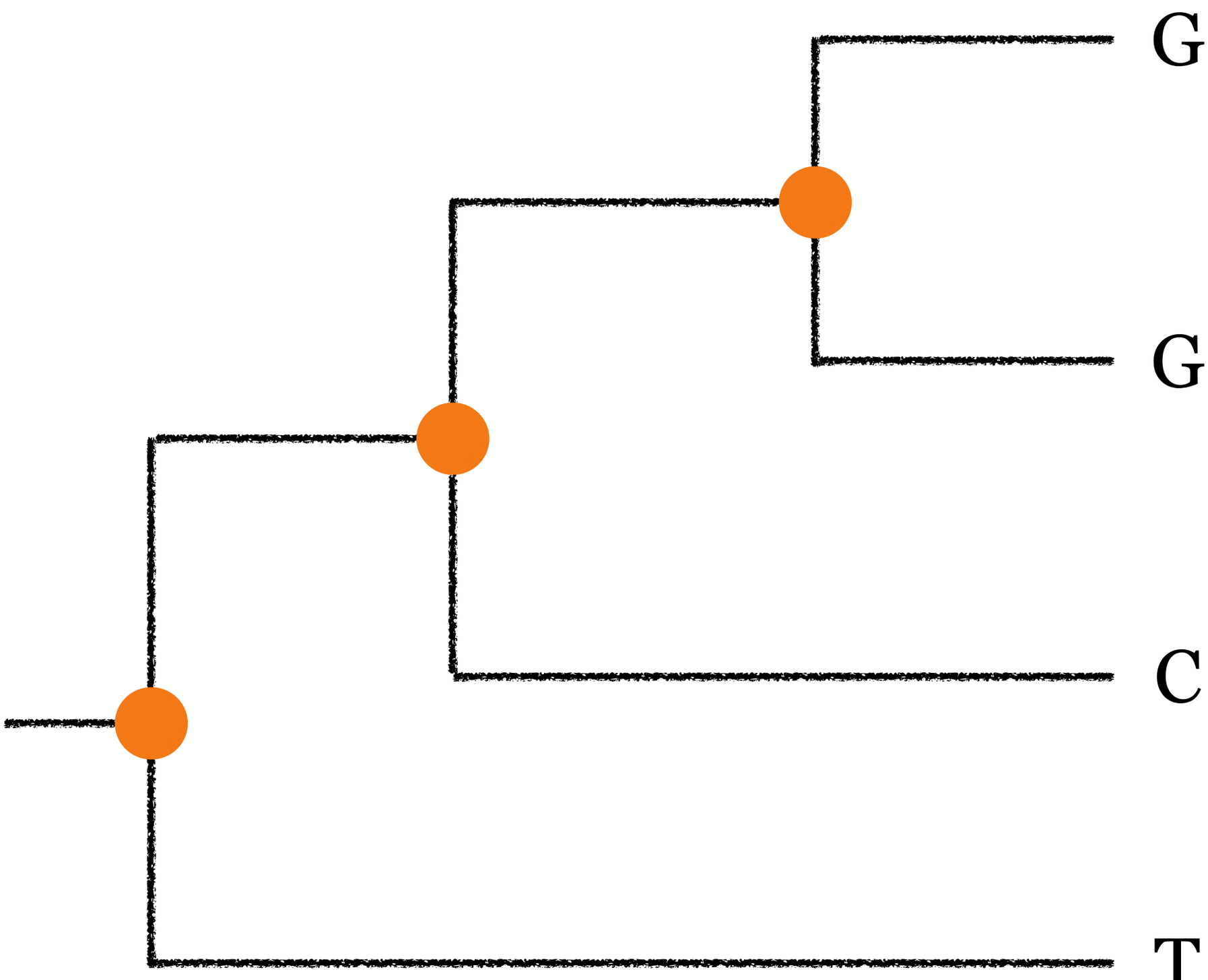
$$\mathbf{P}(t) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathcal{L}_{\mathbf{Q}}\left(\begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \right)$$

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

Calculate the likelihood for this tree

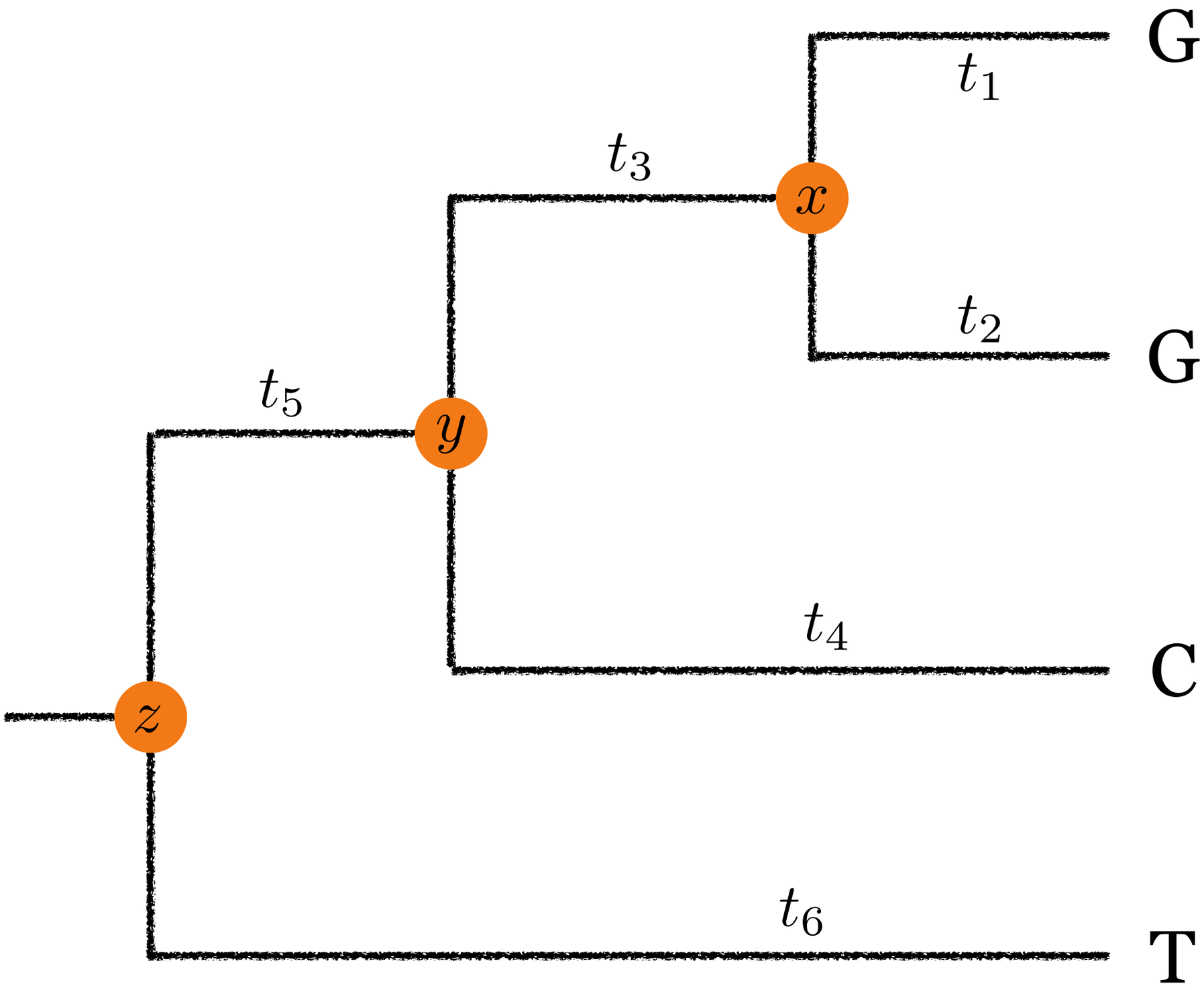


Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree



Depends on parameters:

$\mathbf{Q}$ You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$ Branch lengths

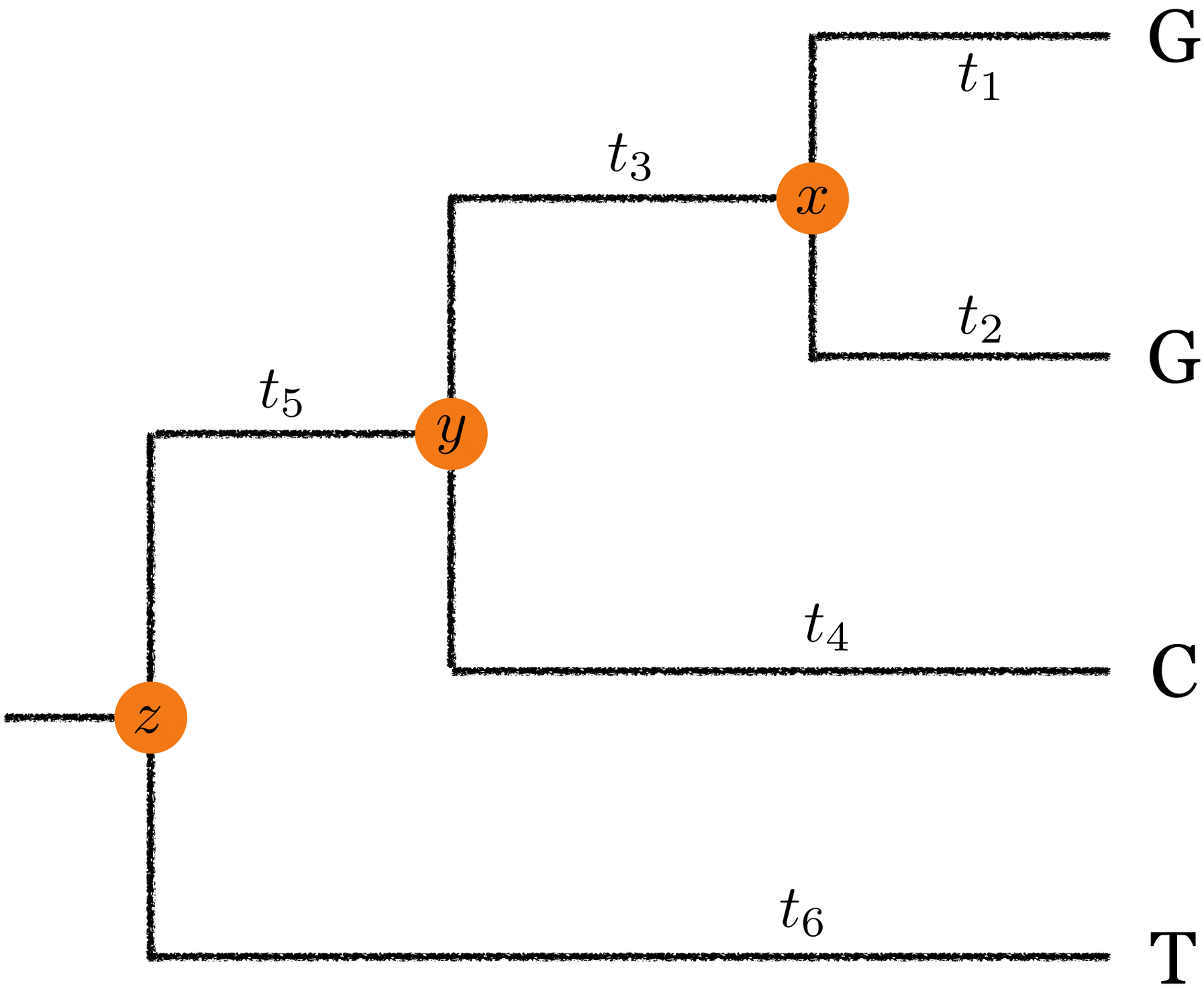
x, y, z Ancestral states

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree



Depends on parameters:

$\mathbf{Q}$ You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$ Branch lengths

x, y, z Ancestral states

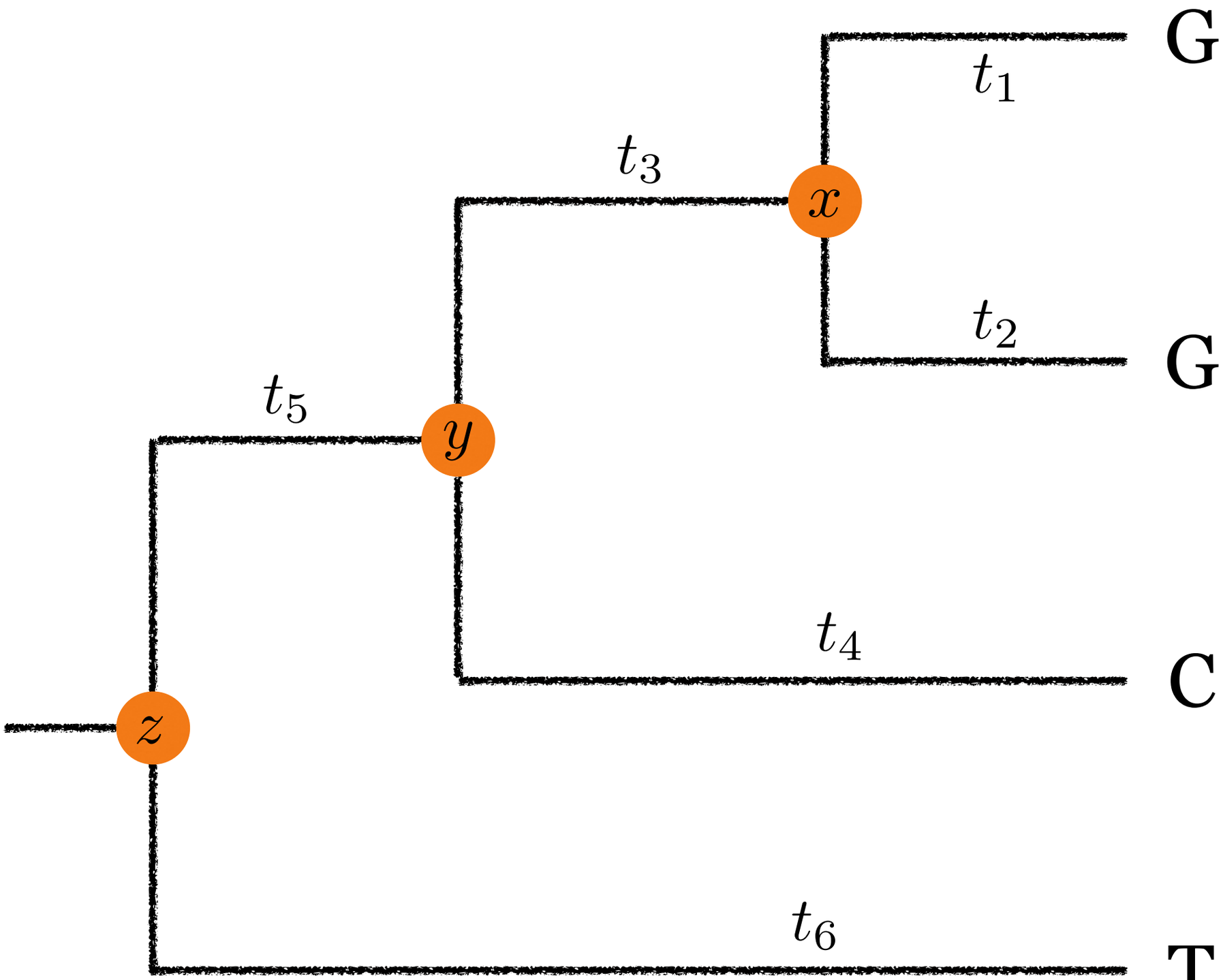
$$\mathcal{L}(T, \mathbf{Q}, \mathbf{t}, x, y, z)$$

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree



Depends on parameters:

$\mathbf{Q}$ You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$ Branch lengths

x, y, z Ancestral states

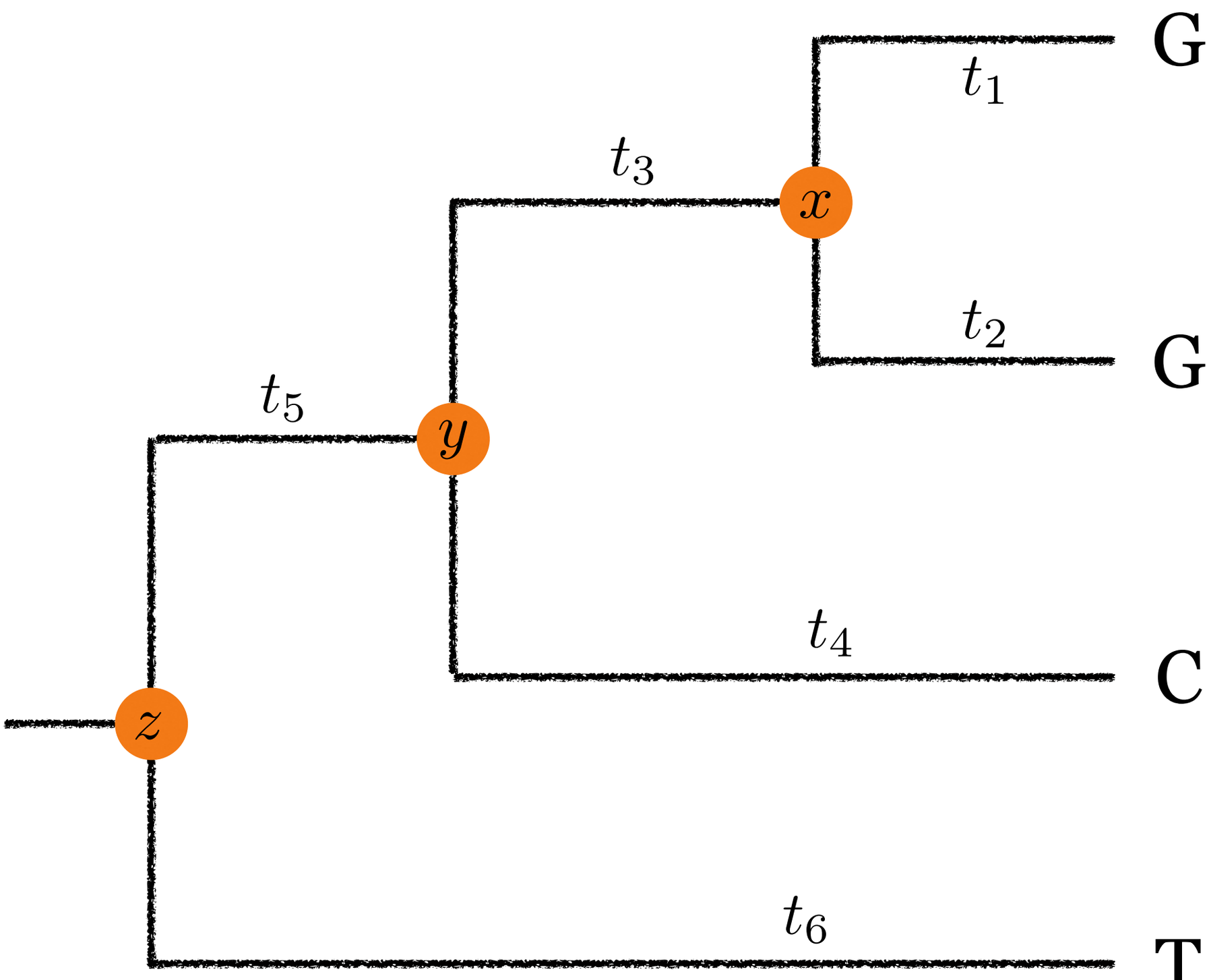
$$\mathcal{L}(T, \mathbf{Q}, \mathbf{t}, x, y, z) = P(GGCT|T, \mathbf{Q}, \mathbf{t}, x, y, z)$$

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

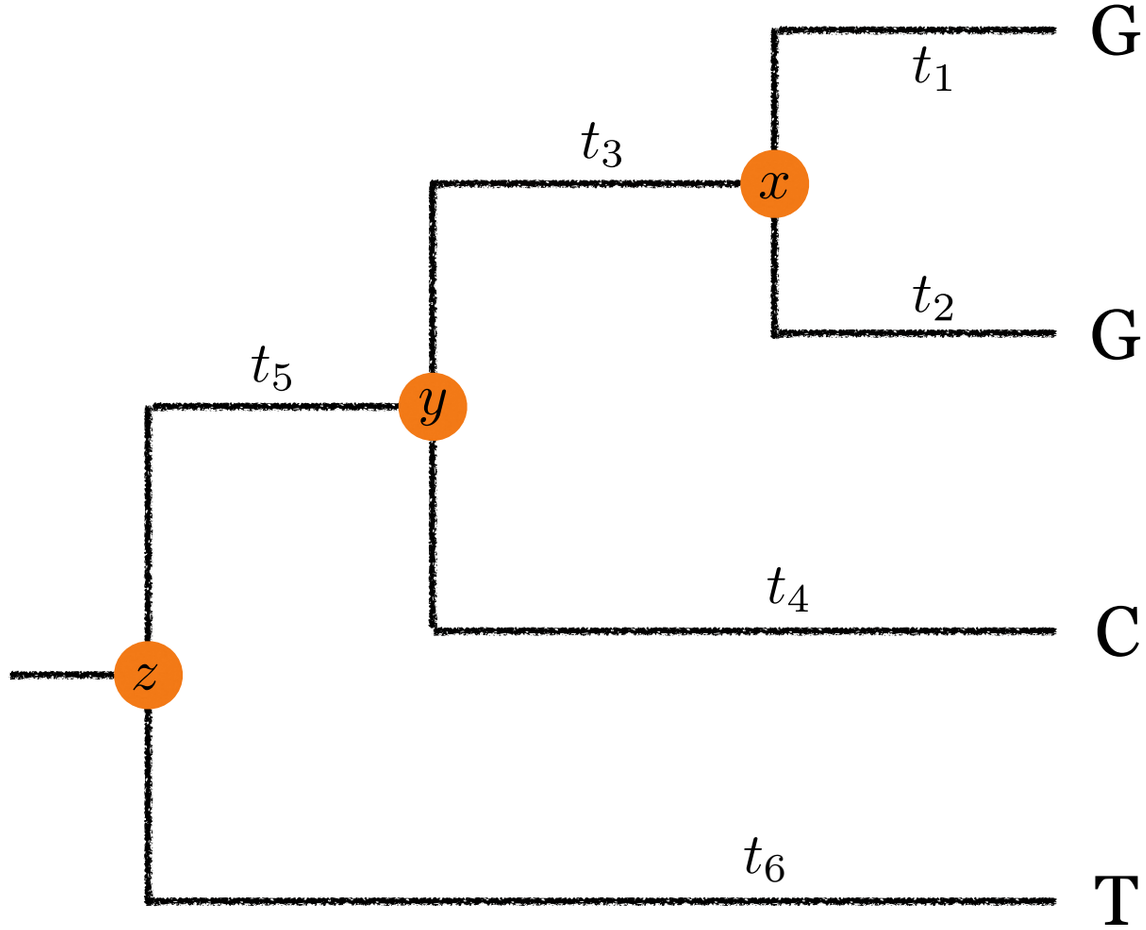


Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree



Assumption 1: The mutation process is the same at every branch of the tree

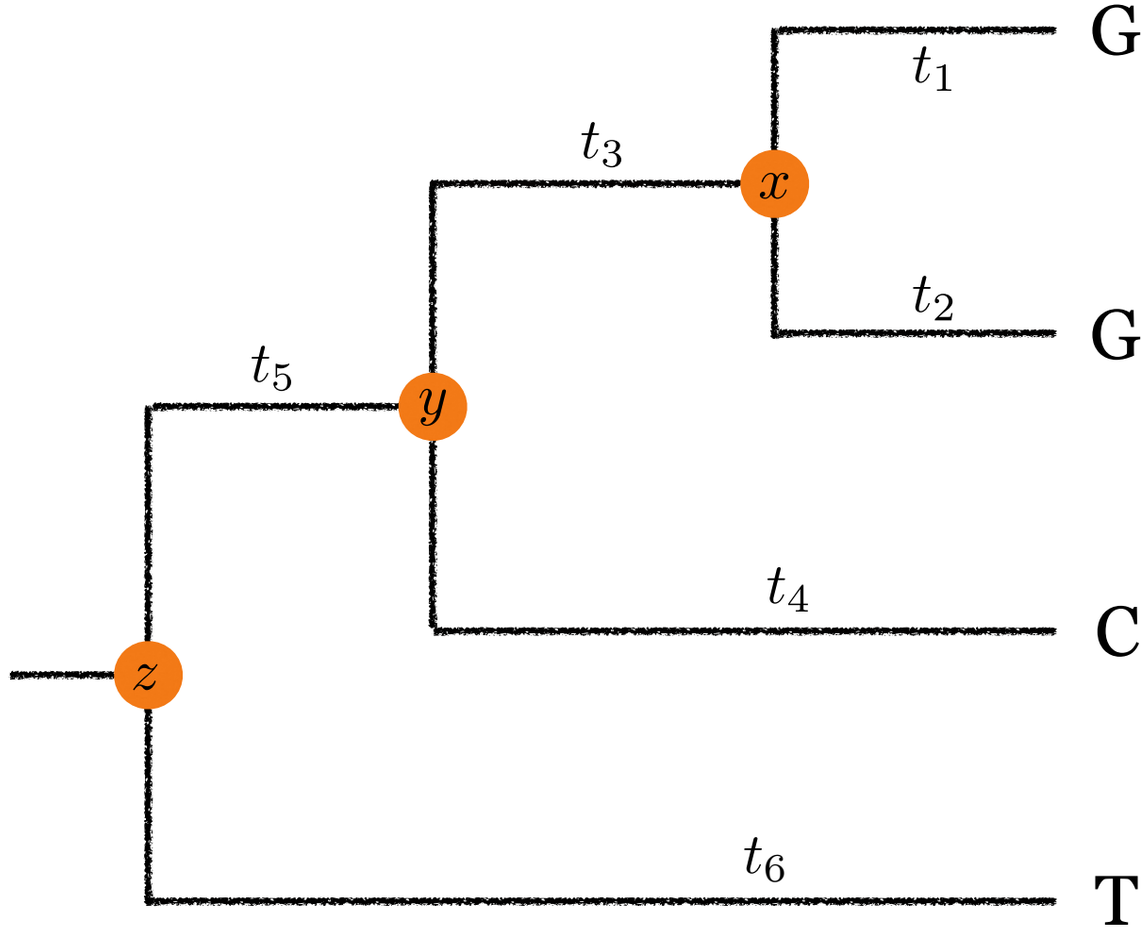
Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{array}{c} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{bmatrix} * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{bmatrix} \end{array} \begin{matrix} A \\ C \\ G \\ T \end{matrix}$$



$$L = \sum_z \sum_y \sum_x \pi(z) P_{t_6}(z, T) P_{t_5}(z, y) P_{t_4}(y, C) P_{t_3}(y, x) P_{t_2}(x, G) P_{t_1}(x, G)$$

Assumption 1: The mutation process is the same at every branch of the tree

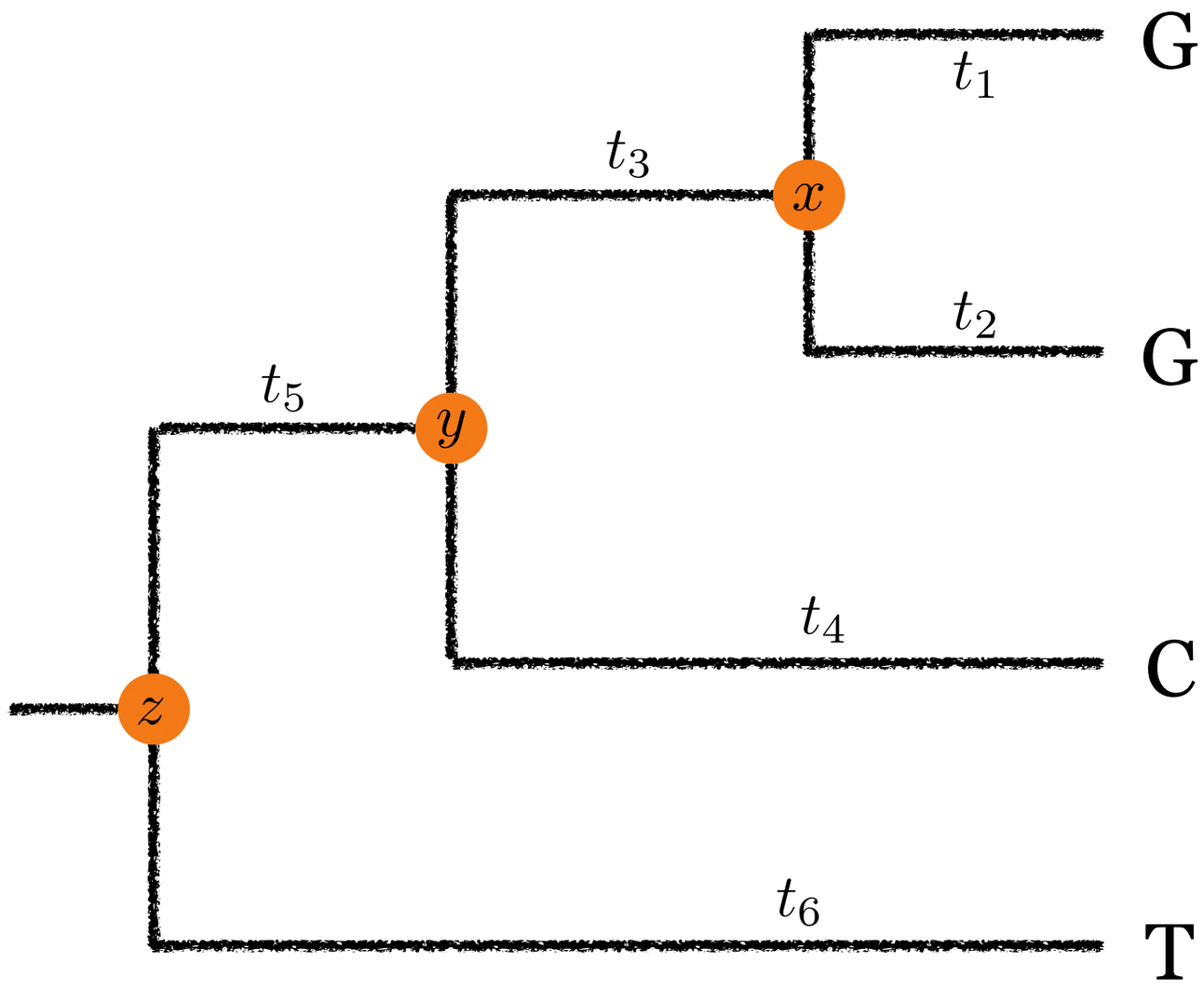
Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{bmatrix} \end{array}$$



$$L = \sum_z \sum_y \sum_x \pi(z) P_{t_6}(z, T) P_{t_5}(z, y) P_{t_4}(y, C) P_{t_3}(y, x) P_{t_2}(x, G) P_{t_1}(x, G)$$

Where do the assumptions play a role?

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

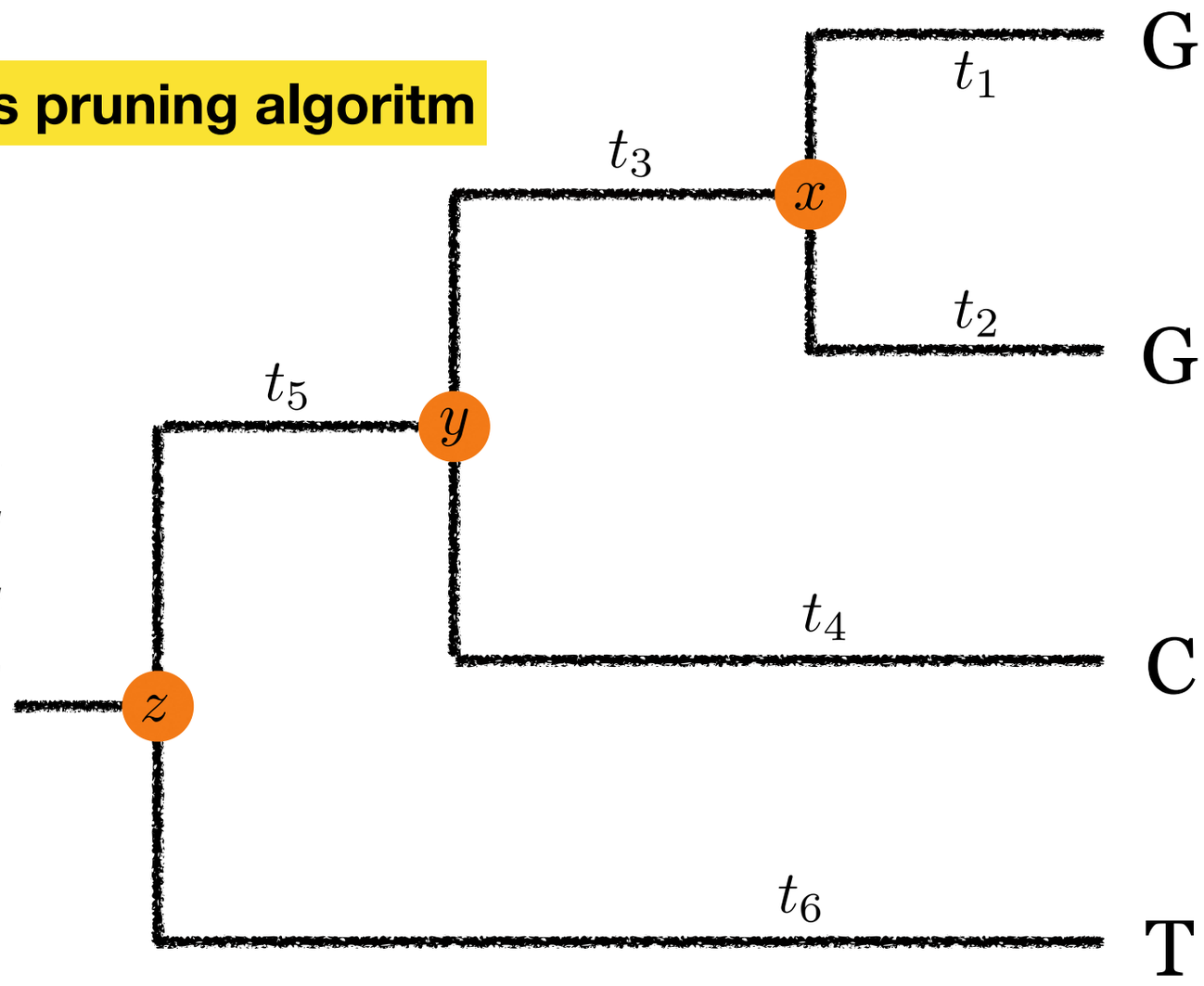
Assumption 3: All sites evolve the same

Calculate the likelihood for this tree

Felsenstein's pruning algoritm

$$\mathbf{P}(t) = e^{\mathbf{Q}\mu t}$$

$$\mathbf{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} * & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & * & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & * & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & * \end{bmatrix} \end{array}$$



$$L = \sum_z \sum_y \sum_x \pi(z) P_{t_6}(z, T) P_{t_5}(z, y) P_{t_4}(y, C) P_{t_3}(y, x) P_{t_2}(x, G) P_{t_1}(x, G)$$

Where do the assumptions play a role?

Assumption 1: The mutation process is the same at every branch of the tree

Assumption 2: We assume sites evolve independently

Assumption 3: All sites evolve the same

Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|cccc|}\hline & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \\ \hline\end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathscr{L}_{\mathbf{Q}}\left(\begin{array}{l} \\ \\ \\ \\ \end{array} \middle| \begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \right)$$

Depends on parameters:

$\mathbf{Q}$ You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$ Branch lengths

x, y, z Ancestral states

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|cccc|}\hline & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \\ \hline\end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathscr{L}_{\mathbf{Q}}\left(\begin{array}{l} \\ \\ \\ \\ \\ \end{array} \middle| \begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \right)$$

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

Depends on parameters:

$\mathbf{Q}$ You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$ Branch lengths

~~x, y, z Ancestral states~~

Average across them

Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|cccc|}\hline & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & & & & \\ \text{C} & & & & \\ \text{G} & & & & \\ \text{T} & & & & \\ \hline\end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathscr{L}_{\mathbf{Q}}\left(\begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \right)$$

Depends on parameters:

$\mathbf{Q}$

You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$

Branch lengths

~~x, y, z~~ Ancestral states

Average across them

Need to optimize

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

➡ At each proposed tree, we maximize Q and t

Maximum likelihood

1. Choose a substitution model

$$\mathbf{P}(t) = \begin{array}{|c|cccc|}\hline & A & C & G & T \\ \hline A & & & & \\ C & & & & \\ G & & & & \\ T & & & & \\ \hline\end{array} = e^{\mathbf{Q}\mu t}$$

2. For a given tree, calculate the likelihood given the data and the substitution model

$$\mathscr{L}_{\mathbf{Q}}\left(\begin{array}{l} \text{AAGTCTAG} \\ \text{AAGTCTAG} \\ \text{AACTCTAG} \\ \text{AATTCTAG} \end{array} \right)$$

Depends on parameters:

$\mathbf{Q}$

You choose which form (each model has its own parameters)

$\mathbf{t} = (t_1, \ldots, t_6)$

Branch lengths

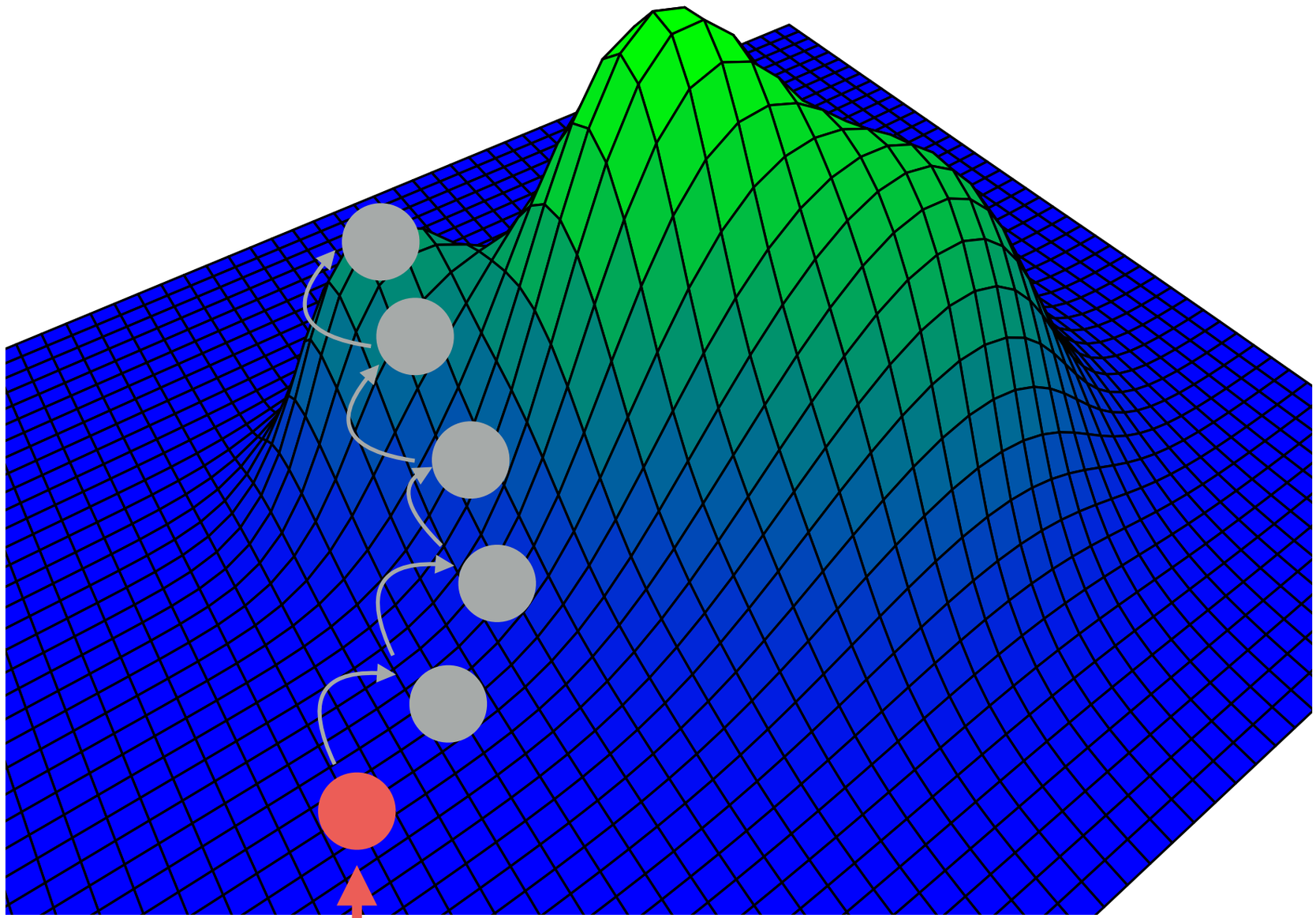
~~x, y, z~~ Ancestral states

Average across them

Need to optimize

3. Search the space of trees using the tree moves (NNI, SPR, TBR) until you find the maximum likelihood tree

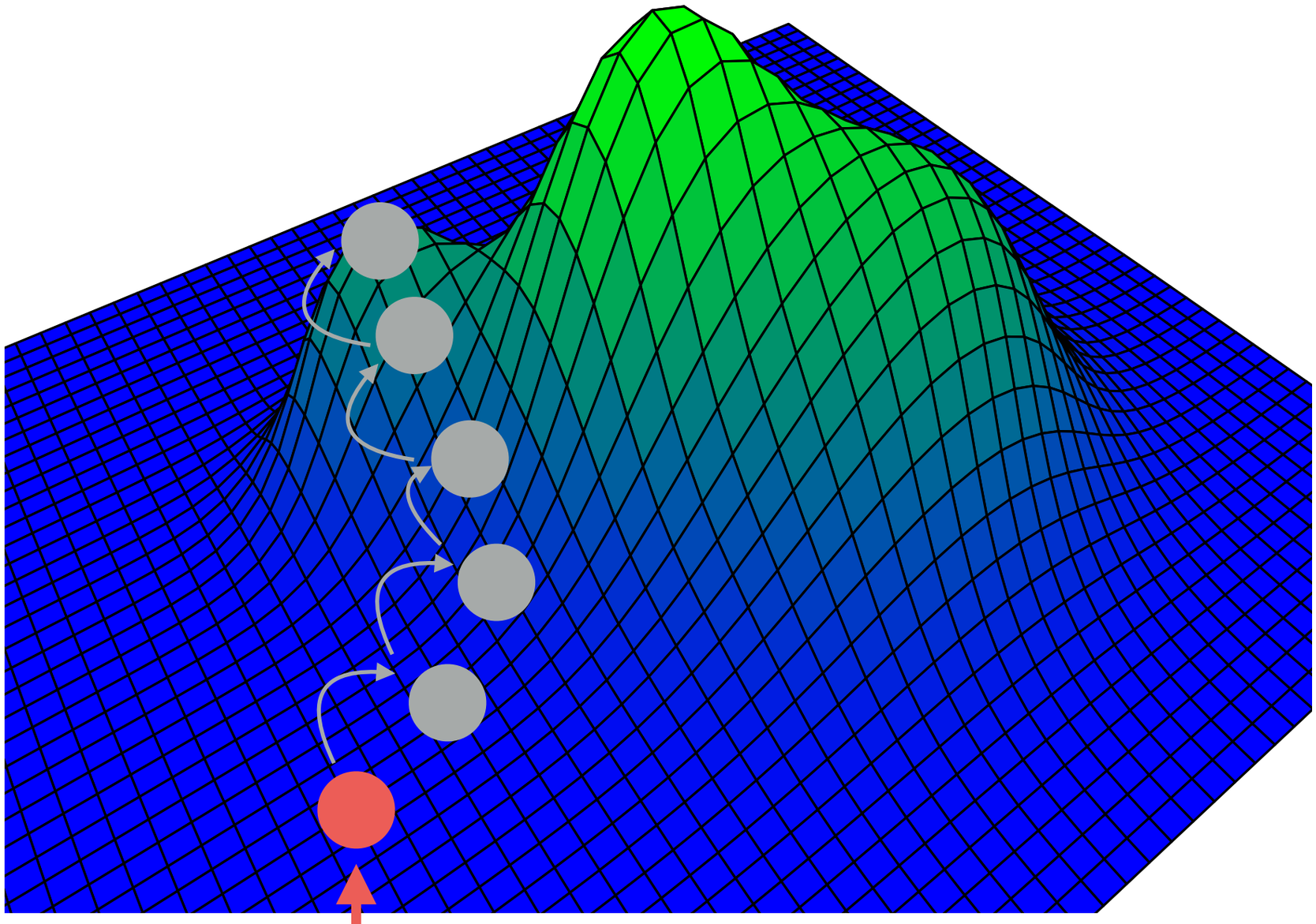
Traverse tree space: finding the MLE



Starting tree

Nearest Neighbor Interchange (NNI)
Subtree Pruning and Regrafting (SPR)
Tree Bisection and Reconnection (TBR)

Traverse tree space: finding the MLE

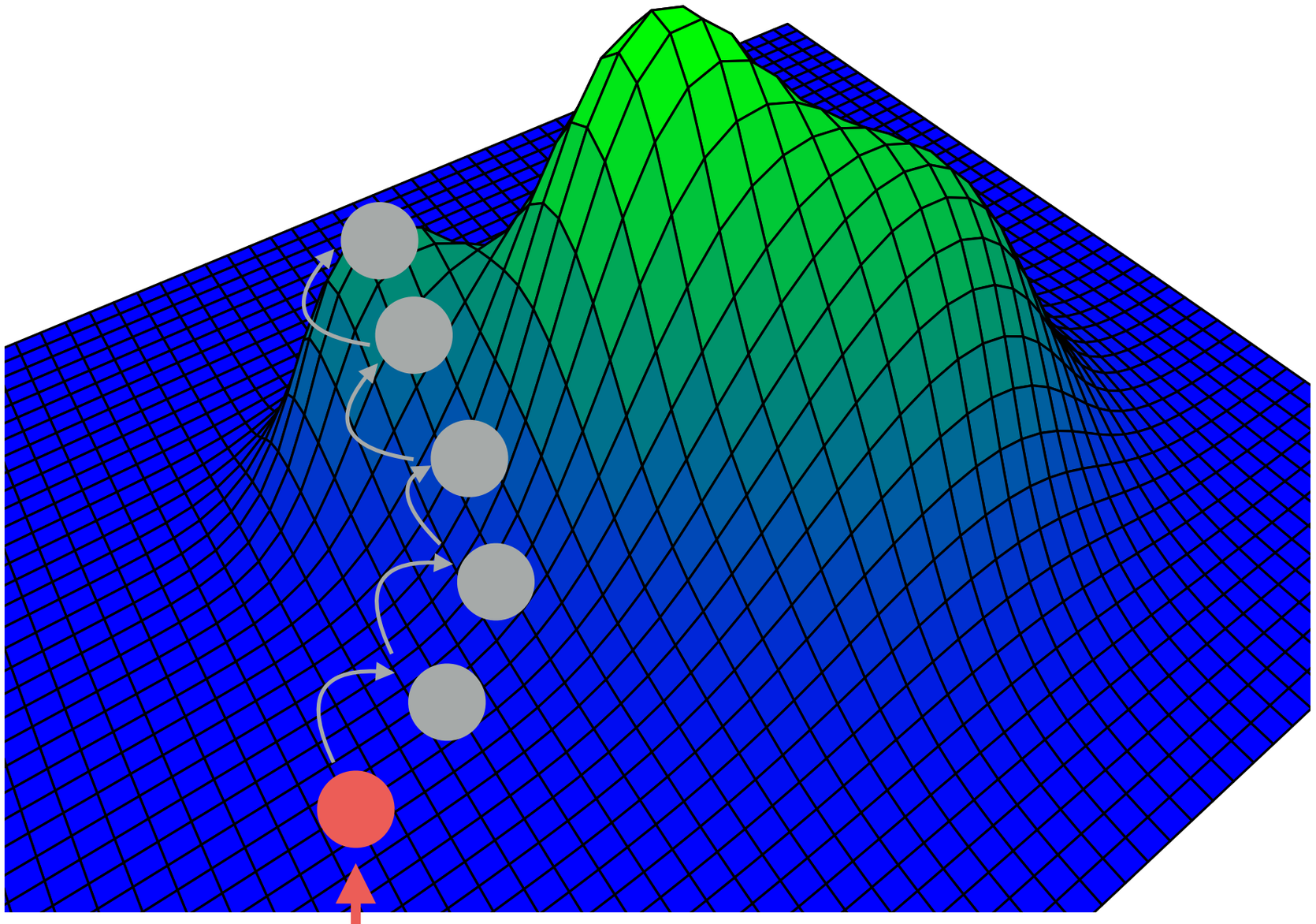


Starting tree

Four things affecting performance:

- Starting tree
- Model chosen
- Data
- Convergence

Traverse tree space: finding the MLE

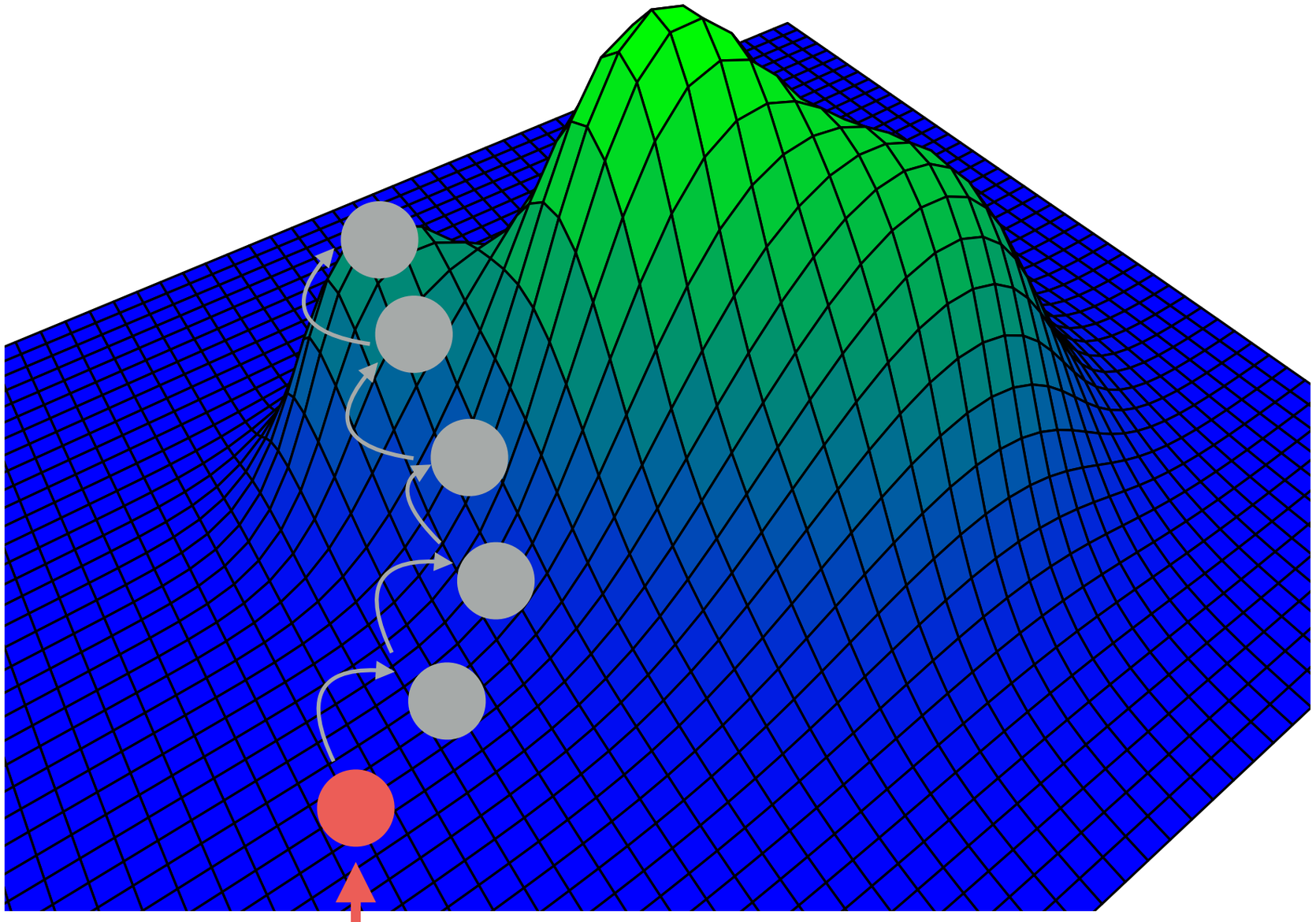


Starting tree

Four things affecting performance:

- Starting tree
  - Affects optimization
  - Get stuck on poor likelihood region
  - Best case: slows down
  - Worst case: suboptimal tree
- Model chosen
- Data
- Convergence

Traverse tree space: finding the MLE

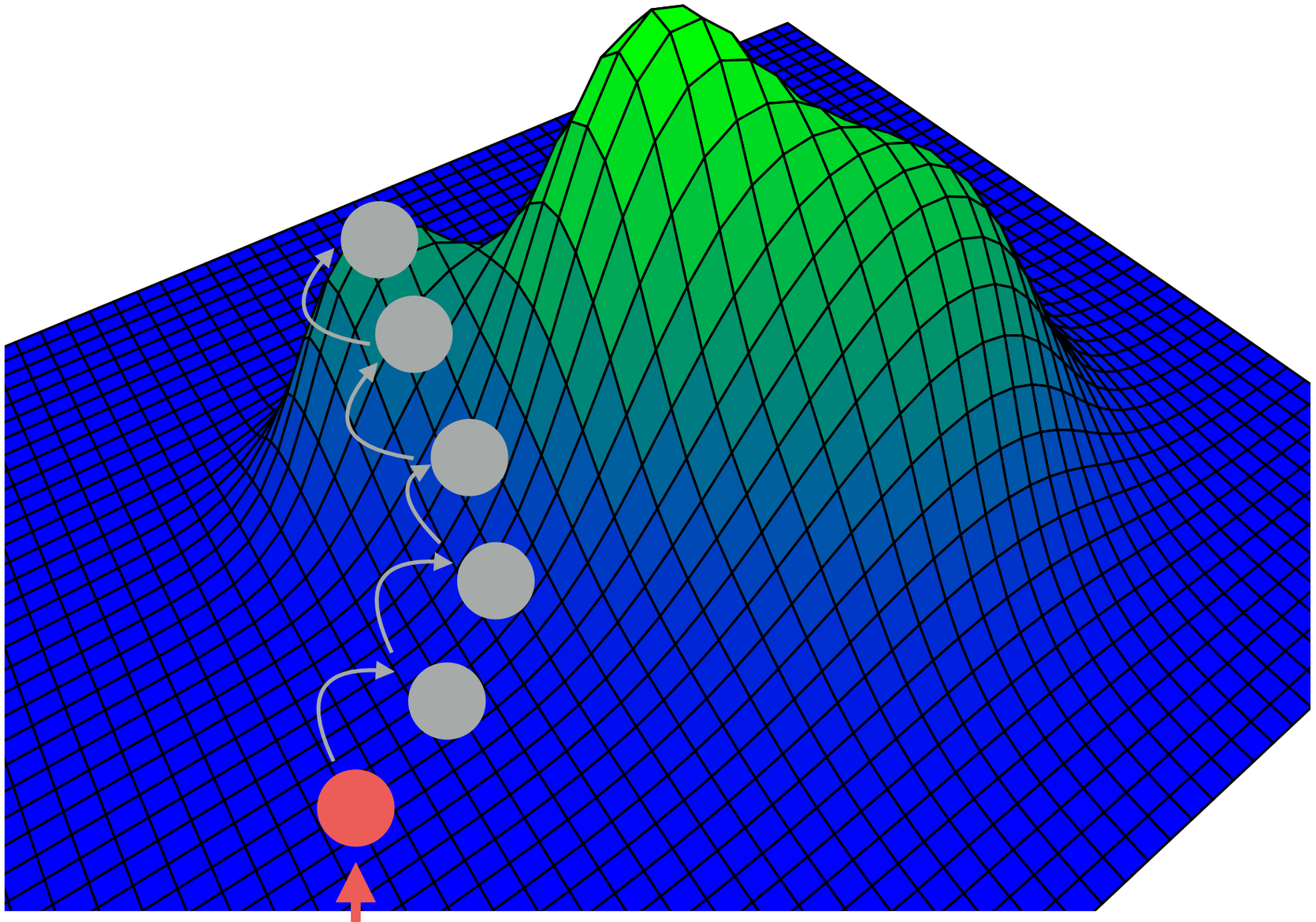


Starting tree

Four things affecting performance:

- Starting tree
  - Affects optimization
  - Get stuck on poor likelihood region
  - Best case: slows down
  - Worst case: suboptimal tree
- Model chosen
  - Affects shape of the surface we optimize
  - You might be optimizing the wrong function
  - Identifiability
- Data
- Convergence

Traverse tree space: finding the MLE



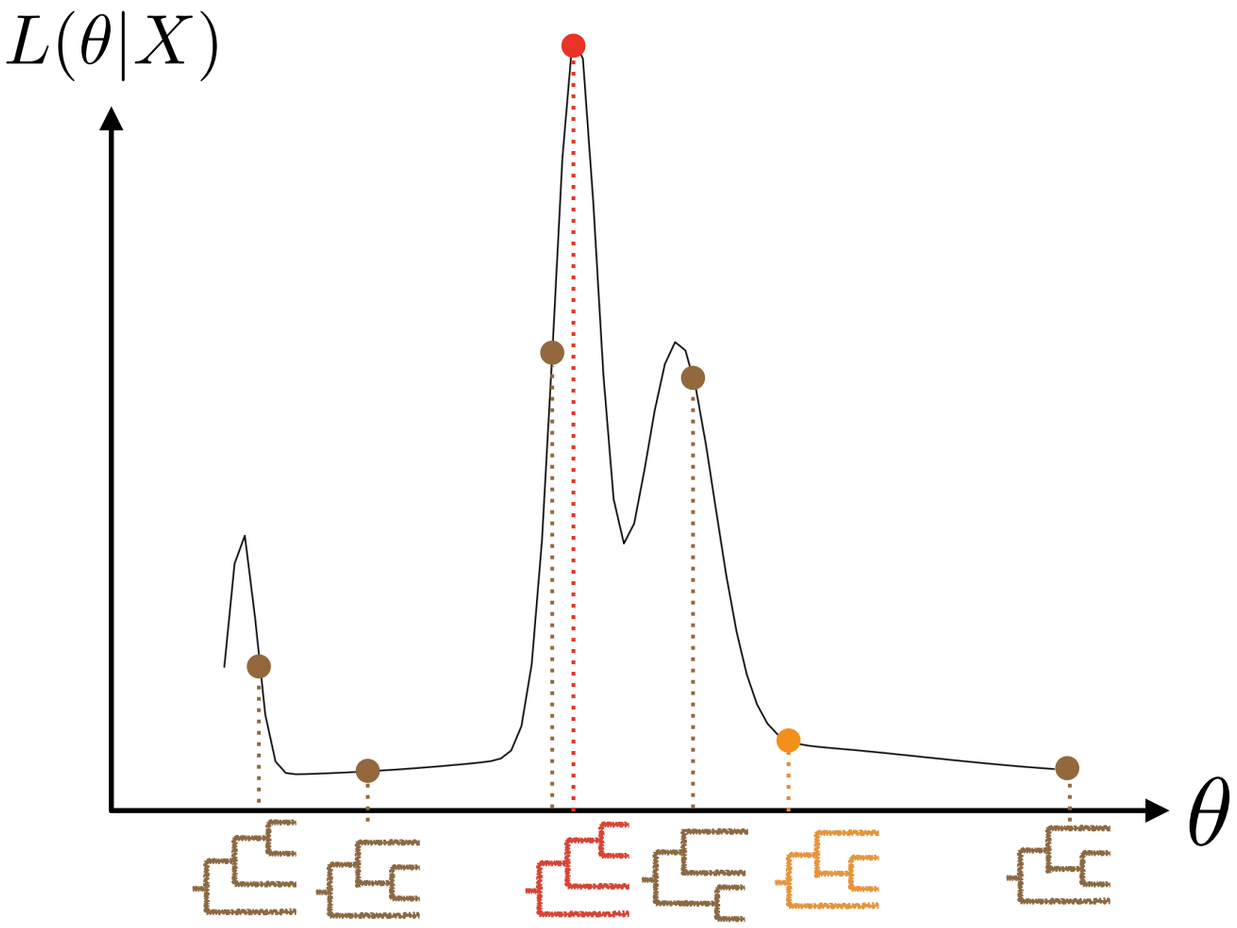
Starting tree

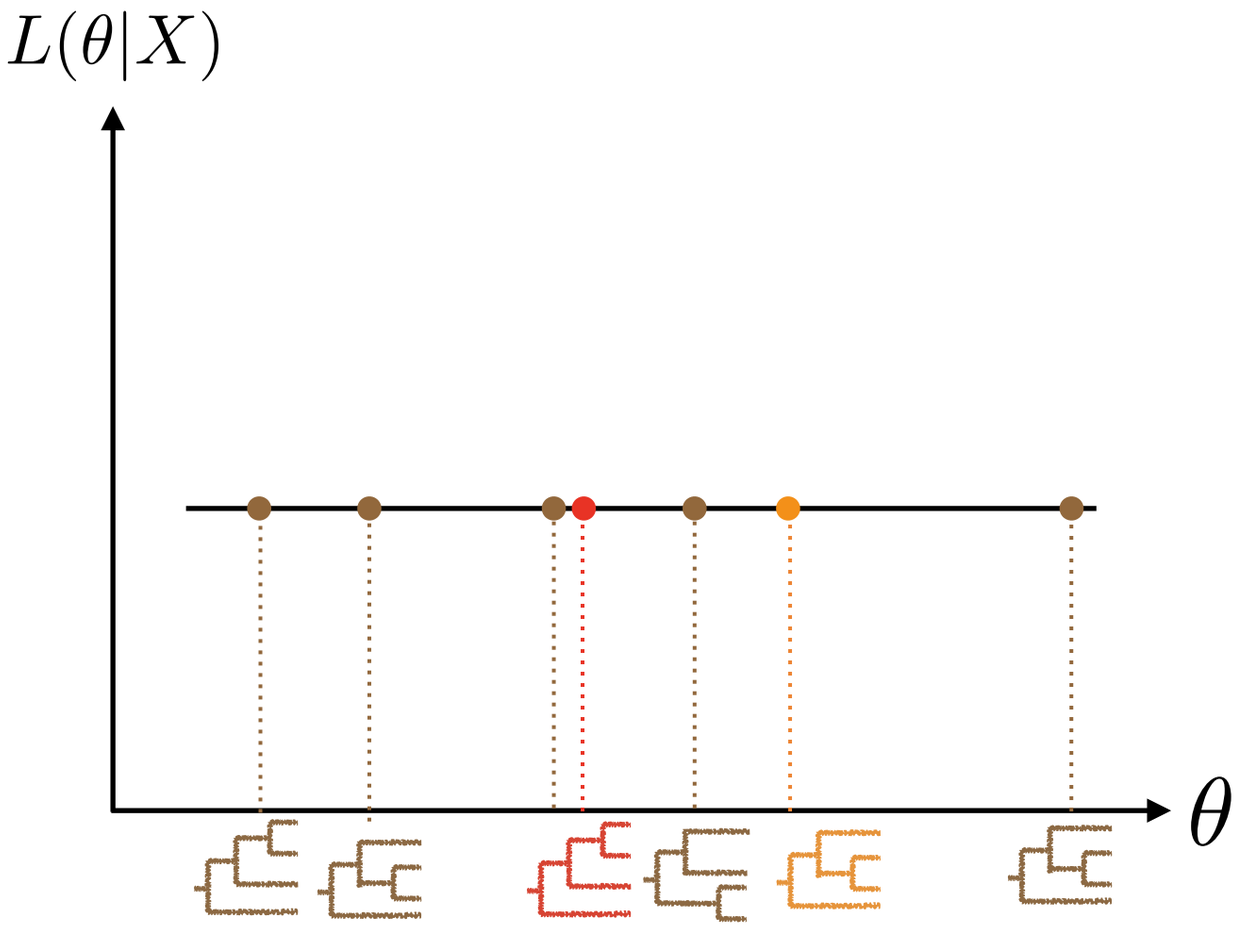
Four things affecting performance:

- Starting tree
  - Affects optimization
  - Get stuck on poor likelihood region
  - Best case: slows down
  - Worst case: suboptimal tree
- Model chosen
  - Affects shape of the surface we optimize
  - You might be optimizing the wrong function
  - Identifiability
- Data
- Convergence

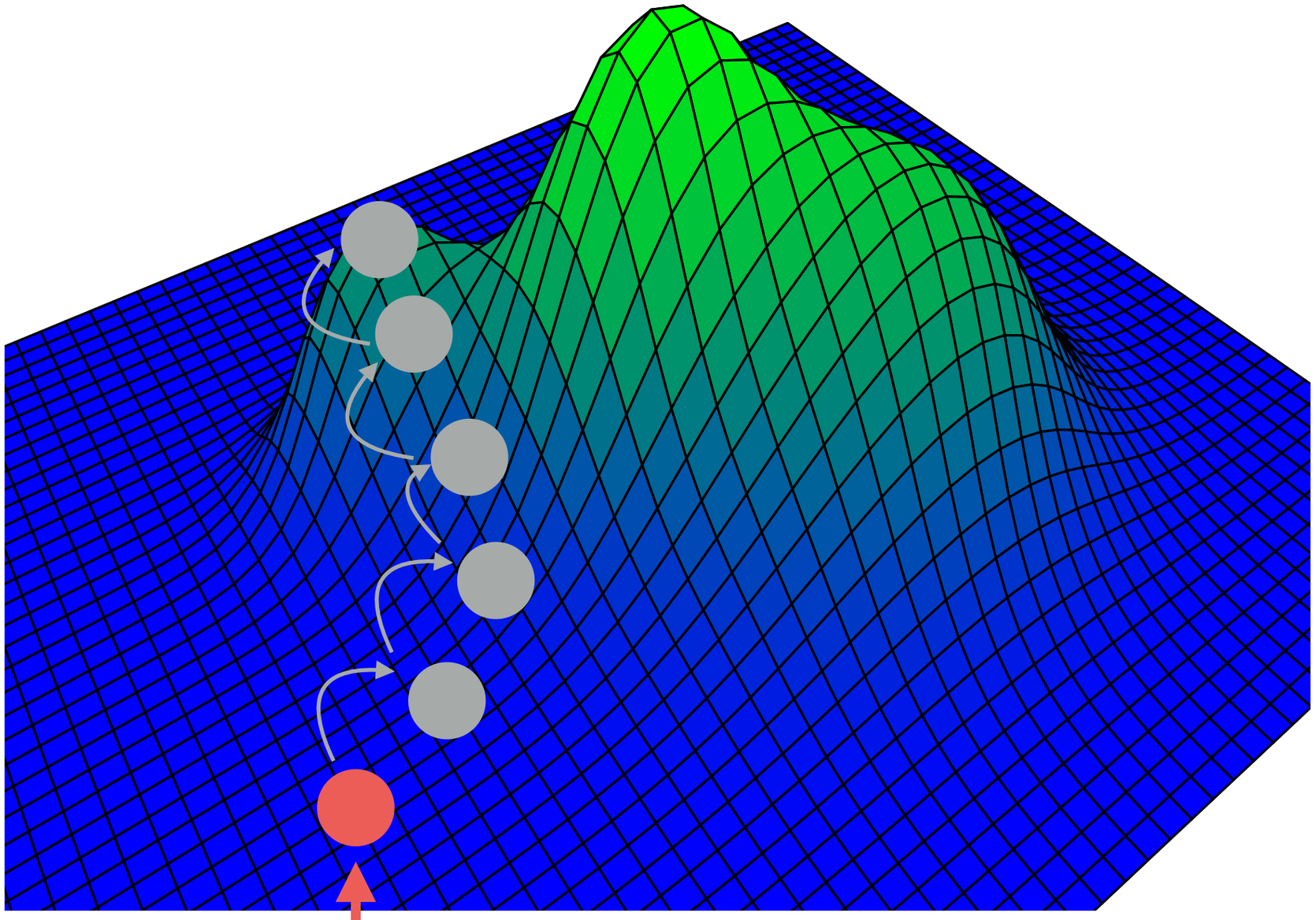
HAL 1.2 "Rough likelihood surface: when analyzing datasets with comparatively few sites and a large number of taxa. The key challenge with such datasets is that 100 distinct ML searches are likely to yield 100 topologically substantially different, but statistically indistinguishable trees."

Identifiability





Traverse tree space: finding the MLE

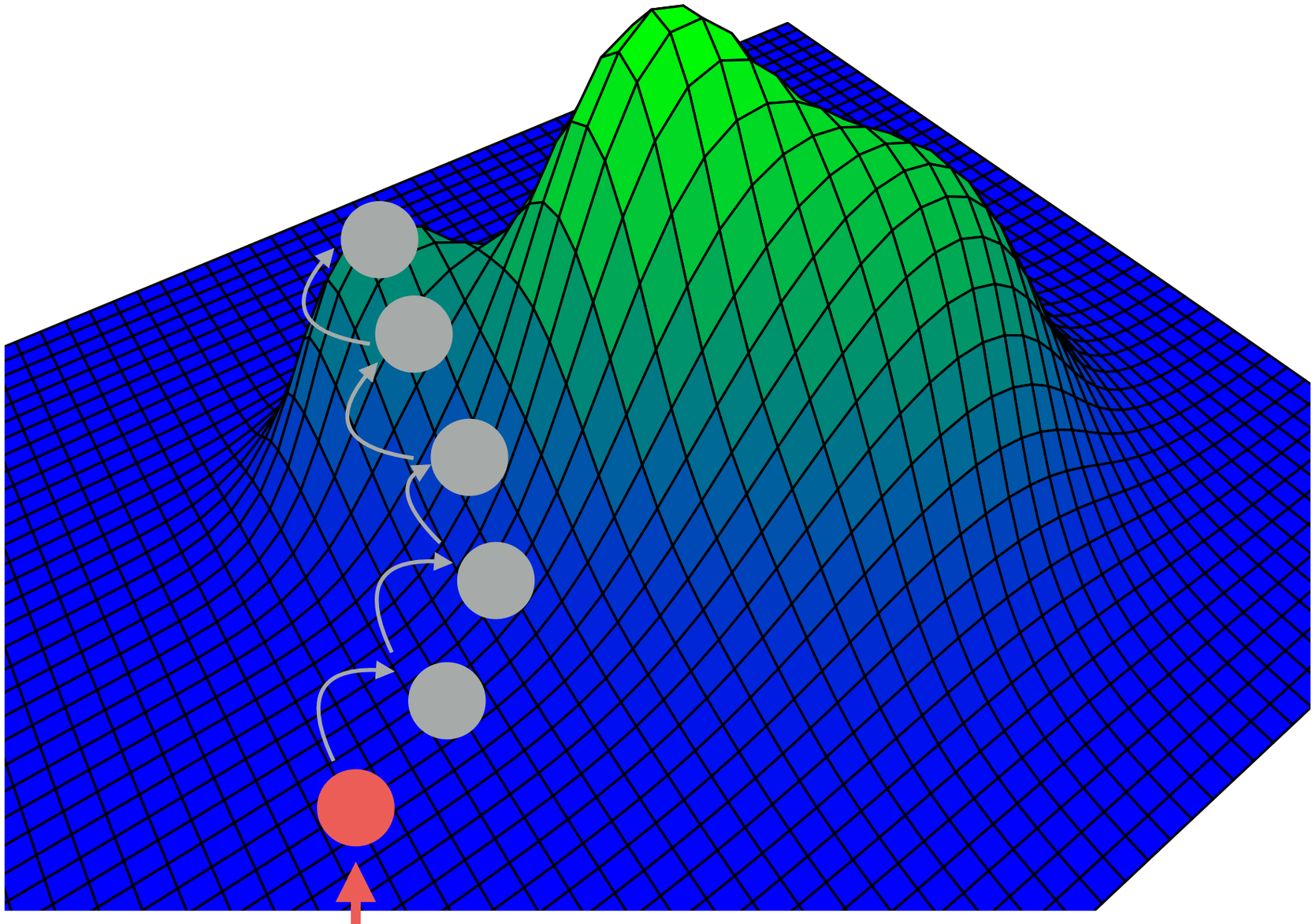


Starting tree

Four things affecting performance:

- Starting tree
  - Affects optimization
  - Get stuck on poor likelihood region
  - Best case: slows down
  - Worst case: suboptimal tree
- Model chosen
  - Affects shape of the surface we optimize
  - You might be optimizing the wrong function
  - Identifiability
- Data
  - Lack of signal (sample size or poorly chosen region)
  - Difference between data and information
  - Identifiability
- Convergence

Traverse tree space: finding the MLE



Starting tree

Four things affecting performance:

- Starting tree
  - Affects optimization
  - Get stuck on poor likelihood region
  - Best case: slows down
  - Worst case: suboptimal tree
- Model chosen
  - Affects shape of the surface we optimize
  - You might be optimizing the wrong function
  - Identifiability
- Data
  - Lack of signal (sample size or poorly chosen region)
  - Difference between data and information
  - Identifiability
- Convergence
  - When do you stop the traversal of tree space?
  - Affects optimization

Statistical Consistency

- Maximum likelihood (and Bayesian), neighbor joining, ME OLS are all statistically consistent methods

- UPGMA and maximum parsimony are not statistically consistent methods

In-class dynamic

- **Time:** 15 minutes
- **Instructions:** Choose a software that does maximum likelihood and follow the tutorial
    - RAxML-NG: HAL 1.3 (github repo)
    - IQ-TREE: tutorial

and create our own reproducible script. **Bonus points** for paying attention to the four important things that affect performance in the chosen method

- **Disclaimer:** I have not done the steps ahead of time to make sure that everything runs smoothly so that we can troubleshoot this pipeline together

- **Options for you:**

1. "I think that I can follow the pipeline by myself or with a small group of peers": you should join the Congregate room

2. "I think I need more one-on-one help to run the commands": you can stay here in the zoom room

Further Reading

- The Contest Between Parsimony and Likelihood
- Phylogenetic analysis using parsimony and likelihood methods
- Comparing Distance-Based Phylogenetic Tree Construction Methods Using An Individual-Based Ecosystem Simulation, EcoSim
- Journal club: An investigation of irreproducibility in maximum likelihood phylogenetic inference

For next class:

- We have three teams: distances, parsimony and likelihood
- Each member of the team will review the pros and cons of their method (make sure to check the “further reading” in lectures and complement it with your own literature search)
- Next class, we will have a friendly discussion among the three teams to highlight the strengths and weaknesses of each method