Study questions:

● **What is a B cell lineage?**

B cell lineage formed by naïve B cell and its variations that have been produced by affinity mutation. Affinity mutation is a process of B cell development that increases affinity for a pathogen-associated antigen. It occurs by multiple rounds of mutations (somatic hypermutations SHM, Darwinian antigen selection).

● **How can studying the immune gene sequences be related to the B cell lineage studies? (Hint: what is the relationship between B cell and BCR?)**

The diversity of immune repertoire conditioned by diversity of BCR. Analyzing BCR diversity (mutations of clones), we can link different B-cells by linking them according to their IG sequences. This will be useful for creation of a lineage of mutations that describes the process of clonal development.

● **Why are the B cell lineage studies potentially valuable in the medical context?**

Studying the process of antibody evolution during pathogen neutralization is essential for understanding clonal selection during the immune response, developing accurate vaccines, discovering therapeutic monoclonal antibodies, or understanding B-cell tumors. The evolutionary approach has been used to quantitatively compare the BCR repertoires of young and aged individuals after influenza vaccination. Furthermore, this antibody approach has been used for clonal reconstruction of BCR repertoires.

● **Why does a "tree" seem a good option to represent a lineage?**

Because it allows us to set the naïve B cell as the root, and by iteratively adding nodes (leaves), we can represent the mutated B cell and visualize the history of the evolutionary changes that the naïve B cell underwent during the process of affinity maturation. Also because it allows us to grow from the root to leaves by adding minimal edge cost.

● **Can we use the classical phylogenetic tree reconstruction algorithms for the B cell lineage?**

Theoretically, yes (if we replace species with BCR sequences with different mutations). Practically, not.

Why? What are the particularities of the B cell lineage trees?

There are several reasons why conventional phylogenetic tree algorithms are not suitable for the reconstruction of BCR lineage trees:

1) In a phylogenetic tree, the root is usually unknown, but in a BCR lineage tree, the root or the BCR sequence of the naive B cell giving rise to the lineage can be frequently predicted with satisfactory accuracy.
2) The observed sequences are represented usually only in the leaves, and the inner nodes represent the relationships amongst sequences while in a BCR lineage tree the observed BCR sequences can be leaves or internal nodes in the tree because different BCR mutations can coexist.
3) IG sequences are under intense selective pressure, and the neutral evolution assumption is invalid.

4) In a phylogenetic tree, the context-dependence of SHM violates the assumption that sites evolve independently and identically.

● **What are the different steps of the Clonaltree algorithm?**

Root - the naive/unmutated BCR sequence

Nodes represent BCR sequences

Weight of edges connecting vertices represents the distance between sequences in terms of mutation, insertion, and deletion operations

Algorithm (Preparation part - construct a graph)

1. ClonalTree performs a **multiple sequence alignment** to consider evolutionary events.
2. Computation of a **similarity-weighted hamming distance** between each pair of sequences. This pairwise distance is then used as **the weight edge** connecting two sequences.

Algorithm (First part – MST with modified Prim's algorithm):

1. Start at the root
2. Add all root's neighbors with minimum edge weight to a priority queue.
3. Iteratively extract from the priority queue the node with the lowest edge weight and highest genotype abundance.
4. If no cycle is formed, the node and the edge will be added to the tree.
5. All the neighbors of the added node with minimum edge weights are included in the priority queue.
6. Repeat steps 3-5 (adding nodes and edges) until all nodes will not be covered.

ClonalTree **adds each node only once** at the priority queue. If a set of edges have the same weight, ClonalTree will choose the one that connects nodes with **high abundance.**

Algorithm (Second part - editing the reconstructed lineage tree)

Adding unobserved intermediate nodes to the tree:

- Traverse the tree in a pre-order manner

- Calculate a distance between sister nodes. If d_{mn} <= d_{mp} and d_{mn} <= d_{np}:

1. Add an unobserved internal node in the tree,

2. Connect it to the observed nodes by direct edge

3. Update edge weights.

Updates for edges weight:

$$d_{pi} = max(d_{pm} - d_{mn}, d_{pn} - d_{mn}, 1)$$

$$d_{im} = d_{pm} - d_{pi}$$

$$d_{in} = d_{pn} - d_{pi}$$

i – internal, **m, n** – sisters, **p** - parent

Detach/reattach sub-trees (performed if it can reduce the depth of the lineage tree by keeping the overall cost):

- consider all branching nodes (nodes with more than one descendent)
- try to detach and reattach each node (under edition condition) to any other node in the lineage tree. If this operation reduces the tree depth:
  1. Accept it;
  2. Examine the resulting lineage tree again for additional edition operations.
- repeat this process until no editing operation can reduce the depth of the tree.

● **What is the originality of the Clonaltree method?**

ClonalTree uses changed Prim's algorithm and then edits the obtained lineage tree.

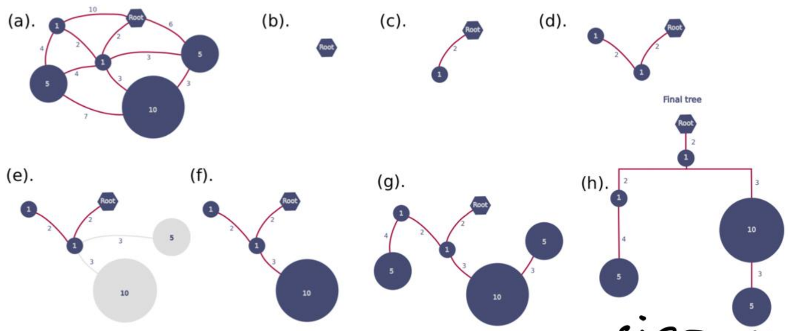
Prim's algorithm changes:

1. In order to decrease the time complexity of the algorithm, Clonaltree adds each node only once at the priority queue.
2. The original Prim's algorithm has only one objective function, which minimizes the sum of edge weights (cost).
   ClonalTree includes a second objective function to maximize genotype abundance.
   If a set of edges have the same weight, it will choose the one that connects nodes with high abundance.

Editing of the obtained lineage tree:

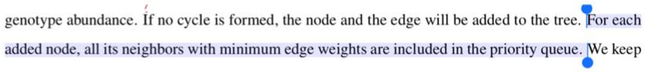
- add unobserved intermediate nodes to the tree
- detach/reattach sub-trees.

There is a little question on the next page (I tried to explain it with the pictures, but we can discuss it in the meeting) :)



size= 4

c) On this step we add noeud 1 to the tree and we should add all the neighbors of this noeud 1 or only one that is connected to another noeud 1 (with edge weight = 2) to the priority queue ?

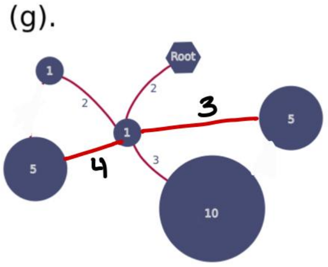


genotype abundance. If no cycle is formed, the node and the edge will be added to the tree. For each added node, all its neighbors with minimum edge weights are included in the priority queue. We keep

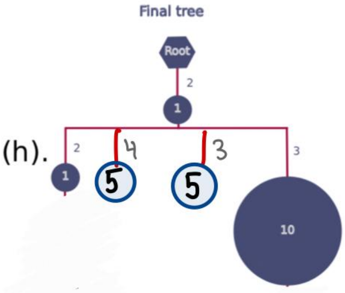
Because then, on the step E we will work also with the neighbors of this first noeud 1 but they have edge weight = 3 (more than 2).

And if add all neighbors, why we won't construct a tree with size 3 (less than size 4)?
I mean, for noeuds 5 we have the same edge weight if we link them with central noeud 1 and if we link them with noeud 1 which on the left side and noeud 10 on the right side? But if we connect it with the central noeud, the depth of tree will be less.

g)



h)



size=3