생물정보학 및 실습 2 실습 보고서 6

2021-20471 박명규
https://github.com/PMKYU98/BioInfo2

Q1 (ba7a)

트리에서 두 leaf 사이 거리 구하여 matrix 작성하기

---

visited boolean 리스트를 사용하여 한 node에서 시작하여 다른 node까지 도달하는 거리를 계산하고 None으로 초기화된 matrix에 채워넣었다. 그 후 None 값이 남아있는 node에 대해서 반복하여 matrix를 완성한다. 마지막으로 internal node에 해당하는 부분을 distance matrix에 제거하여 답을 작성하였다.

Internal node를 제거할 때 처음 입력받은 n 값을 사용하였는데, n개의 leaf라고 문제에서 언급하였기 때문에 가능한 방법이었다. 만약 이것이 없다면 node의 degree를 확인했다면 됐을 것 같다.

Q2 (ba7b)

각 leaf에 대해서 limb length 구하기

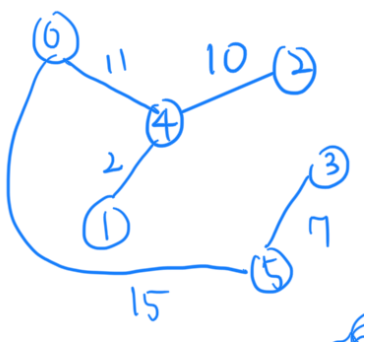
강의 영상에 나온 limb length theorem을 사용하여 모든 i, target leaf, k 쌍에 대해서 계산한 후 최소값을 구하여 출력하는 코드를 작성하였다.

Q3 (ba7c)

Simple tree reconstruction from additive matrix

처음엔 Rosalind에 주어진 pseudocode를 그대로 코드 작성하려 했더니, 중간 중간 변수와 함수를 새로 여럿 추가하면서 흐름을 따라가지 못해 잔뜩 꼬인 결과물이 나와버렸다. 오류가 군데 군데에서 발생하여 하나하나 고치기가 너무 힘들었다.

그래서 과감하게 싹 지워 버리고 distance matrix와 tree에 관련된 필요한 함수를 처음부터 작성하고 정리한 뒤 시작하였다. 이번에 발견된 문제는 internal node 사이의 관계였다. 수업 영상에서 leaf를 삭제하고 internal node를 새로 만들어 붙일때, 다른 leaf에서 거리를 구하여 internal node의 위치를 특정하는 방법이 소개되었었다. 이를 기반으로 코드를 구성하였더니 다음 문제가 발생했다.



Internal node인 4번과 5번 사이에 관계가 고려되지 않고 새로운 5번 node와 0번 node 사이의 거리가 15라는 정보만 사용해서 발생한 문제이다. 따라서 이 작업을 수행할때 i, n, k triplet을 찾은 후에 i와 k 사이에 이미 internal node가 있는지 (path를) 검사하여 새로운 internal node를 거리 정보에 맞게 그려 넣게 되었다.

그 후 테스트에 통과할 수 있었다.

Q4 (ba7d)

UPGMA

3번 문제를 해결할 때 사용한 tree 다루는 함수를 일부 가져와서 문제를 풀었다.

Distance matrix에서 merge를 수행하기 위해 proportional averaging을 진행해야 하는데, 수업 영상의 example 에서는 그런 경우가 나와있지 않아 처음에는 그저 2로 나누면 되는 줄 오해하고 강의 영상을 한참 돌려 보았다. (물론 식으로는 적혀있었다.) 이 부분을 해결하고 나서 다른 부분은 각 node의 age와 그들의 child의 age를 비교하여 tree의 edge를 구하고, 위에서 가져온 함수를 사용하여 tree를 구성했다.
