Elements of Computational Biology
Subject 15: Distance phylogenetics: UPGMA and NJ

Agata Radys, Paweł Cejrowski, Łukasz Myśliński

December 20, 2016

1 Algorithms

1.1 UPGMA (ang. Unweighted Pair Group Method with Arithmetic Mean)

Data: ultrametric matrix d for set L.

Listing 1: UPGMA pseudocode

```
clusters[|L|]
while (clusters.length > 1):
    calculate distances between clusters
        (sum of distances between cluster members
            divided by product of cluster cardinalities)
    find the lowest distance
    merge the closest clusters
```

1.2 NJ

Data: ultrametric matrix d for set L.

Q - matrix: $Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$

Distance from the pair members to the new node:

$d'(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}(\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k))$

$d(g,u) = d(f,g) - d'(f,u)$

Listing 2: NJ pseudocode

```
clusters[|L|]
while (number of clusters > 2):
    calculate Q-matrix
    find the lowest q-distance
    merge the q-closest clusters
    update distances
merge last 2 clusters
```

1.3 Comparing topology

2 Usage

Program was developed in Java programming language without any external libraries. All sources are available on Github and compiled using `Maven`.

Listing 3: Building project

```
git clone git@github.com:MiSSLab/BioComp15.git
mvn package
```

Created Java archive can be run using `JRE`. Sample data can be found in directory `resources/`.

Listing 4: Running project using data1.matrix

```
java -jar -Dfilename="resources/data1.matrix" \
    target/distance-phylogenetics-jar-with-dependencies.jar
```

3 Data formats

3.1 Input

Application requires `CSV` data format and quadratic matrix of distances.

Listing 5: Example data file content

```
a,b,c,d,e
0,8,8,5,3
8,0,3,8,8
8,3,0,8,8
5,8,8,0,5
3,8,8,5,0
```

Labels in header has to be lexicographically sorted, dense vector with every column matching `"[a-z]+"`.

3.2 Output

Resulting trees are printed in `ASCII`-art to the `STDOUT`.

Listing 6: UPGMA output

```
UPGMA(resources/data2.matrix)
[[33.0]]
|-[[28.0]]
|   |---- [[c]]
|   `---- [[d]]
`-[[22.0]]
    |-[[17.0]]
    |   |---- [[a]]
    |   `---- [[b]]
    `---- [[e]]
```

Despite the fact that NJ returns unrooted tree it is presented as a rooted one with particular node choosen as a root.

Listing 7: NJ output

```
NJ(resources/data2.matrix)
[[g]]
|---- [4.75-<-[a]]
|-[7.25-<-[f]]
|   |---- [11.0-<-[c]]
|   `---- [17.0-<-[d]]
`-[4.75-<-[h]]
    |---- [6.75-<-[b]]
    `---- [14.25-<-[e]]
```