

Neighbour-Joining Unit Tests

Hint: Many test values are taken from project Algorithms for Bioinformatics of Alexander Mattheis or the lectures.

Example from: https://en.wikipedia.org/wiki/Neighbor_joining (but other minima chosen)
Terms are taken from several internet sources (lecture slides).

Test 1

Input

| <i>D</i> | a | b | c | d | e |
|----------|---|---|----|----|---|
| a | 0 | 5 | 9 | 9 | 8 |
| b | | 0 | 10 | 10 | 9 |
| c | | | 0 | 8 | 7 |
| d | | | | 0 | 3 |
| e | | | | | 0 |

Iteration 1

Step 1: Calculate neighbor-joining matrix D^* from $N \times N$ distance matrix D

$$D_{i,j}^* = (N - 2) \cdot D_{i,j} - D_{i,J} - D_{J,j}$$

Hint: we know a similar formula from lecture Machine Learning in Life Sciences to find out, if two find out if a matrix is additive

where $D_{i,J} = \sum_{k=1}^N D_{i,k}$ is the total-distance (sum of distances from cluster i to all other clusters)

Step 1.1: Compute total distances

| <i>D</i> | a | b | c | d | e | Σ |
|----------|----|----|----|----|----|----------|
| a | 0 | 5 | 9 | 9 | 8 | 31 |
| b | 5 | 0 | 10 | 10 | 9 | 34 |
| c | 9 | 10 | 0 | 8 | 7 | 34 |
| d | 9 | 10 | 8 | 0 | 3 | 30 |
| e | 8 | 9 | 7 | 3 | 0 | 27 |
| Σ | 31 | 34 | 34 | 30 | 27 | |

Step 1.2: Calculate neighbor-joining matrix

| D^* | a | b | c | d | e |
|-------|---|-----|-----|-----|-----|
| a | | -50 | -38 | -34 | -34 |
| b | | | -38 | -34 | -34 |
| c | | | | -40 | -40 |
| d | | | | | -48 |
| e | | | | | |

first row:

$$D_{a,b}^* = (5 - 2) \cdot 5 - D_{a,B} - D_{A,b} = 3 \cdot 5 - 34 - 31 = -50$$

$$D_{a,c}^* = (5 - 2) \cdot 9 - D_{a,C} - D_{A,c} = 3 \cdot 9 - 34 - 31 = -38$$

$$D_{a,d}^* = (5 - 2) \cdot 9 - D_{a,D} - D_{A,d} = 3 \cdot 9 - 30 - 31 = -34$$

$$D_{a,e}^* = (5 - 2) \cdot 8 - D_{a,E} - D_{A,e} = 3 \cdot 8 - 27 - 31 = -34$$

second row:

$$D_{b,c}^* = (5 - 2) \cdot 10 - D_{b,c} - D_{B,c} = 30 - 34 - 34 = -38$$

$$D_{b,d}^* = (5 - 2) \cdot 10 - D_{b,d} - D_{B,d} = 30 - 30 - 34 = -34$$

$$D_{b,e}^* = (5 - 2) \cdot 9 - D_{b,e} - D_{B,e} = 27 - 27 - 34 = -34$$

third row:

$$D_{c,d}^* = (5 - 2) \cdot 8 - D_{c,d} - D_{C,d} = 24 - 30 - 34 = -40$$

$$D_{c,e}^* = (5 - 2) \cdot 7 - D_{c,e} - D_{C,e} = 21 - 27 - 34 = -40$$

fourth row:

$$D_{d,e}^* = (5 - 2) \cdot 3 - D_{d,e} - D_{D,e} = 9 - 27 - 30 = -48$$

Step 2: Find minimum element in D^* and create new cluster ij

| D^* | a | b | c | d | e |
|-------|---|-----|-----|-----|-----|
| a | | -50 | -38 | -34 | -34 |
| b | | | -38 | -34 | -34 |
| c | | | | -40 | -40 |
| d | | | | | -48 |
| e | | | | | |

$$D_{min} = D_{a,b} = -50 \quad \text{and} \quad ab = a \cup b$$

| D | ab | c | d | e |
|-----|----|---|---|---|
| ab | | | | |
| c | | | | |
| d | | | | |
| e | | | | |

Step 3: Recompute distances

Step 3.1: Pair-members and new cluster (distance in tree)

$$d(i, ij) = \frac{1}{2}(D_{i,j} - \Delta_{i,j}) = \frac{1}{2}\left(D_{i,j} - \frac{D_{i,j} - D_{I,j}}{N - 2}\right)$$

$$d(j, ij) = \frac{1}{2}(D_{i,j} + \Delta_{i,j}) = D_{i,j} - d(i, ij)$$

where $\Delta_{i,j} = \frac{D_{i,j} - D_{I,j}}{N - 2}$ is the ratio between the total difference and the remaining number of iterations

Hint: with these formulae you get better results than in UPGMA, because neighbor-joining does not assume the same evolution rate for both i.e. the branch lengths of merged taxa are different

$$d(a, ab) = \frac{1}{2}\left(D_{a,b} - \frac{D_{a,b} - D_{A,b}}{5 - 2}\right) = \frac{1}{2}\left(5 - \frac{34 - 31}{3}\right) = 2$$

$$d(b, ab) = \frac{1}{2}(D_{a,b} + \Delta_{a,b}) = \frac{1}{2}(5 + 1) = 3 = D_{a,b} - d(a, ab) = 5 - 2$$

Step 3.2: Remaining clusters and new node

$$D_{ij,k} = \frac{(D_{i,k} + D_{j,k} - D_{i,j})}{2}$$

| <i>D</i> | ab | c | d | e |
|----------|----|---|---|---|
| ab | 0 | 7 | 7 | 6 |
| c | | 0 | 8 | 7 |
| d | | | 0 | 3 |
| e | | | | 0 |

$$D_{ab,c} = \frac{1}{2}(D_{a,c} + D_{b,c} - D_{a,b}) = \frac{1}{2}(9 + 10 - 5) = 7$$

$$D_{ab,d} = \frac{1}{2}(D_{a,d} + D_{b,d} - D_{a,b}) = \frac{1}{2}(9 + 10 - 5) = 7$$

$$D_{ab,e} = \frac{1}{2}(D_{a,e} + D_{b,e} - D_{a,b}) = \frac{1}{2}(8 + 9 - 5) = 6$$

Iteration 2

| <i>D</i> | c | d | e | ab |
|----------|---|---|---|----|
| c | 0 | 8 | 7 | 7 |
| d | | 0 | 3 | 7 |
| e | | | 0 | 6 |
| ab | | | | 0 |

Step 1: Calculate neighbor-joining matrix D^* from $N \times N$ distance matrix D

Step 1.1: Compute total distances

| <i>D</i> | c | d | e | ab | Σ |
|----------|----|----|----|----|----------|
| c | 0 | 8 | 7 | 7 | 22 |
| d | 8 | 0 | 3 | 7 | 18 |
| e | 7 | 3 | 0 | 6 | 16 |
| ab | 7 | 7 | 6 | 0 | 20 |
| Σ | 22 | 18 | 16 | 20 | |

Step 1.2: Calculate neighbor-joining matrix

| D^* | c | d | e | ab |
|-------|---|-----|-----|-----|
| c | | -24 | -24 | -28 |
| d | | | -28 | -24 |
| e | | | | -24 |
| ab | | | | |

first row:

$$D_{ab,c}^* = (4 - 2) \cdot 7 - D_{ab,c} - D_{AB,c} = 14 - 20 - 22 = -28$$

$$D_{ab,d}^* = (4 - 2) \cdot 7 - D_{ab,d} - D_{AB,d} = 14 - 20 - 18 = -24$$

$$D_{ab,e}^* = (4 - 2) \cdot 6 - D_{ab,e} - D_{AB,e} = 12 - 20 - 16 = -24$$

second row:

$$D_{c,d}^* = (4 - 2) \cdot 8 - D_{c,D} - D_{C,d} = 16 - 18 - 22 = -24$$

$$D_{c,e}^* = (4 - 2) \cdot 7 - D_{c,E} - D_{C,e} = 14 - 16 - 22 = -24$$

third row:

$$D_{d,e}^* = (4 - 2) \cdot 3 - D_{d,E} - D_{D,e} = 6 - 16 - 18 = -28$$

Step 2: Find minimum element in D^* and create new cluster ij

| D^* | c | d | e | ab |
|-------|---|-----|-----|-----|
| c | | -24 | -24 | -28 |
| d | | | -28 | -24 |
| e | | | | -24 |
| ab | | | | |

Hint: Implementation chooses pair (d, e).

$$D_{min} = D_{d,e} = -28 \quad \text{and} \quad de = d \cup e$$

| D | c | ab | de |
|-----|---|----|----|
| c | | | |
| ab | | | |
| de | | | |

Step 3: Recompute distances

Step 3.1: Pair-members and new cluster (distance in tree)

$$d(d, de) = \frac{1}{2} \left(D_{d,e} - \frac{D_{d,E} - D_{D,e}}{4 - 2} \right) = \frac{1}{2} \left(3 - \frac{16 - 18}{2} \right) = 2$$

$$d(e, de) = \frac{1}{2} (D_{d,e} + \Delta_{d,e}) = \frac{1}{2} (3 - 1) = 1 = D_{d,e} - d(e, de) = 3 - 2$$

Step 3.2: Remaining clusters and new node

| D | c | ab | de |
|-----|---|----|----|
| c | 0 | 7 | 6 |
| ab | | 0 | 5 |
| de | | | 0 |

$$D_{de,c} = \frac{1}{2} (D_{d,c} + D_{e,c} - D_{d,e}) = \frac{1}{2} (8 + 7 - 3) = 6$$

$$D_{de,ab} = \frac{1}{2} (D_{d,ab} + D_{e,ab} - D_{d,e}) = \frac{1}{2} (7 + 6 - 3) = 5$$

Iteration 3

| D | c | ab | de |
|-----|---|----|----|
| c | 0 | 7 | 6 |
| ab | | 0 | 5 |
| de | | | 0 |

Step 1: Calculate neighbor-joining matrix D^* from $N \times N$ distance matrix D

Step 1.1: Compute total distances

| D | c | ab | de | Σ |
|----------|----|----|----|----------|
| c | 0 | 7 | 6 | 13 |
| ab | 7 | 0 | 5 | 12 |
| de | 6 | 5 | 0 | 11 |
| Σ | 13 | 12 | 11 | |

Step 1.2: Calculate neighbor-joining matrix

| D^* | c | ab | de |
|-------|---|----|----|
| c | | | |
| ab | | | |
| de | | | |

first row:

$$D_{c,ab}^* = (3 - 2) \cdot 7 - D_{c,AB} - D_{C,ab} = 7 - 12 - 13 = -18$$

$$D_{c,de}^* = (3 - 2) \cdot 6 - D_{c,DE} - D_{C,de} = 6 - 11 - 13 = -18$$

second row:

$$D_{ab,de}^* = (3 - 2) \cdot 5 - D_{ab,DE} - D_{AB,de} = 5 - 11 - 12 = -18$$

Step 2: Find minimum element in D^* and create new cluster ij

| D^* | c | ab | de |
|-------|---|-----|-----|
| c | | -18 | -18 |
| ab | | | -18 |
| de | | | |

$$D_{min} = D_{c,ab} = -18 \quad \text{and} \quad cab = c \cup ab$$

| D | de | cab |
|-----|----|-----|
| de | | |
| cab | | |

Step 3: Recompute distances

Step 3.1: Pair-members and new cluster (distance in tree)

$$d(c, cab) = \frac{1}{2} \left(D_{c,ab} - \frac{D_{c,AB} - D_{C,ab}}{3 - 2} \right) = \frac{1}{2} \left(7 - \frac{12 - 13}{1} \right) = 4$$

$$d(ab, cab) = \frac{1}{2} (D_{c,ab} + \Delta_{c,ab}) = \frac{1}{2} (7 + (-1)) = 3 = D_{c,ab} - d(c, cab) = 7 - 4$$

Step 3.2: Remaining clusters and new node

| D^* | de | cab |
|-------|----|-----|
| de | 0 | 2 |
| cab | | 0 |

$$D_{cab,de} = \frac{1}{2}(D_{c,de} + D_{ab,de} - D_{c,ab}) = \frac{1}{2}(6 + 5 - 7) = 2$$

Final matrix

| D^* | decab |
|-------|-------|
| decab | 0 |

$$d(i, ij) = \frac{1}{2}(D_{i,j} - \Delta_{i,j}) = \frac{1}{2}\left(D_{i,j} - \frac{D_{i,J} - D_{I,j}}{N - 2}\right)$$

theoretical stuff: (only to see correctness)

$$d(de, decab) = \frac{1}{2}\left(D_{de,cab} - \frac{D_{de,CAB} - D_{DE,cab}}{2 - 2}\right) = \frac{1}{2}\left(2 - \frac{2 - 2}{2 - 2}\right) = \frac{1}{2}(2 - 0) = 1$$

$$d(cab, decab) = \frac{1}{2}(D_{c,ab} + 0) = \frac{1}{2}(2 + 0) = 1 = D_{de,cab} - d(de, decab) = 2 - 1$$

Hint: $0/0 := 0$ in this case

computation:

store last computed distance and divide by 2

but:

we want same result as on Wikipedia and no floating points and so we set

$$d(de, decab) = 2 \quad \text{and} \quad d(cab, decab) = 0$$

(set on last computed distance)

Final-Output

```
((d:2,e:1):1,(c:4,(a:2,b:3):3):1);
```

Wiki result: `((d:2,e:1):2,(c:4,(a:2,b:3):3):0);`