Neighbour-Joining Unit Tests

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

Example from: https://en.wikipedia.org/wiki/Neighbor_joining

Test 1

Input

D	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

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

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

Step 1.1: Compute total distances

D	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					

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

D	ab	c	d	e
ab				
c				
d				
e				

Step 3: Recompute distances

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

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

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

where $\Delta_{i,j} = \frac{D_{i,J} - D_{I,j}}{N-2}$ is the total-distance difference

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

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

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

Step 3.2: Remaining clusters and new node

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

D	**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

D	**ab**	**c**	**d**	**e**
ab	0	7	7	6
c		0	8	7
d			0	3
e				0

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

Step 1.1: Compute total distances

D	**ab**	**c**	**d**	**e**	**Σ**
ab	0	7	7	6	20
c	7	0	8	7	22
d	7	8	0	3	18
e	6	7	3	0	16
Σ	20	22	18	16	

Step 1.2: Calculate neighbor-joining matrix

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

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

$\boldsymbol{D^*}$	ab	c	d	e
ab		-28	-24	-24
c			-24	-24
d				-28
e				

$D_{min} = D_{ab,c} = -28$ and $abc = ab \cup c$

$\boldsymbol{D}$	abc	d	e
abc			
d			
e			

Step 3: Recompute distances

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

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

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

where $\Delta_{i,j} = \frac{D_{i,J} - D_{I,j}}{N-2}$ is the total-distance difference

$$d(ab,abc) = \frac{1}{2}\left(D_{ab,c} - \frac{D_{ab,C} - D_{AB,c}}{4-2}\right) = \frac{1}{2}\left(7 - \frac{22-20}{2}\right) = 3$$

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

Step 3.2: Remaining clusters and new node

$\boldsymbol{D}$	abc	d	e
abc	0	4	3
d		0	3
e			0

$$D_{abc,d} = \frac{1}{2}\left(D_{ab,d} + D_{c,d} - D_{ab,c}\right) = \frac{1}{2}(7 + 8 - 7) = 4$$

$$D_{abc,e} = \frac{1}{2}\left(D_{ab,e} + D_{c,e} - D_{ab,c}\right) = \frac{1}{2}(6 + 7 - 7) = 3$$

Iteration 3

D	abc	d	e
abc	0	4	3
d		0	3
e			0

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

Step 1.1: Compute total distances

D	abc	d	e	Σ
abc	0	4	3	7
d	4	0	3	7
e	3	3	0	6
Σ	7	7	6	

Step 1.2: Calculate neighbor-joining matrix

D^*	abc	d	e
abc		-10	-10
d			-10
e			

first row:

$$D^*_{abc,d} = (3-2) \cdot 4 - D_{abc,D} - D_{ABC,d} = 4 - 7 - 7 = -10$$

$$D^*_{abc,e} = (3-2) \cdot 3 - D_{abc,E} - D_{ABC,e} = 3 - 6 - 7 = -10$$

second row:

$$D^*_{d,e} = (3-2) \cdot 3 - D_{d,E} - D_{D,e} = 3 - 6 - 7 = -10$$

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

D^*	abc	d	e
abc		-10	-10
d			-10
e			

$D_{min} = D_{abc,d} = -10$ and $abc = abcd \cup d$

D	abcd	e
abcd		
e		

Step 3: Recompute distances

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

$$d(abc, abcd) = \frac{1}{2}\left(D_{abc,d} - \frac{D_{abc,D} - D_{ABC,d}}{3-2}\right) = \frac{1}{2}\left(4 - \frac{7-7}{1}\right) = 2$$

$$d(d, abcd) = \frac{1}{2}\left(D_{abc,d} + \Delta_{abc,d}\right) = \frac{1}{2}(4+0) = 2 = D_{abc,d} - d(abc, abcd) = 4 - 2$$

because last round:

$$d(e, abcd) = D_{abc,e} - d(abc, abcd) = 3 - 2 = 1$$

Final-Output

```
((((a:2.0,b:3.0):3,c:4):2,d:2):0,e:1);
```