CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Feb 26

- BWT and FM-Index runtime
- Recap UPGMA
- Neighbor-Joining (NJ)

Notes:

- Office hours TODAY 3-5pm
- Create “cheat-sheet” for midterm
- Lab 5 released right after midterm

BWT and FM-Index runtime

BWT and FM-Index runtime

- Building the FM-Index: dominated by sorting the rotations (cyclic permutations). There are actually linear time algorithms for this, but we will assume a standard sorting algorithm so O(G log G) where G is the length of the reference.

- Creating M, occ, and A are all linear.

- Pattern matching from Lab 4?

BWT and FM-Index runtime

- Building the FM-Index: dominated by sorting the rotations (cyclic permutations). There are actually linear time algorithms for this, but we will assume a standard sorting algorithm so O(G log G) where G is the length of the reference.

- Creating M, occ, and A are all linear.

- Pattern matching: O(n*L)
  - *Linear in the length of the pattern (L)*
  - *Linear in the number of patterns/reads (n)*
  - *Constant in the length of the genome (G)*

Recap UPGMA

Recap questions: discuss with a partner

1) How do we define a tree metric?

2) True or False: every dissimilarity map is a tree metric.

3) How do we define an ultrametric? (both theoretically and intuitively)

4) True or False: every tree metric is an ultrametric.

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

6) What two biological factors might make ultrametric trees an unrealistic assumption?

Recap questions: discuss with a partner

1) How do we define a tree metric?

   A dissimilarity map δ is a tree metric if the sum of the edge weights between A and B is equal to $\delta(A,B)$.

2) True or False: every dissimilarity map is a tree metric.

3) How do we define an ultrametric? (both theoretically and intuitively)

4) True or False: every tree metric is an ultrametric.

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

6) What two biological factors might make ultrametric trees an unrealistic assumption?

Recap questions: discuss with a partner

1) How do we define a tree metric?

   A dissimilarity map δ is a tree metric if the sum of the edge weights between A and B is equal to $\delta(A,B)$.

2) True or False: every dissimilarity map is a tree metric. False!

3) How do we define an ultrametric? (both theoretically and intuitively)

4) True or False: every tree metric is an ultrametric.

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

6) What two biological factors might make ultrametric trees an unrealistic assumption?

Recap questions: discuss with a partner

1) How do we define a tree metric?

   A dissimilarity map δ is a tree metric if the sum of the edge weights between A and B is equal to $\delta(A,B)$.

2) True or False: every dissimilarity map is a tree metric. False!

3) How do we define an ultrametric? (both theoretically and intuitively)
   - The distance from the root to each leaf is the same.
   - 3-point condition: For all distinct A,B,C, $\delta(A,B) \leq \max\{\delta(A,C), \delta(B,C)\}$
   - Intuitively this means out of these three distances, two are equal and one is less.

4) True or False: every tree metric is an ultrametric.

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

6) What two biological factors might make ultrametric trees an unrealistic assumption?

Recap questions: discuss with a partner

1) How do we define a tree metric?

   A dissimilarity map δ is a tree metric if the sum of the edge weights between A and B is equal to $\delta(A,B)$.

2) True or False: every dissimilarity map is a tree metric. False!

3) How do we define an ultrametric? (both theoretically and intuitively)
   - The distance from the root to each leaf is the same.
   - 3-point condition: For all distinct A,B,C, $\delta(A,B) \leq \max\{\delta(A,C), \delta(B,C)\}$
   - Intuitively this means out of these three distances, two are equal and one is less.

4) True or False: every tree metric is an ultrametric. False!

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

6) What two biological factors might make ultrametric trees an unrealistic assumption?

Recap questions: discuss with a partner

1) How do we define a tree metric?

   A dissimilarity map δ is a tree metric if the sum of the edge weights between A and B is equal to $\delta(A,B)$.

2) True or False: every dissimilarity map is a tree metric. False!

3) How do we define an ultrametric? (both theoretically and intuitively)
   - The distance from the root to each leaf is the same.
   - 3-point condition: For all distinct A,B,C, $\delta(A,B) \leq \max\{\delta(A,C), \delta(B,C)\}$
   - Intuitively this means out of these three distances, two are equal and one is less.

4) True or False: every tree metric is an ultrametric. False!

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

   We are assuming that evolution is proportional to time (i.e. the “molecular clock” assumption).

6) What two biological factors might make ultrametric trees an unrealistic assumption?

Recap questions: discuss with a partner

1) How do we define a tree metric?

   A dissimilarity map δ is a tree metric if the sum of the edge weights between A and B is equal to $\delta(A,B)$.

2) True or False: every dissimilarity map is a tree metric. False!

3) How do we define an ultrametric? (both theoretically and intuitively)

   - The distance from the root to each leaf is the same.
   - 3-point condition: For all distinct A,B,C, $\delta(\mathrm{A},\mathrm{B}) \leq \max\{\delta(\mathrm{A},\mathrm{C}), \delta(\mathrm{B},\mathrm{C})\}$
   - Intuitively this means out of these three distances, two are equal and one is less.

4) True or False: every tree metric is an ultrametric. False!

5) What biological assumption(s) are we making when creating ultrametric trees like those produced by UPGMA?

   We are assuming that evolution is proportional to time (i.e. the “molecular clock” assumption).

6) What two biological factors might make ultrametric trees an unrealistic assumption?

   - Mutation rates differ significantly across species.
   - Natural selection (both positive and negative) can change the tempo of evolution.

Bonus questions

- In what scenarios is an ultrametric tree likely a GOOD assumption?

- What is the runtime of UPGMA in terms of the number of samples *n*?

Bonus questions

- In what scenarios is an ultrametric tree likely a GOOD assumption?

  If our samples are from the same species or population, then there has likely been the same amount of evolution from the root to each leaf, so it is (usually) okay to assume time and evolution are proportional.

- What is the runtime of UPGMA in terms of the number of samples *n*?

Bonus questions

- In what scenarios is an ultrametric tree likely a GOOD assumption?

If our samples are from the same species or population, then there has likely been the same amount of evolution from the root to each leaf, so it is (usually) okay to assume time and evolution are proportional.

- What is the runtime of UPGMA in terms of the number of samples *n*?

During each iteration we must do $O(n^2)$ work to compute the new matrix of distances. We merge two nodes each iteration, so we have $O(n)$ iterations total. This gives us a runtime of $O(n^3)$, which can be improved by reusing some distances from the previous iteration.

Next phylogenetic tree method: Neighbor-Joining (NJ)

Notes about UPGMA vs NJ

- NJ was first described in 1987 by Saitou and Nei. Their paper currently has 50,199 citations (an average of over 4 citations a day for the last 30 years!)

Notes about UPGMA vs NJ

- NJ was first described in 1987 by Saitou and Nei. Their paper currently has 50,199 citations (an average of over 4 citations a day for the last 30 years!)

- Both UPGMA and NJ are greedy, polynomial-time clustering algorithms that produce edge weights as well as binary tree topologies.

Notes about UPGMA vs NJ

- NJ was first described in 1987 by Saitou and Nei. Their paper currently has 50,199 citations (an average of over 4 citations a day for the last 30 years!)

- Both UPGMA and NJ are greedy, polynomial-time clustering algorithms that produce edge weights as well as binary tree topologies.

- NJ creates unrooted trees (direction of evolution is not apparent on all branches), while UPGMA creates rooted trees.

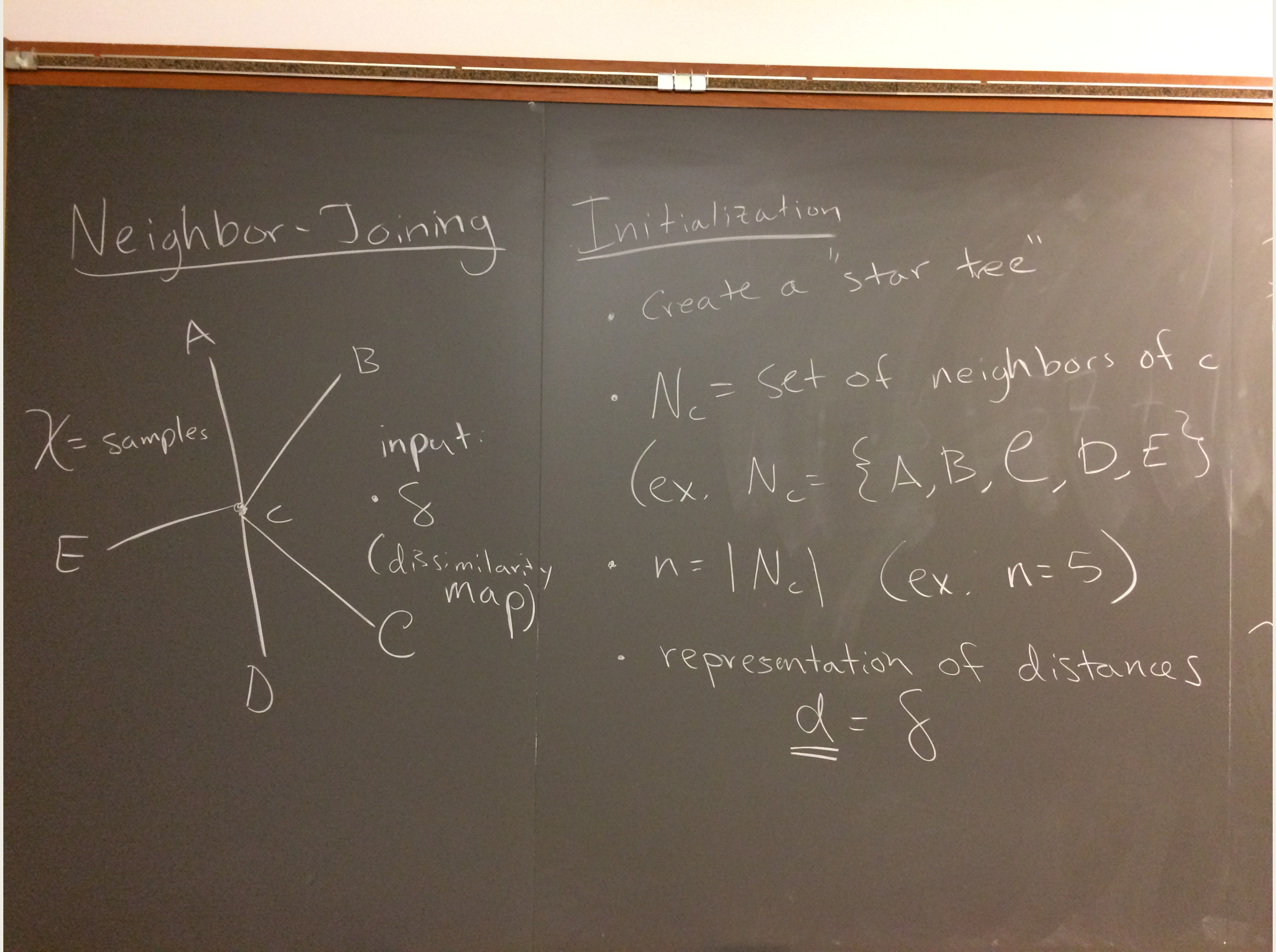
Notes about UPGMA vs NJ

- NJ was first described in 1987 by Saitou and Nei. Their paper currently has 50,199 citations (an average of over 4 citations a day for the last 30 years!)

- Both UPGMA and NJ are greedy, polynomial-time clustering algorithms that produce edge weights as well as binary tree topologies.

- NJ creates unrooted trees (direction of evolution is not apparent on all branches), while UPGMA creates rooted trees.

- NJ is much better for representing multi-species evolution and in general creates more realistic trees that better approximate the original dissimilarity map.



Neighbor-Joining

$\mathcal{X}$ = samples

input:

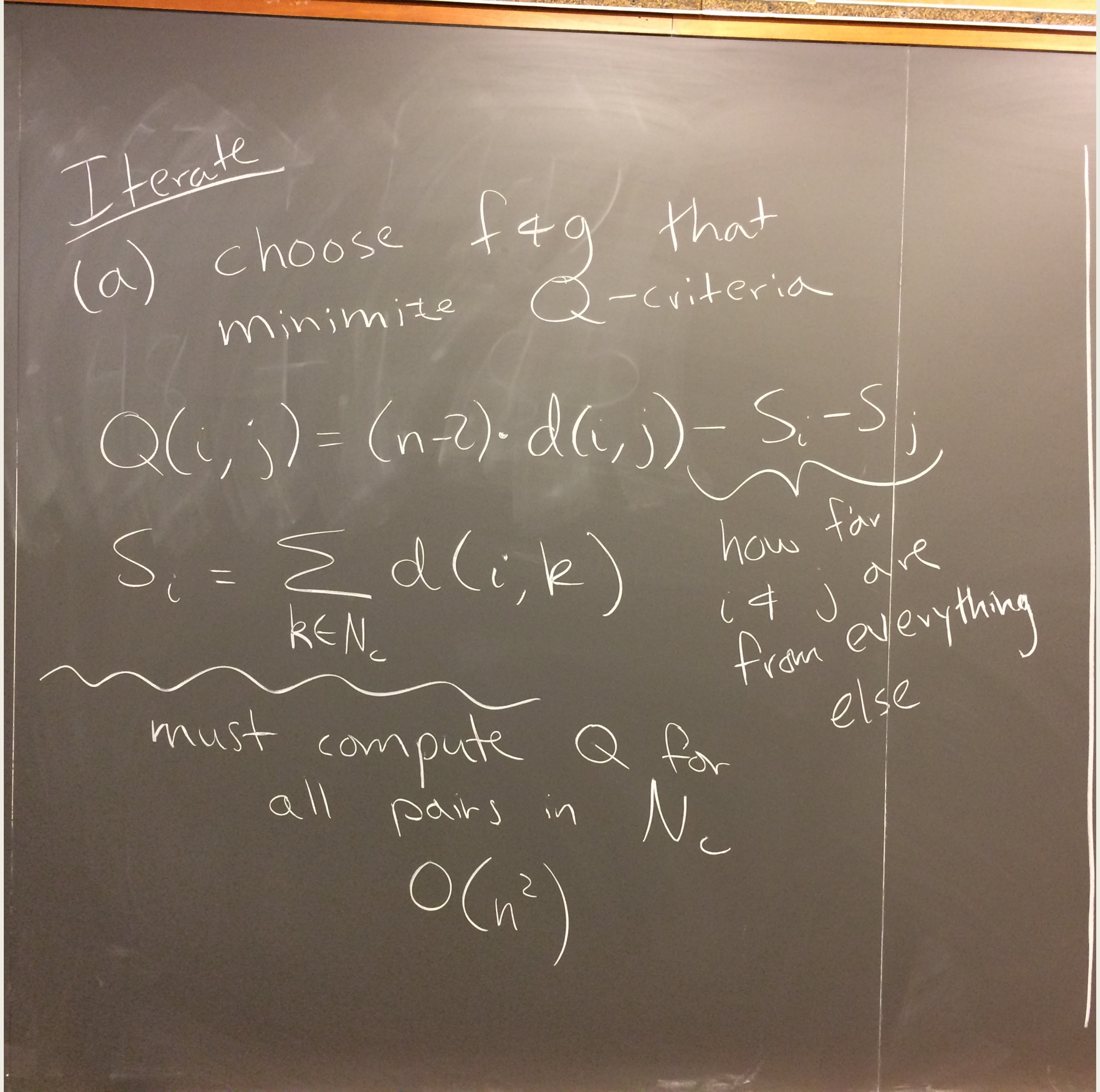
- δ (dissimilarity map)

Initialization

- Create a "star tree"
- N_c = set of neighbors of c

  (ex. $N_c = \{A, B, C, D, E\}$)
- $n = |N_c|$ (ex. $n = 5$)
- representation of distances

  $\underline{\underline{d}} = \delta$



Iterate

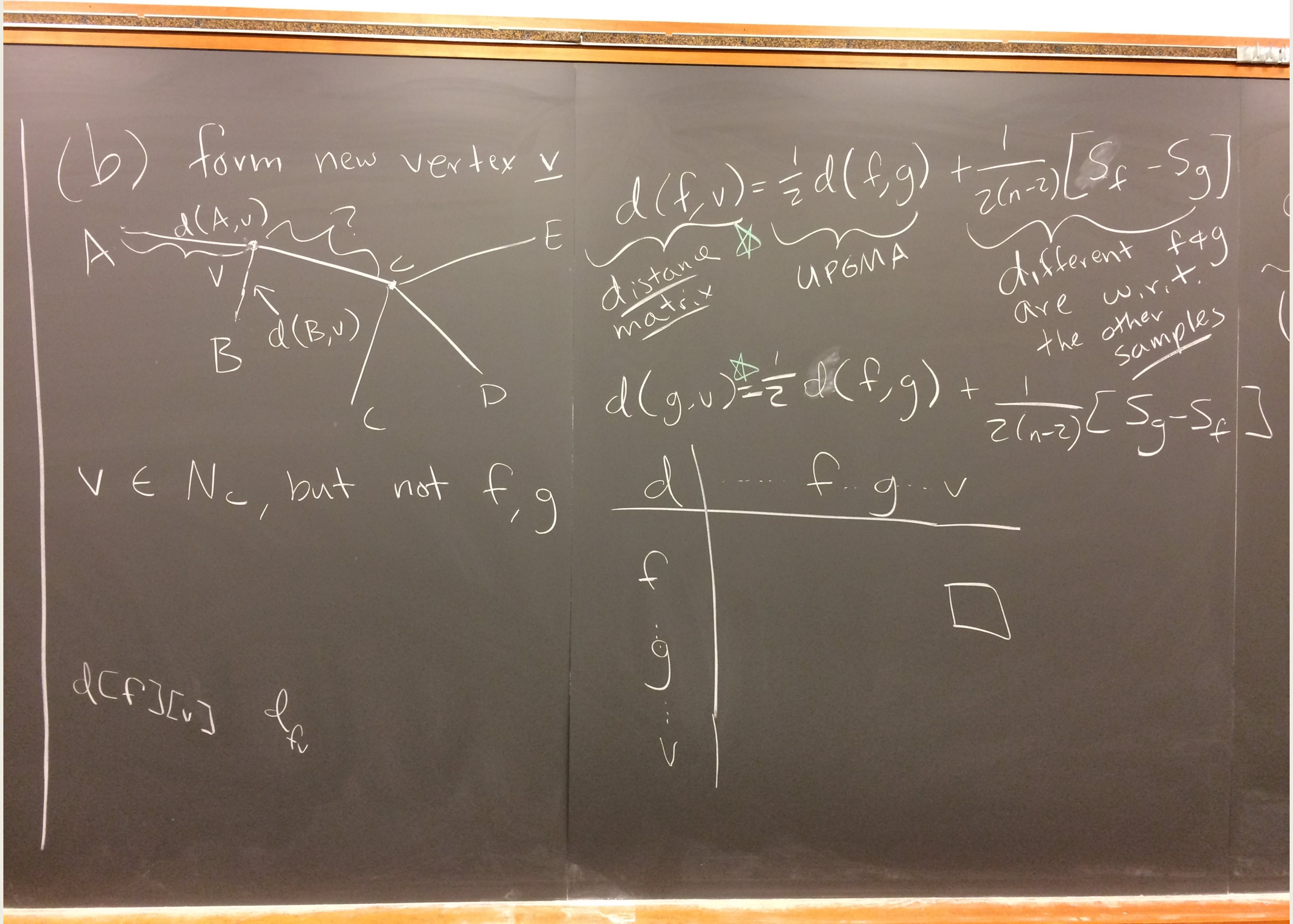
(a) choose f & g that minimize Q-criteria

$$Q(i,j) = (n-2)\cdot d(i,j) \underbrace{- S_i - S_j}_{\text{how far } i \,\&\, j \text{ are from everything else}}$$

$$S_i = \sum_{k \in N_c} d(i,k)$$

must compute Q for all pairs in N_c

$O(n^2)$



(b) form new vertex $\underline{v}$

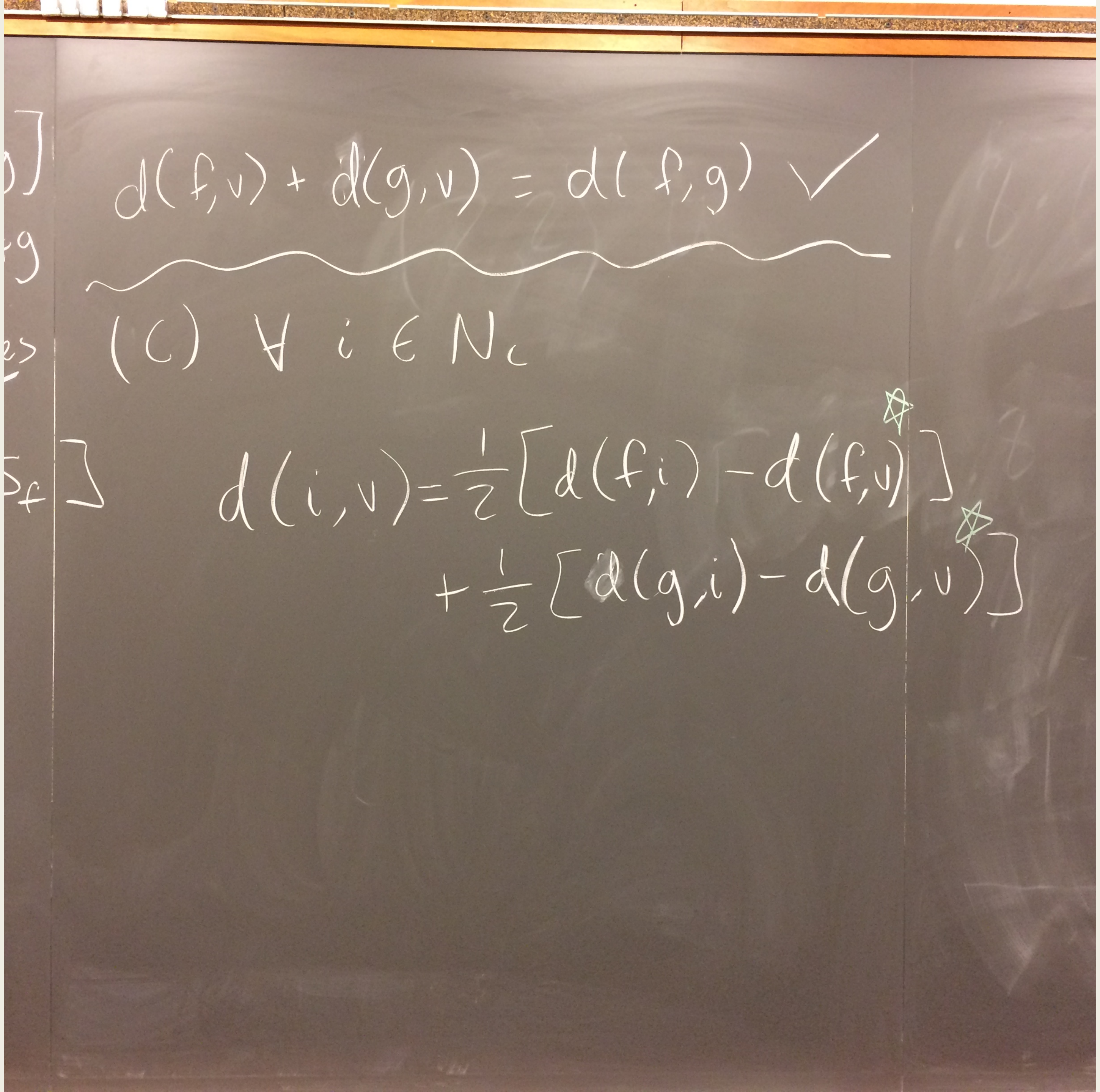
$v \in N_c$, but not f, g

d[f][v] d_{fv}

$$\underbrace{d(f,v)}_{\text{distance matrix}} \overset{\star}{=} \underbrace{\frac{1}{2}d(f,g)}_{\text{UPGMA}} + \underbrace{\frac{1}{2(n-2)}[S_f - S_g]}_{\text{different f\&g are w.r.t. the other samples}}$$

$$d(g,v) \overset{\star}{=} \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}[S_g - S_f]$$

d	... f ... g ... v
f	□
⋮ g	
⋮ v	



)]

g

es

S_f]

$$d(f,v) + d(g,v) = d(f,g) \quad \checkmark$$

(c) $\forall\, i \in N_c$

$$d(i,v) = \frac{1}{2}\left[d(f,i) - d(f,v)^{\star}\right] + \frac{1}{2}\left[d(g,i) - d(g,v)^{\star}\right]$$