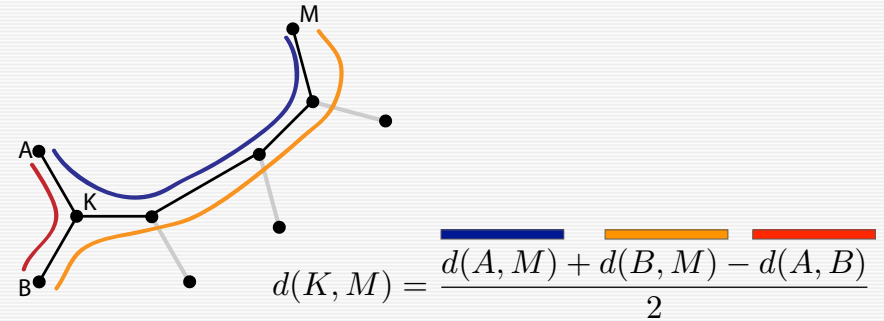


NEIGHBOR JOINING (NJ)

- Saitou and Nei (1987)
- The idea is very similar to clustering
 - Find the two 'nearest' sequences
 - Replace them with their parent, recompute distances and iterate until only two sequences remain

NJ IDEA 1

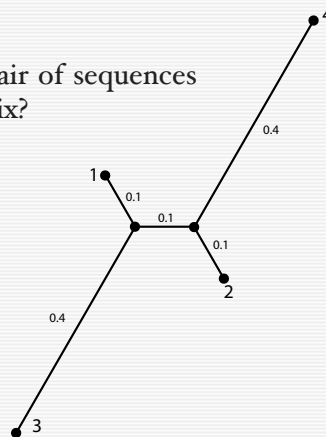
- Computing the distances to the parent of two neighbors.
- Consider computing the distance from two leaves with the same parent to any other leaf.



IDEA 2

- How to decide which two nodes are nearest neighbors, having access to nothing but pairwise distances?
- Is it enough to simply consider the pair of sequences with the smallest distance in the matrix?

	1	2	3	4
1	0	0.3	0.5	0.6
2		0	0.6	0.5
3			0	0.9
4				0



IDEA 2 (CONT.)

- **NOT** enough to look just for the shortest distance pair
- **IS** enough to look at the sequences that are both maximally close to each other and maximally far from the rest of the sequences.
- Define (L is the current number of leaves):

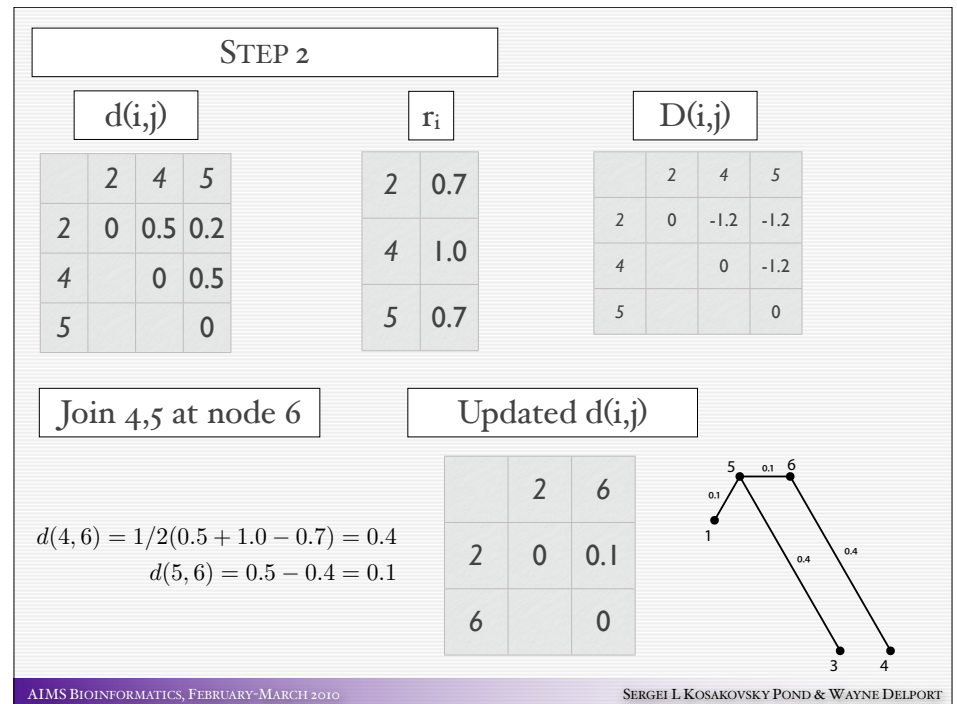
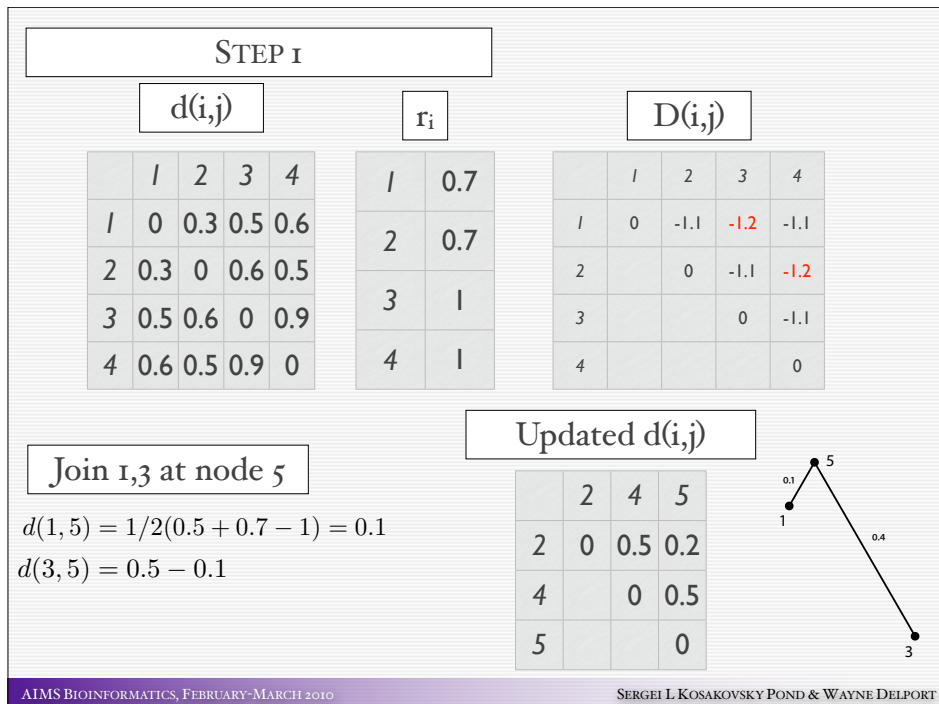
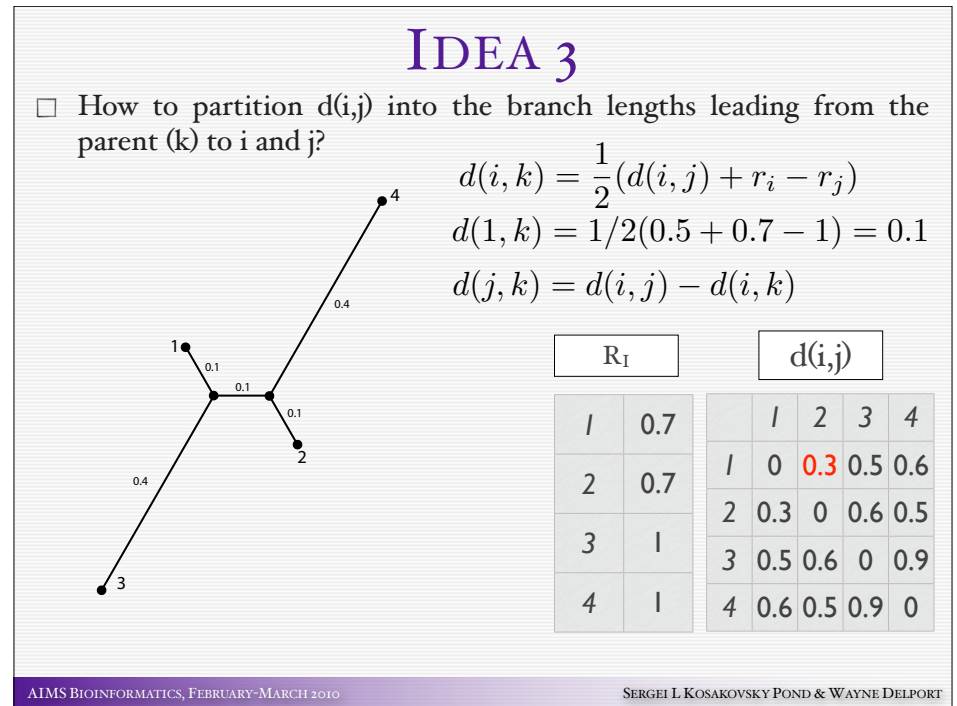
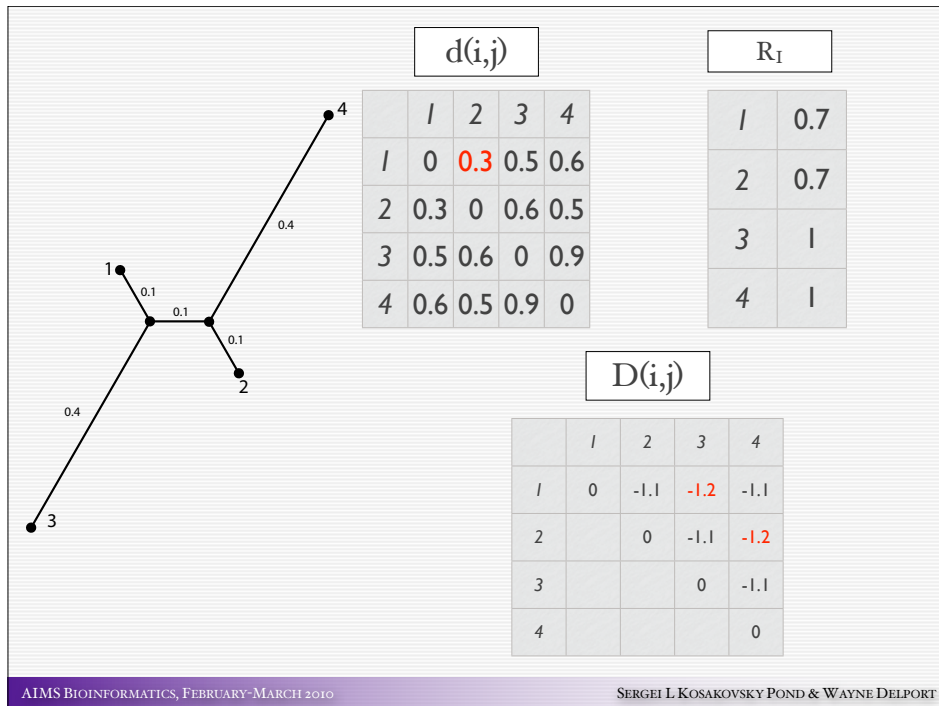
AVERAGE DISTANCE TO OTHER BRANCHES

$$r_i = \frac{1}{|L| - 2} \sum_{k \in L} d(i, k)$$

RE-WEIGHTED DISTANCES

$$D(i, j) = d(i, j) - (r_i + r_j)$$

- The pair with the smallest $D(i, j)$ are closest neighbors.

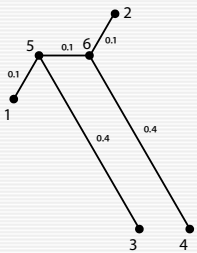


STEP 3

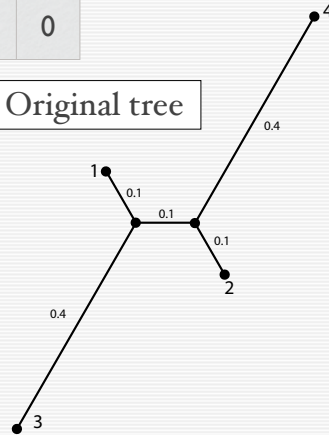
Join the remaining 2 nodes (6 and 2)

	2	6
2	0	0.1
6		0

NJ tree



Original tree



BIOLOGY IS MESSY

- Comparisons of biological sequences very rarely generate additive distance matrices
- NJ can be applied to non-additive matrices and generally performs quite well – many advanced tree search programs take NJ trees as good starting points, for example

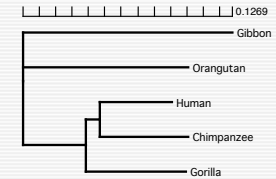
Genetic distances

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	0	0.0978173	0.115615	0.187822	0.21524
Chimpanzee		0	0.12003	0.204948	0.228186
Gorilla			0	0.198166	0.227662
Orangutan				0	0.227143
Gibbon					0

NJ tree distances

	Human	Chimpanzee	Gorilla	Orangutan	Gibbon
Human	0	0.0978173	0.11296	0.190827	0.217544
Chimpanzee		0	0.122685	0.200553	0.22727
Gorilla			0	0.199555	0.226272
Orangutan				0	0.227143
Gibbon					0

NJ tree



NON-ADDITIVE MATRICES

- One can try to find the tree that minimizes an error between the distance matrix $d(i,j)$ and tree-induced pairwise distances $T(i,j)$
 - For example – least squares

$$\min_T \sum_{i,j} (d(i,j) - T(i,j))^2$$

- This problem is NP-hard (need to look at all trees, potentially)
- Difficult to quantify how one tree compares to the other (e.g. if one achieves error 0.1 and the other - 0.105 – are they really that different?)

ALGORITHMIC VS OPTIMALITY BASED TREE RECONSTRUCTION

- Neighbor joining (and some other methods) are **algorithmic** – they produce a single tree from the input.
 - Advantage: fast
 - Disadvantage: have no idea how the found tree compares to the rest $(2N-5)!! - 1$ trees.
- Optimality based criterion search states:
 - Any candidate tree, T and be assigned a score, $s(T)$
 - We seek to minimize (or maximize) $s(T)$ over all possible trees
 - Advantage: compares many trees, gives one an idea of how good the proposed solution is
 - Disadvantage: slow (many trees), still need to explore a combinatorial set of possible solutions.