# Neighbor Joining Algorithm and its applications in Phylogenetic tree construction

Tiansheng Sun
Department of Computer Science, Middlebury College

## Introduction

Where did human come from?

When did HIV jump from primates to humans?


**A phylogenetic tree can answer these questions.**

Phylogenetic trees tell us evolutionary relationships among different entities by looking at their genetic similarities and differences.

| SPECIES | ALIGNMENT |
|---|---|
| Chimp | ACGTAGGCCT |
| Human | ATGTAAGACT |
| Seal | TCGAGAGCAC |
| Whale | TCGAAAGCAT |

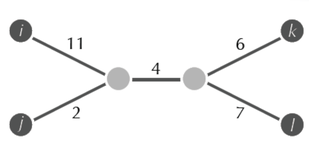
DISTANCE MATRIX

| | Chimp | Human | Seal | Whale |
|---|---|---|---|---|
| Chimp | 0 | 3 | 6 | 4 |
| Human | 3 | 0 | 7 | 5 |
| Seal | 6 | 7 | 0 | 2 |
| Whale | 4 | 5 | 2 | 0 |

Seems like we can contruct a distance matrix where the distance between two spieces is the number of differing symbols between them.
We can then use the distance matrix to make a phylogenetic tree by looking at the smallest value using some algorithms similar to hierarchical clustering?

DISTANCE MATRIX

| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 3 | 6 | 4 |
| B | 3 | 0 | 7 | 5 |
| C | 6 | 7 | 0 | 2 |
| D | 4 | 5 | 2 | 0 |




*Source:* Compeau, P. and Pevzner, P (2018). Bioinfomatics Algorithms: An Active Learning Approach, *364 Active Learning Publishers*

This example suggests that although it looks intuitive that the smallest element in the distance matrix correspond to a pair of neighbors in a revolutionary tree, this is *not* always the case.
Therefore, we need another method to deal with this problem.

*The data for this project is collected from NCBI, access from MEGA. MEGA is also used to align genomes for this project. Graph is drawn manually.*

**Acknowledgement:**

Thank you Professor Linderman for the support of this project.

**References:**

Compeau, P. and Pevzner, P(2018). Bioinfomatics Algorithms: An Active Learning Approach, *353-415 Active Learning Publishers*

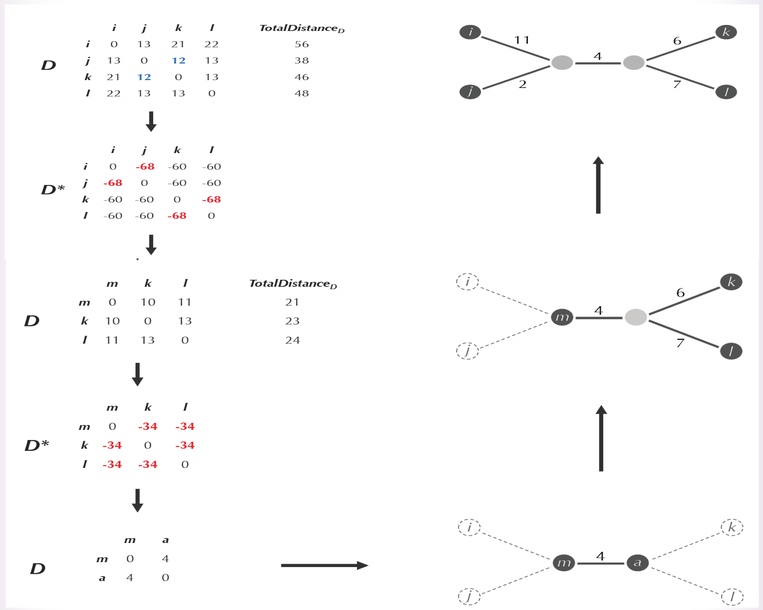
Seitou, N. and Nei M.(1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4(4), 406–425.*

Polygenetic Tree Practical Problems. Acess from *http://bioinformaticsalgorithms.com/software_challenges/Evolution%20Practical%20Challenge.pdf*

## Neighbor Joining Algorithm

Neighbor Joining Algorithm was developed by Naruya Saitou and Masatoshi Nei in 1987 for revolutionary tree construction. It has been sited for over 50,000 times, and is one of the most quoted paper cited in all of science (Compeau and Pevzner, 2018).

Since finding a minimum element in a distance matrix D does not guarantee a pair of neighbors, Neighbor Joining Algorithm transforms distance matrix D into a different matrix D* whose minimum element will definitely yields a pair of neighbors.




*Source:* Compeau, P. and Pevzner, P (2018). Bioinfomatics Algorithms: An Active Learning Approach, *379. Active Learning Publishers*

### Steps:

1. From n*n matrix D constructs n*n matrix D*, such that:
$D^*_{i,j} = (n-2) * D_{i,j} - \text{TotalDistance}_D(i) - \text{TotalDistance}_D(J)$
2. Find the minimum value of D* at row *a* and column *b* and transform the initial n*n matrix into a new (n-1)*(n-1) matrix by replacing *a* and *b* with a new leaf *c*.
3. Calculate the limb length of the new leaf *c* to the deleted leaves *a* and *b, as paired neighbors,* using these functions:
$\Delta_{a,b} = (\text{TotalDistance}_D(i) - \text{TotalDistance}_D(j))/(n-2)$
$\text{LimbLength}(a,c) = 0.5 * (D_{a,b} + \Delta_{a,b})$
$\text{LimbLength}(b,c) = 0.5 * (D_{a,b} - \Delta_{a,b})$
4. Repeat the first three steps until we get a 2*2 distance matrix D which corresponds to a tree with a single edge.
5. Add all pairs of neighbors to our polygenetic tree.

## Biological applications

### Which animal gave us SARs?

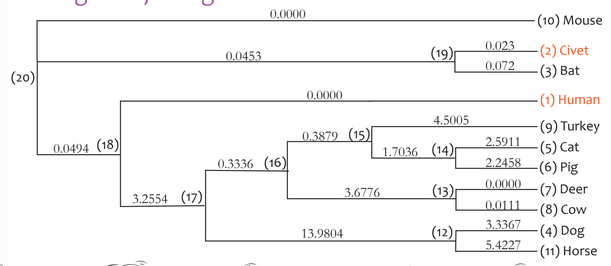
Data

| Accession Number | Animal |
|---|---|
| AY274119 | Human |
| AY304486 | Civet |
| KY417144 | Bat |
| GQ477367 | Dog |
| MG605090 | Cat |
| KC242792 | Pig |
| FJ425190 | Deer |
| DB811784 | Cow |
| EU022526 | Turkey |
| DQ497008 | Mouse |
| LC061273 | Horse |

Output

| | |
|---|---|
| 1(Human)->18:0.000 | 14->15:1.7036 |
| 2(Civet)->19:0.023 | 15->9(Turkey):4.5005 |
| 3(Bat)->19:0.072 | 15->14:1.7036 |
| 4(Dog)->12:3.3367 | 15->16:0.3879 |
| 5(Cat)->14:2.5911 | 16->13:3.6776 |
| 6(Pig)->14:2.2458 | 16->15:0.3879 |
| 7(Deer)->13:0.000 | 16->17:0.3336 |
| 8(Cow)->13:0.0111 | 17->12:13.9804 |
| 9(Turkey)->15:4.5005 | 17->16:0.3336 |
| 10(Mouse)->20:0.000 | 17->18:3.2554 |
| 11(Horse)->12:5.4227 | 18->1(Human):0.000 |
| 12->4(Dog):3.3367 | 18->17:3.2554 |
| 12->11(Horse):5.4227 | 18->20:0.0494 |
| 12->17:13.9804 | 19->2(Civet):0.023 |
| 13->7(Deer):0.000 | 19->3(Bat):0.072 |
| 13->8(Cow):0.0111 | 19->20:0.0453 |
| 13->16:3.6776 | 20->10(Mouse):0.000 |
| 14->5(Cat):2.5911 | 20->18:0.0494 |
| 14->6(Pig):2.2458 | 20->19:0.0453 |

The neighbor-joining tree of coronaviruses from different animals




### Which species of Ebola viruses caused the 2012 Ebola outbreak?

Data

| Accession Number | Virus Species | Date |
|---|---|---|
| KJ660348 | ???? | 2014 |
| FJ217161 | BDBV | 2007 |
| KC545393 | BDBV | 2012 |
| AF272001 | EBOV | 1976 |
| KC242792 | EBOV | 1994 |
| KC589025 | SUDV | 2012 |
| FJ968794 | SUDV | 1976 |
| FJ217162 | TAFV | 1994 |
| AF522874 | RESTV | 1990 |
| FJ621583 | RESTV | 2008 |

Output

| | |
|---|---|
| 1(Gueckedou, Guinea)->18:0.0225 | 13->12:0.6012 |
| 2(Bundibugyo, Uganda)->14:0.6691 | 13->15:0.4609 |
| 3(Isiro, DRC)->14:0.0246 | 14->2(Bundibugyo, Uganda):0.6691 |
| 4(Yambuku, DRC)->17:0.0084 | 14->3(Isiro, DRC):0.0246 |
| 5(Mekouka, Gabon)->17:0.0108 | 14->15:0.2689 |
| 6(Luwero, Uganda)->11:0.0346 | 15->13:0.4609 |
| 7(Sudan)->11:0.0346 | 15->14:0.2689 |
| 8(Tai Forest, Ivory Coast)->16:0.3321 | 15->16:0.3089 |
| 9(Philippines)->12:0.2871 | 16->8(Tai Forest, Ivory Coast):0.3321 |
| 10(Philippines)->12:0.3462 | 16->15:0.3089 |
| 11->6(Luwero, Uganda):0.0346 | 16->18:0.4145 |
| 11->7(Sudan):0.0346 | 17->4(Yambuku, DRC):0.0084 |
| 11->13:1.3754 | 17->5(Mekouka, Gabon):0.0108 |
| 12->9(Philippines):0.2871 | 17->18:0.0106 |
| 12->10(Philippines):0.3462 | 18->1(Gueckedou, Guinea):0.0225 |
| 12->13:0.6012 | 18->16:0.4145 |
| 13->11:1.3754 | 18->17:0.0106 |

The neighbor-joining tree of different species of Ebola Viruses