

Phylogenetics: An (Arduous) *El Dorado* for Parallel and Distributed Computing

Diogo Telmo Neves

Department of Informatics
University of Minho
Braga, Portugal

December 9, 2014

- 1 Introduction
- 2 A Crash Course
- 3 DNA Taxa
- 4 Inference Methods

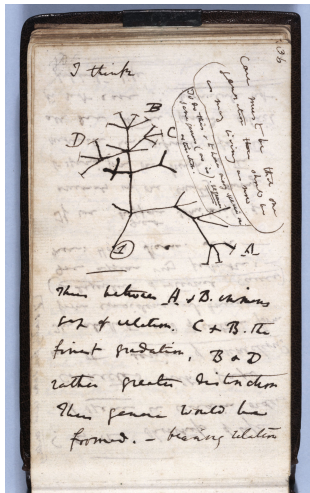
What is Phylogenetics?

- A branch of Biology
- **The study of evolutionary relationships among a set of taxa**
- Taxa does not necessarily means plants or animals
- Nowadays, the principles of phylogenetics are applied on a wide variety of domains

What is the Main Goal of a Phylogenetic Study?

- **To determine a phylogeny**
 - In most cases, a phylogeny is a hypothesis about the evolutionary history of a set of taxa
 - Usually, phylogenies are represented as unrooted trees

Some Examples



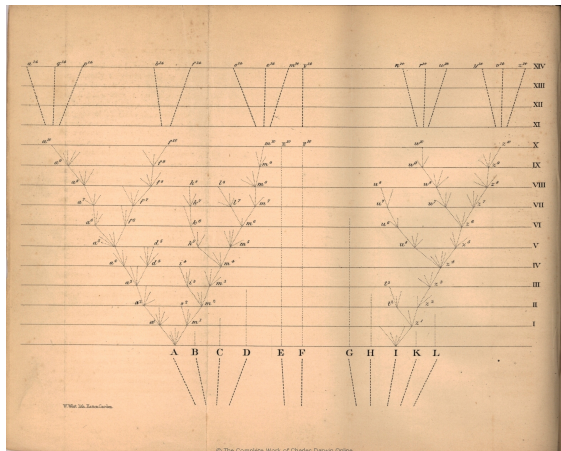
- A Darwin's Tree of Life sketch¹
- Page 36 of Darwin's "B" notebook, mid-July 1837

¹

Image was taken from

http://www.nybg.org/images/press.room/images/exhibition_images/darwins-garden_an_evolutionary_adventure/

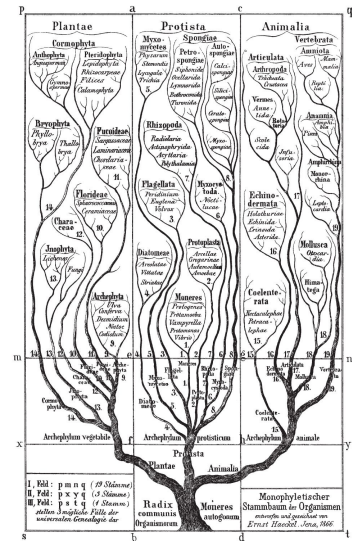
Some Examples



- The **only** image present in "On the Origin of Species" by Charles Darwin, 1859²

² Image was taken from http://darwin-online.org.uk/converted/published/1859-Origin_F373/

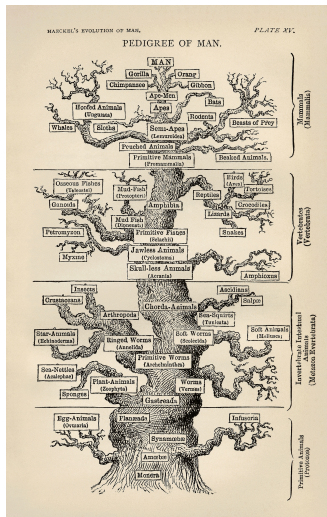
Some Examples



- Tree of Life constructed by Ernst Haeckel in 1866³
- This represents **one of the first attempts to draw an evolutionary tree that included all known life-forms**

³ Image was taken from http://www.evolution-textbook.org/content/free/figures/05_EVOW_Art/

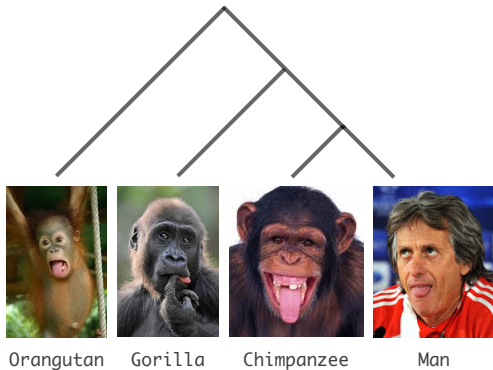
Some Examples



- In 1874 Ernst Haeckel told us that we do have a pedigree...⁴
- Woof, Woof!

⁴ Image was taken from <http://upload.wikimedia.org/wikipedia/commons/d/de/>

Some Examples



- Notice the contradiction that exists between this example and part of the previous one

Why is Phylogenetics Important?

- A phylogeny—the result of a phylogenetic analysis—allows us to get answers that, probably, we would not get with traditional approaches or methods
- Inferring phylogenies is crucial in many domains, such as:
 - in linguistics;
 - in forensics;
 - in cancer research and treatment;
 - in drugs research and design; and
 - so forth (the list is endless)

Exercise

Language	Word
Armenian	gatz
Basque	katu
Dutch	kat
English	cat
Estonian	kass
Finnish	kissa
Icelandic	kottur
Italian	gatto
Norwegian	katt
Polish	kot
Portuguese	gato
Russian	kot
Spanish	gato
Swedish	katt

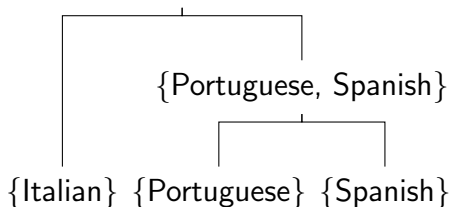
- Can you come up with a subtree using at least three elements from this table?
- Can you come up with another subtree using at least three other elements from this table?
- Can you come up with a tree using all the elements from this table?

Exercise

Language	Word
Armenian	gatz
Basque	katu
Dutch	kat
English	cat
Estonian	kass
Finnish	kissa
Icelandic	kottur
Italian	gatto
Norwegian	katt
Polish	kot
Portuguese	gato
Russian	kot
Spanish	gato
Swedish	katt

(Italian, (Portuguese, Spanish));

{Italian, Portuguese, Spanish}

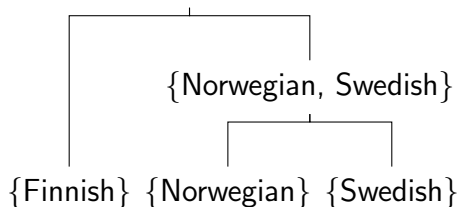


Exercise

Language	Word
Armenian	gatz
Basque	katu
Dutch	kat
English	cat
Estonian	kass
Finnish	kissa
Icelandic	kottur
Italian	gatto
Norwegian	katt
Polish	kot
Portuguese	gato
Russian	kot
Spanish	gato
Swedish	katt

(Finnish, (Norwegian, Swedish));

{Finnish, Norwegian, Swedish}



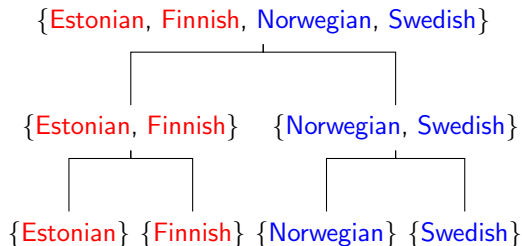
A detailed map of Europe and its surrounding regions, color-coded to distinguish between different countries and territories. The map includes the following labeled areas:

- Greenland Sea** (top left)
- Iceland** (top left, yellow)
- Norwegian Sea** (top center)
- Sweden** (top center, pink)
- Norway** (top center, pink)
- Finland** (top right, green)
- Estonia** (top right, blue)
- Latvia** (top right, pink)
- Lithuania** (top right, yellow)
- Russia** (top right, yellow)
- Belarus** (top right, blue)
- Ukraine** (top right, pink)
- Moldova** (top right, pink)
- Romania** (top right, blue)
- Bulgaria** (top right, green)
- Turkey** (top right, green)
- Cyprus** (top right, yellow)
- Greece** (top right, pink)
- Malta** (top right, yellow)
- North Sea** (center left)
- Denmark** (center left, yellow)
- Scotland** (center left, blue)
- Ireland** (center left, pink)
- Wales** (center left, blue)
- England** (center left, blue)
- Channel Islands** (center left, blue)
- Netherlands** (center left, pink)
- Belgium** (center left, pink)
- Germany** (center left, pink)
- Luxembourg** (center left, pink)
- France** (center left, green)
- Switzerland** (center left, green)
- Austria** (center left, green)
- Czech Republic** (center left, pink)
- Slovakia** (center left, pink)
- Hungary** (center left, yellow)
- Slovenia** (center left, green)
- Croatia** (center left, green)
- Bosnia and Herzegovina** (center left, green)
- Serbia** (center left, pink)
- Macedonia** (center left, pink)
- Albania** (center left, pink)
- Italy** (center left, yellow)
- Sicily** (center left, yellow)
- Sardinia** (center left, yellow)
- Corse** (center left, green)
- Spain** (center left, blue)
- Portugal** (center left, blue)
- Bay of Biscay** (center left)
- Balears** (center left, blue)
- Mediterranean Sea** (bottom center)
- Morocco** (bottom left)
- Algeria** (bottom left)
- Tunisia** (bottom left)
- Black Sea** (bottom right)

Exercise

Language	Word
Armenian	gatz
Basque	katu
Dutch	kat
English	cat
Estonian	kass
Finnish	kissa
Icelandic	kottur
Italian	gatto
Norwegian	katt
Polish	kot
Portuguese	gato
Russian	kot
Spanish	gato
Swedish	katt

((Estonian, Finnish), (Norwegian, Swedish));



Exercise

- **Disclaimer:** Analyzing one word says to little about the evolution and relationships of languages!

A Crash Course

Introduction

- As seen, phylogenetics can be used in several domains, not only in (Computational) Biology
- Hereafter, we will focus on using DNA (not RNA, nor proteins, nor any other type of data)
- DNA is a nucleic acid, which is made from monomers known as nucleotides
- Among other elements, a nucleotide has a nucleobase
- The nucleobases are:
 - **A**denine
 - **C**ytosine
 - **G**uanine
 - **T**hymine
- **A**denine and **G**uanine are Purines
- **C**ytosine and **T**hymine are Pyrimidines

A Crash Course

Introduction

Primate	(Excerpt of) DNA Sequence
Man	TGGTCCTGCTGTCCTCTCCTGGCGCCCTGGGCGCGAGCGGATGTCC
Neanderthal	TGGTCCTGCAGTCCTCTCCTGGCGCCCCGGGCGCGAGCGGTTGTCC
Chimpanzee	TGATCCTGCAGTCCTCTTCTGGCGCCCTGGGCGCGTGCGGTTGTCC
Gorilla	TGGACCTGCAGTCATCTTCTGCCC GCCCGAGCGCTTGCCGATGTCC
Orangutan	ACAACCTGCACTCCTATTCTGCCGAGCCGGGCGCGTGGCAAAGTCC

- How can we establish relationships between the given taxa?
 - We need a metric!
- Which is the complexity to compare the given taxa?
 - It will depend on the selected metric
 - It will depend on the (DNA) substitution model (we will discuss this later on)

A Crash Course

Hamming Distance: A Naïve Metric

- The Hamming distance⁵ between two fragments of DNA, with equal length and from different species, is given by the number of characters—sites—at which the corresponding nucleobases are different
- Examples:

Primate	(Excerpt of) DNA Sequence
Man	TGGTCCTGCTGTCCTCTCCTGGCGCCCTGGGCGCGAGCGGATGTCC
Neanderthal	TGGTCCTGCAGTCCTCTCCTGGCGCCCCGGGCGCGAGCGGTGTCC

Primate	(Excerpt of) DNA Sequence
Man	TGGTCCTGCTGTCCTCTCCTGGCGCCCTGGGCGCGAGCGGATGTCC
Chimpanzee	TGATCCTGCAGTCCTCTTCTGGCGCCCTGGGCGCGTGC GGTTGTCC

⁵ Richard W. Hamming. *Error-Detecting and Error-Correcting Codes*. The Bell System Technical Journal,

A Crash Course

From Hamming Distance to Pairwise Distance Matrix

- A distance matrix can be constructed from all pairwise sequences distances

	M	N	C	G	O
M	0	3	5	12	17
N	3	0	4	11	16
C	5	4	0	11	14
G	12	11	11	0	14
O	17	16	14	14	0

- A phylogeny can now be constructed by running an agglomerative (bottom-up) hierarchical clustering algorithm over the distance matrix
- Notice that the matrix is symmetric and the values of the main diagonal are zeros
 - So, one only needs less than half-matrix

A Crash Course

UPGMA Distance Method

- **U**nweighted **P**air-**G**roup **M**ethod with **A**rithmetic mean is a distance method that can be used to construct a phylogeny from a distance matrix
- UPGMA applies an agglomerative hierarchical clustering algorithm over the distance matrix
- UPGMA assumes that lineages have evolved at a constant rate (which is not true!)
- As a consequence of assuming that the evolution occurred at a constant rate, the leaves of a phylogenetic tree are equally distance from the root (again, this is not true!)

A Crash Course

UPGMA Distance Method

- Select the pair with lowest distance (in the case of a tie just select one of the pairs) and form a cluster

	M	N	C	G
N	3			
C	5	4		
G	12	11	11	
O	17	16	14	14

$$d(M, N) = d(N, M) = 3 \Rightarrow d(N, (N, M)) = d(M, (N, M)) = 3/2 = 1.5$$

- The matrix has to be recomputed (the new cluster is added, each species of the new cluster is removed, and distances to the new cluster have to be computed)

A Crash Course

UPGMA Distance Method

	M	N	C	G
N	3			
C	5	4		
G	12	11	11	
O	17	16	14	14

$$d(C, (M, N)) = (d(C, M) + d(C, N))/2$$

$$d(C, (M, N)) = (5 + 4)/2 = 4.5$$

$$d(G, (M, N)) = (d(G, M) + d(G, N))/2$$

$$d(G, (M, N)) = (12 + 11)/2 = 11.5$$

$$d(O, (M, N)) = (d(O, M) + d(O, N))/2$$

$$d(O, (M, N)) = (17 + 16)/2 = 16.5$$

	(M, N)	C	G
C	4.5		
G	11.5	11	
O	16.5	14	14

- The process is repeated until there is no more clusters to be computed

A Crash Course

UPGMA Distance Method

- And the matrices of the remaining iterations are:

	(M, N)	C	G
C	4.5		
G	11.5	11	
O	16.5	14	14

	(C, (M, N))	G
G	$(11.5 + 11) / 2 = 11.25$	
O	$(16.5 + 14) / 2 = 15.25$	14

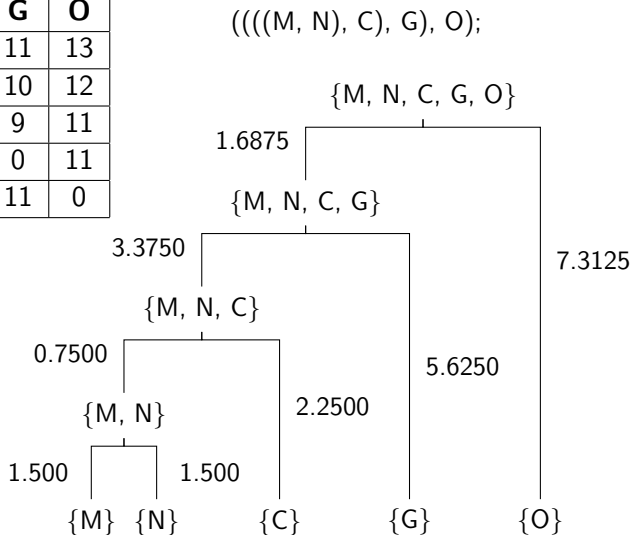
	(G, (C, (M, N)))
O	$(15.25 + 14) / 2 = 14.625$

A Crash Course

Exercise

	M	N	C	G	O
M	0	3	4	11	13
N	3	0	3	10	12
C	4	3	0	9	11
G	11	10	9	0	11
O	13	12	11	11	0

- What is “wrong” with this tree?



A Crash Course

Questions?

- When was the agglomerative clustering algorithm applied?
- Should DNA pairwise comparisons be done using the Hamming Distance metric?
 - If not, Is there better alternatives?
- Is there a better method than UPGMA?

Multiple Sequence Alignment (MSA)

Introduction

- Usually, DNA sequences are unaligned
- Therefore, before being used in an analysis the DNA sequences should be aligned first
- But... **why should one align DNA sequences?**
 - So far, we have seen one reason: to discover evolutionary relationships among a set of sequences
 - But it is also important to match structural regions as well functional regions (i.e., some DNA regions form a structure—for instance, a molecule—while other regions are responsible for performing specific functions)

Multiple Sequence Alignment (MSA)

Introduction

- Typically, one wants to align multiple sequences at the same time
- There are several tools to perform **M**ultiple **S**equence **A**lignment (MSA), such as:
 - MAFFT
(<http://mafft.cbrc.jp/alignment/server/index.html>)
 - **B**asic **L**ocal **A**lignment **S**earch **T**ool
(<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>)
 - and many others
(http://en.wikipedia.org/wiki/List_of_sequence_alignment_software)
- It is important to notice that the problem of finding the optimal MSA is NP-hard

Multiple Sequence Alignment (MSA)

Example

- We start by creating a FASTA⁶ file:

>Man

TGGTCCTGCTGTCCTCTCCTGGCGCCCTGGGCGCGAGCGGATGTCC

>Neanderthal

TGGTCCTGCAGTCCTCTCCTGGCGCCCCGGGCGCGAGCGGTTGTCC

>Chimpanzee

TGATCCTGCAGTCCTCTTCTGGCGCCCTGGGCGCGTGCGGTTGTCC

>Gorilla

TGGACCTGCAGTCATCTTCTGCCCCGCCGAGCGCTTGCCGATGTCC

>Orangutan

ACAACCTGCACTCCTATTCTGCCGAGCCGGGCGCGTGGCAAAGTCC

⁶For more details about the FASTA format see
http://en.wikipedia.org/wiki/FASTA_format

Multiple Sequence Alignment (MSA)

Example

- Then we align the sequences using a MSA tool
 - For instance, if we use MAFFT then we issue the following command:

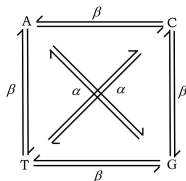
```
$mafft --auto primates.in > primates.out
```

where: “primates.in” is the FASTA file that was previously created (i.e., the DNA sequences to be aligned)
- The MSA will be written to “primates.out” file
- The MSA file can then be used as input to an inference tool (e.g., FastTree, or RAxML) in order to get a tree

Substitution Models

Introduction

- Within the evolution process it is usual to observe DNA mutations
- There are several types of mutations, for instance point mutations
 - A point mutation is due to a transition (i.e., $A \leftrightarrow G$, or $C \leftrightarrow T$) or a transversion (purine \leftrightarrow pyrimidine)
 - As a curiosity, some point mutations may cause cancer
- Point mutations can also be due to an insertion or a deletion
- The following figure represents all possible transitions and transversions



Substitution Models

Jukes and Cantor, 1969 (JC69)

- The JC69 (**J**ukes and **C**antor, 19**69**) makes **no** distinction between transitions and transversions, which means that the mutation rate— μ —is the only parameter of this model
 - The JC69 is, therefore, one of the simplest substitution models

Substitution Models

Jukes and Cantor, 1969 (JC69)

- As with other models, the JC69 model assumes that all nucleobases are equally frequent, which means that:

$$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$$

- And the rate matrix is given by:

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

- Under the JC69 model the evolutionary distance between two sequences is given by:

$$\hat{d} = -\frac{3}{4} \ln(1 - \frac{4}{3}p)$$

where: p is the proportion of sites that differ

Substitution Models

Generalised Time-Reversible (GTR)

- The **Generalised Time-Reversible** model (Tavaré 1986) is a neutral, independent, finite-sites, time-reversible model
- The frequency vector— $\Pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ —is pretty different from the counterpart of JC69, **the frequency of nucleobases is computed per each site**

Substitution Models

Generalised Time-Reversible (GTR)

- The GTR rate matrix is given by:

$$Q = \begin{pmatrix} -(x_1 + x_2 + x_3) & \frac{\pi_1 x_1}{\pi_2} & \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_1 x_3}{\pi_4} \\ x_1 & -\left(\frac{\pi_1 x_1}{\pi_2} + x_4 + x_5\right) & \frac{\pi_2 x_4}{\pi_3} & \frac{\pi_2 x_5}{\pi_4} \\ x_2 & x_4 & -\left(\frac{\pi_1 x_2}{\pi_3} + \frac{\pi_2 x_4}{\pi_3} + x_6\right) & \frac{\pi_3 x_6}{\pi_4} \\ x_3 & x_5 & x_6 & -\left(\frac{\pi_1 x_3}{\pi_4} + \frac{\pi_2 x_5}{\pi_4} + \frac{\pi_3 x_6}{\pi_4}\right) \end{pmatrix}$$

- So, when using four sites—nucleobases: A, C, G, and T—the GTR model requires 6 substitution rate parameters— $(x_1, x_2, x_3, x_4, x_5, x_6)$ —and 4 equilibrium base frequency parameters— $(\pi_1, \pi_2, \pi_3, \pi_4)$ —(see previous slide)

Inference Methods

Introduction

- Enumerating trees:

#Taxa n	#Trees $(2n - 5)!!$
2	1
3	1
4	3
5	15
10	2027025
20	221643095476700000000

- Tree searching, just like MSA searching, is a NP-hard problem
- Moreover, the tree search space has a factorial growth

Inference Methods

Introduction

- Therefore, programs that search for an optimal tree use heuristics
- Those heuristics try to solve a NP-hard problem by performing local searches
- The main goal is to find a tree that better explains the taxa under some criterion, such as:
 - Minimum tree distance;
 - Minimum number of evolutionary events; or
 - Maximum likelihood of a tree

Inference Methods

Distance

- UPGMA
- Neighbor joining
 - The most widely used distance method
 - Starts by creating a star tree: all leaves are (directly) connected to the same node
 - Then, the tree is iteratively expanded by pairing “neighboring” taxa, based on a distance matrix
 - It is possible to exploit parallelism during the construction of the tree
 - There are several implementations available
- Minimum evolution
 - For instance, the Jukes-Cantor model of evolution
- ... and so forth

Inference Methods

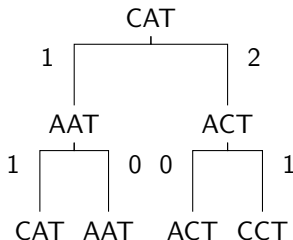
Maximum Parsimony (MP)

- The goal of **Maximum Parsimony (MP)** is to find a tree that better explains the given taxa using the minimum number of evolutionary events
 - This is known as the Large Parsimony problem
 - MP, like MSA, is a NP-hard problem
 - Exploiting parallelism within this problem may yield significant speedups
- MP relies on:
 - Changes at different positions of the sequences are independent; and
 - Changes in different parts of the tree are independent

Inference Methods

Maximum Parsimony (MP)

- The parsimony score of a tree is useful to find which tree better explains the given taxa
 - This is known as the Small Parsimony problem
 - Depending on the size of the tree, parallelism exploitation may contribute to improve performance
- Example using the aforementioned Hamming Distance:



The parsimony score is 5

Inference Methods

Maximum Parsimony (MP)

- There are better algorithms to compute the parsimony score of a tree than using the Hamming Distance, for instance:
 - The Fitch's algorithm; and
 - The Sankoff's algorithm

Inference Methods

Maximum Likelihood (ML)

- The basic idea of the **Maximum Likelihood (ML)** estimation method is to compute a phylogenetic tree T , together with edge lengths ω , that maximizes the likelihood, as follows:

$$L(T) = P(A \mid T, \omega, M)$$

where:

- A is a MSA; and
 - M is a (DNA) substitution model (see DNA Section)
- ML, like MSA and MP, is also a NP-hard problem
 - Again, parallelism exploitation may contribute to improve performance

Inference Methods

Maximum Likelihood (ML)

- ML Programs with source code freely available:
 - FastTree
(<http://www.microbesonline.org/fasttree/>)
 - RAxML
(<https://github.com/stamatak/standard-RAxML>)
 - PhyML
(<https://github.com/stephaneguindon/phyml>)
 - ... and many others

?