Models of Sequence Evolution With Selection

Eva Stukenbrock & Julien Dutheil

eva.stukenbrock@mpi-marburg.mpg.de

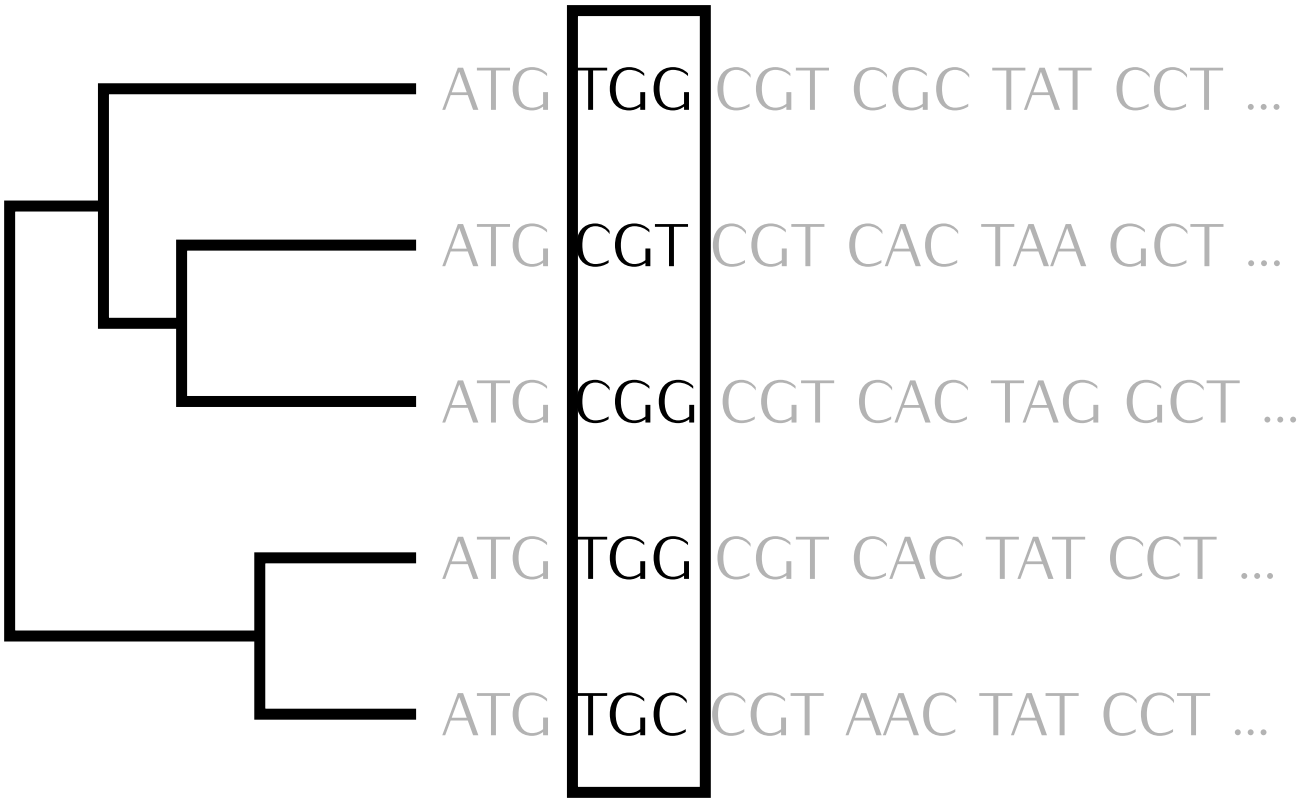
julien.dutheil@mpi-marburg.mpg.de

Max Planck Institute for Terrestrial Microbiology – Marburg

20 February 2013

Modeling the evolution of a sequence alignment

Modeling the evolution of a sequence alignment



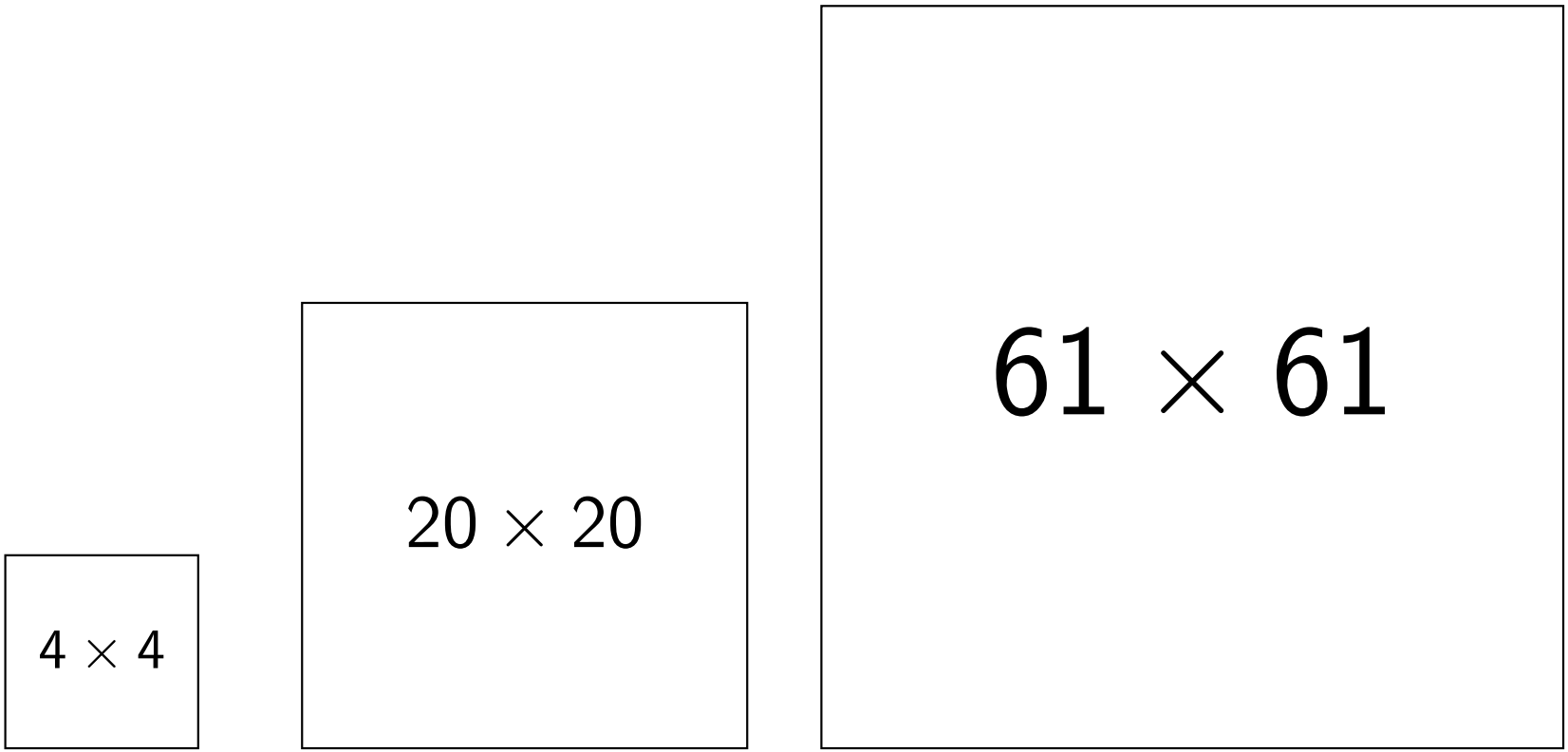
- We assume that all columns (sites) are independent

Modeling the evolution on a branch

- Mutations can occur at any time, with a given rate

Modeling the evolution on a branch

- Mutations can occur at any time, with a given rate
- The probability of each type of mutation is given by a matrix:



Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications...
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications...
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters
  - We consider only 2 types of mutations: synonymous and non-synonymous. The ratio of the two is noted *omega*

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications...
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters
  - We consider only 2 types of mutations: synonymous and non-synonymous. The ratio of the two is noted *omega*
  - We allow nucleotide transitions and transversions to occur at a distinct rate. The ratio of the two is noted *kappa*

Models for codon evolution

- A codon model would have $61 \times 61 = 3721$ parameters (this is typically more than the actual data!)
- We need simplifications...
  - We consider that two mutations cannot occur at the same time, so codon mutation involving more than one change are discarded, this leaves 526 parameters
  - We consider only 2 types of mutations: synonymous and non-synonymous. The ratio of the two is noted *omega*
  - We allow nucleotide transitions and transversions to occur at a distinct rate. The ratio of the two is noted *kappa*
- We can therefore express all mutation probabilities with only two parameters

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)
2. Multiplying all probabilities for all branches for one site

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)
2. Multiplying all probabilities for all branches for one site
3. Summing over all possible ancestral states

Likelihood of an alignment

With this framework we can compute the probability of a data set given a mutation model by:

1. Computing the probability for one branch for one site (requires the exponential of the mutation matrix)
2. Multiplying all probabilities for all branches for one site
3. Summing over all possible ancestral states
4. Multiplying for all sites in the alignment

Estimating parameters

- The probability of the data given a set of parameters is called likelihood

Estimating parameters

- The probability of the data given a set of parameters is called likelihood
- Estimating parameters consists in finding their value which maximize the likelihood (Maximum likelihood estimates, mle)

Estimating parameters

- The probability of the data given a set of parameters is called likelihood
- Estimating parameters consists in finding their value which maximize the likelihood (Maximum likelihood estimates, mle)

Note

Even though this is not a requirement, in practice, codon models are not used for phylogenetic inference. The phylogenetic tree / sequence genealogy is assumed to be known.