

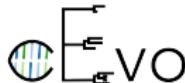
Computational Biology

Lecturers:
Tanja Stadler, Carsten Magnus & Tim Vaughan

Teaching Assistants:
Jūlija Pečerska, Jérémie Sciré,
Sarah Nadeau & Marc Manceau

Computational Evolution
Department of Biosystems Science and Engineering

HS 2019



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Alignment, BLAST Questions

- ② What are the weaknesses and strengths of the different alignment methods (dot-matrix method, Smith-Waterman, Needleman-Wunsch, BLAST)?
- ② With which alignment methods do you get an optimal alignment?
- ② Do you obtain the same alignments when using different scoring schemes in the Smith-Waterman and Needleman-Wunsch algorithms?

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

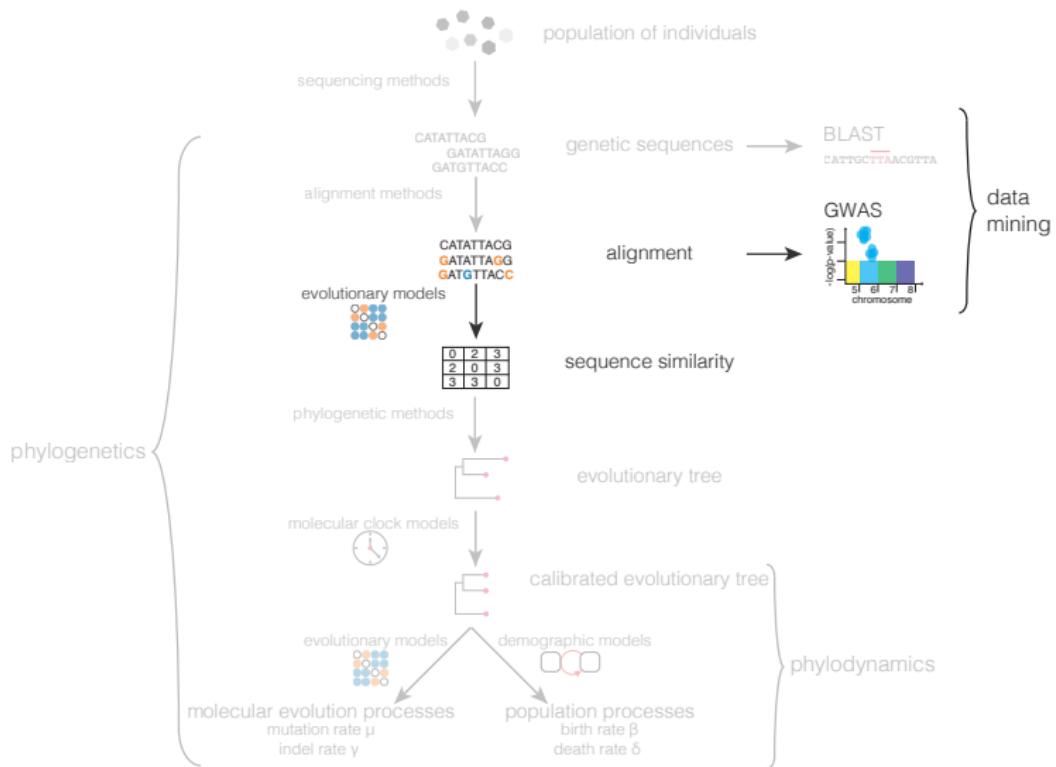
Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Overview and outline of lecture 03



03: Nucleotide substitution models

- GWAS
- Measures of variability
- Sequence distance
- Definition of distance
- Nucleotide substitution as a Markov chain
- Substitution rate matrices
- Markov models of nucleotide substitution: JC69

References

03: Nucleotide substitution models

GWAS
Measures of variability
Sequence distance
Definition of distance
Nucleotide substitution as a Markov chain
Substitution rate matrices
Markov models of nucleotide substitution: JC69

References

Genome Wide Association Studies.

Genome wide association studies

Research question: Can we identify genetic risk factors for common diseases?

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

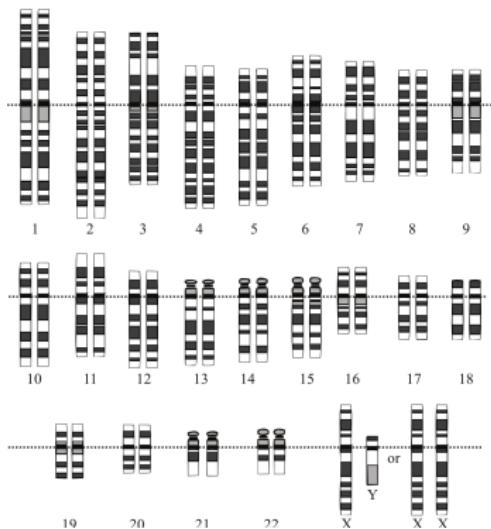
Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Genome wide association studies

Research question: Can we identify genetic risk factors for common diseases?



The Human Genome

en.wikipedia.org/wiki/Human_genome

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

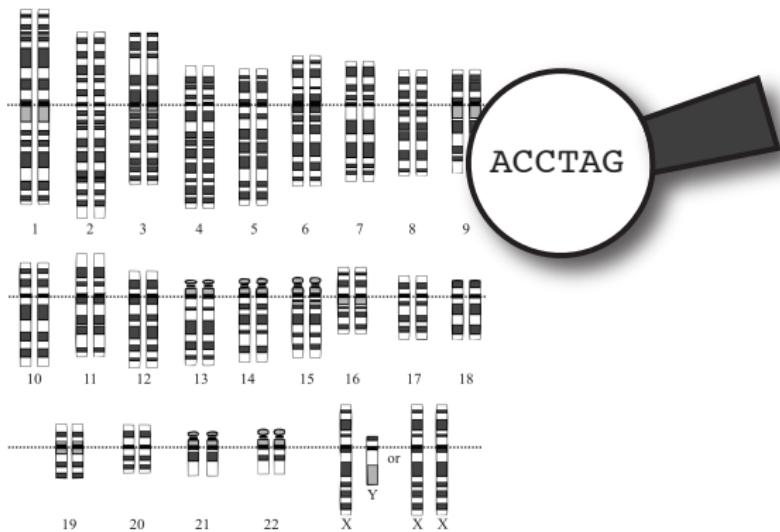
Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Genome wide association studies

Research question: Can we identify genetic risk factors for common diseases?



The Human Genome

en.wikipedia.org/wiki/Human_genome

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

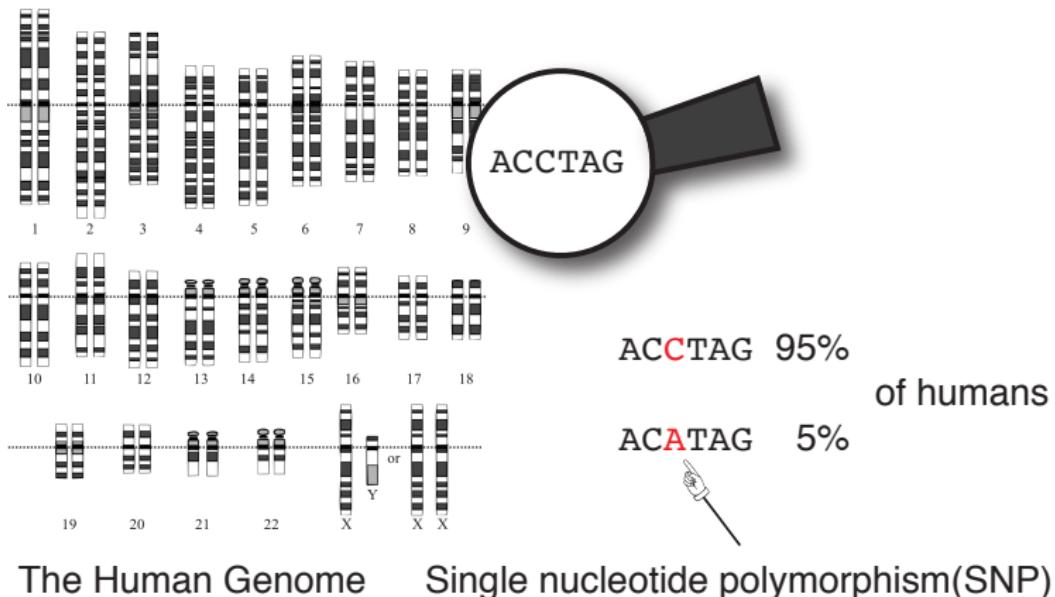
Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Genome wide association studies

Research question: Can we identify genetic risk factors for common diseases?

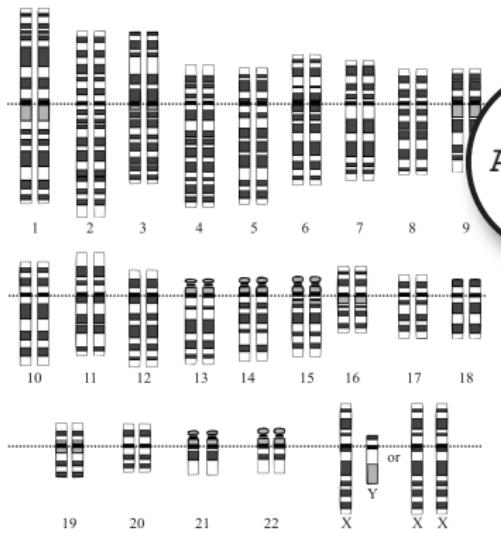


en.wikipedia.org/wiki/Human_genome

- 03: Nucleotide substitution models
- GWAS
- Measures of variability
- Sequence distance
- Definition of distance
- Nucleotide substitution as a Markov chain
- Substitution rate matrices
- Markov models of nucleotide substitution: JC69
- References

Genome wide association studies

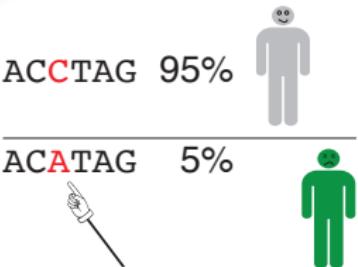
Research question: Can we identify genetic risk factors for common diseases?



The Human Genome

Single nucleotide polymorphism(SNP)

en.wikipedia.org/wiki/Human_genome



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

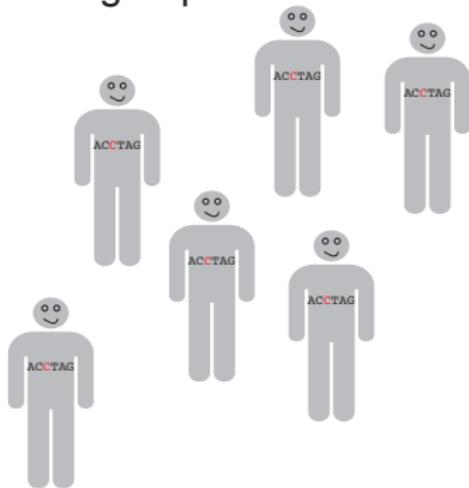
References

Genome wide association studies

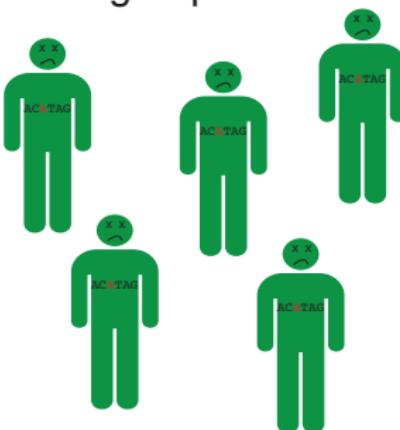
Case-control setup – The most commonly used strategy:

- ▶ two large groups of individuals: one healthy control group, and one group with a certain disease (e.g. diabetes)

Control group



Case group



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

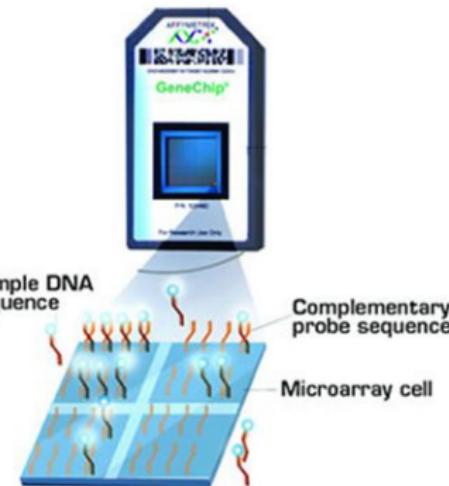
Markov models of nucleotide substitution: JC69

References

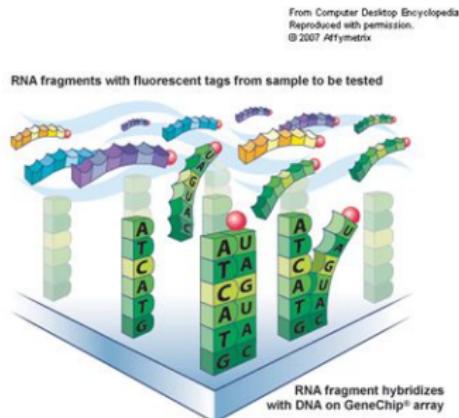
Genome wide association studies

Case-control setup – The most commonly used strategy:

- ▶ two large groups of individuals: one healthy control group, and one group with a certain disease (e.g. diabetes)
- ▶ all individuals are genotyped for the majority of known SNP locations (Illumina, Affymetrix)



• (Liu 2007)



• (Affymetrix)

adapted from <http://slideplayer.com/slide/8465760/>

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Genome wide association studies

Statistical analysis

- ▶ case control setup: two groups (i) patients with disease, (ii) healthy patients
- ▶ for each SNP, count the number of healthy individuals without this specific mutation, H_N , and with the mutation, H_S , as well as the number of diseased individuals without the mutation, D_N , and with the mutation, D_S

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Genome wide association studies

Statistical analysis

- ▶ case control setup: two groups (i) patients with disease, (ii) healthy patients
- ▶ for each SNP, count the number of healthy individuals without this specific mutation, H_N , and with the mutation, H_S , as well as the number of diseased individuals without the mutation, D_N , and with the mutation, D_S
- ▶ calculate the odds ratio (OR) of diseased and healthy people with respect to the abundance of each SNP

$$\text{OR} = \frac{\text{odds of having the disease amongst individuals with minor variant on SNP position}}{\text{odds of having the disease amongst individuals with major variant on SNP position}}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Genome wide association studies

Statistical analysis

- ▶ case control setup: two groups (i) patients with disease, (ii) healthy patients
- ▶ for each SNP, count the number of healthy individuals without this specific mutation, H_N , and with the mutation, H_S , as well as the number of diseased individuals without the mutation, D_N , and with the mutation, D_S
- ▶ calculate the odds ratio (OR) of diseased and healthy people with respect to the abundance of each SNP

$OR = \frac{\text{odds of having the disease amongst individuals with minor variant on SNP position}}{\text{odds of having the disease amongst individuals with major variant on SNP position}}$

$$= \frac{D_S/H_S}{D_N/H_N}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Genome wide association studies: Example

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Observed

	case	control	total
minor	2104	2676	4780
major	1896	3324	5220
total	4000	6000	10 000

Toolbox: Statistical significance

- The odds ratio can only inform us about a potential association between a SNP and a genetic disease.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Toolbox: Statistical significance

- ▶ The odds ratio can only inform us about a potential association between a SNP and a genetic disease.
- ▶ How can we test for statistical significance?
 - ▶ Contingency table tests (Fisher's exact test, Pearson's χ^2 test) to calculate the p-value

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

Toolbox: Statistical significance

- ▶ The odds ratio can only inform us about a potential association between a SNP and a genetic disease.
- ▶ How can we test for statistical significance?
 - ▶ Contingency table tests (Fisher's exact test, Pearson's χ^2 test) to calculate the p-value

Definition: p-value

Given a random variable X and a realisation (observation) x . Let us assume a null hypothesis \mathcal{H}_0 , which is a statement on the distribution of X . The p-value is the probability of observing x or a more extreme realisation under the assumption the null hypothesis was true.

$$\text{p-value} := P(X = x \text{ or more extreme} | \mathcal{H}_0)$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Statistical significance

Definition: p-value

Given a random variable X and a realisation (observation) x . Let us assume a null hypothesis \mathcal{H}_0 , which is a statement on the distribution of X . The p-value is the probability of observing x or a more extreme realisation under the assumption the null hypothesis was true.

$$\text{p-value} := P(X = x \text{ or more extreme} | \mathcal{H}_0)$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Statistical significance

- ▶ The odds ratio can only inform us about a potential association between a SNP and a genetic disease.
- ▶ How can we test for statistical significance?
 - ▶ Contingency table tests (Fisher's exact test, Pearson's χ^2 test) to calculate the p-value

Definition: p-value

Given a random variable X and a realisation (observation) x . Let us assume a null hypothesis \mathcal{H}_0 , which is a statement on the distribution of X . The p-value is the probability of observing x or a more extreme realisation under the assumption the null hypothesis was true.

$$\text{p-value} := P(X = x \text{ or more extreme} | \mathcal{H}_0)$$

Definition: Significance level

The significance level, α , is a value that one defines before performing a statistical test. If the p-value lies below this significance level, the observed result of the random experiment cannot support the null hypothesis which leads to rejection of the null hypothesis.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

Toolbox: Fisher's exact test

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Fisher's exact test: Statistical test to examine the significance of the association between two kinds of classifications (here: being diseased, having minor variant).

Contingency table:

		B	B ₁	B ₂	
		A	a	b	a+b
		A ₁			
		A ₂	c	d	c+d
		a+c	b+d	$\frac{n}{a+b+c+d}$	

Toolbox: Fisher's exact test

Fisher's exact test: Statistical test to examine the significance of the association between two kinds of classifications (here: being diseased, having minor variant).

Contingency table:

		B ₁	B ₂	
		a	b	a+b
		c	d	c+d
A ₁		a	b	a+b
A ₂		c	d	c+d
		a+c	b+d	$\frac{n}{a+b+c+d}$

Hypothesis:

We want to test whether class A is linked to class B.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Fisher's exact test

Fisher's exact test: Statistical test to examine the significance of the association between two kinds of classifications (here: being diseased, having minor variant).

Contingency table:

		B		
		B ₁	B ₂	
A	A ₁	a	b	a+b
	A ₂	c	d	c+d
		a+c	b+d	$\frac{n}{a+b+c+d}$

Hypothesis:

We want to test whether class A is linked to class B.

Thus our null hypothesis is:

\mathcal{H}_0 : The number of individuals expressing both A₁ and B₁ is based on chance.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Hypergeometric distribution

To be able to phrase the null hypothesis in terms of a probability distribution, we need to consider the hypergeometric distribution, which is best explained in terms of an urn experiment. Assume an urn with r red and s black balls.

k balls are drawn without replacement. How is the number of red balls among the k drawn balls, R_k , distributed?



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Hypergeometric distribution

To be able to phrase the null hypothesis in terms of a probability distribution, we need to consider the hypergeometric distribution, which is best explained in terms of an urn experiment. Assume an urn with r red and s black balls.

k balls are drawn without replacement. How is the number of red balls among the k drawn balls, R_k , distributed?



We need to calculate this probability for $i \in \{0, 1, \dots, k\}$:

$$P(R_k = i) =$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Hypergeometric distribution

To be able to phrase the null hypothesis in terms of a probability distribution, we need to consider the hypergeometric distribution, which is best explained in terms of an urn experiment. Assume an urn with r red and s black balls.

k balls are drawn without replacement. How is the number of red balls among the k drawn balls, R_k , distributed?



We need to calculate this probability for $i \in \{0, 1, \dots, k\}$:

$$P(R_k = i) = \frac{\binom{r}{i} \binom{s}{k-i}}{\binom{r+s}{k}}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Fisher's exact test II

Contingency table:

		B	B ₁	B ₂	
		A			
A ₁	a	b	a+b		
	c	d	c+d		
		a+c	b+d	$n =$ $a+b+c+d$	

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Fisher's exact test II

Contingency table:

		B ₁	B ₂	
		a	b	a+b
		c	d	c+d
A ₁		a	b	a+b
A ₂		c	d	c+d
	a+c	b+d	$n =$ $a+b+c+d$	

Hypothesis:

We want to test whether class A is linked to class B

Thus our null hypothesis is:

\mathcal{H}_0 : The number of individuals expressing both A_1 and B_1 is based on chance.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Fisher's exact test II

Contingency table:

A \ B	B ₁	B ₂	
A ₁			a+b
A ₂			c+d
	a+c	b+d	$n = a+b+c+d$

Hypothesis:

We want to test whether class A is linked to class B

Thus our null hypothesis is:

\mathcal{H}_0 : The number of individuals expressing both A_1 and B_1 is based on chance.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Fisher's exact test II

Contingency table:

		B ₁	B ₂	
				a+b
A ₁				
A ₂				c+d
		a+c	b+d	$n =$ $a+b+c+d$

Hypothesis:

We want to test whether class A is linked to class B

Thus our null hypothesis is:

\mathcal{H}_0 : The number of individuals expressing both A_1 and B_1 is based on chance.

☞ The correct formulation of the null hypothesis is:

\mathcal{H}_0 : The random variable of the number of individuals expressing both A_1 and B_1 is distributed according to a hypergeometric distribution.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Fisher's exact test II

Contingency table:

		B	B ₁	B ₂	
		A			
A ₁					a+b
					c+d
	a+c	b+d		n=	a+b+c+d

Hypothesis:

We want to test whether class A is linked to class B

Thus our null hypothesis is:

\mathcal{H}_0 : The number of individuals expressing both A_1 and B_1 is based on chance.

☞ The correct formulation of the null hypothesis is:

\mathcal{H}_0 : The random variable of the number of individuals expressing both A_1 and B_1 is distributed according to a hypergeometric distribution.

$$\text{p-value} = \sum_{i=a}^{a+b} \frac{\binom{a+b}{i} \binom{c+d}{a+c-i}}{\binom{n}{a+c}}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Toolbox: Pearson's χ^2 -test

- ▶ Fisher's exact test only works for small numbers. Problem: $\binom{n}{k}$ cannot be calculated correctly for bigger numbers.
- ▶ Pearson invented the χ^2 -test to account for bigger numbers.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Pearson's χ^2 -test

- ▶ Fisher's exact test only works for small numbers. Problem: $\binom{n}{k}$ cannot be calculated correctly for bigger numbers.
- ▶ Pearson invented the χ^2 -test to account for bigger numbers.

One calculates the deviance between observed and expected numbers based on a hypergeometric distribution:

e.g. # expected cases with minor variant = $10\ 000 \frac{4000}{10\ 000} \frac{4780}{10\ 000}$

Observed = $(O_{ij})_{i,j \in \{1,2\}}$

	case	control	total
minor	2104	2676	4780
major	1896	3324	5220
total	4000	6000	10 000

Expected = $(E_{ij})_{i,j \in \{1,2\}}$

	case	control	total
			4780
			5220
	4000	6000	10 000

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Toolbox: Pearson's χ^2 -test

Observed = $(O_{ij})_{i,j \in \{1,2\}}$

	case	control	total
minor	2104	2676	4780
major	1896	3324	5220
total	4000	6000	10 000

Expected = $(E_{ij})_{i,j \in \{1,2\}}$

	case	control	total
minor	1912	2868	4780
major	2088	3132	5220
total	4000	6000	10 000

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

To test whether this result is significant, we calculate the following test statistic:

$$\chi^2 = \sum_{i,j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Pearson proved that this statistic is χ^2 -distributed.

Toolbox: Pearson's χ^2 -test

Observed = $(O_{ij})_{i,j \in \{1,2\}}$

	case	control	total
minor	2104	2676	4780
major	1896	3324	5220
total	4000	6000	10 000

Expected = $(E_{ij})_{i,j \in \{1,2\}}$

	case	control	total
minor	1912	2868	4780
major	2088	3132	5220
total	4000	6000	10 000

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

To test whether this result is significant, we calculate the following test statistic:

$$x = \sum_{i,j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Pearson proved that this statistic is χ^2 -distributed.

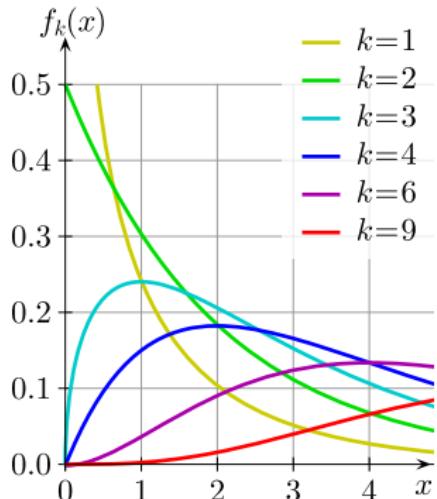
The p-value is then the probability of a χ^2 -distributed random variable being $\geq x$:

In our case $x = 61$ and thus p-value = 5.055×10^{-15}

Toolbox: The χ^2 distribution

Assume n i.i.d random variables $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. The quantities

$X^2 := \sum_{i=1}^n \frac{(Y_i - \mu)^2}{\sigma^2}$ are then χ^2 - distributed with n degrees of freedom.



- ▶ It can be proven that many statistics are approximately χ_n^2 -distributed.
- ▶ This simplifies hypothesis testing or estimation of confidence intervals.
- ▶ $P(\chi_n^2 = x)$ is a short (and sloppy) notation which means: the probability that a test statistic S which is χ_n^2 -distributed, is $\geq x$, i.e. $P(\chi_n^2 = x) = P(S \geq x)$

further reading: [Sokal and Rohlf, 2012]

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

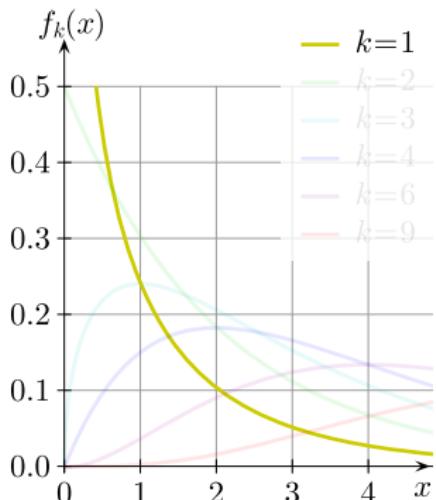
Markov models of nucleotide substitution: JC69

References

Toolbox: The χ^2 distribution

Assume n i.i.d random variables $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. The quantities

$X^2 := \sum_{i=1}^n \frac{(Y_i - \mu)^2}{\sigma^2}$ are then χ^2 - distributed with n degrees of freedom.



- ▶ It can be proven that many statistics are approximately χ_n^2 -distributed.
- ▶ This simplifies hypothesis testing or estimation of confidence intervals.
- ▶ $P(\chi_n^2 = x)$ is a short (and sloppy) notation which means: the probability that a test statistic S which is χ_n^2 -distributed, is $\geq x$, i.e. $P(\chi_n^2 = x) = P(S \geq x)$

further reading: [Sokal and Rohlf, 2012]

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

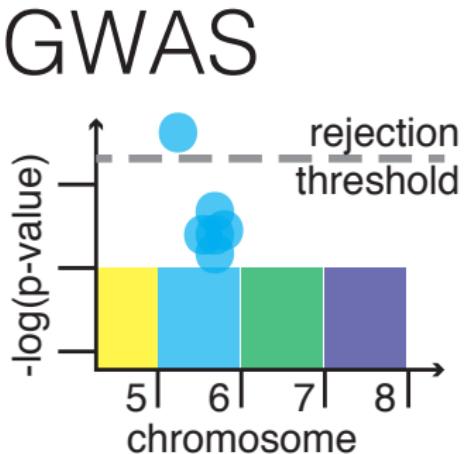
Markov models of nucleotide substitution: JC69

References

Genome wide association studies

Typical results

- ▶ graph with the SNP positions (sorted according to chromosome) on the x-axis and $-\log(p\text{-value})$ on the y-axis
- ▶ SNPs (as well as chromosomes) with extremely low p-values could be associated with the specific disease



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

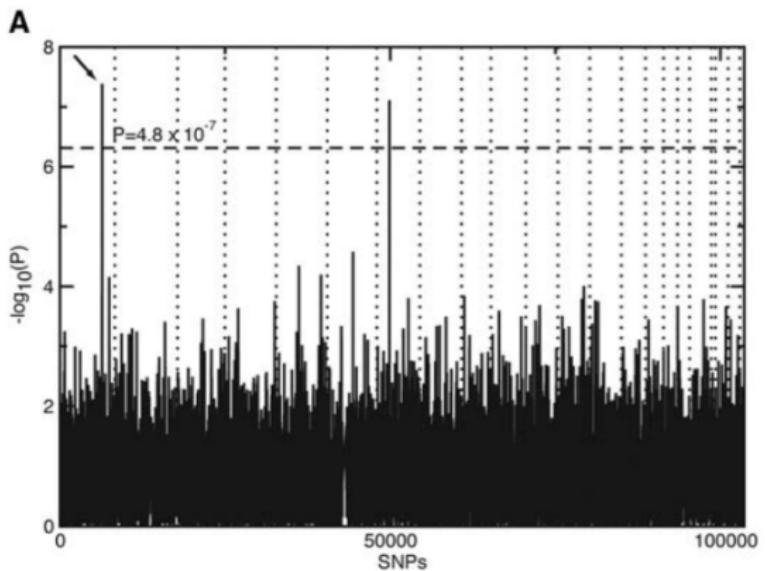
Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

The first GWAS

- ▶ first GWAS used to identify SNPs in complement factor H gene associated with age related macular degeneration
- ▶ 96 affected individuals and 50 healthy controls [Klein et al., 2005]



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GWAS today

- ▶ GWAS approaches received some criticism due to

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GWAS today

- ▶ GWAS approaches received some criticism due to
 - ▶ missing quality control steps, i.e. tests for possible biases for tested group structure needed

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GWAS today

- ▶ GWAS approaches received some criticism due to
 - ▶ missing quality control steps, i.e. tests for possible biases for tested group structure needed
 - ▶ multiple testing, i.e. correction for multiple testing needed (Bonferroni-correction)

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GWAS today

- ▶ GWAS approaches received some criticism due to
 - ▶ missing quality control steps, i.e. tests for possible biases for tested group structure needed
 - ▶ multiple testing, i.e. correction for multiple testing needed (Bonferroni-correction)
 - ▶ correlations only between single SNPs but not between genes tested

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GWAS today

- ▶ GWAS approaches received some criticism due to
 - ▶ missing quality control steps, i.e. tests for possible biases for tested group structure needed
 - ▶ multiple testing, i.e. correction for multiple testing needed (Bonferroni-correction)
 - ▶ correlations only between single SNPs but not between genes tested
- ▶ However, as of 2011, 1 200 GWAS have been conducted; over 200 diseases and traits, and almost 4,000 SNP associations have been found. GWAS approaches have been extended to address the above listed criticism.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

GWAS today

- ▶ GWAS approaches received some criticism due to
 - ▶ missing quality control steps, i.e. tests for possible biases for tested group structure needed
 - ▶ multiple testing, i.e. correction for multiple testing needed (Bonferroni-correction)
 - ▶ correlations only between single SNPs but not between genes tested
- ▶ However, as of 2011, 1 200 GWAS have been conducted; over 200 diseases and traits, and almost 4,000 SNP associations have been found. GWAS approaches have been extended to address the above listed criticism.

Further reading:

- ▶ [Pearson and Manolio, 2008]
- ▶ [Bush and Moore, 2012]
- ▶ <https://www.ebi.ac.uk/gwas/>
- ▶ <https://easygwas.ethz.ch/>

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

03: Nucleotide substitution models

- GWAS
- Measures of variability
- Sequence distance
- Definition of distance
- Nucleotide substitution as a Markov chain
- Substitution rate matrices
- Markov models of nucleotide substitution: JC69

References

Quantifying variation between sequences.

Motivation

We want to infer phylogenetic trees. Phylogenies depict the genetic relatedness of sequences:

A C T A G C T G human

A G T T G C T G chimpanzee

A C T T G A T G gorilla

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

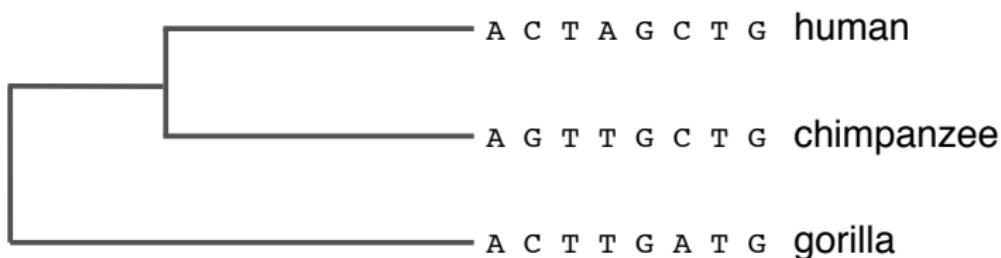
Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Motivation

We want to infer phylogenetic trees. Phylogenies depict the genetic relatedness of sequences:



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

How to measure the sequences differences

Example:

ATTACGAC

TCTACGAC

- ▶ Hamming distance:
 - ▶ number of segregating sites
 - ▶ count the sites that vary
 - ▶ in our example 2
- ▶ p-distance:
 - ▶ $\frac{\text{number of segregating sites}}{\text{sequence length}}$
 - ▶ in our example $2/8=0.25$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

How to measure the sequences differences

Example:

ATTACGAC

TCTACGAC

- ▶ Hamming distance:
 - ▶ number of segregating sites
 - ▶ count the sites that vary
 - ▶ in our example 2
- ▶ p-distance:
 - ▶ $\frac{\text{number of segregating sites}}{\text{sequence length}}$
 - ▶ in our example $2/8=0.25$

Example from last lecture: *triosephosphate isomerase*:



NGTTDQVDKIVKILNEGQIASTDVVEVVVSPPYVFLPVVKSQLRPEIQVAAQNCW

||.....!..!..|!|..|..!... .|||||. | ..!|..!|||...! |||||!



NGDKASIADLCKVLTTGPLNAD _TEVVVGCAPYTLARSQLPDSVCVAQNCY

Hamming-distance 35

$p = 0.636$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

03: Nucleotide substitution models

- GWAS
- Measures of variability
- Sequence distance
 - Definition of distance
 - Nucleotide substitution as a Markov chain
 - Substitution rate matrices
 - Markov models of nucleotide substitution: JC69

References

Molecular evolution models.

The fundamental problem

A C T T G A T G



A C T A G C T G

taxon 1

A G T T G C T G

taxon 2

A C T T G A T G

taxon 3

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

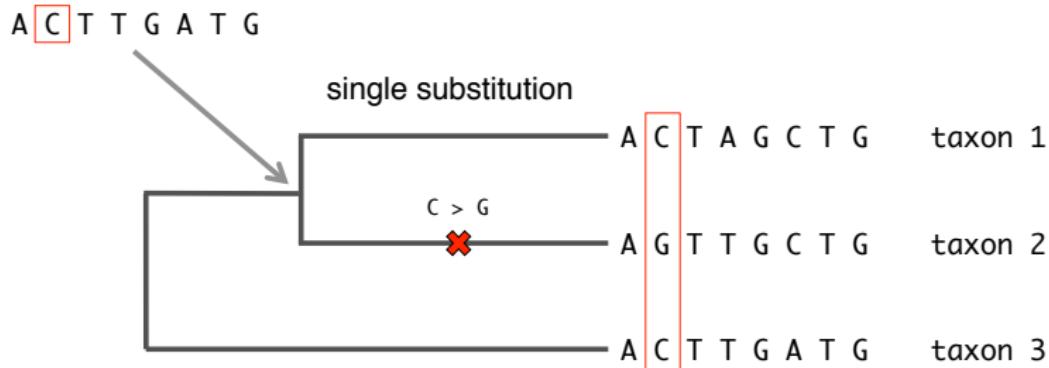
Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

The fundamental problem



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

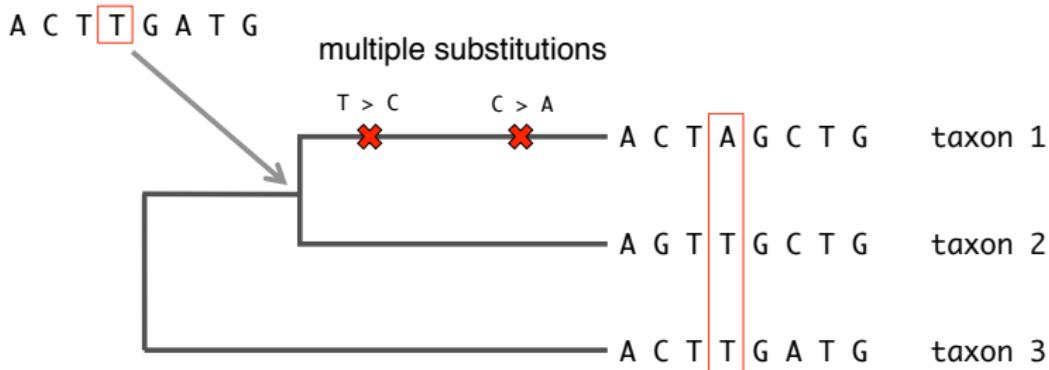
Substitution rate matrices

Markov models of

nucleotide substitution:
JC69

References

The fundamental problem



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

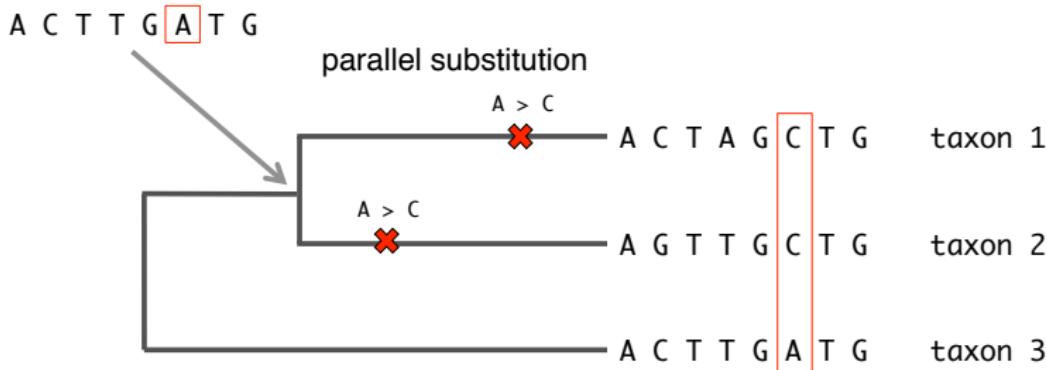
Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

The fundamental problem



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

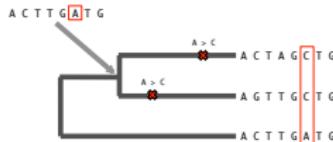
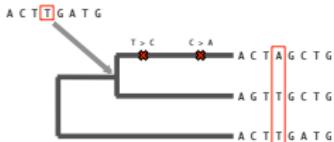
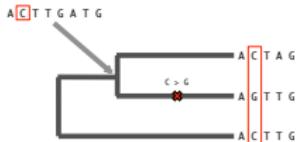
Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

The fundamental problem



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Problem of phylogenetics:

We observe sequences but not their evolutionary history. Thus we have to take **all** possible evolutionary trajectories into account.

How can we account for the evolutionary steps in between?

Sequence distance

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Definition: Distance between two sequences

The distance between two sequences is the expected number of nucleotide substitutions per site

[Yang, 2014].

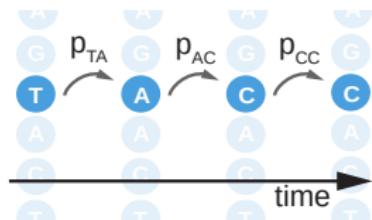
- ▶ This definition includes all (non-visible) evolutionary steps in between two sequences.
- ▶ We need to employ a mathematical model for estimating the distance, as we do not know the evolutionary steps. This model must include
 - ▶ (stochastic) process modelling the substitution through time
 - ▶ substitution rates

Nucleotide substitutions as a Markov chain

Definition of a Markov chain (see also [Ross, 1996])

stochastic process, i.e. a series of random experiments through time

Nucleotide substitutions as MC



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

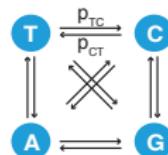
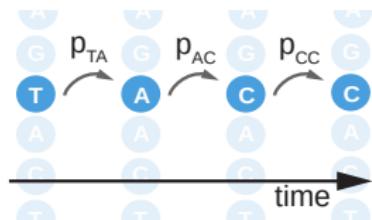
Nucleotide substitutions as a Markov chain

Definition of a Markov chain (see also [Ross, 1996])

stochastic process, i.e. a series of random experiments through time

lives on a **state space** and jumps to the different states

Nucleotide substitutions as MC



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

Nucleotide substitutions as a Markov chain

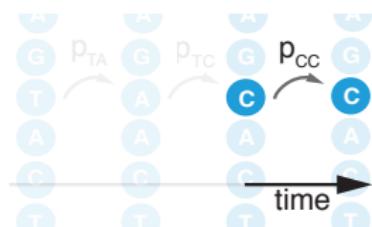
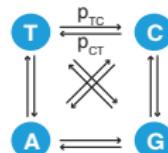
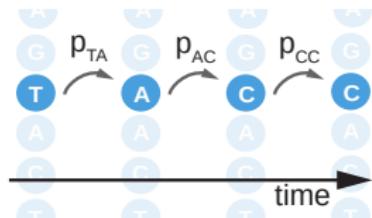
Definition of a Markov chain (see also [Ross, 1996])

stochastic process, i.e. a series of random experiments through time

lives on a **state space** and jumps to the different states

memorylessness: the probability of jumping to a state only depends on the current state

Nucleotide substitutions as MC



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Mathematical toolbox: Markov chains

Definition of a Markov chain (consult [Ross, 1996] for further information on MC):

1. Stochastic process, i.e. a series of random experiments through time
 - ▶ a series of random variables $(X_t)_{t \in \mathcal{T}}$. If \mathcal{T} is discrete, the Markov chain is called **discrete**, otherwise **continuous**.
2. lives on a state space and jumps to the different states
 - ▶ state space \mathcal{S}
3. **memorylessness:** the probability of jumping to a state only depends on the current state
 - ▶ $P(X_{t_{n+1}} = x_{t_{n+1}} | X_{t_n} = x_{t_n}, X_{t_{n-1}} = x_{t_{n-1}}, \dots) = P(X_{t_{n+1}} = x_{t_{n+1}} | X_{t_n} = x_{t_n})$

If the transition probabilities on the state space do not change over time, the Markov chain is called **time homogenous**.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

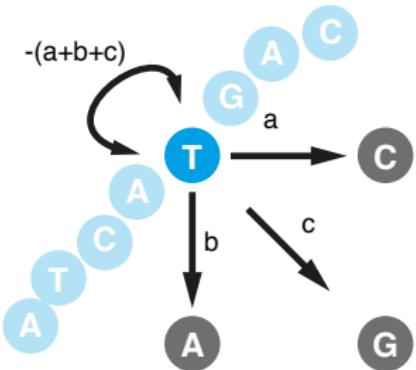
JC69

References

Nucleotide substitution as a Markov chain

State space of each nucleotide position: $\mathcal{S} = \{T, C, A, G\}$

Example: Assume the process is at state T



Substitution rate matrix:

$$Q = \begin{pmatrix} T & C & A & G \\ T & -(a+b+c) & a & b & c \\ C & d & -(d+e+f) & e & f \\ A & g & h & -(g+h+i) & i \\ G & j & k & l & -(j+k+l) \end{pmatrix}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

Rates versus probabilities

Rate: measures events per time unit

- ▶ deterministic, fixed quantity
- ▶ examples: birth rate (offspring per year), substitution rate (nucleotide changes per unit time)
- ▶ describes averages

Probability: measure of chance that a random event occurs

- ▶ stochastic
- ▶ examples: obtaining 6 when throwing a die, time to substitution (when seen as a random process)
- ▶ describes an exact event

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

From rates to probabilities

Let α be the **rate** of an event E happening per unit of time. The **probability** that this event happens in a very small (infinitesimally small) time step Δt is defined as $\alpha\Delta t$. We denote the time until the event happens with the random variable X. The probability that the event does not happen in Δt is:

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

From rates to probabilities

Let α be the **rate** of an event E happening per unit of time. The **probability** that this event happens in a very small (infinitesimally small) time step Δt is defined as $\alpha\Delta t$. We denote the time until the event happens with the random variable X. The probability that the event does not happen in Δt is:

$$P(X > \Delta t) = 1 - \alpha\Delta t$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

From rates to probabilities

Let α be the **rate** of an event E happening per unit of time. The **probability** that this event happens in a very small (infinitesimally small) time step Δt is defined as $\alpha\Delta t$. We denote the time until the event happens with the random variable X . The probability that the event does not happen in Δt is:

$$P(X > \Delta t) = 1 - \alpha\Delta t$$

Let τ be a time with $\tau = k\Delta t$. We can divide the probability that E does not happen in τ into k time intervals in which the event does not happen:

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

From rates to probabilities

Let α be the **rate** of an event E happening per unit of time. The **probability** that this event happens in a very small (infinitesimally small) time step Δt is defined as $\alpha\Delta t$. We denote the time until the event happens with the random variable X . The probability that the event does not happen in Δt is:

$$P(X > \Delta t) = 1 - \alpha\Delta t$$

Let τ be a time with $\tau = k\Delta t$. We can divide the probability that E does not happen in τ into k time intervals in which the event does not happen:

$$P(X > \tau) = (1 - \alpha\Delta t)^k = (1 - \alpha\Delta t)^{\tau/\Delta t} \xrightarrow[\Delta t \rightarrow 0]{} e^{-\alpha\tau}$$

The limit at the end of the latter equation holds true, because of the definition of the exponential function $e^x = \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

From rates to probabilities

Let α be the **rate** of an event E happening per unit of time. The **probability** that this event happens in a very small (infinitesimally small) time step Δt is defined as $\alpha\Delta t$. We denote the time until the event happens with the random variable X . The probability that the event does not happen in Δt is:

$$P(X > \Delta t) = 1 - \alpha\Delta t$$

Let τ be a time with $\tau = k\Delta t$. We can divide the probability that E does not happen in τ into k time intervals in which the event does not happen:

$$P(X > \tau) = (1 - \alpha\Delta t)^k = (1 - \alpha\Delta t)^{\tau/\Delta t} \xrightarrow{\Delta t \rightarrow 0} e^{-\alpha\tau}$$

The limit at the end of the latter equation holds true, because of the definition of the exponential function $e^x = \lim_{n \rightarrow \infty} (1 + \frac{x}{n})^n$.

$$\Rightarrow P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

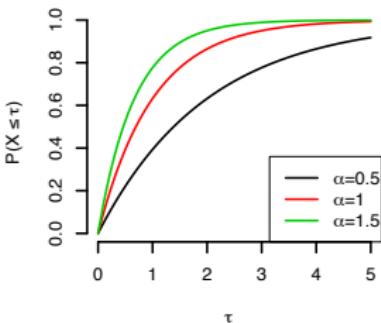
Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

From rates to probabilities

$P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$ is the **cumulative density function**



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

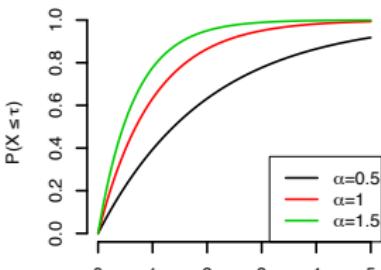
Substitution rate matrices

Markov models of nucleotide substitution: JC69

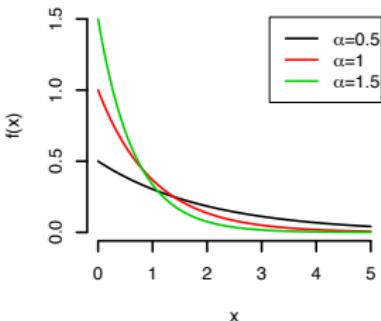
References

From rates to probabilities

$P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$ is the **cumulative density function**



$f(x) = \frac{dP}{dt}(x) = \alpha e^{-\alpha x}$ is the **probability density function** of an **exponential distribution**



03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

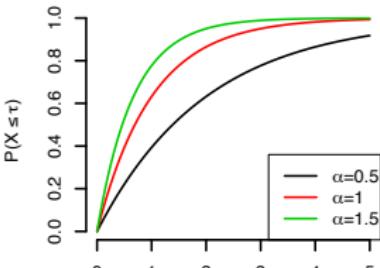
Substitution rate matrices

Markov models of nucleotide substitution: JC69

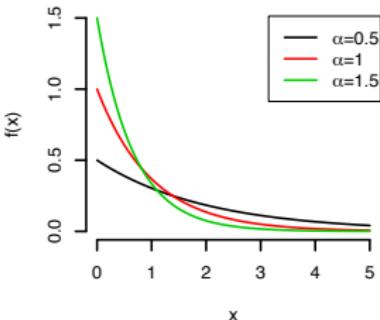
References

From rates to probabilities

$P(0 \leq X \leq \tau) = 1 - e^{-\alpha\tau}$ is the **cumulative density function**



$f(x) = \frac{dP}{dt}(x) = \alpha e^{-\alpha x}$ is the **probability density function** of an **exponential distribution**



An event occurring with rate α means that it occurs after an exponentially distributed waiting time with parameter α .

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

From rate matrix to transition probabilities

Let $P(t) = (p_{ij}(t))_{i,j \in S}$ be the *transition probability matrix* with all probabilities that given the Markov chain is in state i at time 0, the Markov chain will be in state j at time t . When we look at an infinitesimally small time step Δt , in which only one event can happen, we can calculate the transition probability at time $t + \Delta t$ with the same trick as used for the transformation from rate to probabilities:

$$P(t + \Delta t) = P(t) + P(t)Q\Delta t$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

From rate matrix to transition probabilities

Let $P(t) = (p_{ij}(t))_{i,j \in S}$ be the *transition probability matrix* with all probabilities that given the Markov chain is in state i at time 0, the Markov chain will be in state j at time t . When we look at an infinitesimally small time step Δt , in which only one event can happen, we can calculate the transition probability at time $t + \Delta t$ with the same trick as used for the transformation from rate to probabilities:

$$P(t + \Delta t) = P(t) + P(t)Q\Delta t \quad \Leftrightarrow \quad \frac{P(t + \Delta t) - P(t)}{\Delta t} = P(t)Q$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

From rate matrix to transition probabilities

Let $P(t) = (p_{ij}(t))_{i,j \in S}$ be the *transition probability matrix* with all probabilities that given the Markov chain is in state i at time 0, the Markov chain will be in state j at time t . When we look at an infinitesimally small time step Δt , in which only one event can happen, we can calculate the transition probability at time $t + \Delta t$ with the same trick as used for the transformation from rate to probabilities:

$$P(t + \Delta t) = P(t) + P(t)Q\Delta t \quad \Leftrightarrow \quad \frac{P(t + \Delta t) - P(t)}{\Delta t} = P(t)Q$$

Now we let $\Delta t \rightarrow 0$, thus it follows:

$$\lim_{\Delta t \rightarrow 0} \frac{P(t + \Delta t) - P(t)}{\Delta t} = \frac{dP}{dt}(t) = P(t)Q$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

From rate matrix to transition probabilities

Let $P(t) = (p_{ij}(t))_{i,j \in S}$ be the *transition probability matrix* with all probabilities that given the Markov chain is in state i at time 0, the Markov chain will be in state j at time t . When we look at an infinitesimally small time step Δt , in which only one event can happen, we can calculate the transition probability at time $t + \Delta t$ with the same trick as used for the transformation from rate to probabilities:

$$P(t + \Delta t) = P(t) + P(t)Q\Delta t \quad \Leftrightarrow \quad \frac{P(t + \Delta t) - P(t)}{\Delta t} = P(t)Q$$

Now we let $\Delta t \rightarrow 0$, thus it follows:

$$\lim_{\Delta t \rightarrow 0} \frac{P(t + \Delta t) - P(t)}{\Delta t} = \frac{dP}{dt}(t) = P(t)Q$$

This is a differential equation with the solution:

$$P(t) = e^{Qt}$$

!!The substitution rate matrix defines the transition probabilities!!

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References

Meaning of $P(t) = e^{Qt}$

As just derived:

$$P(t) = e^{Qt}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Meaning of $P(t) = e^{Qt}$

As just derived:

$$P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Meaning of $P(t) = e^{Qt}$

As just derived:

$$P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!}$$

Every possible series of states is visited in time t !

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

- One can obtain the same result using the theorem of Chapman-Kolmogorov

Why Markov chains are a great model for nucleotide substitutions

- ▶ memorylessness: a nucleotide substitution happens independently from the substitution history at this site
- ▶ substitution rate matrix defines the transition probabilities
- ▶ the transition probabilities take into account every possible substitution path

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Why Markov chains are a great model for nucleotide substitutions

- ▶ memorylessness: a nucleotide substitution happens independently from the substitution history at this site
- ▶ substitution rate matrix defines the transition probabilities
- ▶ the transition probabilities take into account every possible substitution path

- ▶ applying theories of linear algebra we can calculate the transition probability matrix according to:
$$P(t) = e^{Qt} = U \text{diag}(e^{\epsilon_1 t}, e^{\epsilon_2 t}, e^{\epsilon_3 t}, e^{\epsilon_4 t}) U^{-1}$$
 - ▶ cookbook recipe provided on moodle
 - ▶ for further information on how to derive this formula please consult a textbook on linear algebra

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

03: Nucleotide substitution models

- GWAS
- Measures of variability
- Sequence distance
- Definition of distance
- Nucleotide substitution as a Markov chain
- Substitution rate matrices
- Markov models of nucleotide substitution: JC69

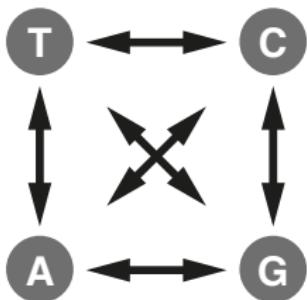
References

Substitution rate matrices and their properties.

The easiest substitution model: JC69

JC69:

- ▶ named after TH Jukes, CR Cantor: Evolution of protein molecules. 1969 [Jukes and Cantor, 1969].
- ▶ all substitution have the same rate, λ



Substitution rates:

$$\begin{matrix} & \text{T} & \text{C} & \text{A} & \text{G} \\ \text{T} & \cdot & \lambda & \lambda & \lambda \\ \text{C} & \lambda & \cdot & \lambda & \lambda \\ \text{A} & \lambda & \lambda & \cdot & \lambda \\ \text{G} & \lambda & \lambda & \lambda & \cdot \end{matrix}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

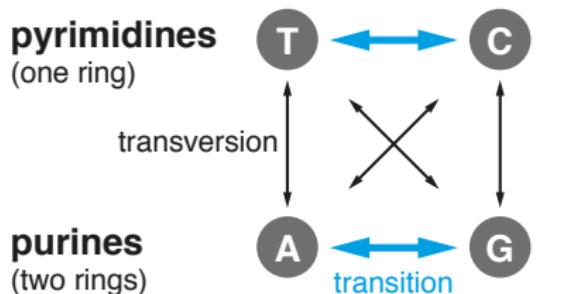
Markov models of nucleotide substitution: JC69

References

Accounting for transition/transversion: K80

K80:

- ▶ named after M Kimura: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. 1980. [Kimura, 1980]
- ▶ transitions happen at rate α , transversions at rate β



Substitution rates:

	T	C	A	G
T	.	α	β	β
C	α	.	β	β
A	β	β	.	α
G	β	β	α	.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

Including equilibrium frequencies: TN93 and HKY

TN93:

- ▶ named after [Tamura and Nei, 1993]
- ▶ transitions between T \leftrightarrow C happen at rate $\alpha_1 \times$ (nucleotide equilibrium frequency)
- ▶ transitions from A \leftrightarrow G happen at rate $\alpha_2 \times$ (nucleotide equilibrium frequency)
- ▶ transversions happen at rate $\beta \times$ (nucleotide equilibrium frequency)

Substitution rates:

$$\begin{array}{cccc}
 & T & C & A & G \\
 T & \cdot & \alpha_1 \pi_C & \beta \pi_A & \beta \pi_G \\
 C & \alpha_1 \pi_T & \cdot & \beta \pi_A & \beta \pi_G \\
 A & \beta \pi_T & \beta \pi_C & \cdot & \alpha_2 \pi_G \\
 G & \beta \pi_T & \beta \pi_C & \alpha_2 \pi_A & \cdot
 \end{array}$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Including equilibrium frequencies: TN93 and HKY

TN93:

- ▶ named after [Tamura and Nei, 1993]
- ▶ transitions between T \leftrightarrow C happen at rate $\alpha_1 \times$ (nucleotide equilibrium frequency)
- ▶ transitions from A \leftrightarrow G happen at rate $\alpha_2 \times$ (nucleotide equilibrium frequency)
- ▶ transversions happen at rate $\beta \times$ (nucleotide equilibrium frequency)

Substitution rates:

$$\begin{array}{cccc} & T & C & A & G \\ T & \cdot & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ C & \alpha_1\pi_T & \cdot & \beta\pi_A & \beta\pi_G \\ A & \beta\pi_T & \beta\pi_C & \cdot & \alpha_2\pi_G \\ G & \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & \cdot \end{array}$$

▶ Note that the special case of $\alpha_1 = \alpha_2$ was described earlier [Hasegawa et al., 1984] and is named HKY

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

A more general substitution model: GTR

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GTR (REV):

- ▶ generalised time-reversible model
- ▶ based on three papers:
[Tavaré, 1986, Yang, 1994, Zharkikh, 1994]

Substitution rates:

$$\begin{array}{cccc}
 & T & C & A & G \\
 T & \cdot & a\pi_C & b\pi_A & c\pi_G \\
 C & a\pi_T & \cdot & d\pi_A & e\pi_G \\
 A & b\pi_T & d\pi_C & \cdot & f\pi_G \\
 G & c\pi_T & e\pi_C & f\pi_A & \cdot
 \end{array}
 \quad
 \begin{array}{l}
 + \text{ quite flexible} \\
 + \text{ time-reversible} \\
 - \text{ not completely general}
 \end{array}$$

Time reversibility

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

$$\begin{pmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{pmatrix}$$

$$= \begin{pmatrix} \cdot & a & b & c \\ a & \cdot & d & e \\ b & d & \cdot & f \\ c & e & f & \cdot \end{pmatrix} \cdot \begin{pmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{pmatrix}$$

The most general substitution model

UNREST:

- ▶ unrestricted model first described in [Yang, 1994]
- ▶ each substitution has a (different) rate

Substitution rates:

$$\begin{array}{cccc} & T & C & A & G \\ T & \cdot & a & b & c \\ C & d & \cdot & e & f \\ A & g & h & \cdot & i \\ G & j & k & l & \cdot \end{array}$$

- + most general case
- + all other models are special cases of UNREST
- mathematically very complicated and not handy to use
- not time-reversible

- ▶ Further models described in [Yang, 1994]

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

Substitution models

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

model	parameters	description
JC69	1	all substitutions have the same rate
K80	2	accounts for transition and transversions
HKY	2+3*	distinction between transition and transversions, including equilibrium frequencies
TN93	3+3*	different rates for transitions
GTR	6+3*	general, but still time-reversible
UNREST	12	most general, not time-reversible

* equilibrium frequencies of nucleotides

Calculating transition probabilities and sequence distance.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

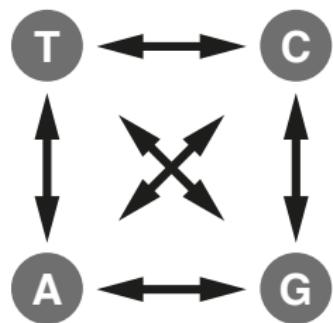
Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References



Substitution rates:

$$Q = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$

JC69: Approximation for small time, t

If t is very small in comparison to λ , we can use the first two terms of the Taylor expansion $P(t) = e^{Qt} = I + \sum_{i=1}^{\infty} \frac{(Qt)^i}{i!}$ to get

$$\rightarrow p_{i,j}(t) \approx \lambda t \text{ and } p_{i,i}(t) \approx 1 - 3\lambda t$$

Example: Assume a region of the human genome evolves according to JC69 at the rate $\lambda = 2.2/3 \times 10^{-9} \frac{\text{substitutions per site}}{\text{year}}$. The probability that starting with a T, after $t = 10^6$ years, we observe a C is

$$p_{TC} = \lambda t = 7.3 \times 10^{-4}$$

and that it still is a T is

$$p_{TT} = 1 - 3\lambda t = 0.9978.$$

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

What we learnt today

- ▶ case control setup in a GWAS
- ▶ sequence distance included non-visible substitution steps in between
- ▶ we model sequence evolution with a Markov chain framework
 - ▶ memoryless
 - ▶ toolbox: from rates to probabilities
 - ▶ sequence evolution model is defined by its substitution rate matrix
- ▶ rate assumptions in different substitution models:
 - ▶ JC69: all substitutions have the same rate
 - ▶ K80: transition and transversions have different rates
 - ▶ other (more general) models: [Yang, 2014]
- ▶ approximation of distance for JC69

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution: JC69

References

GWAS, molecular evolution: Questions

- ① In a GWAS, why can you not reject your null hypothesis if the p-value is $< \alpha$?
- ② Why is the Markov Chain model a good model for sequence evolution?
- ③ Why is it not advisable to reconstruct a phylogeny based on the Hamming distance?

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:
JC69

References

References |

- Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12):e1002822–11.
- Hasegawa, M., Yano, T., and Klishino, H. (1984). A New Molecular Clock of Mitochondrial-Dna and the Evolution of Hominoids. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences*, 60(4):95–98.
- Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism.*, pages 21–123.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, NY)*, 308(5720):385–389.
- Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *Jama-Journal of the American Medical Association*, 299(11):1335–1344.
- Ross, S. M. (1996). *Stochastic Processes. Second edition*. Wiley.
- Sokal, R. and Rohlf, F. (2012). *Biometry. Fourth Edition*. W.H. Freeman and Company.
- Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some mathematical questions in biology—DNA sequence analysis (New York, 1984)*, pages 57–86. Amer. Math. Soc., Providence, RI.
- Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39(1):105–111.
- Yang, Z. (2014). *Molecular Evolution – A Statistical Approach*. Oxford University Press.
- Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of molecular evolution*, 39(3):315–329.

03: Nucleotide substitution models

GWAS

Measures of variability

Sequence distance

Definition of distance

Nucleotide substitution as a Markov chain

Substitution rate matrices

Markov models of nucleotide substitution:

JC69

References