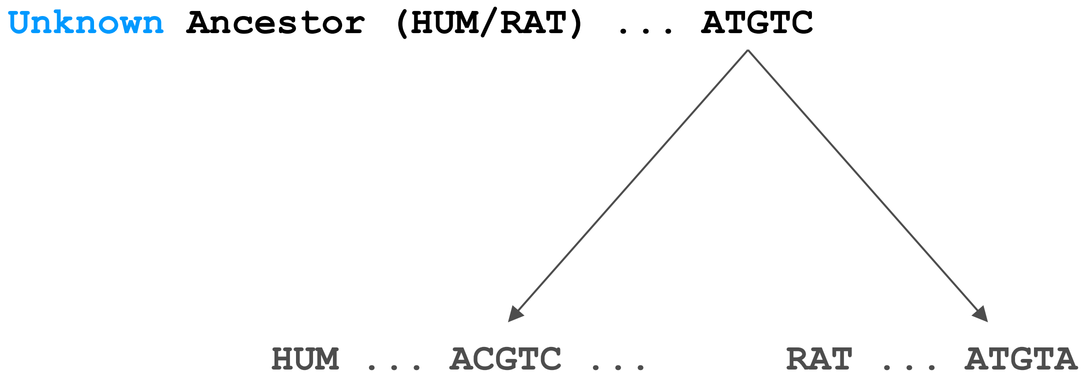
Evolutionary Distances



Unknown Ancestor (HUM/RAT) ... ATGTC

HUM ... ACGTC ...

RAT ... ATGTA

Genomics and Bioinformatics

Chapter 11

Evolution is the interplay of mutation and selection

attgcgatgca	attgcgatgca			
attgccatgca	attgccatgca			neutral
aatgcgatgca	aatgcgatgca	aatgcgatgca	aatgcgatgca	advantageous
attgcaatgca				disadvantageous
~~attgcgatgct~~	~~attgcgatgct~~			lethal

Mutation generates diversity
Selection affects the frequency of sequences in the gene pool

Evolution is directed in time

attgcgatgca attgcgatgca

Ancestor sequences

attgccatgca attgccatgca

aatgcgatgca aatgcgatgca aatgcgatgca

attgcaatgca

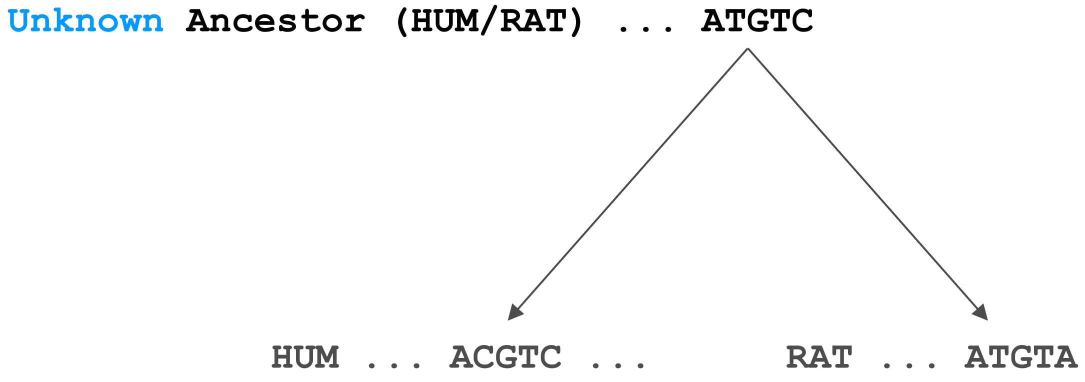
many descendants of the same ancestor

Many sequences that we observe today go back to the same ancestral sequence. They are called homologous sequences.

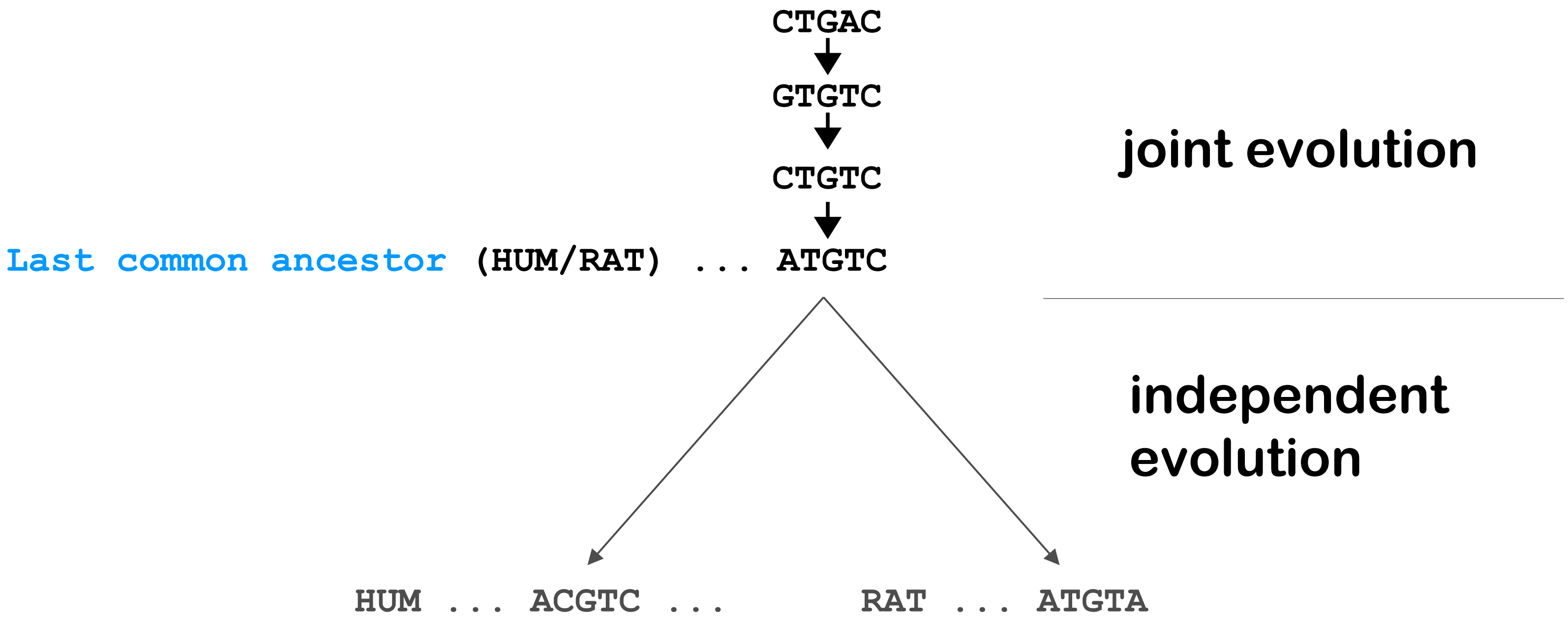
Homologous sequences we observe today are brothers rather than parents and children

```
HUM ...ACGTCAAGGCCGCCTGGGGCAACGTCAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGAGAATGTTCC...
RAT ...ATGTAAGCCCCGGCTCTGCCCATGTAAGGTCAAGGCTCACGGCAAGAAGGTTGCTGATCCAAAGCTGC...
```

They share common ancestors
Almost always the ancestor is unknown



For every pair of homologous sequences there exists a last common ancestor (LCA)



CTGAC

GTGTC

CTGTC

Last common ancestor (HUM/RAT) ... ATGTC

joint evolution

independent evolution

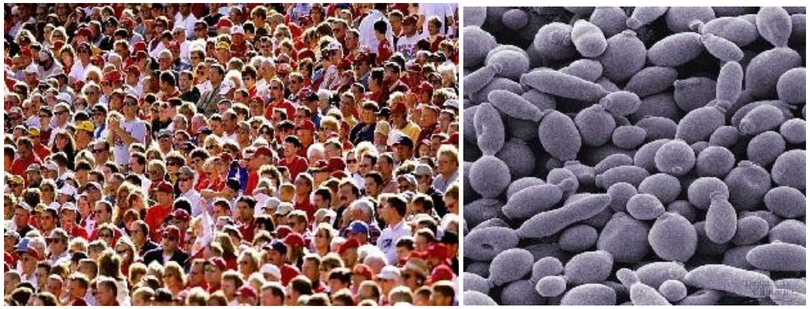
HUM ... ACGTC ...

RAT ... ATGTA

For some pairs the last common ancestor is recent, for some it lived long ago

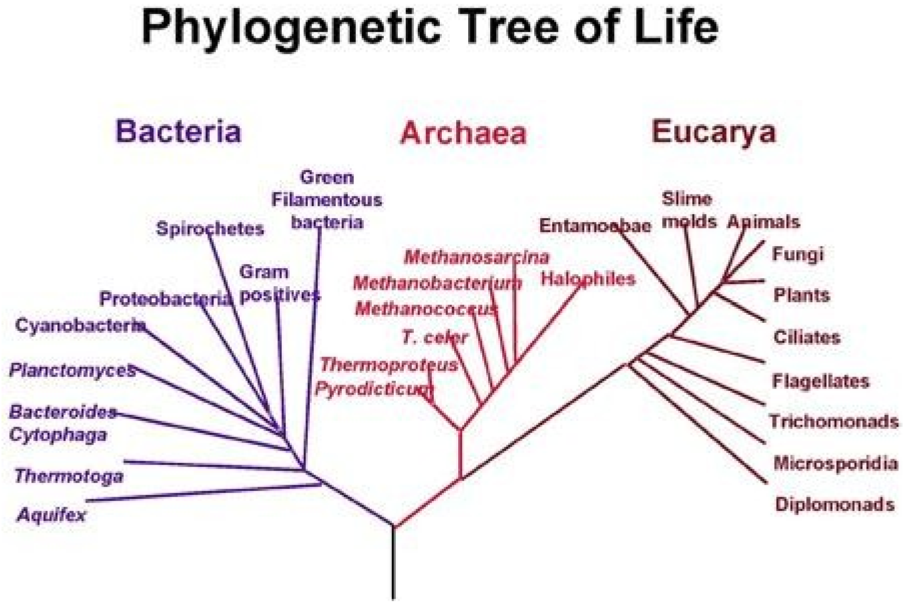
LCA of man and chimp lived about 6 million years ago

That's recent



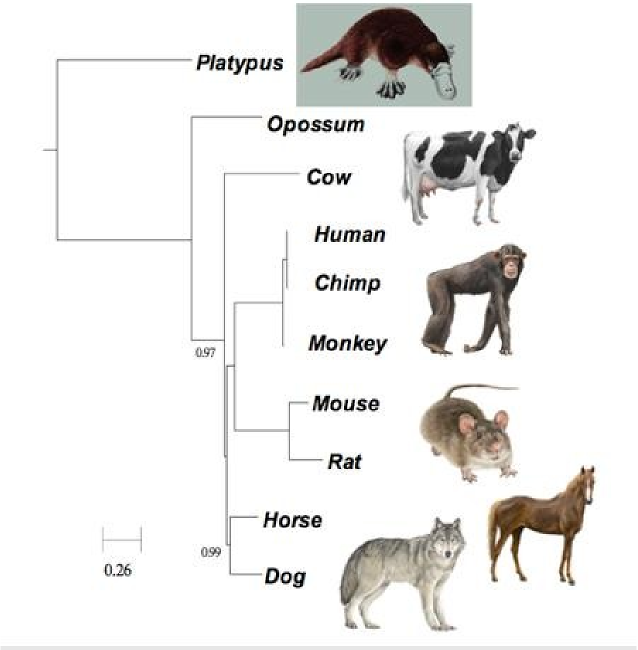
LCA of man and yeast lived about 1 billion years ago

All genomes on earth are believed to go back to common ancestors



Some genomes have a more recent common ancestor than others. Their genomes are more similar.

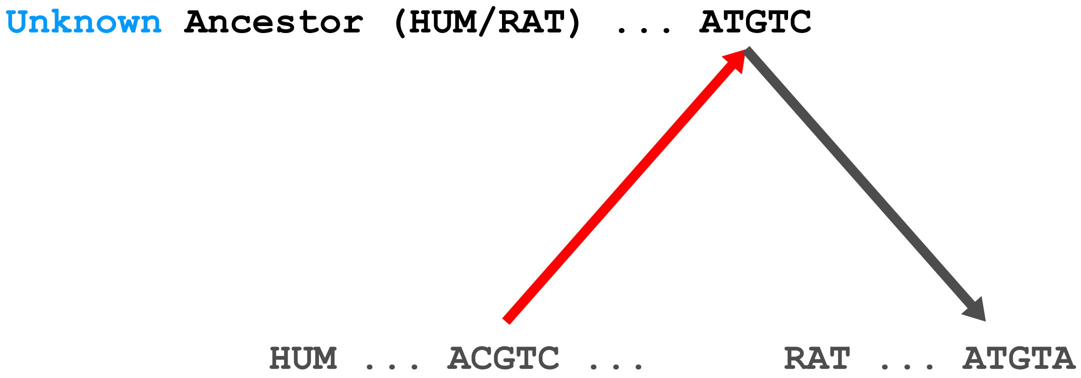
The more recent the common ancestor the less mismatches accumulated in the sequences



Organism	Mismatches compared to human sequence
Chimp	0
Mouse	9
Fruit Fly	29
Wheat	43
Yeast	51

Cytochrome C: An enzyme needed for respiration

The mismatches we observe can result from mutations in both evolutionary branches



Unknown Ancestor (HUM/RAT) ... ATGTC

HUM ... ACGTC ...

RAT ... ATGTA

```
HUM ... ACGTC ...
        ATGTC
RAT ... ATGTA ...
```

We can reverse the mutations in one branch and concatenate both branches resulting in a “mutation history” that transfers one observed sequence into another

The more mutations the more expected mismatches

GAGTATGGTGCGGAGAGAAT

A sequence of length 20

GACTACGGTGAAGAGAGAAC

In 20 million years 5 random positions mutated

This resulted in 5 mismatch positions

How many mismatches do you expect after 40 million years?

The same sequence position can mutate multiple times

GAGTATGGTGCGGAGAGAAT

A sequence of length 20

GACTACGGTGAAGAGAGAAC

In 20 million years 5 random positions mutated

CACTAGGGTGCAGAGCGACC

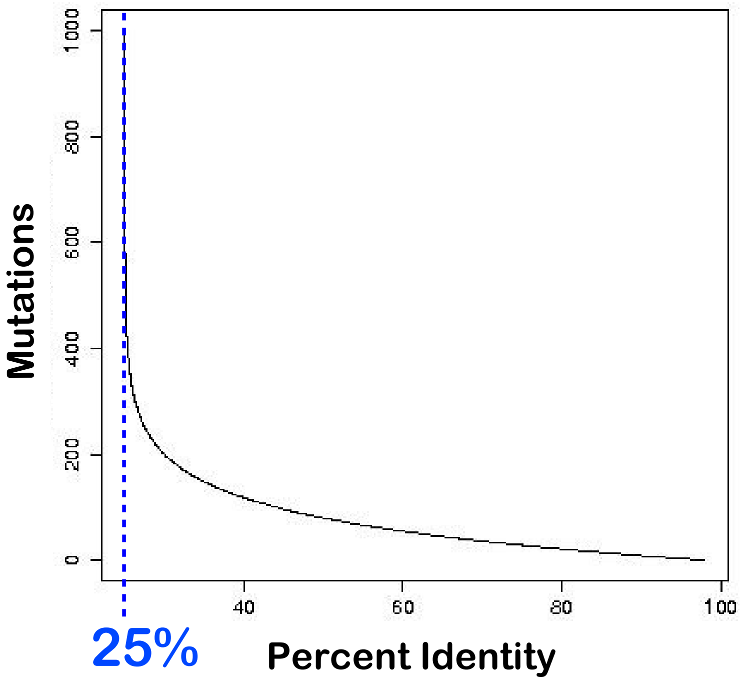
In another 20 million years 5 more random positions mutated

CACTAGGGTGCAGAGCGACC

The 10 mutations lead to only 7 mismatch positions

How many mismatch positions do you expect after 50.000 random mutations?

The percentage of mismatches does not grow linearly with the number of mutations



After thousands of mutations the sequences will be totally randomized

With 4 letters you expect 25% identity of two unrelated sequences by chance

How can we get a formula for the curve above?

We count A positions in sequence 2

```
HUM ...ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC...
RAT ...ATGTAAGCCCCGGCTCTGCCCAGGTCAAGGCTCACGGCAAGAAGGTTGCTGATGCCCTGGCCAAAGCTGC...
```

Count all positions:

absolute frequency: 15
relative frequency: 0.21

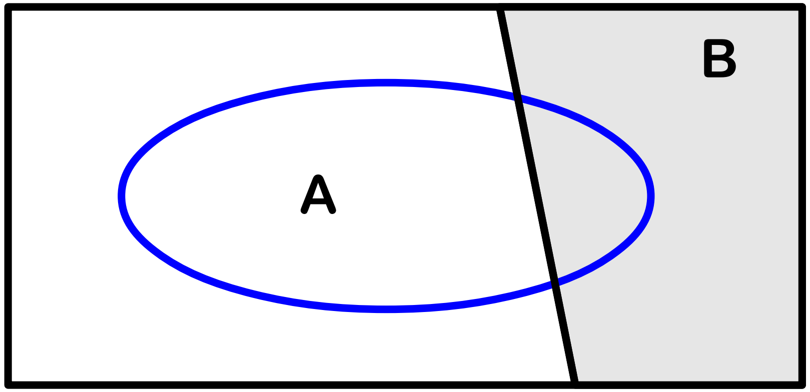
```
HUM ...ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC...
RAT ...ATGTAAGCCCCGGCTCTGCCCAGGTCAAGGCTCACGGCAAGAAGGTTGCTGATGCCCTGGCCAAAGCTGC...
```

Count only positions with an A in sequence 1:

absolute frequency: 12
relative frequency: 0.75

Restricting a data set changes relative frequencies

$$r(A|B) = \frac{h(A \cap B)}{h(B)} = \frac{r(A \cap B)}{r(B)}$$



A: all data points with property 1

B: all data points with property 2

h: absolute frequency

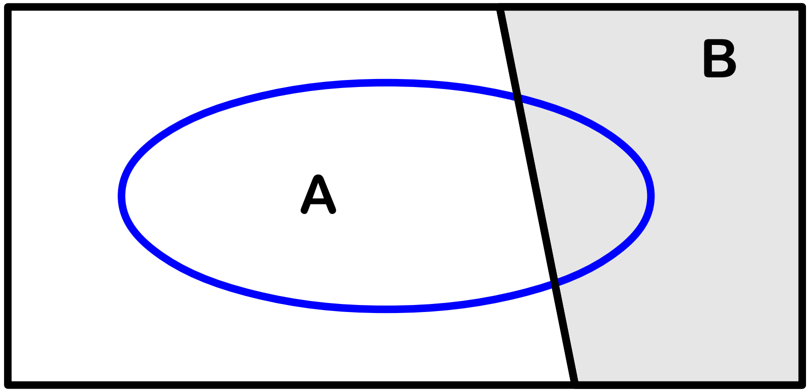
r: relative frequency

A|B: A restricted to B

The probabilistic analogue to data set restriction is conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B



$$P(A \cap B) = P(A)\,P(B|A) = P(B)\,P(A|B)$$

Two events A and B are independent if conditioning on B does not change the probability of A

$$\boxed{P(A|B) = P(A)}$$

From

$$P(A \cap B) = P(A)\,P(B|A) = P(B)\,P(A|B)$$

we get the multiplication rule for independent events

$$P(A \cap B) = P(A)\,P(B)$$

One random experiment that depends on the outcome of another

```
1. experiment $X \sim (1/4,1/4,1/4,1/4)$
2. experiment
   If $X$ = „A“:
         $Y \sim (0.7,0.1,0.1,0.1)$
   If $X$ =„C“:
         $Y \sim (0.1,0.7,0.1,0.1)$
   If $X$ = „G“:
         $Y \sim (0.1,0.1,0.7,0.1)$
   If $X$=„T“:
         $Y \sim (0.1,0.1,0.1,0.7)$
```

The outcome of X determines the distribution of Y

Dependent random experiments can be described by conditional probabilities

1. experiment X~(1/4,1/4,1/4,1/4)

2. experiment

If X = „A“:

Y~ (0.7,0.1,0.1,0.1)

If X =„C“:

Y~ (0.1,0.7,0.1,0.1)

If X = „G“:

Y~ (0.1,0.1,0.7,0.1)

If X=„T“:

Y~ (0.1,0.1,0.1,0.7)

P(X=A and Y=A)

= P(X=A) P(Y=A | X=A)

= 0.25 x 0.7

= 0.175

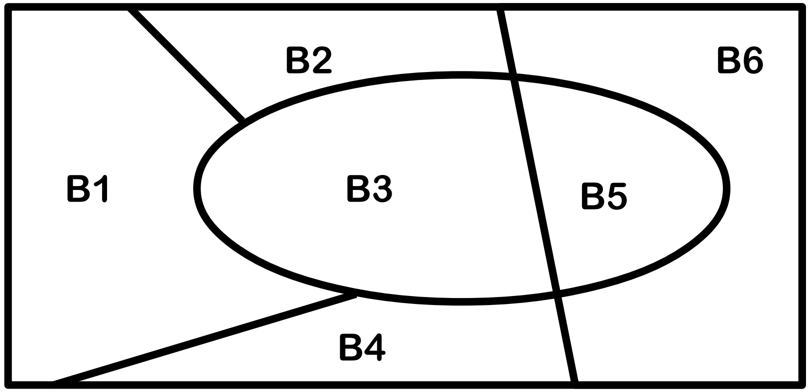
≠ P(X=A) P(Y=A)

= 0.25 x 0.25

= 0.0625

$$P(A \cap B) = P(A)\,P(B|A) = P(B)\,P(A|B)$$

A partition is a family of disjoint sets whose union is the full space



$$B_1, \dots, B_n$$

$$B_i \subset \Omega$$

$$B_i \cap B_j = \emptyset$$

$$B_1 \cup \cdots \cup B_n = \Omega$$

Law of total probability

For an event A and a partition B_1, ..., B_n we have

$$A = (A \cap B_1) \cup \ldots, \cup (A \cap B_n)$$

Since the sets are disjoint we have

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_n)$$

and

$$\boxed{P(A) = \sum_{i=1}^{n} P(A|B_i)\, P(B_i)}$$

Dependent random experiments can be described by conditional probabilities

```
1. experiment X~(1/4,1/4,1/4,1/4)
2. experiment
   If X = „A“:
         Y~ (0.7,0.1,0.1,0.1)
   If X =„C“:
         Y~ (0.1,0.7,0.1,0.1)
   If X = „G“:
         Y~ (0.1,0.1,0.7,0.1)
   If X=„T“:
         Y~ (0.1,0.1,0.1,0.7)
```

P(Y=A)=

P(X=A) P(Y=A|X=A)
+
P(X=C) P(Y=A|X=C)
+
P(X=G) P(Y=A|X=G)
+
P(X=T) P(Y=A|X=T)

= 0.25x0.7 + 3x0.25x0.1

= 0.25

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\, P(B_i)$$

Bayes' Theorem

We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}$$

$$= \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|A^c)\,P(A^c)}$$

… by the law of total probability
note that (A, A^c) is a partition

Bayes Theorem transforms the term P(A|B) into terms involving P(B|A)

Law of inverse probability

Zonks

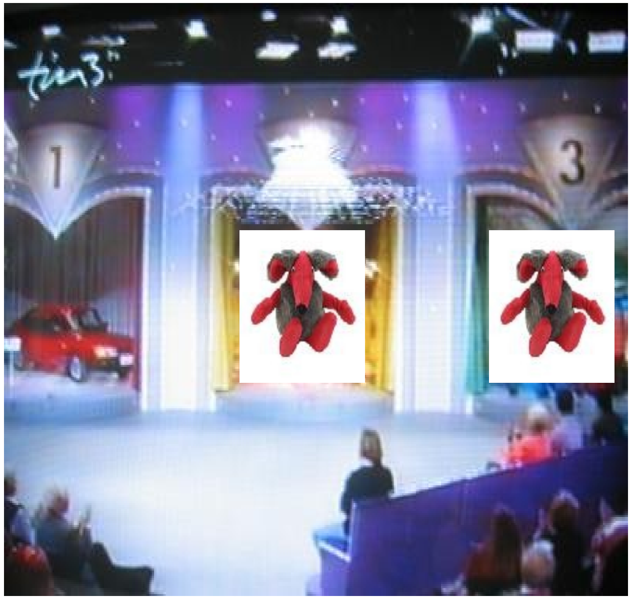
The candidate can choose between 3 closed doors

Behind one of the doors is a new car, behind the two others are Zonks (stuffed animal ca. 3 Euro)

Candidate picks door 2

We have

P(door 2 = car) = 1/3



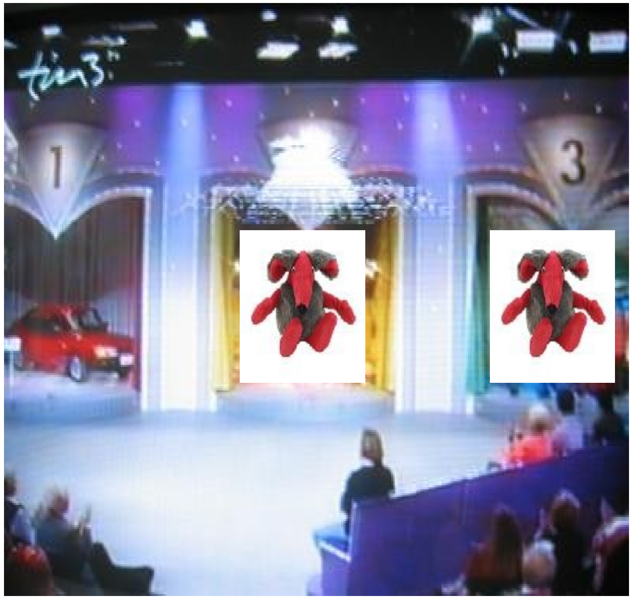
The show proceeds always in the same pattern

The host opens a door with a Zonk that the candidate has not chosen

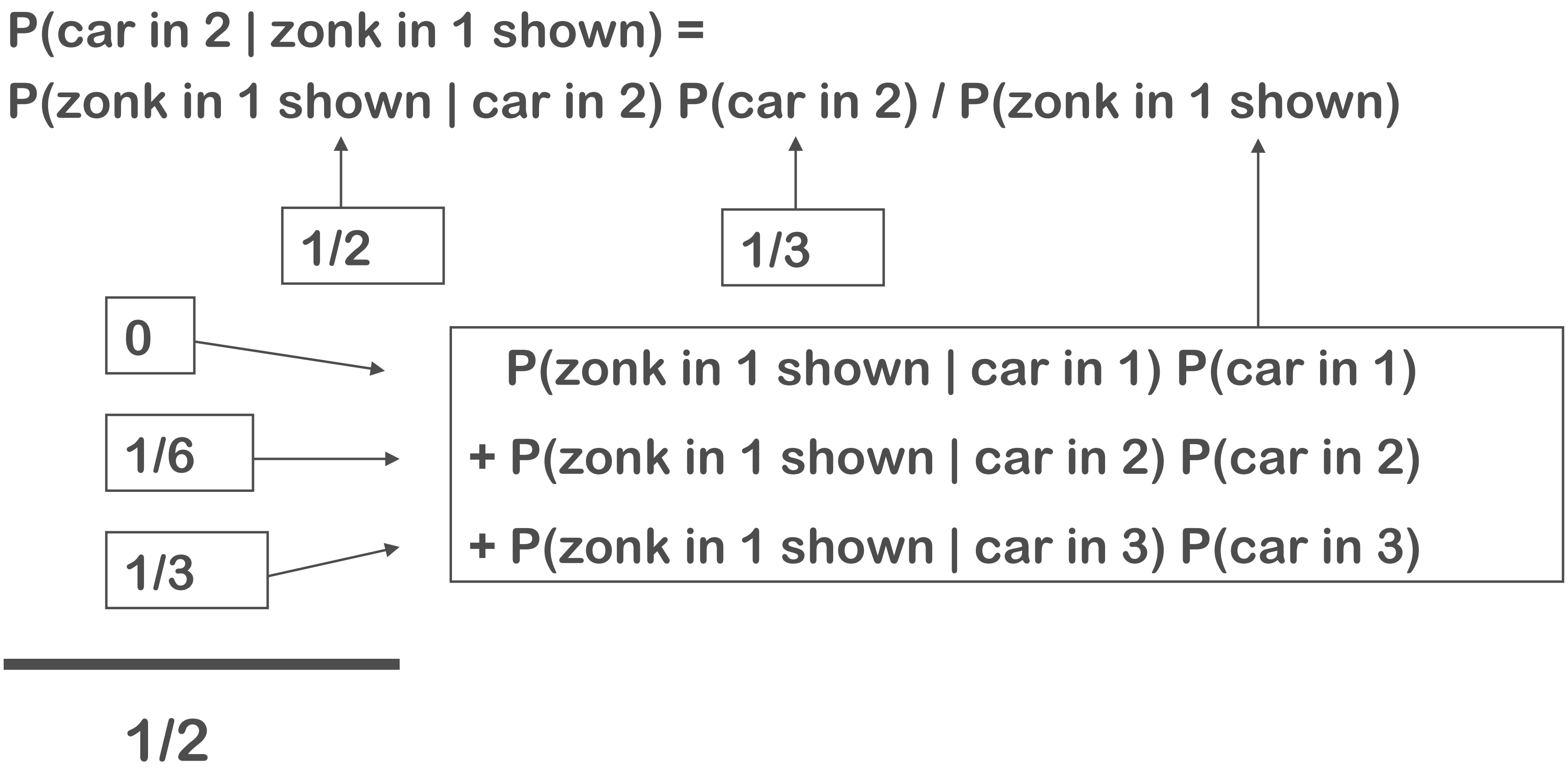
(Note that this is always possible)

With only two doors left he asks the candidate whether he wants to switch doors

Do you switch?



Here is what Bayes' rule has to say



P(car in 2 | zonk in 1 shown) =

P(zonk in 1 shown | car in 2) P(car in 2) / P(zonk in 1 shown)

1/2

1/3

0

1/6

1/3

P(zonk in 1 shown | car in 1) P(car in 1)

+ P(zonk in 1 shown | car in 2) P(car in 2)

+ P(zonk in 1 shown | car in 3) P(car in 3)

1/2

P(car in 2 | zonk in 1 shown) = 1/3

P(car in 3 | zonk in 1 shown) =

1- P(car in 2 | zonk in 1 shown) = 2/3 !

You can double your odds by changing doors

Before door 1 was opened:

door:	1	2	3
P [car]:	1/3	1/3	1/3

After door 1 was opened:

door:	1	2	3
P [car]:	0	1/3	2/3

Think of a lethal painfully killing disease

There is a test for the disease

It has
Sensitivity 0.95 - P(test positive | you are ill)
Specificity 0.98 - P(test negative | you are fine)

1 out 100,000 people have the disease

You are tested and the test is positive

SHIT!!!!!

What is the probability that you have the disease?

New hope using Bayes' theorem

.95 sensitivity: $P(T+|D+)=.95$
.98 specificity: $P(T-|D-) =.98$
or $P(T+|D-)=.02$

Bayes Theorem:

$$P(D+|T+) = P(T+|D+)P(D+) / P(T+)$$

where $P(T+) = P(T+|D+)P(D+) + P(T+|D-)P(D-)$

$$P(D+|T+)=(.95 \times .00001)/(.95 \times .00001 + .02 \times .99999)$$
$$= .00047$$

In spite of the positive test the probability that you are ill is less then 0.05 %

The concept of conditional probability can be extended to conditional distributions

For two discrete random variables X and Y we have

$$P(X = x|Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

If the distribution of X is

$X \sim (p_1, \ldots, p_n)$ **with values** $(x_1, \ldots, x_n)$,

the conditional distribution of X given Y=y is given by

$$X|\{Y = y\} \sim (q_1, \ldots, q_n) \quad q_i = \frac{P(X=x_i, Y=y)}{P(Y=y)}$$

The law of total probability allows us to calculate non conditional distributions from conditional distributions

If the distribution of Y is

$Y \sim (r_1, \ldots, r_n)$ **with values** $(y_1, \ldots, y_n)$

and the conditional probabilities of X are given by

$$P(X = x_i|\, Y = y_j) = q_{ij}$$

Since { Y=y$_i$ }, i=1,...n, is a partition we have

$$P(X = x_i) = \sum_{j=1}^{n} P(X = x_i|Y = y_j)P(Y = y_j)$$

or the distribution of X is given by

$X \sim (p_1, \ldots, p_n)$ **where** $p_i = \sum_j q_{ij}\, r_j$

Example with two binary variables

X~(0.3,0.7)

Y| X=1~ (0.25,0.75)

Y| X=0 ~ (0.75,0.25)

What is the distribution of Y?

P(Y=0)= ?

P(Y=1)= ?

P(Y=0) = P(X=0)P(Y=0|X=0)

+ P(X=1)P(Y=0|X=1)

= 0.3 x 0.75 + 0.7 x 0.25

= 0.4

P(Y=1) = 1 – P(Y=0)

= 0.6

Y~(0.4,0.6)

Two random variables can have the same distribution without being independent

```
1. experiment X~(1/4,1/4,1/4,1/4)
2. experiment
   If X = „A“:
         Y~ (0.7,0.1,0.1,0.1)
   If X =„C“:
         Y~ (0.1,0.7,0.1,0.1)
   If X = „G“:
         Y~ (0.1,0.1,0.7,0.1)
   If X=„T“:
         Y~ (0.1,0.1,0.1,0.7)
```

X and Y have the same distribution:

X~(1/4,1/4,1/4,1/4)

Y~(1/4,1/4,1/4,1/4)

X and Y are not independent

$P(Y=T|X=T) = 0.7$

$P(Y=T) = 0.25$

The n x n matrix of probabilities P(X=x,Y=y) defines the joint probability of X and Y

For two random variables with values in (A,C,G,T) we can write their joint distribution as

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix}$$

where for example m_{23} = P(X=C, Y=G)

The joint probability is a discrete probability distribution on the space of all possible combinations of outcomes (x_i,y_j)

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix}$$

For example m_{23} = P((X,Y)=(C,G))

Note that: $\sum_{i,j} m_{ij} = 1$

The joint distribution of independent random variables has a product form

$$X \sim (p_1, \ldots, p_n)$$

$$Y \sim (r_1, \ldots, r_n)$$

$$m_{ij} = p_i \, r_j$$

Dependent random variables yields a different joint distribution

```
1. experiment X~(1/4,1/4,1/4,1/4)
2. experiment
   If X = „A“:
          Y~ (0.7,0.1,0.1,0.1)
   If X =„C“:
          Y~ (0.1,0.7,0.1,0.1)
   If X = „G“:
          Y~ (0.1,0.1,0.7,0.1)
   If X=„T“:
          Y~ (0.1,0.1,0.1,0.7)
```

$$M = \begin{pmatrix} 0.175 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.175 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.175 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.175 \end{pmatrix}$$

The marginal distributions of X and Y do not determine their joint distribution

In both simulations X and Y have the same distributions:

X~(1/4,1/4,1/4,1/4)

Y~(1/4,1/4,1/4,1/4)

We call these distributions the marginal distributions of X and Y

However the joint distributions are different:

$$M = \begin{pmatrix} 0.175 & 0.025 & 0.025 & 0.025 \\ 0.025 & 0.175 & 0.025 & 0.025 \\ 0.025 & 0.025 & 0.175 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.175 \end{pmatrix} \quad M = \begin{pmatrix} 0.0625 & 0.0625 & 0.0625 & 0.0625 \\ 0.0625 & 0.0625 & 0.0625 & 0.0625 \\ 0.0625 & 0.0625 & 0.0625 & 0.0625 \\ 0.0625 & 0.0625 & 0.0625 & 0.0625 \end{pmatrix}$$

dependent independent

The joint distribution determines the marginal distributions

Given the joint distribution of X and Y

$$M = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{pmatrix} \qquad (X, Y) \sim M$$

Both {X=x_1}, … {X=x_n} and {Y=y_1},…,{Y=y_n} are partitions and the law of total probability defines the marginal distributions

$$P(X = x_i) = \sum_{y_j} P(X = x_i, Y = y_j) = \sum_j m_{ij}$$

Sum over the columns

$$P(Y = y_j) = \sum_{x_i} P(X = x_i, Y = y_j) = \sum_i m_{ij}$$

Sum over the rows

Dependent random experiments can be described by one marginal distribution and a transition matrix

```
1. experiment X~(1/4,1/4,1/4,1/4)
2. experiment
   If X = „A“:
         Y~ (0.7,0.1,0.1,0.1)
   If X =„C“:
         Y~ (0.1,0.7,0.1,0.1)
   If X = „G“:
         Y~ (0.1,0.1,0.7,0.1)
   If X=„T“:
         Y~ (0.1,0.1,0.1,0.7)
```

Marginal distribution of X

$$X \sim (p_1, \ldots, p_n)$$

A transition matrix P

$$P = \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \\ \end{array}$$

$$P = \left(\begin{array}{cccc} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{array} \right)$$

P encodes the design of the second experiment given the outcome of the first

Every column is a design

If the random variables are independent all columns of the transition matrix are the same

$$P = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

No matter what the outcome of X is. Y is always simulated by the same experimental design.

The columns of a transition matrix are (conditional) distributions

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix}$$

$$p_{ij} = P(Y = y_j | X = x_i)$$

All columns sum up to 1

$\sum_i p_{ij} = 1$ **for all j**

Multiplying the first marginal distribution vector with the transition matrix yields the second marginal distribution

$$X \sim (p_1, \ldots, p_n) \qquad Y \sim (q_1, \ldots, q_n)$$

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix} = (p_1, p_2, p_3, p_4) \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix}$$

$$\begin{aligned} q_j &= P(Y = y_i) \\ &= \sum_k P(Y = y_j | X = x_k) \, P(X = x_k) \\ &= \sum_k p_{kj} \, p_k \end{aligned}$$

We can model the evolutionary mutation process by a transition matrix

GAGTATGGTGCGGAGAGAAT

GACTACGGTGAAGAGAGAAC

CACTAGGGTGCAGAGCGACC

CACTAGGGTGCAGAGCGACC

For each position we iteratively apply a transition matrix of the form

$$P = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

where a is the probability of "no mutation" and 3 x b the probability of a mutation

We can define a unit of evolutionary divergence through a transition matrix

Normalize the transition matrix such that the expected number of mismatches is 1%

$$P(X = Y) = 0.99$$

This can be done by using the following matrix

$$P = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix} \qquad \begin{aligned} a &= 0.99 \\ b &= 1/300 \end{aligned}$$

Two sequences simulated by this model are 1 PAM apart
PAM = Point Accepted Mutations

We can apply the transition matrix to a sequence position by position

ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT
GGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCACGTCAAGGCCGCACGT
CAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCACGTCA
AGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC

To generate the second sequence:

Draw a random number for every position in the sequence

- **If it is larger than 0.01 the base in the second sequence is identical to the first sequence**
- **Else it is one of the other 3 bases with equal probability**

$$P = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

$$a = 0.99$$

$$b = 1/300$$

ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT
GGGGCAAGGTTGGCGCGCACGGGGAGCATGGTGCGGAGGCCCTGGAGAATGTTCCACGTCAAGGCCGCACGT
CAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCACTGGAGAATGTTCCACGTCA
AGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC

In average 1% of the positions will be mutated

1 PAM of evolution has operated on the first sequence

We can apply the transition matrix again to the second sequence to generate a third one

ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT
GGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCACGTCAAGGCCGCACGT
CAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCACGTCA
AGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC

↓

$$P = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

1 PAM of mutations

ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT
GGGGCAAGGTTGGCGCGCACGGGGAGCATGGTGCGGAGGCCCTGGAGAATGTTCCACGTCAAGGCCGCACGT
CAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCACTGGAGAATGTTCCACGTCA
AGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC

↓

$$P = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}$$

1 PAM of mutations

ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT
GGGGCAAGGTTGGCGCGCACGGGGAGCATGGTGCGGGGGCCCTGGAGAATGTTCCACGTCAAGGCCGCACGT
CAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCACTGGAGAATGTTCCACGTCA
AGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGAATGGTGCGGAGGCCCTGGAGAATGTTCC

The 1st and the 3rd sequence are 2 PAM apart

The transition matrix determines the expected number of match positions

Original sequence: X(0)

1x evolved sequence: X(1)

2x evolved sequence: X(2)

L: length of sequence

M(i,j): number of match positions between X(i) and X(j)

Expected number of matches:

$P(X_i(0)=X_i(1)) = 0.99$ These sequences are 1PAM apart

$E(M(0,1)) = 0.99 \times L$ 99% identity

$P(X_i(1)=X_i(2)) = 0.99$ These sequences are 1PAM apart

$E(M(1,2)) = 0.99 \times L$ 99% identity

What about the number of matches between X(0) and X(2)?

$P(X_i(0)=X_i(2)) = ?$ These sequences are 2PAM apart

$E(M(0,2))= ?$

Two PAM of mutations yield on average less than 2% mismatches

GAGTATGGTGCGGAGAGAAT

GACTACGGTGAAGAGAGAAC

CACTAGGGTGCAGAGCGACC

CACTAGGGTGCAGAGCGACC

In the example the transition matrix is normalized to 25% mutations instead of 1%

The same position can mutate several times

It can even remutate
A → C → A

We will on average have less than 2% mismatches

$P(X_i(0)=X_i(2)) > 0.98$

There is a conceptual difference between a mutation and a mismatch

GAGTATGGTGCGGAGAGAAT

GACTACGGTGAAGAGAGAAC

CACTAGGGTGCAGAGCGACC

CACTAGGGTGCAGAGCGACC

Mutations are historic events. They can not be observed directly

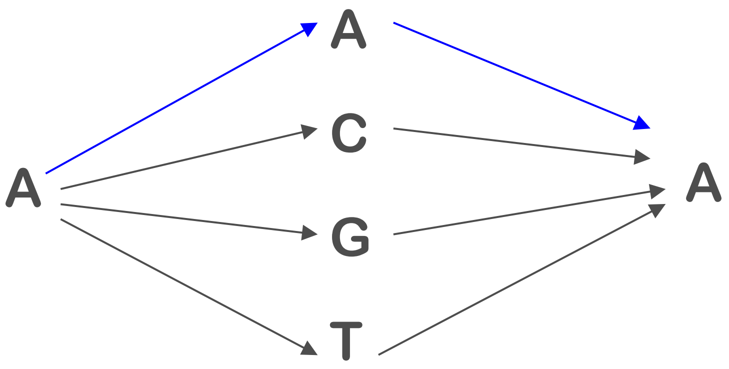
Mismatches are observations caused by mutations

2 PAM lead to an expected 2% mutations

But 2% mutations lead to somewhat less than 2 % mismatches

How do we calculate the expected number of mismatches after 2 PAM ?

Two step transitions



How do we get from A to A in 2 steps?

In the first step we need to go from A to somewhere

In the second step we need to go from somewhere to A

We have 4 possible paths:

A → A → A
A → C → A
A → G → A
A → T → A

A → A → A is much more likely than the other 3 paths but nevertheless, these also yield X(0)=A and X(2)=A

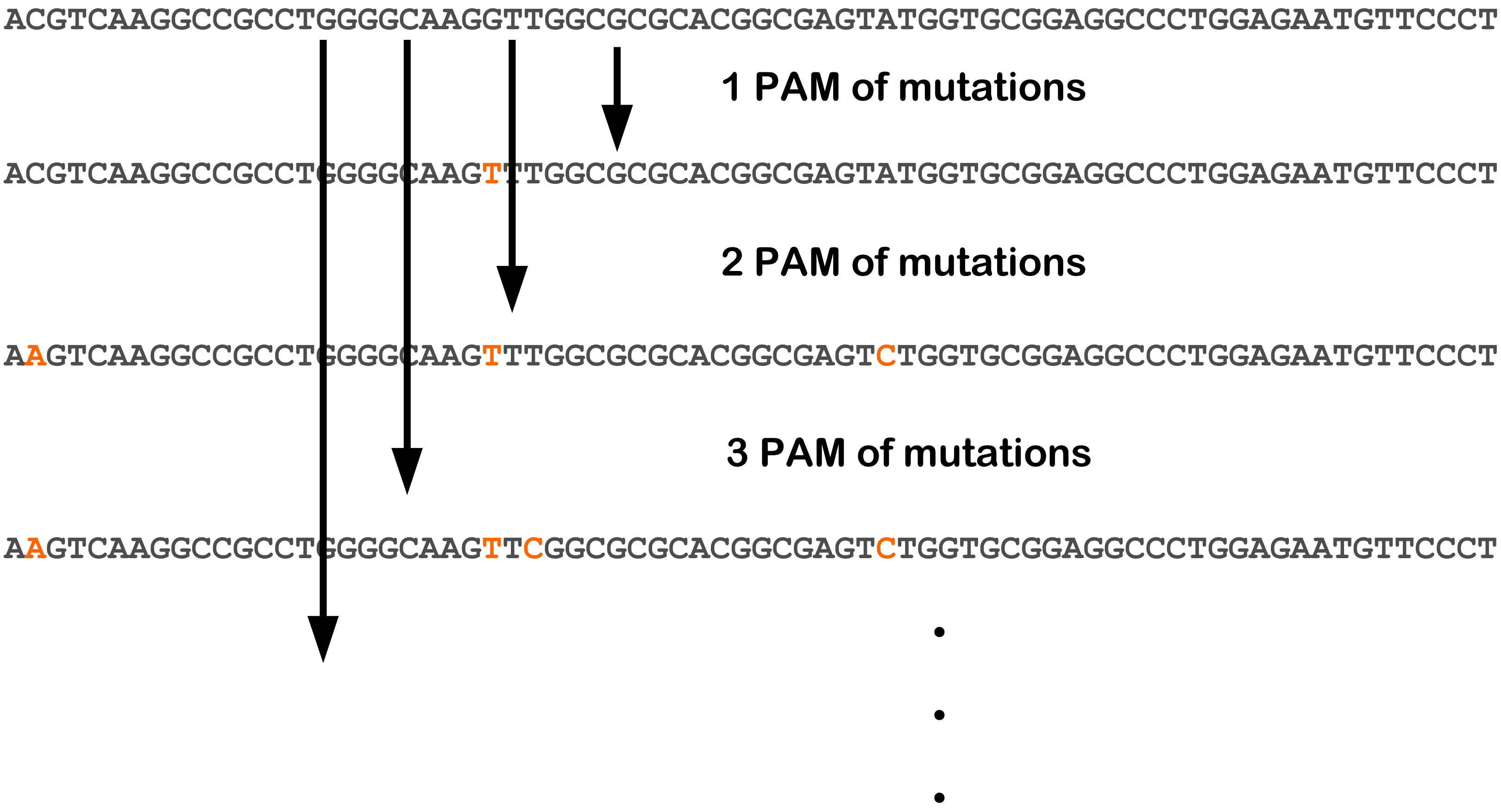
We can get a 2 step transition matrix by matrix multiplication

The sets {X(1)=y} for all possible values of y are a partition
Hence:

$$\begin{aligned}
P(X(2)=j|X(0)=i) &= \sum_y P(X(2)=j, X(1)=y|X(0)=i) \\
&= \sum_y P(X(2)=j|X(1)=y)\, P(X(1)=y|X(0)=i) \\
P_{ij}^{(2)} &= \sum_y P_{iy}\, P_{yj}
\end{aligned}$$

matrix multiplication

This random mutation process can be iterated



ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT

1 PAM of mutations

ACGTCAAGGCCGCCTGGGGCAAGTTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCCCT

2 PAM of mutations

AAGTCAAGGCCGCCTGGGGCAAGTTTGGCGCGCACGGCGAGTCTGGTGCGGAGGCCCTGGAGAATGTTCCCT

3 PAM of mutations

AAGTCAAGGCCGCCTGGGGCAAGTTCGGCGCGCACGGCGAGTCTGGTGCGGAGGCCCTGGAGAATGTTCCCT

.
.
.

The iterated application of a transition matrix is called a Markov chain

A Markov chain is a sequence of random experiments defined by

(a) a start point X(0)
(b) a transition matrix P

If X(t) = i, use column i of P as the distribution used to generate X(t+1)

Note that the distribution of X(t+1) only depends on the current state X(t) and not on the past (X(t-1), X(t-2), …)

A Markov chain has no memory

The powers of P give multistep transition matrices

$P^{(n)} = \underbrace{P \times P \times P \times \ldots \times P}_{\text{n-times}} = P^{n}$ gives us an n step transition matrix

For the evolutionary transition matrix the diagonal entries of $P^{(n)}$ give us the probability of a match after n iterations

$$P = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix} \qquad \begin{array}{c} a = 0.99 \\ b = 1/300 \end{array}$$

$P(X(0)=X(1)) = P_{1,1} = 0.99$

$P(X(0)=X(10)) = P^{(10)}_{1,1} = 0.904$

$P(X(0)=X(100)) = P^{(100)}_{1,1} = 0.446$

$P(X(0)=X(1000)) = P^{(1000)}_{1,1} = 0.25$

The Chapman Kolmogrov equation connects different multistep matrices with each other

$$P^{(n+m)} = P^{(n)}\, P^{(m)}$$

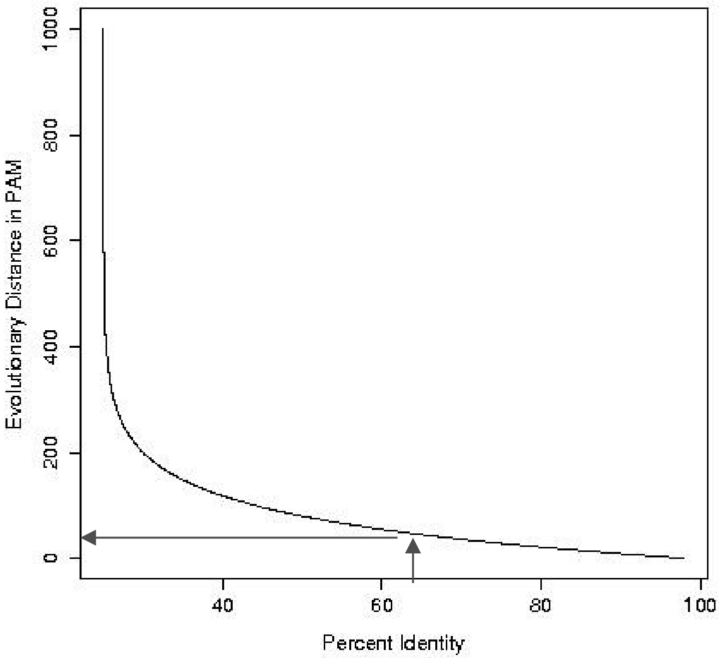
To get from i to j in n+m steps we need to go from i to somewhere in the first n steps and then from somewhere to j in the following m steps

How many PAM of evolution separate this human sequence segment from its rat homologue?

```
HUM ...ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC...
RAT ...ATGTAAGCCCCGGCTCTGCCCAGGTCAAGGCTCACGGCAAGAAGGTTGCTGATGCCCTGGCCAAAGCTGC...
       1011010001111110010101111000011011111101101010111011011111110011010101
```

lengths of segments:	70
match positions :	45
percent identity:	64.2

The sequence of transition matrices P(n) relates PAM mutation to percent sequence identity



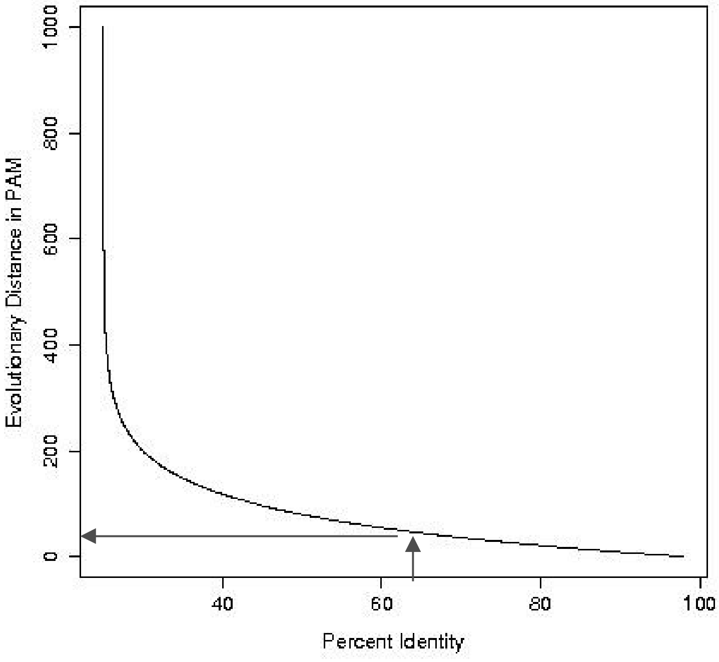
For every PAM distance we have a transition matrix

P: 1 PAM
$P^{(2)}$: 2 PAM
$P^{(3)}$: 3 PAM
…

The diagonal entries of each multistep matrix $P^{(n)}$ give the expected proportion of matches after n PAM of mutation

We expect 64,2% identity after about 47 PAM

How many years does evolution need for a PAM of mutation?



We have estimated the evolutionary distance of the sequences to 47 PAM

How many million years does this correspond to?

Unfortunately, this is more difficult to answer, since evolution is not proceeding with a constant speed at all times and at all genome positions

More theory is needed

→ Molecular Clock

The 47 PAM model is not the only model that can produce 45 matches

With 47 PAM random mutations we expect 64% identity and this is what we have observed for the two sequences of man and rat

50 PAM in contrast lead to an expectation of 63% instead of 64% sequence identity

Is the observation of 64% in contradiction to the assumption that the sequences could be 50 PAM apart?

63% percent is only the expected value. The real outcome of the mutation process could be a little bit less or a little bit more ...

… for example 64%

We need to proceed from the expected number of match positions to the distribution of this number

$S^{(47)}$ = number of match positions after 47 PAM

$S^{(47)}$ is a random variable. What is its distribution?

After 47 PAM of mutations the probability that a sequence position is a match position is

$p = P(X(0)=X(47))$

$= 0.64$

The sequence segments have n=70 positions

What is the distribution of the number of matches?

The number of match positions follows a binomial distribution

```
HUM ...ACGTCAAGGCCGCCTGGGGCAAGGTTGGCGCGCACGGCGAGTATGGTGCGGAGGCCCTGGAGAATGTTCC...
RAT ...ATGTAAGCCCCGGCTCTGCCCAGGTCAAGGCTCACGGCAAGAAGGTTGCTGATGCCCTGGCCAAAGCTGC...
       1011010001111110010101111000011011111101101010111011011111110011010101
```

Observation: 45 matches over 70 positions

$S^{(47)}$ is the sum of 70 Bernoulli experiments with parameter p = 0.64

Hence $S^{(47)}$ ~ Bin (0.64,70)

$$P(S^{(47)} = k) = \binom{n}{k}\, p^k\, (1-p)^{n-k}$$

We can calculate the likelihood of the 47 PAM model given the observation of 45 match positions

$$P(S^{(47)} = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

n=70 the lengths of the segments
k=45 the observed match positions
p=0.64 the probability of a match after 47 PAM

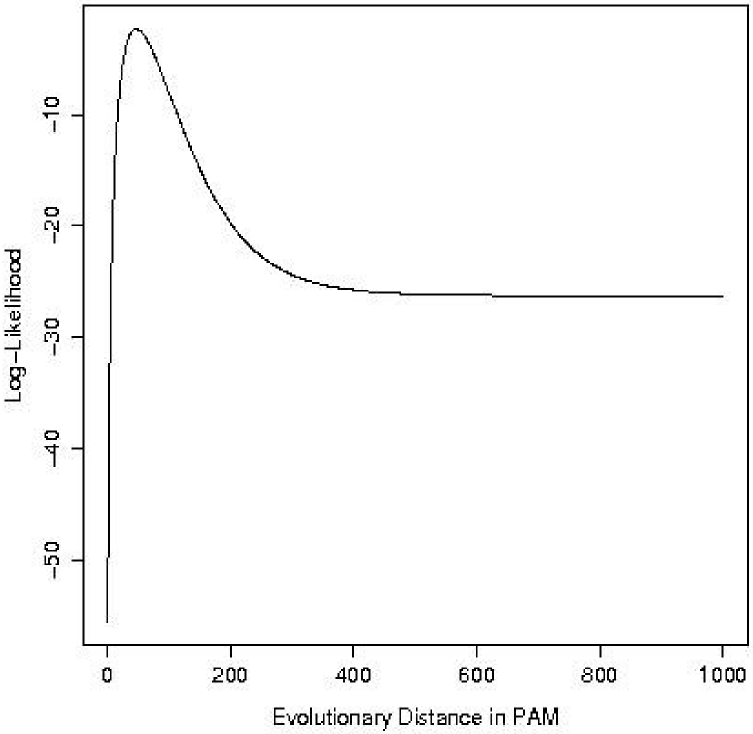
$$\binom{70}{45} 0.64^{45} \times 0.36^{25} \approx 0.1$$

How does this compare to the probability of 45 matches after 50 PAM or 100 PAM ?

Likelihoods for many possible PAM distances

1 PAM Model:	$\binom{70}{45}\ 0.99^{45} \times 0.01^{25} \approx 4.3e-32$
10 PAM Model:	$\binom{70}{45}\ 0.90^{45} \times 0.1^{25} \approx 5.6e-9$
47 PAM Model:	$\binom{70}{45}\ 0.64^{45} \times 0.36^{25} \approx 0.098$
50 PAM Model:	$\binom{70}{45}\ 0.63^{45} \times 0.37^{25} \approx 0.096$
100 PAM Model:	$\binom{70}{45}\ 0.44^{45} \times 0.66^{25} \approx 0.0003$
1000 PAM Model:	$\binom{70}{45}\ 0.25^{45} \times 0.75^{25} \approx 3.9e-12$

We can draw the log-likelihood as a function of the PAM distance



This plot is also called a log-likelihood landscape

The likelihood peaks at 47 PAM

47 PAM is the Maximum Likelihood Estimate (MLE)

Small PAM distances can not explain the small number of 45 matches at all

How precise is the estimate 47 PAM?

We can compare pairs of models using likelihood ratios

Likelihood of the 47 PAM Model: $\binom{70}{45}\, 0.64^{45} \times 0.36^{25} \approx 0.098$

Likelihood of the 100 PAM Model: $\binom{70}{45}\, 0.44^{45} \times 0.66^{25} \approx 0.0003$

Likelihood ratio: 326.7

Number of white balls: 8.35

There is hardly any evidence that any model between 40-60 PAM explains the data better than another model from this interval

PAM(47) vs. PAM(50)	**LR=1.05**	**(0.06 white balls)**
PAM(47) vs. PAM(60)	**LR=1.77**	**(0.83 white balls)**
PAM(47) vs. PAM(40)	**LR=1.29**	**(0.37 white balls)**
PAM(47) vs. PAM(70)	**LR=4.72**	**(2.24 white balls)**
PAM(47) vs. PAM(80)	**LR=16.3**	**(4.03 white balls)**
PAM(47) vs. PAM(100)	**LR=326**	**(8.35 white balls)**
PAM(47) vs. PAM(1000)		**(34 white balls)**
PAM(47) vs. PAM(1)		**(76.8 white balls)**

The sequences are between 40 and 60 PAM apart
… more should not be said

End of Chapter 11