

A photograph of two cheetahs in a savanna landscape. One cheetah is in the foreground, looking towards the camera, while another is in the background, looking to the left. The background is a vast, open plain with dry, yellowish-brown grass under a clear sky.

# The impact of Analysis of Algorithms on Bioinformatics

Gaston H. Gonnet

Informatik, ETH, Zurich

Analysis of Algorithms, Maresias, April 16, 2008

# Abstract

In principle, Analysis of Algorithms and Bioinformatics share few tools and methods. This is only true when we look at the surface, deeper inspection shows many points of convergence, in particular in asymptotic analysis and model development. We would also like to stress the importance and usefulness of Maximum Likelihood for modelling in bioinformatics and the relation to problems in Analysis of Algorithms.

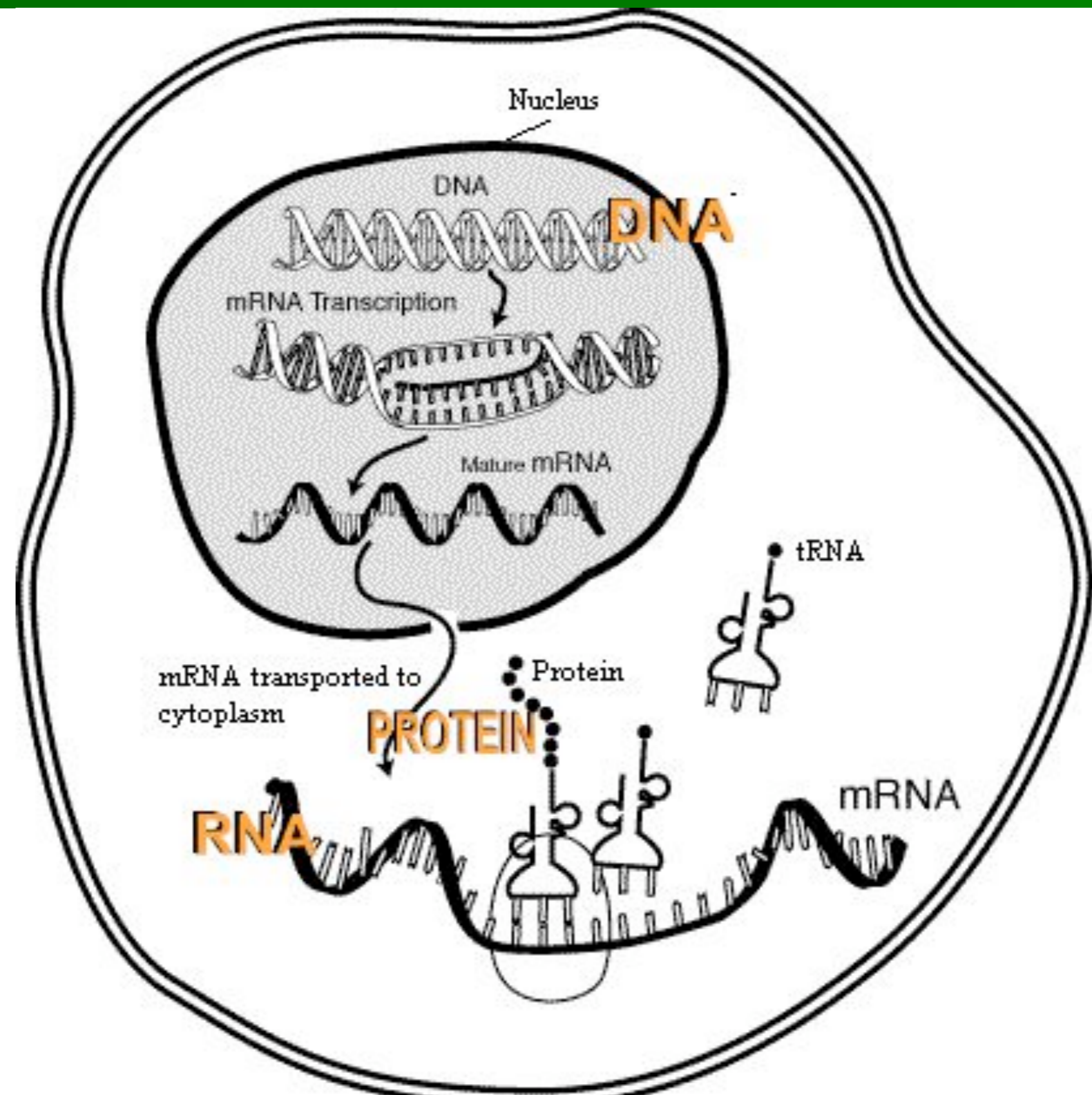
# Bertioga to Sao Sebastiao in 1971



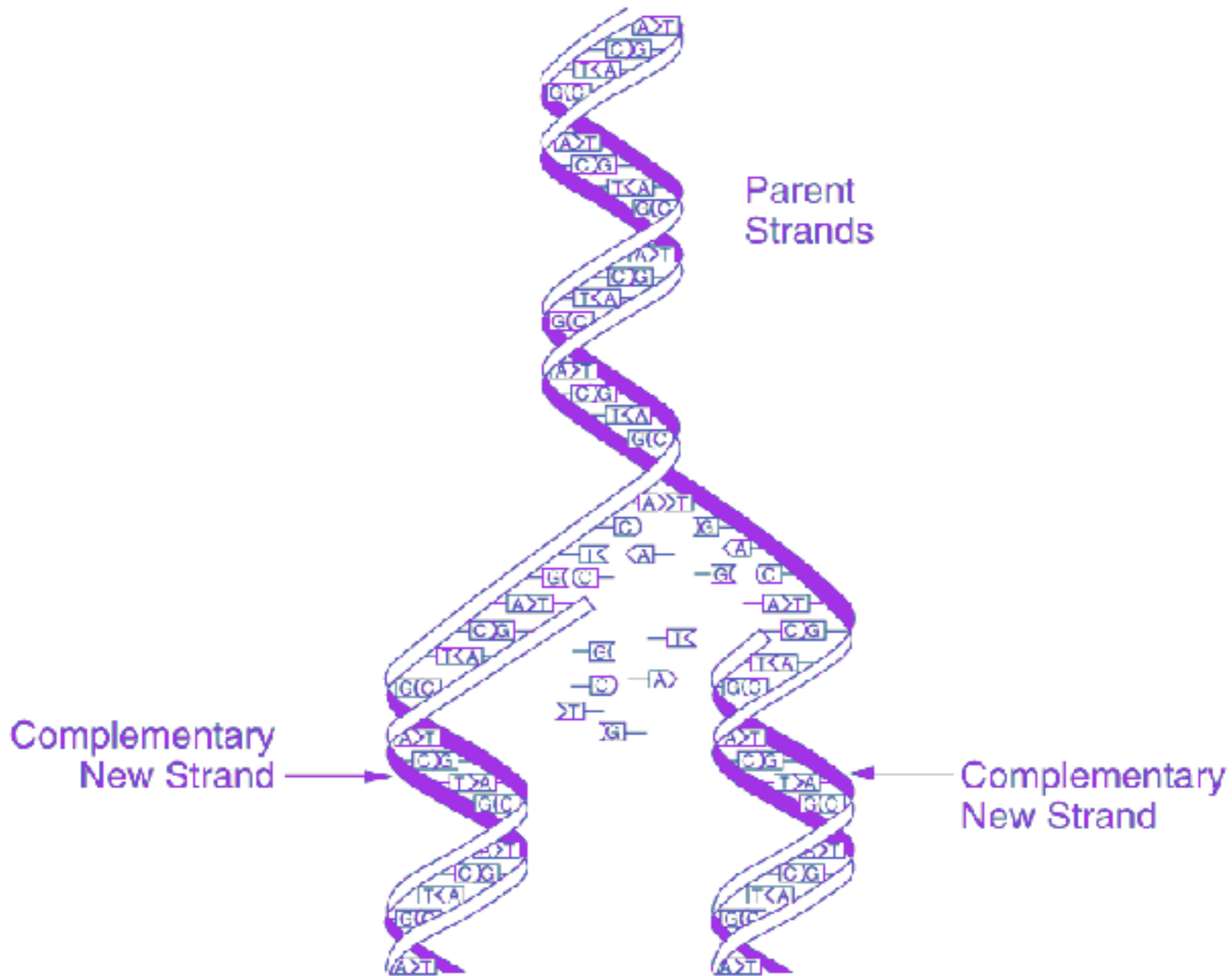
# Bertioga to Sao Sebastiao in 1971



# Central Dogma of Molecular Evolution



# Double stranded DNA is reproduced



# Modelling

Analysis of Algorithms gives us a natural ability to model and analyze processes.

What makes a good model?

Must capture the essence of the process

As simple as possible, but not simpler

Realistic in terms of the application

Analyzable

# Modelling: Closed form vs computational

Simple models may allow closed form solutions, more realistic (complicated) models may only allow numerical solutions

A closed form solution gives you insight!

Numerical computation gives you results which can be used.



# Mistakes happen during DNA replication

Most mistakes are harmful, give the organism a disadvantage and it does not survive/compete.

Some mistakes are irrelevant, i.e. do not cause any difference. Rarely they remain in the population.

Some mistakes are helpful they either improve the organism or adapt it better to the environment. These are very likely to survive in the population.

# Mistakes modeled as a Markovian process

The occurrence and complicated acceptance of DNA mutations is modeled as a Markov process  
This is known to be flawed, but still is the best model for DNA/protein evolution

$M =$

	A	C	G	T
A	0.93	0.01	0.07	0.01
C	0.02	0.95	0.02	0.02
G	0.03	0.01	0.88	0.01
T	0.02	0.03	0.03	0.96

# Mutation matrices

$$Mp_0 = p_1$$

$M$  defines a unit of mutation

$$M^\infty p = f$$

Infinite mutation results in the natural (default) frequencies

$$Mf = f$$

$f$  is the eigenvector with eigenvalue 1 of  $M$

# Mutation matrices (II)

$$M^d = e^{dQ}$$

$Q$  is the rate (differential equations of transitions) matrix

$$Q = U \Lambda U^{-1}$$

Eigenvalue/eigenvector decomposition of  $Q$

$$M^d = U e^{d\Lambda} U^{-1}$$

$$\lambda_1 = 0, \quad U_1 = f$$

from  $Mf=f$

$$\lambda_i < 0, \quad i > 1$$

reaches steady state

# The principle of Molecular evolution



**Dog DNA**

aactgagcggtt...



**Elephant DNA**

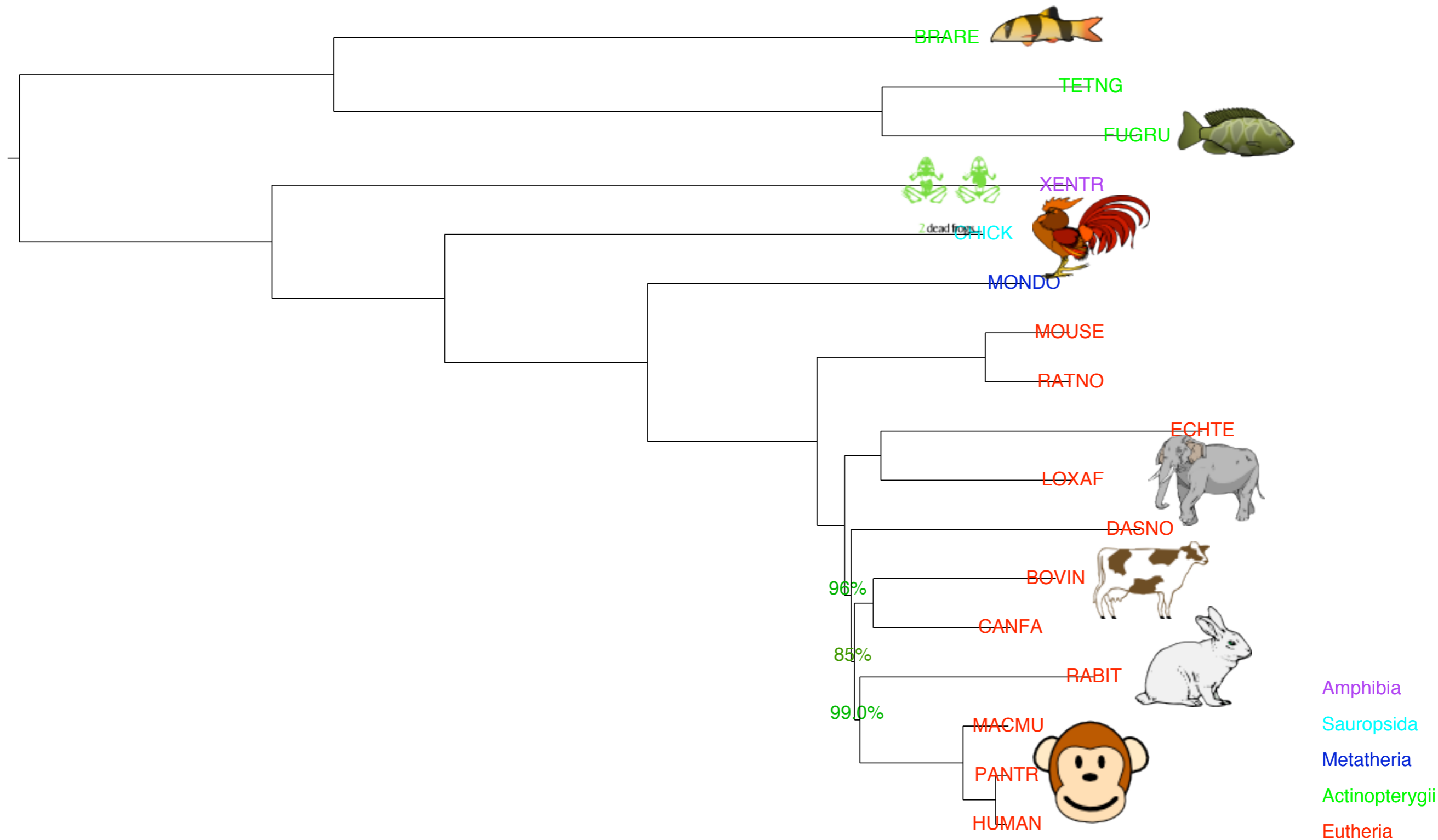
aactgaccggtt...



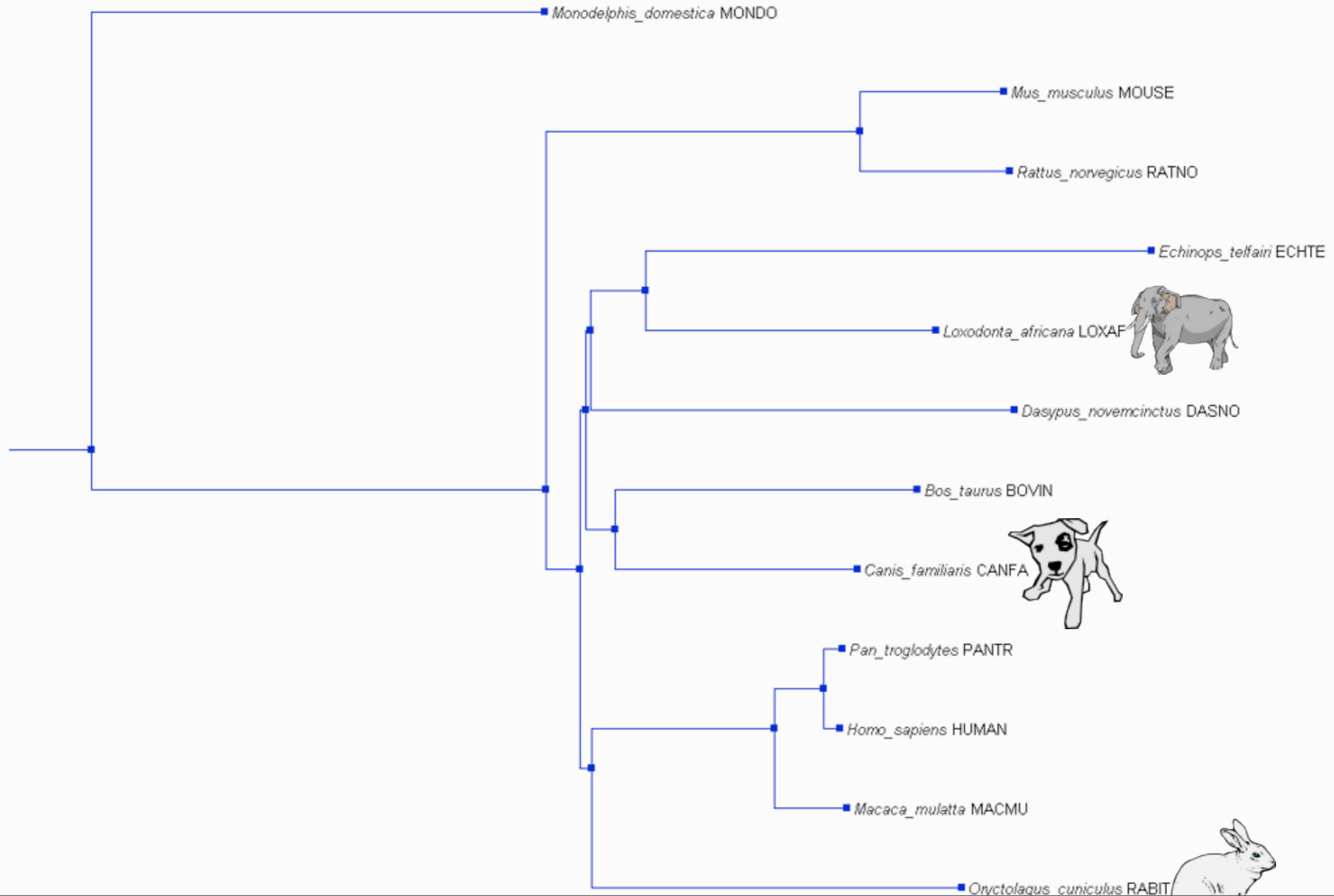
**Rabbit DNA**

aactgaccggtt...

# Phylogenetic tree of 17 vertebrates

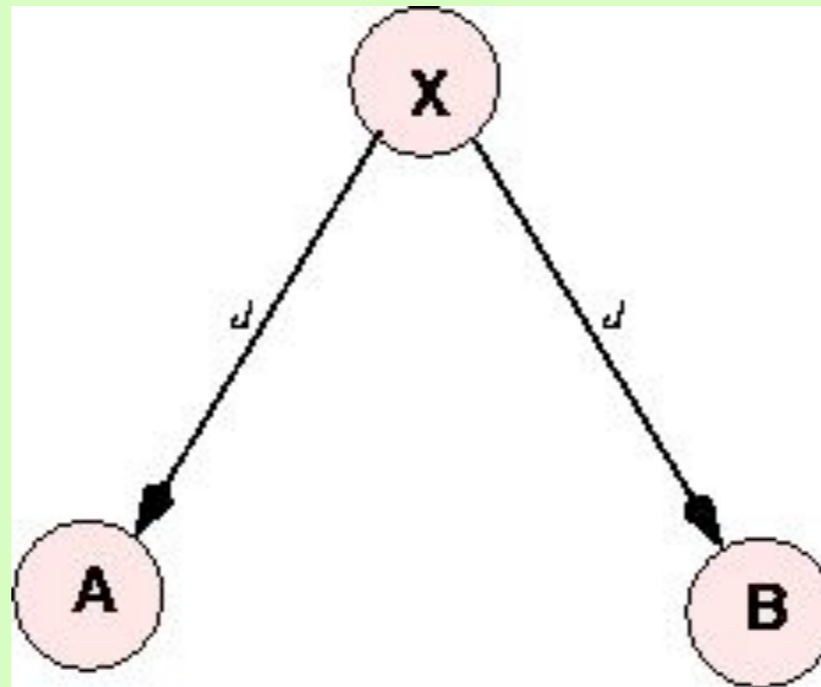


# Tree of mammals



# Probabilities vs likelihoods

Some event



Over all  $X$ ,  $A$   
and  $B$  defines  
a probability  
space

For particular  
data, as a  
function of  $d$ ,  
defines a  
likelihood



# Maximum likelihood (I)

How to estimate parameters by Maximum likelihood?

Compute the likelihood, or log of the likelihood, and maximize

$$L(\theta) = \text{Prob}\{\text{event depending on } \theta\}$$

$$L(\theta) = \prod_i \text{Prob}\{i^{\text{th}} \text{ event depending on } \theta\}$$

$$\ln(L(\theta)) = \sum_i \ln(\text{Prob}\{i^{\text{th}} \text{ event depending on } \theta\})$$

# Maximum likelihood (II)

$$\max(L(\theta)) = L(\hat{\theta})$$

$$\frac{L'(\hat{\theta})}{L(\hat{\theta})} = 0$$

$$\frac{L''(\hat{\theta})}{L(\hat{\theta})} = -\frac{1}{\sigma^2(\hat{\theta})}$$

Also applicable to vectors with the usual matrix interpretations

# Maximum likelihood (III)

Completely analogous to the asymptotic estimation of integrals based on the approximation of the maximum by

$$a_0 e^{-a_1 x^2}$$

# Maximum likelihood (IV)

The maximum likelihood estimators are:

Unbiased

Most efficient (of the unbiased estimators, the ones with smallest variance)

Normally distributed

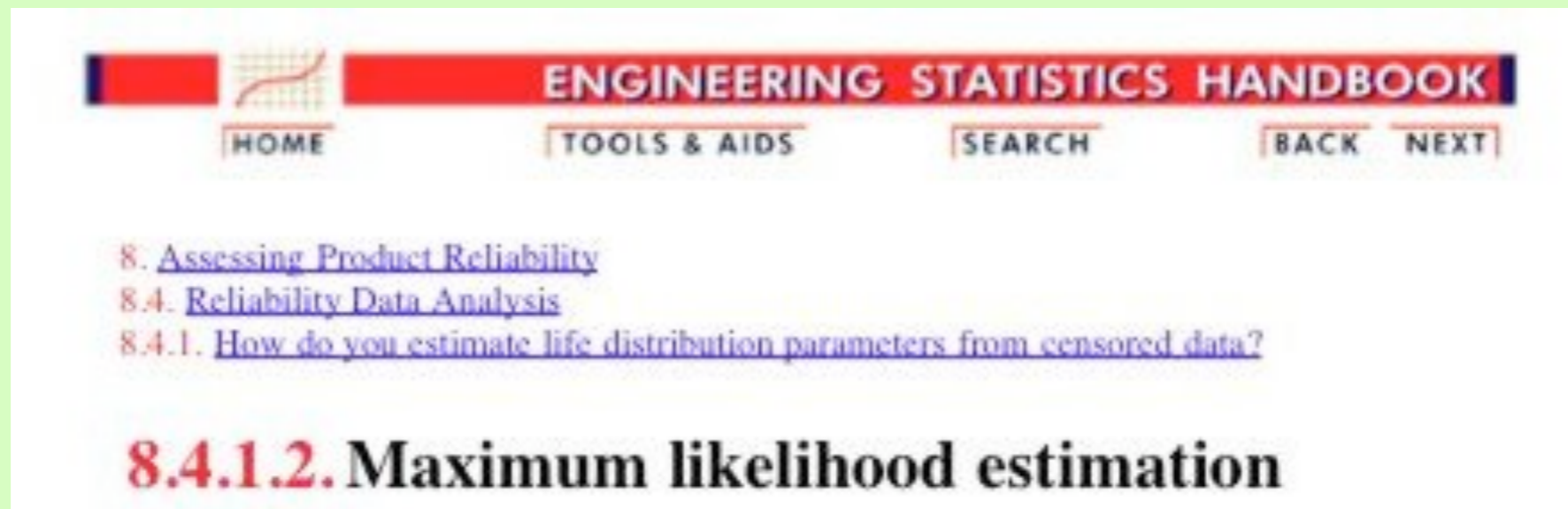
# Maximum likelihood (V)

This is ideal for symbolic/numeric computation

Complicated problems/models can be stated in their most natural form

The literature usually warns against the difficulty of computing derivatives and solving non-linear equations (maximum) ????

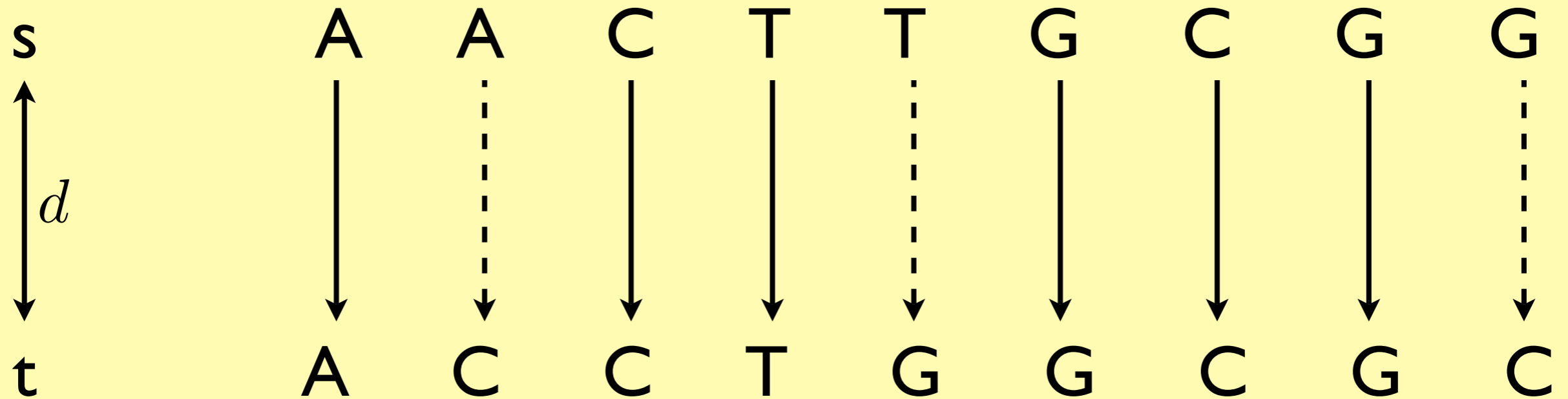
# Some people have not discovered symbolic computation yet...



There are only two drawbacks to MLE's, but they are important ones:

- With small numbers of failures (less than 5, and sometimes less than 10 is small), MLE's can be heavily biased and the large sample optimality properties do not apply
- Calculating MLE's often requires specialized software for solving complex non-linear equations. This is less of a problem as time goes by, as more statistical packages are upgrading to contain MLE analysis capability every year.

# Inter sequence distance estimation by ML



$$L(d) = \prod_i (M^d)_{s_i, t_i}$$

$$\ln(L(d)) = \sum_i \ln((M^d)_{s_i, t_i})$$

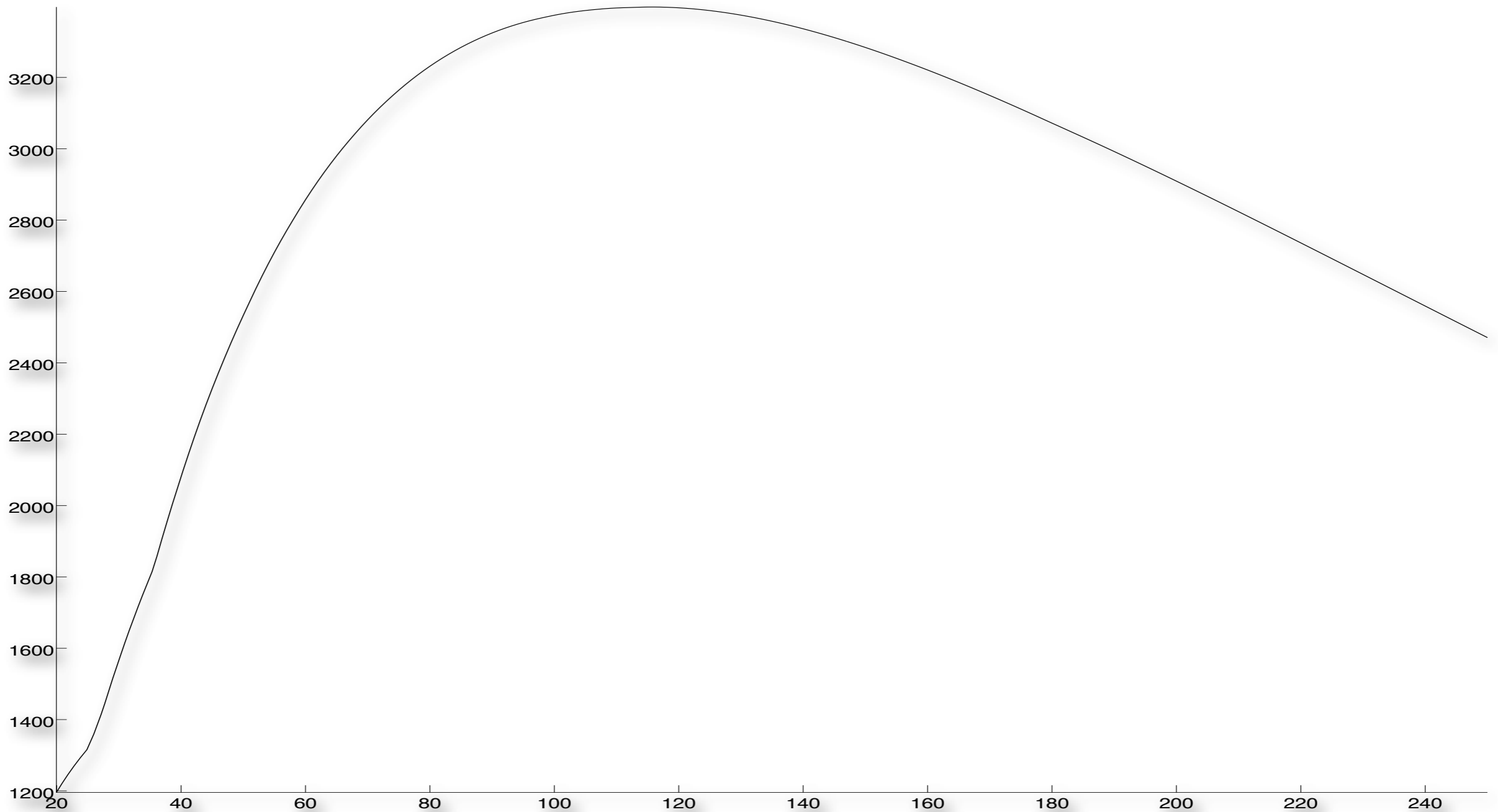
# Inter sequence distance estimation by ML

$$\ln(L(d)) = \sum_i \ln((M^d)_{s_i, t_i})$$

This is normally called the score of an alignment and it is used (with some normalization) by the dynamic programming algorithm for sequence alignment



# Inter sequence distance estimation by ML



Score (likelihood) vs PAM distance for a particular protein alignment

# Estimation of deletion costs by ML

The Zipfian model of indels postulates that indels have a probability given by:

$$\Pr\{\text{indel of length } k\} = c_0(d) \frac{1}{\zeta(\theta) k^\theta}$$

where the first term is the probability of opening an indel and the second gives the distribution of indels according to length

# Estimation of deletion costs by ML (II)

Empirically:

$$\ln(c_0(d)) = d_0 + d_1 \ln(d)$$

which means that the score of an indel is modeled by the formula:

$$\ln(\text{indel length } k) = d_0 + d_1 \ln(d) - \theta \ln k$$

a model with 3 unknown parameters

# Estimation of deletion costs by ML (III)

Collecting information from gaps in real alignments (thousands of them) we can fit these parameters by maximum likelihood



# Evolution happens at very different speeds

Some proteins in bacteria are 80% identical to those in humans (3,000,000,000 years)

Most proteins in mammals are about 80% identical (200,000,000 years)

Humans and chimpanzees are about 99% identical at the protein level (5,000,000 years)

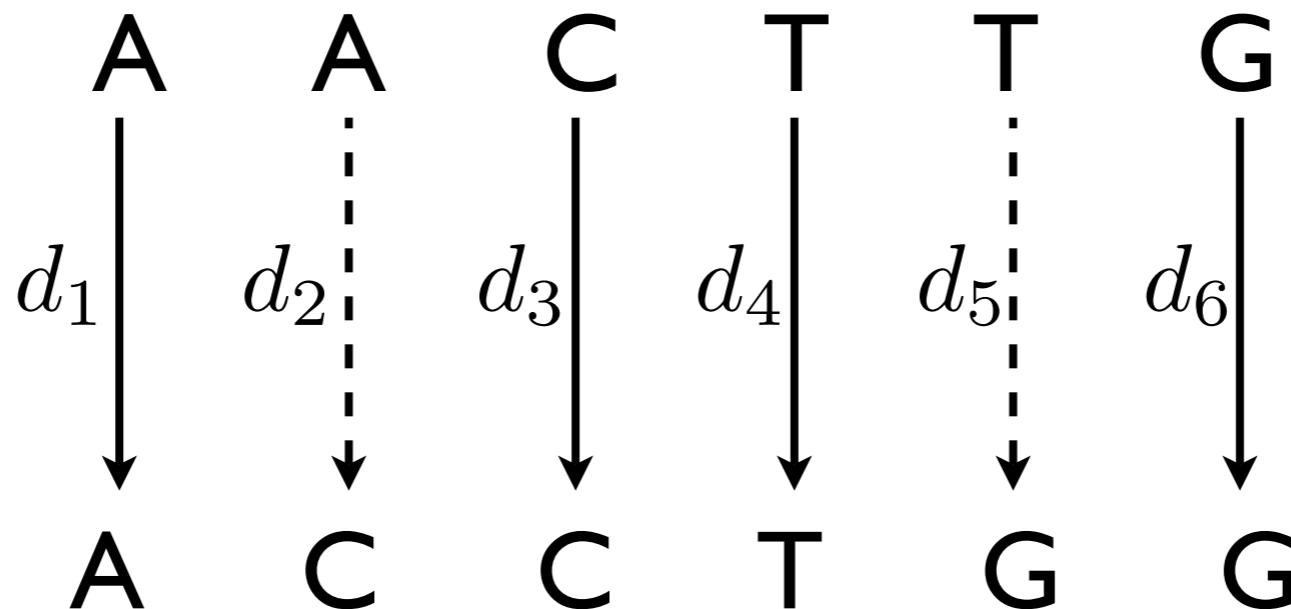
Mitochondrial DNA is more than 99% identical in all humans (200,000 years)

The HIV virus has mutated about 10% in the last 20 years

# Variable evolution rates (I)

It is recognized as biologically appropriate that different positions evolve at different rates  
Modeled with a gamma distribution (no good reason, but not a terrible idea either)

$$d_i \in \Gamma(k, \theta)$$



# Variable evolution rates (II)

The Gamma distribution is only defined over positive values and has two parameters

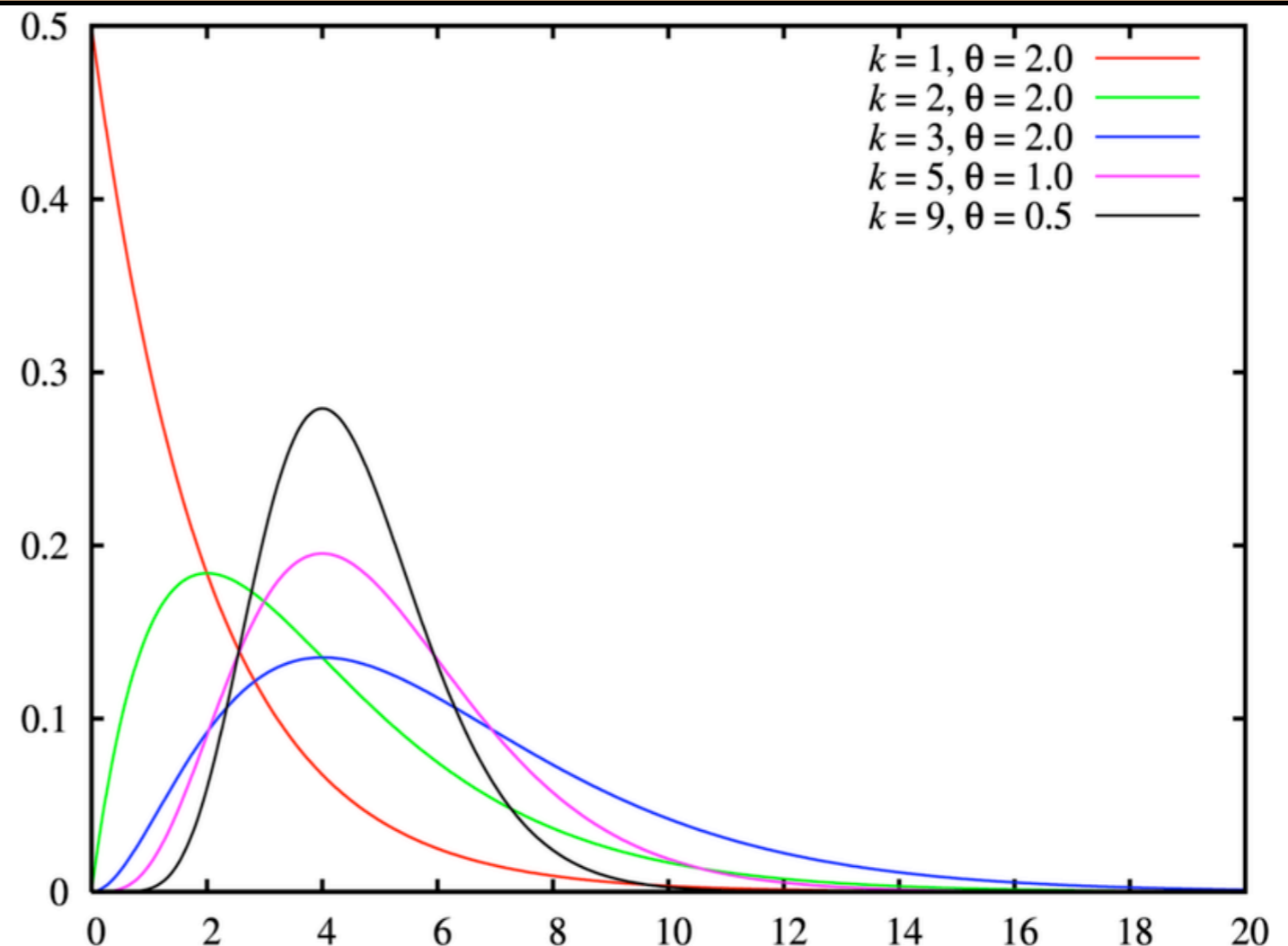
It can be shaped from an exponential distribution to an almost normal one

$$p(x) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k}$$

$$E[x] = k\theta$$

$$\sigma^2(x) = k\theta^2$$

$$\begin{aligned} mgf(t) &= \int_0^{\infty} e^{tx} p(x) dx \\ &= (1 - \theta t)^{-k} \end{aligned}$$





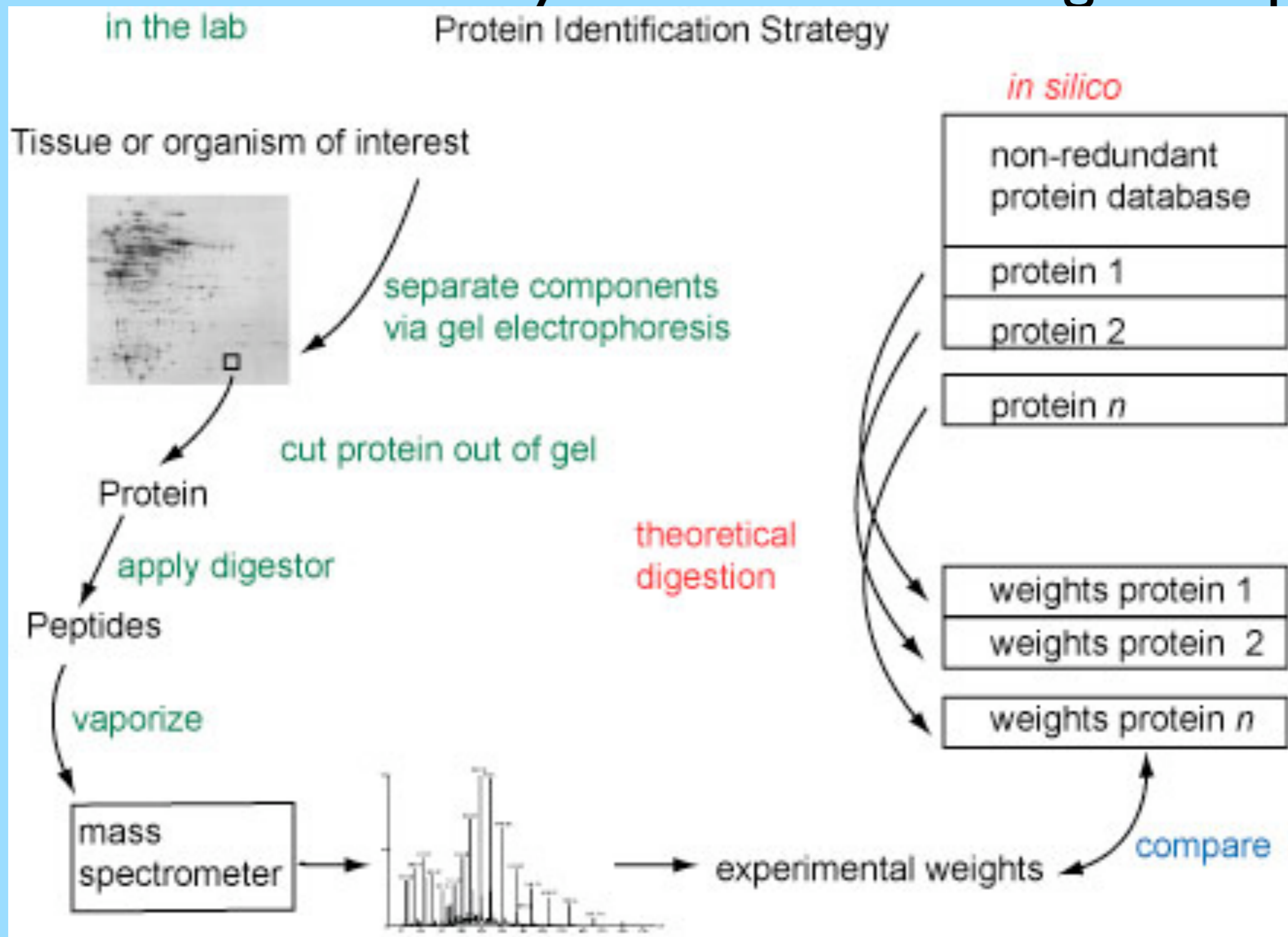
# Variable evolution rates (III)

The expected transitions rates are the result of the combined event of selecting a distance (with gamma distribution) and an evolution transition

$$\begin{aligned} E[M^x] &= \int_0^\infty p(x) M^x dx \\ &= U \left( \int_0^\infty p(x) e^{\Lambda x} dx \right) U^{-1} \\ &= U \int_0^\infty m g f(\Lambda) U^{-1} \\ &= U (I - \theta \Lambda)^{-k} U^{-1} \end{aligned}$$

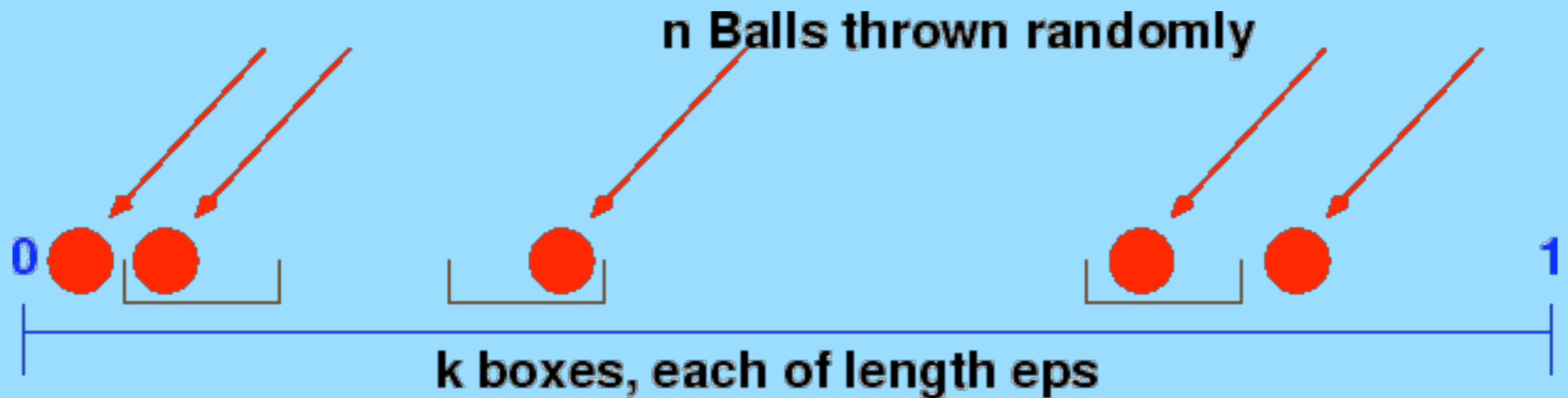
# Molecular weight fingerprinting

## Protein identification by the mass of its digested parts



# The model for comparison

How to model the approximate match of  $k$  of the weights



$\Pr\{k,n,eps\}$  = Probability of  $k$  boxes with at least one ball each

# The model for comparison - generating function

All the distribution events are captured in the generating function:

$$G_{k,n,\epsilon} = (a_1\epsilon + a_2\epsilon + \dots + a_k\epsilon + b(1 - k\epsilon))^n$$

$a_i$  corresponds to a ball falling in box  $i$  and  $b$  corresponds to a ball falling outside all boxes

We want to find the coefficient of all terms having all the  $a_i$  to some positive power

# The model for comparison - generating function

For example, for  $k=2$

$$G_{2,n,\epsilon} = (a_1\epsilon + a_2\epsilon + b(1 - 2\epsilon))^n$$

$$\begin{aligned} G_{2,n,\epsilon}^* &= G_{2,n,\epsilon} - G_{2,n,\epsilon|a_1=0} \\ &= (a_1\epsilon + a_2\epsilon + b(1 - 2\epsilon))^n - (a_2\epsilon + b(1 - 2\epsilon))^n \end{aligned}$$

$$G_{2,n,\epsilon}^{**} = G_{2,n,\epsilon}^* - G_{2,n,\epsilon|a_2=0}^*$$

$$P_{2,n,\epsilon} = 1 - 2(1 - \epsilon)^n + (1 - 2\epsilon)^n$$

# The model for comparison - generating function

$$P_{k,n,\epsilon} = \sum_{i=0}^k (-1)^i \binom{k}{i} (1 - i\epsilon)^n$$

$$P_{k,n,\epsilon} = (1 - e^{-n\epsilon})^k \left( 1 + \frac{k n \epsilon (e^{n\epsilon} - k)}{2(e^{n\epsilon} - 1)^2} \epsilon + O(\epsilon^2) \right)$$

# Bioinformatics is fun



# How diverse are humans?

A SNP (pronounced snip) is a position in our DNA where at least 1% of the population shows a difference. (Single Nucleotide Polymorphism)

SNPs are responsible for most of the human diversity. They are also the cause all of the genetic diseases.

There are about 300,000 SNPs in the human population.

Easy-to-find SNPs are called markers. Some easy to test markers are the main tool for genetic fingerprinting (e.g. paternity tests)



# Paternity testing

Locus	Mom	Child	Dad
D8S1179	13/14	14/16	13/16
TH01	7/9	8/9	7/8
CSF1PO	10/11	7/10	7
AR21UY	7/11	11/17	18/21
...			

# DHL founder, Larry Hillblom's case

Died in a seaplane crash.

No family, University of California to receive his estate

Various women in several different countries made claims that he was the father of their children.

4 Children were proved to come from the same father and they received their rightly share.

the END