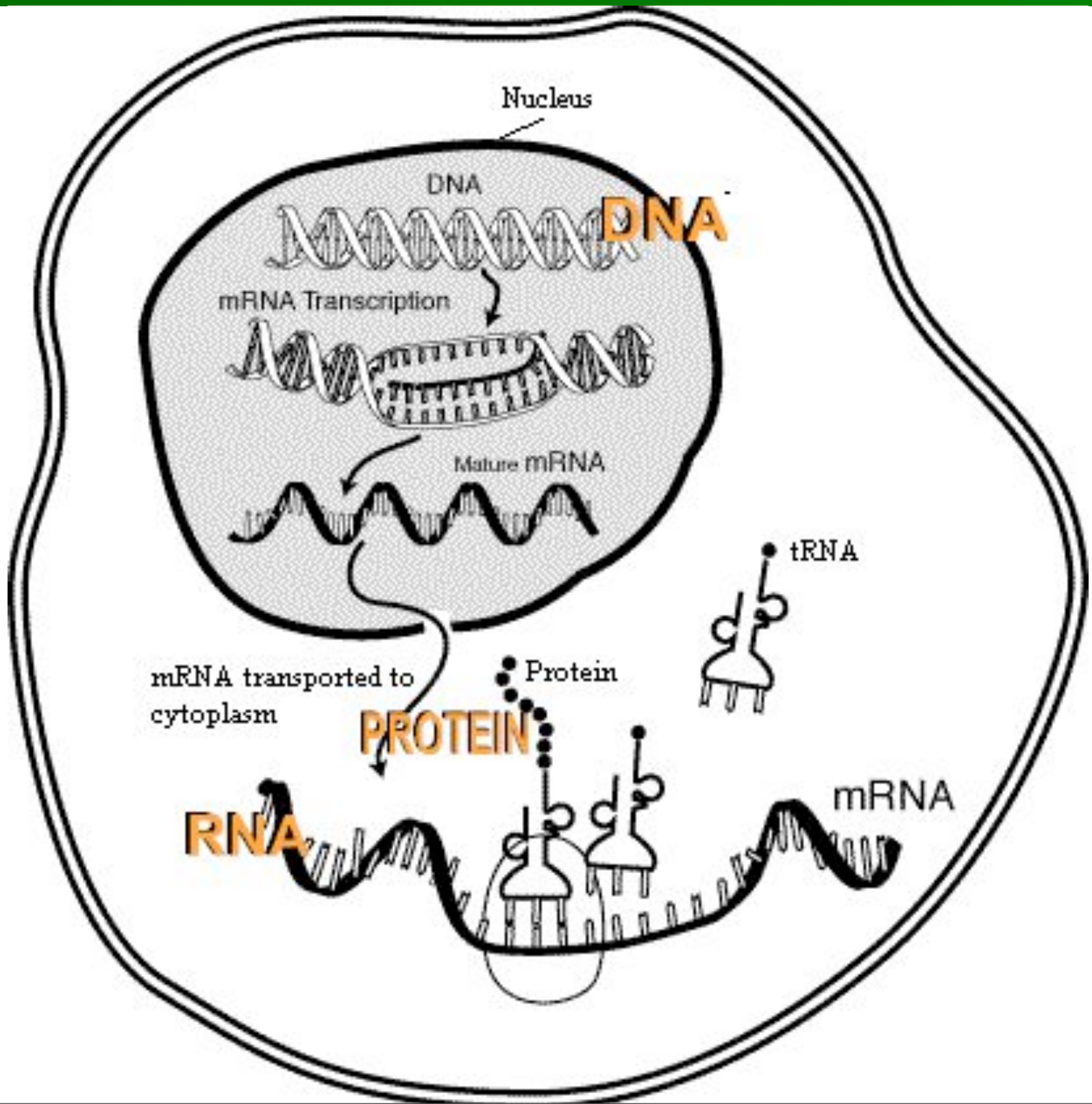
Computational Mathematics serving/impacting Bioinformatics

Gaston H. Gonnet

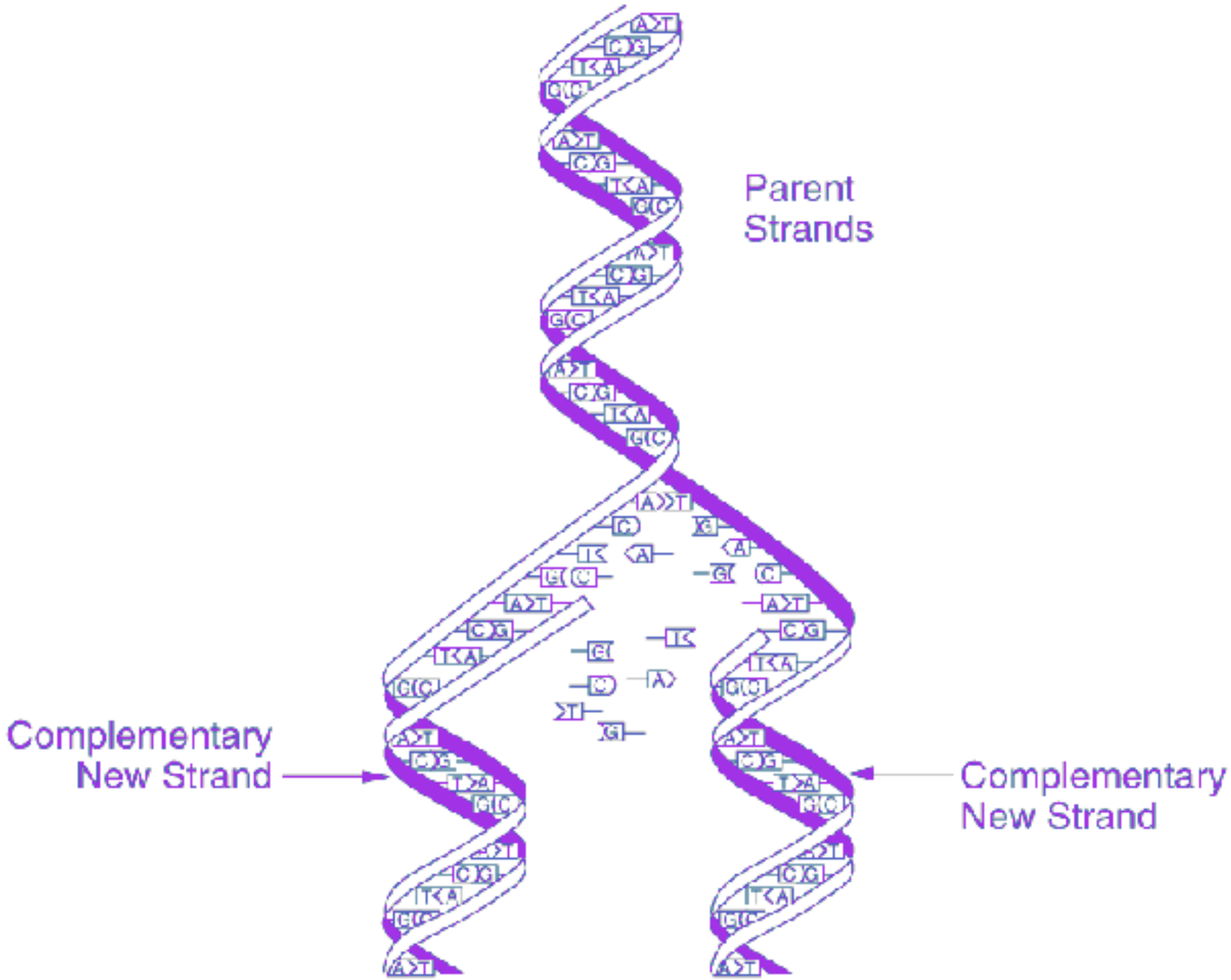
Informatik, ETH, Zurich

FoCM, Hong Kong, June 20th, 2008

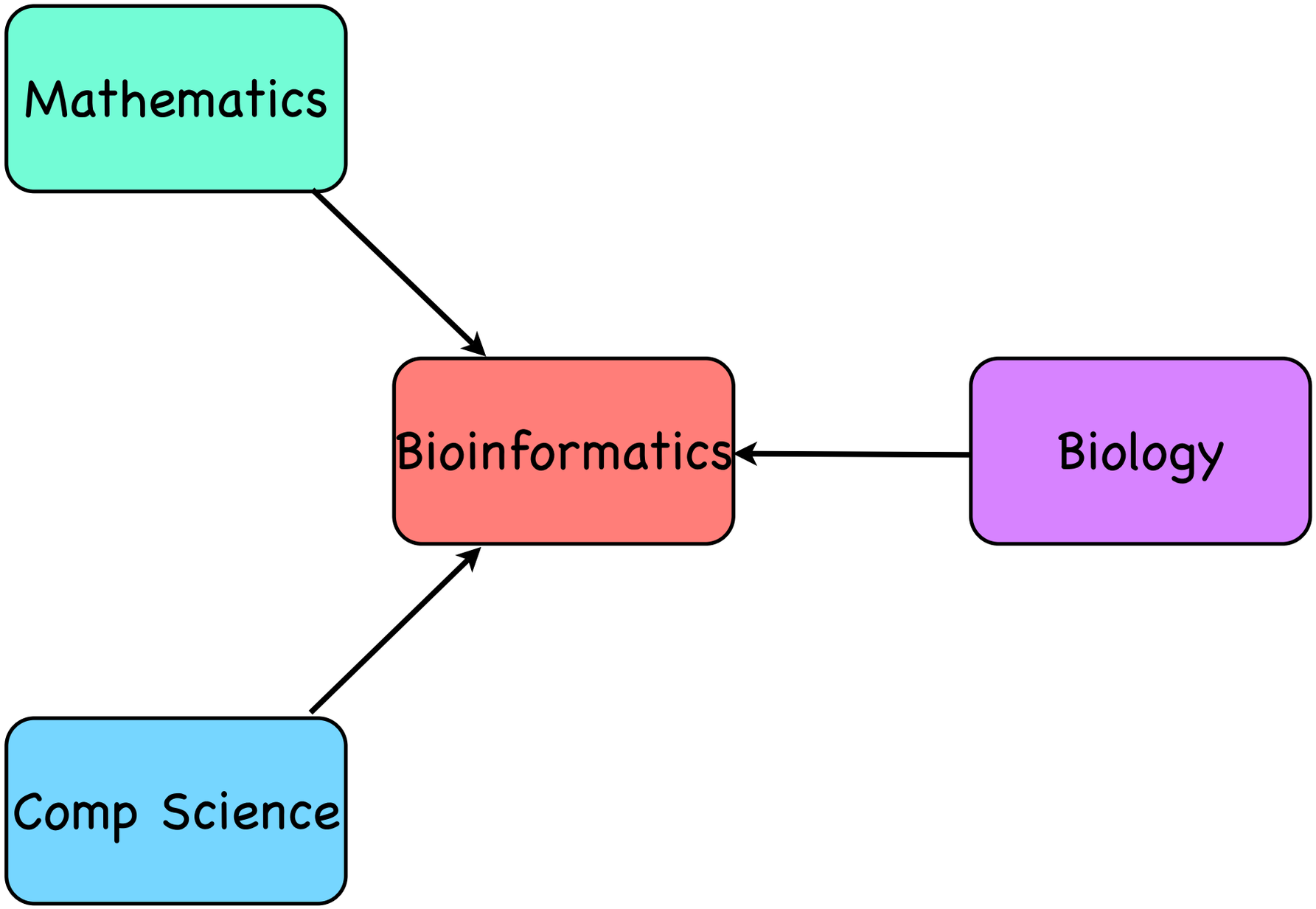
Central Dogma of Molecular Evolution



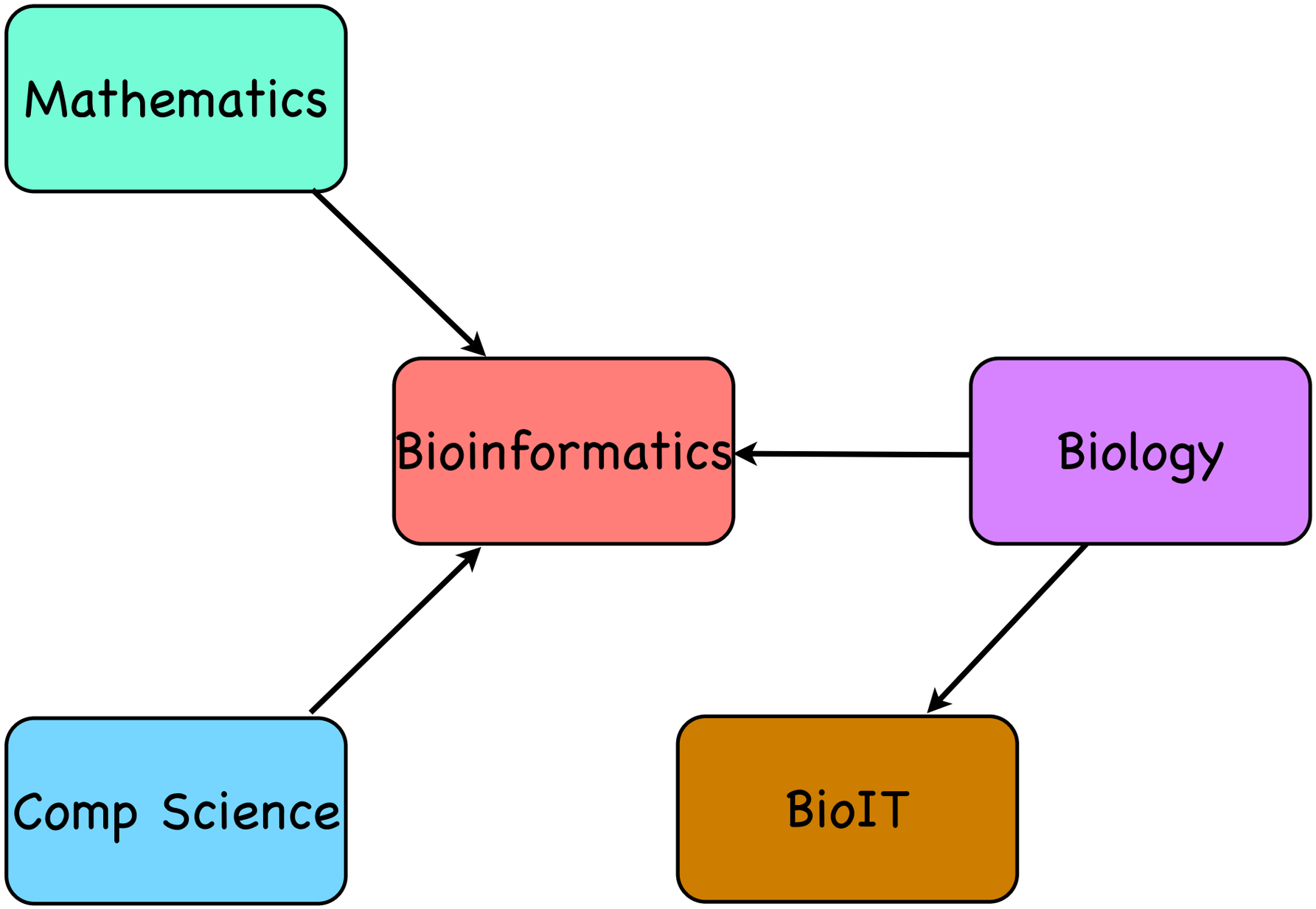
Double stranded DNA is reproduced



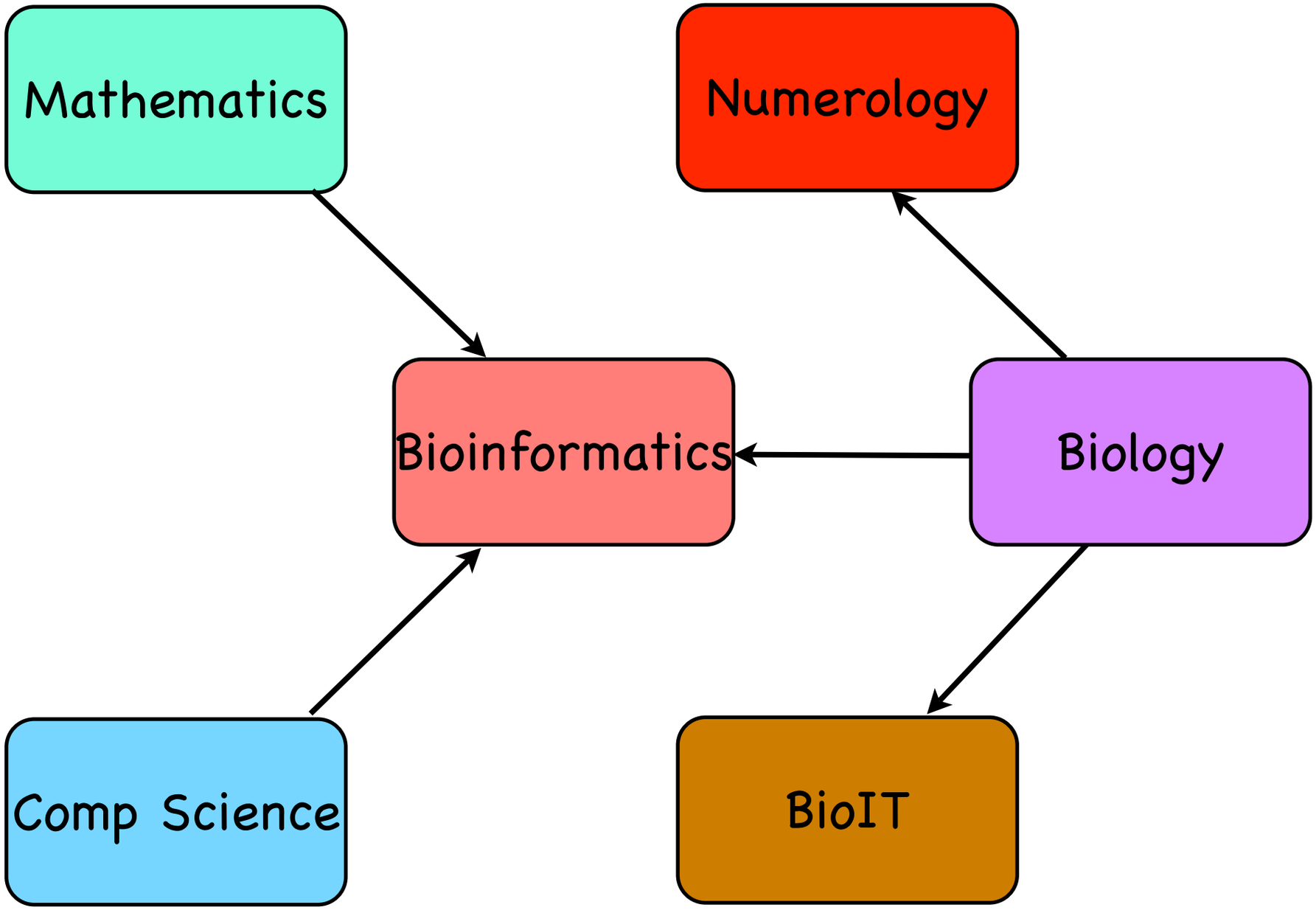
What is Bioinformatics?



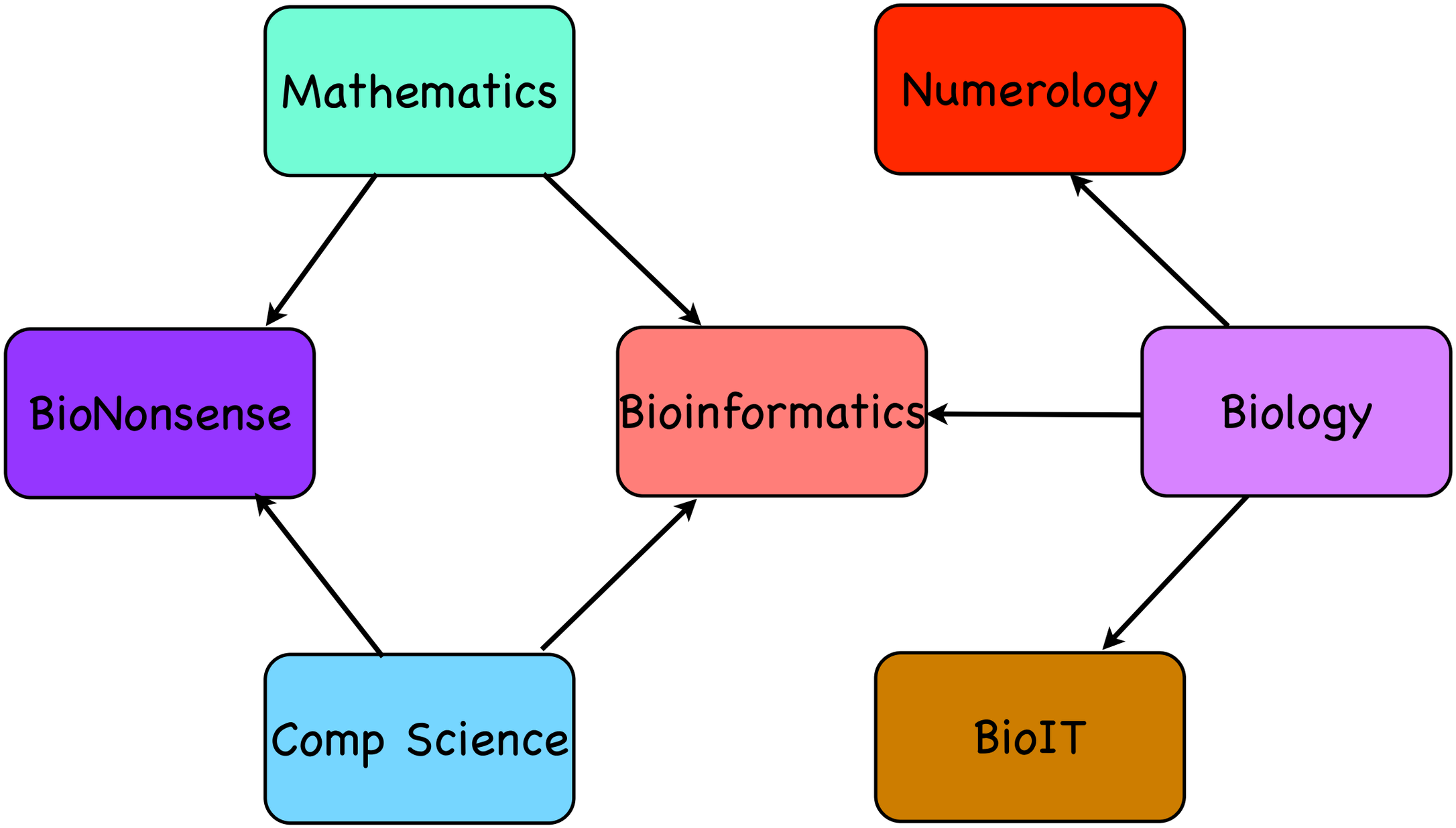
What is Bioinformatics?



What is Bioinformatics?



What is Bioinformatics?



Modelling

Modelling

Modelling is an essential part of Bioinformatics. Almost every computation/prediction/estimation is based on a model of nature.

Modelling

Modelling is an essential part of Bioinformatics. Almost every computation/prediction/estimation is based on a model of nature.

What makes a good model?

Modelling

Modelling is an essential part of Bioinformatics. Almost every computation/prediction/estimation is based on a model of nature.

What makes a **good** model?

Must capture the essence of the process

Modelling

Modelling is an essential part of Bioinformatics. Almost every computation/prediction/estimation is based on a model of nature.

What makes a **good** model?

- Must capture the essence of the process
- As simple as possible, but not simpler

Modelling

Modelling is an essential part of Bioinformatics. Almost every computation/prediction/estimation is based on a model of nature.

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

Modelling

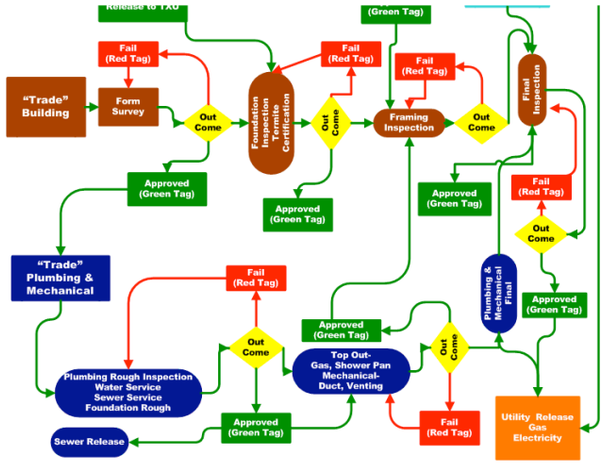
Modelling is an essential part of Bioinformatics. Almost every computation/prediction/estimation is based on a model of nature.

What makes a **good** model?

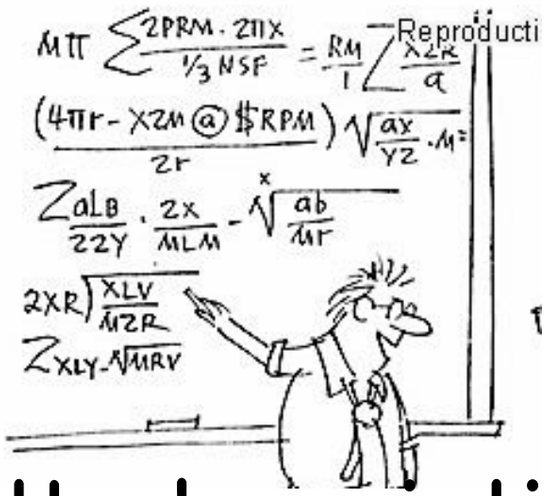
- Must capture the essence of the process
- As simple as possible, but not simpler
- Realistic in terms of the application
- Analyzable

Modelling: big picture

design a model



Process

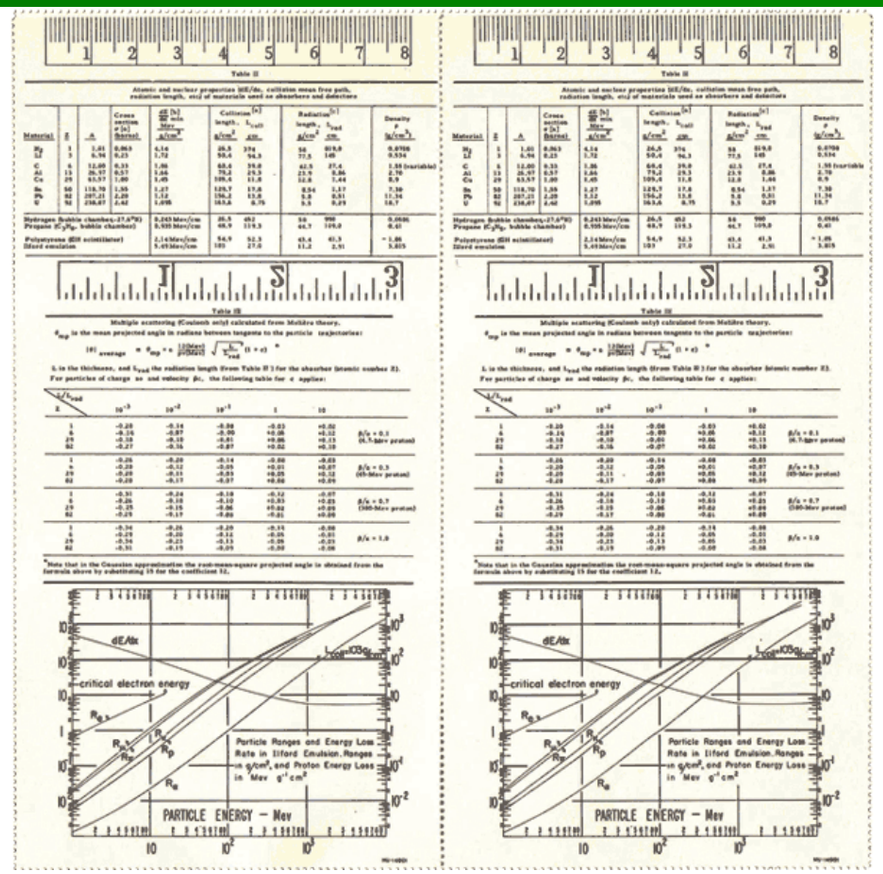


Math description

model has parameters $\alpha, \phi, [a_0, a_1, a_2, ...]$

Modelling: big picture (II)

model describes some real data



parameters are set to fit the data in the best possible way

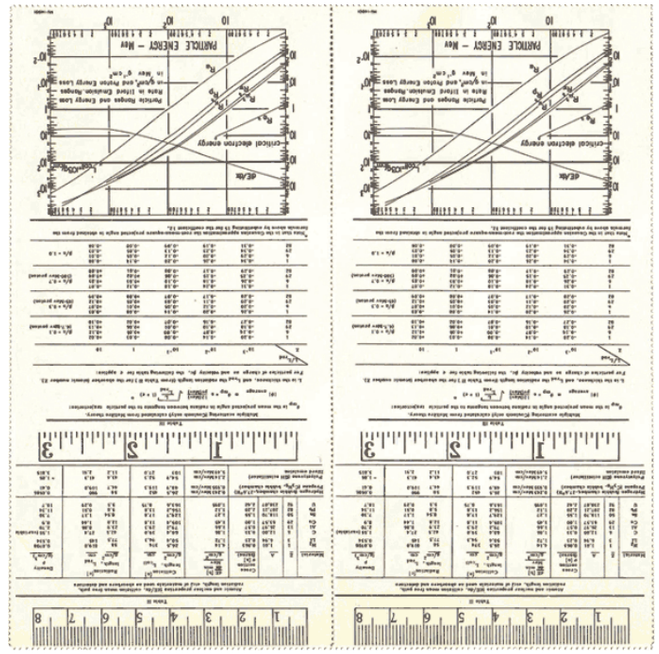
$$\alpha, \phi, [a_0, a_1, a_2, \ldots]$$

Usually by Maximum Likelihood

In closed form or numerically

Modelling: big picture (III)

model and parameters are validated with independent data



model is ready for prediction/ computation

Modelling: Closed form vs computational

Modelling: Closed form vs computational

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

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!

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.

Modelling: where the best contributions happen

Modelling: where the best contributions happen

We are very comfortable with abstractions

Modelling: where the best contributions happen

We are very comfortable with abstractions

Good understanding of powers and limits of modelling

Modelling: where the best contributions happen

We are very comfortable with abstractions

Good understanding of powers and limits of modelling

Can do the math

Modelling: where the best contributions happen

We are very comfortable with abstractions

Good understanding of powers and limits of modelling

Can do the math

Can do the computations

Mistakes happen during DNA replication

Mistakes happen during DNA replication

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

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. They rarely remain in the population.

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. They rarely 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

$$Mp_0 = p_1$$

$$M^\infty p = f$$

$$Mf = f$$

Mutation matrices

$$Mp_0 = p_1$$

M defines a unit of mutation

$$M^\infty p = f$$

$$Mf = f$$

Mutation matrices

$$Mp_0 = p_1$$

M defines a unit of mutation

$$M^\infty p = f$$

Infinite mutation results in the natural (default) frequencies

$$Mf = f$$

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)

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

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

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

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

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

Mutation matrices (II)

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

Q is the rate (differential equations of transitions) matrix

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

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

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

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

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

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

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

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

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

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

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

from $Mf=f$

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

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

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

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

from $Mf=f$

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

reaches steady state

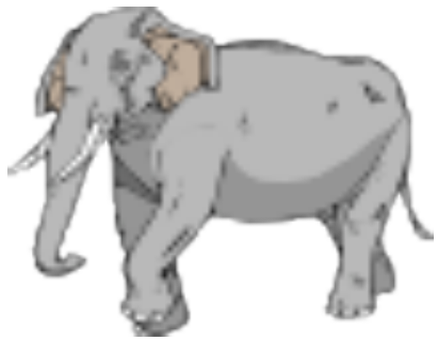
The principle of Molecular evolution

The principle of Molecular evolution

The principle of Molecular evolution

The principle of Molecular evolution

The principle of Molecular evolution

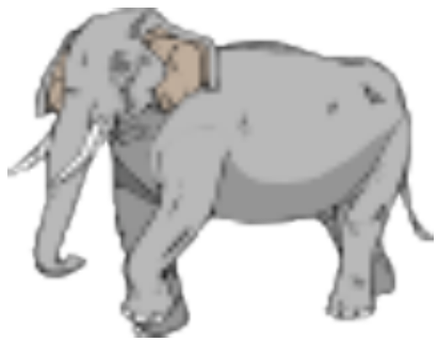


Dog DNA

aactgagcggtt...

The principle of Molecular evolution

Dog DNA
aactgagcggtt...

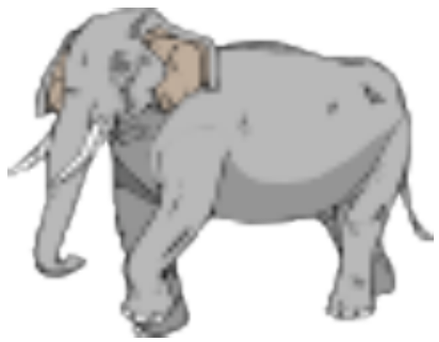


Elephant DNA
aactgacccggtt...

The principle of Molecular evolution

Dog DNA

aactgagcggtt...



Elephant DNA

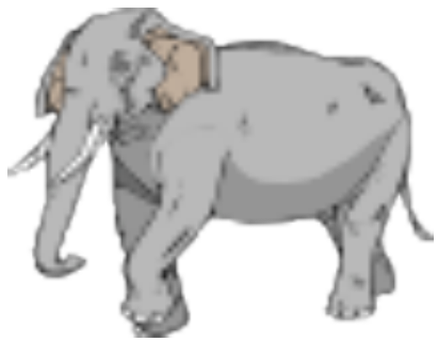
aactgacccggtt...

Rabbit DNA

aactgaccggtt...

The principle of Molecular evolution

Dog DNA
aactgagcggtt...



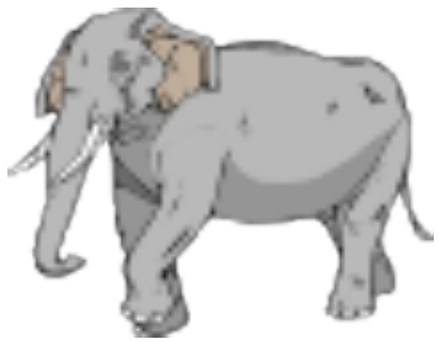
Elephant DNA
aactgacccggtt...

Rabbit DNA
aactgaccggtt...

The principle of Molecular evolution

Dog DNA

aactgagcggtt...



Elephant DNA

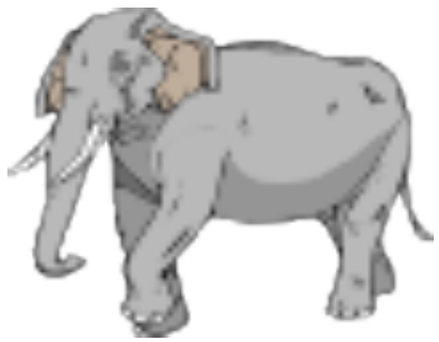
aactgacccggtt...

Rabbit DNA

aactgaccggtt...

The principle of Molecular evolution

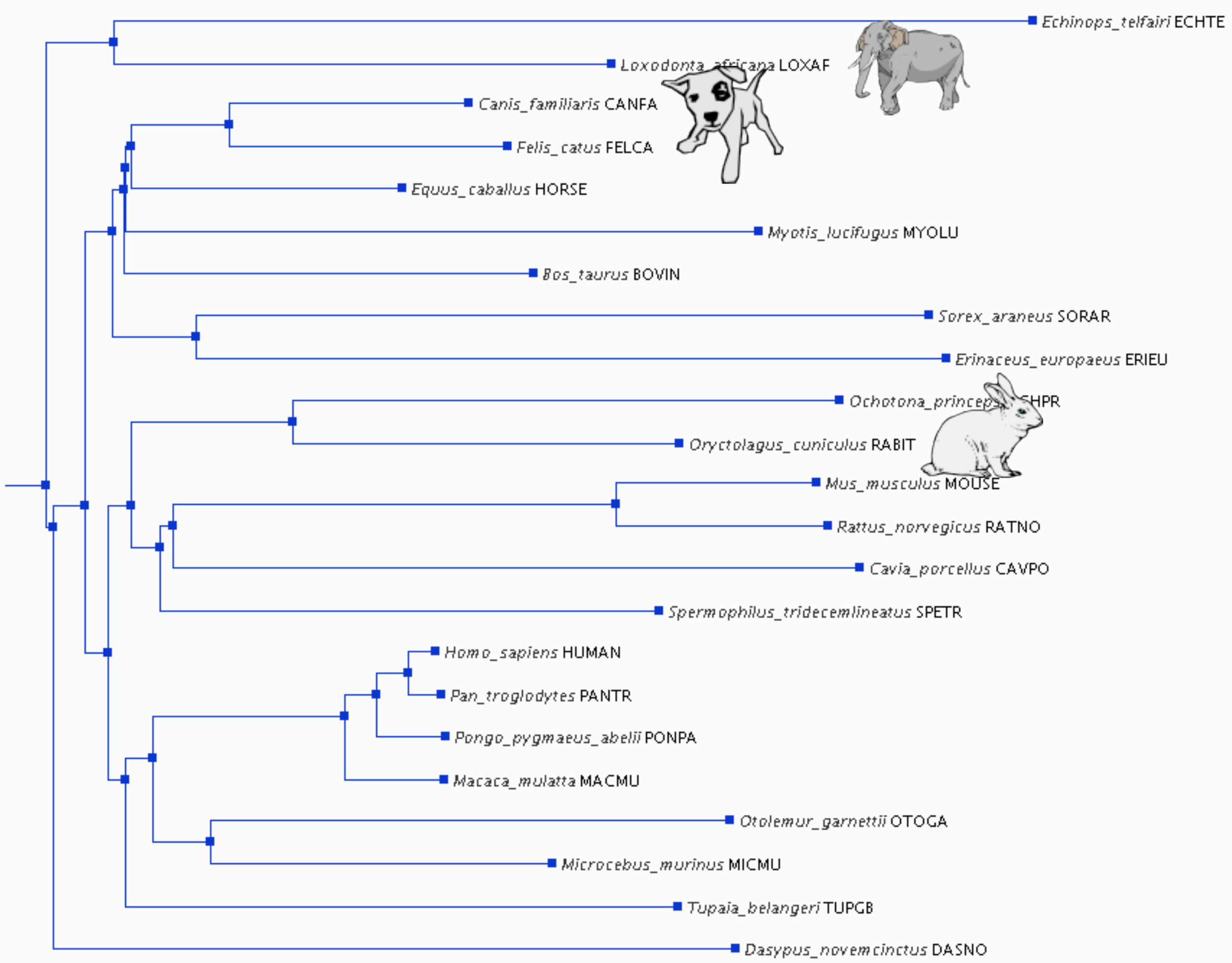
Dog DNA
aactgagcggtt...



Elephant DNA
aactgacccggtt...

Rabbit DNA
aactgaccggtt...

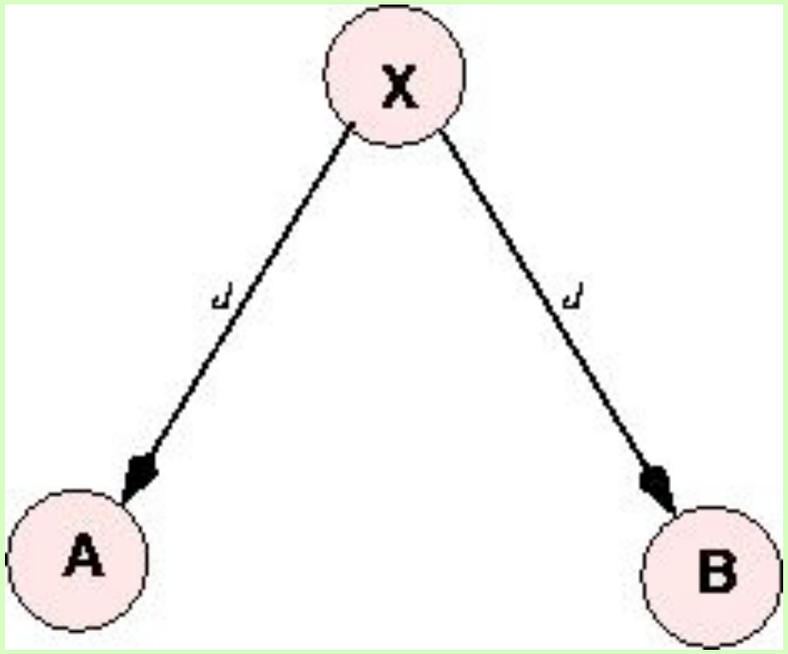
Tree of mammals



Probabilities vs likelihoods

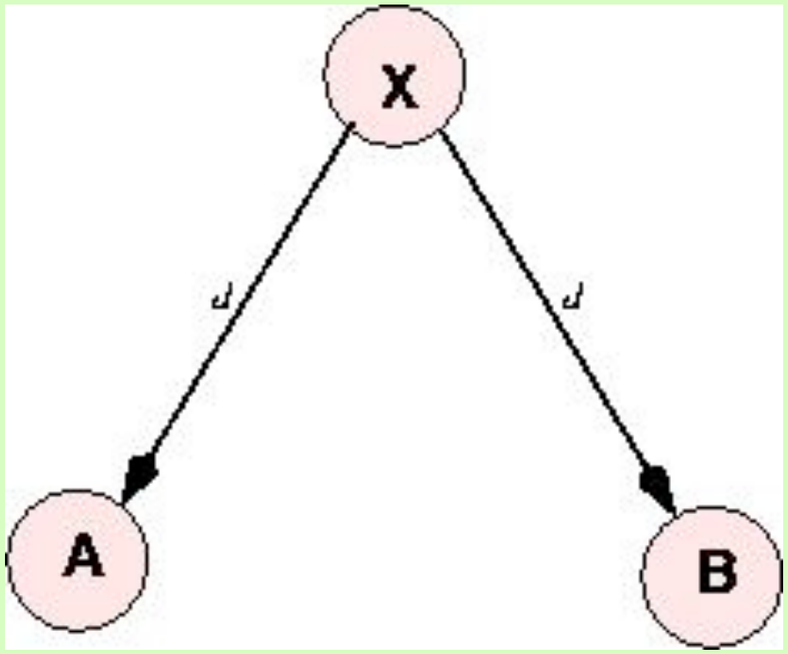
Probabilities vs likelihoods

Some event



Probabilities vs likelihoods

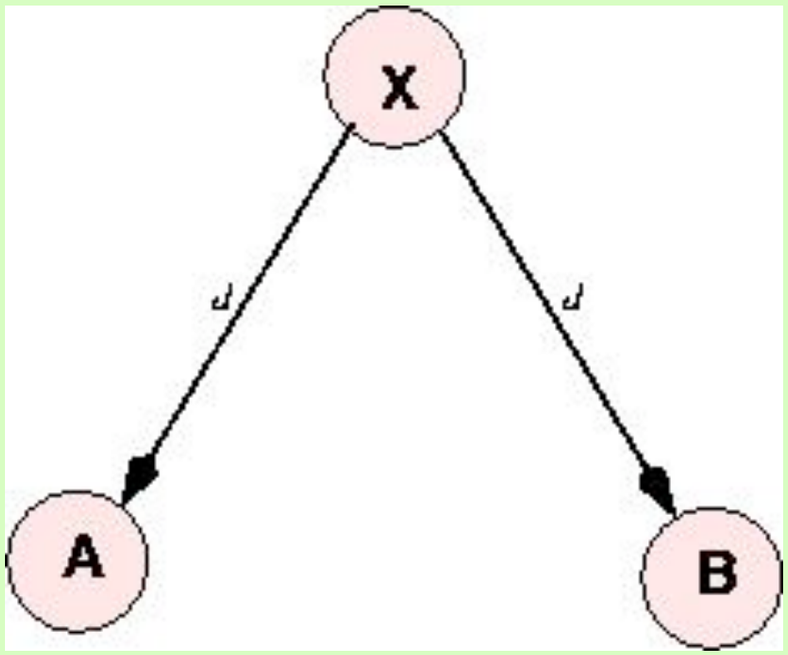
Some event



Over all X, A and B defines a probability space

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

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

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

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

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)

The maximum likelihood estimators are:

Maximum likelihood (III)

The maximum likelihood estimators are:

Unbiased

Maximum likelihood (III)

The maximum likelihood estimators are:

Unbiased

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

Maximum likelihood (III)

The maximum likelihood estimators are:

Unbiased

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

Normally distributed

Maximum likelihood (IV)

Maximum likelihood (IV)

This is ideal for symbolic/numeric computation

Maximum likelihood (IV)

This is ideal for symbolic/numeric computation

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

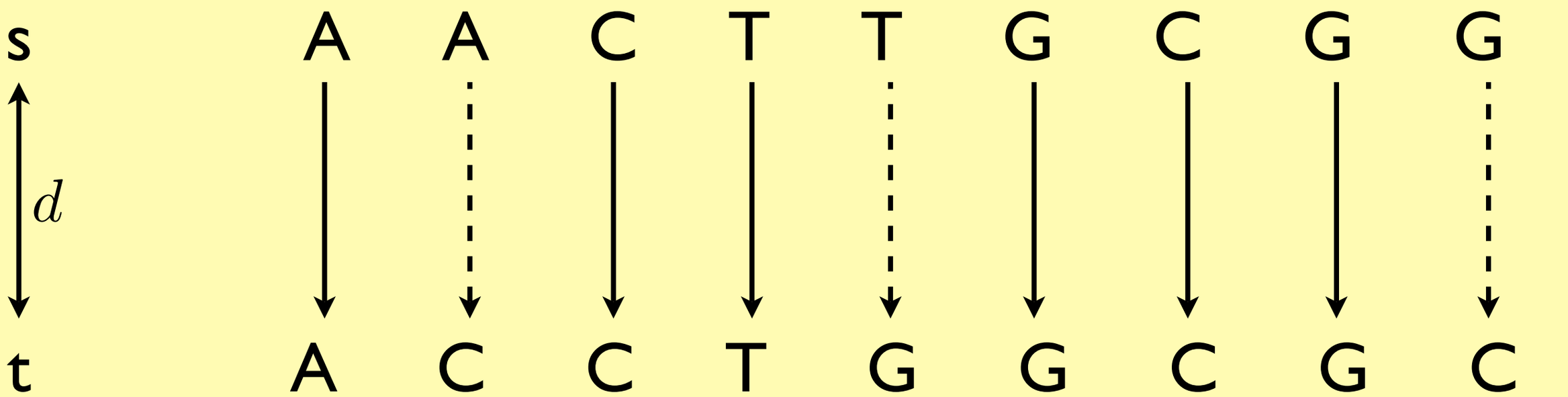
Maximum likelihood (IV)

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) ????

Inter sequence distance estimation by ML



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

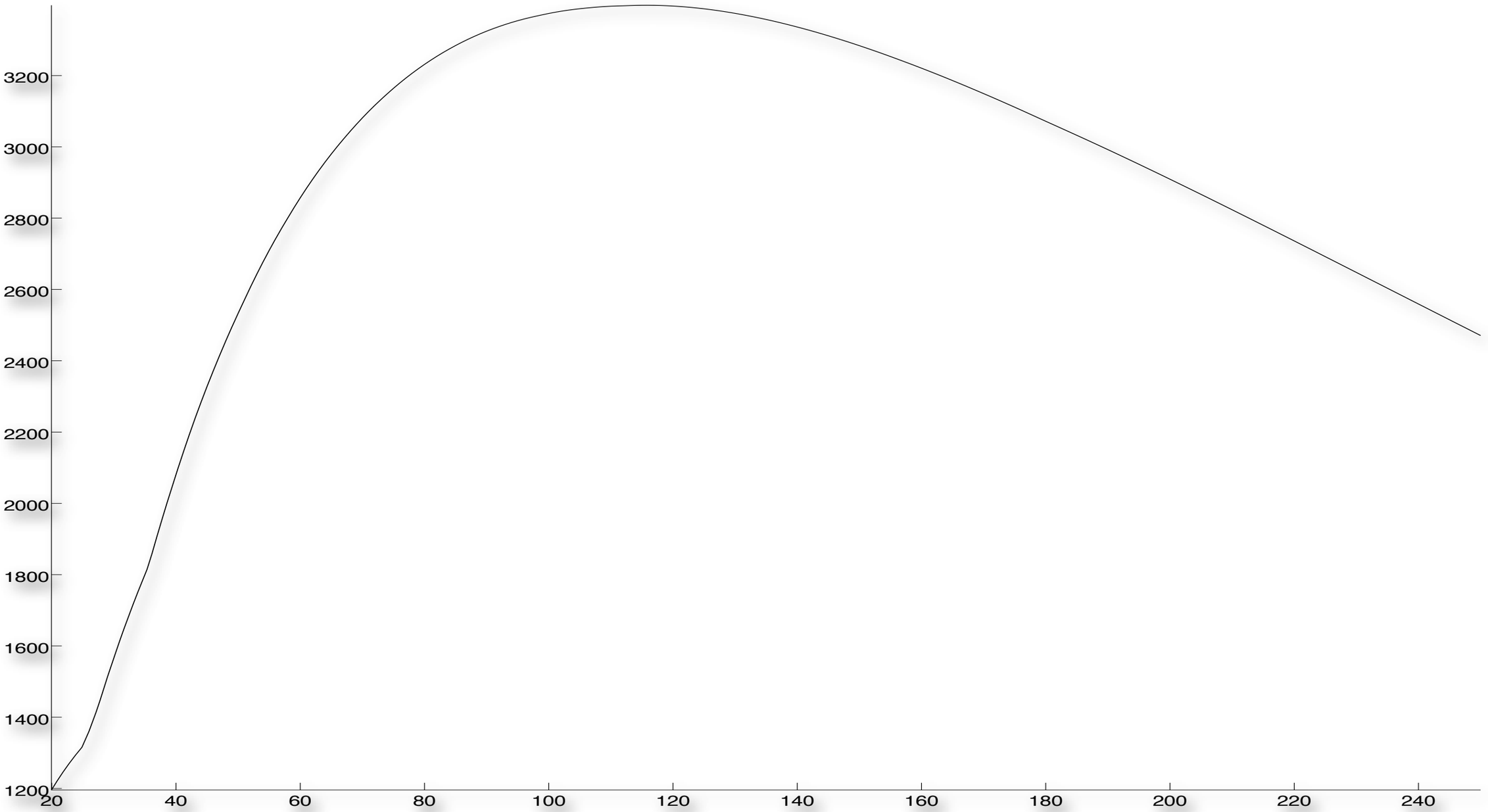
$$\ln(L(d)) = \sum_i \ln((M^d)_{s_i,t_i})$$

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:

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

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

Input (part of 108 Mb of it)

ML := ML + ln(1-Ps(27.3)) * 8312 +ln(Ps(27.3)) * 441+ln(Pl(6))+ln(Pl(3))+ln(Pl(7))+ln(Pl(10))+ln(Pl(1))+ln(Pl(5))+ln(Pl(2))+ln(Pl(6))+ln(Pl(14))+ln(Pl(1))+ln(Pl(24))+ln(Pl(2))+ln(Pl(14))+ln(Pl(18))+ln(Pl(4))+ln(Pl(11))+ln(Pl(3))+ln(Pl(7))+ln(Pl(10))+ln(Pl(12))+ln(Pl(4))+ln(Pl(2))+ln(Pl(27))+ln(Pl(7))+ln(Pl(1))+ln(Pl(8))+ln(Pl(12))+ln(Pl(1))+ln(Pl(10))+ln(Pl(6))+ln(Pl(24))+ln(Pl(24))+ln(Pl(3))+ln(Pl(1))+ln(Pl(2))+ln(Pl(7))+ln(Pl(5))+ln(Pl(36))+ln(Pl(42))+ln(Pl(4))+ln(Pl(27))+ln(Pl(1))+ln(Pl(11))+ln(Pl(1))+ln(Pl(9))+ln(Pl(1))+ln(Pl(20))+ln(Pl(2))+ln(Pl(21))+ln(Pl(14))+ln(Pl(1))+ln(Pl(17))+ln(Pl(3))+ln(Pl(13))+ln(Pl(34))+ln(Pl(3))+ln(Pl(4))+ln(Pl(4))+ln(Pl(2))+ln(Pl(2))+ln(Pl(3))+ln(Pl(11))+ln(Pl(3))+ln(Pl(5))+ln(Pl(3))+ln(Pl(4))+ln(Pl(11))+ln(Pl(9))+ln(Pl(1))+ln(Pl(3))+ln(Pl(9))+ln(Pl(9))+ln(Pl(1))+ln(Pl(5))+ln(Pl(3))+ln(Pl(39))+ln(Pl(24))+ln(Pl(27))+ln(Pl(10))+ln(Pl(6))+ln(Pl(10))+ln(Pl(43))+ln(Pl(3))+ln(Pl(3))+ln(Pl(5))+ln(Pl(1))+ln(Pl(3))+ln(Pl(4))+ln(Pl(2))+ln(Pl(1))+ln(Pl(2))+ln(Pl(19))+ln(Pl(1))+ln(Pl(5))+ln(Pl(3))+ln(Pl(3))+ln(Pl(9))+ln(Pl(23))+ln(Pl(15))+ln(Pl(10))+ln(Pl(45))+ln(Pl(7))+ln(Pl(12))+ln(Pl(16))+ln(Pl(2))+ln(Pl(2))+ln(Pl(3))+ln(Pl(3))+ln(Pl(8))+ln(Pl(9))+ln(Pl(1))+ln(Pl(12))+ln(Pl(23))+ln(Pl(1))+ln(Pl(1))+ln(Pl(1))+ln(Pl(4))+ln(Pl(21))+ln(Pl(5))+ln(Pl(21))+ln(Pl(7))+ln(Pl(6))+ln(Pl(4))+ln(Pl(6))+ln(Pl(27))+ln(Pl(15))+ln(Pl(2))+ln(Pl(4))+ln(Pl(5))+ln(Pl(11))+ln(Pl(14))+ln(Pl(10))+ln(Pl(11))+ln(Pl(11))+ln(Pl(6))+ln(Pl(2))+ln(Pl(24))+ln(Pl(7))+ln(Pl(8))+ln(Pl(24))+ln(Pl(15))+ln(Pl(29))+ln(Pl(52))+ln(Pl(9))+ln(Pl(31))+ln(Pl(10))+ln(Pl(4))+ln(Pl(3))+ln(Pl(26))+ln(Pl(11))+ln(Pl(1))+ln(Pl(2))+ln(Pl(1))+ln(Pl(15))+ln(Pl(6))+ln(Pl(3))+ln(Pl(15))+ln(Pl(12))+ln(Pl(13))+ln(Pl(16))+ln(Pl(4))+ln(Pl(2))+ln(Pl(1))+ln(Pl(1))+ln(Pl(8))+ln(Pl(43))+ln(Pl(5))+ln(Pl(9))+ln(Pl(7))+ln(Pl(17))+ln(Pl(19))+ln(Pl(39))+ln(Pl(3))+ln(Pl(3))+ln(Pl(5))+ln(Pl(31))+ln(Pl(7))+ln(Pl(4))+ln(Pl(8))+ln(Pl(21))+ln(Pl(6))+ln(Pl(7))+ln(Pl(2))+ln(Pl(9))+ln(Pl(3))+ln(Pl(4))+ln(Pl(37))+ln(Pl(5))+ln(Pl(28))+ln(Pl(9))+ln(Pl(18))+ln(Pl(6))+ln(Pl(17))+ln(Pl(8))+ln(Pl(11))+ln(Pl(1))+ln(Pl(5))+ln(Pl(2))+ln(Pl(1))+ln(Pl(2))+ln(Pl(1))+ln(Pl(1))+ln(Pl(2))+ln(Pl(2))+ln(Pl(2))+ln(Pl(1))+ln(Pl(6))+ln(Pl(2))+ln(Pl(3))+ln(Pl(93))+ln(Pl(14))+ln(Pl(1))+ln(Pl(1))+ln(Pl(25))+ln(Pl(4))+ln(Pl(6))+ln(Pl(6))+ln(Pl(7))+ln(Pl(4))+ln(Pl(4))+ln(Pl(4))+ln(Pl(51))+ln(Pl(55))+ln(Pl(1))+ln(Pl(5))+ln(Pl(6))+ln(Pl(4))+ln(Pl(1))+ln(Pl(6))+ln(Pl(6))+ln(Pl(17))+ln(Pl(15))+ln(Pl(5))+ln(Pl(5))+ln(Pl(3))+ln(Pl(9))+ln(Pl(3))+ln(Pl(1))+ln(Pl(1))+ln(Pl(5))+ln(Pl(4))+ln(Pl(6))+ln(Pl(10))+ln(Pl(1))+ln(Pl(1))+ln(Pl(1))+ln(Pl(1))+ln(Pl(3))+ln(Pl(2))+ln(Pl(5))+ln(Pl(2))+ln(Pl(12))+ln(Pl(13))+ln(Pl(2))+ln(Pl(24))+ln(Pl(14))+ln(Pl(6))+ln(Pl(8))+ln(Pl(18))+ln(Pl(4))+ln(Pl(4))+ln(Pl(19))+ln(Pl(19))+ln(Pl(2))+ln(Pl(68))+ln(Pl(15))+ln(Pl(2))+ln(Pl(8))+ln(Pl(12))+ln(Pl(12))+ln(Pl(7))+ln(Pl(8))+ln(Pl(14))+ln(Pl(1))+ln(Pl(5))+ln(Pl(11))+ln(Pl(18))+ln(Pl(6))+ln(Pl(25))+ln(Pl(8))+ln(Pl(9))+ln(Pl(4))+ln(Pl(20))+ln(Pl(22))+ln(Pl(4))+ln(Pl(18))+ln(Pl(27))+ln(Pl(6))+ln(Pl(22))+ln(Pl(7))+ln(Pl(3))+ln(Pl(2))+ln(Pl(3))+ln(Pl(7))+ln(Pl(14))+ln(Pl(1))+ln(Pl(3))+ln(Pl(5))+ln(Pl(1))+ln(Pl(2))+ln(Pl(5))+ln(Pl(5))+ln(Pl(30))+ln(Pl(30))+ln(Pl(6))+ln(Pl(4))+ln(Pl(17))+ln(Pl(6))+ln(Pl(30))+ln(Pl(17))+ln(Pl(17))+ln(Pl(21))+ln(Pl(2))+ln(Pl(15))+ln(Pl(1))+ln(Pl(3))+ln(Pl(2))+ln(Pl(3))+ln(Pl(1))+ln(Pl(17))+ln(Pl(1))+ln(Pl(1))+ln(Pl(1))+ln(Pl(10))+ln(Pl(11))+ln(Pl(1))+ln(Pl(1))+ln(Pl(9))+ln(Pl(2))+ln(Pl(24))+ln(Pl(2))+ln(Pl(10))+ln(Pl(6))+ln(Pl(1))+ln(Pl(7))+ln(Pl(1))+ln(Pl(3))+ln(Pl(1))+ln(Pl(4))+ln(Pl(1))+ln(Pl(1))+ln(Pl(5))+ln(Pl(1))+ln(Pl(48))+ln(Pl(3))+ln(Pl(1))+ln(Pl(1))+ln(Pl(3))+ln(Pl(1))+ln(Pl(9))+ln(Pl(2))+ln(Pl(203))+ln(Pl(4))+ln(Pl(6))+ln(Pl(2))+ln(Pl(5))+ln(Pl(7))+ln(Pl(16))+ln(Pl(1))+ln(Pl(21))+ln(Pl(2))+ln(Pl(2))+ln(Pl(4))+ln(Pl(3))+ln(Pl(14))+ln(Pl(6))+ln(Pl(14))+ln(Pl(5))+ln(Pl(2))+ln(Pl(7))+ln(Pl(41))+ln(Pl(1))+ln(Pl(3))+ln(Pl(4))+ln(Pl(1))+ln(Pl(37))+ln(Pl(10))+ln(Pl(24))+ln(Pl(2))+ln(Pl(9))+ln(Pl(15))+ln(Pl(1))+ln(Pl(26))+ln(Pl(2))+ln(Pl(15))+ln(Pl(19))+ln(Pl(1))+ln(Pl(7))+ln(Pl(10))+ln(Pl(1))+ln(Pl(6))+ln(Pl(3))+ln(Pl(2))+ln(Pl(3))+ln(Pl(1))+ln(Pl(9))+ln(Pl(1))+ln(Pl(4))+ln(Pl(1))+ln(Pl(10))+ln(Pl(5))+ln(Pl(5))+ln(Pl(2))+ln(Pl(2))+ln(Pl(17))+ln(Pl(18))+ln(Pl(18))+ln(Pl(6))+ln(Pl(3))+ln(Pl(22))+ln(Pl(18))+ln(Pl(10))+ln(Pl(1))+ln(Pl(25))+ln(Pl(17))+ln(Pl(5))+ln(Pl(14))+ln(Pl(1))+ln(Pl(17))+ln(Pl(16))+ln(Pl(18))+ln(Pl(18))+ln(Pl

Evolution happens at very different speeds

Evolution happens at very different speeds

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

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)

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)

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)

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)

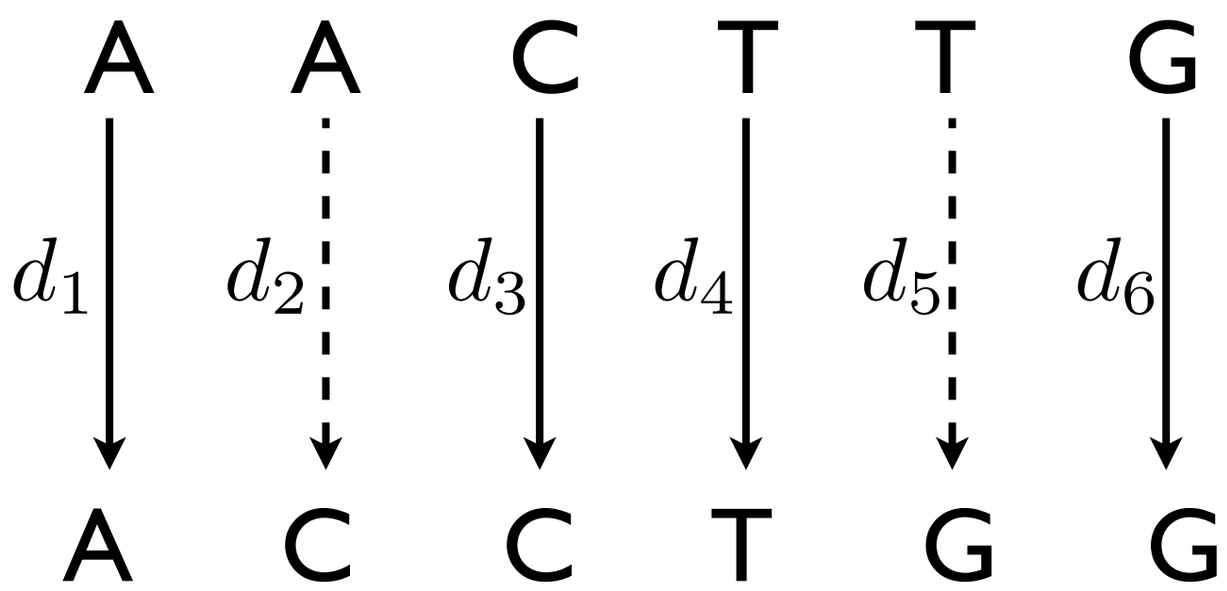
Mitochondrial DNA is more than 99% identical in all humans (200,000 years)

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

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

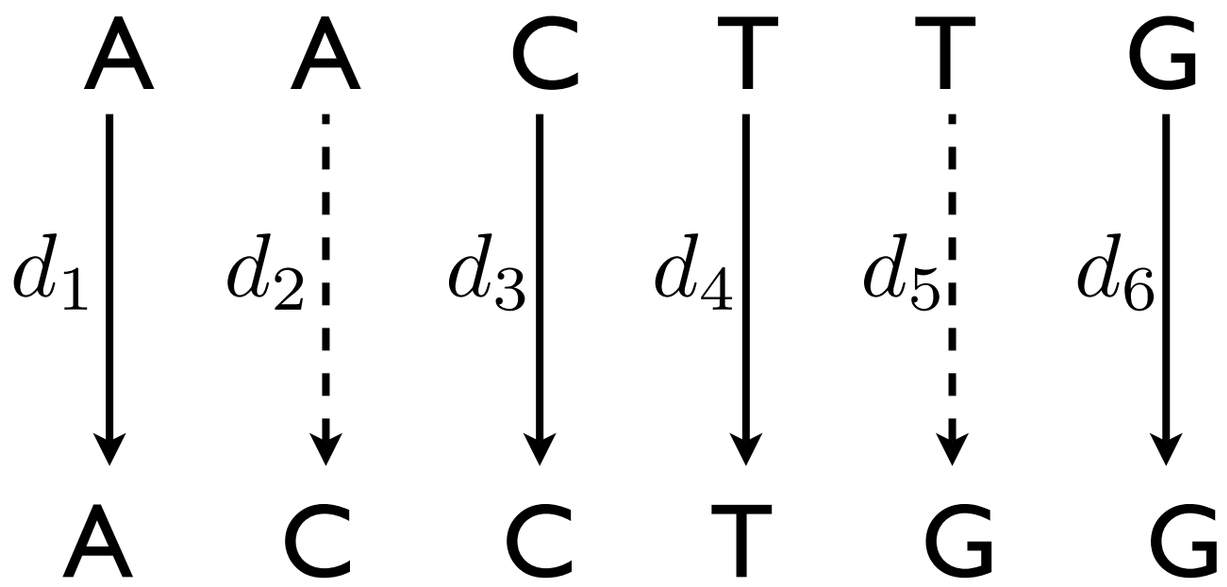


Variable evolution rates (I)

It is recognized as biologically appropriate that different positions evolve at different rates

Modeled with a gamma distribution (no particularly 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

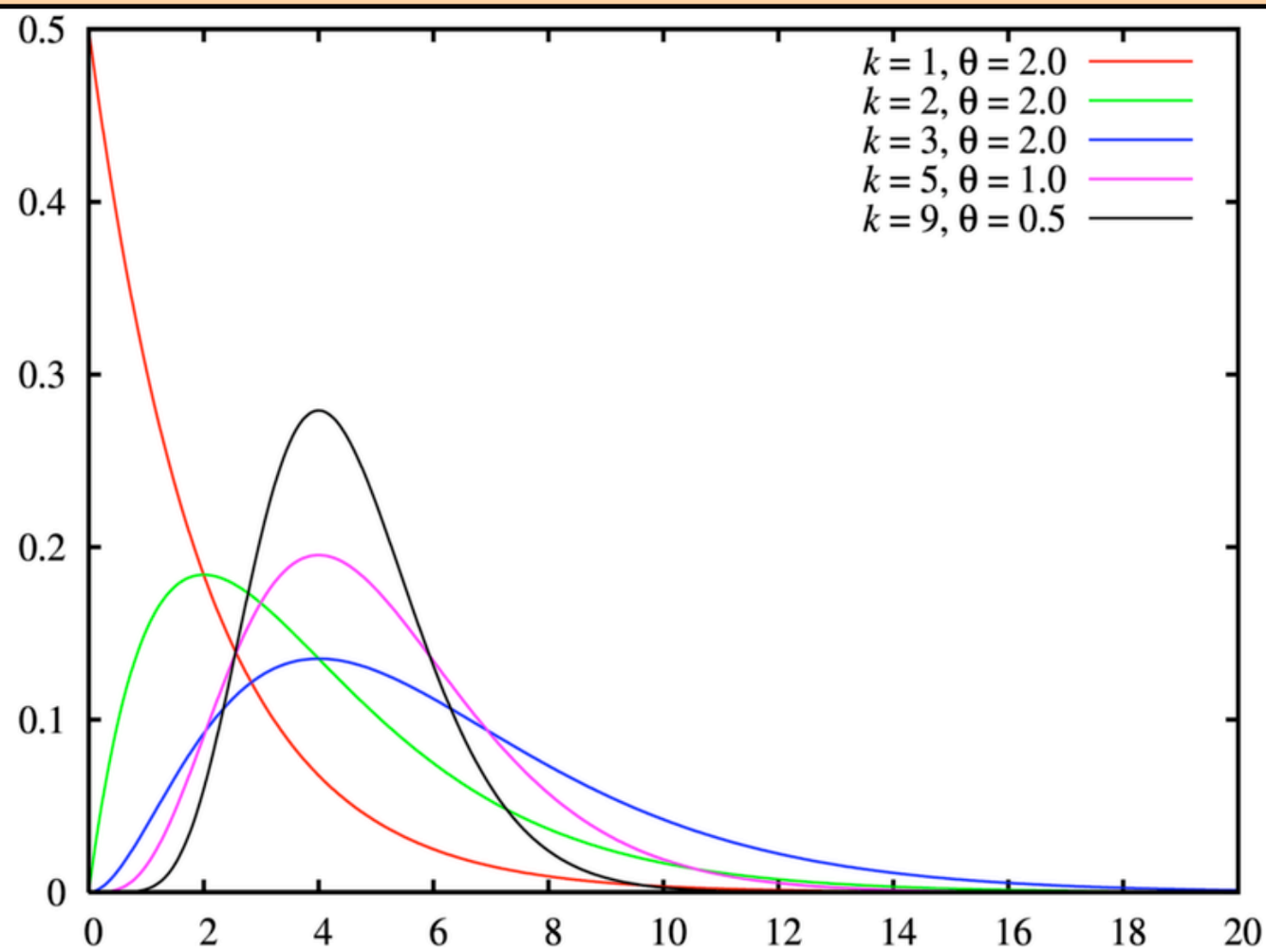
$$p(x) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k}$$

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

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

$$mgf(t) = \int_0^\infty e^{tx} p(x) dx$$

$$= (1 - \theta t)^{-k}$$



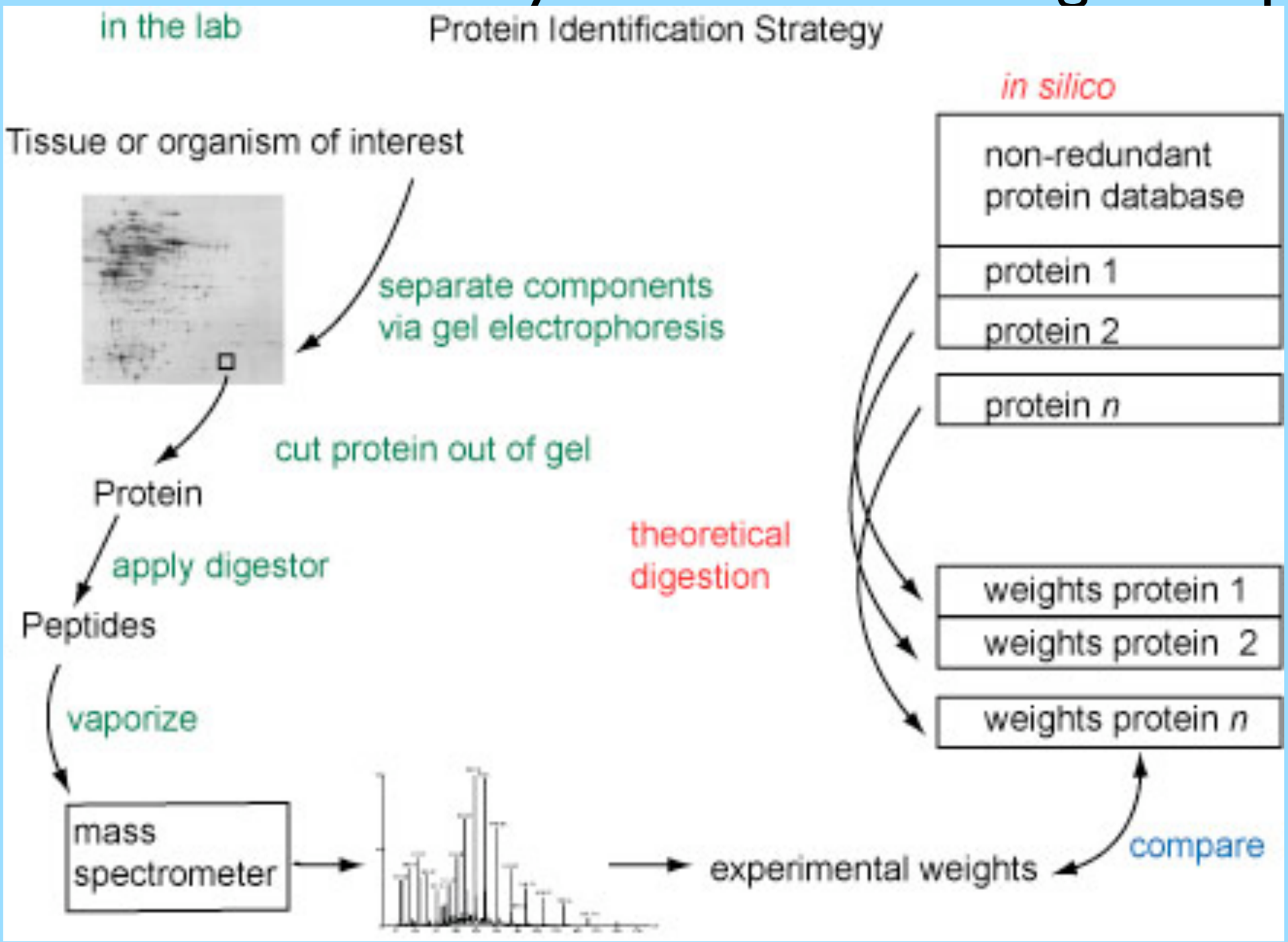
Variable evolution rates (III)

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

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

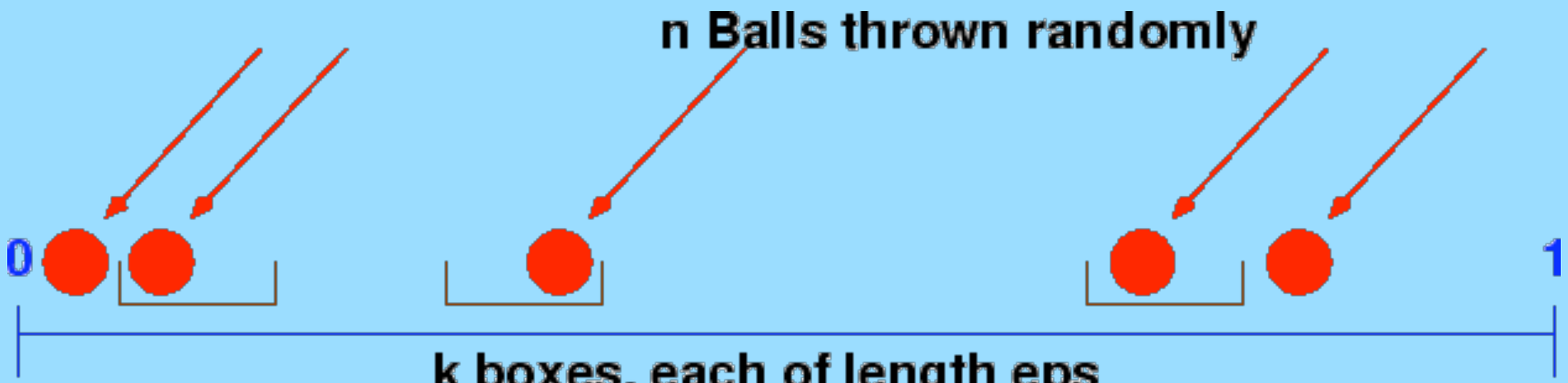
Molecular weight fingerprinting

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:

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

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

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

$$G^*_{2,n,\epsilon} = G_{2,n,\epsilon} - G_{2,n,\epsilon}|_{a_1=0}$$

$$= (a_1\epsilon + a_2\epsilon + b(1-2\epsilon))^n - (a_2\epsilon + b(1-2\epsilon))^n$$

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

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

The model for comparison - generating function

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

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

(recently solved by Daniele Gardy with a much better expansion)

How diverse are humans?

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)

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.

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.

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
...			

Ancestry testing

Simplest model: each locus has its own probability, independent of the others

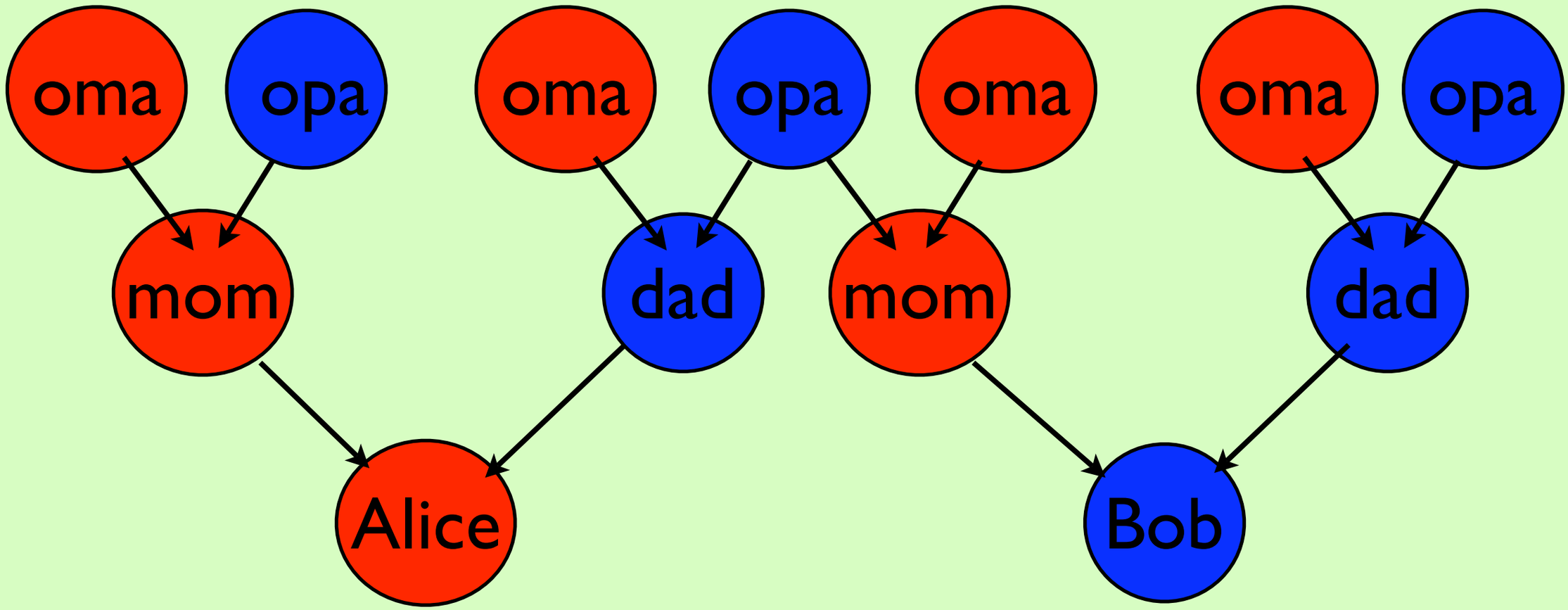
good for simple paternity testing

More realistic model: loci which are close to each other are highly correlated

necessary for more complicated cases

Ancestry testing (II)

Example: do Alice and Bob share one grandparent?



DHL founder, Larry Hillblom's case

DHL founder, Larry Hillblom's case

Died in a seaplane crash.

DHL founder, Larry Hillblom's case

Died in a seaplane crash.

No family, University of California to receive his estate

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.

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 proven to come from the same father and they received their rightly share.

the END