Selectome, looking for Darwinian selection in the Tree of Life

Marc Robinson-Rechavi, Heinz Stockinger and Nicolas Salamin

Department of Ecology and Evolution, University of Lausanne

And

Swiss Institute of Bioinformatics

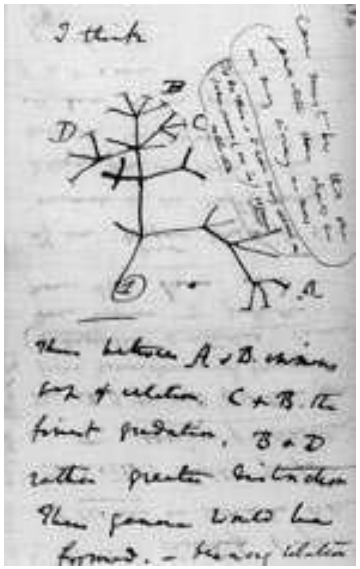
HP2C meeting, Lugano

March 16-17, 2010

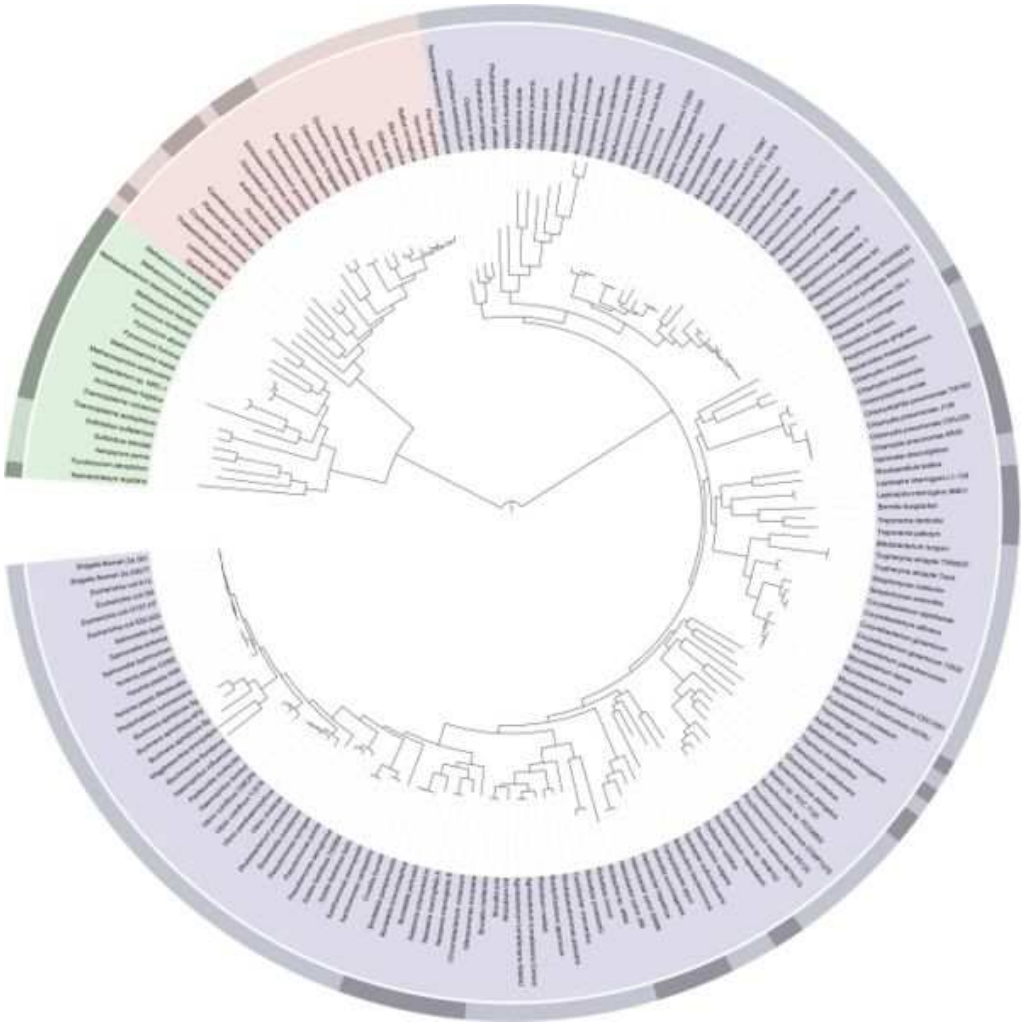
Evolution

One area of biology is interested to understand how life on earth was shaped by evolution.

- Natural selection, described first by Darwin (1859), is the main evolutionary force acting on natural variation.
- Selective pressure from the environment drive evolution by allowing the fittest to leave more offsprings.

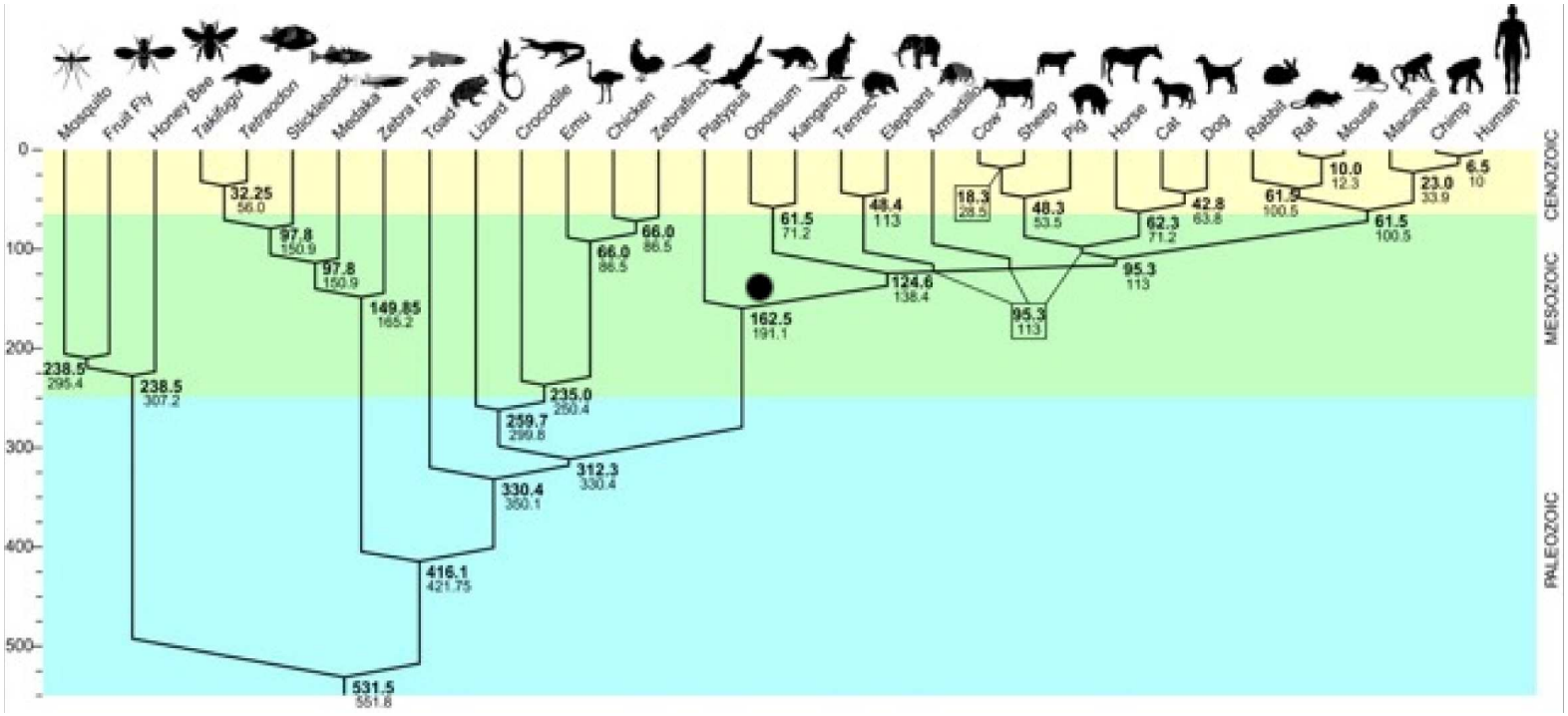


Tree of Life
from Darwin to today



Evolutionary theory predicts that species share common ancestor

Phylogenetic tree



Mutations and selection

Environment creates selective pressure

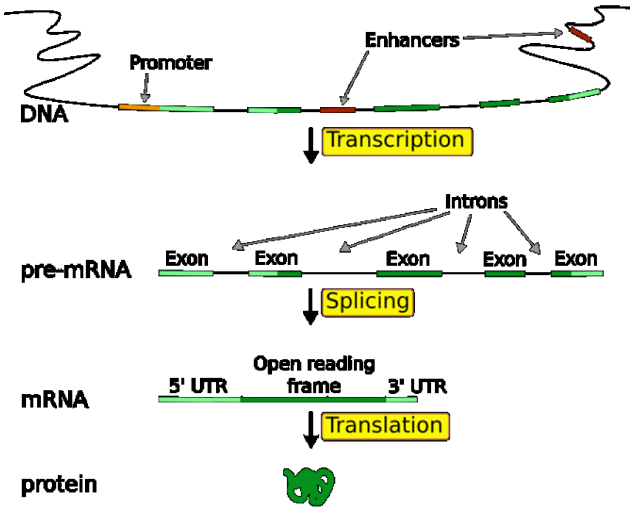
- creates random mutations in genome that can modify phenotype (biochemical function, gene expression, . . .)

Effects of these mutations

- most have negative effect (i.e. deleterious); will be removed by purifying selection
- some have no effect; may be fixed under the neutral process of drift
- others are beneficial; will be kept by positive selection (also called Darwinian, adaptive, or directional selection)

Positive selection is the mechanism of adaptation to the environment. It important to find its trace in genomes.

Selective pressure on proteins



species 1	A	L	P	H	Y	Protein
	GCC	CTT	CCT	CAT	TAT	DNA
species 2	A	R	P	H	Y	Protein
	GCC	CGT	CCT	CAT	TAC	DNA

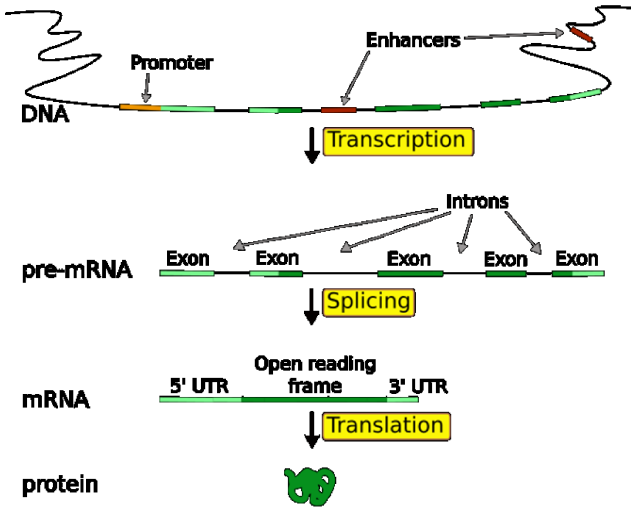
Mutation rate, genetic drift and time:

$$dS = \frac{\text{Number of synonymous changes}}{\text{Number of synonymous sites}}$$

Mutation rate, genetic drift, time and selection pressure:

$$dN = \frac{\text{Number of non-synonymous changes}}{\text{Number of non-synonymous sites}}$$

Selective pressure on proteins



species 1	A	L	P	H	Y	Protein
	GCC	CTT	CCT	CAT	TAT	DNA
species 2	A	R	P	H	Y	Protein
	GCC	CGT	CCT	CAT	TAC	DNA

Mutation rate, genetic drift and time:

$$dS = \frac{\text{Number of synonymous changes}}{\text{Number of synonymous sites}}$$

Mutation rate, genetic drift, time and selection pressure:

$$dN = \frac{\text{Number of non-synonymous changes}}{\text{Number of non-synonymous sites}}$$

Meaning of ω

The parameter $\omega = dN/dS$ gives the strength of Darwinian selection.

- if amino acid change is neutral $\Rightarrow$ will be fixed at same rate as synonymous mutation, so $\omega = 1$
- if amino acid change is deleterious $\Rightarrow$ purifying selection reduce its fixation rate so $\omega < 1$
- if amino acid change is selectively advantageous $\Rightarrow$ will be fixed at higher rate than synonymous mutation, so $\omega > 1$

Looking for Darwinian selection

Vol 437|20 October 2005|doi:10.1038/nature04240 nature

LETTERS

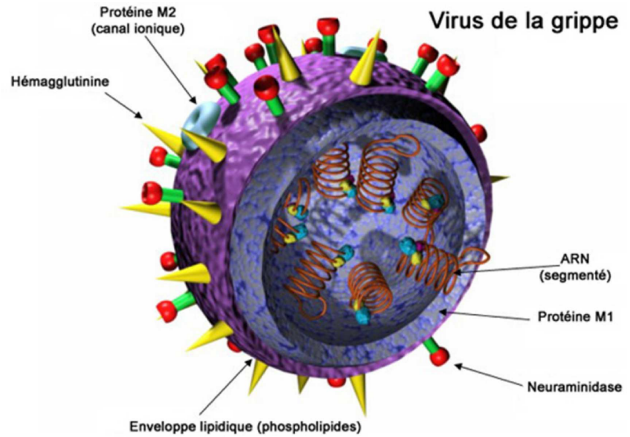
Natural selection on protein-coding genes in the human genome

Carlos D. Bustamante[1], Adi Fledel-Alon[1], Scott Williamson[1], Rasmus Nielsen[1,2], Melissa Todd Hubisz[1], Stephen Glanowski[3], David M. Tanenbaum[3], Thomas J. White[4], John J. Sninsky[4], Ryan D. Hernandez[1], Daniel Civello[4], Mark D. Adams[5], Michele Cargill[4]* & Andrew G. Clark[6]*

Comparisons of DNA polymorphism within species to divergence between species enables the discovery of molecular adaptation in evolutionarily constrained genes as well as the differentiation of weak from strong purifying selection[1–4]. The extent to which weak negative and positive darwinian selection have driven the molecular evolution of different species varies greatly[5–16], with some species, such as *Drosophila melanogaster*, showing strong evidence of pervasive positive selection[6–9], and others, such as the selfing weed *Arabidopsis thaliana*, showing an excess of deleterious variation within local populations[9,10]. Here we contrast patterns of coding sequence polymorphism identified by direct sequencing

some form of coding nucleotide variability either within human subjects or between humans and a chimpanzee. A total of 34,099 fixed synonymous differences between all humans in our sample and the chimpanzee yield a genomic average synonymous divergence of $\overline{d_S} = 1.02\%$. Correspondingly, we found 20,467 non-synonymous differences ($\overline{d_N} = 0.242\%$) across 11.81 megabases (Mb) of aligned coding DNA. We also discovered 15,750 synonymous and 14,311 non-synonymous SNPs among the human subjects, yielding average synonymous and non-synonymous SNP densities of $\overline{p_S} = 0.470\%$ and $\overline{p_N} = 0.169\%$. We note that the ratio of non-synonymous to synonymous differences (23.76%) is smaller than the ratio of non-

Genes and Darwinian selection



Known genes involved in arm-race

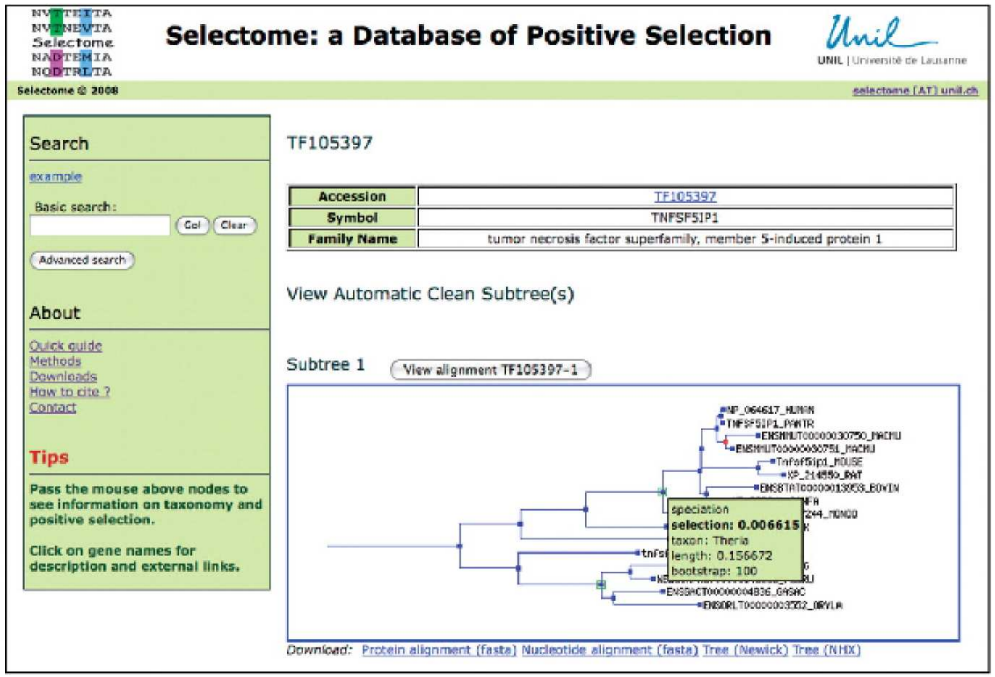
- virus and bacteria: HIV proteins (env, gag, pol) or H1N1 influenza
- MHC to recognize any kind of external peptides in our body
- genes involved in sexual reproduction (sperm lysin in abalone, protamine P1 in primates)
- genes of perseption and digestion (e.g. olfactory receptor, lysosyme)

We need a systematic assessment of genes and lineages affected by Darwinian selection.

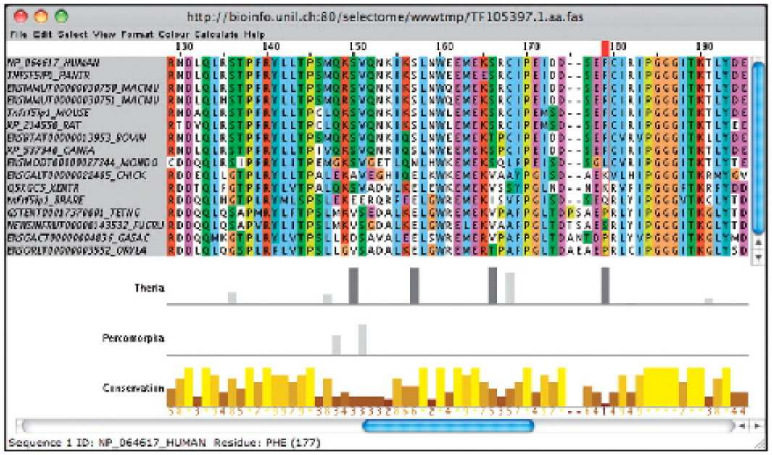
http://selectome.unil.ch

Selectome database

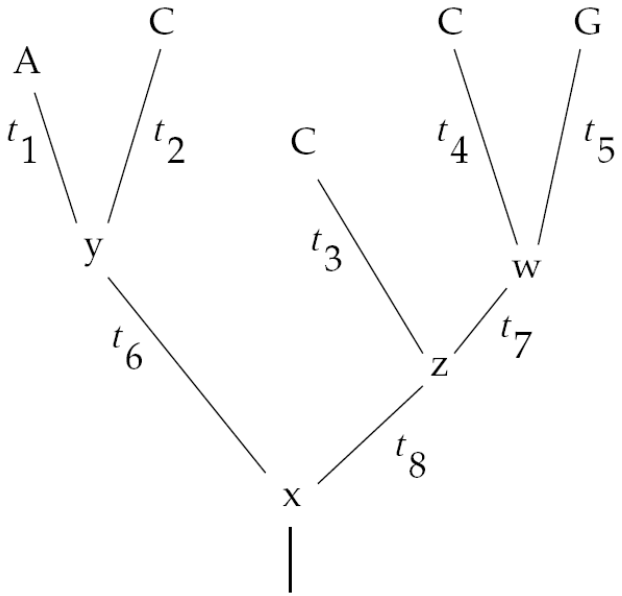
A



B



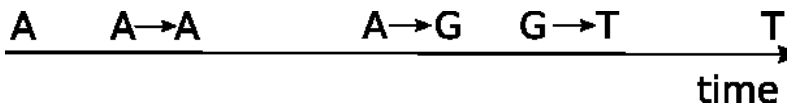
Modeling molecular evolution



How to infer evolutionary changes along each branch of a tree

- Markov model for transitions between states
- Maximum likelihood estimation of parameters
- Dynamic programming algorithm

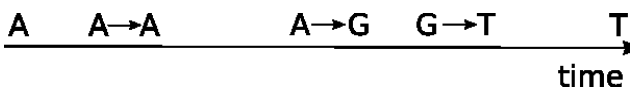
Number of mutation events



Substitution events occur according to a continuous time Markov chain. The number of these events along a branch has a Poisson distribution:

$$Prob[k \text{ events}] = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

- μ is the rate of mutations
- expected number of events in time t is μt



Markov process

The Markov chain is characterized by its generator matrix $Q = \{q_{ij}\}$, where q_{ij} is the instantaneous rate of change from nucleotides i to j when $\Delta t \rightarrow 0$, that is

$$Pr\{X(t+\Delta t) = j | X(t) = i\} = q_{ij}\Delta t$$

The Q matrix fully determines the dynamics of the Markov chain.

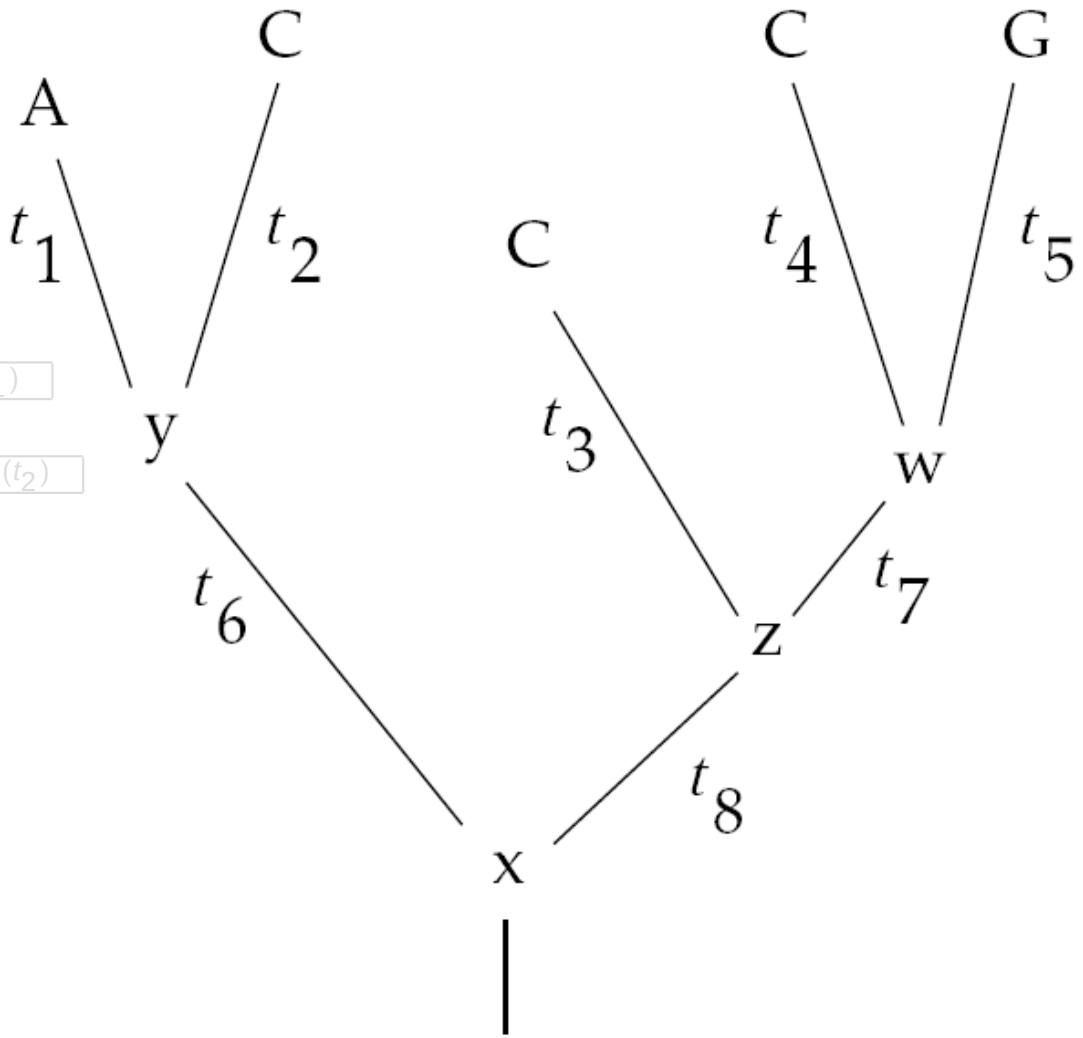
It specifies, in particular, the transition-probability matrix over any time $t > 0$, $P(t) = \{p_{ij}(t)\}$ where

$$p_{ij}(t+s) = Pr\{X(t+s) = j | X(t) = i\} = \sum^{k} p_{ik}(t)p_{kj}(s)$$

with the relationship

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

ML calculations for one site



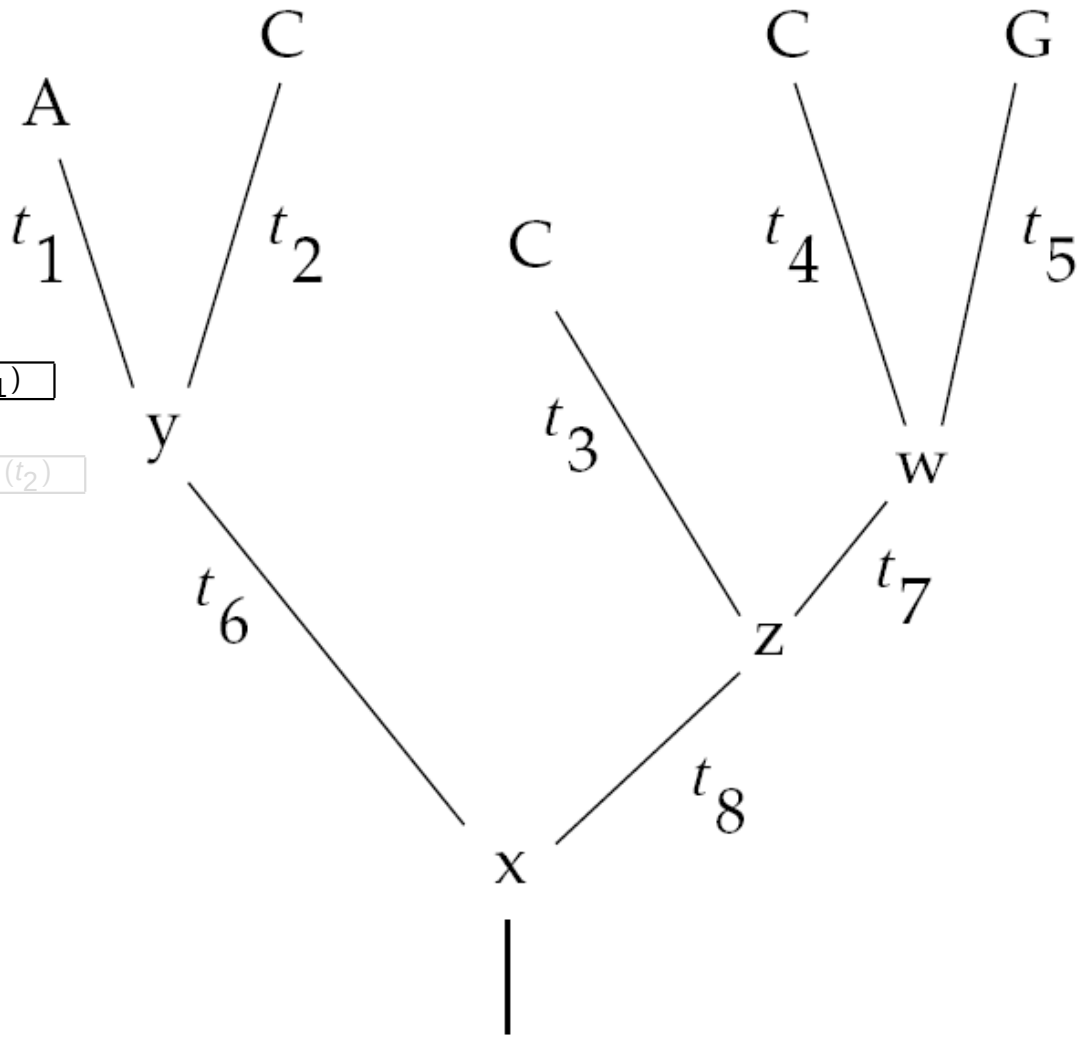
$$\begin{array}{|c|c|c|c|} \hline P_{AA}(t_1) & P_{CA}(t_1) & P_{GA}(t_1) & P_{TA}(t_1) \\ \hline \end{array} * \begin{array}{|c|c|c|c|} \hline P_{AC}(t_2) & P_{CC}(t_2) & P_{GC}(t_2) & P_{TC}(t_2) \\ \hline \end{array}$$

ML calculations for one site

$P_{AA}(t_1)$	$P_{CA}(t_1)$	$P_{GA}(t_1)$	$P_{TA}(t_1)$

*

$P_{AC}(t_2)$	$P_{CC}(t_2)$	$P_{GC}(t_2)$	$P_{TC}(t_2)$

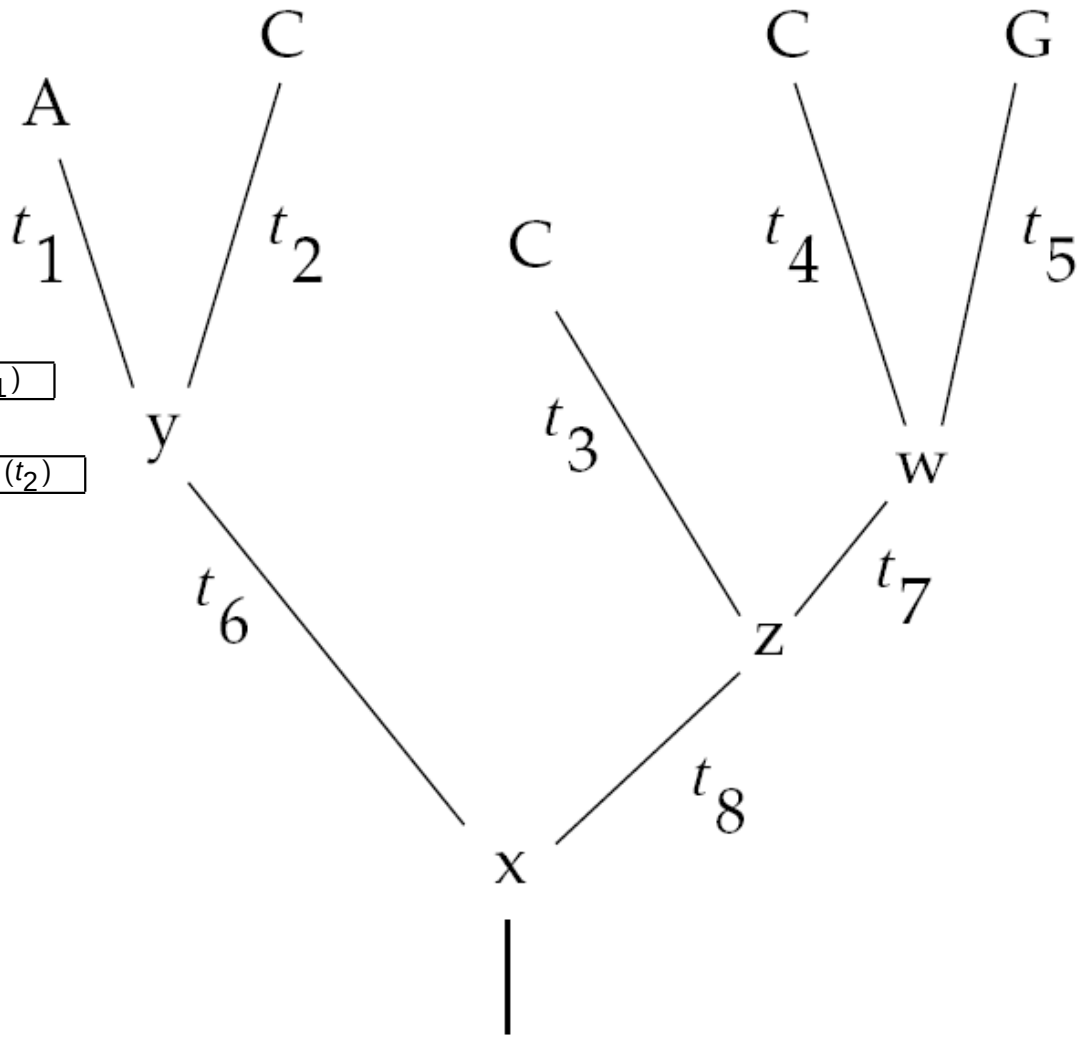


ML calculations for one site

$P_{AA}(t_1)$	$P_{CA}(t_1)$	$P_{GA}(t_1)$	$P_{TA}(t_1)$

*

$P_{AC}(t_2)$	$P_{CC}(t_2)$	$P_{GC}(t_2)$	$P_{TC}(t_2)$

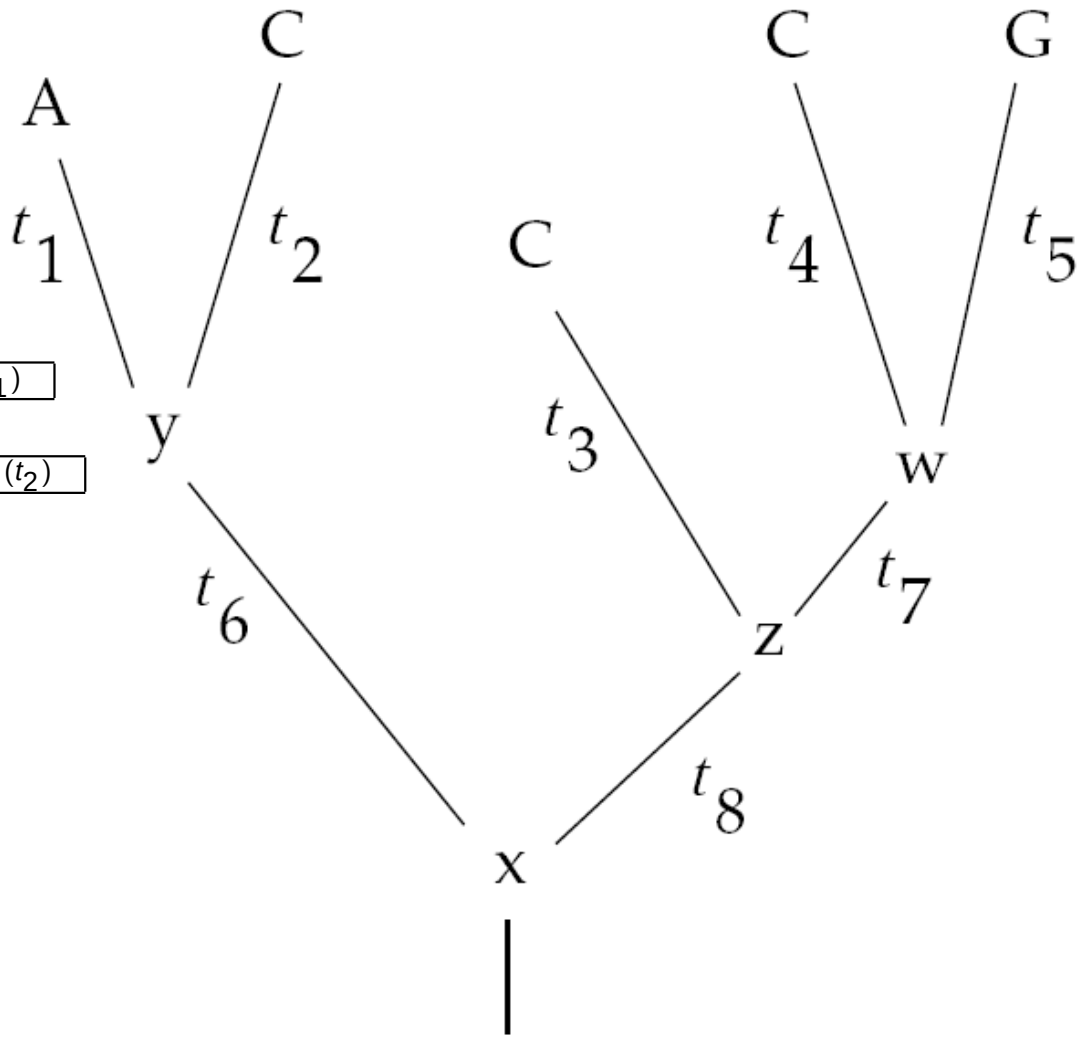


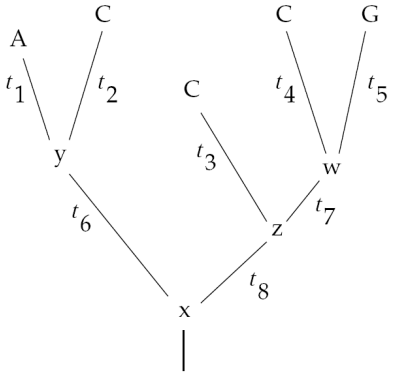
ML calculations for one site

$P_{AA}(t_1)$	$P_{CA}(t_1)$	$P_{GA}(t_1)$	$P_{TA}(t_1)$

*

$P_{AC}(t_2)$	$P_{CC}(t_2)$	$P_{GC}(t_2)$	$P_{TC}(t_2)$





ML calculations for one site

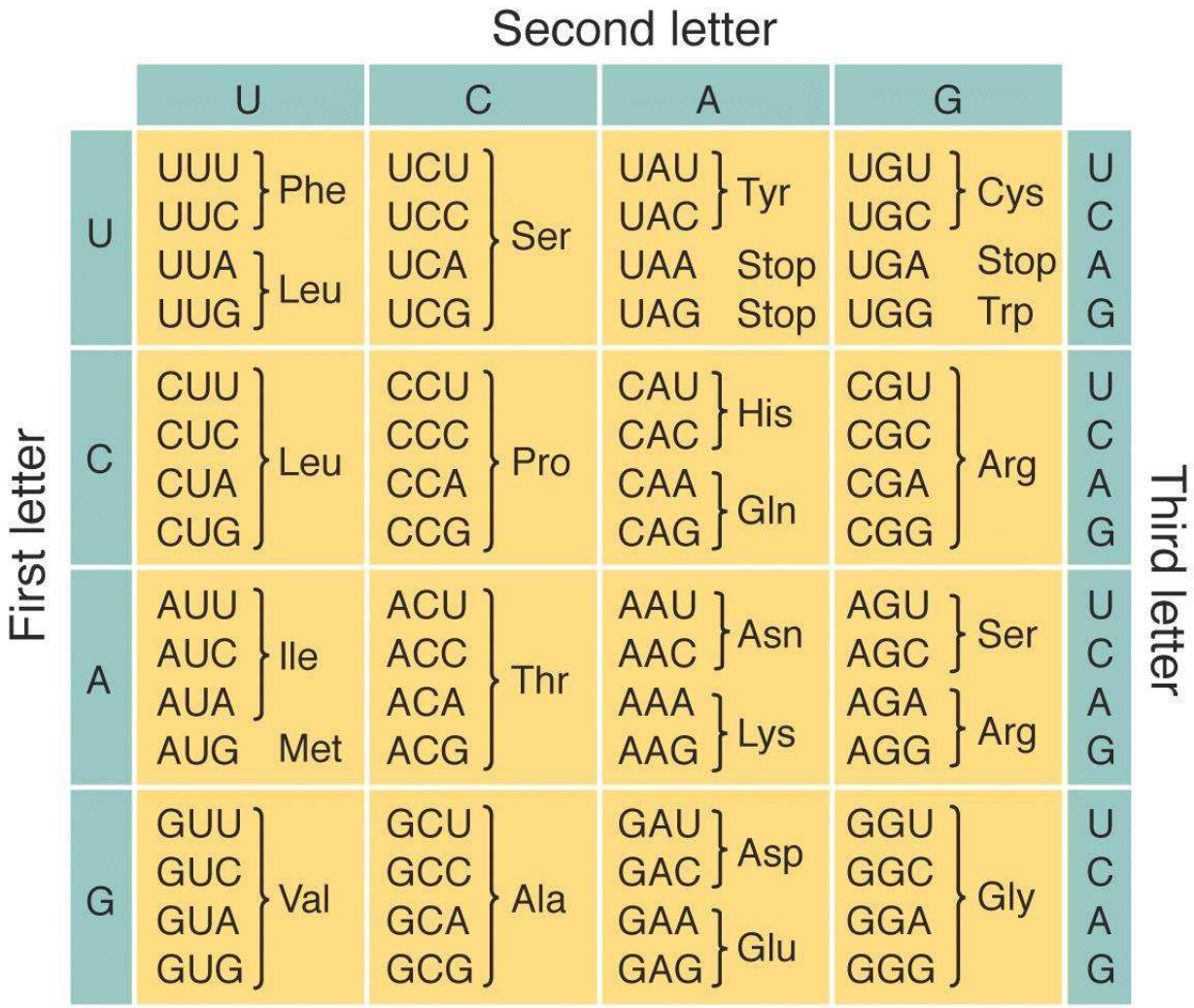
For the full tree, we get

$$Prob(A,C,C,C,G,x,y,z,w|T,Q) = \sum_{x}^{ACGT} \sum_{y}^{ACGT} \sum_{z}^{ACGT} \sum_{w}^{ACGT}$$

$$Prob(x)P_{xy}(t_6)P_{yA}(t_1)P_{yC}(t_2)$$
$$P_{xz}(t_8)P_{zC}(t_3)$$
$$P_{zw}(t_7)P_{wC}(t_4)P_{wG}(t_3)$$

Dynamic programming can speed up this calculation by some margin.

Codon data: from 4 to 61 states



Codon models

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \pi_j, & \text{for synonymous transversion} \\ \kappa\pi_j, & \text{for synonymous transition} \\ \omega\pi_j, & \text{for non-synonymous transversion} \\ \omega\kappa\pi_j, & \text{for non-synonymous transition} \end{cases}$$

where

- κ is the transition/transversion ratio
- π_j is the frequency of codon j
- ω measures selective pressure on amino acid

Q is 61 by 61 sparse matrix with null values known *a priori*.

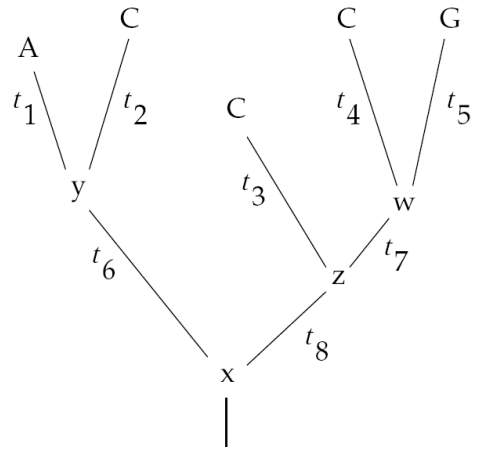
Problem: Q has to be exponentiated every time one of the $n-1$ branch length changes (n is nb of species).

Branch site codon model

More complex model

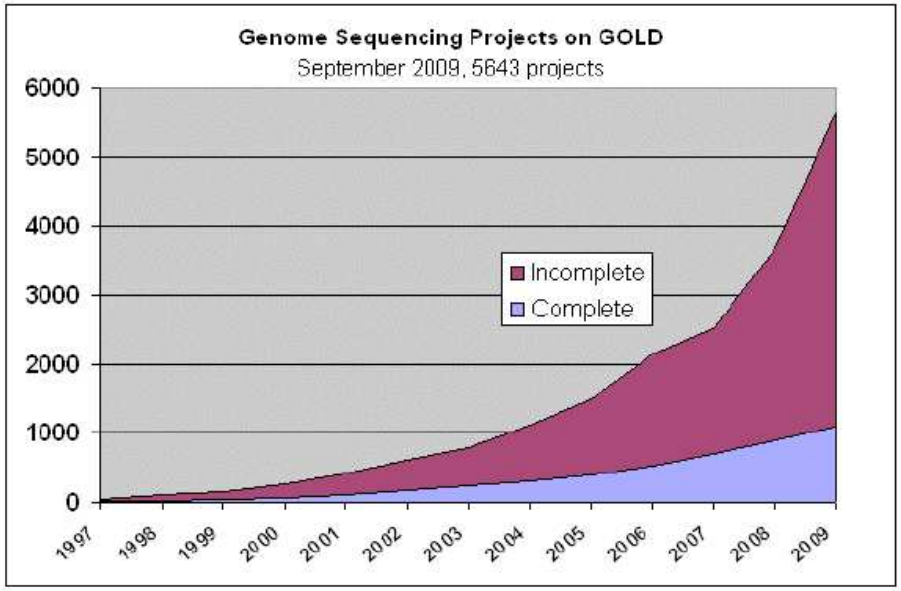
- foreground where positive selection occurs
- background where neutral or purifying selection occurs

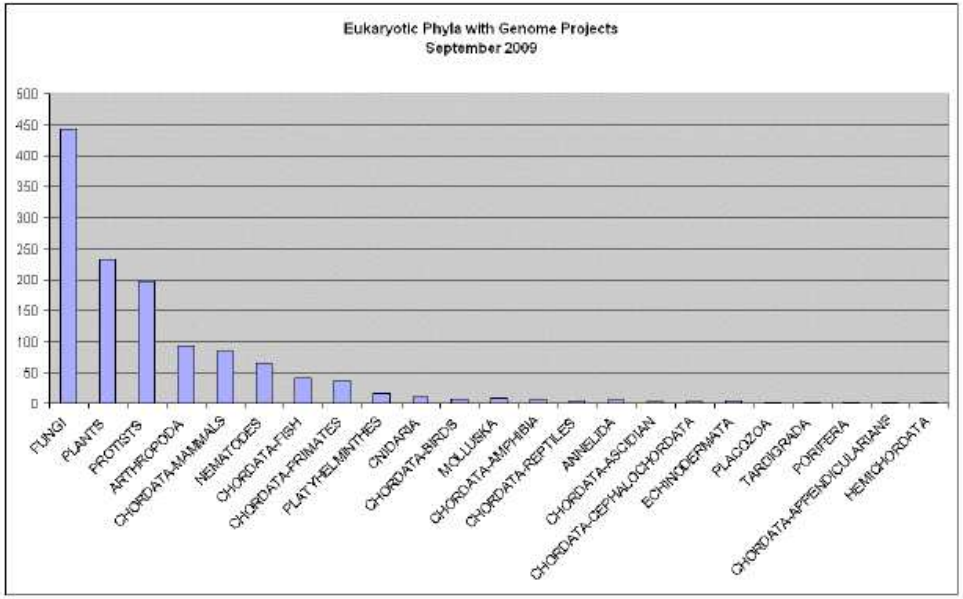
Class	Proportion	Background ω	Foreground ω
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$
2a	$\frac{(1-p_0-p_1)p_0}{(p_0+p_1)}$	$0 < \omega_0 < 1$	$\omega_2 > 1$
2b	$\frac{(1-p_0-p_1)p_1}{(p_0+p_1)}$	$\omega_1 = 1$	$\omega_2 > 1$



Compare this with a null model where ω_2 is fixed with Likelihood ratio test.

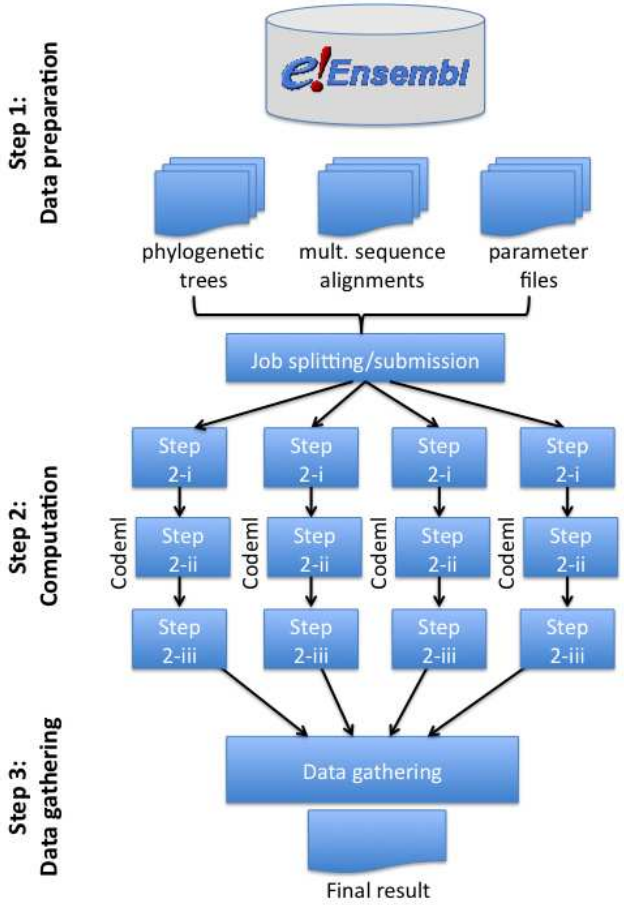
Limitations for Selectome





The size of the datasets to analyze will increase dramatically and exponentially.

Computational workflow

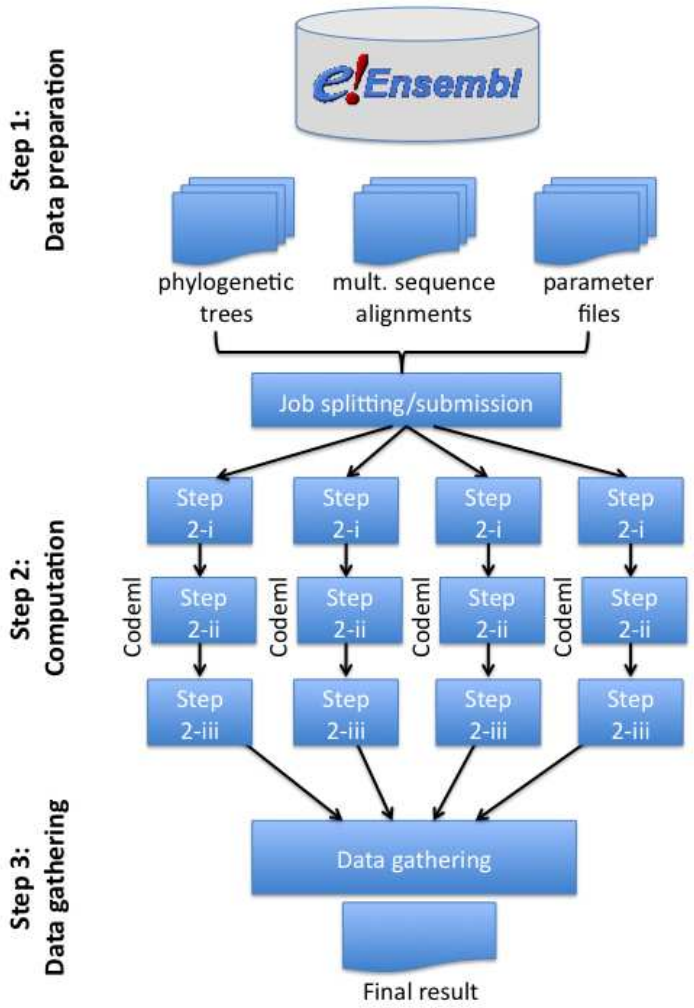


- for each gene (ca. 20,000 genes)
- for null and alternative models
  - for each site of a sequence (ca. 1000 on average)
    - calculate the likelihood at each node
    - $n - 1$ nodes for n species (ca. 30 species)
- repeat every 2-3 months when Ensembl is updated.

Impossible to compute Darwinian selection for all genes of interest with current implementation before the information is outdated.

We have a CPU bound problem in the estimation of the prevalence of selection in biology.

Streamlining calculations



- step 2 use completely independent data sets and is embarrassingly parallel computations
- further parallelization
  - steps 2-i and 2-ii have the potential to reuse the likelihood calculations done for one edge.
  - step 2-iii involves empirical Bayesian estimation of sites under selection, which can be optimized by setting appropriate prior distributions and reusing likelihood calculations efficiently from steps 2-i and 2-ii.

Costs: higher memory usage, and non independence between computations.

Code optimization

Actual code is made to analyse *n* species for one gene under a single model at a time. It was never made for genome wide studies

Using Newton-Raphson type of algorithms, it estimates

- codon model parameters (κ, different ω and p_i)
- branch lengths

It takes on average 20 minutes per job for ca. 30 species.
We want to analyse ca. 20,000 genes for data that is increasing exponentially.

Code optimization

Take also advantage of the optimization experience of the RAxML project

- SSE3-based optimization of the likelihood kernel
- exploitation of fine-grain loop-level parallelism using Pthreads and MPI (scales up to 8 or 16 cores on DNA models $O(4^2)$ vs $O(61^2)$)
- once model parameters are optimized, coarse-grained model of parallelism by
  - storing probability vectors
  - gather operation on the codon data distributed in a cyclic way to the threads during fined grained phase
- NUMA-specific data locality issues (e.g. impact of first touch policies) specifically for required memory of per-branch probability vector pairs assignment and accession

Groups involved

Marc Robinson-Rechavi (DEE-UNIL, SIB)
Sébastien Moretti
Walid Gharib

Nicolas Salamin (DEE-UNIL, SIB)
Maryam Zaheri

Heinz Stockinger (SIB)
Thierry Schuepbach (SIB)

Hannes Schabauer (HP2C)

Ziheng Yang (UCL)
Alexis Stamatakis (U. Tech. Munich)