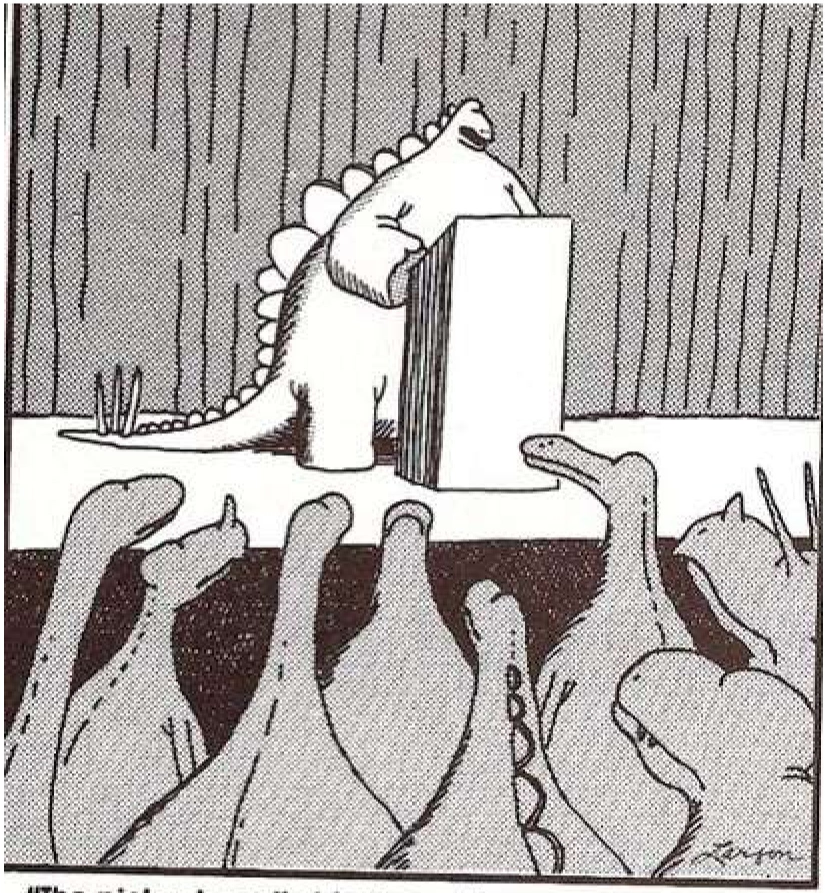
Molecular evolution

Pradipta Ray,

BIOL 6385 / BMEN 6389

The University of Texas at Dallas

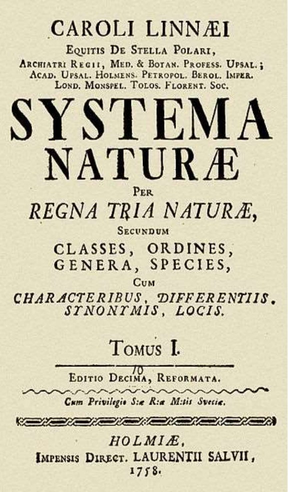
(some material based on content by PR in Eric Xing's 10-810 Carnegie Mellon class)



"The picture's pretty bleak, gentlemen. … The world's climates are changing, the mammals are taking over, and we all have a brain about the size of a walnut."

Far side, Larson

Brief early timeline



CAROLI LINNÆI
EQUITIS DE STELLA POLARI,
ARCHIATRI REGII, MED. & BOTAN. PROFESS. UPSAL.;
ACAD. UPSAL. HOLMENS. PETROPOL. BEROL. IMPER.
LOND. MONSPEL. TOLOS. FLORENT. SOC.

SYSTEMA
NATURÆ
PER
REGNA TRIA NATURÆ,
SECUNDUM
CLASSES, ORDINES,
GENERA, SPECIES,
CUM
CHARACTERIBUS, DIFFERENTIIS,
SYNONYMIS, LOCIS.

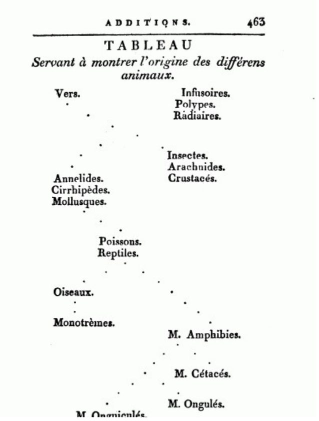
TOMUS I.

10
EDITIO DECIMA, REFORMATA.

Cum Privilegio S:æ R:æ M:tis Sveciæ.

HOLMIÆ,
IMPENSIS DIRECT. LAURENTII SALVII,
1758.

1735 : Linnaeus
Classification of living (and non-living) things

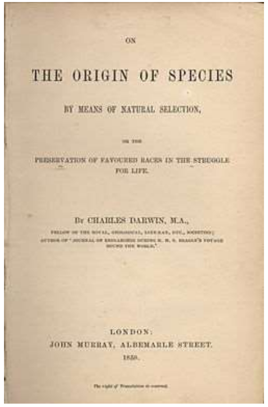


ADDITIONS. 463

TABLEAU
Servant à montrer l'origine des différens animaux.

Vers.
Infusoires.
Polypes.
Radiaires.
Insectes.
Arachnides.
Crustacés.
Annelides.
Cirrhipèdes.
Mollusques.
Poissons.
Reptiles.
Oiseaux.
Monotrèmes.
M. Amphibies.
M. Cétacés.
M. Ongulés.
M. Onguiculés.

1809: Lamarck
First theory of transmutation of species



ON
THE ORIGIN OF SPECIES
BY MEANS OF NATURAL SELECTION,
OR THE
PRESERVATION OF FAVOURED RACES IN THE STRUGGLE
FOR LIFE.

BY CHARLES DARWIN, M.A.,
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE
ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

The right of Translation is reserved.

1859: Darwin
Theory of evolution and natural selection

40 Vegmeaktes

Versuche
über
Pflanzen-Hybriden
von
Gregor Mendel
(Vorgelegt in den Sitzungen vom 8. Februar u. 8. März 1865)
Einleitende Bemerkungen

[illegible]

Courtesy of American Philosophical Society, Curt Stern Papers. Noncommercial, educational use only.

1865: Mendel
Laws of Inheritance ,
rediscovered 1900

Saltationism

Speciation is the result of abrupt large genetic changes

Biometric school

Continuous genetic variation underlies continuous phenome

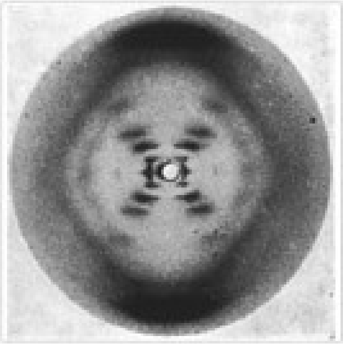
Later chronological developments

- George Nuttal mixed sera and antisera from different species to determine “blood relationships”:
  - More closely related species exhibit stronger cross-reactions between sera and antisera

- Morgan and fruit flies
  - Chromosomes, laws of heredity and trait propagation, recombination and cross over

Double helix

- In 1953, James Watson and Francis Crick proposed the double-helix model of DNA structure
  - Based on X ray diffraction performed by Rosalind Franklin
- Mechanism of genetic transfer revealed



wikipedia.org

Human evolution

- Humans were thought to be monophylletic, and only distantly related to the great apes

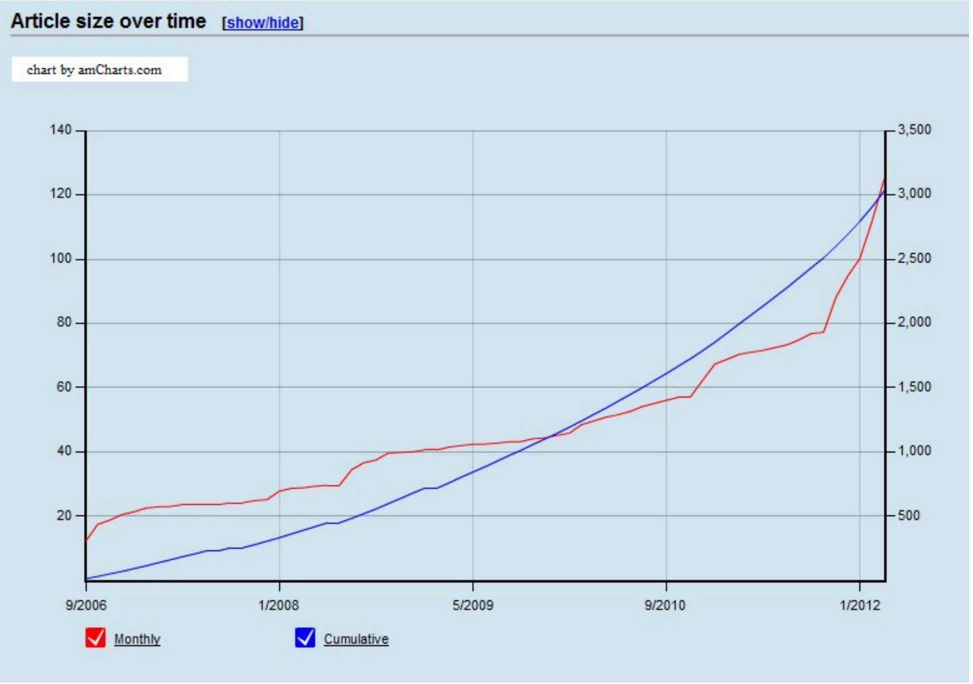
- Sarich & Wilson (1967) cross reacted serum albumin between primates
  - Humans, gorilla and chimpanzee were genetically equidistant from the orang-utan

Sequencing explosion

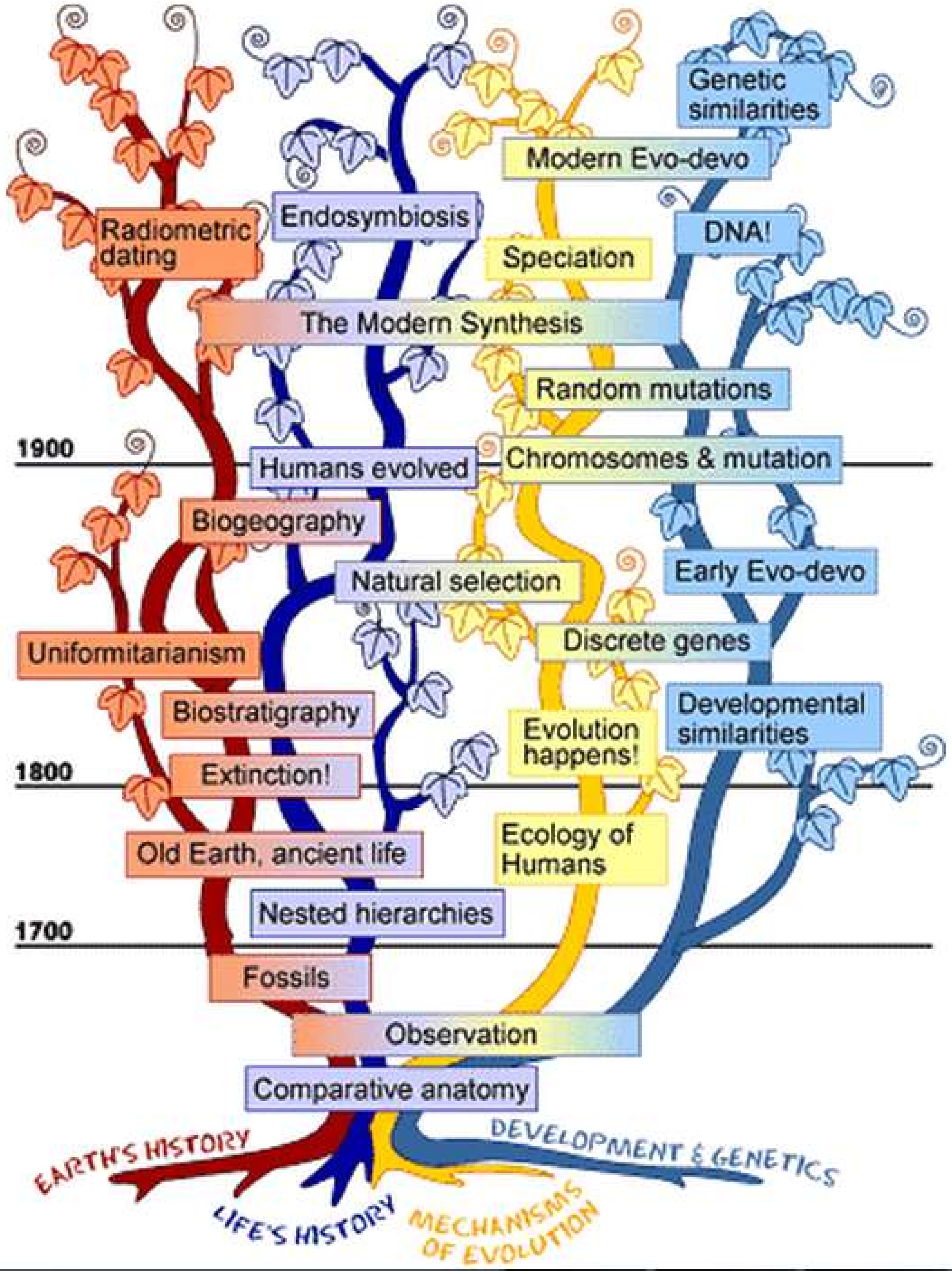
- Real “explosion” of information on molecular evolution since the advent of PCR: (1983)
  - Nucleotide could now be sequenced based on PCR → cloning → chromatography / die based sequencing
- Can sequence DNA from samples thousands of years old (ancient DNA analysis : Neanderthal and Woolly Mammoth genome)

No of sequenced genomes

- Wikipedia article size of "List of sequenced eukaryotic genomes"
  - Not a perfect correlation, but still …



wikipedia.org



Genetic similarities

Modern Evo-devo

Radiometric dating

Endosymbiosis

DNA!

Speciation

The Modern Synthesis

Random mutations

1900

Humans evolved

Chromosomes & mutation

Biogeography

Natural selection

Early Evo-devo

Discrete genes

Uniformitarianism

Biostratigraphy

Developmental similarities

Evolution happens!

1800

Extinction!

Old Earth, ancient life

Ecology of Humans

Nested hierarchies

1700

Fossils

Observation

Comparative anatomy

EARTH'S HISTORY

LIFE'S HISTORY

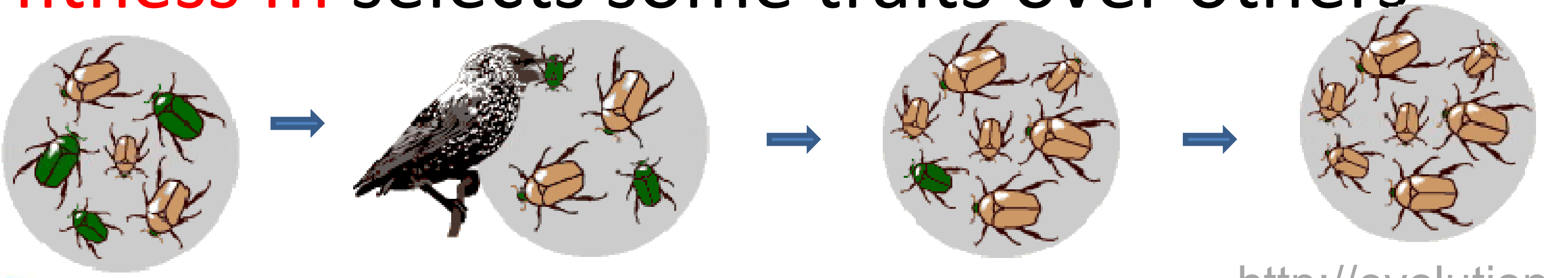
MECHANISMS OF EVOLUTION

DEVELOPMENT & GENETICS

http://evolution.berkeley.edu/evosite/

Natural selection

- Small discrete genetic changes causes organisms to be different at the individual level
- Natural selection : Some changes are more important for survival or lineage propagation based on environmental and other factors : fitness fn selects some traits over others

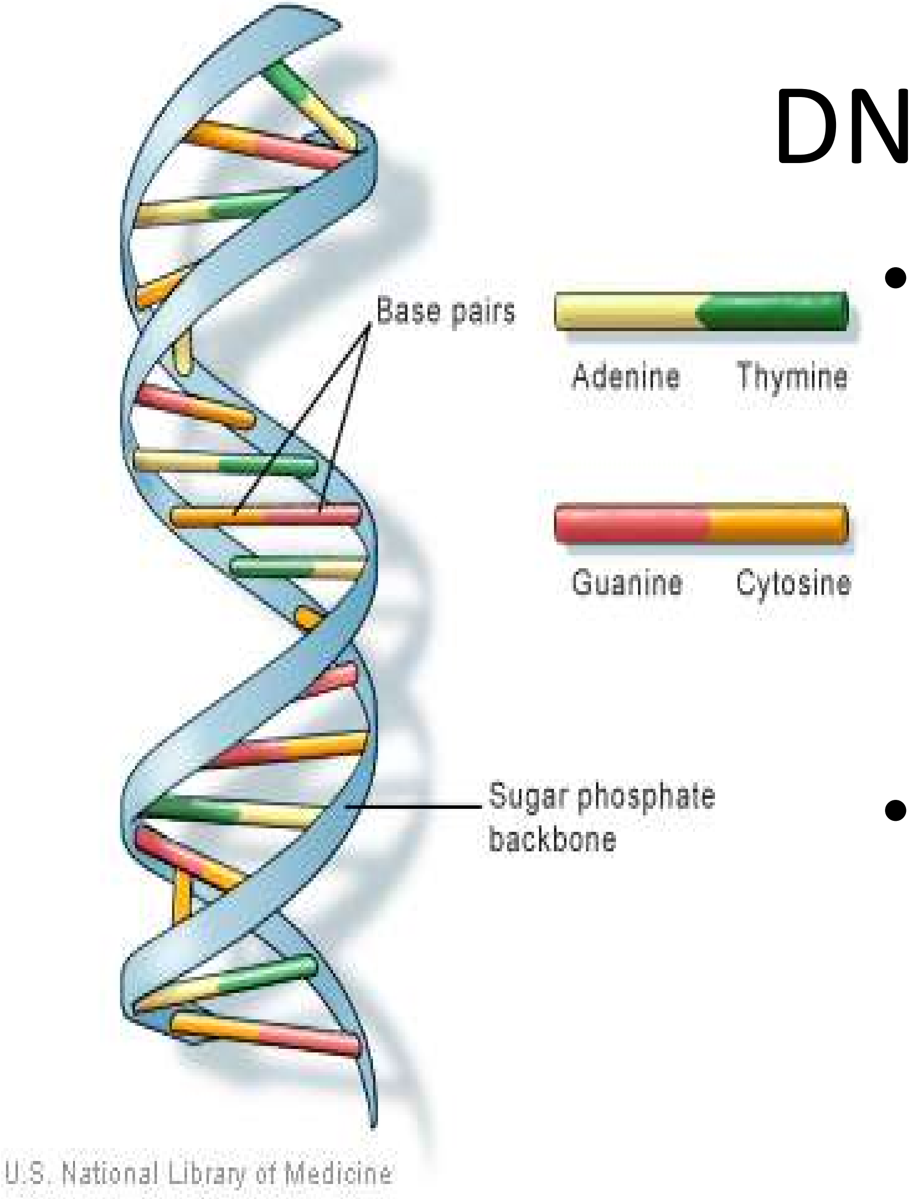


http://evolution.berkeley.edu/evosite/

Speciation

- Differences in accumulated genetic changes in sub-populations can cause them to become reproductively isolated : causing speciation
- Can be influenced by different kinds of environmental factors
  - physical isolation of populations due to geological events
  - quickly changing environment (eg extinction of another species) changing the nature of selectional forces
  - faster mutation rate due to positive selection or environmental factors like radiation

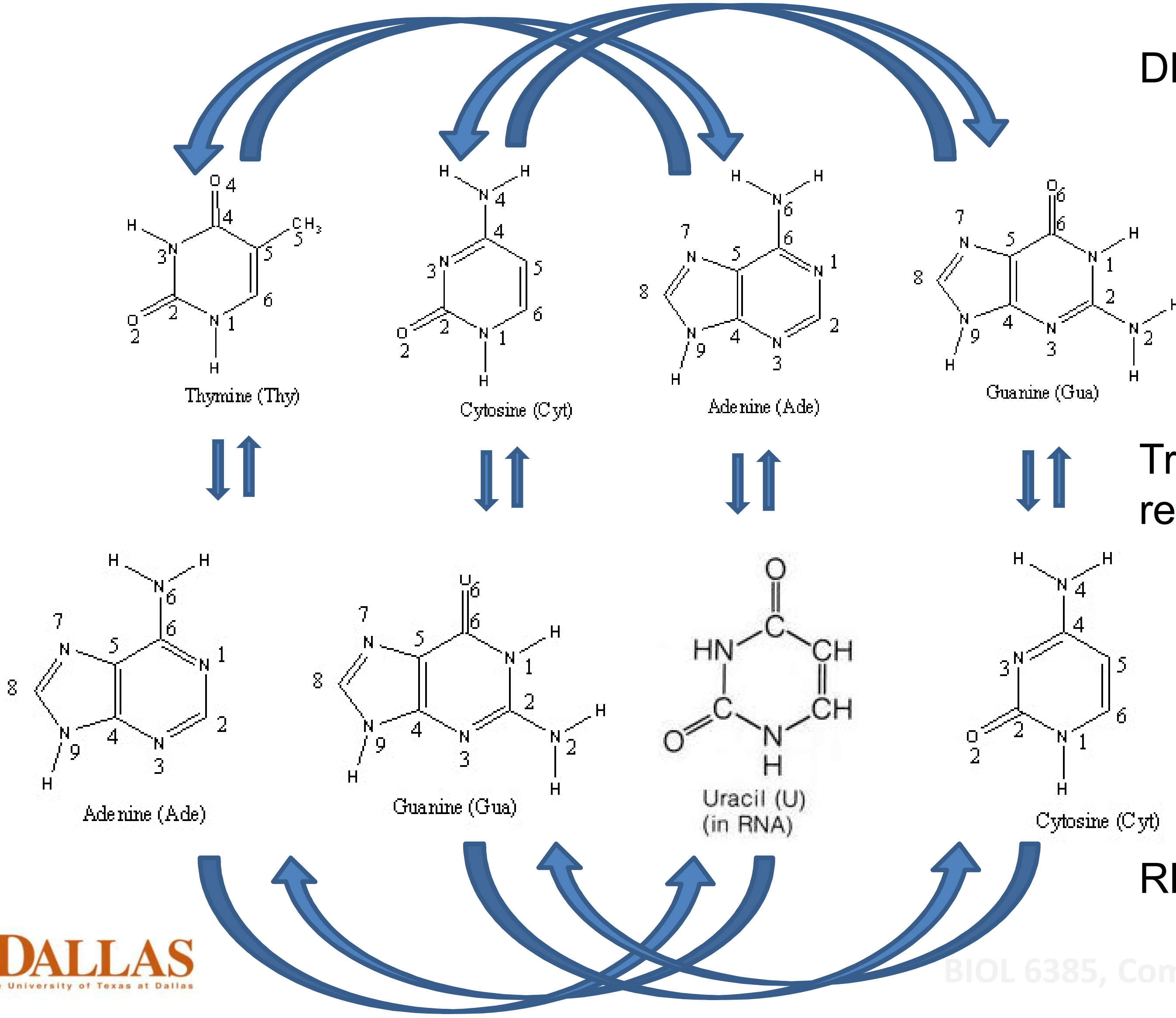
DNA



U.S. National Library of Medicine

- Genetic material arranged in several double stranded chromosomes in the nucleus of each cell
- Combined genetic material is called the genome

Component nucleic acids



DNA complement

Transcription /
rev transcription

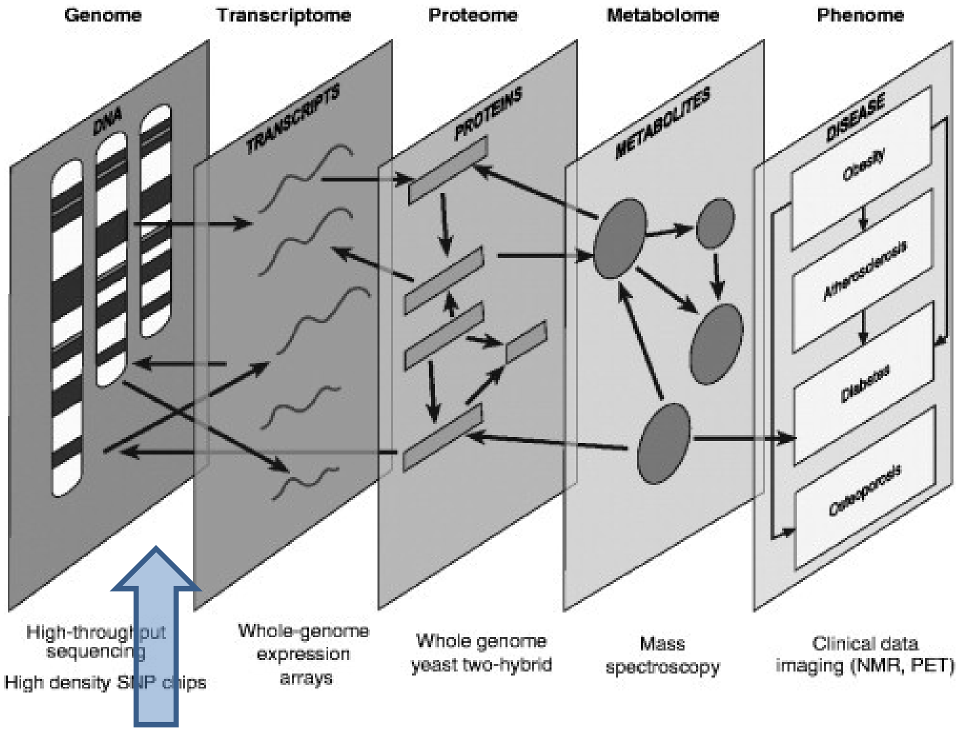
RNA complement

-“Ome”s

Stability increases across environment, condition, cell type, organism, etc

Pattern breaking

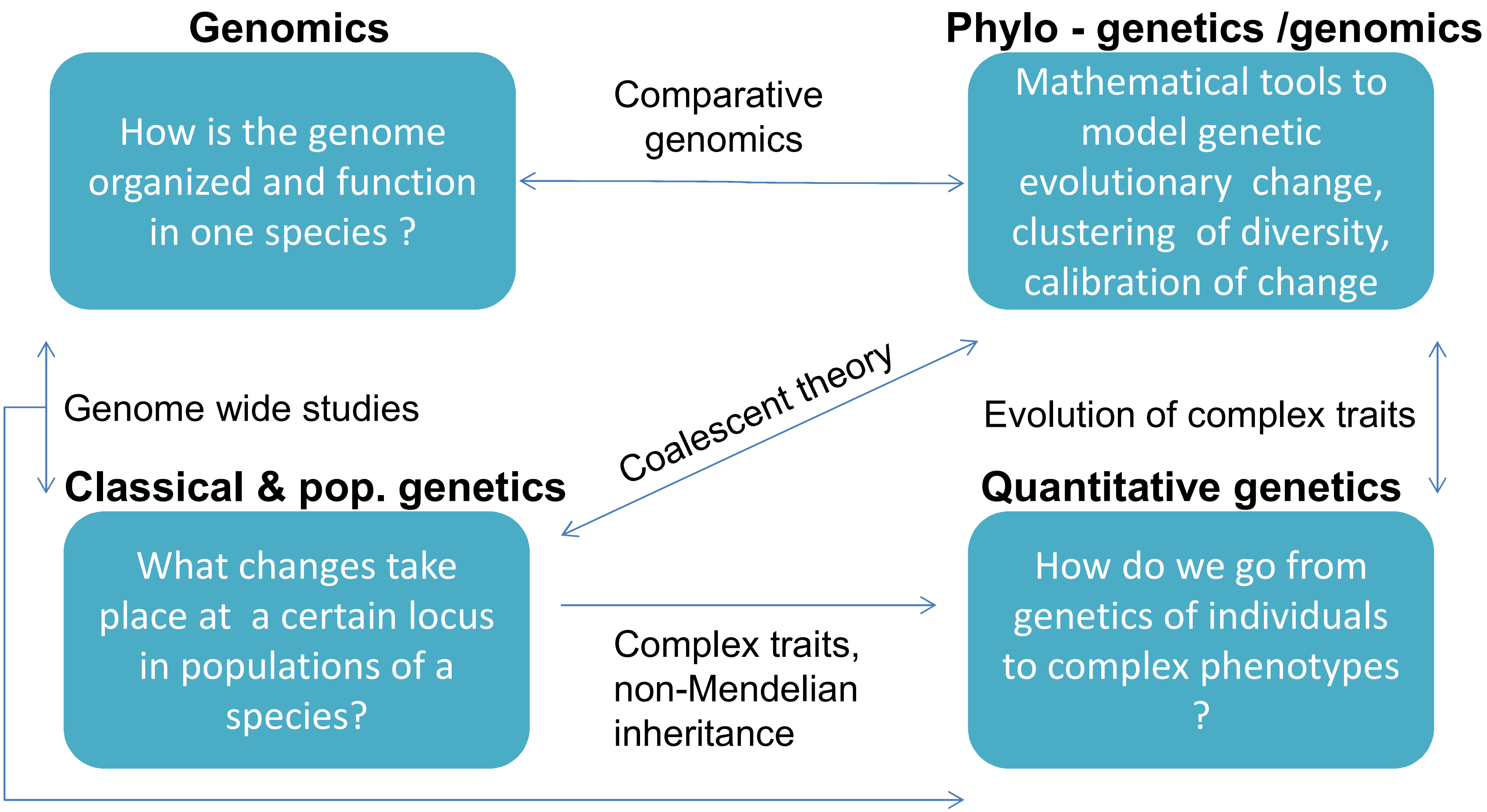
- Study of evolution
  - how the genome changes over generations & species
  - how such changes affect successive “ome”s



Another layer : **epigenome** : inherited traits which cannot be fully explained by the genome

Farber & Lusis, Adv in Genetics, Vol 60

Broad fields of study



Genomics

How is the genome organized and function in one species ?

Comparative genomics

Phylo - genetics /genomics

Mathematical tools to model genetic evolutionary change, clustering of diversity, calibration of change

Genome wide studies

Coalescent theory

Evolution of complex traits

Classical & pop. genetics

What changes take place at a certain locus in populations of a species?

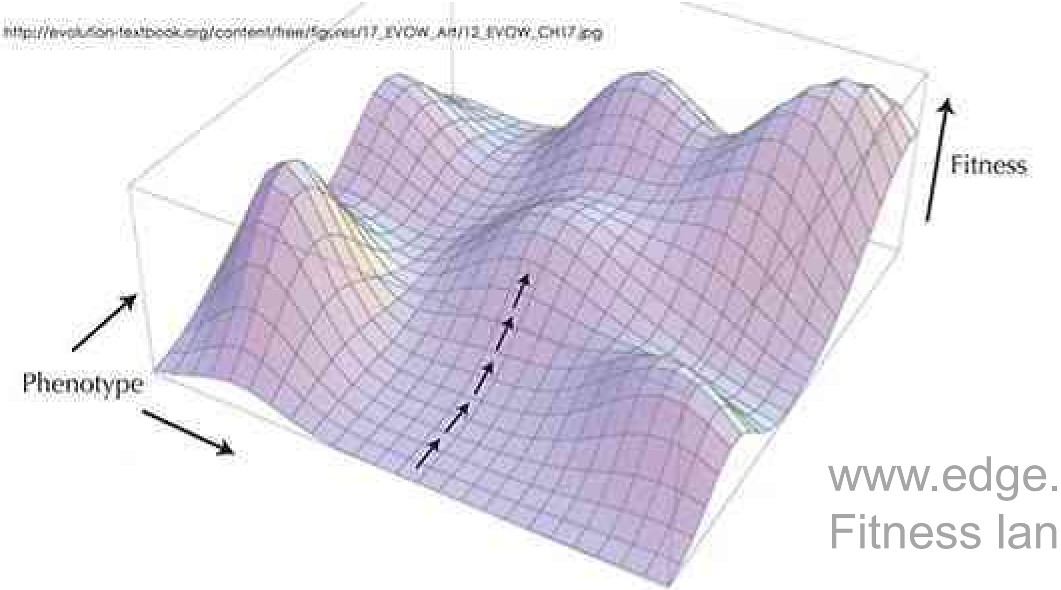
Complex traits, non-Mendelian inheritance

Quantitative genetics

How do we go from genetics of individuals to complex phenotypes ?

Evolution as an optimization process

- Gaming the “fitness function”
- Risks of over playing the system
- Reversibility of evolution ?



http://evolution-textbook.org/content/free/figures/17_EVOW_Art/12_EVOW_CH17.jpg

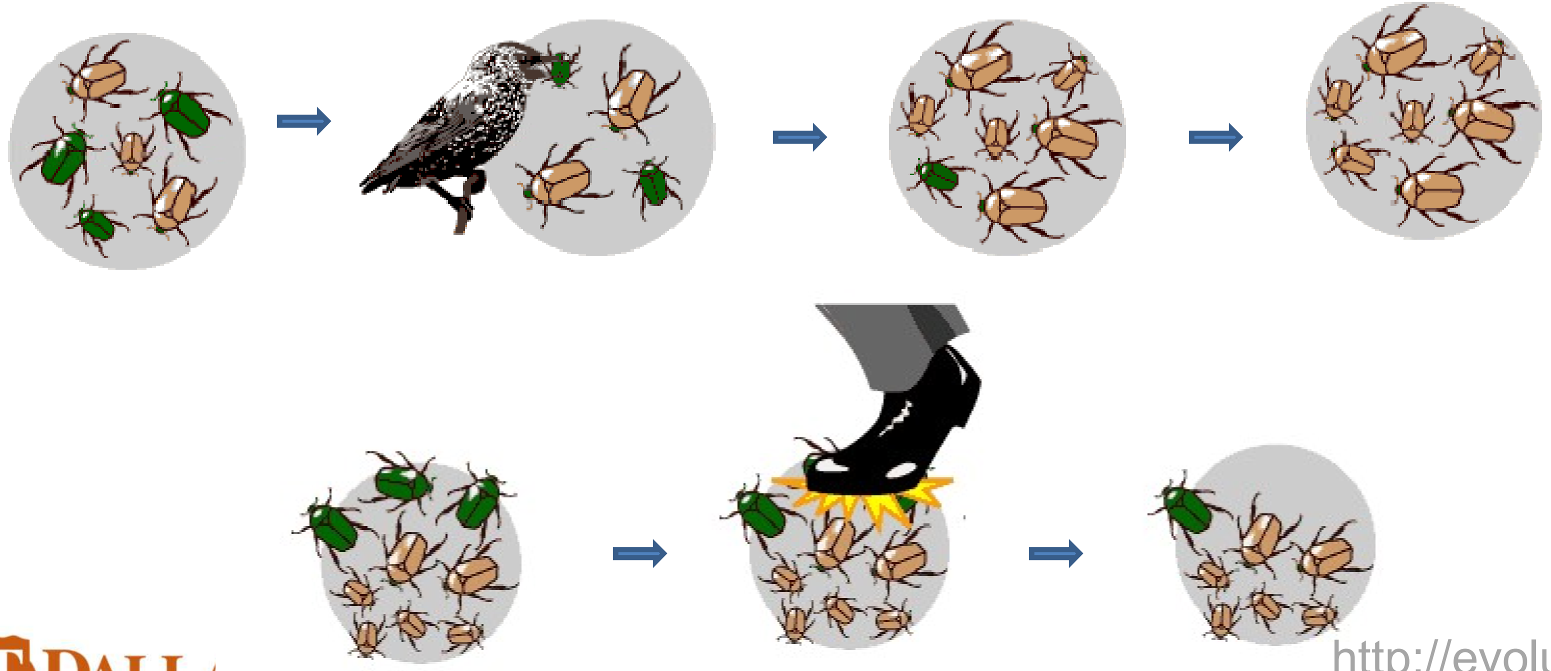
www.edge.org
Fitness landscapes, S. Brand

Phylogenetics

- How single nucleotides and other genomic entities change over time
  - Substitution matrices
- Cluster a group of genes or organisms based on their similarity to each other [alignment answers a related question]
- Analyze the nature of such changes
- Calibrate the rate of change

Evolution as a stochastic process

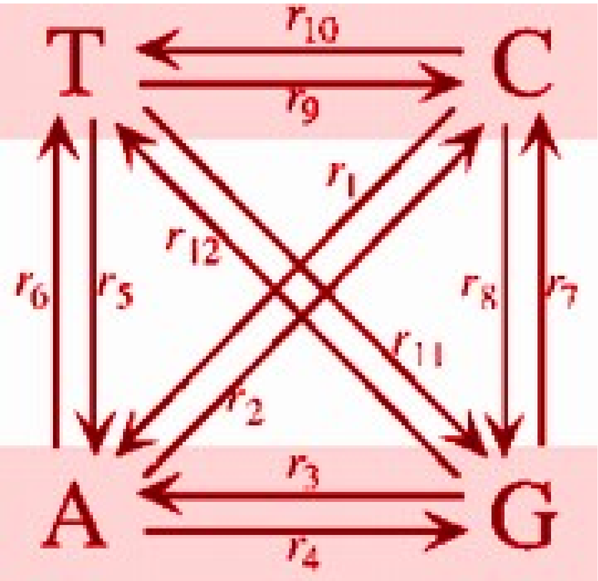
- Forces of optimization (selection) compete with completely random forces to shape our genomes



http://evolution.berkeley.edu/evosite/

Substitution

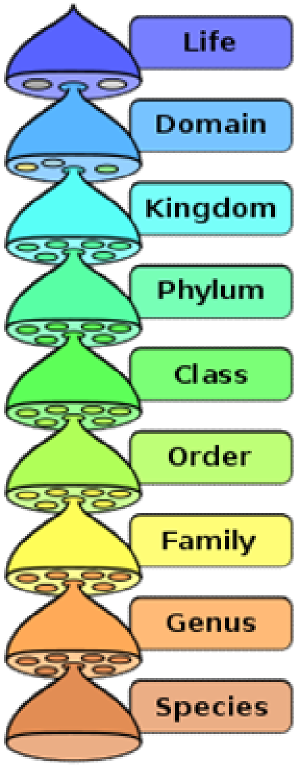
- What are the rates at which nucleotides / AAs / codons change into each other ?
- Can we calculate the probability of an A turning into a G over a time period of t ?
- What kind of assumptions can we make about such stochastic processes ?



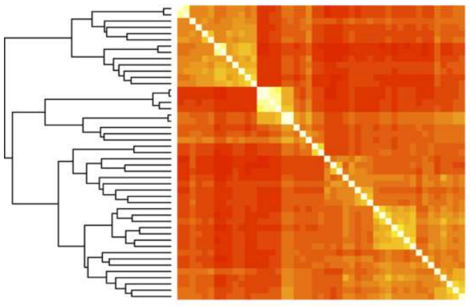
library.caltech.edu

Systematics

- Cladistics / taxonomy : do organisms / genomic entities (like duplicated genes) grouped together based on genomic similarity reflect shared evolutionary history ?
- How to build dendrograms based on pairwise (or otherwise) differences ?



wikipedia.org



Saw et al, Stand Gen Sci 6:1

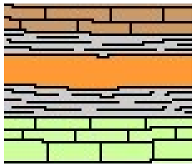
Nature of genomic changes

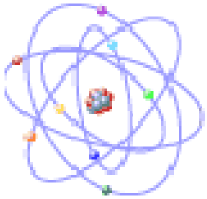
- Are the changes just random (neutral) ? Are they based on selectional forces acting on the genome ? How to quantify ?

- Neutral Theory (Kimura) : Vast majority of changes are neutral

Calibration of genomic changes

- Controversial assumption in evolutionary theory
  - Mutations (typically mostly neutral ones) in some genomic sequences and proteins take place at regular clock-like intervals
  - Can be calibrated against fossil record : using stratigraphy, radiocarbon dating, molecular clock





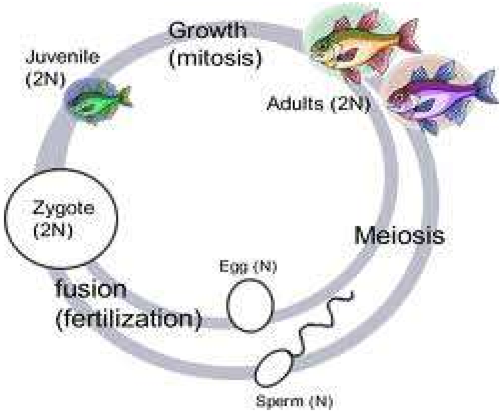
http://evolution.berkeley.edu/evosite/

Asexual reproduction

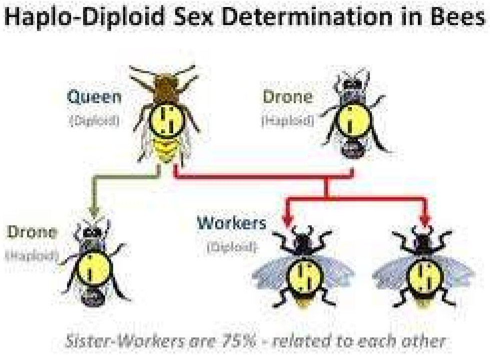
- Children are clones of parents
- Genetic diversity
  - errors during cloning (mutation)
  - lateral gene transfer
    - conjugation – direct transfer of genetic material between individuals
    - transformation – uptake of exogenous DNA
    - transduction – transfer of genetic material between individuals through 3rd party (like virus)

Sexual reproduction

- Individual has 2 copies of each chromosome : one from each parent (homologous chr)
- 2 genders : haploid, diploid and ploidy reduction
- Other complicated mechanisms



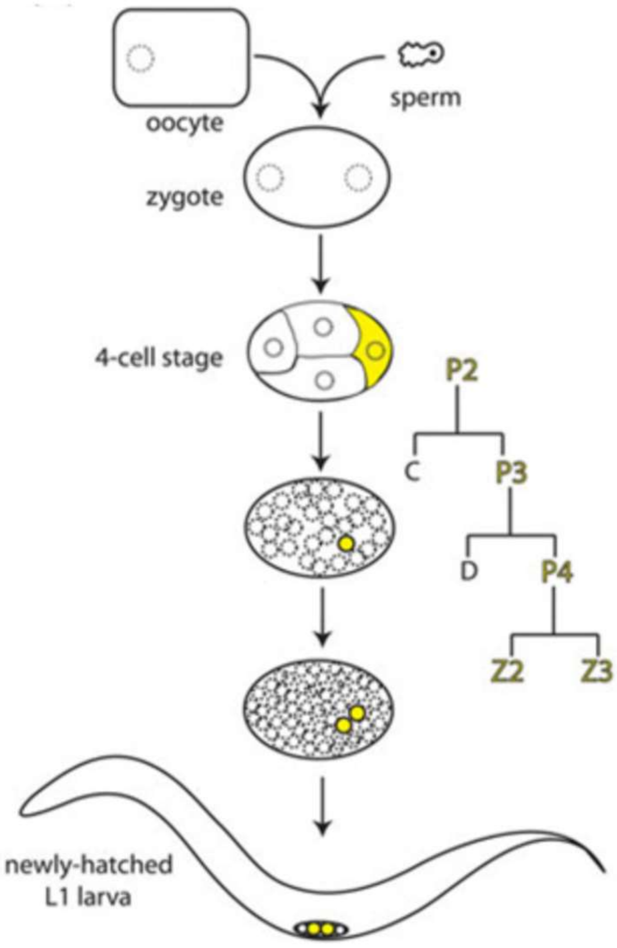
biologycorner.com


Haplo-Diploid Sex Determination in Bees
Queen
(Diploid)
Drone
(Haploid)
Drone
(Haploid)
Workers
(Diploid)
Sister-Workers are 75% - related to each other

ccsbio

Germ line and soma

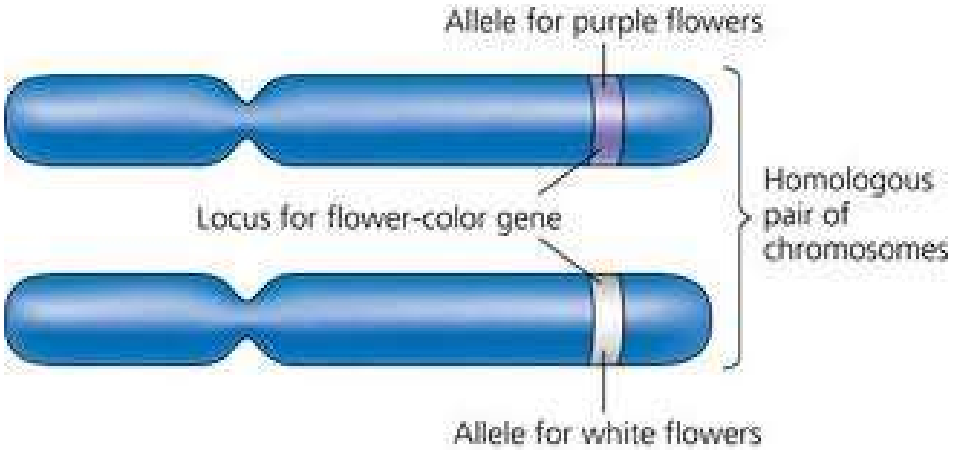
- DNA needed for homeostasis, metabolism, producing offspring
- Composition of DNA can change : changes to DNA in the **germ line** are transmitted to offspring
- Non-germ line evolution : evolution of cancer



wormbook.org

Alleles

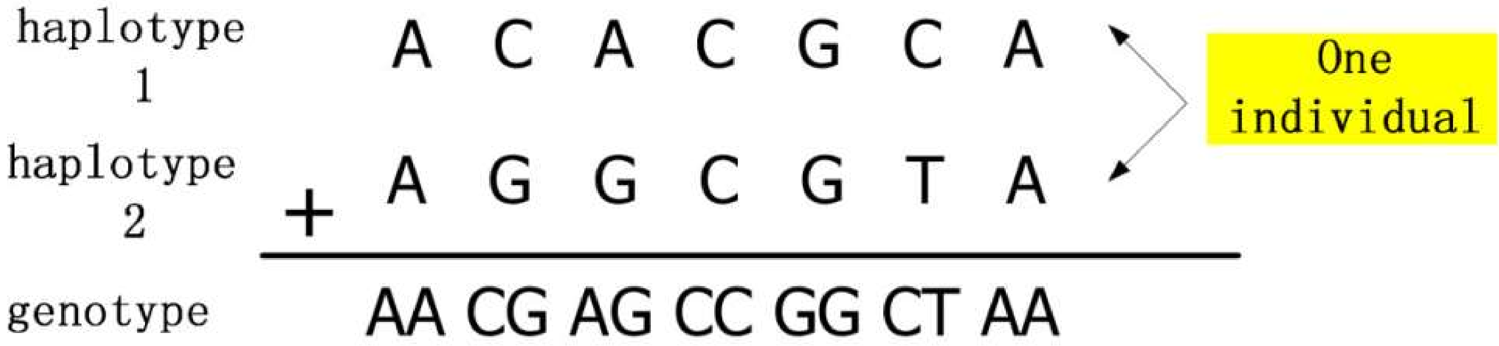
- One of many variants of a genetic locus
- Organism, wrt an allele :
  - hemizygous : only one copy of chromosome
  - homozygous : both copies have same allele
  - heterozygous : copies have different alleles


Allele for purple flowers
Locus for flower-color gene
Homologous pair of chromosomes
Allele for white flowers

Rozaini Othman

Haplotype vs genotype

- When we know the allelic composition of multiple alleles in an individual, can we partially reconstruct the chromosomes ?



haplotype 1 A C A C G C A

haplotype 2 + A G G C G T A

genotype AA CG AG CC GG CT AA

Zhou and Wang *BMC Bioinformatics* 2007 **8**:484

Added aspects of sexual reproduction

- Sexual selection : gender specific selective forces on top of existing environmental selective forces (co – evolution)
- Sex determination : Sex chromosome

	Male	Female
Sex chromosome (pair config)	XY	XX
Sex chromosome (pair config)	WW	ZW
Haplodiploidy (total no of chr)	N	2N

Changing nucleotide composition

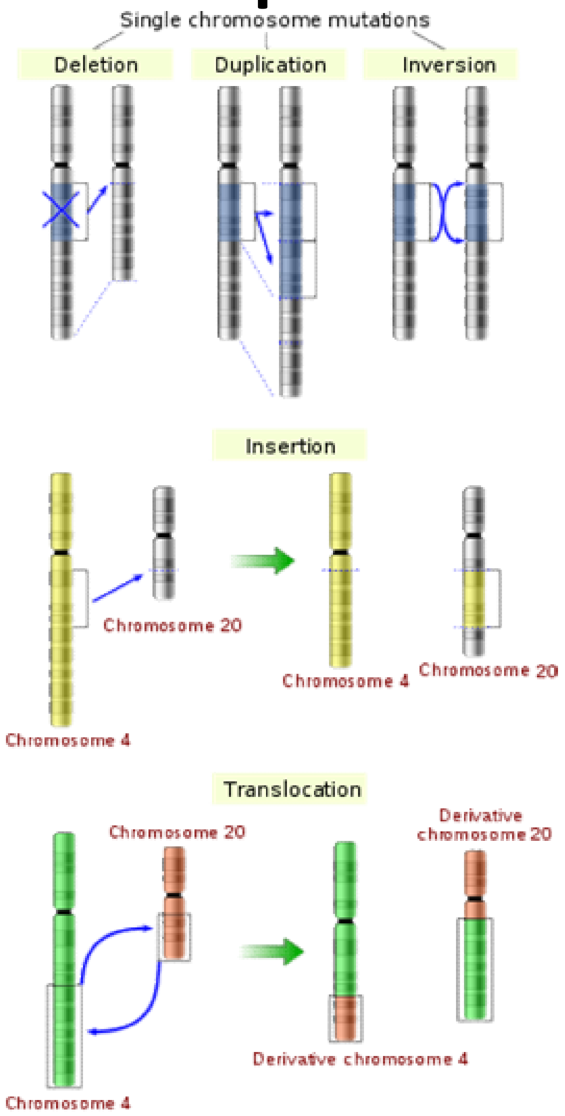
Point mutation

Translocation

Insertion

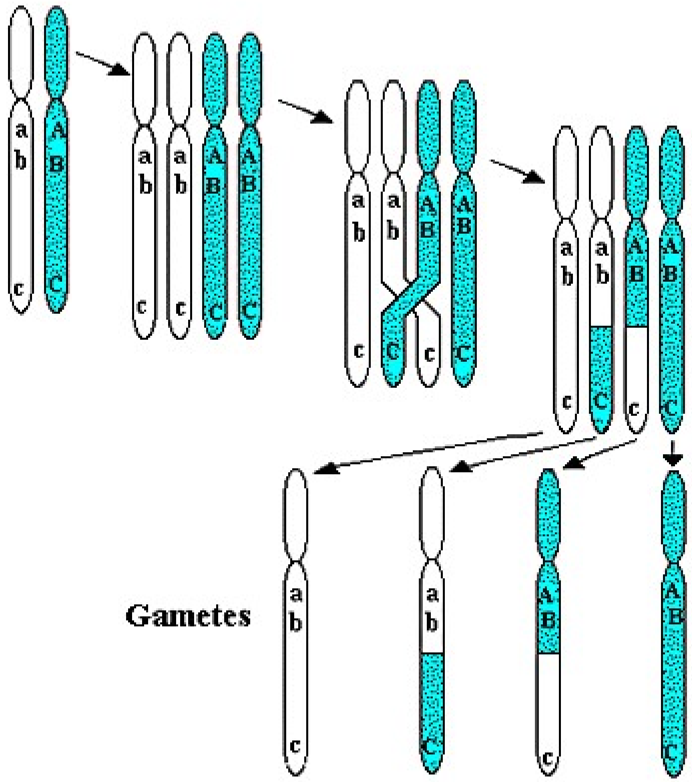
Duplication

Deletion



wikipedia.org

Recombination



Crossing-over and recombination during meiosis

Okay, lets get to the math !

Substitution models

- At the simplest level, we study how a single nucleotide changes over time
- We build genome wide models of evolution in a bottom – up manner based on this.
- Alternatively, directly model evolution of higher granularity genomic units (like codons).

Stochastic process

- Formulation : set of indexed random variables

$$\{ F_{X_t}(x) \text{ or } F(x,t) \mid t \in T \}$$

- Categories :

Examples of SPs :	Continuous X_t	Discrete X_t
Continuous t		
Discrete t		

Stochastic process : what

- Notion of how a RV "evolves"
  - T may not be time, it may be complicated : like t = (x, y)

- Why isnt t just a parameter in the RV ?

$$F_X(x) \leftarrow g(x, \theta)$$

$$F_{X_t}(x) \leftarrow g(x, t, \theta)$$

Stationarity & homogeneity

$$F_{X_{t_1}, \ldots, X_{t_k}}(x_{t_1}, \ldots, x_{t_k})$$

homogenous

$$= F_{X_{t_1+\tau}, \ldots, X_{t_k+\tau}}(x_{t_1}, \ldots, x_{t_k})$$

$$P_{X_0, X_{12}, X_{15}}(A, G, G) = P_{X_{30}, X_{42}, X_{45}}(A, G, G)$$

stationary

(discrete valued)

$$F_{X_{t+s} - X_s} = F_{X_t}$$

$$P(X_{10} = A \mid X_5 = G) = P(X_{30} = A \mid X_{25} = G)$$

Is $\wp$ a fn of t ?

But evolutionary parameters change with time !

- Selectional forces change with time for example

- Piecewise homogenous and stationary processes are still possible ! (over short evolutionary time)

Continuous time Markov Process

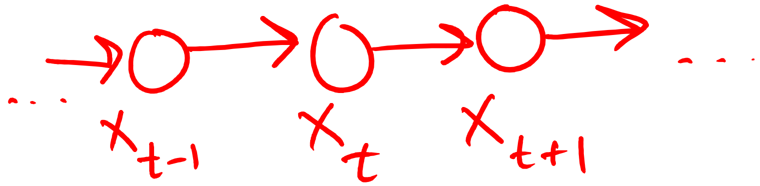
- Markov Chains and Continuous Time Markov Processes are both Markovian
  - Future is conditionally independent of the past, given the present

$$P(X_{t_{k+1}} = x_{t_{k+1}} \mid X_{t_k} = x_{t_k}, X_{t_{k-1}} = x_{t_{k-1}}, \ldots, X_{t_1} = x_{t_1})$$

$$= P(X_{t_{k+1}} = x_{t_{k+1}} \mid X_{t_k} = x_{t_k})$$

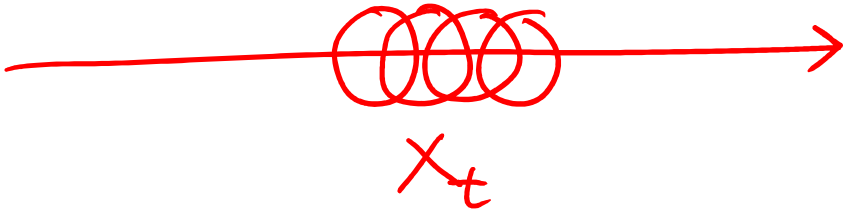
$$[\text{for } t_1 < t_2 < t_3 < \ldots < t_{k-1} < t_k < t_{k+1}]$$

MC



- Finite or countably infinite index
- Discrete valued
- Markovian

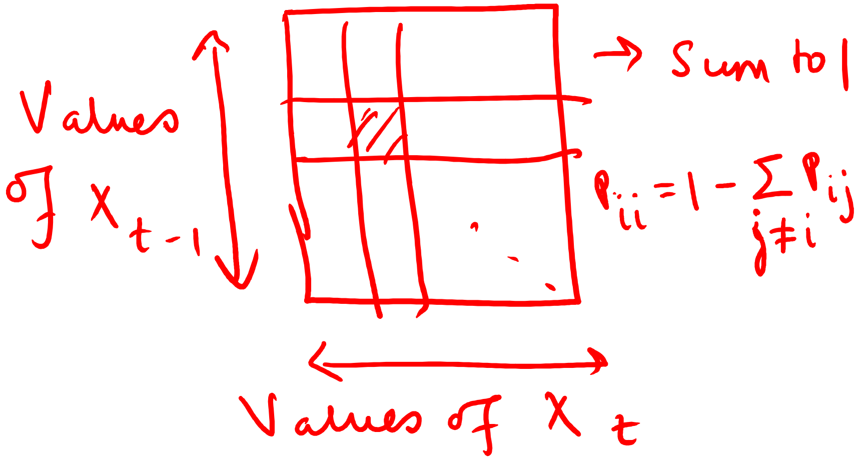
CTMP



- Uncountably infinite index
- Discrete valued
- Markovian

MC

- Parameterization



$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

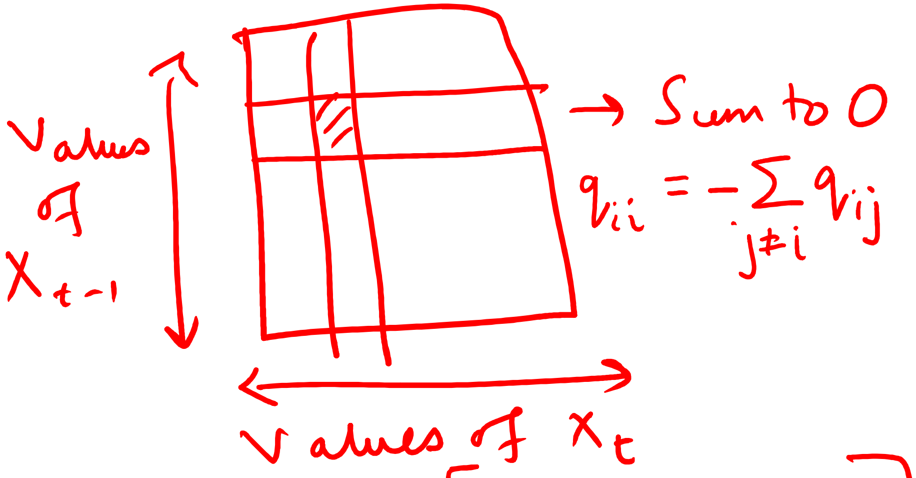
$$p_{ij} = P(X_t = j \mid X_{t-1} = i)$$

TRANSITION MATRIX

TRANSITION PROB. MATRIX

CTMP

- Parameterization

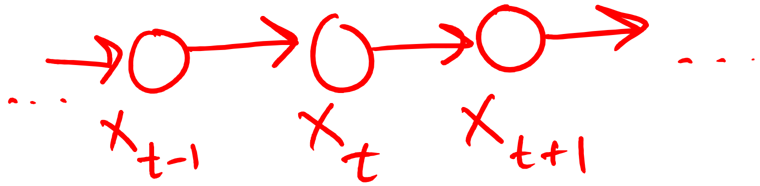


$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

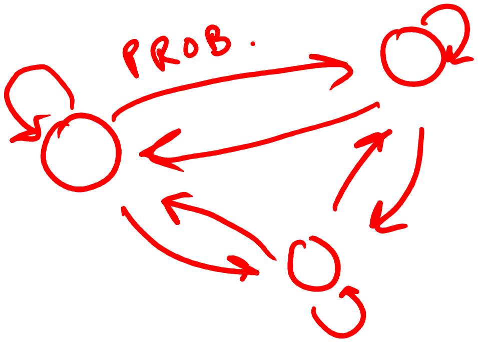
$$q_{ij} = \lim_{h \to 0} \left[\frac{P(X_{t+h} = j \mid X_t = i)}{h} \right]$$

INSTANTANEOUS RATE MATRIX

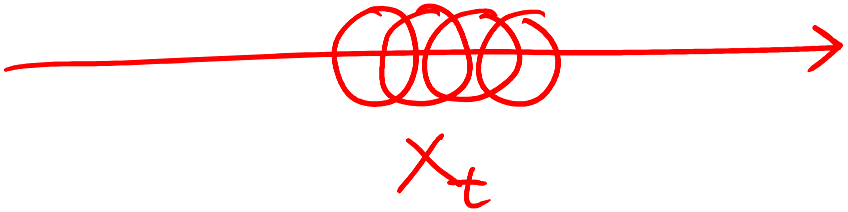
MC



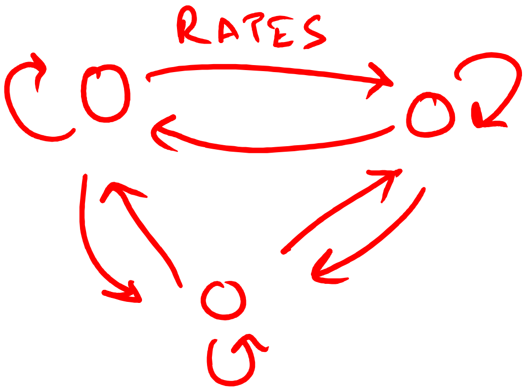
- State space diagram



CTMP

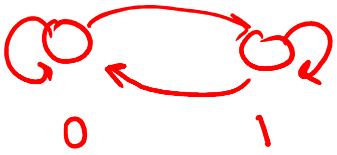


- State space diagram



HOW TO GET PROBABILITIES BY FIXING TIME?

Rates to probabilities



$$\mathcal{Q} = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix} \qquad P(t) = \begin{bmatrix} p_{00}(t) & p_{01}(t) \\ p_{10}(t) & p_{11}(t) \end{bmatrix}$$

$$G = P_t \cdot \mathcal{Q} = \begin{bmatrix} p_{00}(t) & p_{01}(t) \\ p_{10}(t) & p_{11}(t) \end{bmatrix} \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix}$$

$$= \begin{bmatrix} p_{00}(t)q_{00} + p_{01}(t)q_{10} & p_{00}(t)q_{01} + p_{01}(t)q_{11} \\ p_{10}(t)q_{00} + p_{11}(t)q_{10} & p_{10}(t)q_{01} + p_{11}(t)q_{11} \end{bmatrix}$$

$$G_{0,1} = \lim_{h \to 0} \frac{p_{00}(t)\,[\,P(X_{t+h}=1 \mid X_t=0)\,]}{h} + \lim_{h\to 0} \frac{p_{01}(t)\,[\,p(X_{t+h}=1 \mid X_t=1)\,]}{h}$$

$= \text{Rate of change of } P(0 \to 1)$

$$P_t \cdot \mathcal{Q} = P'_t$$

Formulation

$$P'(t) = P(t) \cdot g$$

Let, $P(t) = e^{g \cdot t}$

$$P'(t) = g \cdot e^{g \cdot t} = g P(t)$$

$$P(t) = c \cdot e^{gt}$$

← BOUNDARY COND.

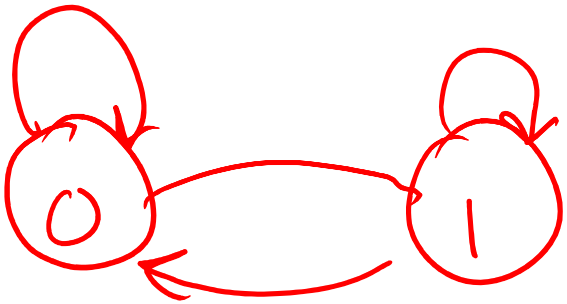
Calculating P(t)

$$P(t) = e^{\mathcal{Q} \cdot t}$$

$$= I + \frac{\mathcal{Q}\, t}{1!} + \frac{\mathcal{Q}^2 t^2}{2!} + \ldots$$

How about a closed form soln?

[Why do we care]



$$Q = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & -\mu & \mu \\ 1 & \mu & -\mu \end{array}$$

$$\begin{bmatrix} P'_{00}(t) & P'_{01}(t) \\ P'_{10}(t) & P'_{11}(t) \end{bmatrix} = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix} \begin{bmatrix} -\mu & \mu \\ \mu & -\mu \end{bmatrix}$$

$$P'_{01}(t) = P_{00}(t)\cdot\mu + (-\mu)P_{01}(t)$$

$$= (1 - P_{01}(t))\cdot\mu$$

$$+ (-\mu)\,P_{01}(t)$$

$\longleftarrow$ OBTAIN BY SOLVING SIMULTANEOUS EQN.

$$P'_{01}(t) + 2\mu P_{01}(t) = \mu$$

$$\boxed{f'(x) + p(t)f(x) = q(t)}$$

INTEGRATING FACTOR $\quad u(t) = e^{\int p(t)\,dt} = e^{2\mu t}$

$$P_{01}(t) = \frac{\int u(t)\,q(t)\,dt + c}{u(t)}$$

$$= \frac{\int e^{2\mu t}\,\mu\,dt + C}{e^{2\mu t}}$$

$$= \frac{\frac{1}{2}\int e^{2\mu t}\,d(2\mu t) + C}{e^{2\mu t}}$$

$$= \left(\frac{1}{2} e^{2\mu t} + c\right) / e^{2\mu t}$$

Boundary condition

$t = 0, \quad P_{01}(t) = 0$

$0 = \frac{1}{2} + c \cdot 1$

$c = -1/2$

How about $t = \infty$?

Solve all but one and use normalization fact

$$P_{01}(t) = \frac{1}{2} - \frac{1}{2} e^{-2\mu t}$$

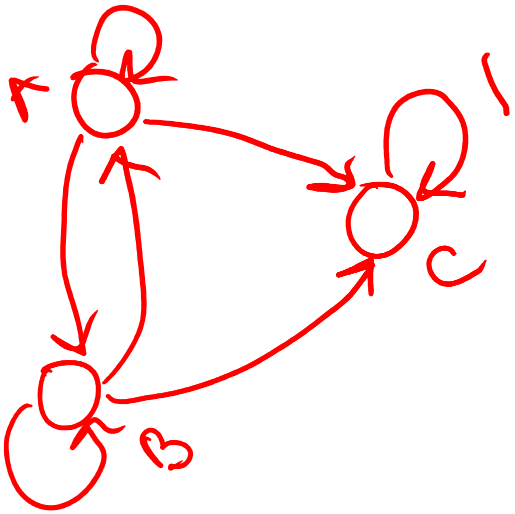
$$P_{00}(t) = 1 - P_{01}(t)$$

$$= \frac{1}{2} + \frac{1}{2} e^{-2\mu t}$$

Burning your bridges

- Can we come back to the states we are in ?
  - ever ? [short term analysis]
  - with the same rate that we go out of it ? [long term analysis]

In a Markov Chain,

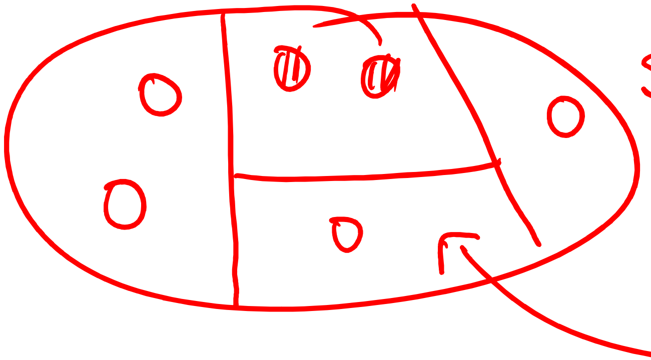


Irreducibility

If $P(X_{t+\tau} = j \mid X_t = i) > 0$ for some τ, j is accessible from i

$$i A j$$

$i A j$ & $j A i \iff i C j$ [i communicates w/ j]



Set of all states

RATE MATRIX WILL INDUCE A COMMUNICABILITY PARTITION

SINGLE PARTITION = IRREDUCIBLE

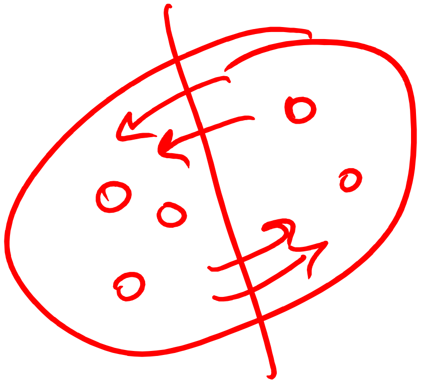
$\Updownarrow$

ST. DISTR.

Detailed balance

- Is the "flow" of probability balanced ?
- Is the process time reversible ?
  - Can we use Bayes Rule to flip the X0 and Xt ?

$$\pi_i q_{ij} = \pi_j q_{ji} \quad \forall i,j$$



" Circulating back "

Long run probabilities

- Equilibrium probability : we expect to see such nucleotide probabilities in current species

$$\pi_i = \lim_{t \to \infty} P(X_t = i)$$

$$\pi P(t) = \pi$$

$$\pi Q = 0$$

$$\pi_j q_j = \sum_{i \neq j} \pi_i q_{ij}$$

$$\sum \pi_i = 1$$

SYMMETRY!

Sometimes, lesser is better

- Jukes Cantor ‘69

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} \\ \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}-\frac{1}{4}e^{-t\mu} & \frac{1}{4}+\frac{3}{4}e^{-t\mu} \end{pmatrix}$$

Jukes, T.H. and C.R. Cantor. (1969)
Evolution of Protein Molecules, pp. 21-132.
Academic Press, New York.

Confounding factor

- mu and t
  - higher time, lower mutation rate
  - lower time, higher mutation rate

$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

Using symmetry

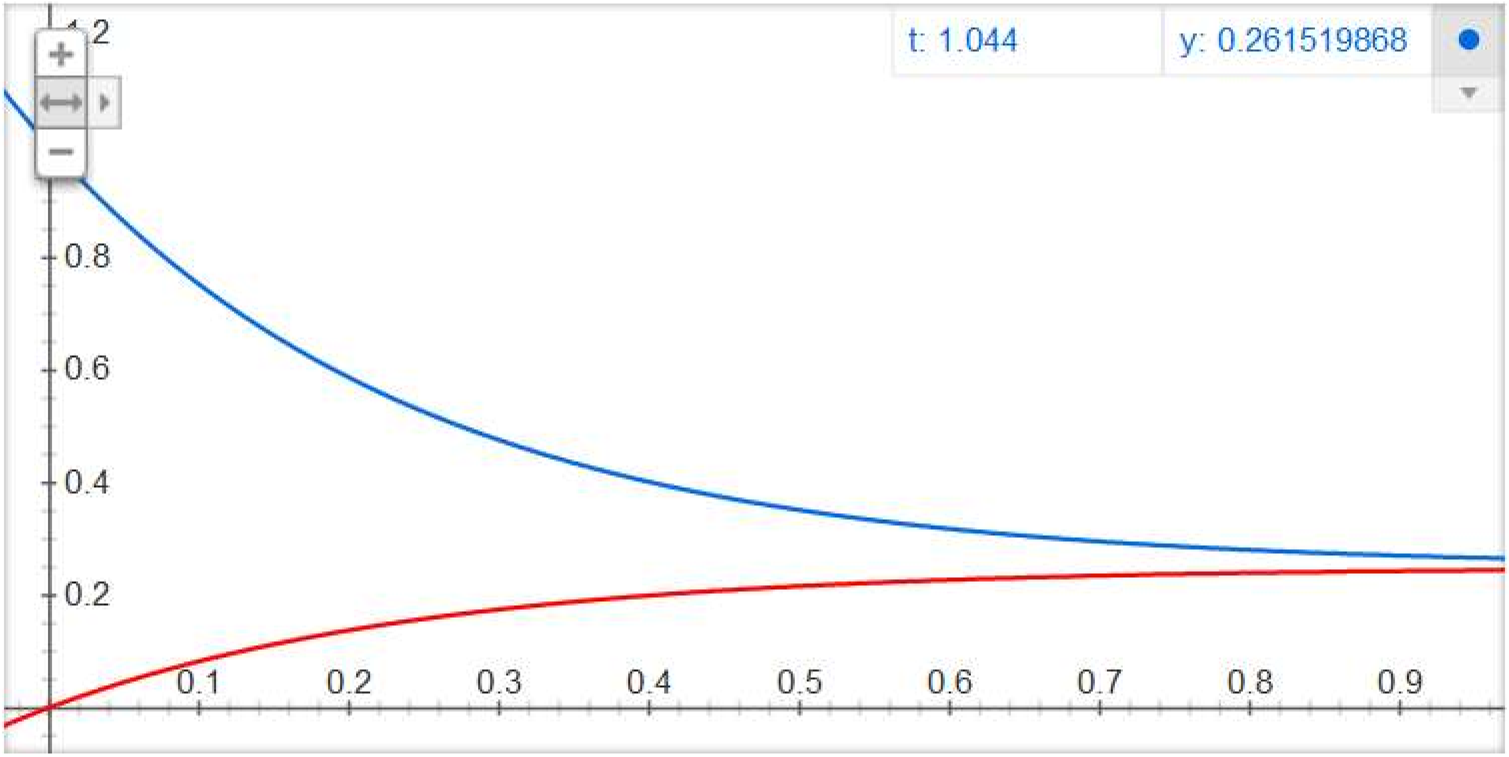
- Are A, T, G, C s interchangable ?
  - then the equilibrium probabilities are 0.25

- How many functions of t and mu are there anyway ? (shrink the matrix for simultaneous eqns)

$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

The nature of transition probabilities

Graph for 0.25*(1+3*exp((-4)*1*t)), 0.25*(1-exp((-4)*1*t))



- What are the equilibrium frequencies ?

Transitions vs transversions

- Purine (A, G)
- Pyrimidine (C, T)
- Transition : purine to purine, or pyrimidine to pyrimidine
- 2 / 3 SNP are transitions

Kimura ‘80

- Purines vs pyrimidines

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix} \times \mu$$

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16, 111-120.

UT DALLAS The University of Texas at Dallas

Felsenstein ‘81

- Equilibrium frequencies modelled

$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix}$$

$$P_{ij}(\nu) = \begin{cases} \pi_i + (1 - \pi_i)\, e^{-\beta\nu} & \text{if } i = j \\ \pi_j \left(1 - e^{-\beta\nu}\right) & \text{if } i \neq j \end{cases}$$

$$\beta = 1/(1 - \pi_A^2 - \pi_C^2 - \pi_G^2 - \pi_T^2)$$

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution, 17, 368-376.

HKY 85

- K80 + F81

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

$$P_{AC}(\nu, \kappa, \pi) = \pi_C \left(1.0 - e^{-\beta\nu}\right)$$

$$P_{AT}(\nu, \kappa, \pi) = \pi_T \left(1.0 - e^{-\beta\nu}\right)$$

$$P_{AG}(\nu, \kappa, \pi) = \left[\pi_G \left(\pi_A + \pi_G + (\pi_C + \pi_T)e^{-\beta\nu}\right) - \pi_G e^{-(1+(\pi_A+\pi_G)(\kappa-1.0))\beta\nu}\right] / \left(\pi_A + \pi_G\right)$$

$$P_{AA}(\nu, \kappa, \pi) = \left[\pi_A \left(\pi_A + \pi_G + (\pi_C + \pi_T)e^{-\beta\nu}\right) + \pi_G e^{-(1+(\pi_A+\pi_G)(\kappa-1.0))\beta\nu}\right] / (\pi_A + \pi_G)$$

$$\beta = \frac{1}{2(\pi_A + \pi_G)(\pi_C + \pi_T) + 2\kappa[(\pi_A\pi_G) + (\pi_C\pi_T)]}$$

Hasegawa, M., H. Kishino, and T. Yano. (1985) Dating of human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22, 160-174.

Generalized time reversible (GTR)

$$Q = \begin{pmatrix} -(x_1 + x_2 + x_3) & \frac{\pi_1 x_1}{\pi_2} & \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_1 x_3}{\pi_4} \\ x_1 & -(\frac{\pi_1 x_1}{\pi_2} + x_4 + x_5) & \frac{\pi_2 x_4}{\pi_3} & \frac{\pi_2 x_5}{\pi_4} \\ x_2 & x_4 & -(\frac{\pi_1 x_2}{\pi_3} + \frac{\pi_2 x_4}{\pi_3} + x_6) & \frac{\pi_3 x_6}{\pi_4} \\ x_3 & x_5 & x_6 & -(\frac{\pi_1 x_3}{\pi_4} + \frac{\pi_2 x_5}{\pi_4} + \frac{\pi_3 x_6}{\pi_4}) \end{pmatrix}$$

Time vs real time

- Is the “t” real time ?
- How can we figure out the scale of change in real time ?
  - Coming up, when we study phylogenies

Modelling higher granularity genomic entities

- Proteins
  - Dayhoff and other models
- Codons
  - Synonymous vs non synonymous change

Empirical models

- Empirical models may not have a “rate matrix”

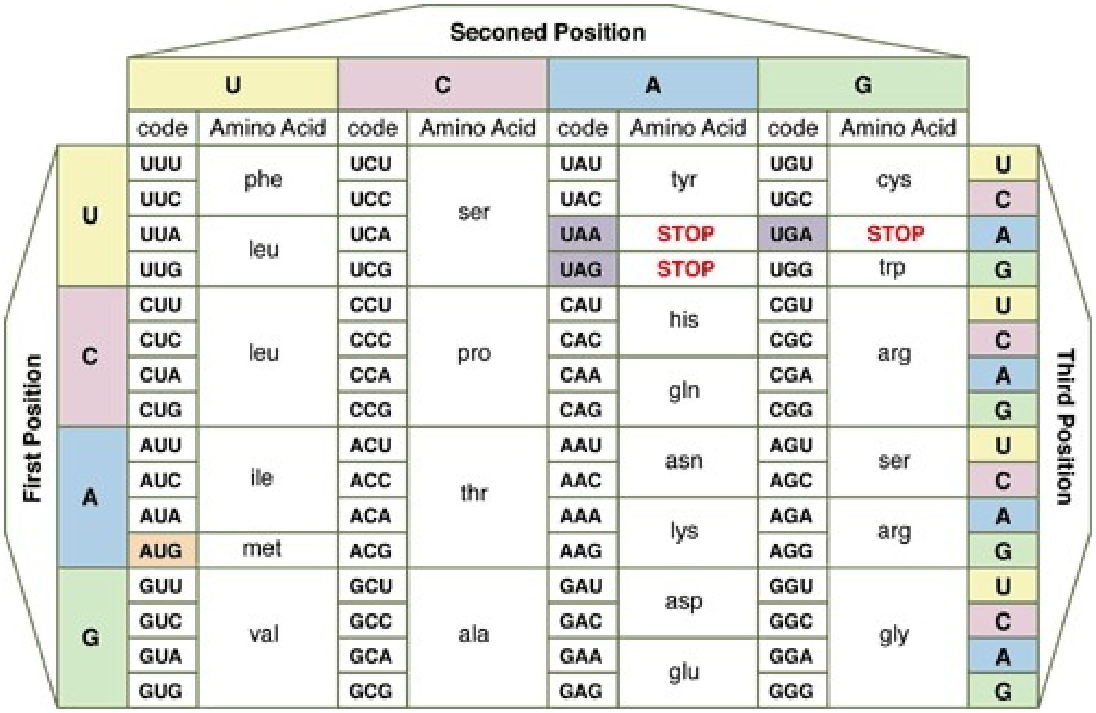
$$l(t) = \sum_i \sum_j n_{ij} \log \left[\pi_i P_{ij}(t) \right]$$

$$\begin{tikzcd} \text{TIME} \arrow[r, bend left] & \text{PARAMETER} \arrow[l, bend left] \end{tikzcd}$$

TRANSITION PROBABILITY

Codon table

- Synonymous & non synonymous mutations



	Seconed Position								
	U		C		A		G		
First Position	code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Goldman & Yang, 1994

- Bottom up modelling :

$$
\begin{aligned}
q_{ij} &= 0 \quad [\text{i \& j differ by 2 or 3 codon positions}] \\
&= \pi_j \; [\text{ differ by 1 syn. tranversion}] \\
&= \kappa \pi_j \; [\text{differ by 1 syn. transition}] \\
&= \omega \pi_j \; [\text{differ by 1 non.syn. transversion}] \\
&= \omega \kappa \pi_j \; [\text{differ by 1 non syn. transition}]
\end{aligned}
$$

Selection

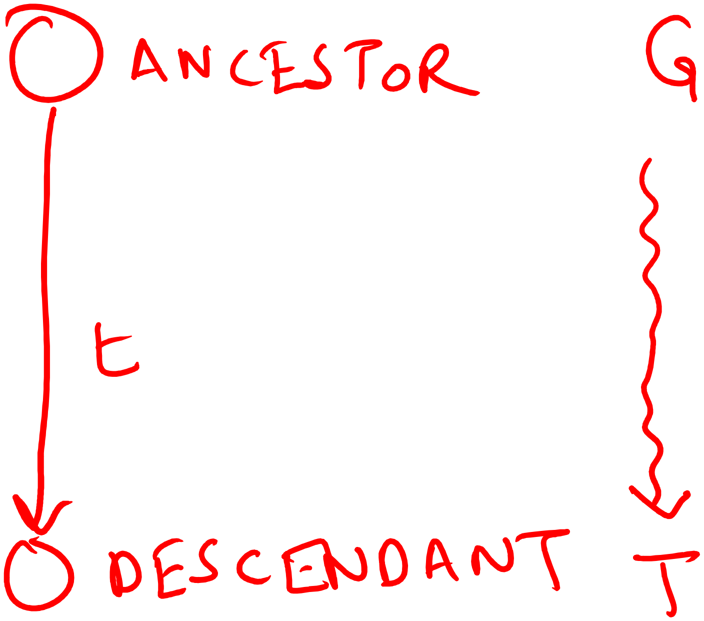
- Most generally :
  - Biasing model to one form of change over another

- Happens at every level :
  - Nucleotide (Transition vs transversion)
  - Nucleotide in the context of a Codon (Synonymous vs non synonymous)
  - Codon (some classes of amino acids may be interchangable)

Selection

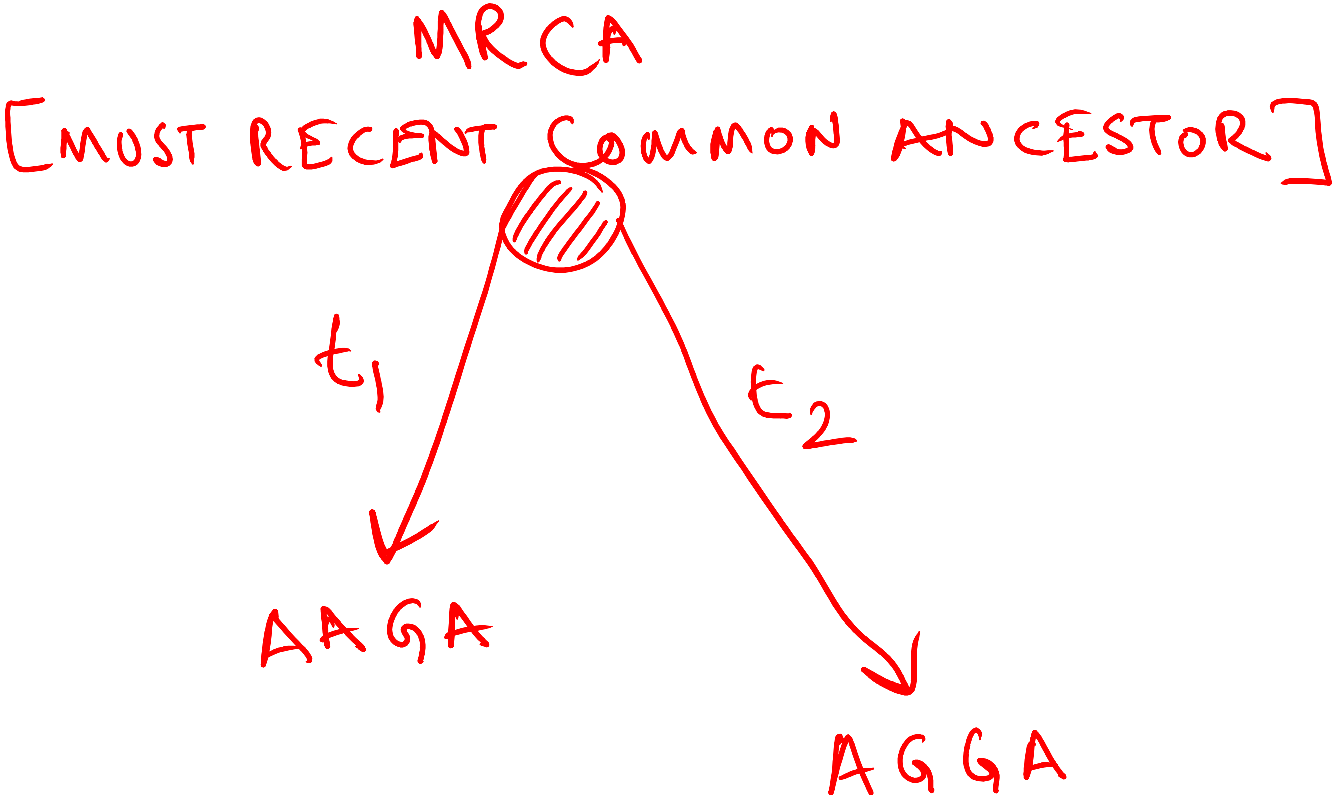
- We will talk more about selection and how it shapes our genomes after we study evolutionary trees

Modelling a lineage

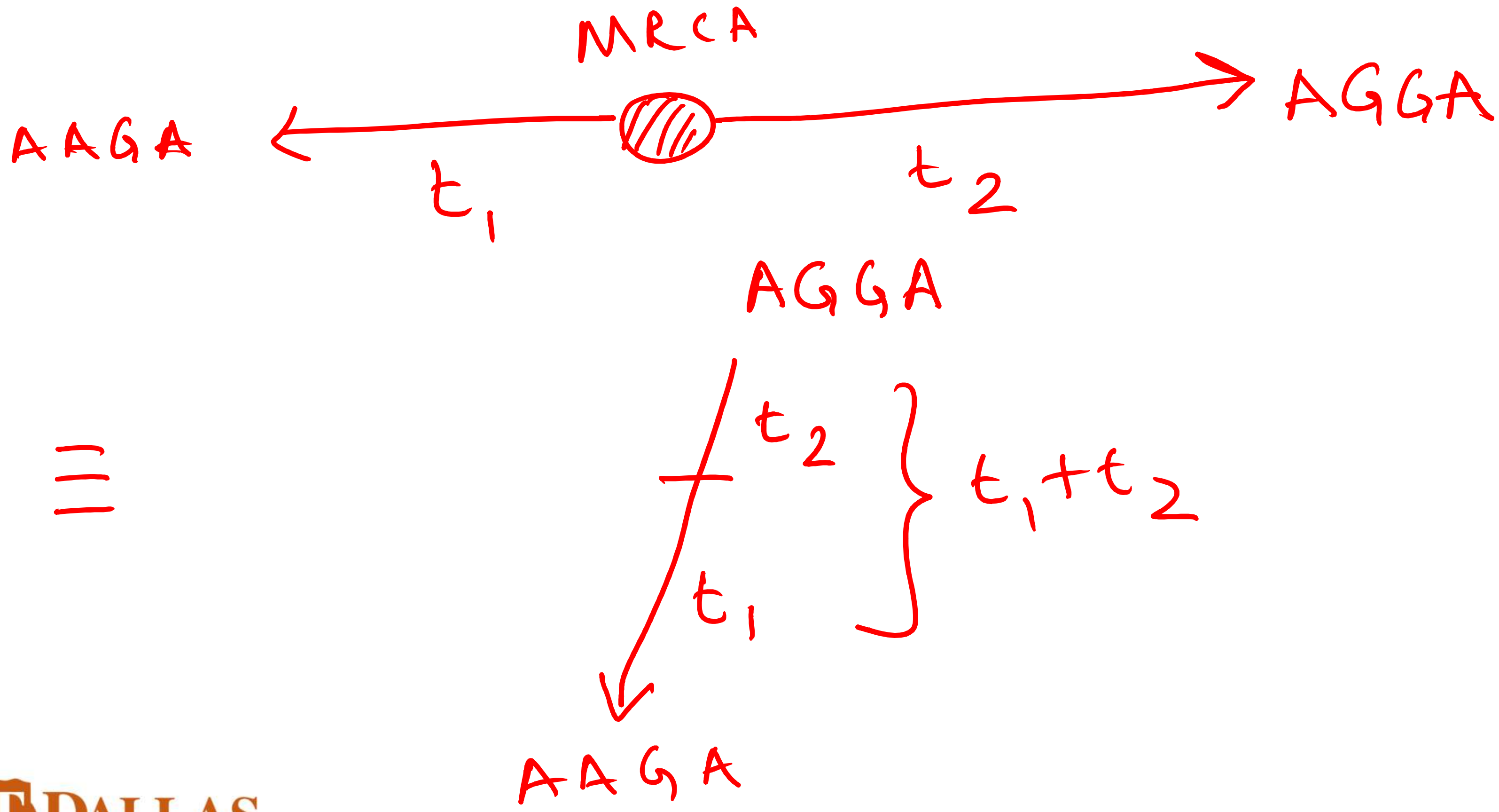


- What's the catch ?

Modelling two extant species



Modelling two extant species



Why can we do this ?

- Is it because they are :
  - Markovian ?
  - Or because they are memoryless ?
  - [✓ Or because they are time reversible ?]

All together now …

- Why just model a single lineage and forces acting on it ?
- Why not take into account all the species that branched off from that lineage ?
  - The more the merrier, in statistics
  - Which is where phylogenies come in !

Acknowledgements

- Eric Xing
- Howard Seltman