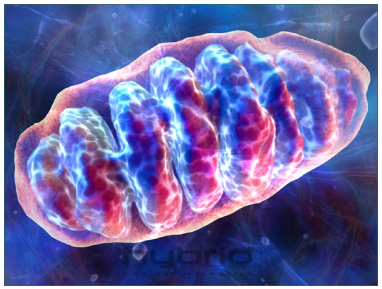
computational bioscience program

university of colorado school of medicine

Evolution of Proteins: Proteins 7350

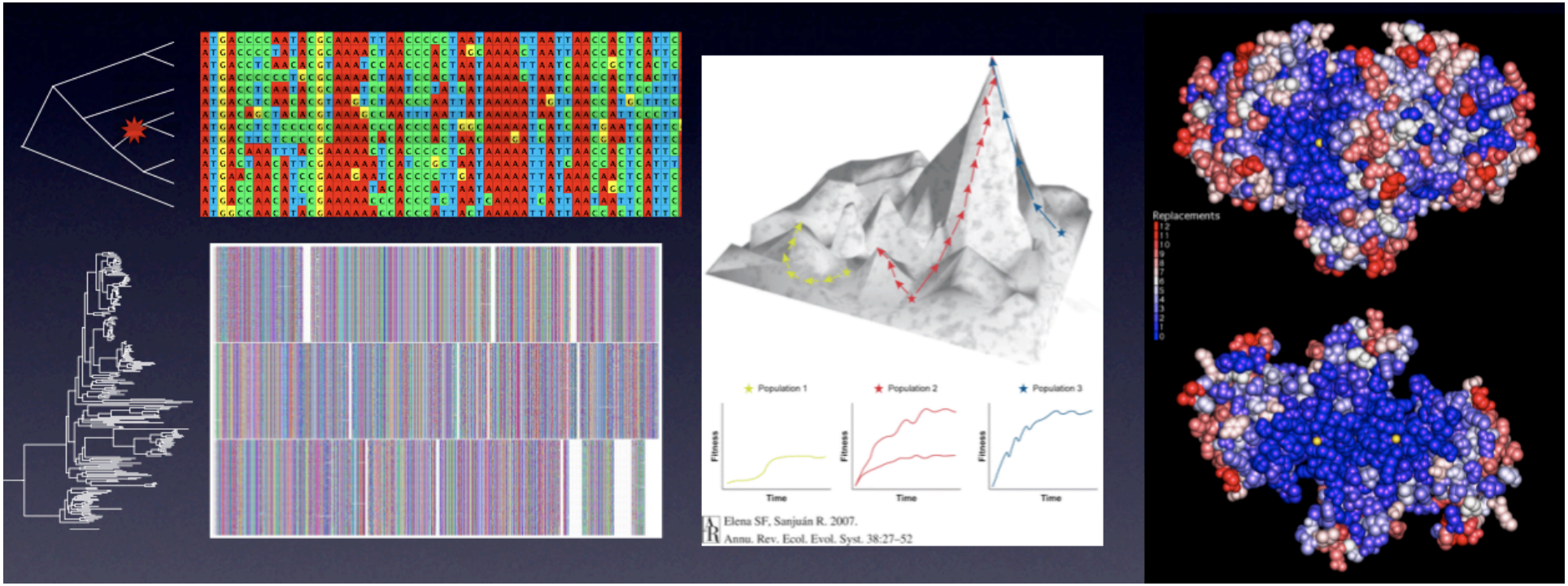
Pollock_ProteinEvol5.ppt



Biochemistry and Molecular Genetics
Computational Bioscience Program
Consortium for Comparative Genomics
University of Colorado School of Medicine

David.Pollock@uchsc.edu
www.EvolutionaryGenomics.com

Evolution of Proteins



Jason de Koning

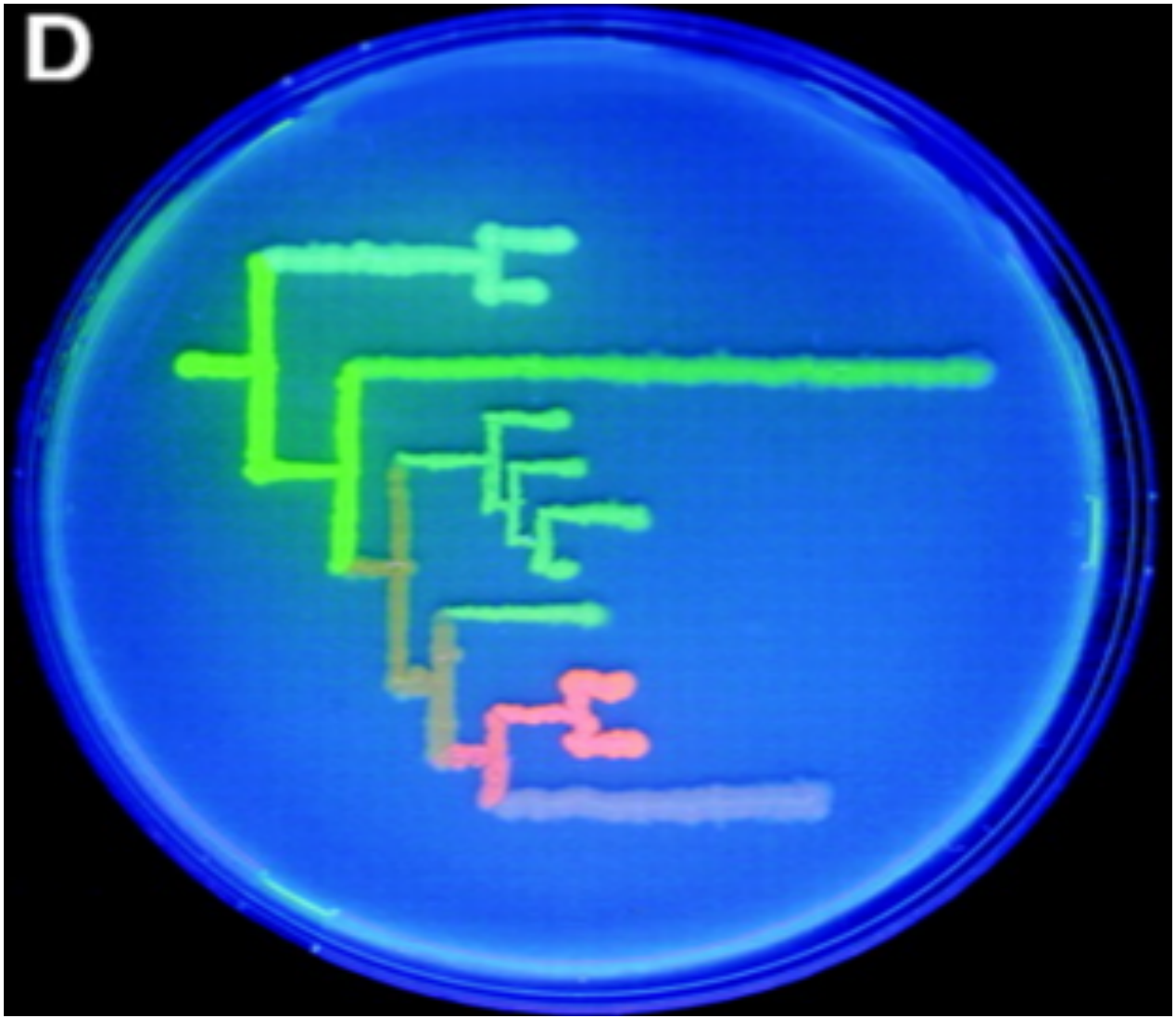
Description

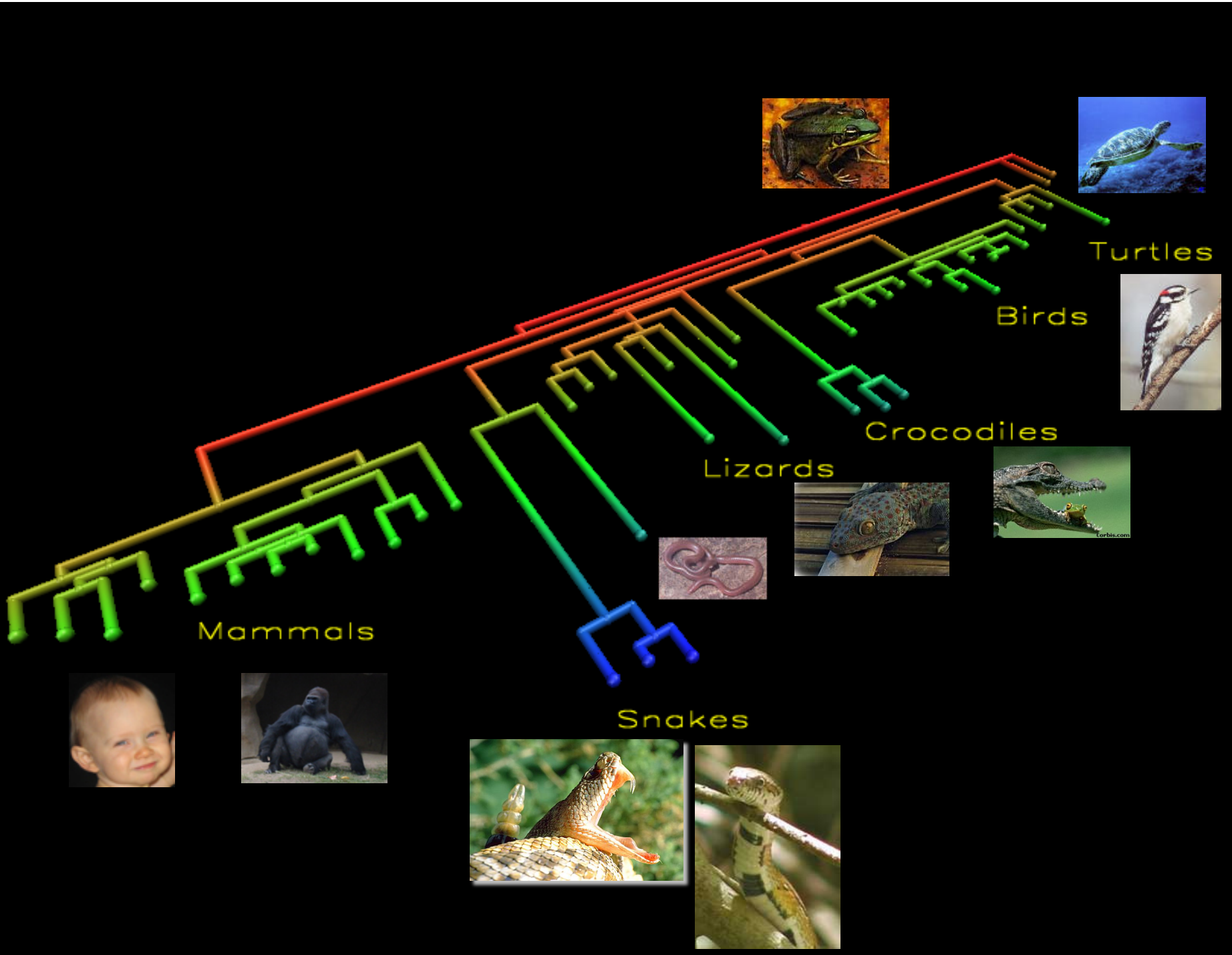
- Focus on protein structure, sequence, and functional evolution
- Subjects covered will include structural comparison and prediction, biochemical adaptation, evolution of protein complexes…

Topics (continued)

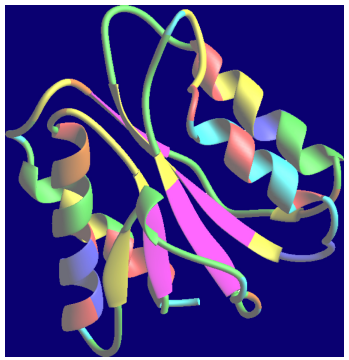
- …Probabilistic methods for detecting patterns of sequence evolution, effects of population structure on protein evolution, lattice and other computational models of protein evolution, protein folding and energetics, mutagenesis experiments, directed evolution, coevolutionary interactions within and between proteins, and detection of adaptation, diversifying selection and functional divergence.

Reconstruction of Ancestral Function




Turtles
Birds
Crocodiles
Lizards
Mammals
Snakes

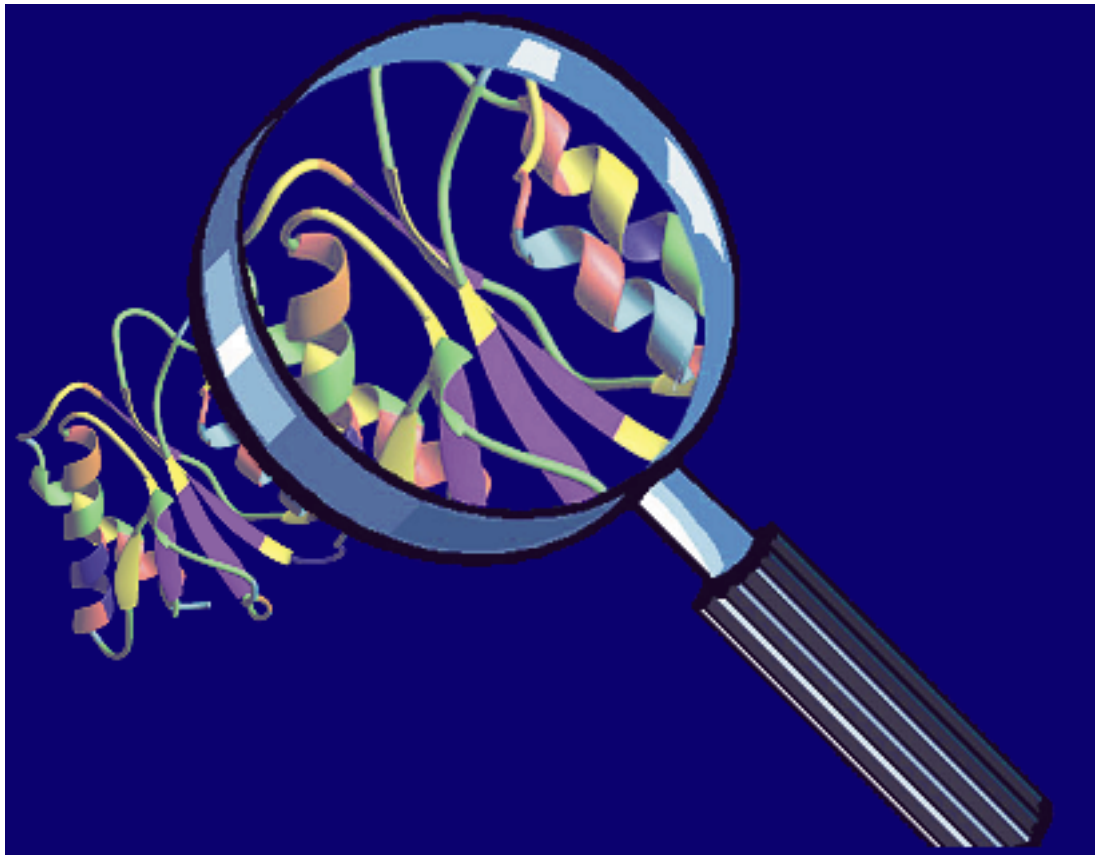
How do You Understand a New Protein?



Structural and Functional Studies

Experimental (NMR, X-tallography…)

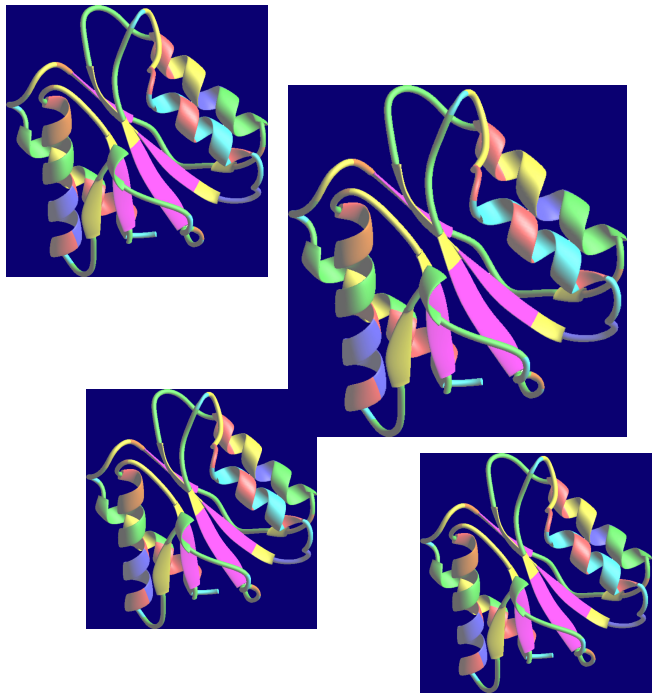
Computational (structure prediction…)



Comparative Sequence Analysis

Looking at sets of sequences

A common but **wrong** assumption: sequences are a **random** sample from the set of all possible sequences



Mouse: …TLSPGLKIVSNPL…
Rat: …TLTPGLKLVSDTL…
Baboon: …TVSPGLRIVSDGV…
Chimp: …TISPGLVIVSENL…

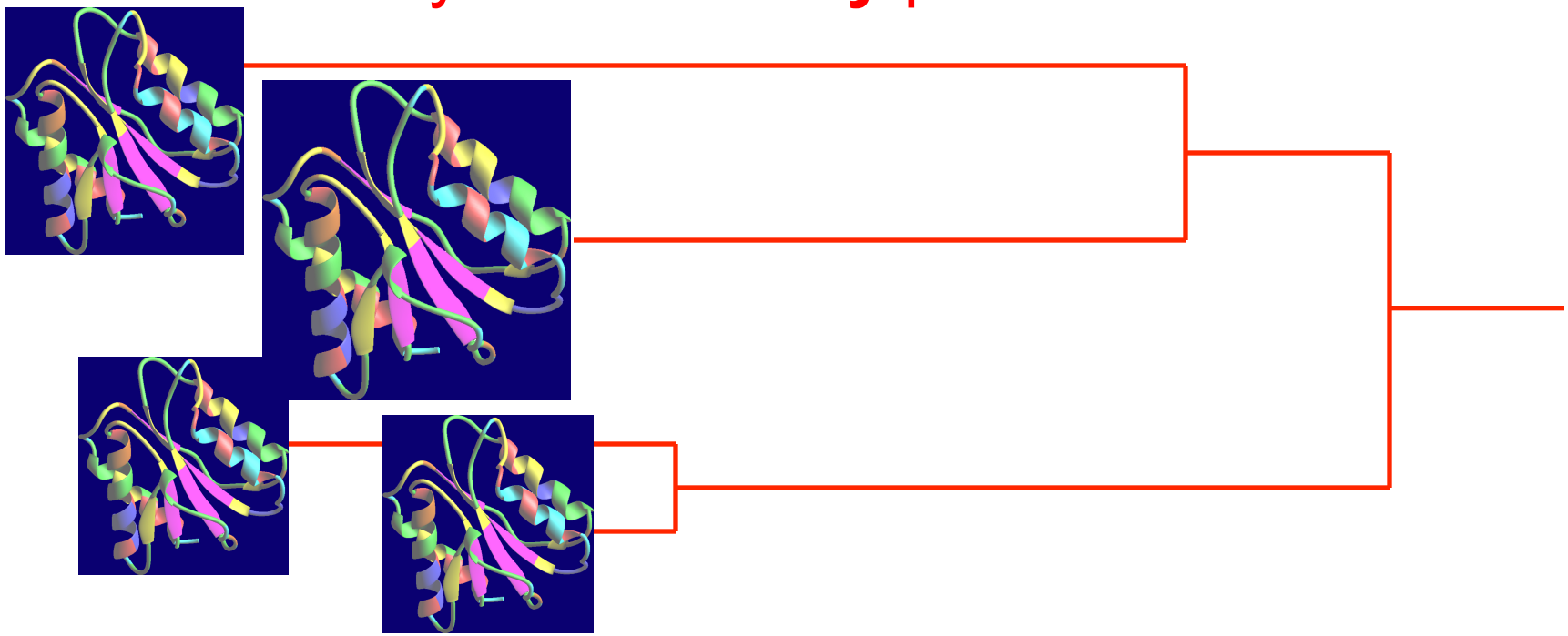
Conserved proline

Variable "High entropy"

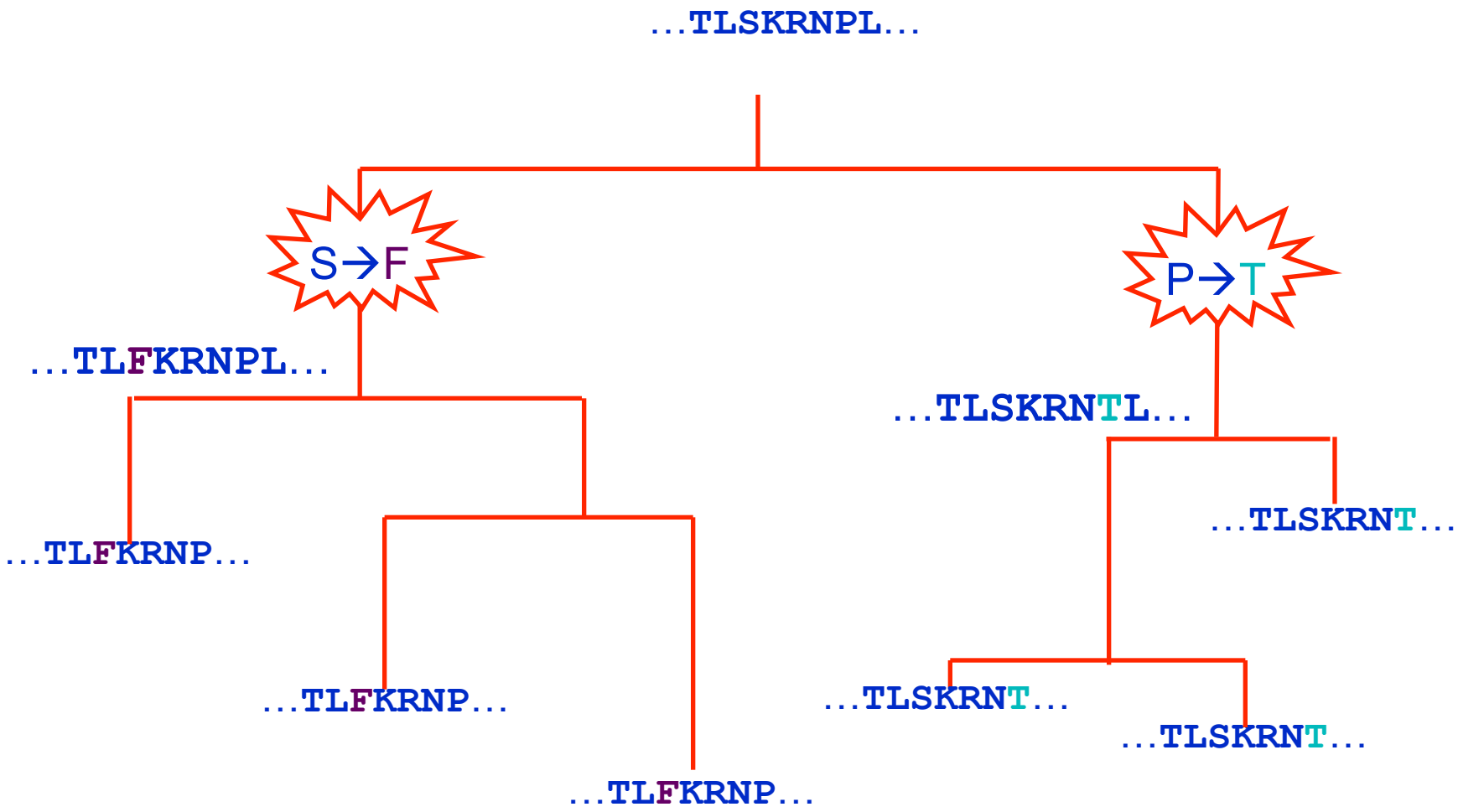
Comparative Sequence Analysis

Looking at sets of sequences

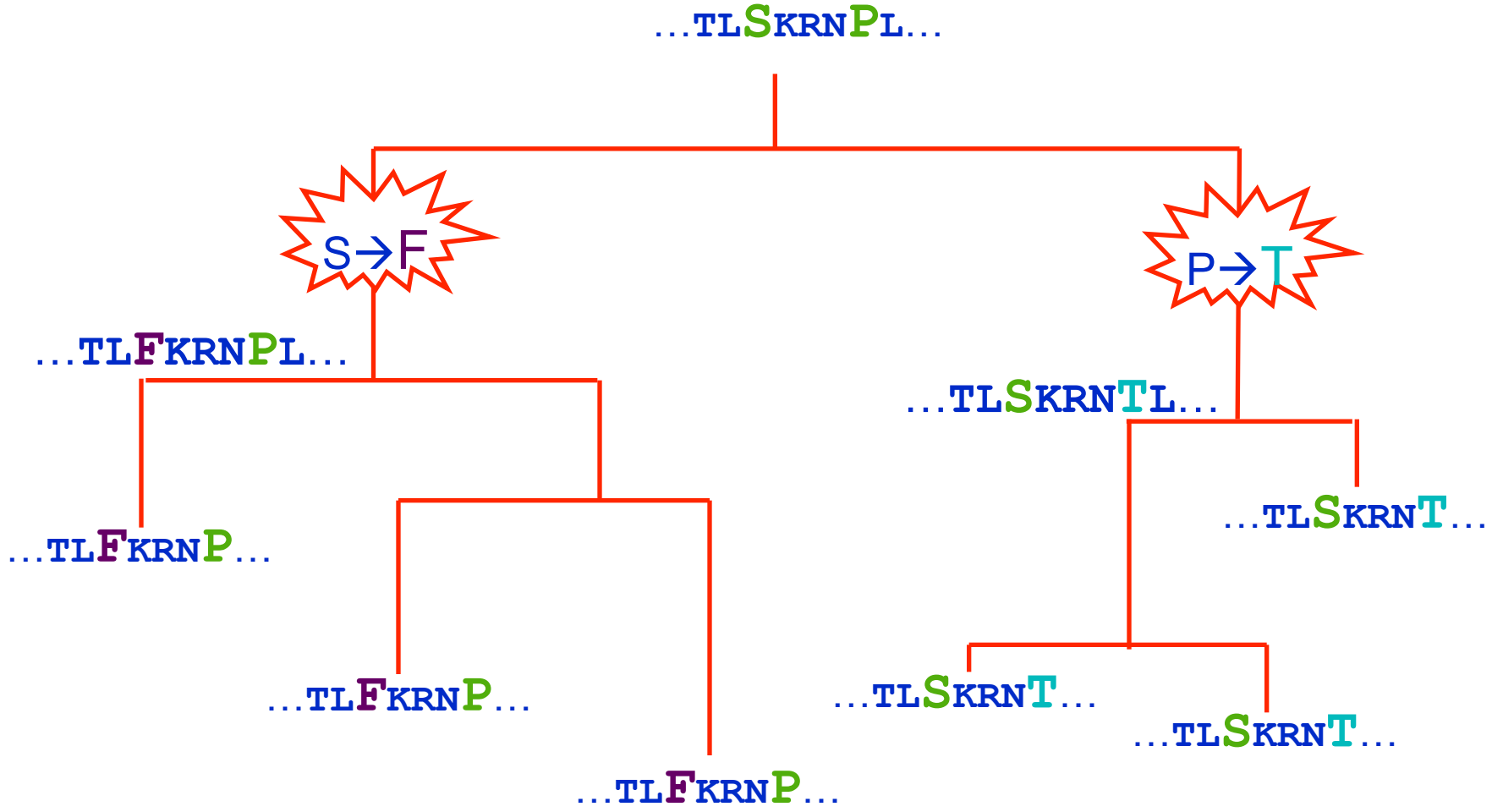
In reality, proteins are related
by **evolutionary** process



Confounding Effect of Evolution



Confounding Effect of Evolution



Everytime there is an F, there is a P!

Everytime there is an S, there is a T!

Ways to Deal with This…

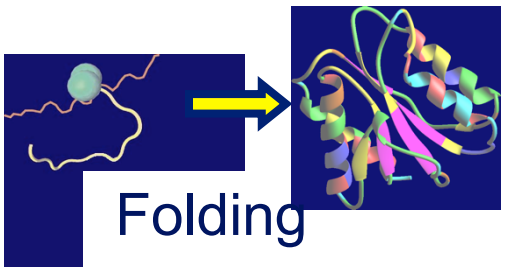
Most common: Ignorance is Bliss

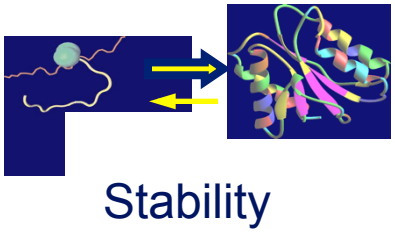
Some: Try to estimate the extent of the confounding (Mirny, Atchley)

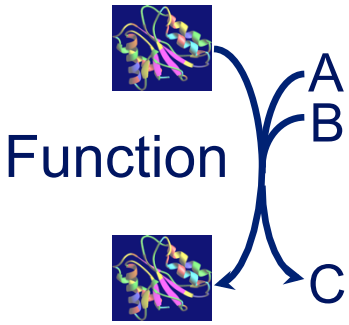
Remove the confounding (Maxygen)

Include evolution explicitly in the model (Goldstein, Pollock, Goldman, Thorne, …)

Fitness Selection Stochastic Realizations

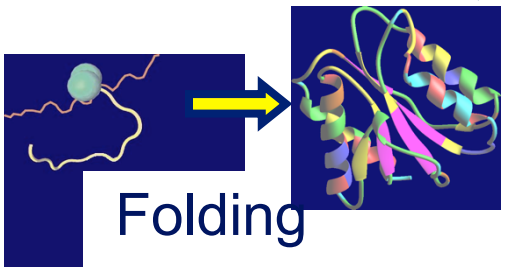




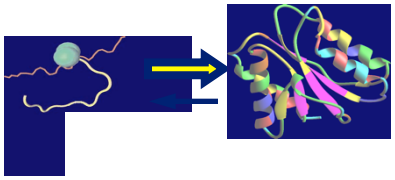


Mouse: …**TLSPGLKIVSNPL**…
Rat: …**TLTPGLKLVSDTL**…
Baboon: …**TVSPGLRIVSDGV**…
Chimp: …**TISPGLVIVSENL**…

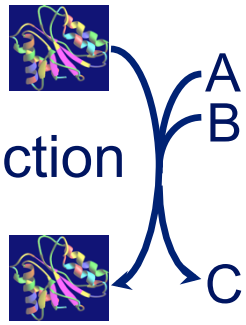
Understanding Model Data



Folding



Stability



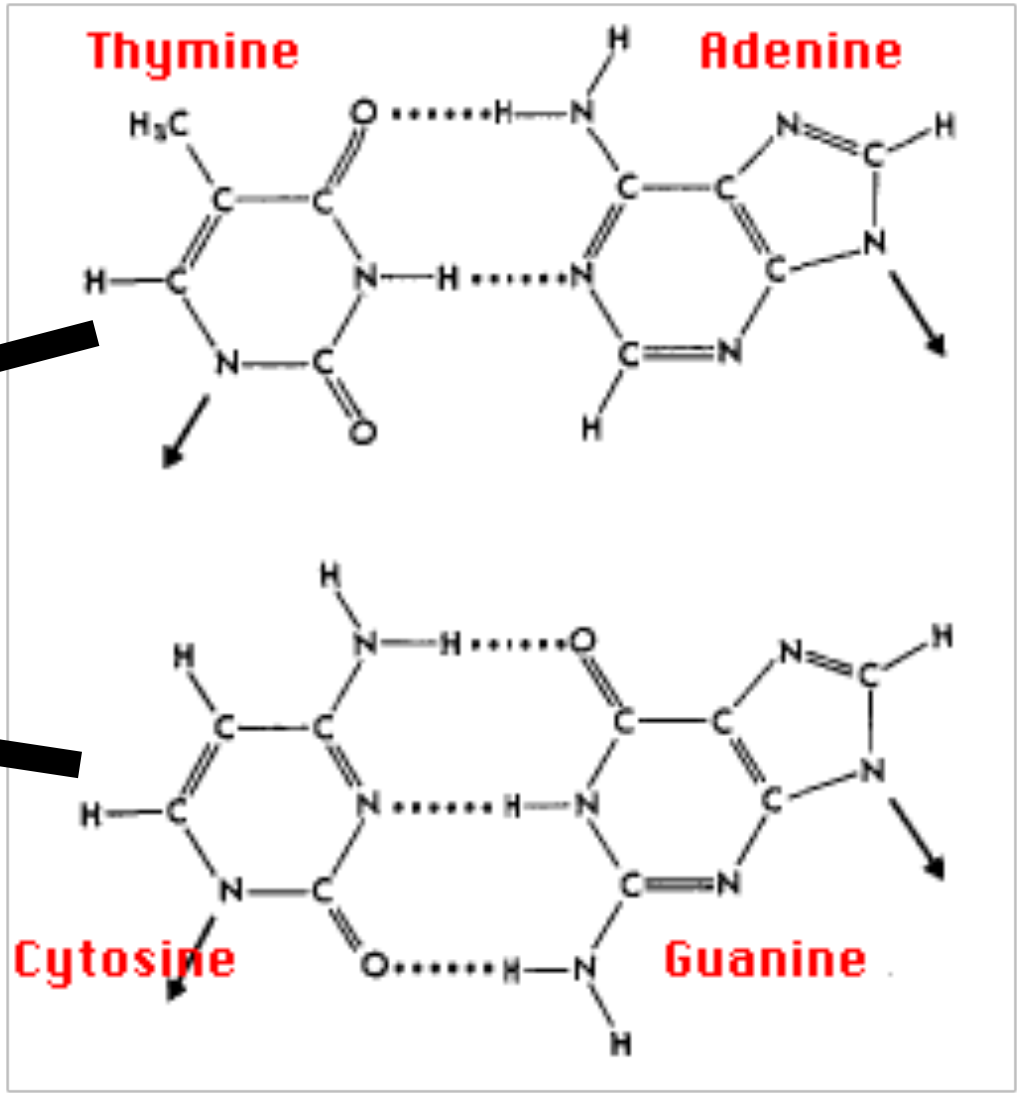
Function

A

B

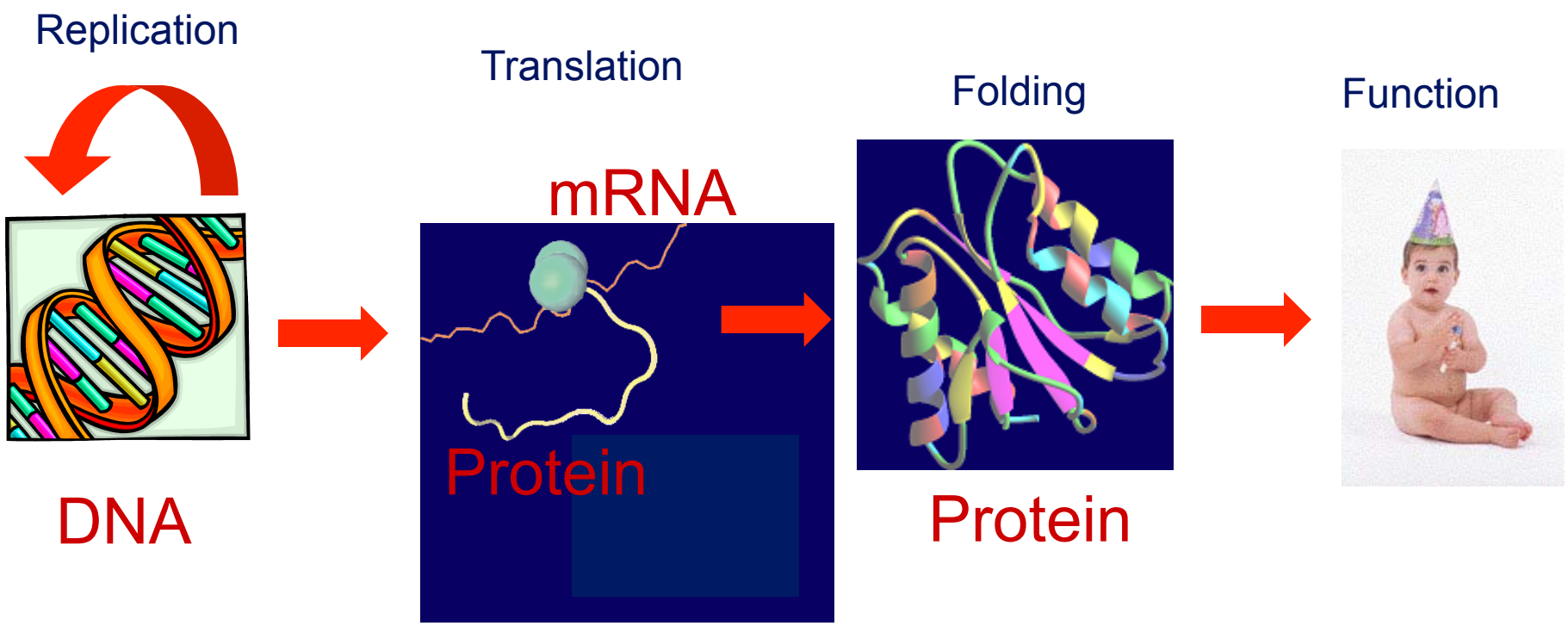
C

Mouse: …**TLSPGLKIVSNPL**…
Rat: …**TLTPGLKLVSDTL**…
Baboon: …**TVSPGLRIVSDGV**…
Chimp: …**TISPGLVIVSENL**…

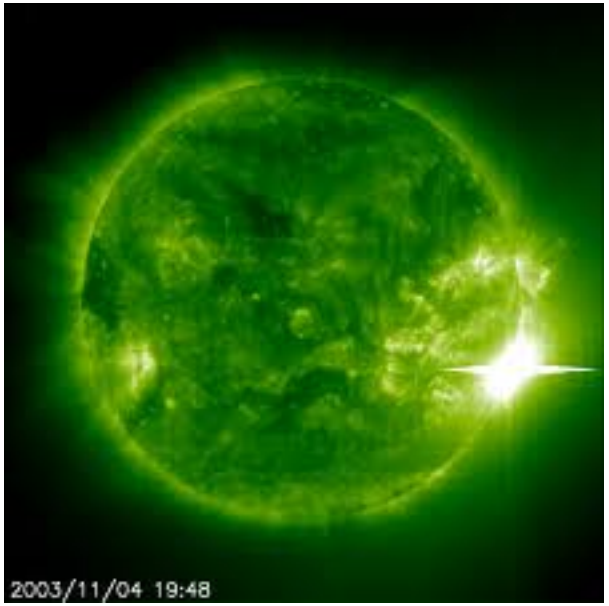


DNA

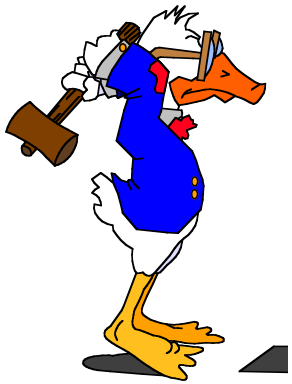
What does DNA do?

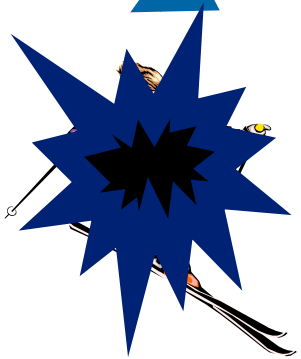


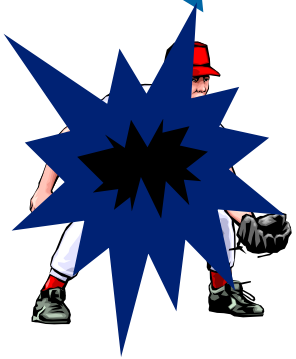
Mutations result in genetic variation

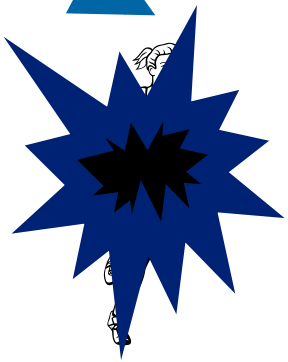


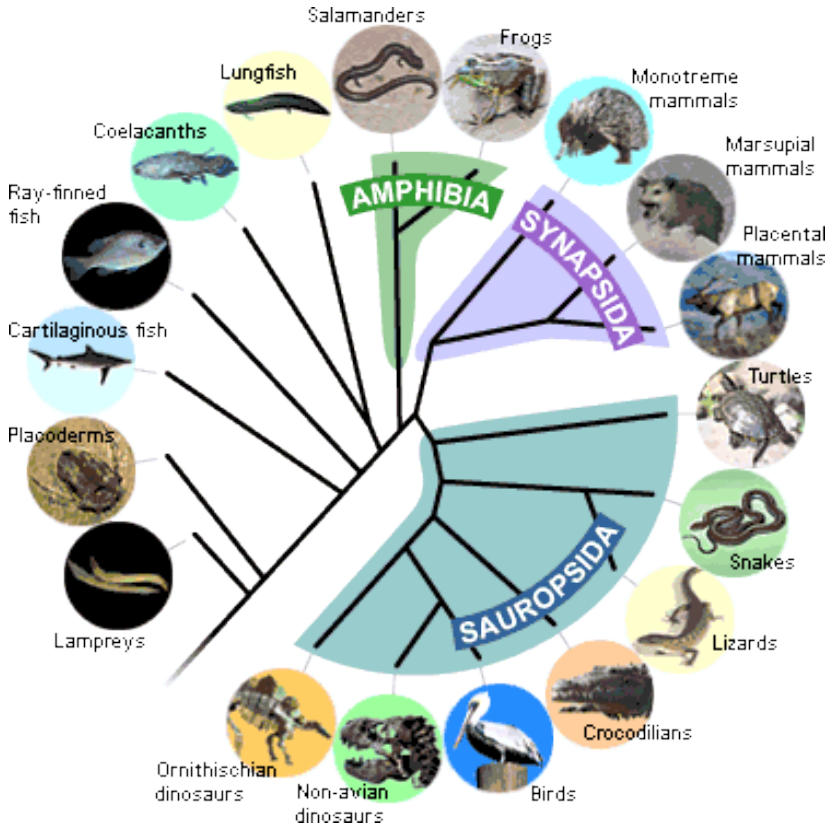
Selective Pressure

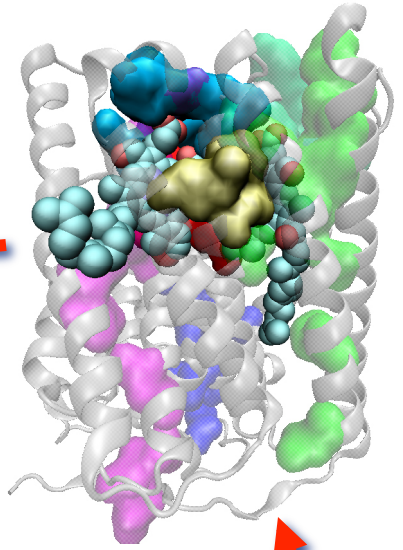


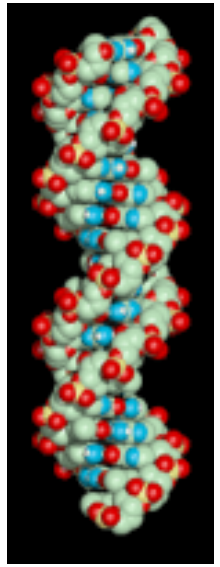




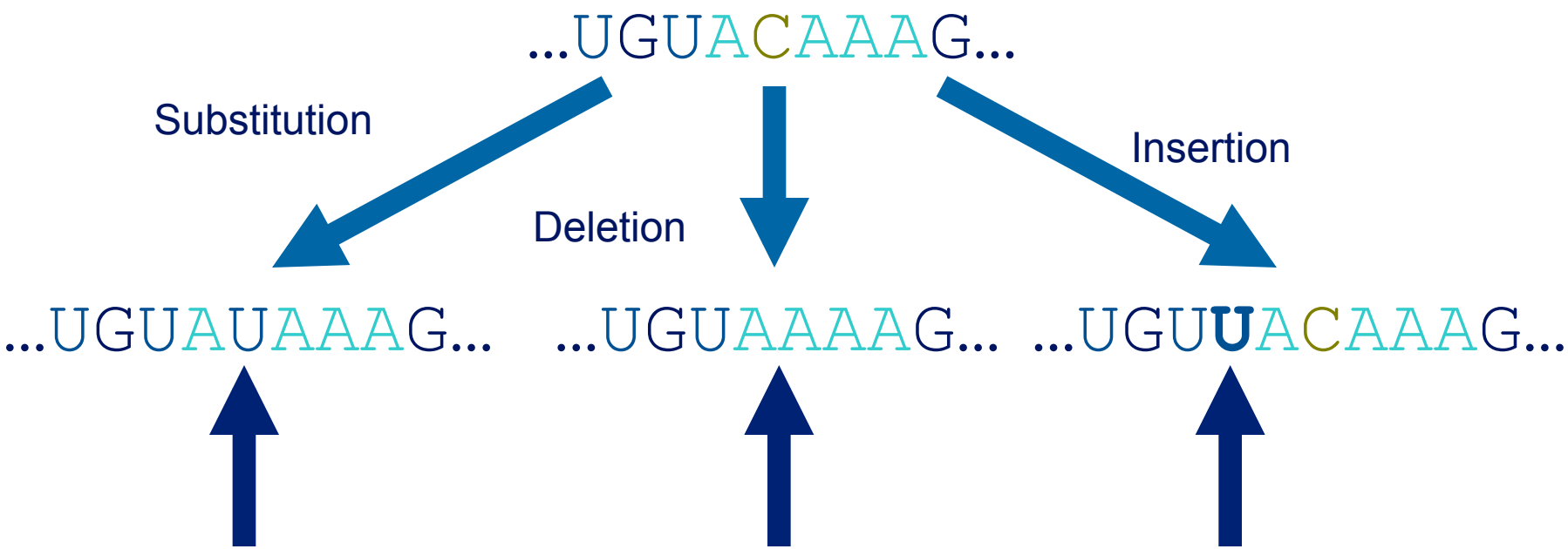



Salamanders
Frogs
Lungfish
Monotreme mammals
Coelacanths
Marsupial mammals
Ray-finned fish
AMPHIBIA
SYNAPSIDA
Placental mammals
Cartilaginous fish
Turtles
Placoderms
SAUROPSIDA
Snakes
Lizards
Lampreys
Crocodilians
Ornithischian dinosaurs
Non-avian dinosaurs
Birds





Genetic changes

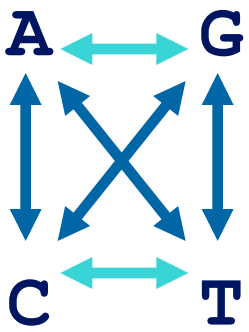


Substitutions Can Be:

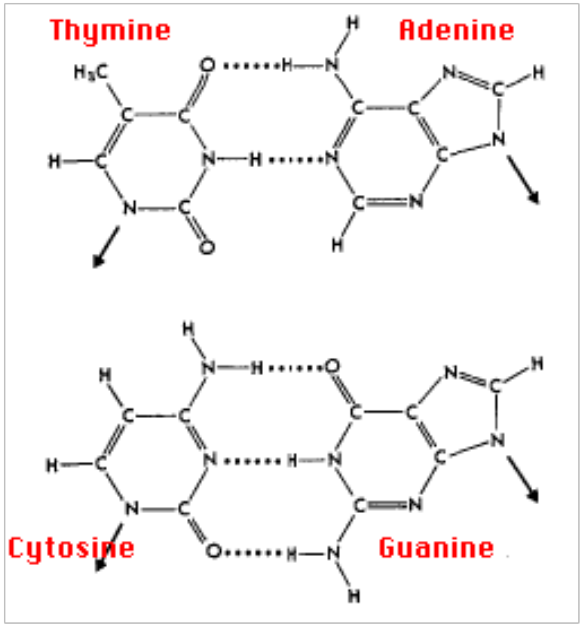
Transitions

Purines:

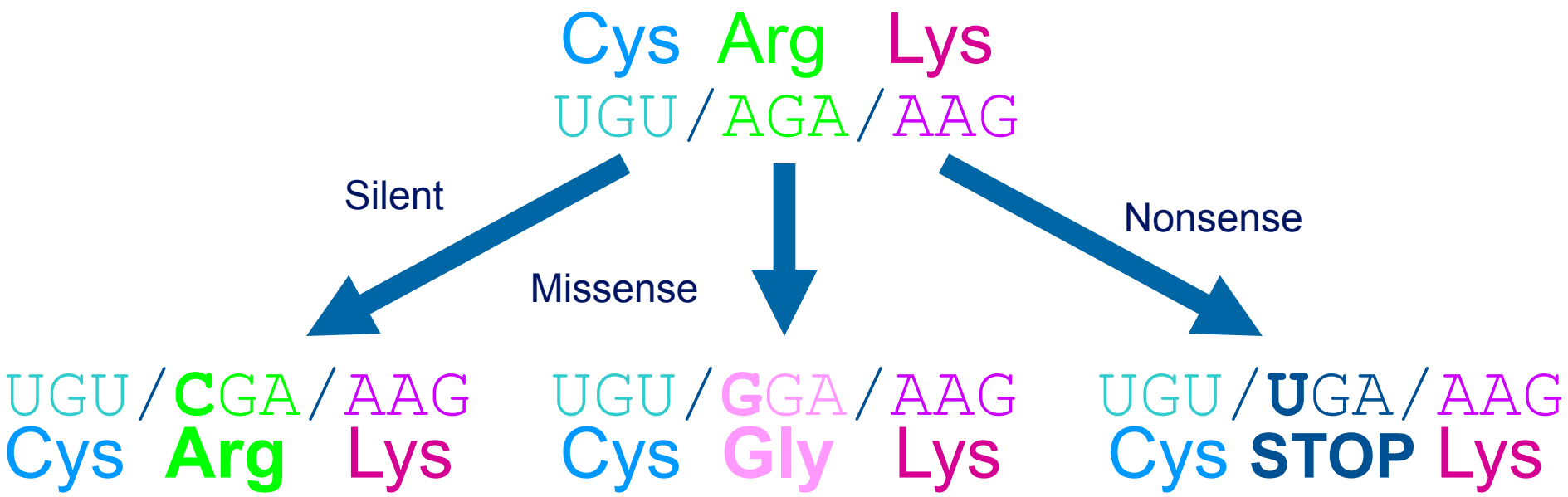
Pyrimidines:



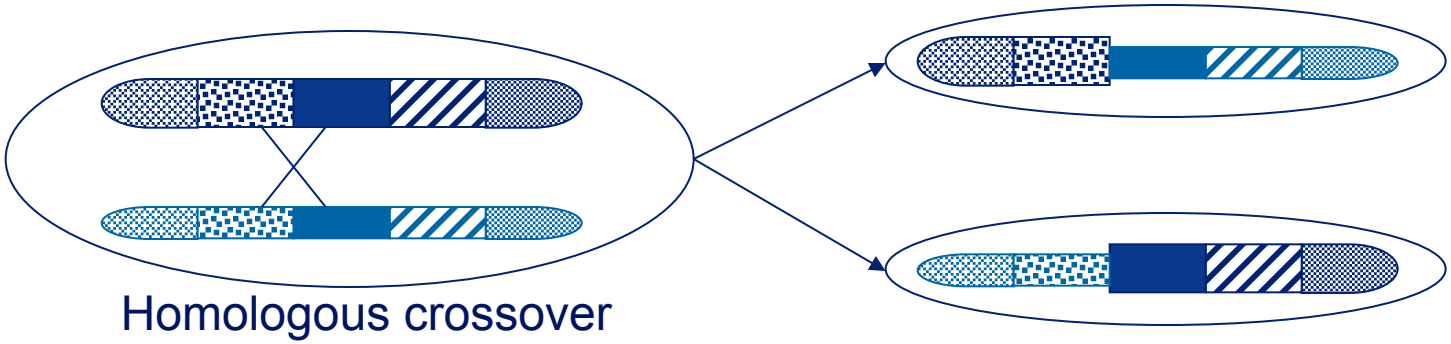
Transversions



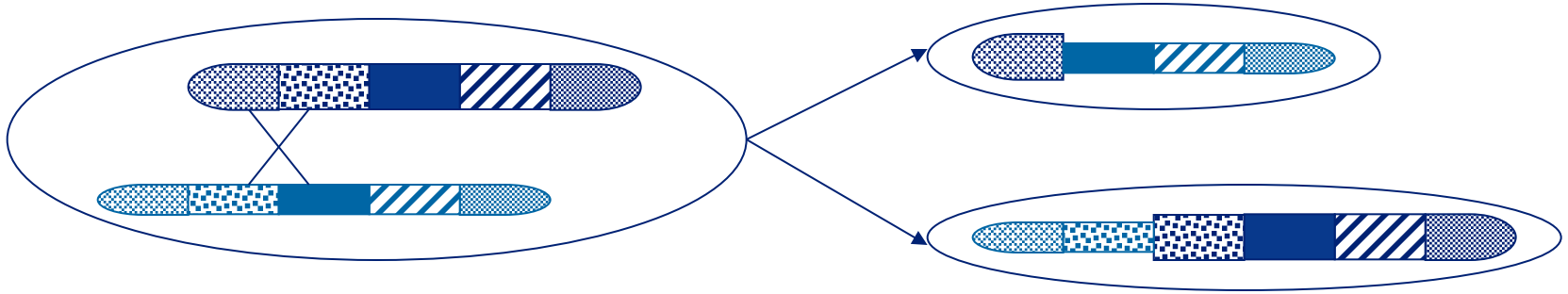
Substitutions in coding regions can be:



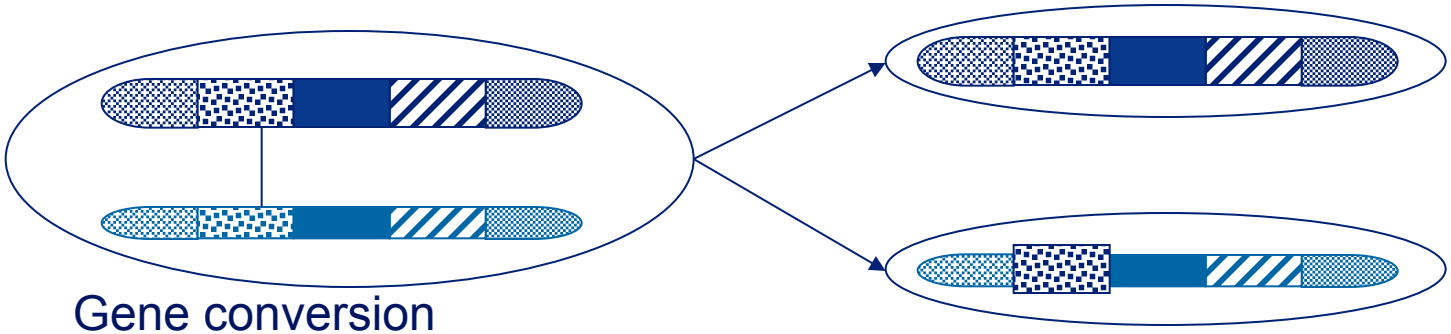
First position: 4% of all changes silent
Second position: no changes silent
Third position: 70% of all changes silent (wobble position)



Homologous crossover



Uneven crossover leading to gene deletion and duplication



Gene conversion

Fate of a duplicated gene

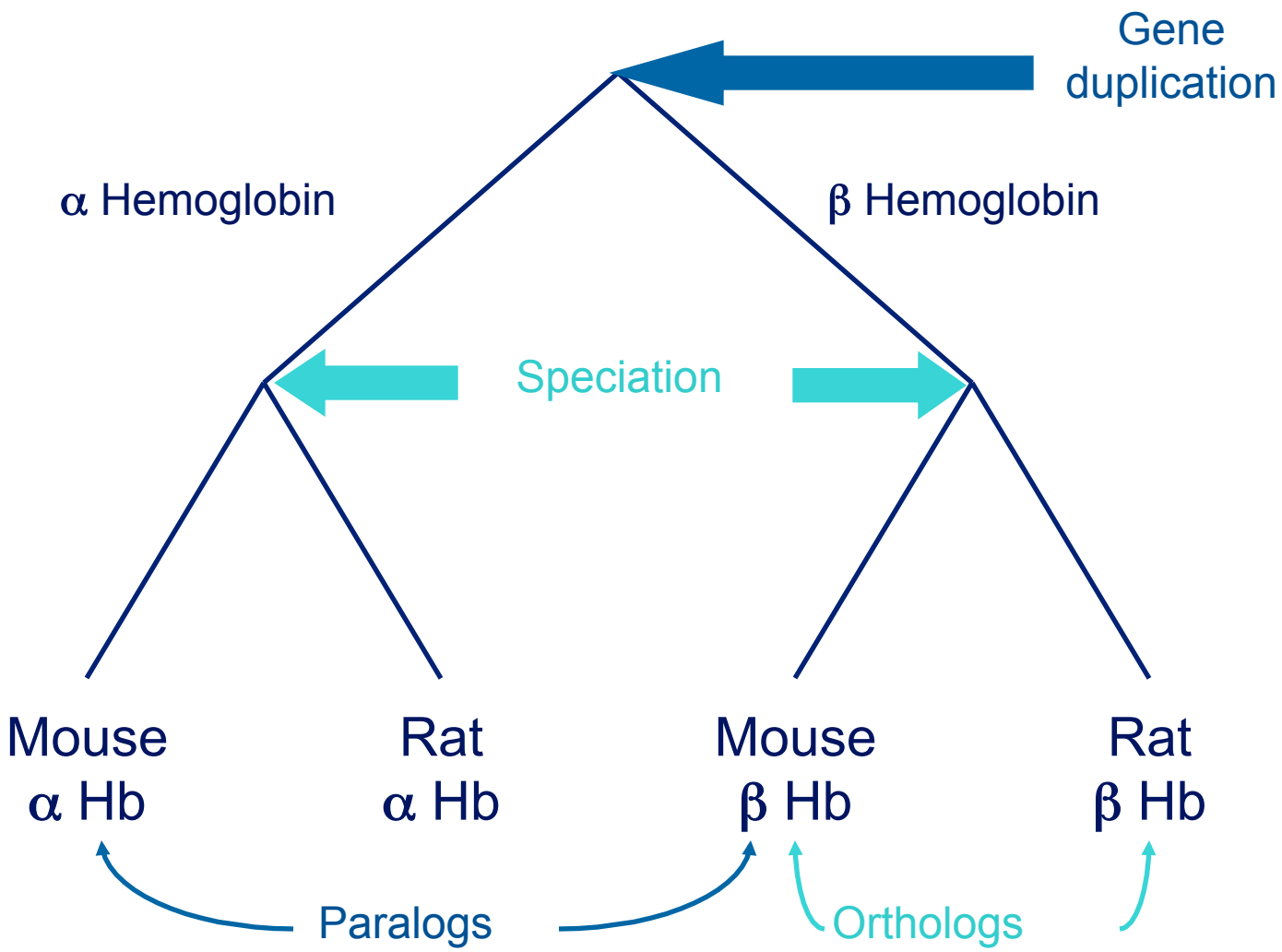
Keep on doing whatever it originally was doing

Lose ability to do anything (become a pseudogene)

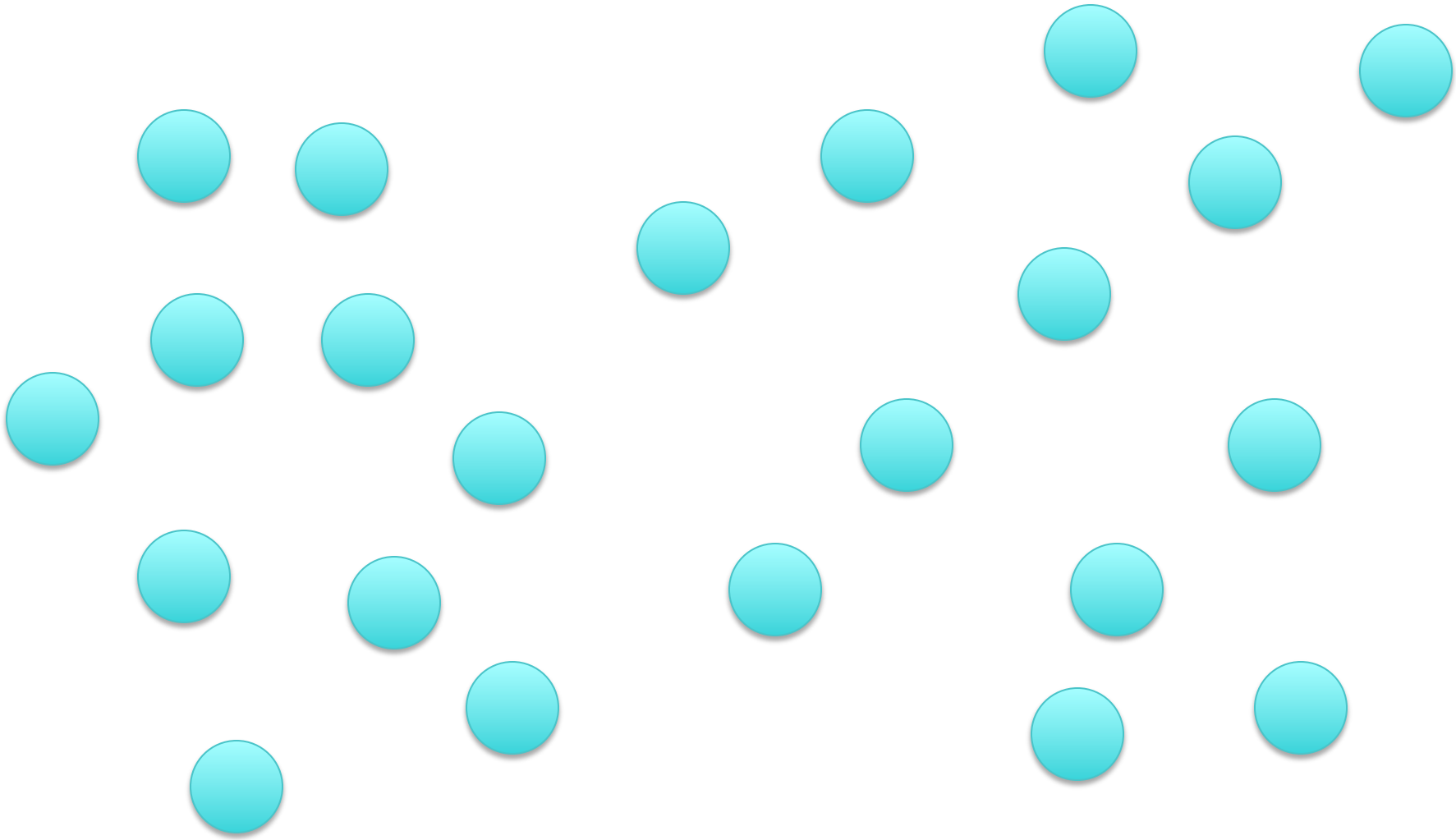
Learn to do something new (neofunctionalization)

Split old functions among new genes (subfunctionalization)

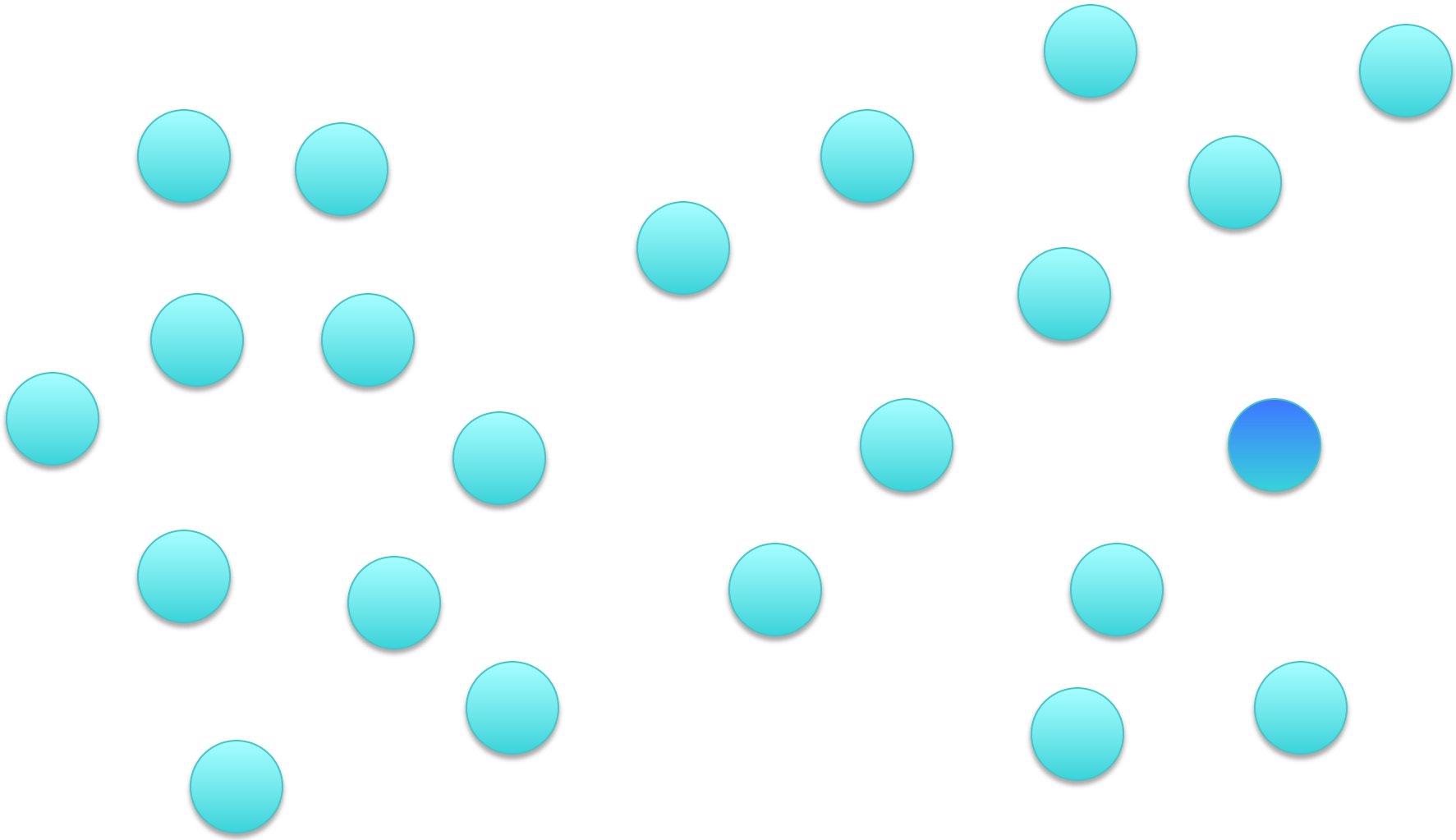
Homologies


Gene duplication
α Hemoglobin
β Hemoglobin
Speciation
Mouse α Hb
Rat α Hb
Mouse β Hb
Rat β Hb
Paralogs
Orthologs

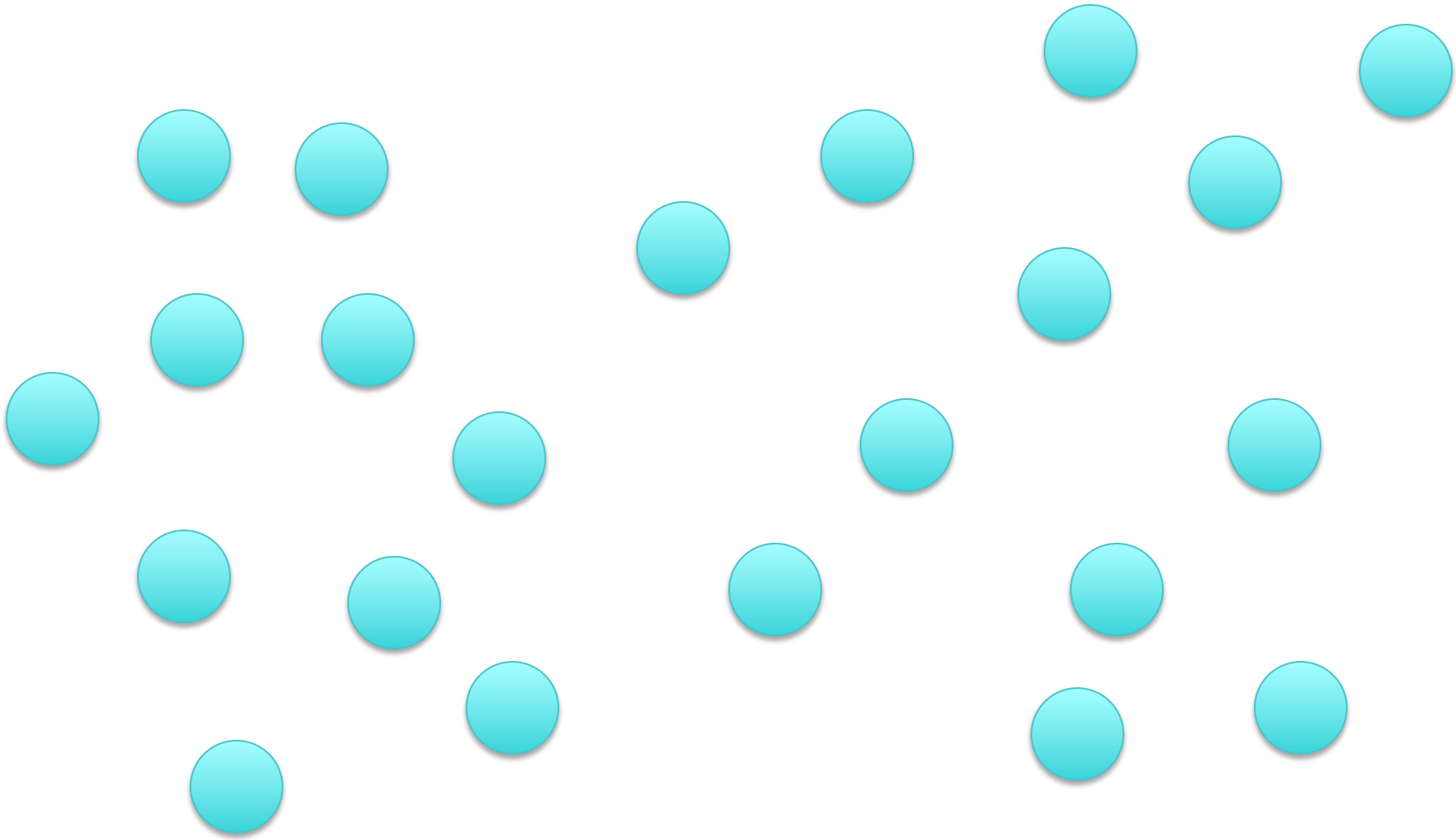
Initial Population



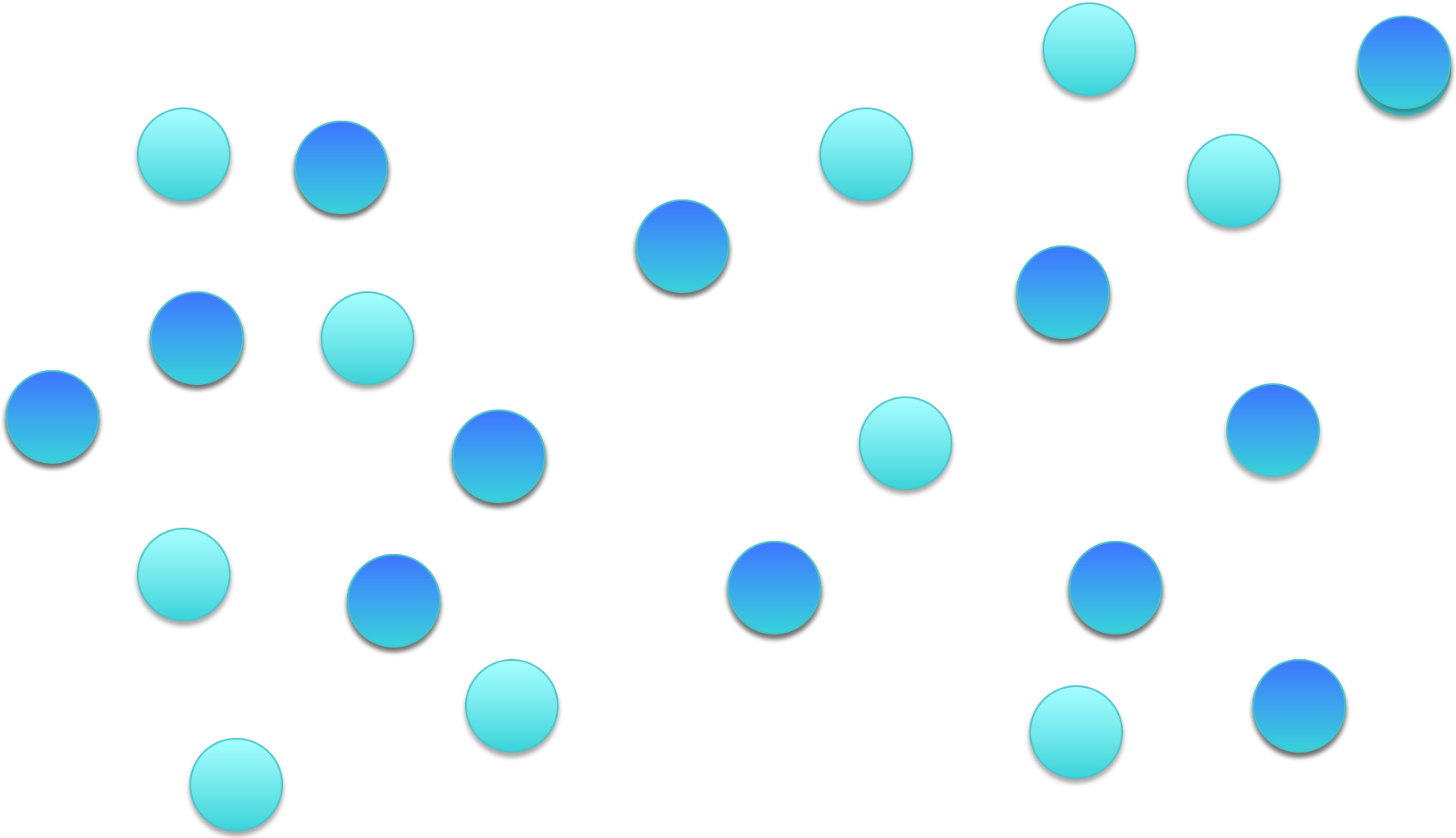
Mistakes are Made



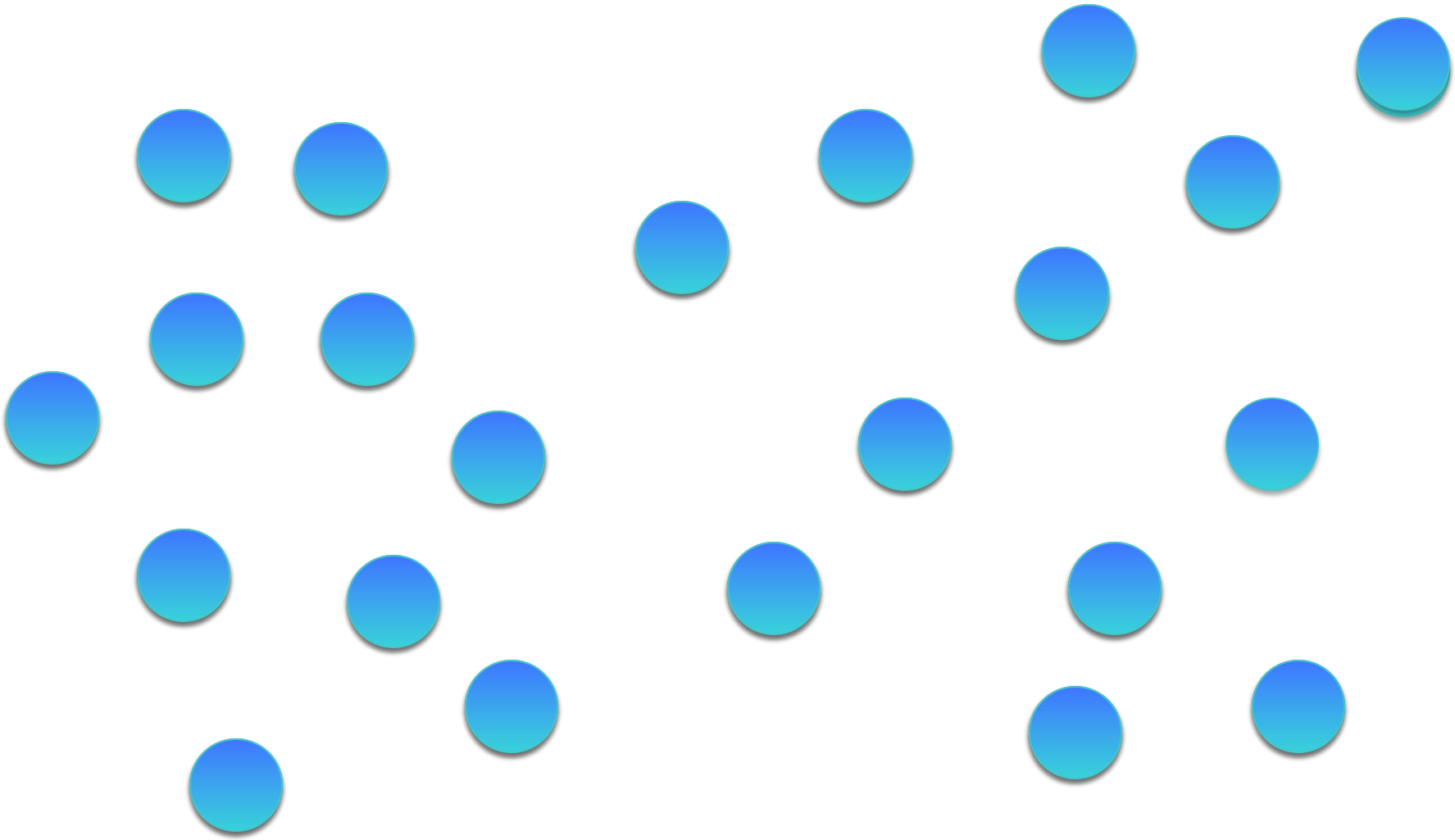
Elimination



Polymorphism



Fixation



Selection

Differences in fitness (capacity for fertile offspring)

1 gene
2 alleles (variations), **A** and **B**
3 genotypes (diploid organism): **AA**, **AB**, **BB**

Genotype	Fitness
AA	$\omega_{AA} = 1$ (wild type)
AB	$\omega_{AB} = 1 + S_{AB}$
BB	$\omega_{BB} = 1 + S_{BB}$

$S > 0$ advantageous
$S < 0$ unfavorable
$S \sim 0$ neutral

Evolution of Gene Frequencies

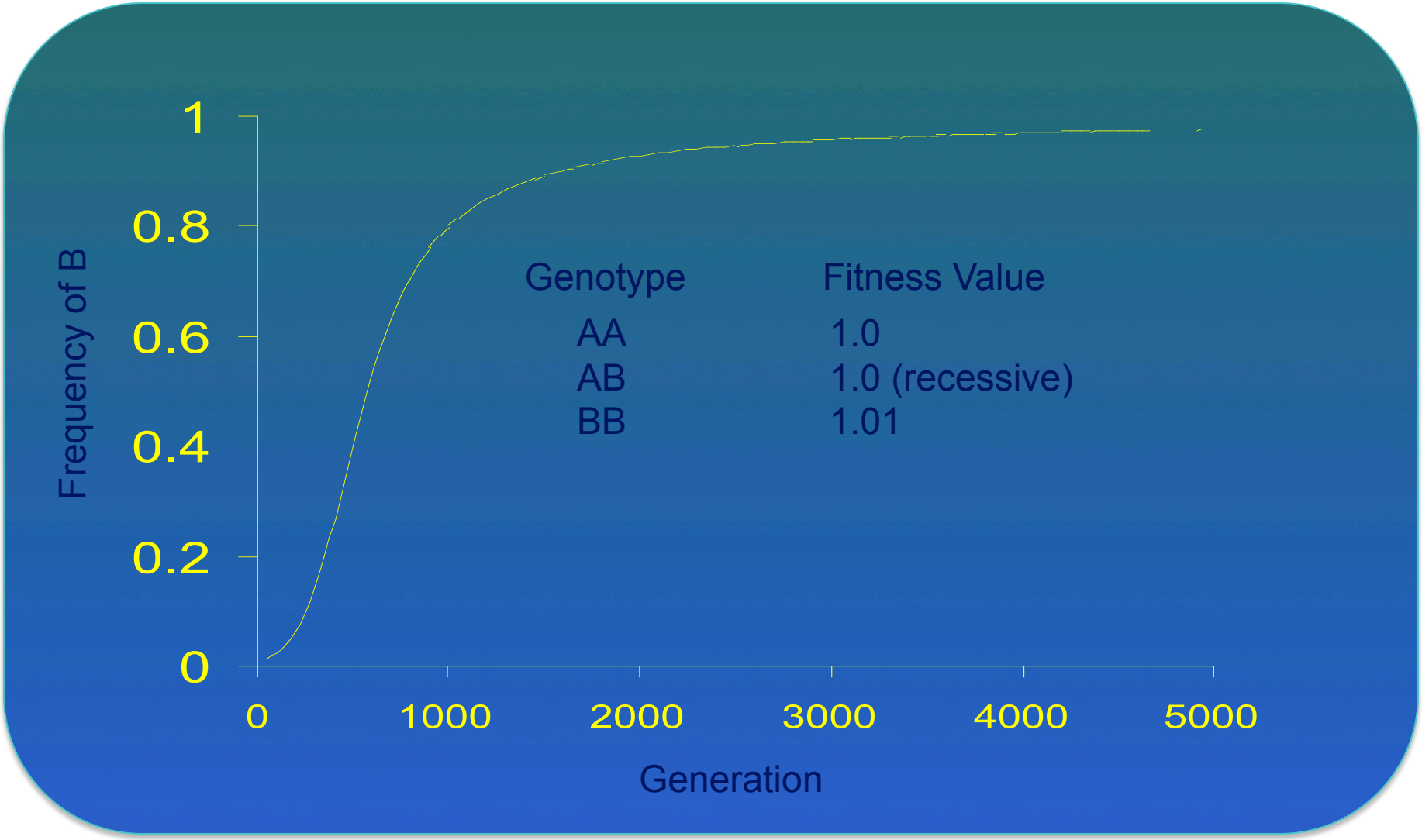
q = frequency of **B**
p = (1-q) = frequency of **A**

∞ population: differential equation for p', q'

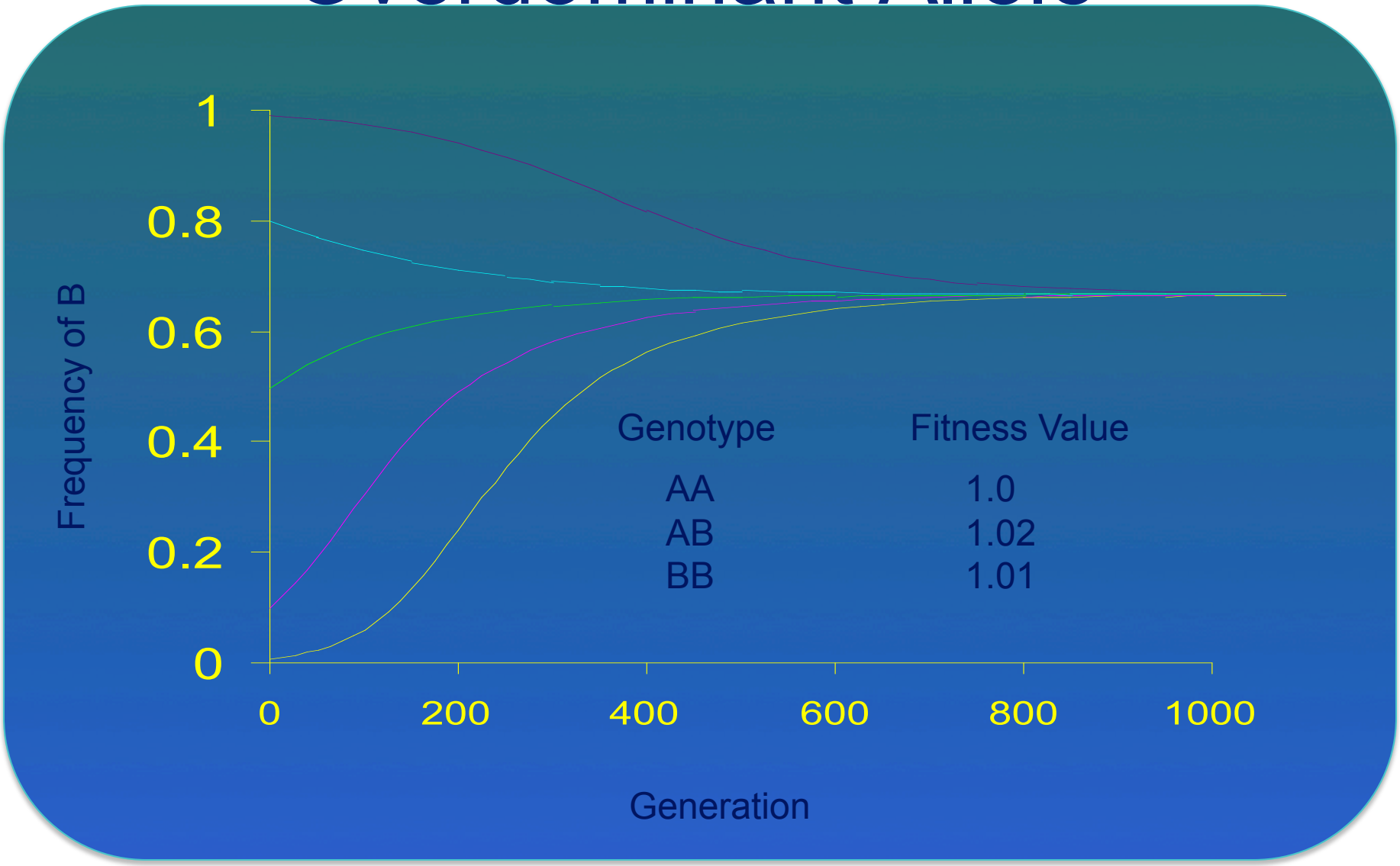
q(next generation)

$$= q(\text{this generation}) + \frac{pq[ps_{AB} + q(s_{BB}-s_{AB})]}{p^2 + 2pq(s_{AB}+1) + q^2(s_{BB}+1)}$$

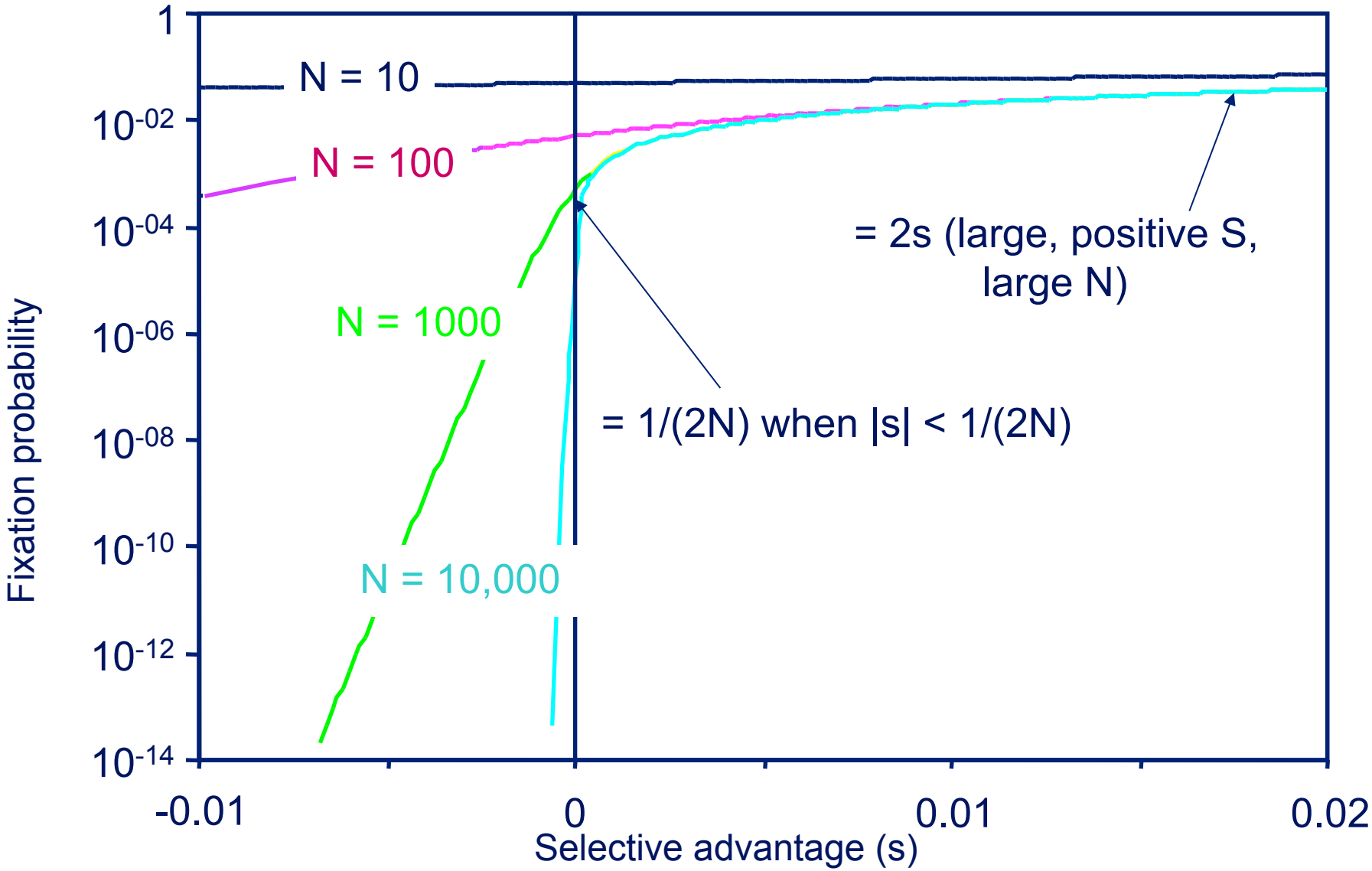
Fixation of an Advantageous Recessive Allele (s=0.01)



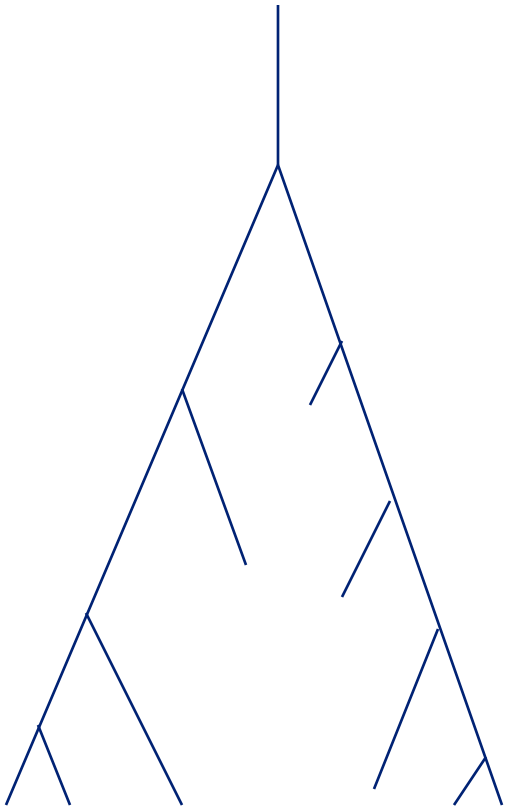
Equilibration of an Overdominant Allele

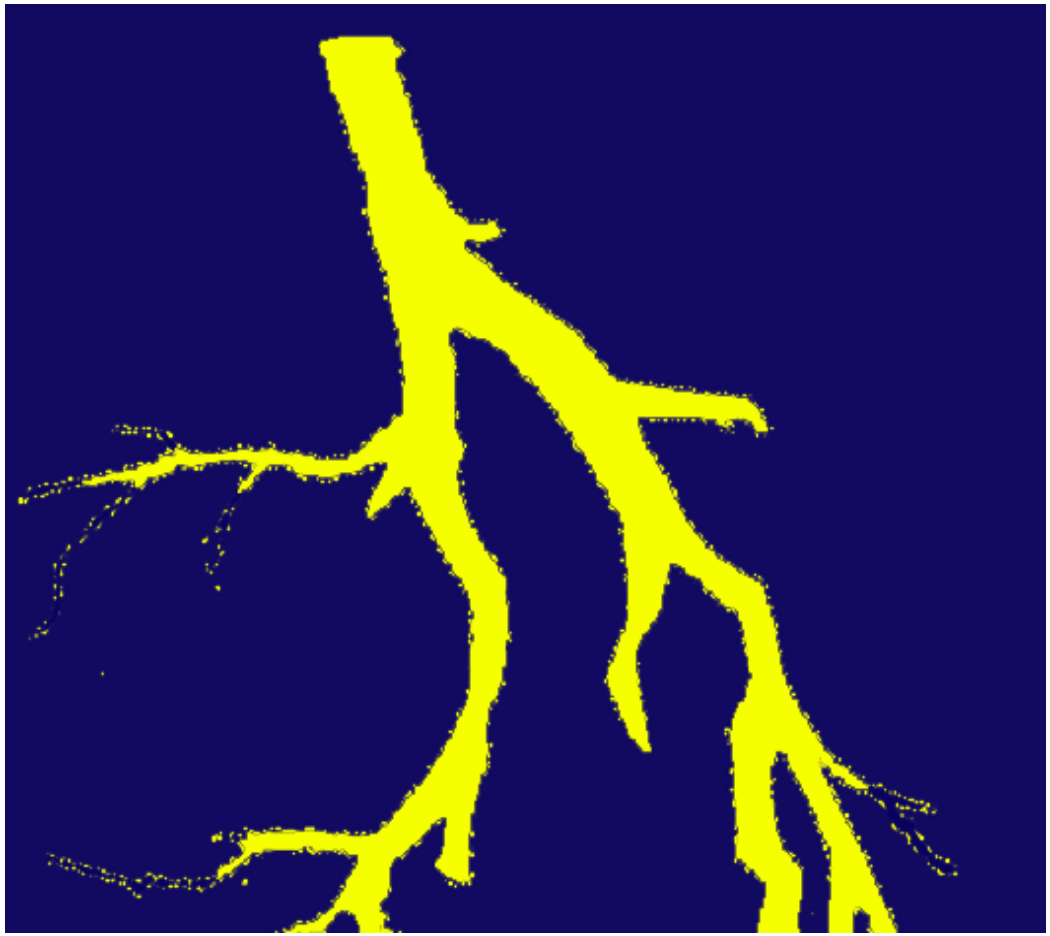


Probability of fixation = $\frac{1-e^{-2s}}{1-e^{-2Ns}}$



Real phylogenetic trees





Different Rates of Substitutions

DNA substitution rate depends on

location in the genome
coding or non-coding
synonymous or non-synonymous
identity and location on protein

Non-coding regions, coding region synonymous substitutions

~ 3-4 x 10^{-9} substitutions/site year

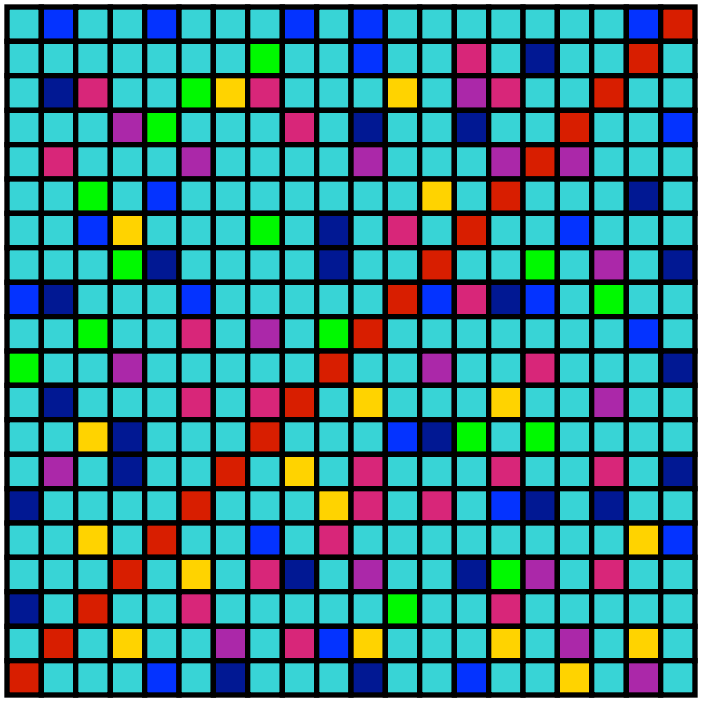
Coding regions, non-synonymous substitutions

Histones	~0
Insulin	0.2 x 10^{-9}
Myoglobin	0.57 x 10^{-9}
γ Interferon	2.59 x 10^{-9}
Relaxin	3.06 x 10^{-9}

Interpreting Evolutionary Changes Requires a Model

…IGTLS…

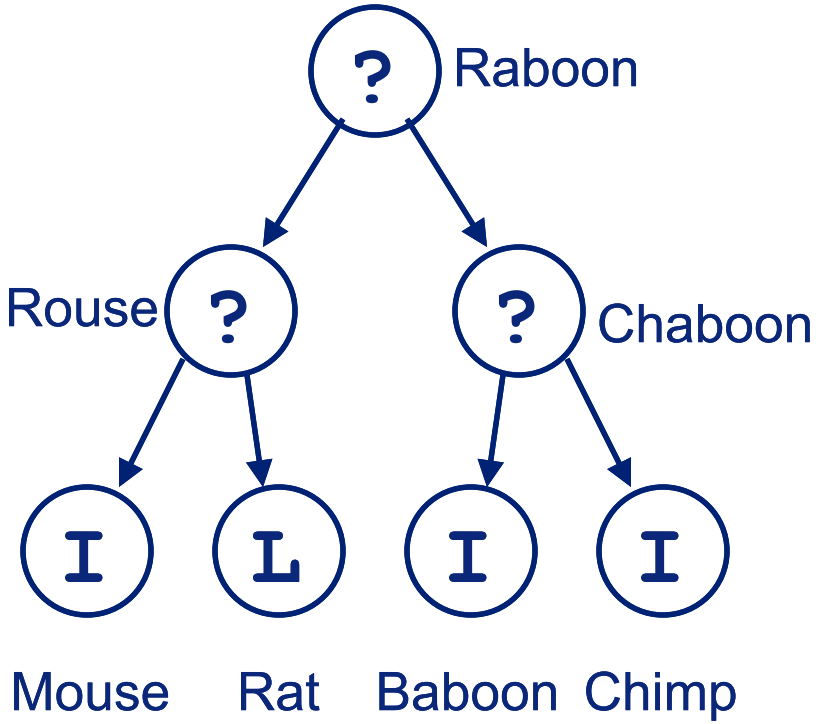
…IGRLS…



In evolution:
what is the rate R(T→R) at which **T**s become **R**s?

e.g. 0.00005 / my

Using Current Sequences to Develop the *Evolutionary Model*



Mouse: …TLSPGLKIVSNPL…
Rat: …TLTPGLKLVSDTL…
Baboon: …TVSPGLRIVSDGV…
Chimp: …TISPGLVIVSENL…

Each location

We need to find the best model for the data

Find the Best Model Using Statistical Methods

In the absence of other information, the best model is the one that maximizes the probability that the data would result **IF** the model were correct

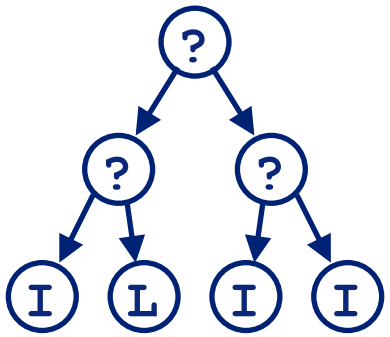
Rev. Thomas Bayes
(1702-1761)

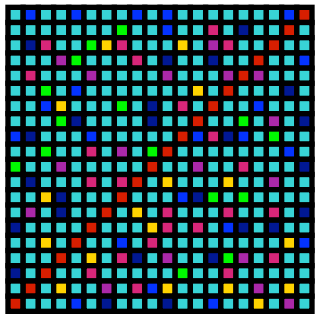
Maximizing the Probability that the Data would Result if the Model were Correct

Maximize **Log Likelihood** or **Posterior Probability**

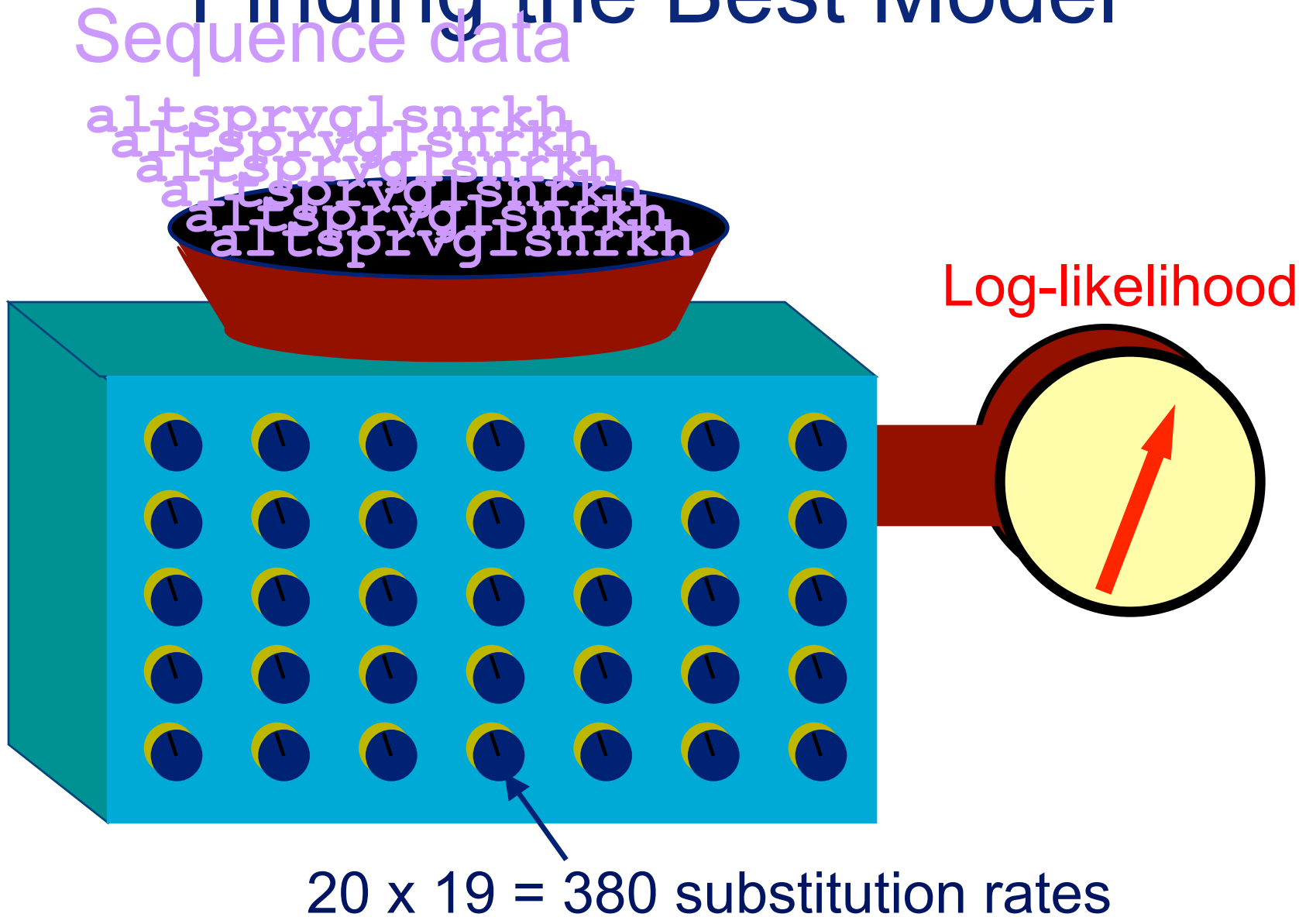
Log{P(Observed data|Evolutionary Model)}

$$= \sum_{\text{locations}} \log \{ P(\text{tree} \mid \text{data}) \}$$

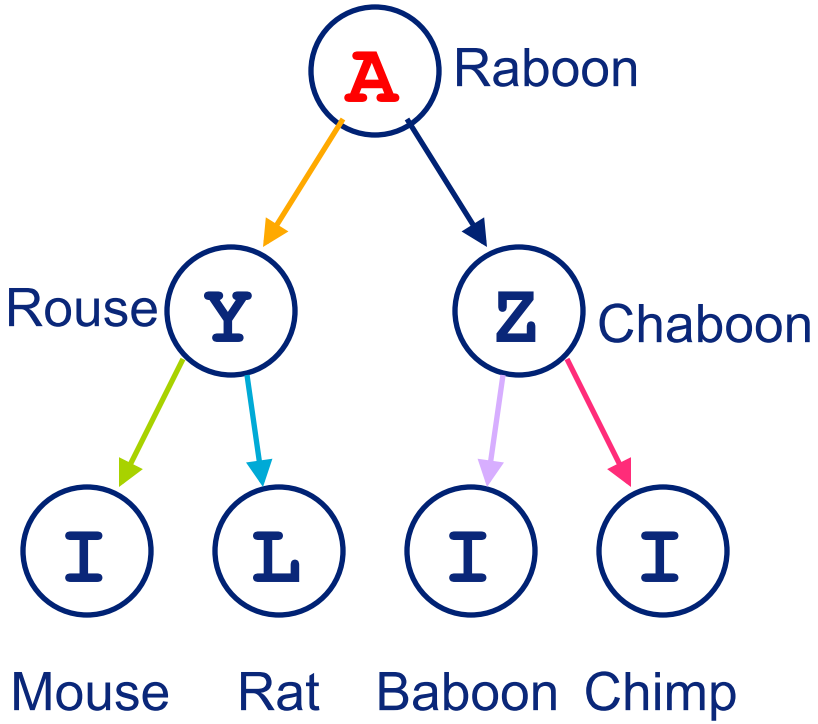




Finding the Best Model

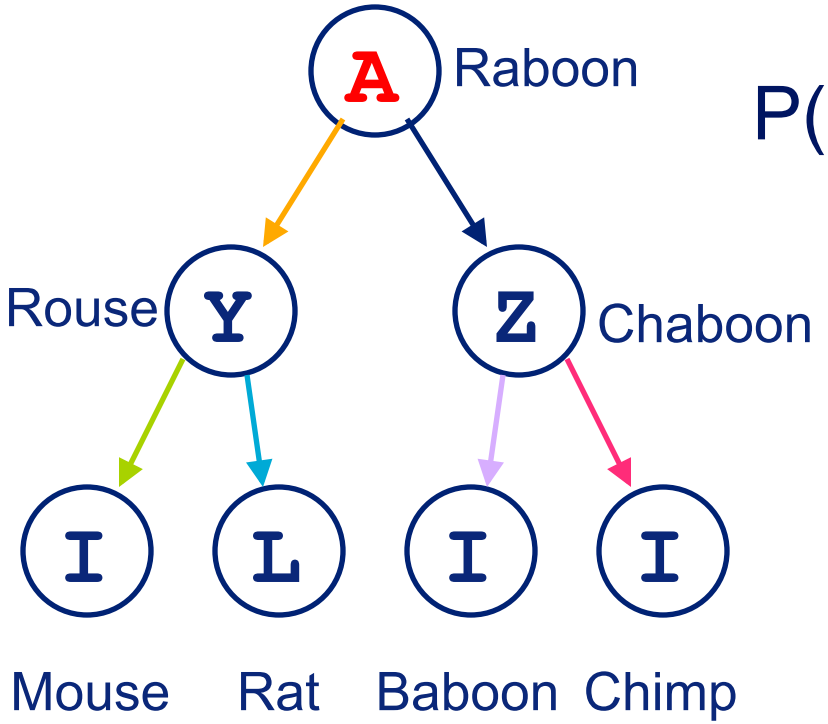


Reconstruction of Ancestral Proteins



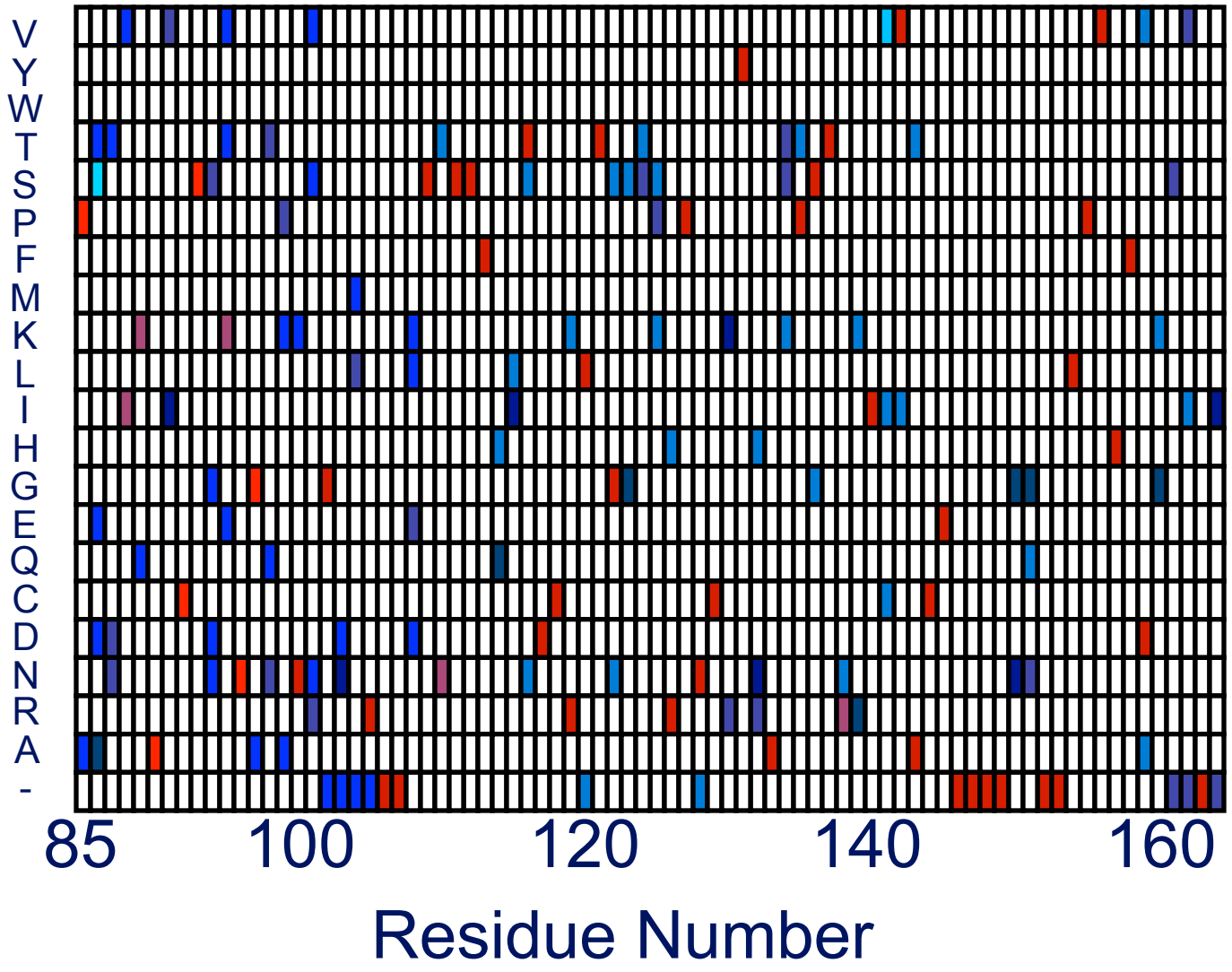
What is the probability that the Raboon had an A at this position?

Reconstruction of Ancestral Proteins

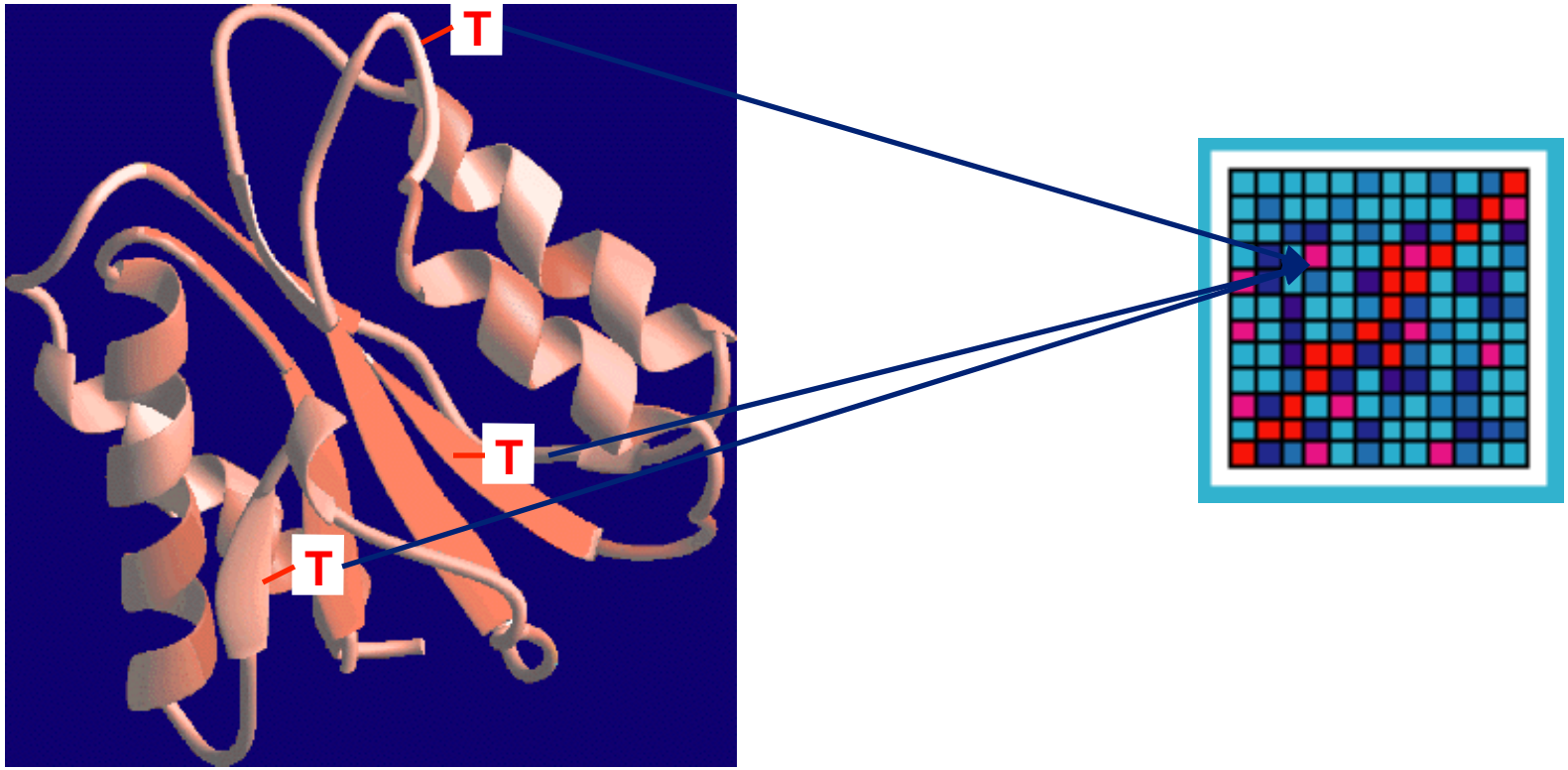


$$P(\text{Raboon had an A} \mid \text{model}) = \frac{\sum_{\text{All possible paths starting with A}} P(\text{Data} \mid \text{Model})}{\sum_{\text{All possible paths}} P(\text{Data} \mid \text{Model})}$$

Probabilistic Reconstruction

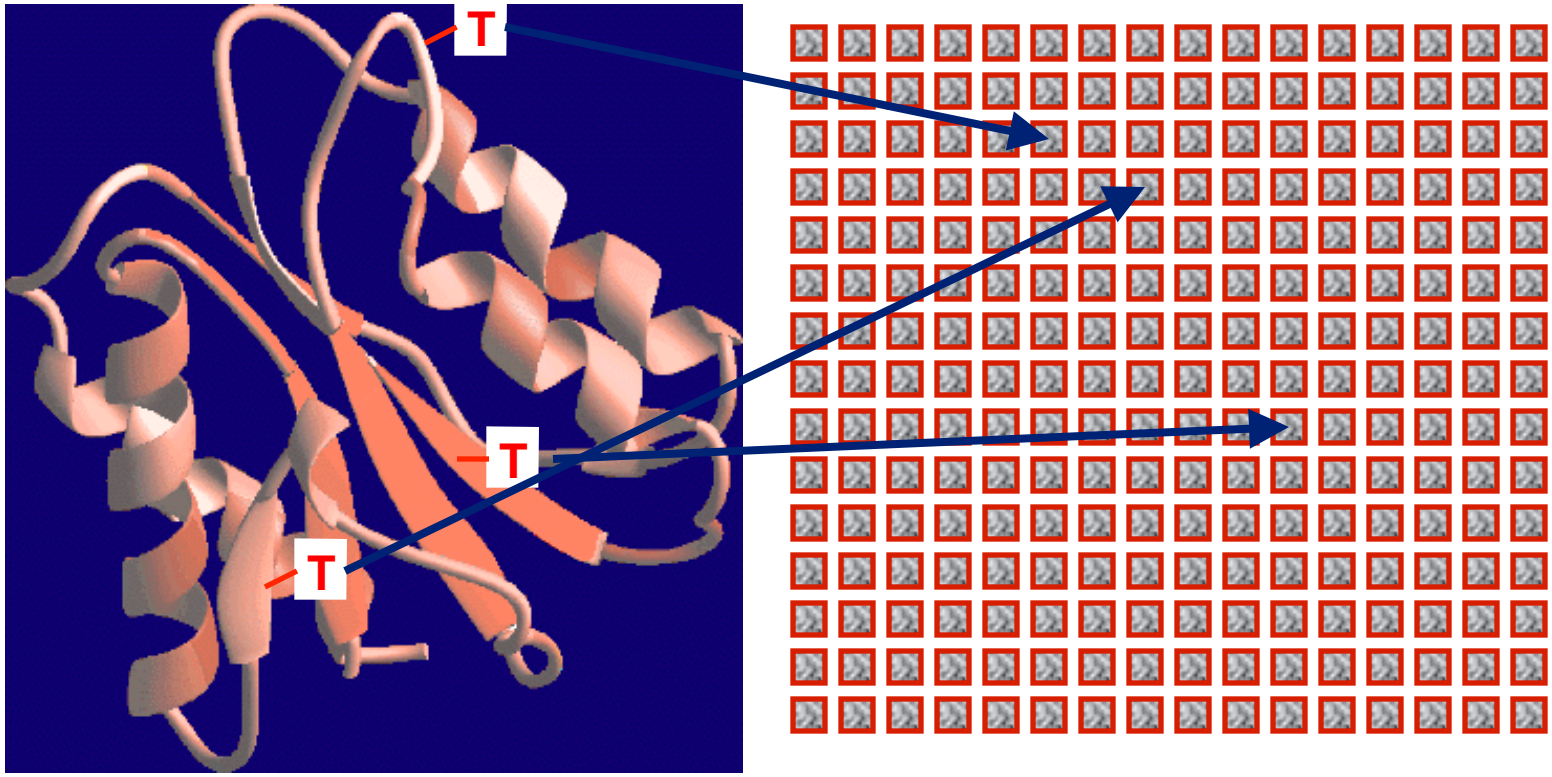


Assumption: R(T→S) is the Same For All Locations



Same for: inside, outside, helix, sheet, coil, active site, dimerization site ...

We Would Like Separate Substitution Matrices for Each Location

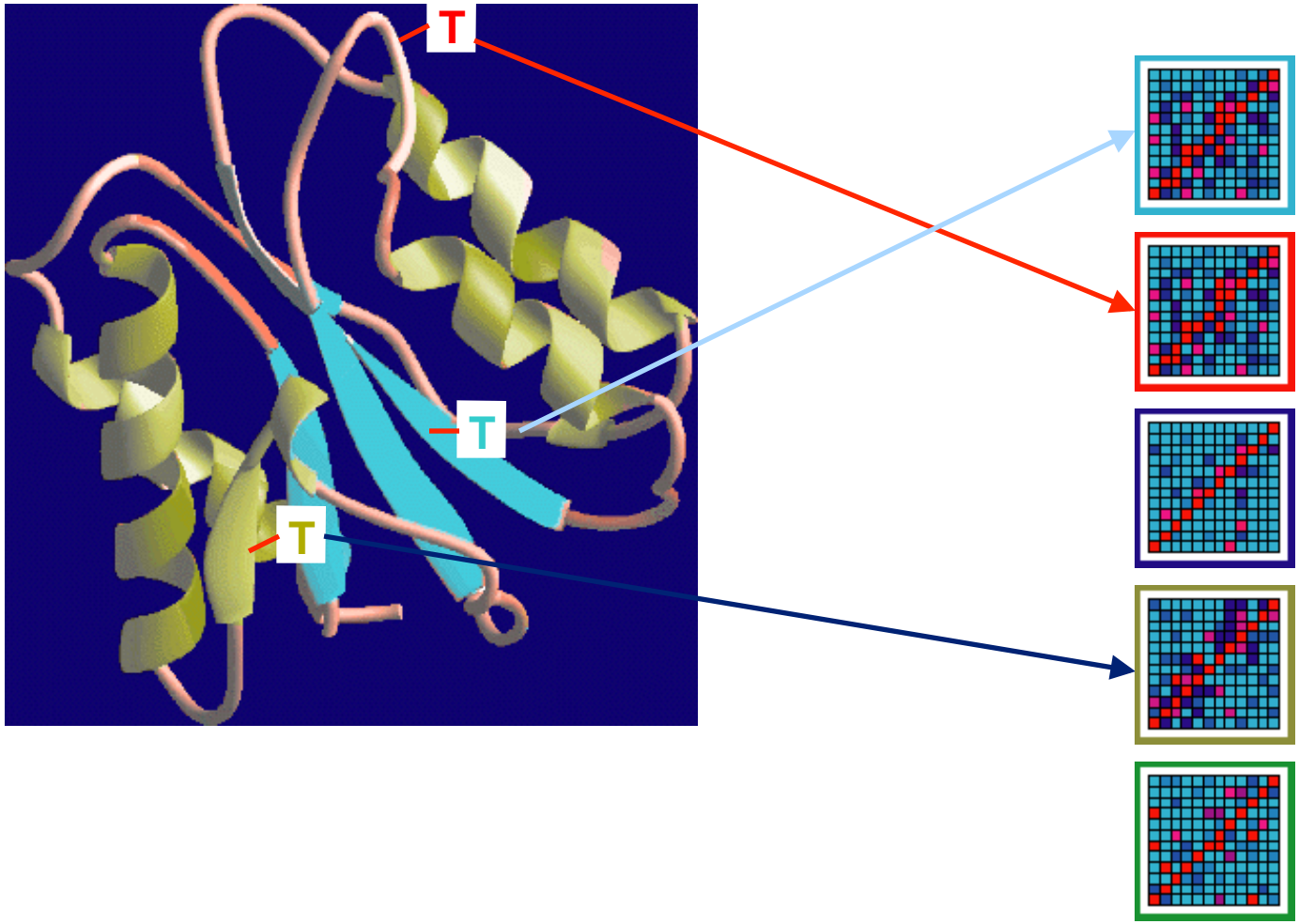


380N adjustable parameters!

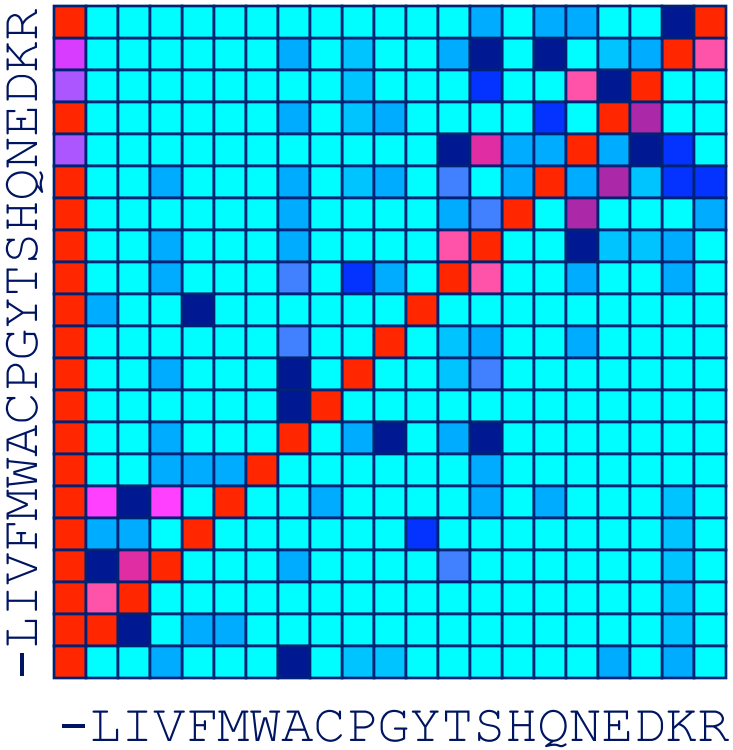
N is the number of residue positions

Proteins have Structure

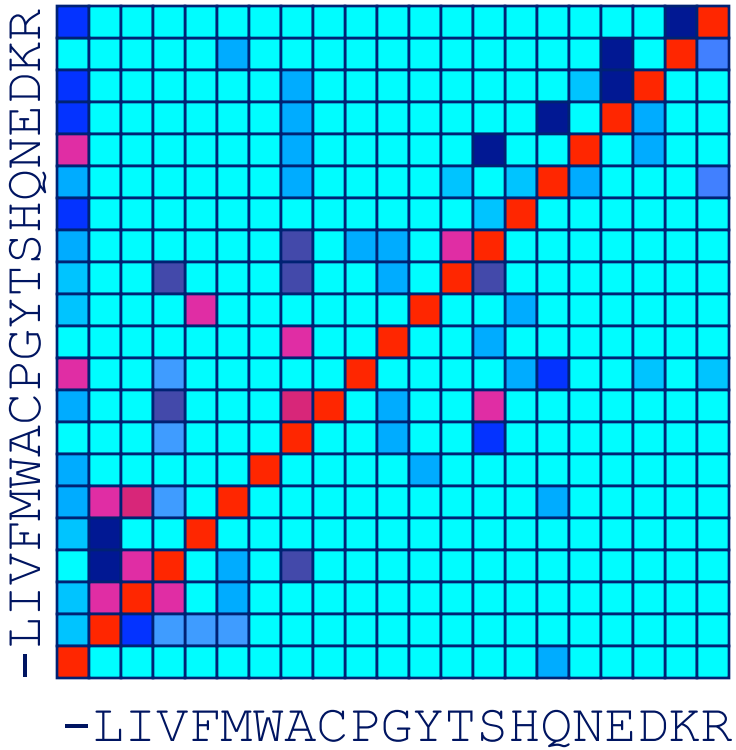
Different Matrices for Different Local Structures



Exposed Coil

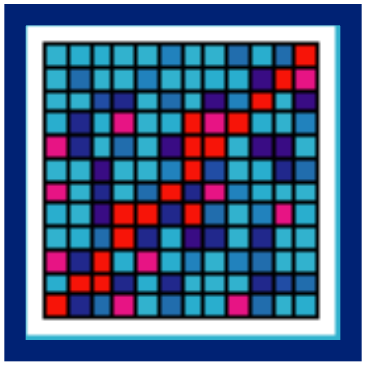


Buried Helix

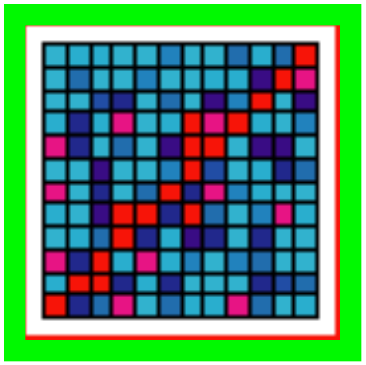


Note difference in gap creation

Buried Helix

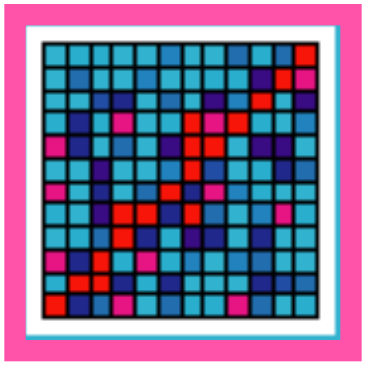


Buried Sheet

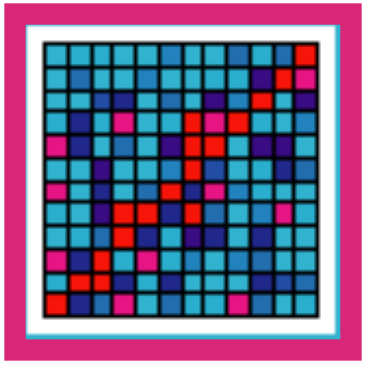


Buried Turn

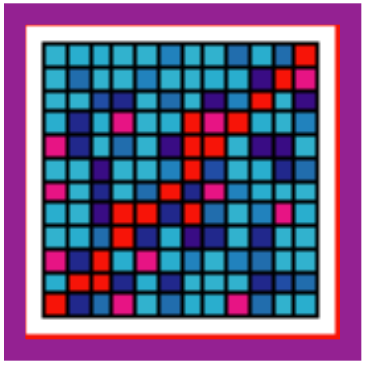
Buried Coil



Exposed Helix

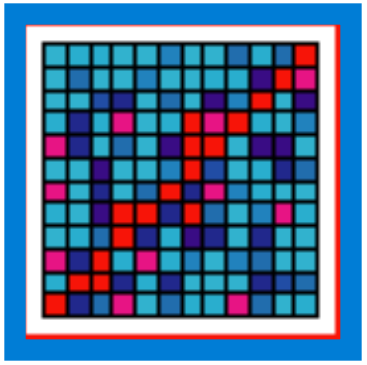


Exposed Sheet

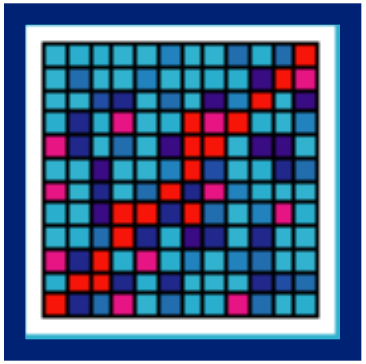


Exposed Turn

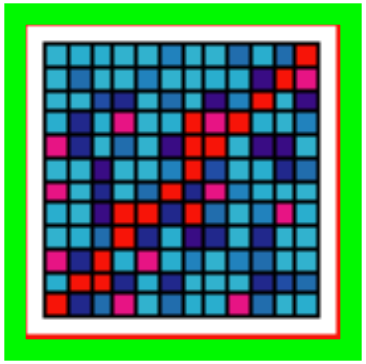
Exposed Coil



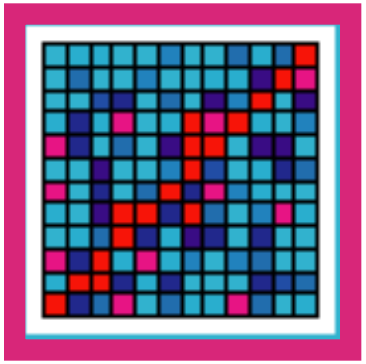
Buried Mesophile



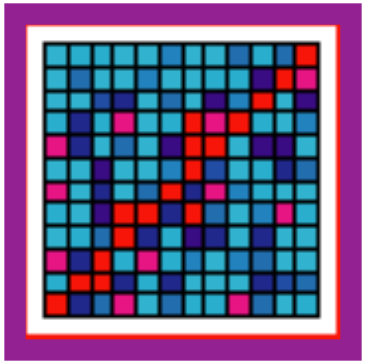
Buried Thermophile



Exposed Mesophile



Exposed Thermophile



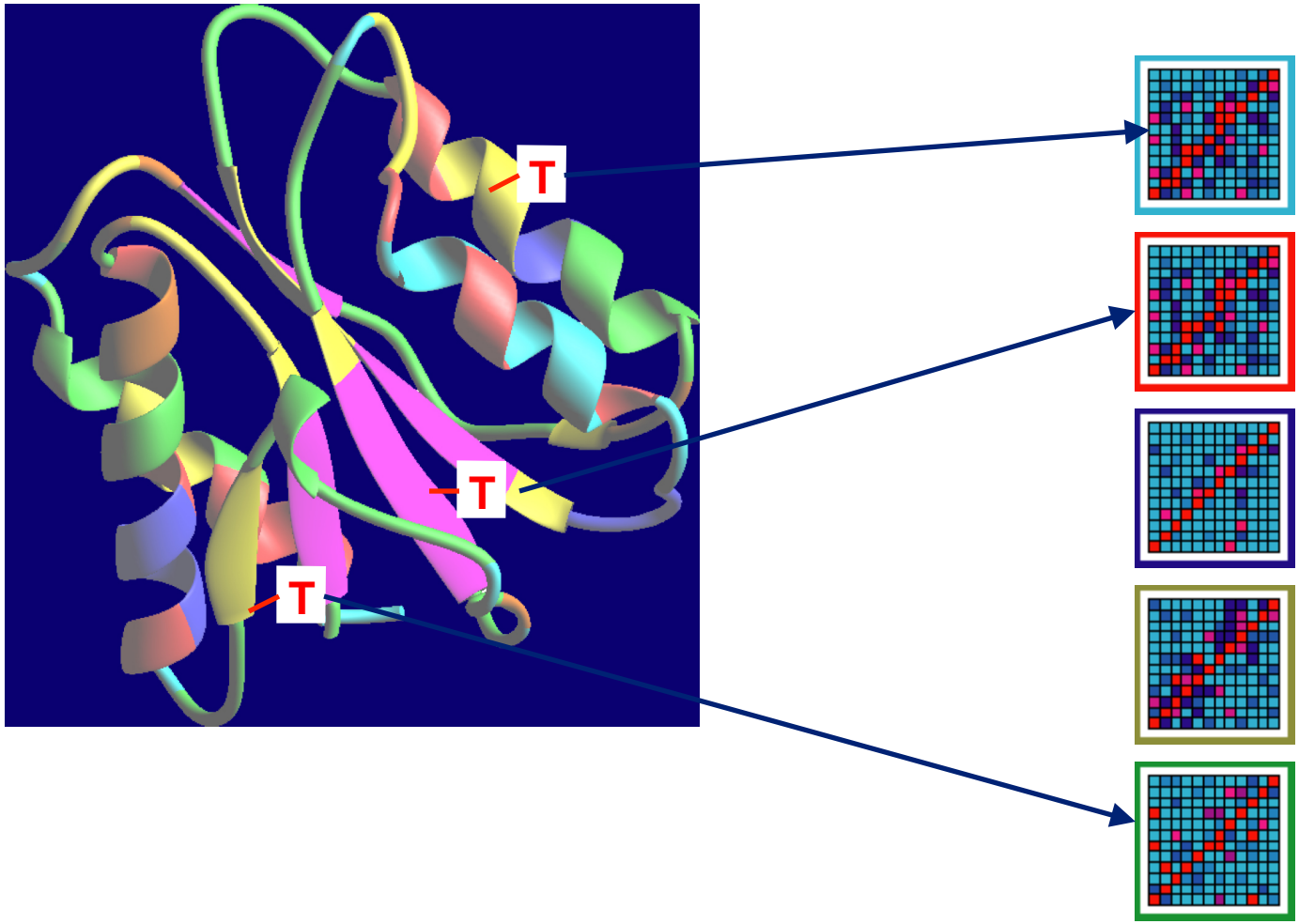
Is This Enough?

Assumes all locations in a given local structure evolve identically

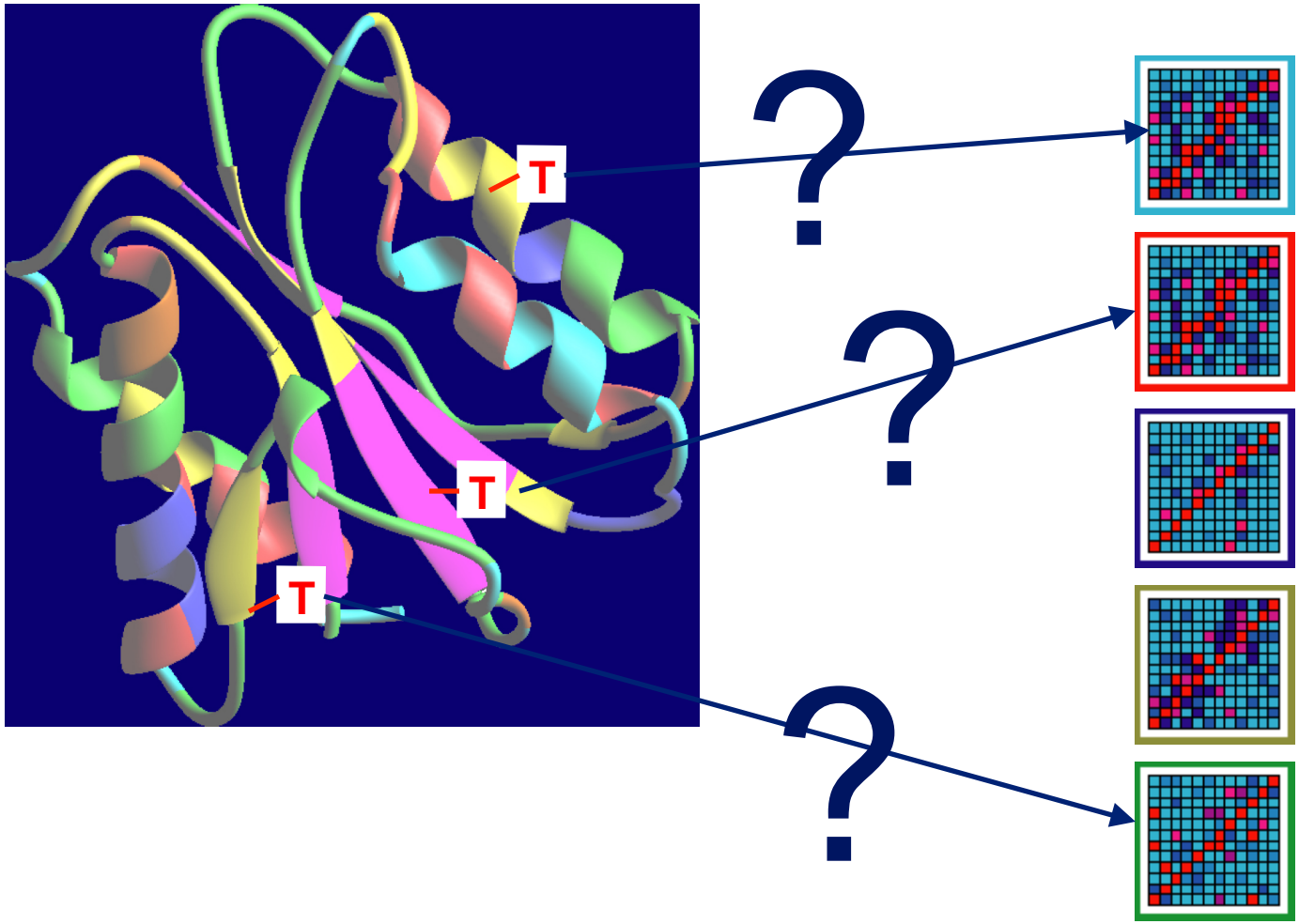
- Ignores complex nature of structural constraints
- Ignores functional constraints
  - active sites
  - dimerization sites
- Ignores any other type of selective pressure
- Designation between local structure categories somewhat arbitrary
- What about proteins of unknown structure?

Different “Site Classes”

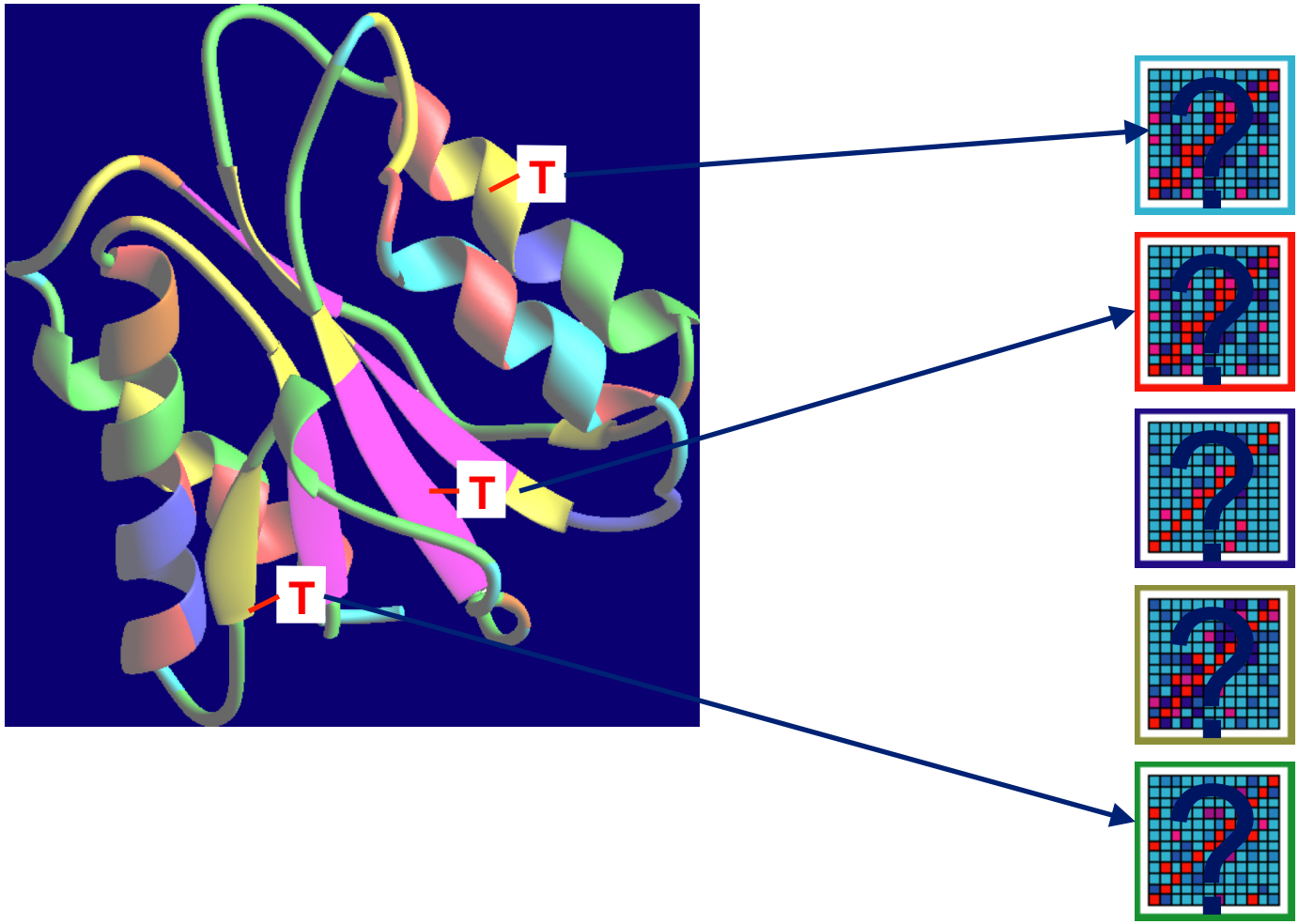
Each with its own matrix



We Don’t Know Which Locations Belong to Which Site Classes…



…Or the Matrices Corresponding to These Site Classes



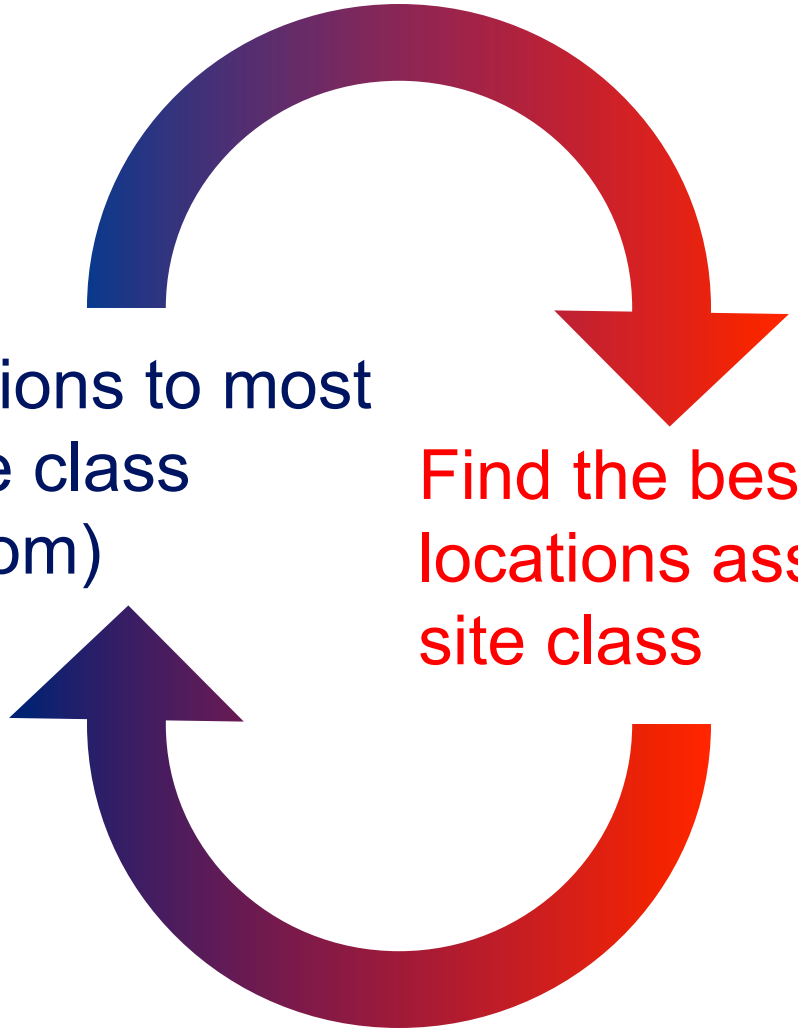
If we knew which locations in the protein belonged to which site classes, our troubles would be over

What is the best model (max Log Likelihood) for the locations in this site class.

If we knew what the set of models were, our troubles would be over

Which model fits each location best (max Log (Likelihood x P(that model))?

Solution: Iterate



Assign all locations to most appropriate site class (at first at random)

Find the best model for the locations assigned to each site class

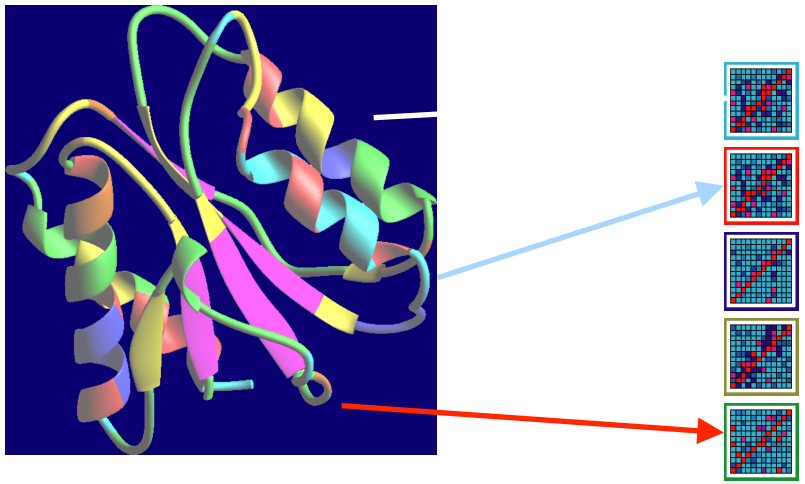
Don't know:

- Substitution models
- Which location fits which model

Site Class	Presence	Overall rate	R(K →F)	R(F →K)	Common AA
2	18%	Slow	Moderate	Rare	Aromatics
3	26%	Moderate	Moderate	Slow	Hydrophobes
4	32%	Fast	Moderate	Rapid	Hydrophiles
5	18%	Very Fast	Fast	Speedy	Flexible

Can Identify:

- **Different types of selective pressure**
- **Which locations under which type of pressure**

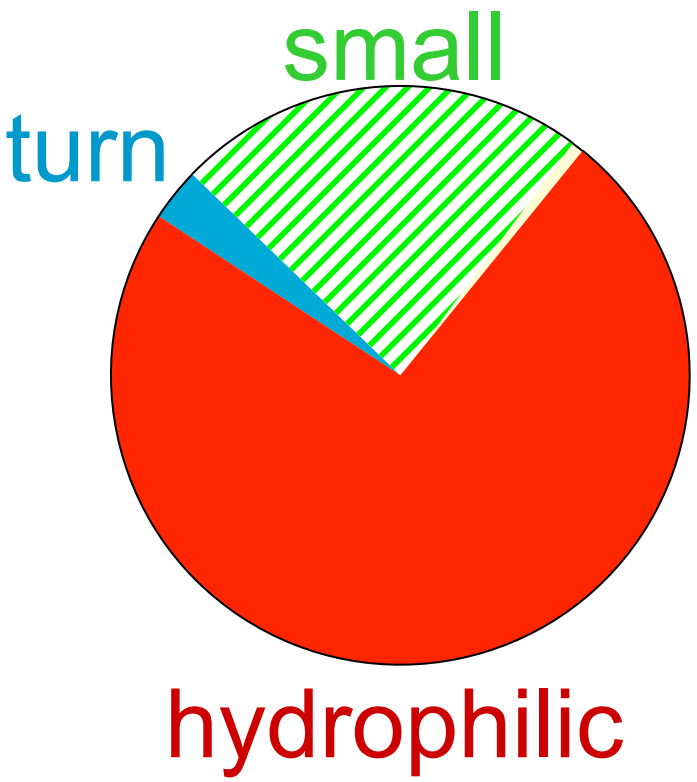


- **Locations under distinctive selective pressure**
- **Changes in selective pressure**
- **Selective pressure that depends upon subclass**
  (identity of ligand, location in cell, etc.)

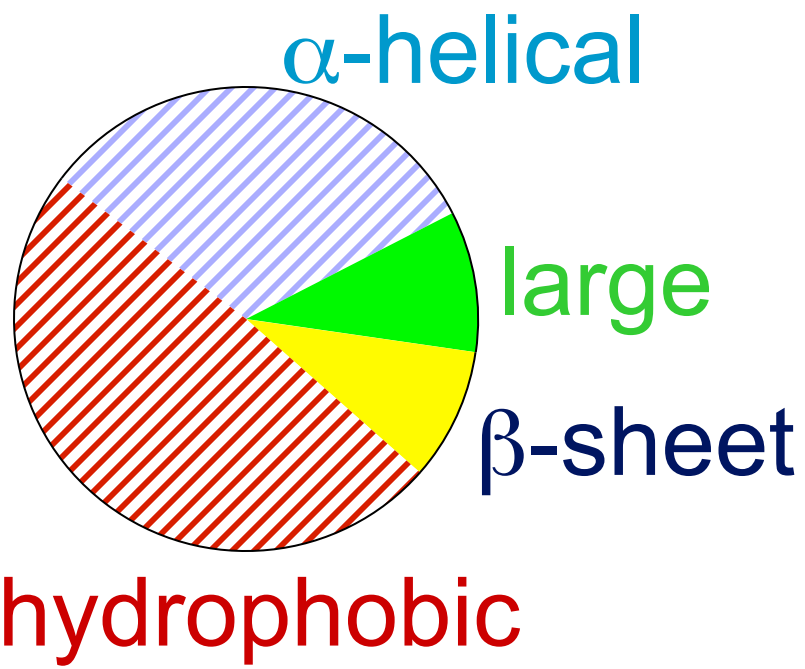
Exposed Locations

Properties of Common Amino Acids

Faster-varying



Slower-varying

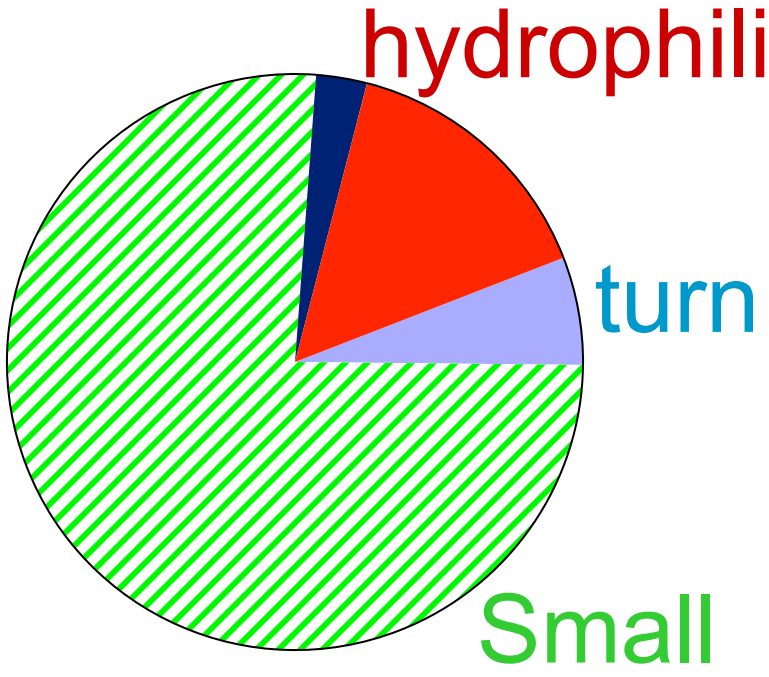


Buried Locations

Properties of Common Amino Acids

Faster-varying

Slower-varying



Two Extreme Views of Evolution

Adaptionists
(Dawkins, etc.)

Every day, in every way,
I'm getting better and better!
- Emile Coue

Neutralists
(Kimura, Gould)

"Nearly-neutral model"

When We Observe Something…

Adaptionists:

- If it exists, in must be an adaptation.
- Why is it necessary/helpful/useful for survival?
- What is its purpose?

Neutralists:

- Random fixation of chance event
- Stochastic processes
- Can reflect number of possibilities (sequence entropy)

Of Course Adaptation Occurs

High selective pressure
Large populations

Of Course Neutral Drift Occurs

Low selective pressure
Small populations (bottlenecks)

~10^{20} Mutations, 10,000 Accepted:
Chance or Necessity?

Adaptionists:

10^{20} unfavorable mutations accepted with probability 0
10,000 positive mutations accepted with probability 1

Neutralists:

10^{20} unfavorable mutations accepted with probability 0
10^{10} neutral mutations accepted with probability 10^{-6}
100 positive mutations accepted with probability 1

Result: 99% of observed mutations are neutral

These numbers, like 64% of all statistics, are made up.

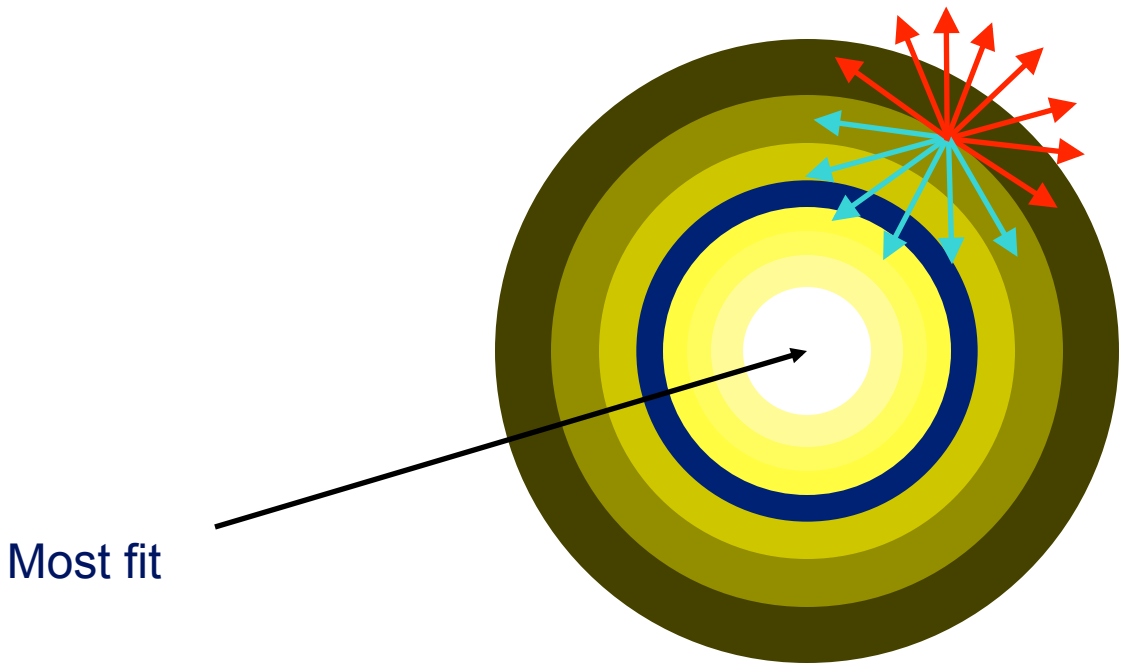
Why is it Difficult to Tell?

- Changes are "neutral" if $|s| < 1/2N_e$
  well below what we can measure in the lab
  not contradicted by DNA, protein plasticity

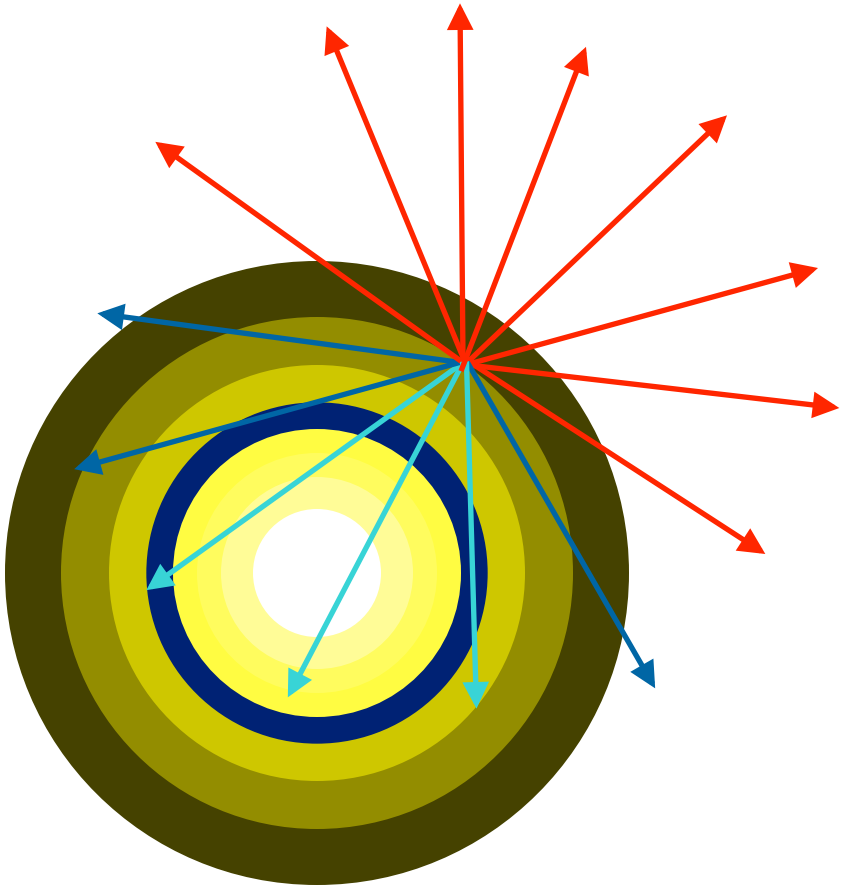
- Many observations are consistent with both models…

  Example: regions that matter "less" (non-coding regions, etc) change faster

Sequence Space



Sequence Space



Reason for Neutral Theory

- Large degree of polymorphism
- High rate of substitutions
- Existence of molecular clock ….

Neutrality and the Molecular Clock?

Adaptive substitutions (s >1/2N):

Population size N, mutation rate μ

$2N$ μ mutations per year

For adaptive mutations

probability of fixation = $2s$

Rate of substitutions = mutation rate * P(fixation)

= $4N$ μs (proportional to N)

Neutral substitutions ($|s|$<1/2N):

Population size N, mutation rate μ

$2N$μ mutations per year

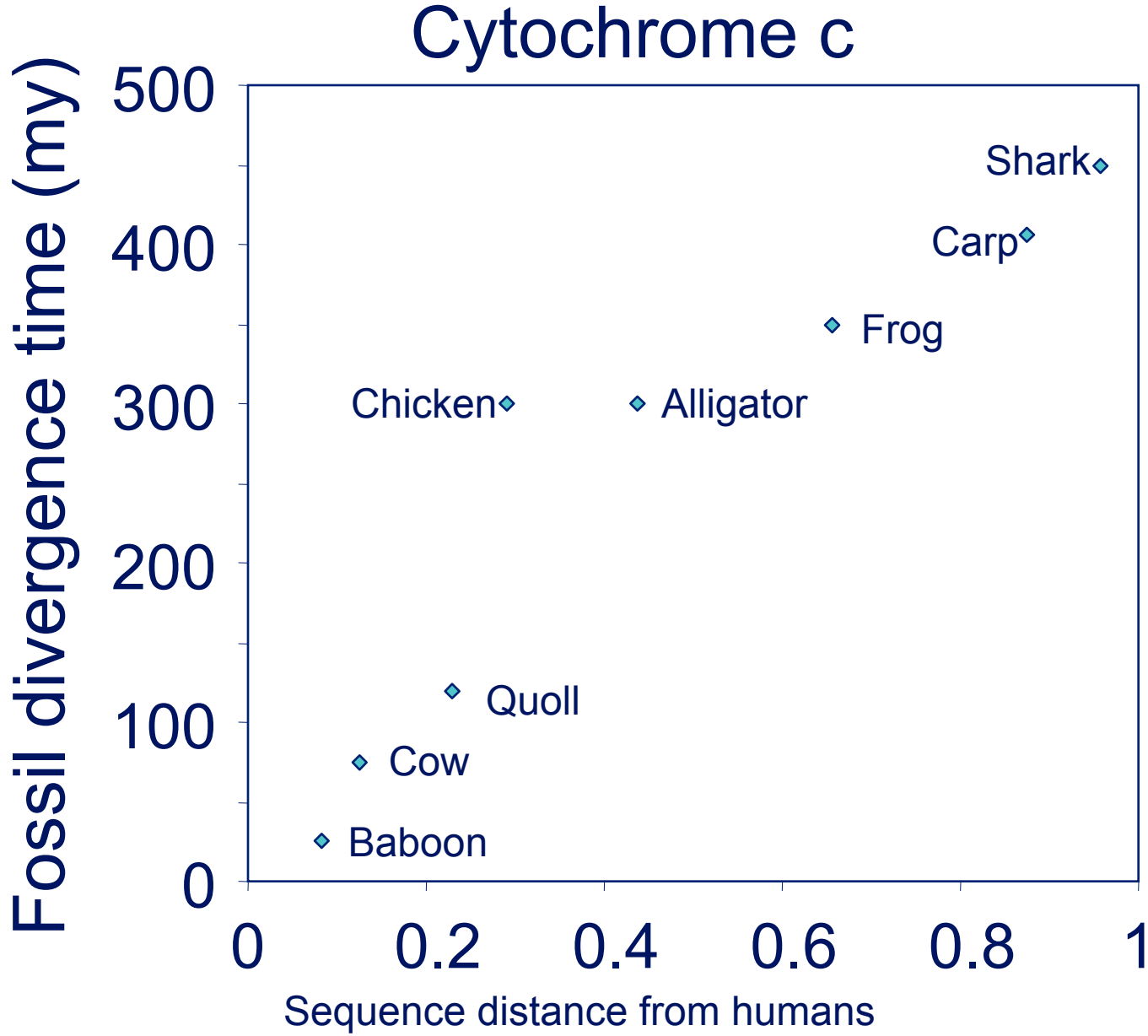
For neutral mutations,

probability of fixation = $1/2N$

Rate of substitutions = mutation rate * P(fixation)

= μ (independent of N)

Evidence for the Molecular Clock



The Molecular Clock is Not Constant

Adaptionists: Ahha!

Neutralists: Other effects:

- If mutations due to germ-line replication, rate should depend upon generation time
- Rate of mutations may depend on metabolic rate (free radicals)
- DNA repair efficiency

Panglossian Paradigm:

"It is demonstrable," said he, "that things cannot be otherwise than as they are; for as all things have been created for some end, they must necessarily be created for the best end. Observe, for instance, the nose is formed for spectacles, therefore we wear spectacles. The legs are visibly designed for stockings, accordingly we wear stockings…"

Voltaire's *Candide*