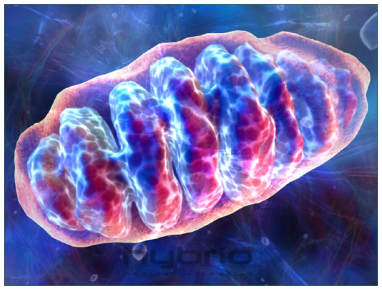
computational bioscience program

university of colorado school of medicine

Evolution of Proteins 2: Proteins 7350

Pollock_ProteinEvol6.ppt

Slides with unpublished data are deleted



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

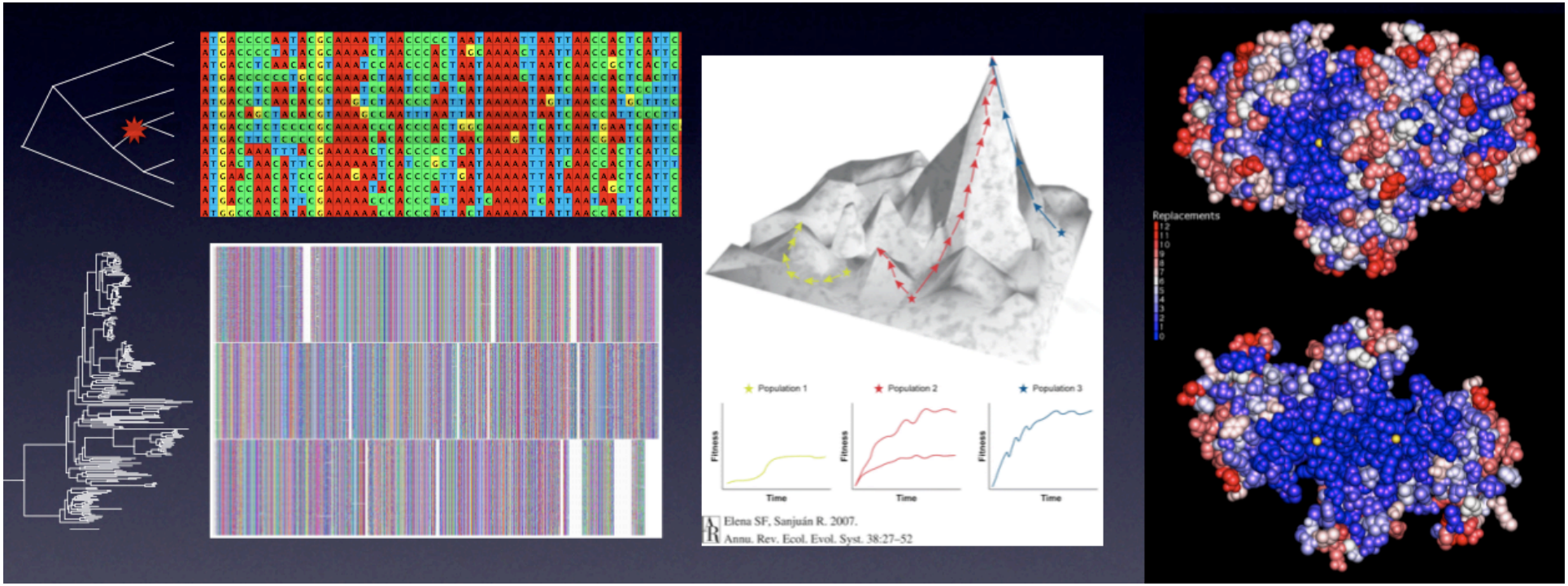
David.Pollock@uchsc.edu
www.EvolutionaryGenomics.com

Evolution Genomics Biodiversity

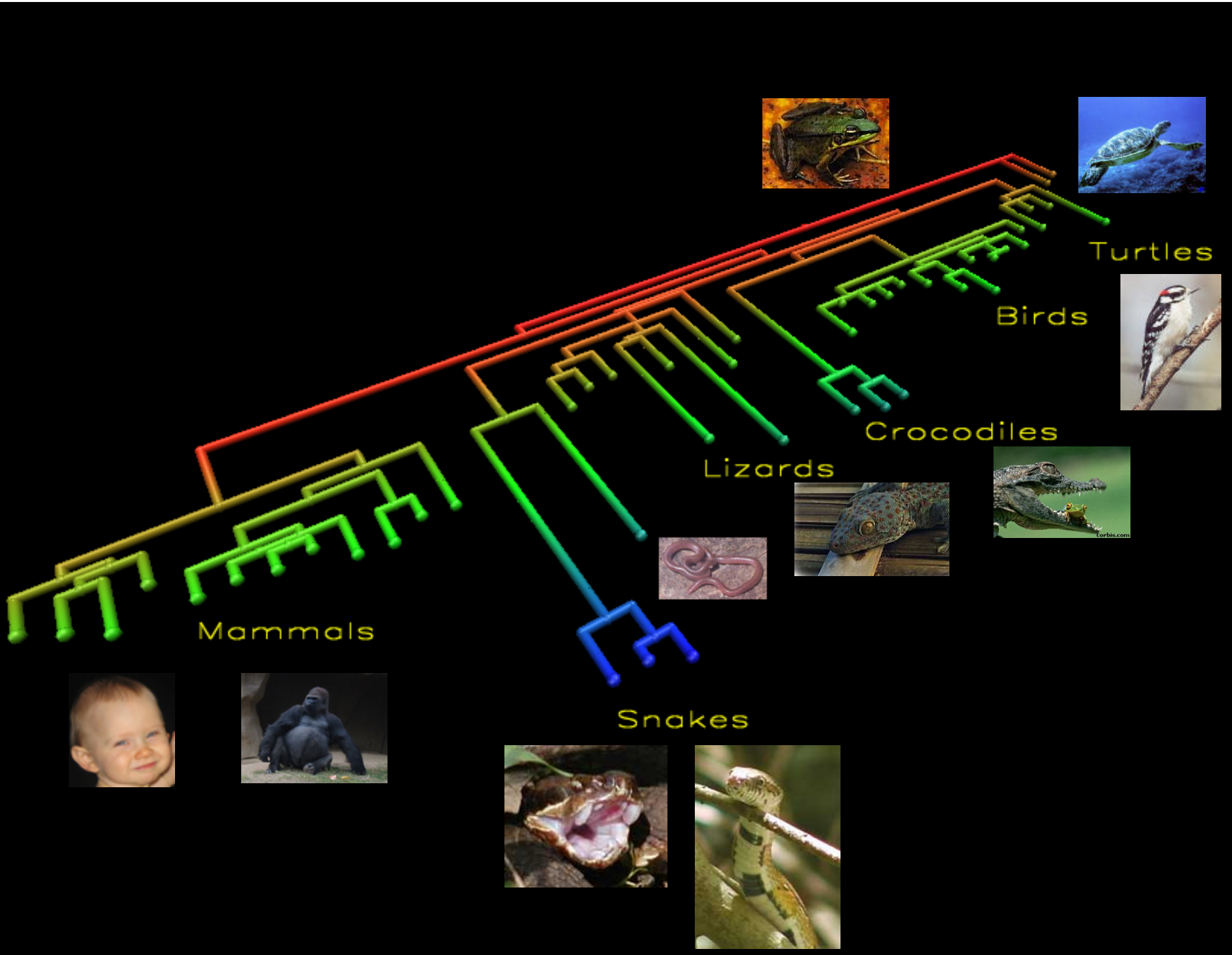
Overview

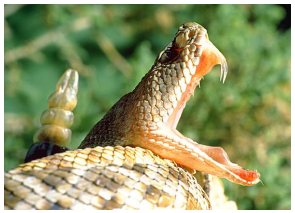
- Explanation of what we do
- Mitochondrial genomics: a pilot project
- PLEX: Context-dependent evolutionary genomics in a practical time frame
- Coevolution of transcription factors and their binding sites: an example

Evolution of Proteins

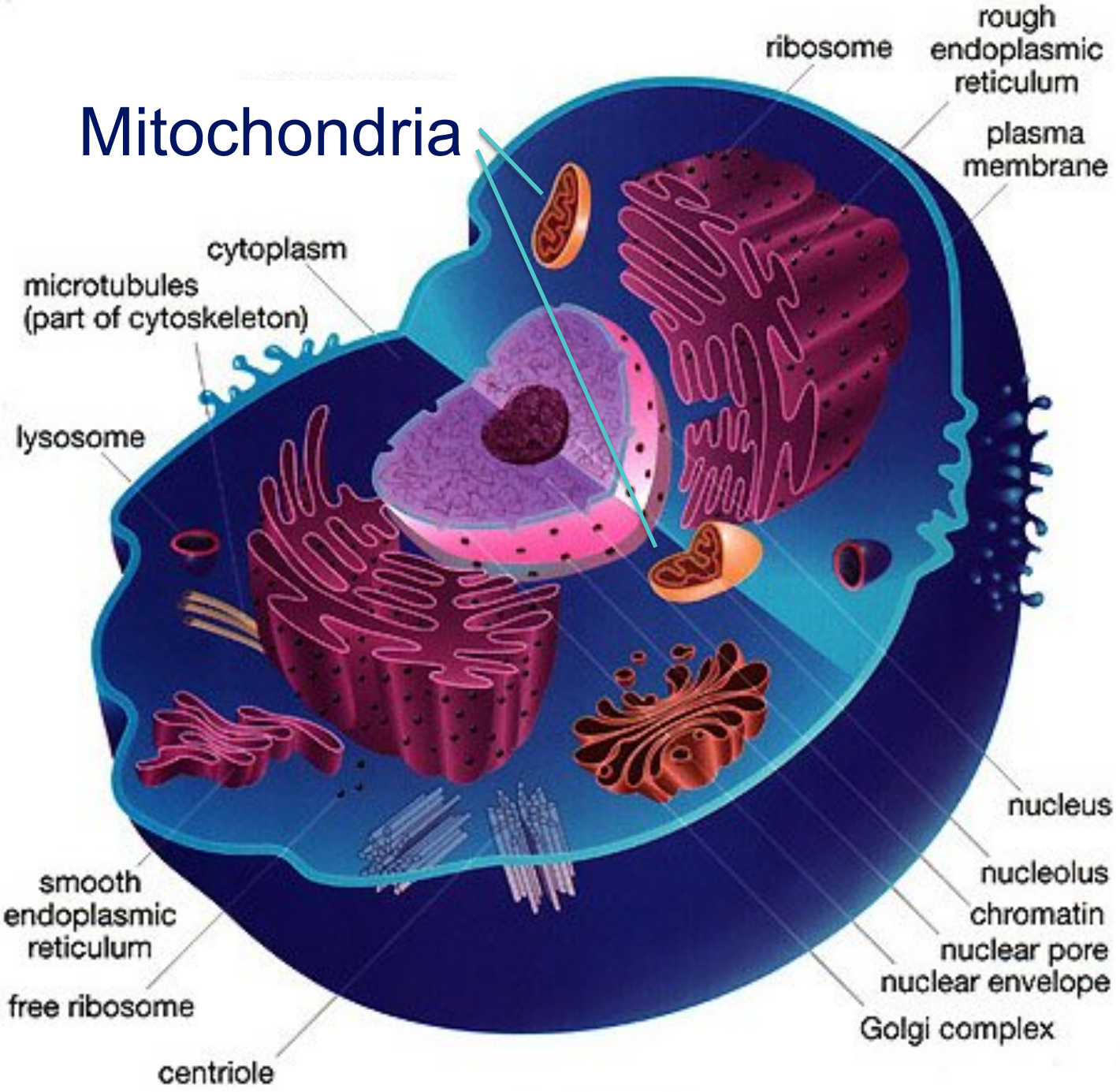


Jason de Koning

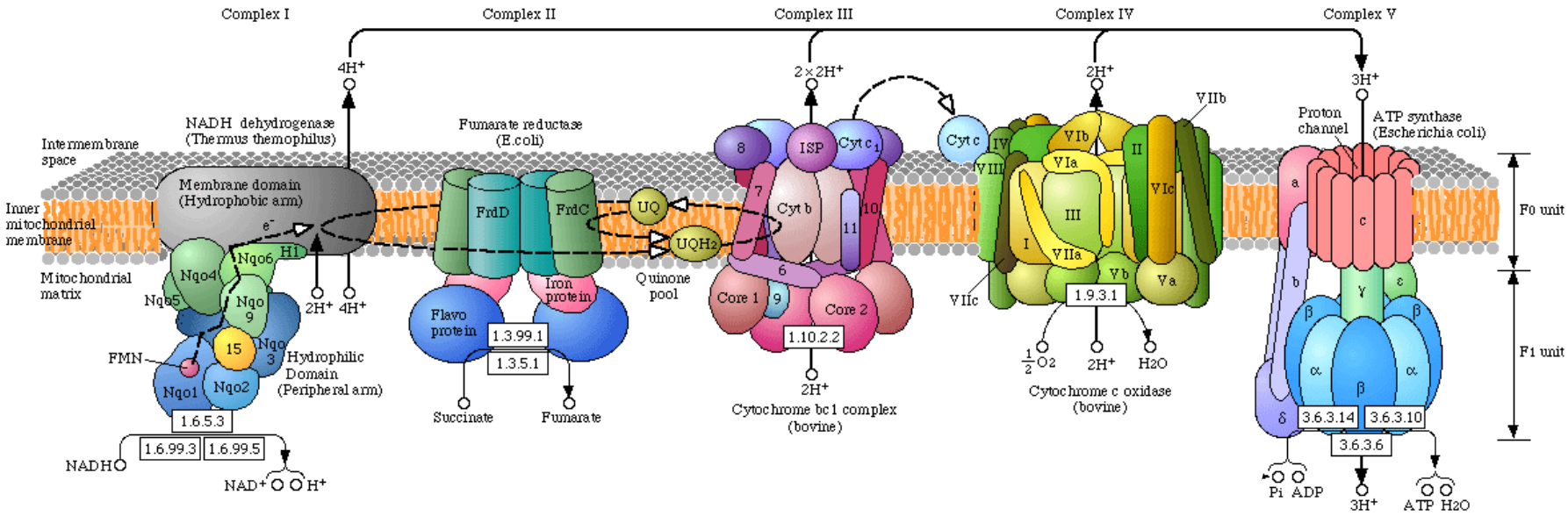

Turtles
Birds
Crocodiles
Lizards
Mammals
Snakes



Mitochondria



Oxidative Phosphorylation



Nuclear	35	4	10	10	12
mtDNA:	7	0	1	3	2

Molecular Evolution

Structure, Function & Rates

- Conserved sites correspond to structurally or functionally important residues
- Changes in evolutionary rates correspond to:
  - Loss of function
  - Altered function (functional divergence)
- Synonymous rates (dS) are compared to non-synonymous rates (dN) as a “neutral” standard.

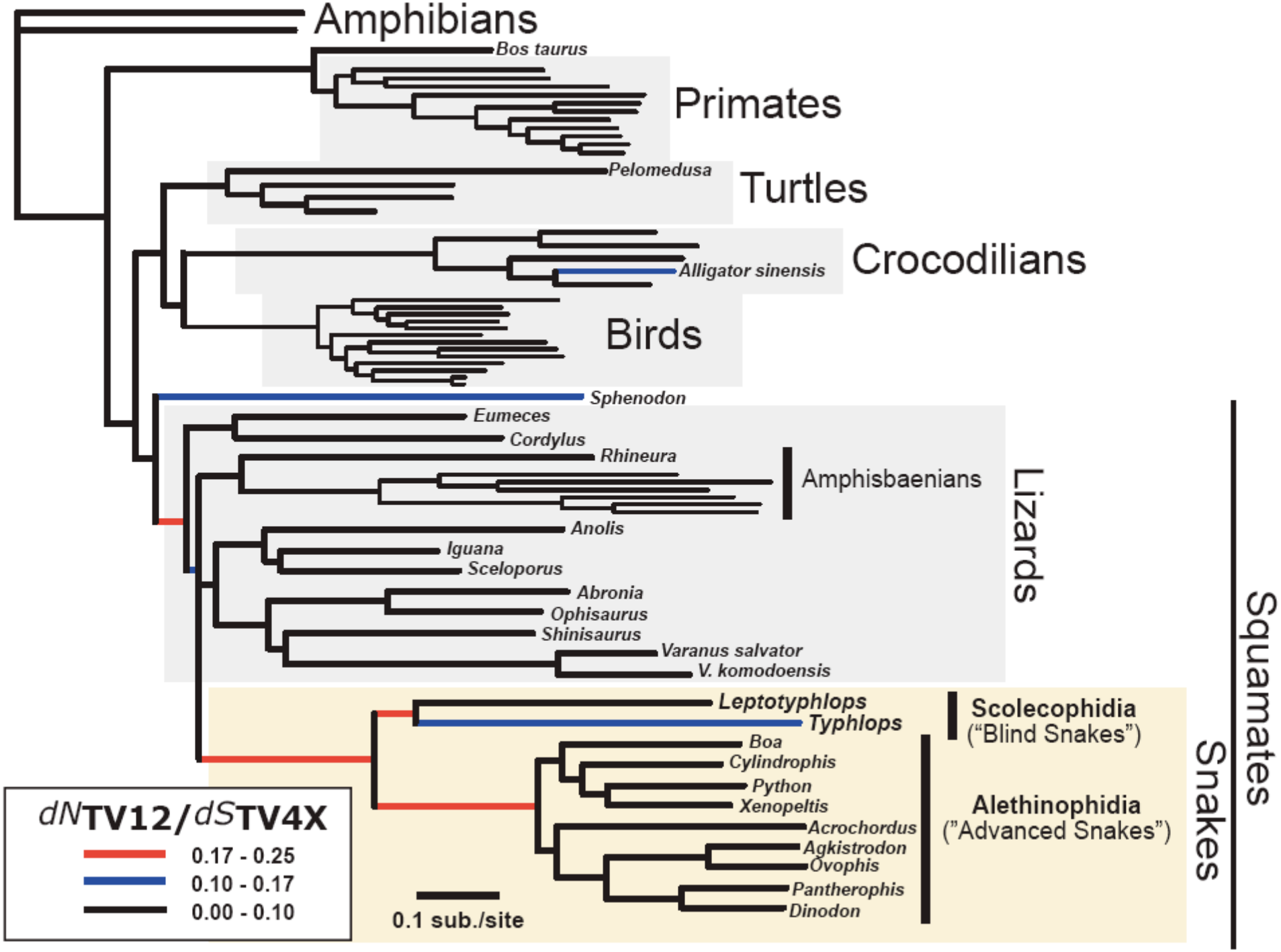
$$\mu N * \frac{1}{N} = \mu$$

- Convergence is a sign of adaptive importance, and is rare at the molecular level
- Coevolution is usually distributed among many sites, often weak

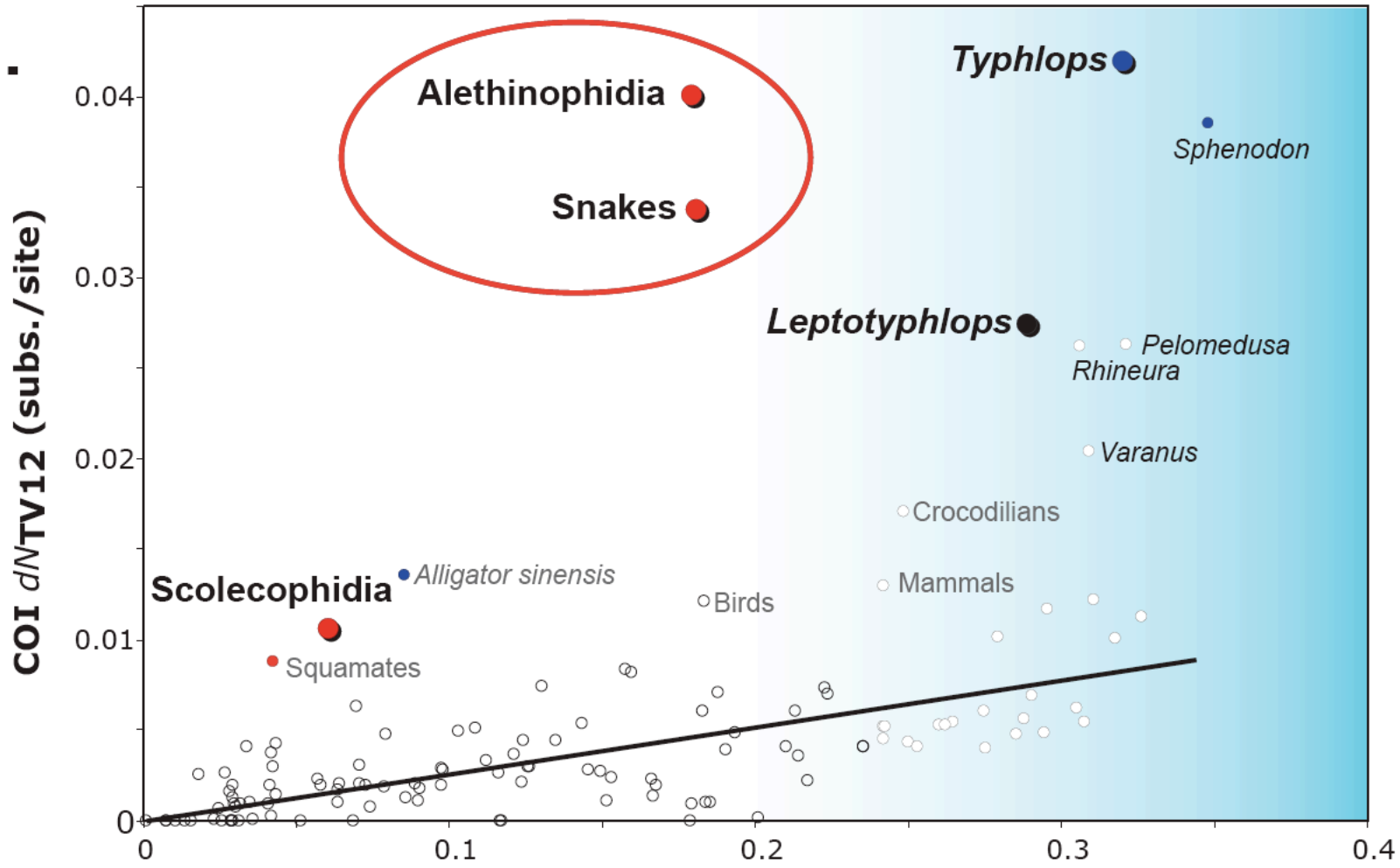
Positive Selection in Snake Mitochondria

- Standard programs strongly indicate positive selection in proteins throughout the mitochondria
  - Especially cytochrome oxidase subunit I and cytochrome b (the hearts of complex IV and III)

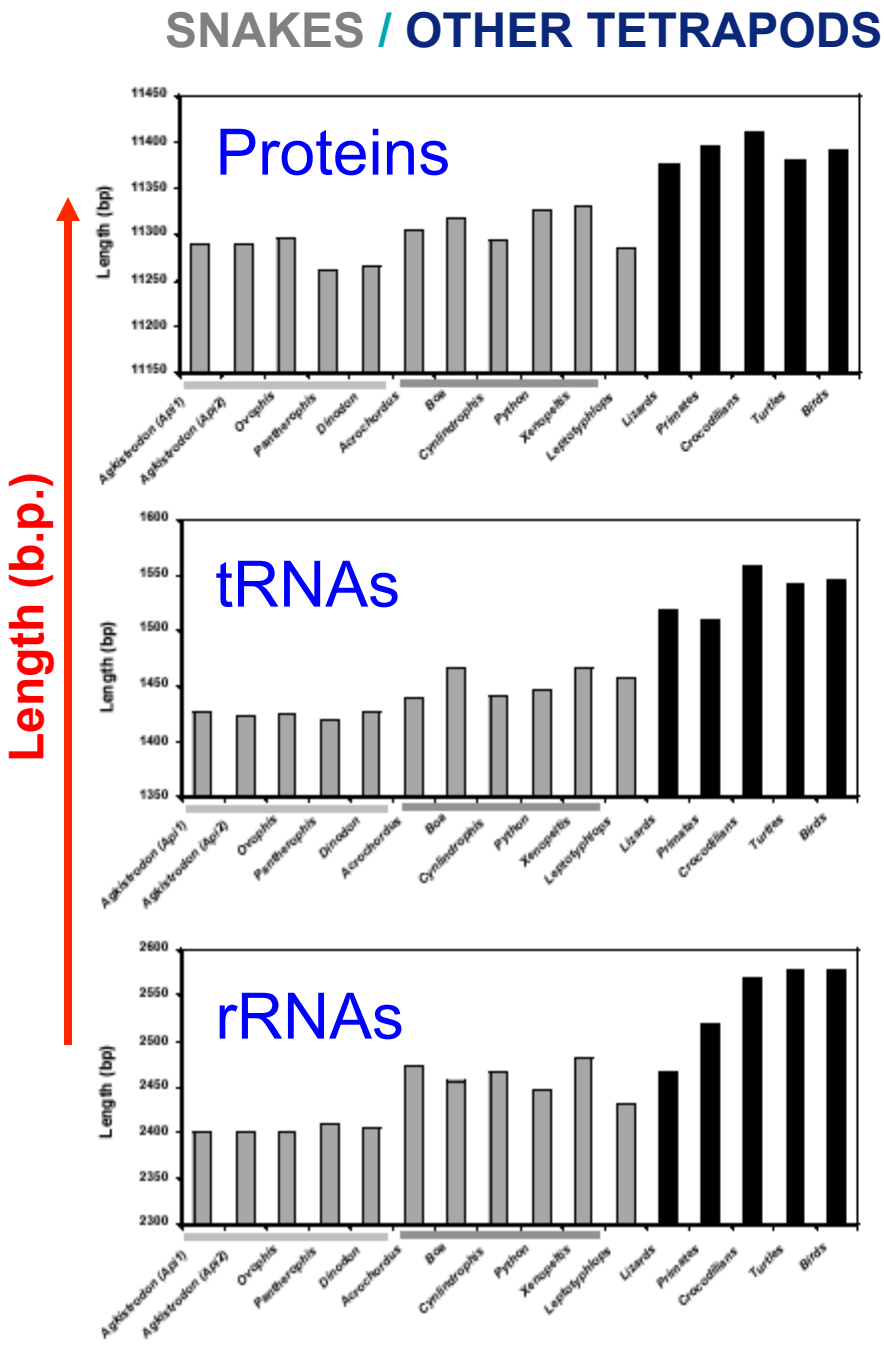
- Concern over “saturation”, inaccurate models
  - Focus on transversions


Amphibians
Bos taurus
Primates
Pelomedusa
Turtles
Alligator sinensis
Crocodilians
Birds
Sphenodon
Eumeces
Cordylus
Rhineura
Amphisbaenians
Anolis
Iguana
Sceloporus
Abronia
Ophisaurus
Shinisaurus
Varanus salvator
V. komodoensis
Lizards
Leptotyphlops
Typhlops
Scolecophidia
("Blind Snakes")
Boa
Cylindrophis
Python
Xenopeltis
Acrochordus
Agkistrodon
Ovophis
Pantherophis
Dinodon
Alethinophidia
("Advanced Snakes")
Snakes
Squamates
dNTV12/dSTV4X
0.17 - 0.25
0.10 - 0.17
0.00 - 0.10
0.1 sub./site

Excess Replacements in COI

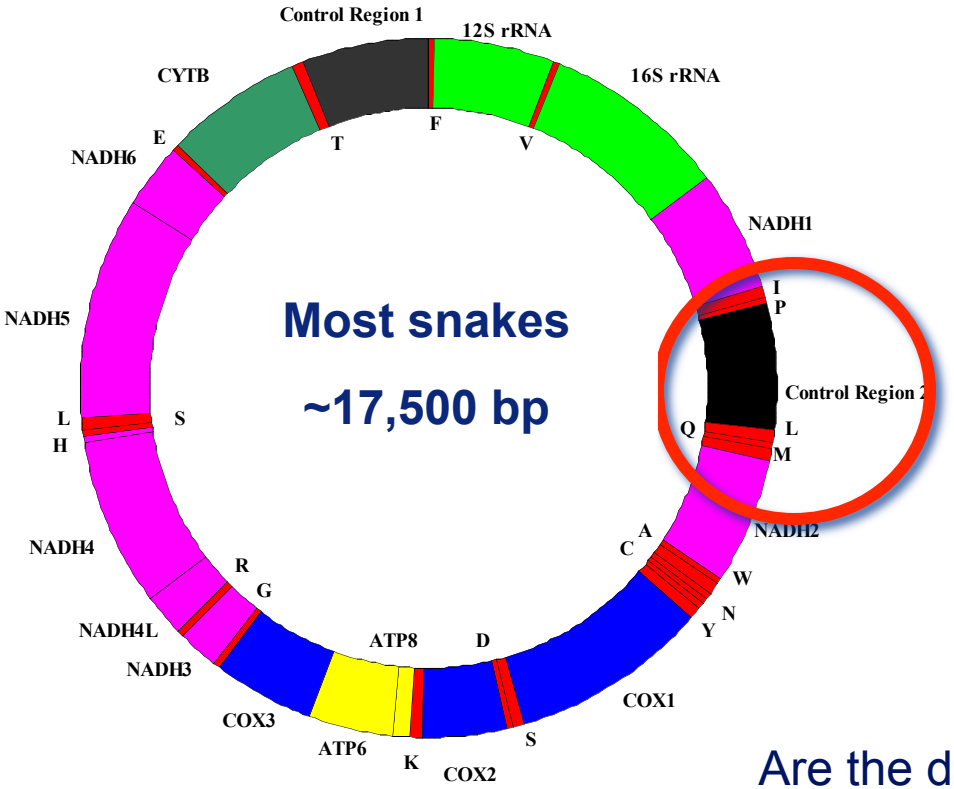


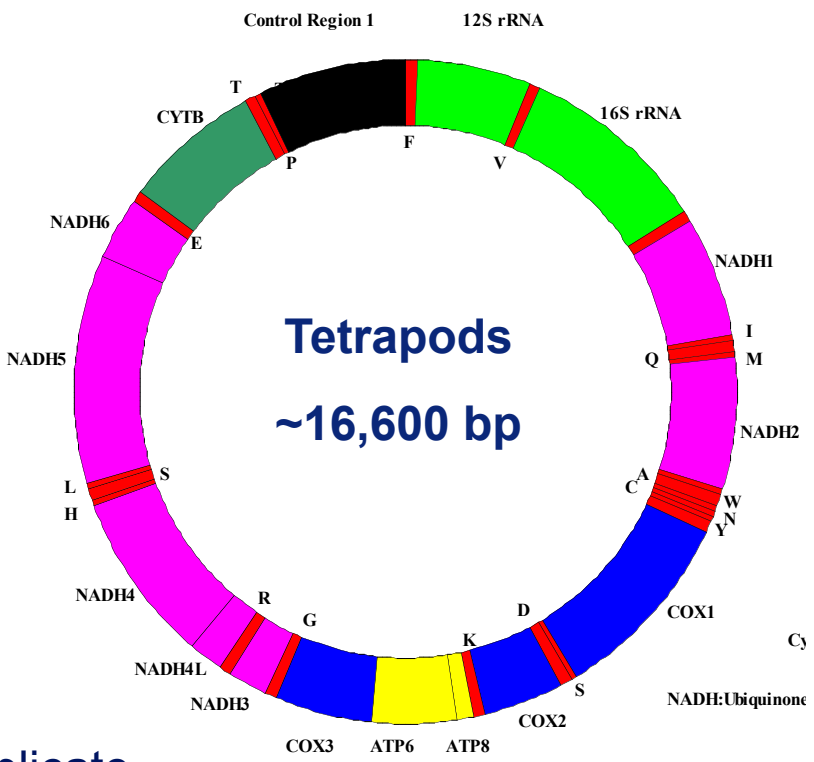
- There is very little in tetrapod mtDNA that is not functional
  - Implies selection to remove junk
- Most snake mtDNA genes are short
  - Implies even stronger selection to reduce excess nucleotide length



Duplicate Control Regions (CR) in Most Snake mtDNA

Origin of genome replication and bidirectional transcription initiation



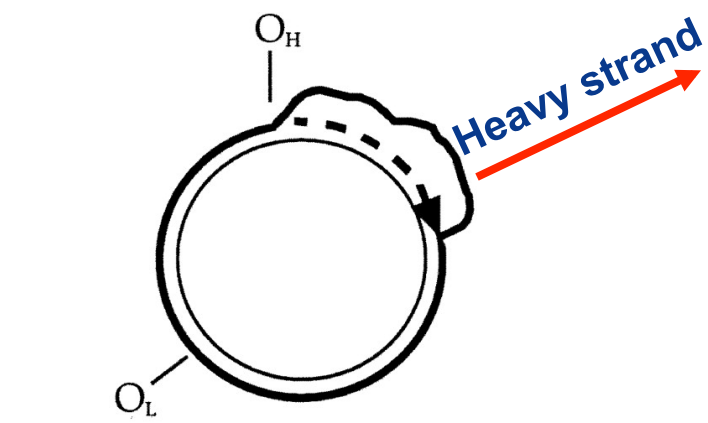


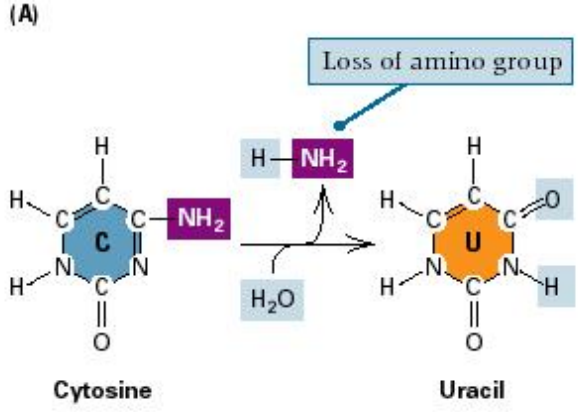
An extra ~1000+ bp of DNA

Concerted evolution (96-100% identical)

Are the duplicate control regions functional? Is there adaptive relevance?

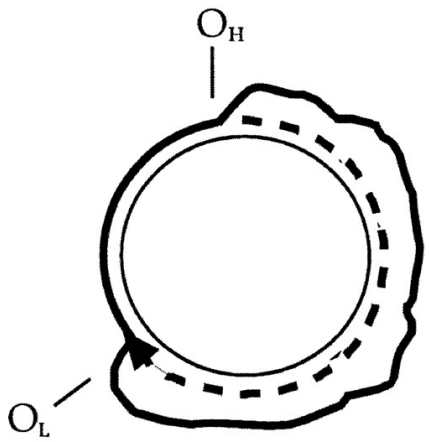
Typical Mitochondrial Genome Replication (single control region)

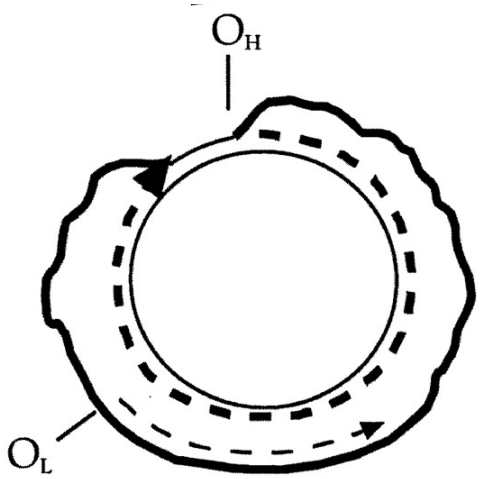




C → U → T

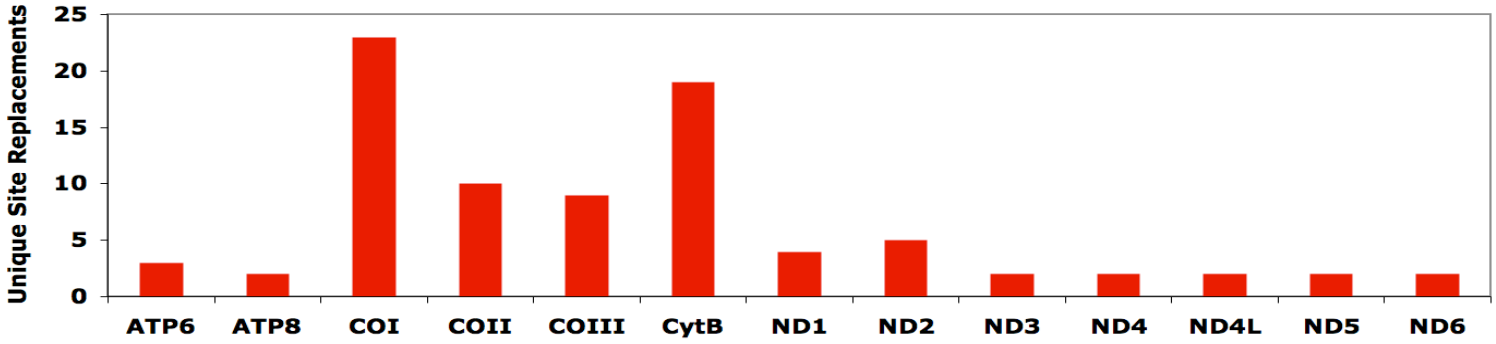
A → H → G



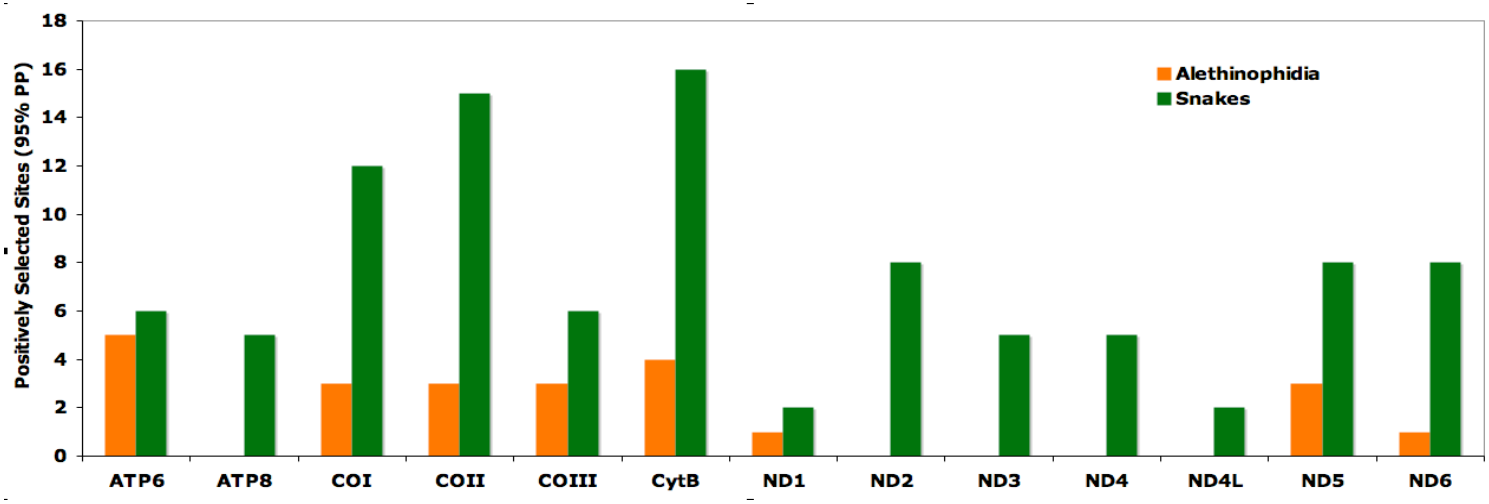


Accelerated Evolution Early in Snake Evolution at Conserved Sites and Genome-wide

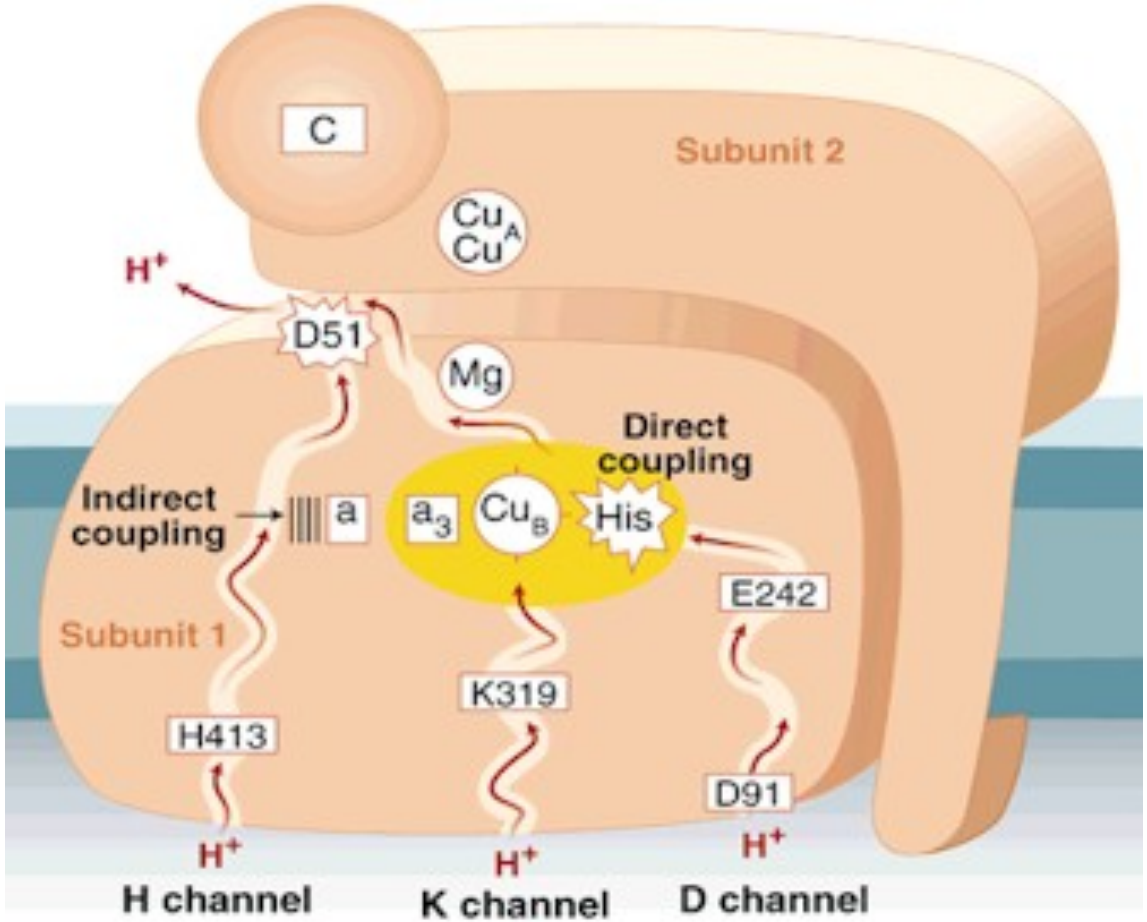
Replacements at sites otherwise conserved across tetrapods



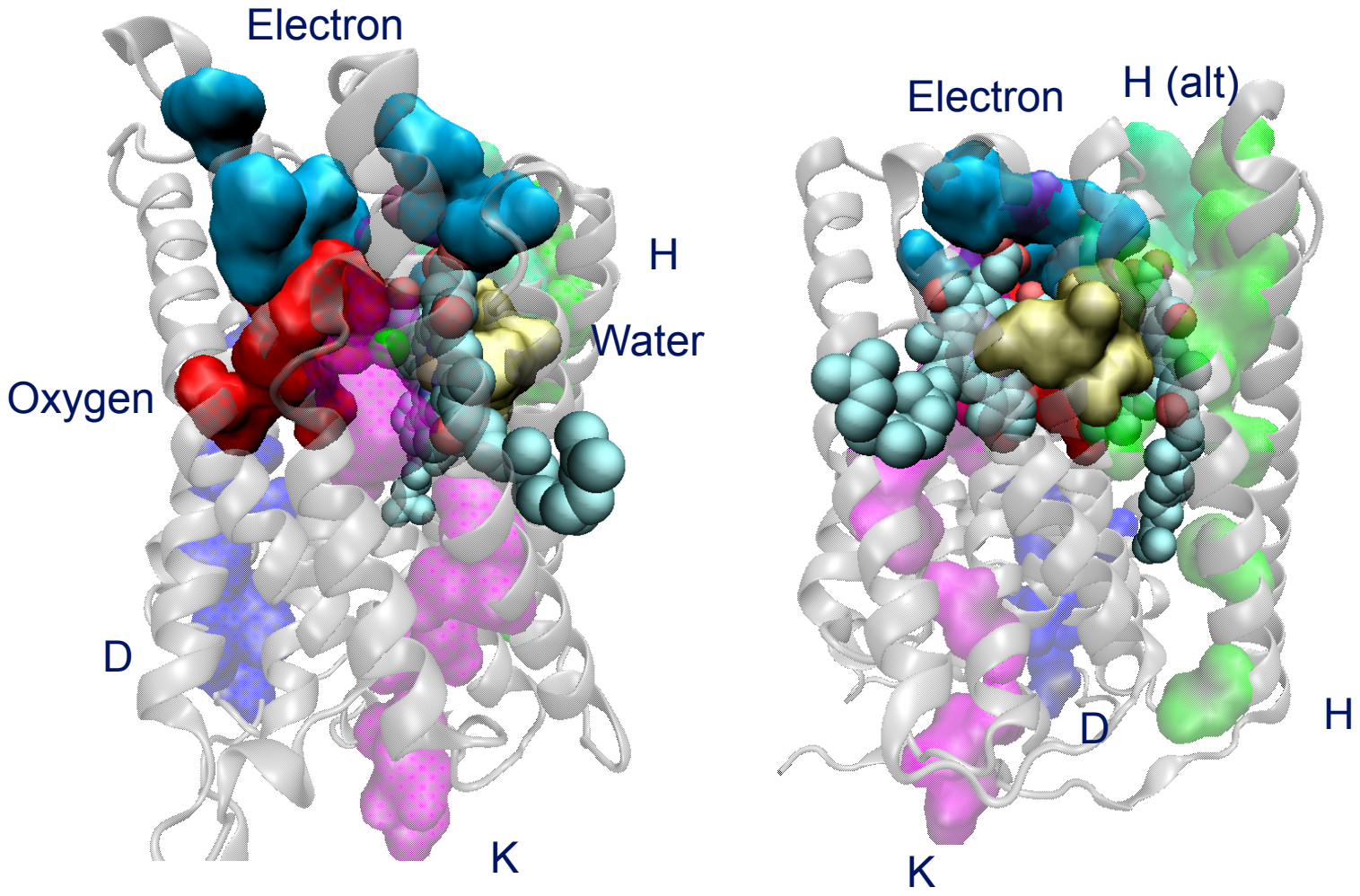
Positively selected sites along Alethinophidian and Ancestral Snake branches



Predicted Function of Channels



COI Functional Regions



Unique Sites

- Altered in snakes
- Otherwise conserved across most tetrapods
- Focus on sites most likely to be functionally relevant (limit numbers)
- Associated with coevolving site pairs
  - Coevolution in snake mtDNA is very high
    - 22% of site pairs at $p<0.01$

Unique Residue Clusters

Cluster Number	Residues	Cα Distance	Location
1	**35L** – 37I – 54Y	5.0 Å*, 10.6 Å*	H Channel
2	**443Y** – 447Y	6.7 Å*	H Channel
3	**256A** – 258V	5.6 Å*	K Channel
4	**266E** – 267P	3.8 Å*	K Channel
5	26A – 108S	11.9 Å	D Channel
6	205G – 231Y	6.3 Å*	O_2 Delivery
7	**299V – 301T**	5.5 Å*	O_2 Delivery
8	194L – 281G	6.9 Å*	O_2 Delivery

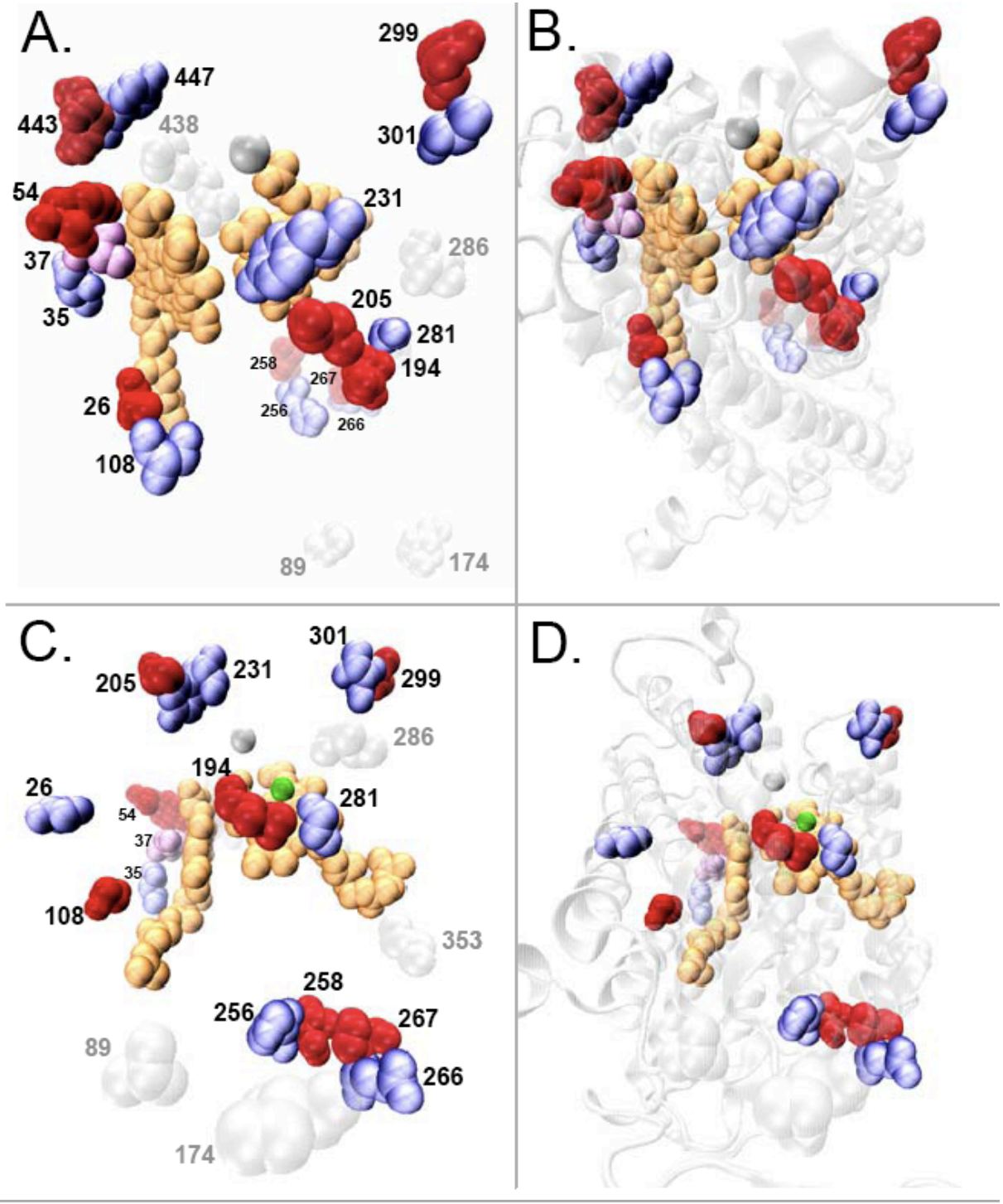
Coevolution is Usually Distributed

- Not usually strong pairwise
- Oftentimes adjacent to function or binding
  - Pollock and Wang, 2005, 2007
  - Yeang and Haussler, 2007

Evolution Genomics Biodiversity

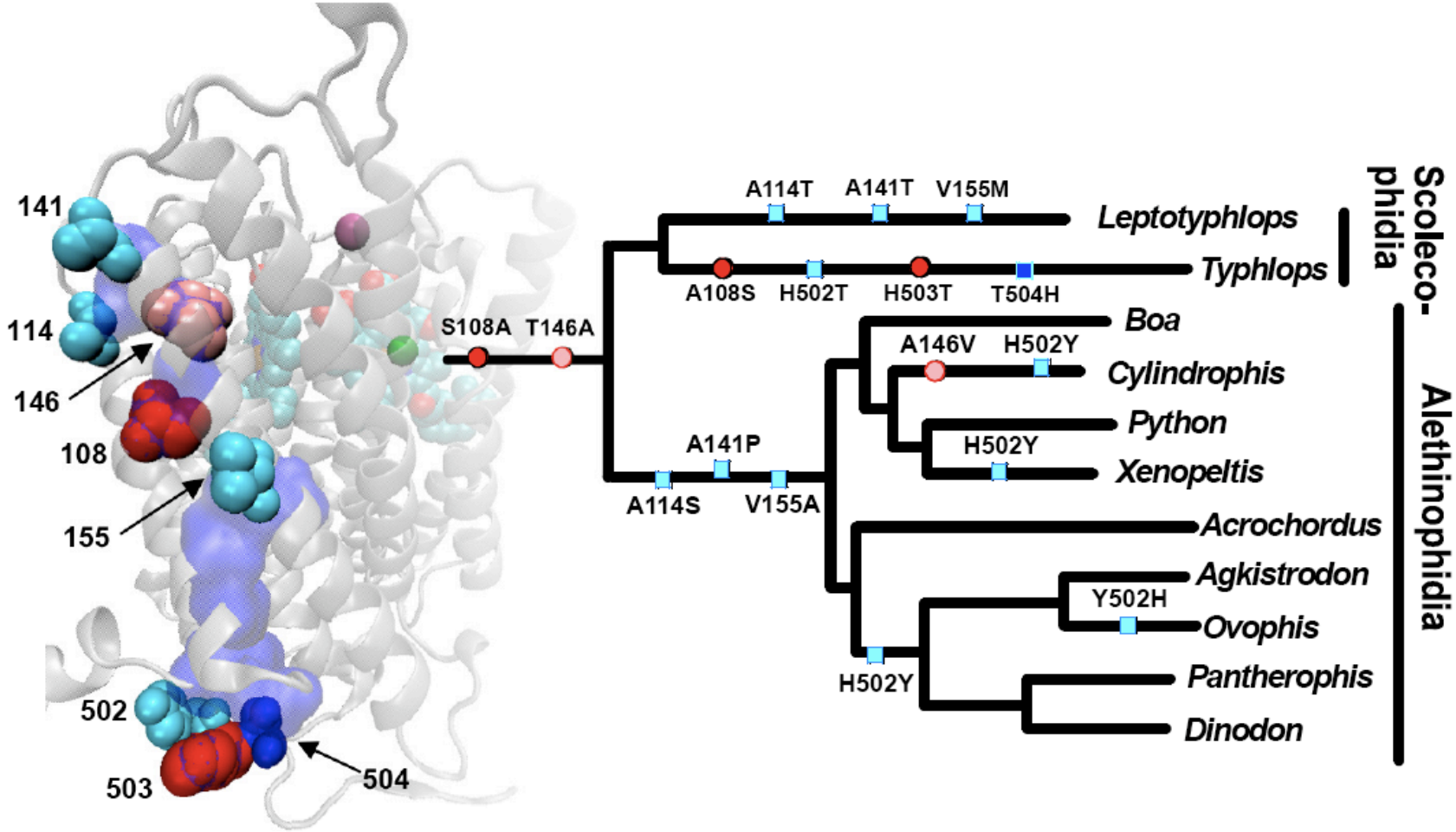
Coevolution in COI

Physically paired unique substitutions



Channel D (Direct Coupling)

Loss of Polarity, then Recovery?



Unique Sites

Changes in proton channels

Changes adjacent to channels

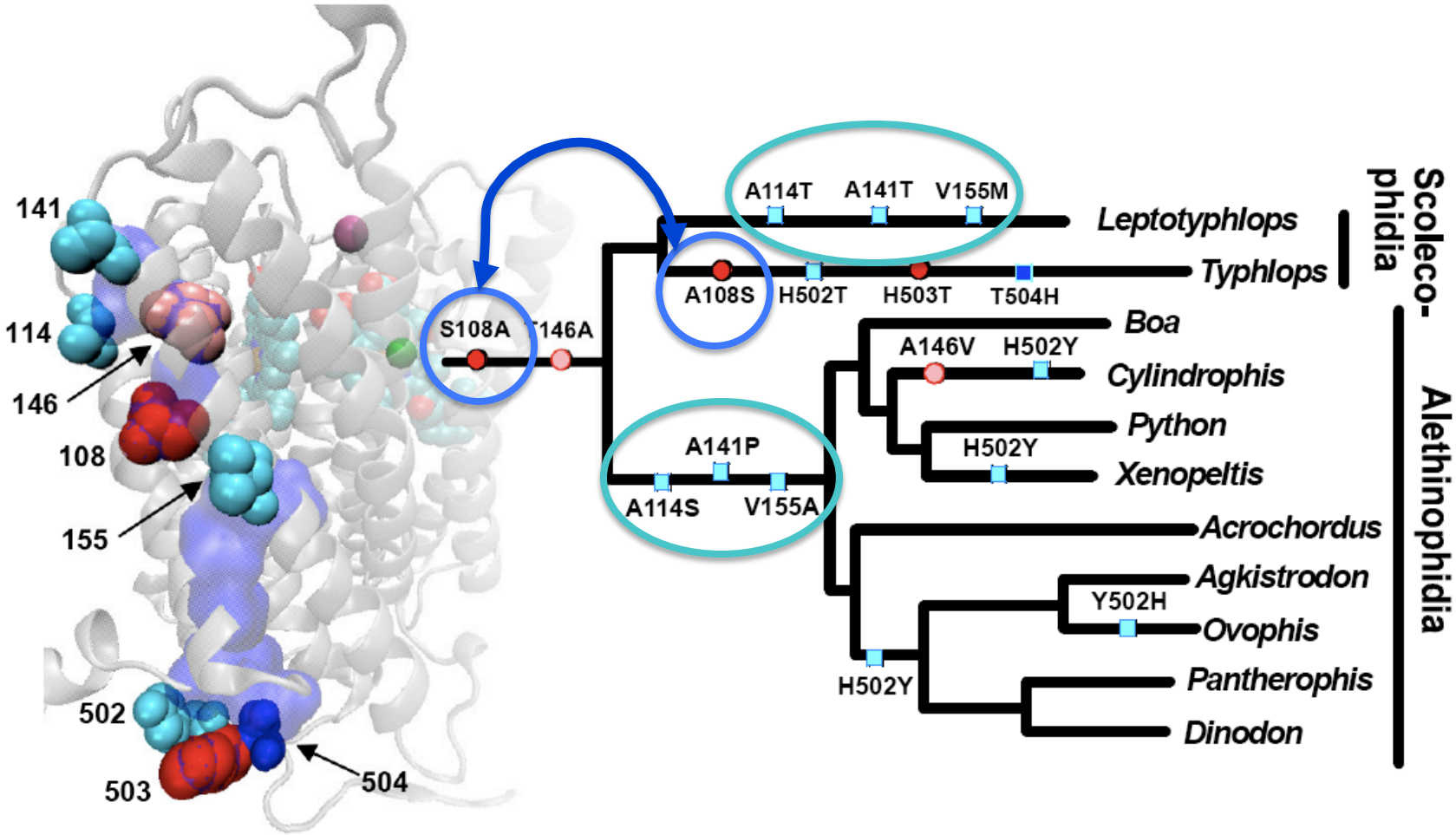
Non-Unique Sites

Changes in proton channels

Changes adjacent to channels

Channel D (Direct Coupling)

Convergence and Reversion



Unique Sites

Changes in proton channels

Changes adjacent to channels

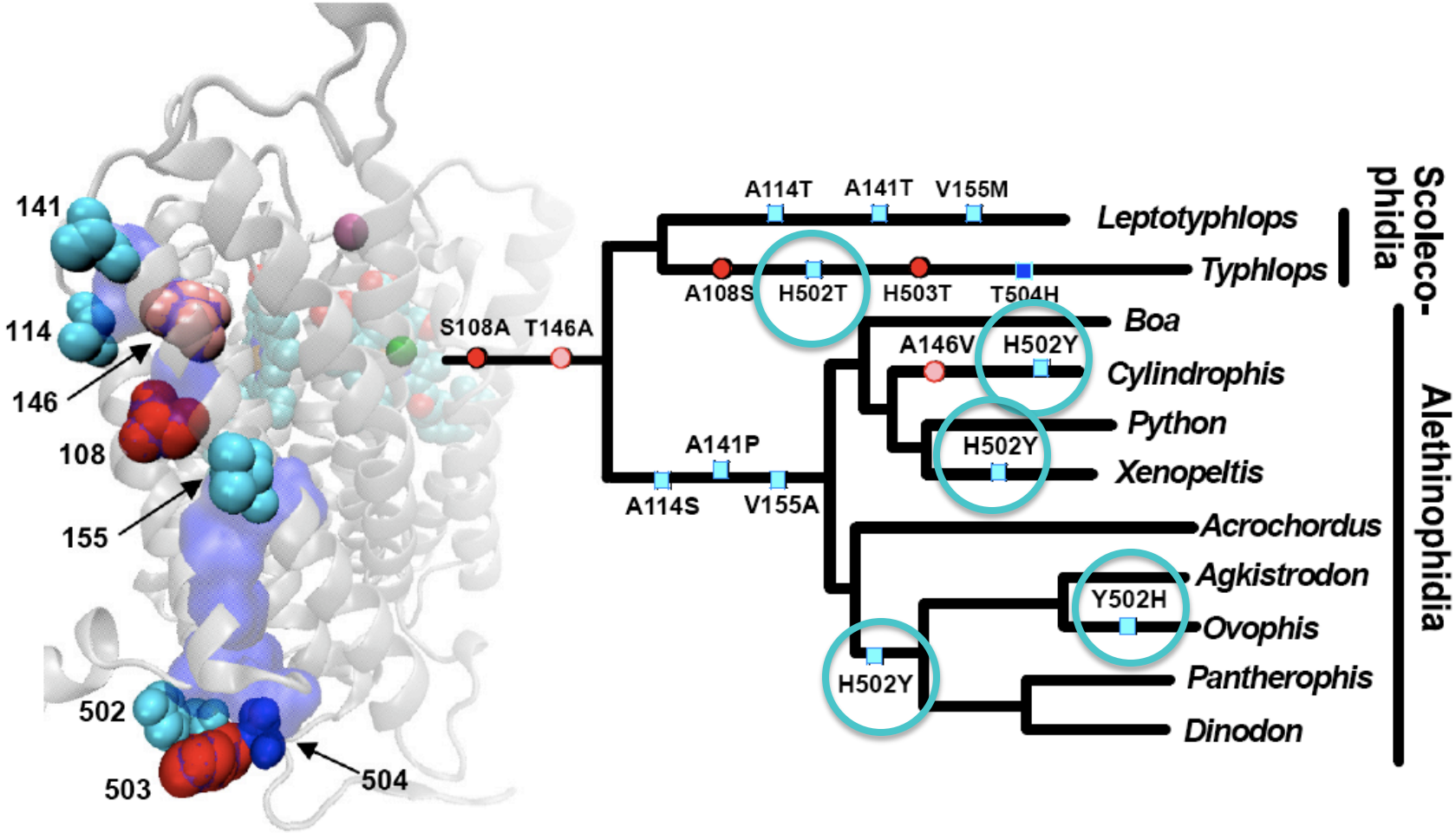
Non-Unique Sites

Changes in proton channels

Changes adjacent to channels

Channel D (Direct Coupling)

Repeated Convergence (and a reversion)



Unique Sites

Changes in proton channels

Changes adjacent to channels

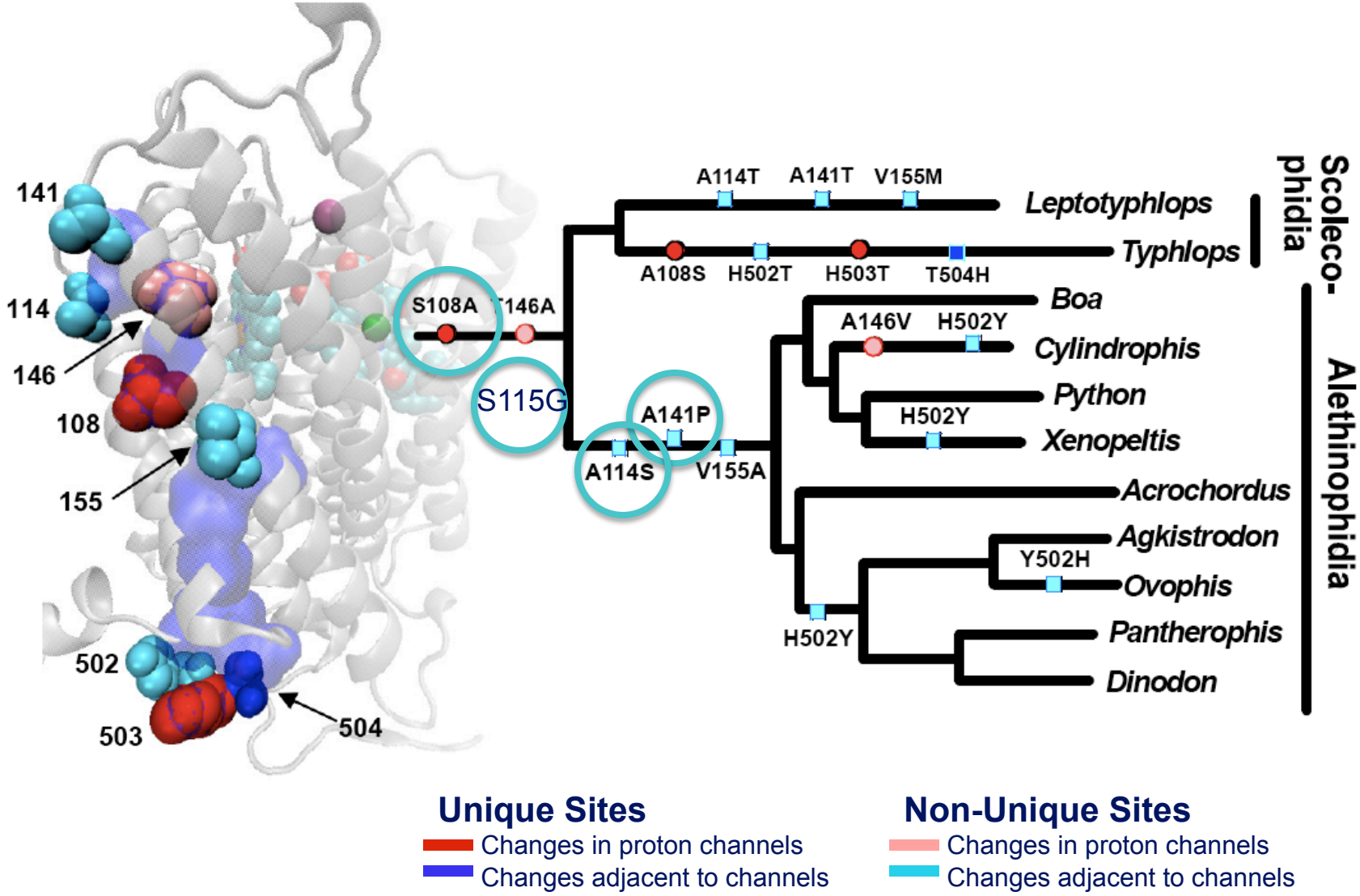
Non-Unique Sites

Changes in proton channels

Changes adjacent to channels

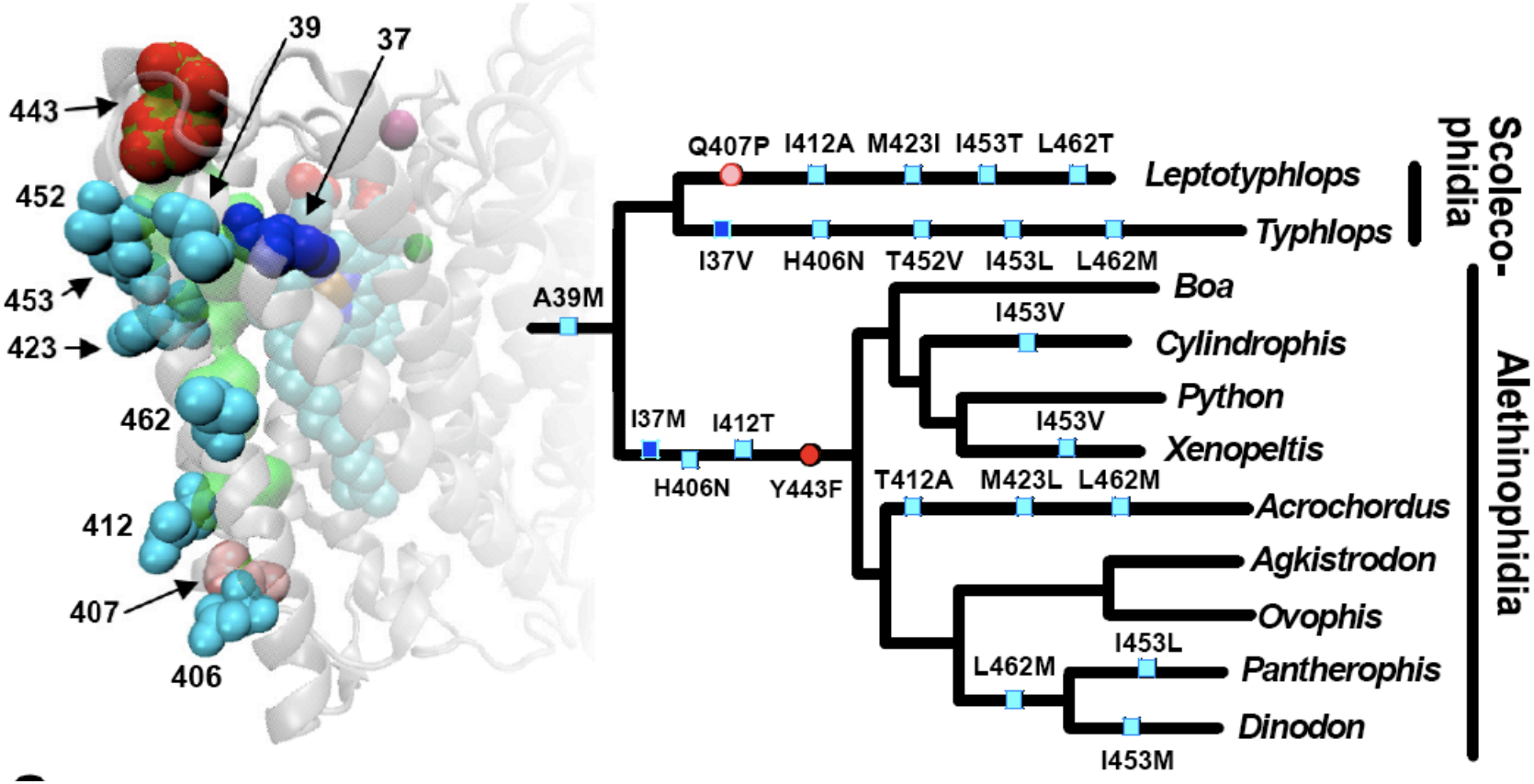
Channel D in *Rhineura*

a distantly-related legless tubular squamate



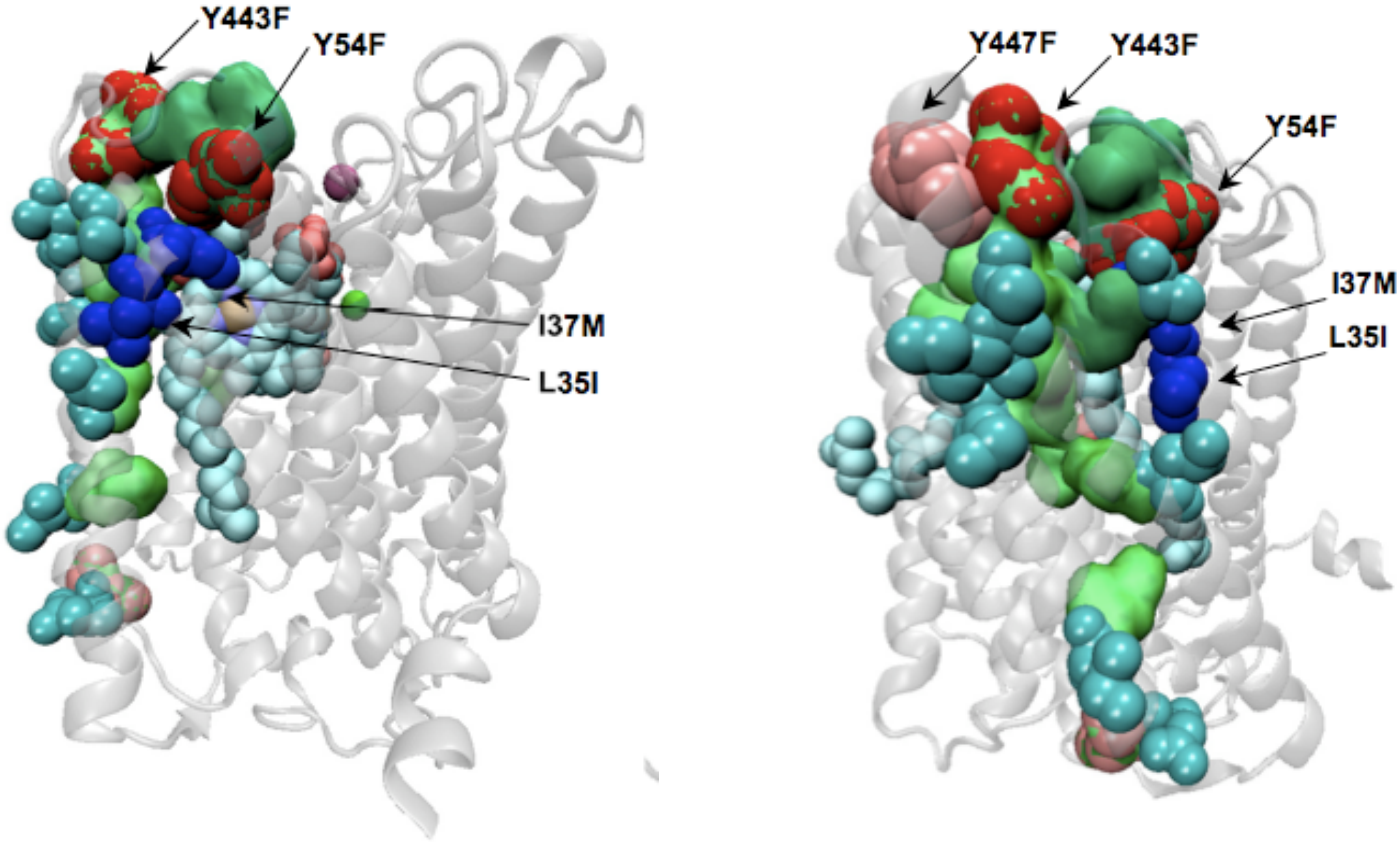
Channel H: Exclusive Pumping, Indirect Coupling

Controversial function is shut down



Controversy over channel H

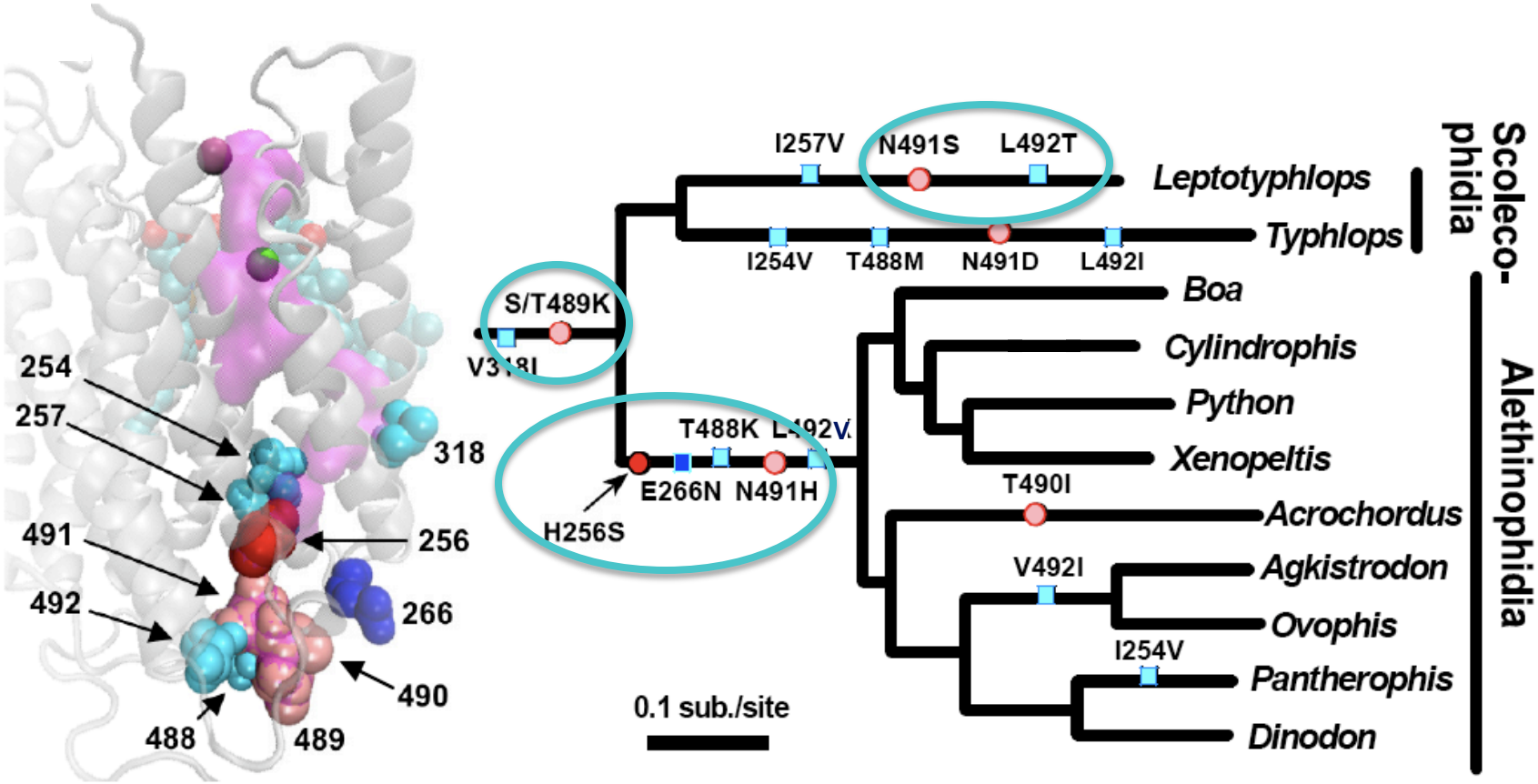
- *Is it really used?*
- *Which of two different paths is it?*



Snakes go out of their way to completely destroy all possible outlets of the alternative Channel H; tyrosine (Y) to Phenylalanine (F) substitutions are usually quite rare

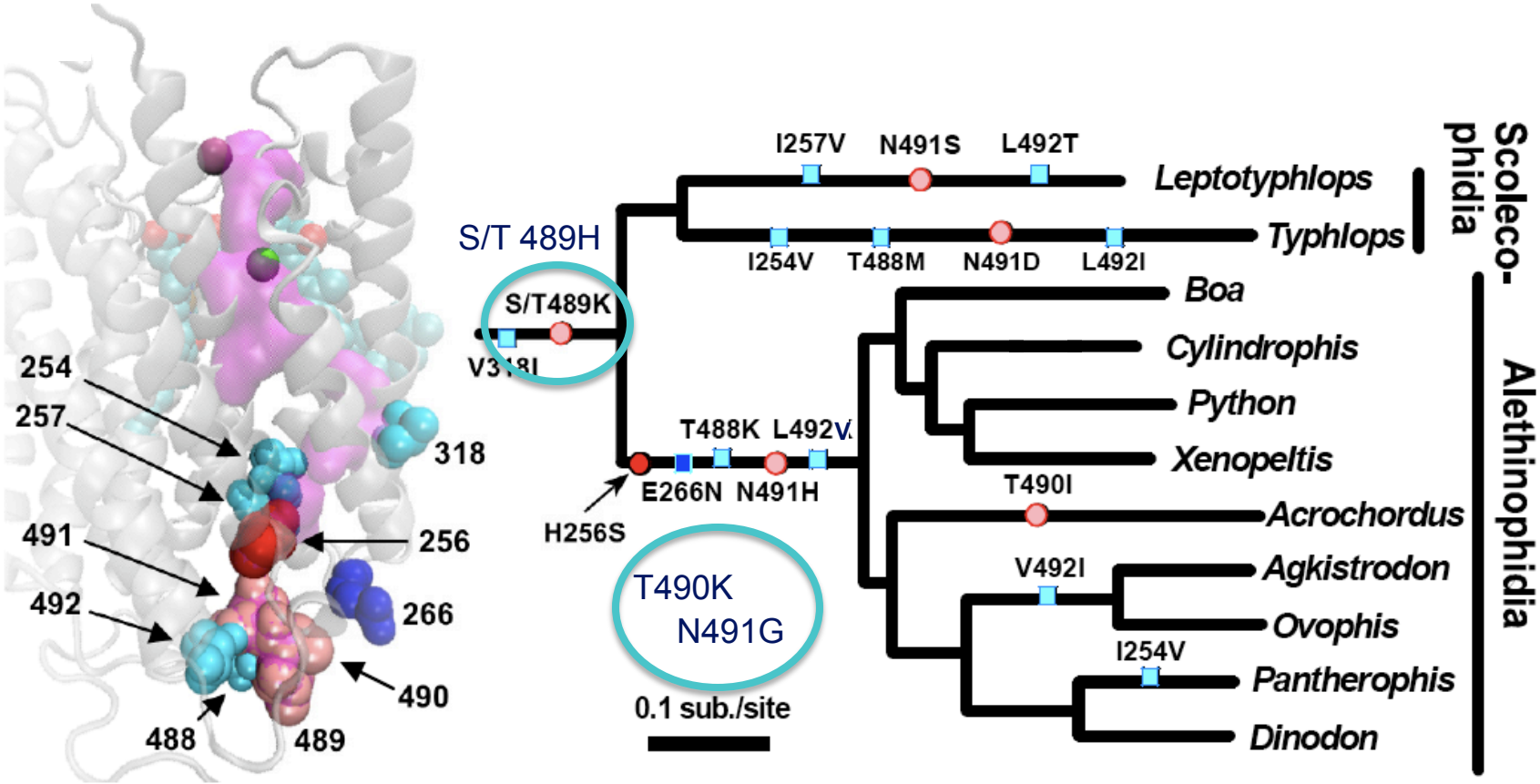
Channel K: Proton Delivery to Reaction Center

Not shut down; increase in positive charge at entrance



Channel K in *Rhineura*

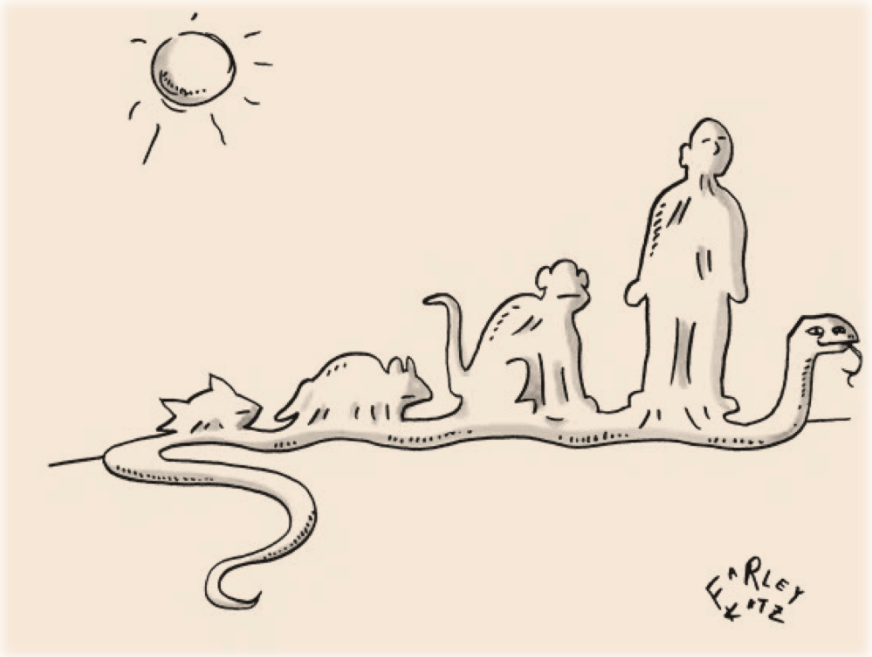
Not shut down; increase in positive charge at entrance

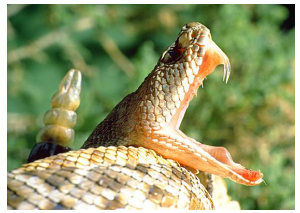


Snakes a Model for Extreme Adaptation

Metabolism – Physiology - Venom

- Aerobic Metabolism
  - One of the lowest basal metabolic rates
  - Highest fluctuation between basal and max
  - Fluctuations of 40-fold in 48 hours





Snakes a Model for Extreme Adaptation

Metabolism – Physiology - Venom

- Aerobic Metabolism
- Physiological Remodeling to Digest Prey
  - Heart muscle - may enlarge 50%
  - Liver – may enlarge 100%
  - Gut - may enlarge 100 - 150%

Secor & Diamond, Nature, 1998

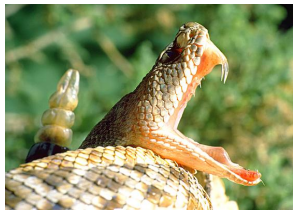
progress

A vertebrate model of extreme physiological regulation

Stephen M. Secor & Jared Diamond

Department of Physiology, University of California Medical School, Los Angeles, California 90095-1751, USA

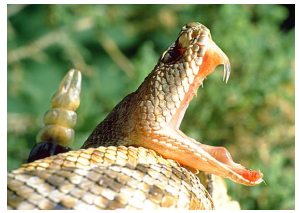
Investigation of vertebrate regulatory biology is restricted by the modest response amplitudes in mammalian model species that derive from a lifestyle of frequent small meals. By contrast, ambush-hunting snakes eat huge meals after long intervals. In juvenile pythons during feeding, there are large and rapid increases in metabolism and secretion, in the activation of enzymes and transporter proteins, and in tissue growth. These responses enable an economic hypothesis concerning the evolution of regulation to be tested. Combined with other experimental advantages, these features recommend juvenile pythons as the equivalent of a squid axon in vertebrate regulatory biology.



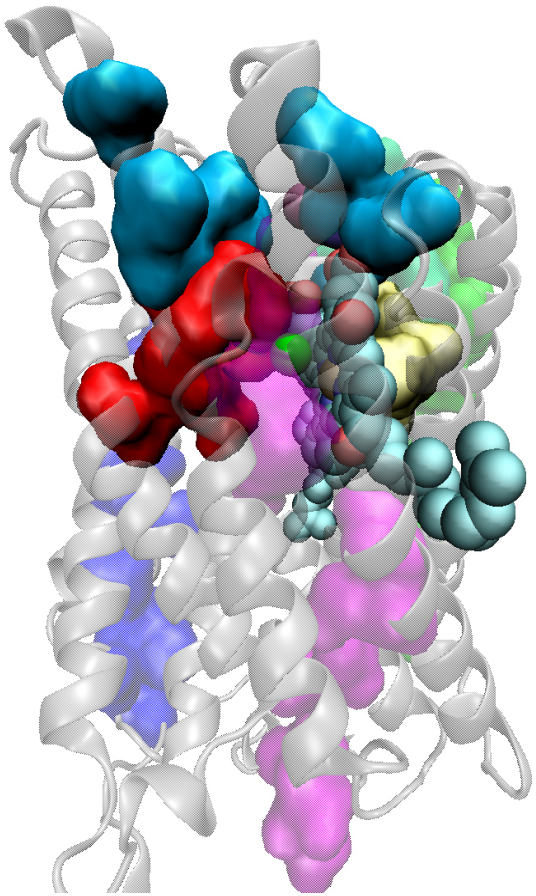
Snakes a Model for Extreme Adaptation

Metabolism – Physiology - Venom

- Aerobic Metabolism
- Physiological Remodeling
- Venom
  - Diverse arsenal of deadly venom proteins
  - Widespread adaptive evolution of venom proteins



Massive Multi-Protein Adaptation

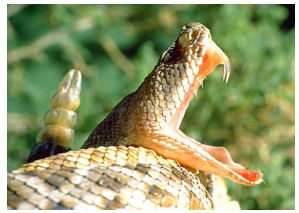


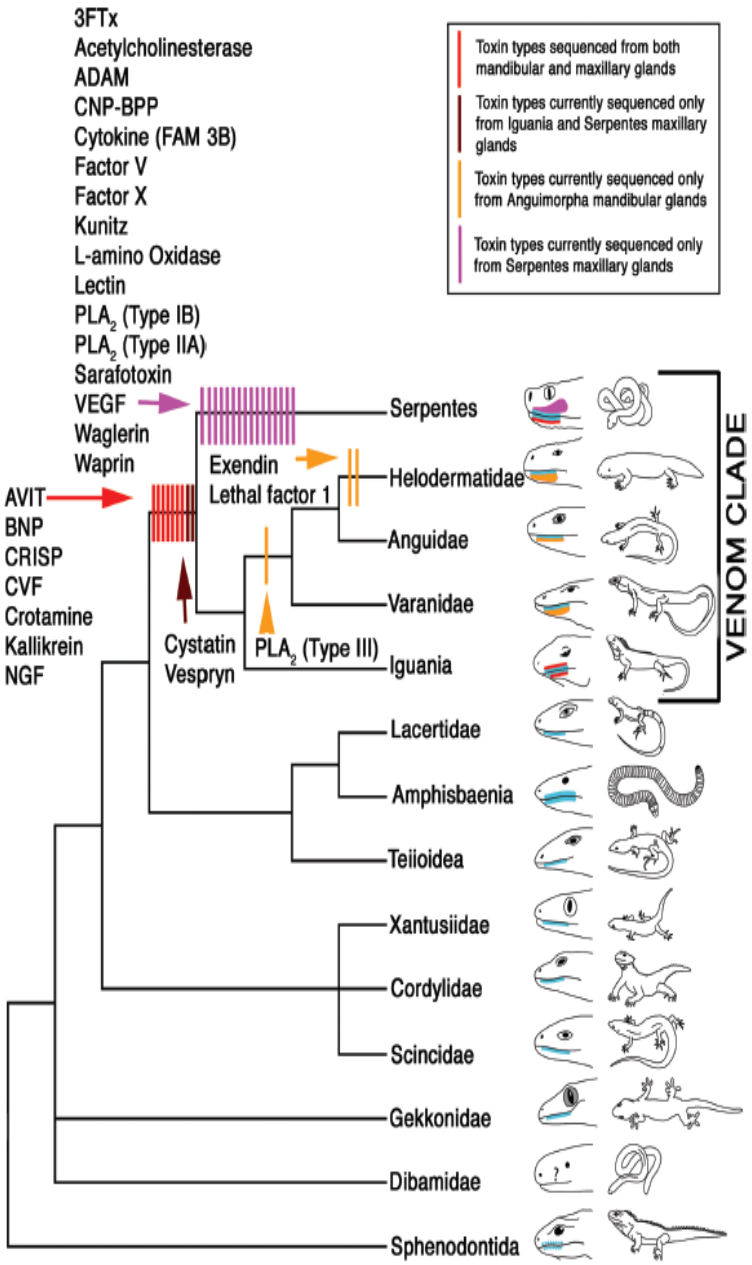
- ✓Most extreme adaptation known in metabolic proteins
- ✓Molecular coevolution – best example known
- ✓Molecular convergence
- ✓Shift in mitochondrial function
  - ✓Increase proton flow to the reaction center?
- ✓Likely important for metabolic fluctuations in snakes
- ✓Microevolutinary event → macroevolutionary adaptation

Snakes a Model for Extreme Adaptation

Metabolism – Physiology - Venom

- Aerobic Metabolism
- Physiological Remodeling
- Venom
- Evolutionary History
  - => Fossorial and inactive
  - => Terrestrial and capable of switching from inactive to very active





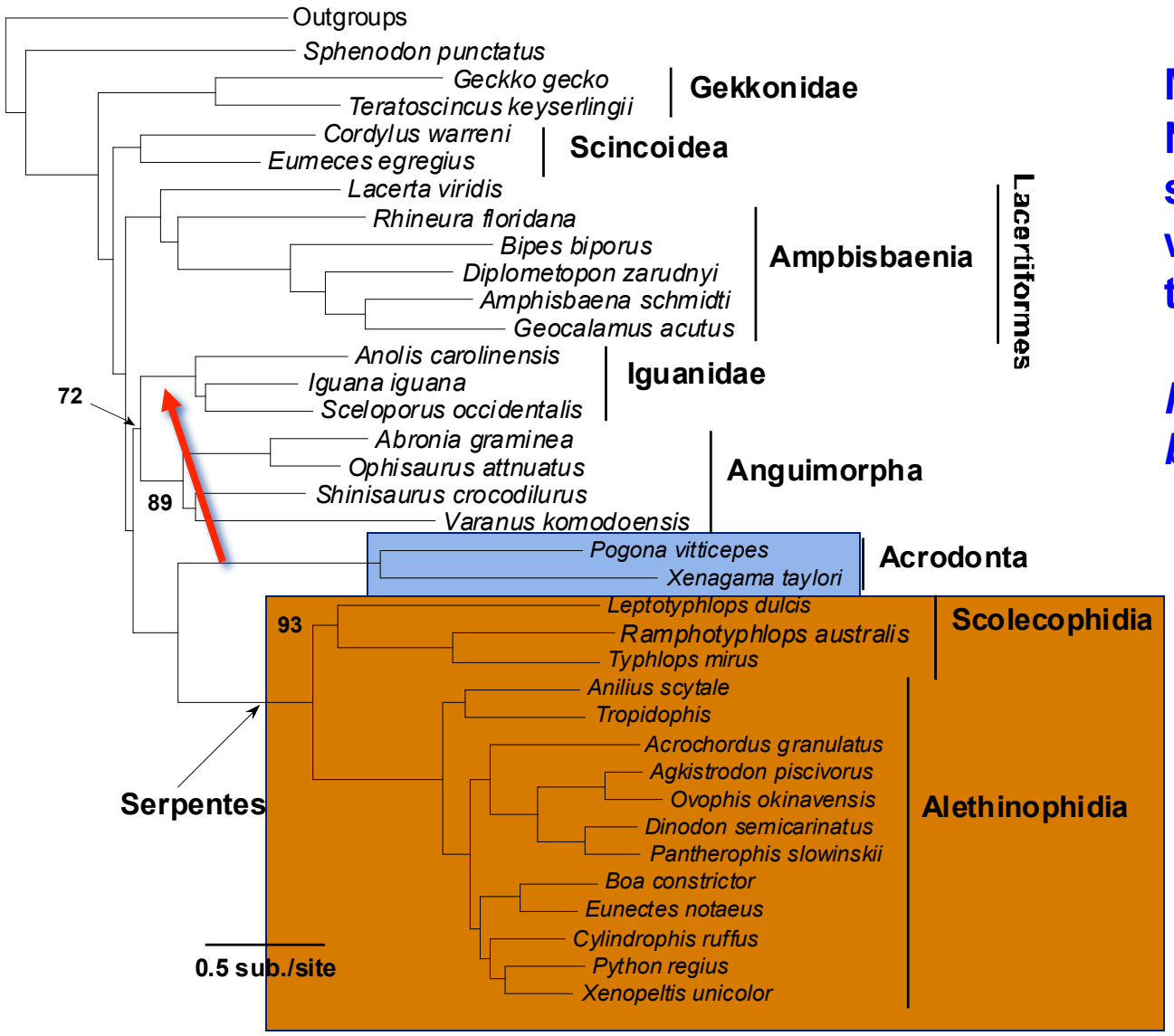
Did Venom Play a Role?

Snake venom genes were present (and expressed in salivary glands) in lizards *PRIOR* to snake evolution

A broad arsenal of amazingly toxic proteins evolved only in some snakes

Venom is also one of the main known causes of accelerated or diversifying evolution

Snake / Lizard Phylogeny

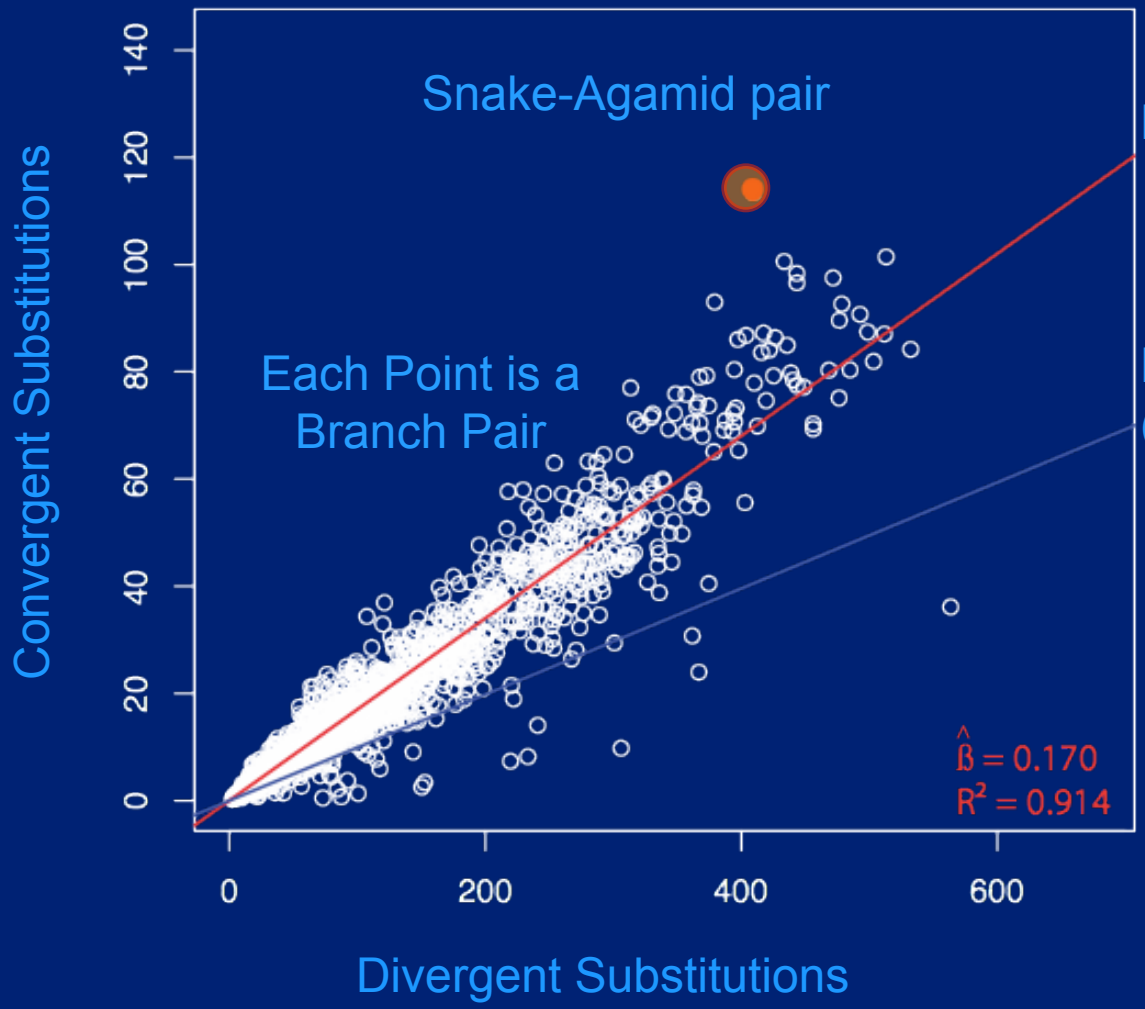


Morphological and Nuclear Data strongly disagree with mtDNA-based trees

Note: mtDNA tree is based on 12,000 bp!

Snakes

Excess Convergence in Mitochondrial Proteins



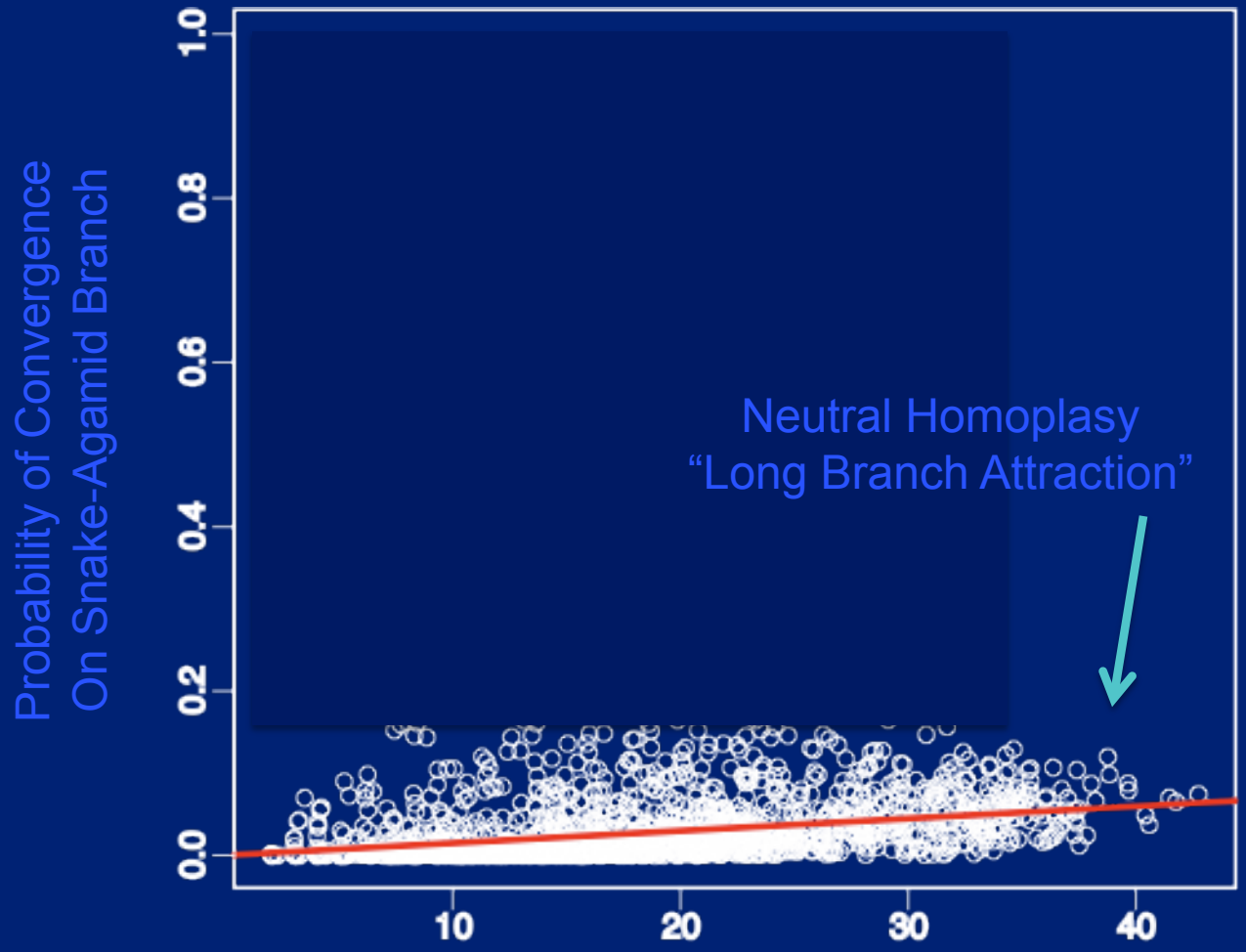
~114 convergent residues (Bayesian posterior integration)

>44 above empirical expectation

~30 with high posterior support

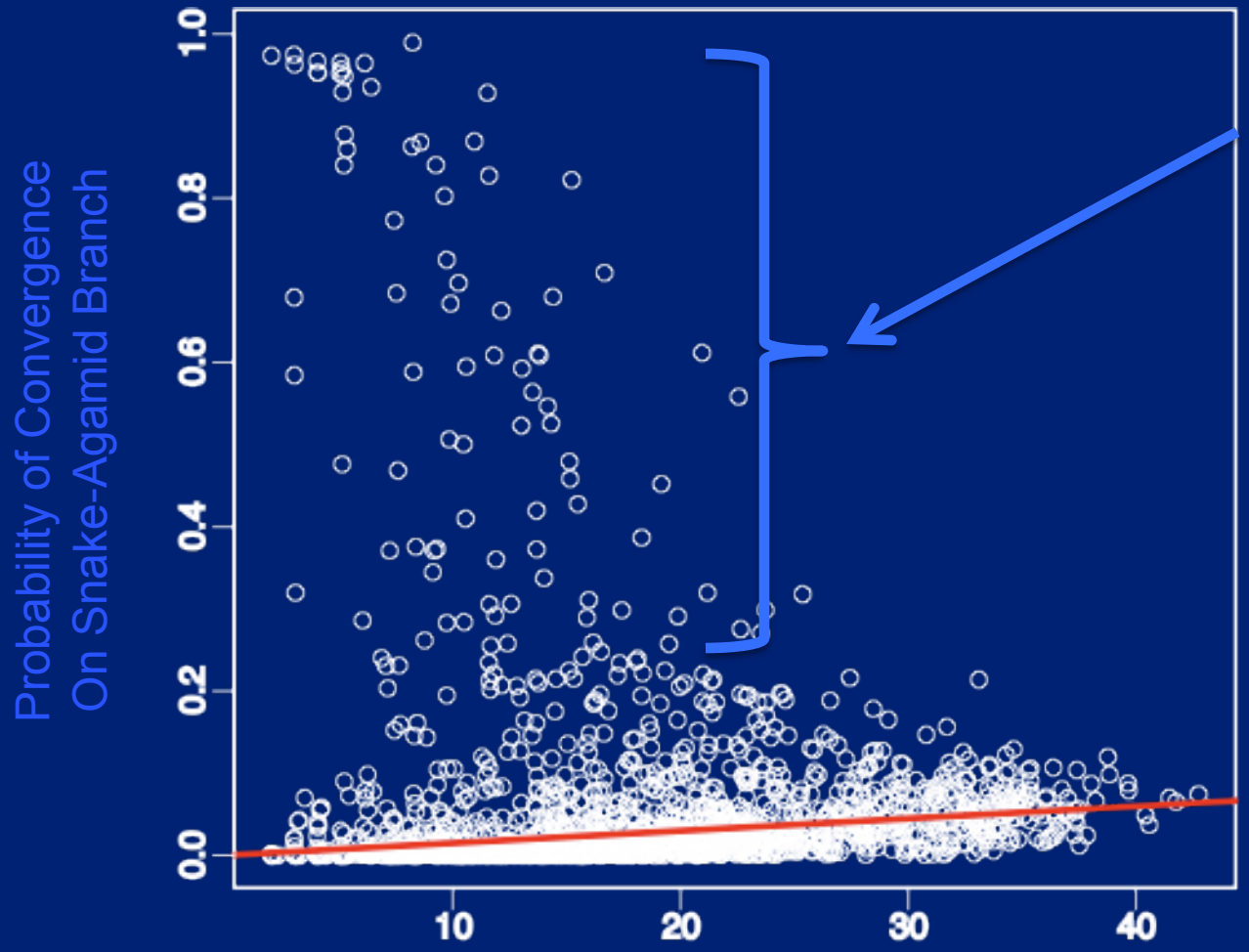
Fast Sites Converge a Little Bit

Predicted by Neutral Convergence



Converged Sites Evolved Slowly

Consistent with Adaptive Convergence



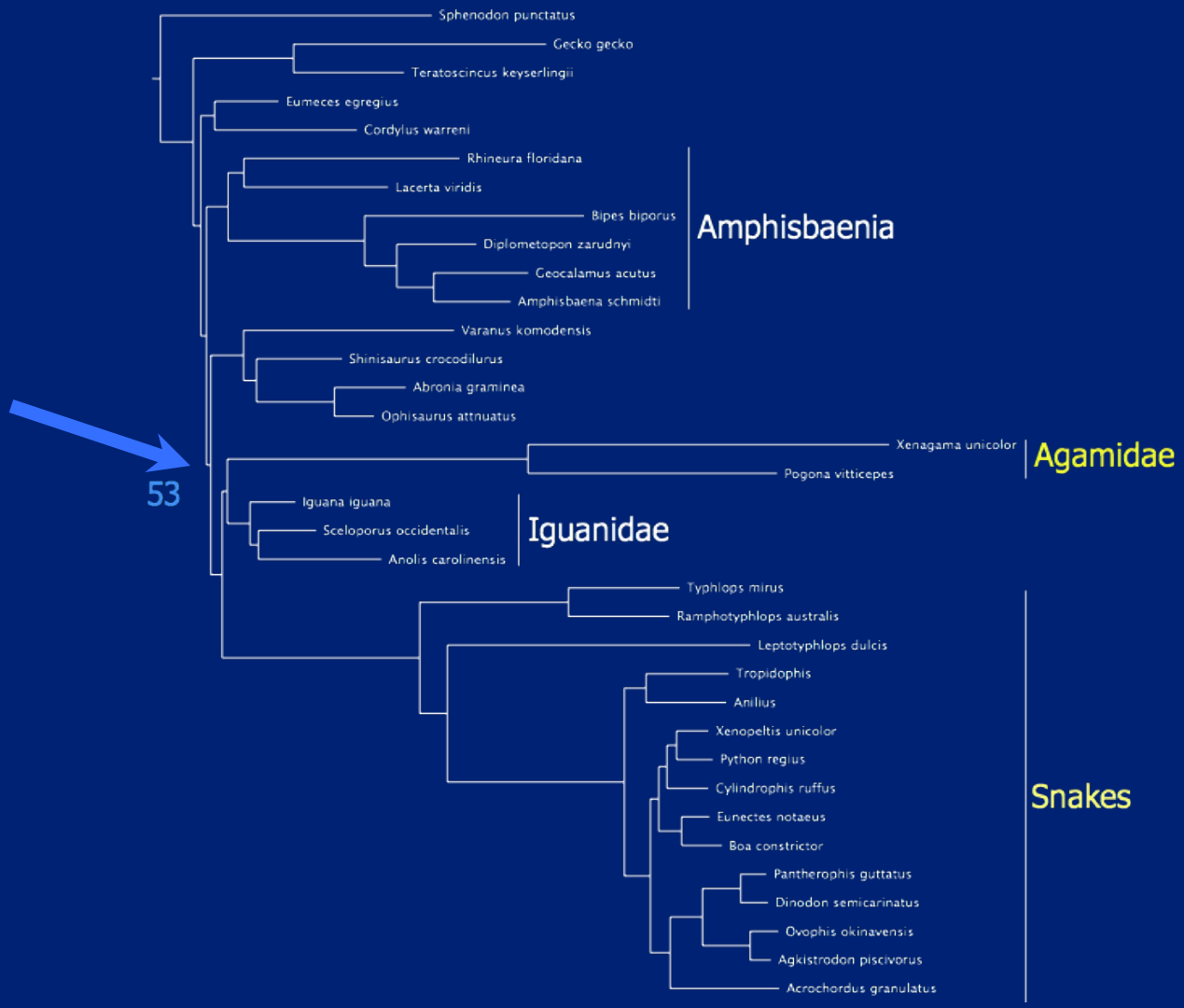
Convergence at Highly Conserved Sites

Screening Convergent Sites Restores Nuclear Tree

top 5% of convergent sites were screened

Iguanids and Agamids Join as Sister Taxa

Latent Signal for Correct Tree is Present



Ruggedness, Dimensionality, and Changing Landscapes

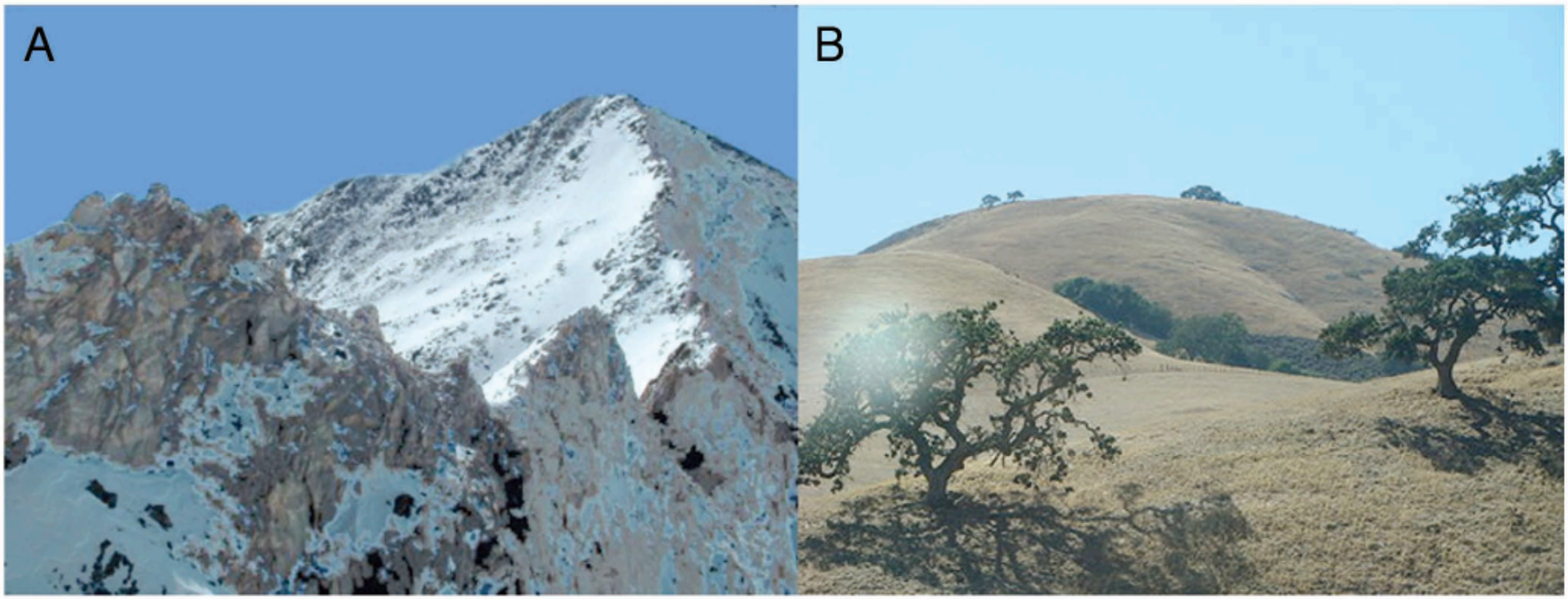


Figure 1. Alternative views of potential protein adaptive landscapes. In (A), the protein adaptive landscape is viewed as being like an arrêt ridge, with only a single narrow path leading from the current adaptive peak in the foreground to a new adaptive peak in the distance. This landscape is conducive to convergence. In (B), the adaptive landscape is viewed as being like rolling hills, with many alternative routes to nearby adaptive hilltops that are not substantially different from one another. With so many alternative paths and alternative similar hilltops, under this scenario sequences would be unlikely to converge (i.e., follow the same path) even under similar adaptive pressure.

PLEX: Context-dependent evolutionary genomics in a practical time frame

- Large phylogenomic datasets now common
  - Parametric inference with realistic evolutionary models is (was) computationally burdensome
- MCMC + data augmentation of ancestral states and substitution histories can be extremely fast
  - Augmentation step is (was) a major performance bottleneck (>99% of computation)
- Order of magnitude speed improvements and excellent scaling can be achieved
  - partially sampling substitution histories

Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles

Nicolas Rodrigue[a,1], Hervé Philippe[b], and Nicolas Lartillot[b]

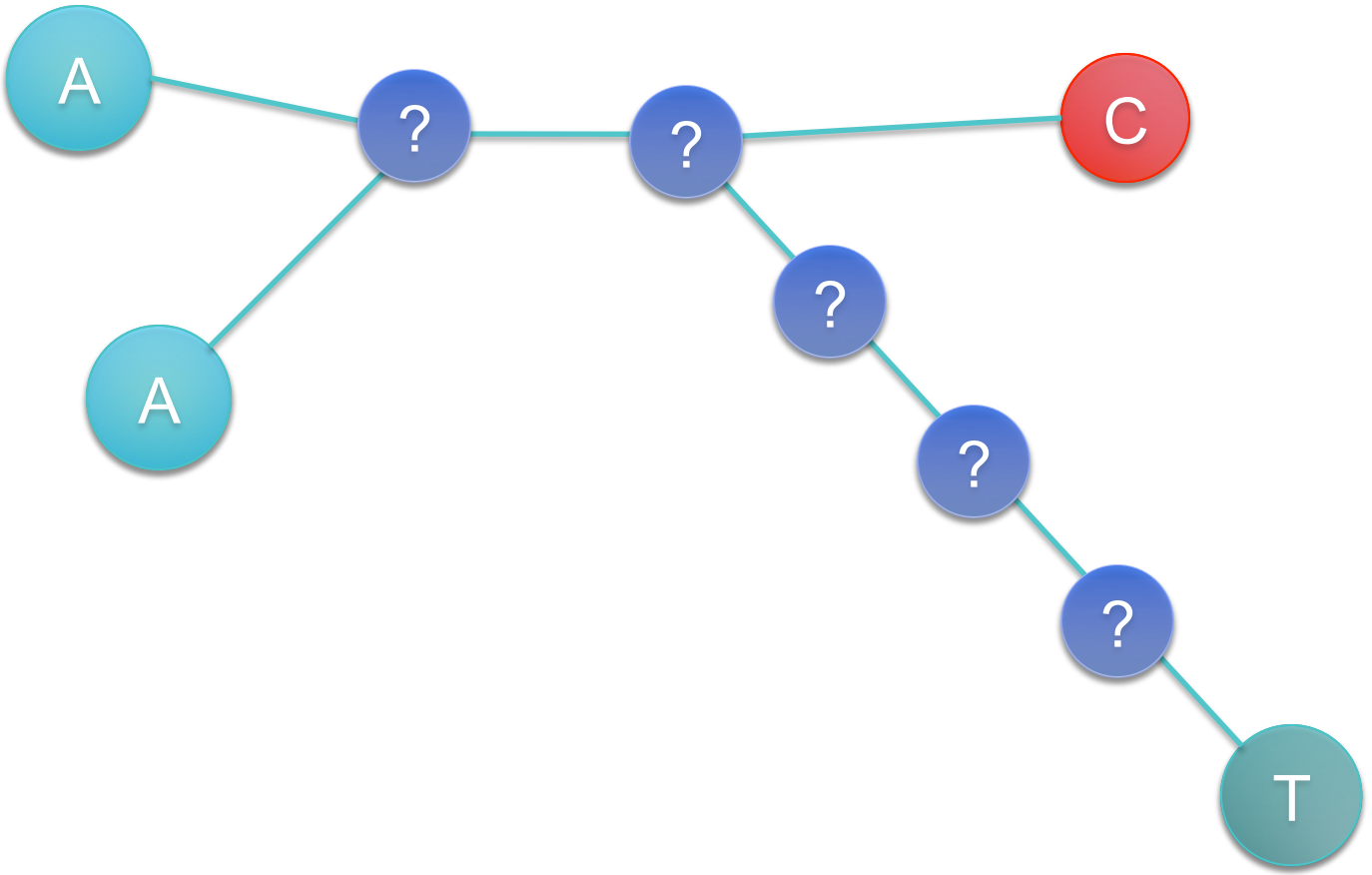
[a]Department of Biology, University of Ottawa, Ottawa, Ontario, K1N 6N5 Canada; and [b]Department of Biochemistry, Centre Robert Cedergren, Université de Montréal, Montréal, Québec, H3C 3J7 Canada

Edited by David M. Hillis, University of Texas at Austin, Austin, TX, and approved January 27, 2010 (received for review September 24, 2009)

!!

in some cases we study here. Although the empirical mixture approaches can provide less taxing models, the Bayes factors reported above (computed using pruning-based sampling) still required over 2 months of CPU time. It is thus of interest to advance further computational methods, both to ameliorate our current data-augmentation-based sampler and to bridge this type of MCMC sampling with our thermodynamic integration methods.

Time Complexity of Integrated Likelihood Calculations on a Phylogeny



Time Complexity of Integrated Likelihood Calculations on a Phylogeny

(N states, b branches between nodes, s sites)

$$O\left(N^4 + N^3 b + N^2 bs\right)$$

Substitution Histories

Ancestral States

Calculation gets overwhelming with increased complexity

Spatial Variation (many rate matrices)
Gradient Mixture Models
Context dependence

Temporal Variation
Markov-modulated codon models
Switching selection regimes

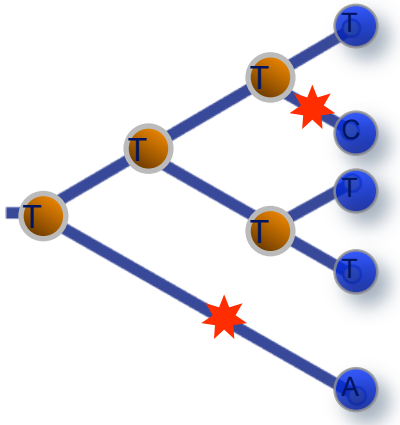
$$O\left(N^4 s + N^3 bs + N^2 bs\right)$$

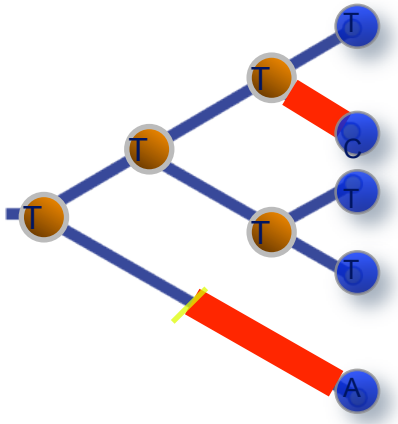
$$O\left(N^4 + N^3 b + N^2 bs\right)$$

N is very large

(e.g., 183 x 183)

Time Complexity Using Data Augmentation





Complete sampling in continuous time (Nielsen, Rodrigue, Lartillot)

Partial sampling in continuous time (de Koning et al. 2010)

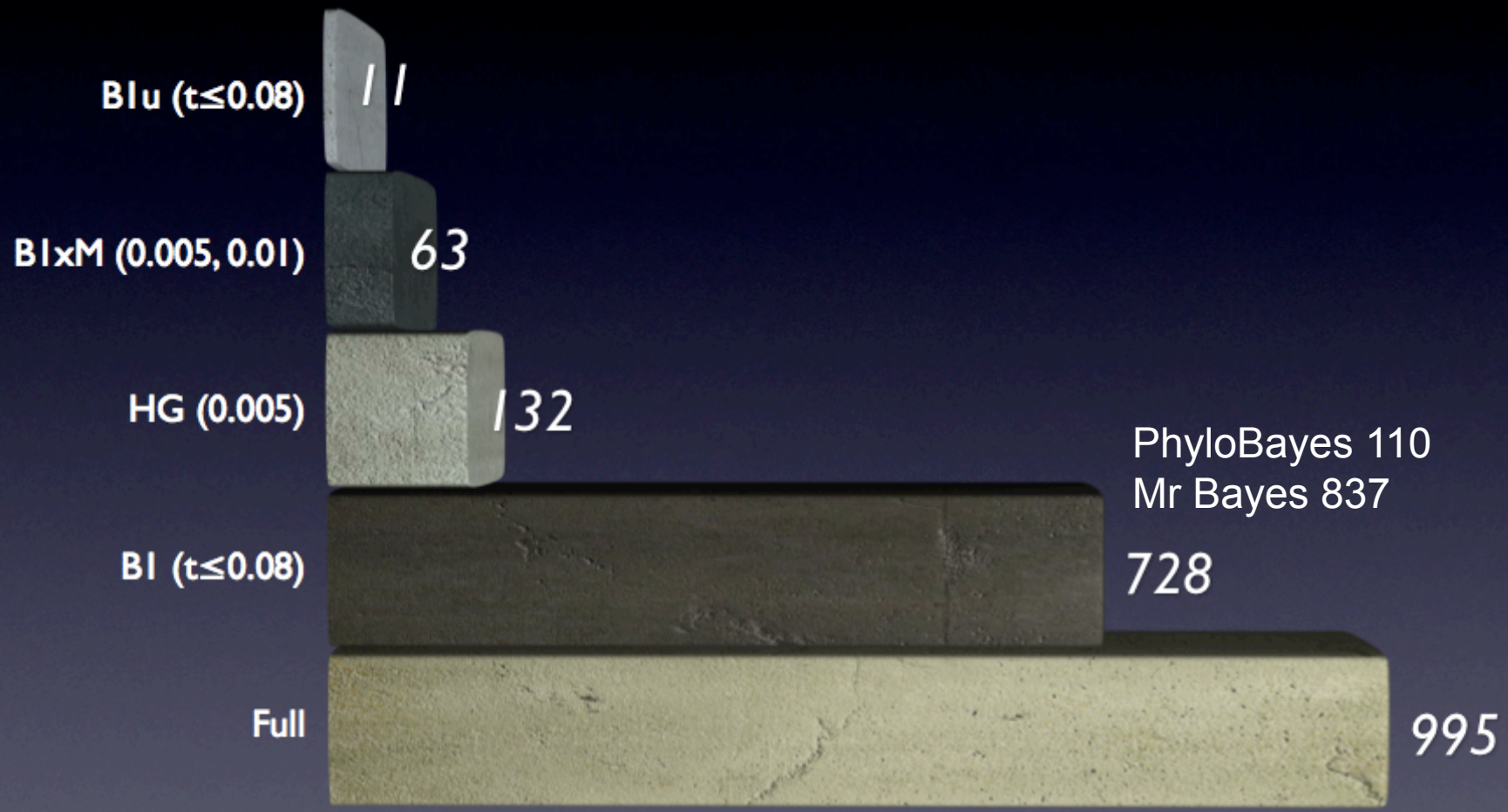
"Don't need to use fully integrated likelihood calculations"

"Don't need to fully sample the timing of substitution either"

Likelihood: $O\left(N^2\right)$ $O\left(N^2\right)$

Sampler: $O\left(N^4 + N^3 b + N^2 bs + Nbs\right) + more$ $O\left(Nb's\right),\ b' >= b$

Likelihood Analysis at the Speed of Parsimony



Time to analyse 224 taxon dataset, GTR model(100,000 generations of MCMC)

Dramatically Improved Scaling

Model	Standard Likelihood	BIu Speedup
DNA	1000 sec	11 sec (100x)
Amino Acid	7.5 hrs	17 sec (1600x)
Codon	2 months	12 min (7000x)

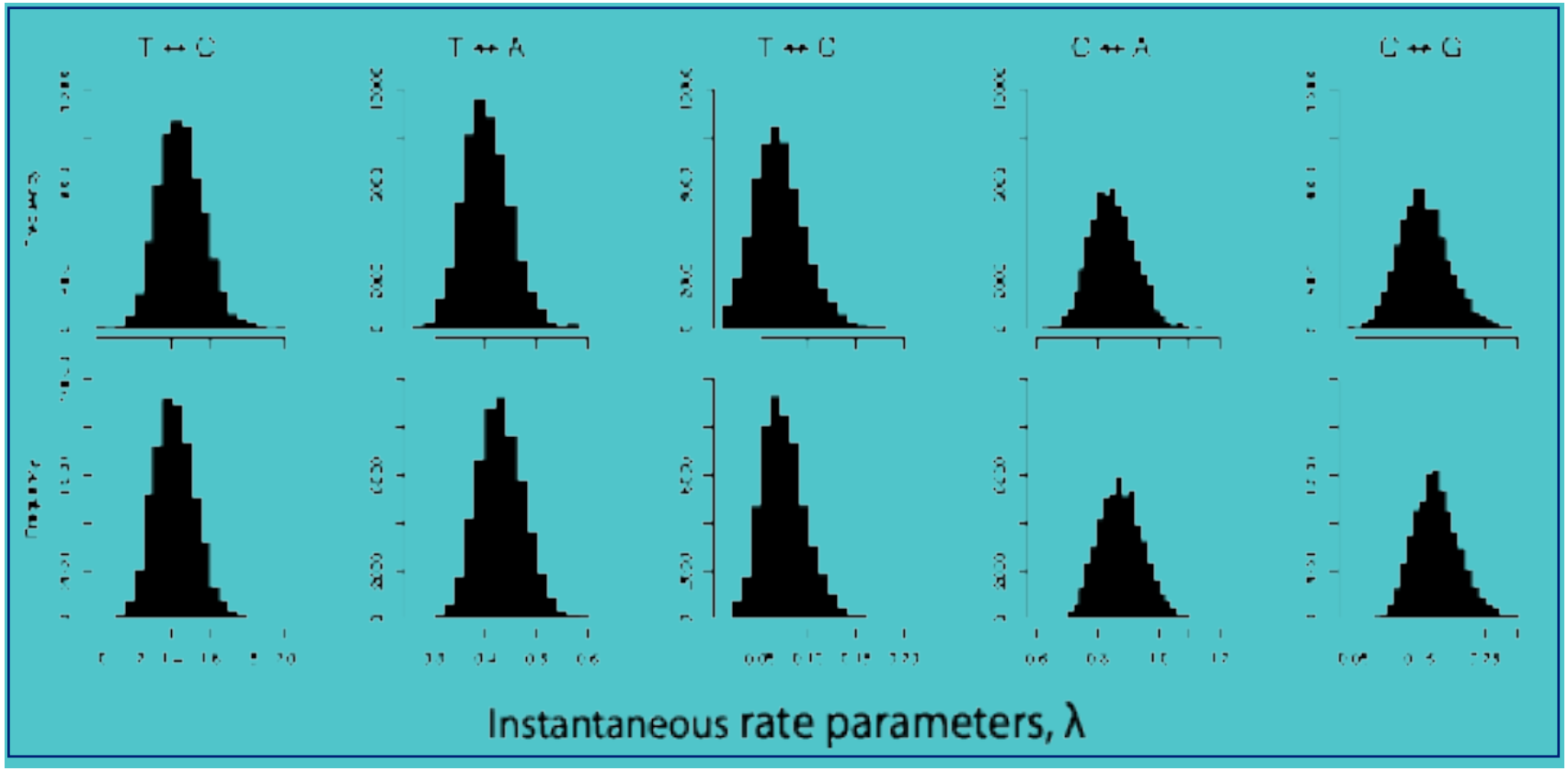
Much better than using “exotic” computation strategies:
GPU speedup is only ~100x for codon models
(Suchard & Rambaut 2009)

Posterior Parameter Distributions

GTR, mammalian *cyt-b*

Full

B1u



	Full	*B1u* (t≤0.08)	*B1u* (t≤0.02)
10 taxa	-7,586.47	-7,586.90	
224 taxa	-99,391.12	-99,542.03	-99,446.88

Evaluating Changes in Rate and Branch Length

- Classic likelihood calculations

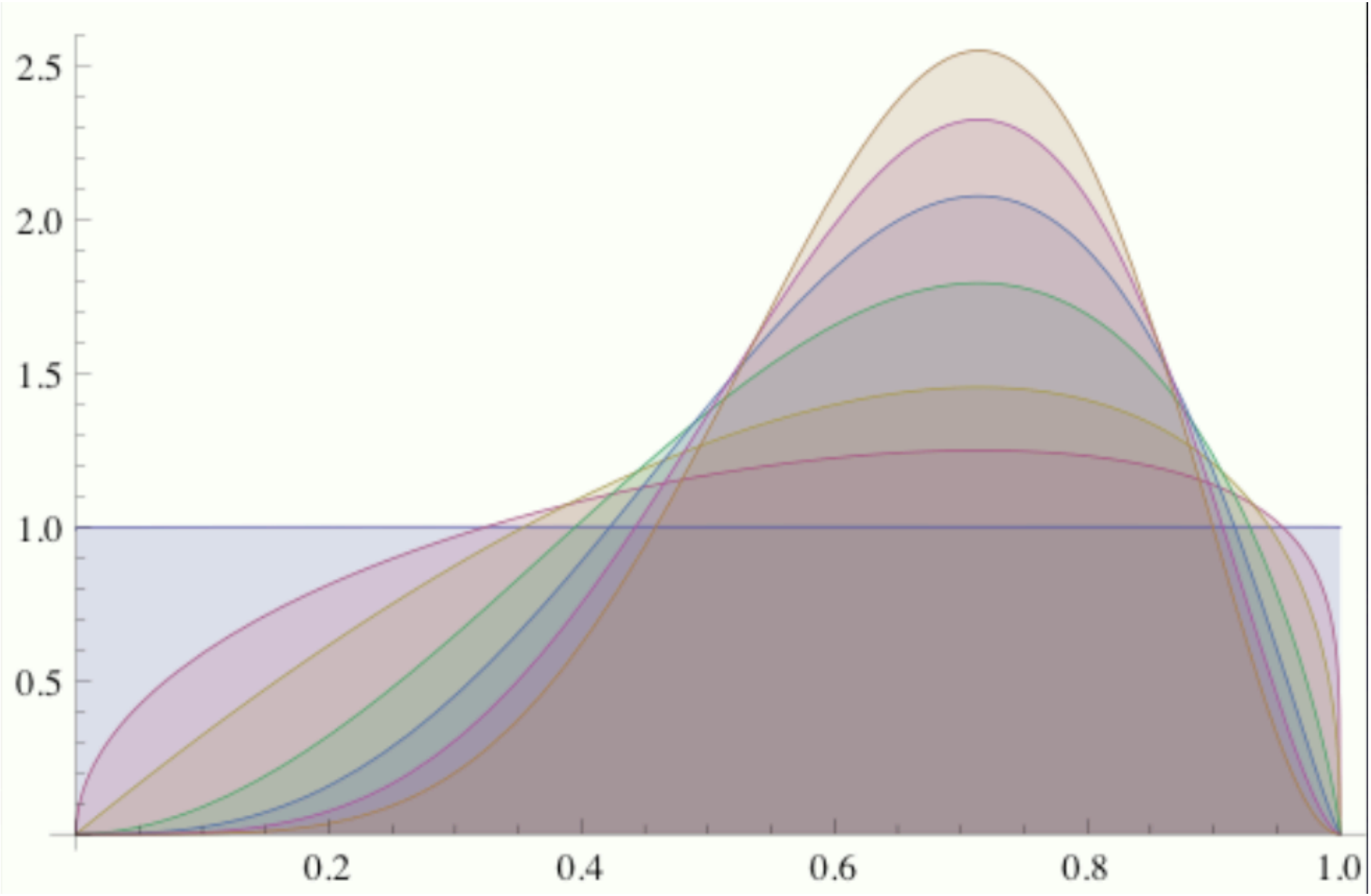
$$O\left(N^4 + N^3 b + N^2 b\right)$$

- Sampled substitution histories

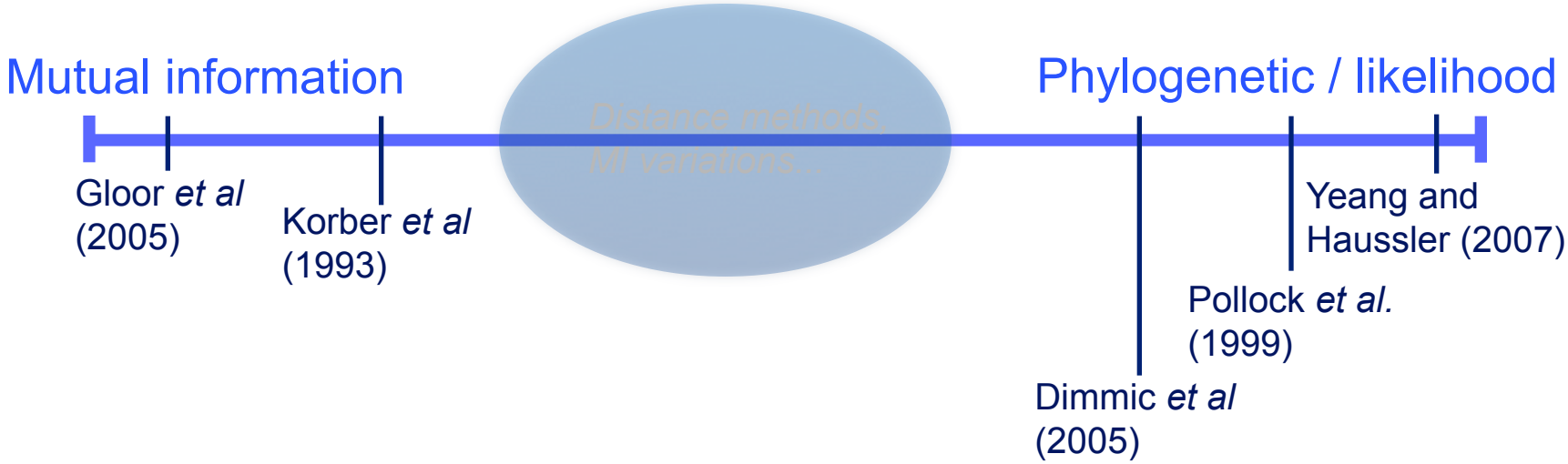
$$O(1)$$

  - Branch lengths and rate parameters can be evaluated separately and have analytically solvable posterior distributions

Thermodynamic Integration



Pairwise Coevolution Approaches



Tree and model ignorant
Fast and easy

Popular!

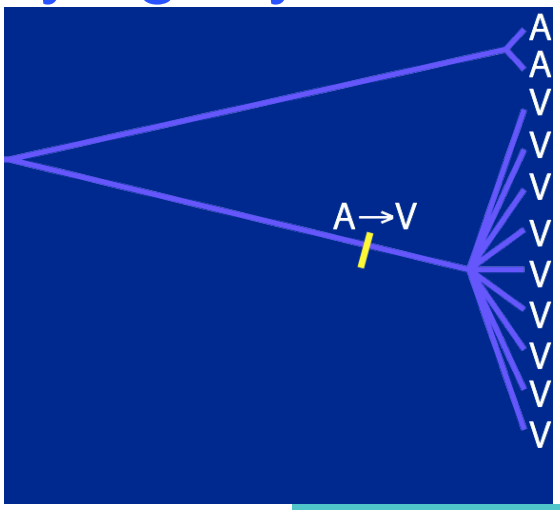
Phylogenetic & model-based
Slow and hard

Feared!

$$MI = \sum_i \sum_j P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$$

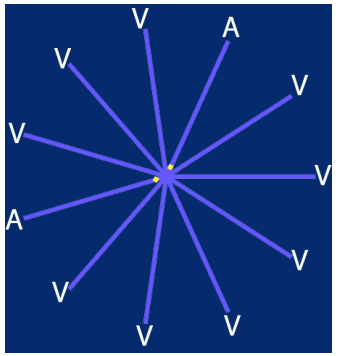
Mutual information methods are misled by:

(1) Phylogeny

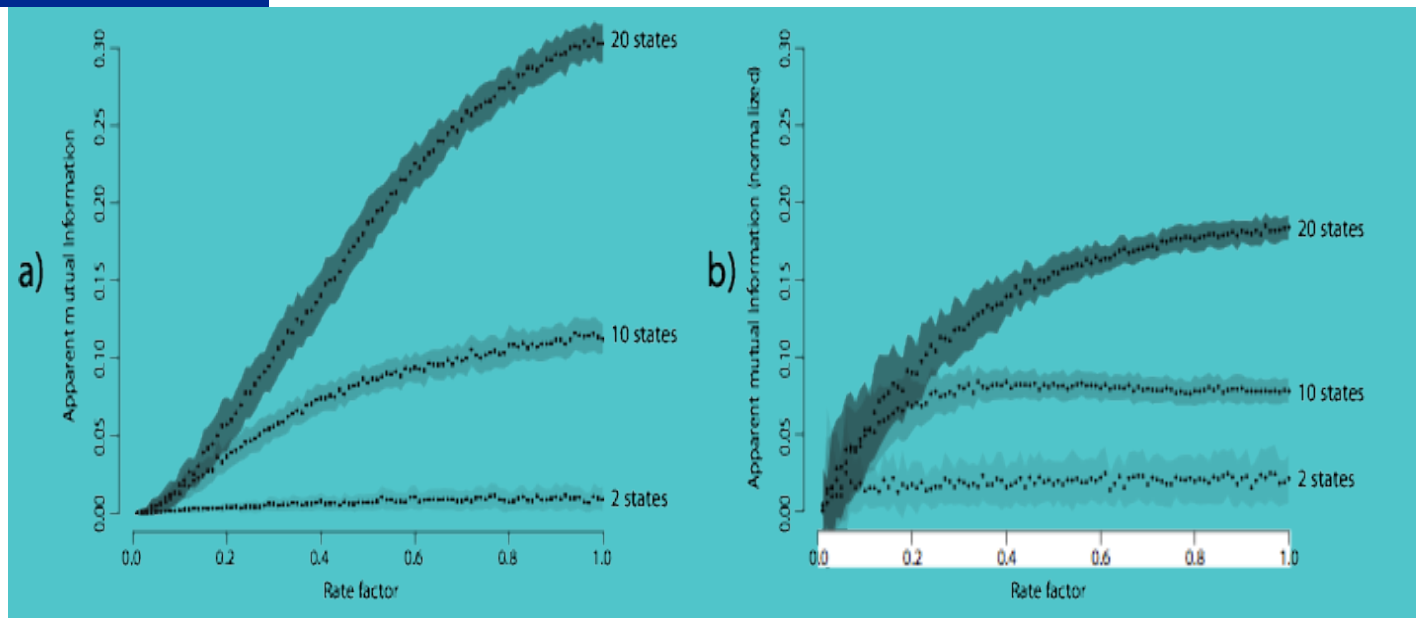


Observed Frequencies
V: 9/11 = 82%
A: 2/11 = 18%

Actual Time Spent
V: 3.3/6.9=48%
A: 3.6/6.9=52%



(2) Molecular Evolution



Phylogenetically-integrated MI (*pMI*)

$$P(z) = \int_0^{t_{b'}} \frac{P(x \to z|s)\,P(z \to y|(t_{b'} - s))}{P(x \to y|t_{b'})}\,ds =$$

$$\begin{cases} \frac{1}{\Lambda_x - \Lambda_y} + \frac{e^{\Lambda_y t_{b'}} t_{b'}}{e^{\Lambda_y t_{b'}} - e^{\Lambda_x t_{b'}}} & z = x \\ t_{b'} - \frac{1}{\Lambda_x - \Lambda_y} - \frac{e^{\Lambda_y t_{b'}} t_{b'}}{e^{\Lambda_y t_{b'}} - e^{\Lambda_x t_{b'}}} & z = y \\ t_{b'} & z = x = y \\ 0 & otherwise \end{cases}$$

$$P(z_i, z_j) = \int_0^{t_{b'}} \frac{P(x_i \to z_i|s)\,P(z_i \to y_i|(t_{b'} - s))}{P(x_i \to y_i|t_{b'})} \frac{P(x_j \to z_j|s)\,P(z_j \to y_j|(t_{b'} - s))}{P(x_j \to y_j|t_{b'})}\,ds$$

=24 cases

- Unrestricted *non-reversible* amino-acid substitution with gamma-distributed rate variation among sites
- Posterior-predictive null distribution for automated significance testing
- Fast! Roughly 6,200 pairs of sites per second (Yeang and Haussler, 2008: 29.35 seconds per site pair on a slower CPU):
  - approximate corrected speed-up about 100,000X

Work in Progress

- Context dependent nucleotide substitution
- Amino acid mixture models (dependent rates)
- Overlaid nucleotide and fitness models
  - Gradient mixture models
- Whole molecule fitness
  - Transcription factors and binding sites
  - Protein stability and function

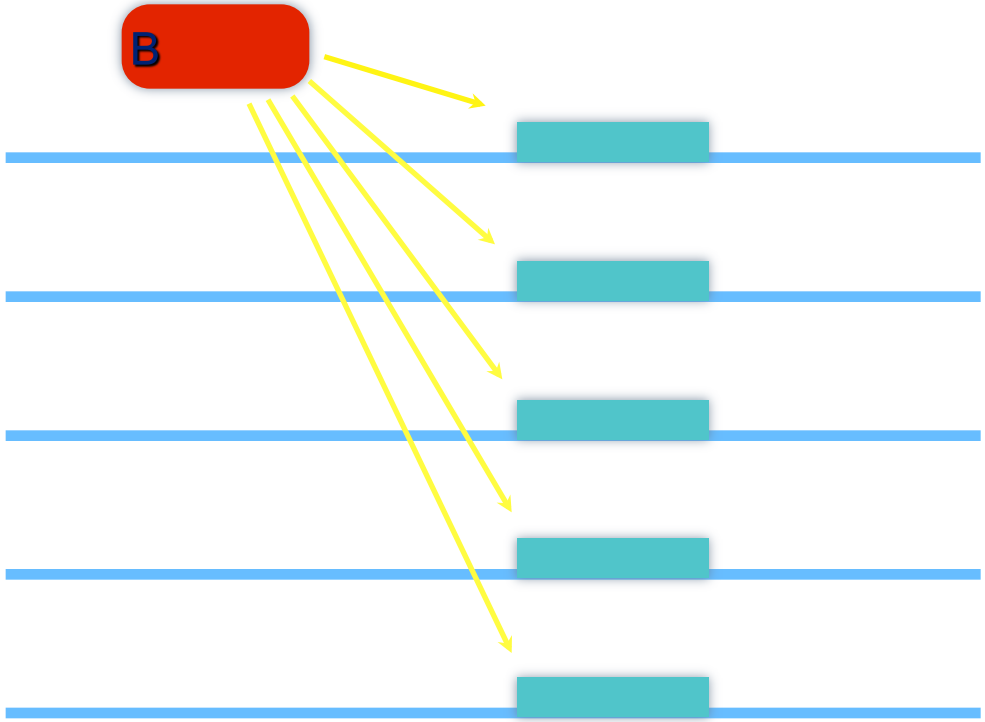
Summary

- Partial sampling of substitution histories with B1u integration eliminates the most burdensome aspect of MCMC based phylogenomic analysis
- Accuracy is high; precision can be tuned by decreasing the threshold of branch bisection
- Should largely alleviate the pressure for convenience-motivated simplifications

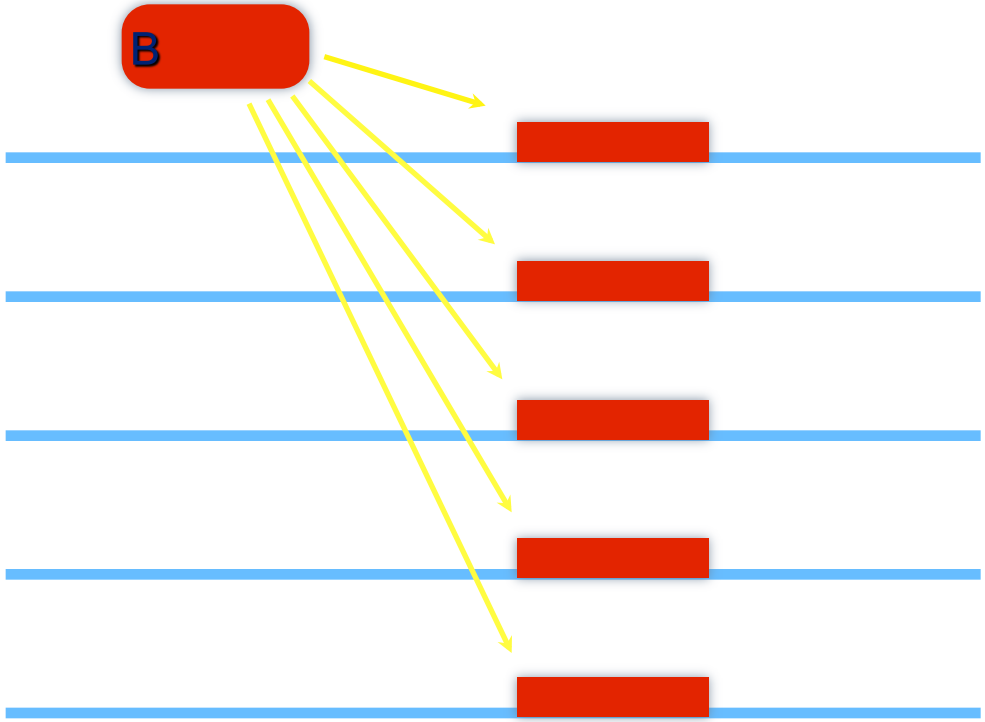
Acknowledgements: Jason de Koning, Wanjun Gu, Todd Castoe; Richard Goldstein, Nicolas Rodrigue

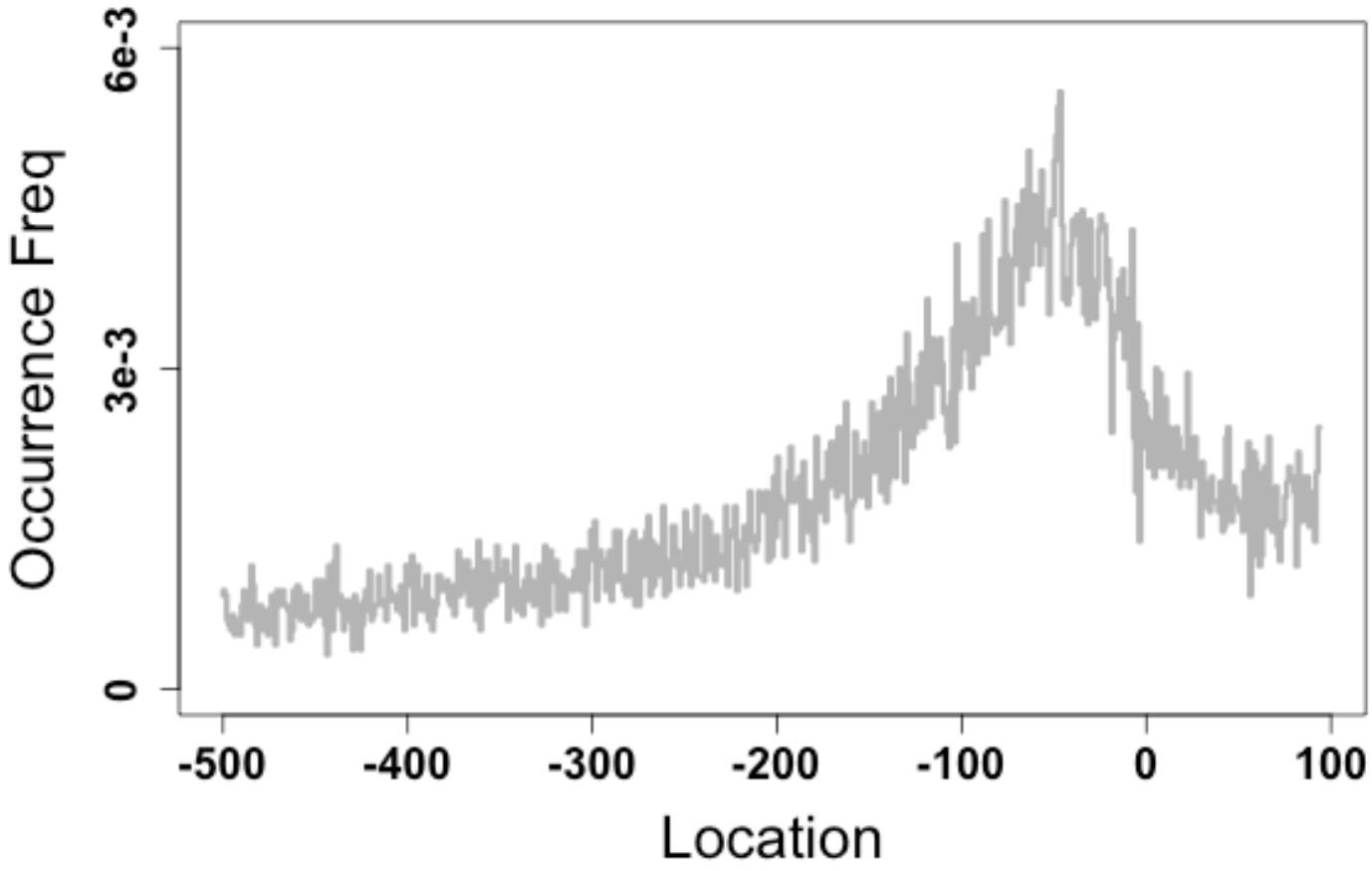
TF binding modifications

TF binding modifications



TF binding modifications




Occurrence Freq
0
3e-3
6e-3
-500
-400
-300
-200
-100
0
100
Location

Adaptation, Coevolution and Convergence

- Normal non-Adaptive Evolution
- Adaptive evolution drives a different mode of coevolution and convergence