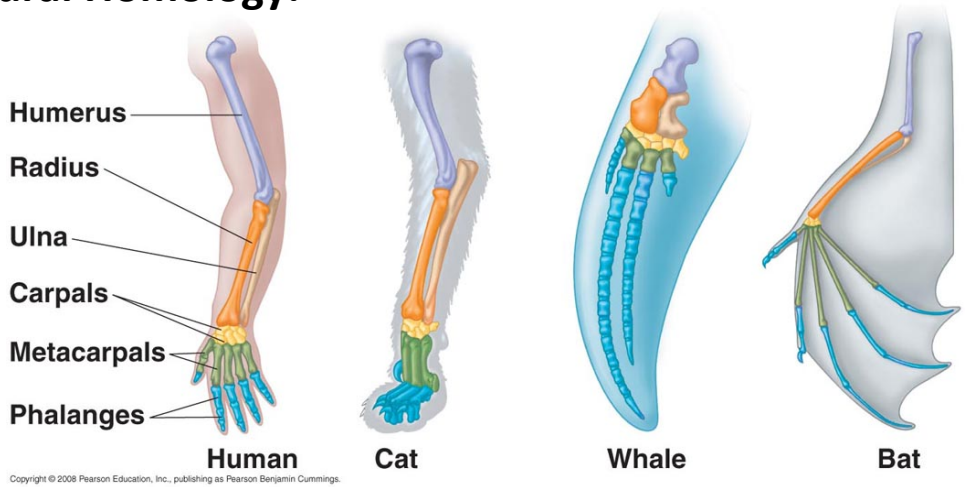
Evolutionary Conservation

Homology

Structural Homology:



Sequence Homology (Alignment):

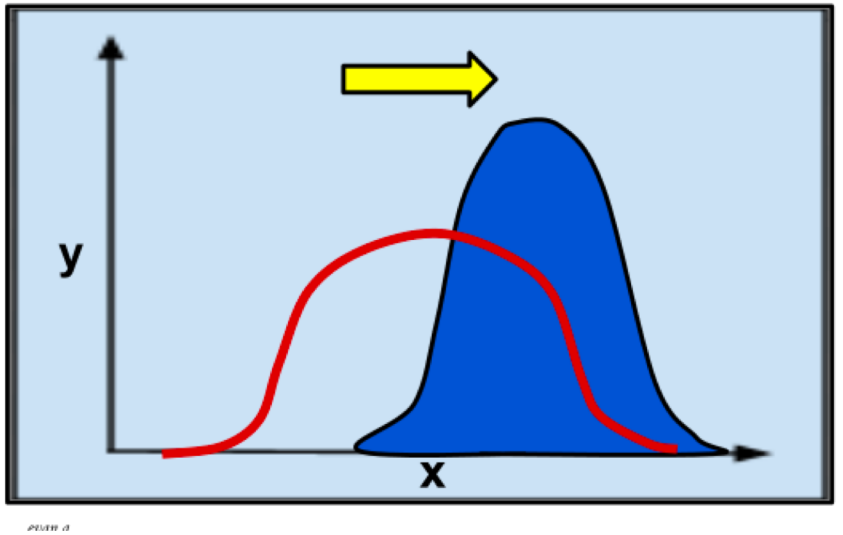
```
Human KKASKPKKAASKAPTKKPKATPKATPVKKAKKKLAAT
Mouse KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKPAAT
  Rat KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKPAAT
  Cow KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKPAAT
Chimp KKAAKPKKAASKAPSKKPKATPKATPVKKAKKKLAAT
      ***:**********:******************** ***
```

Natural Selection

In populations:

- *When the selection coefficient of an allele > 0*
- *Leads to the increase in the frequency of alleles that provide fitness advantages* ***-> fixation.***

- *These fixed alleles then are maintained by* ***purifying selection.***



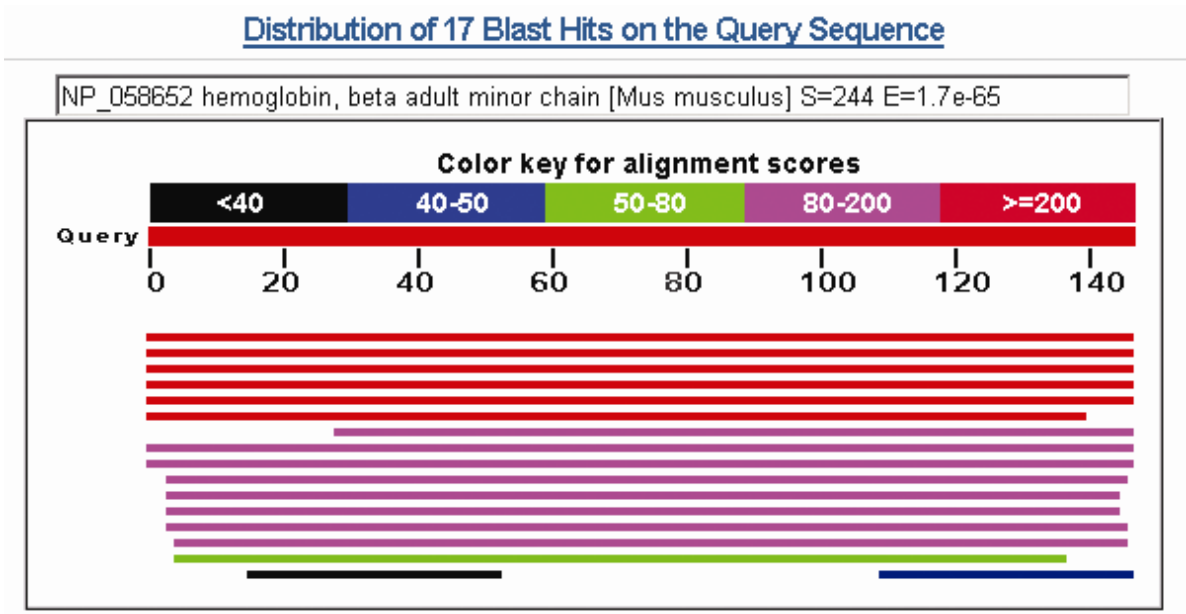
wikipedia; Ealbert17, CC BY-SA 4.0

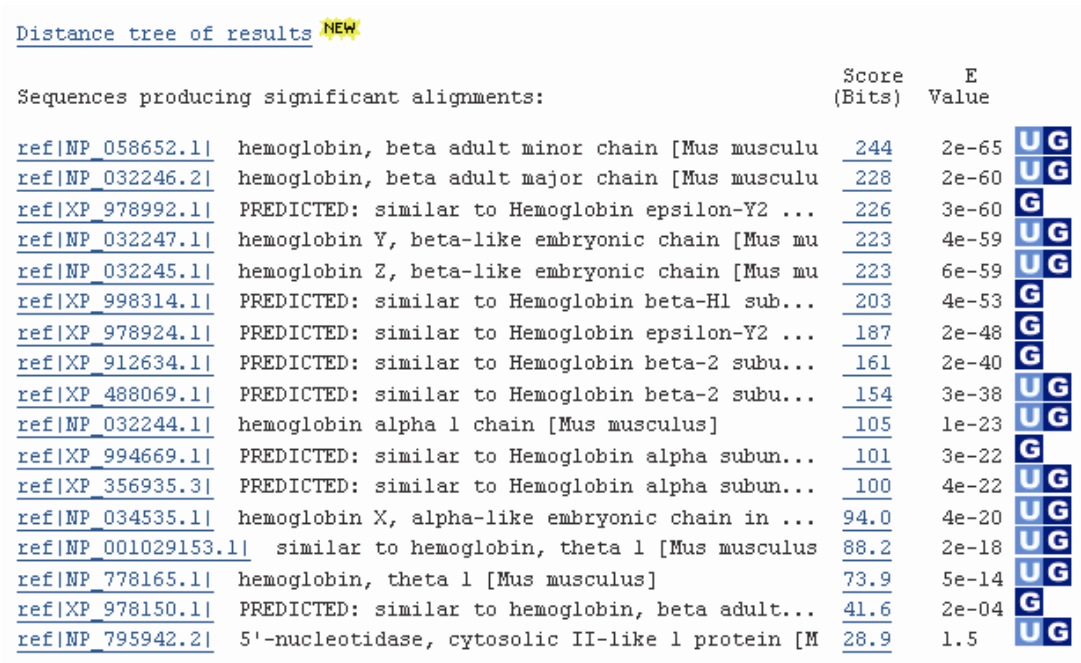
Selection coefficient	**Interpretation**	**Effect on population frequency of allele**
$s > 0$	advantageous	Increases
$s < 0$	deleterious	Decreases
$s = 0$	neutral	No predictable effect (except when drift is strong)

Sequence Similarity

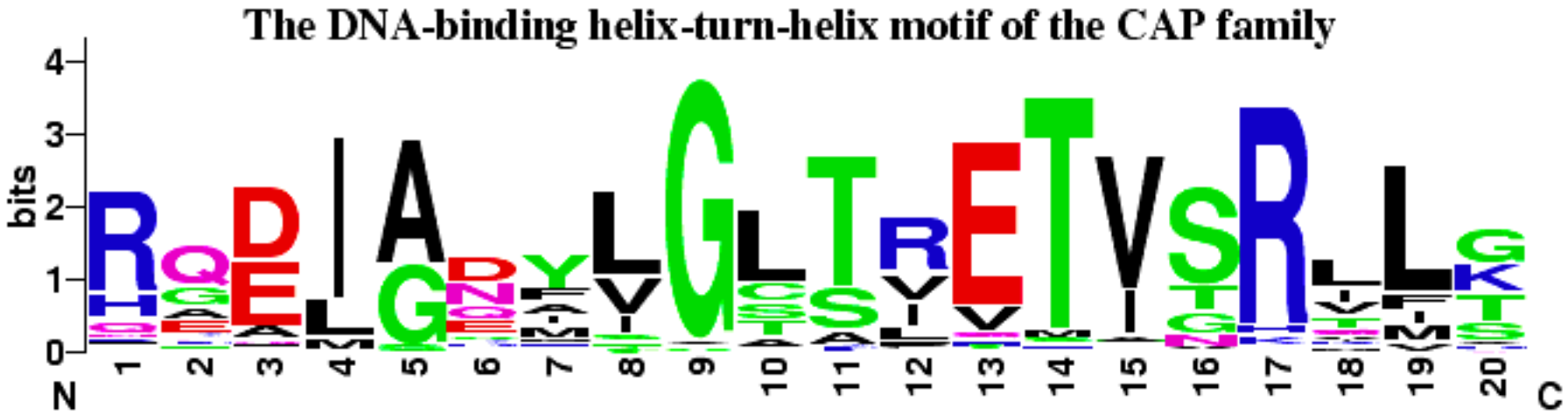
Across genomes:

- *Natural selection maintains sequence similarity*
- ***Sequence similarity maintains function.***





Conservation at Sites in An Alignment

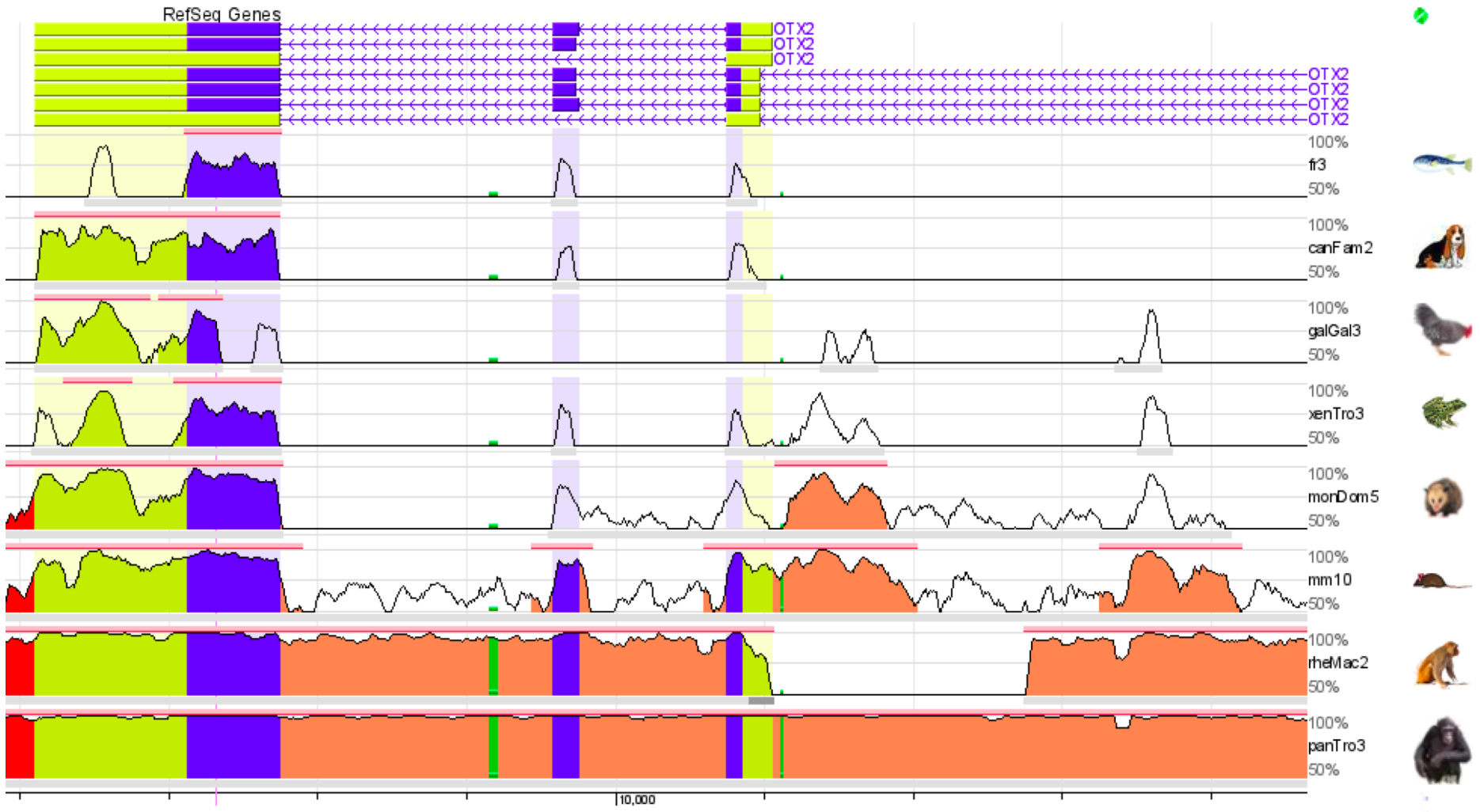


http://weblogo.berkeley.edu/examples.html

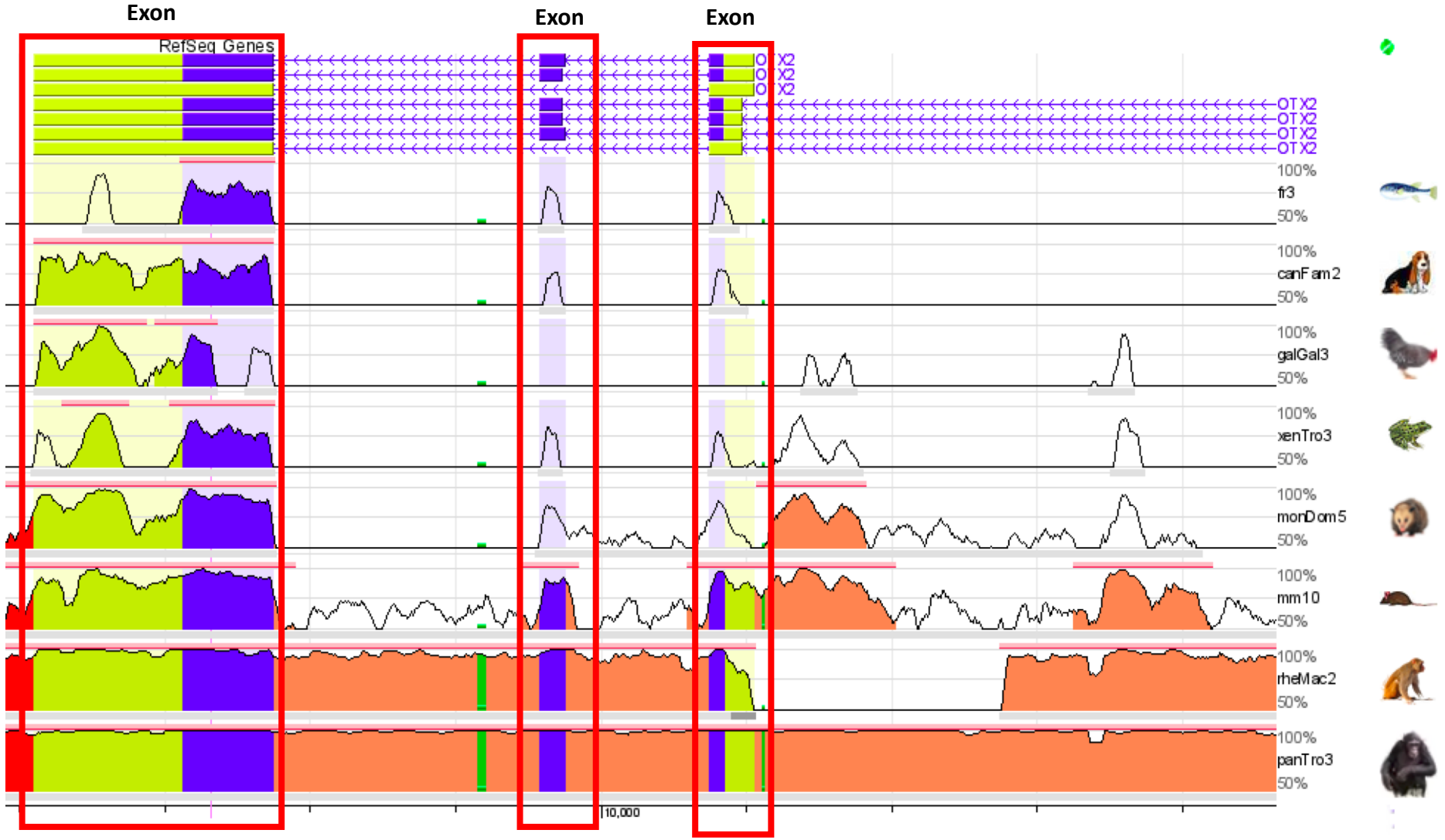
Catabolite Activator Protein is a transcription promoter that binds to >100 sites in the *E. coli* genome.

Positions 11-14, 17, and 20 are critical to the sequence specific binding motif and are more ***conserved***.

Conservation Across Genomes in the OTX2 Gene



Conservation Across Genomes in the OTX2 Gene



Alignment

The most fundamental problem in genomics.

Seeks to establish a mapping between the letters of a set of sequences

Fairly straightforward with closely related taxa

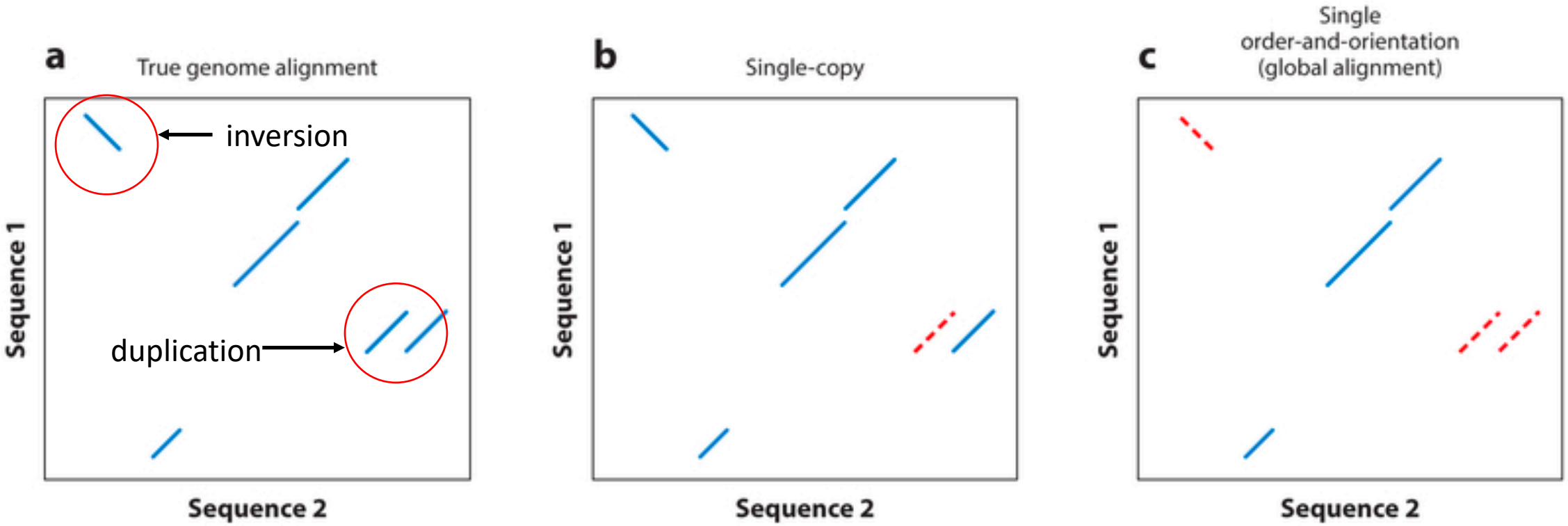
Early attempts relied on global alignments.

Smith & Waterman (1981) specialized these to produce local alignments, which optimize aligned subsequences.

But at large evolutionary distances, genomes contain more complicated differences.

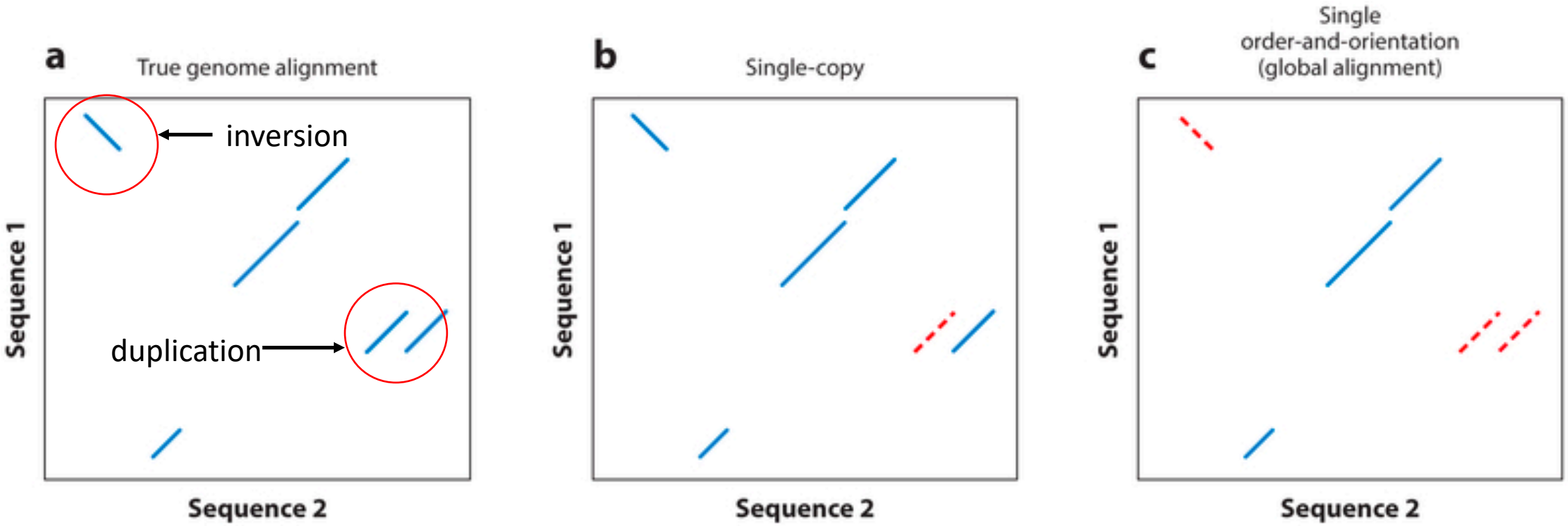
- *Duplications*
- *Inversions*
- *Transpositions*
- *Rearrangements*

Challenges with Genome Alignments



Armstrong et al. 2019. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci.

Challenges with Genome Alignments



Armstrong et al. 2019. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci.

Global alignment fails to detect inversions and duplications.

The Single-Copy Heuristic

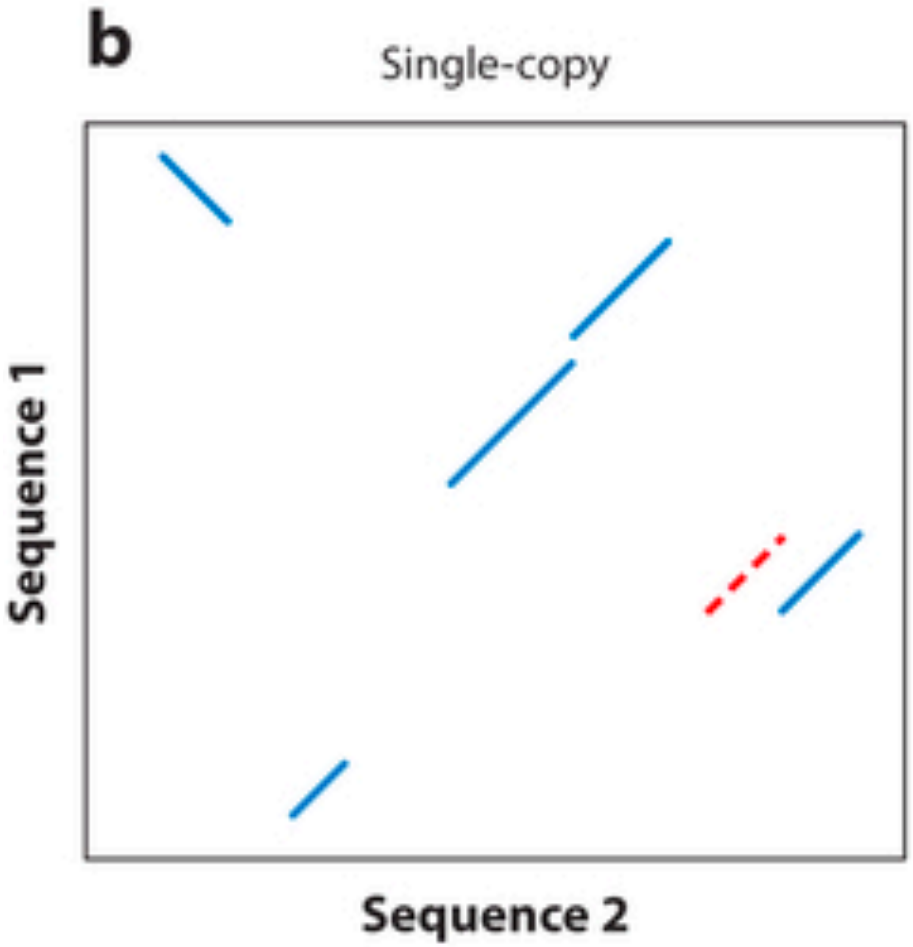
Chooses the single best target alignment for each region

Simplicity:

- Based on alignment score or percent identity

Overly simplistic:

- As we know, the reciprocal best-fit does not guarantee finding an ortholog.
- Single-copy alignments assume one-to-one orthology.
- Can miss some divergent duplications



Armstrong et al. 2019. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci.

BLASTZ (and LASTZ)

1. Remove lineage-specific interspersed repeats from both sequences.
2. For all pairs of spaces 12-mers (one from each sequence) that are identical except perhaps for one transition, do the following:
   a) Extend the induced alignment in each direction, not allowing gaps. Stop extending when the score decreases more than some threshold.
   b) If the gap-free alignment scores more than 3000, then
      a) Repeat the extension step, but allow for gaps.
      b) Retain the alignment scores above 5000.
3. Between each pair of adjacent alignments from step 2, repeat step 2, but using a more sensitive seeding procedure (e.g. 7-mer) and lower score thresholds
4. Adjust sequence positions in the resulting alignments to make them refer to the original sequences.
5. Filter the alignments as appropriate for particular purposes

BLASTZ can require that matching regions occur in the same order and orientation in both sequences.

Uses a scoring matrix:

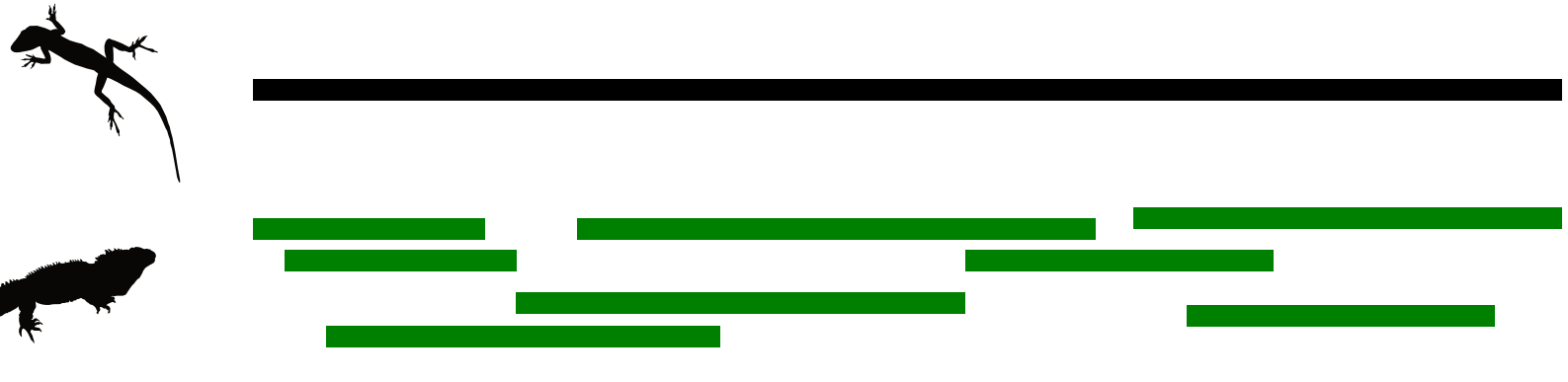
	A	C	G	T
A	91	-114	-31	-123
C	-114	100	-125	-31
G	-31	-125	-100	-114
T	-123	-31	-114	91

Scoring matrix penalizes nucleotide substitutions and makes it harder to trigger a gapped alignment.

Schwartz et al. 2002. Human-Mouse Alignments with BLASTZ. *Genome Research.*
Harris. 2007. LASTZ. *PhD Thesis.*

Local Alignments with LASTZ

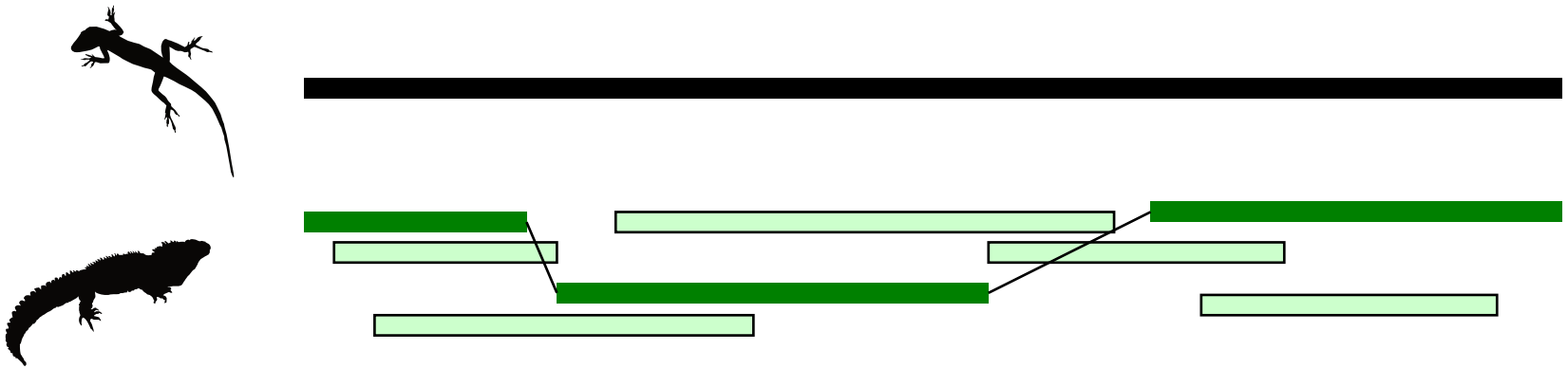
Anolis ***reference*** (AnoCar2.0)



Sphenodon ***query***

Gemmell et al. 2020. The genome of the tuatara reveals features of ancient amniote evolution. *Nature.*

Chaining: form gapless blocks

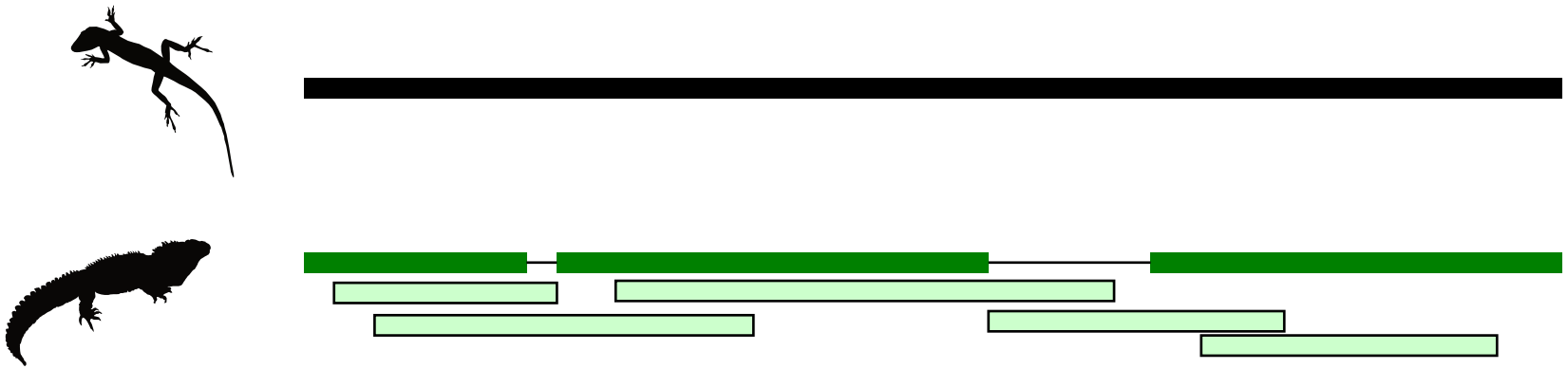


Maximal scoring combinations of local alignments
Maintains a single order and orientation
Filters our spurious alignments

- these often form short low scoring chains anyway.

Gemmell et al. 2020. The genome of the tuatara reveals features of ancient amniote evolution. *Nature.*

Netting: rank and collect highest scoring chains



These chains must cover the reference genome only once.

Netting makes finding high-confidence rearrangements easier.

The result is called a ***pairwise syntentic net.***

> *Kent et al. 2002. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. PNAS.*

Gemmell et al. 2020. The genome of the tuatara reveals features of ancient amniote evolution. *Nature.*

Pairwise Genome Aligners

Program	Reference	Description
MUMmer	Marçais et al. 2018. MUMmer4: a fast and versatile genome alignment system. *PLOS Comput. Biol*.	Fast aligner relying on maximal unique matches from a query sequence to a reference sequence; recent versions remove the colinearity restriction of the first version and improve the speed
Chains and nets	Kent et al. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *PNAS*	Combines fragmented local alignments into larger, high-scoring chains, which are arranged into hierarchical nets representing rearrangements
Shuffle-LAGAN	Brudno et al. Glocal alignment: finding rearrangements during alignment. *Bioinformatics*	A glocal (global + local) aligner that is less restrictive than global alignment but still enforces monotonicity of the blocks relative to one sequence

Multiple Alignment

- Often it is necessary to consider the alignment between a set of more than two sequences.
- A multiple alignment is an equivalence relation on a set of sequences.
- Multiple aligners estimate orthology or homology between bases.
- Alignments are partitioned into columns by the equivalence class.
  - Every base is related to all bases in its column
  - No two bases in different columns are related
- Progressive alignment is a useful heuristic method
  - A guide tree is used that represents the known phylogeny of the sequences
  - The most closely related sequences are aligned first
  - the resulting alignment is itself aligned to other sequences or alignments
  - follows the structure of the guide tree.

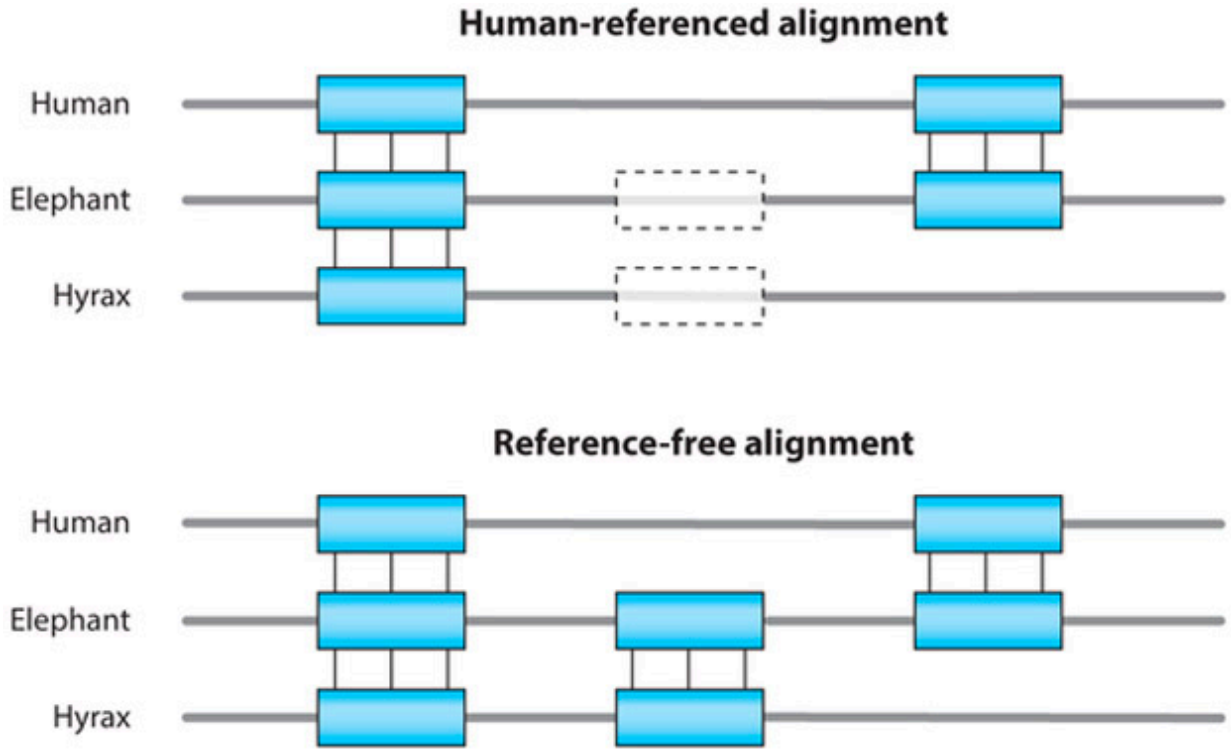
Using References in Multiple Alignments

Can be a useful heuristic

Bases the multiple alignment on a single reference.

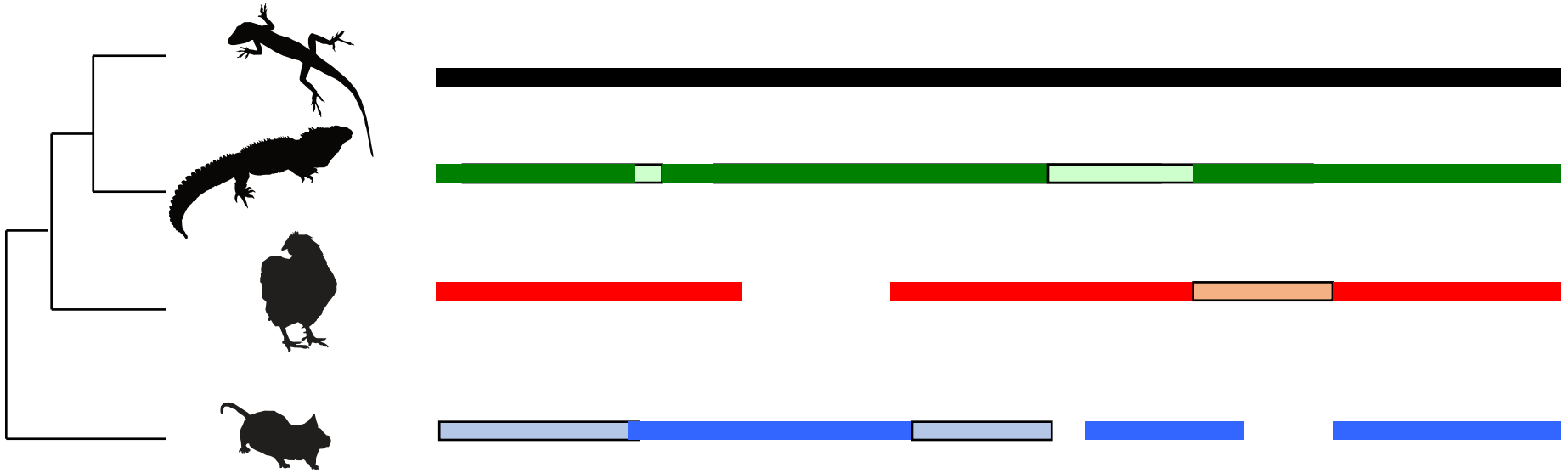
- All other sequences are aligned pairwise to the reference
- Then they are combined to form the multiple alignment

However, information relating to genomes distant to the reference is lost.



Armstrong et al. 2019. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci.

MULTIZ: create multiple alignment



- 28 vertebrates
- ~270Mbp of aligned sequence
- ~25% of the *Anolis* lizard genome = HOMOLOGY

Gemmell et al. 2020. The genome of the tuatara reveals features of ancient amniote evolution. *Nature.*

Multiple Genome Aligners

Program	Reference	Description
MULTIZ	Blanchette et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*.	Multiple alignment based on pairwise alignment from every genome to a single reference
progressiveMauve	Darling et al. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE*	Progressive aligner that attempts to remove anchors causing small rearrangements by optimizing a breakpoint-weighted score
Cactus	Paten et al. 2011. Cactus: algorithms for genome multiple sequence alignment. *Genome Res*.	Graph-based aligner that attempts to remove anchors representing small rearrangements

So you've constructed a multiple genome alignment

Now what?

Why are they useful?

So you've constructed a multiple genome alignment

Now what?

Why are they useful?

Identifying functional elements*!

So you've constructed a multiple genome alignment

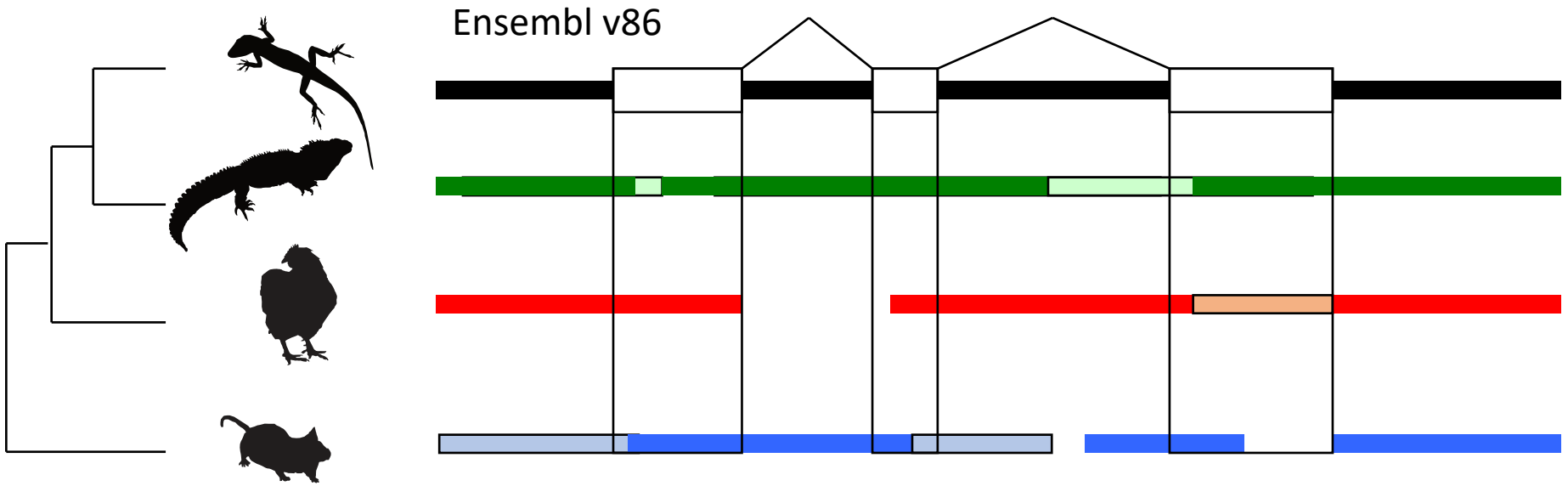
Now what?

Why are they useful?

Identifying functional elements*!

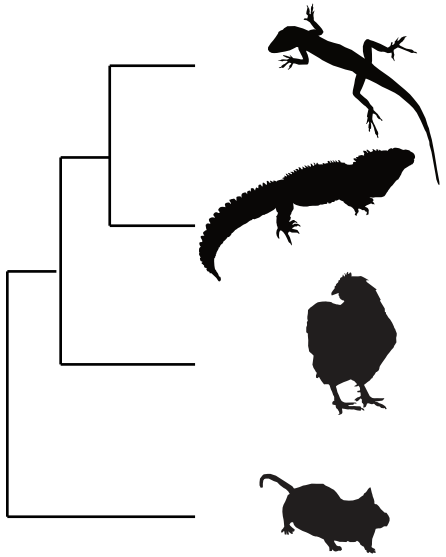
*Parts of the genome maintained by selection.

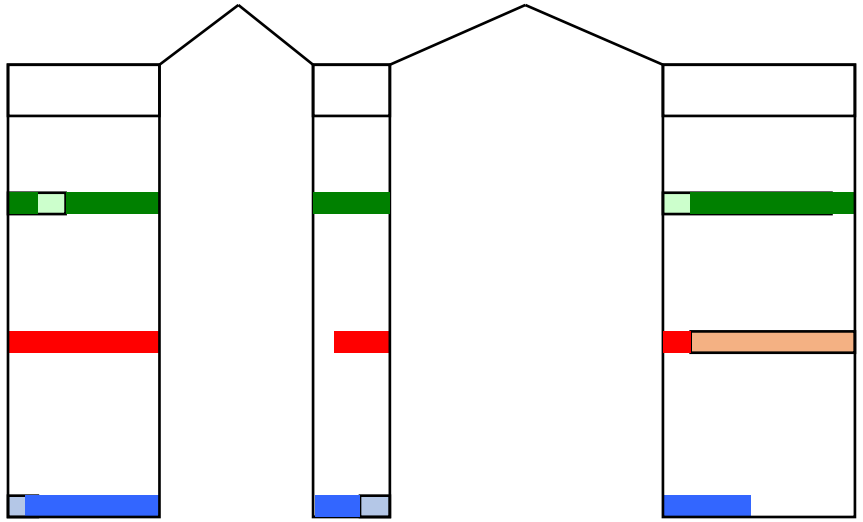
Extract annotations using reference



Gemmell et al. 2020. The genome of the tuatara reveals features of ancient amniote evolution. *Nature.*

Extract annotations using reference





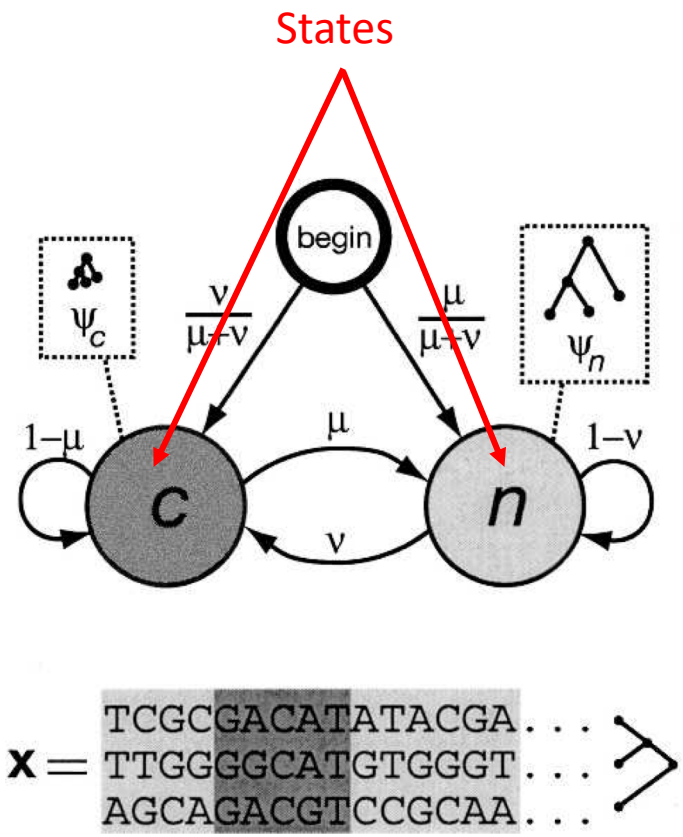
Gemmell et al. 2020. The genome of the tuatara reveals features of ancient amniote evolution. *Nature.*

phastCons

Identifies conserved elements in multiple genome alignments.

Uses a phylogenetic hidden Markov model (phylo-HMM)

There are two states for a genomic region in the phyloHMM.

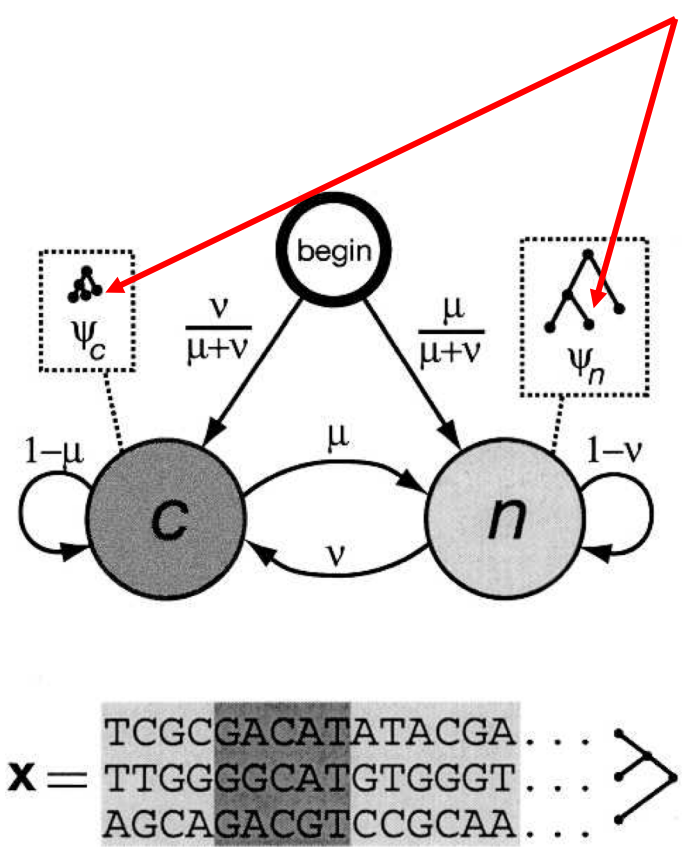


Siepel et al. 2005.Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research.*

phastCons

Identifies conserved elements in multiple genome alignments.

Uses a phylogenetic hidden Markov model (phylo-HMM)



Each state is associated with a phylogenetic model

These are identical except for a scaling parameter applied to Ψ_c.

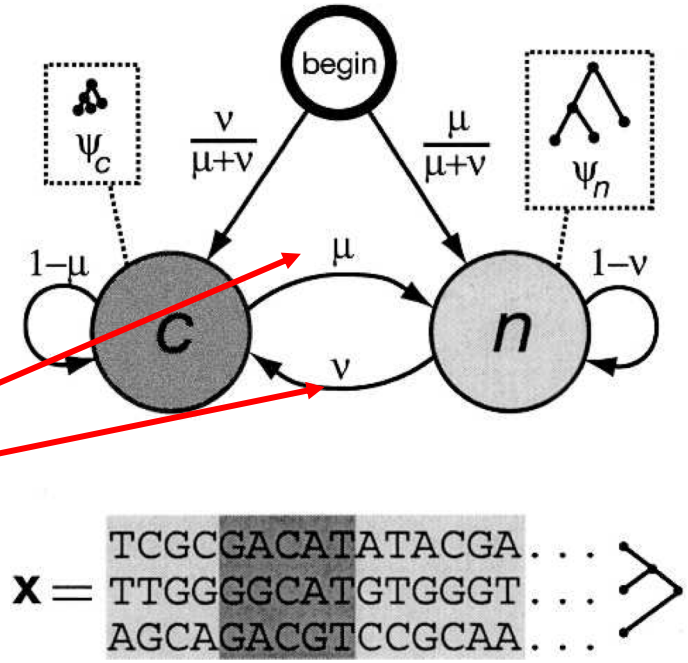
This represents the average substitution rate in conserved regions as a fraction of the average substitution rate in nonconserved regions.

Siepel et al. 2005.Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research.*

phastCons

Identifies conserved elements in multiple genome alignments.

Uses a phylogenetic hidden Markov model (phylo-HMM)



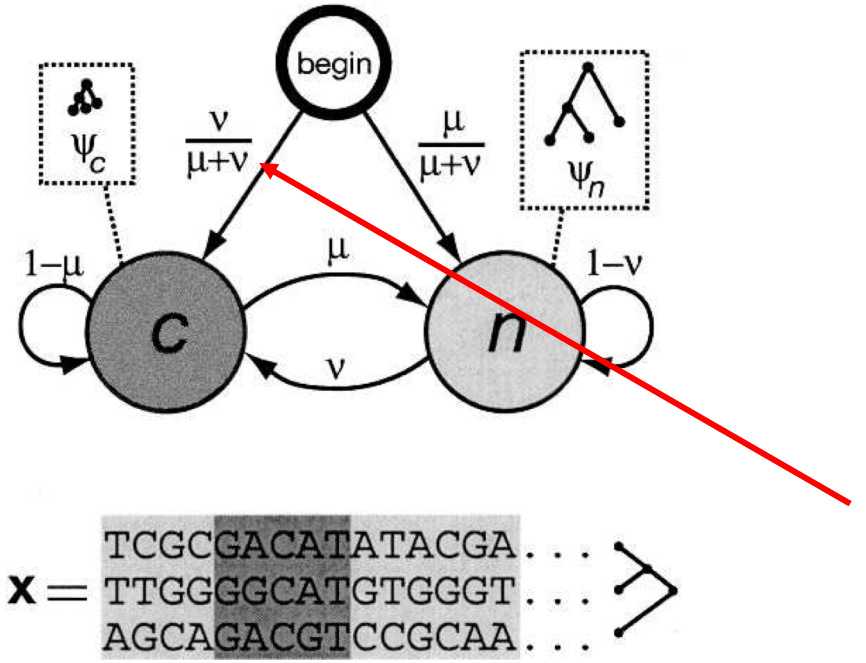
Two parameters define the state-transition probabilities.

Siepel et al. 2005.Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research.*

phastCons

Identifies conserved elements in multiple genome alignments.

Uses a phylogenetic hidden Markov model (phylo-HMM)



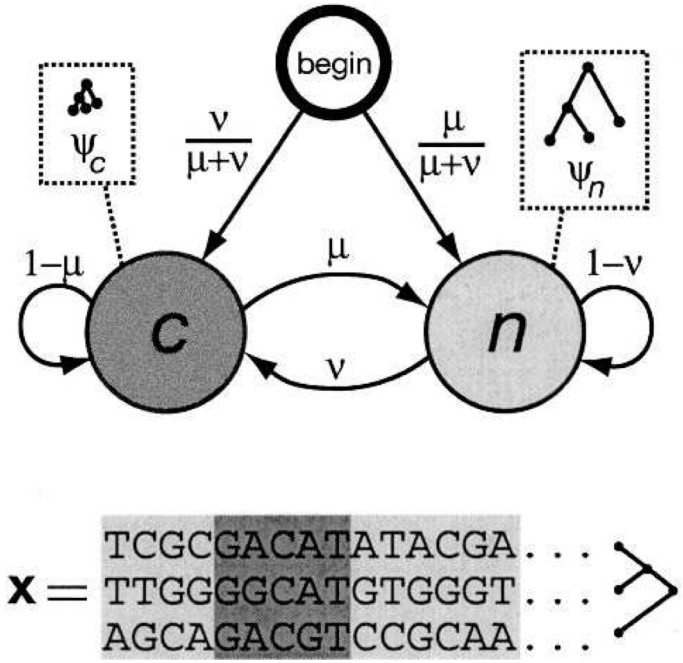
The probability of visiting each state first is set to the probability of that state at equilibrium.

Siepel et al. 2005.Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research.*

phastCons

Identifies conserved elements in multiple genome alignments.

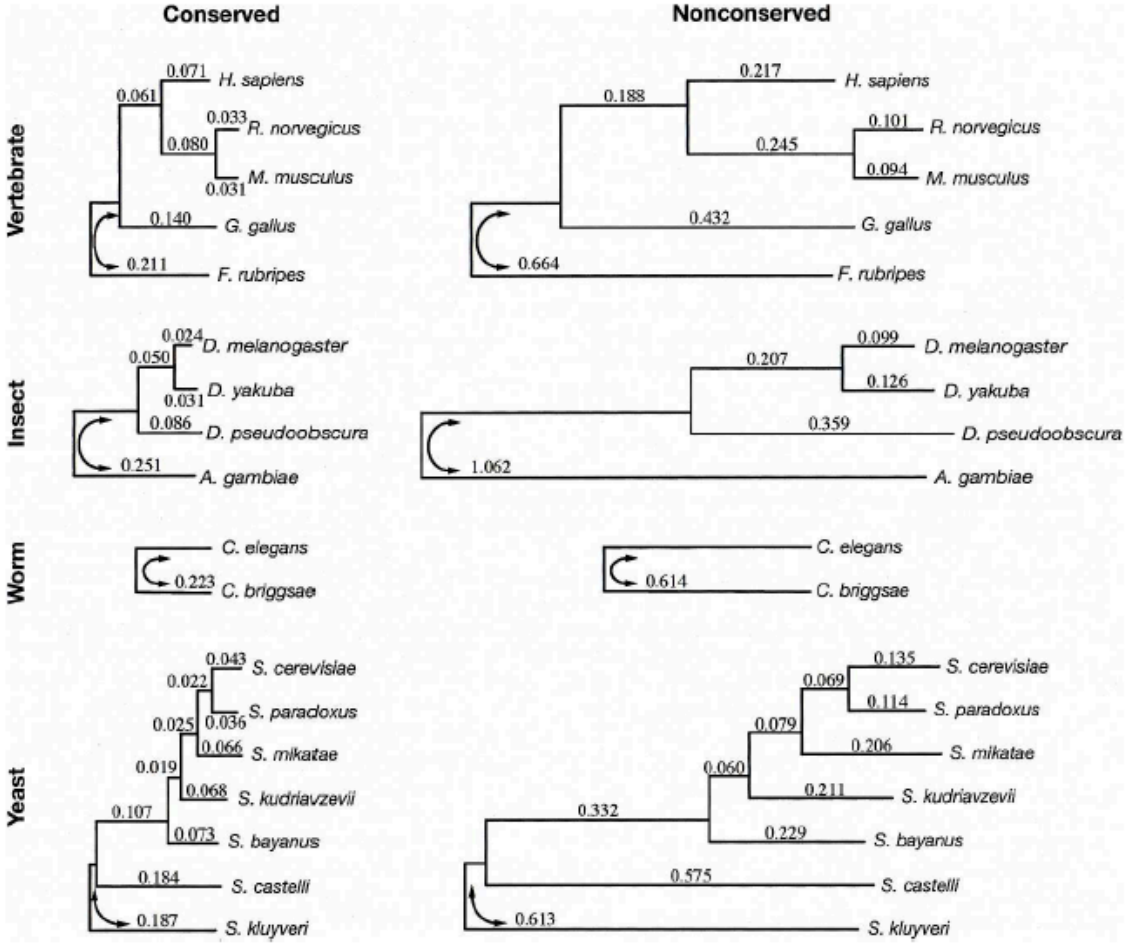
Uses a phylogenetic hidden Markov model (phylo-HMM)

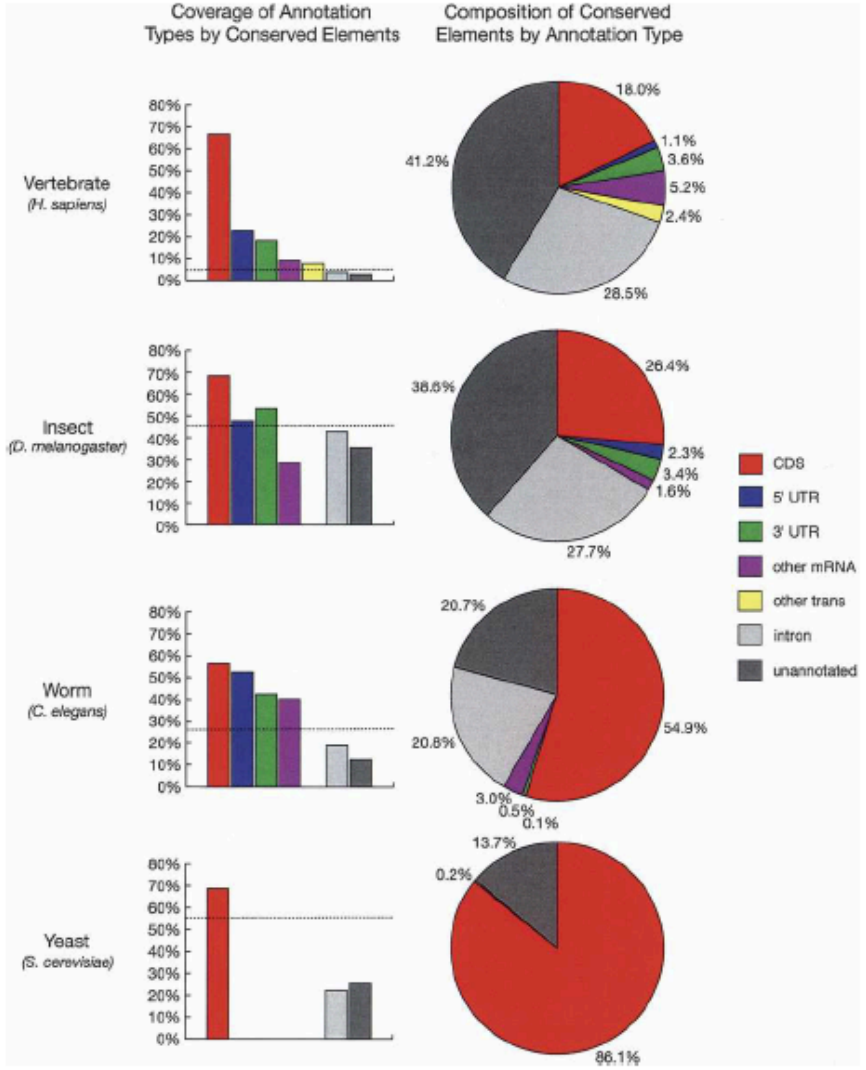


The model is a probabilistic machine that generates a multiple alignment consisting of alternating sequences of conserved and nonconserved alignment columns

Siepel et al. 2005.Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research.*

Using phastCons to identify conserved regions in eukaryote genomes

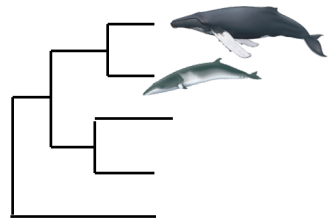




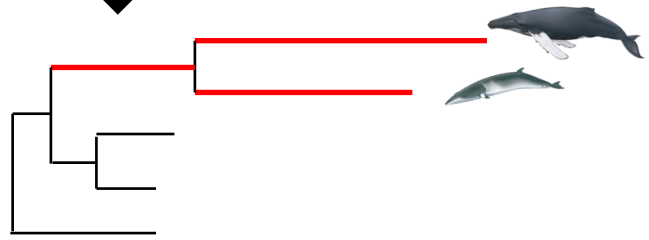
Siepel et al. 2005.Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research.*

What about estimating parts of the genome that *accelerated* in certain lineages over evolutionary time?

Neutral evolutionary model:



Alternative accelerated model:

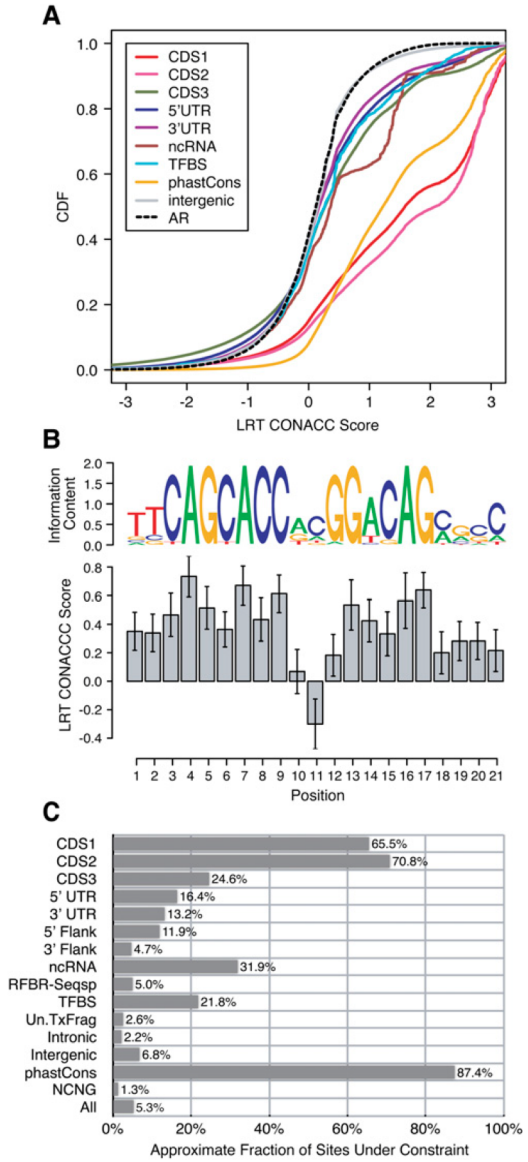


PhyloP

phyloP assigns a “conservation score” to all sites based on the p-value of the LRT.

These scores as a function of genomic position within 52 transcription factor binding sites.

Sites under constraint (purifying selection) in each annotation feature in an alignment of 36 mammals.



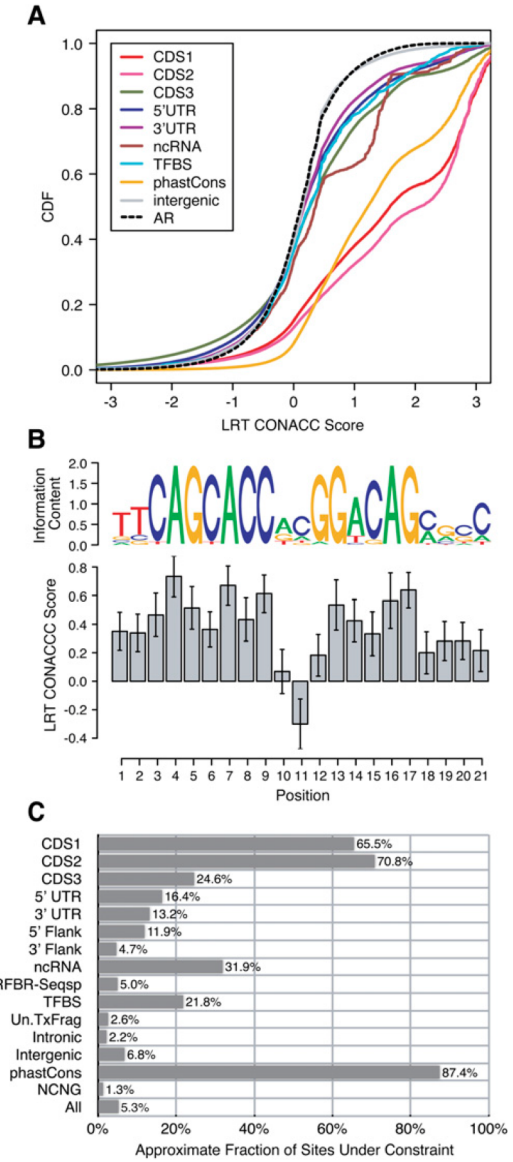
Pollard et al. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research.*

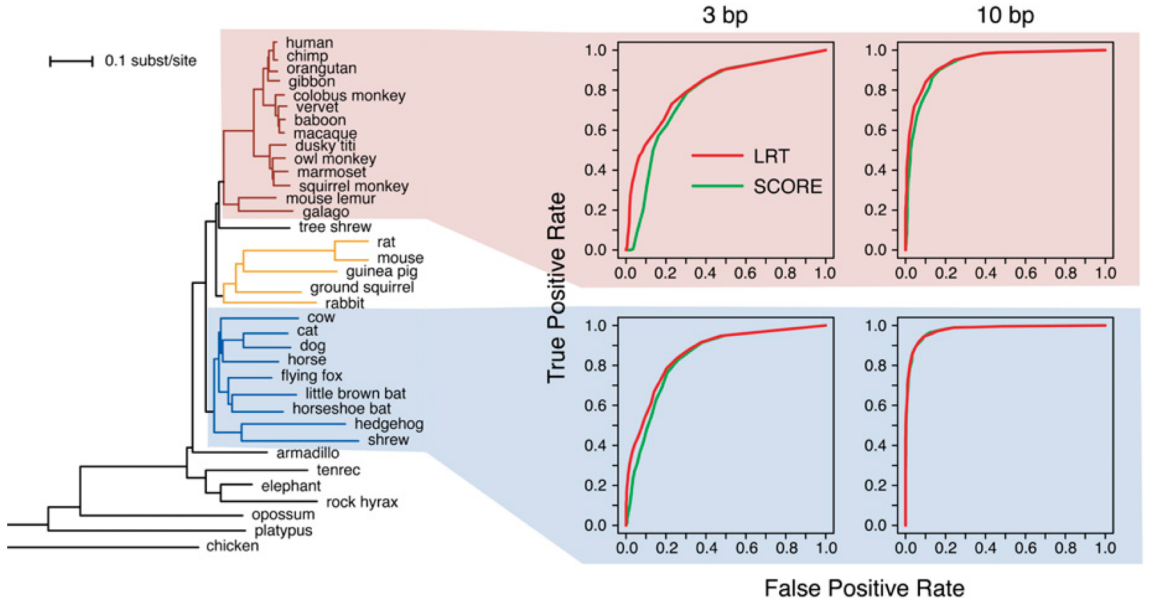
PhyloP

phyloP assigns a “conservation score” to all sites based on the p-value of the LRT.

These scores as a function of genomic position within 52 transcription factor binding sites.

Sites under constraint (purifying selection) in each annotation feature in an alignment of 36 mammals.



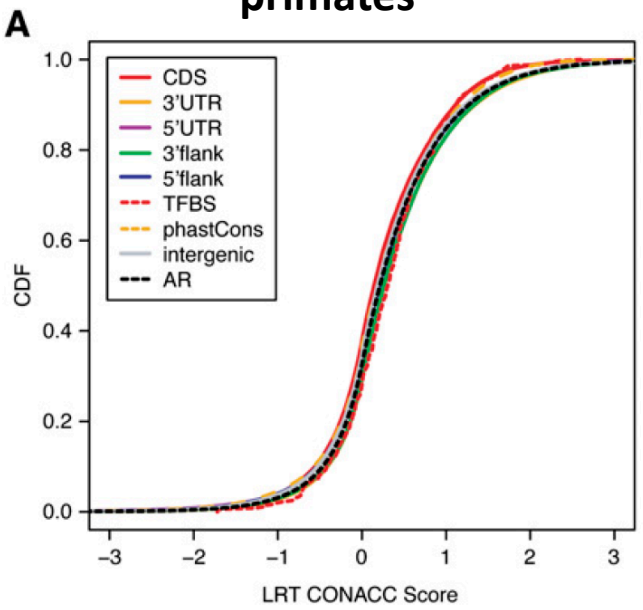


Receiver operating characteristic (ROC) curves show the ability of phyloP to detect shifts in the substitution rates (particularly with the LRT) of primates and laurasiatherians.

Pollard et al. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research.*

PhyloP

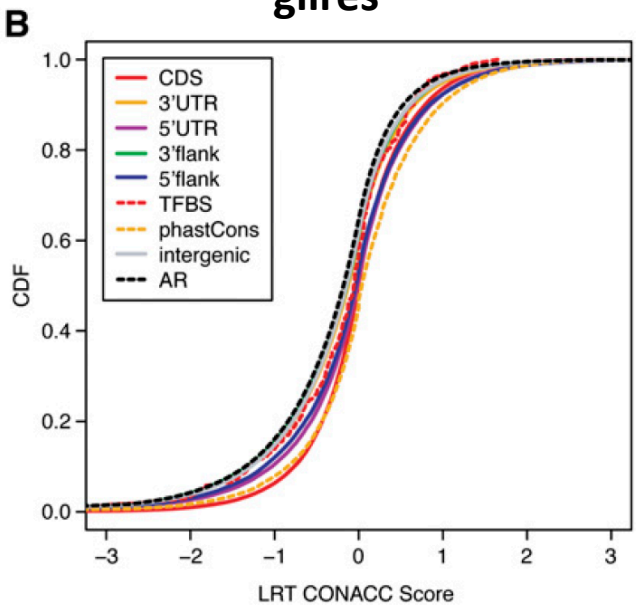
primates



phyloP found similar CONACC scores across all annotation features in primates.

This suggests a lack of a shift in the rates in these regions during primate evolution.

glires



In contrast, phyloP found major differences in the CONACC scores of different features in the glires clade.

This may be due to stronger ***purifying selection*** in glires.

Pollard et al. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research.*