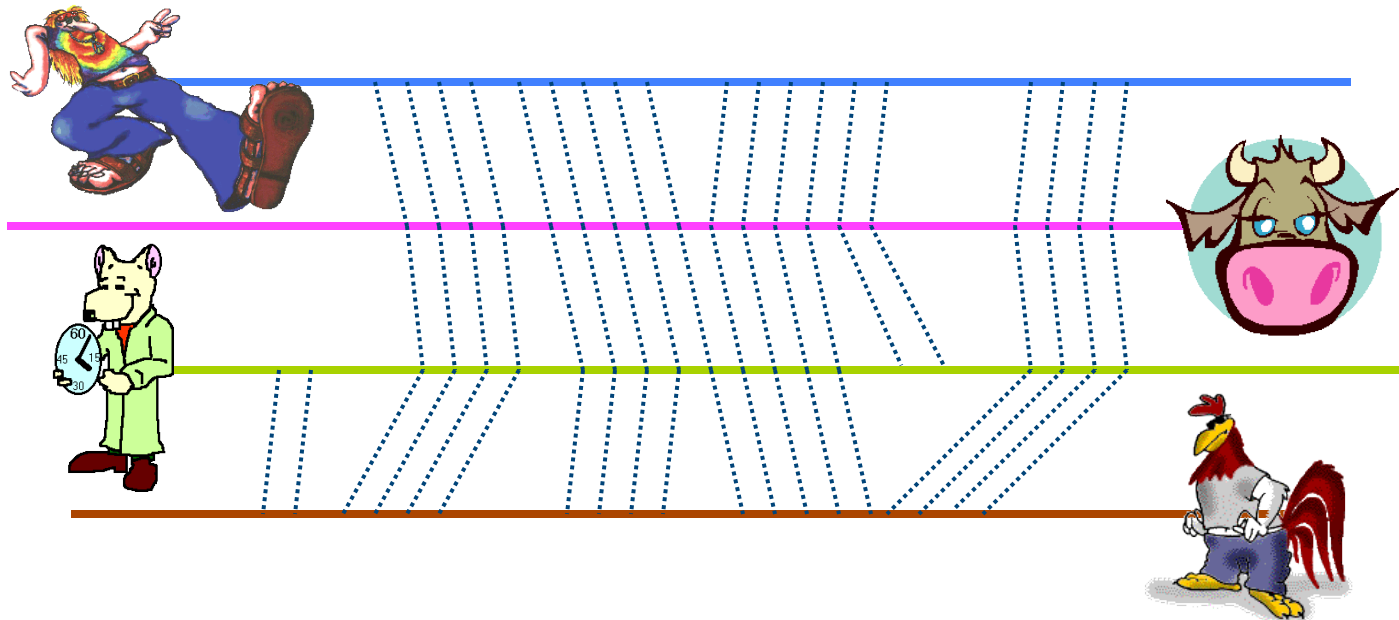




# Multiple Sequence Alignment

## Gene Finding, Conserved Elements

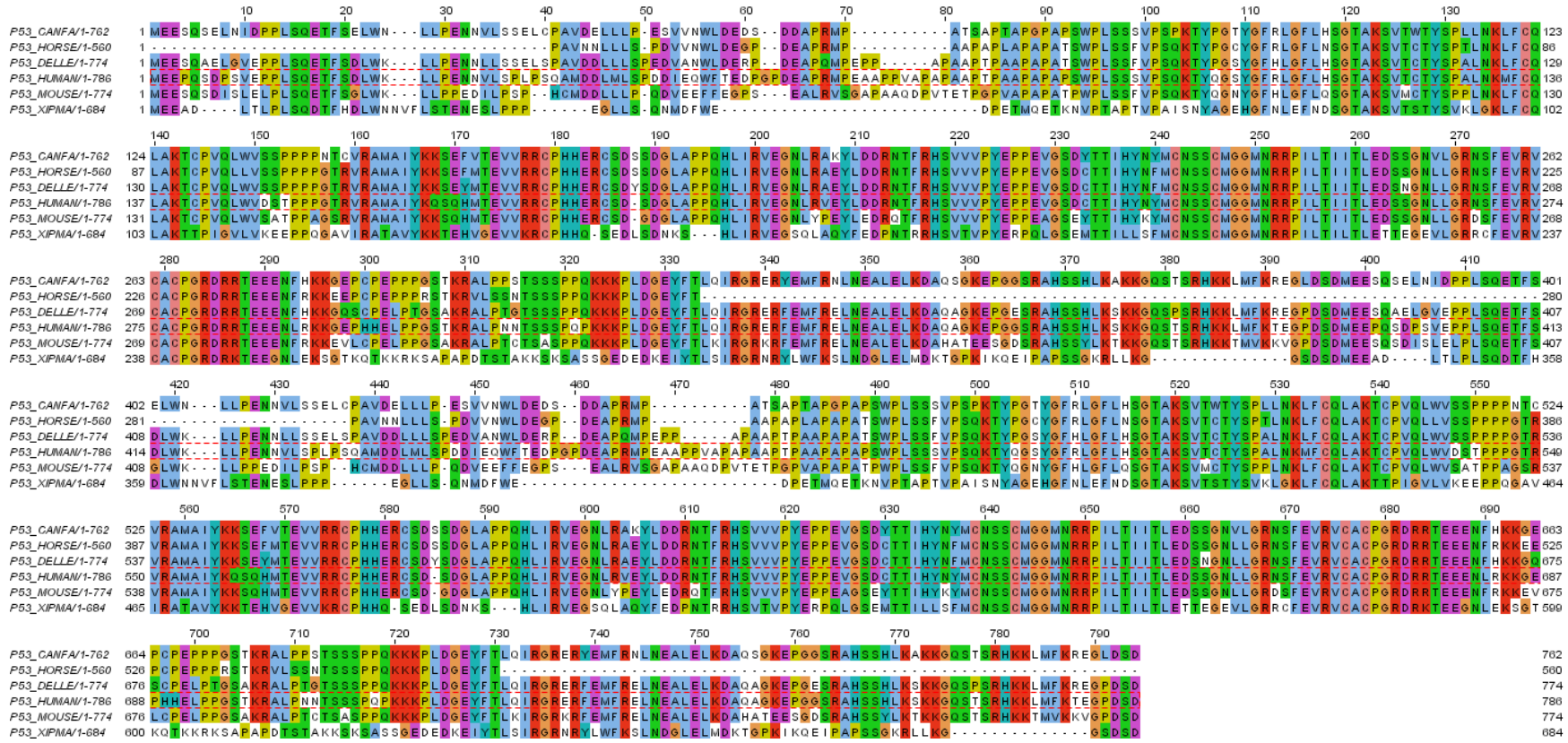




# Definition

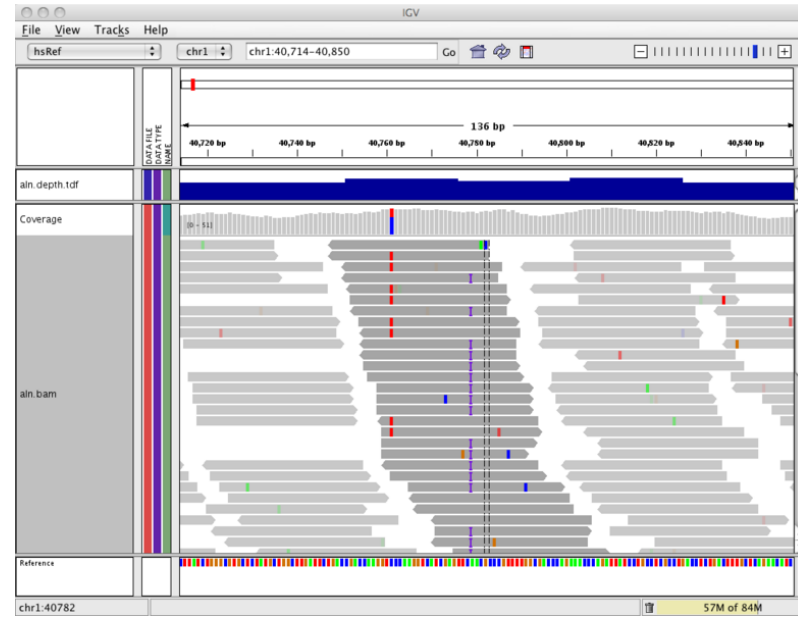
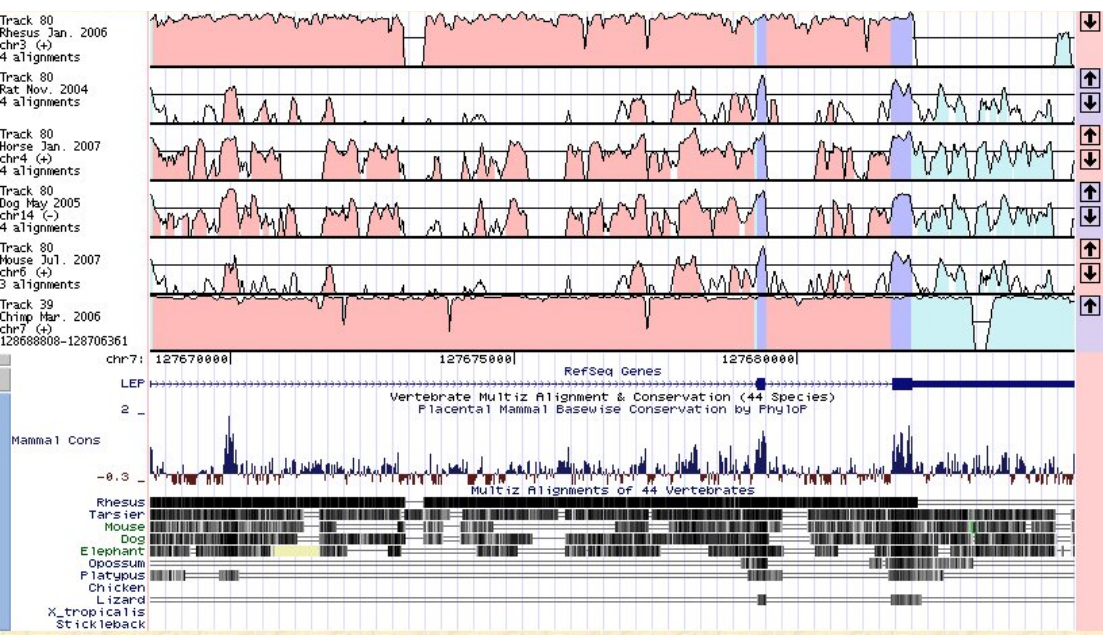
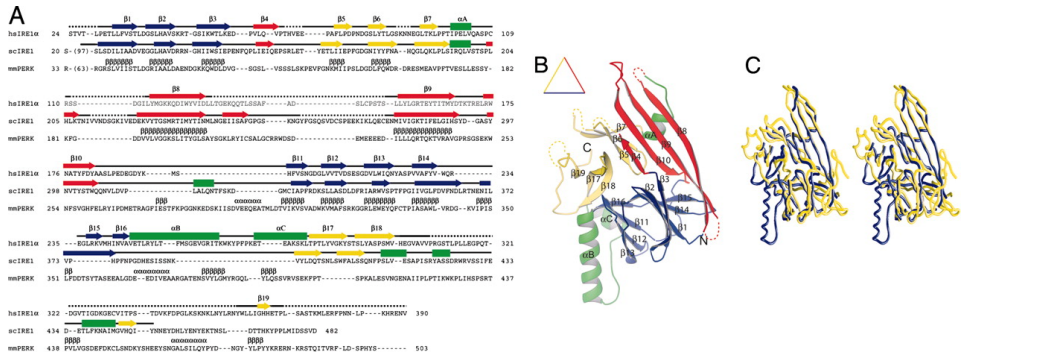
Given N sequences  $x^1, x^2, \dots, x^N$ :

- Insert gaps (-) in each sequence  $x^i$ , such that
  - All sequences have the same length L
  - Score of the global map is maximum

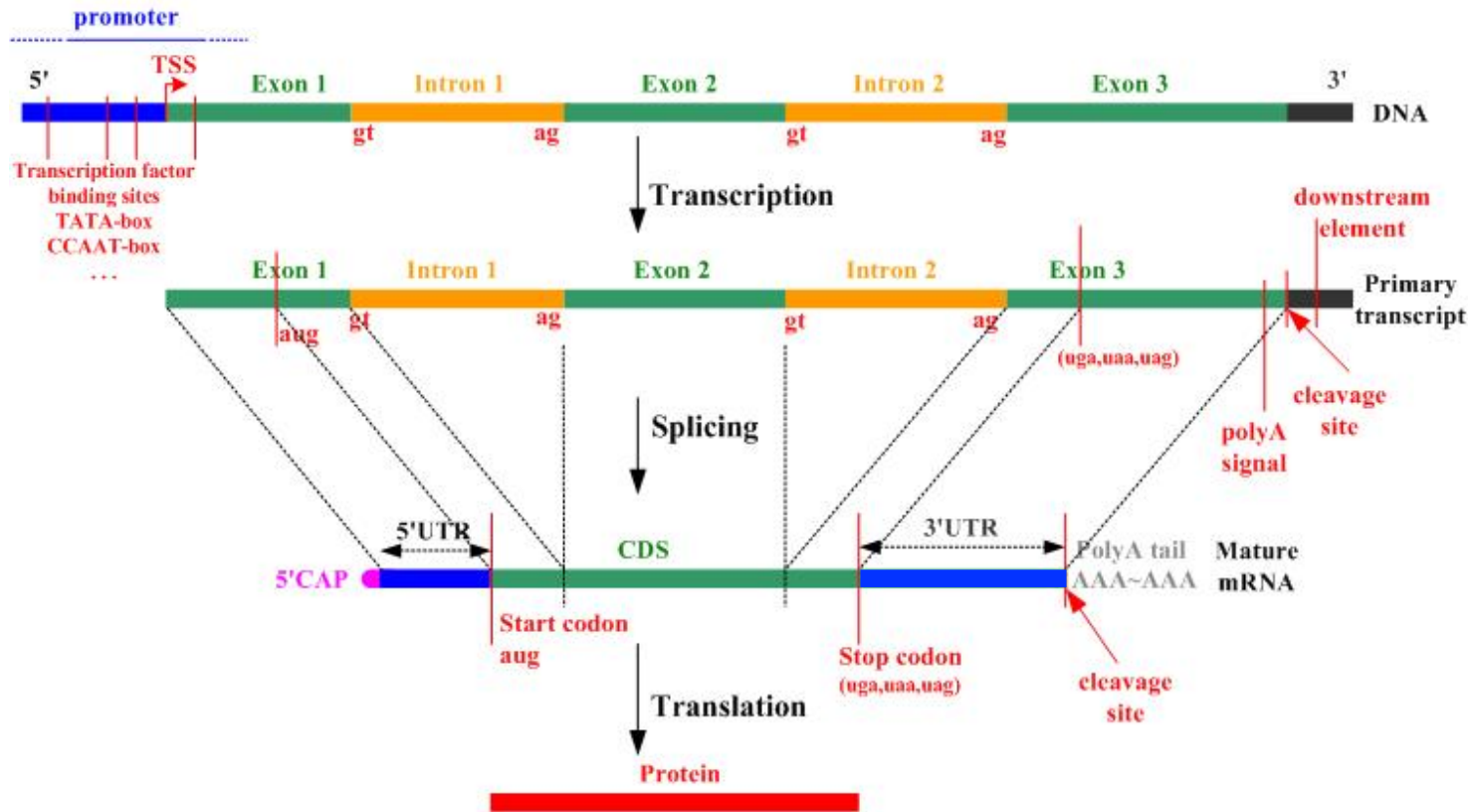




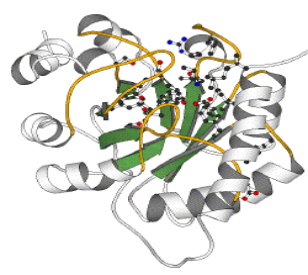
# Applications



# Gene structure



exon = protein-coding  
intron = non-coding



**Codon:**  
A triplet of nucleotides that is converted to one amino acid



# Gene Finding

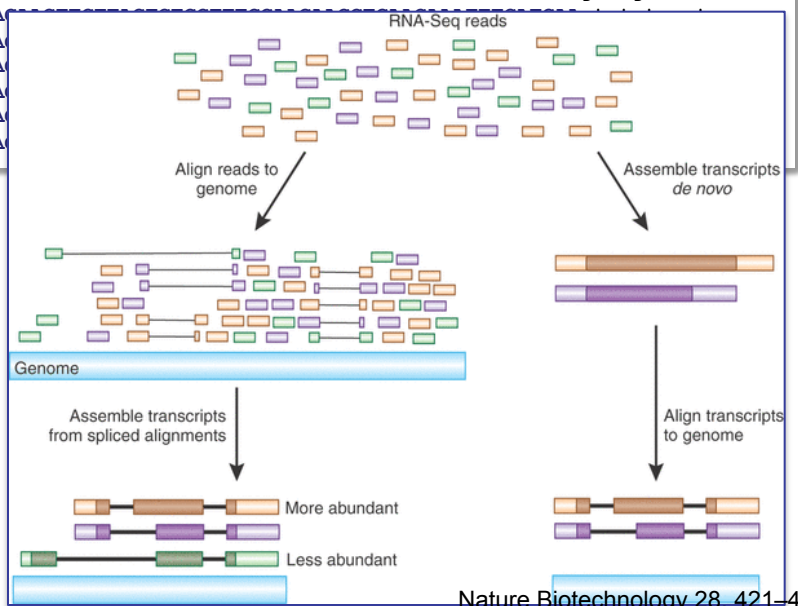


- Classes of Gene predictors

- Ab initio*: Only look at the genomic DNA of target genome
  - De novo*: Target genome + aligned informant genome(s)

Human	t t t c t t a g A C T T T A A A G C T G T C A A G C C G T G T T C T A G A T A A A A T A A G T A T T G G A C A A C T T G T T A G T C T C T T T C C A A C A A C C T G A A C A A A T T T G A T G A A g t a t g t a c c t a
Macaque	t t t c t t a g A C T T T A A A G C T G T C A A G C C G T G T T C T A G A T A A A A T A A G T A T T G G A C A A C T T G T T A G T C T C T T T C C A A C A A C C T G A A C A A A T T T G A T G A A g t a t g t a c c t a
Mouse	t t g c t t a g A C T T T A A A G T T G T C A A G C C G G T T C T T G A T A A A A T A A G T A T T G G A C A A C T T G T T A G T C T T C T T T C C A A C A A C C T G A A C A A A T T T G A T G A A g t a t g t a - c c a
Rat	t t g c t t a g A C T T T A A A G T T G T C A A G C C G T G T T C T T G A T A A A A T A A G T A T T G G A C A A C T T A T T A G T C T T C T T T C C A A C A A C C T G A A C A A A T T T G A T G A A g t a t g t a c c c a
Rabbit	t - - a t t a g A C T T T A A A G C T G T C A A G C C G T G T T C T A G A T A A A A T A A G T A T T G G C A A C T T A T T A G T C T C T T T C C A A C A A C C T G A A C A A A T T T G A T G A A g t a t g t a c c t a
Dog	t - c a t t a g A C T T T A A A G C T G T C A A G C C G T G T T C T G G A T A A A A T A A G T A T T G G A C A A C T T G T T A G T C T C T T T C C A A C A A C C T G A A C A A A T T C G A T G A A g t a t g t a c c t a
Cow	t - c a t t a g A C T T T G A A G C T A T C A A G C C G T G T T C T G G A T A A A A T A A G T A T T G G A
Armadillo	g c a - - t a g A C C T T A A A A C T G T C A A G C C G T G T T T T A G A T A A A A T A A G T A T T G G A
Elephant	g c t - t t a g A C T T T A A A A C T G T C C A G C C G T G T T C T T G A T A A A A T A A G T A T T G G A
Tenrec	t c - c t t a g A C T T T A A A A C T T T C G A G C C G G G T T C T A G A T A A A A T A A G T A T T G G A
Opossum	- - - t t t a g A C C T T A A A A C T G T C A A G C C G T G T T C T A G A T A A A A T A A G C A C T G G A
Chicken	- - - t t a g A C C T T A A A A C T G T C A A G C A A A G T T C T A G A T A A A A T A A G T A C T G G A

- RNA-seq based approaches





# Using Comparative Information

Alignment 1  
Seq1: human  
Seq2: macaque  
Reg id: 75  
Reg length: 100  
Plot min: 50  
Regions: 7

Alignment 2  
Seq1: human  
Seq2: pig  
Reg id: 75  
Reg length: 100  
Plot min: 50  
Regions: 6

Alignment 3  
Seq1: human  
Seq2: rabbit  
Reg id: 75  
Reg length: 100  
Plot min: 50  
Regions: 4

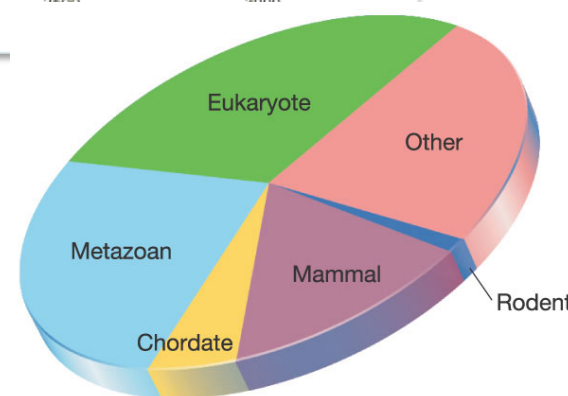
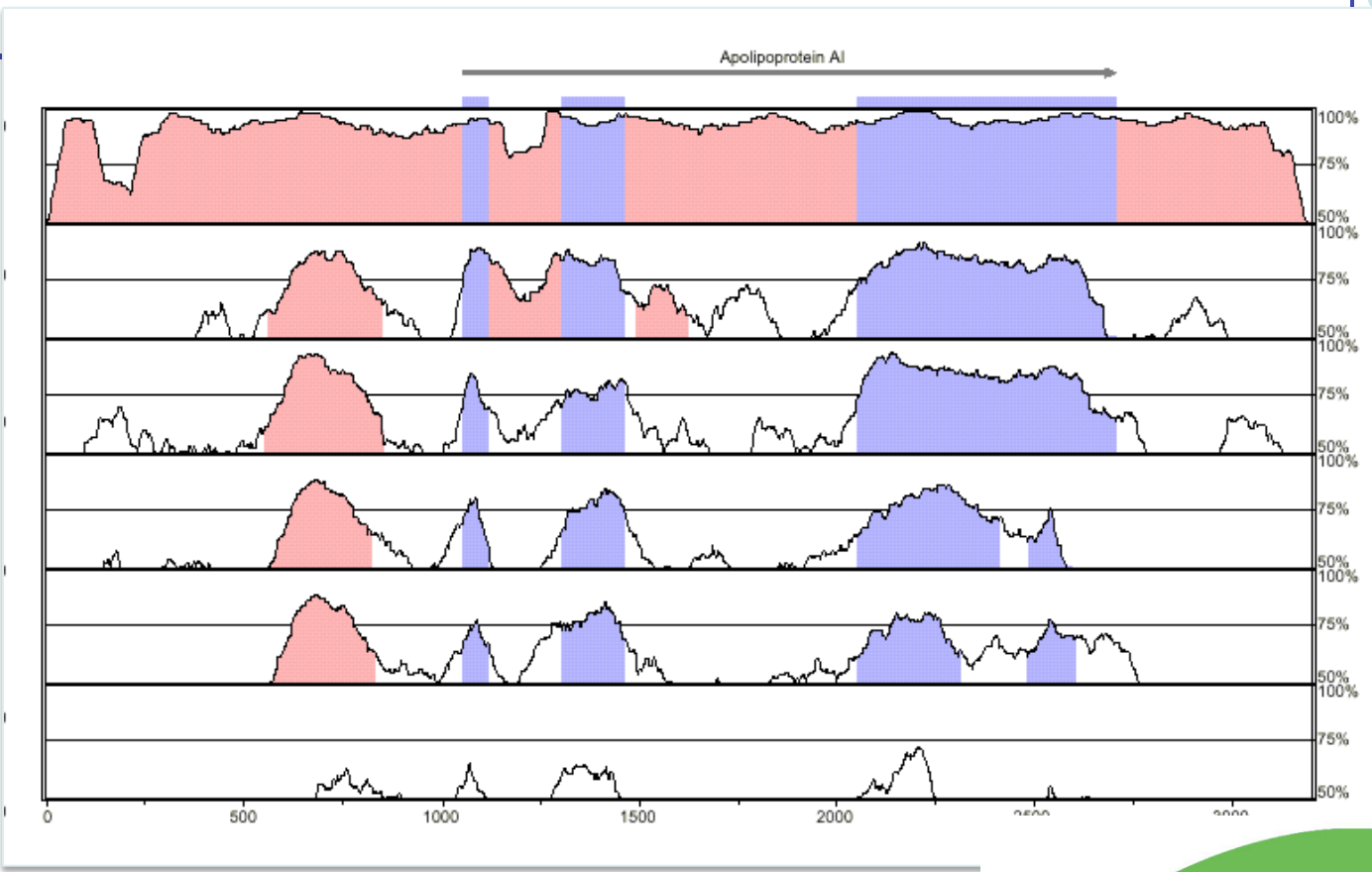
Alignment 4  
Seq1: human  
Seq2: mouse  
Reg id: 75  
Reg length: 100  
Plot min: 50  
Regions: 5

Alignment 5  
Seq1: human  
Seq2: rat  
Reg id: 75  
Reg length: 100  
Plot min: 50  
Regions: 5

Alignment 6  
Seq1: human  
Seq2: chicken  
Reg id: 75  
Reg length: 100  
Plot min: 50  
Regions: 0

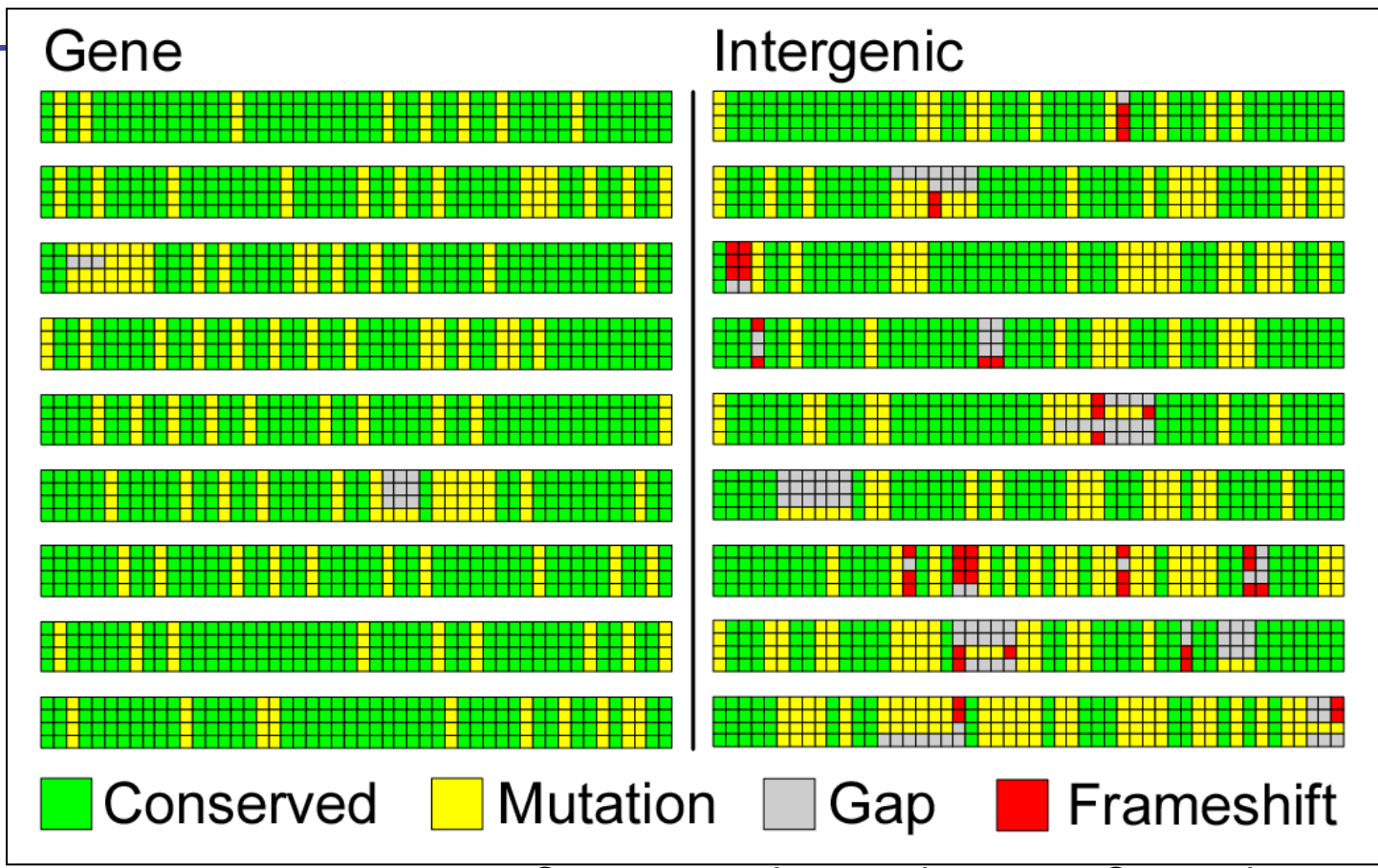
Resolution: 4  
Window size: 100  
Start: 1

■ Exon  
■ UTR  
■ CNS





# Patterns of Conservation



	Genes	Intergenic	Separation
<span style="color: yellow;">■</span> Mutations	30%	58%	→ 2-fold
<span style="color: grey;">■</span> Gaps	1.3%	14%	→ 10-fold
<span style="color: red;">■</span> Frameshifts	0.14%	10.2%	→ 75-fold



# Scoring Function: Sum Of Pairs

**Definition:** Induced pairwise alignment

A pairwise alignment induced by the multiple alignment

Example:

**x:** AC-GCGG-C  
**y:** AC-GC-GAG  
**z:** GCCGC-GAG

Induces:

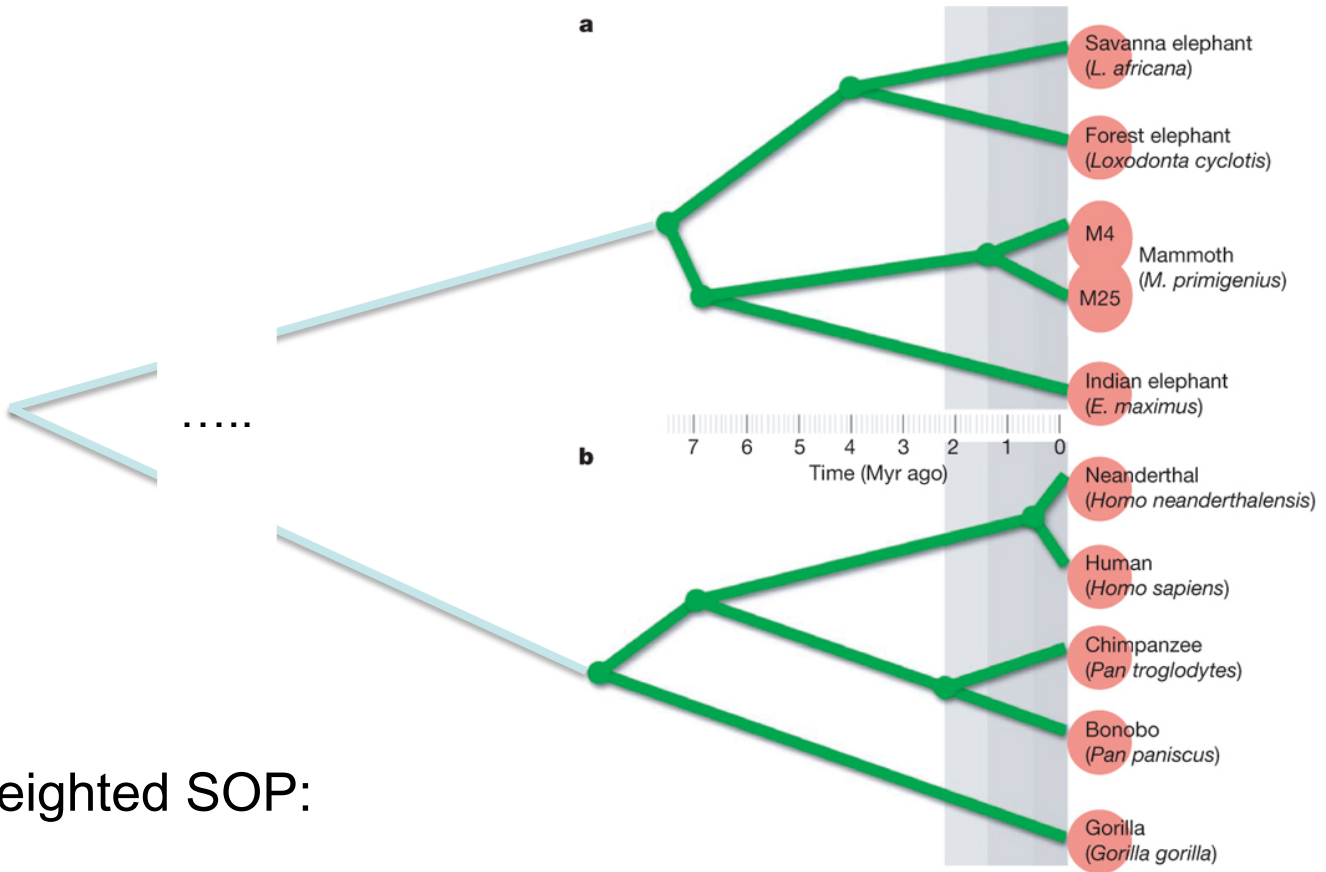
**x:** ACGCGG-C;    **x:** AC-GCGG-C;    **y:** AC-GCGAG  
**y:** ACGC-GAC;    **z:** GCCGC-GAG;    **z:** GCCGCGAG





# Sum Of Pairs (cont'd)

- Heuristic way to incorporate evolution tree:



- Weighted SOP:

$$S(m) = \sum_{k < l} w_{kl} s(m^k, m^l)$$



# A Profile Representation

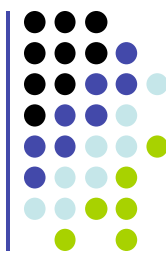
	-	A	G	G	C	T	A	T	C	A	C	C	T	G
T	A	G	-	C	T	A	C	C	A	-	-	-	-	G
C	A	G	-	C	T	A	C	C	A	-	-	-	-	G
C	A	G	-	C	T	A	T	C	A	C	-	-	G	G
C	A	G	-	C	T	A	T	C	G	C	-	-	G	G
A	0	1	0	0	0	0	1	0	0	.8	0	0	0	0
C	.6	0	0	0	1	0	0	.4	1	0	.6	.2	0	0
G	0	0	1	.2	0	0	0	0	0	.2	0	0	.4	1
T	.2	0	0	0	0	1	0	.6	0	0	0	0	.2	0
-	.2	0	0	.8	0	0	0	0	0	0	.4	.8	.4	0

- Given a multiple alignment  $M = m_1 \dots m_n$ 
  - Replace each column  $m_i$  with profile entry  $p_i$ 
    - Frequency of each letter in  $\Sigma$
    - # gaps
    - Optional: # gap openings, extensions, closings
  - Can think of this as a “likelihood” of each letter in each position



# Multiple Sequence Alignments

## Algorithms



# Multidimensional DP

Generalization of Needleman-Wunsh:

$$S(m) = \sum_i S(m_i)$$

(sum of column scores)

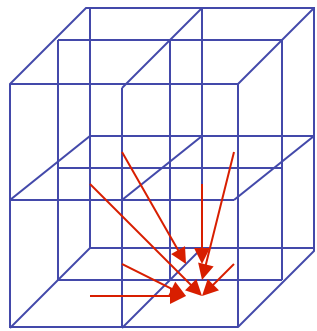
$F(i_1, i_2, \dots, i_N)$ : Optimal alignment up to  $(i_1, \dots, i_N)$

$F(i_1, i_2, \dots, i_N) = \max_{(\text{all neighbors of cube})} (F(\text{nbr}) + S(\text{nbr}))$



# Multidimensional DP

- Example: in 3D (three sequences):
- 7 neighbors/cell



$$F(i,j,k) = \max\{ F(i-1, j-1, k-1) + S(x_i, x_j, x_k), \\ F(i-1, j-1, k) + S(x_i, x_j, -), \\ F(i-1, j, k-1) + S(x_i, -, x_k), \\ F(i-1, j, k) + S(x_i, -, -), \\ F(i, j-1, k-1) + S(-, x_j, x_k), \\ F(i, j-1, k) + S(-, x_j, -), \\ F(i, j, k-1) + S(-, -, x_k) \}$$



# Multidimensional DP

Running Time:

1. Size of matrix:  $L^N$ ;

Where  $L$  = length of each sequence

$N$  = number of sequences

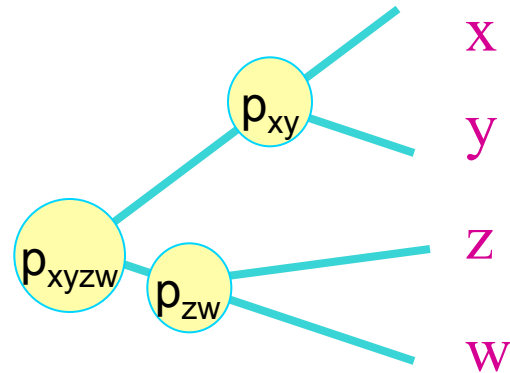
2. Neighbors/cell:  $2^N - 1$

Therefore.....  $O(2^N L^N)$





# Progressive Alignment



- When evolutionary tree is known:
  - Align closest first, in the order of the tree
  - In each step, align two sequences  $x$ ,  $y$ , or profiles  $p_x$ ,  $p_y$ , to generate a new alignment with associated profile  $p_{\text{result}}$

Weighted version:

- Tree edges have weights, proportional to the divergence in that edge
- New profile is a weighted average of two old profiles



# Progressive Alignment

X

## Example

Profile: (A, C, G, T, -)

$$p_x = (0.8, 0.2, 0, 0, 0)$$

$$p_y = (0.6, 0, 0, 0, 0.4)$$

- When evolutionary tree is known:
  - Align closest first, in the order of the tree
  - In each step, align two sequence alignment with associated profile

$$s(p_x, p_y) = 0.8*0.6*s(A, A) + 0.2*0.6*s(C, A) + 0.8*0.4*s(A, -) + 0.2*0.4*s(C, -)$$

Result:  $p_{xy} = (0.7, 0.1, 0, 0, 0.2)$

Weighted version:

- Tree edges have weights, proportional to the divergence in that edge
- New profile is a weighted average of two old profiles

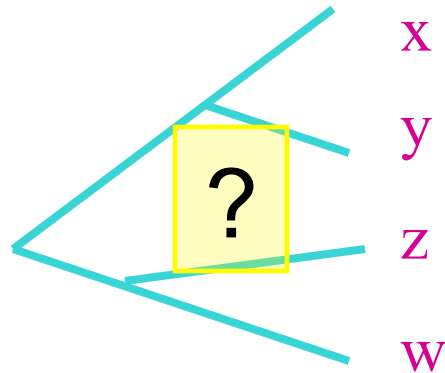
$$s(p_x, -) = 0.8*1.0*s(A, -) + 0.2*1.0*s(C, -)$$

Result:  $p_{x-} = (0.4, 0.1, 0, 0, 0.5)$

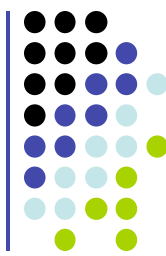




# Progressive Alignment



- When evolutionary tree is unknown:
  - Perform all pairwise alignments
  - Define distance matrix  $D$ , where  $D(x, y)$  is a measure of evolutionary distance, based on pairwise alignment
  - Construct a tree (*UPGMA / Neighbor Joining / Other methods*)
  - Align on the tree



# Heuristics to improve alignments

- Iterative refinement schemes
- A\*-based search
- Consistency
- Simulated Annealing
- ...



# Iterative Refinement

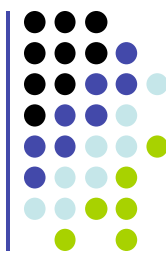
One problem of progressive alignment:

- Initial alignments are “frozen” even when new evidence comes

## Example:

**x:**      **GAAGTT**  
**y:**      **GAC-TT**      **Frozen!**

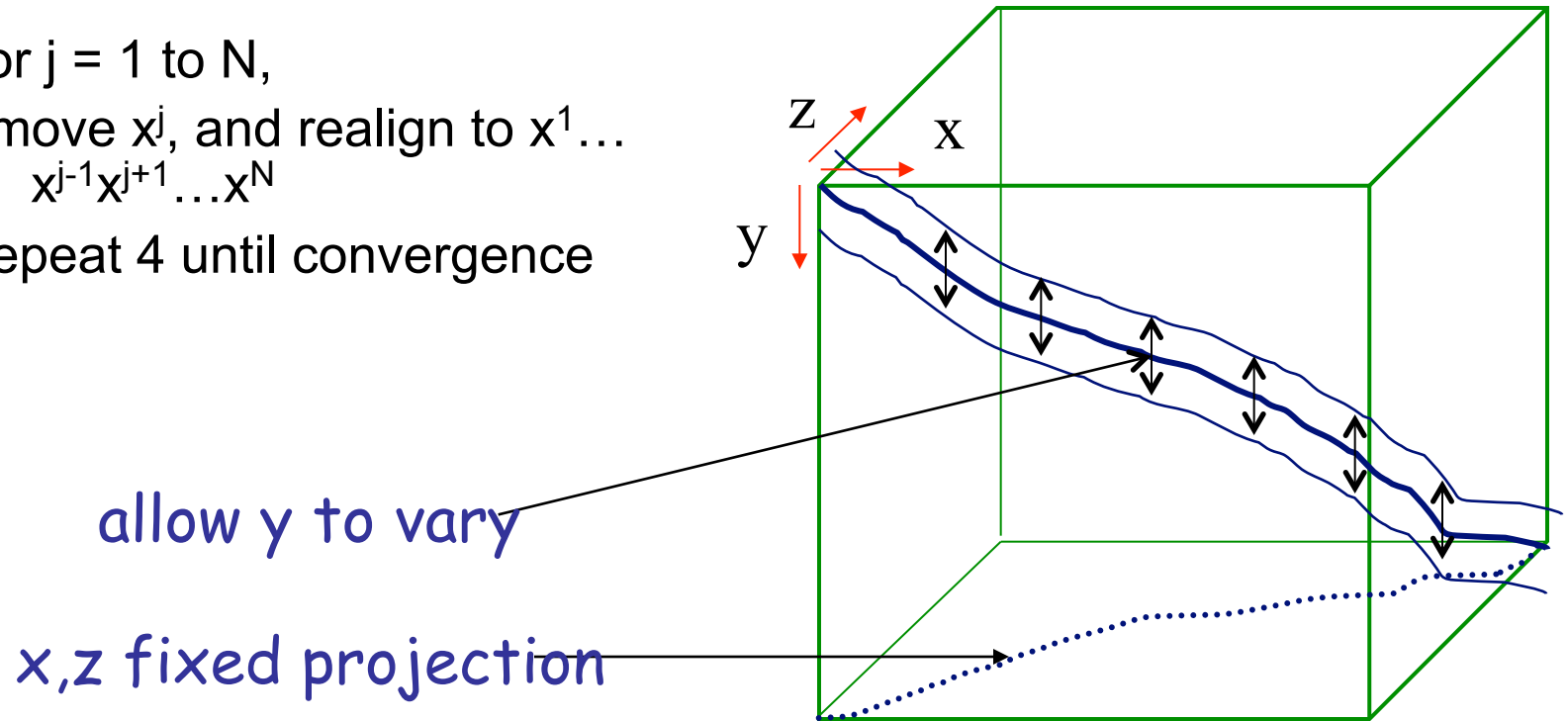
**z:**      **GAACTG**  
**w:**      **GTACTG**      **Now clear correct y = GA-CTT**



# Iterative Refinement

## Algorithm (Barton-Stenberg):

1. For  $j = 1$  to  $N$ ,  
Remove  $x^j$ , and realign to  $x^1 \dots$   
 $x^{j-1} x^{j+1} \dots x^N$
2. Repeat 4 until convergence





# Iterative Refinement

Example: align (x,y), (z,w), (xy, zw):

```
x:      GAAGTTA  
y:      GAC-TTA  
z:      GAACTGA  
w:      GTACTGA
```

After realigning y:

```
x:      GAAGTTA  
y:      G-ACTTA  
z:      GAACTGA  
w:      GTACTGA
```

**+ 3 matches**



# Iterative Refinement

Example not handled well:

**x :**      **GAAGTTA**

**y<sub>1</sub> :**    **GAC-TTA**

**y<sub>2</sub> :**    **GAC-TTA**

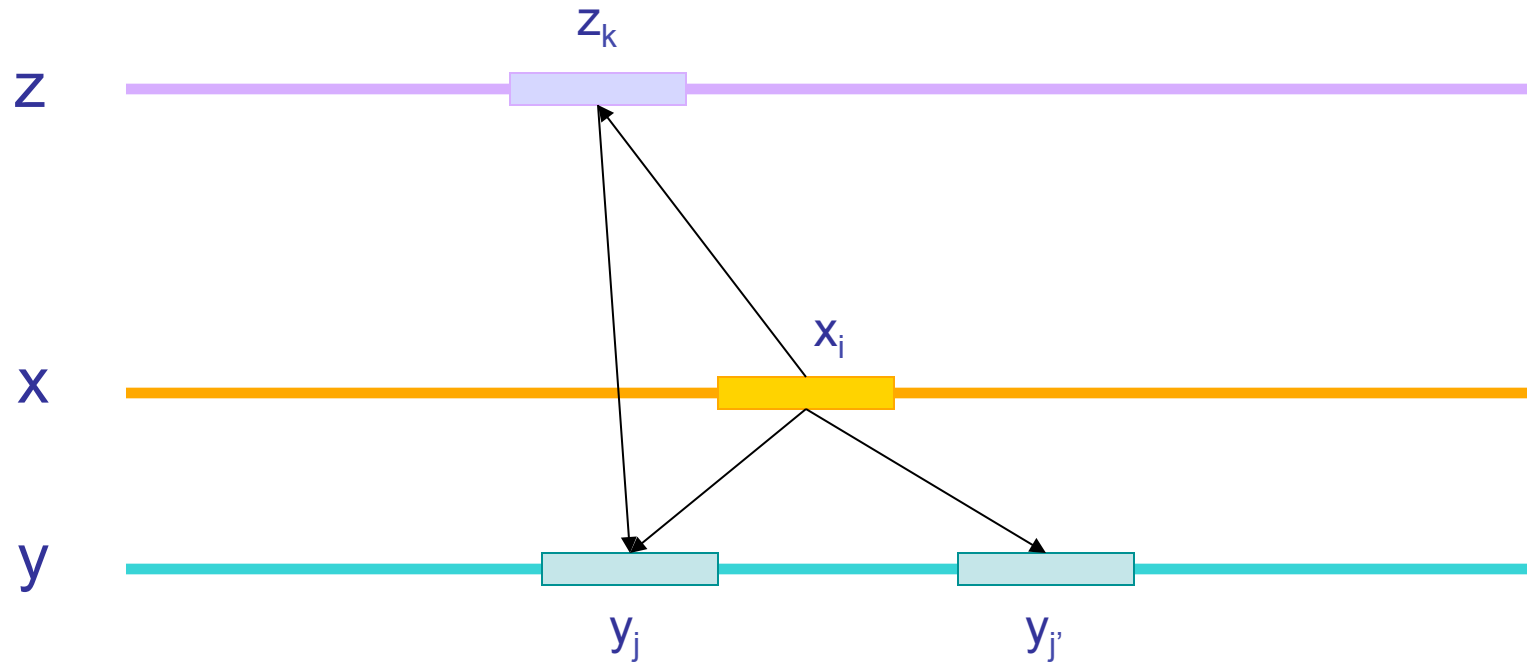
**y<sub>3</sub> :**    **GAC-TTA**

**z :**      **GAACTGA**

**w :**      **GTACTGA**

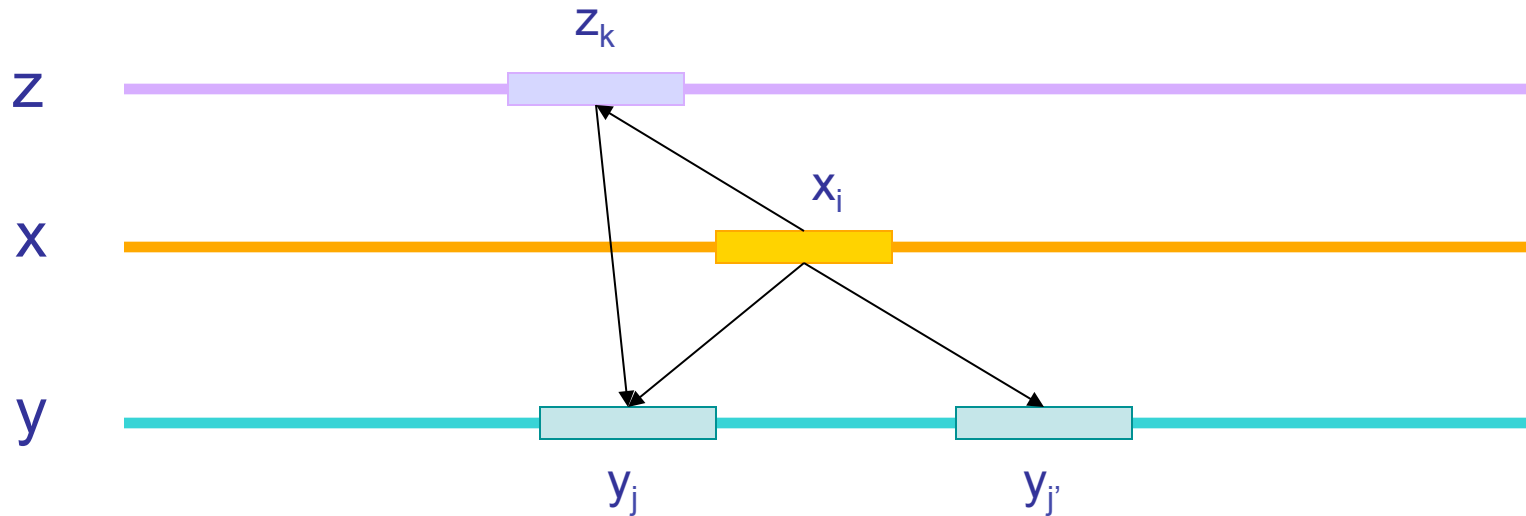
Realigning any single  $y_i$   
changes nothing

# Consistency





# Consistency



Basic method for applying consistency

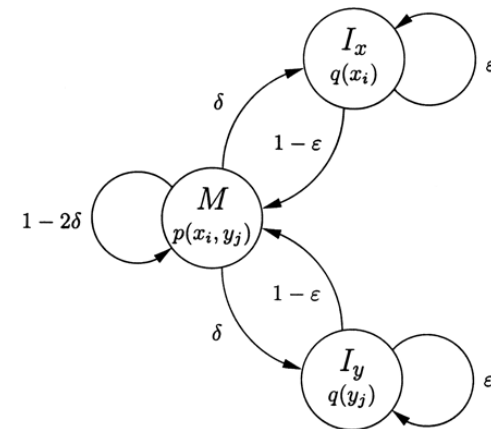
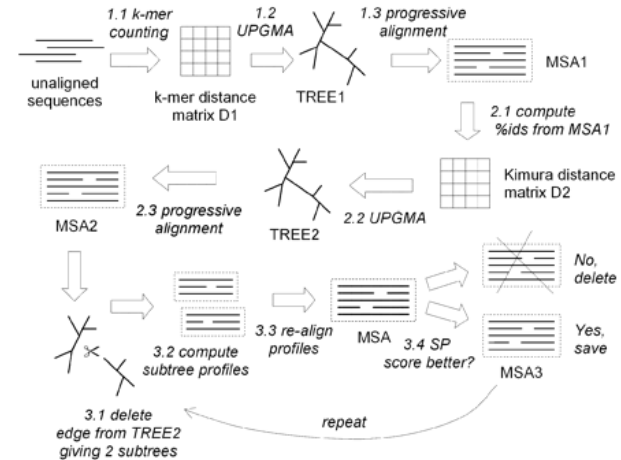
- Compute all pairs of alignments  $xy$ ,  $xz$ ,  $yz$ , ...
- When aligning  $x$ ,  $y$  during progressive alignment,
  - For each  $(x_i, y_j)$ , let  $s(x_i, y_j) = \text{function\_of}(x_i, y_j, a_{xz}, a_{yz})$
  - Align  $x$  and  $y$  with DP using the modified  $s(.,.)$  function





# Real-world protein aligners

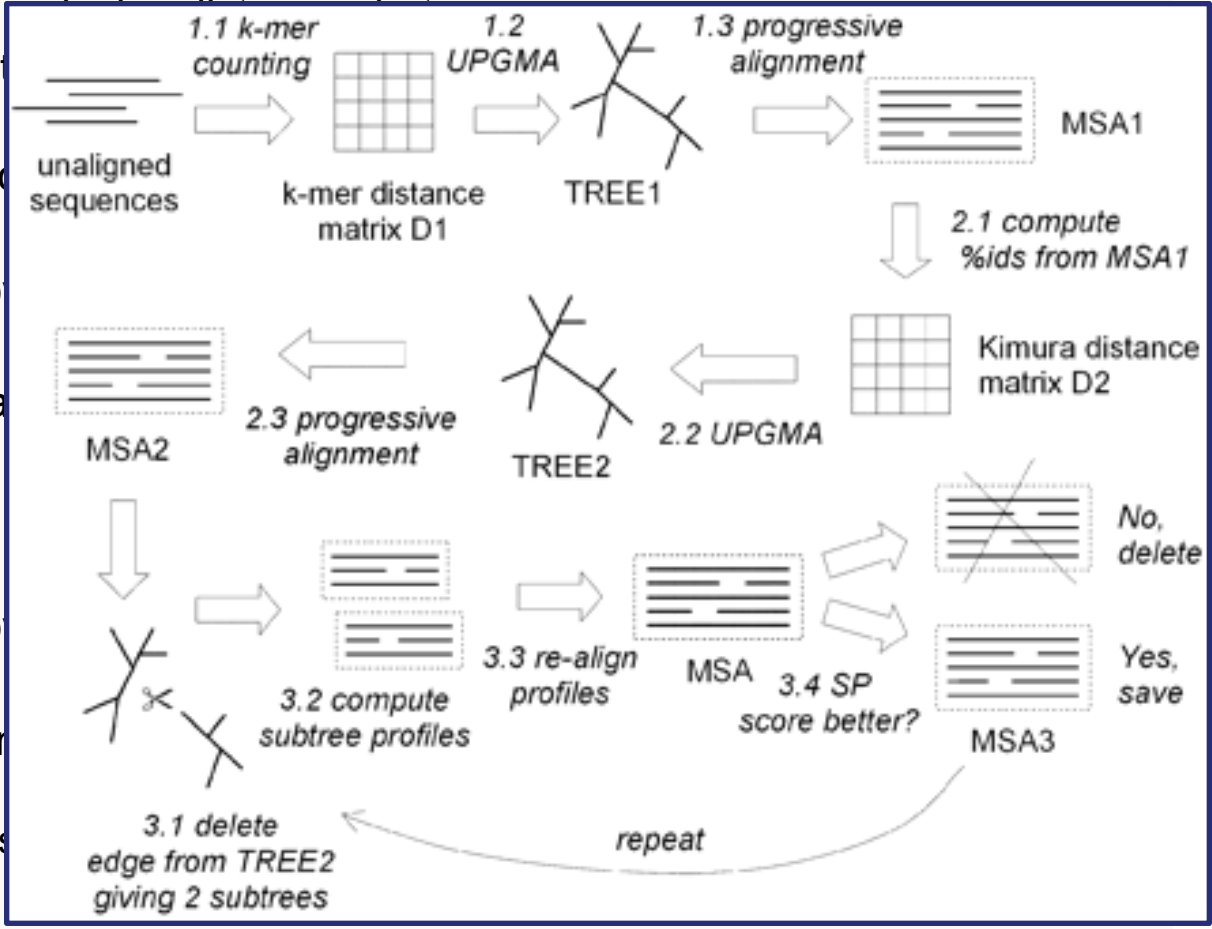
- MUSCLE
  - High throughput
  - One of the best in accuracy
- ProbCons
  - High accuracy
  - Reasonable speed





# MUSCLE at a glance

1. Fast measurement of all pairwise distances
  - $D_{\text{DRAFT}}(x, y)$  defined in terms of the number of mismatches between  $x$  and  $y$
2. Build tree  $T_{\text{DRAFT}}$  based on  $D_{\text{DRAFT}}$
3. Progressive alignment of sequences according to  $T_{\text{DRAFT}}$
4. Measure new Kimura-based distance  $D_{\text{KIMURA}}$  between sequences
5. Build tree  $T$  based on  $D_{\text{KIMURA}}$
6. Progressive alignment of sequences according to  $T$
7. Iterative refinement; for  $n$  iterations
  - *Tree Partitioning*: Split  $T$  into two subtrees
  - If new alignment  $M'$  has a better SP score than  $M$ , then





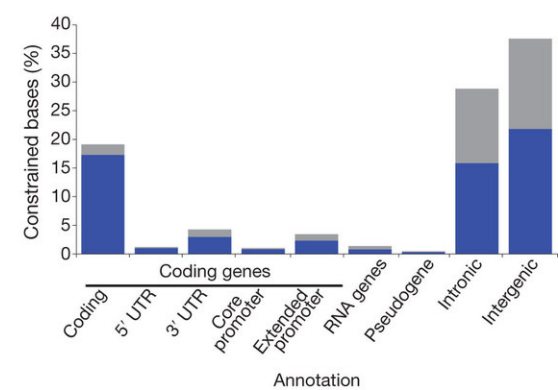
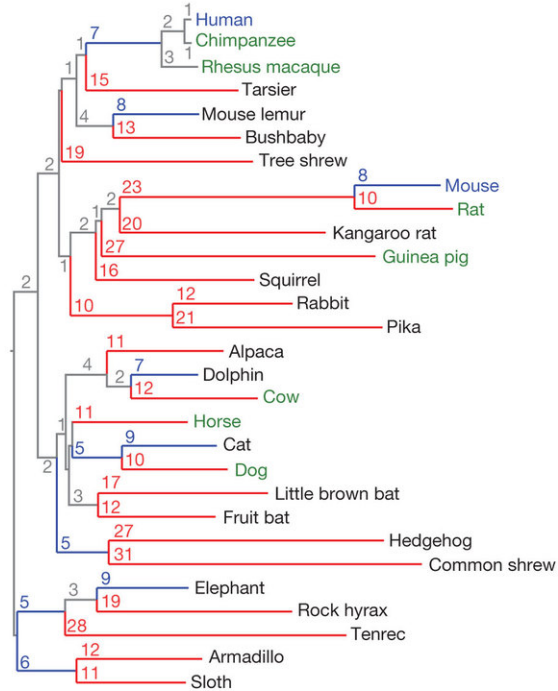
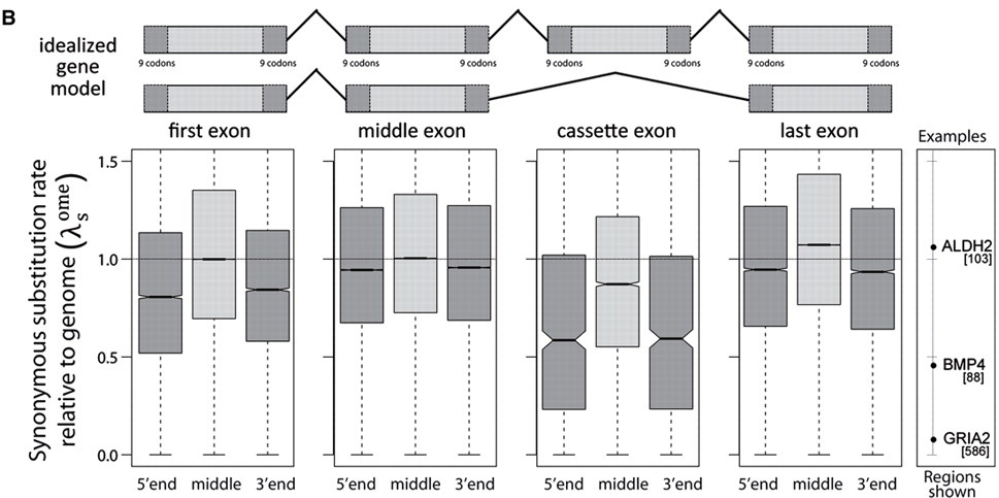
# PROBCONS at a glance

1. Computation of all posterior matrices  $M_{xy}$  :  $M_{xy}(i, j) = \text{Prob}(x_i \sim y_j)$ , using a HMM
2. Re-estimation of posterior matrices  $M'_{xy}$  with **probabilistic consistency**
  - $M'_{xy}(i, j) = 1/N \sum_{\text{sequence } z} \sum_k M_{xz}(i, k) \times M_{yz}(j, k); \quad M'_{xy} = \text{Avg}_z(M_{xz} M_{zy})$
3. Compute for every pair  $x, y$ , the maximum expected accuracy alignment
  - $A_{xy}$ : alignment that maximizes  $\sum_{\text{aligned } (i, j) \text{ in } A} M'_{xy}(i, j)$
  - Define  $E(x, y) = \sum_{\text{aligned } (i, j) \text{ in } A_{xy}} M'_{xy}(i, j)$
4. Build tree  $T$  with hierarchical clustering using similarity measure  $E(x, y)$
5. Progressive alignment on  $T$  to maximize  $E(.,.)$
6. Iterative refinement; for many rounds, do:
  - **Randomized Partitioning**: Split sequences in  $M$  in two subsets by flipping a coin for each sequence and realign the two resulting profiles

# Mammalian alignments

	<i>ALDH2</i> <sup>[103]</sup>	<i>BMP4</i> <sup>[88]</sup>	<i>GRIA2</i> <sup>[586]</sup>
Ancestor AA	L N R L A D L I E	L Q S G E E E E E	E S T N E F G I F
Ancestor DNA	CTG AAC CGC CTG GCT GAT CTG ATT GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Human	CTG AAC CGC CTG <b>GGC</b> GAT CTG <b>ATG</b> GAG	<b>GGT</b> CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Chimp	CTG AAC <b>GGG</b> CTG GCT GAT CTG <b>ATG</b> GAG	<b>GGT</b> CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Rhesus	CTG AAC <b>GGG</b> CTG GCT GAT CTG <b>ATG</b> GAG	<b>GGT</b> CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Tarsier	CTG AAC <b>GGG</b> CTG <b>GGC</b> GAT CTG ATT GAG	<b>GGT</b> CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Mouse lemur	CTG AAC CGC CTG <b>GGC</b> GAT CTG <b>ATG</b> GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Bushbaby	CTG AAC CGC CTG GCT GAT CTG ATT <b>GA</b>	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Treeshrew	CTG <b>TA</b> <b>CG</b> CTG GCT GAT CTG ATT <b>GA</b>	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Mouse	<b>ATG</b> <b>TAC</b> <b>GGA</b> <b>ATG</b> <b>GGG</b> GAT <b>CTG</b> ATT <b>GA</b>	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Rat	<b>ATG</b> <b>TAC</b> <b>GGA</b> <b>ATG</b> GCT GAT <b>CTC</b> <b>ATG</b> <b>GA</b>	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Kangaroo rat	CTG <b>TA</b> <b>CG</b> CTG <b>GGC</b> GAT CTG <b>ATG</b> GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Guinea pig	<b>ATA</b> AAC CGC CTG <b>GA</b> GAT <b>CTG</b> <b>ATG</b> GAG	<b>GGG</b> CAG TCT GGG <b>GA</b> GAG <b>GA</b> GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Squirrel	<b>ATG</b> <b>TAC</b> <b>GGG</b> CTG GCT GAT CTG <b>ATG</b> GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Rabbit	CTG <b>TA</b> <b>CG</b> CTG GCT <b>GA</b> CTG ATT GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Pika	CTG <b>TA</b> <b>CG</b> CTG GCT <b>GA</b> CTG ATT GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Alpaca	CTG AAC CGC CTG GCT GAT CTG ATT GAG	CTC CAG TCT GGG GAG GAG GAG GA ---A	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Dolphin	CTG AAC CGC CTG GCT GAT CTG ATT GAG	<b>GGT</b> CAG TCT GGG GAG GAG GAG GG ---G	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Cow	CTG AAC CGC CTG GCT GAT CTG ATT GAG	<b>GGT</b> CAG TCT GGG GAG GAG <b>GA</b> GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Horse	CTG AAC CGC <b>ATG</b> GCT GAT CTG ATT <b>GA</b>	<b>GGT</b> CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Cat	CTG AAC CGC CTG <b>GCT</b> GAT CTG ATT GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Dog	CTG AAC CGC CTG <b>GCA</b> GAT CTG ATT GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Microbat	CTG AAC CGC CTG <b>GGC</b> <b>GA</b> CTG ATT GAG	CTC CAG TCT GGG GAG GAG <b>GA</b> GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Megabat	CTG AAC <b>GGG</b> CTG GCT <b>GA</b> CTG <b>ATG</b> GAG	CTC CAG TCT GGG GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Hedgehog	CTG AAC CGC CTG GCT <b>GA</b> <b>CTT</b> ATT GAG	CTC CAG TCT GGG <b>GA</b> GAG GAG GAA <b>TTT</b>	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Shrew	-----	CTC CAG TCT <b>GGG</b> GAG GAG GAG GAA GAG	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Elephant	-----	CTC CAG TCT GGG GAG GAG GAG GAA ---	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Rock hyrax	-----	CTC CAG TCT GGG GAG GAG GAG GAA ---	G-A TCA <b>GGG</b> -AT GAA TTT GGG ATT TTT
Tenrec	CTG AAC <b>GA</b> <b>ATG</b> <b>GGC</b> GAT CTG ATT <b>GA</b>	CTC CAG TCT GGG GAG GAG GAG GAA ---	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Armadillo	CTG AAC CGC CTG <b>GGC</b> <b>GA</b> <b>ATG</b> GAG	CTC CAG TCT GGG GAG GAG GAG GAA ---	GAA TCA ACT AAT GAA TTT GGG ATT TTT
Sloth	CTG AAC <b>GA</b> CTG <b>GGC</b> <b>GG</b> <b>GG</b> ATT <b>GA</b>	<b>GGG</b> CAG TCT GGG GAG GAG GAG GAA ---	GAA TCA ACT AAT GAA TTT GGG ATT TTT

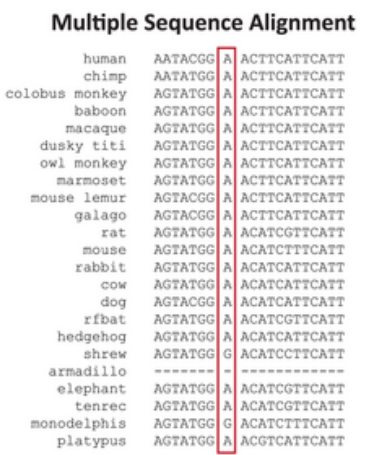
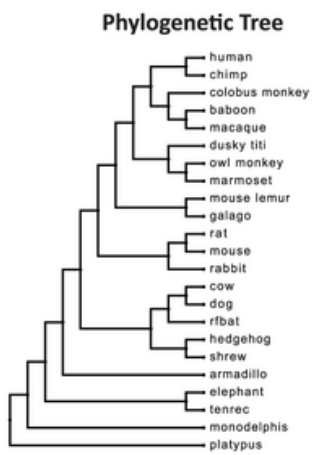
$\lambda_s^{ome} = 1.1$                        $\lambda_s^{ome} = 0.5$                        $\lambda_s^{ome} = 0.1$



**References**

- [Lindblad-Toh et al. Nature 478:476-482, 2011](#)
- [Lin et al. Genome Research 21:1916-1928, 2011](#)

# Genome Evolutionary Rate Profiling (GERP)



1. Compute position-specific RS scores

-1.4 2.7 1.9 1.3 3.6 -3.3 2.4 1.8 1.2 2.5 -1.7 1.3 2.4 3.9 2.5 -3.1

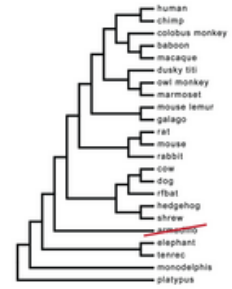
2. Generate candidate elements



3. Select final elements by p-value

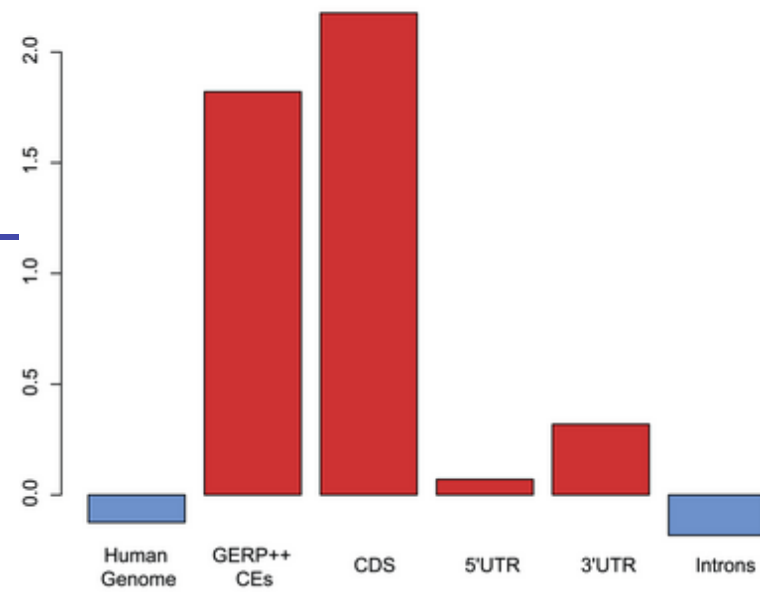


ML Tree Scaling Factor  $r = 0.7$

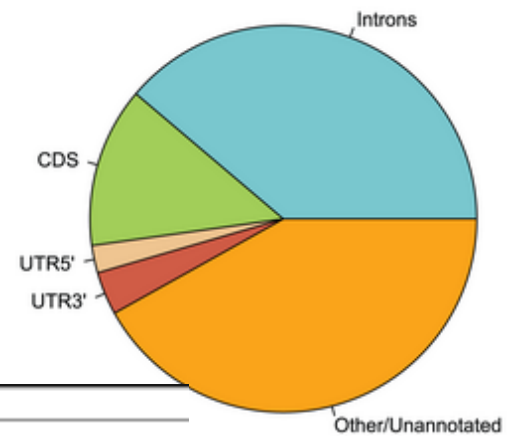


RS Score = 1.14  
(Neutral - Estimated)

A. Average Position Conservation Score



B. Composition of Constrained Elements



Annotation	% Coverage by CEs
Exons	84.6%
Introns	6.9%
UTR5'	23.7%
UTR3'	33.9%
ncRNA	10.1%