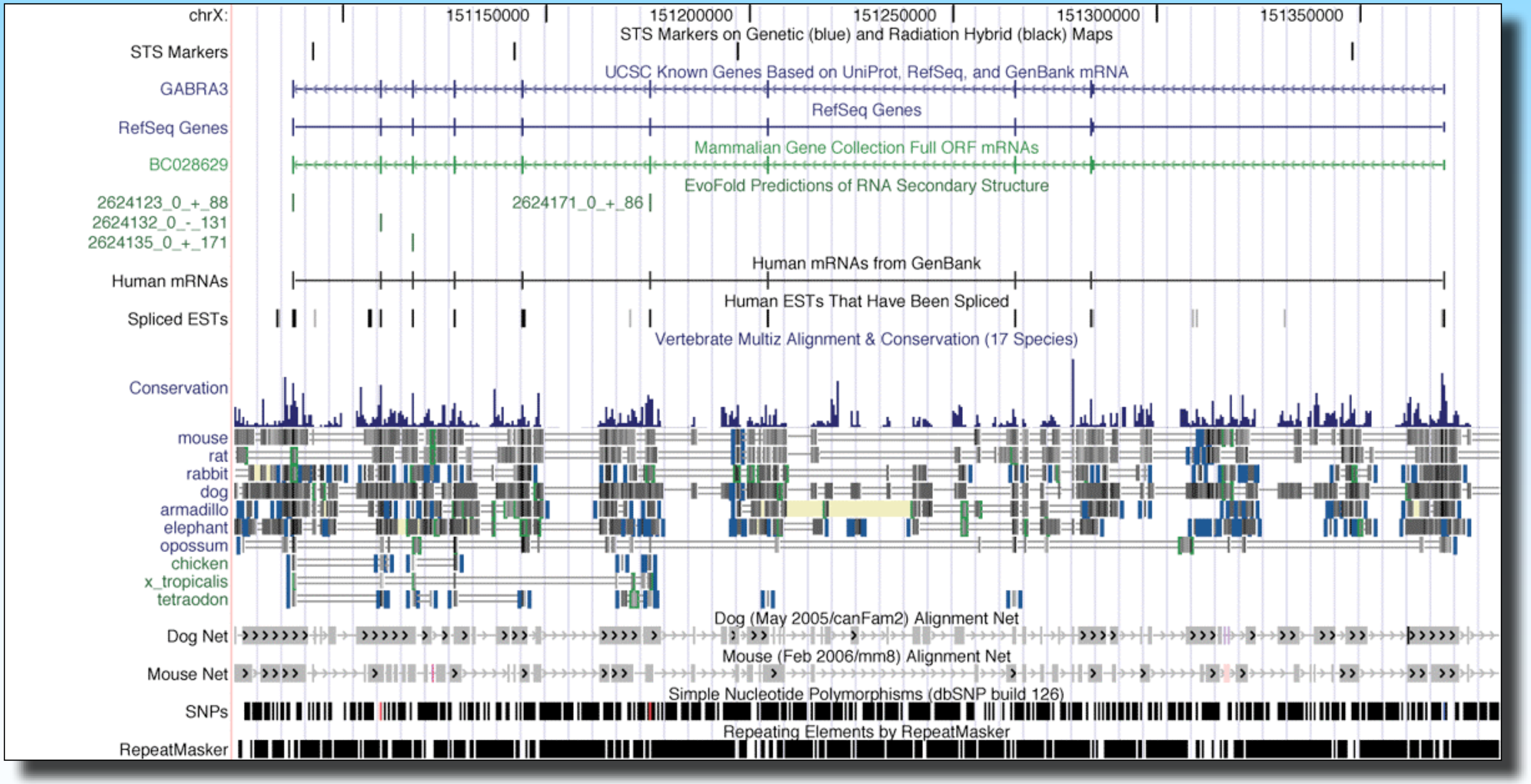
Whole Genome Analysis and Annotation

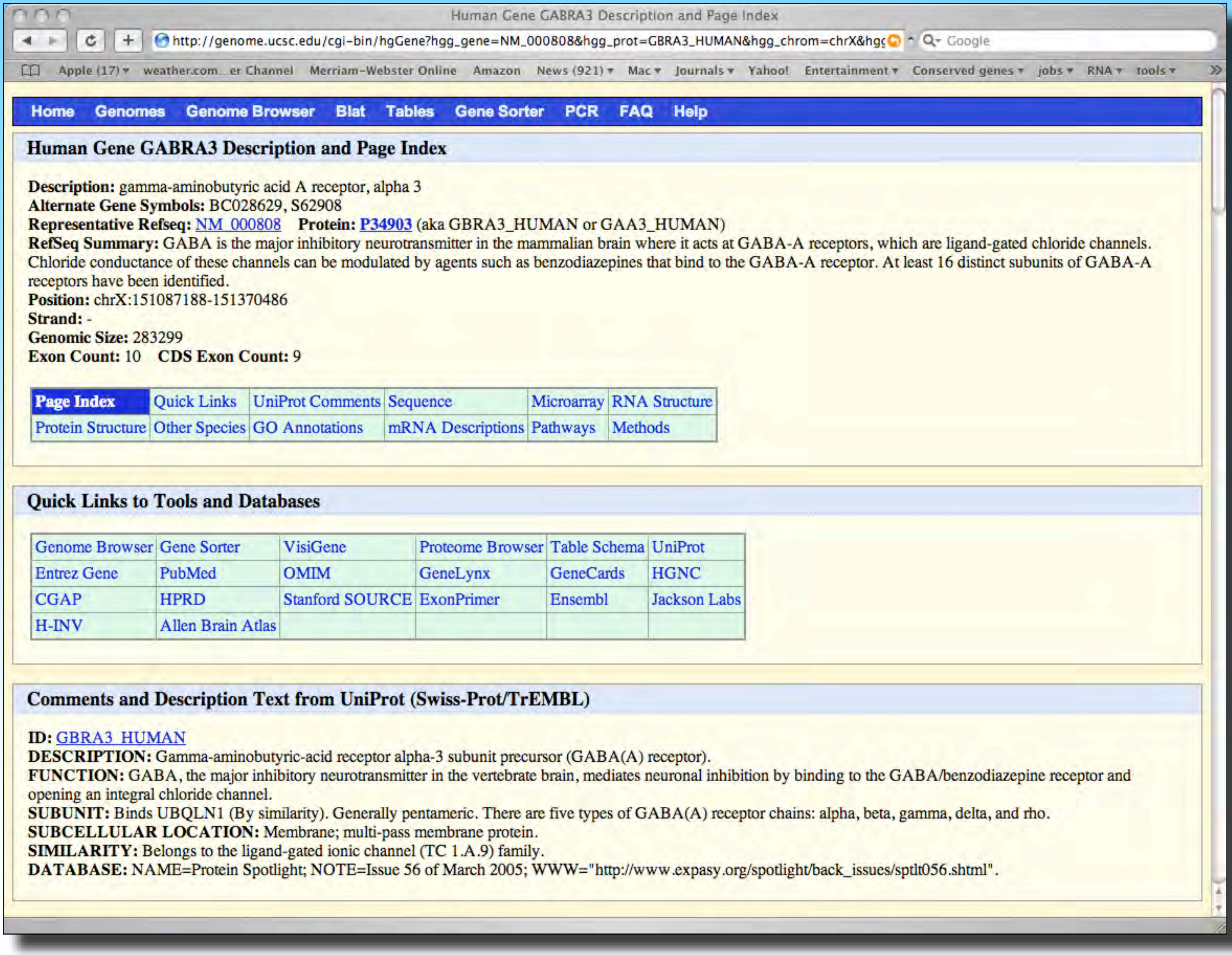
Adam Siepel

Biological Statistics & Computational Biology
Cornell University

The Challenge

Genome Browsers


chrX:
151150000
151200000
151250000
151300000
151350000
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
GABRA3
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC028629
EvoFold Predictions of RNA Secondary Structure
2624123_0_+_88
2624132_0_-_131
2624135_0_+_171
2624171_0_+_86
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
mouse
rat
rabbit
dog
armadillo
elephant
opossum
chicken
x_tropicalis
tetraodon
Dog (May 2005/canFam2) Alignment Net
Dog Net
Mouse (Feb 2006/mm8) Alignment Net
Mouse Net
Simple Nucleotide Polymorphisms (dbSNP build 126)
SNPs
Repeating Elements by RepeatMasker
RepeatMasker



Human Gene GABRA3 Description and Page Index

Description: gamma-aminobutyric acid A receptor, alpha 3
Alternate Gene Symbols: BC028629, S62908
Representative Refseq: NM_000808 **Protein:** P34903 (aka GBRA3_HUMAN or GAA3_HUMAN)
RefSeq Summary: GABA is the major inhibitory neurotransmitter in the mammalian brain where it acts at GABA-A receptors, which are ligand-gated chloride channels. Chloride conductance of these channels can be modulated by agents such as benzodiazepines that bind to the GABA-A receptor. At least 16 distinct subunits of GABA-A receptors have been identified.
Position: chrX:151087188-151370486
Strand: -
Genomic Size: 283299
Exon Count: 10 **CDS Exon Count:** 9

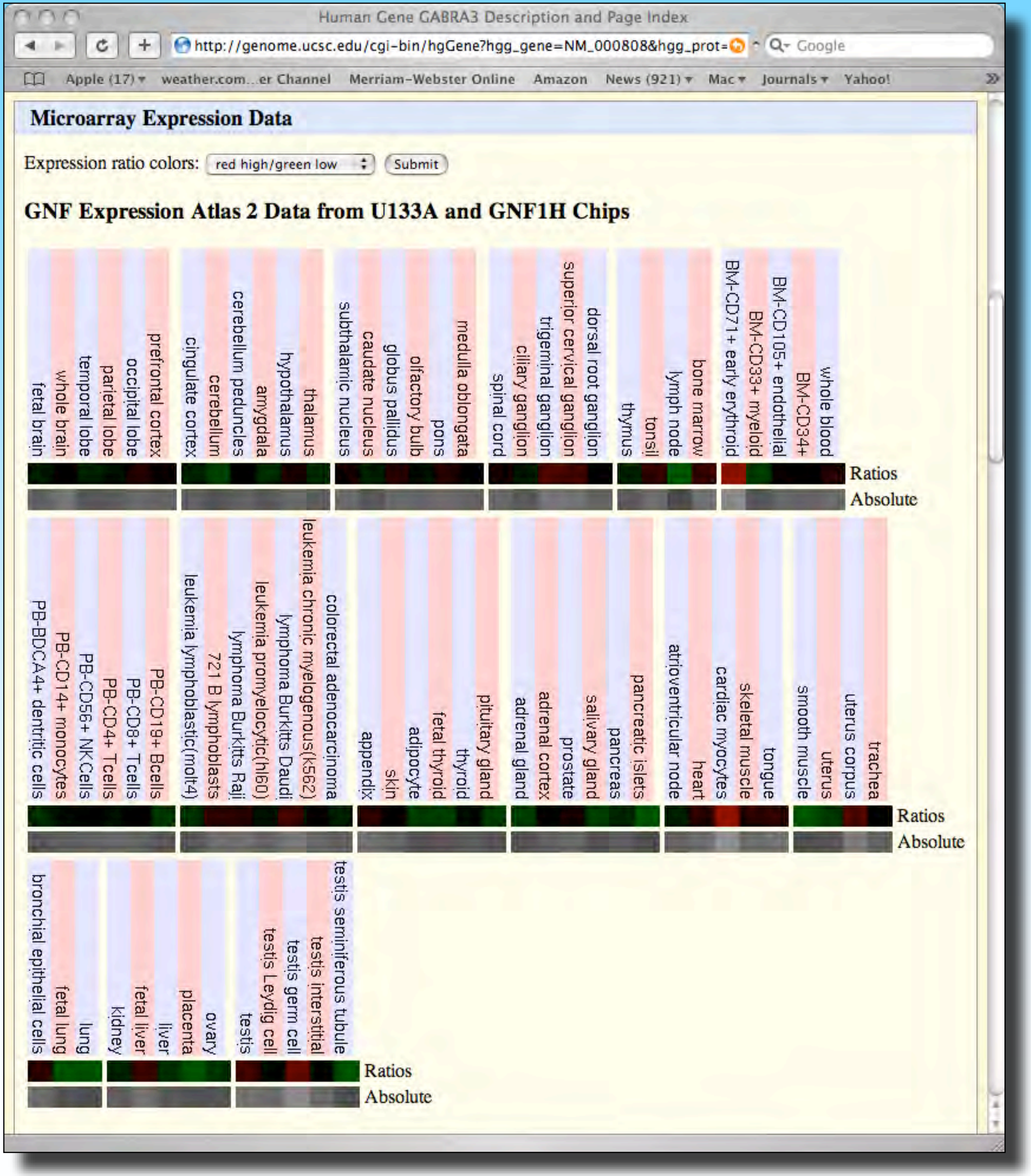
Page Index	Quick Links	UniProt Comments	Sequence	Microarray	RNA Structure
Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways	Methods

Quick Links to Tools and Databases

Genome Browser	Gene Sorter	VisiGene	Proteome Browser	Table Schema	UniProt
Entrez Gene	PubMed	OMIM	GeneLynx	GeneCards	HGNC
CGAP	HPRD	Stanford SOURCE	ExonPrimer	Ensembl	Jackson Labs
H-INV	Allen Brain Atlas				

Comments and Description Text from UniProt (Swiss-Prot/TrEMBL)

ID: GBRA3_HUMAN
DESCRIPTION: Gamma-aminobutyric-acid receptor alpha-3 subunit precursor (GABA(A) receptor).
FUNCTION: GABA, the major inhibitory neurotransmitter in the vertebrate brain, mediates neuronal inhibition by binding to the GABA/benzodiazepine receptor and opening an integral chloride channel.
SUBUNIT: Binds UBQLN1 (By similarity). Generally pentameric. There are five types of GABA(A) receptor chains: alpha, beta, gamma, delta, and rho.
SUBCELLULAR LOCATION: Membrane; multi-pass membrane protein.
SIMILARITY: Belongs to the ligand-gated ionic channel (TC 1.A.9) family.
DATABASE: NAME=Protein Spotlight; NOTE=Issue 56 of March 2005; WWW="http://www.expasy.org/spotlight/back_issues/sptlt056.shtml".



Human Gene GABRA3 Description and Page Index

http://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=NM_000808&hgg_prot=

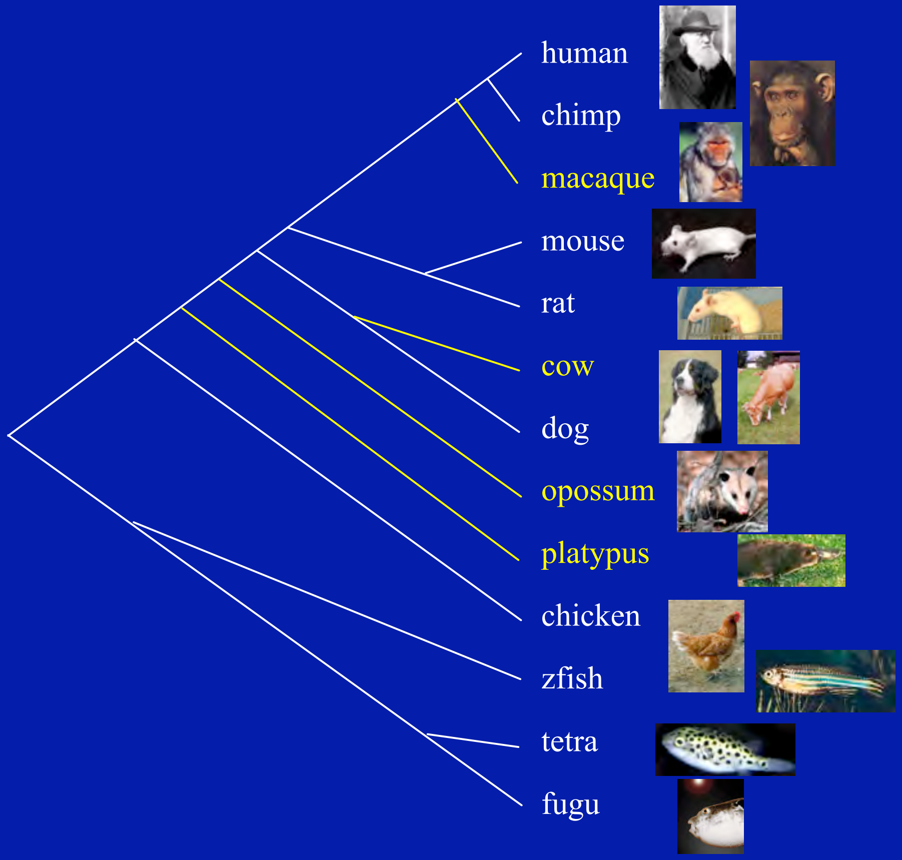
Microarray Expression Data

Expression ratio colors: red high/green low Submit

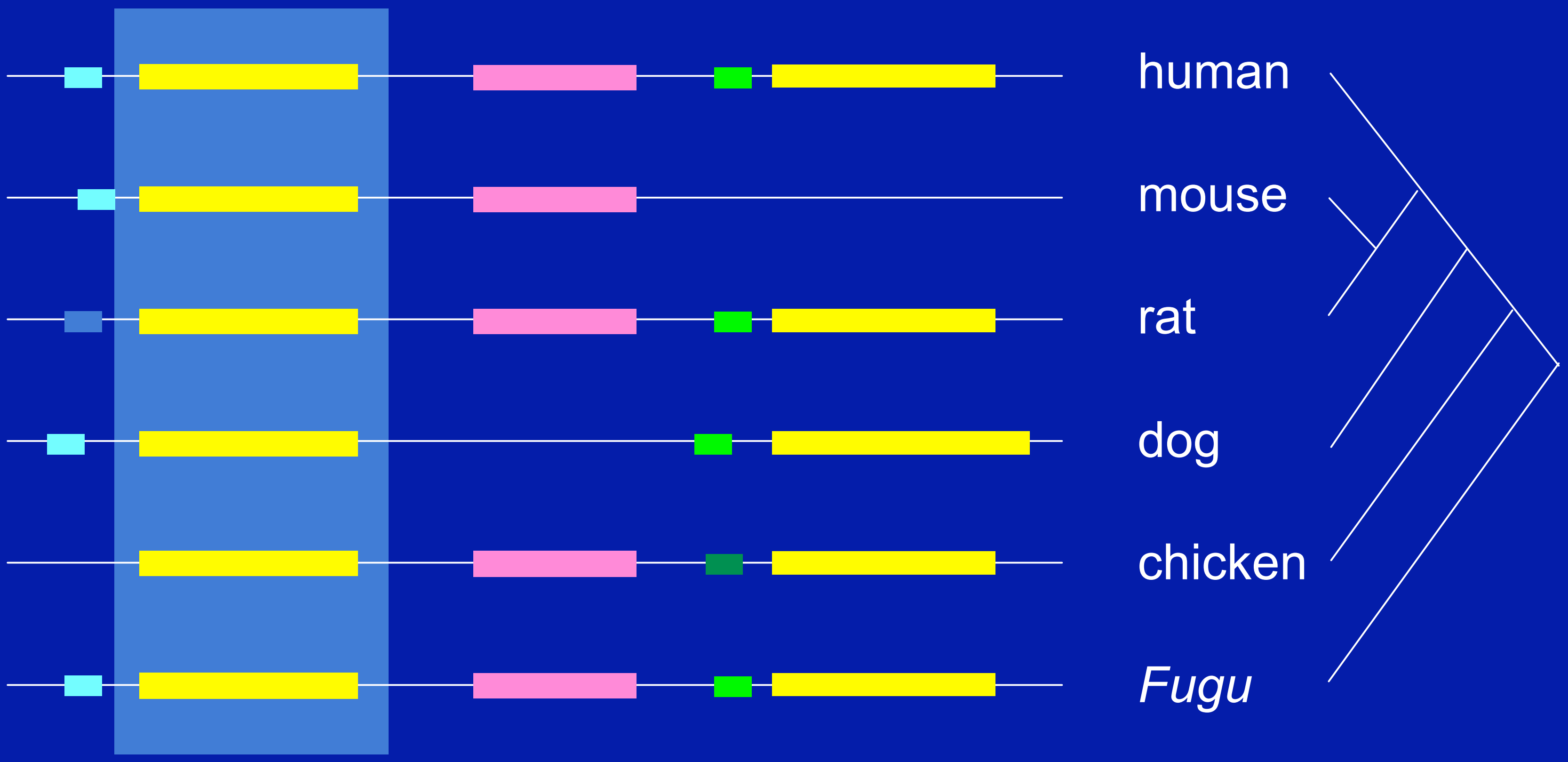
GNF Expression Atlas 2 Data from U133A and GNF1H Chips

Whole Genome Analysis

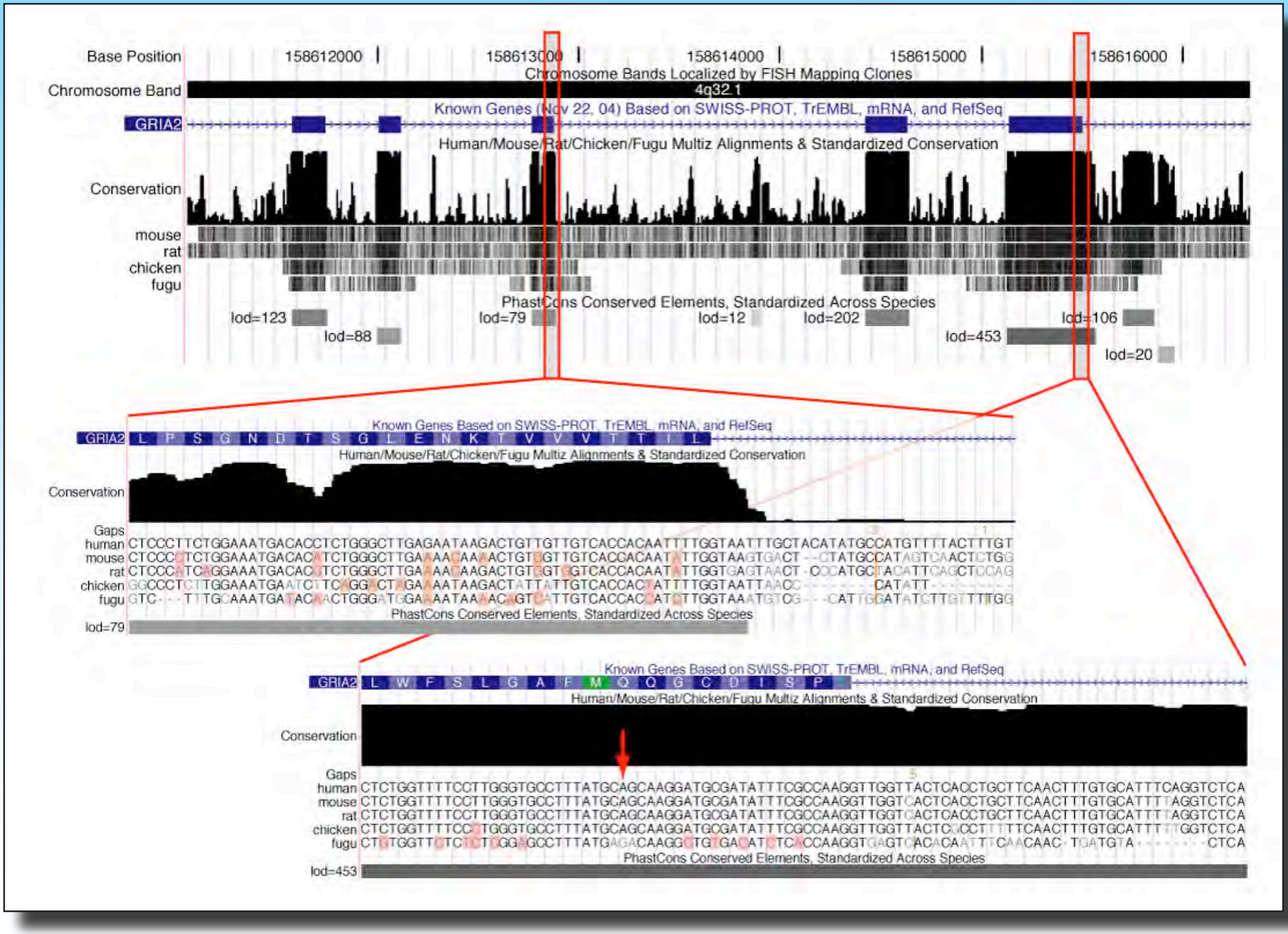
Comparative Analysis of Complete Mammalian Genomes



Detection of Functional Elements



Conservation Track

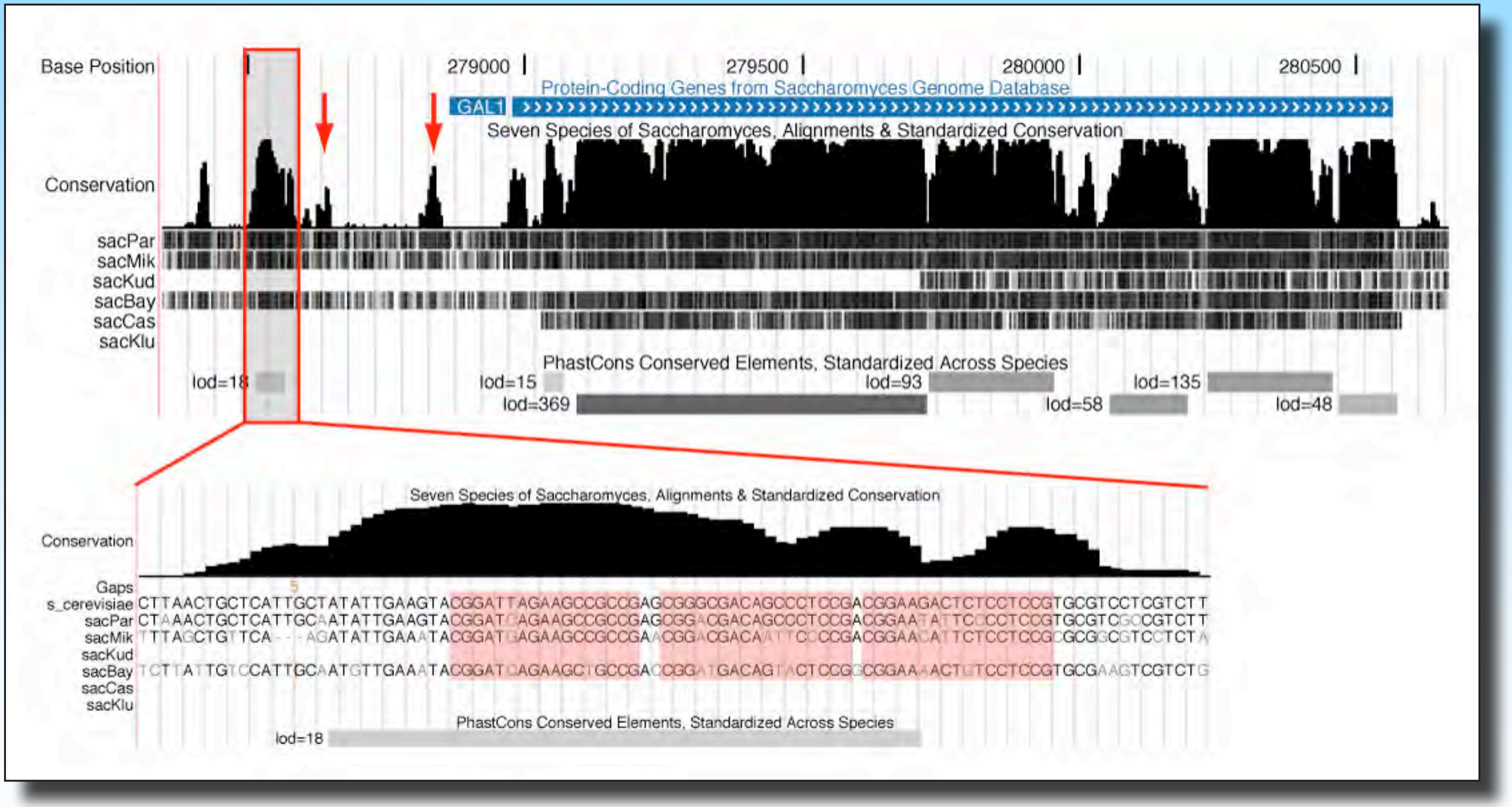

Base Position
158612000
158613000
158614000
158615000
158616000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
4q32.1
GRIA2
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=123
lod=88
lod=79
lod=12
lod=202
lod=453
lod=106
lod=20
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2 L P S G N D T S G L E N K T V V V T T I L
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
PhastCons Conserved Elements, Standardized Across Species
lod=79
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2 L W F S L G A F M Q Q G C D I S P
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
PhastCons Conserved Elements, Standardized Across Species
lod=453

```
  human CTCCCTTCTGGAAATGACACCTCTGGGCTTGAGAATAAGACTGTTGTTGTCACCACAATTTTGGTAATTTGCTACATATGCCATGTTTTACTTTGT
  mouse CTCCCCTCTGGAAATGACACATCTGGGCTTGAAAACAAAACTGTCGTTGTCACCACAATATTGGTAAGTGACT--CTATGCCATAGTCAACTCTGG
    rat CTCCCATCAGGAAATGACACCTCTGGGCTTGAAAACAAGACTGTGGTGGTCACCACAATATTGGTGAGTAACT-CCCATGCTACATTCAGCTCCAG
chicken GGCCCTCTTTGGAAATGAATCTTCAGGACTAGAAAATAAGACTATTATTGTCACCACTATTTTGGTAATTAACC---------CATATT----------
   fugu GTC---TTTGCAAATGATACAACTGGGATGGAAAATAAAACAGTCATTGTCACCACCATCTTGGTAAATGTCG---CATTGGATATCTTGTTTTGG
```

```
  human CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTTACTCACCTGCTTCAACTTTGTGCATTTCAGGTCTCA
  mouse CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTCACTCACCTGCTTCAACTTTGTGCATTTTAGGTCTCA
    rat CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTCACTCACCTGCTTCAACTTTGTGCATTTTAGGTCTCA
chicken CTCTGGTTTTCCCTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTTACTCGCCTTTTTCAACTTTGTGCATTTTTGGTCTCA
   fugu CTGTGGTTCTCTCTCGGAGCCTTTATGAGACAAGGCTGTGACATCTCACCAAGGTGAGTCACACAATTTCAACAAC-TGATGTA---------CTCA
```

Siepel, Bejerano, Pedersen, et al., *Genome Res*, 2005

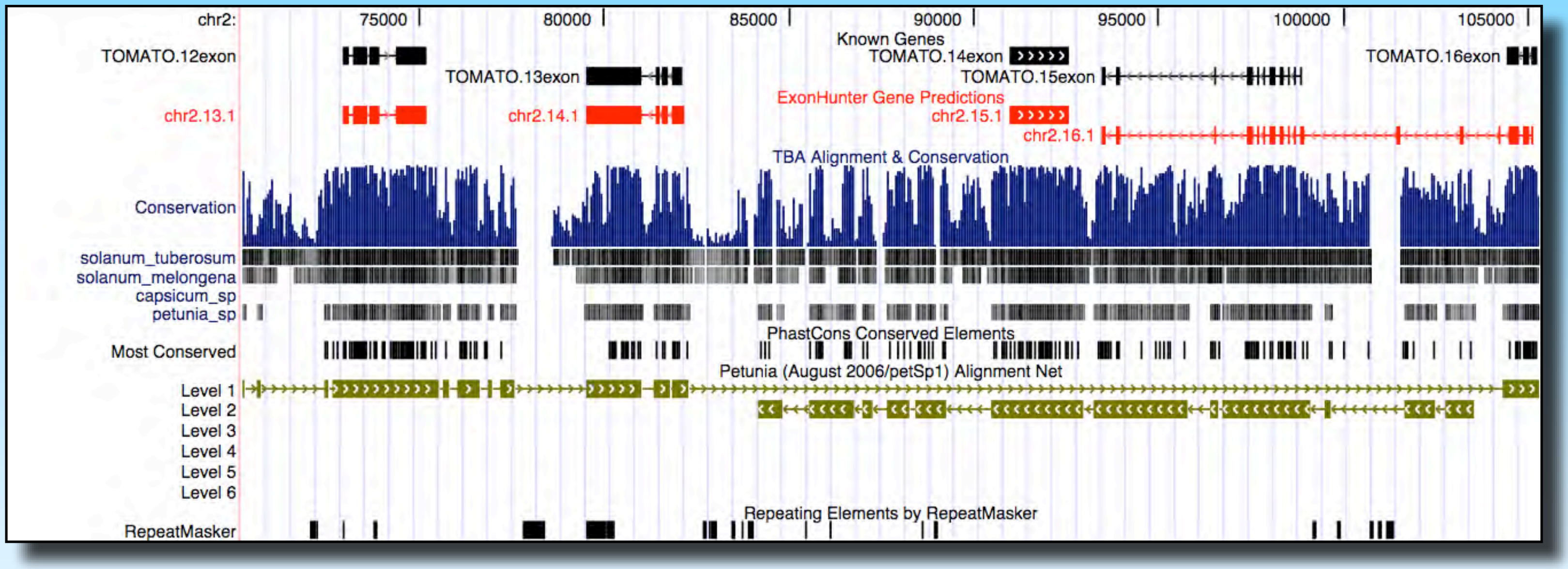
Conservation Track: *GAL1*

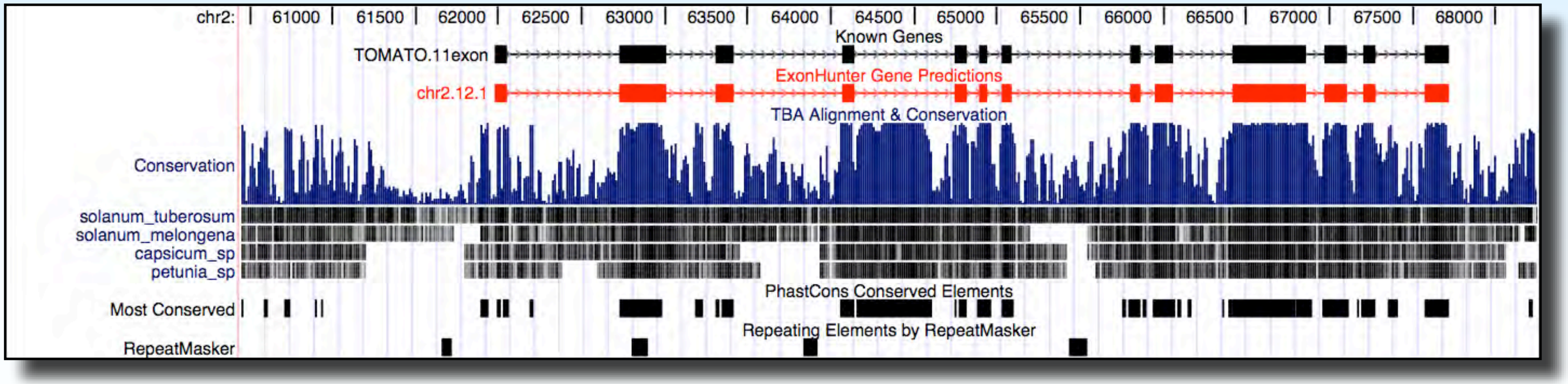

Base Position
279000
279500
280000
280500
Protein-Coding Genes from Saccharomyces Genome Database
GAL1
Seven Species of Saccharomyces, Alignments & Standardized Conservation
Conservation
sacPar
sacMik
sacKud
sacBay
sacCas
sacKlu
PhastCons Conserved Elements, Standardized Across Species
lod=18
lod=15
lod=369
lod=93
lod=58
lod=135
lod=48
Seven Species of Saccharomyces, Alignments & Standardized Conservation
Conservation
Gaps
5
PhastCons Conserved Elements, Standardized Across Species
lod=18

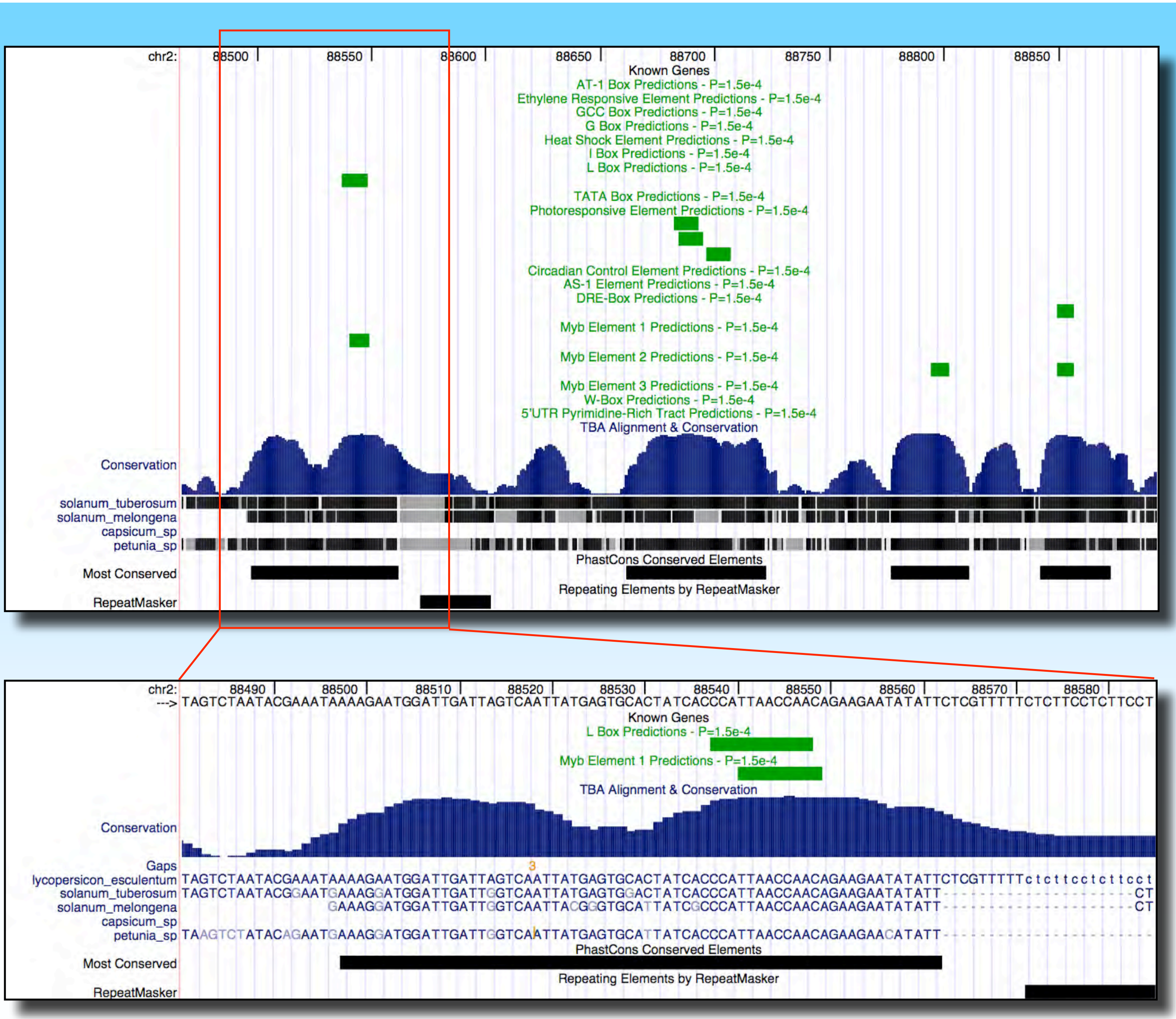
```
s_cerevisiae CTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCCTCGTCTT
      sacPar CTAAACTGCTCATTGCAATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCGCCGTCTT
      sacMik TTTAGCTGTTCA---AGATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCGTCCTCTA
      sacKud
      sacBay TCTTATTGTCCATTGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAGTCGTCTG
      sacCas
      sacKlu
```

Siepel, Bejerano, Pedersen, et al., *Genome Res*, 2005

Solanaceae Browser

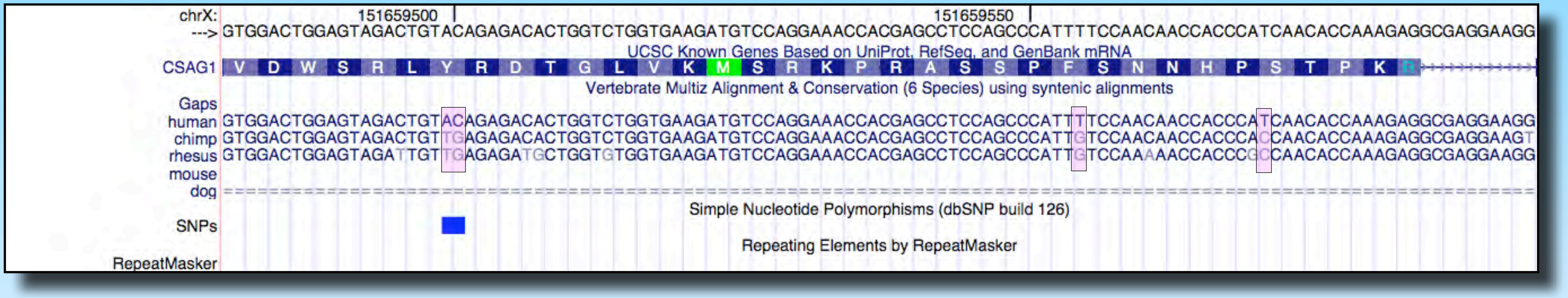




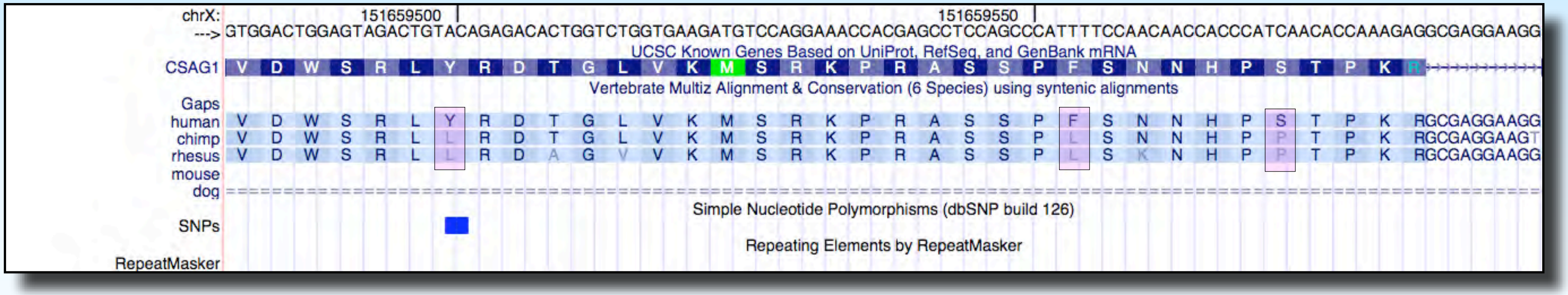

chr2:
88500
88550
88600
88650
88700
88750
88800
88850
Known Genes
AT-1 Box Predictions - P=1.5e-4
Ethylene Responsive Element Predictions - P=1.5e-4
GCC Box Predictions - P=1.5e-4
G Box Predictions - P=1.5e-4
Heat Shock Element Predictions - P=1.5e-4
I Box Predictions - P=1.5e-4
L Box Predictions - P=1.5e-4
TATA Box Predictions - P=1.5e-4
Photoresponsive Element Predictions - P=1.5e-4
Circadian Control Element Predictions - P=1.5e-4
AS-1 Element Predictions - P=1.5e-4
DRE-Box Predictions - P=1.5e-4
Myb Element 1 Predictions - P=1.5e-4
Myb Element 2 Predictions - P=1.5e-4
Myb Element 3 Predictions - P=1.5e-4
W-Box Predictions - P=1.5e-4
5'UTR Pyrimidine-Rich Tract Predictions - P=1.5e-4
TBA Alignment & Conservation
Conservation
solanum_tuberosum
solanum_melongena
capsicum_sp
petunia_sp
PhastCons Conserved Elements
Most Conserved
Repeating Elements by RepeatMasker
RepeatMasker
chr2:
88490
88500
88510
88520
88530
88540
88550
88560
88570
88580
---> TAGTCTAATACGAAATAAAAGAATGGATTGATTAGTCAATTATGAGTGCACTATCACCCATTAACCAACAGAAGAATATATTCTCGTTTTTCTCTTCCTCTTCCT
Known Genes
L Box Predictions - P=1.5e-4
Myb Element 1 Predictions - P=1.5e-4
TBA Alignment & Conservation
Conservation
Gaps
3
lycopersicon_esculentum TAGTCTAATACGAAATAAAAGAATGGATTGATTAGTCAATTATGAGTGCACTATCACCCATTAACCAACAGAAGAATATATTCTCGTTTTTctcttcctcttcct
solanum_tuberosum TAGTCTAATACGGAATGAAAGGATGGATTGATTGGTCAATTATGAGTGGACTATCACCCATTAACCAACAGAAGAATATATT-----------------CT
solanum_melongena GAAAGGATGGATTGATTGGTCAATTACGGGTGCATTATCGCCCATTAACCAACAGAAGAATATATT-----------------CT
capsicum_sp
petunia_sp TAAGTCTATACAGAATGAAAGGATGGATTGATTGGTCAATTATGAGTGCATTATCACCCATTAACCAACAGAAGAACATATT-------------------
PhastCons Conserved Elements
Most Conserved
Repeating Elements by RepeatMasker
RepeatMasker

Whole Genome Analysis

Possible Positive Selection



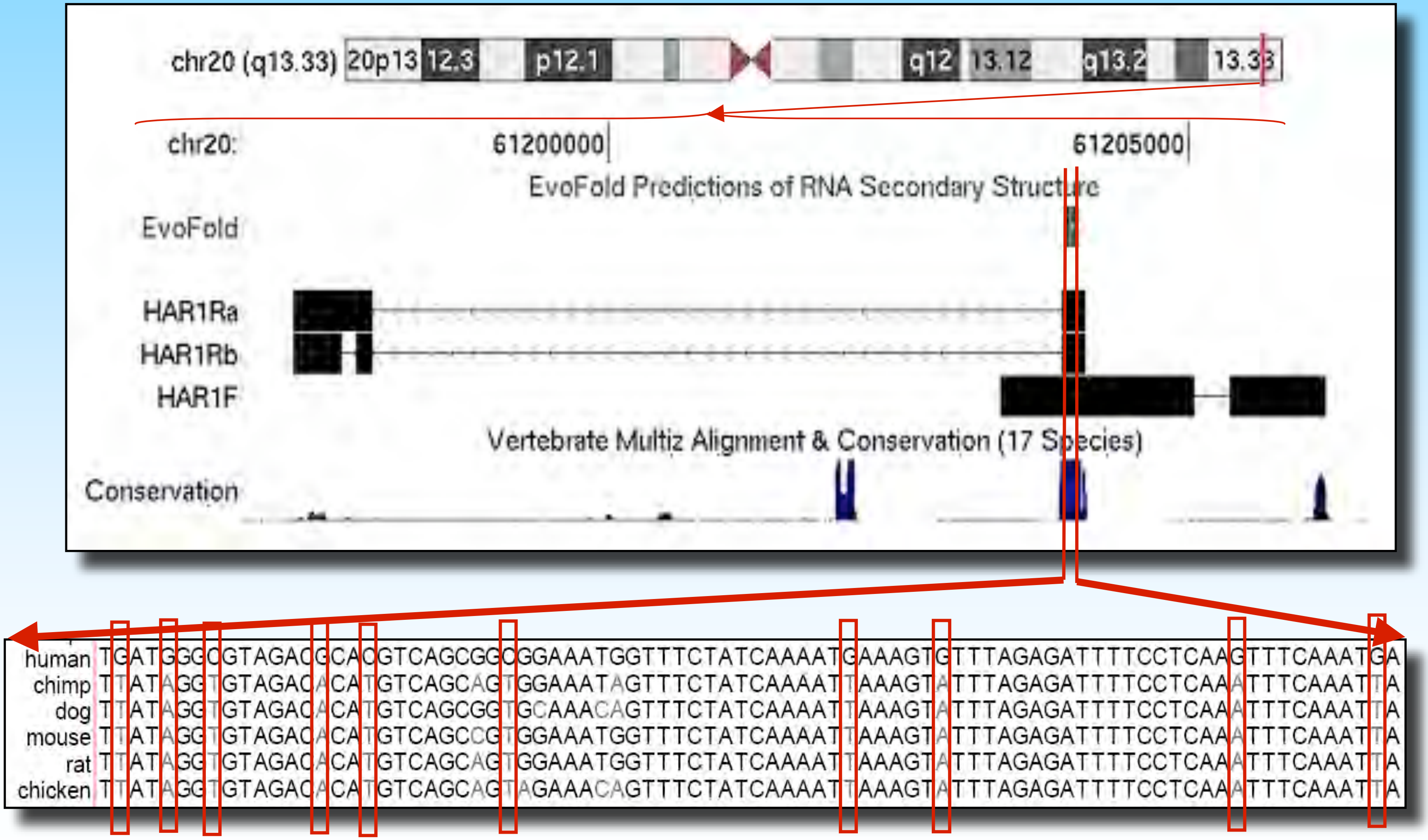
```
      chrX:            151659500 |                                      151659550 |
       ---> GTGGACTGGAGTAGACTGTACAGAGACACTGGTCTGGTGAAGATGTCCAGGAAACCACGAGCCTCCAGCCCATTTTCCAACAACCACCCATCAACACCAAAGAGGCGAGGAAGG
                                  UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
      CSAG1  V  D  W  S  R  L  Y  R  D  T  G  L  V  K  M  S  R  K  P  R  A  S  S  P  F  S  N  N  H  P  S  T  P  K  R
                          Vertebrate Multiz Alignment & Conservation (6 Species) using syntenic alignments
       Gaps
      human GTGGACTGGAGTAGACTGTACAGAGACACTGGTCTGGTGAAGATGTCCAGGAAACCACGAGCCTCCAGCCCATTTTCCAACAACCACCCATCAACACCAAAGAGGCGAGGAAGG
      chimp GTGGACTGGAGTAGACTGTTGAGAGACACTGGTCTGGTGAAGATGTCCAGGAAACCACGAGCCTCCAGCCCATTGTCCAACAACCACCCACCAACACCAAAGAGGCGAGGAAGT
     rhesus GTGGACTGGAGTAGATTGTTGAGAGATGCTGGTGTGGTGAAGATGTCCAGGAAACCACGAGCCTCCAGCCCATTGTCCAAAAACCACCCGCCAACACCAAAGAGGCGAGGAAGG
      mouse
        dog ====================================================================================================================
                               Simple Nucleotide Polymorphisms (dbSNP build 126)
       SNPs
                                   Repeating Elements by RepeatMasker
RepeatMasker
```



```
      chrX:            151659500 |                                      151659550 |
       ---> GTGGACTGGAGTAGACTGTACAGAGACACTGGTCTGGTGAAGATGTCCAGGAAACCACGAGCCTCCAGCCCATTTTCCAACAACCACCCATCAACACCAAAGAGGCGAGGAAGG
                                  UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
      CSAG1  V  D  W  S  R  L  Y  R  D  T  G  L  V  K  M  S  R  K  P  R  A  S  S  P  F  S  N  N  H  P  S  T  P  K  R
                          Vertebrate Multiz Alignment & Conservation (6 Species) using syntenic alignments
       Gaps
      human  V  D  W  S  R  L  Y  R  D  T  G  L  V  K  M  S  R  K  P  R  A  S  S  P  F  S  N  N  H  P  S  T  P  K  RGCGAGGAAGG
      chimp  V  D  W  S  R  L  L  R  D  T  G  L  V  K  M  S  R  K  P  R  A  S  S  P  L  S  N  N  H  P  P  T  P  K  RGCGAGGAAGT
     rhesus  V  D  W  S  R  L  L  R  D  A  G  V  V  K  M  S  R  K  P  R  A  S  S  P  L  S  K  N  H  P  P  T  P  K  RGCGAGGAAGG
      mouse
        dog ====================================================================================================================
                               Simple Nucleotide Polymorphisms (dbSNP build 126)
       SNPs
                                   Repeating Elements by RepeatMasker
RepeatMasker
```

Chondrosarcoma associated gene 1 isoform a

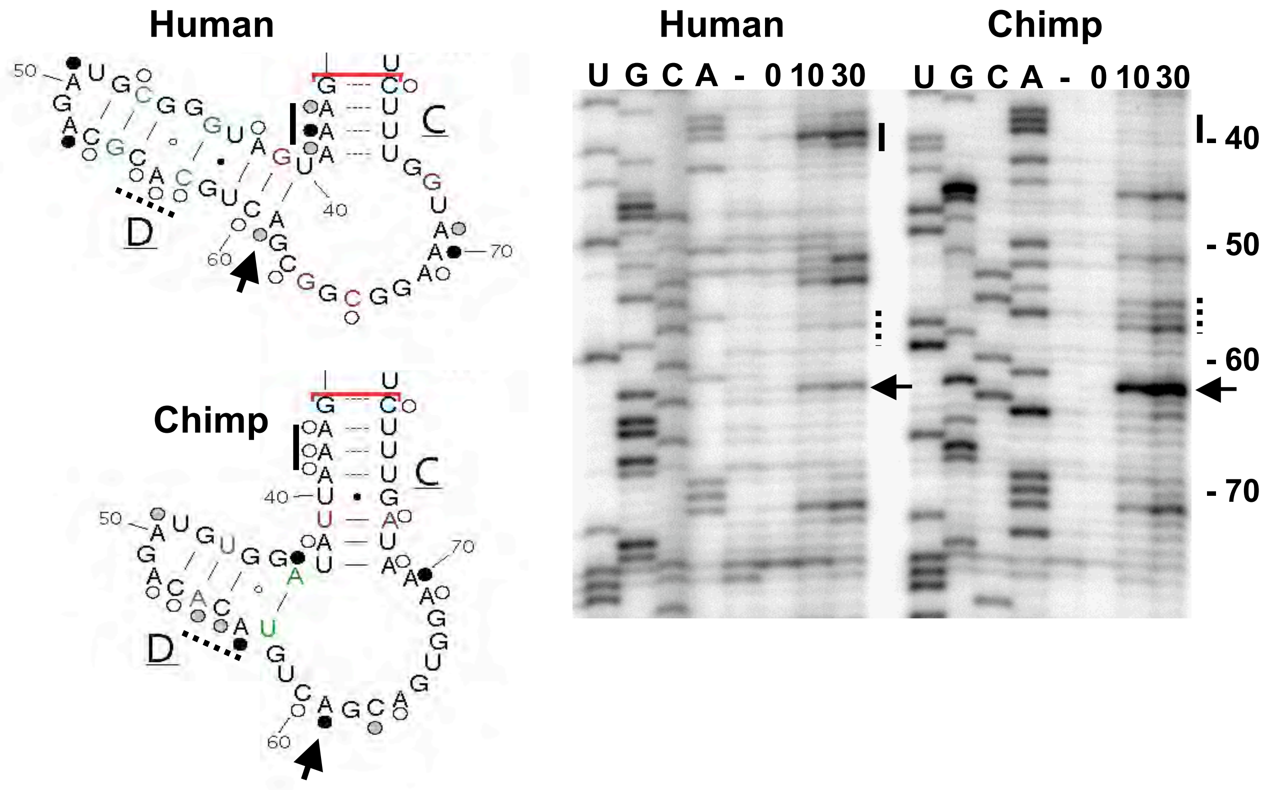
“Human Accelerated Region 1” (HAR1)



```
  human TGATGGGCGTAGACGCACGTCAGCGGCGGAAATGGTTTCTATCAAAATGAAAGTGTTTAGAGATTTTCCTCAAGTTTCAAATGA
  chimp TTATAGGTGTAGACACATGTCAGCAGTGGAAATAGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
    dog TTATAGGTGTAGACACATGTCAGCGGTGCAAACAGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
  mouse TTATAGGTGTAGACACATGTCAGCCGTGGAAATGGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
    rat TTATAGGTGTAGACACATGTCAGCAGTGGAAATGGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
chicken TTATAGGTGTAGACACATGTCAGCAGTAGAAACAGTTTCTATCAAAATTAAAGTATTTAGAGATTTTCCTCAAATTTCAAATTA
```

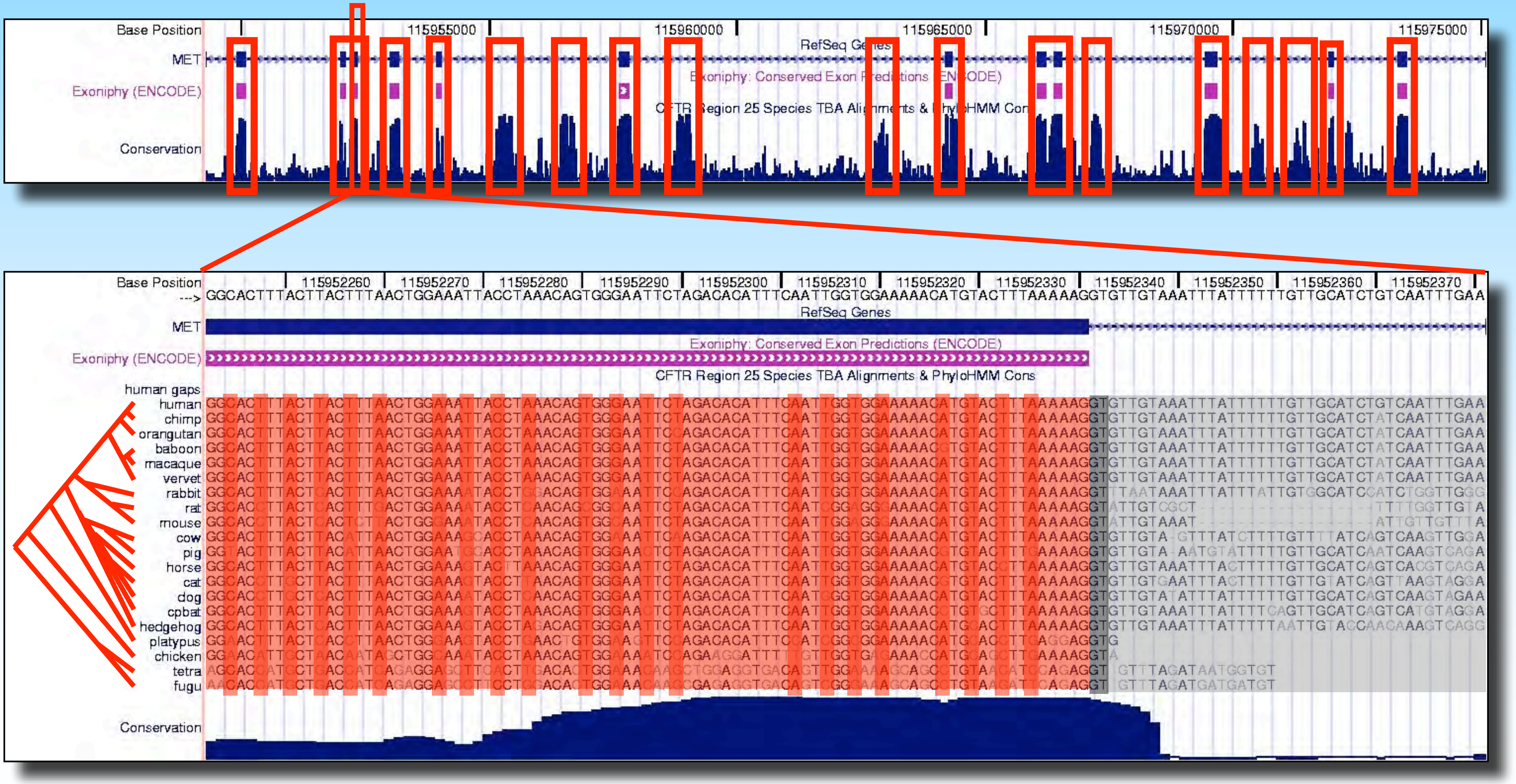
Pollard, Salama, et al., Nature, 2006

New Human RNA Structure



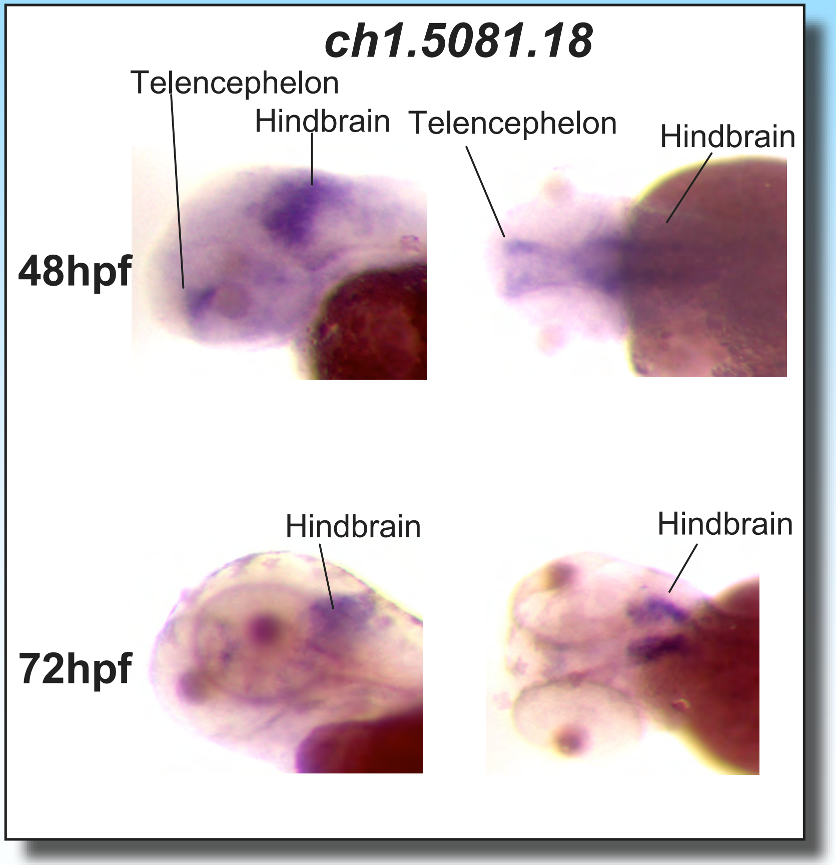
Pollard, Salama, et al., Nature, 2006

Exon Predictions

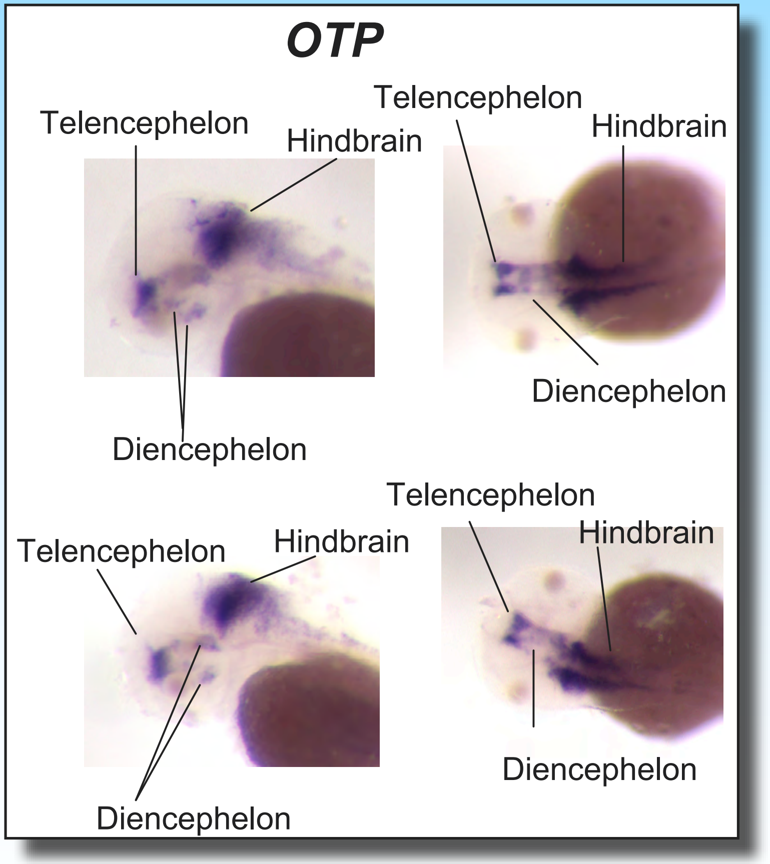


Data from E. Green & colleagues (Thomas et al., *Nature* 2003)

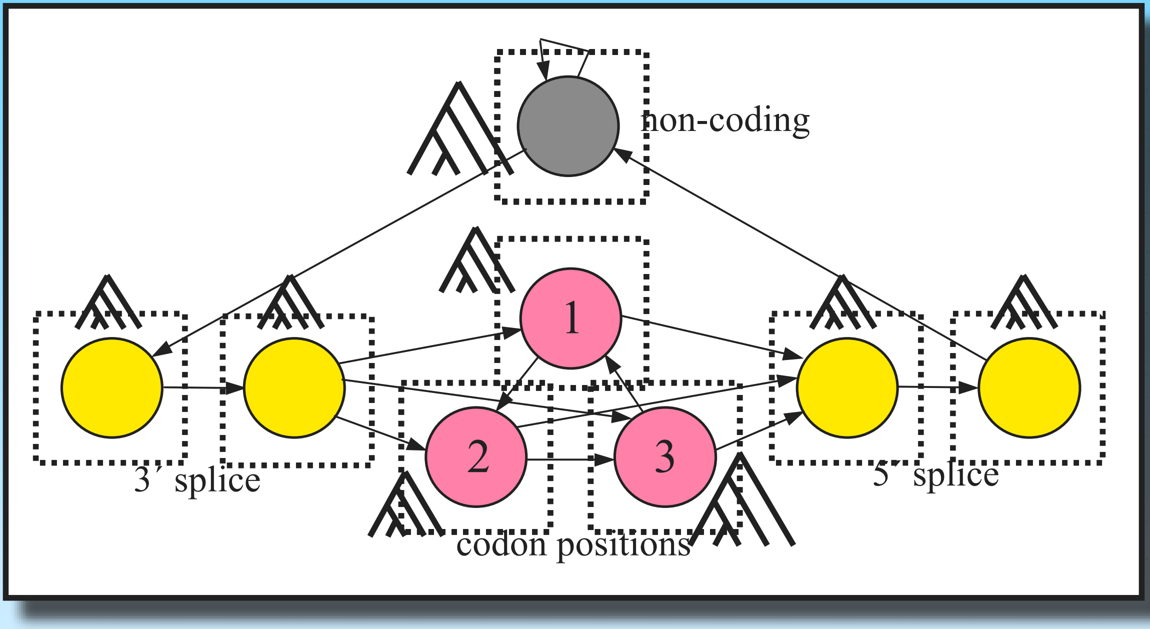
Whole Mount *in situ* Hybridizations to Zebra Fish Embryos



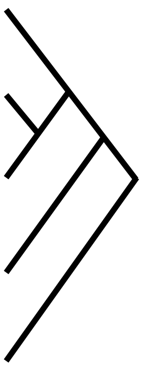
ch1.5081.18



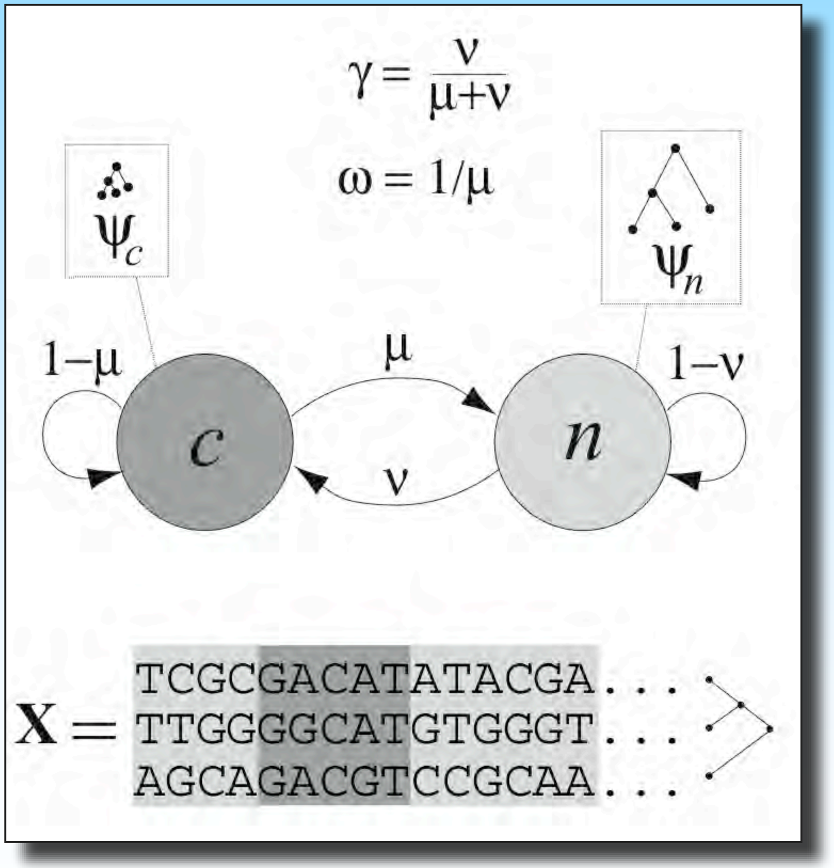
OTP



C	T	C	A	G	C	A	G	…	G	A	G	G	T	A	A	G	human
T	T	A	A	G	C	A	G	…	G	A	A	G	T	G	T	G	mouse
T	T	A	A	G	C	A	G	…	G	A	A	G	T	G	T	T	rat
A	A	T	A	G	C	A	A	…	G	A	G	G	T	C	C	A	dog
C	A	C	A	G	C	A	A	…	G	A	G	G	T	C	A	A	chicken



Phylo-HMM Used by PhastCons



Siepel, Bejerano, Pedersen, et al., *Genome Res*, 2005

Introduction to Hidden Markov Models, Phylogenetic Models, and Phylo-HMMs

OHMAN THE OREGONIAN ©2000
SOMETIMES, I WISH THEY HAD NEVER MAPPED THE HUMAN GENOME...

A Markov Model (Chain)

- Suppose ***Z*** = ($Z_1, ..., Z_L$) is a sequence of cloudy (Z_i = 0) or sunny (Z_i = 1) days
- We could assume days are iid with probability theta of sun but cloudy and sunny days occur in *runs*
- We can capture the correlation between successive days by assuming a first-order Markov model:

$$P(Z_1, \dots, Z_L) = P(Z_1)P(Z_2|Z_1)P(Z_3|Z_2)\cdots P(Z_L|Z_{L-1})$$

instead of complete independence:

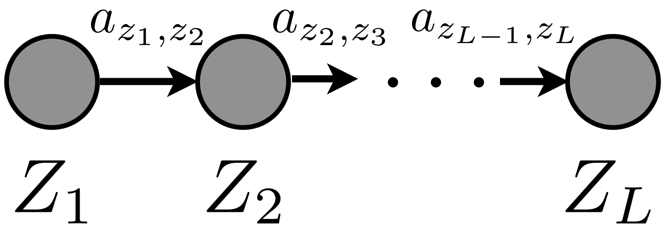
$$P(Z_1, \dots, Z_L) = P(Z_1)\cdots P(Z_L)$$

Three Views

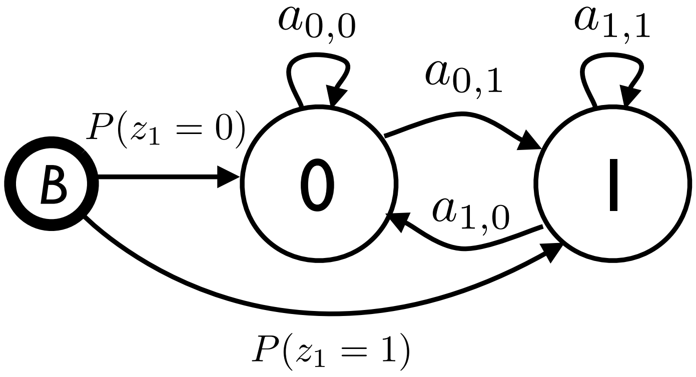
1. $$P(\mathbf{z}) = P(z_1) \prod_{i=2}^{L} a_{z_{i-1},z_i}$$

where $a_{c,d} = P(z_i = d | z_{i-1} = c)$

2.

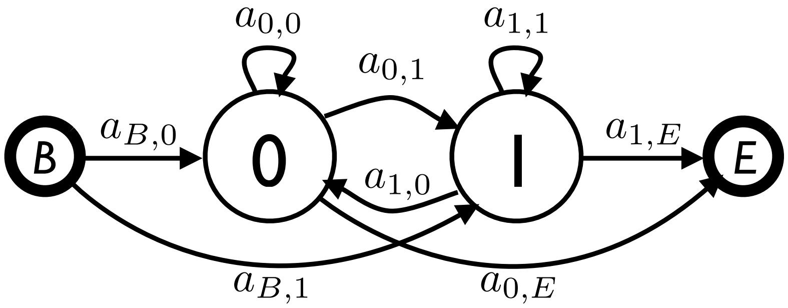


3.



Process Interpretation

- Let's add an *end state* and *cap* the sequence with z_0 = *B*, z_{L+1} = *E*, e.g. $\mathbf{z}$ = *B*00001 1000*E*



- This is a probabilistic machine that generates sequences of any length. It is a stochastic finite state machine and defines a *grammar*.
- We can now simply say: $P(\mathbf{z}) = \prod_{i=0}^{L} a_{z_i, z_{i+1}}$

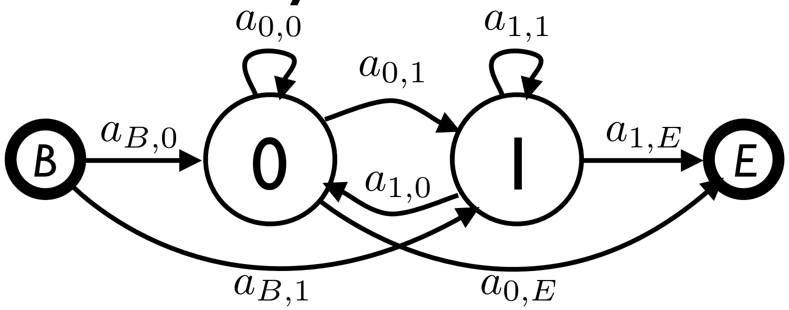
  P($\mathbf{z}$) is a probability distribution over all sequences (for given alphabet).

A Hidden Markov Model

- Let $\boldsymbol{X} = (X_1, ..., X_L)$ indicate whether AS bikes on day *i* ($X_i = 1$) or not ($X_i = 0$)
- Suppose AS bikes on day *i* with probability theta$_0$ = 0.25 if it is cloudy ($Z_i = 0$) and with probability theta$_1$ = 0.75 if it is sunny ($Z_i =1$)
- Further suppose the Z_is are *hidden;* we see only $\boldsymbol{X} = (X_1, ..., X_L)$
- This *hidden Markov model* is a mixture model in which the Z_is are correlated
- We call $\boldsymbol{Z} = (Z_1, ..., Z_L)$ the *path*

HMM, cont.

- ***Z*** is determined by the Markov chain:

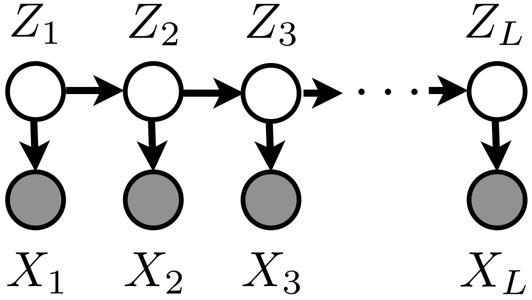


- The joint probability of ***X*** and ***Z*** is:

$$P(\mathbf{x}, \mathbf{z}) = P(\mathbf{z})P(\mathbf{x}|\mathbf{z}) = a_{B,z_1} \prod_{i=1}^{L} e_{z_i,x_i} a_{z_i,z_{i+1}}$$

where $e_{z_i,x_i} = P(x_i|z_i)$

- The *X*



Parameters of the Model

- Transition parameters: a_{s_1,s_2} for all $s_1, s_2 \in S \cup \{B, E\}$
- Emission parameters: $e_{s,x}$ for all $s \in S$, $x \in \mathcal{A}$
- The transition parameters define conditional distributions for state *s*₂ at position *i* given state *s*₁ at position *i*-1
- The emission parameters define conditional distributions over observation *x* given state *s*, both at position *i*
- The observations can be anything!

Key Questions

- Given the model (parameter values) and a sequence ***X***, what is the most likely path?

$$\hat{\mathbf{z}} = \mathrm{argmax}_{\mathbf{z}} P(\mathbf{x}, \mathbf{z})$$

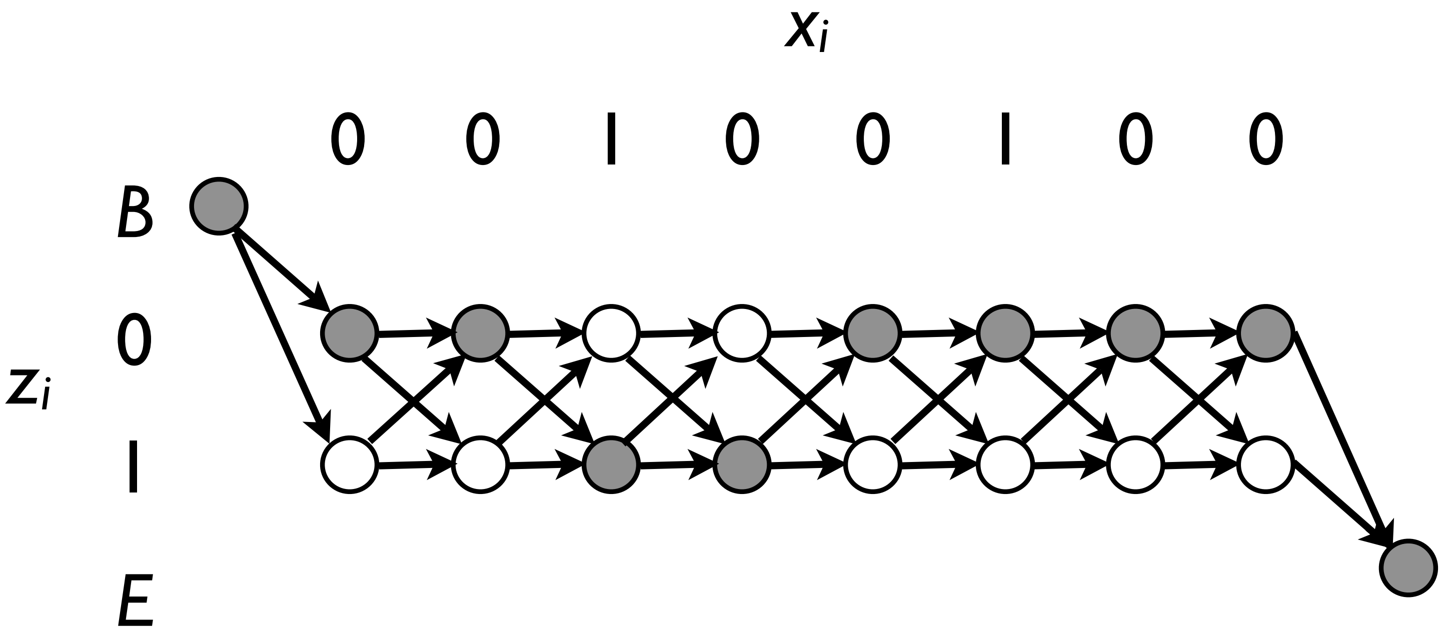
- What is the likelihood of the sequence?

$$P(\mathbf{x}) = \sum_{z} P(\mathbf{x}, \mathbf{z})$$

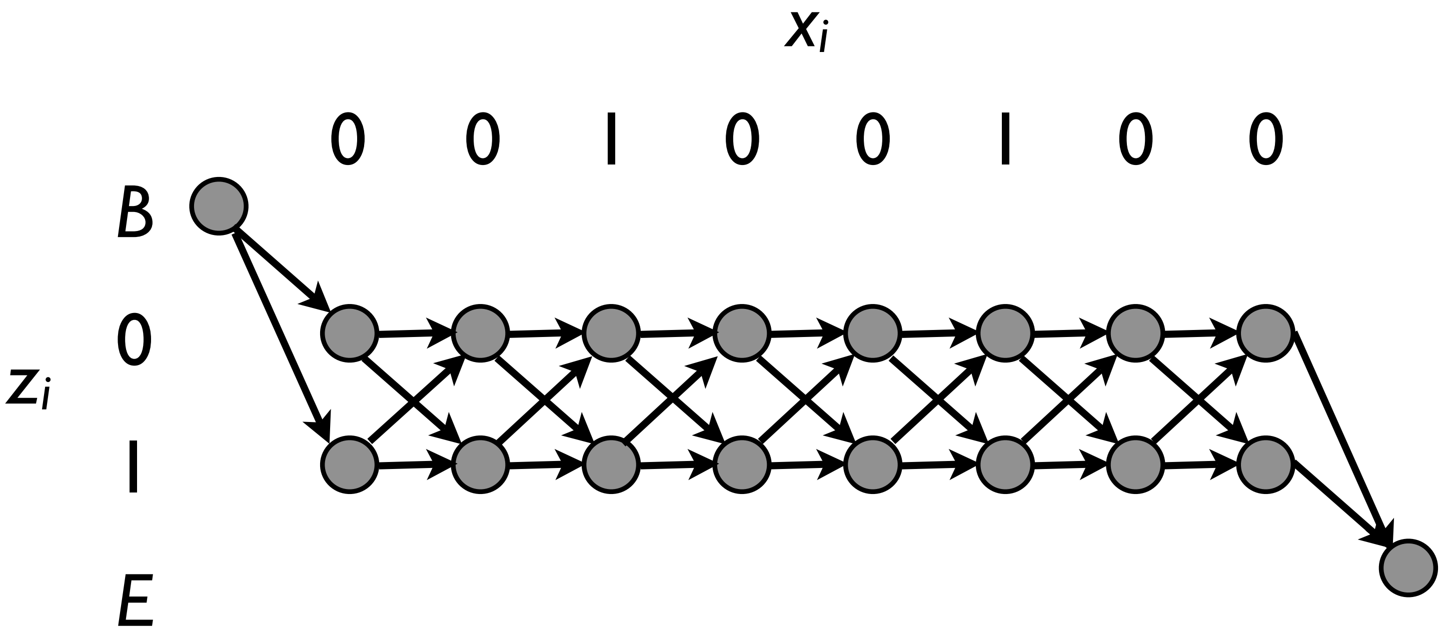
- What is the posterior probability of Z_i given ***X***

- What is the maximum likelihood estimate of all parameters?

Graph Interpretation of Most Likely Path



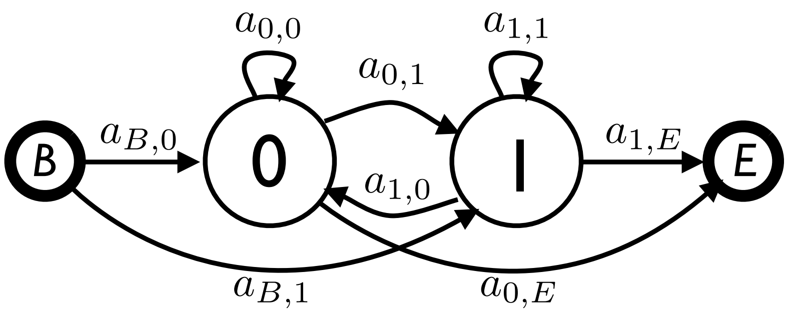
Graph Interpretation of Probability of ***x***



Viterbi Algorithm for Most Likely Path

- Let $v_{i,j}$ be the weight of the most likely path for $(x_1, ..., x_i)$ that ends in state j
- Base case: $v_{0,B} = 1$, $v_{i,B} = 0$ for $i > 0$
- Recurrence: $v_{i,j} = e_{x_i,j} \max_k v_{i-1,k} a_{k,j}$
- Termination: $P(\mathbf{x}, \hat{\mathbf{z}}) = \max_k v_{L,k} a_{k,E}$
- Keep back-pointers for traceback, as in alignment
- See Durbin et al. for algorithm

Example

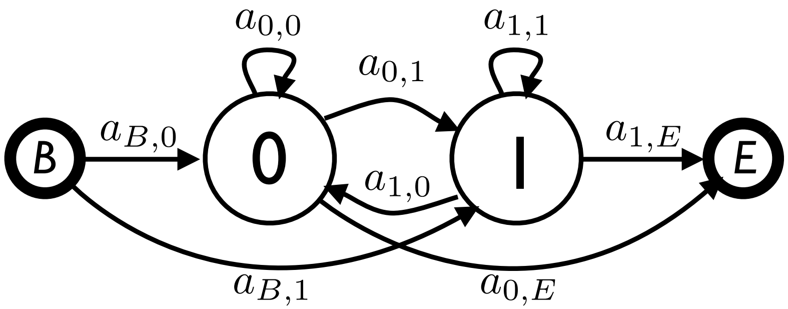


$$P(x_i = 1 | z_i = 0) = 0.25$$
$$P(x_i = 1 | z_i = 1) = 0.75$$

Z = ? ? ? ? ? ? ? ? ? ? ? ?

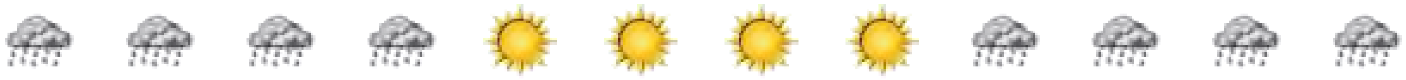
X = 0 1 0 0 1 1 0 1 0 0 1 0

Example



$$P(x_i = 1 | z_i = 0) = 0.25$$

$$P(x_i = 1 | z_i = 1) = 0.75$$



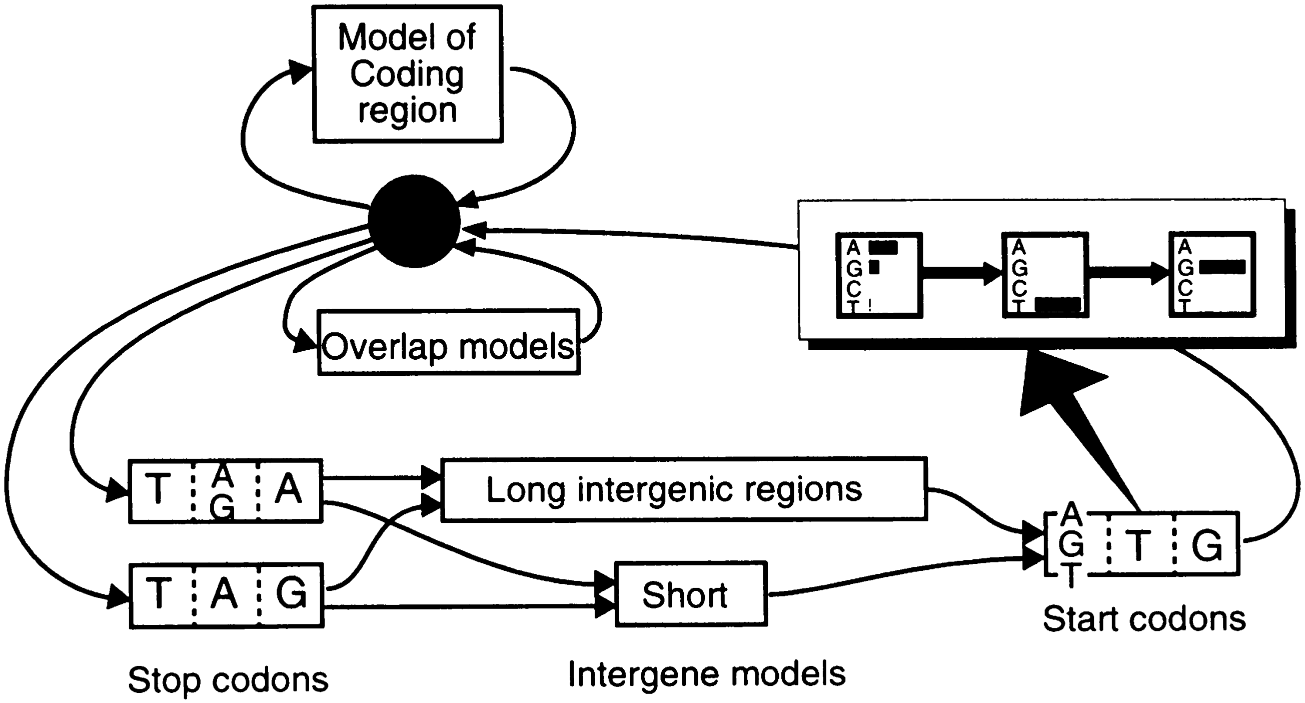
Z = 0 0 0 0 1 1 1 1 0 0 0 0

X = 0 1 0 0 1 1 0 1 0 0 1 0

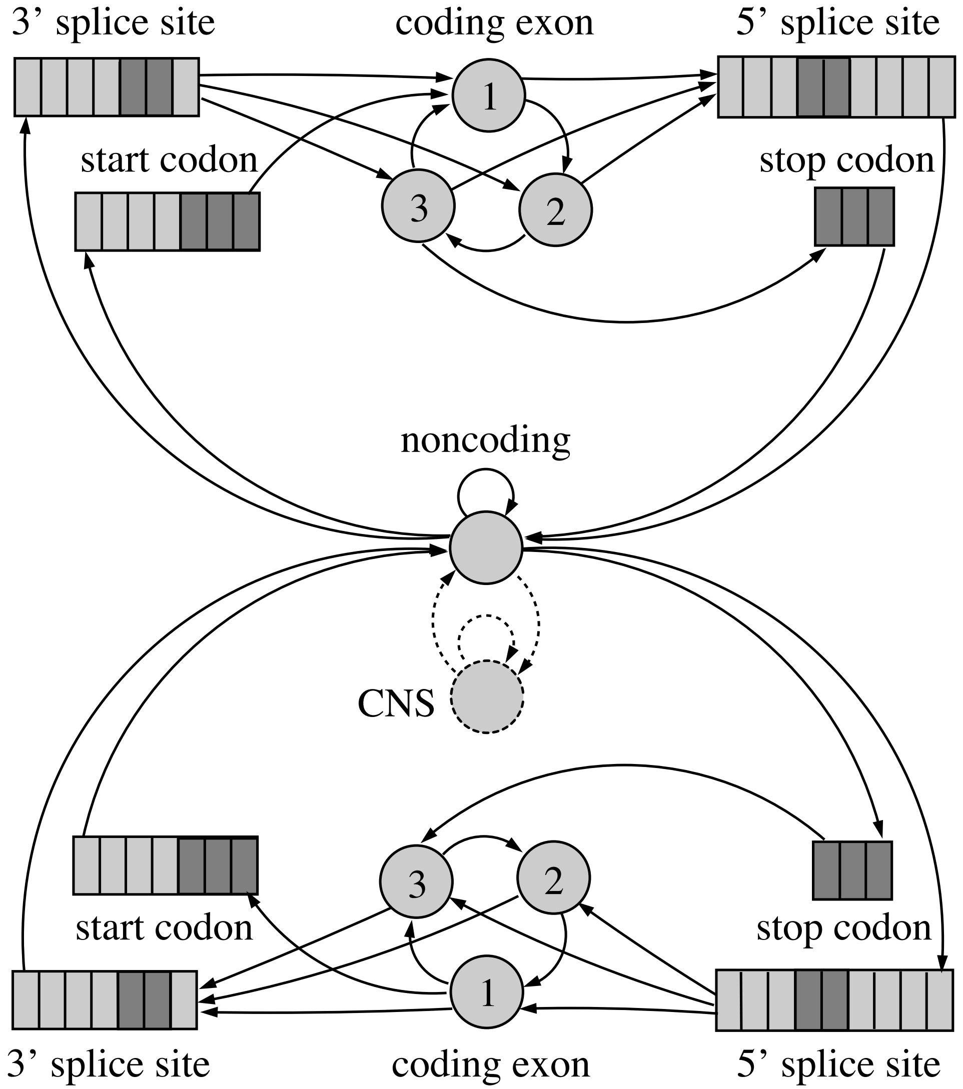
Why HMMs Are Cool

- Extremely general and flexible models for sequence modeling
- Efficient tools for *parsing* sequences
- Also proper probability models: allow maximum likelihood parameter estimation, likelihood ratio tests, etc.
- Inherently *modular*, accommodating of complexity
- In many cases, strike an ideal balance between simplicity and expressiveness

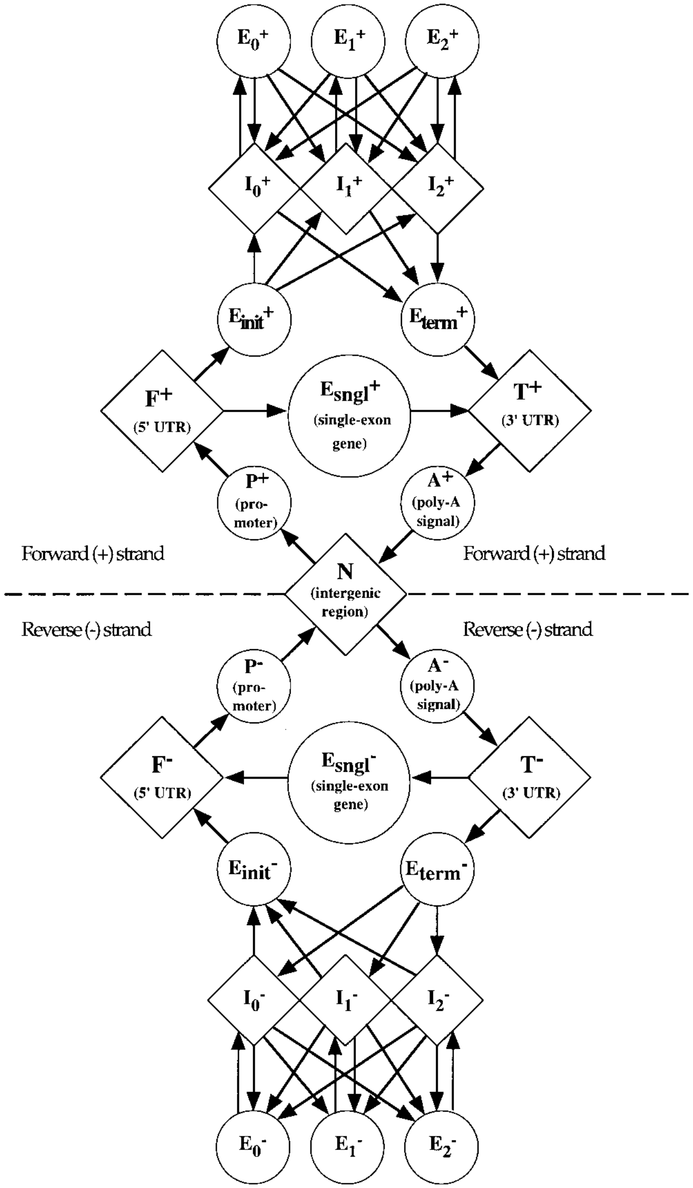
Some Applications In Bioinformatics



Krogh, Mian & Haussler, 1994

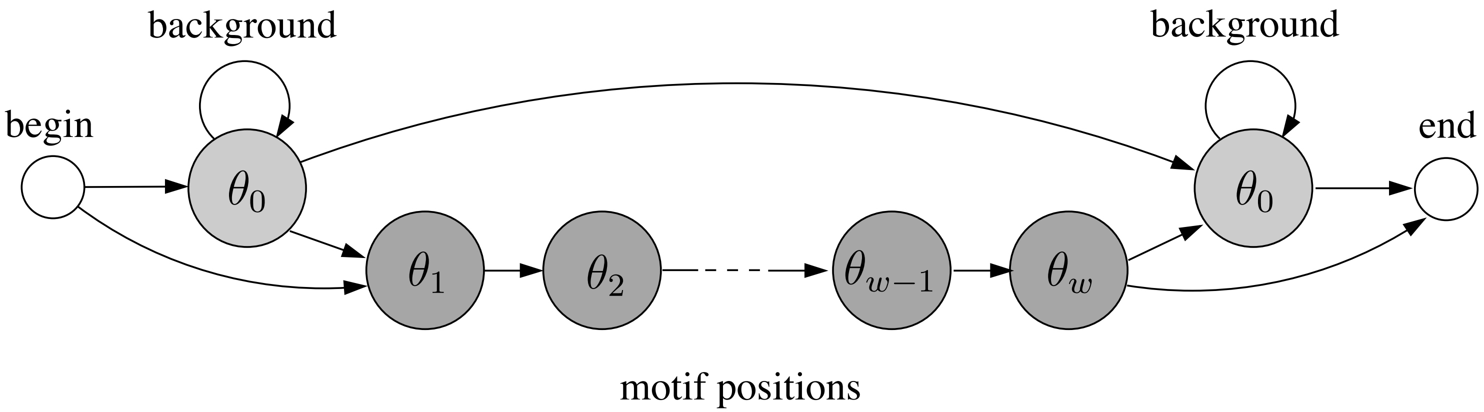


Siepel & Haussler, 2004

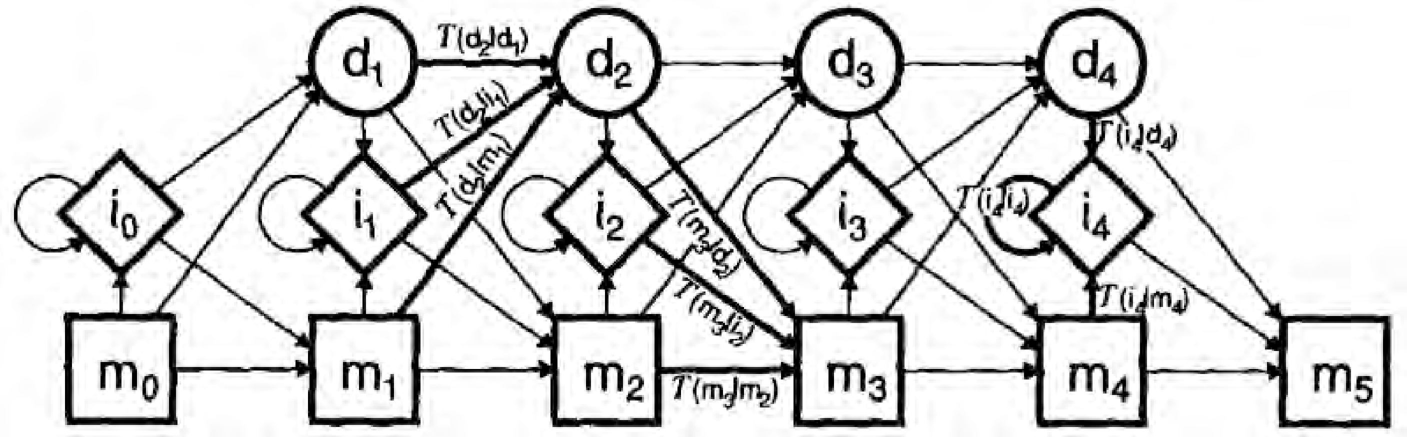


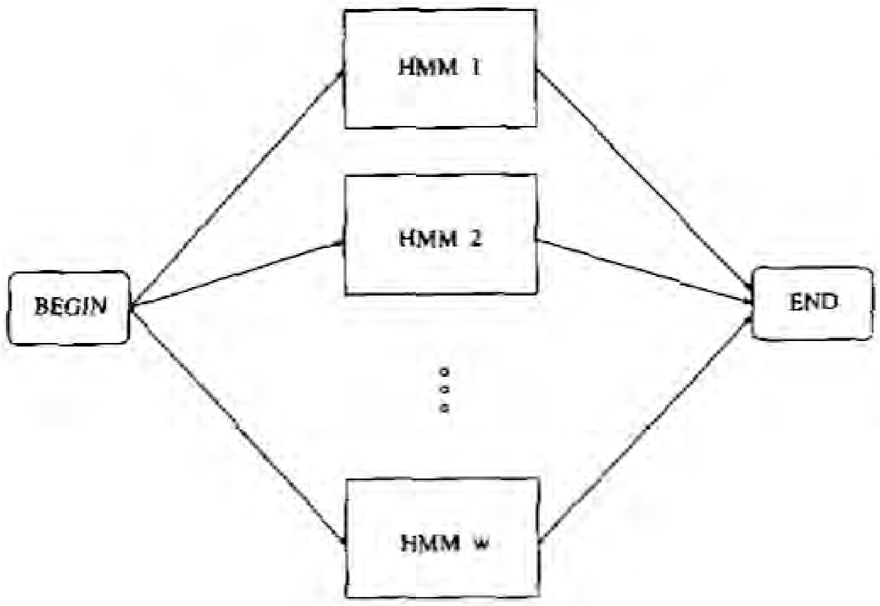
Burge & Karlin, 1997

HMMs Generalize Motif Models



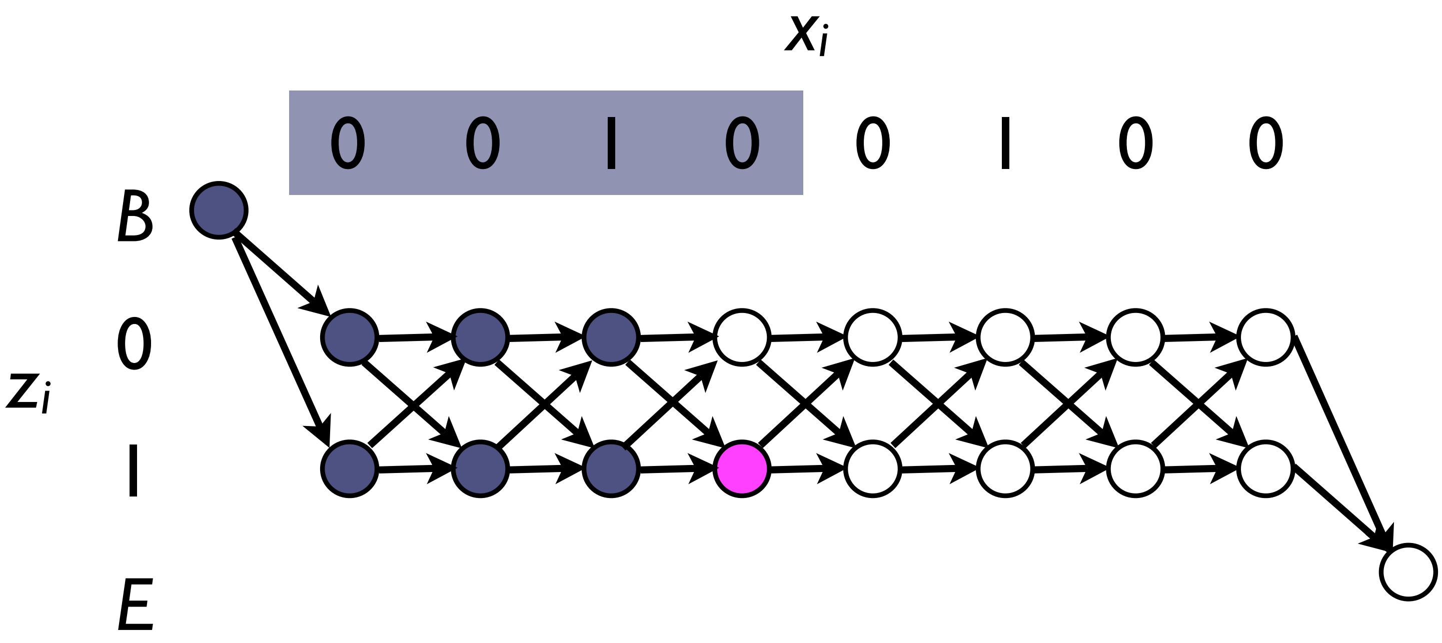
phastMotif





Krogh et al., 1994

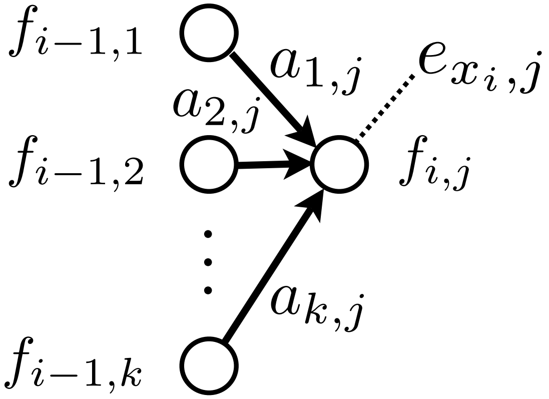
Forward Algorithm



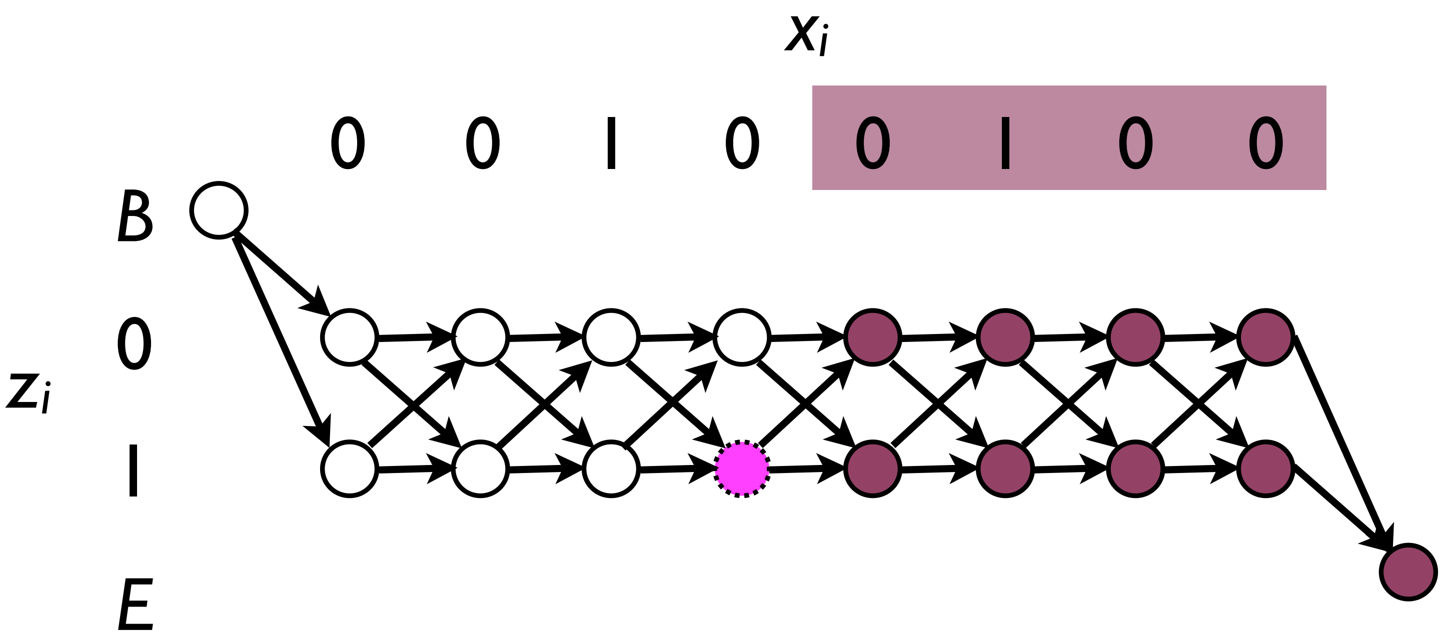
$$f_{4,1} = P(x_1, \ldots, x_4, z_4 = 1)$$

Forward Algorithm

- Let $f_{i,j}$ be the (marginal) probability of (x_1, ..., x_i) *and* $z_i = j$: $f_{i,j} = P(x_1, \ldots, x_i, z_i = j)$
- Base case: $f_{0,B} = 1$, $f_{i,B} = 0$ for $i > 0$
- Recurrence: $f_{i,j} = e_{x_i,j} \sum_k f_{i-1,k} a_{k,j}$
- Termination: $P(\mathbf{x}) = \sum_k f_{L,k} a_{k,E}$



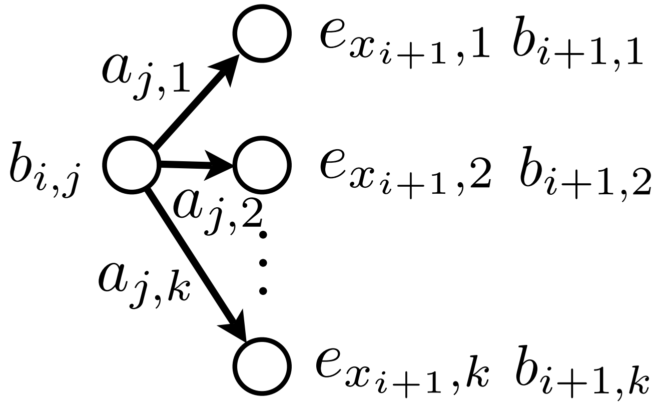
Backward Algorithm



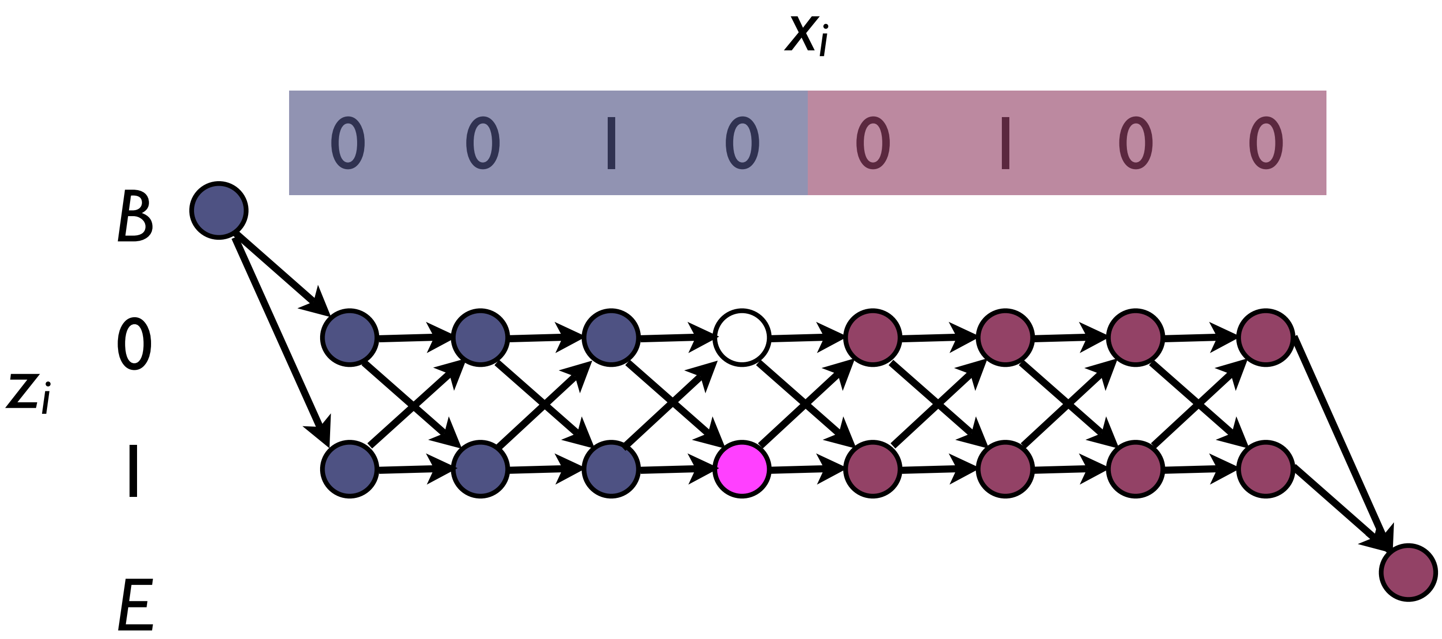
$$b_{4,1} = P(x_5, \ldots, x_L | z_4 = 1)$$

Backward Algorithm

- Let *$b_{i,j}$* be the (marginal) probability of (x_{i+1}, ..., x_L) *given z_i = j*: $b_{i,j} = P(x_{i+1}, \ldots, x_L | z_i = j)$
- Base case: $b_{L,j}$ = $a_{j,E}$ for all states *j*
- Recurrence: $b_{i,j} = \sum_k a_{j,k} e_{x_{i+1},k} b_{i+1,k}$
- Termination: $P(\mathbf{x}) = \sum_k a_{B,k} e_{x_1,k} b_{1,k}$

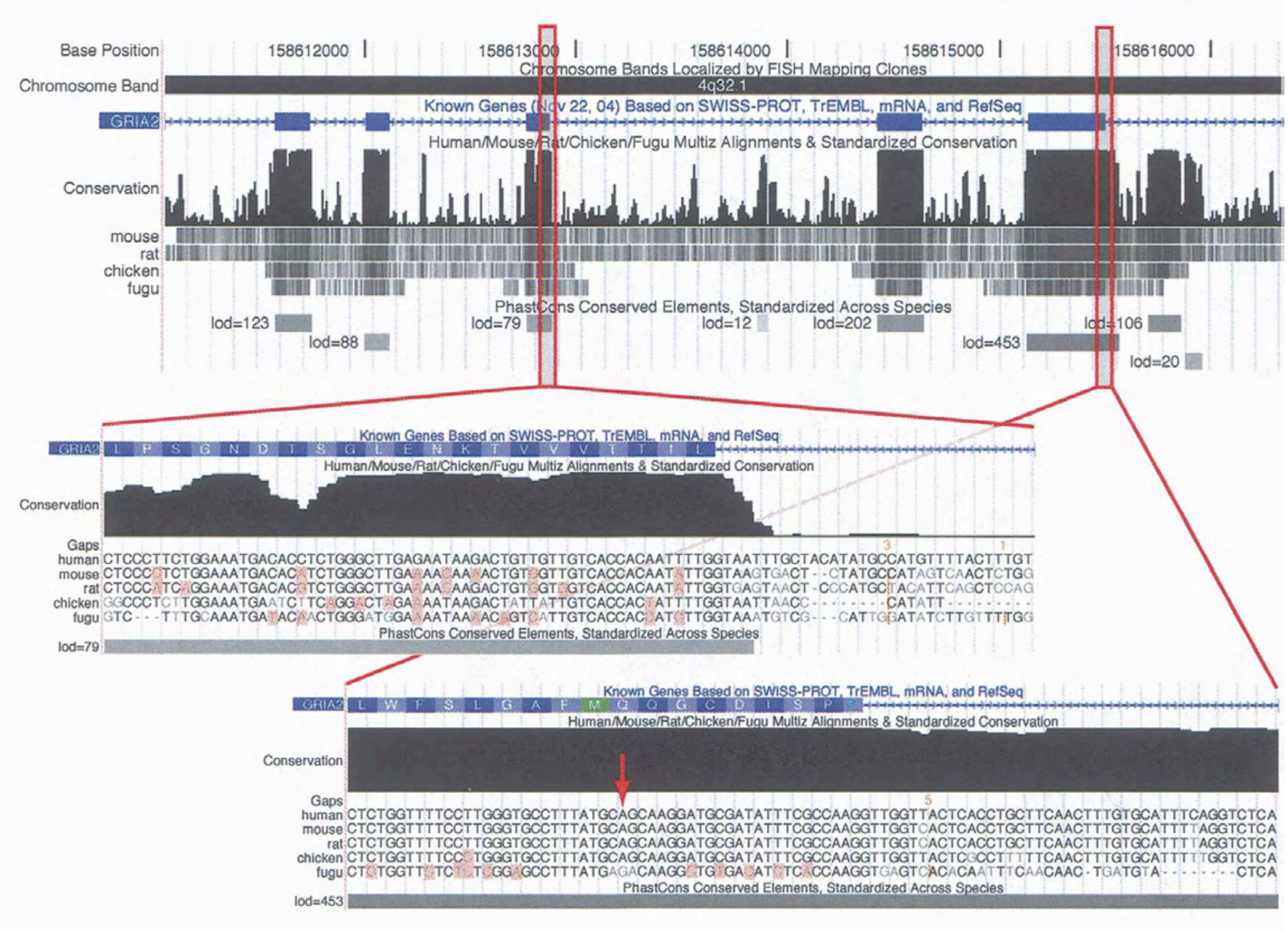

$a_{j,1}$
$b_{i,j}$
$a_{j,2}$
$a_{j,k}$
$e_{x_{i+1},1}\ b_{i+1,1}$
$e_{x_{i+1},2}\ b_{i+1,2}$
$e_{x_{i+1},k}\ b_{i+1,k}$

Forward/Backward



$$P(z_4 = 1|\mathbf{x}) = \frac{P(x_1, \ldots, x_4, z_4 = 1)P(x_5, \ldots, x_L|z_4 = 1)}{P(\mathbf{x})} = \frac{f_{4,1}b_{4,1}}{P(\mathbf{x})}$$

Real-world Use


Base Position
158612000
158613000
158614000
158615000
158616000
Chromosome Band
Chromosome Bands Localized by FISH Mapping Clones
4q32.1
Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
mouse
rat
chicken
fugu
PhastCons Conserved Elements, Standardized Across Species
lod=123
lod=88
lod=79
lod=12
lod=202
lod=453
lod=106
lod=20
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2
L P S G N D T S G L E N K T V V V T T I L
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
3
1
PhastCons Conserved Elements, Standardized Across Species
lod=79
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
GRIA2
L W F S L G A F M Q Q G C D I S P
Human/Mouse/Rat/Chicken/Fugu Multiz Alignments & Standardized Conservation
Conservation
Gaps
5
PhastCons Conserved Elements, Standardized Across Species
lod=453

```
  human CTCCCTTCTGGAAATGACACCTCTGGGCTTGAGAATAAGACTGTTGTTGTCACCACAATTTTGGTAATTTGCTACATATGCCATGTTTTACTTTGT
  mouse CTCCCCTCTGGAAATGACACATCTGGGCTTGAAAACAAAACTGTCGTTGTCACCACAATATTGGTAAGTGACT--CTATGCCATAGTCAACTCTGG
    rat CTCCCATCAGGAAATGACACGTCTGGGCTTGAAAACAAGACTGTCGTCGTCACCACAATATTGGTGAGTAACT-CCCATGCTACATTCAGCTCCAG
chicken GGCCCTCTTGGAAATGAATCTTCAGGACTAGAAAATAAGACTATTATTGTCACCACTATTTTGGTAATTAACC--------CATATT--------
   fugu GTC---TTTGCAAATGATACAACTGGGATGGAAAATAAAACAGTCATTGTCACCACCATCTTGGTAAATGTCG---CATTGGATATCTTGTTTTGG
```

```
  human CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTTACTCACCTGCTTCAACTTTGTGCATTTCAGGTCTCA
  mouse CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTCACTCACCTGCTTCAACTTTGTGCATTTTAGGTCTCA
    rat CTCTGGTTTTCCTTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTCACTCACCTGCTTCAACTTTGTGCATTTTAGGTCTCA
chicken CTCTGGTTTTCCCTGGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCCAAGGTTGGTTACTCGCCTTTTTCAACTTTGTGCATTTTTGGTCTCA
   fugu CTCTGGTTCTCCTCTCGGAGCCTTTATGAGACAAGGCTGTGACATCTCACCAAGGTGAGTCACACAATTTCAACAAC-TGATGTA--------CTCA
```

Typical Phylogeny

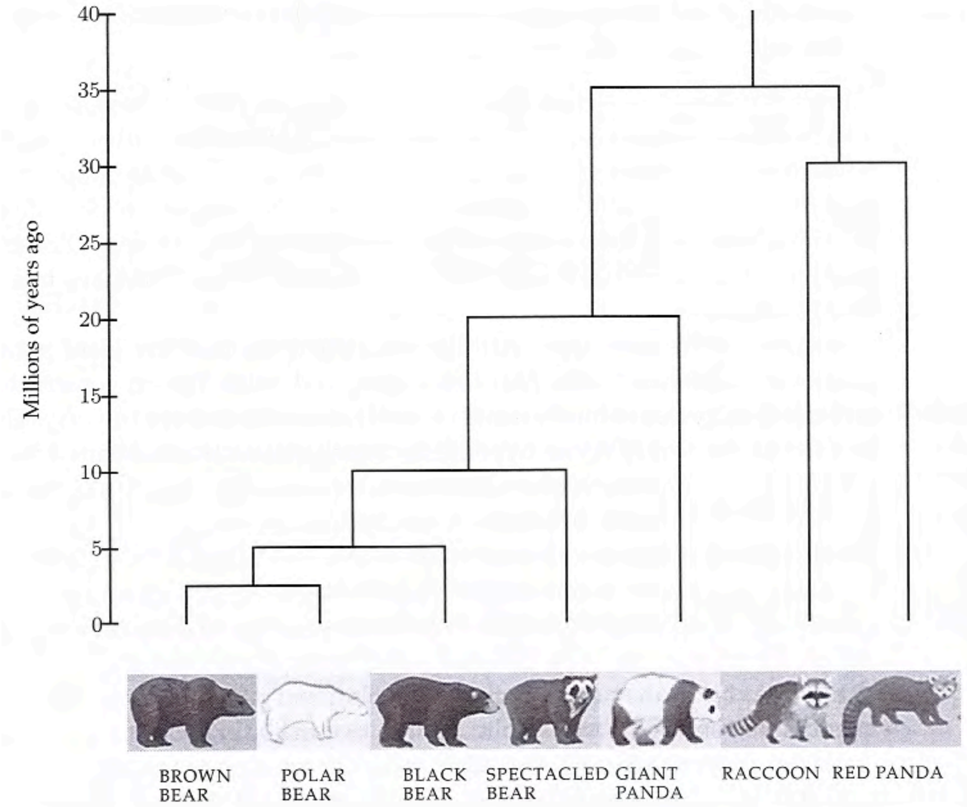
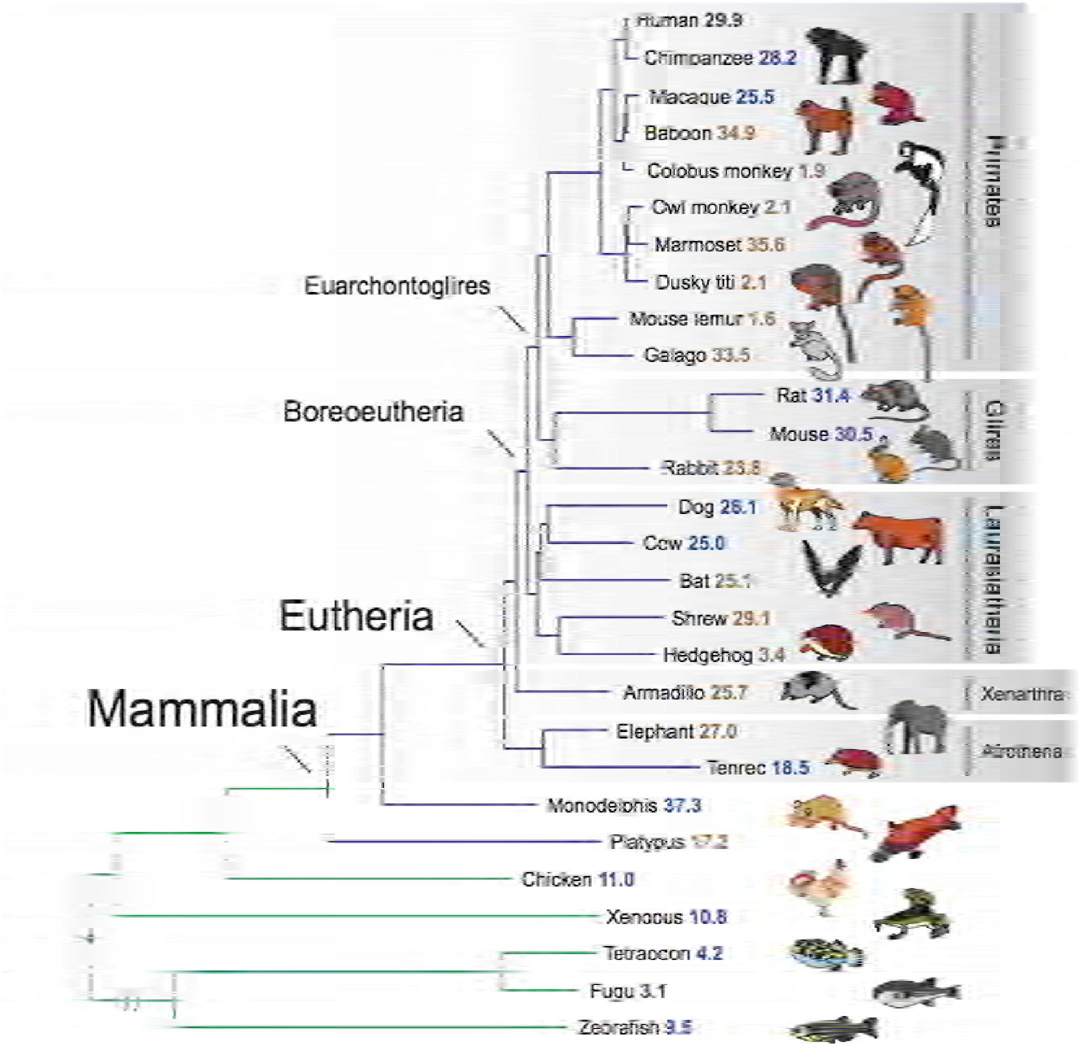


Figure 10.7 An evolutionary tree showing the divergence of raccoons and bears. Despite their difference in size and shape, these two families are closely related.

Recent Vertebrate Phylogeny



Questions

- What is the tree?
- What were the ancestral states (genomes, genes, etc.)?
- When did the divergences occur?
- What is the process?
- Where are the genes?
- ...

The Data

- Originally, morphological “characters” such as number of toes, shape of tooth
- Continuous traits
- DNA or amino acid sequences*
- Gene order or copy number
- Gene expression patterns
- Networks
- ...

General Approaches

- Parsimony: search for tree and ancestral states requiring the fewest events
- Distance matrices: define distance function on taxa, find tree that best approximates matrix of pairwise distances
- Statistical: define probabilistic model, perform ML or Bayesian inference
- Other approaches: compatibility, quartet methods, phylogenetic invariants, Hadamard methods, ...

Parsimony for Sequences

- Given a multiple alignment ***X*** and a tree *T*, let U_T(***X***) be the minimum number of changes (substitutions) along the branches of *T* required to explain *X*
- If U_T(***X***$_i$) is the minimum number of changes for column *i* of ***X***, then

$$U_T(\mathbf{X}) = \sum_i U_T(\mathbf{X}_i)$$

- We seek the best-scoring tree,

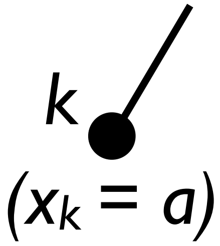
$$\hat{T} = \mathrm{argmin}_T U_T(\mathbf{X})$$

- Ancestral sequences reconstructed in passing

Sankoff's Algorithm

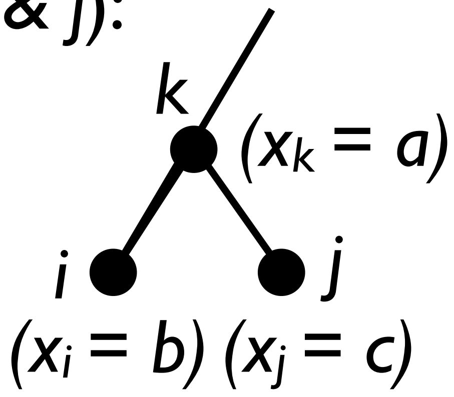
- Let x_k be the base at node *k*. Let $S_k(a)$ be min. no. changes beneath *k*, given $x_k = a$
- Base case (leaf *k*):

$$S_k(a) = \begin{cases} 0 & x_k = a \\ \infty & \text{otherwise} \end{cases}$$



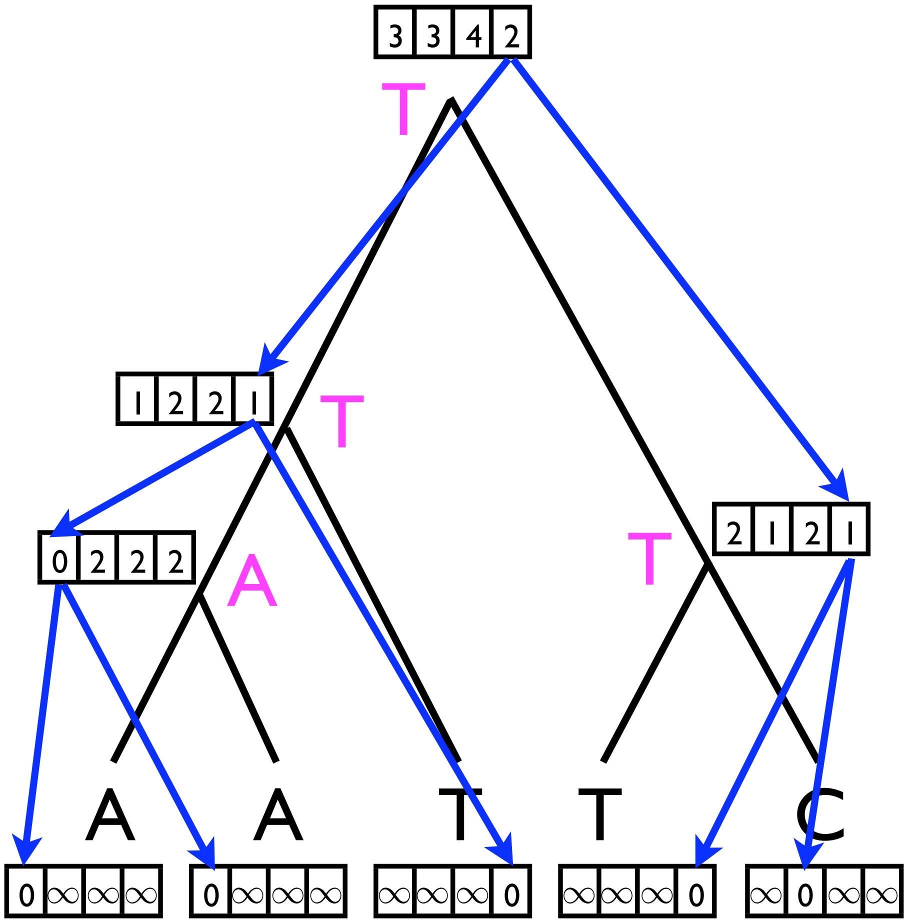
- Recurrence (ancestor *k*, children *i* & *j*):

$$S_k(a) = \min_b \left(S_i(b) + I(a \neq b)\right) + \min_c \left(S_j(c) + I(a \neq c)\right)$$



- Termination: $S_{\text{tree}} = \min_a S_{\text{root}}(a)$

Parsimony Example



Problems with Parsimony

- Incapable of dealing with multiple hits. Especially a problem with long branches
- Not a natural framework for addressing the correlation between “weights” and branch lengths
- Not consistent!
- We would like a statistical approach

Poisson Processes

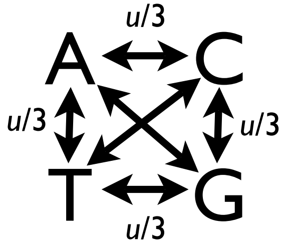
- Let *f*(*x*|*t*) denote the probability of *x* events in an interval of length *t*
- Suppose *f*(*x*|*t*) obeys the *Poisson postulates*:

1. $f(1|t) = \lambda t + o(t) \quad [\lambda > 0,\ \lim_{t \to 0} o(t)/t = 0]$
2. $\sum_{x=2}^{\infty} f(x|t) = o(t)$
3. The numbers of events in nonoverlapping intervals are independent

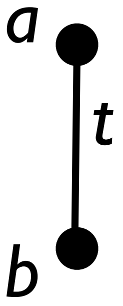
- Then *x* has a Poisson distribution:

$$f(x|t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

Jukes-Cantor Model

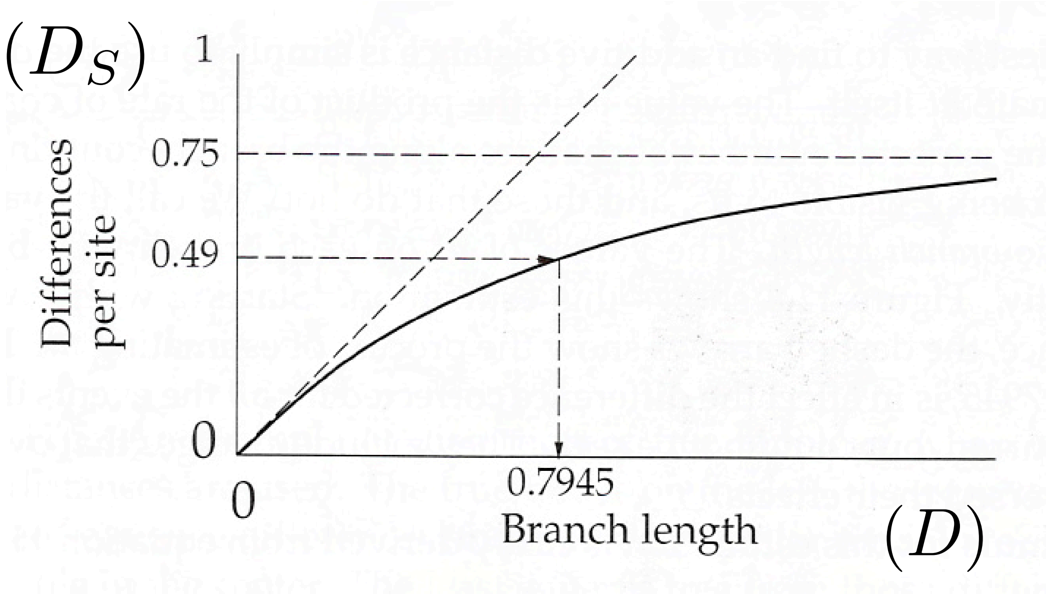


- Suppose DNA substitutions occur by a Poisson process
- *Some* change occurs at rate 4*u*/3. A new base is randomly drawn from the four possibilities.
- On a branch of length *t*, the probability of 0 events is: $e^{-4ut/3}$
- The probability of $\geq$1 events is: $1 - e^{-4ut/3}$
- The probability of *b*|*a* is thus:



$$P(b|a,t) = \begin{cases} e^{-4ut/3} + \frac{1}{4}(1 - e^{-4ut/3}) = \frac{1}{4}(1 + 3e^{-4ut/3}) & b = a \\ \frac{1}{4}(1 - e^{-4ut/3}) & b \neq a \end{cases}$$

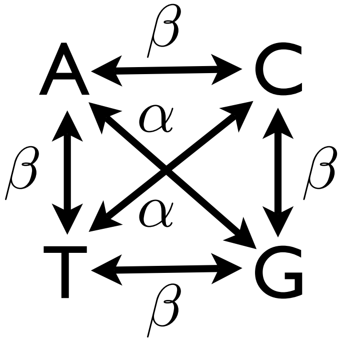
Jukes-Cantor, cont.



$$D = \hat{u}t = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D_S\right)$$

Jukes & Cantor, 1969; Felsenstein, 2004

Kimura's Model



- Distinguishes between transitions and transversions
- Scaling constraint: $\alpha + 2\beta = 1$

  This implies: $\alpha = \dfrac{R}{R+1}, \quad \beta = \dfrac{1}{2(R+1)} \qquad \left[R = \dfrac{\alpha}{2\beta}\right]$
- It can be shown that:

$$P(\text{transition}|t) = \frac{1}{4} - \frac{1}{2}\exp\left(-\frac{2R-1}{R+1}t\right) + \frac{1}{4}\exp\left(-\frac{2}{R+1}t\right)$$

$$P(\text{transversion}|t) = \frac{1}{2} - \frac{1}{2}\exp\left(-\frac{2}{R+1}t\right)$$

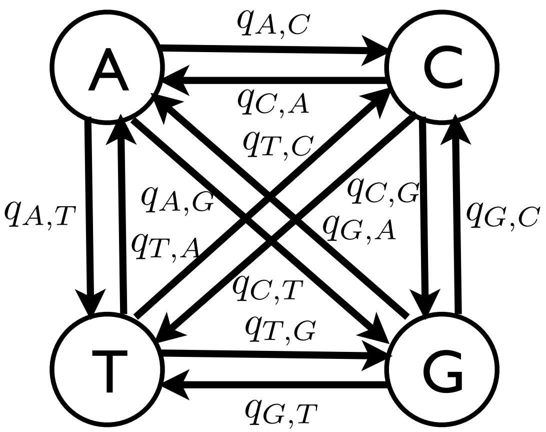
- These relationships are also invertible

Kimura, 1980

Some Other (DNA) Models

- Felsenstein, 1981 (F81): Rates proportional to equilibrium frequencies $(\pi_A, \pi_C, \pi_G, \pi_T)$
- Felsenstein, 1984 (F84): Rates proportional to equilibrium frequencies, transition/ transversion bias
- Hasegawa-Kishino-Yano, 1985 (HKY85): Similar to F84 but different parameterization
- TN93: Generalizes both F84 & HKY85, allows for unequal A-G and C-T transition biases
- ...

A General Framework



$$\mathbf{Q} = \begin{pmatrix} -q_{A,C} - q_{A,G} - q_{A,T} & q_{A,C} & q_{A,G} & q_{A,T} \\ q_{C,A} & -q_{C,A} - q_{C,G} - q_{C,T} & q_{C,G} & q_{C,T} \\ q_{G,A} & q_{G,C} & -q_{G,A} - q_{G,C} - q_{G,T} & q_{G,T} \\ q_{T,A} & q_{T,C} & q_{T,G} & -q_{T,A} - q_{T,C} - q_{T,G} \end{pmatrix}$$

Subject to: $\displaystyle\sum_{a,b:a\neq b} \pi_a q_{a,b} = 1$

Time-Reversibility

- The process is *reversible* if, for all *a* and *b*,

  $$\pi_a q_{a,b} = \pi_b q_{b,a}$$

  where π_x is the equilibrium frequency of base *x*
- This is *not* the same as requiring **Q** to be symmetric, but it does impose a kind of symmetry on the process
- At equilibrium, the expected numbers of *a*-to-*b* and *b*-to-*a* substitutions will be equal
- Reversibility has nice mathematical properties and in most cases is not strongly contradicted by real biological data

The REV (GTR) Model

- The most general reversible model is:

$$\mathbf{Q}_{\text{REV}} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & f\pi_T \\ b\pi_A & d\pi_C & - & g\pi_T \\ c\pi_A & f\pi_C & g\pi_G & - \end{pmatrix}$$

- This model has eight free parameters (accounting for constraints) and a stationary distribution of $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$
- In practice, $\boldsymbol{\pi}$ is often taken to be equal to the observed relative frequencies and the other five parameters are estimated by ML

Others are Special Cases

$$\mathbf{Q}_{\text{JC}} = \begin{pmatrix} - & u/3 & u/3 & u/3 \\ u/3 & - & u/3 & u/3 \\ u/3 & u/3 & - & u/3 \\ u/3 & u/3 & u/3 & - \end{pmatrix} \qquad \boldsymbol{\pi} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

$$\mathbf{Q}_{\text{K2P}} = \begin{pmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{pmatrix} \qquad \boldsymbol{\pi} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$$

$$\mathbf{Q}_{\text{HKY}} = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix} \qquad \boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$$

Computing Probabilities

- Suppose *discrete* Markov process with transition matrix ***A***
- Let ***P***(*k*) be the matrix of conditional probabilities after *k* steps. That is, $\boldsymbol{P}_{a,b}(k)$ = P(*b*|*a,k*). Note ***P***(0) = ***I***
- Recall that ***P***(*k*) = ***P***(*k*-1)***A***, so that ***P***(*k*) = $\boldsymbol{A}^k$ (because $P(b|a,k) = \sum_c P(c|a,k-1)a_{c,b}$)
- Therefore:

$$\begin{aligned}\Delta\mathbf{P}(k) &= \mathbf{P}(k) - \mathbf{P}(k-1) \\ &= \mathbf{P}(k-1)\mathbf{A} - \mathbf{P}(k-1) \\ &= \mathbf{P}(k-1)(\mathbf{A} - \mathbf{I})\end{aligned}$$

Continuous Analog

- Suppose each step represents a tiny segment *dt* of a branch of length t, so *k* = *t* / *dt*. What happens as *dt* approaches 0?
- It can be shown that ***P***(*t*) is continuous, and that a differential equation analogous to the above arises:

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q}$$

- This equation has solution:

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \mathbf{I} + \mathbf{Q}t + \frac{\mathbf{Q}^2t^2}{2} + \frac{\mathbf{Q}^3t^3}{6} + \cdots$$

$$= \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n t^n}{n!}$$

Diagonalization

- In practice, we *diagonalize* ***Q***:

$$\mathbf{Q} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$$

- Now:

$$\begin{aligned}\mathbf{P}(t) &= \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n t^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{(\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1})^n t^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{\mathbf{U}\boldsymbol{\Lambda}^n\mathbf{U}^{-1} t^n}{n!} \\ &= \mathbf{U}e^{\boldsymbol{\Lambda}t}\mathbf{U}^{-1}\end{aligned}$$

Computing Likelihoods

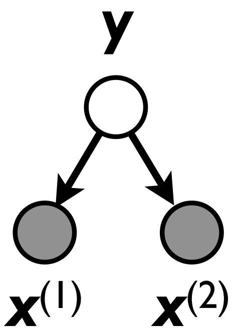
- Suppose $\boldsymbol{X}$ is a (gapless) alignment of $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, with $\boldsymbol{X}_i$ as the ith column.

$\boldsymbol{X}_i$

$\boldsymbol{x}^{(1)}$ = `AATCGGTACGA...`

$\boldsymbol{x}^{(2)}$ = `ATTCAGCACGT...`

- The sequences are derived from an unobserved ancestral sequence $\boldsymbol{y}$



- Assuming independence,

$$P(\mathbf{X}|\mathbf{Q},t,\boldsymbol{\pi}) = \prod_{i=1}^{L} P(\mathbf{X}_i|\mathbf{Q},t,\boldsymbol{\pi}) = \prod_{i=1}^{L} \sum_{y_i} P(x_i^{(1)}, x_i^{(2)}, y_i|\mathbf{Q},t,\boldsymbol{\pi})$$

- Assuming stationarity,

$$P(x_i^{(1)}, x_i^{(2)}, y_i|\mathbf{Q},t,\boldsymbol{\pi}) = \pi_{y_i} P(x_i^{(1)}|y_i,\mathbf{Q},t) P(x_i^{(2)}|y_i,\mathbf{Q},t)$$

Likelihoods, cont.

- Now suppose ***X*** is a *multiple* alignment of sequences related by a (known) phylogeny

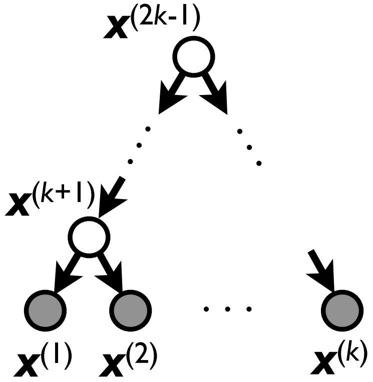
$\boldsymbol{X}_i$

$\boldsymbol{x}^{(1)} =$ AATCGGTACGA...

$\boldsymbol{x}^{(2)} =$ ATTCAGCACGT...

⋮

$\boldsymbol{x}^{(k)} =$ GTTGACTATGA...



- $P(x_i^{(1)}, ..., x_i^{(2k-1)})$ is a product over branches:

$$P\left(x_i^{(1)}, \ldots, x_i^{(2k-1)}\right) = \pi_{x_i^{(2k-1)}} \prod_{j=1}^{2k-2} P\left(x_i^{(j)} \mid x_i^{\mathrm{parent}(j)}, t_j\right)$$

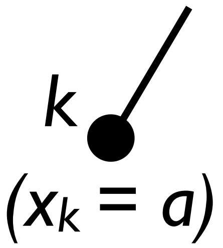
- But we need:

$$P\left(x_i^{(1)}, \ldots, x_i^{(k)}\right) = \sum_{x_i^{(k+1)}, \ldots, x_i^{(2k-1)}} P\left(x_i^{(1)}, \ldots, x_i^{(2k-1)}\right)$$

Recall: Sankoff's Algorithm

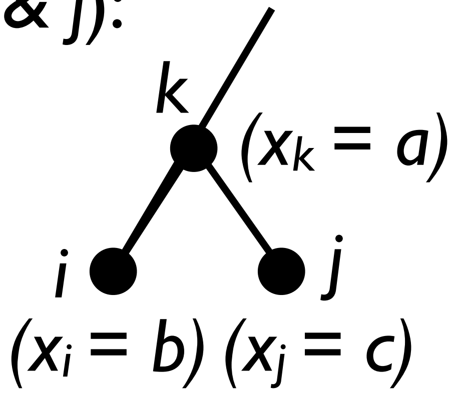
- Let x_k be the base at node k. Let $S_k(a)$ be min. no. changes beneath k, given $x_k = a$
- Base case (leaf k):

$$S_k(a) = \begin{cases} 0 & x_k = a \\ \infty & \text{otherwise} \end{cases}$$



- Recurrence (ancestor k, children i & j):

$$\begin{aligned} S_k(a) = & \min_b \left(S_i(b) + w(a \to b)\right) \\ & + \min_c \left(S_j(c) + w(a \to c)\right) \end{aligned}$$

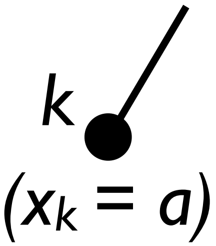


- Termination: $S_{\text{tree}} = \min_a S_{\text{root}}(a)$

Felsenstein's Algorithm

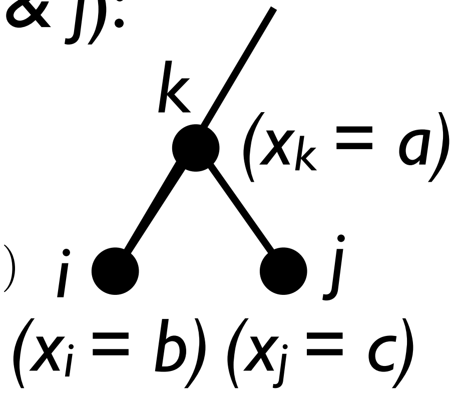
- Let $P(x^{(\underline{k})} \mid x^{(k)} = a)$ be the probability of the *observed bases* beneath node k, given $x^{(k)} = a$
- Base case (leaf k):

$$P(x^{(\underline{k})}|x^{(k)} = a) = \begin{cases} 1 & x^{(k)} = a \\ 0 & \text{otherwise} \end{cases}$$



- Recurrence (ancestor k, children i & j):

$$P(x^{(\underline{k})}|x^{(k)} = a) = \sum_b P(x^{(\underline{i})}|x^{(i)} = b)P(b|a, t_i)$$
$$\times \sum_c P(x^{(\underline{j})}|x^{(j)} = c)P(c|a, t_j)$$



- Termination:

$$P(x^{(1)}, \ldots, x^{(k)}) = \sum_a \pi_a P(x^{(\underline{2k-1})}|x^{(2k-1)} = a)$$

Estimating Parameters

- We now have an efficient way to compute the likelihood of a given phylogenetic model,

$$P(\mathbf{X}|\mathcal{T}, \mathbf{t}, \boldsymbol{\pi}, \mathbf{Q})$$

- If we fix the tree $\mathcal{T}$, ML estimation of the other parameters is a standard nonlinear optimization problem:

$$(\hat{\mathbf{t}}, \hat{\boldsymbol{\pi}}, \hat{\mathbf{Q}}) = \arg\max_{\mathbf{t}, \boldsymbol{\pi}, \mathbf{Q}} P(\mathbf{X}|\mathcal{T}, \mathbf{t}, \boldsymbol{\pi}, \mathbf{Q})$$

- It can be solved numerically using well-known algorithms (e.g., quasi-Newton methods)

Finding the Tree

- Unfortunately, finding the tree is still hard.
- Like with parsimony, we use heuristic or branch-and-bound methods to search the space of trees. We compute a likelihood for each tree and keep the best one.
- Unlike with parsimony, we have to solve a nonlinear optimization problem for each tree!
- Divide-and-conquer heuristics can be useful, because the search space for small trees is manageable

Posterior Probabilities

- What is the posterior distribution of bases at the root? By Bayes' rule:

$$P(x^{(2k-1)} = a | x^{(1)}, \ldots, x^{(k)}) = \frac{P(x^{(1)}, \ldots, x^{(k)} | x^{(2k-1)} = a)\pi_a}{P(x^{(1)}, \ldots, x^{(k)})}$$

- We have already computed the numerator and the denominator! (Felsenstein's algorithm)
- With reversibility, we can root the tree at any node and compute the posterior distribution
- Possible to compute simultaneously for all nodes using an "inside/outside" algorithm resembling the forward/backward algorithm

Non-nucleotide Models

- Can define ***Q*** in terms of codons, amino acids, paired nucleotides in RNA structures
- Codon models are especially useful. They can be parameterized in terms of a nonsynonymous/synonymous rate ratio ω.
- Estimates of this parameter imply negative selection, positive selection, or neutral evolution
- Likelihood ratio tests for positive selection can be constructed