Stochastic Processes

Katie Pollard

BMI 206
http://docpollard.org/bmi206
November 23, 2016

Stochastic processes: random series

- Random variables X_t where t=1,2,3,…
- Typically t indexes time, but in bioinformatics it is often position along a sequence.
- Examples of stochastic process models:
  - Possion process (continuous time, count events)
    - X_t is number of events in interval 0 to t
    - Events in disjoint intervals are independent
  - Brownian motion (continuous time, position)
    - $X_t - X_s$ is displacement in interval s to t
    - Displacements in disjoint intervals are independent and normally distributed
  - Markov process (discrete time)

Markov Processes

Markov processes: limited dependence

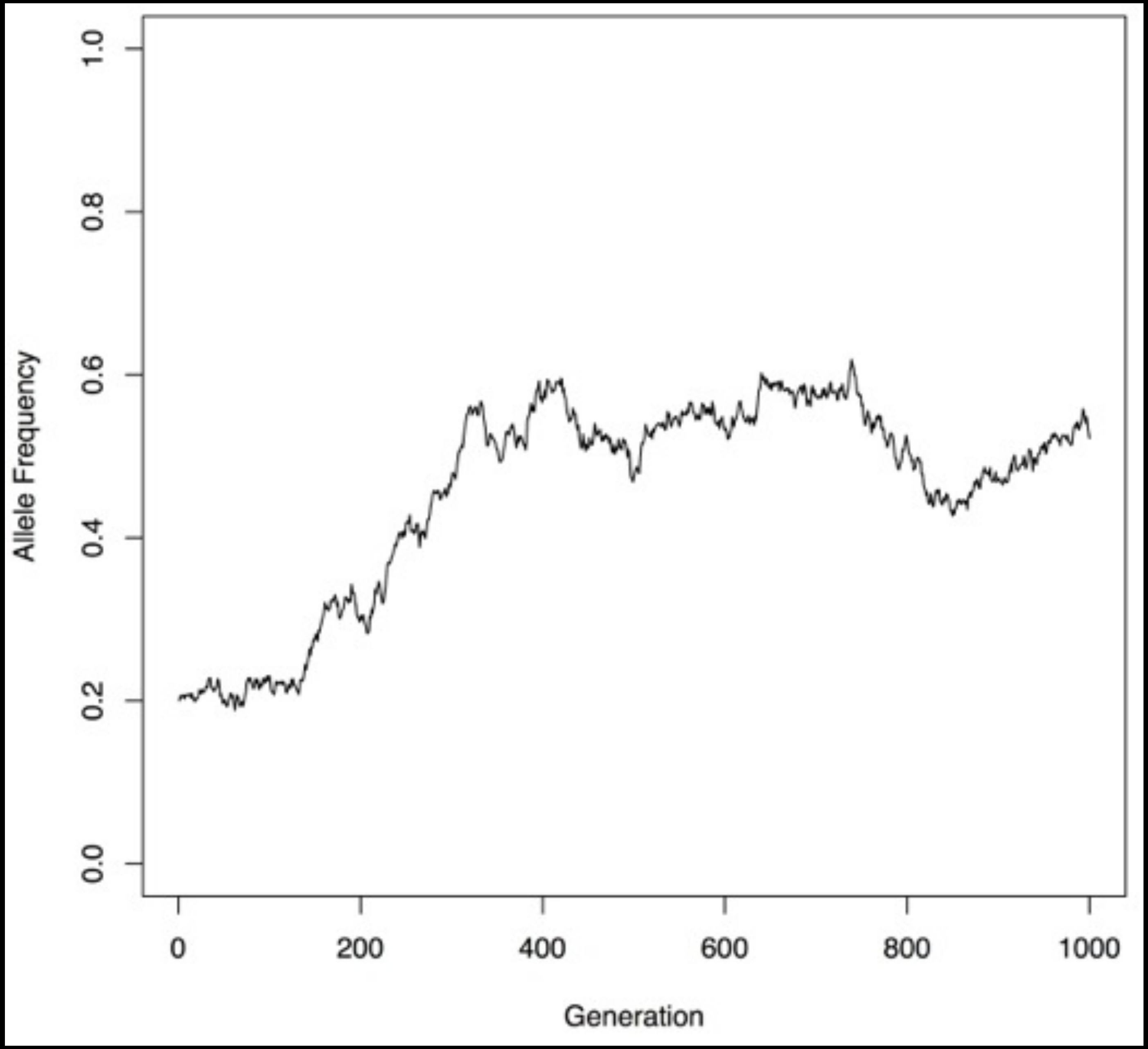
- Stochastic processes that transition through states
- Transition matrix parameterizes state changes
- Memoryless: Next step depends only on previous step(s) and not the whole history

$$P(Xt+1 \mid X1,X2,\ldots,Xt)=P(Xt+1 \mid Xt)$$ First order

- Model of serial dependence:
  - Classic application: stock market
  - Markov chain monte carlo algorithms estimate using simulations with temporal dependence
  - Bioinformatics applications: dependence along a sequence or flow on a network

Allele frequency over time

- Wright-Fisher model: Fixed population of 2N chromosomes, binomial trials for next generation
- The number of derived alleles in generation t is X_t

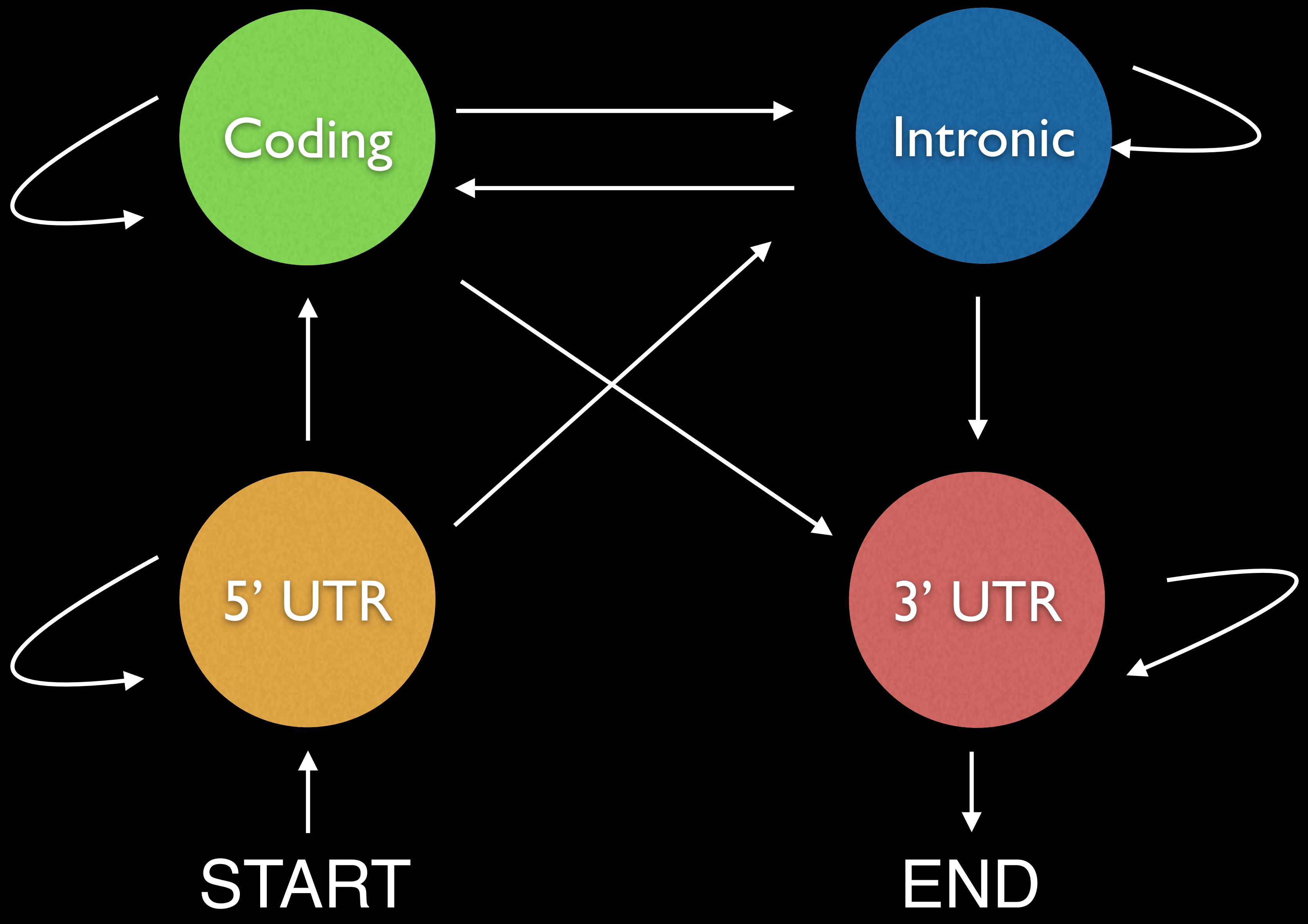


This is a Markov process

What are the states?

What is the transition matrix?

Gene Annotation



- Model of probability of being in each state along a genic region of DNA. Step one nucleotide at a time.
- Each arrow is associated with a transition probability.

OrthoMCL: Markov Clustering

Ortholog detection is a clustering problem.

DISTANCE MATRIX:

1. All vs. all BLASTP of protein sequences
2. Reciprocal best hits within each genome are candidate paralog pairs
3. Reciprocal best hits between genomes are candidate ortholog pairs
4. Filter: E-value<1e-5, % match length > 50%
5. Similarity = normalized -log10(BLASTP p-value)

OrthoMCL: Li et al. Genome Research (2003); MCL: Enright et al. NAR (2002)

OrthoMCL: Markov Clustering

MARKOV CLUSTERING:

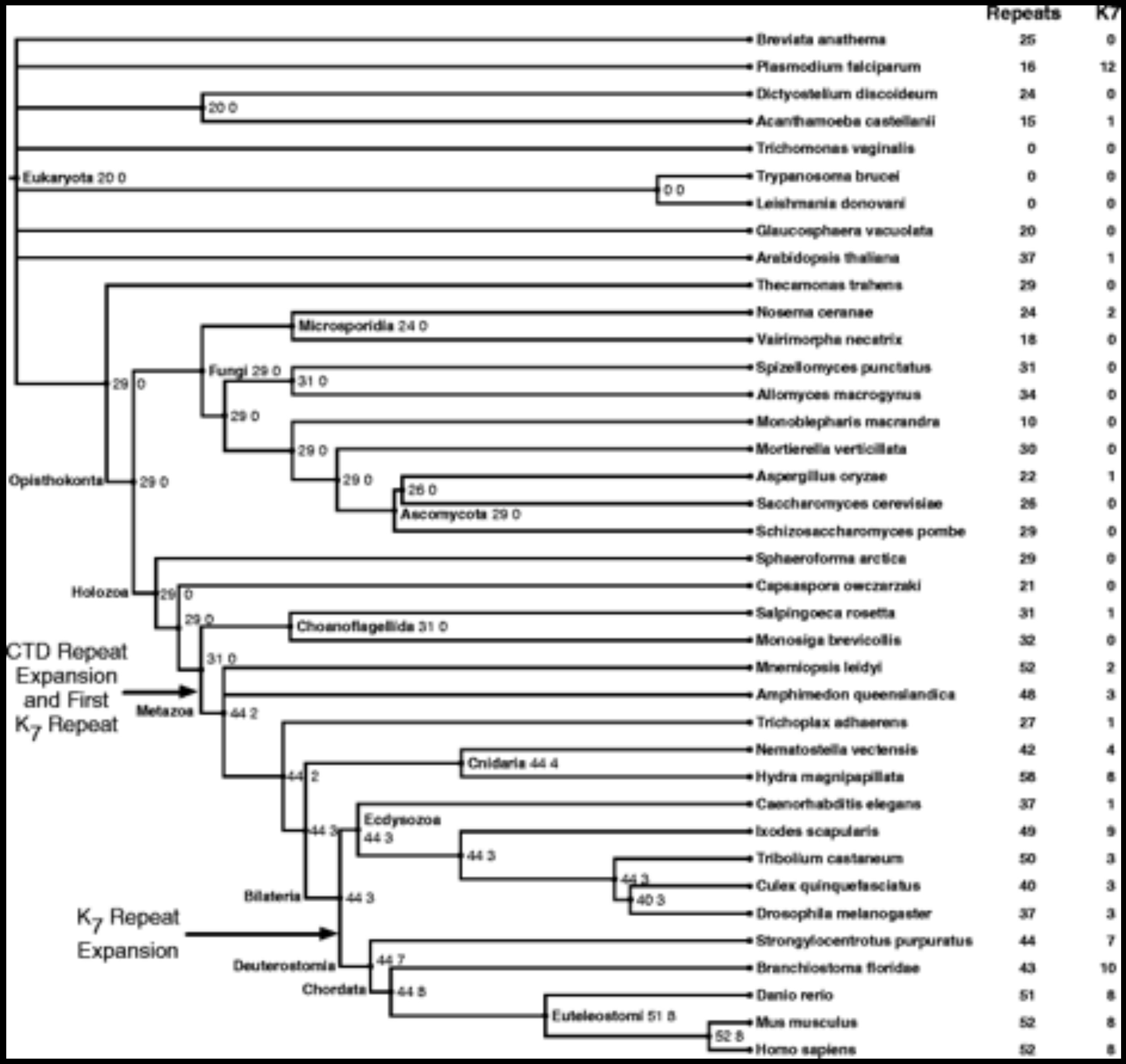
1. Protein similarity matrix is a weighted network.
2. Simulates random walks on the network graph.
3. Defines protein families as clusters of nodes that are frequently visited from each other, whereas nodes in different clusters are not.
4. Inflation is a tuning parameter that determines the length of the random walks.
5. Algorithm is just simple operations on the similarity matrix (expansion & inflation).

OrthoMCL: Li et al. Genome Research (2003); MCL: Enright et al. NAR (2002)

Continuous Time Markov Models

Birth-Death Processes

- Events are gains (births) and losses (deaths)
- A Poisson process where the probability of an event depends on the number of events that already occurred
- This makes it a continuous time Markov process



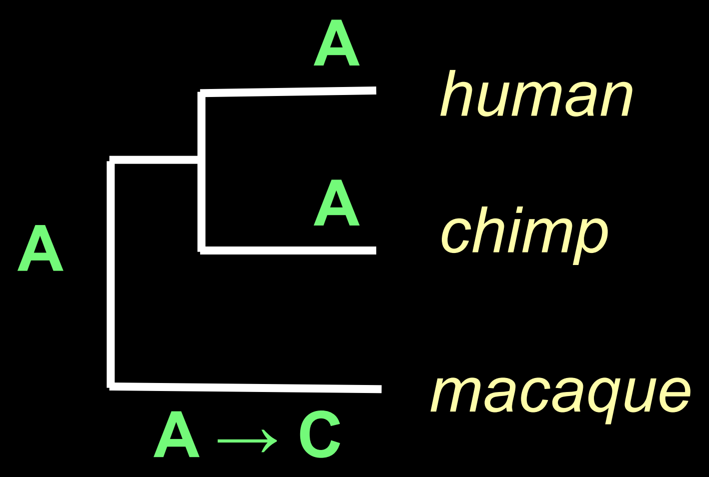
	Repeats	K7
Breviata anathema	25	0
Plasmodium falciparum	16	12
Dictyostelium discoideum	24	0
Acanthamoeba castellanii	15	1
Trichomonas vaginalis	0	0
Trypanosoma brucei	0	0
Leishmania donovani	0	0
Glaucosphaera vacuolata	20	0
Arabidopsis thaliana	37	1
Thecamonas trahens	29	0
Nosema ceranae	24	2
Vairimorpha necatrix	18	0
Spizellomyces punctatus	31	0
Allomyces macrogynus	34	0
Monoblepharis macrandra	10	0
Mortierella verticillata	30	0
Aspergillus oryzae	22	1
Saccharomyces cerevisiae	26	0
Schizosaccharomyces pombe	29	0
Sphaeroforma arctica	29	0
Capsaspora owczarzaki	21	0
Salpingoeca rosetta	31	1
Monosiga brevicollis	32	0
Mnemiopsis leidyi	52	2
Amphimedon queenslandica	48	3
Trichoplax adhaerens	27	1
Nematostella vectensis	42	4
Hydra magnipapillata	58	6
Caenorhabditis elegans	37	1
Ixodes scapularis	49	9
Tribolium castaneum	50	3
Culex quinquefasciatus	40	3
Drosophila melanogaster	37	3
Strongylocentrotus purpuratus	44	7
Branchiostoma floridae	43	10
Danio rerio	51	8
Mus musculus	52	8
Homo sapiens	52	8

Example: gene family evolution by duplication and deletion

CAFE software package

Simonti et al. (2015) BMC Evol Bio
De Bie et al. (2006) Bioinformatics

Modeling Molecular Evolution



$$
\mathbf{Q} = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & -0.87 & 0.17 & 0.53 & 0.17 \\ C & 0.24 & -1.20 & 0.18 & 0.78 \\ G & 0.77 & 0.18 & -1.19 & 0.24 \\ T & 0.17 & 0.53 & 0.17 & -0.87 \end{array}
$$

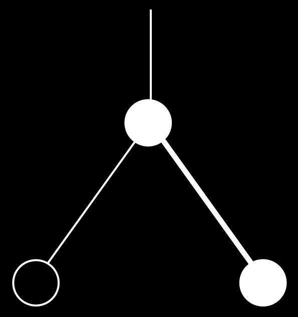
Rate Matrix

$$
\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) = (0.3, 0.2, 0.2, 0.3)
$$

Equilibrium Frequencies

Continuous-Time Markov Models

- Substitutions follow a continuous-time Markov model along each branch.

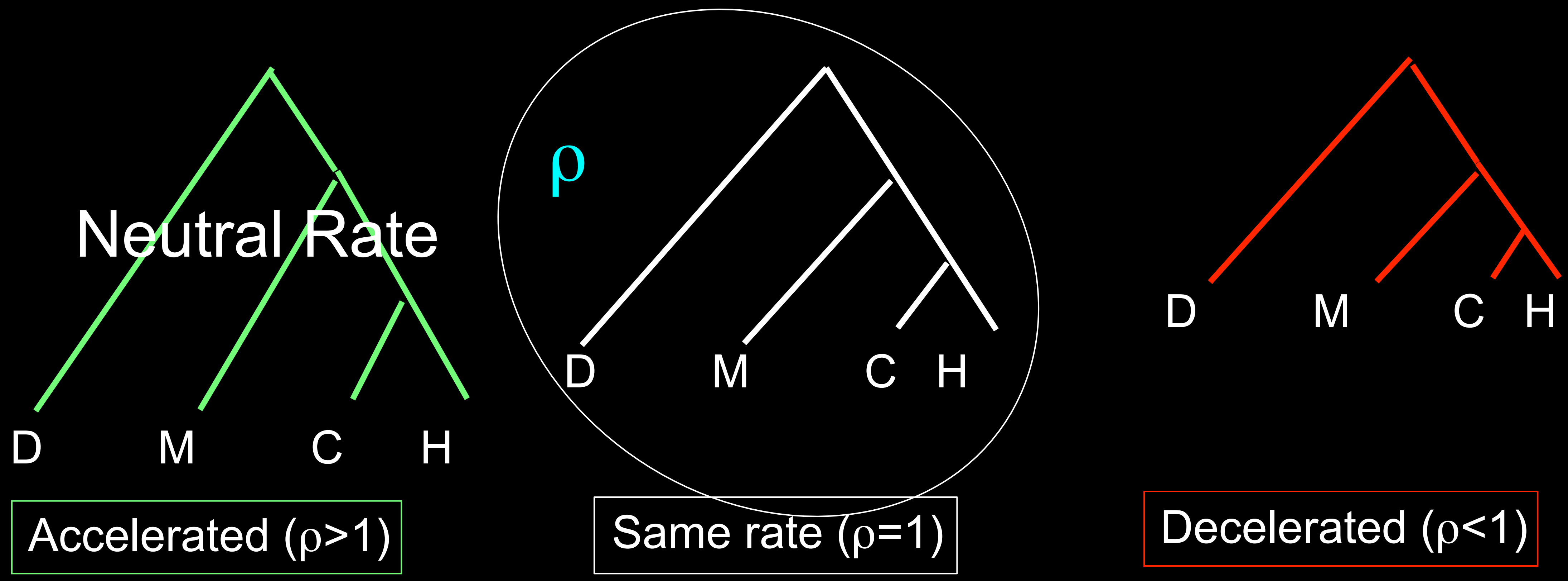


$$P_{\alpha\beta}(t) = [\exp(Qt)]_{\alpha\beta} =: \sum_n \frac{(Qt)^n}{n!}$$

$$\langle diag(Q)|\pi\rangle = -1$$

- P(t) is the probability of a change in time t.
- Q is the instantaneous rate of change, typically scaled so that t = subs/site.

Finding Fast and Slow Regions



phyloP: Compare models with statistical tests (e.g., LRT)

H0: ρ = 1

CON: ρ < 1

ACC: ρ > 1

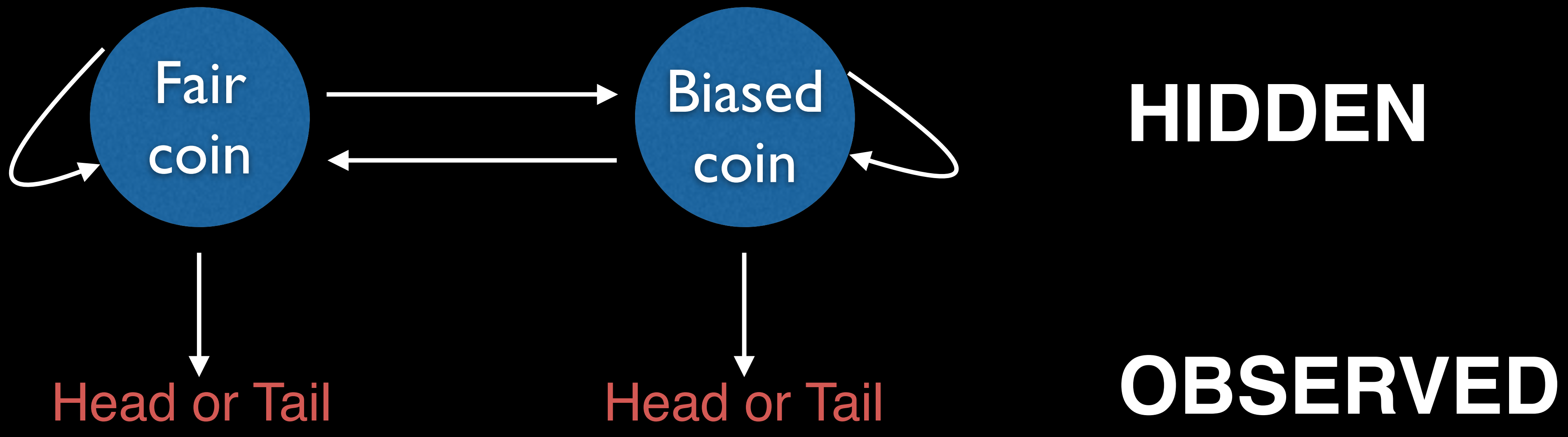
PhastCons: Use these models with HMMs to annotate conserved regions

Pollard et al. (2009) Genome Research
Siepel et al. (2005) Genome Research

Hidden Markov Models

Hidden Markov Models (HMMs)

- States emit data with different probabilities
- Sequence of states is not observed
- Only observe the emitted data



Fair coin

Biased coin

HIDDEN

Head or Tail

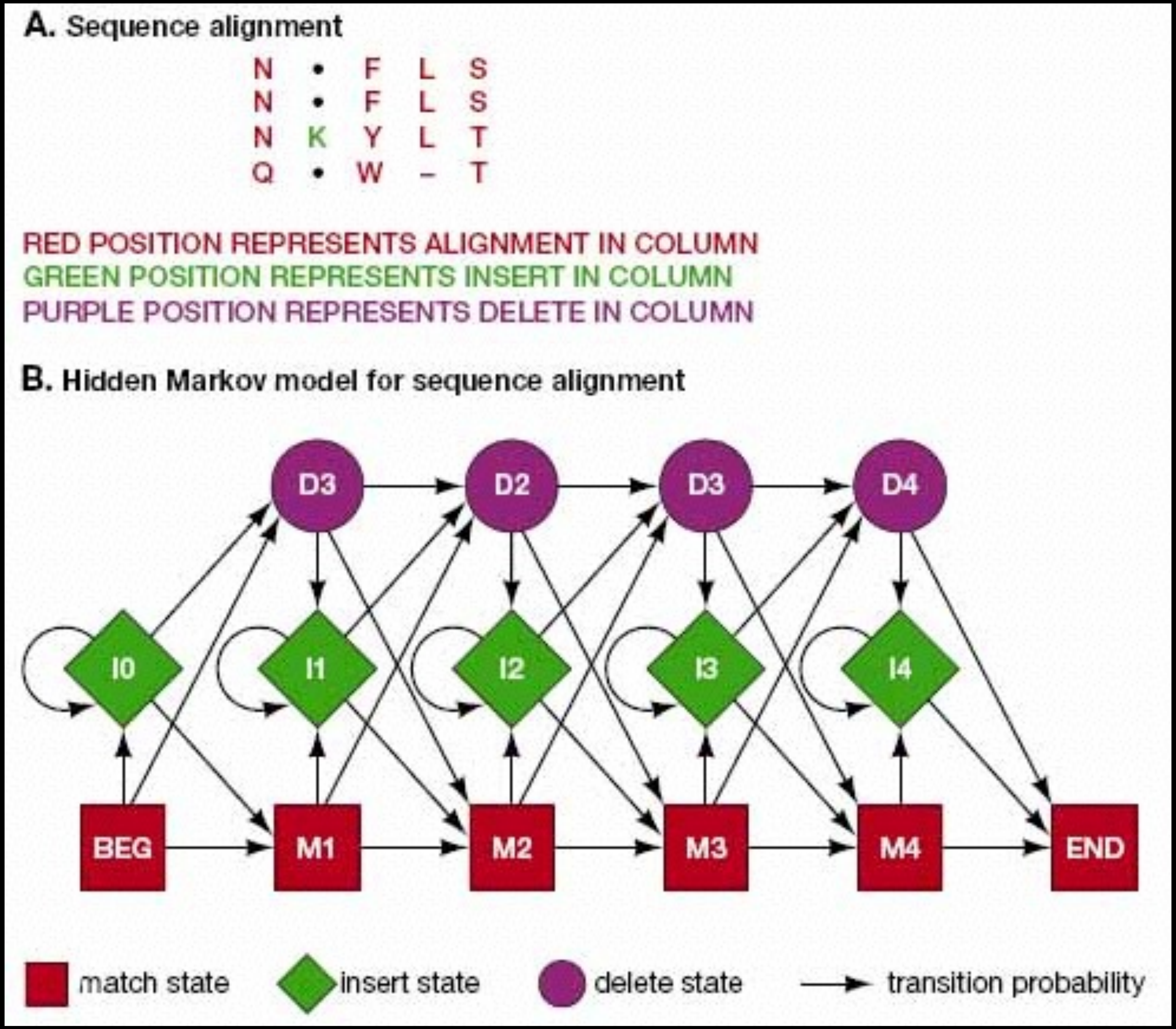
Head or Tail

OBSERVED

- Parameters: transition matrix plus emission probabilities for each state
- Goal: Estimate model parameters and the hidden sequence of states

Sequence alignment with profile HMMs

- Probabilistic models of protein family evolution
- Used for search, alignment, summary



A. Sequence alignment

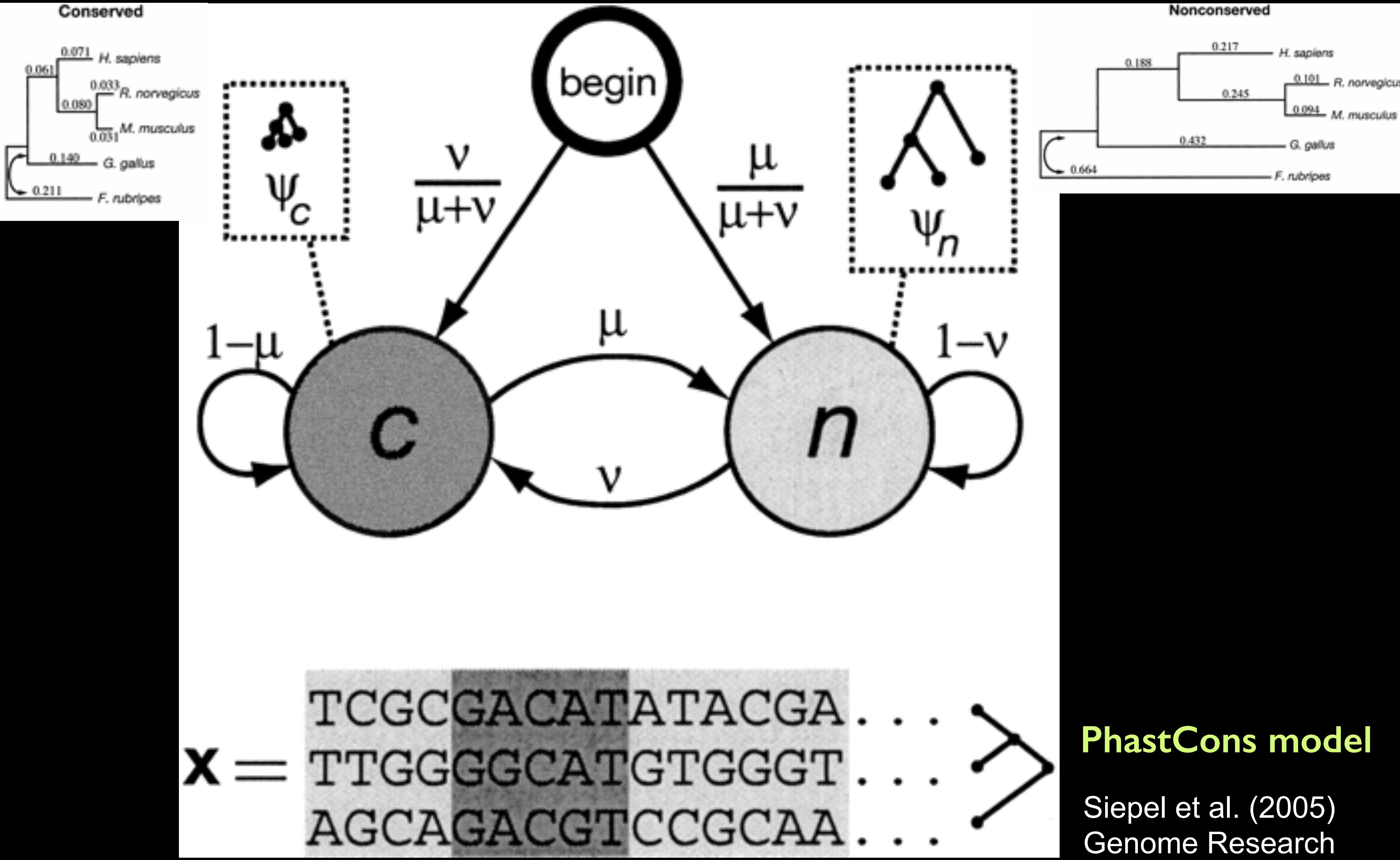
```
N • F L S
N • F L S
N K Y L T
Q • W - T
```

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

B. Hidden Markov model for sequence alignment

Sean Eddy & colleagues
Figure: David Mount

Phylogenetic Hidden Markov Model



PhastCons model

Siepel et al. (2005)
Genome Research

Stochastic Context Free Grammars

Stochastic Context Free Grammars

- RNA genes form 3D structures with pairs of nucleotides that are not next to each other in the sequence physically interacting.
- Chompsky and other computational linguists developed grammar theory.
- SCFGs enable the modeling of palindromes.
- Emit pairs of sequences from the inside out, e.g.,

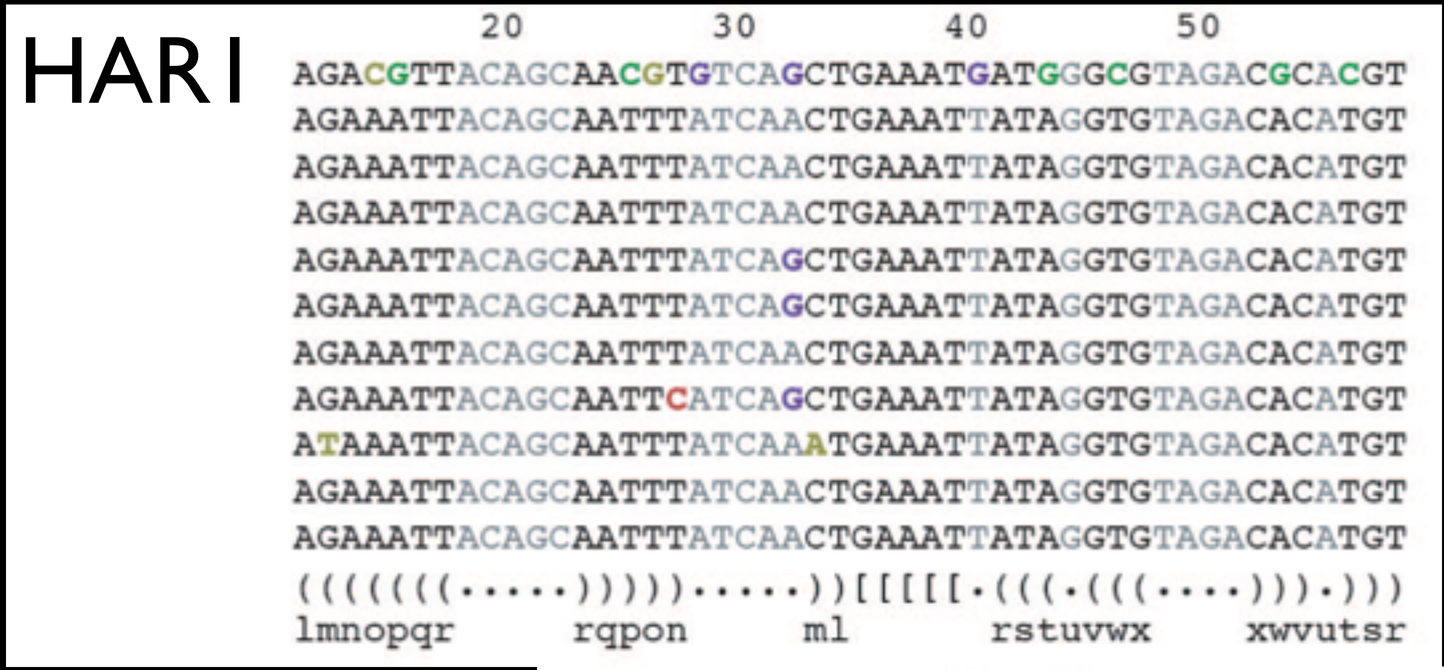
  grammar: S ⟶ aSa | bSb | aa | bb

  path: S, aSa, aSa, bSb, aa

  produces the sequence: aabaabaa
- Enables probabilistic models of RNA gene evolution

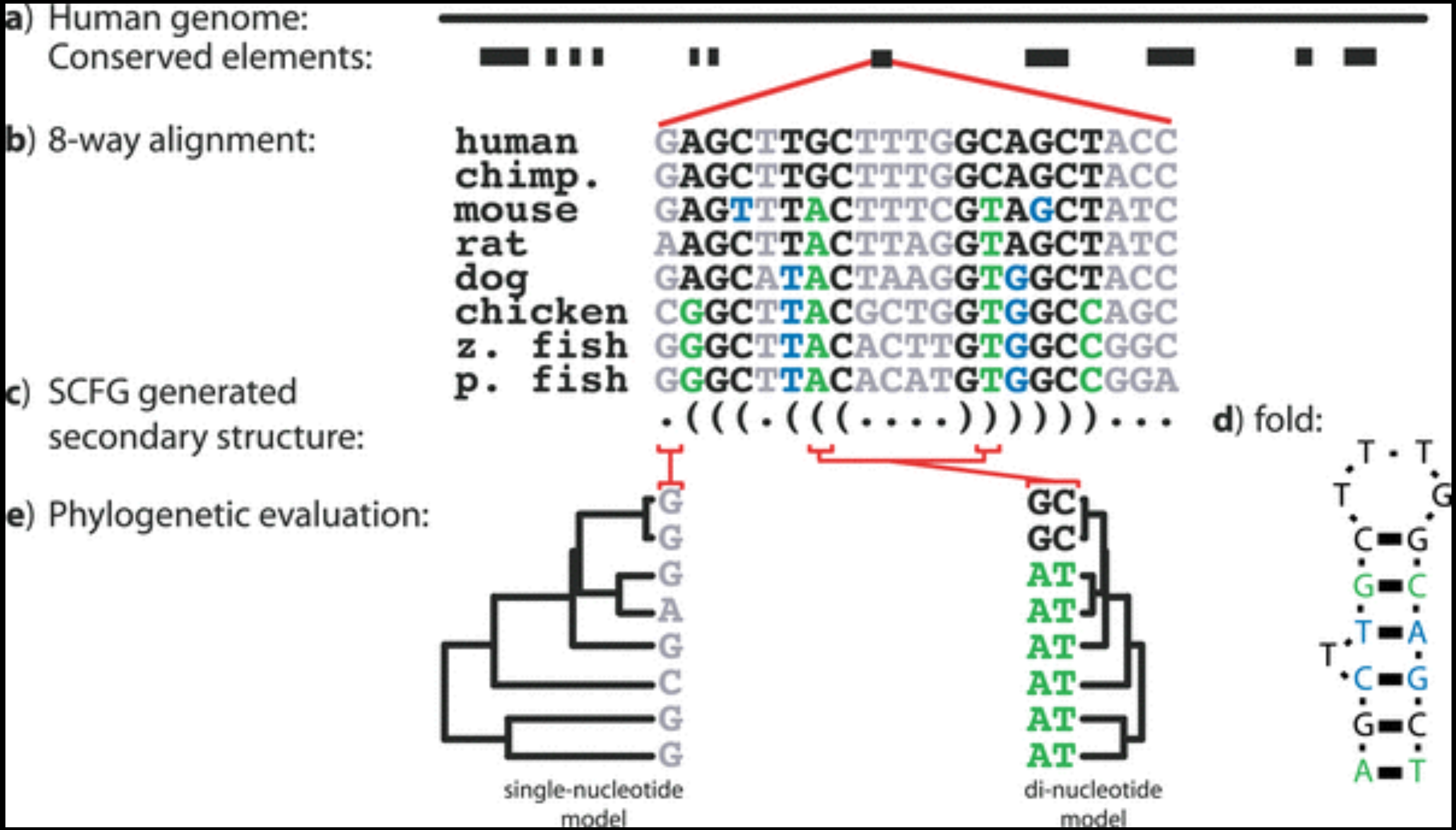
EvoFold: RNA Gene Prediction

- Conserved stems
- Structure preserving substitutions
- Stochastic context free grammar



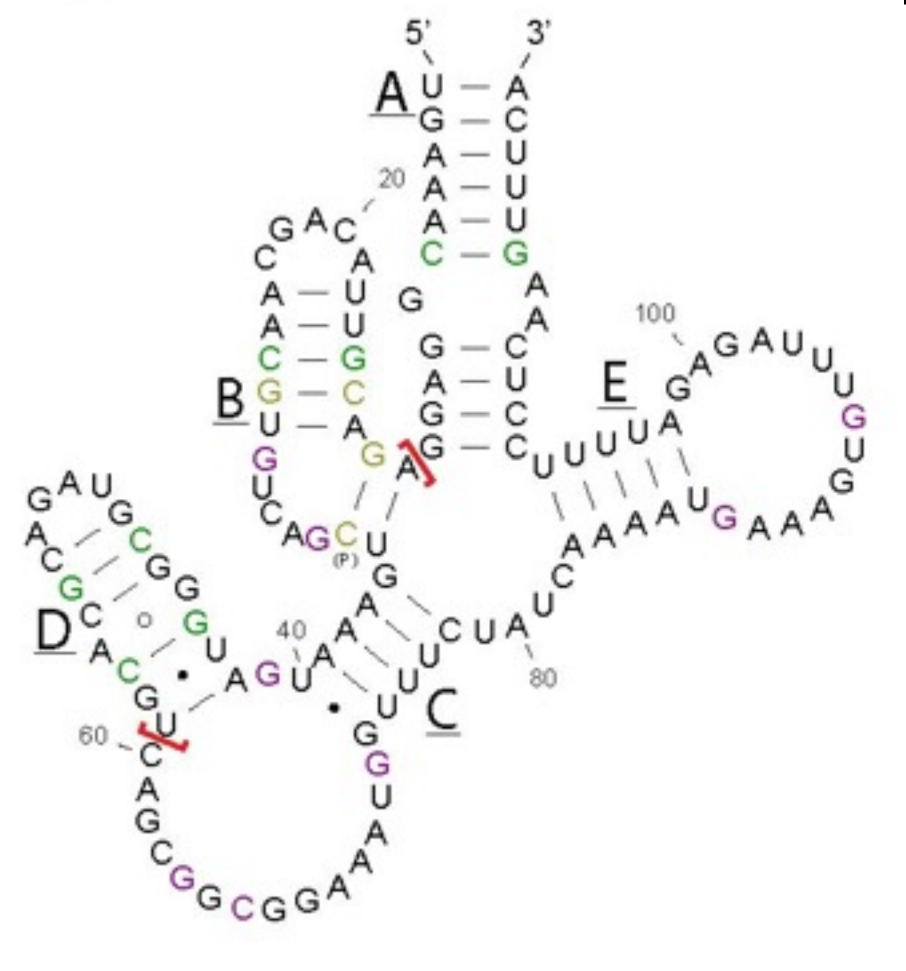
HAR1

```
            20        30        40        50
AGACGTTACAGCAACGTGTCAGCTGAAATGATGGGCGTAGACGCACGT
AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT
ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT
(((((((.....))))).....))[[[[[.(((.(((....))).)))
lmnopqr     rqpon     ml      rstuvwx      xwvutsr
```



```
human    GAGCTTGCTTTGGCAGCTACC
chimp.   GAGCTTGCTTTGGCAGCTACC
mouse    GAGTTTACTTTCGTAGCTATC
rat      AAGCTTACTTAGGTAGCTATC
dog      GAGCATACTAAGGTGGCTACC
chicken  CGGCTTACGCTGGTGGCCAGC
z. fish  GGGCTTACACTTGTGGCCGGC
p. fish  GGGCTTACACATGTGGCCGGA
         .(((.(((....))))))...
```

```
G  GC
G  GC
G  AT
A  AT
G  AT
C  AT
G  AT
G  AT
```



Pedersen et al. PLoS Comp Bio 2006; Pollard et al., Nature 2006

Additional Slides

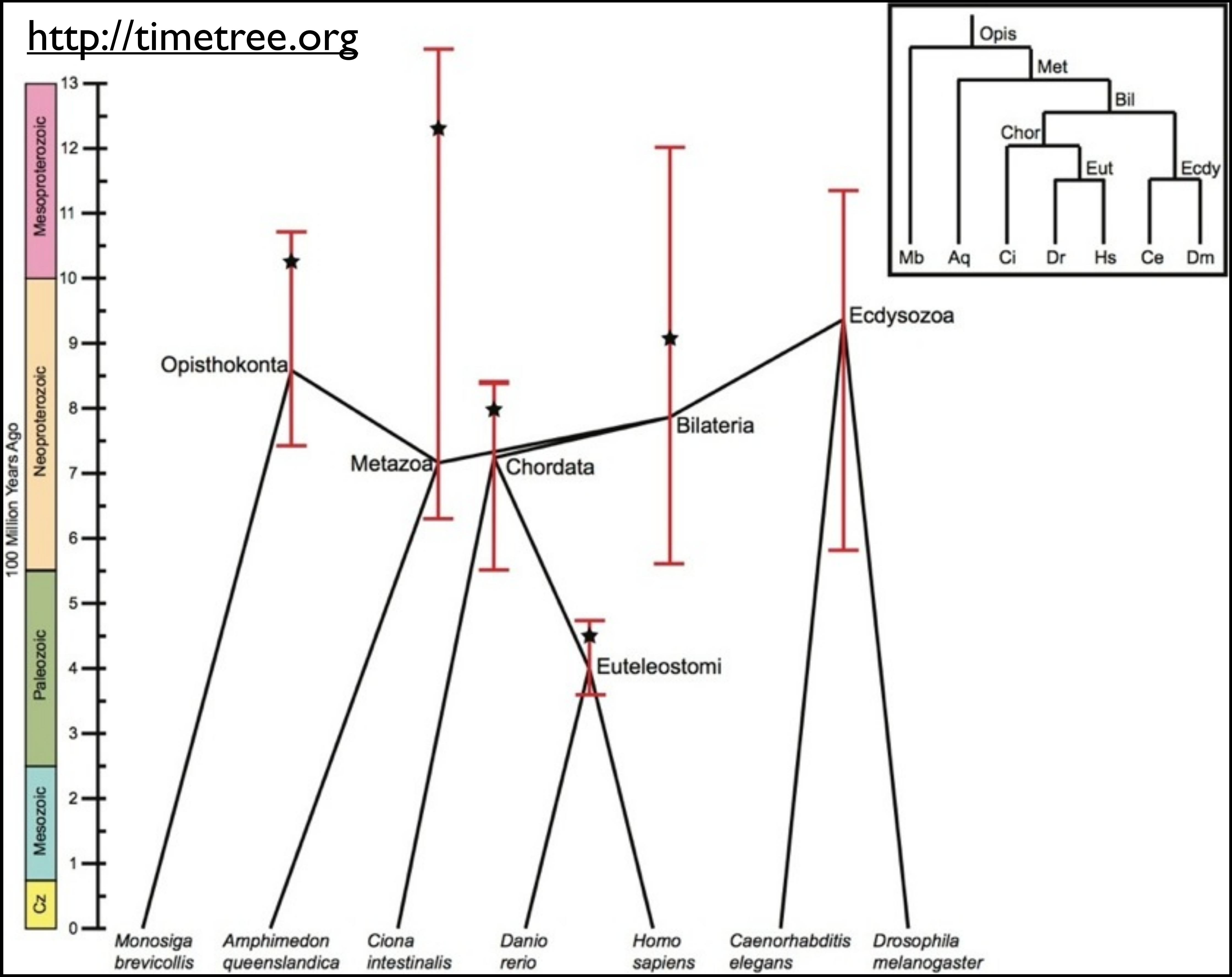
What is a gene family?

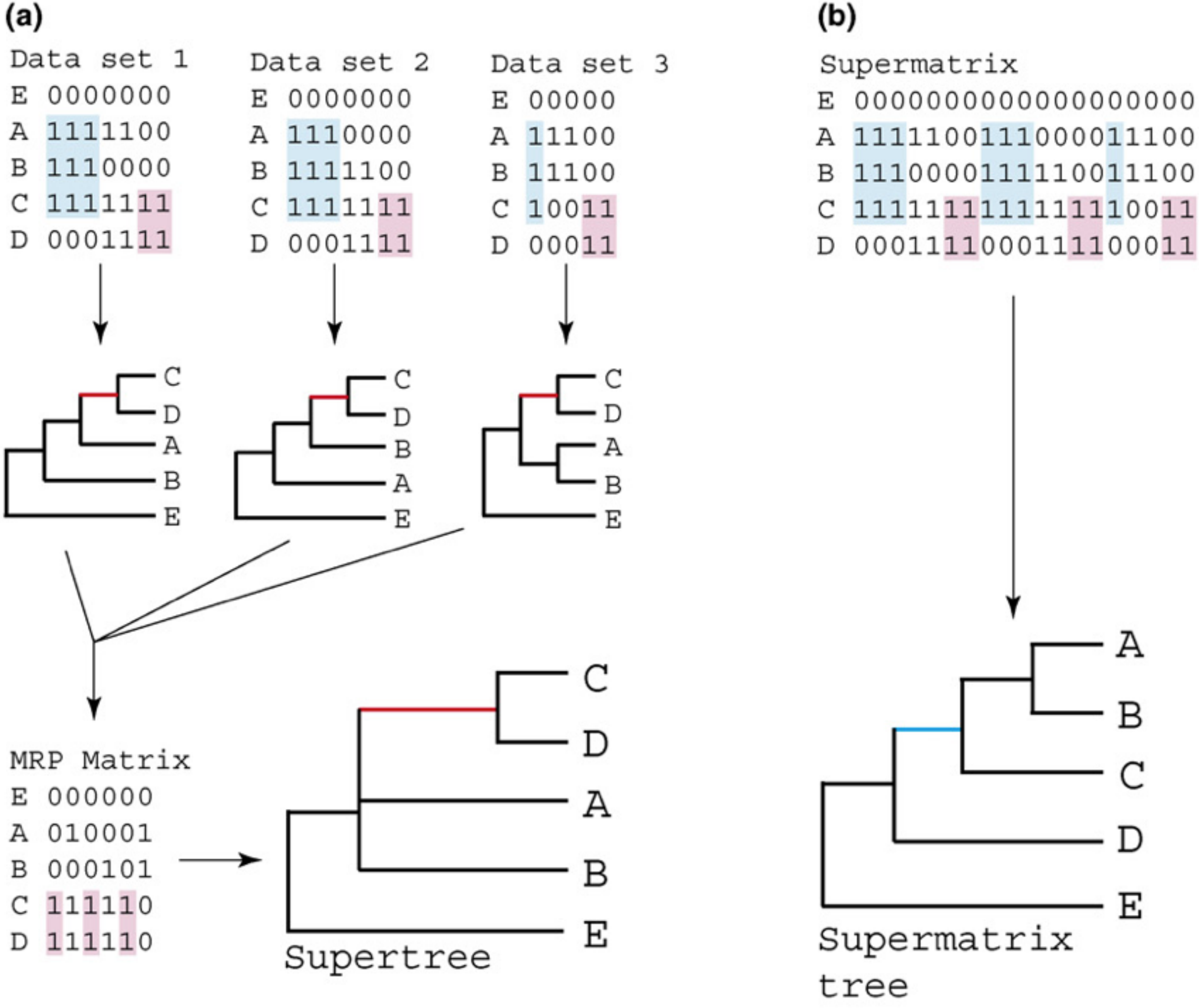
- A collection of DNA sequences encoding gene products (protein/RNA) with shared function and/or evolutionary history
    - Shared history is known as homology
    - Homology does not imply shared function
    - Shared function does not require homology
- Typically from multiple species
- Can include more than one member per species (paralogs vs. orthologs)

Building Trees

- What sequence to use?
  - species tree: neutral sites (e.g, 4-fold degenerate sites or ancestral repeats)
  - gene tree: usually all sites
- Many choices of algorithm
- How to combine data from multiple genes?
  - combine separate trees: super tree
  - combine data, build one tree: super matrix

Uncertainty in Trees





(a)

Data set 1
E 0000000
A 1111100
B 1110000
C 1111111
D 0001111

Data set 2
E 0000000
A 1110000
B 1111100
C 1111111
D 0001111

Data set 3
E 00000
A 11100
B 11100
C 10011
D 00011

MRP Matrix
E 000000
A 010001
B 000101
C 111110
D 111110

Supertree

(b)

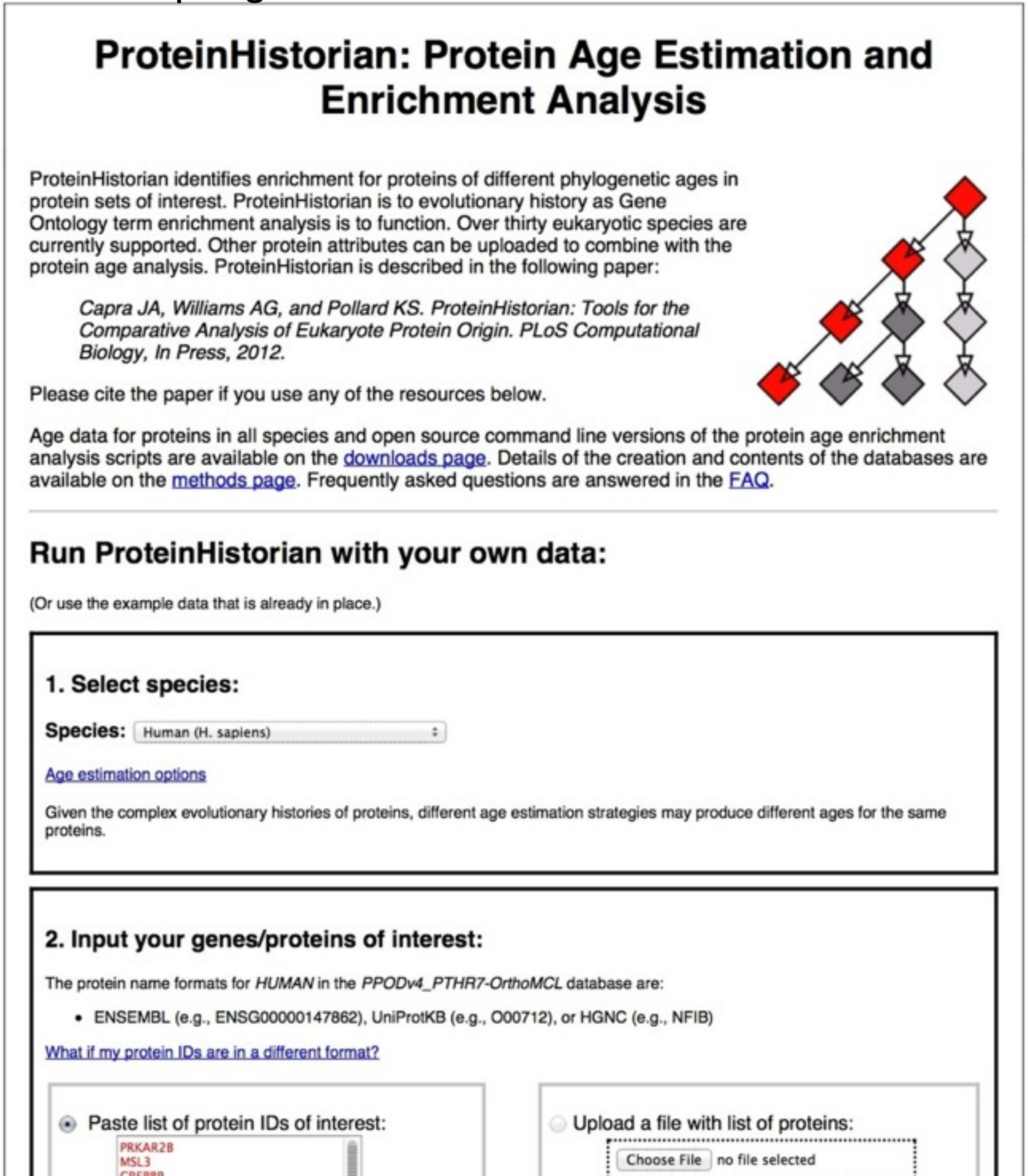
Supermatrix
E 0000000000000000000
A 1111100111000011100
B 1110000111110011100
C 1111111111111110011
D 0001111000111100011

Supermatrix tree

TRENDS in Ecology & Evolution

Queiroz & Gatesy, 2006

http://lighthouse.ucsf.edu/ProteinHistorian/



ProteinHistorian: Protein Age Estimation and Enrichment Analysis

ProteinHistorian identifies enrichment for proteins of different phylogenetic ages in protein sets of interest. ProteinHistorian is to evolutionary history as Gene Ontology term enrichment analysis is to function. Over thirty eukaryotic species are currently supported. Other protein attributes can be uploaded to combine with the protein age analysis. ProteinHistorian is described in the following paper:

> *Capra JA, Williams AG, and Pollard KS. ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. PLoS Computational Biology, In Press, 2012.*

Please cite the paper if you use any of the resources below.

Age data for proteins in all species and open source command line versions of the protein age enrichment analysis scripts are available on the downloads page. Details of the creation and contents of the databases are available on the methods page. Frequently asked questions are answered in the FAQ.

Run ProteinHistorian with your own data:

(Or use the example data that is already in place.)

1. Select species:

Species: Human (H. sapiens)

Age estimation options

Given the complex evolutionary histories of proteins, different age estimation strategies may produce different ages for the same proteins.

2. Input your genes/proteins of interest:

The protein name formats for *HUMAN* in the *PPODv4_PTHR7-OrthoMCL* database are:

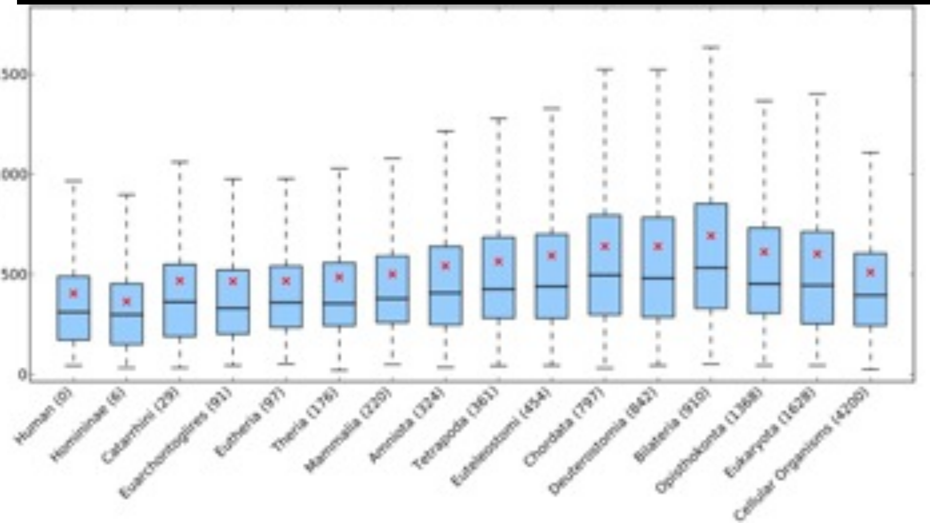
- ENSEMBL (e.g., ENSG00000147862), UniProtKB (e.g., O00712), or HGNC (e.g., NFIB)

What if my protein IDs are in a different format?

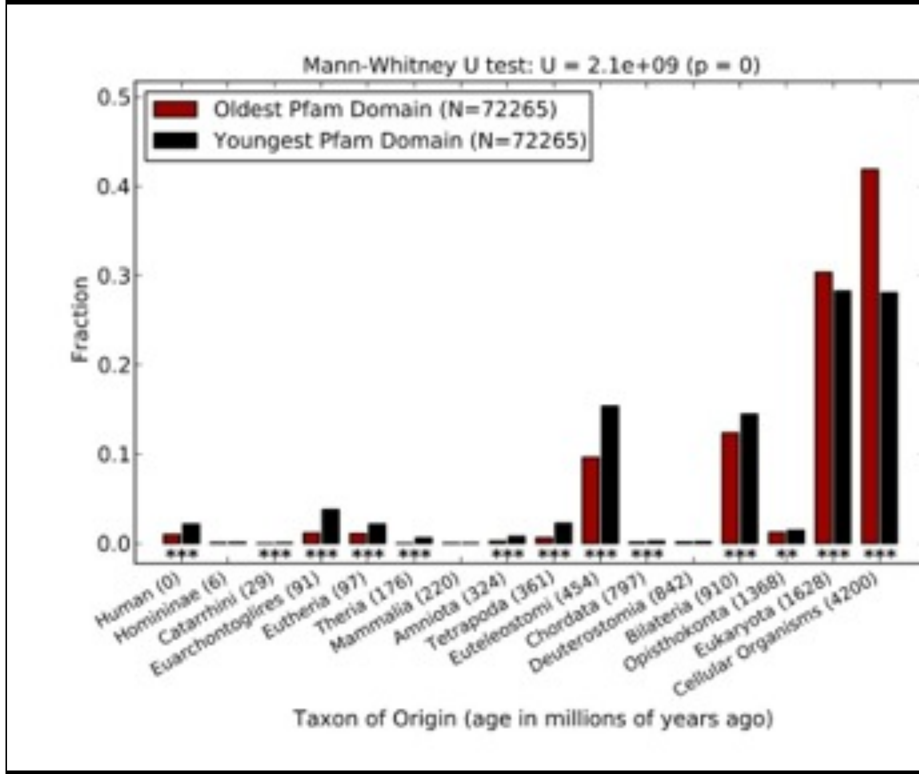
● Paste list of protein IDs of interest:

PRKAR2B
MSL3

○ Upload a file with list of proteins:

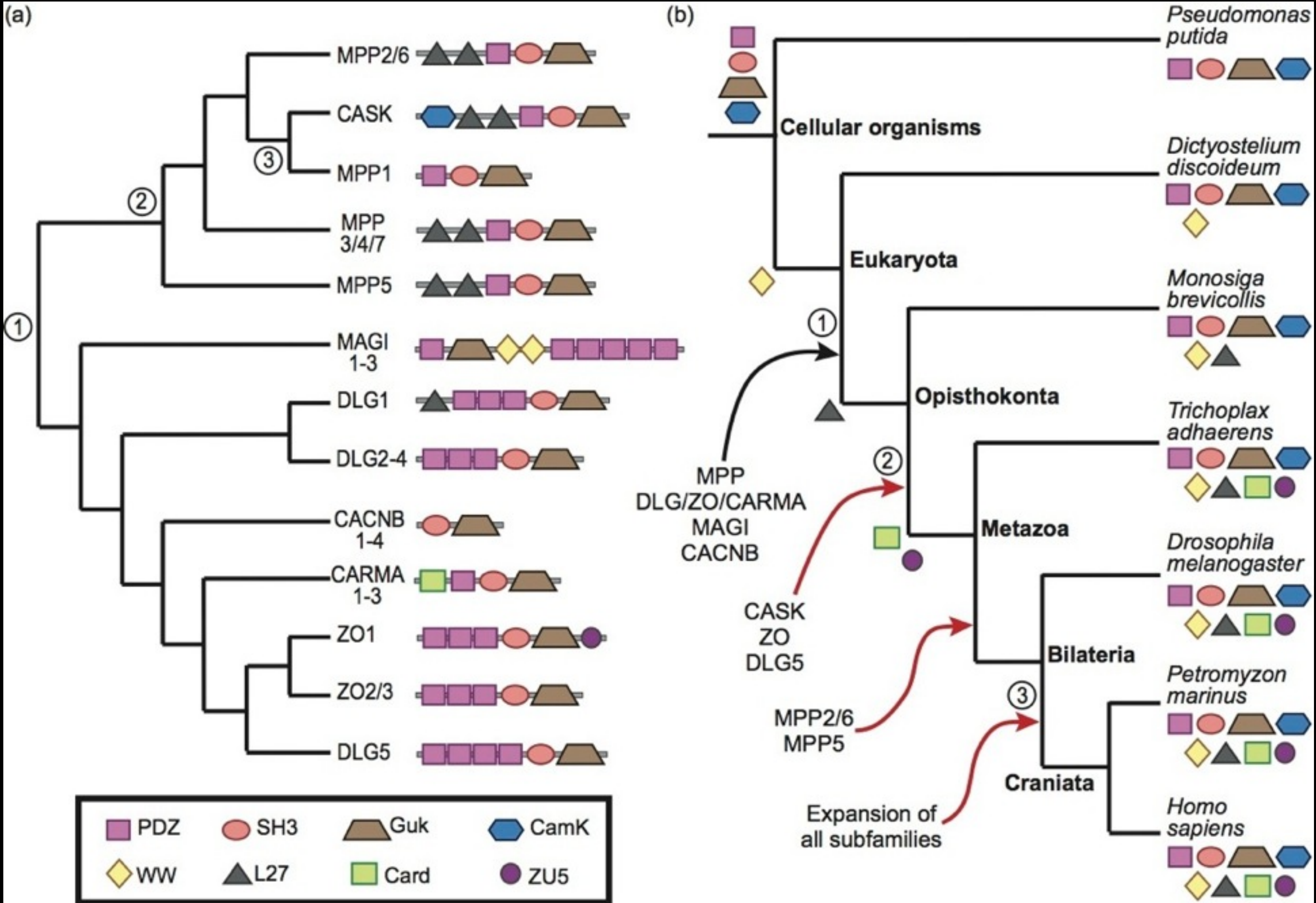


Age vs. length



Domain Age

MAGUK Superfamily History


(a)
MPP2/6
CASK
MPP1
MPP 3/4/7
MPP5
MAGI 1-3
DLG1
DLG2-4
CACNB 1-4
CARMA 1-3
ZO1
ZO2/3
DLG5
1
2
3
PDZ
SH3
Guk
CamK
WW
L27
Card
ZU5
(b)
Pseudomonas putida
Cellular organisms
Dictyostelium discoideum
Eukaryota
Monosiga brevicollis
Opisthokonta
Trichoplax adhaerens
Metazoa
Drosophila melanogaster
Bilateria
Petromyzon marinus
Craniata
Homo sapiens
1
2
3
MPP
DLG/ZO/CARMA
MAGI
CACNB
CASK
ZO
DLG5
MPP2/6
MPP5
Expansion of all subfamilies