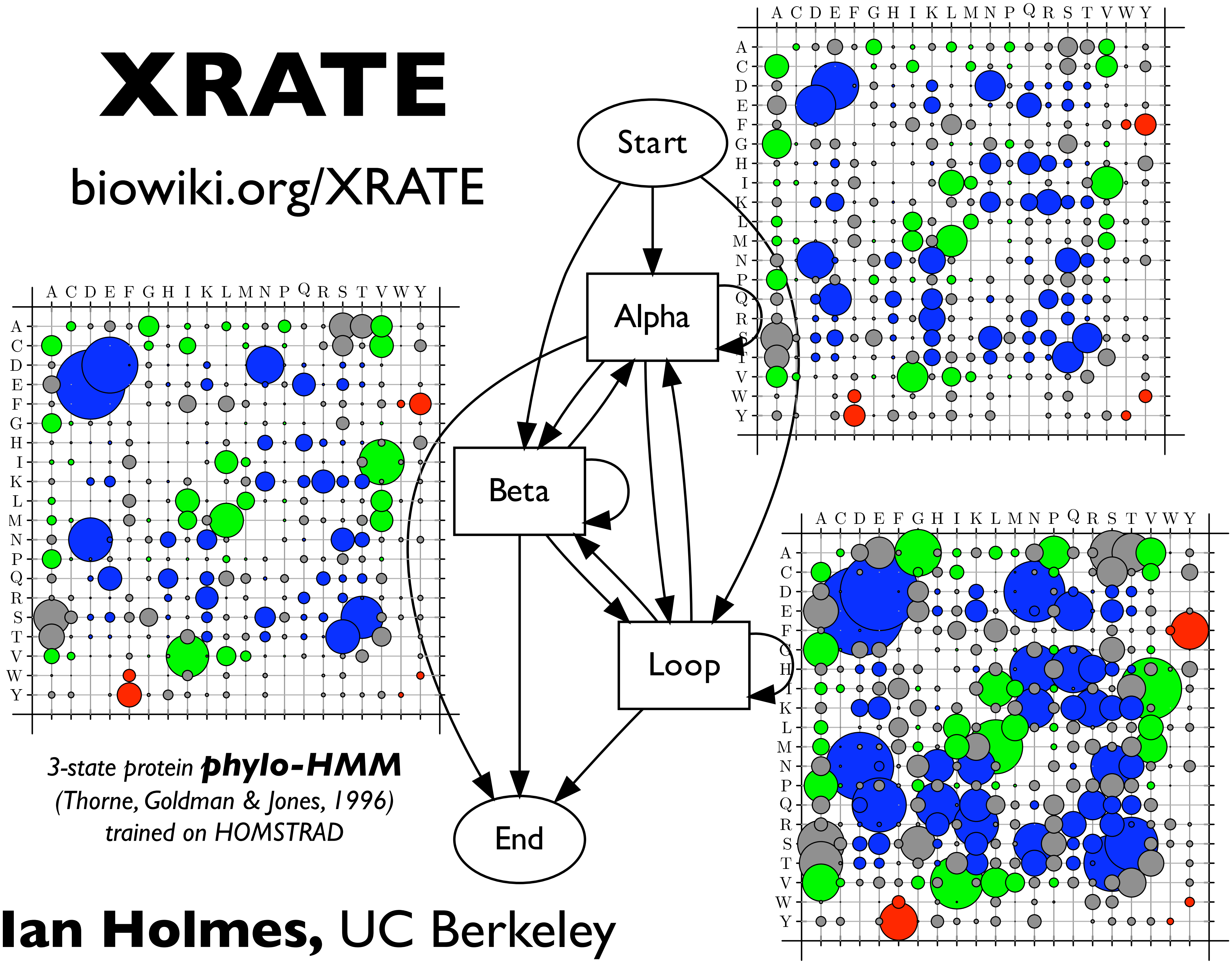
XRATE

biowiki.org/XRATE

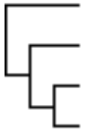

Start
Alpha
Beta
Loop
End
A C D E F G H I K L M N P Q R S T V W Y
A C D E F G H I K L M N P Q R S T V W Y

3-state protein ***phylo-HMM***
(Thorne, Goldman & Jones, 1996)
trained on HOMSTRAD

Ian Holmes, UC Berkeley

HMM emitting columns of amino acids

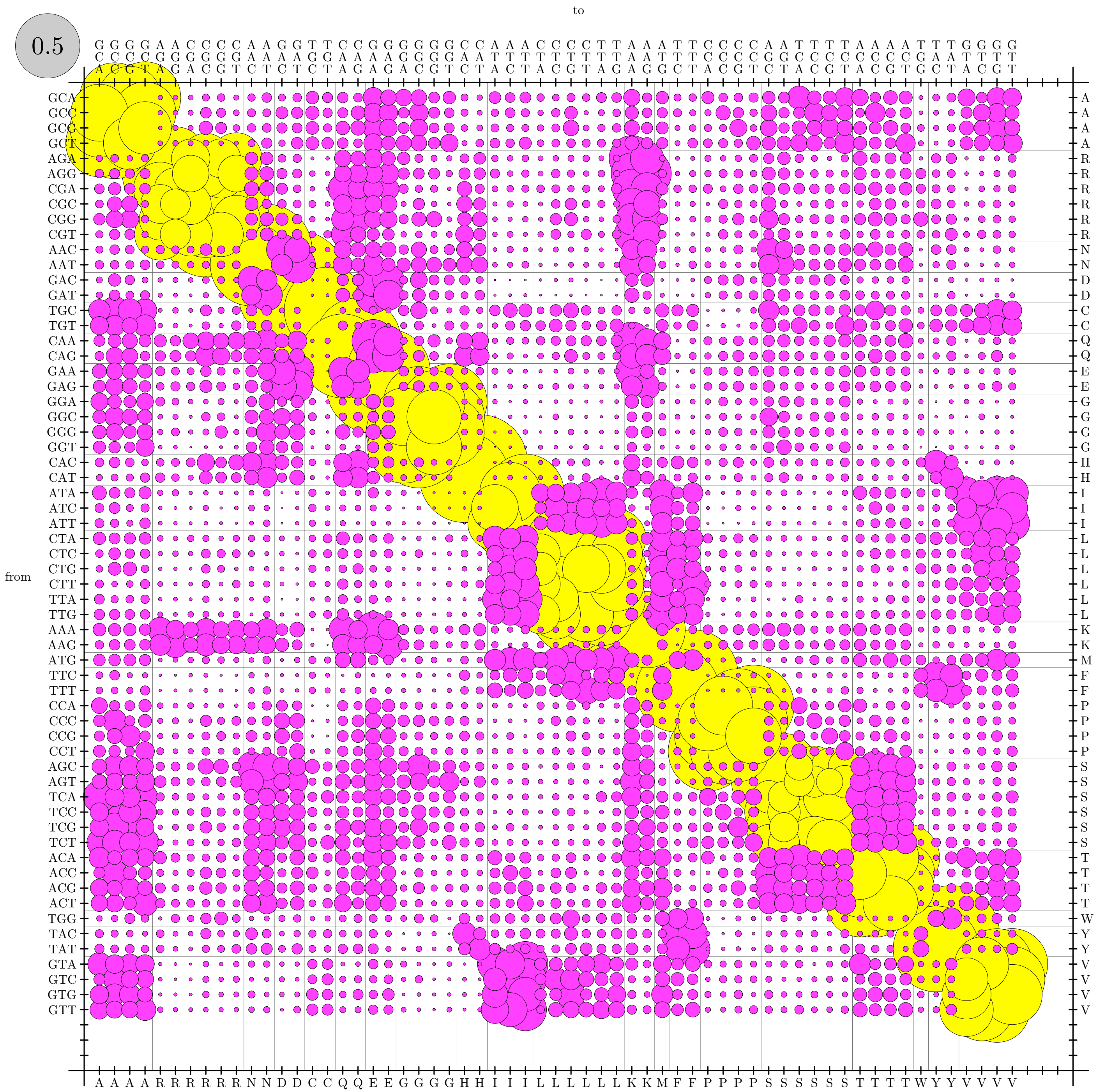
Ricin alignment;
colors denote secondary structure



*Finite continuous-time Markov chain +
tree + HMM = Felsenstein + Viterbi*

Thorne, Goldman, Jones; MBE, 1996

Codons



Empirical 61x61 matrix trained on PANDIT (7738 protein families)

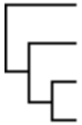
Kosiol, Holmes & **Goldman,** MBE 2007

Parametric d_N/d_S models correctly reproduce PAML

Heger, Ponting & Holmes, *in prep*

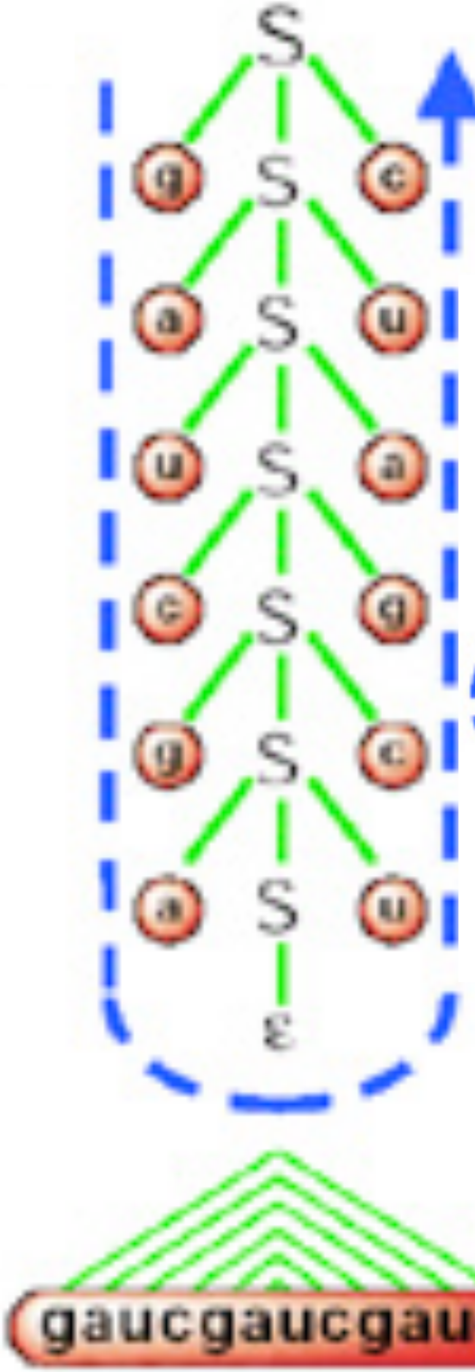
HMM emitting exons and introns

Intron flanked by two exons (simple example)

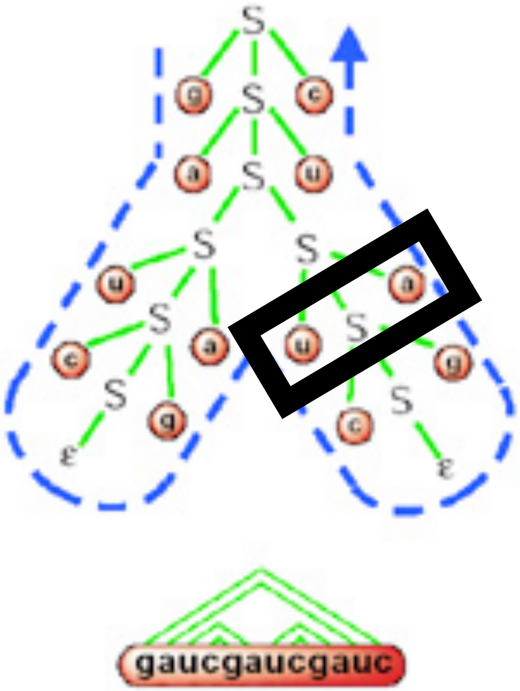


e.g. Pedersen & Hein, Bioinf., 2003
Siepel & Haussler, 2004

RNA



Context-free RNA parse tree
Searls, 2002



Phylo-SCFGs:

PFOLD

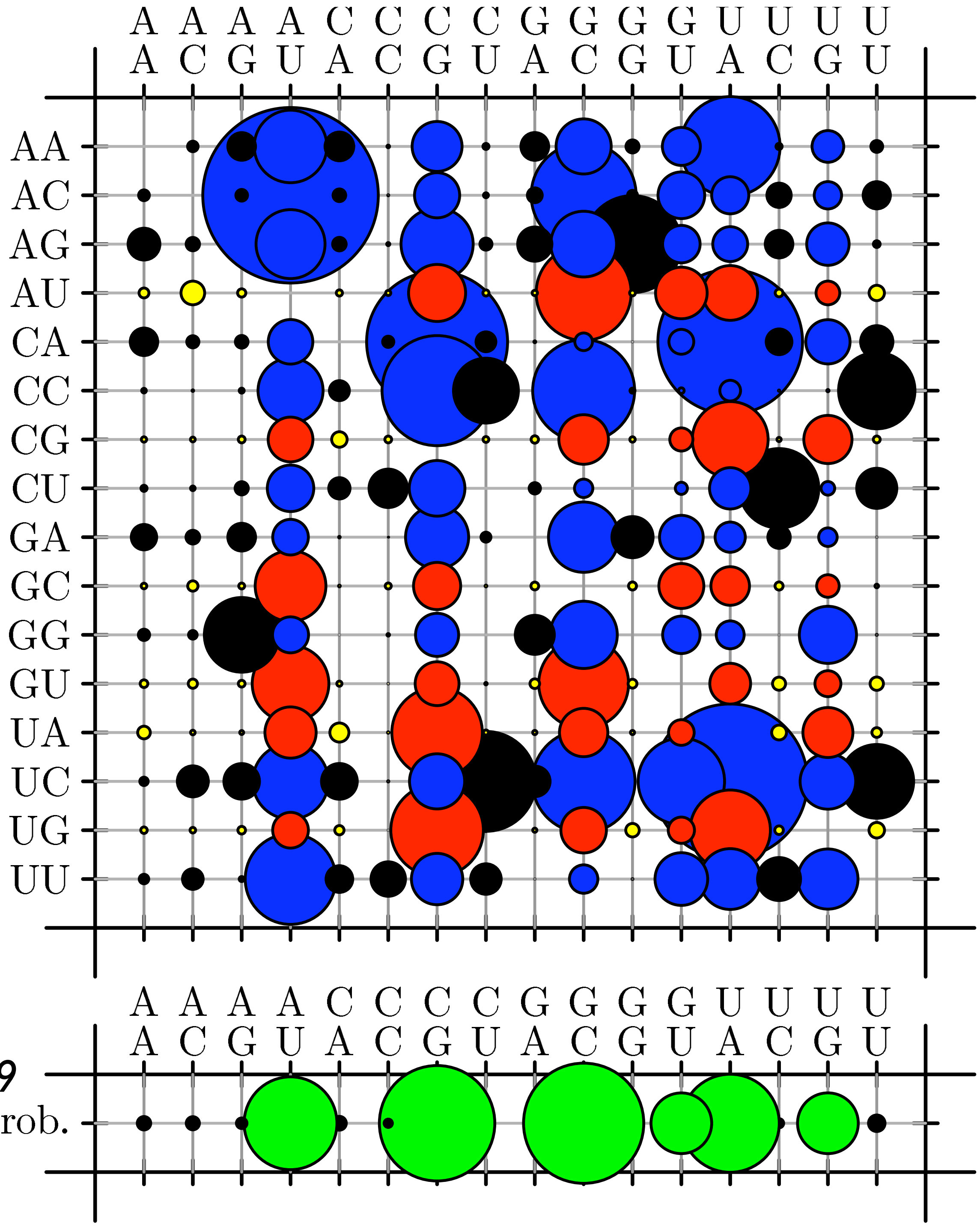
Knudsen & Hein, 1999

EVOFOLD

Pedersen et al, 2004

ClosingBp

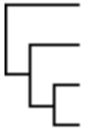
Bradley et al, 2008



Trained on rRNA (Dowell & Eddy, 2006)

SCFG emitting basepaired columns

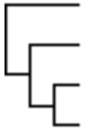
Hammerhead ribozyme



PFOLD
Knudsen & Hein, Bioinformatics, 1999

SCFG emitting basepaired columns

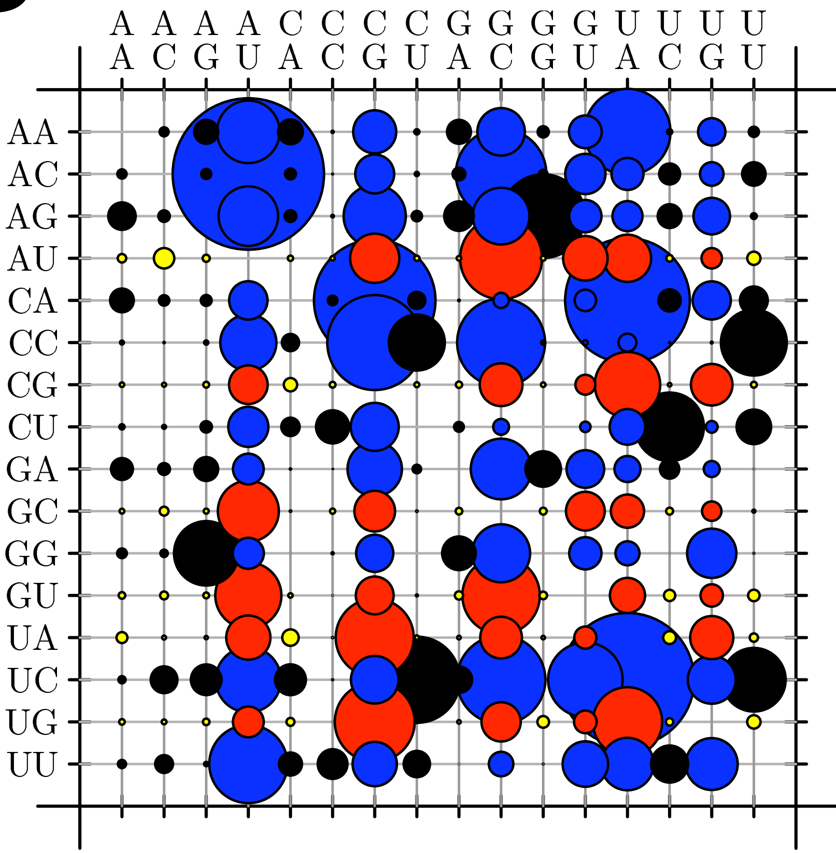
Hammerhead ribozyme

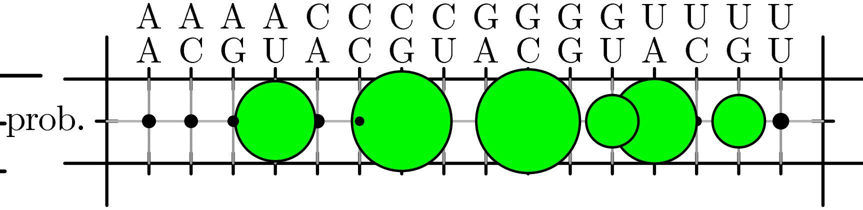


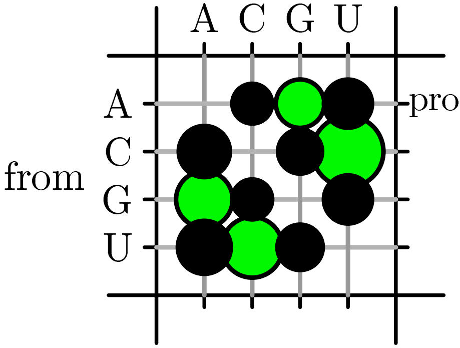
PFOLD
Knudsen & Hein, Bioinformatics, 1999

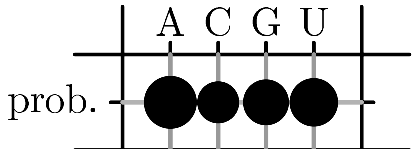
PFOLD phylo-grammar

```
S   →  L (0.131)
     | B (0.869)

F   →  lnuc F* rnuc
F*  →  F (0.788)
     | B (0.212)

L   →  F (0.105)
     | U (0.895)

B   →  L S

U   →  nuc U*
U*  →  End
```









PFOLD in xrate format

```
S   → L (0.131)
    | B (0.869)

F   → lnuc F* rnuc
F*  → F (0.788)
    | B (0.212)

L   → F (0.105)
    | U (0.895)

B   → L S

U   → nuc U*
U*  → End
```

```
;; state S: the initial state. Goes to L or B
(transform (from (S)) (to (L)) (prob 0.131))
(transform (from (S)) (to (B)) (prob 0.869))

;; state F: emits a covarying base pair
(transform (from (F)) (to (LNUC F* RNUC)))
(transform (from (F*)) (to (F)) (prob 0.788))
(transform (from (F*)) (to (B)) (prob 0.212))

;; state L: goes to U (unpaired) or F (paired)
(transform (from (L)) (to (F)) (prob 0.105))
(transform (from (L)) (to (U)) (prob 0.895))

;; state B: generates a bifurcation
(transform (from (B)) (to (L S)))

;; state U: emits a single unpaired base
(transform (from (U)) (to (NUC U*)))
(transform (from (U*)) (to ()) (prob 1))
```

Bioinformatics motivation

- **Analyze multiple sequence alignments**
  - **measure** evolutionary rates in various contexts
  - **annotate** ncRNAs, CDS's, motifs, pseudogenes, ...

- **Develop versatile models, algorithms, tools**
  - *"phylo-grammars"*
  - Expectation Maximization, "phylo-EM"
  - Extensible: gap characters, lineage-specific rates, arbitrary grammars/models, parametric (c.f. HyPhy), ...

Phylo-Grammar literature

- Felsenstein & Churchill. 1996.
  Three-state phylo-HMM (rates .3, 2, 10).
- Thorne, Goldman & Jones. 1996.
  Protein secondary structure phylo-HMM.
- Knudsen & Hein. 1999.
  Phylo-SCFG for RNA structure prediction.
- Pedersen & Hein. 2003.
  Phylo-HMM for gene prediction.
- Siepel & Haussler. 2004.
  Dinucleotide (CpG) null model.

The Expectation Maximization algorithm for estimating substitution rate matrices from multiple alignments with phylogenetic trees

a.k.a.

“Phylo-EM”

Phylo-EM (definitions)

$$\theta = \{\pi, \mathbf{R}\}$$

Parameters

$$p_i(t) = P(x(t) = i)$$

x(t) = state at time t

$$\begin{aligned} \frac{d}{dt}\mathbf{p}(t) &= \mathbf{R}\mathbf{p} \\ \mathbf{p}(0) &= \pi \end{aligned}$$

Equation of state

$$\begin{aligned} \mathbf{p}(t) &= \pi\mathbf{M}(t) \\ \mathbf{M}(t) &= \exp(\mathbf{R}t) \end{aligned}$$

Matrix exponential

$$= \mathbf{U}\exp(\mathbf{D}t)\mathbf{U}^{-1}$$

Diagonal form

$$\mathbf{R} = \mathbf{U}^{-1}\mathbf{D}\mathbf{U}$$

Phylo-EM (derivation)

$$\begin{aligned}
\theta^{(n+1)} &= \text{argmax}_\theta \; \mathcal{E}(\theta|\theta^{(n)}) \\
\mathcal{E}(\theta|\theta^{(n)}) &= \sum_x P(x|y,\theta^{(n)}) \log P(x,y|\theta)
\end{aligned}$$

(general form of EM algorithm)

$$\begin{aligned}
y &= \text{present-day sequences (observed)} \\
x &= \text{ancestral sequences (unobserved)}
\end{aligned}$$

$$\mathcal{E}(\theta|\theta^{(n)}) = \sum_i \left(S_i \log \pi_i + D_i R_{ii} + \sum_{j \neq i} C_{ij} \log R_{ij} \right)$$

$$\begin{aligned}
S_i &= E\left[\#\text{ of ancestral residues in state } i\right] \\
D_i &= E\left[\#\text{ of residues} \times \text{time spent in state } i\right] \\
C_{ij} &= E\left[\#\text{ of mutations } i \rightarrow j\right]
\end{aligned}$$

Phylo-EM (algorithm)

$$R_{ij} \quad \leftarrow \quad \frac{C_{ij}}{D_i}$$

(iterate to convergence)

$$R_{ij} \quad = \quad \text{Rate of substitution } i \to j$$

(continuous-time Markov chain)

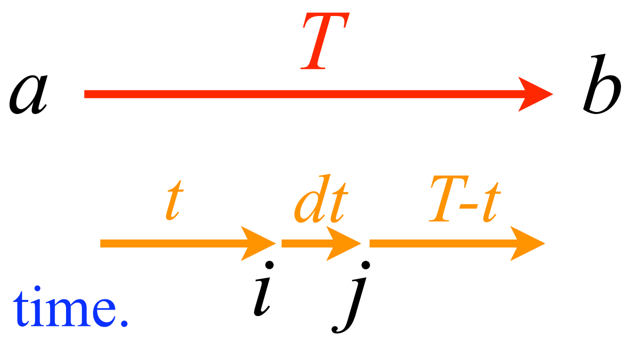
$$C_{ij} \quad = \quad E[\text{number of } i \to j \text{ substitutions}]$$

$$D_i \quad = \quad E[\text{dwell time in state } i]$$

posterior expectations: $E[x] \equiv \langle x \rangle_{P(\text{history}|\text{data},\mathbf{R})}$

On branch of length T, transition $a \to b$ is observed. Expected number of $i \to j$ substitution events: $C_{ij}(a,b,T)$

$$\begin{aligned}\mathbf{R} &= \mathbf{U}\Lambda\mathbf{U}^{-1}\\ \exp(\mathbf{R}t) &= \mathbf{U}\exp(\Lambda t)\mathbf{U}^{-1}\end{aligned}$$



Diagonalize rate matrix. Integrate over substitution event time.

$$\begin{aligned}C_{ij}(a,b,T) &= \frac{1}{\exp(\mathbf{R}T)_{ab}}\int_0^T \exp(\mathbf{R}t)_{ai}\ (R_{ij}dt)\ \exp(\mathbf{R}(T-t))_{jb}\\ &= \frac{R_{ij}}{\exp(\mathbf{R}T)_{ab}}\sum_{k=1}^{N} U_{ak}U^{-1}_{ki}\sum_{l=1}^{N} U_{jl}U^{-1}_{lb}\mathcal{J}_{kl}(T)\end{aligned}$$

Integral for dwell time $D_i(a,b,T)$ can similarly be expressed in terms of...

$$\mathcal{J}_{kl}(T) = \begin{cases} T\exp(\lambda_k T) & \text{if } \lambda_k = \lambda_l \\ (\exp(\lambda_k T) - \exp(\lambda_l T))/(\lambda_k - \lambda_l) & \text{if } \lambda_k \neq \lambda_l \end{cases}$$

Loop & stem rates

- Can formulate evolutionary questions as phylo-grammar parameterization problems

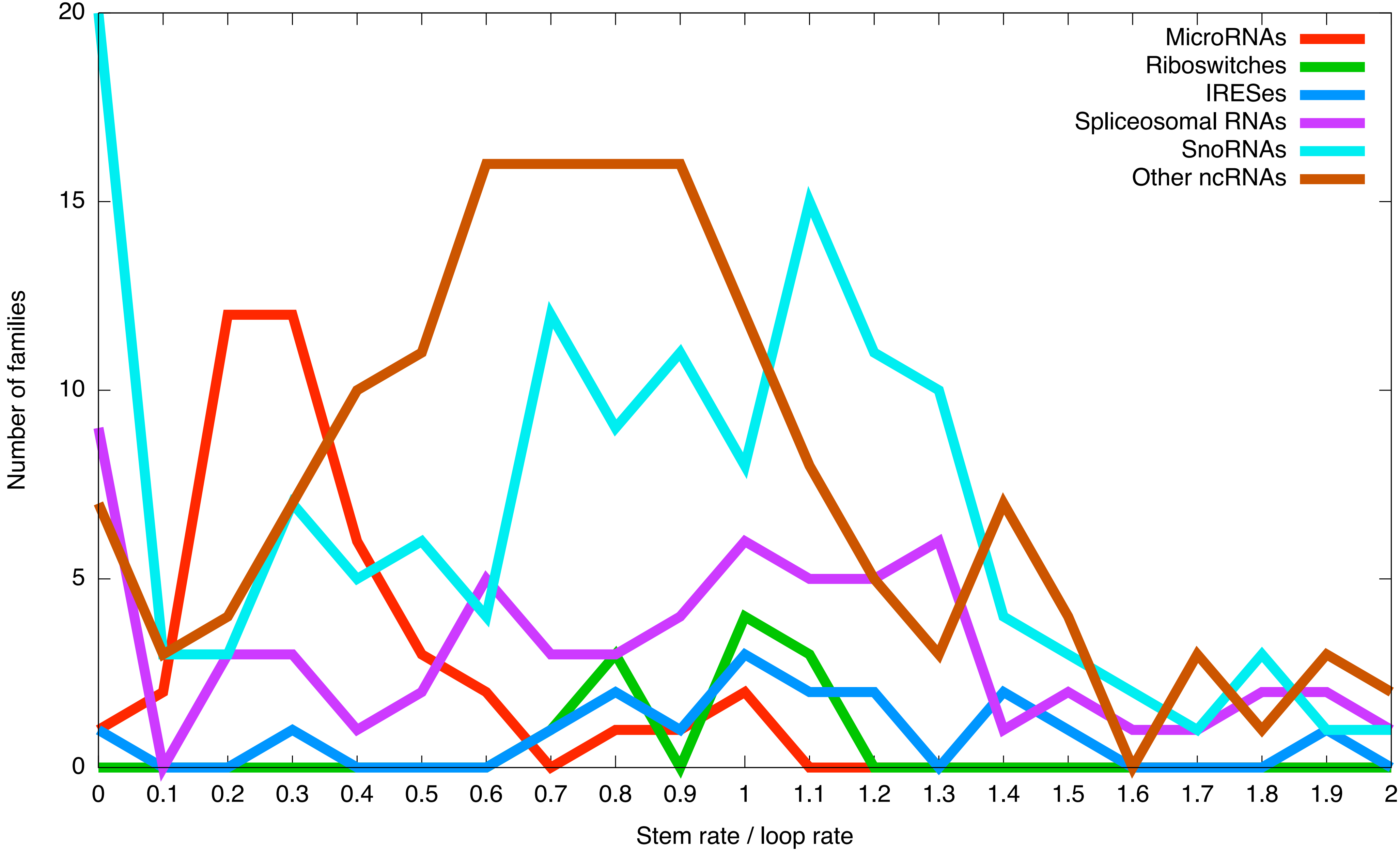
  e.g. "what is the ratio of substitution rates in loops compared to stems?"

- 323 RFAM families in 5 classes (miRNA, U*, sno*, IRES, riboswitch)

- Use XRate to estimate trees (Jukes-Cantor)

- Fit scaling factors to loop & stem matrices

miRNA stems evolve slower than other ncRNAs



Trained on RFAM

Limitations of phylo-EM

- **Advantages of phylo-EM**
  - Converges very quickly
  - Counts are useful in themselves
- **Disadvantages**
  - Gets stuck in local maxima
  - Sensitive to initial seed
  - Point estimate; no “error bars”
- **An MCMC equivalent would be nice**

EM-flavored MCMC

1. Sample θ from $g(\theta|\theta^{(n)})$:

$$\begin{aligned} g(\theta|\theta^{(n)}) &= \frac{\exp\left(\mathcal{E}(\theta|\theta^{(n)})\right)}{Z} \\ Z &= \int \exp\left(\mathcal{E}(\theta'|\theta^{(n)})\right) d\theta' \end{aligned}$$

2. Accept new θ with Hastings probability

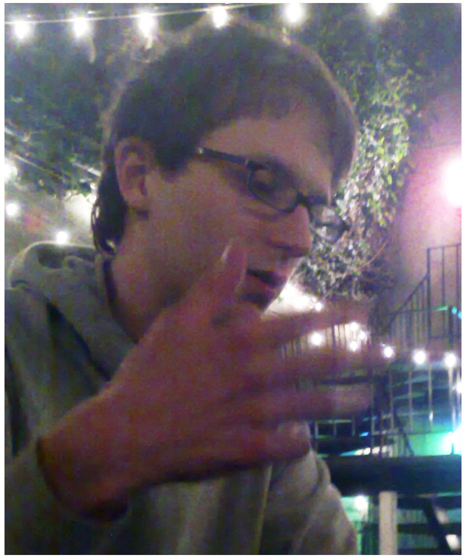
$$h(\theta,\theta^{(n)}) = \frac{P(y|\theta)}{P(y|\theta^{(n)})}\frac{g(\theta^{(n)}|\theta)}{g(\theta|\theta^{(n)})} = \frac{P(y|\theta)}{P(y|\theta^{(n)})}\exp\left(\mathcal{E}(\theta^{(n)}|\theta) - \mathcal{E}(\theta|\theta^{(n)})\right)$$

3. If accept, set $\theta^{(n+1)} \leftarrow \theta$
   If reject, set $\theta^{(n+1)} \leftarrow \theta^{(n)}$

Code's at **biowiki.org**, thanks to...

Robert Bradley

Andrew Uzilov



Lars Barquist

Collaborators

Marc Suchard
Nick Goldman
Carolin Kosiol
Chris Ponting
Andreas Heger
Sue Celniker
Mike Eisen

Mitchell Skinner

and also...

Oscar Westesson
Avinash Varadarajan
Yuri Bendaña
Pete Klosterman
Sharon Chao

Funding

NHGRI