Evolutionary Models for Multiple Sequence Alignment

CBB/CS 261

B. Majoros

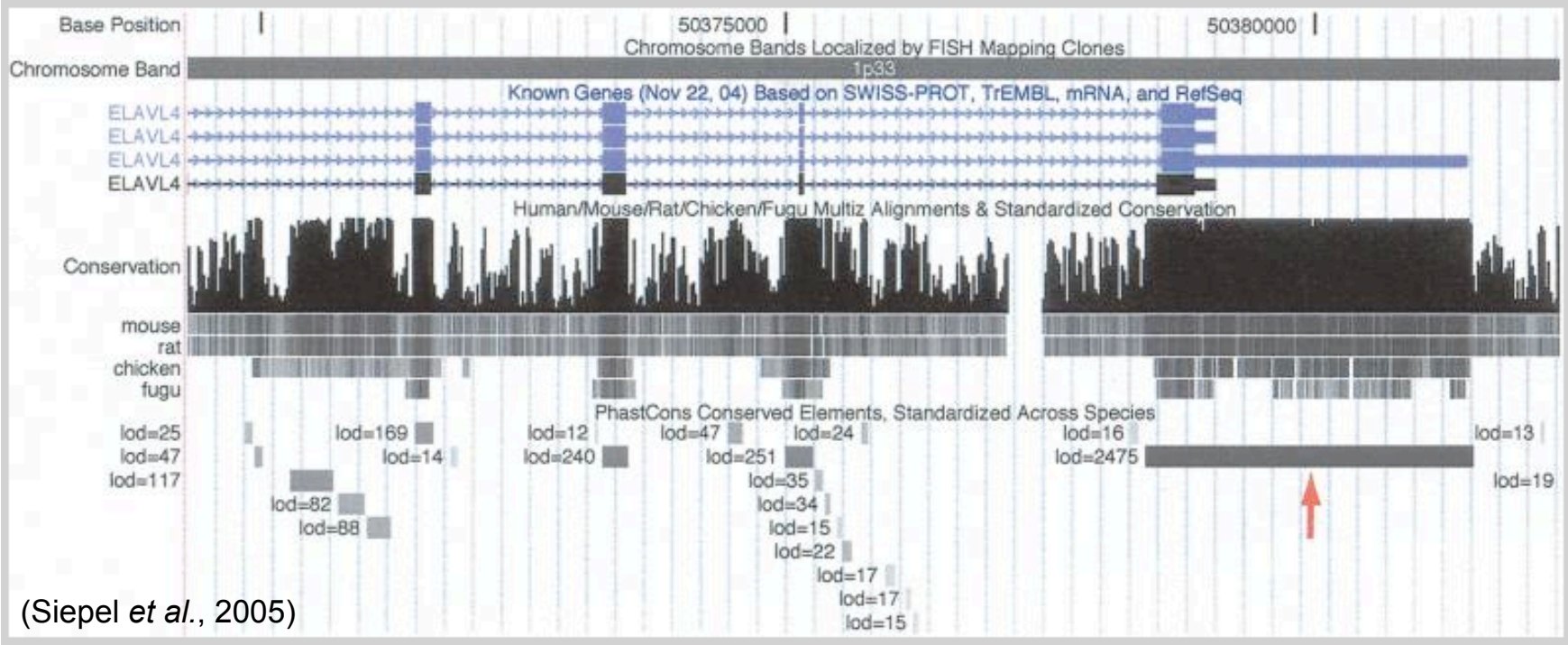
Duke
UNIVERSITY

Part I

Evolutionary Sequence Models

The Utility of Evolutionary Models

Evolutionary sequence models make use of the assumption that *natural selection operates more strongly on some genomic features* than others (i.e., functional versus non-functional DNA elements), resulting in a detectable bias in sequence conservation for the features of interest.

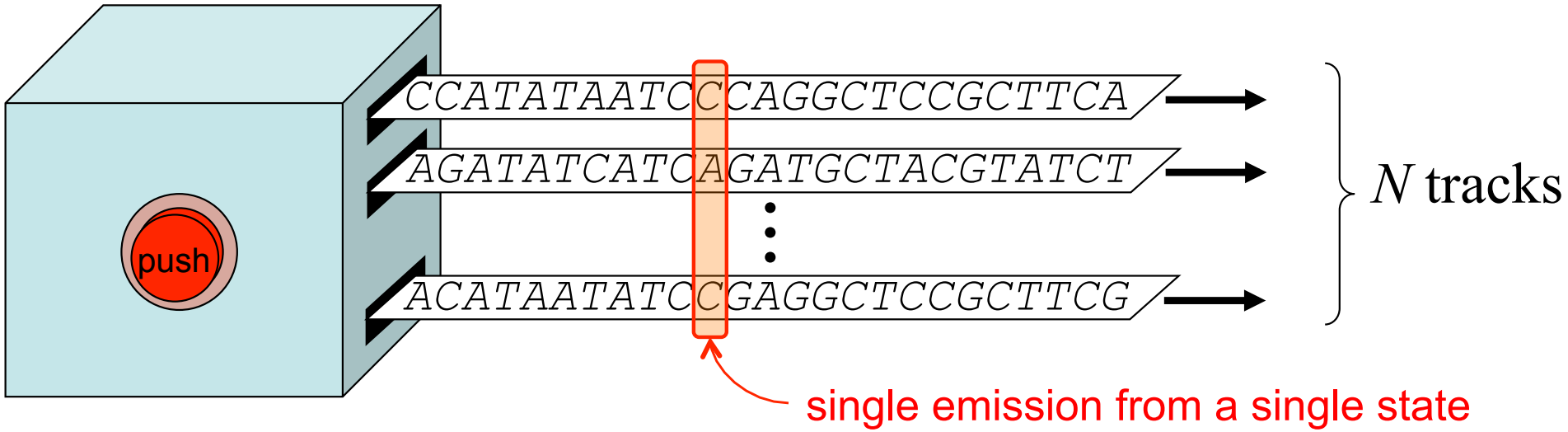


(Siepel *et al.*, 2005)

More generally, *conservation patterns* may differ between *levels* of DNA organization (i.e., amino acids in coding segments, versus individual nucleotides in conserved noncoding elements).

Duke UNIVERSITY

Recall: Multivariate HMMs



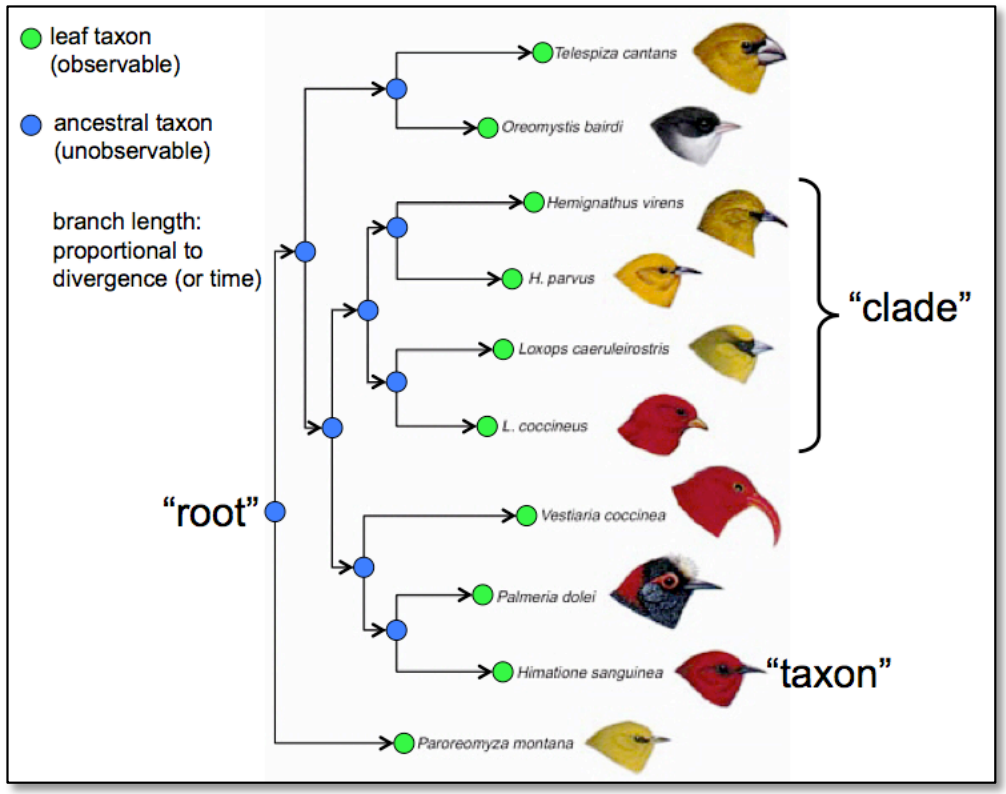
Each state emits N residues, one per track—i.e., a column of a multiple alignment.

Thus, each state must have a model for the joint distribution of the tracks (to represent emission probabilities).

Q: how might we model dependencies between tracks?

Non-independence of Sequences

Due to their *common ancestry*, genomic sequences for related taxa are not independent. We can control for that non-independence by explicitly modeling their dependence structure using a *phylogenetic tree*:



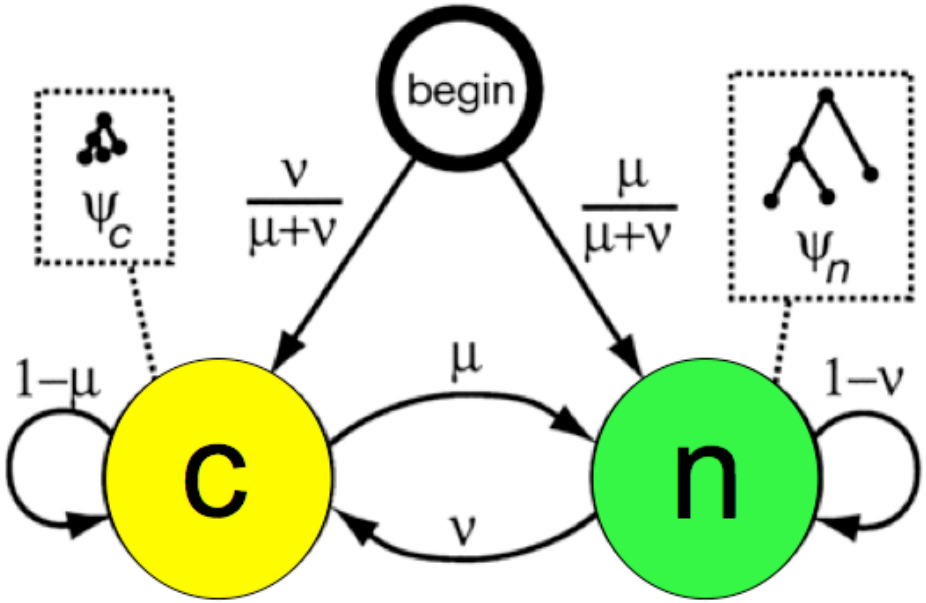
Branch lengths represent *evolutionary distance*, which conflates the distinct phenomena of *elapsed time* and *substitution rate* (as well as *selection* and *drift*).

We will see later that a phylogenetic tree (or "*phylogeny*") can be interpreted as a special type of *Bayesian network*, in which sequence conservation probabilities are expressed as a function of the branch lengths.

Duke UNIVERSITY

PhyloHMMs

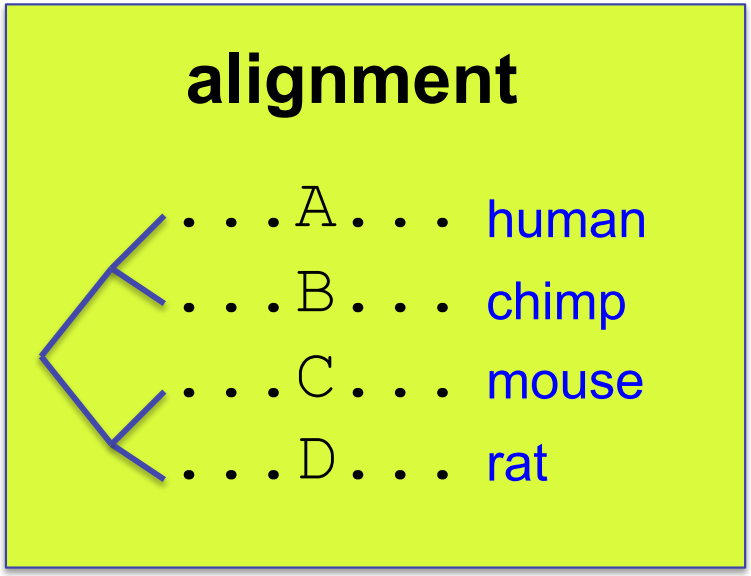
A *PhyloHMM* is a discrete multivariate HMM in which each state q_i has an associated *evolution model* ψ_i describing the expected *rates* and *patterns* of evolution in the class of features represented by that state.

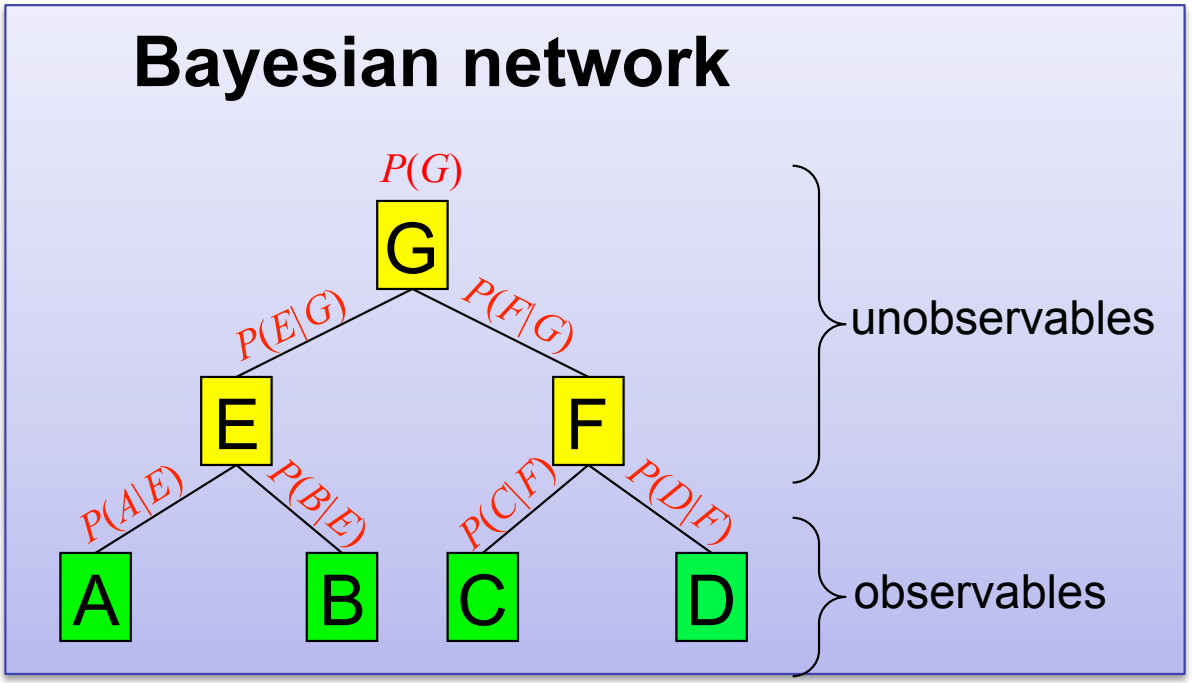


$\mathbf{x} =$ TCGCGACATATACGA . . .
TTGGGGCATGTGGGT . . .
AGCAGACGTCCGCAA . . .

(Siepel *et al.*, 2005)

Evaluating the Emission Probability





$$P(A,B,C,D) = \sum_{G,E,F} P(G)P(E|G)P(A|E)P(B|E)P(F|G)P(C|F)P(D|F)$$

$$P(observables) = \sum_{unobservables} \left(P(root) \prod_{\substack{nonroot \\ v}} P(v|parent(v)) \right)$$

Duke UNIVERSITY

A Recursion for the Emission Likelihood

The likelihood can be computed using a recursion known as *Felsenstein's pruning algorithm*:

$$L_u(a) = \begin{cases} \delta(u,a) & \text{if } u \text{ is a leaf} \\ \prod_{c \in children(u)} \sum_{b \in \alpha} L_c(b) P(c = b \mid u = a) & \text{otherwise} \end{cases}$$

$$= P(\text{descendents of } u | u=a).$$

$P(c=b|u=a)$ = probability of observing b in the child, given that we observe a in the parent. We can model this using a matrix of substitution probabilities, parameterized by the evolutionary time t that has passed between the ancestor and descendant taxa:

$$\mathbf{P}(t) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \overbrace{\begin{array}{cccc} A & C & G & T \\ \end{array}}^{\text{descendant}} \begin{bmatrix} p_{A\to A} & p_{A\to C} & p_{A\to G} & p_{A\to T} \\ p_{C\to A} & p_{C\to C} & p_{C\to G} & p_{C\to T} \\ p_{G\to A} & p_{G\to C} & p_{G\to G} & p_{G\to T} \\ p_{T\to A} & p_{T\to C} & p_{T\to G} & p_{T\to T} \end{bmatrix} \begin{array}{c} A \\ C \\ G \\ T \end{array} \overbrace{\quad}^{\text{ancestor}}$$

Substitution Matrix vs. Rate Matrix

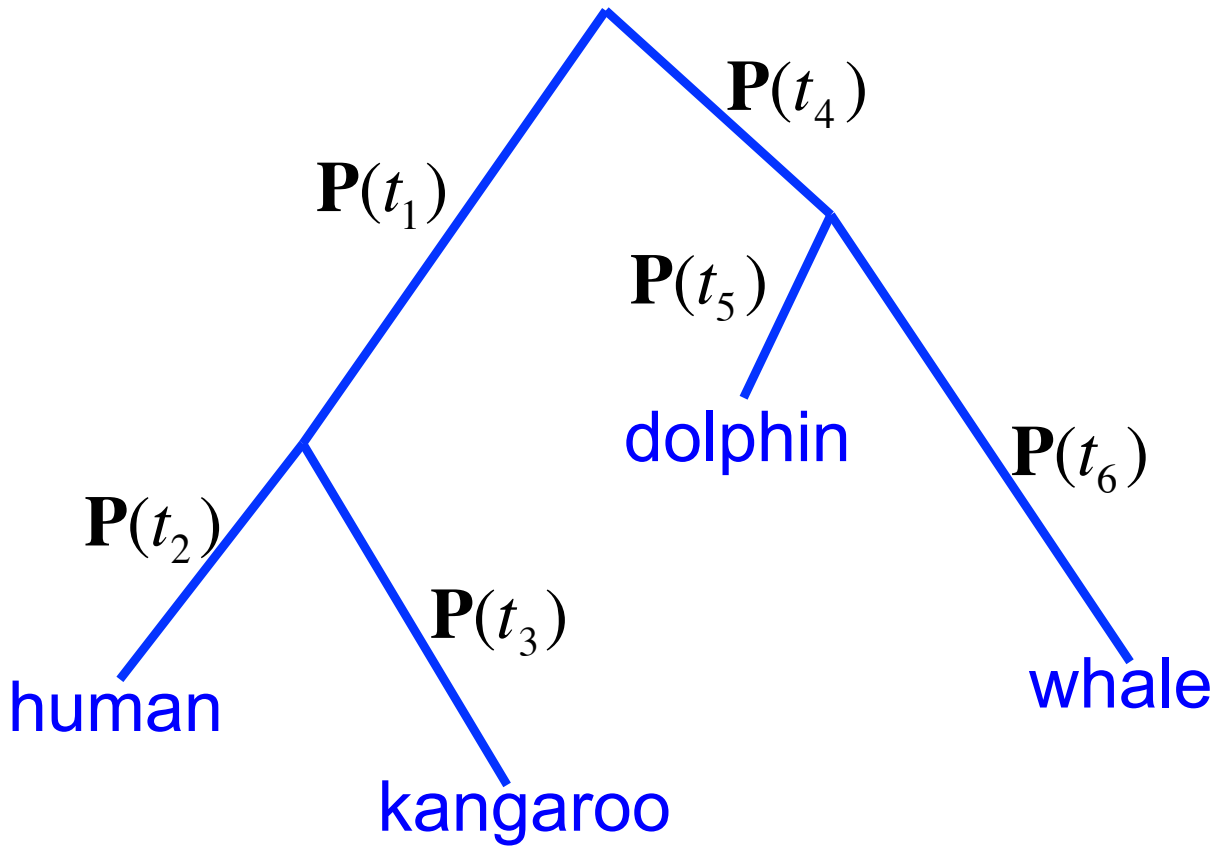
There is an important distinction between a *substitution matrix* and an *instantaneous rate matrix*.

$\mathbf{Q}$ = Instantaneous rate matrix. Gives instantaneous rates of substitutions (not parameterized by time).

$\mathbf{P}(t)$ = Substitution matrix. Gives the probabilities of substitutions for a specific branch length, t ("time").

Given $\mathbf{Q}$ and a set of phylogeny branch lengths $\{t_i\}$, we can compute a substitution matrix $\mathbf{P}(t_i)$ for each branch...

One Q, Many P(t)'s



$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Continuous-time Markov Chains (CTMCs)

Substitution models are typically based on continuous-time Markov chains. The *Markov property* for CTMCs states that:

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$$

$$\frac{d\mathbf{P}(t)}{dt} = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t+\Delta t) - \mathbf{P}(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t)\mathbf{P}(\Delta t) - \mathbf{P}(t)\mathbf{I}}{\Delta t}$$

$$= \mathbf{P}(t) \lim_{\Delta t \to 0} \underbrace{\frac{\mathbf{P}(\Delta t) - \mathbf{P}(0)}{\Delta t}}_{\text{does not depend on } t} = \mathbf{P}(t)\mathbf{Q}$$

We can derive an *instantaneous rate matrix* $\mathbf{Q}$ from $\mathbf{P}(t)$, where we make use of the fact that $\mathbf{P}(0)=\mathbf{I}$.

The solution to this differential equation is:

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n t^n}{n!} = I + t\mathbf{Q} + \frac{\mathbf{Q}^2 t^2}{2!} + \ldots$$

$e^{\mathbf{Q}t}$ (the "*matrix exponential*") denotes a Taylor expansion, which we can solve via *spectral (eigenvector) decomposition*:

$$\mathbf{Q} = \mathbf{G}\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}\mathbf{G}^{-1} \Rightarrow \mathbf{P}(t) = \mathbf{G}\begin{bmatrix} e^{t\lambda_1} & & & \\ & e^{t\lambda_2} & & \\ & & \ddots & \\ & & & e^{t\lambda_n} \end{bmatrix}\mathbf{G}^{-1}$$

Desirable Properties of Substitution Matrices

Reversibility: ("detailed balance")

$$\forall_{ij}\left(\pi_i \mathbf{P}_{ij}(t) = \pi_j \mathbf{P}_{ji}(t)\right)$$

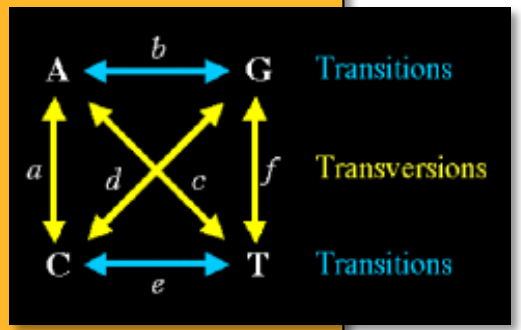
where π_i is the equilibrium frequency of base i.

Transition/Transversion Discrimination:

Using different parameters for transition and transversion rates.

Transitions: `purine↔purine`, `pyrimidine↔pyrimidine`

Transversions: `purine↔pyrimidine`



Some Common Forms for Q

Jukes-Cantor:

$$\mathbf{Q}_{JK} = \begin{bmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{bmatrix}$$

Kimura:

$$\mathbf{Q}_{K2P} = \begin{bmatrix} - & \beta & \alpha & \beta \\ \beta & - & \beta & \alpha \\ \alpha & \beta & - & \beta \\ \beta & \alpha & \beta & - \end{bmatrix}$$

these assume uniform equilibrium frequencies!

Felsenstein:

$$\mathbf{Q}_{FEL} = \begin{bmatrix} - & \alpha\pi_C & \alpha\pi_G & \alpha\pi_T \\ \alpha\pi_A & - & \alpha\pi_G & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_C & - & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_C & \alpha\pi_G & - \end{bmatrix}$$

Hasegawa, Kishino, Yano:

$$\mathbf{Q}_{HKY} = \begin{bmatrix} - & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & - & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & - & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & - \end{bmatrix}$$

General reversible model:

$$\mathbf{Q}_{REV} = \begin{bmatrix} - & \beta\pi_C & \alpha\pi_G & \chi\pi_T \\ \beta\pi_A & - & \kappa\pi_G & \omega\pi_T \\ \alpha\pi_A & \kappa\pi_C & - & \tau\pi_T \\ \chi\pi_A & \omega\pi_C & \tau\pi_G & - \end{bmatrix}$$

Duke UNIVERSITY

Part II

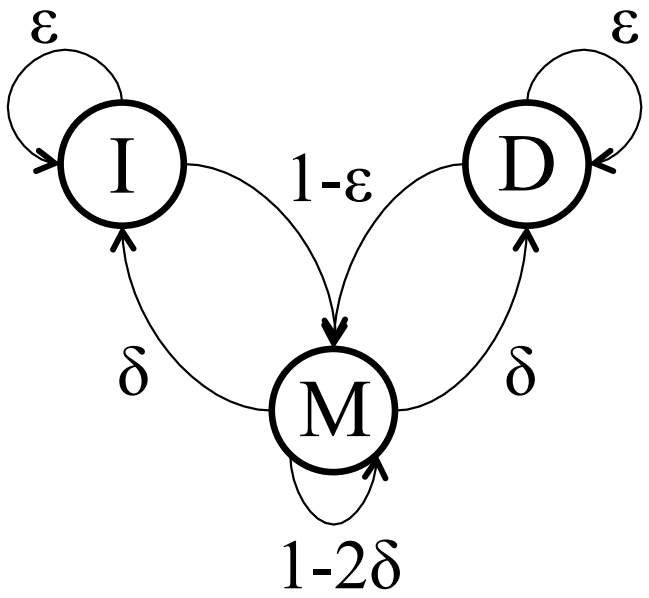
Multiple Alignment

Recall: Pairwise alignment with PHMMs

- **Emission** probabilities assess similarity between aligned residues

BLOSUM62

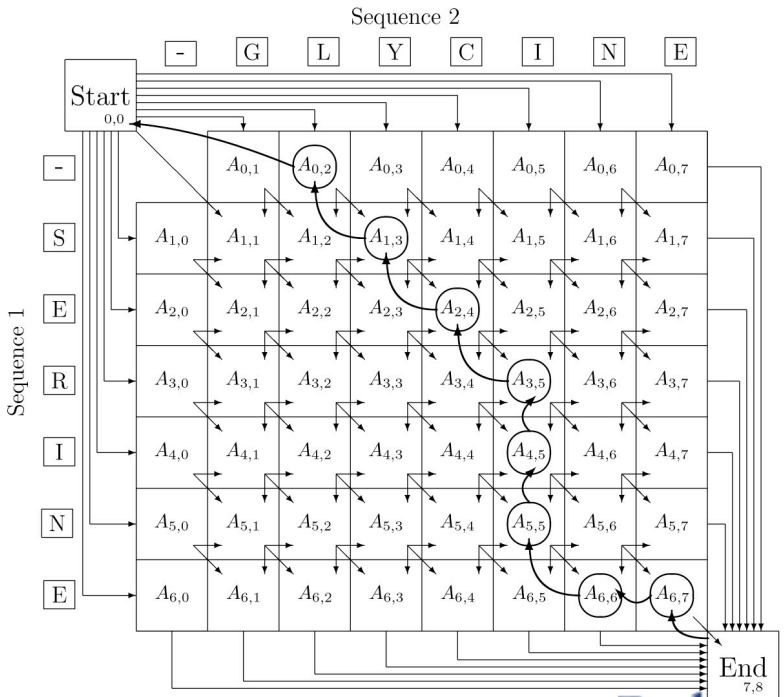
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W



- **Transition** probabilities can be used to penalize gaps

$$\delta = \text{gap open} \qquad \varepsilon = \text{gap extend}$$

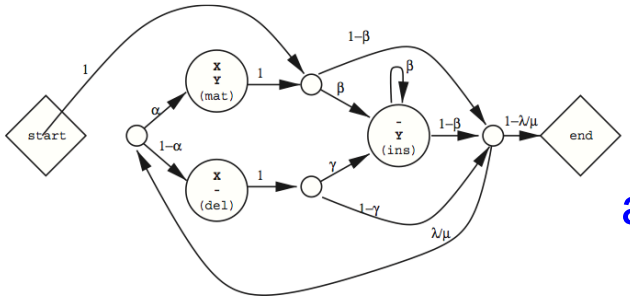
- **Viterbi** decoding finds the optimal alignment



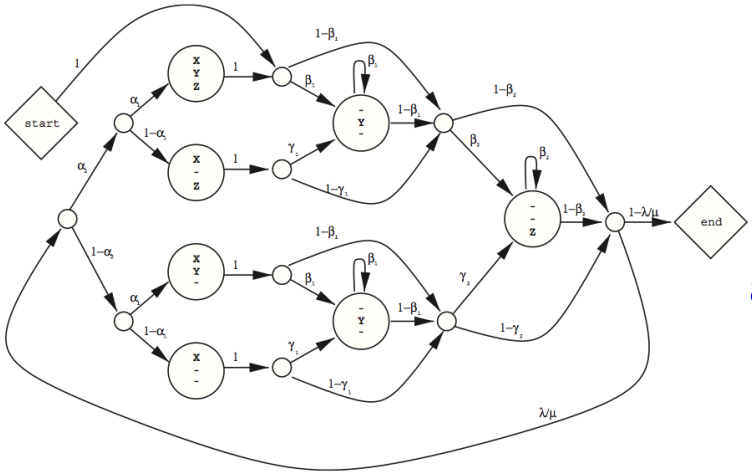
– – S E R I N E –
G L Y C I – – N E

Duke UNIVERSITY

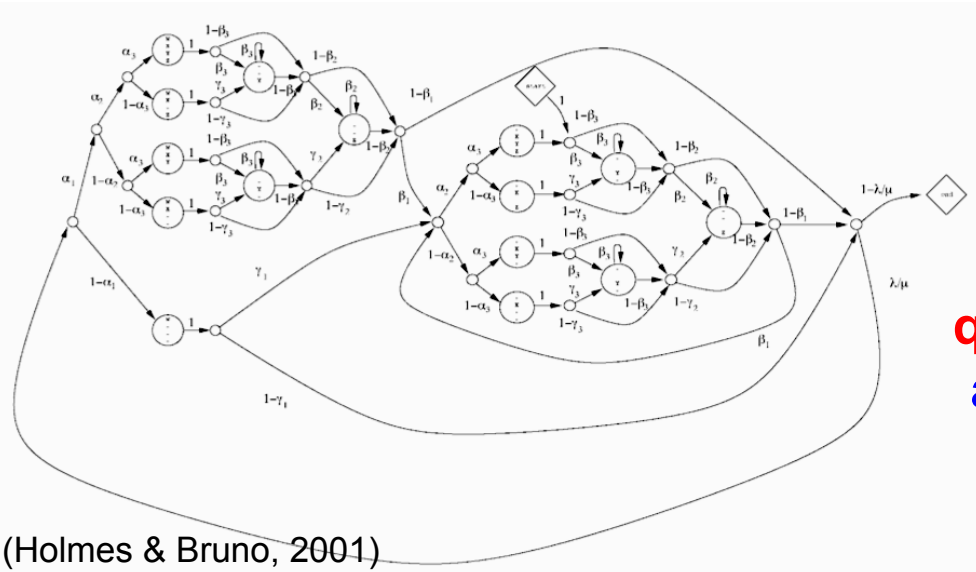
Pair HMM ➔ Triple HMM ➔ Quadruple HMM ?



pair HMM
aligns 2 species



triple HMM
aligns 3 species



quadruple HMM
aligns 4 species

(Holmes & Bruno, 2001)

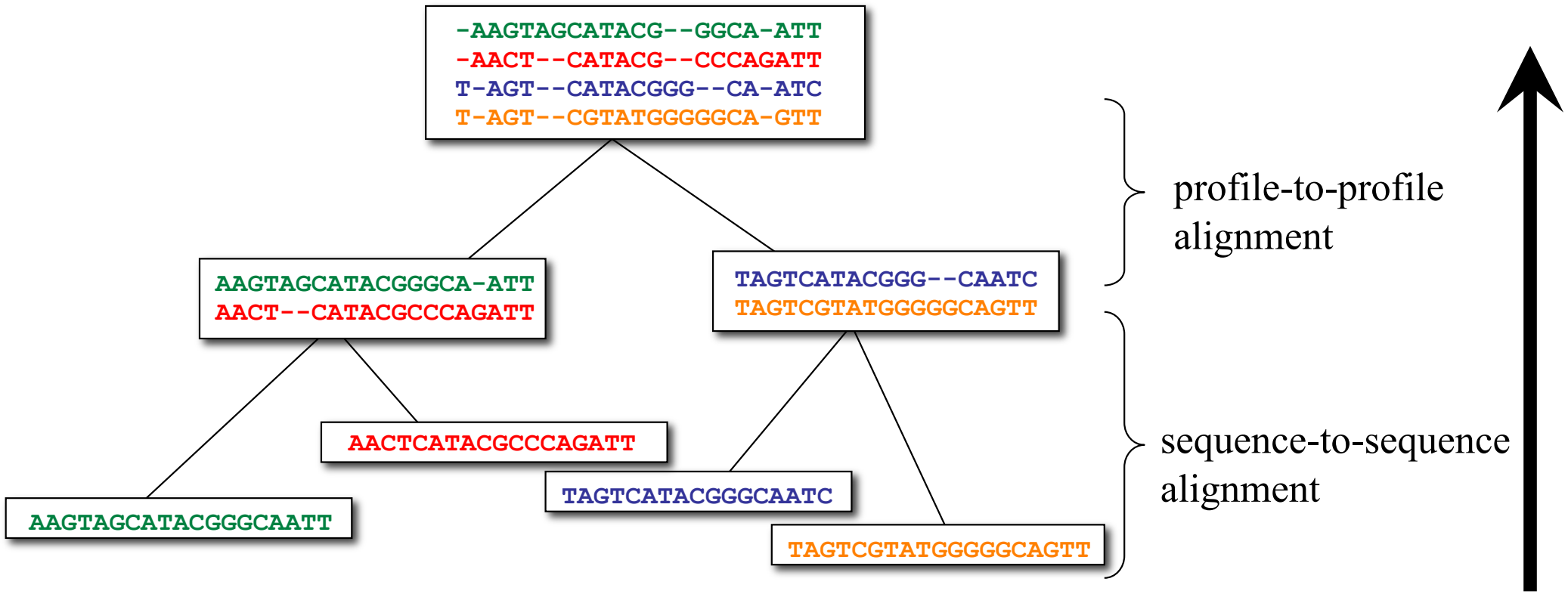
for 100bp sequences...

Sequences	Cells	Memory
2	9801	40k
3	970,299	3.7M
4	96,059,601	366M
5	9,509,900,499	35G
6	9.41E+11	3.4T
7	9.32E+13	339T
8	9.28E+15	33P
9	9.14E+17	3.2E
10	9.04E+19	314E

#gallons of water in the Pacific ocean

Duke UNIVERSITY

Progressive Alignment: One Pair at a Time



- First, align the leaves of the tree (using a Pair HMM)
- Then align ancestral taxa, using either a "consensus" sequence for ancestors, or averaging over all pairs of leaf residues

---

A more principled approach: model the ancestral sequences explicitly, using a probabilistic evolutionary model...

Duke UNIVERSITY

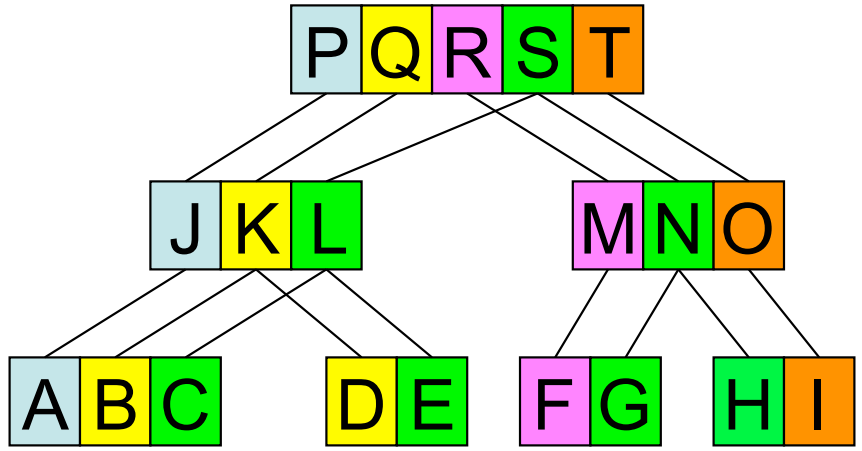
Networks of Residues

Problem:

Align the sequences ABC, DE, FG, and HI.

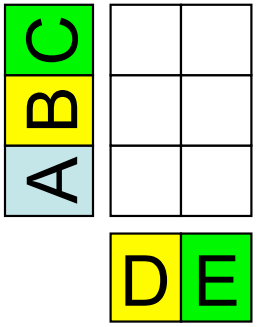
Solution:

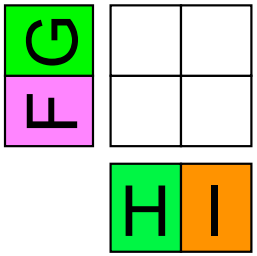
```
AB-C-
-D-E-
--FG-
---HI
```

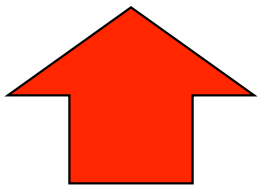


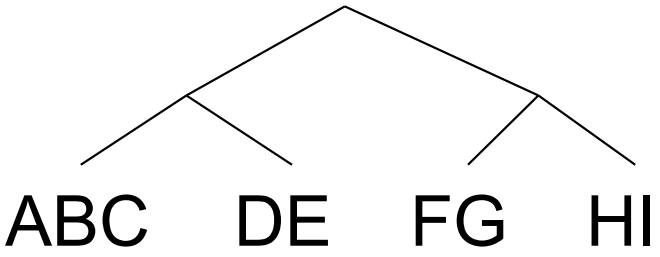
The multiple alignment problem is precisely the problem of inferring the network of residue homologies—i.e., the evolutionary history of each base.

Building the Network

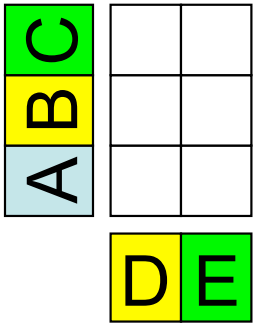


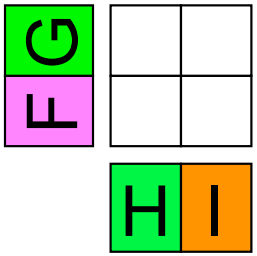




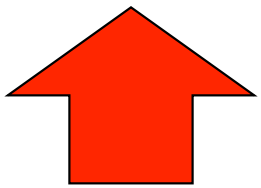


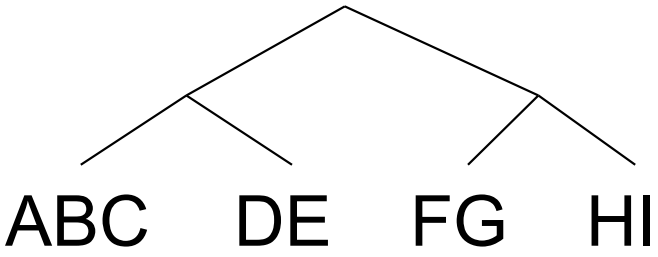
Building the Network



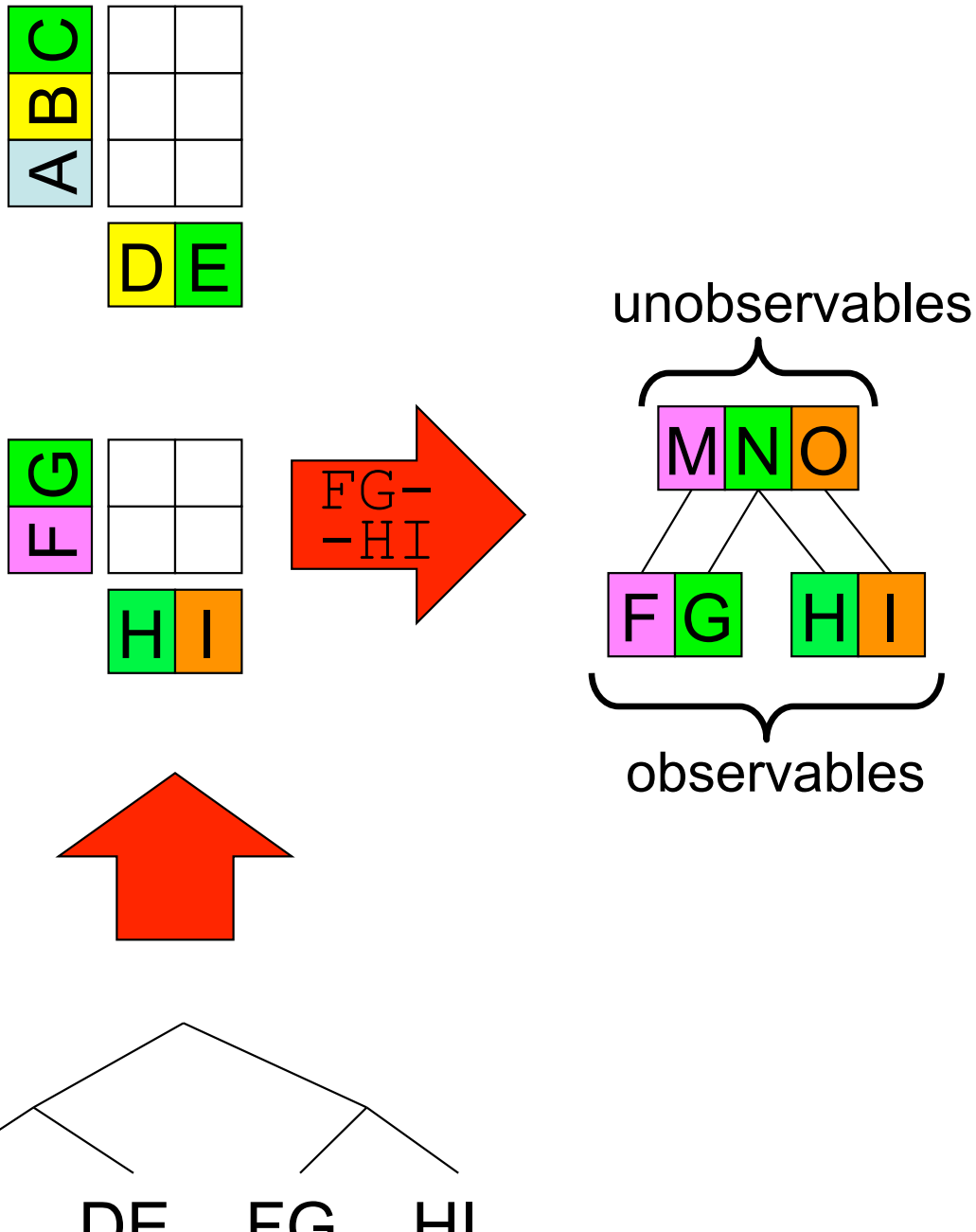


FG-
-HI

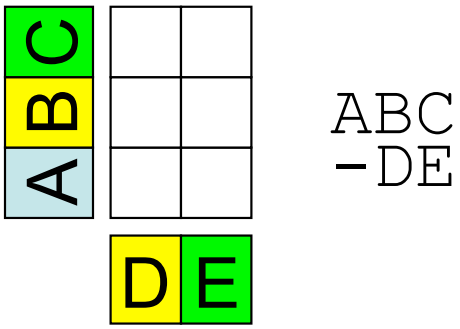




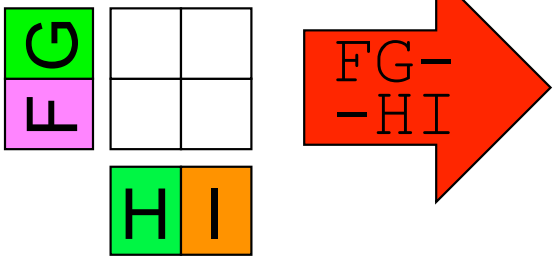
Building the Network

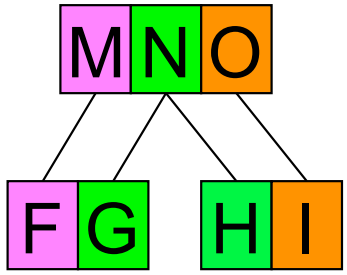


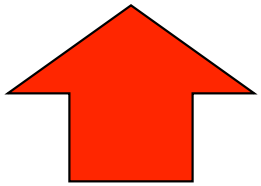
Building the Network

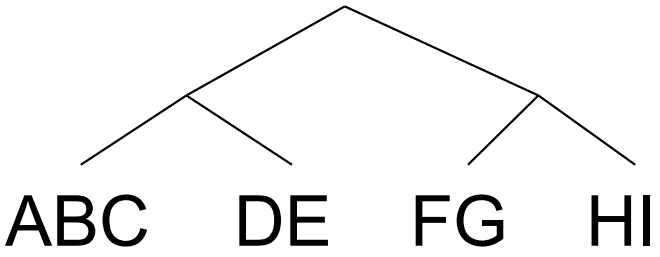


ABC
-DE



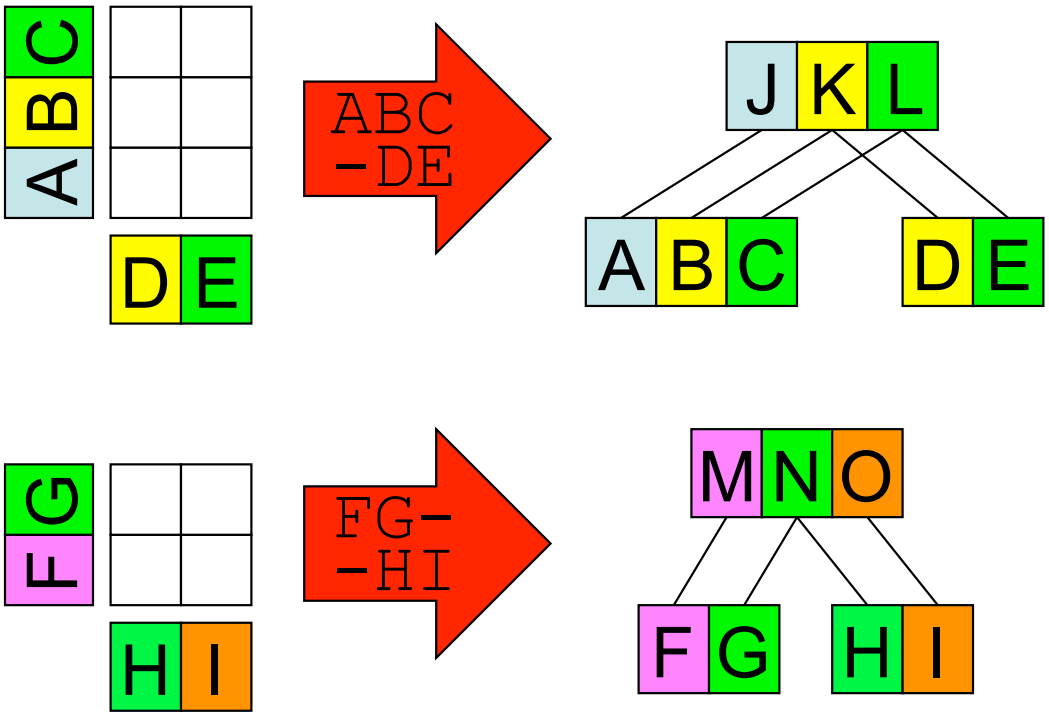




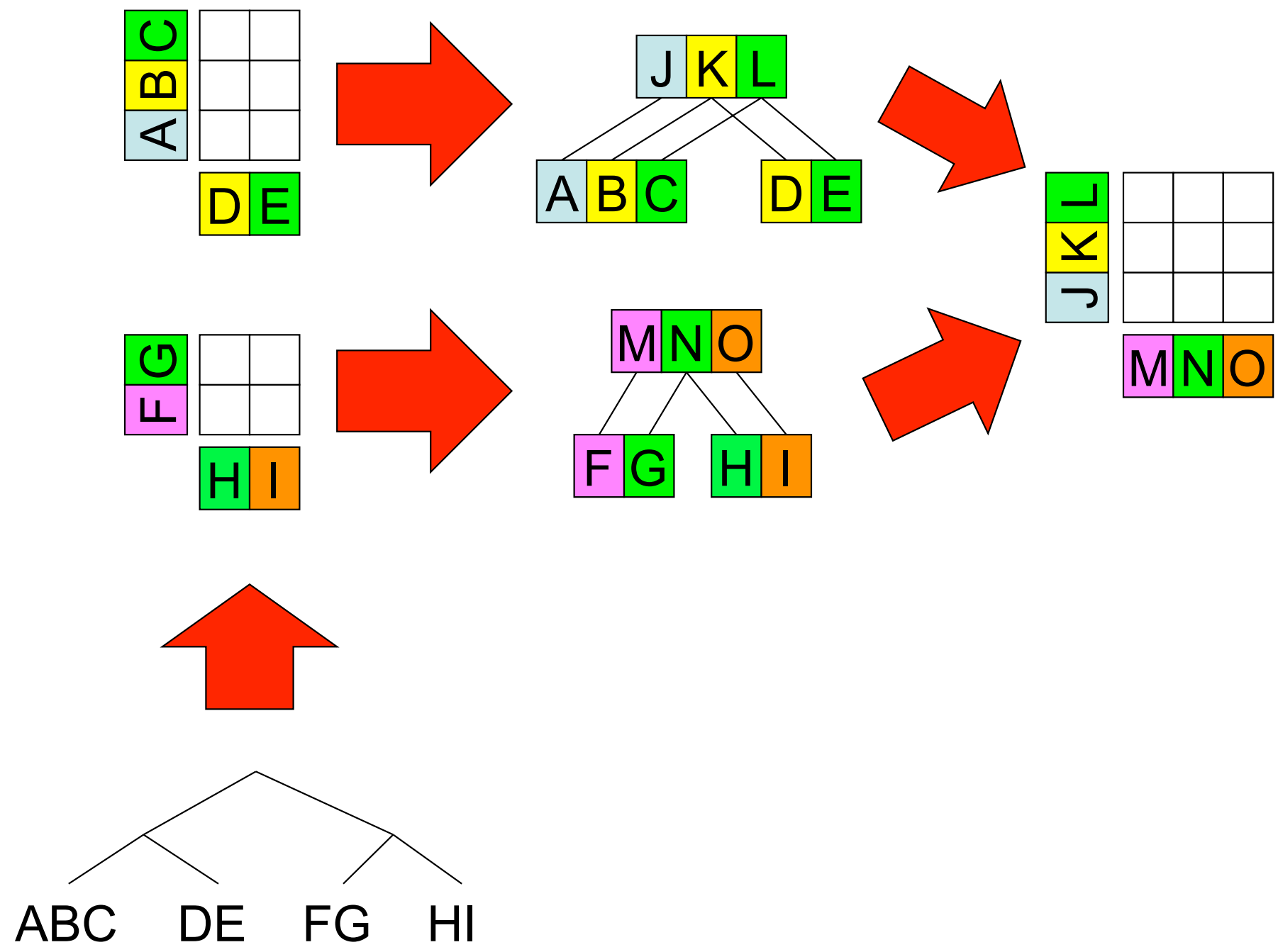


Duke
UNIVERSITY

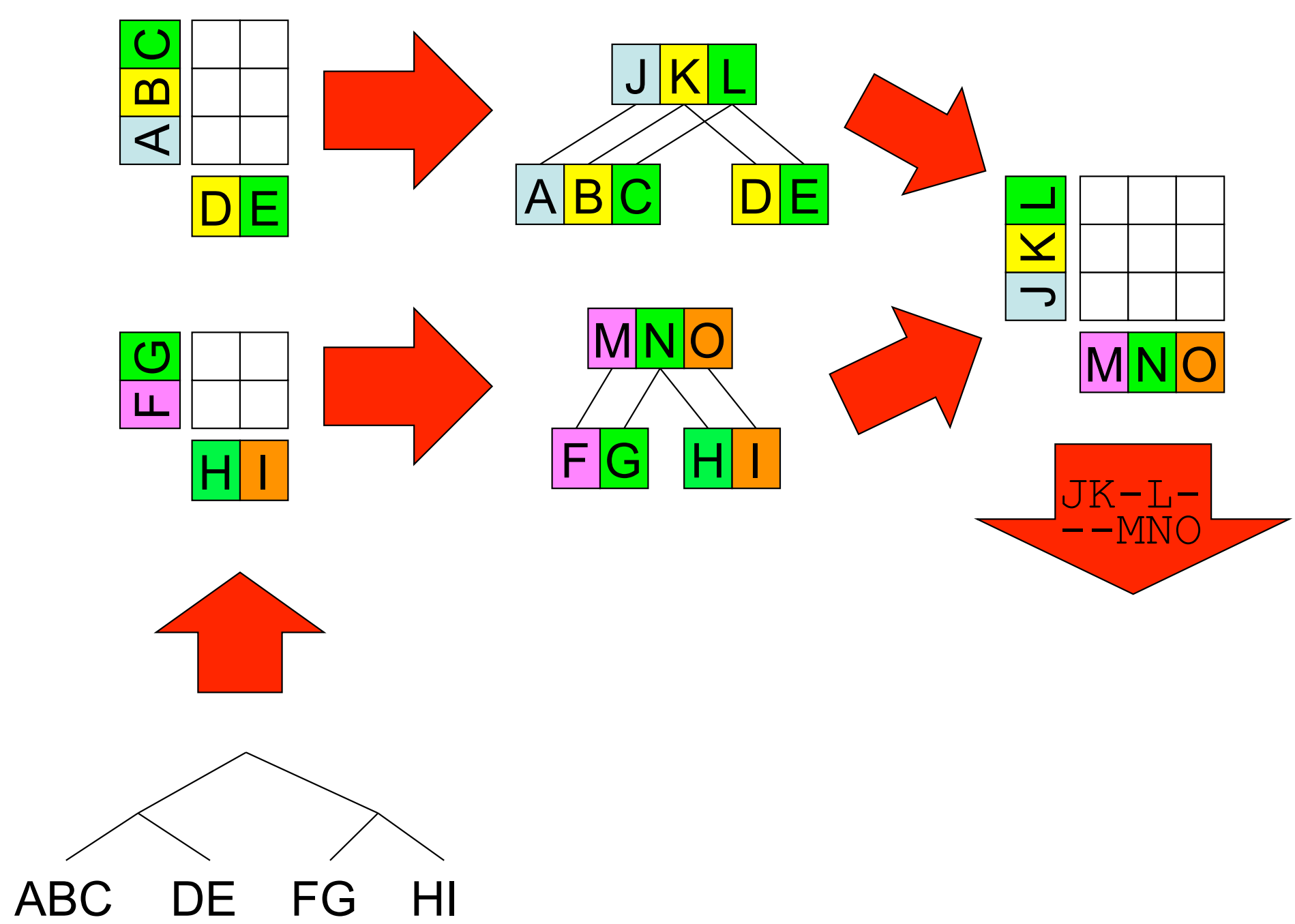
Building the Network



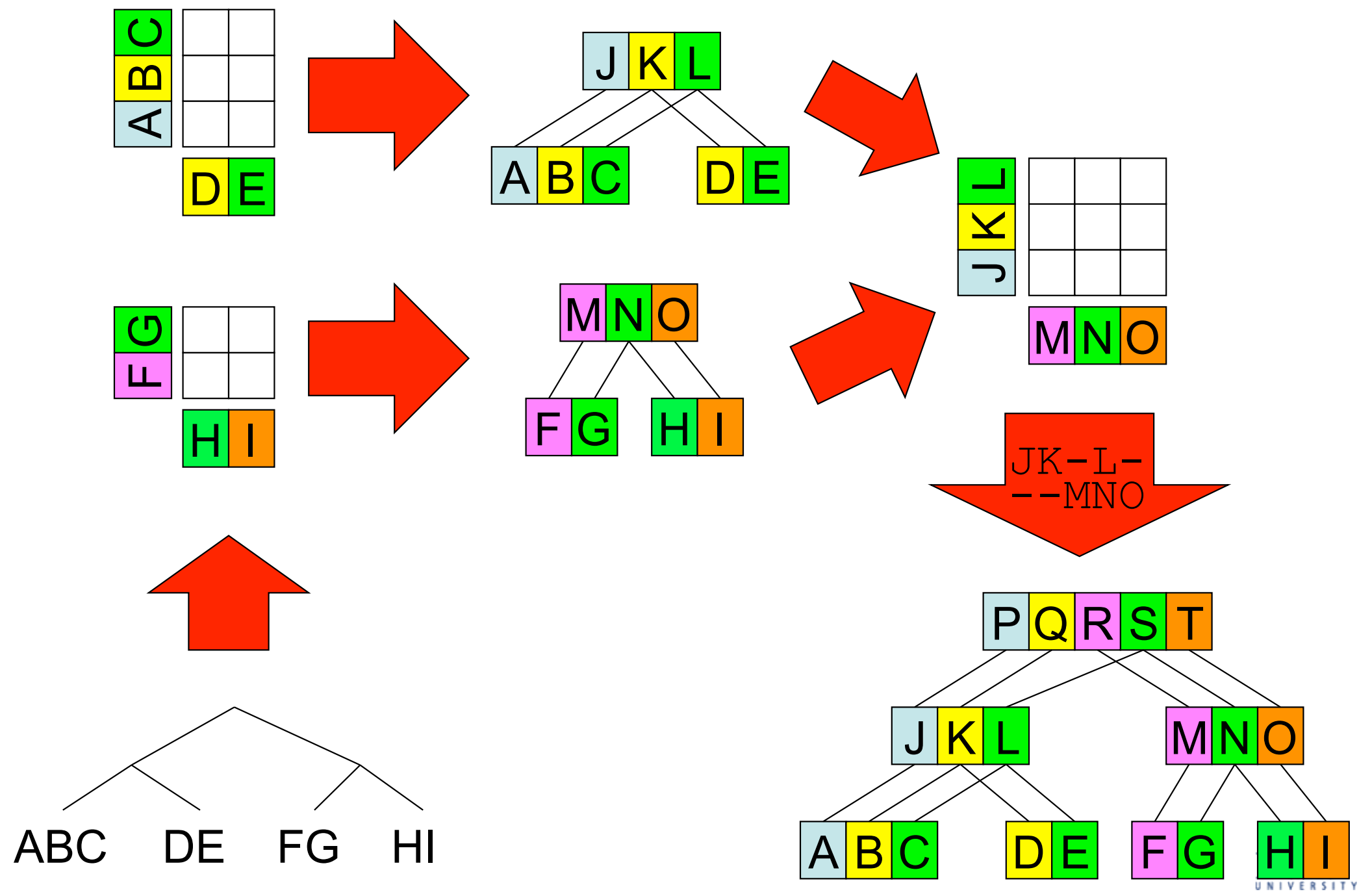
Building the Network



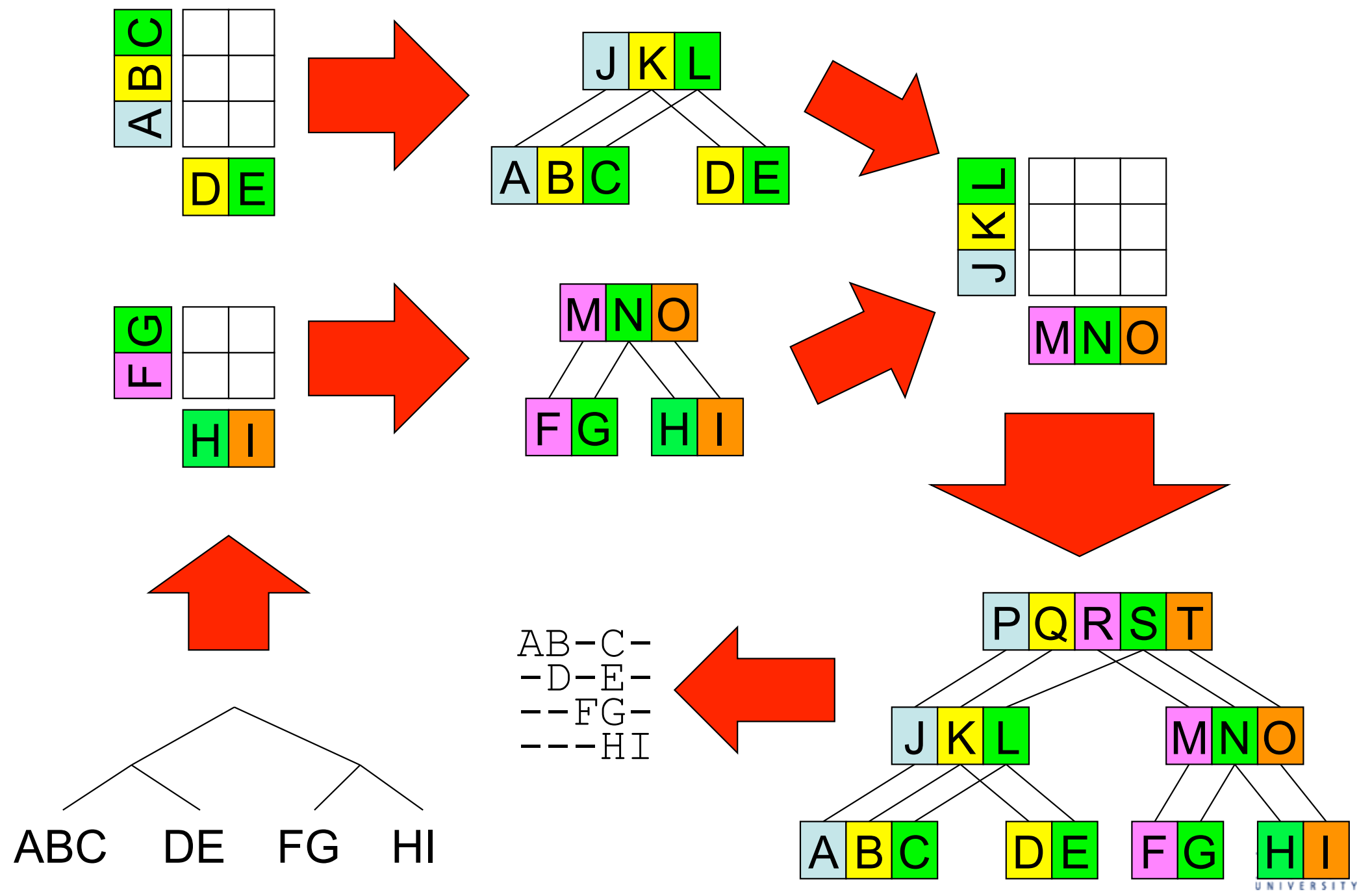
Building the Network



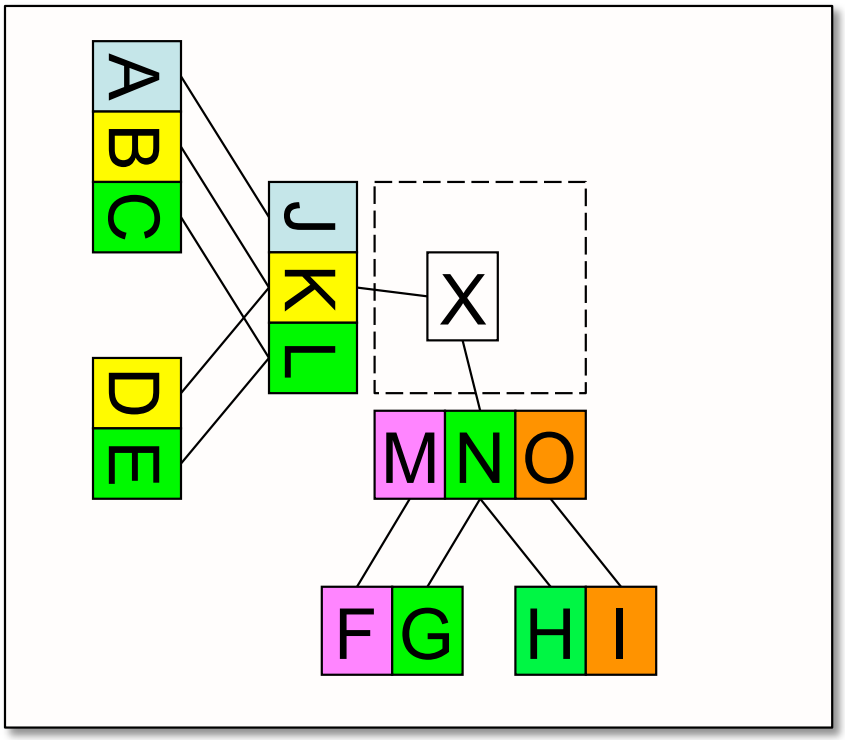
Building the Network

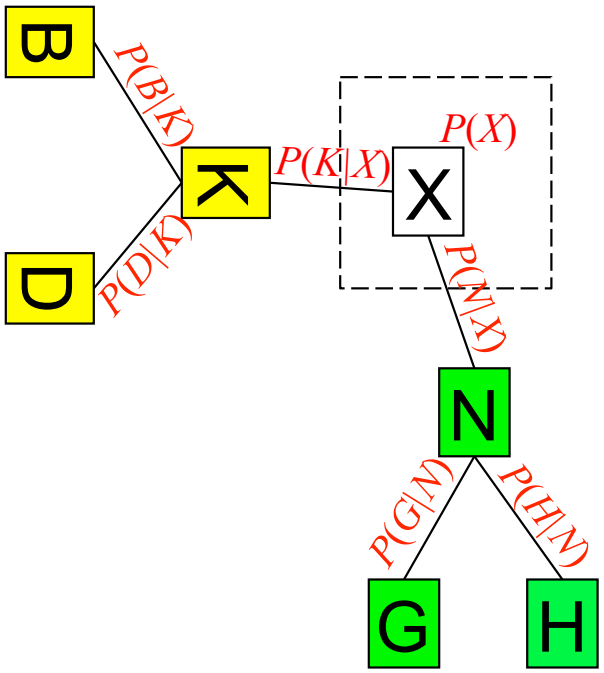


Building the Network



Evaluating Emission Probabilities

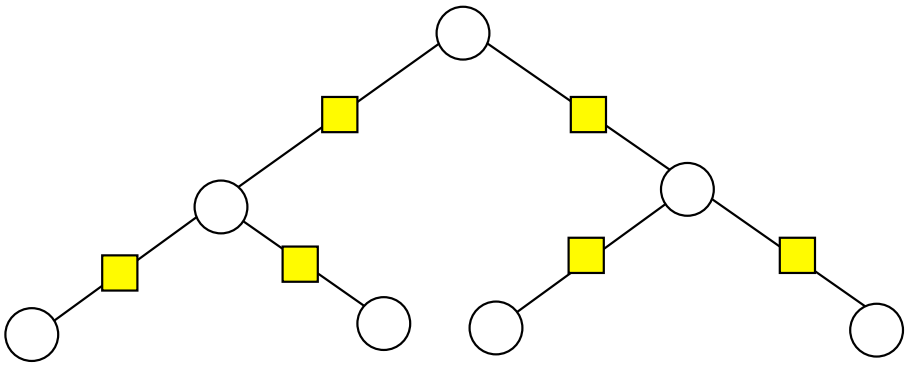




$$P(B,D,G,H) = \sum_{X,K,N} P(X)P(K|X)P(B|K)P(D|K)P(N|X)P(G|N)P(H|N)$$

$$P(observables) = \sum_{unobservables} \left(P(root) \prod_{\substack{nonroot \\ v}} P(v|parent(v)) \right)$$

Sampling Alignments



= a sequence S

= (B,H), a "Branch-HMM" (transducer) B describing the evolutionary process whereby the child evolves from the parent, and the actual *indel history* H which is a specific realization of this process (a "draw")

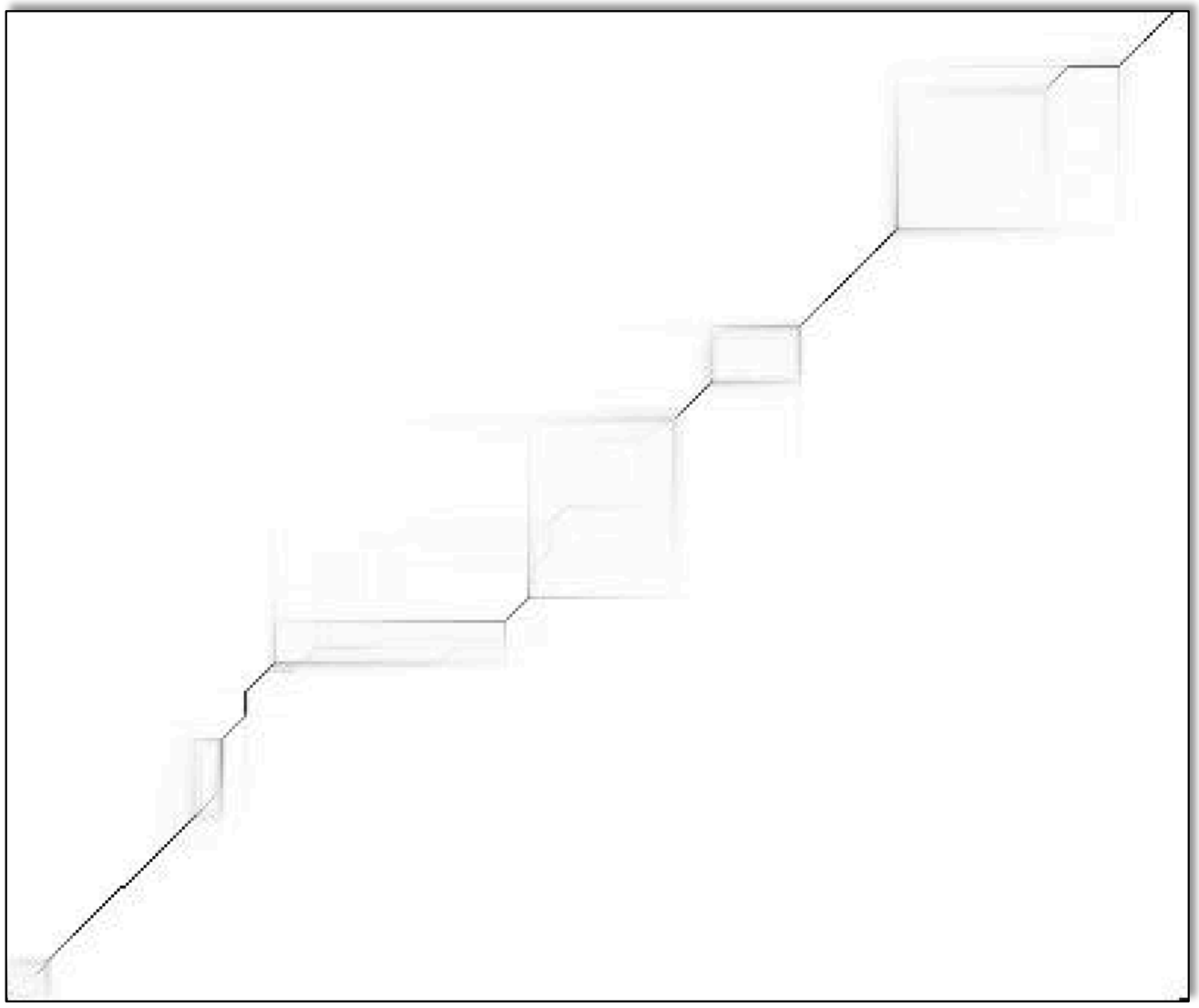
Sampling of alignments proceeds by sampling pairwise "branch alignments" (or "indel histories") H that live within the yellow squares. *An indel history is simply a path through a Pair HMM.*

Sampling branch alignments is simple: just sample from a PHMM via *Forward* or *Backward*:

$$P(\mathbf{y}_k|\mathbf{y}_{k-1},i,j,S_1,S_2) = \begin{cases} \dfrac{B_{i+1,j+1,\mathbf{y}_k} P_t(\mathbf{y}_k|\mathbf{y}_{k-1}) P_e(s_{i,1},s_{j,2}|\mathbf{y}_k)}{B_{i,j,\mathbf{y}_{k-1}}} & \text{if } \mathbf{y}_k \in Q_M \\ \dfrac{B_{i,j+1,\mathbf{y}_k} P_t(\mathbf{y}_k|\mathbf{y}_{k-1}) P_e(-,s_{j,2}|\mathbf{y}_k)}{B_{i,j,\mathbf{y}_{k-1}}} & \text{if } \mathbf{y}_k \in Q_I \\ \dfrac{B_{i+1,j,\mathbf{y}_k} P_t(\mathbf{y}_k|\mathbf{y}_{k-1}) P_e(s_{i,1},-|\mathbf{y}_k)}{B_{i,j,\mathbf{y}_{k-1}}} & \text{if } \mathbf{y}_k \in Q_D \end{cases}$$

Duke UNIVERSITY

Posterior Alignment Matrix



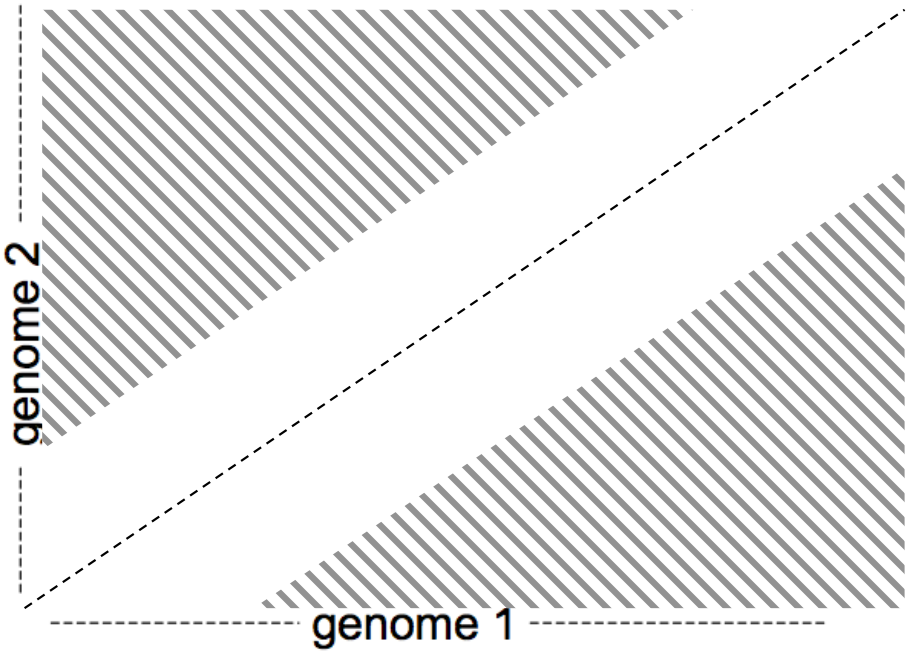
sequence 2...

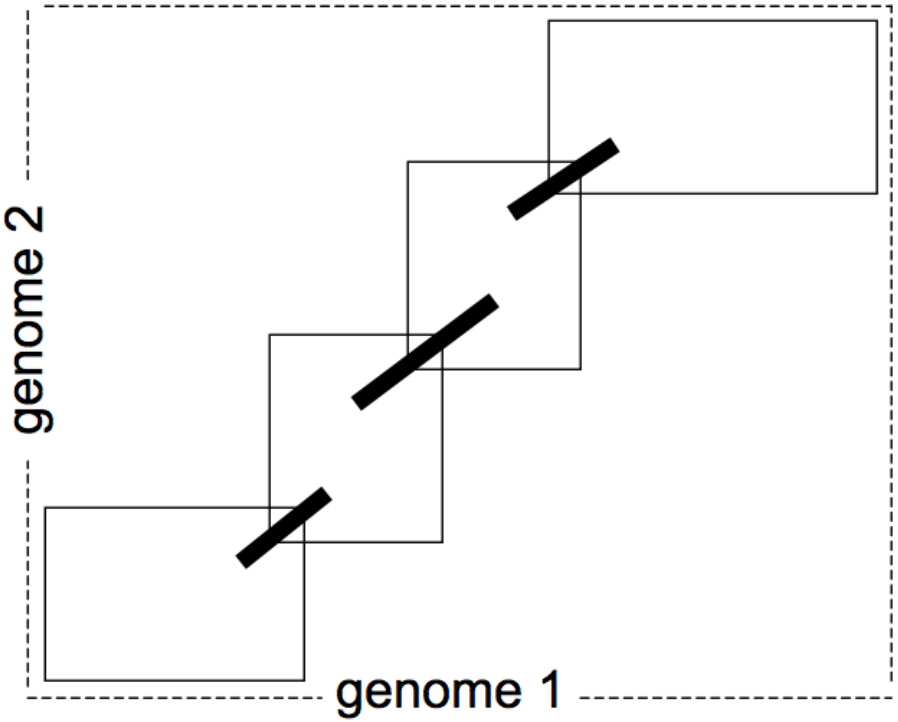
sequence 1...

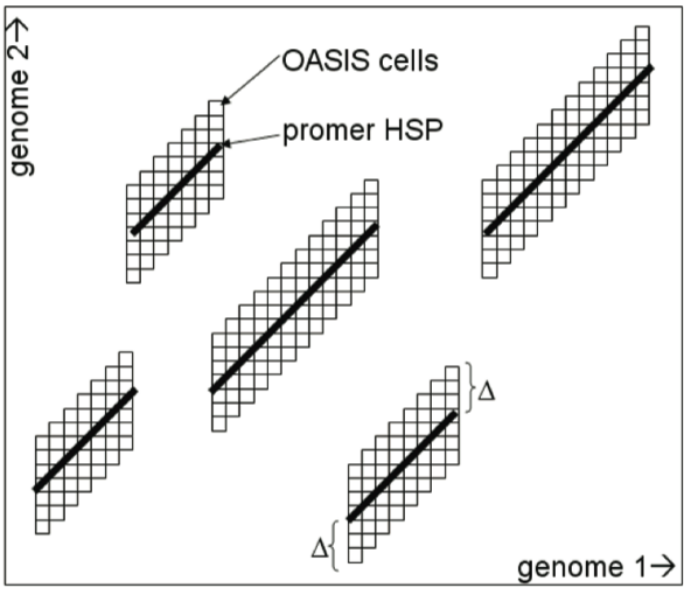
pixel intensity = posterior probability of a match in that cell

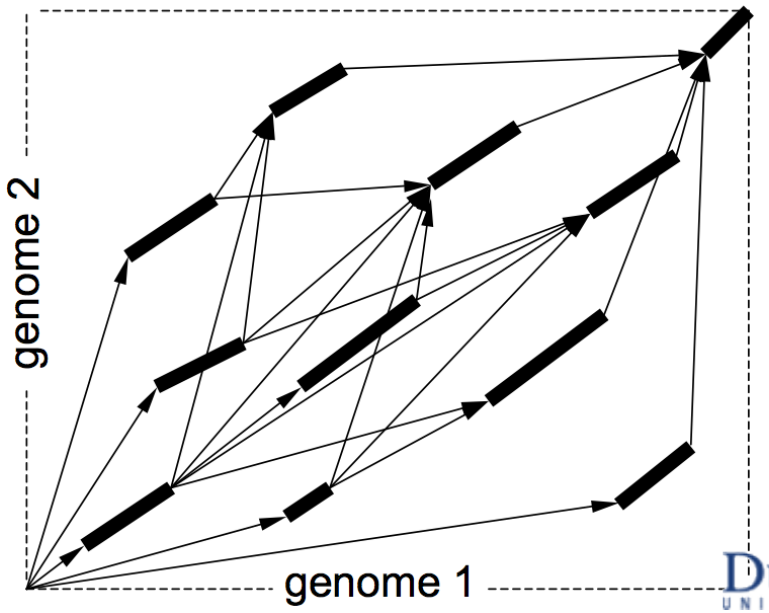
(posterior probability: conditional on the full input sequences)

Banding: Reduce the Search Space





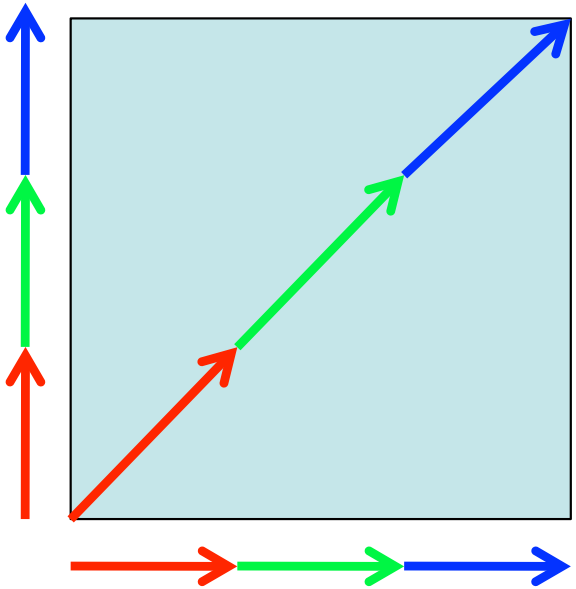


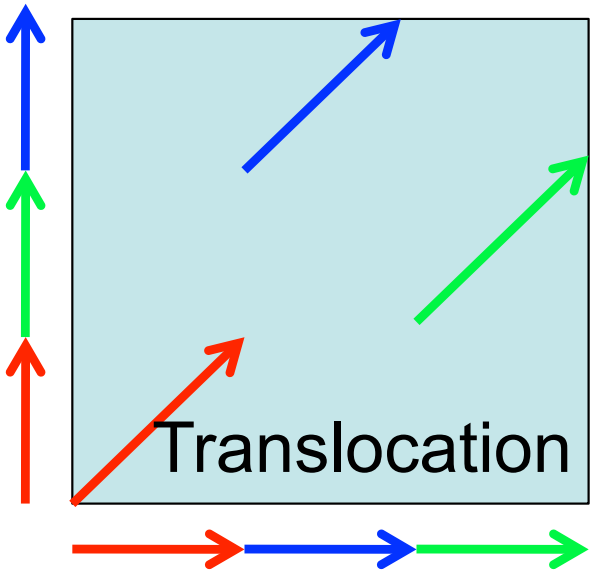


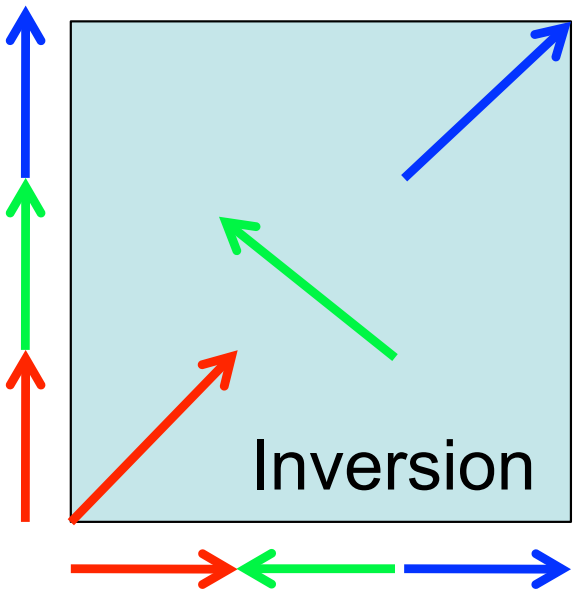
Duke UNIVERSITY

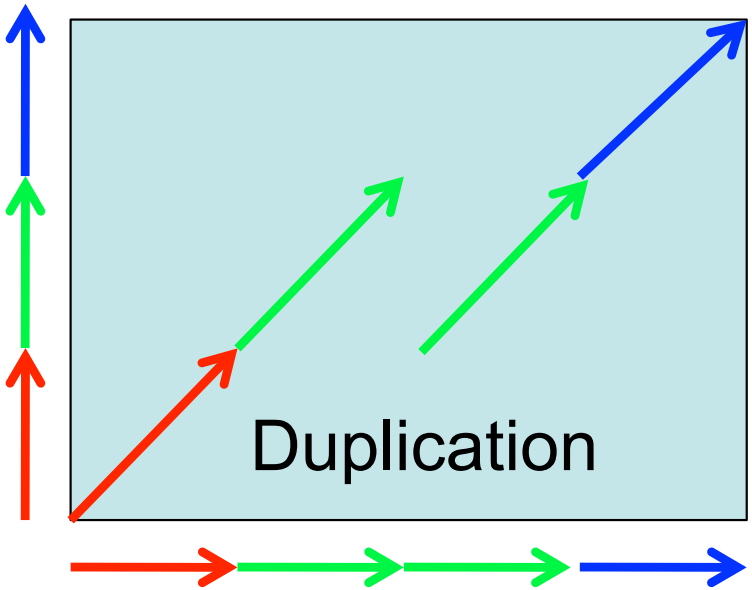
Block Rearrangements are a Problem!

The simple case:









Summary

- *Optimal MSA computation is intractible* in the general case
- *Progressive alignment* is more tractible, but is greedy
- *Iterative refinement* attempts to undo greedy decisions
- *PairHMM's* provide a principled way to perform pairwise steps
- *Felsenstein's algorithm* computes likelihoods on phylogenies
- Substitution models can use *continuous-time Markov chains*
- Large-scale *rearrangements* are a problem
- *Banding* can improve alignment speed