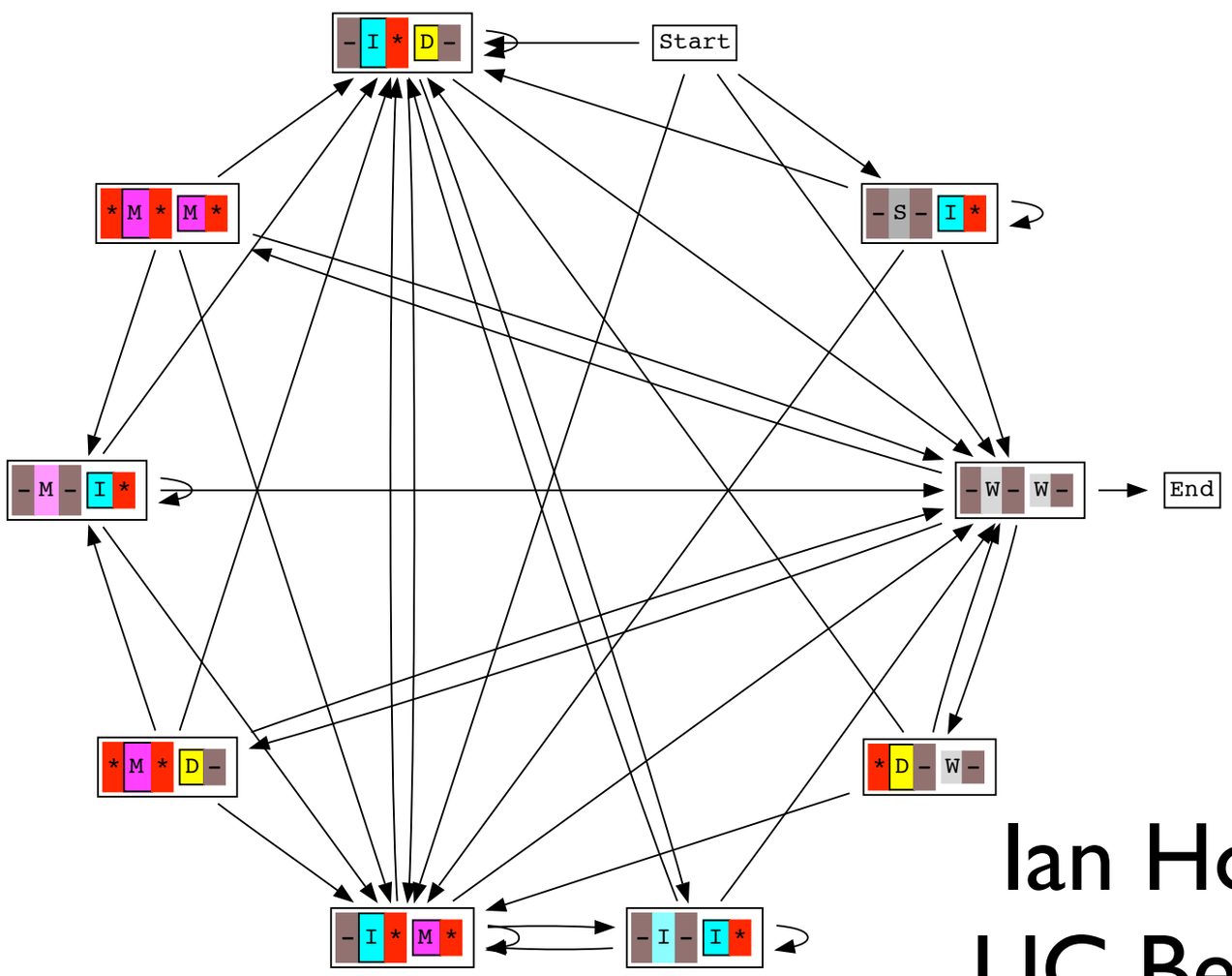
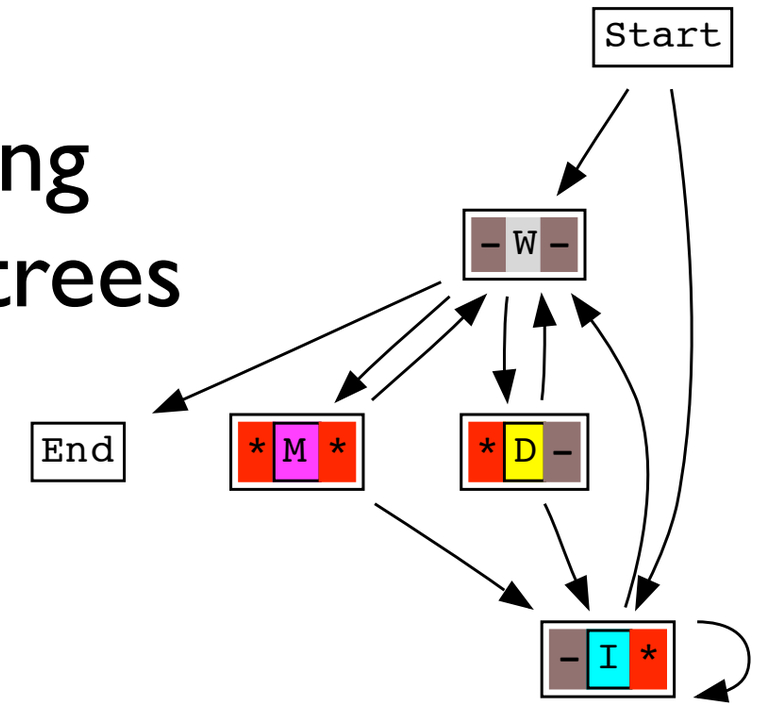
Workshop Announcement

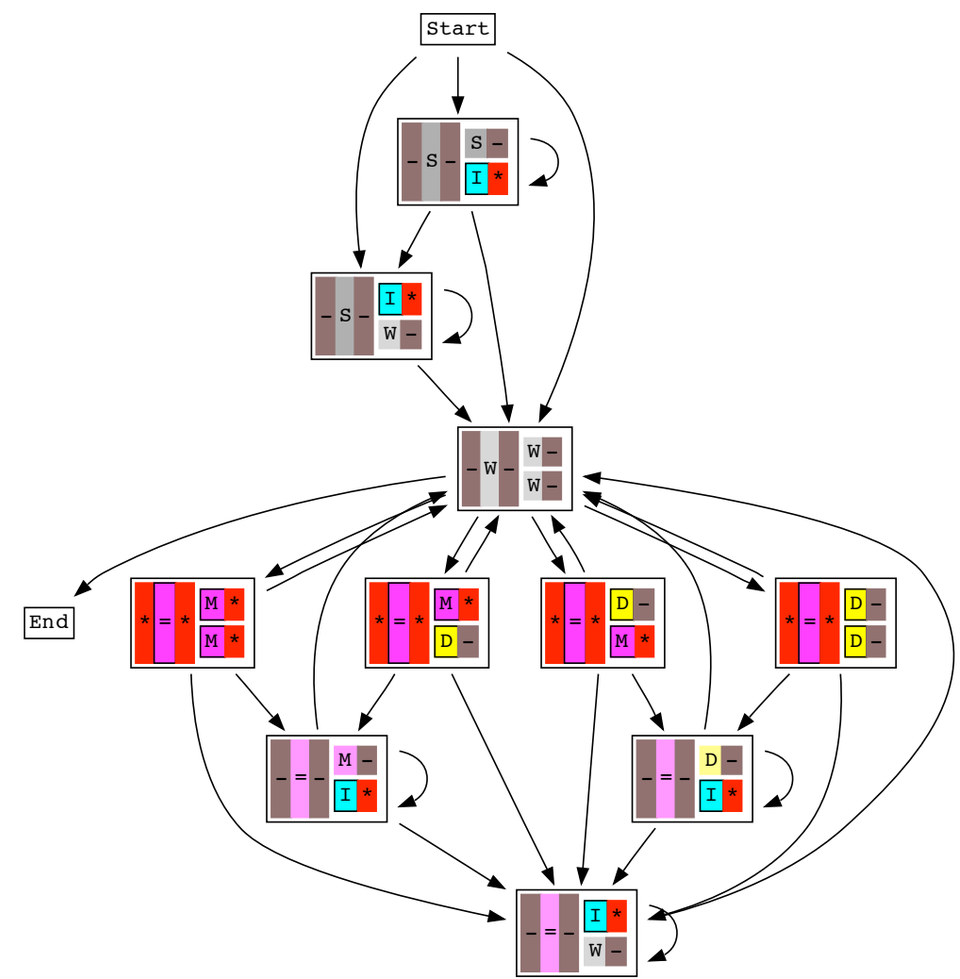
- **“Inference in Sequence Evolution”**
- Mathematical Biosciences Institute, Ohio State University
- First quarter of 2010 (exact date TBA)
- Organizers: I.Holmes & G.Lunter

Transducers

A probabilistic framework for modeling insertions & deletions on phylogenetic trees

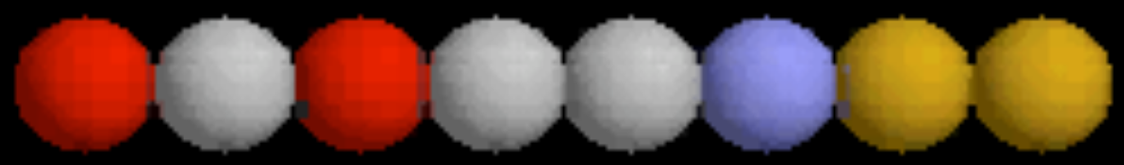






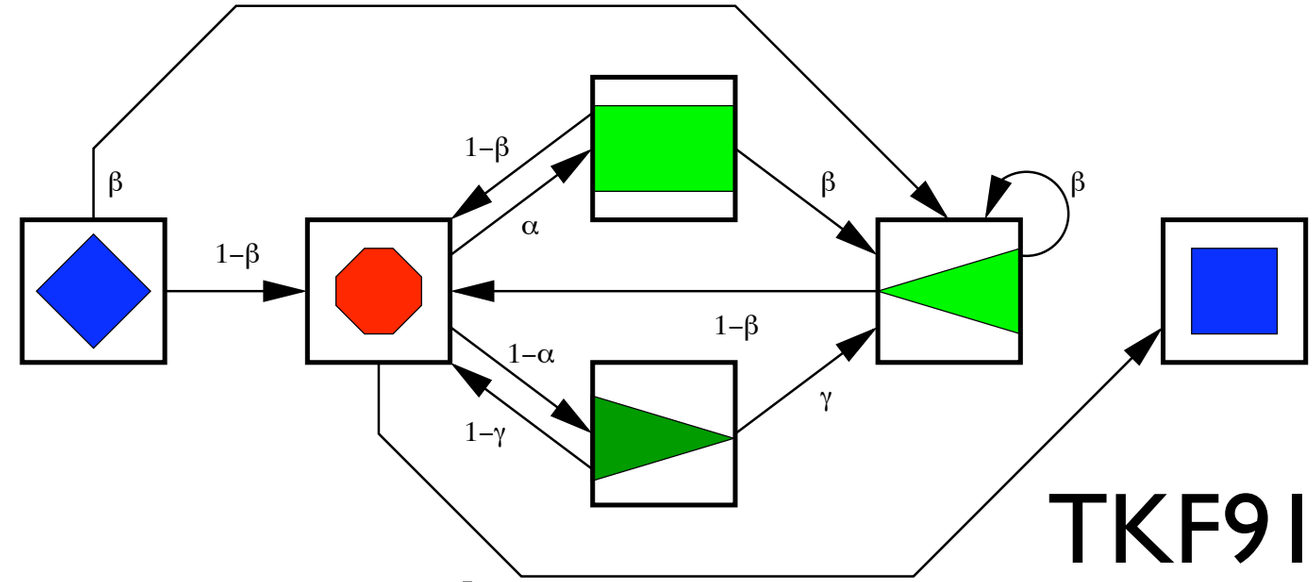
Ian Holmes
UC Berkeley
(Univ. Oxford)

Phylo-Alignment



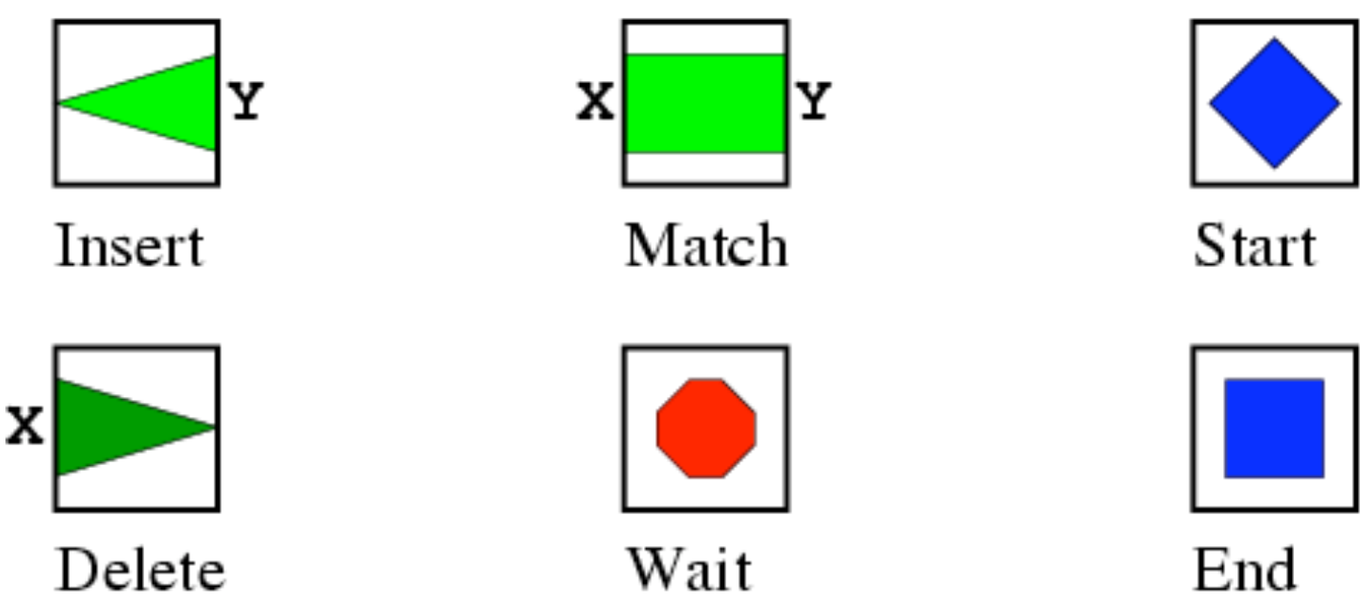
Phylo-Alignment

Transducers



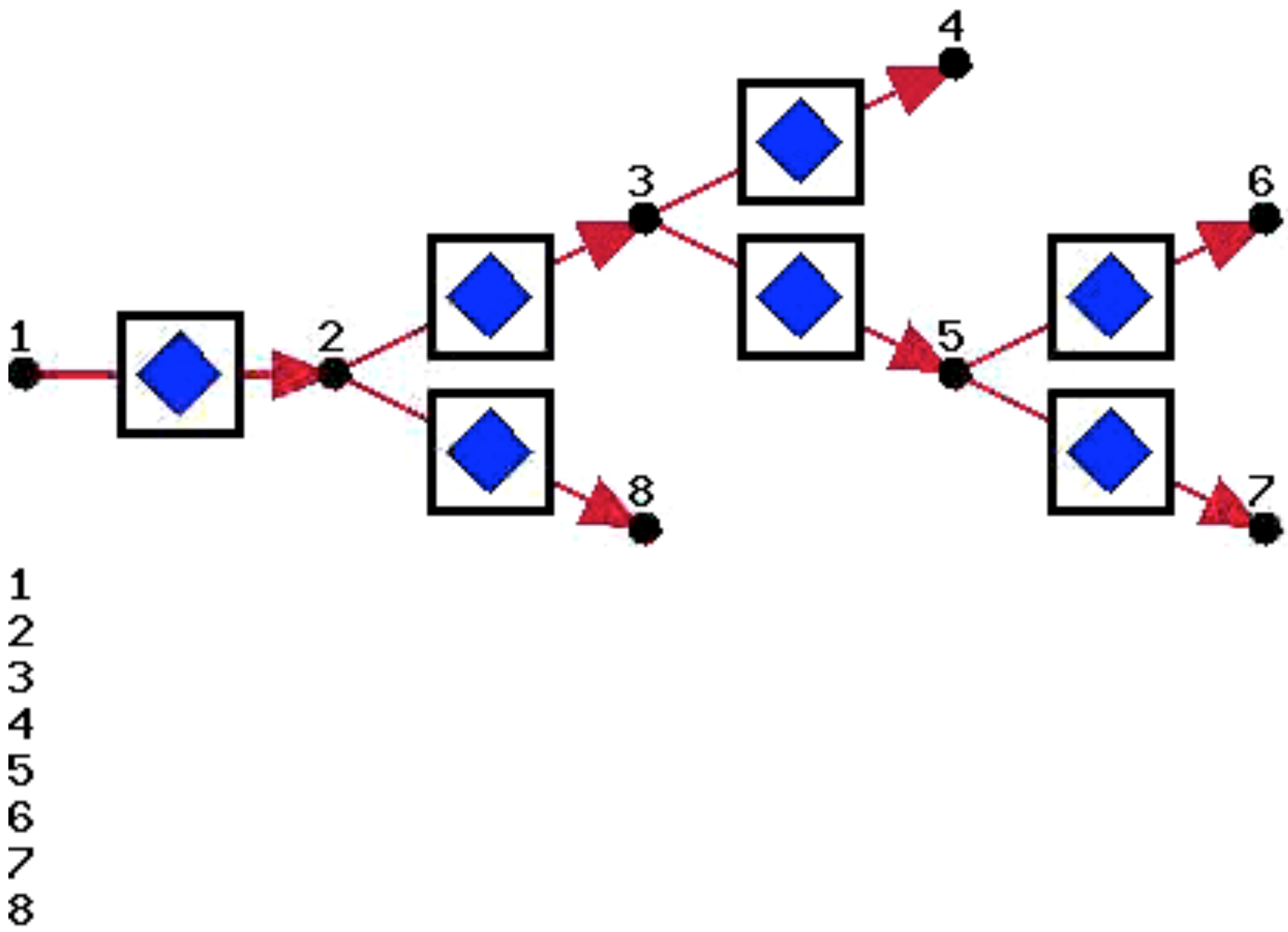
TKF91

- Operation of evolving a sequence along one branch of a phylogenetic tree
- Represent as a finite state machine
- Input "tape" X, output "tape" Y



Pair HMM: $P(X,Y)$
Transducer: $P(Y|X)$

String transducers on a tree

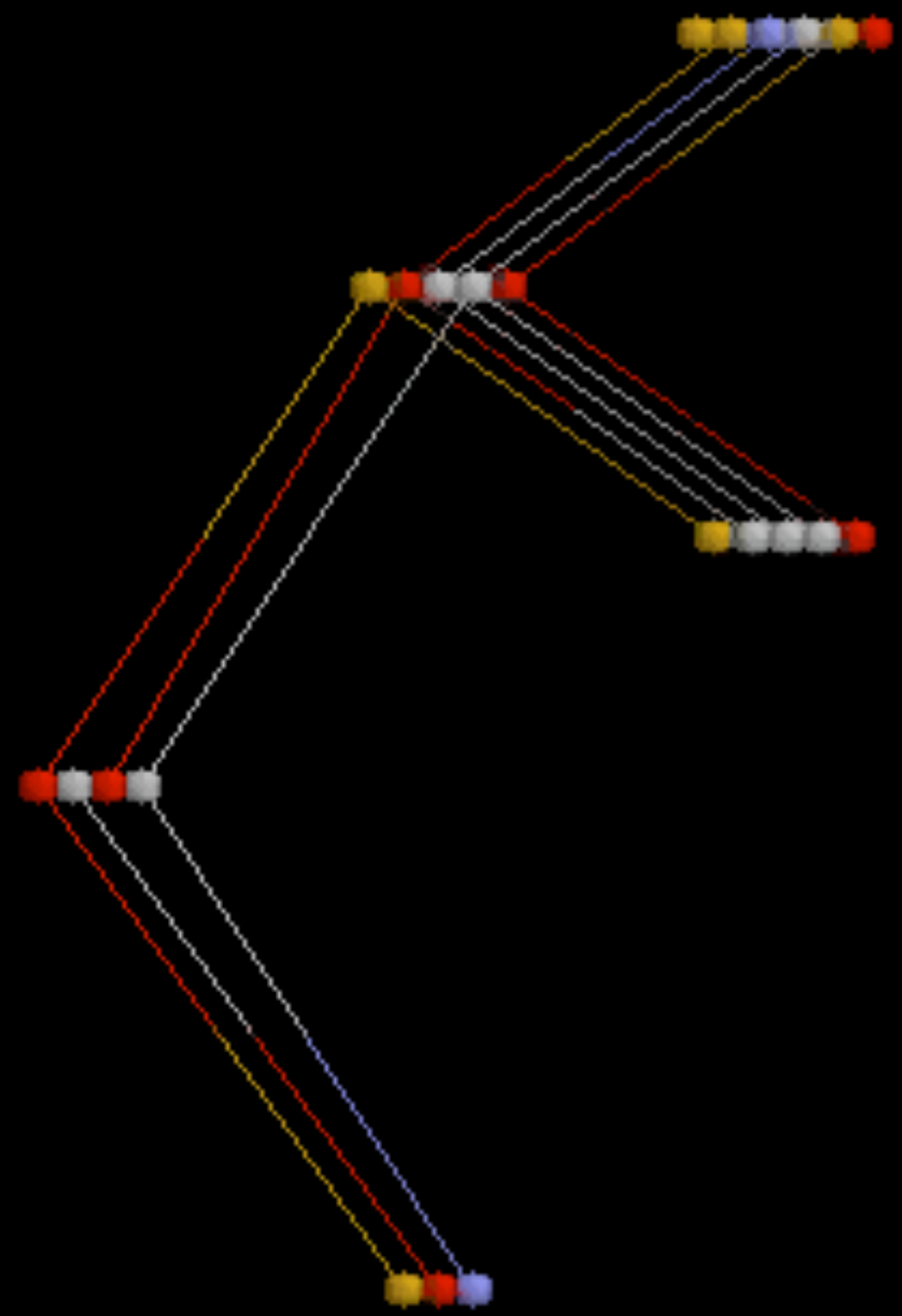


Mohri, Computational Linguistics, 1997
Hein, PSB, 2001; Holmes & Bruno, Bioinf., 2001
Holmes, Bioinformatics, 2003

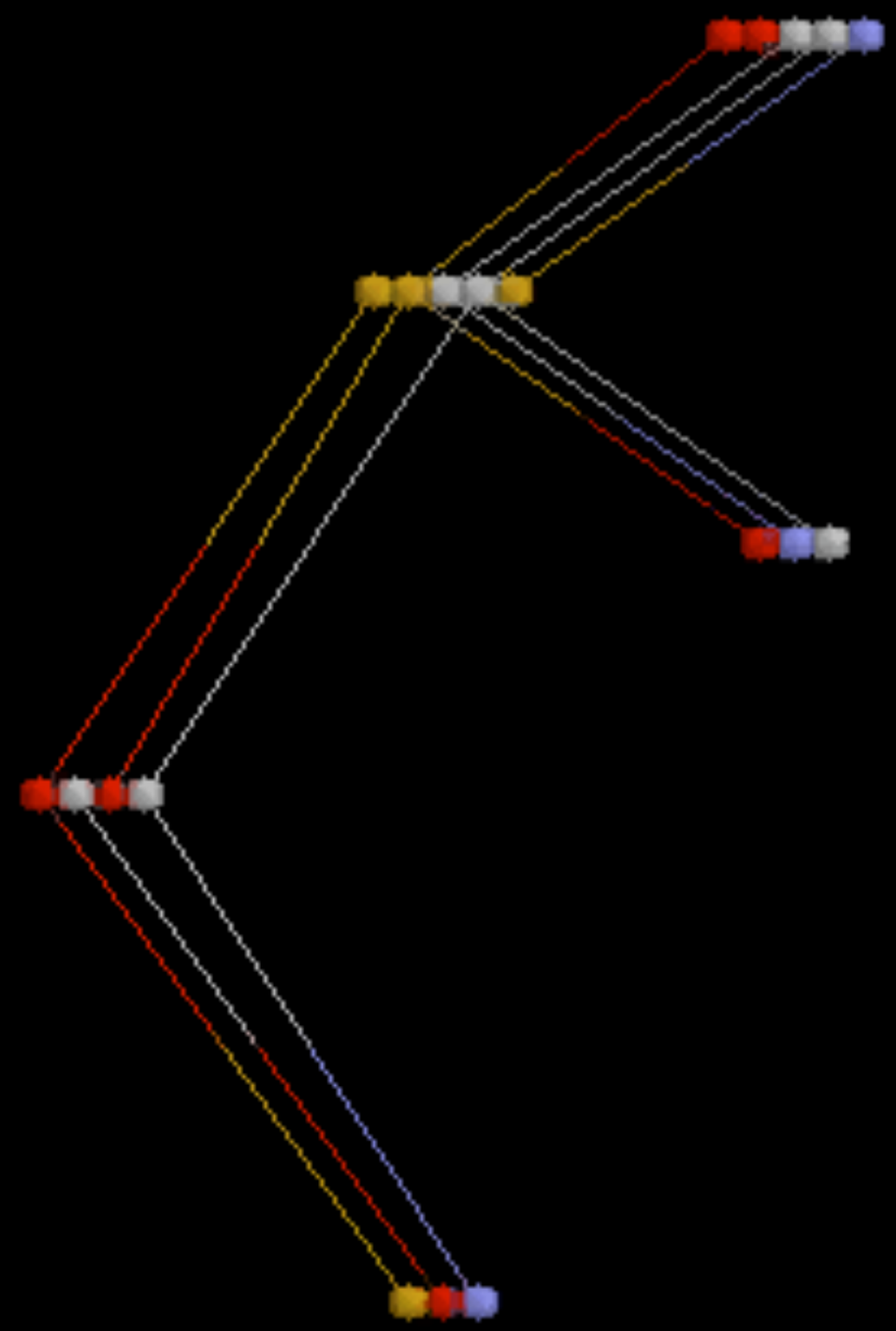
also Bishop & Thompson;
Thorne, Kishino & Felsenstein;
Steel; Lunter, Miklos; Kim & Sinha;
Satija, Pachter; Paten; Haussler ...

Curse of dimensionality

- Number of states in composite transducer increases exponentially with number of taxa
- So does the number of cells in the DP matrix
- Solution: **Markov Chain Monte Carlo**
  - Hold some parts of the state path fixed, while resampling other parts *(Gibbs sampling)*



Sampling one branch at a time



Sampling one node at a time

tkfalign *vs* Clustal (BAliBase)

BAliBASE subcategory	Prog.	Ref.	Iter.	CLUSTALW
Equidistant, similar lengths; high ID ($> 35\%$)	0.775	0.784	0.774	0.884
Equidistant, similar lengths; medium ID ($20\% - 40\%$)	0.673	0.689	0.693	0.790
Equidistant, similar lengths; low ID ($< 25\%$)	0.654	0.658	0.669	0.787
Close family ($> 25\%$) plus "orphan" outliers ($< 20\%$)	0.814	0.827	0.839	0.928
Divergent subfamilies ($< 20\%$ between subfamilies)	0.481	0.525	0.528	0.693
Long gaps at the ends: N/C terminal extensions	0.348	0.359	0.372	0.672
Long gaps in the middle: Insertions	0.573	0.603	0.622	0.789

Prog. = Progressive alignment
Ref. = Refinement
Iter. = MCMC + refinement

Holmes & Bruno, 2001

tkfalign *vs* Clustal (BAliBase)

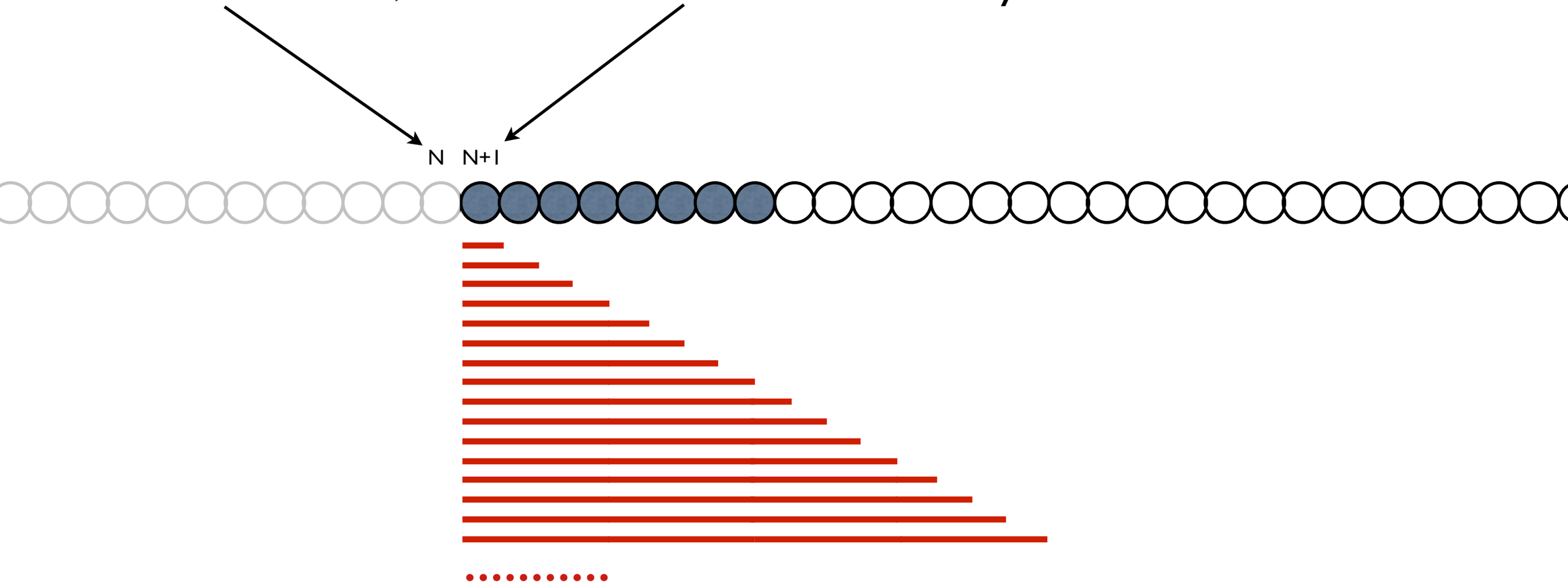
BAliBASE subcategory	Prog.	Ref.	Iter.	CLUSTALW
Equidistant, similar lengths; high ID ($> 35\%$)	0.775	0.784	0.774	0.884
Equidistant, similar lengths; medium ID ($20\% - 40\%$)	0.673	0.689	0.693	0.790
Equidistant, similar lengths; low ID ($< 25\%$)	0.654	0.658	0.669	0.787
Close family ($> 25\%$) plus "orphan" outliers ($< 20\%$)	0.814	0.827	0.839	0.928
Divergent subfamilies ($< 20\%$ between subfamilies)	0.481	0.525	0.528	0.693
Long gaps at the ends: N/C terminal extensions	0.348	0.359	0.372	0.672
Long gaps in the middle: Insertions	0.573	0.603	0.622	0.789

Prog. = Progressive alignment
Ref. = Refinement
Iter. = MCMC + refinement

Holmes & Bruno, 2001

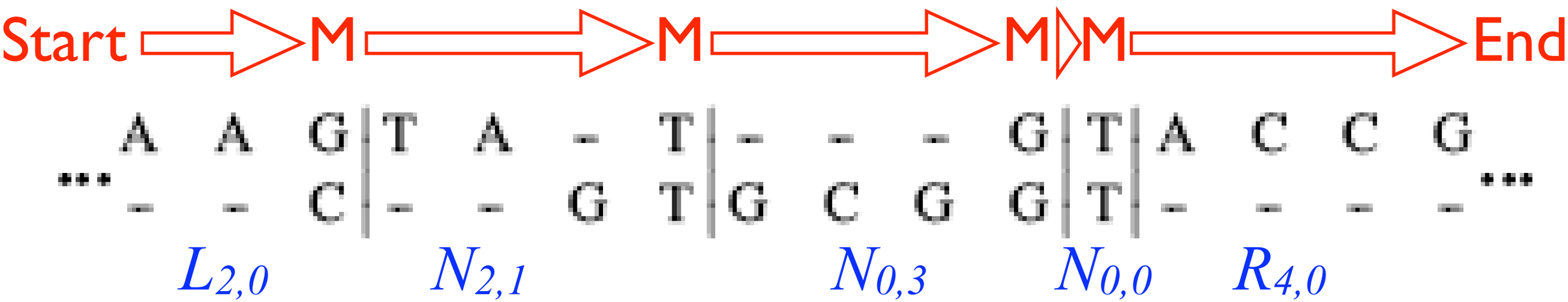
The “Long Indel” Model

Deletion events are “attached” to the leftmost residue that they remove.
If residue N survives, then residue N+1 is unaffected by deletion events from the left



Miklòs, Lunter & Holmes, 2004

Independent "chop zones"



$$
\begin{aligned}
L_{2,0} &\times p_t(\mathrm{G} \rightarrow \mathrm{C}) \\
&\times N_{2,1} \times q(\mathrm{G}) p_t(\mathrm{T} \rightarrow \mathrm{T}) \\
&\times N_{0,3} \times q(\mathrm{G}) q(\mathrm{C}) q(\mathrm{G}) p_t(\mathrm{G} \rightarrow \mathrm{G}) \\
&\times N_{0,0} \times p_t(\mathrm{T} \rightarrow \mathrm{T}) \\
&\times R_{4,0}
\end{aligned}
$$

Table 1
Symbolic Representation of the Four Types of Chop Zone According to Whether They Adjoin the Left and/or Right Flanking Sequence (•••) and Notation for Their Probabilities

$L_{ij} = P\Big($	$\cdots$	$\#^i$	$-^j$	M $\Big\|\Big)$	$N_{ij} = P\Big($	$\#^i$	$-^j$	M $\Big\|\Big)$
		$-^i$	$\#^j$	M		$-^i$	$\#^j$	M
$R_{ij} = P\Big($	$\#^i$	$-^j$	$\cdots\Big)$		$B_{ij} = P\Big($	$\cdots$ $\#^i$	$-^j$	$\cdots\Big)$
	$-^i$	$\#^j$				$-^i$	$\#^j$	

NOTE.—These probabilities are conditional on observing the i (or $i+1$) ancestral nucleotides. The # signs represent unaligned residues; M pairs represent aligned residues, and vertical bars represent chop zone boundaries.

Miklòs, Lunter & Holmes, 2004

Generalized Pair HMM

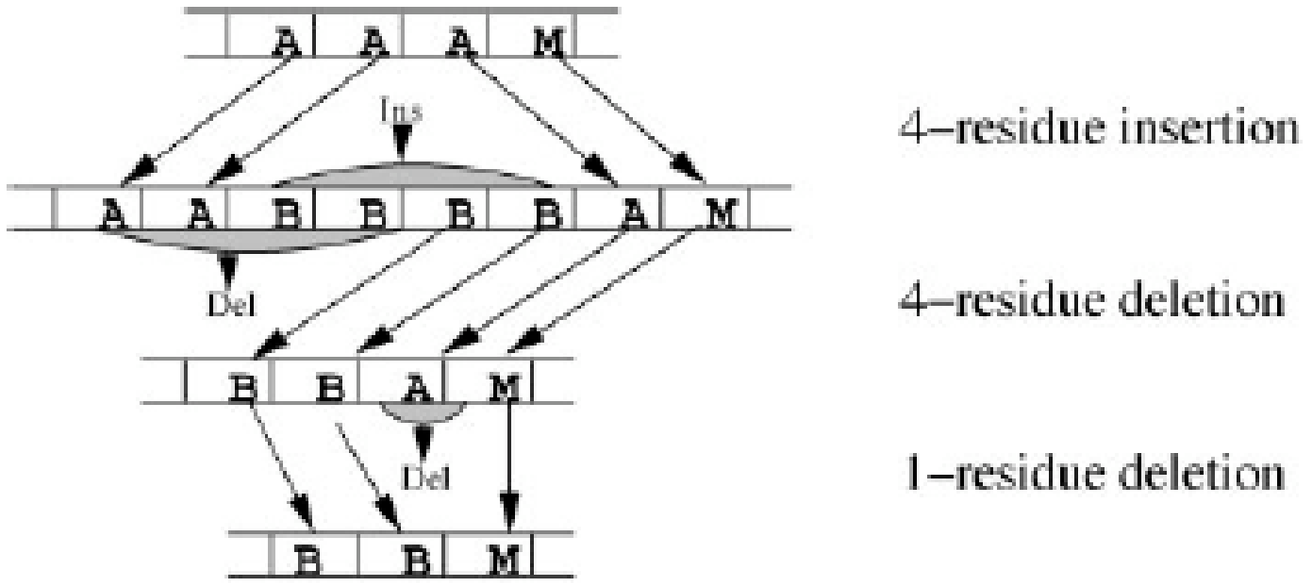


FIG. 1.—An example three-event trajectory for a zone that changes length from four resides to three (outcome $B^i_{3,2}$; see section titled *Algorithm*). By definition, the final ancestral residue in the zone (the M) cannot be deleted, whereas every other ancestral residue (the A's) *must* be deleted.

Gap penalties calculated by direct enumeration of trajectories (shown left: $N_{3,2}$)

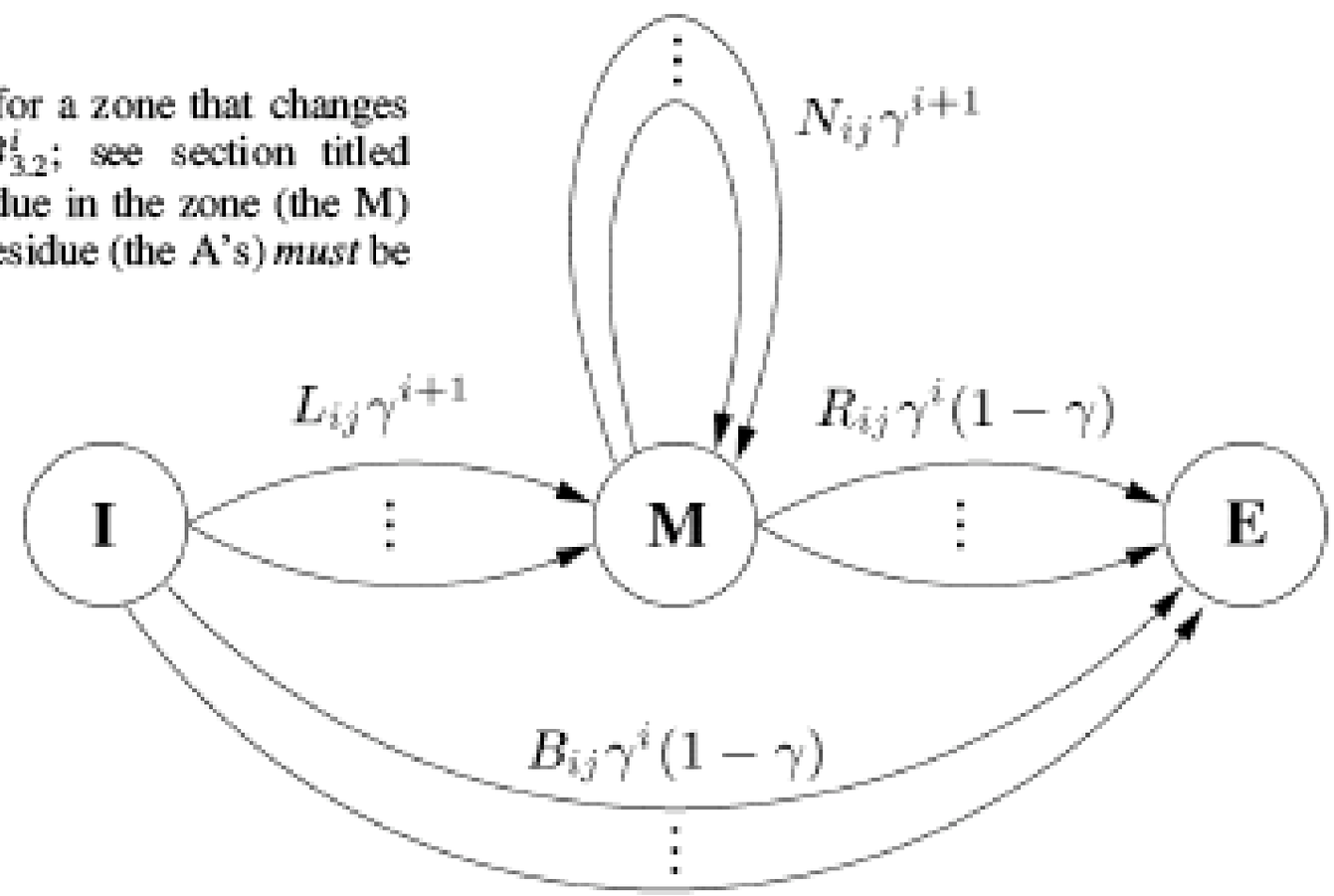


FIG. 3.—A hidden Markov model formulation of the long indel model. The emission probabilities (associated to transitions) are not included. The parameter $\gamma = \lambda_1/\mu_1$ is the parameter governing the geometric equilibrium length distribution.

Shown right: the generalized Pair HMM for the "long indel" model

Long Indel *vs* Gotoh (Homstrad)

Table 3
Performance of Alignment Methods, as Measured by Alignment Accuracy or "Overlap," the Percentage of Alignment Columns Identical to Those of the HOMSTRAD Structural Alignments

Alignment Method	Training Set Optimization[a]	Test Set Overlap (%)
TKF91	ML	73.8
TKF92	ML	75.9
Gotoh (BLOSUM62)	NCBI defaults	80.9
Long indel	ML	81.1
Long indel, mixed geometric	Accuracy	82.1
Gotoh (BLOSUM62)	Accuracy	82.2

[a] Parameters were optimized over a training set to maximize either likelihood or overlap. In addition, for the Gotoh algorithm we used NCBI (National Center for Biotechnology Information) defaults for gap opening and gap extension parameters.

Miklòs, Lunter & Holmes, 2004

“Every good work of software starts by scratching a developer’s personal itch” - Eric Raymond

2001 (Holmes & Bruno)
MCMC based on TKF91. **Poor performance due to affine gaps, rate variation**

2002 (Holmes & Rubin)
EM algorithm for estimating substitution rates & in particular rate variation

2003 (Holmes)
General algorithm for transducer composition on phylogenetic trees

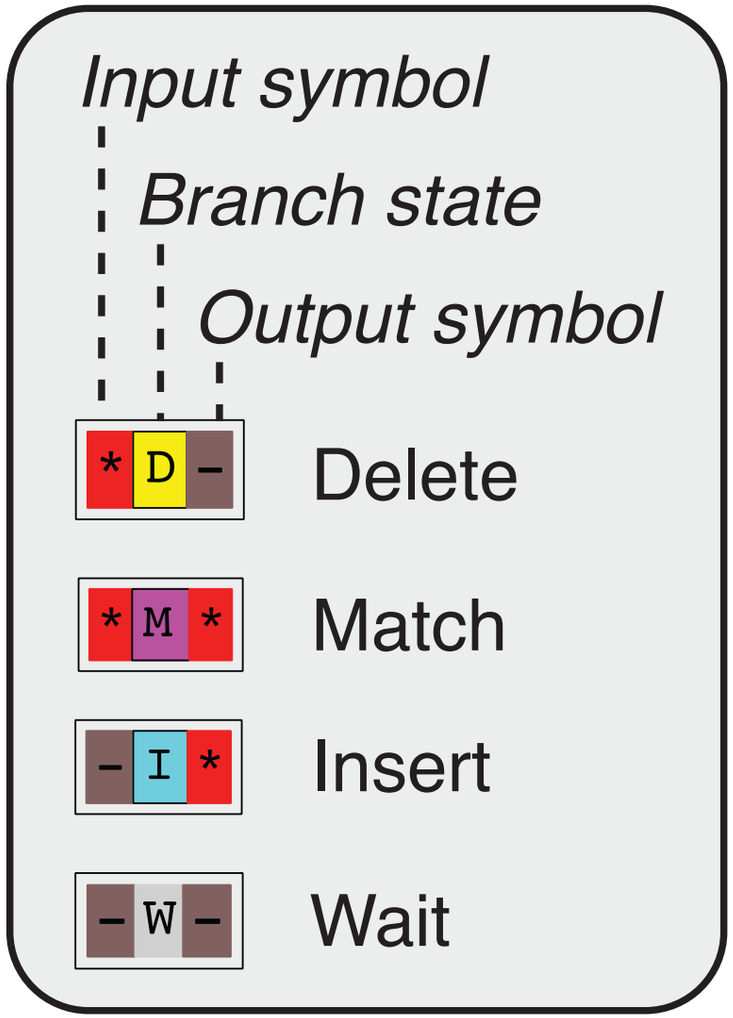
2004 (Miklòs, Lunter & Holmes)
“Long Indel” model: affine-gap Pair HMMs from evolutionary models

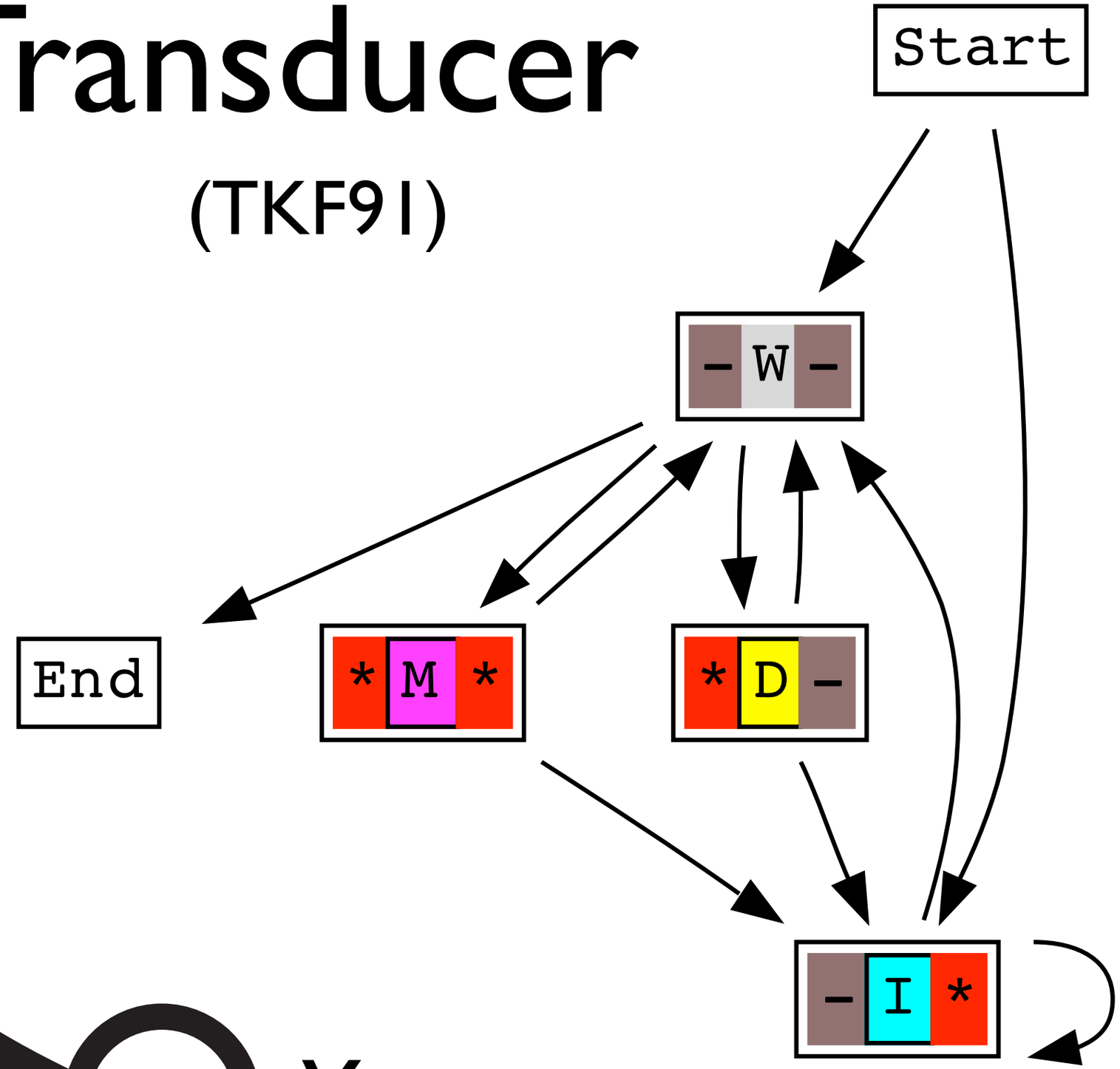
2005-2006 (Holmes; Klosterman *et al*)
(Started extending transducer theory to RNA sequence analysis & SCFGs)
Enormous difficulties debugging transducer composition & sampling algorithms!

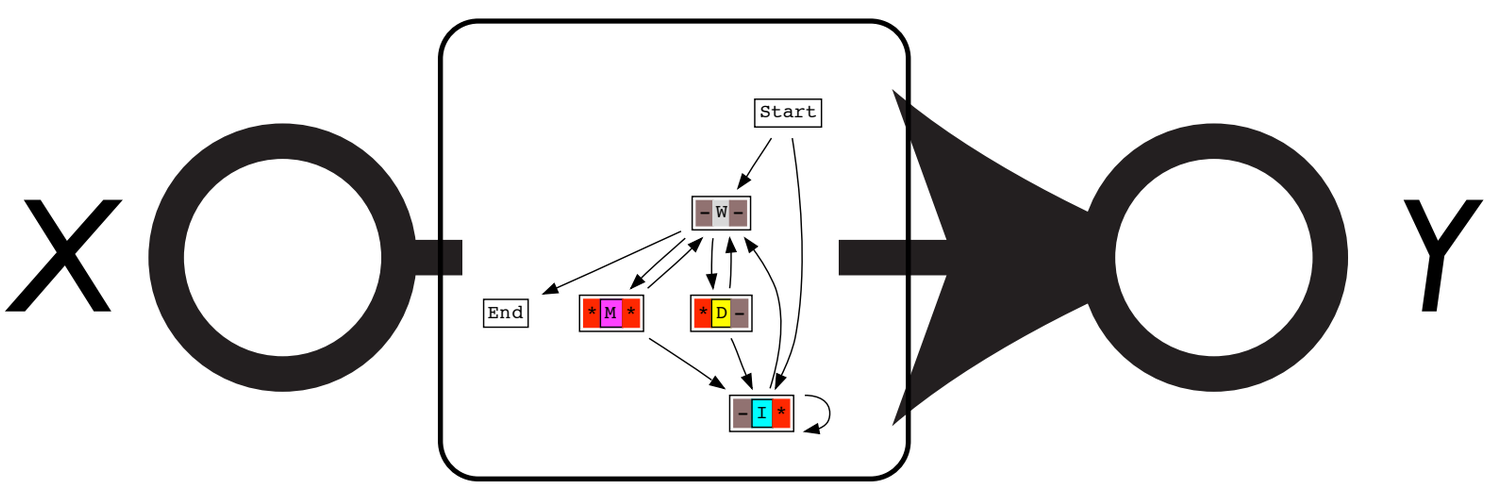
2007 (Holmes)
Phylocomposer

Transducer

(TKF91)

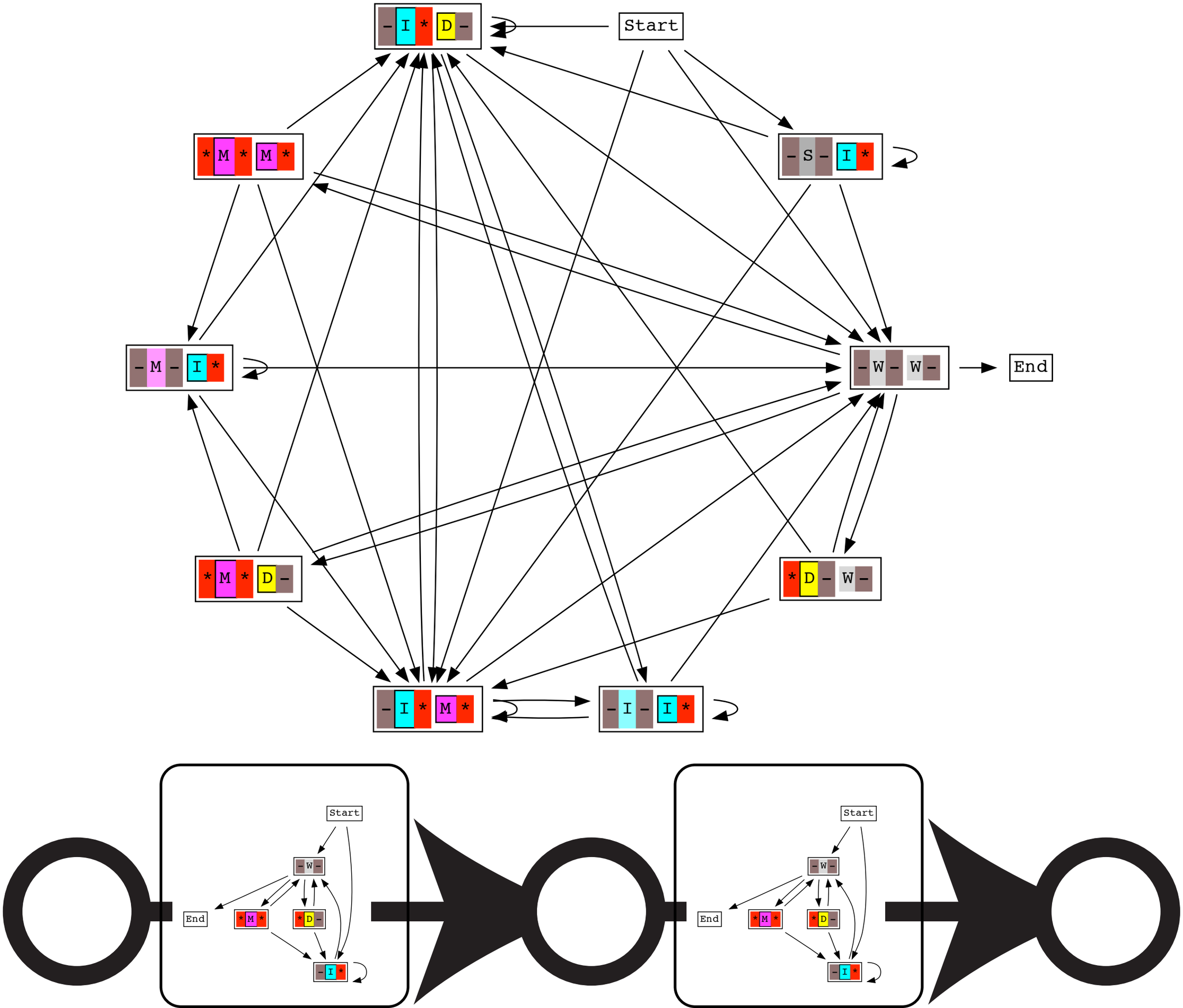


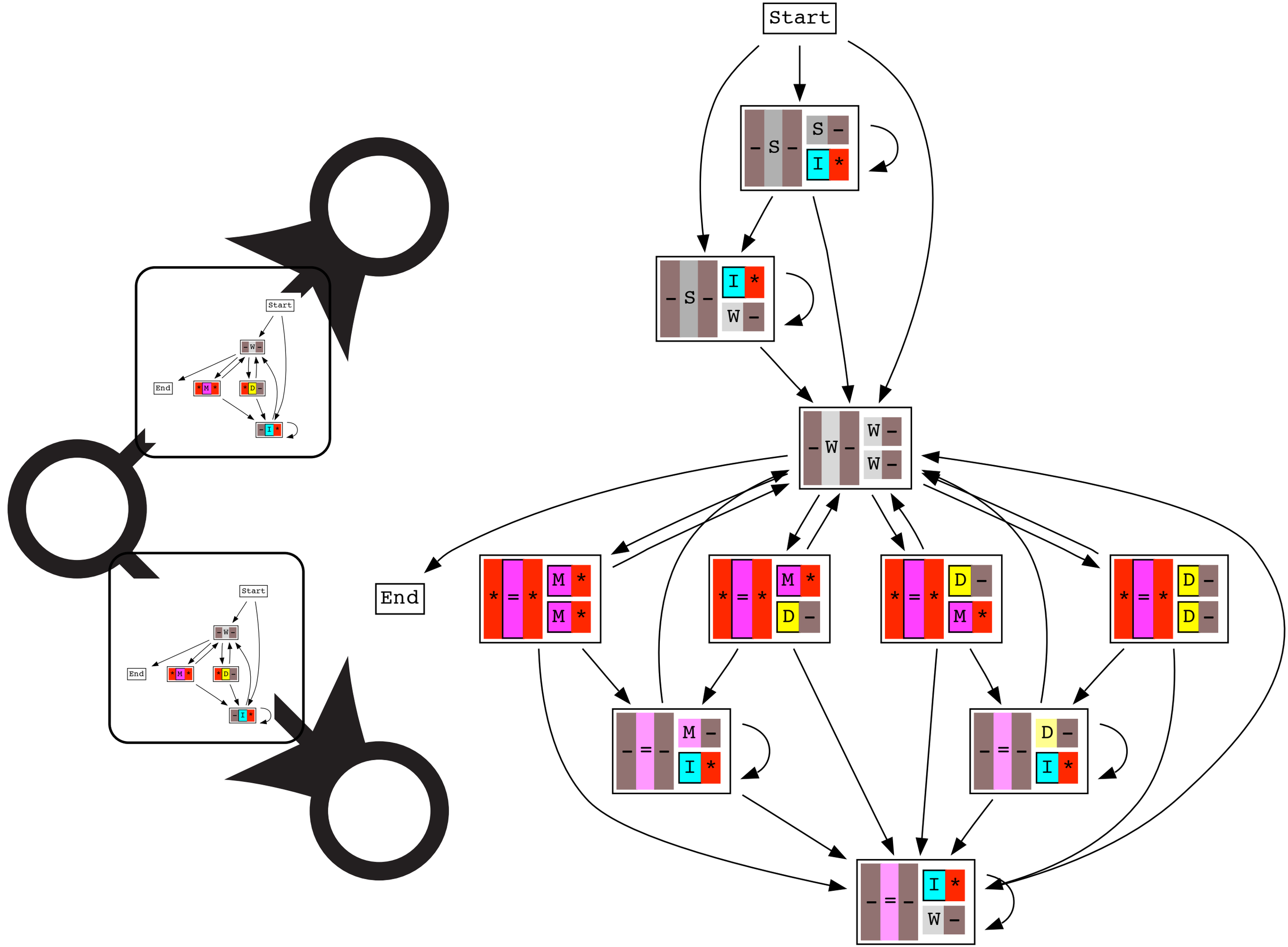


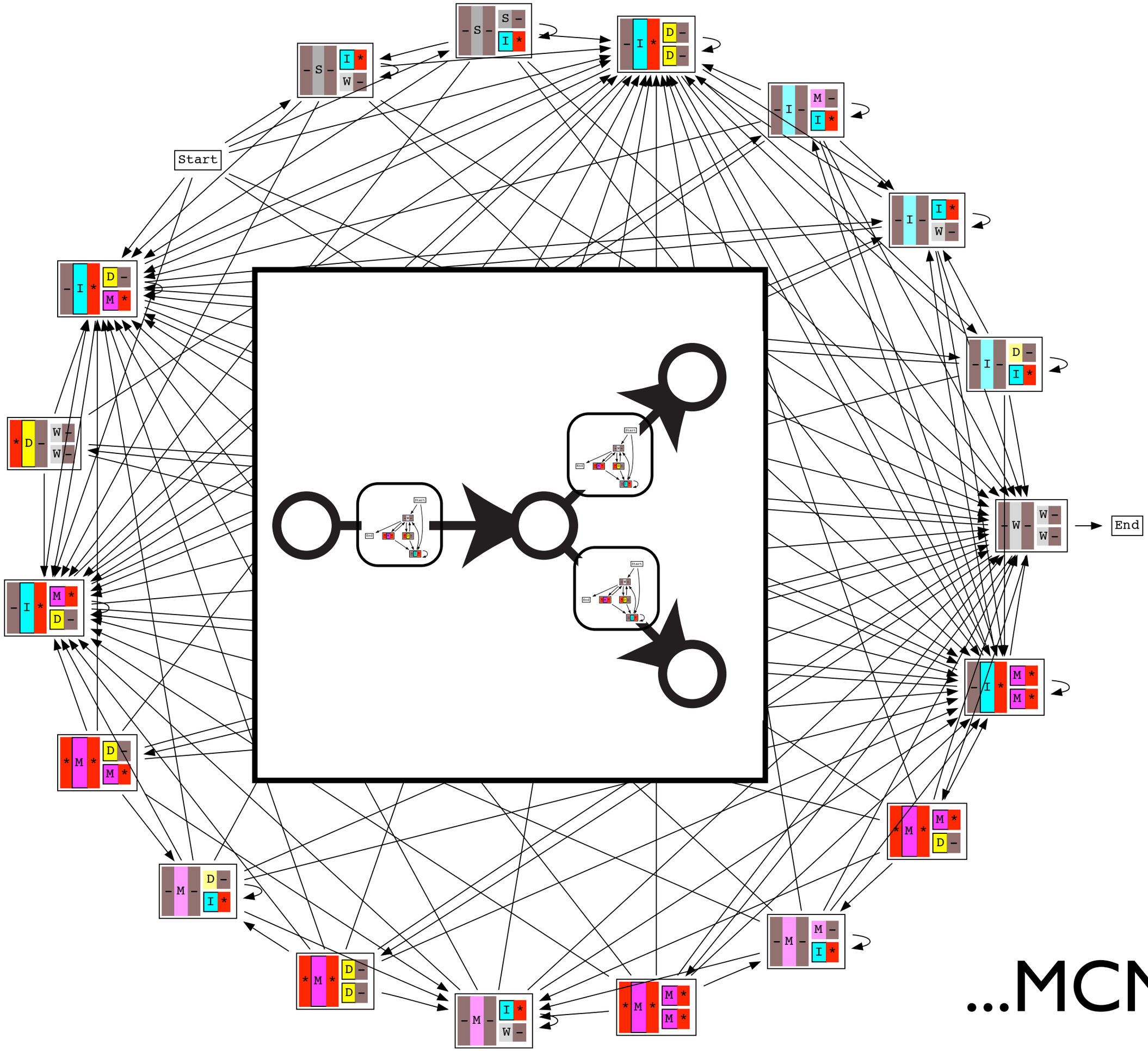


Pair HMM: $P(X,Y)$

Transducer: $P(Y|X)$

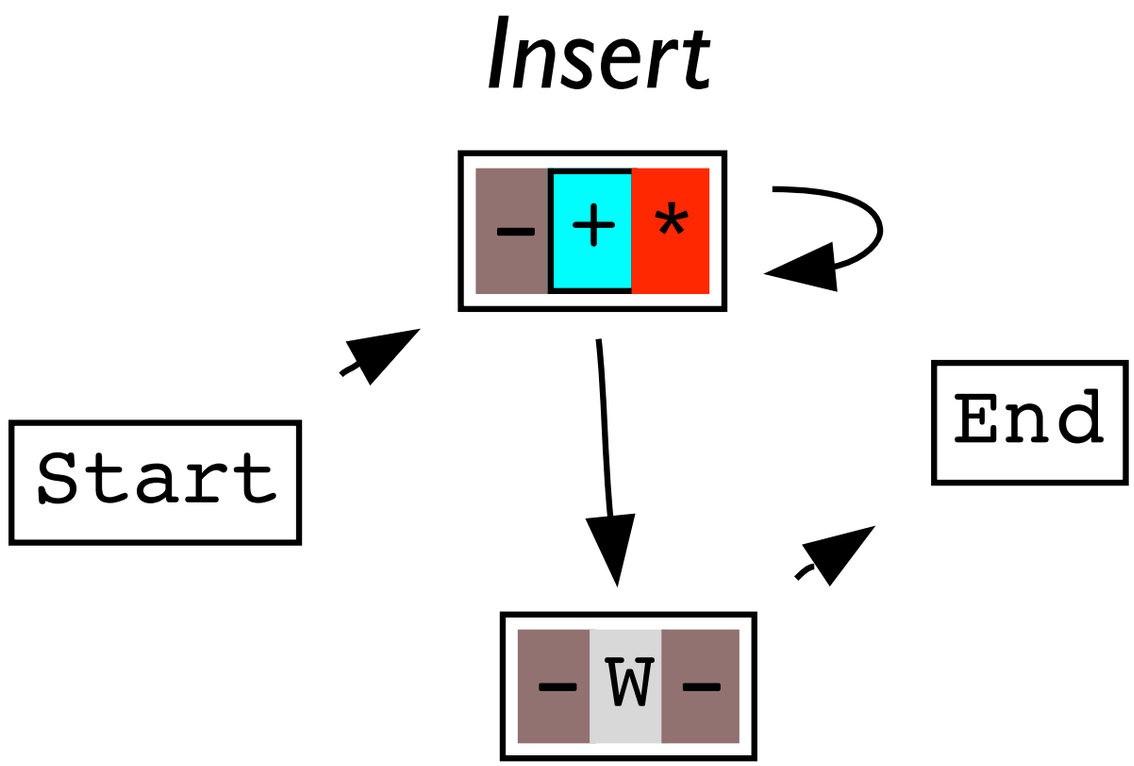

Start
End
Start
End
Start
End


Start
- S -
S -
I *
- S -
I *
W -
- W -
W -
W -
End
* = *
M *
M *
* = *
M *
D -
* = *
D -
M *
* = *
D -
D -
- = -
M -
I *
- = -
D -
I *
- = -
I *
W -
Start
- W -
End
* M *
* D -
- I *
Start
- W -
End
* M *
* D -
- I *



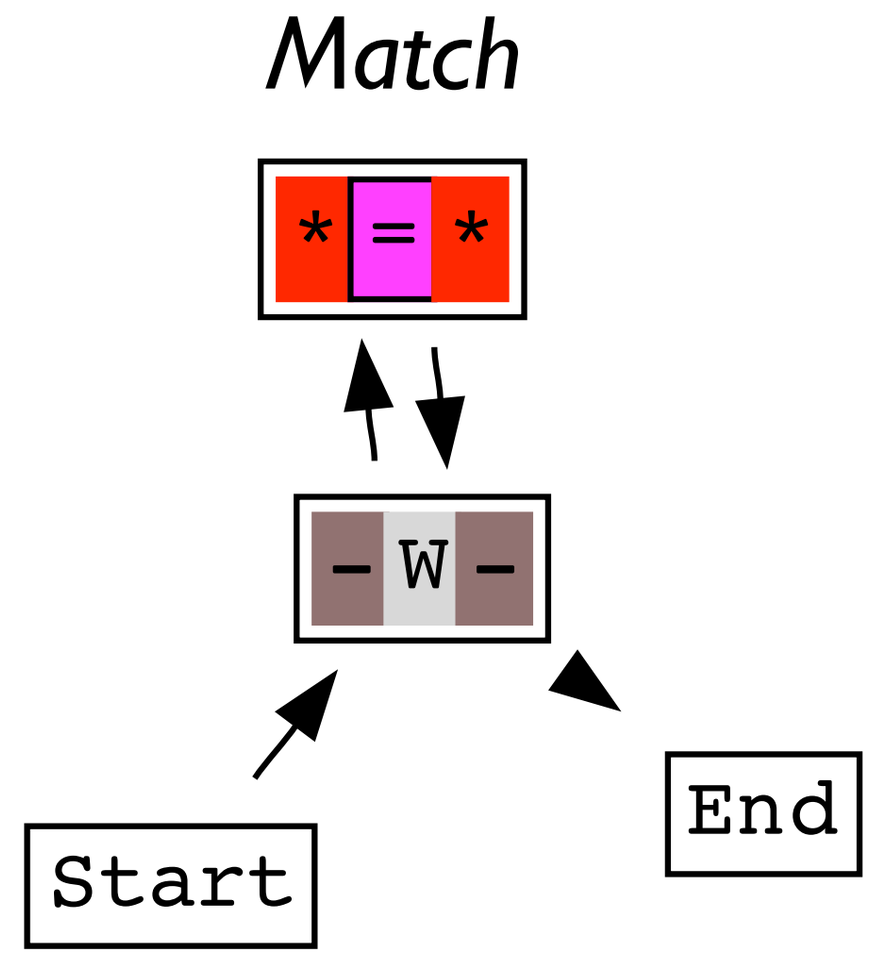
...MCMC

Singleton transducer

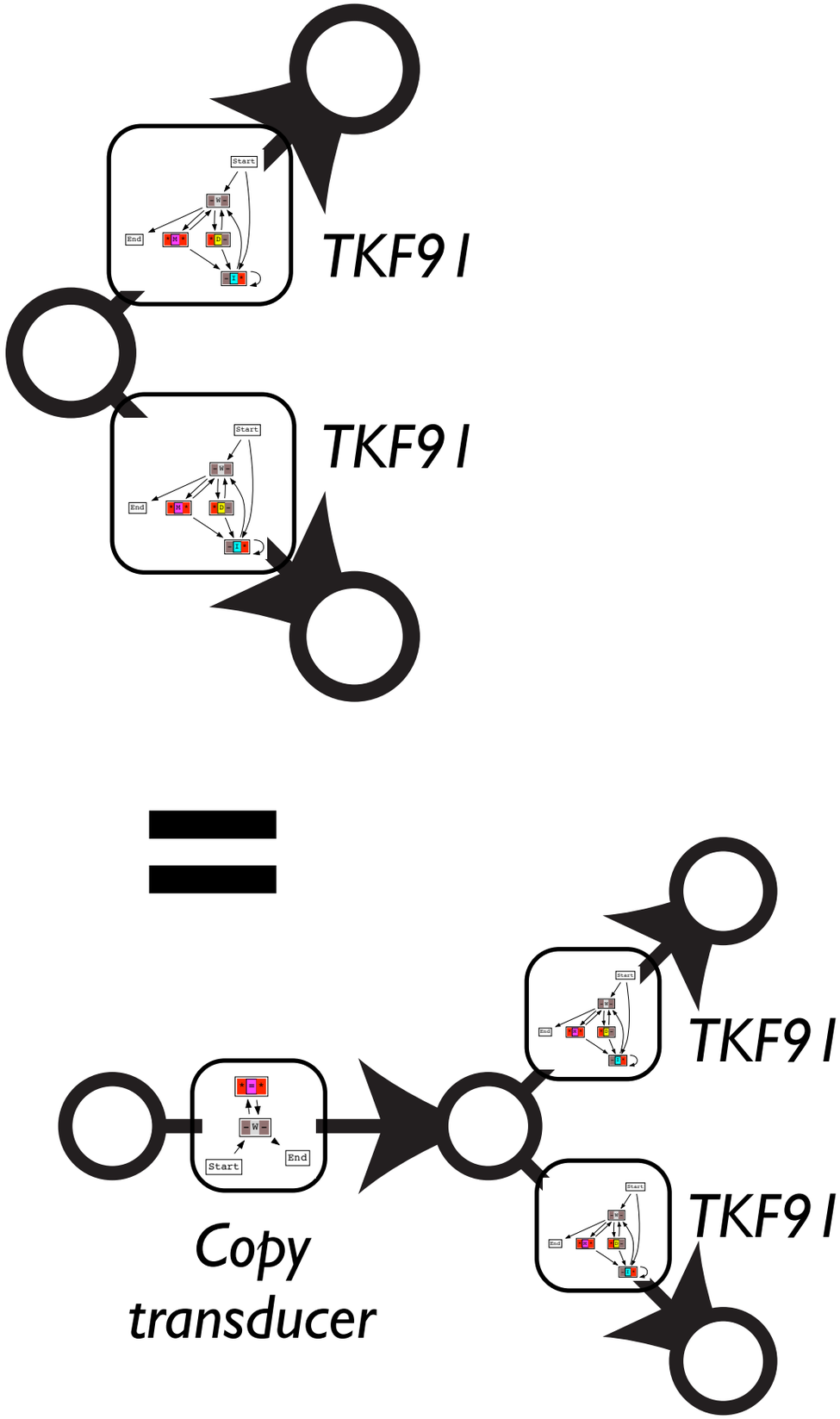


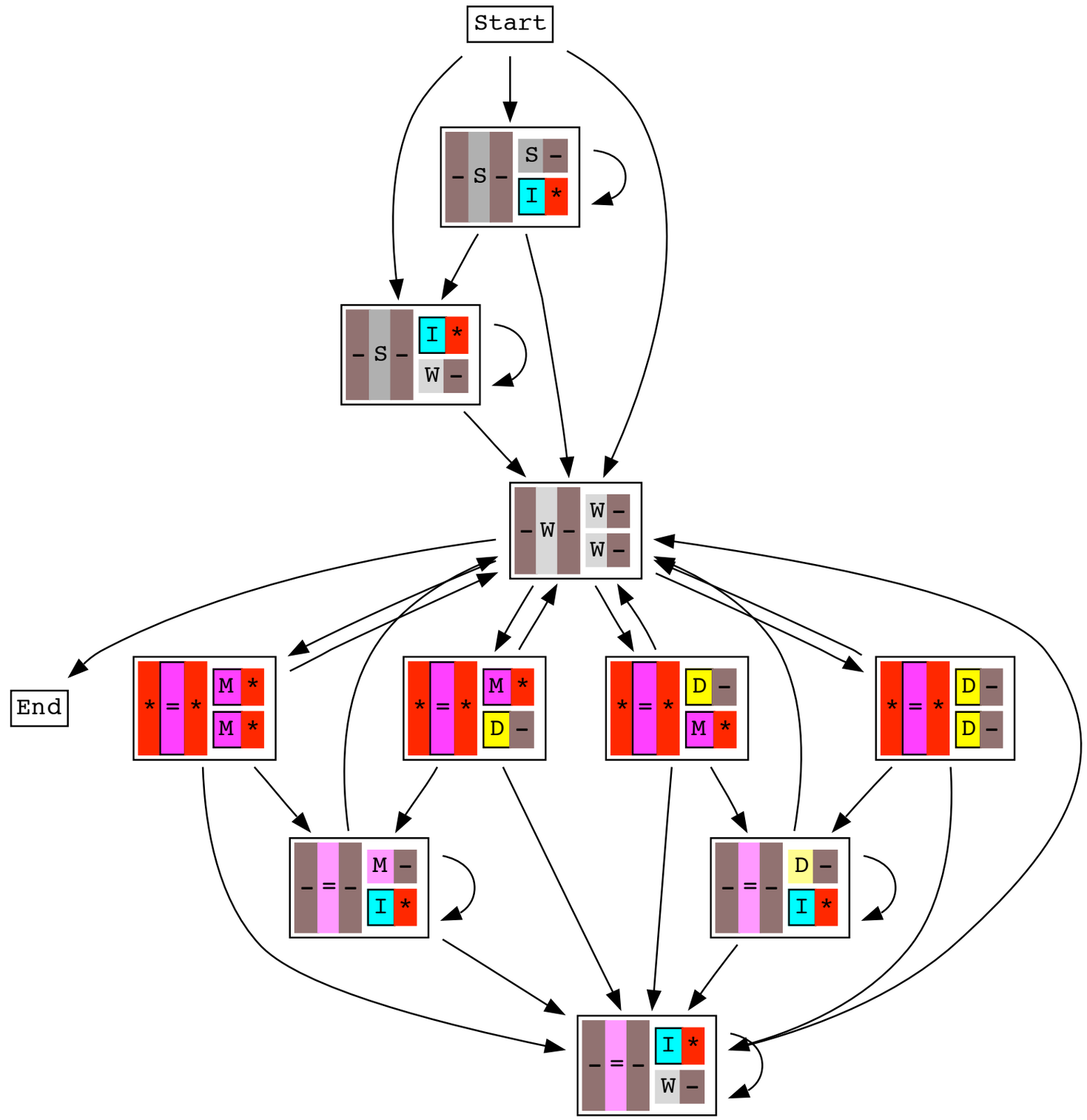
(Equilibrium distribution of TKF91 model)

Copy transducer



(Zero-length branch in TKF91 model)


TKF91
TKF91
=
Copy
transducer
TKF91
TKF91

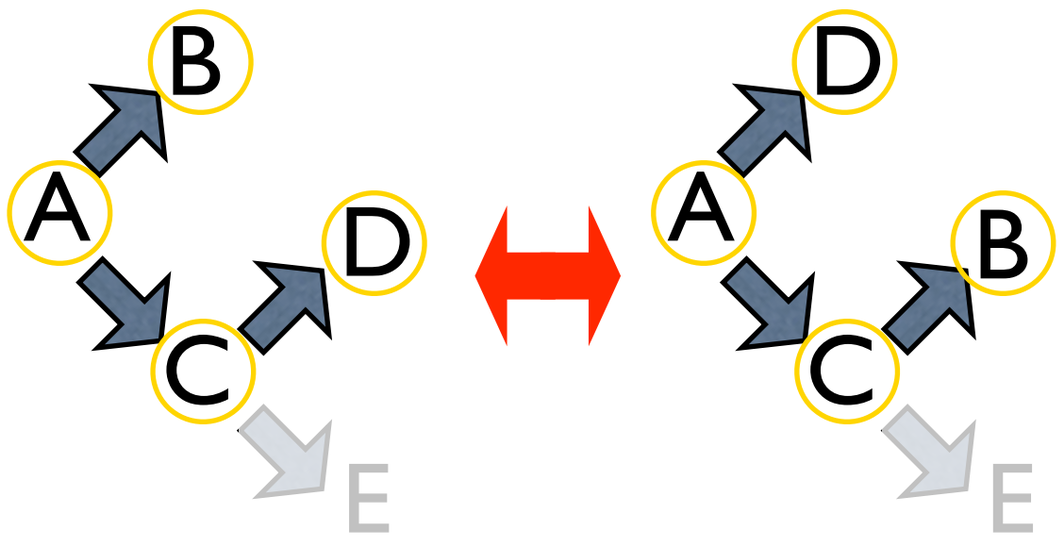

Start
End

HMMoC adapter

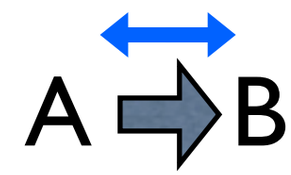
- Gerton Lunter's **HMMoC**
- Hidden Markov Model Compiler
- Speedup factor 10^2-10^3

Handel MCMC moves

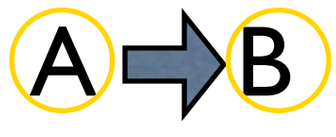
`tkfalign`
`handalign`



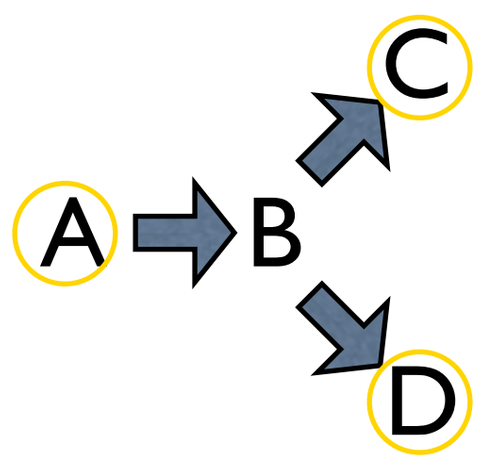
Try a change in tree topology (*aunt attack*); sample alignment ABCD for both configurations



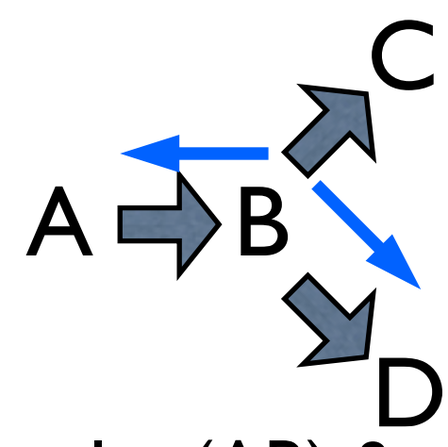
Sample t(AB) (*branch scaling*)



Sample alignment AB (*branch realignment*)

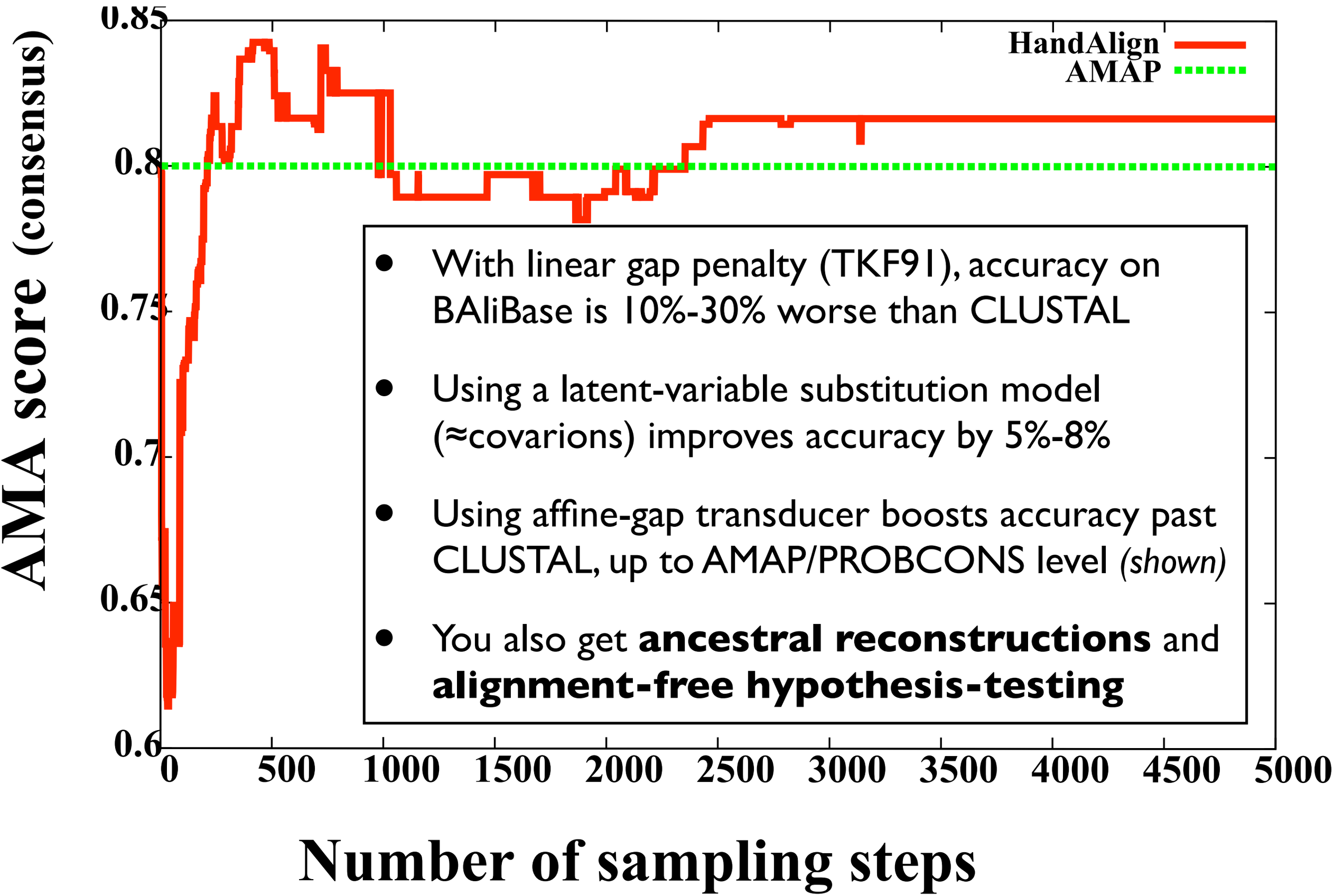


Sample alignment ABCD and sequence at B (*node realignment*)



Sample t(AB) & t(BD), keeping the sum constant (*branch sliding*)

MCMC Statistical Alignment



- With linear gap penalty (TKF91), accuracy on BAliBase is 10%-30% worse than CLUSTAL
- Using a latent-variable substitution model (≈covarions) improves accuracy by 5%-8%
- Using affine-gap transducer boosts accuracy past CLUSTAL, up to AMAP/PROBCONS level *(shown)*
- You also get **ancestral reconstructions** and **alignment-free hypothesis-testing**