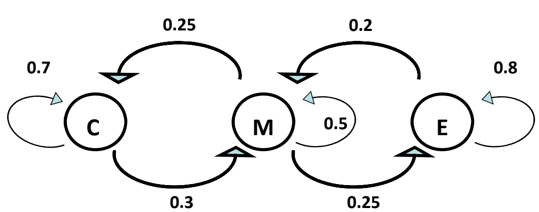
HIDDEN MARKOV MODELS
Profile HMMs

11/10/2022

How HMMs solve the problems with PSSMs

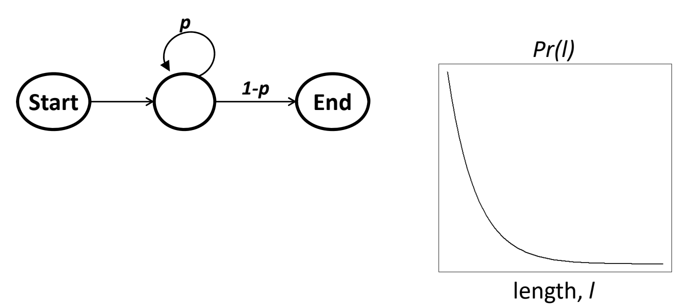
- Do not capture positional dependencies
- Hard to recognize pattern instances that contain indels
- ➢ Variable length motifs
- Do not handle boundary detection problems well

Hidden Markov Models can emit variable length sequences



i	C	M	E
π_i	0.5	0	0.5
$e_i(H)$	0.3	0.9	0.2
$e_i(L)$	0.7	0.1	0.8

Caveat emptor: Model topology influences the distribution of lengths of emitted sequences

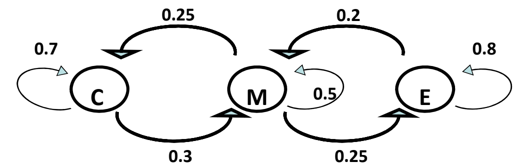


How HMMs solve the problems with PSSMs

- Do not capture positional dependencies
- Hard to recognize pattern instances that contain indels
- Variable length motifs
- ➢ Do not handle boundary detection problems well

Boundary detection:

- Label sequences using Viterbi or posterior decoding
- State path gives pattern boundaries



HHLHLHLHLLLHHLHLHHHHHHHLHHHHHHHHHHHHLHLHLLHLHHLH..

CCCCCCCCCCCCCCCCMMMMMMMMMMMMMMMMMMMMEEEEEEEEEEEE..

How HMMs solve the problems with PSSMs

- ➢ Do not capture positional dependencies
- Hard to recognize pattern instances that contain indels
- Variable length motifs
- Do not handle boundary detection problems well

A PSSM for the WEIRD motif

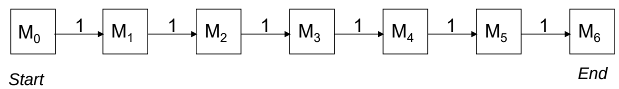
WEIRD
WEIRD
WEIQH
WEIRD
WEIQH

→

D	0.08	0.08	0.08	0.08	0.33
E	0.08	0.54	0.08	0.08	0.08
H	0.08	0.08	0.08	0.08	0.25
I	0.08	0.08	0.54	0.08	0.08
Q	0.08	0.08	0.08	0.25	0.08
R	0.08	0.08	0.08	0.33	0.08
W	0.54	0.08	0.08	0.08	0.08

$$q[x,j] = \frac{c[x,j] + b}{k + b|\Sigma|}$$

An HMM that is equivalent to a PSSM for the WEIRD motif



Emission probabilities

	M1	M2	M3	M4	M5
D	0.08	0.08	0.08	0.08	0.33
E	0.08	0.54	0.08	0.08	0.08
H	0.08	0.08	0.08	0.08	0.25
I	0.08	0.08	0.54	0.08	0.08
Q	0.08	0.08	0.08	0.25	0.08
R	0.08	0.08	0.08	0.33	0.08
W	0.54	0.08	0.08	0.08	0.08

All transitions have probability 1. The probability of an emitted sequence depends on the emission probabilities only.

Does not capture positional dependencies

WEIRD
WEIRD
WEIQH
WEIRD
WEIQH

→

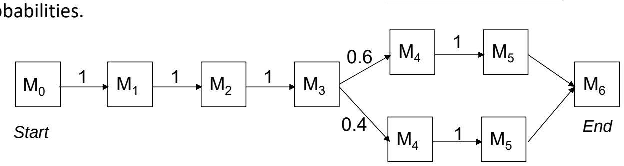
D	0.08	0.08	0.08	0.08	0.33
E	0.08	0.54	0.08	0.08	0.08
H	0.08	0.08	0.08	0.08	0.25
I	0.08	0.08	0.54	0.08	0.08
Q	0.08	0.08	0.08	0.25	0.08
R	0.08	0.08	0.08	0.33	0.08
W	0.54	0.08	0.08	0.08	0.08

Note: We never see QD or RH, only RD and QH.
But, P(RH) = P(QD) = 0.083, while P(QH) = 0.063

Branching topologies can model positional dependencies

	M4	M5
D	0.095	0.43
H	0.095	0.095
Q	0.095	0.095
R	0.43	0.095

The probability of an emitted sequence depends on the *emission and transition* probabilities.



	M1	M2	M3
D	0.08	0.08	0.08
E	0.08	0.54	0.08
H	0.08	0.08	0.08
I	0.08	0.08	0.54
Q	0.08	0.08	0.08
R	0.08	0.08	0.08
W	0.54	0.08	0.08

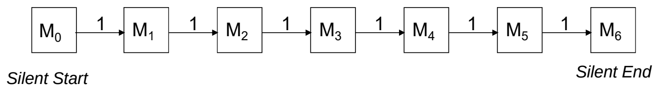
	M4	M5
D	0.095	0.095
H	0.095	0.395
Q	0.395	0.095
R	0.095	0.095

How HMMs solve the problems with PSSMs

- Do not capture positional dependencies
- ➢ Hard to recognize pattern instances that contain indels
- Variable length motifs
- Do not handle boundary detection problems well

Given a PSSM, one can construct an equivalent HMM, where the emission probabilities are those given by the columns of the PSSM. This model does not allow for patterns with indels or position specific dependencies.

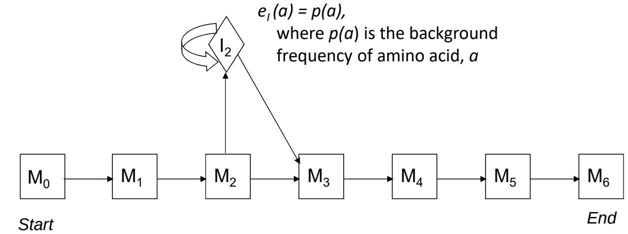
The probability of an emitted sequence depends on the emission probabilities only.



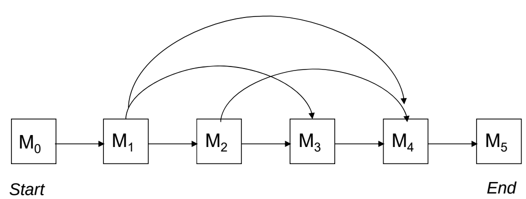
Emission probabilities

	M1	M2	M3	M4	M5
D	0.08	0.08	0.08	0.08	0.33
E	0.08	0.54	0.08	0.08	0.08
H	0.08	0.08	0.08	0.08	0.25
I	0.08	0.08	0.54	0.08	0.08
Q	0.08	0.08	0.08	0.25	0.08
R	0.08	0.08	0.08	0.33	0.08
W	0.54	0.08	0.08	0.08	0.08

$e_I(a) = p(a)$,
where *$p(a)$* is the background frequency of amino acid, *a*

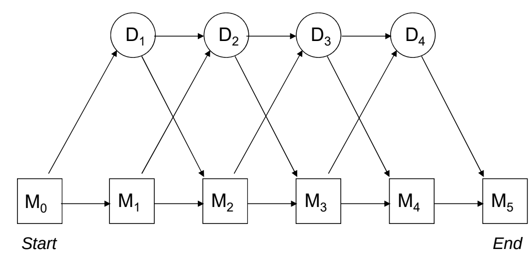


Insertions can be accounted for by adding insertion states. The self-loop allows for insertions of length greater than 1.

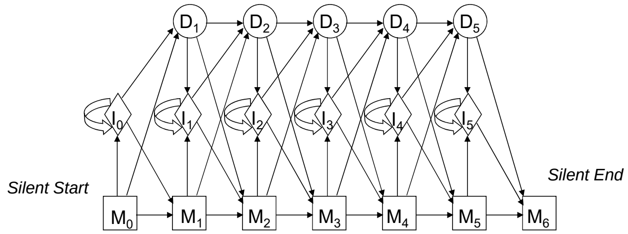


Deletions can be model by adding arcs that jump over all one or more states. However, the number of such arcs will grow exponentially with the number of Match states

$e_D(-) = 1$
$e_D(a) = 0$, for all symbols *a* in *Σ*



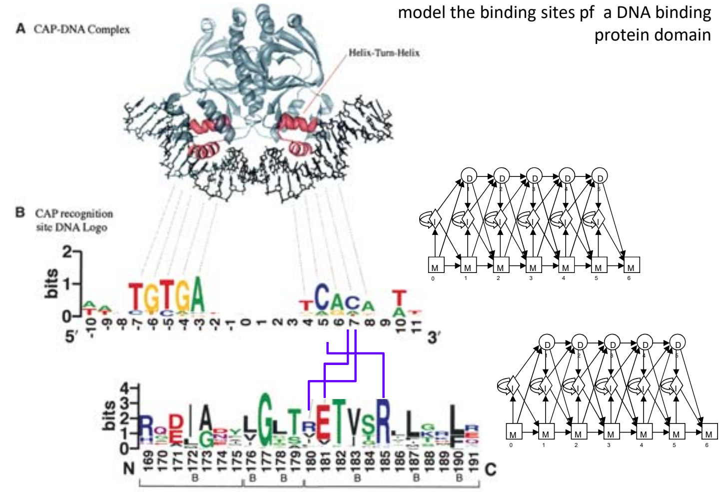
Deletions can be modeled more efficiently using explicit Deletion states. Note that a path can pass through state D_i or M_i but not both, A Profile HMM (Krogh et al, 1994) combines these features, providing a flexible, generic model for modeling conserved motifs.



Number of match states: average length of the motif.

Note that I_0 allows for sequence preceding the motif and I_5 allows for sequence following the motif.

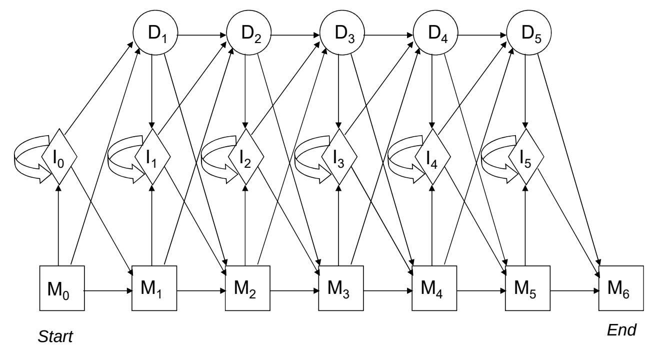
This example shows a pair HMM;s used to model the binding sites pf a DNA binding protein domain



Crooks *et al., Genome Research,* 2004

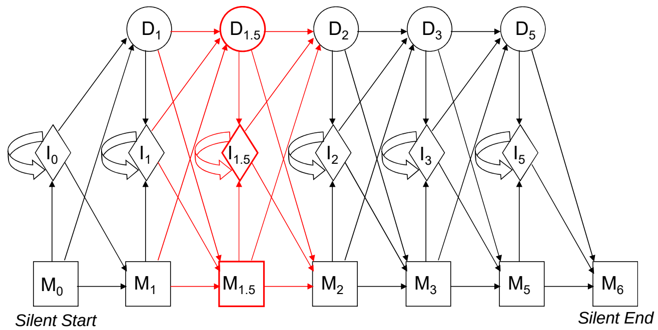
Multiple sequence alignments and Profile HMMs

1. Instantiating Profile HMM parameters with a multiple alignment
   - Aligned sequences are labeled data
   - columns in the alignment correspond to positions in the model
   - Use counts to calculate emission and transmission probabilities (MLE)
2. Using a Profile HMM to align sequences
   - Unaligned sequences are unlabeled data
   - Use Baum Welch to discover the motif (i.e., to learn model parameters)
   - Use Viterbi or Posterior decoding to align the sequences



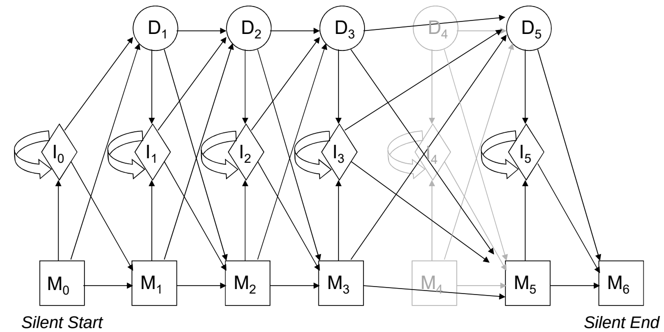
Profile HMM

Number of match states: average sequence length of the motif



Model surgery: If more than 50% of sequences pass through I_j,

- add $M_{j'}$, $I_{j'}$, $D_{j'}$ between column j and column j+1,
- connect column j to column j' & column j' to column j+1
- and re-estimate the parameters



Model surgery: If more than 50% of sequences pass through D_j,

- remove M_j, I_j, D_j,
- connect column j-1 to column j+1 and
- re-estimate the parameters

... RLSKIISMFQAHIRGYLIRKAYKRGYQARCLLK ...
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ...
.. GWQKRVRGWIVIVRRNFKKKRNEKLSATAZZZZZYQ ...
... MKRSQVVKQEKAARKVQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...

Discovery

... RLSKIISMIQAHIRGYLIRKAYKRGYQARCLLK ..
... RNKHAIAVIWAFWLVQSSFRGYQAGSKARRELK ..
... GWIQKRVRGWIVIRRNFKKKRNEKLSATAZZZZZYQ
... MKRSQVVKQEKAARKIQKFWRGHRVQHNQR ...
... QEEVSAIIIQRAYRRYLLKQKVKILRVQSS ...

Modeling

Recognition

.. GWQKRVRGWIVIVRRNQVNQAAVTIQRWYRCQVQRRRAGFKKKRNEKLSATAZZZZZ

	Recognition	Discovery & parameter inference	Parameter instantiation (MLE)
Data	Unlabeled	Unlabeled training data	Labeled training data
PSSMs (no gaps)	$\mathcal{S}(t,o) = \sum_{i=1}^{w} \mathcal{S}[t[o+i],i],$	Gibbs sampler	$P[x,i] = \frac{q[x,i]}{p_x}$ $q[x,j] = \frac{c[x,i]+b}{k+b\cdot\lvert\Sigma\rvert}$
HMMs (gaps)	P(O|λ) • Forward or Backward Viterbi decoding • Most likely path, Q* • Viterbi Posterior decoding • Most likely states, $\hat{q}_1 \ldots \hat{q}_T$ • Forward, Backward	Baum Welch • Forward • Backward	$\frac{\left(\sum_{d=1}^{k} A_{ij}^{d}\right)+b}{\sum_{j'\in\mathcal{N}(i)}\left(\left(\sum_{d=1}^{k} A_{ij'}^{d}\right)+b\right)}$ $\frac{\sum_{d=1}^{k} \mathcal{E}_i^d(\sigma)+b}{\sum_{\alpha\in\Sigma}\left(\sum_{d=1}^{k} e_i^d(\alpha)+b\right)}$