A worked example: We illustrate the process of using Profile HMM's for multiple sequence alignment with the following example. We are given unlabeled sequences as input.

1. WRCCTGC
2. WCCGGCC
3. QWCGCC
4. WRCCCGC
5. NWHCCGC

These sequences represent five instances of a biological pattern. The mean sequence length is 6.8, suggesting a Profile HMM topology with seven Match states (as well as silent start and end states), seven Deletion states, and eight Insertion states (Fig. 5.15a). The parameters of this model are estimated from the input sequences using Baum Welch Expectation Maximization. Once the parameters have been instantiated, the motif can be determined using Viterbi or posterior decoding. Decoding generates a state path for each sequence, for example

1. $M_0\ M_1\ I_1\ M_2\ M_3\ M_4\ M_5\ D_6\ M_7\ M_8$
2. $M_0\ M_1\ M_2\ M_3\ M_4\ M_5\ M_6\ M_7\ M_8$
3. $M_0\ I_0\ M_1\ M_2\ D_3\ D_4\ M_5\ M_6\ M_7\ M_8$
4. $M_0\ M_1\ I_1\ M_2\ M_3\ M_4\ M_5\ D_6\ M_7\ M_8$
5. $M_0\ I_0\ M_1\ I_1\ M_2\ M_3\ D_4\ M_5\ D_6\ M_7\ M_8$

Aligning each state path with the associated sequence results in labeled sequences:

1. W R C C T G _ C
   $M_1\ I_1\ M_2\ M_3\ M_4\ M_5\ D_6\ M_7$
2. W C C G G C C
   $M_1\ M_2\ M_3\ M_4\ M_5\ M_6\ M_7$
3. Q W C _ _ G C C
   $I_0\ M_1\ M_2\ D_3\ D_4\ M_5\ M_6\ M_7$
4. W R C C C G _ C
   $M_1\ I_1\ M_2\ M_3\ M_4\ M_5\ D_6\ M_7$
5. N W H C C _ G _ C
   $I_0\ M_1\ I_1\ M_2\ M_3\ D_4\ M_5\ D_6\ M_7$

Note that a gap symbol ('_') is inserted in the sequence at locations that correspond to a Deletion state in the state path.

Placing symbols emitted by the same M or I state into the same column, results in the following multiple alignment. Gap symbols that are labeled with deletion state D_i are inserted into the M_i column.

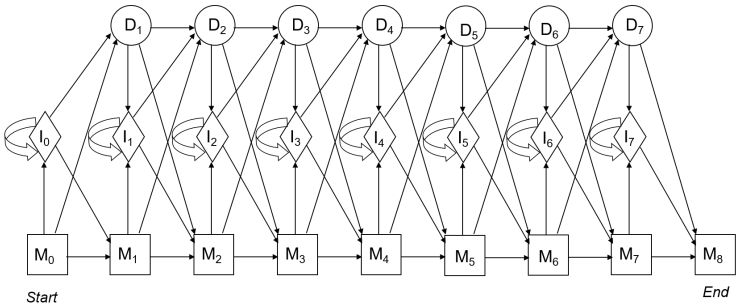
I_0	M_1	I_1	M_2	M_3	M_4	M_5	M_6	M_7
-	W	R	C	C	T	G	-	C
-	W	-	C	C	G	G	C	C
Q	W	-	C	-	-	G	C	C
-	W	R	C	C	C	G	-	C
N	W	H	C	C	-	G	-	C

We next consider whether *model surgery* is needed to refine the model. The M_6 column is sparsely populated, with gaps in more than half of the sequences, suggesting that the sixth column may not represent a conserved site in the pattern. This can be addressed by deleting states M_6, I_6 and D_6 (Fig. 5.15b). In contrast, since only two out of five sequences have a gap in column I_1, this column may represent an integral component of the pattern. An additional column must be added between the M_1 and M_2 columns to represent this element of the pattern (Fig. 5.15c).

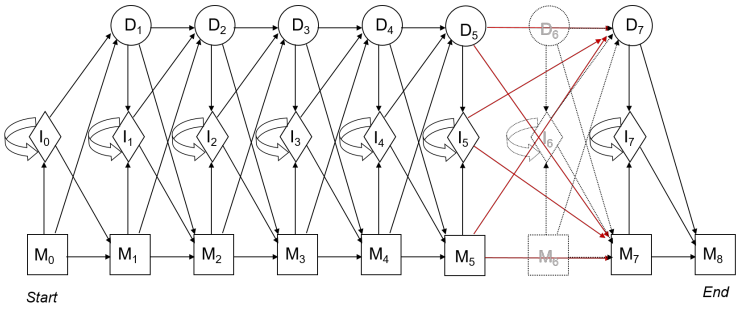
Following model surgery, parameter values must be re-estimated. If the changes to the model are substantial, then new estimates of π, a_{ij} and $e_i(\sigma)$ are obtained by applying Baum Welch to the unlabeled input sequences. Alternatively, if we believe that the multiple alignment is essentially correct, but that a small number of columns have been wrongly interpreted, the sequences can be relabeled according to the changes in the column headers. In this example, this would involve relabeling the penultimate letter `C` in Sequences 2 and 3 with state I_5. In Sequences 1, 4, and 5, residues labeled I_1 would instead be labeled $M_{1.5}$. A gap symbol labeled $D_{1.5}$ would be inserted following the `W` in Sequences 2 and 3:

1. `W` `R` `C` `C` `T` `G` `C`
   M_1 $M_{1.5}$ M_2 M_3 M_4 M_5 M_7
2. `W` `_` `C` `C` `G` `G` `C` `C`
   M_1 $D_{1.5}$ M_2 M_3 M_4 M_5 I_6 M_7
3. `Q` `W` `_` `C` `_` `_` `G` `C` `C`
   I_0 M_1 $D_{1.5}$ M_2 D_3 D_4 M_5 I_6 M_7
4. `W` `R` `C` `C` `C` `G` `C`
   M_1 $M_{1.5}$ M_2 M_3 M_4 M_5 M_7
5. `N` `W` `H` `C` `C` `_` `G` `C`
   I_0 M_1 $M_{1.5}$ M_2 M_3 D_4 M_5 M_7

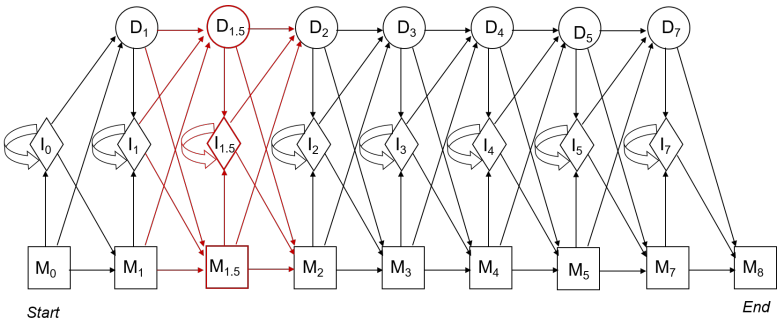
Once new labels have been assigned, the parameter values can be re-estimated using the standard approach for labeled sequences described on pp. 110-112.



(a) Profile HMM for a motif with 7 conserved sites



(b) Model surgery: removing states



(c) Model surgery: adding states

Figure 5.15: (a) A Profile HMM of length 7. (b) Model Surgery: States M_6, I_6 and D_6 are deleted because three out of five sequences have gaps in Column 6. The transitions into and out of these states are also deleted and replaced by transitions connecting states in Column 5 to states in Column 7. (c) Model Surgery: An additional column (states $M_{1.5}$, $I_{1.5}$ and $D_{1.5}$) has been added between Column 1 and Column 2 because three out of five sequences have a residue labeled I_1. The transitions connecting states in Columns 1 and 2 are replaced with transitions connecting Columns 1 and 1.5 and transitions connecting Columns 1.5 and 2.

