

1 **S3 Appendix: Technical details of maximum likelihood estimation**

2 *General formulas*

3 Instead of getting transition probabilities, we may prefer to get a direct maximum
4 likelihood estimate of a more fundamental parameter value that provides a more
5 interpretable result. Furthermore, by involving a stricter structure in our formulation,
6 parameterized by less parameters, our results should be less prone to error from overfitting
7 the data. We will achieve this by means of a technique resembling what is done to estimate
8 the parameterization of the continuous distributions for the emission matrix.

9 For every transition factor indexed by u , we introduce a parameter $p_{s+1,u}$, a
10 function, $g_u : \mathbb{N} \rightarrow \mathbb{N}$, a function $h_u : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$, and a function $\psi_u : \mathbb{R} \rightarrow \mathbb{R}^{N \times N}$. The
11 form of these functions depends on the form of the random variable being modeled by the
12 transition factor, and we will describe this in more detail below. For now we define:

$$13 \quad p_{s+1,u} = \frac{\sum_r \sum_t \sum_i \sum_j \xi_{ij}^{(s,r,t,u)} h_u(i,j)}{\sum_r \sum_t \sum_i \gamma^{(s,r,t,u)} g_u(i)} \quad (59)$$

$$14 \quad \mathcal{T}^{(s+1,u)} = \psi_u(p_{s+1,u}) \quad (60)$$

15 Similarly for our initial conditions we let:

$$16 \quad \hat{p}_{s+1,u} = \frac{\sum_r \sum_t \sum_i \sum_j \xi_{ij}^{(s,r,-1,u)} \hat{h}_u(i,j)}{\sum_r \sum_t \sum_i \gamma^{(s,r,-1,u)} \hat{g}_u(i)} \quad (61)$$

$$17 \quad \tau^{(s+1,u)} = \hat{\psi}_u(\hat{p}_{s+1,u}) \quad (62)$$

18 Now we can define our factor specific functions.

19 *Detachment rate estimation*

20 We start with the detachment rate because it is the easiest to describe. It is defined
21 by a Bernoulli random variable, which determines the rate at which a peptide enters a
22 detached state instead of remaining unchanged.

23 We can only measure the detachment rate from states with remaining amino acids.
24 We therefore set $g_u(i)$ to 1 for those states, and 0 for the remaining state. To pick up the
25 transition rate into these states, we set $h_u(i, j)$ to 1 if i has remaining amino acids *and* j
26 indicates the detached state. We set it to 0 for all other combinations of i and j , which
27 should never occur as they should have been forbidden in the definition of the transition
28 factor in the previous iteration of Baum-Welch, $T^{(s,u)}$.

29 This is a weighted maximum likelihood estimate of a Bernoulli random variable.
30 We use the resulting value of $p_{s+1,u}$ to define the corresponding transition factor. ψ_u then
31 defines $T_{ij}^{(s+1,u)}$ to be $p_{s+1,u}$ when i has remaining amino acids *and* j indicates the detached
32 state, and to be $1 - p_{s+1,u}$ when $i = j$ and they are not in the detached state. When both
33 are in the detached state, $T_{ij}^{(s+1,u)}$ is 1, and for all other combinations of i and j , we set
34 $T_{ij}^{(s+1,u)}$ to 0.

35 *N-terminal blocking rate estimation*

36 N-terminal blocking behavior is also defined by Bernoulli random variables. There
37 are two of these rates, and we start by defining our factor-specific functions for the cyclic
38 blocking behavior, which defines the rate at which a peptide moves into a blocked state
39 instead of remaining unchanged.

40 We can only measure this rate from the unblocked states. Let $g_u(i) = 1$ in those
41 states, and 0 for all other states. $h_u(i, j) = 1$ if i represents an unblocked state and j
42 indicates the specifically corresponding blocked state. $h_u(i, j) = 0$ for all other
43 combinations of i and j , though these should never occur under a correct implementation
44 of this algorithm.

45 This again is a weighted maximum likelihood estimate of a Bernoulli random
46 variable, and we can describe the action of ψ_u . $T_{ij}^{(s+1,u)}$ should be $p_{s+1,u}$ when i is an
47 unblocked state and j is the corresponding blocked state, while it should be $1 - p_{s+1,u}$
48 when $i = j$ and they are in an unblocked state. When both are in the same unblocked state,
49 $T_{ij}^{(s+1,u)}$ is 1, and for all other combinations of i and j , we set $T_{ij}^{(s+1,u)}$ to 0.

50 For the matrix factor representing the initial N-terminal blocking, we analyze the
51 data in exactly the same way, but just use a different parameter and matrix to track our
52 results, in order to allow the initial blocking rate to be different from the cyclic one.

53 *Dye destruction rate estimation*

54 An assumption of equal exposure to both chemical failure and photobleaching of
55 the fluorophores means we should treat this as a binomial random variable. As with N-
56 terminal blocking rates, we have one rate that identifies the behavior before sequencing
57 starts (the missing fluorophore rate), and another for destruction during sequencing (the
58 dye loss rate). These will again be mostly equivalent in their analysis, though tracked with
59 different variables to enforce a separation. We show here the analysis for the dye loss rate.
60 We also wish to emphasize that we track these separately, and with separate functions g_u ,
61 h_u , and ψ_u , for each color of fluorophore.

62 We note that states with more fluorophores of the color being analyzed provide
63 more evidence of the rate of fluorophore loss. With this in mind, $g_u(i)$ is set to the number
64 of fluorophores of the color of interest in state i . We then let $h_u(i, j) = g_u(i) - g_u(j)$ if
65 $j \leq i$, and otherwise set it to 0 (for transitions which should never occur).

66 To construct $T^{(s+1,u)}$, we let $T_{ij}^{(s+1,u)} = B(h_u(i,j); g_u(i), p_{s+1,u})$ if $j \leq i$, where B
 67 represents the parameterized probability mass function of the binomial distribution. This
 68 expands to:

$$69 \quad T_{ij}^{(s+1,u)} = \binom{g_u(i)}{g_u(j)} (p_{s+1,u})^{h_u(i,j)} (1 - p_{s+1,u})^{g_u(j)} \quad (63)$$

70 When $j \leq i$. When $j > i$, we let $T_{ij}^{(s+1,u)} = 0$.

71 There is a key difference in the analysis of the missing fluorophore, or dud-dye,
 72 rate. A bias is introduced because peptides with all fluorophores missing in every color will
 73 not be visible and are therefore absent from the data. A correction for this bias is discussed
 74 later in this text.

75 *Edman failure rate estimation*

76 Our factored transitions which manage Edman degradation are defined by a
 77 Bernoulli random variable, much like the factors managing the detachment rate or the two
 78 forms of the N-terminal blocking rates. While there is an additional complication from the
 79 probabilistic loss of a fluorophore in the case of a successful Edman degradation, we can
 80 safely ignore this until we need to construct $T^{(s+1,u)}$. When determining $p_{s+1,u}$, we need
 81 only concern ourselves with whether Edman degradation failed.

82 We need to omit states with a blocked N-terminus. For all states with a blocked N-
 83 terminus, $g_u(i) = 0$. For states with an unblocked N-terminus, $g_u(i) = 1$. We let
 84 $h_u(i,j) = 0$ if either or both of i or j represent a state with a blocked N-terminus. We also
 85 set it to 0 for all invalid combinations of i and j that should never occur; this relationship
 86 is complicated and peptide dependent. Combinations of i and j are valid when $i = j$, or
 87 when removing the N-terminal amino acid from state i can result in state j . In the second
 88 case, state j represents either the same combination of fluorophore counts of different

89 colors but with one less amino acid, *or* it represents that combination less one amino acid
 90 *and* minus one fluorophore, *of the specific color* of fluorophore which may or may not be
 91 present on the amino acid being removed. This is invalid when the N-terminal amino acid
 92 cannot be labeled and in that case should be zero.

93 In any case, when $i = j$, we let $h_u(i, j) = 0$, as this is an Edman success. For all
 94 other *valid* combinations of i and j , we let $h_u(i, j) = 1$. This will give a weighted maximum
 95 likelihood estimate of the probability of an Edman failure event.

96 To define $T^{(s+1,u)}$, we let $T_{ij}^{(s+1,u)} = 1 - p_{s+1,u}$ when $i = j$. For $i \neq j$ (assuming a
 97 valid transition), if state i does not have a possibility of a fluorophore on its N-terminal
 98 amino acid, then $T_{ij}^{(s+1,u)} = p_{s+1,u}$. If a fluorophore is possible on its N-terminal amino
 99 acid, the computation is a bit more involved.

100 In [16], we derived a formula for the probability of fluorophore removal with
 101 Edman degradation, which we reiterate here under a slightly different symbolic
 102 representation. Let $\lambda(i)$ represent the number of labelable amino acids which can take a
 103 fluorophore of the same color that the N-terminal amino acid may have when in state i . Let
 104 $G(i)$ represent the number of fluorophores of the same color the N-terminal amino acid
 105 may have, when in state i . We note that this second function is similar in form to the variant
 106 of $g_u(i)$ described under “Dye destruction rate estimation.” Now, in the case where $i \neq j$
 107 (for a valid transition) and state i *does* have a possibility of an N-terminal amino acid, and
 108 j , in relation to i , represents an amino acid removal, we let:

$$109 \quad T_{ij}^{(s+1,u)} = \frac{G(i)}{\lambda(i)} p_{s+1,u} \quad (64)$$

110 Then, for the case where $i \neq j$ (for a valid transition) and state i has a possibility of
 111 an N-terminal amino acid, and j , in relation to i , represents no amino acid removal, we then
 112 let:

113
$$T_{ij}^{(s+1,u)} = \left(1 - \frac{G(i)}{\lambda(i)}\right) p_{s+1,u} \quad (65)$$

114 Finally, for invalid transitions, we let $T_{ij}^{(s+1,u)} = 0$.

115 ***Additional technical discussion of weighted parameter estimation***

116 We note that, with the exception of Edman degradation, all of the functions $g_u(i)$
117 and $h_u(i, j)$ are trivial to compute when the cumulative forward and backward probability
118 results are indexed by a higher-order tensor. Then, these functions become a simple
119 extraction of an index of that higher-order tensor.

120