

S4 Appendix: Technical details of bias correction

When a peptide is missing all its fluorophores it is not visible and does not get sequenced. Correcting for this issue involves considerable additional complexity. We start by noting that this introduces a statistical dependency between missing fluorophore rates of different colors of fluorophores. This is because a count of zero fluorophores of one color only eliminates the peptide from the dataset if there are also zero fluorophores of every other color. We find it clearer to first define $\tau^{(s+1,u)}$. As in the case for the dye loss rate, let $g_u(i)$ represent the number of fluorophores of the color of interest in state i , and let $h_u(i, j) = g_u(i) - g_u(j)$ if $j \leq i$, and otherwise set it to zero.

Now however, instead of defining $\tau^{(s+1,u)}$ with a parameterized binomial distribution as we did before for $T^{(s+1,u)}$, we use a modified binomial distribution where the entries represent the probability of the transition *conditioned on* the probability of the peptide being observable. Letting \hat{C} represent the set of all pre-sub-transition indices corresponding to missing fluorophore rates of different fluorophore colors, this equation is given by:

$$\tau_{ij}^{(s+1,u)} = \frac{B(h_u(i, j); g_u(i), \hat{p}_{s+1,u})}{1 - \prod_{\hat{u} \in \hat{C}} B(0; g_{\hat{u}}(i), \hat{p}_{s+1,\hat{u}})} \quad (66)$$

We of course require, as before, $j \leq i$, and otherwise set $\tau_{ij}^{(s+1,u)}$ to zero. Additionally, we note that $\tau_{ij}^{(s+1,u)}$ should be zero for the state i such that $g_u(i) = 0$ when $\hat{u} \in \hat{C}$ (the state with no fluorophores of any color).

With other forms of error, we could point to existing and well understood formulas for their maximum likelihood estimates, as they can be viewed either as binomial or Bernoulli random variables. That will not work in this case, and we must explicitly derive

23 this result. We consider the colors of fluorophore together, due to their statistical
 24 dependencies. Then we seek to maximize the likelihood given by:

$$25 \quad L = \prod_i \prod_j \left(\frac{\prod_{\hat{u} \in \hat{C}} B(h_{\hat{u}}(i, j); g_{\hat{u}}(i), \hat{p}_{s+1, \hat{u}})}{\prod_{\hat{u} \in \hat{C}} B(0; g_{\hat{u}}(i), \hat{p}_{s+1, \hat{u}})} \right)^{\Xi_{ij}^{(s, r, -1, \hat{C})}} \quad (67)$$

26 Assuming i and j vary over only their valid ranges, and where we let $\Xi_{ij}^{(s, r, -1, \hat{C})}$
 27 represent the probability of a transition from state i to state j given $Y_{1:T} = y_{1:T}$ and $\theta^{(s)}$
 28 (for iteration s , sequence r , time step t) when considering only the subset of sub-transitions
 29 contained in U , the set of all missing fluorophore related sub-sequences. In mathematical
 30 notation:

$$31 \quad \Xi_{ij}^{(s, r, t, U)} = \prod_{u \in U} \xi^{(s, r, t, u)} \quad (68)$$

32 We now note a useful simplification to our formula. We previously defined our
 33 initial conditions for $f^{(s, -1)}$ so that it would be one for the perfectly labeled and non-
 34 blocked state, and zero everywhere else. The only pre-sub-transition other than the missing
 35 fluorophore rate is the transition for initial N-blocking. The status of the N-terminus is
 36 irrelevant to the missing fluorophore rate, thus we consider the blocked and unblocked
 37 states together. We also ignore all states with missing fluorophores as irrelevant. We name
 38 the remaining state I . Then we can reduce our likelihood equation to:

$$39 \quad L = \prod_j \left(\frac{\prod_{\hat{u} \in \hat{C}} B(h_{\hat{u}}(I, j); g_{\hat{u}}(I), \hat{p}_{s+1, \hat{u}})}{\prod_{\hat{u} \in \hat{C}} B(0; g_{\hat{u}}(I), \hat{p}_{s+1, \hat{u}})} \right)^{\Xi_{Ij}^{(s, r, -1, \hat{C})}} \quad (69)$$

40 Expanding our equation for the likelihood we get:

$$41 \quad L = \prod_j \left(\frac{\prod_{\hat{u} \in \hat{C}} \left(\frac{g_{\hat{u}}(I)}{g_{\hat{u}}(j)} \right) (\hat{p}_{s+1, \hat{u}})^{h_{\hat{u}}(I, j)} (1 - \hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(j)}}{1 - \prod_{\hat{u} \in \hat{C}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)}} \right)^{\Xi_{Ij}^{(s, r, -1, \hat{C})}} \quad (70)$$

42 We now take the logarithm, which makes the equation easier to work with while
 43 preserving order:

$$\begin{aligned}
 44 \quad \log(L) &= \sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} \left(\sum_{\hat{u} \in \hat{c}} \left(\log \left(\frac{g_{\hat{u}}(I)}{g_{\hat{u}}(j)} \right) + h_{\hat{u}}(I, j) \log(\hat{p}_{s+1, \hat{u}}) \right. \right. \\
 45 \quad &\quad \left. \left. + g_{\hat{u}}(j) \log(1 - \hat{p}_{s+1, \hat{u}}) \right) - \log \left(1 - \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)} \right) \right) \\
 46 & \tag{71}
 \end{aligned}$$

47 Noting that this equation is symmetric with respect to \hat{u} , we need only solve for
 48 one result to maximize for every $\hat{p}_{s+1, \hat{u}}$. We then take the derivative with respect to a choice
 49 of $\hat{p}_{s+1, \hat{u}}$ and set it to zero to search for extrema.

$$50 \quad 0 = \sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} \left(\frac{h_u(I, j)}{\hat{p}_{s+1, u}} - \frac{g_u(j)}{1 - \hat{p}_{s+1, u}} + \frac{g_u(I) \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)}}{\hat{p}_{s+1, u} \left(1 - \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)} \right)} \right) \tag{72}$$

51 This reduces to:

$$\begin{aligned}
 52 \quad 0 &= \sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} \left(h_u(I, j) - (\hat{p}_{s+1, u}) g_u(I) \right. \\
 53 \quad &\quad \left. - (1 - \hat{p}_{s+1, u}) g_u(I) \left(1 - \left(1 - \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)} \right) \right) \right) \\
 54 & \tag{73}
 \end{aligned}$$

55 And reduces again to:

$$56 \quad 0 = \sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} \left(g_u(j) - \frac{(1 - \hat{p}_{s+1, u}) g_u(I)}{1 - \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)}} \right) \tag{74}$$

57 If we rearrange, we find:

$$58 \quad \frac{1 - \hat{p}_{s+1, u}}{1 - \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s+1, \hat{u}})^{g_{\hat{u}}(I)}} = \frac{\sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} g_u(j)}{g_u(I)} \tag{75}$$

59 This is a multivariate generalization of the solution to a geometric sequence and is
60 therefore a generalization of the associated inverse problem. This also constitutes root-
61 finding of a polynomial of arbitrary order, and thus is unlikely to have a closed form
62 solution. If we assume large $g_{\hat{u}}(I)$ or small $\hat{p}_{s+1,\hat{u}}$ for at least one color of fluorophore,
63 then this can be approximated by:

$$64 \quad \hat{p}_{s+1,u} \approx \frac{\sum_j \Xi_{Ij}^{(s,r,-1,\hat{C})} h_u(I,j)}{g_u(I)} \quad (76)$$

65 This is the maximum likelihood estimate for an ordinary binomial random variable.
66 We use this to demonstrate that the extremum we've found is a maximum, in place of
67 applying the second derivative test as would ordinarily be done.

68 We need a way to approximately solve this when $g_{\hat{u}}(I)$ is small or $\hat{p}_{s+1,\hat{u}}$ is large.
69 We suggest an iterative method, where we plug the left-hand side result into the right-hand
70 side on each iteration. We iterate over z , writing our new equation as:

$$71 \quad \hat{p}_{s+1,u}^{(z+1)} = 1 - \frac{\left(1 - \prod_{\hat{u} \in \hat{C}} \left(\hat{p}_{s+1,\hat{u}}^{(z)}\right)^{g_{\hat{u}}(I)}\right) \sum_j \Xi_{Ij}^{(s,r,-1,\hat{C})} g_u(j)}{g_u(I)} \quad (77)$$

72 We need to prove that this iteration converges. The multivariate nature of this
73 problem requires us to consider this in a multivariate vector space of dimension $|\hat{C}|$. We
74 also set some requirements. Firstly, this equation is clearly nonsense unless:

$$75 \quad \sum_{u \in \hat{C}} g_u(I) \geq 2 \quad (78)$$

76 We also require a constraint on the solution:

$$77 \quad 0 < \hat{p}_{s+1,\hat{u}} < 1 \quad (79)$$

78 For convenience, we now introduce the following variables:

$$79 \quad P_{s+1}^{(z)} = \prod_{\hat{u} \in \hat{C}} \left(\hat{p}_{s+1,\hat{u}}^{(z)}\right)^{g_{\hat{u}}(I)} \quad (80)$$

$$80 \quad P_{s+1} = \prod_{\hat{u} \in \hat{C}} \left(\hat{p}_{s+1,\hat{u}}\right)^{g_{\hat{u}}(I)} \quad (81)$$

$$81 \quad 0 < S_{s+1,u} = \frac{1 - \hat{p}_{s+1,u}}{1 - P_{s+1}} = \frac{\sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} g_u(j)}{g_u(I)} < 1 \quad (82)$$

82 This gives us the equation:

$$83 \quad \hat{p}_{s+1,u}^{(z+1)} = 1 - (1 - P_{s+1}^{(z)}) S_{s+1,u} \quad (83)$$

84 Equivalently:

$$85 \quad \hat{p}_{s+1,u}^{(z+1)} = 1 - S_{s+1,u} + S_{s+1,u} P_{s+1}^{(z)} \quad (84)$$

86 If we have a vector of $\hat{p}_{s+1,u}^{(z)}$ such that $\hat{p}_{s+1,u}^{(z)} = P_{s+1}$, then clearly from this
 87 equation we will get the correct result in the next iteration, and $\hat{p}_{s+1,u}^{(z+1)} = \hat{p}_{s+1,\hat{u}}$. More
 88 generally, due to the linear relationship of these terms, and noting that $0 < S_{s+1,u} < 1$, if
 89 $P_{s+1}^{(z)} = P_{s+1} + \epsilon$ for $|\epsilon| \ll 1$, then there exists $|\delta| < |\epsilon|$ such that $\hat{p}_{s+1,u}^{(z+1)} = \hat{p}_{s+1,\hat{u}} + \delta$.
 90 Therefore, to prove convergence of every $\hat{p}_{s+1,u}^{(z)}$ to $\hat{p}_{s+1,u}$ for all $u \in \hat{C}$ it is sufficient to
 91 prove convergence of $P_{s+1}^{(z)}$ to P_{s+1} .

92 There is a recurrence relation of $P_{s+1}^{(z+1)}$ in terms of $P_{s+1}^{(z)}$, which is:

$$93 \quad P_{s+1}^{(z+1)} = \prod_{u \in \hat{C}} \left(1 - S_{s+1,u} + S_{s+1,u} P_{s+1}^{(z)} \right)^{g_u(I)} \quad (85)$$

94 The right-hand side is a polynomial expression of order $\sum_{u \in \hat{C}} g_u(I)$ with all positive
 95 coefficients. It follows that its derivatives of every order up to and including $\sum_{u \in \hat{C}} g_u(I)$
 96 are strictly positive. Our previous requirement that $\sum_{u \in \hat{C}} g_u(I) \geq 2$ implies that the
 97 polynomial is strictly positive and has strictly positive first and second derivatives for $0 <$
 98 $P_{s+1}^{(z)} < 1$. The strictly positive second derivative in this range guarantees that no more than
 99 two fixed points are possible in the given range. There is by definition a fixed-point solution
 100 $0 < P_{s+1} < 1$, which is the solution we want our iteration to converge towards. There is
 101 also another fixed point at 1.

102 Noting that for $P_{s+1}^{(z)} = 0$ we get $P_{s+1}^{(z+1)} > 0$, it follows that:

$$103 \quad P_{s+1}^{(z+1)} > P_{s+1}^{(z)} \text{ for } 0 < P_{s+1}^{(z)} < P_{s+1} \quad (86)$$

104 A double root at P_{s+1} is incompatible with the fixed point at 1. Therefore:

105
$$P_{s+1}^{(z+1)} < P_{s+1}^{(z)} \text{ for } P_{s+1} < P_{s+1}^{(z)} < 1 \quad (87)$$

106 Using the strictly positive first derivative, we also find that both:

107
$$P_{s+1}^{(z+1)} < P_{s+1} \text{ for } 0 < P_{s+1}^{(z)} < P_{s+1} \quad (88)$$

108
$$P_{s+1}^{(z+1)} > P_{s+1} \text{ for } P_{s+1} < P_{s+1}^{(z)} < 1 \quad (89)$$

109 These four results can be combined to prove the relation:

110
$$\left| P_{s+1}^{(z+1)} - P_{s+1} \right| < \left| P_{s+1}^{(z)} - P_{s+1} \right| \quad (90)$$

111 Therefore, given any $0 < P_{s+1}^{(0)} < 1$ our iteration eventually converges towards the
 112 desired result.

113 We then fold this iteration into the outer iteration used to improve all estimates in
 114 the Baum-Welch algorithm, which gives the equation:

115
$$\hat{p}_{s+1,u} = 1 - \frac{\left(1 - \prod_{\hat{u} \in \hat{c}} (\hat{p}_{s,\hat{u}})^{g_{\hat{u}}(I)}\right) \sum_j \Xi_{Ij}^{(s,r,-1,\hat{c})} g_u(j)}{g_u(I)} \quad (91)$$

116 This is equivalent to imputing missing data using the parameter estimates in the
 117 previous iteration, as described in the nontechnical summary of this bias correction. We
 118 also note that it is tempting to apply a root finding method from an open-source library,
 119 which would likely converge faster in practice. However, our method allows us to
 120 accommodate cross-parameter effects by default, which is a desirable property. There is of
 121 course a valid concern that cross-parameter effects could in some way invalidate this proof,
 122 but we have found in practice that this does not appear to be the case.

123