Are Profile Hidden Markov Models Identifiable?

Srilakshmi Pattabiraman
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois
sp16@illinois.edu

Tandy Warnow
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois
warnow@illinois.edu

ABSTRACT

Profile Hidden Markov Models (HMMs) are graphical models that can be used to produce finite length sequences from a distribution. In fact, although they were only introduced for bioinformatics 25 years ago (by Haussler et al., Hawaii International Conference on Systems Science 1993), they are arguably the most commonly used statistical model in bioinformatics, with multiple applications, including protein structure and function prediction, classifications of novel proteins into existing protein families and superfamilies, metagenomics, and multiple sequence alignment. The standard use of profile HMMs in bioinformatics has two steps: first a profile HMM is built for a collection of molecular sequences (which may not be in a multiple sequence alignment), and then the profile HMM is used in some subsequent analysis of new molecular sequences. The construction of the profile thus is itself a statistical estimation problem, since any given set of sequences might potentially fit more than one model well. Hence a basic question about profile HMMs is whether they are *statistically identifiable*, which means that no two profile HMMs can produce the same distribution on finite length sequences. Indeed, statistical identifiability is a fundamental aspect of any statistical model, and yet it is not known whether profile HMMs are statistically identifiable. In this paper, we report on preliminary results towards characterizing the statistical identifiability of profile HMMs in one of the standard forms used in bioinformatics.

CCS CONCEPTS

• **Applied computing → Molecular sequence analysis**;

KEYWORDS

Profile Hidden Markov Models, statistical identifiability, molecular sequence analysis

ACM Reference Format:
Srilakshmi Pattabiraman and Tandy Warnow. 2018. Are Profile Hidden Markov Models Identifiable?. In *ACM-BCB '18: 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, August 29–September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3233547.3233563

ACM-BCB'18, August 29–September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233563

1 INTRODUCTION

Profile Hidden Markov Models (HMMs) are arguably the most common statistical models in bioinformatics. Originally introduced by Haussler and colleagues in [10, 12], and then expanded later in many subsequent texts [4–6, 9, 11, 21, 25], profile HMMs are now used in many analytical steps in biological sequence analysis [15, 17–19, 22].

Profile Hidden Markov models are graphical models with match states, insertion states, and deletion states; and the match and insertion states emit letters from an underlying alphabet Σ (*i.e.*, Σ may be the 20 amino acids, the four nucleotides, or some other set of symbols). In the standard form presented in [4] (widely in use in bioinformatics applications), each profile Hidden Markov model has a single start state and a single end state, and every path through the model produces a string from Σ^*. The topology of this standard model as seen in Figure 1 shows directed edges between certain pairs of states, and each such directed edge has a non-zero transition probability.

In this paper, we address the question of statistical identifiability of profile Hidden Markov models, which in essence asks whether the model is reconstructible given the probability distribution it defines [23]. Thus, if there are two sets of parameters of the model that generate the same joint distribution, then the model is not identifiable. Note that if a model is not identifiable, then it is impossible for any algorithm designed to estimate the model from a finite dataset to be *statistically consistent*: that is, it is not possible for the method to converge in probability to the true model with increasing amounts of data.

Statistical identifiability is a basic property of statistical models, and is the subject of rigorous study [1–3, 7, 13, 16, 20]. Indeed, the importance of identifiability is evident in the following quotes: "Unidentifiable models are pathological, usually due to conceptual error in model formulation" [24] and "Many statisticians frown on the use of under-identified models: if a parameter is not identifiable, two or more values are indistinguishable, no matter how much data you have" [8]. However, to the best of our knowledge, nothing has yet been established about the statistical identifiability of profile Hidden Markov Models (HMMs), although the question of identifiability of parameters in HMMs more generally has also been specifically addressed [14].

In this paper, we partially characterize the conditions under which profile HMMs are statistically identifiable. Our study includes a characterization of identifiable profile HMMs when no deletion states are permitted but also shows two profile HMMs in the standard format that define the same probability distribution. Hence, we show that profile HMMs are not identifiable. We conclude our study with a discussion of the implications of this research and future directions.

2 RESULTS

2.1 Preliminary material and notation

The question we address in this paper is whether profile HMMs (in this standard format, as described in Figure 1) are statistically identifiable. We present profile HMMs for modeling collections of DNA sequences (i.e., strings over $\{A, C, T, G\}$), which is one of their uses; however, the results we present here are independent of the choice of alphabet. As shown in Figure 1, the standard topology profile HMM with n match states has a single begin state "begin" and a single end state "end"; every path through the profile HMM thus begins and ends at these states. The standard profile HMM also has n match states, $n+1$ insertion states, and n deletion states. Every match state M_j (with $j \in [n]$) emits a letter A, T, G, or C according to some fixed but unknown probability distribution $\mathcal{P}_j$. All insertion states $I_{j'}$ (with $j' \in \{0\} \cup [n]$) emit a letter with the same known distribution $\mathcal{P}_{ins}$. Note therefore that the emission probabilities can be different for different match states, but all insertion states have the same emission probabilities. Finally, the deletion states are "silent" (i.e., they do not emit any letters), and are denoted by D_j. The probability of transition from one state to another is represented by the positive values on the edges (also referred to as edge weights); hence, the sum of the weights on the edges leaving any given node is 1.0. Note that under this standard profile HMM, once you know the number of match states you also know the entire topology.

We introduce some notation to simplify the rest of the exposition. We let x_i denote the transition probability from M_{i-1} to M_i (with M_0 denoting the start state and M_{n+1} denoting the end state) and y_i denote the transition probability from I_{i-1} to M_i. We let z^i_Y denote the emission probability of letter Y from match state M_i, *i.e.*, $\mathbb{P}[Y|\text{match state } = i] = z^i_Y$. We use $\mathbb{P}_{ins}[j]$ to denote the emission probability of letter $j \in \{A, C, T, G\}$ at the insertion states, and constrain all insertion states to have the same emission probability distribution.

Let $*$ denote an arbitrary length string. Thus, $A*$ denotes all sequences that begin with A. Let ? denote an arbitrary letter, and let $?^{[k]}$ denote k contiguous arbitrary letters. Thus, $?A*$ denotes all sequences whose second letter is A. Let p_S denote $\mathbb{P}[$ sequence $S]$, the probability of the model emitting sequence S, and let $p_{\mathcal{S}}$ denote the probability of emitting all sequences in the set $\mathcal{S}$. We drop the stylized notation when the set $\mathcal{S}$ is clear from context.

2.2 No deletion nodes

Here, we consider the standard profile HMM topology with the probability of transitioning to any deletion node being 0. In other words, we consider a profile HMM topology without deletion nodes, as shown in Figure 2.

Consider the path that begins at the start state and ends at M_i and that only goes through match states; the probability of picking that path is denoted by $p^{(\text{match}:i)}$, and is easily seen to be $\prod_{k=1}^{i} x_k$. Note that this is the only path with i edges that begins at the start state and ends at M_i. Similarly, the probability of picking the path from the start state to I_j that passes only through match states is denoted by $p^{(\text{insrt}:j)}$ and is equal to $\prod_{k=1}^{j-1} x_k \cdot (1 - x_j)$. As before, this is the only path with $j+1$ edges that begins at the start state and ends at I_j.

Theorem 2.1. *Consider a standard profile HMM topology with n match states and no deletion states. Then, the model is identifiable if and only if no match state has the same emission probability distribution as the insertion states.*

Proof. $\Leftarrow$: We begin by proving that if no match state has the same distribution as the insertions states, then the model is identifiable. Note that when there are no deletion states, the length of the shortest sequence with non-zero probability of being generated is the number of match states. Hence, given the distribution of sequences defined by a profile HMM that has no deletion states, we immediately know the number of match states, and hence also the topology. We will show that we can use the topology of the profile HMM to compute all the numerical parameters, and hence define the entire model, once we are given the distribution of strings defined by the model.

So let the length of the shortest sequence (with non-zero probability) be n. We provide the proof of identifiability for the case where all nucleotides have equal probability of being generated at the insertion states (*i.e.*, $\mathbb{P}_{ins}[A] = \mathbb{P}_{ins}[C] = \mathbb{P}_{ins}[T] = \mathbb{P}_{ins}[C] = \frac{1}{4}$). For the more general case where the emission probabilities at the insertion states are different, the proof is a simple modification of the one provided below.

We now show how to compute the emission probabilities z^i_A, z^i_T, z^i_G, and z^i_C. Note that the probability that a string of length n generated by this model has an A in the i^{th} position is given by

$$p_{?^{[i-1]}A?^{[n-i]}} = z^i_A \prod_{j=1}^{n} x_j \tag{1}$$

and hence

$$z^i_A = \frac{p_{?^{[i-1]}A?^{[n-i]}}}{\sum_{X \in \{A,T,G,C\}} p_{?^{[i-1]}X?^{[n-i]}}}. \tag{2}$$

The next equation follows since every path that emits an A as the first letter either goes through the first match state or through the first insertion state:

$$p_{A*} = x_1 z^1_A + (1 - x_1)\frac{1}{4}. \tag{3}$$

We will refer to this equation as the 0^{th} system; note that it is a linear system in one variable, x_1. Furthermore, if not all z^1_X (for $X \in \{A, C, T, G\}$) are equal to $\frac{1}{4}$ then there is a unique solution for x_1; however, if all are equal to $\frac{1}{4}$ then every value for x_1 is a solution. Also, the same equations hold where A is replaced by the other nucleotides.

Recall that y_i is the transition probability from I_{i-1} to M_i. Consider the probability of a string that has A as its second letter. Equations (4) and (5) below (both straightforward to establish) will be referred to jointly as the "1^{st} system":

$$p_{?A*} = x_1\left(x_2 z^2_A + (1 - x_2)\frac{1}{4}\right) + (1 - x_1)\left(y_1 z^1_A + (1 - y_1)\frac{1}{4}\right) \tag{4}$$

$$p_{?T*} = x_1\left(x_2 z^2_T + (1 - x_2)\frac{1}{4}\right) + (1 - x_1)\left(y_1 z^1_T + (1 - y_1)\frac{1}{4}\right) \tag{5}$$

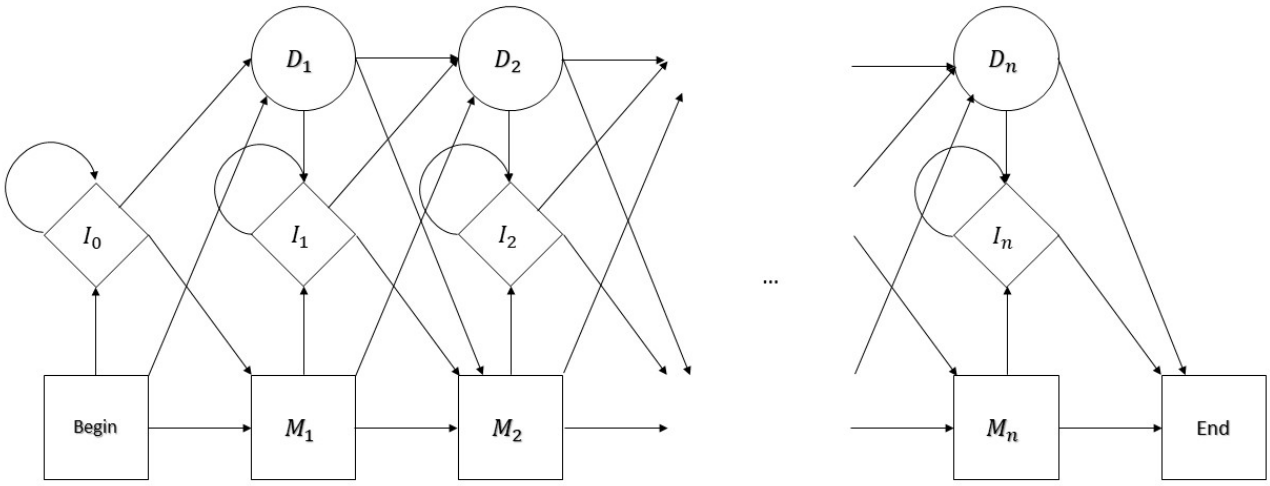

Figure 1: The topology of the standard profile Hidden Markov Model (according to [4]) with n match states. Note that only certain pairs of nodes are connected by edges; every such edge has strictly positive transition probability, and the sum of the transition probabilities on the edges leaving any single node is 1. The match states (denoted by M) and insertion states (denoted by I) emit letters from an underlying alphabet Σ, and hence have associated emission probabilities for each letter in Σ. The deletion states (denoted by D) are silent and do not emit anything. Each such profile HMM is a generative model, since every path from the start state to the end state produces a string from Σ^*. Hence each profile HMM defines a probability distribution on Σ^*.

The 1st system of equations given by Equations (4) and (5) is linear in (x_2, y_1) as long as Equation (3) is solved. The system can be written as $p^{(1)} = M^{(1)} w^{(1)}$ where

$$M^{(1)} = \begin{bmatrix} x_1(z_A^2 - 1/4) & (1-x_1)(z_A^1 - 1/4) \\ x_1(z_T^2 - 1/4) & (1-x_1)(z_T^1 - 1/4) \end{bmatrix}, \tag{6}$$

$$w^{(1)} = \begin{bmatrix} x_2 \\ y_1 \end{bmatrix}, \text{ and} \tag{7}$$

$$p^{(1)} = \begin{bmatrix} p_{?A*} - (1/4) \\ p_{?T*} - (1/4) \end{bmatrix}. \tag{8}$$

Without loss of generality, let's assume $z_A^1 \neq 0$, and $z_A^1, z_T^1 \neq \frac{1}{4}$; this implies that $M_{12}^{(1)}, M_{22}^{(1)}$ are always non-zero. When any of the other entries of $M^{(1)}$ are zero, y_1 is trivially obtained. Furthermore, using the equation for $p_{?X*}$ for the letter X such that $z_X^2 \neq \frac{1}{4}$, x_2 can be computed. Thus, the only case left to be considered is when $z_A^1, z_T^1, z_A^2, z_T^2 \neq 0$. However, x_2, y_1 can be computed using Equations (4) and (5) when $M^{(1)}$ is invertible, which holds when

$$\frac{z_T^1 - 1/4}{z_T^2 - 1/4} \neq \frac{z_A^1 - 1/4}{z_A^2 - 1/4}.$$

When $M^{(1)}$ is singular, we append the system with another equation linear in (x_2, y_1). To that end, the probability of generating sequences of the form $AA*$ is given by:

$$p_{AA*} = x_1 z_A^1 x_2 z_A^2 + x_1 z_A^1 (1-x_2)\frac{1}{4} + (1-x_1)\frac{1}{4} y_1 z_A^1 + (1-x_1)\frac{1}{4}(1-y_1)\frac{1}{4}, \tag{9}$$

Rearranging,

$$p_{AA*} = x_1 z_A^1 \left(z_A^2 - \frac{1}{4}\right) x_2 + (1-x_1)\frac{1}{4}\left(z_A^1 - \frac{1}{4}\right) y_1 + x_1 z_A^1 \frac{1}{4} + (1-x_1)(\frac{1}{4})^2 \tag{10}$$

Consider the system $p^{(1)'} = M^{(1)'} w^{(1)}$ formed by appending Equation (10) to the 1st system of equations:

$$M^{(1)'} = \begin{bmatrix} x_1(z_A^2 - 1/4) & (1-x_1)(z_A^1 - 1/4) \\ x_1 z_A^1 (z_A^2 - 1/4) & (1-x_1)(1/4)(z_A^1 - 1/4) \end{bmatrix}, \tag{11}$$

$$w^{(1)} = \begin{bmatrix} x_2 \\ y_1 \end{bmatrix}, \text{ and} \tag{12}$$

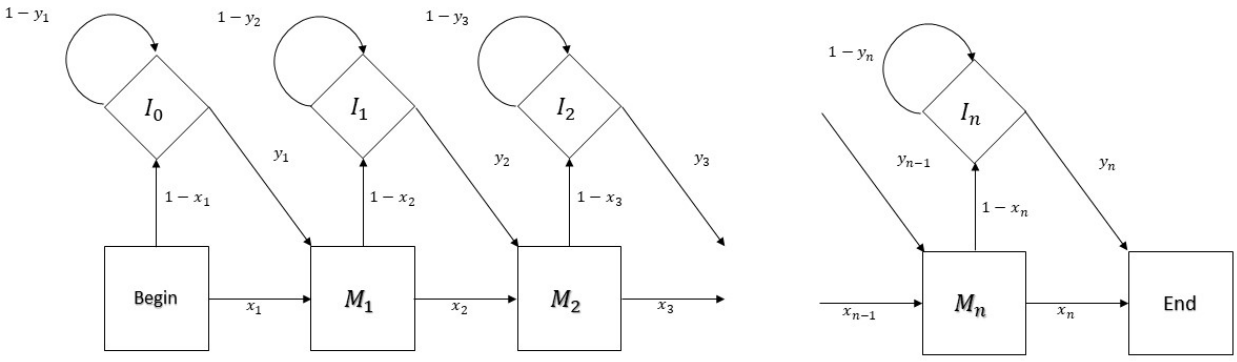

Figure 2: The standard profile HMM topology with n match states and no deletion nodes.

$$p^{(1)'} = \begin{bmatrix} p_{?A*} - (1/4) \\ p_{AA*} - x_1 z_A^1 (1/4) - (1-x_1)(1/4)^2 \end{bmatrix}. \tag{13}$$

$M^{(1)'}$ is invertible if $z_A^1, z_A^2 \neq \frac{1}{4}$, which is the assumption that we began with. Thus, the appended system can be used to compute x_2, y_1 whenever the 1st system is rank deficient.

We let $g_m(B)$ denote the probability of generating a string s whose m^{th} letter is B (where $B \in \{A, C, T, G\}$), but subject to the constraints that (a) $s[m]$ is not generated by M_m or I_{m-1}, and (b) $s[m-1]$ is not generated by I_{m-2}. Then, for all $B \in \{A, C, T, G\}$, $p_{?^{[m-1]}B*}$ (the probability that a randomly generated string s has $s[m] = B$) satisfies:

$$p_{?^{[m-1]}B*} = g_m(B) + p^{(\text{match}:m-1)} \left(x_m z_B^m + (1-x_m)\frac{1}{4} \right) + p^{(\text{insrt}:m-2)} \left(y_{m-1} z_B^{m-1} + (1-y_{m-1})\frac{1}{4} \right). \tag{14}$$

Thus, the $(m-1)^{\text{th}}$ system is given by Equations (15) and (16):

$$p_{?^{[m-1]}A*} = g_m(A) + p^{(\text{match}:m-1)} \left(x_m z_A^m + (1-x_m)\frac{1}{4} \right) + p^{(\text{insrt}:m-2)} \left(y_{m-1} z_A^{m-1} + (1-y_{m-1})\frac{1}{4} \right), \tag{15}$$

$$p_{?^{[m-1]}T*} = g_m(T) + p^{(\text{match}:m-1)} \left(x_m z_T^m + (1-x_m)\frac{1}{4} \right) + p^{(\text{insrt}:m-2)} \left(y_{m-1} z_T^{m-1} + (1-y_{m-1})\frac{1}{4} \right) \tag{16}$$

Thus, the $(m-1)^{\text{th}}$ system of equations is linear in variables x_m, y_{m-1}. Furthermore, the matrix $M^{(m-1)}$ associated with Equations (15) and (16), when written as $p^{(m-1)} = M^{(m-1)} w^{(m-1)}$, where $w^{(m-1)} = [x_m \quad y_{m-1}]^{\text{T}}$, is given by:

$$M^{(m-1)} = c_0 \begin{bmatrix} x_{m-1}(z_A^m - 1/4) & (1-x_{m-1})(z_A^{m-1} - 1/4) \\ x_{m-1}(z_T^m - 1/4) & (1-x_{m-1})(z_T^{m-1} - 1/4) \end{bmatrix}, \tag{17}$$

where $c_0 = \left(\prod_{i=1}^{m-2} x_i \right)$.

When $M^{(m-1)}$ is not invertible, consider the strings that have A in the $(m-1)^{\text{th}}$ and m^{th} positions. Thus, when the equation obtained by expressing the probability of generating the string $?^{[m-2]}AA*$ in terms of the transition probabilities and the emission probabilities is appended to the system, the new matrix $M^{(m-1)'}$ associated with

the system $p^{(m-1)'} = M^{(m-1)'} w^{(m-1)}$ is given by

$$M^{(m-1)'} = \begin{bmatrix} x_{m-1}(z_A^m - 1/4) & (1-x_{m-1})(z_A^{m-1} - 1/4) \\ x_{m-1} z_A^{m-1}(z_A^m - 1/4) & (1-x_{m-1})(1/4)(z_A^{m-1} - 1/4) \end{bmatrix}. \tag{18}$$

Following the argument presented for the 1st system, we conclude that $w^{(m-1)}$ can be computed.

To find y_n, consider all sequences of length $n+1$.

$$p_{[?]^{n+1}} = \sum_{i=1}^{n} (1-x_i) y_i \left(\prod_{j=1}^{n} x_j \right) / x_i. \tag{19}$$

Thus, y_n is obtained from this equation.

We point out that the letters A, T are representatives. In general, we pick the letters that give unique solutions to the systems of linear equations that are obtained in the proof. Hence, we have proved that if each of the match states is different (in distribution) from the insertion states, then the model is identifiable.

$\Rightarrow$: We now prove the other direction. We show that if the emission probabilities for a match state are identical (in distribution) to the insertion states, then the profile HMM is not identifiable. Specifically, we show (Figure 3) two different profile Hidden Markov models (each with a single match state) where the emission probability distribution for the match state is identical to that of the insertion states, and for which the two models define the same distribution on strings. In both models shown in Figure 3,

$$p_A = x_1 \frac{1}{4} x_2, \tag{20}$$

$$p_{AA} = (1-x_1)\frac{1}{4} y_1 \frac{1}{4} x_2 + x_1 \frac{1}{4}(1-x_2)\frac{1}{4} y_2, \tag{21}$$

and

$$\begin{aligned} p_{A^{[n]}} &= x_1 \frac{1}{4}(1-x_2)\frac{1}{4}(1-y_2)^{n-2}\frac{1}{4^{n-2}} y_2 \\ &+ (1-x_1)\frac{1}{4}(1-y_1)^{n-2}\frac{1}{4^{n-2}} y_1 \frac{1}{4} x_2 \\ &+ \sum_{n_1+n_2=n-3} (1-x_1)\frac{1}{4}(1-y_1)^{n_1}\frac{1}{4^{n_1}} y_1 \frac{1}{4}(1-x_2)\frac{1}{4}(1-y_2)^{n_2}\frac{1}{4^{n_2}} y_2, \end{aligned} \tag{22}$$

where $n \geq 3$ and $n_1, n_2 \geq 0$ in Equation (22). Thus, the profile HMM with one match state whose emission probability distribution is identical to that of the insertion states is not identifiable.

This proof can be extended to show that a profile HMM with no deletion nodes and arbitrary number of match states that has at least one match state whose distribution is identical to that of the insertion states is not identifiable. Consider a profile HMM with n match states as depicted by Model 1 in Figure 4. Without loss of generality, we assume that match state M_3 has the same distribution as that of the insertion states. Note that the highlighted region is exactly the toy example that we described above, so that Model 1 and Model 2 have identical sequence distributions. □

2.3 The standard profile HMM with one match state

We begin with a proof of non-identifiability of the standard profile HMM with one match state. We then identify the parameters that can be computed uniquely for the the standard profile HMM with one match state.

Theorem 2.2. *The standard profile HMM topology with one match state is non-identifiable.*

Proof. Consider the two models as shown in Figure 5. The emission distribution at the match state is the same across the models and is equal to $\{z_A = a, z_T = t, z_G = g, z_C = c\}$. The emission distribution at both the insertion states is equal to $\mathbb{P}_{ins}[A] = \mathbb{P}_{ins}[C] = \mathbb{P}_{ins}[T] = \mathbb{P}_{ins}[C] = \frac{1}{4}$. Let $\alpha_i^{'}, i \in \{1, \ldots, 12\}$, denote the transmission probabilities for Model 1 and $\alpha_i^{''}, i \in \{1, \ldots, 12\}$, denote the transmission probabilities for Model 2, both as shown in Figure 5. Let $X_1 X_2 \ldots X_k$ be an arbitrary DNA sequence of length k; we will show that the two models emit this sequence with the same probability. Let $p^{'}_{X_1 X_2 \ldots X_k}$ denote the probability with which Model 1 emits the sequence $X_1 X_2 \ldots X_k$; $p^{''}_{X_1 X_2 \ldots X_k}$ denotes the the probability with which Model 2 emits the same sequence. When $k = 1$ we obtain:

$$p^{''}_{X_1} - p^{'}_{X_1} = \frac{1}{4}(\alpha_3^{''}\alpha_{10}^{''}\alpha_{11}^{''} - \alpha_3^{'}\alpha_{10}^{'}\alpha_{11}^{'} + \alpha_9^{''}\alpha_{12}^{''}\alpha_6^{''} - \alpha_9^{'}\alpha_{12}^{'}\alpha_6^{'}) = 0. \tag{23}$$

When $k = 2$ we obtain:

$$\begin{aligned} & p^{''}_{X_1X_2} - p^{'}_{X_1X_2} \\ &= \left(\frac{1}{4}\right)^2 (0.3(\alpha_3^{''}\alpha_{10}^{''}\alpha_{11}^{''} - \alpha_3^{'}\alpha_{10}^{'}\alpha_{11}^{'}) + 0.6(\alpha_3^{''}\alpha_{10}^{''}\alpha_{12}^{''} - \alpha_3^{'}\alpha_{10}^{'}\alpha_{12}^{'}) \\ &\qquad + 0.4(\alpha_9^{''}\alpha_{12}^{''}\alpha_6^{''} - \alpha_9^{'}\alpha_{12}^{'}\alpha_6^{'})) \\ &= 0. \end{aligned} \tag{24}$$

Finally, for $k \geq 3$ we obtain:

$$\begin{aligned} & p^{''}_{X_1X_2\ldots X_k} - p^{'}_{X_1X_2\ldots X_k} \\ &= (\alpha_3^{''}\alpha_{10}^{''}\alpha_{11}^{''} - \alpha_3^{'}\alpha_{10}^{'}\alpha_{11}^{'})0.3^{k-1} + (\alpha_9^{''}\alpha_{12}^{''}\alpha_6^{''} - \alpha_9^{'}\alpha_{12}^{'}\alpha_6^{'})0.4^{k-1} \\ &\qquad + \left(\frac{1}{4}\right)^k (\alpha_3^{''}\alpha_{10}^{''}\alpha_{12}^{''} - \alpha_3^{'}\alpha_{10}^{'}\alpha_{12}^{'}) \sum_{n_1+n_2=k-2} 0.3^{n_1} 0.4^{n_2} \\ &= \frac{9}{400}(0.4^{k-1} - 0.3^{k-1}) - \frac{9}{400}(0.4^{k-1} - 0.3^{k-1}) = 0, \end{aligned} \tag{25}$$

□

Theorem 2.3. *Consider the standard profile HMM topology with one match state. If the model topology is given, then some (but perhaps not all) of the transition probabilities can be identified if the emission probabilities at the match state are not equal to that of the insertion states.*

Proof. Consider the standard profile HMM topology with one match state as shown in Figure 6. We will show that under the assumption of the theorem, $\alpha_2, \alpha_4, \alpha_6, \alpha_7$, and α_8 can be computed uniquely. Further, if $\alpha_2 \neq \alpha_6, \alpha_3\alpha_5 \neq \alpha_7\alpha_1$, then α_1 and the emission probabilities at the match state can be determined. We provide the proof for the case wherein $\mathbb{P}_{ins}[A] = \mathbb{P}_{ins}[C] = \mathbb{P}_{ins}[T] = \mathbb{P}_{ins}[C] = \frac{1}{4}$. For the more general case where the emission probabilities of the letters at the insertion states are different, the proof is a modification of the one provided. Let $\alpha_i, i \in \{1, \ldots, 12\}$, denote the transmission probabilities as shown in Figure 6. Consider a sequence of length one. The letter is emitted either by insertion

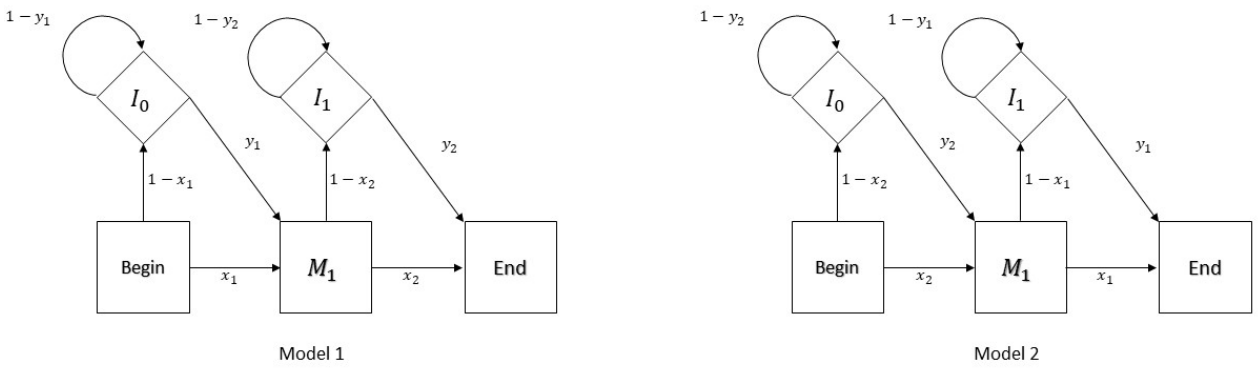

Figure 3: Two profile HMMs that have the same sequence distribution.

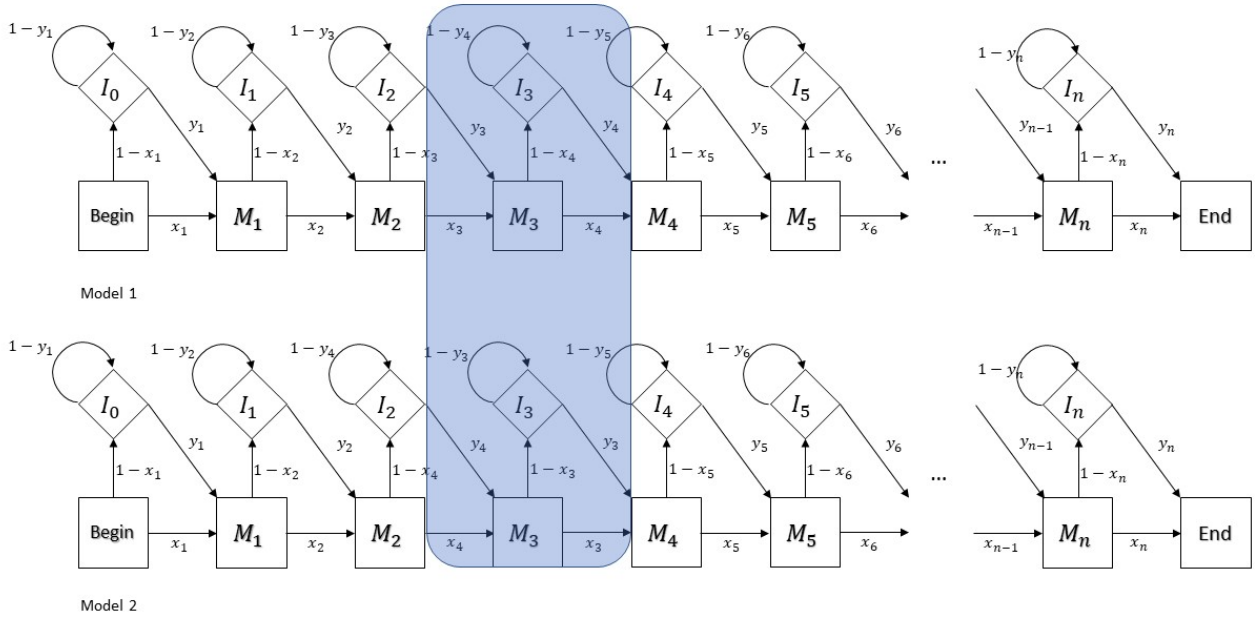

Figure 4: Two profile HMMs with n match states (but no deletion states) that have the same sequence distribution.

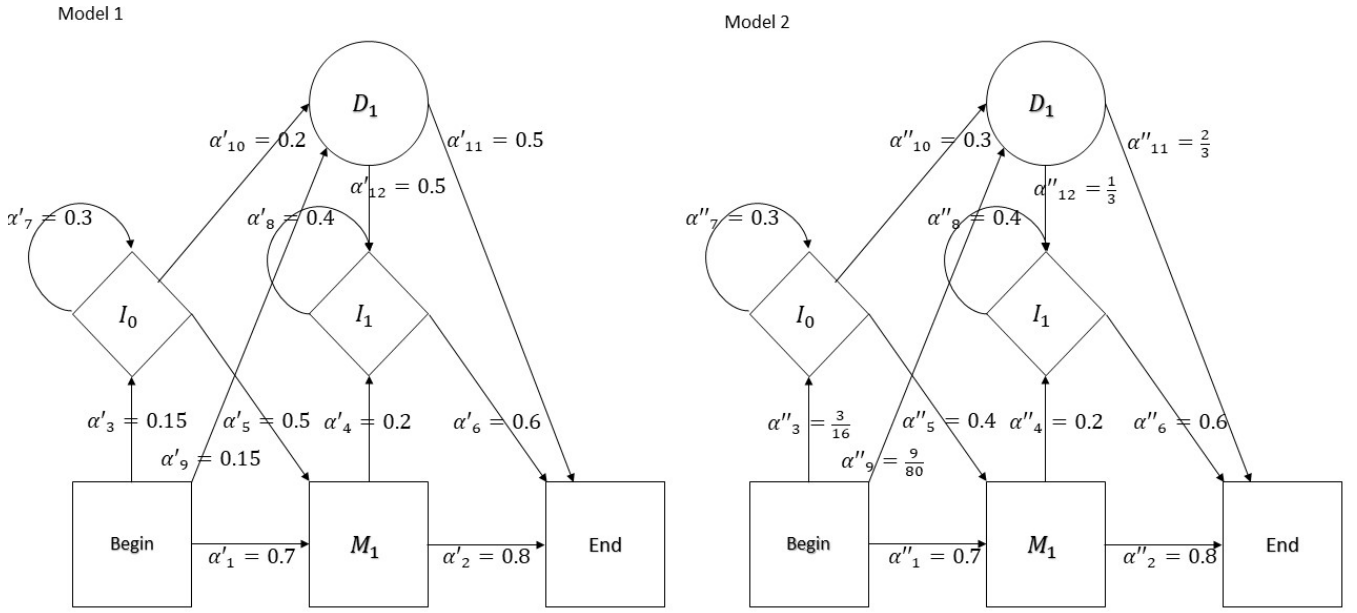

Figure 5: Two standard profile HMMs with one match state that define the same distribution on sequences, establishing that profile HMMs in the standard format are not identifiable (see Theorem 2.2).

states I_0 or I_1, or by the match state M_1. Thus the probability of emitting the letter B is given by

$$p_B = \alpha_3 \frac{1}{4} \alpha_{10} \alpha_{11} + \alpha_1 z^1_B \alpha_2 + \alpha_9 \alpha_{12} \frac{1}{4} \alpha_6. \tag{26}$$

Assume without loss of generality that $z^1_A \neq z^1_T$. Therefore,

$$p_A - p_T = \alpha_1 (z^1_A - z^1_T) \alpha_2 \neq 0. \tag{27}$$

We now consider sequences that begin with a particular letter B. Again, the first letter is generated either by insertion states I_0 or I_1, or by the match state M_1. Thus,

$$p_{B*} = \alpha_1 z^1_B + \alpha_3 \frac{1}{4} + \alpha_9 \alpha_{12} \frac{1}{4}. \tag{28}$$

Therefore,

$$p_{A*} - p_{T*} = \alpha_1 (z^1_A - z^1_T) \neq 0. \tag{29}$$

Dividing (27) by (29), we find

$$\alpha_2 = \frac{p_A - p_T}{p_{A*} - p_{T*}}. \tag{30}$$

Thus, $\alpha_4 = 1 - \alpha_2$ is also computed. Consider all sequences of the form $B_1 B_2$.

$$p_{B_1 B_2} = \alpha_1 z^1_{B_1} \alpha_4 \frac{1}{4} \alpha_6 + \alpha_3 \frac{1}{4} \alpha_7 \frac{1}{4} \alpha_{10} \alpha_{11} + \alpha_3 \frac{1}{4} \alpha_5 z^1_{B_2} \alpha_2 + \alpha_3 \frac{1}{4} \alpha_{10} \alpha_{12} \frac{1}{4} \alpha_6 + \alpha_9 \alpha_{12} \frac{1}{4} \alpha_8 \frac{1}{4} \alpha_6. \tag{31}$$

Therefore,

$$p_{AA} - p_{TA} = \alpha_1 (z^1_A - z^1_T) \alpha_4 \frac{1}{4} \alpha_6 \neq 0, \tag{32}$$

$$p_{AA} - p_{AT} = \alpha_3 \frac{1}{4} \alpha_5 (z^1_A - z^1_T) \alpha_2 \neq 0. \tag{33}$$

We now consider sequences that begin with two letters $B_1 B_2$.

$$p_{B_1 B_2 *} = \alpha_1 z^1_{B_1} \alpha_4 \frac{1}{4} + \alpha_3 \frac{1}{4} \alpha_7 \frac{1}{4} + \alpha_3 \frac{1}{4} \alpha_5 z^1_{B_2} + \alpha_3 \frac{1}{4} \alpha_{10} \alpha_{12} \frac{1}{4} + \alpha_9 \alpha_{12} \frac{1}{4} \alpha_8 \frac{1}{4}. \tag{34}$$

Therefore,

$$p_{AA*} - p_{TA*} = \alpha_1 (z^1_A - z^1_T) \frac{1}{4} \alpha_4 \neq 0. \tag{35}$$

Dividing (32) by (35), we find

$$\alpha_6 = \frac{p_{AA} - p_{TA}}{p_{AA*} - p_{TA*}}. \tag{36}$$

Thus, $\alpha_8 = 1 - \alpha_6$ is also computed. We now consider all sequences of form $B_1 B_2 B_3$.

$$p_{AAA} - p_{AAT} = \alpha_3 \frac{1}{4} \alpha_7 \frac{1}{4} \alpha_5 (z^1_A - z^1_T) \alpha_2 \tag{37}$$

Dividing (37) by (33), we get

$$\alpha_7 = 4 \frac{p_{AAA} - p_{AAT}}{p_{AA} - p_{AT}} \tag{38}$$

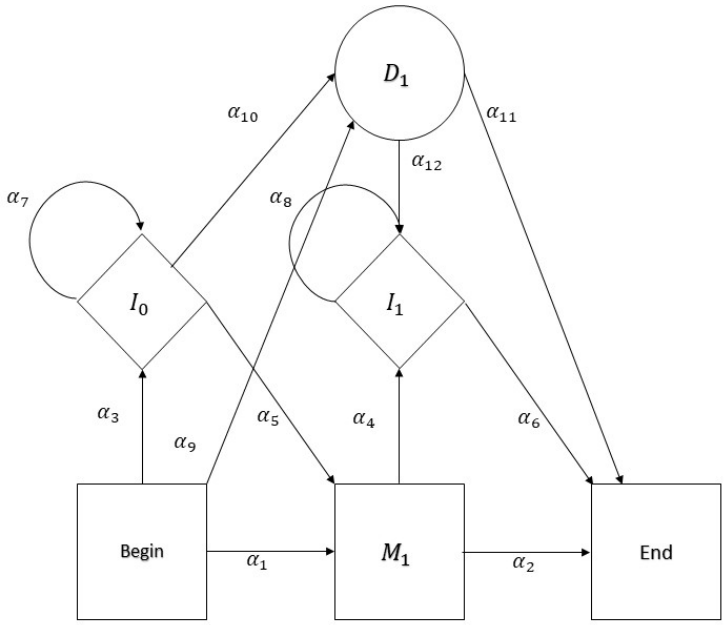

Figure 6: The standard profile HMM with one match state.

Dividing (33) by (27), we find that

$$\alpha_2\alpha_3 = \alpha_1 4\frac{p_{AA} - p_{AT}}{p_A - p_T} \tag{39}$$

Let p_ϵ denote the probability of not emitting any letter. To find $\alpha_1, \alpha_9, \alpha_{11}$, consider the following equations:

$$p_? = \alpha_1\alpha_2 + \alpha_3\alpha_{10}\alpha_{11} + \alpha_9\alpha_{12}\alpha_6 \tag{40}$$

$$p_{[?]^2} = \alpha_1\alpha_4\alpha_6 + \alpha_3\alpha_7\alpha_{10}\alpha_{11} + \alpha_3\alpha_5\alpha_2 + \alpha_3\alpha_{10}\alpha_{12}\alpha_6 + \alpha_9\alpha_{12}\alpha_8\alpha_6 \tag{41}$$

$$\alpha_{10} = 1 - \alpha_5 - \alpha_7 \tag{42}$$

$$\alpha_{12} = 1 - \alpha_{11} \tag{43}$$

$$\alpha_{11} = \frac{p_\epsilon}{\alpha_9} \tag{44}$$

Substituting equations (39), (42), (43) and (44) in (40) and (41), we obtain the following:

$$p_? = \alpha_1\alpha_2 + (\alpha_3(1-\alpha_7) - \gamma\alpha_1)\frac{p_\epsilon}{\alpha_9} + (\alpha_9 - p_\epsilon)\alpha_6 \tag{45}$$

$$p_{[?]^2} = \alpha_4\alpha_6\alpha_1 + \alpha_7\frac{p_\epsilon}{\alpha_9}(\alpha_3(1-\alpha_7) - \gamma\alpha_1) + \alpha_6\alpha_8(\alpha_9 - p_\epsilon) + \alpha_6\left(1 - \frac{p_\epsilon}{\alpha_9}\right)(\alpha_3(1-\alpha_7) - \gamma\alpha_1) + \gamma\alpha_2\alpha_1 \tag{46}$$

where $\gamma = 4\frac{p_{AA}-p_{AT}}{p_A-p_T}$. Thus, equations (45), (46) together with the equation $\alpha_1 + \alpha_3 + \alpha_9 = 1$ form a system of three equations in three variables (α_1, α_3, and α_9). Wolfram|Alpha returns a unique solution for α_1 and two pairs of solutions for (α_3, α_9) under the condition that $\alpha_1\alpha_7 \neq \alpha_3\alpha_5$ and $\alpha_2 \neq \alpha_3$. Since α_1 is unique, we can compute the emission distribution.

Equations (47), (48), and (49) are obtained from (27).

$$p_A - p_T = \alpha_1(z_A^1 - z_T^1)\alpha_2, \tag{47}$$

$$p_A - p_G = \alpha_1(z_A^1 - z_G^1)\alpha_2, \tag{48}$$

$$p_A - p_C = \alpha_1(z_A^1 - z_C^1)\alpha_2, \tag{49}$$

$$1 = z_A^1 + z_T^1 + z_G^1 + z_C^1. \tag{50}$$

Equations (47), (48), (49), and (50) together are a linear system of 4 equations with 4 unknowns, and can be expressed as $Mz^1 = p$ where

$$M = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{51}$$

$$z^1 = \begin{bmatrix} z_A^1 \\ z_T^1 \\ z_G^1 \\ z_C^1 \end{bmatrix}, \text{ and} \tag{52}$$

$$p = \begin{bmatrix} (p_A - p_T)/(\alpha_1\alpha_2) \\ (p_A - p_G)/(\alpha_1\alpha_2) \\ (p_A - p_C)/(\alpha_1\alpha_2) \\ (1)/(\alpha_1\alpha_2). \end{bmatrix} \tag{53}$$

Since M is invertible, z^1 can be obtained. □

2.4 Estimating parameters from finite data

Identifiability results establish what can be known from the true distribution, but do not directly imply that a statistically consistent method is possible. Here we describe how to estimate what can be estimated from data for the standard model, modified so that there are no deletion nodes, and discuss the amount of data that are needed to estimate the true topology and the numeric parameters (within some error threshold) with high probability.

One could leverage the ideas used in our proof techniques to reconstruct the model using empirical joint distributions obtained from the data. However, since the number of paths doubles from one system of equations to the next, such an approach is not efficient. Yet, some parameters of the model can still be estimated efficiently using our techniques. For example, the number of match states, and therefore the topology can be estimated from the shortest string produced.

Suppose we had N independent sequences that were generated by a specific profile HMM with n match states and no deletion states. The probability of not observing any sequence of length n is given by

$$\begin{aligned}\mathbb{P}[\text{all sequences have length } >n] &= \left(1 - \prod_{i=1}^{n} x_i\right)^N \\ &\leq \left(1 - x_{\min}^n\right)^N \\ &\leq \exp\{-x_{\min}^n N\}, \qquad (54)\end{aligned}$$

where $x_{\min} = \min_{1\leq i\leq n} x_i$. The probability of error decays exponentially with the number of sequences. Thus, if the transition probabilities from one match state to the next were all bounded from below, then a finite number $N' = \frac{1}{x_{\min}^n} \log\left(\frac{1}{\delta}\right)$ of independently generated sequences are sufficient for reconstructing the topology with confidence at least $1 - \delta$.

Other parameters such as emission probabilities of the match state, and a constant number of transition probabilities x_i's and y_i's can also be computed efficiently from the empirical distributions of sequences,- and their errors can be bounded.

3 CONCLUSION

In this text, we made the first strides towards completely characterizing the identifiability of profile hidden Markov models. We analyzed identifiability for the case where there are no deletion states, but otherwise all the properties of the standard model hold. For this case, Theorem 2.1 shows that the model is identifiable if and only if no match state has the same emission probability distribution as the insertion states. Further, we analyzed the question of identifiability for the special case of only one match state under the standard topology, and proved that it is not identifiable. In particular, we presented two models with different transition probabilities and showed that the probability of emitting any particular sequence is the same for the two models. This in turn implies that the standard profile HMM is non-identifiable. For the model with the standard topology and one match state, we also identified the parameters that can be computed uniquely. Characterizing partial identifiability for the standard topology with an unknown number of match states is still open.

4 ACKNOWLEDGMENTS

This research was supported by National Science Foundation grants ABI-1458652 and III:AF:1513629 to TW. This research began as a final project by the first author for the course Computer Science 581: Algorithmic Computational Genomics, taught by the second author at the University of Illinois at Urbana-Champaign in Spring 2018.

REFERENCES

[1] Frederick A Matsen, Elchanan Mossel, and Mike Steel. 2008. Mixed-Up Trees: The Structure of Phylogenetic Mixtures. *Bulletin of mathematical biology* 70 (2008), 1115–39. Issue 4.
[2] Elizabeth S Allman, Catherine Matias, and John A Rhodes. 2009. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* (2009), 3099–3132.
[3] Joseph T Chang. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences* 137, 1 (1996), 51–73.
[4] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.
[5] Ingo Ebersberger, Sascha Strauss, and Arndt von Haeseler. 2009. HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC evolutionary biology* 9, 1 (2009), 157.
[6] Sean R. Eddy. 1998. Profile hidden Markov models. *Bioinformatics (Oxford, England)* 14, 9 (1998), 755–763.
[7] Steven N Evans and Philip B Stark. 2002. Inverse problems as statistics. *Inverse problems* 18, 4 (2002), R55.
[8] David Freedman. 2005. *Statistical Models: Theory and Practice*. Cambridge University Press.
[9] Torben Friedrich, Birgit Pils, Thomas Dandekar, Jörg Schultz, and Tobias Müller. 2006. Modelling interaction sites in protein domains with interaction profile hidden Markov models. *Bioinformatics* 22, 23 (2006), 2851–2857.
[10] David Haussler, Anders Krogh, I. Saira Mian, and Kimmen Sjölander. 1993. Protein Modeling using Hidden Markov Models: Analysis of Globins. In *Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences*.
[11] Timo Koski. 2001. *Hidden Markov models for bioinformatics*. Vol. 2. Springer Science & Business Media.
[12] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of molecular biology* 235, 5 (1994), 1501–1531.
[13] Colby Long and Laura Kubatko. 2017. Identifiability and reconstructibility of species phylogenies under a modified coalescent. *arXiv preprint arXiv:1701.06871* (2017).
[14] Rachel J MacKAY. 2002. Estimating the order of a hidden Markov model. *Canadian Journal of Statistics* 30, 4 (2002), 573–589.
[15] Siavash Mirarab, Nam phuong Nguyen, and Tandy Warnow. 2012. SEPP: SATé-enabled phylogenetic placement. In *Pacific Symposium on Biocomputing*. 247–58.
[16] Elchanan Mossel and Sebastien Roch. 2012. Phylogenetic mixtures: concentration of measure in the large-tree limit. *The Annals of Applied Probability* 22, 6 (2012), 2429–2459.
[17] Nam-phuong Nguyen, Siavash Mirarab, Keerthana Kumar, and Tandy Warnow. 2015. Ultra-large alignments using phylogeny aware profiles. *Genome Biology* 16, 124 (2015). https://doi.org/10.1186/s13059-015-0688-z A preliminary version appeared in the Proceedings RECOMB 2015.
[18] Nam-phuong Nguyen, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. 2014. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* 30, 24 (2014), 3548–3555. https://doi.org/10.1093/bioinformatics/btu721
[19] Nam-phuong Nguyen, Michael Nute, Siavash Mirarab, and Tandy Warnow. 2016. HIPPI: highly accurate protein family classification with ensembles of hidden Markov models. *BMC Bioinformatics* 17 (Suppl 10) (2016), 765. Special issue for RECOMB-CG 2016.
[20] Judea Pearl and Michael Tarsi. 1986. Structuring causal trees. *Journal of Complexity* 2, 1 (1986), 60–77.
[21] Benjamin Schuster-Böckler and Alex Bateman. 2007. An introduction to hidden Markov models. *Current protocols in bioinformatics* 18, 1 (2007), A.3A.1–A.3A.9.
[22] Kimmen Sjölander. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 20, 2 (2004), 170–179.
[23] A. W. van der Vaart. 1998. *Asymptotic Statistics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802256
[24] Ziheng Yang. 2006. *Computational Molecular Evolution*. Oxford University Press.
[25] Zemin Zhang and William I Wood. 2003. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* 19, 2 (2003), 307–308.