# Duality between the local score of one sequence and constrained Hidden Markov Model

Sabine Mercier, Gregory Nuel

**myjournal manuscript No.**
(will be inserted by the editor)

# Duality between the local score of one sequence and constrained Hidden Markov Model

**Sabine Mercier · Grégory Nuel**

Received: date / Accepted: date

**Abstract** We are interested here in a theoretical and practical approach for detecting atypical segments in a multi-state sequence. We prove in this article that the segmentation approach through an underlying constrained Hidden Markov Model (HMM) is equivalent to using the maximum scoring subsequence (also called local score), when the latter uses an appropriate rescaled scoring function. This equivalence allows results from both HMM or local score to be transposed into each other. We propose an adaptation of the standard forward-backward algorithm which provides exact estimates of posterior probabilities in a linear time. Additionally it can provide posterior probabilities on the segment length and starting/ending indexes. We explain how this equivalence allows one to manage ambiguous or uncertain sequence letters and to construct relevant scoring functions. We illustrate our approach by considering the TM-tendency scoring function.

**Keywords** Local score · maximum scoring subsequence · HMM · posterior distribution · forward/backward · biological sequence analysis

## 1 Introduction

Since the development of biological sequence databases in the 80s, the extraction of information from such an enormous amount of data has been the subject of great interest. Considering the size of the data, sequence analysis has been largely developed with both mathematical and computing challenges.

Sabine Mercier
Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse 2 Jean Jaurès
E-mail: mercier@univ-tlse2.fr

Grégory Nuel
Laboratoire de Probabilités Statistique Modélisation (LPSM), CNRS 8001, Sorbonne Université
E-mail: gregory.nuel@math.cnrs.fr

Extracting information from biological sequences includes a large range of areas such as Markov models (Durbin et al., 1998), similarity (Pearson, 2013), local pairwise or multiple alignments (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Altschul et al., 1990; Sievers et al., 2011), words counts (Robin et al., 2005), segmentation (Keith, 2008; Devillers et al., 2011; Luong et al., 2013), including statistical significance that is omnipresent in biological sequence analysis (Karlin, 2005; Mitrophanov and Borodovsky, 2006).

In this context, it is often interesting to point out homogenous regions in sequence data. For that purpose, two principal techniques are largely used: Hidden Markov Models (HMMs) and Score-Based approaches.

**HMMs:** HMMs consist in two structures, an underlying model for unobserved state defined on finite dimensional spaces and one for the observations. These models have been extensively used in biological sequence analysis in a variety of problems in molecular biology such as pairwise and multiple sequence alignments (Arribas-Gil et al., 2012), gene prediction (Munch and Krogh, 2006; Krogh et al., 2001), classification (Won et al., 2007), and many others. See Yoon (2009) for a tutorial review of HMMs and their applications where three types of HMMs are principally presented, profile-HMMs, pair-HMMs, and context-sensitive models.

Note that HMM are usually very specific and adapted to the context of the study and the kind of subsequence to be highlighted. For example, see Krogh et al. (2001) for transmembrane helices research; or Borodovsky and McIninch (1993) for a HMM modeling protein-coding gene. Note that there also exists a large number of HMM variants to meet the needs of various applications (see Yoon, 2009, for more explanation). Usually the Viterby algorithm is used (Viterbi, 1967) to find the most probable state sequence among all possible ones conditional on the observed sequence and is then considered as deterministically correct. Patterns in the state sequence are then found by analyzing this most probable sequence. In Aston and Martin (2007) the authors propose a method based on forward and backward quantities (Durbin et al., 1998) to compute distributions of pattern in the hidden state sequence of a HMM conditional on the observed sequence which is more robust to the uncertainty of the hidden state sequence.

**Score-Based approaches:** In biological sequence analysis, a real value called *score*, and denoted by $f$, is assigned to each component of a sequence. These scores reflect a physico-chemical property of the component which depends on the studied context (Kyte and Doolittle, 1982; Dayhoff et al., 1978; Zhao and London, 2006). The idea is to find regions of the sequence where the cumulated score is significantly high. For that purpose, one possible approach consists of scanning the sequence using sliding windows of a given length $\ell$ and calculating a cumulative score for each window. This leads to a graphical representation where the maximal region of length $\ell$ can easily be observed. While this approach is appealing for sequence analysis, its major drawback is that it needs to make a choice on $\ell$.

The alternative to sliding windows, when there is no "natural" value for $\ell$, is the local score. The local score (see Karlin and Altschul, 1990, or Eq. (1)

in Section 2) is the maximal cumulative score that can be found over every segment of any position and any length in the sequence. It corresponds to the maximal level of the desired signal that can be locally found in a sequence. The term of "local score" derives from the one of "score of local alignment" of two sequences compared to the one of "global score alignment" used for example in Waterman (1995). In his book, Waterman gathers the results available at that time on the mathematical behavior of the local score of sequence alignment. Due to the complexity of the mathematical tool, first results stand on the distribution of the local score for one sequence. This can bring confusion, and this is also increased by the fact that the BLAST software (Basic Alignment Search Tool) used a generalization of the work of Karlin *et al.* for one sequence to compute the distribution of the score of the maximum similarity segments for sequence alignments. There also exists a researcher community which uses the appellation "maximum scoring subsequence" as in Ruzzo and Tompa (1999). In fact, from our point of view, the term of "maximal scoring subsegment" should be more appropriate as one is interested in this context in the contiguous part of the sequence. From now on we will use the term of "local score" to refer to the maximal scoring subsegment of one sequence.

Basically, all sliding window approaches can be replaced by their local score counterparts with several dramatic advantages: no window size to choose, efficient dynamic programming algorithms, and many probabilistic/statistical results (see below). If one knows the target length of the atypical regions to be detected, it remains reasonable to use sliding windows. In some contexts, sliding windows can seem to be a little bit easier to interpret graphically than the cumulative score of the local score function. The definition of the local score implies that the scoring functions used are composed of both positive and negative scores. Indeed, if only positive scores are possible the whole sequence will always achieve the maximal possible score without highlighting the local region as expected.

Finally, scan Statistics are also a largely used tool to extract atypical regions in sequences. They are defined as the maximal cumulative occurrences of a given event in sliding windows of fixed length: like for example the occurrence of palindromes in a nucleic sequence ; or the number of CpG islands (see Takai and Jones, 2002, 2003, for a scan statistic approach) (see Guéguen, 2005, for a HMM approach). The sliding window approach uses a scoring function which can have negative and/or positive scores whereas scan statistics are based on counts of an observed event and thus scores are only positive, often modeled by binomial or Poisson distributions. Scan statistic approaches has been largely studied since J. Naus first published on the problem in the 1960s (Naus, 1982; Glaz et al., 2001). See Chen and Glaz (2016) or Zhao and Glaz (2017) for recent developments on the topic.

*Our contribution*

The main purpose of the present work is to establish a duality between a generative constrained HMM framework and a Score-Based one. We prove that

under specified hypotheses, HMMs and Score-Based approaches can in fact be equivalent in the sense that they provide the same segmentation distributions. This main result brings very interesting corollaries and applications using the fact that results and or information from both approaches can be transferred to the other: we demonstrate how we have derived several interesting probabilistic results for the local score approach including scoring function rescaling; we present how an adaptation of the forward and backward algorithm allows to compute the full posterior distribution including location, length, etc. of local score segments, based on a given sequence observation; how score ambiguity or weighted observations can be considered. Supervised or unsupervised score learning is also possible but will not take place in this article but in further works focusing more on the statistical aspects of the problem.

**Related work:** In Aston and Martin (2007) the authors develop a computational method allowing to obtain the distribution of patterns (in fact, regular expressions) in the unobserved sequence of hidden states of a HMM conditional to the observed sequence. The approach is based on deterministic finite state automata and Markov chain embedding techniques combined with the computation of the posterior hidden state distribution as a heterogeneous Markov chain. The purpose of this interesting contribution is to study the occurrence of patterns of interest (number, positions) in the hidden sequence such as 0101010101 (example of pattern from Aston and Martin, 2007). The problem they consider is therefore quite different from our objective of establishing the posterior distribution of atypical segment location.

In Wolfsheimer et al. (2012) the authors proposed a work for the classical pairwise alignment and pair HMMs. Our contribution can therefore be considered as an extension of these previous results to the local score of one sequence.

**Outline of the paper:** Section 2 is devoted to the local score with a state of the art and the probabilization of the segmentation space using the Gibbs measure. Section 3 is devoted to the generating model and the segmentation corresponding to the hidden states of the underlying constrained HMM. The complete connection between the use of the local score and the generating model to highlight atypical segment in a given sequence is established in Section 4. Section 5 presents different applications of the possibilities that the equivalence can offer. Technical details including the Forward and Backward quantities are given in Appendix A. Implementation of the main examples can be found in Appendix B. Appendix C proposes a computation function to rescale the scoring function.

## 2 Local score

### 2.1 State of art for the local score

The local score, usually noted $H_n$, of a random sequence of length $n$, has been widely studied since its definition in the 1990s. The establishment of the

distribution of the local score can be assessed in different contexts depending on the average score $\mathbb{E}[f]$ (negative, positive or equal to 0), the sequence length $n$ (short, medium or long sequences), and how the sequences are modeled (by a sequence of Independent and Identically Distributed variables –i.i.d.–) or by a Markov chain). Each context uses very different mathematical tools to establish results.

In Karlin and Altschul (1990) and Karlin and Dembo (1992) the authors show that for a sequence of i.i.d. random variables, a scoring function with a negative expected score function, $\mathbb{E}[f] < 0$, and possible positive scores, $\mathbb{P}(f > 0) > 0$, that the distribution of the local score has asymptotically a Gumbel distribution. In practice, the sequence length $n$ must be higher than $10^3$ for this asymptotic approximation to be valid. An improvement of this result is proposed in Mercier et al. (2003) taking into account additive and correcting terms. In 2001 (see Mercier and Daudin, 2001), an exact method is proposed using Markov chain theory, that allows to establish the exact $p$-value of $H_n$ in the i.i.d. model, whatever the average score is. In theory, the exact method is applicable whatever the length $n$, but in practice it is accurate for small sequences with $n$ up to $10^3$ for an acceptable computational time. In the special case where $\mathbb{E}[f] = 0$, the authors of Daudin et al. (2003) demonstrate an asymptotic behavior of the local score significance based on Brownian motion theory. See Lagnoux et al. (2017) for a review and illustrations in the case of i.i.d. sequences.

When the score sequence is assumed to be Markovian rather than i.i.d., there are fewer results. In Nuel (2006) and Hassenforder and Mercier (2007) the authors established the exact distribution of the local score of a Markov chain whatever the average score, but in practice, this result needs a maximal sequence length of several hundreds of components. For the local score of one Markov chain with a non positive mean score, the convergence of the distribution of the local score $H_n$ to a Gumbel distribution is confirmed using simulations (see Robelin, 2005; Guedj et al., 2006; Fariello et al., 2017) and demonstrated for a random scoring function in Karlin and Dembo (1992). An improvement of this last result is proposed in Grusea and Mercier (2020) for a lattice scoring function.

Some research has been done on the length of the local score realization. Motivated by sequence comparison, in Arratia and Waterman (1989), the authors considered the longest segment of a random sequence $X$ taking values in $\{0; 1\}$ with a proportion of 1 being larger than a given threshold. In a more general context, assuming $\mathbb{E}[X] < 0$ and $\mathbb{P}(X > 0) > 0$, Dembo and Karlin (1991a,b) proved an asymptotic behavior of the length of this optimal segment when the length of the sequence goes to infinity. In another context, (Karlin and Ost, 1988) established a classical extremal type limit law for the length of common words among a set of random sequences. Reinert and Waterman (2007) also gives a result on the distribution of the length of the longest exact match between two random sequences. In Chabriac et al. (2014), using Brownian motion theory, the authors established an asymptotic distribution on the pair, local score and local score length.

Statistical results on the location of local score segment are scarce. In Lagnoux et al. (2015), the authors established that for very long sequences, the local score is realized in the last Lindley excursion of the score sequence (that is the sequence of the corresponding score of each component using a given score scale (see Mercier and Daudin, 2001; Lagnoux et al., 2015, 2017, for a more detailed definition)) with an asymptotic probability around 1/3. But this last result is not accurate enough to know exactly where the local score segment begins. In Lagnoux et al. (2019), an asymptotic density for the index position of the local score in the sequence is exposed. These results, which are based on Brownian motion theory, can be useful for sequence lengths larger than $10^5$.

Finally, let us point out that the algorithmic aspects of the problem also attracted attention. See Altschul et al. (1990) for the well known BLAST software or Needleman and Wunsch (1970); Smith and Waterman (1981) for the original algorithms of pairwise global and local score research, or Ruzzo and Tompa (1999).

## 2.2 Local score and Segmentation

Let $\mathbb{A} = A_1 \dots A_n \in \mathcal{A}^n$ be a given sequence (typically, $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ for a DNA sequence, or the set of the 20 amino acids for proteins) and $f : \mathcal{A} \to \mathbb{R}$ be a scoring function. Let us define the local score (see Karlin and Altschul, 1990) of the sequence $\mathbb{A}$ based on the scoring function $f$ as:

$$H_f(\mathbb{A}) \stackrel{\text{def}}{=} \max_{[i,j]} \sum_{k \in [i,j]} f(A_k) \tag{1}$$

where $[i, j]$ could possibly be empty (hence $H_f(\mathbb{A}) \geq 0$). Usually, the local score is denoted $H_n$ for a sequence of length $n$. We need in the sequel to focus on the scoring function $f$ used to compute the local score instead of the length of the sequence, and we voluntary change here the notation.

Let us define a segmentation sequence $S = S_1 S_2 \dots S_n \in \mathcal{S} \stackrel{\text{def}}{=} \{S \in \{1, 2, 3\}^n, S_i - S_{i-1} \in \{0, 1\} \text{ for } i = 1, \dots, n\}$ with $S_0 = 1$ by convention.

**Remark 1** *We can define a bijection between the set $\mathcal{S}$ of segmentation and the set of segments $I$ include in $\{1, \dots, n\}$. For all $S \in \mathcal{S}$, we can define $I$ the indices for which the segmentation values are equal to 2. Conversely, for any interval (possibly empty) the associated segmentation takes value 1 before a segment of interest, 2 in the segment, and 3 after the segment.*

*In the following example, we take the segment $I = [2, 4]$ for a length sequence $n = 5$ and we give the corresponding segmentation $S$. Notation as $S_I$ (respectively $I_S$) will be used in the sequel to indicate the corresponding segmentation (resp. segment) of a given segment $I$ (resp. segmentation $S$).*

$$\begin{aligned} \textit{Index} &= 1\ 2\ 3\ 4\ 5 \\ S \quad &= 1\ 2\ 2\ 2\ 3 \end{aligned}$$

Extending our model to situations where there is more than one atypical segment is possible and this can be clearly seen as a great improvement for the local score approach. In practice, this extension is quite straightforward since we just have to expand the hidden space of $S_i$ (e.g. $S_i \in \{1,2,3,4,5\}$ for up to two segments, the atypical states corresponding to $S_i = 2$ or 4). The further computational complexity would simply increase linearly with the number of atypical segments. We chose to present the case standing on exactly one segment to make the presentation clearer. An application with several atypical segments is presented in Lefebvre et al. (2020).

For any sequence $\mathbb{A} \in \mathcal{A}^n$ and segmentation $S \in \mathcal{S} \subset \{1,2,3\}^n$, let us define:

$$H_f(S|\mathbb{A}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathbf{1}_{\{S_i=2\}} f(A_i) = \sum_{i \in I_S} f(A_i)$$

and we therefore have $H_f(\mathbb{A}) = \max_{S \in \mathcal{S}} H_f(S|\mathbb{A})$.

## 2.3 Gibbs Distribution

The local score focuses on the segment that realizes the maximum scoring subsegment. But several segments can realize the local score. Moreover one biologist can also be interested in some suboptimal segments. If we consider that the "energy" of each subsegment location is given by its associated scoring, it is therefore natural to probabilize the segmentation space on the basis of this "energy". In statistical physics, this is classically achieved by introducing a Gibbs' (or Boltzman's) distribution where the probability of a configuration of the system is proportional to the energy in a log-scale. We hence propose

$$\forall I = [i,j] \in \mathcal{I}, \quad \mathbb{P}^{(f,T)}_{\text{Gibbs}}(I|\mathbb{A}) \propto \exp\left(\frac{1}{T}\sum_{k=i}^{j} f(A_k)\right) \tag{2}$$

where $\sum_{k=i}^{j} f(A_k)$ is the subsegment scoring, the energy, and the parameter $T > 0$ is called the temperature. Note that $\mathbb{P}^{(f,T)}_{\text{Gibbs}}(I|\mathbb{A})$ tends towards a Dirac distribution when $T \to 0$, for which we recover the classical local score that focuses only on the maximum scoring subsegment without considering the suboptimal segments; and $\mathbb{P}^{(f,T)}_{\text{Gibbs}}(I|\mathbb{A})$ tends towards a uniform distribution when $T \to \infty$. The temperature $T$ therefore is a contrast parameter.

To illustrate the Gibbs' distribution on the segments, let us consider a simulated sequence of length $n = 40$ taking its values over the four nucleotide alphabet $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ and the following scoring function $f(\mathtt{A}) = f(\mathtt{C}) = -2$ and $f(\mathtt{G}) = f(\mathtt{T}) = +1$. Figure 1 represents the probabilities of the 821 different segments for different temperatures $T = 0.1$ (top left Panel), $T = 0.6$ (top right Panel), $T = 2$ (bottom left Panel) and $T = 4$ (bottom right Panel). The local score segment is highlighted in the top left Panel as the only one with non negligible probability. In the top right Panel, suboptimal segments start appearing. For a larger $T$, the probability landscape becomes richer (bottom

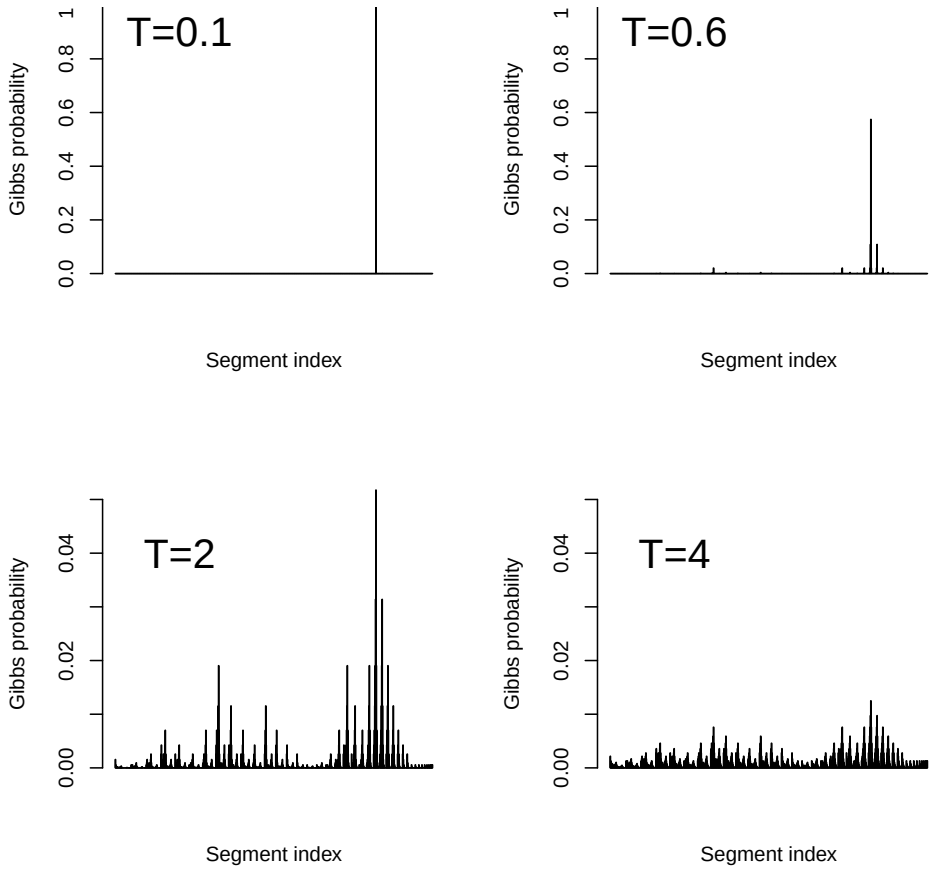


**Fig. 1** Gibbs' distributions of the 821 segments of a sequence of length $n = 40$ for different temperatures: Top left Panel with $T = 0.1$ ; Top right Panel with $T = 0.6$ ; Bottom left Panel with $T = 2$ and bottom right with $T = 4$. Note the scale change of the $x$ axis from the top panels to the bottom panels.

panels with a change of $x$ axis scale). The question is then to choose an adapted $T$ value to highlight *interesting* information of the *whole* sequence. Section 4 explains how to choose $T$ in a canonical way under certain hypotheses.

## 3 Generating Model

Let us now consider an alternative formulation of the problem through a Hidden Markov Model. It implies an underlying generating model that allows probabilities on the space of sequence segments.

Let $q_0(\cdot)$ and $q_1(\cdot)$ be two multinomial distributions over $\mathcal{A}$. For any $\mathbb{A} \in \mathcal{A}^n$ and $S \in \mathcal{S} \subset \{1,2,3\}^n$ we define:

$$\mathbb{P}^{q_0,q_1}_{\text{HMM}}(\mathbb{A}|S) \overset{\text{def}}{=} \prod_{i=1}^{n} q_0(A_i)^{\mathbf{1}_{\{S_i \neq 2\}}} q_1(A_i)^{\mathbf{1}_{\{S_i=2\}}} \tag{3}$$

with $\mathbb{P}(A_i|S_i \neq 2) = q_0(A_i)$ and $\mathbb{P}(A_i|S_i = 2) = q_1(A_i)$ ($q_0$ to generate $A_i$ when $S_i \in \{1,3\}$ and $q_1$ to generate $A_i$ when $S_i = 2$). When non-ambiguous, we will drop "HMM" and "$q_0, q_1$" for the sake of simplification.

If we assume that all $S \in \mathcal{S}$ are equiprobable we can deduce from (3) the probability of a segmentation conditional to the observed sequence as follows:

$$\mathbb{P}^{q_0,q_1}_{\text{HMM}}(S|\mathbb{A}) = \frac{1}{Z} \times \mathbb{P}^{q_0,q_1}_{\text{HMM}}(\mathbb{A}|S) \tag{4}$$

where the normalization factor $Z$ is given by:

$$Z \overset{\text{def}}{=} \sum_{S \in \mathcal{S}} \mathbb{P}^{q_0,q_1}_{\text{HMM}}(\mathbb{A}|S) = \sum_{S \in \{1,2,3\}^n} \mathbf{1}_{\{S \in \mathcal{S}\}} \mathbb{P}^{q_0,q_1}_{\text{HMM}}(\mathbb{A}|S) \tag{5}$$

This assumption seems reasonable as we do not have any prior information on the atypical segment.

With such a uniform prior on the segmentation space $\mathcal{S}$, it is clear that the prior probability of having no atypical segment is proportional to 1 ($S = 11\ldots1$ being the only sequence with no atypical segment) while the prior probability of having one atypical segment is proportional to $\binom{n+1}{2}$ (the cardinal of $\mathcal{S} \setminus \{11\ldots1\}$). As a consequence, the model with a uniform prior on $\mathcal{S}$ will always favor the presence of one atypical segment, even in the case where there is none. In order to avoid this issue, it is necessary to account for the number of configurations of both events in the prior distribution. See Section A.2 for details.

One should note that the Markov chain defined by $(\mathbb{A}, S)$ is constrained. Indeed, as we can see in the definition of the set $\mathcal{S}$ the transitions between the states are given. It is a very simple HMM model with only three states but because of this constraint inference differs from classical HMM inference by few details (see Appendix A for more explanation).

*Posterior probabilities*

Inspired from classical HMM inference (see Rabiner, 1989; Durbin et al., 1998), we adapted the so-called forward and backward quantities denoted by $F_j(k)$ and $B_j(k)$ for $j = 1, \ldots, n$ and $k \in \{1,2,3\}$ (see Appendix A.2 for definitions). We can exploit the forward and backward quantities to compute quantities of interest such as $\mathbb{P}(S_j = k|\mathbb{A})$ for $1 \leq j \leq n$, $\mathbb{P}(S_{j-1} = k, S_j = \ell|\mathbb{A})$ for $1 < j \leq n$.

– $\mathbb{P}(S_j = k|\mathbb{A})$ is very interesting for $j = n$ since it provides the posterior probability of having zero or one segment generated by $q_1$

$$\mathbb{P}(\text{no segment}|\mathbb{A}) = \mathbb{P}(S_n = 1|\mathbb{A})$$

$$\mathbb{P}(\text{one segment}|\mathbb{A}) = \mathbb{P}(S_n = 2 \text{ or } 3|\mathbb{A}) \ .$$

– $\mathbb{P}(S_{j-1} = k, S_j = \ell|\mathbb{A})$ with $k = 1$ and $\ell = 2$ (resp. $k = 2$ and $\ell = 3$) corresponds to the posterior probability that the segment starts (resp. ends) in position $j > 1$ (resp. $j - 1$) but we need to normalize by the posterior probability of having a non-empty segment:

$$\mathbb{P}(\text{segment starts at time } 1|\mathbb{A}) = \frac{\mathbb{P}(S_1 = 2|\mathbb{A})}{\mathbb{P}(S_n = 2 \text{ or } 3|\mathbb{A})}$$
$$\mathbb{P}(\text{segment starts at time } j|\mathbb{A}) = \frac{\mathbb{P}(S_{j-1} = 1, S_j = 2|\mathbb{A})}{\mathbb{P}(S_n = 2 \text{ or } 3|\mathbb{A})} \text{ for } 1 < j \leq n \ .$$

$$\mathbb{P}(\text{segment ends at time } n|\mathbb{A}) = \frac{\mathbb{P}(S_n = 2|\mathbb{A})}{\mathbb{P}(S_n = 2 \text{ or } 3|\mathbb{A})}$$
$$\mathbb{P}(\text{segment ends at time } \ell|\mathbb{A}) = \frac{\mathbb{P}(S_\ell = 2, S_{\ell+1} = 3|\mathbb{A})}{\mathbb{P}(S_n = 2 \text{ or } 3|\mathbb{A})} \text{ for } 1 \leq \ell < n \ .$$

Different implementations are proposed in Appendix B depending on sequence length, short sequence (see Appendix B.1) or long sequences (see Appendix B.2). We also propose in Appendix B.3 an implementation to compute the posterior probability that a segment has a given length. Illustration of the computation of those quantities is given in Section 5.1.

## 4 Equivalence between the two probability distributions

Let us consider a sequence $\mathbb{A} \in \mathcal{A}^n$. Denote by $\mathcal{M}_\mathcal{A}$ the set of multinomial distributions on $\mathcal{A}$. Due to Remark 1 we use the notation $\mathbb{P}^{f,1}_{\text{Gibbs}}(S|\mathbb{A})$ instead of $\mathbb{P}^{f,1}_{\text{Gibbs}}(I|\mathbb{A})$.

**Theorem 1**

$$\forall (q_0, q_1) \in \mathcal{M}^2_\mathcal{A}, \quad \exists!\ \sigma : \mathcal{A} \to \mathbb{R} \ \textit{such as}\ \mathbb{P}^{q_0,q_1}_{HMM}(S|\mathbb{A}) \overset{\forall S \in \mathcal{S}}{=} \mathbb{P}^{\sigma,1}_{Gibbs}(S|\mathbb{A})$$

*and* $\sigma(a) \overset{\forall a \in \mathcal{A}}{=} \log\left(\frac{q_1(a)}{q_0(a)}\right)$.

*Proof* Supposing the uniform distribution on $\mathcal{S}$ such as $\mathbb{P}(S) = 1/|\mathcal{S}|$ for all $S \in \mathcal{S}$, we then have with $Z$ defined in (5)

$$\mathbb{P}^{q_0,q_1}_{\text{HMM}}(S|\mathbb{A}) = \frac{1}{Z}\mathbb{P}(\mathbb{A}|S) \propto \frac{\mathbb{P}(\mathbb{A}|S)}{\mathbb{P}(\mathbb{A}|S = 1 \ldots 1)}$$

and

$$\frac{\mathbb{P}(\mathbb{A}|S)}{\mathbb{P}(\mathbb{A}|S=1\dots1)} = \frac{\prod_{i=1}^{n} q_0(A_i)^{\mathbf{1}_{\{S_i\neq 2\}}} q_1(A_i)^{\mathbf{1}_{\{S_i=2\}}}}{\prod_{i=1}^{n} q_0(A_i)}$$
$$= \prod_{i,S_i=2} \frac{q_1(A_i)}{q_0(A_i)} = \exp\left(\sum_{i,S_i=2} \log \frac{q_1(A_i)}{q_0(A_i)}\right) .$$

The function defined as $\sigma(a) \overset{\forall a\in\mathcal{A}}{=} \log(q_1(a)/q_0(a))$ verifies the hypothesis. The unicity can be proved as follows. Let $f : \mathcal{A} \to \mathbb{R}$ such as $\mathbb{P}(S|\mathbb{A}) \overset{\forall S\in\mathcal{S}}{\propto} \exp\left(\sum_{i,S_i=2} f(A_i)\right)$. Then $\forall a \in \mathcal{A}$ define the segmentation such as $S_i = 2$ iff $i = \inf\{1 \leq j \leq n : A_j = a\}$, $S_j = 1$ for $0 \leq j \leq i-1$ and $S_j = 3$ for $i+1 \leq j \leq n+1$. Applying the hypothesis for those segmentations, we get $\forall a \in \mathcal{A}$, $\exp\left(\log \frac{q_1(A_i)}{q_0(A_i)}\right) = f(A_i)$ that leads to $f = \sigma$.

**Theorem 2**
*$\forall q_0 \in \mathcal{M}_{\mathcal{A}}$ and $\forall f : \mathcal{A} \to \mathbb{R}$ verifying $\sum_a q_0(a)f(a) < 0$ and $\exists a, f(a) > 0$,*

$$\exists!\ T > 0 \text{ and } q_1 \in \mathcal{M}_{\mathcal{A}}, \text{ such that } \mathbb{P}^{q_0,q_1}_{HMM}(S|\mathbb{A}) \overset{\forall S\in\mathcal{S}}{=} \mathbb{P}^{f,T}_{Gibbs}(I_S|\mathbb{A})$$

*and $q_1(a) \overset{\forall a\in\mathcal{A}}{=} q_0(a)\exp(f(a)/T)$. We will call $f/T$ the rescaled scoring function according to $f$.*

*Proof* Let us consider the following Lemma.

**Lemma 1** *For any scoring function $f : \mathcal{A} \to \mathbb{R}$ and any background frequency $q_0(\cdot)$ the two following properties are equivalent:*

*i)* $\exists!\ \rho > 0, \sum_a q_0(a)\exp(\rho f(a)) = 1$
*ii)* $\sum_a q_0(a)f(a) < 0$ *and* $\exists a, f(a) > 0$

*Proof* Consider $g : [0,\infty) \to \mathbb{R}$ with $g(r) = \sum_a q_0(a)\exp(rf(a)) = \mathbb{E}_{q_0}[e^{rf}]$. Using elementary analysis and an argument of convexity, the equivalence between the two assumptions is easy to prove.

Taking $T = 1/\rho$ and $q_1 = q_0\exp(f/T) \in \mathcal{M}_{\mathcal{A}}$ leads easily to the result.

**Remark 2**
– *The two hypotheses, $\sum_a q_0(a)f(a) < 0$ and $\exists a, f(a) > 0$ are classical in the context of local score studies (see Karlin and Altschul, 1990) and are usually verified in practice using a classical data base composition of amino acids and usual scoring scales (see `http://web.expasy.org/protscales`).*
– *It can be shown using a similar proof as for Lemma 1 that $\sigma$ defined in Theorem 1 verifies $\sum_a q_0(a)f(a) < 0$ and $\exists a, f(a) > 0$.*

Theorem 1 leads to a corollary given in Subsection 5.2 that allows to compute the scoring of ambiguous letters. An application of Theorem 1 allows also to learn scoring functions from data. This last feature will be illustrated in a forthcoming work. An example of application of Theorem 2 is proposed on real proteins in Section 5.3.

| | $P(S_i = 1\|\mathbb{A})$ | $P(S_i = 2\|\mathbb{A})$ | $P(S_i = 3\|\mathbb{A})$ |
|---|---|---|---|
| $i = 1$ | 0.801 | 0.199 | $< 10^{-8}$ |
| $i = 2$ | 0.699 | 0.266 | 0.035 |
| $i = 3$ | 0455 | 0.489 | 0.056 |
| $i = 4$ | 0.325 | 0.550 | 0.125 |
| $i = 5$ | 0.266 | 0.463 | 0.271 |
| $i = 6$ | 0.133 | 0.532 | 0.335 |
| $i = 7$ | 0.072 | 0.455 | 0.473 |
| $i = 8$ | 0.056 | 0.213 | 0.731 |
| $i = 9$ | 0.035 | 0.128 | 0.838 |
| $i = 10$ | $< 10^{-8}$ | 0.114 | 0.886 |

**Table 1** Marginal posterior probabilities $\mathbb{P}(S_i = k|\mathbb{A})$ for the three states $k = 1, 2, 3$ at each index $i = 1, \ldots, n$.

## 5 Applications

We present here four applications to illustrate the results in the previous section. The computation of posterior probabilities on simulated sequences is presented in Section 5.1. First we propose a toy example for a sequence of several components. The question of a long sequence and possible overflows in the computation is also illustrated and practical computation methods are proposed. Section 5.2 presents how the HMM approach can allow one to compute scoring functions for ambiguous letters. We illustrate in Section 5.3 how the TM-tendency scoring function of Zhao and London can be rescaled in order to give a better mathematical interpretability to the result allowing the intuitive existence of a distribution of the component of the sequence in atypical state. Section 5.4 focuses on an illustration for the detection of sequences with atypical segment.

### 5.1 Posterior probabilities on simulated sequences

#### *5.1.1 A binary toy example*

Here, we consider the sequence $\mathbb{A} = 1011011001 \in \{0, 1\}^{10}$ to illustrate the computation. Appendix B presents computation details and implementation. The simulated sequence length is only $n = 10$ for simplicity. The scoring function has been chosen as follows: $f(0) = -2$ and $f(1) = 1$. Figure 2 (left Panel) represents the marginal probabilities of the sequence positions to be in State 2 ; computation of posterior probabilities for starting and ending indices is presented in the right Panel. Table 5.1.1 gives the values of the marginal posterior probabilities $\mathbb{P}(S_i = k|\mathbb{A})$ for the three states $k = 1, 2, 3$ at each index $i = 1, \ldots, n$. See Appendix A.2 for the link between posterior probabilities and the forward and backward quantities. As explained in Section A.2 we use here the following initialization for the backward quantities: $B_n(1) = 0$ and $B_n(2) = B_n(3) = 1$.

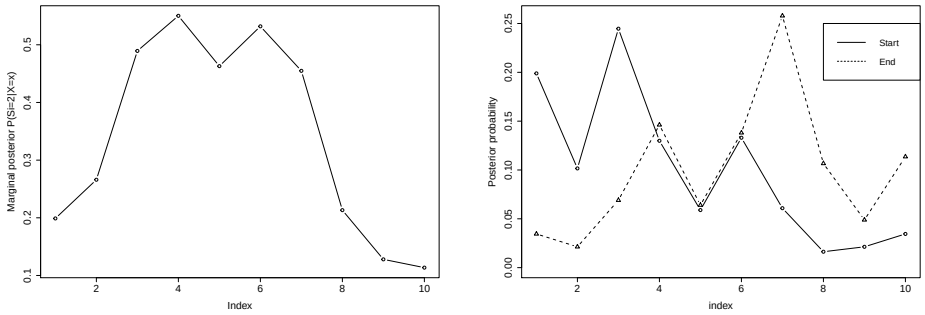


**Fig. 2** Toy example ($\mathcal{A} = \{0, 1\}$, $n = 10$, $\mathbb{A} = 1011011001$, with the following scoring function $f$: $f(0) = -2$ and $f(1) = 1$. Left Panel: Marginal posterior distribution of segmentation State 2. Right Panel: Segment start/end posterior distribution .

### *5.1.2 A practical implementation for long sequence*

When $n$ grows large, we might face under or overflow issues (small real numbers rounded to 0.0 in floating point arithmetic, see also Appendix B.2 for more details on the overflow issues). One way to overcome this problem is for example to perform all computations in log-scale. We propose here another way that consists of rescaling the forward and backward quantities so that each of them sums one (see details in Appendix B.2).

For illustration we simulated a sequence length is $n = 300$ taking its values in $\{1, 2, 3, 4\}$ with $q_0 = (0.4, 0.2, 0.2, 0.2)$ and $q_1 = (0.2, 0.4, 0.3, 0.1)$. A true segment is inserted in $I = [100, 150]$. The scoring function used corresponds to the one of Theorem 1, $\sigma = \log(q_0/q_1)$ (and thus $T = 1$), that leads to: $\sigma(1) = -0.693$, $\sigma(2) = 0.693$, $\sigma(3) = 0.405$, $\sigma(4) = -0.693$. Figure 3 represents the marginal posterior probabilities of the sequence component to be in State 2, that is in the inserted segments, $\mathbb{P}(S_j = 2|\mathbb{A}) = F_j(2)B_j(2)/Z$. In this figure, the inserted segment is clearly highlighted, and its position and length seem correct. Computation of marginal posterior probabilities for starting and ending indices of the inserted segment is presented in Figure 4 (left Panel). Finally, posterior length distribution of this segment is given in Figure 4 (right Panel).

Computation for a sequence length greater than $10^4$ is also very fast since the complexity is in $\mathcal{O}(n)$, and recovers correctly a small atypical segment of length as small as 100 (data not shown). Figure 5 illustrates the computation for $n = 40000$ and a true segment inserted in $I = [10000, 11000]$.

## 5.2 IUPAC ambiguous DNA code

Theorem 1 implies that the score of ambiguous letters can be defined in a canonical way.

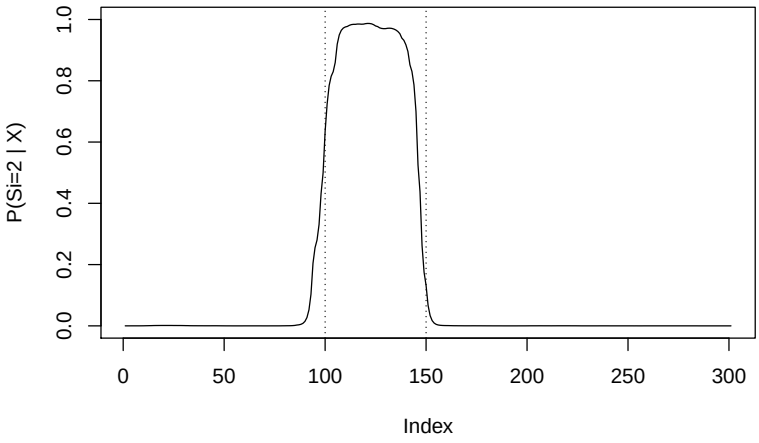


**Fig. 3** Marginal posterior distribution of segmentation State 2 for the simple simulation with $n = 300$ and $I = [100, 150]$, $q_0 = (0.4, 0.2, 0.2, 0.2)$, $q_1 = (0.2, 0.4, 0.3, 0.1)$ and the scoring function of Theorem 1 $\sigma = \log(q_0/q_1)$.

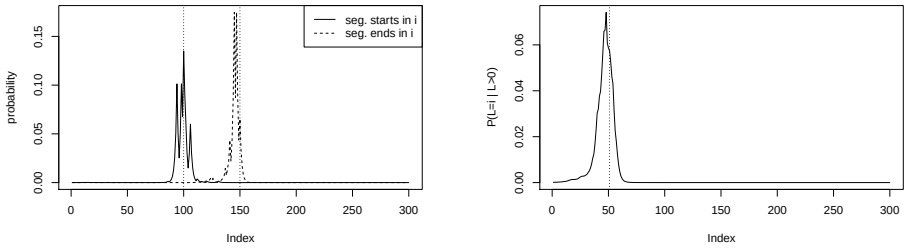


**Fig. 4** Segment start/end posterior distribution (left Panel) and length (right Panel) for the simple simulation with $n = 300$ and $I = [100, 150]$, $q_0 = (0.4, 0.2, 0.2, 0.2)$, $q_1 = (0.2, 0.4, 0.3, 0.1)$ and $f = \log(q_0/q_1)$.

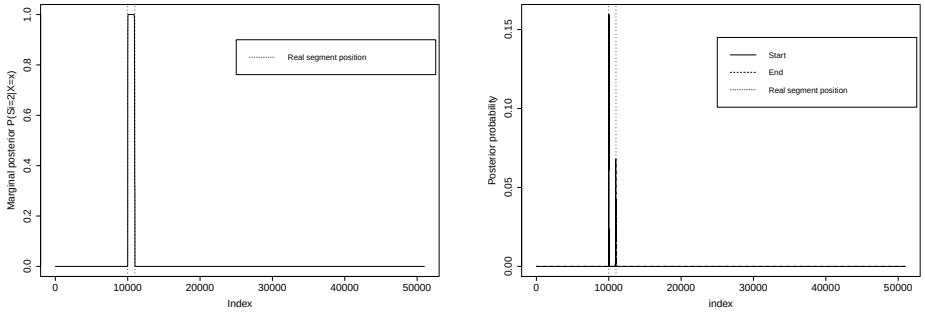


**Fig. 5** Long sequence example: Marginal posterior distribution of segmentation State 2 (left Panel) and Segment start/end posterior distribution (right Panel).

**Corollary 1** *The duality between the generating model and the scoring function implies that for* $(a \neq b)$

$$\sigma(a \text{ or } b) = \log\left(\frac{q_1(a)+q_1(b)}{q_0(a)+q_0(b)}\right) .$$

*Proof* Theorem 1 implies that $\sigma(a \text{ or } b) = \log\left(\frac{q_1(a \text{ or } b)}{q_0(a \text{ or } b)}\right) = \log\left(\frac{q_1(a)+q_1(b)}{q_0(a)+q_0(b)}\right)$.

One can note that the score deduced from the generating model which allows a mathematically correct interpretation is *neither* $\sigma(a)+\sigma(b)$ *nor* $q_0(a)\sigma(a)+q_0(b)\sigma(b)$ as one would be likely to expect.

Let us consider the IUPAC ambiguous DNA code. If we start from a scoring function $f$ it would be reasonable to expect the local score approach to be equivalent to some generating model, and hence to have $\mathbb{P}(S|\mathbb{A}) \propto \exp(H_f(\mathbb{A}|S))$. Unfortunately, in general $q_1(a) = q_0(a)\exp(f(a))$ does not define a probability distribution. However, thanks to Theorem 2, we know that there exists a unique rescaling $T$ so that

$$q_1(a) = q_0(a)\exp(f(a)/T)$$

defines a probability distribution. And Corollary 1 defines a score for ambiguous letters. Let us present a numerical illustration of these two results on the IUPAC ambiguous DNA code. Let us consider the four nucleotides $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ with the following scoring function $f(\mathtt{A}) = -2$, $f(\mathtt{C}) = -1$, $f(\mathtt{G}) = 0$, $f(\mathtt{T}) = +1$ and two different background distributions:

$$q_0^U = (0.25, 0.25, 0.25, 0.25) \text{ and } q_0^{\mathtt{GC}} = (0.1, 0.4, 0.4, 0.1) .$$

These two distributions verify the two hypothesis of a non-positive average score and possible non-negative scores.

For $q_0^U = (0.25, 0.25, 0.25, 0.25)$, we get $T = 1.13$, and

$$\sigma(\mathtt{A}) = -1.76,\ \sigma(\mathtt{C}) = -0.88,\ \sigma(\mathtt{G}) = 0,\ \sigma(\mathtt{T}) = +0.88 .$$

For $q_0^{GC} = (0.1, 0.4, 0.4, 0.1)$, we get $T = 0.61$, and

$$\sigma(\mathtt{A}) = -3.29,\ \sigma(\mathtt{C}) = -1.65,\ \sigma(\mathtt{G}) = 0,\ \sigma(\mathtt{T}) = +1.65 .$$

Figure 6 represents the plot of the initial scoring values for the four nucleotides versus the $\sigma$ values for both background distributions $q_0^U$ and $q_0^{\mathtt{GC}}$. Table 2 gives how the corresponding scores should be for the four nucleotides and the ambiguous components for each background distribution. For ambiguous letters, we have for example $\sigma(\mathtt{M}) = \sigma(\mathtt{A} \text{ or } \mathtt{C}) = -1.23$ that greatly differs from $\sigma(\mathtt{A}) + \sigma(\mathtt{C}) = -1.76 + (-0.88) = -2.64$.

Let us now consider the two following distributions $q_0$ and $q_1$ defined by: $q_0(x) = 0.213$ for $x = \mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}$, $q_0(x) = 0.021$ for $x = \mathtt{M},\mathtt{R},\mathtt{W},\mathtt{S},\mathtt{Y},\mathtt{K}$, $q_0(x) =$

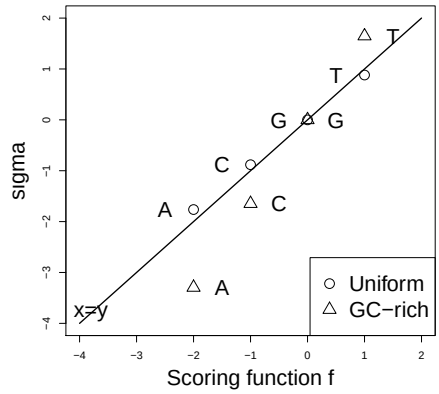


**Fig. 6** Initial scoring values for the four nucleotides versus the $\sigma$ values for both background distributions $q_0^U$ (Uniform) and $q_0^{\mathtt{GC}}$ (GC-rich).

0.004 for $x = \mathtt{V}, \mathtt{H}, \mathtt{D}, \mathtt{B}$ and $q_0(\mathtt{N}) = 0.006$; and $q_1(x) = 0.056$ for $x = \mathtt{A}, \mathtt{T}$, $q_1(x) = 0.301$ for $x = \mathtt{C}, \mathtt{G}$ and

$$q_1 = (0.009, 0.046, 0.022, 0.019, 0.034, 0.034, 0.007, 0.017, 0.033, 0.036, 0.029)$$

for $\mathtt{M}$,$\mathtt{R}$,$\mathtt{W}$,$\mathtt{S}$,$\mathtt{Y}$,$\mathtt{K}$,$\mathtt{V}$,$\mathtt{H}$,$\mathtt{D}$,$\mathtt{B}$,$\mathtt{N}$. We deduce from Theorem 1 the scoring function $\sigma(x) = \log(q_1(x)/q_0(x))$ for $x = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$. We calculate $\sigma$ for the other letters using Corollary 1. We also consider another scoring function $\tilde{\sigma}$, equal to $\sigma$ for $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ and we calculate its value for the ambiguous letters using the following way: for example $\tilde{\sigma}(a \text{ or } b) = \sigma(a) + \sigma(b)$. Figure 7 presents a comparison of the two scoring functions and highlights that the scores can be quite different for the ambiguous letters.

Moreover, we consider the usual approach of highlighting atypical segments using local score approach and we propose here an experiment to illustrate the impact of the scoring function choice for the ambiguous letters on the detection performance. We simulate 5000 sequences of length 100 in the IUPAC alphabet under the upper distribution $q_0$. We insert a segment of 20 letters using the distribution $q_1$. We calculate the local score for each sequence using both scoring functions $\sigma$ and $\tilde{\sigma}$ and we highlight the segment which realizes the local score in both cases. Allowing an error of 3 (*resp.* 5) in the positions of the beginning and the ending index position, the scoring function $\sigma$ detects 10% (*resp.* 22%) of the correct inserted segments and $\tilde{\sigma}$ only detects 1% (*resp.* 3%) of them. Using the scores computed by the application of Corollary 1 for the ambiguous letters instead of the usual way brings a much better performance. In this example, we can see that the sensitivity is not very good (22%) even with the use of the appropriate scoring values for the ambiguous letters. This can be explained by the fact that the two distributions may be not discriminative enough to easily highlight the inserted segments.

**Table 2**

| IUPAC Code | A | C | G | T/U | | |
|---|---|---|---|---|---|---|
| Meaning | A | C | G | T | | |
| $f$ | -2 | -1 | 0 | +1 | | |
| $\sigma$ for $q_0^U$ | $-1.76$ | $-0.88$ | 0.00 | 0.88 | | |
| $\sigma$ for $q_0^{GC}$ | $-3.29$ | $-1.65$ | 0.00 | 1.65 | | |
| IUPAC Code | M | R | W | S | Y | K |
| Meaning | A or C | A or G | A or T | C or G | C or T | G or T |
| $\sigma$ for $q_0^U$ | -1.23 | -0.54 | 0.26 | -0.35 | 0.34 | 0.54 |
| $\sigma$ for $q_0^{GC}$ | -1.82 | -0.21 | 0.96 | -0.52 | 0.18 | 0.61 |
| IUPAC Code | V | H | D | B | N | |
| Meaning | no T | no G | no C | no A | A,C,G or T | |
| $\sigma$ for $q_0^U$ | -0.64 | 0.00 | 0.18 | 0.24 | 0.00 | |
| $\sigma$ for $q_0^{GC}$ | -0.63 | 0.00 | 0.43 | 0.1 | 0.00 | |

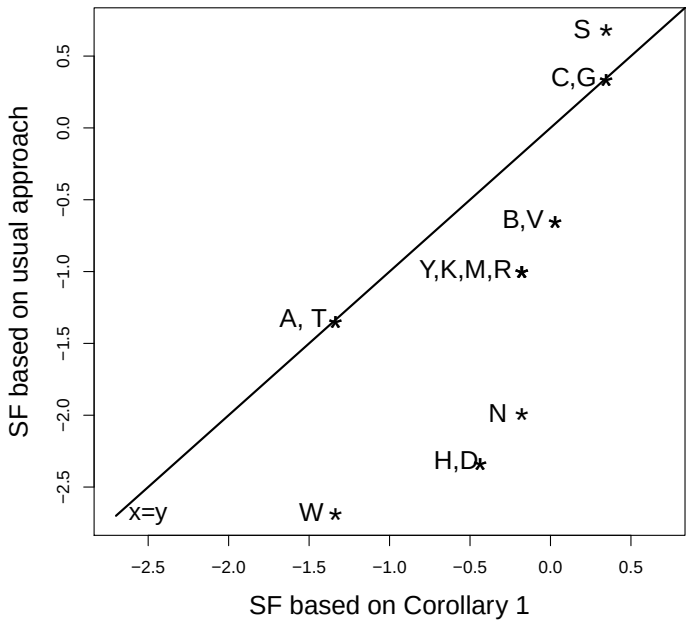


**Fig. 7** Scoring function (SF) $\sigma$ using Theorem 1 for $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ and Corollary 1 for the ambiguous letters. And $\tilde{\sigma}$ using Theorem 1 for $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ and the following usual approach $\tilde{\sigma}(a \text{ or } b) = \sigma(a) + \sigma(b)$ for the other components.

**Remark 3** *Uncertainty (*e.g. *in protein crystallisation, Next Generation Sequencing – NGS –), can also be taken into account. If uncertainty is expressed with weights $w_a, w_b > 0$ for observations $a$ and $b$, we can use the score:*

$$\log\left(\frac{w_a q_1(a) + w_b q_1(b)}{w_a q_0(a) + w_b q_0(b)}\right)$$

**Table 3** Zhao and London TM-tendency scale ($f_{\mathrm{ZL}}$) (see Zhao and London, 2006). Distribution $q_0$ (in %) for 56 TM proteins without their TM regions and Zhao and London score rescaled ($\sigma_{\mathrm{ZL}}$) using $q_0$ distribution.

| | A | C | D | E | F | G | H | I | K | L |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_{\mathrm{ZL}}$ | 0.38 | -0.30 | -3.27 | -2.90 | 1.98 | -0.19 | -1.44 | 1.97 | -3.46 | 1.82 |
| $q_0$ | 5.489 | 2.539 | 5.472 | 5.604 | 3.380 | 6.103 | 2.624 | 5.384 | 4.923 | 10.049 |
| $\sigma_{\mathrm{ZL}}$ | 0.128 | -0.101 | -1.102 | -0.977 | 0.667 | -0.064 | -0.485 | 0.664 | -1.166 | 0.613 |
| | M | N | P | Q | R | S | T | V | W | Y |
| $f_{\mathrm{ZL}}$ | 1.40 | -1.62 | -1.44 | -1.84 | -2.57 | -0.53 | -0.32 | 1.46 | 1.53 | 0.49 |
| $q_0$ | 1.970 | 5.367 | 5.814 | 3.855 | 5.716 | 8.903 | 6.055 | 5.726 | 1.159 | 3.862 |
| $\sigma_{\mathrm{ZL}}$ | 0.472 | -0.546 | -0.485 | -0.620 | -0.866 | -0.179 | -0.108 | 0.492 | 0.516 | 0.165 |

*to ensure the two probabilized spaces to be equivalent and the existence of a distribution from which the components of the atypical segment are derived. These scores are clearly different from $w_a\sigma(a)+w_b\sigma(b)$, which would typically be used in the scoring system.*

*We can also define a canonical score for any weighted profile. Let $w = (w_a)_{a\in\mathcal{A}}$ with $w_a > 0$ be a profile, coming for example from a multiple alignment. Then*

$$\sigma(w) = \log\left(\frac{\sum_a w_a q_1(a)}{\sum_a w_a q_0(a)}\right) \neq \sum_a w_a\sigma(a) = \sum_a w_a \log\left(\frac{q_1(a)}{q_0(a)}\right) \ .$$

### 5.3 Rescaling Zhao and London scale

This subsection illustrates how a given scoring function can induce the emission probabilities of the generative model. Let us consider the TM-tendency scale of Zhao and London (see Zhao and London, 2006), denoted by $f_{\mathrm{ZL}}$ and given in Table 3. This score function is presented as a refined "hydrophobicity"-type TM sequence prediction scale, that should approach the theoretical limit of accuracy, and seems more accurate than the well-known Kyte and Doolittle hydrophobic scale (see Kyte and Doolittle, 1982). Let us consider the 56 human soluble and helical transmembrane (TM) proteins extracted from the protein data base UniProtKB/Swiss-Prot (`http://www.uniprot.org/uniprot` with request "name:"transmembrane protein" soluble helical AND reviewed:yes"). Leaving aside the TM regions for each protein, and keeping only the non-TM regions of the sequences, we derive the distribution $q_0$ (see Table 3). We verify that $\mathbb{E}[f_{\mathrm{ZL}}] \simeq -0.50 < 0$.

Using Theorem 2 we compute the parameter $T \simeq 2.97$ from which we deduce $\sigma_{\mathrm{ZL}} = f_{\mathrm{ZL}}/T$ defined as the *rescaled scoring function* and given in Table 3. Note that the rescaling does not affect the sign of the scores nor their relative values of the initial scale thus an amino acid considered as hydrophobic or hydrophilic is still considered as such.

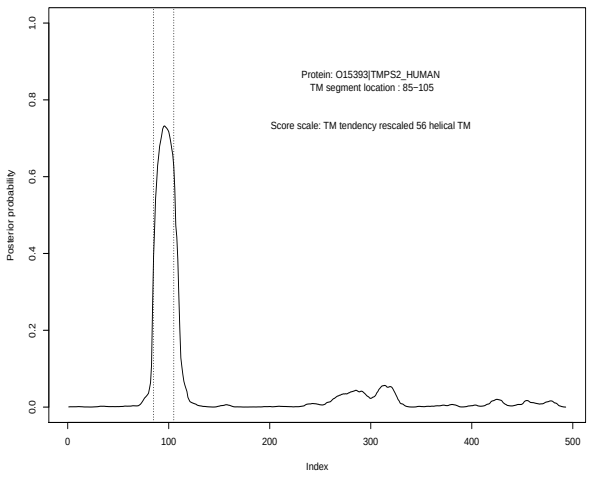


**Fig. 8** Marginal posterior probability that $S_i$ in Sate 2 using the rescaled Zhao and London score function.

Let us now consider `sp|O15393|TMPS2_HUMAN`, a supplementary human TM protein with a transmembrane segment at positions $85 - 105$. Let us apply the generating model using the rescaled scoring function of Zhao and London. Figure 8 gives the marginal posterior probabilities of being in State 2. The upper probabilities are around $\simeq 0.75$ and correspond to the TM segment. The start of the highlighted TM region is accurate and its end is very close to the real one. Figure 9 gives segment start/end posterior probabilities (left Panel) and posterior probabilities for length (right Panel). The length is relatively accurate even if the probabilities are not very high : maximal posterior probability for the segment starting index is found at index 84 (instead of 85), maximal posterior probability for the segment ending index is found at index 106 (instead of 105), and maximal posterior probability for segment length is found to be 23 (instead of 21). Those probabilities are computed as explained in Section A.2 with the following initialization for the backward quantities $B_n(1) = 0$ and $B_n(2) = B_n(3) = 1$.

*Deriving $q_1(.)$ distribution from Zhao and London scoring function*: Some scoring functions exist which have been established by biologists with experimental design such as the one proposed in Zhao and London (2006). Intuitively, pointing out exceptional segment with scoring functions means that such a segment has been a realization of a generating model different than the one of background of the sequence. Given a scoring function $f$ constructed by empirical biologist's experiments or which is well considered by biologists, with a background distribution $q_0$, allows one by Theorem 2 to compute $T$ and $q_1$. The distribution $q_1$ correspond to the one under which atypical segments are generated, and for which the two approaches, local score and HMM can be

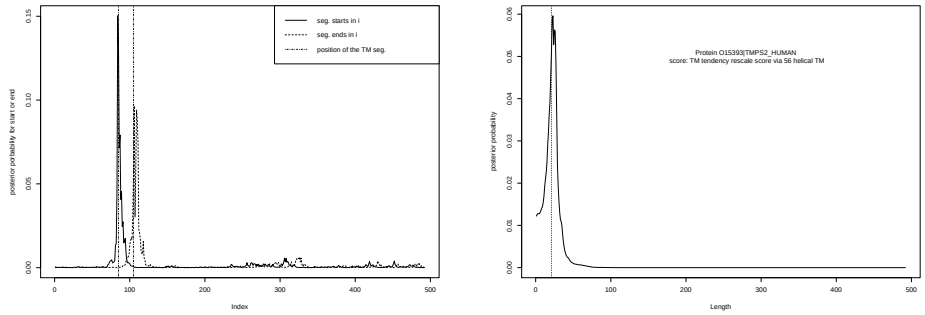


**Fig. 9** Segment start/end posterior probabilities (left Panel) and posterior probability there exists a segment of a given length (right Panel) using the Zhao and London rescaled scores.

**Table 4** Distribution, given in percentage, of TM segment deduced from the Zhao and London TM-tendency scale (ZL) Zhao and London (2006).

| | A | C | D | E | F | G | H | I | K | L |
|---|---|---|---|---|---|---|---|---|---|---|
| $q_1$ | 6.239 | 2.295 | 1.818 | 2.109 | 6.587 | 5.724 | 1.615 | 10.457 | 1.534 | 18.556 |
| | M | N | P | Q | R | S | T | V | W | Y |
| $q_1$ | 3.158 | 3.109 | 3.579 | 2.074 | 2.404 | 7.447 | 5.436 | 9.365 | 1.941 | 4.555 |

combined. Using $\sigma_{\mathrm{ZL}} = f_{\mathrm{ZL}}/T$ to highlight atypical segments instead of the previous one $f_{\mathrm{ZL}}$ assure the intuitive existence of a component distribution in atypical regions. The distribution of TM segments related to the scoring function of Zhao and London is given in Table 4.

Moreover, the knowledge of scoring functions, deduced from biological experiments and well considered by biologists, and the application of Theorem 2 allow to propose a distribution of the atypical regions which could be difficult to catch with insufficient data.

## 5.4 Detection of sequence with atypical segment

In this section, we use the following initialization for the backward quantities $B_n(1) = 1$ and $B_n(2) = B_n(3) = 1/\binom{n+1}{2}$ (see Section A.2). We simulate under the distribution $q_0 = (0.4, 0.2, 0.2, 0.2)$, $10^3$ sequences of different length $n$ taking values in $\{1, 2, 3, 4\}$. We also simulate $10^3$ sequences of length $n$ under $q_0$ distribution with an exchange segment of length $\ell$ simulated using the distribution $q_1 = (0.2, 0.4, 0.3, 0.1)$. We consider the following three statistics: i) the $p$-value of the local score $\mathbb{P}(H_n \geqslant h)$ with $h$ the observed local score of the sequence; ii) $\mathbb{P}(S_n \neq 1|\mathbb{A})$; iii) $M(S_2) = \max_{i=1,\dots,n+1} \mathbb{P}(S_i = 2)$. We compute the local score value as defined in 1, using the scoring function $\sigma$ of Theorem

**Table 5** AUROC of the three proposed statistics.

| $n$ | 100 | 300 | 1000 | 1000 | 1000 | 5000 | $10^4$ |
|---|---|---|---|---|---|---|---|
| $\ell$ | 10 | 50 | 10 | 20 | 50 | 50 | 50 |
| $\mathbb{P}(H_n \geqslant .)$ | 0.760 | 0.984 | 0.588 | 0.769 | 0.963 | 0.909 | 0.924 |
| $\mathbb{P}(S_n \neq 1)$ | 0.776 | 0.990 | 0.597 | 0.773 | 0.973 | 0.923 | 0.937 |
| $M(S_2)$ | 0.756 | 0.984 | 0.569 | 0.750 | 0.965 | 0.911 | 0.929 |

1, and we use the exact method for independent and identically distributed model (see Mercier and Daudin, 2001) to compute the corresponding $p$-values.

We can see on Figure 10 the distribution of the three statistics under $H_0$ (no atypical segment) and $H_1$ (one atypical segment of length $\ell = 50$ – top Panel – and $\ell = 100$ – bottom Panel). For the $p$-value, we observe that $p$-values under $H_0$ are distributed uniformly on $[0, 1]$ while they are close to 0 under $H_1$. For the maximum marginal probability of State 2, the distribution under $H_0$ has a peak around 0.25 while the distribution under $H_1$ is concentrated in the proximity of 1.0. Finally, our HMM-based Bayes factor (the probability of having one atypical segment) is logically close to 0.0 under $H_0$ and close to 1.0 under $H_1$. All three statistics clearly have a good discriminative power (see Table 5 for the corresponding AUROC) with a slight numerical advantage to the Bayes factor, but the latter also has the interest in providing an easier interpretation and canonical choice of thresholding (*e.g.* using a natural threshold of 0.5 for example).

We also compute the AUROC (see Xavier et al., 2011) for the three statistics. The AUROC is a measure of the discriminative power of a statistic for a binary response variable. An AUROC of 0.5 is pure noise, AUROC $> 0.7$ considered acceptable, AUROC $> 0.9$ considered excellent. The values are gathered in Table 5. We can see in this table that the three statistics achieve a similar performance with a little adventage for $\mathbb{P}(S_n \neq 1|\mathbb{A})$.

## 6 Conclusion

By considering the set of all possible segments and a Gibbs distribution on this set, we also take into account suboptimal segments and not only on the optimal segment which realizes the maximal score. This probabilization of the segmentation space allows us to establish a duality between our constrained HMM approach and the maximum scoring subsegment with a rescaled scoring function. We prove that the scoring function to be used is unique, given a studied context, in order to bring a mathematically correct interpretation and a background distribution from which the highlighted segments derived. This duality provides many potential opportunities.

The HMM approach allows a rapid (linear) implementation to extract the usual information that is expected: the probability that the sequence contains

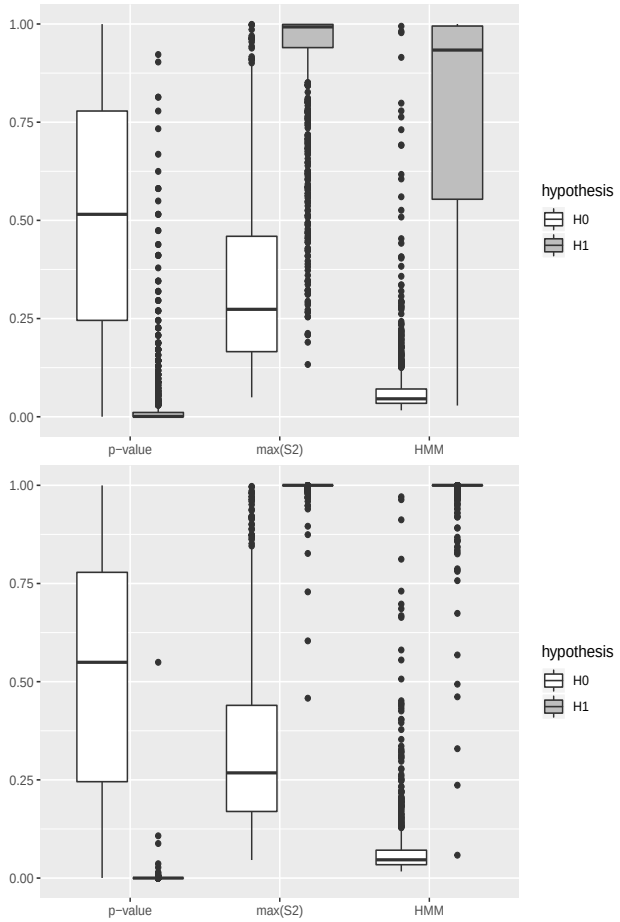


**Fig. 10** Box-plots of the three statistics: the $p$-value of the local score ($\mathbb{P}(H_n \geqslant .)$), the HMM approach ($\mathbb{P}(S_n \neq 1|\mathbb{A})$) and the $\max(S_2)$ statistic ($\max_{i=1,\dots,n+1} \mathbb{P}(S_i = 2)$) under $H_0$ (no atypical segment) and $H_1$ (one atypical segment of length). Top Panel: $\ell = 50$ and $n = 1000$; Bottom Panel: $\ell = 100$ and $n = 1000$).

an atypical segment; the probabilities for the position of the atypical segment and other information on the length of the segment. The computation is easy and fast even for very long sequences with lengths greater than $10^5$ or more.

Biologists can adapt the classical scoring functions by rescaling them. That way, the natural approach to test the presence of an atypical segment in a sequence which deals with a natural alternative hypothesis defined as $H_1$: "There exists $1 \leq i_0 \leq j_0 \leq n$ such as $A_{i_0}, \dots, A_{j_0}$ are distributed with a distribution $q_1 \neq q_0$." (with $H_0$: "$A_1, \dots, A_n$ are $q_0$ distributed") is verified.

Moreover, this can be applied to sequences with an uncertain position (which is often the case in the NGS era) for which a canonical score can be derived from the original scoring function.

For example, in Ruzzo and Tompa (1999) the authors propose an efficient algorithm which allows in a linear time to find all the suboptimal segments in a sequence. Using HMM for such a purpose seems quite a challenging problem, whereas an adaptation to this Ruzzo and Tompa's algorithm in our approach can be considered and then transferred to the segmentation space.

The model we propose here assumes the presence of exactly one segment of interest inserted in each sequence. But this must not be considered as a limitation. Extending our constrained approach to $K$ subsegments is quite straightforward. This suggests that the local score might as well be extended to $K$ subsegments (for $K = 2$, a local score on 2 segments could be defined by $\max_{0\leqslant i\leqslant j<k\leqslant \ell\leqslant n}\left(\sum_{m=i}^{j} f(A_m) + \sum_{r=i}^{j} f(A_r)\right)$). By exploiting the duality of the two notions it is therefore easily to provide an efficient algorithm to generalize such a local score.

In Lefebvre et al. (2020) an approach is proposed, based on the work developed in this present article, to learn by an unsupervised method a scoring function with confidence interval. Another perspective of this work consists in learning scoring functions: from a given set of sequences known for containing atypical segments, the corresponding scoring function deduced from the background and the atypical state distributions can be rapidly deduced using a maximum likelihood computation (EM or direct optimization) both allowing one to estimate the new scoring function but also establish confidence intervals, hypothesis testing etc... If using the HMM method to learn a scoring function is not new, a very interesting point of our method is that it can allow inference and confidence intervals to be established. But these extensions are left for further work.

## Appendices

### A Notations and results

#### *A.1 Potentials*

Let us define for $k \in \{1, 2, 3\}$, $\phi_1(S_1 = k|\theta) \stackrel{\text{def}}{=} \pi(1, k)e(a_1)^{\mathbf{1}_{k=2}}$, equally denoted $\phi_1(k|\theta)$ or $\phi_1(k)$ to lighten the writing, and

$$\phi_i(S_{i-1} = j, S_i = k|\theta) \stackrel{\text{def}}{=} \pi(j, k)e(a_i)^{\mathbf{1}_{k=2}} \text{ for } i = 2, \ldots, n$$

equally denoted $\phi_i(j, k|\theta)$ or $\phi_i(j, k)$ to lighten the writing, with

$$e(a) = \frac{q_1(a)}{q_0(a)} \quad \text{or} \quad e(a) = \exp\left(\frac{f(a)}{T}\right) \quad \text{or} \quad e(a) = \exp(\sigma(a)) \ ,$$

$$\pi = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

and with $\theta = (q_0, q_1)$ or $\theta = (f, T)$ depending on the chosen approach. With this potential, we define:

$$\mathbb{Q}(\mathbb{A} = a|\theta) \stackrel{\text{def}}{=} \sum_{S} \prod_{i=1}^{n} \phi_i(C_i|\theta) \text{ and } \mathbb{Q}(\mathbb{A} = a, S_i = k|\theta) \stackrel{\text{def}}{=} \sum_{S, S_i = k} \prod_{i=1}^{n} \phi_i(C_i|\theta)$$

where $S = (S_i)_{1\leq i\leq n} \in \{1,2,3\}^n$, $C_1 = \{S_1\}$ and $C_i = \{S_{i-1}, S_i\}$ for $i = 2,\dots,n$. In the HMM case we have:

$$\mathbb{Q}(\mathbb{A} = a|\theta) = \frac{\sum_S \mathbb{P}(\mathbb{A} = a, S|\theta)}{\prod_{i=1}^n q_0(a_i)} = \frac{\sum_S \mathbb{P}(\mathbb{A} = a|S;\theta) \times \mathbb{P}(S)}{\prod_{i=1}^n q_0(a_i)}$$

$$\text{with} \quad \mathbb{P}(S) = \mathbf{1}_{\{S_1=1 \text{ or } 2\}} \prod_{i=2}^n \pi(S_{i-1}, S_i)$$

$$\mathbb{Q}(\mathbb{A} = a|\theta) = \frac{\sum_{S\in\mathcal{S}} \mathbb{P}(\mathbb{A} = a|S;\theta)}{\prod_{i=1}^n q_0(a_i)} \ .$$

In the score case we have:

$$\mathbb{Q}(\mathbb{A} = a|\theta) = \sum_{S\in\mathcal{S}} \exp\left(H(S|\mathbb{A} = a)\right) \ .$$

The given matrix $\Pi$ is in fact not strictly speaking a probability transition matrix. We voluntary omit to talk about an adding terminal "end" state in the hidden state space to simplify the presentation. The matrix $\Pi$ should have been

$$\pi = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and $\mathbb{P}(S_1 = 1) = 0.5$, $\mathbb{P}(S_1 = 2) = 0.5$, $\mathbb{P}(S_1 = 3) = P(S_1 = 4) = 0$. This allows to have equiprobability of the segmentation : for example $\mathbb{P}(S_1 \dots S_5 = 12223) = 0.5^5$ and $\mathbb{P}(S_1 \dots S_5 = 11122) = 0.5^5$ whereas with $\Pi$ with only 3 states and $\Pi_3 = (0,0,1)$ brings $\mathbb{P}(S_1 \dots S_5 = 12223) = 0.5^4$ and $\mathbb{P}(S_1 \dots S_5 = 11122) = 0.5^5$ which are not equal. Omitting the ending absorbing state simplifies the implementation and works as the same. In fact equiprobability of the segmentation is not necessary to detect atypical segments but it allows the connection to classical scoring approaches.

## *A.2 Forward and backward*

Classical uses of forward and backward quantities correspond to HMM probabilities (see Durbin et al., 1998). Here, we adapt the forward and backward definition using Bayesian network potentials (see Koller and Friedman, 2009) as it allows us to take into account constraints given in the transition $\pi$. Our forward and backward definition also verify suitable recursions, slightly different from the ones classically used in HMM. The probabilities of interest are recovered based on ratios.

**Forward**

We define $F_1 \overset{\text{def}}{=} \phi_1$ and for $i = 2,\dots,n$: $F_i(S_i = k|\theta) \overset{\text{def}}{=} \sum_{S_{1:i-1}, S_i=k} \prod_{u=1}^i \phi_u(C_u|\theta)$, equally denoted $F_i(k,\theta)$ or $F_i(k)$. We can easily prove by induction that:

$$F_i(k) = \sum_j F_{i-1}(j)\phi_i(j,k)$$

with our specific structure of $\pi$ this gives:

$$F_i(1) = F_{i-1}(1), \quad F_i(2) = (F_{i-1}(1) + F_{i-1}(2)) \times e(x_i), \quad F_i(3) = F_{i-1}(2) + F_{i-1}(3) \ .$$

The interesting thing is that:

$$F_n(S_n = k) = \mathbb{Q}(S_n = k, \mathbb{A} = a|\theta) = \sum_{S, S_n=k} \prod_{i=1}^n \phi_i(C_i|\theta) \quad \text{hence} \quad \sum_k F_n(k) = \mathbb{Q}(\mathbb{A} = a|\theta)$$

**Backward**

We can also define in a similar way the backward quantities: $B_n \overset{\text{def}}{=} 1$ and for $i = n, \dots, 2$:

$$B_{i-1}(S_{i-1} = j|\theta) \overset{\text{def}}{=} \sum_{S_{i-1}=j, S_{i:n}} \prod_{u=i}^{n} \phi_u(C_u|\theta)$$

equally denoted $B_i(k)$. Again, by induction we easily get: $B_{i-1}(j) = \sum_k \phi_i(j,k) B_i(k)$. With our specific structure of $\pi$ this gives:

$$B_{i-1}(1) = B_i(1) + B_i(2)e(a_i), \quad B_{i-1}(2) = B_i(2)e(a_i) + B_i(3), \quad B_{i-1}(3) = B_i(3) \ .$$

We clearly have $\phi_1(S_1)B_1(k) = \mathbb{Q}(S_1 = k, \mathbb{A} = a|\theta)$. But more interestingly we have:

$$F_i(k)B_i(k) = \mathbb{Q}(S_i = k, \mathbb{A} = a|\theta) \quad \text{and}$$

$$F_{i-1}(j)\phi_i(j,k)B_i(k) = \mathbb{Q}(S_{i-1} = j, S_i = k, \mathbb{A} = a|\theta) \ .$$

Probabilities of interest can be recovered by the fact that

$$\mathbb{P}(S_i = k|\mathbb{A} = a, \theta) = \frac{\mathbb{Q}(S_i = k, \mathbb{A} = a|\theta)}{\mathbb{Q}(\mathbb{A} = a|\theta)} = \frac{F_i(k)B_i(k)}{\sum_\ell F_i(\ell)B_i(\ell)} \ .$$

These results correspond to the assumption that all segmentation are equiprobable. As explained in Section 3, such hypothesis will favor the model with one atypical segment over the one where there is none. Fortunately, this issue can easily be corrected for any prior $\tau \in [0,1]$ of having an atypical segment by setting $B_n(1) = (1-\tau)$ and $B_n(2) = B_n(3) = \tau / \binom{n+1}{2}$ at the beginning of the backward recursion. The posterior distribution of having an atypical segment is then $(F_n(2)B_n(2) + F_n(3)B_n(3))/(F_n(1)B_n(1) + F_n(2)B_n(2) + F_n(3)B_n(3))$. When working conditionally to the existence of an atypical segment which posterior location or length must be computed, one simply uses $B_n(1) = 0$ and $B_n(2) = B_n(3) = 1$ to ensure that the configuration with no atypical segment will not be considered.

## *A.3 Marginal posterior probabilities*

We can also exploit the forward and backward quantities to compute the posterior probabilities given in Section 3.We can easily obtain marginal distribution from the forward and backward quantities but we have to decide conditioning to what. Indeed, if the backward recursions are computed with $B_n(1) = B_n(2) = B_n(3) = 1$, we condition only on $\{\mathbb{A} = a\}$. If we use $B_n(1) = 0$ and $B_n(2) = B_n(3) = 1$ we condition on $\{\mathbb{A} = a, S_n \in \{2,3\}\}$ which means that $S = 1, \dots, 1$ is not possible anymore. It is typically the condition we need when computing marginal start/end segment distributions. Setting up $B_n(1) = 0$ and $B_n(2) = B_n(3) = 1$ (ie. "No atypical segment" impossible) we get with $\mathbb{P}(\text{ev}) = \sum_k F_n(k) \times B_n(k)$:

$$\mathbb{P}(\text{segment starts at index 1}) = \mathbb{P}(S_1 = 2) = \frac{F_1(2) \times B_1(2)}{\mathbb{P}(\text{ev})}$$

$$\begin{aligned}\mathbb{P}(\text{segment starts at index i}) &= \mathbb{P}(S_{i-1} = 1, S_i = 2) \\ &= \frac{\sum_{S_{i-1}=1} F_{i-1}(S_{i-1} = 1) \times \phi_i(S_{i-1} = 1, S_i = 2) \times B_i(S_i = 2)}{\mathbb{P}(\text{ev})} \\ &= \frac{F_{i-1}(1) \times \phi_i(1,2) \times B_i(2)}{\mathbb{P}(\text{ev})}\end{aligned}$$

$$\mathbb{P}(\text{segment stops at index i}) = \mathbb{P}(S_i = 2, S_{i+1} = 3) = \frac{F_i(2) \times \overbrace{\phi_i(2,3)}^{1} \times B_{i+1}(3)}{\mathbb{P}(\text{ev})}$$

$$\mathbb{P}(\text{segment stops at index n}) = \mathbb{P}(Sn = 2) = \frac{F_n(2) \times \overbrace{B_n(2)}^{1}}{\mathbb{P}(\text{ev})}$$

## B Implementation: a simple binary toy example

Let us consider a binary sequence and a scoring function $f(0) = -2$ and $f(1) = +1$. The following code lines in Appendix B.1 first rescale the scoring function (see Appendix C for the rescaling function used). Then we compute the forward and backward quantities from which marginal posterior probabilities are deduced: $(\mathbb{P}(S_i = k))_{1 \leq i \leq n}$ for $k = 1, 2, 3$ ; and also posterior probabilities for a segment to start or to end at a given index. A practical implementation is given for large sequence length in appendix B.2. We also propose an implementation in R language to compute posterior probabilities for the length of the atypical segment in Appendix B.3.

### *B.1 Marginal posterior probabilities*

```
f=c(-2,1)
x=c(1,0,1,1,0,1,1,0,0,1)
temp=scaling(f)$$temp
ev=exp(f[x+1]/temp)
# ev=q1[x+1]/q0[x+1]
# Forward
n=length(x)
Fw=matrix(nrow=n,ncol=3)
Fw[1,]=c(1.0,ev[1],0.0)
for (i in 2:n) {
  Fw[i,1]=Fw[i-1,1]
  Fw[i,2]=(Fw[i-1,1]+Fw[i-1,2])*ev[i]
  Fw[i,3]=Fw[i-1,2]+Fw[i-1,3]}
# Backward
Bk=matrix(nrow=n,ncol=3)
# Initialization for uniform prior on the segmentation space
# Bk[n,]=c(1.0,1.0,1.0)
# on the segmentation space with 0 vs 1 atypical segment
# Bk[n,]=c(1,1/choose(n+1,2),1/choose(n+1,2))
# on segmentation with one atypical segment
Bk[n,]=c(0.0,1.0,1.0)
for (i in n:2) {
  Bk[i-1,1]=Bk[i,1]+ev[i]*Bk[i,2]
  Bk[i-1,2]=ev[i]*Bk[i,2]+Bk[i,3]
  Bk[i-1,3]=Bk[i,3]}
post_marginal=Fw*Bk/apply(Fw*Bk,1,sum)
pev=sum(Fw[n,]*Bk[n,])
post.begin=rep(NA,n)
post.start[1]=ev[1]*Bk[1,2]/pev
for (i in 2:n) post.start[i]=Fw[i-1,1]*ev[i]*Bk[i,2]/pev
post_end=rep(NA,n)
for (i in 1:(n-1)) post_end[i]=Fw[i,2]*Bk[i+1,3]/pev
post_end[n]=Fw[n,2]/pev
```

### *B.2 Practical implementation: forward and backward rescaling*

When $n$ grows large, we face underflow issues. Indeed, as we can see in the definitions of the forward and backward quantities in Section A.2, the quantities stand on sums and products. The number of sums and products to be computed depends on the length of the sequence. In floating point arithmetic, a real number greater than $A$ or less than $1/A$ (the value of $A > 0$ depends on the system, typically $A = 10^{320}$ with double precision in C language) are set to Inf (overflow) or 0 (underflow). Since our forward and backward quantities can

easily get out of this range (for example $F_n(2) = 10^{-5000}$) we need a rescaling mechanism thereafter provided.

One way to overcome this problem is to rescale the forward and backward quantities so that each of them sums one.

$$\tilde{F}_j(k) = F_j(k)/\exp(L_j) \text{ with } L_j = \log(\sum_k F_j(k)) \quad \sum_k \tilde{F}_j(k) = 1$$

$$\tilde{B}_j(k) = B_j(k)/\exp(M_j) \text{ with } M_j = \log(\sum_k B_j(k)) \quad \sum_k \tilde{B}_j(k) = 1$$

```
f=c(-2,1)
N=100
eta=0.2
set.seed(42)
x=c(
  sample(0:1,replace=TRUE,size=N,prob=c(1-eta,eta)),
  sample(0:1,replace=TRUE,size=2*N,prob=c(eta,1-eta)),
  sample(0:1,replace=TRUE,size=N,prob=c(1-eta,eta)))
temp=scaling(f)$temp
ev=exp(f[x+1]/temp)
# ev=q1[x+1]/q0[x+1] for an HMM approach
# Forward
n=length(x)
L=rep(NA,n)
Fw=matrix(nrow=n,ncol=3)
Fw[1,]=c(1.0,ev[1],0.0)
# Normalization
tmp=sum(Fw[1,]); Fw[1,]=Fw[1,]/tmp; L[1]=log(tmp)
for (i in 2:n) {
  Fw[i,1]=Fw[i-1,1]
  Fw[i,2]=(Fw[i-1,1]+Fw[i-1,2])*ev[i]
  Fw[i,3]=Fw[i-1,2]+Fw[i-1,3]
  # Normalization
  tmp=sum(Fw[i,]); Fw[i,]=Fw[i,]/tmp; L[i]=L[i-1]+log(tmp)
}
# Backward
M=rep(NA,n)
Bk=matrix(nrow=n,ncol=3) # on a B j(k)=B[j,k] # init
Bk[n,]=c(0.0,1.0,1.0)
# Normalization
tmp=sum(Bk[n,]); Bk[n,]=Bk[n,]/tmp; M[n]=log(tmp)
for (i in n:2) {
  Bk[i-1,1]=Bk[i,1]+ev[i]*Bk[i,2]
  Bk[i-1,2]=ev[i]*Bk[i,2]+Bk[i,3]
  Bk[i-1,3]=Bk[i,3]
  # Normalization
  tmp=sum(Bk[i-1,]); Bk[i-1,]=Bk[i-1,]/tmp; M[i-1]=M[i]+log(tmp) }
# verif
# log(apply(Fw*Bk,1,sum))+L+M
post_marginal=Fw*Bk/apply(Fw*Bk,1,sum)
lpev=L[n]+M[n]+log(sum(Fw[n,]*Bk[n,]))
post.start=rep(NA,n)
post.start[1]=ev[1]*Bk[1,2]*exp(M[1]-lpev)
for (j in 2:n) post.start[j]=Fw[j-1,1]*ev[j]*Bk[j,2]*exp(L[j-1]+M[j]-lpev)
post_end=rep(NA,n)
for (j in 1:(n-1)) post_end[j]=Fw[j,2]*Bk[j+1,3]*exp(L[j]+M[j+1]-lpev)
post_end[n]=Fw[n,2]*exp(L[n]-lpev)
```

### *B.3 Length of the atypical segment*

The idea is to replace $e(x)$ by $e(x) \times z$ where $z$ is a univariate dummy variable. We hence have:

$$\sum_S \prod_{i=1}^{n} \phi_i(C_i|\theta) = \sum_{k \geqslant 0} (\mathbb{A} = a, W = k|\theta) z^k$$

where $W$ is the length of the atypical segment ($W = 0$ when there is no segment). Function "`Mono_polyS4`" can be given upon request

```
# Back to the simplest example
f=c(-2,1)
x=c(1,0,1,1,0,1,1,0,0,1)
temp=scaling(f)$\$$temp
ev=exp(f[x+1]/temp)
# ev=q1[x+1]/q0[x+1] for an HMM approach
max\deg=10
# Forward
z=mono(max\deg)
n=length(x)
Fw=array(lapply(rep(0,3*n),const,degree=max\deg),dim=c(n,3))
Fw[1,1][[1]]=const(1.0,max\deg)
Fw[1,2][[1]]=ev[1]*z
Fw[1,3][[1]]=const(0.0,max\deg)
for (i in 2:n) {
  Fw[i,1][[1]]=Fw[i-1,1][[1]]
  Fw[i,2][[1]]=(Fw[i-1,1][[1]]+Fw[i-1,2][[1]])*(ev[i]*z)
  Fw[i,3][[1]]=Fw[i-1,2][[1]]+Fw[i-1,3][[1]]}
res=Reduce('+',Fw[n,])
length\dist=res@coef[-1]/sum(res@coef[-1])
```

## C Rescaling function

```
scaling=function(score,q0=NULL,rho\int=c(1e-10,1e10),tol=1e-10)
{
  logsumexp=function(l) {
    i=which.max(l);
   res=l[i]+log1p(sum(exp(l[-i]-l[i])));
   if (is.nan(res)) res=-Inf;
   return(res);
  }
  if (is.null(q0)) q0=rep(1/length(score),length(score))
  test=c(sum(q0*score)<0,sum(score>0)>0)
  if (sum(test)<2) stop("conditions␣are␣not␣valid")
# numerical optimization of the temperature
  opt=NULL
while (is.null(opt)) {
  try({opt=uniroot(f=function(rho) return(logsumexp(log(q0)+rho*score)),
        rho\int,tol=tol)},silent=TRUE)
  rho\int[1]=10*rho\int[1]
}
rho=opt$root
temp=1/(opt$root)
  q1=q0*exp(rho*score)
```

```
22  cat('q1=', q1)
23    sigma=log(q1/q0)
24          # rbind(sigma,log(q1/q0))
25    return(list(q0=q0,score=score,temp=temp,q1=q1,sigma=sigma))
26  }
```

## References


Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Richard Arratia and Michael S Waterman. The erdos-rényi strong law for pattern matching with a given proportion of mismatches. *The Annals of Probability*, pages 1152–1169, 1989.

Ana Arribas-Gil, Catherine Matias, et al. A context dependent pair hidden markov model for statistical alignment. *Stat Appl Genet Mol Biol*, 11(1):5, 2012.

John A. D. Aston and Donald E. K. Martin. Distributions associated with general runs and patterns in hidden markov models. *The Annals of Applied Statistics*, 1(2):585–611, 2007.

Mark Borodovsky and James McIninch. Genmark: Parallel gene recognition for both DNA strands. *Computers and Chemistry*, 17(2), 1993.

Claudie Chabriac, Agnès Lagnoux, Sabine Mercier, and Pierre Vallois. Elements related to the largest complete excursion of a reflected bm stopped at a fixed time. application to local score. *Stochastic Processes and their Applications*, 124(12):4202–4223, 2014.

Jie Chen and Jospeh Glaz. Scan statistics for monitoring data modeled by a negative binomial distribution. *Communications in Statistics-Theory and Methods Ser. A.*, 45(6): 1632–1642, 2016.

Jean-Jacques Daudin, Marie Pierre Etienne, and Pierre Vallois. Asymptotic behavior of the local score of independent and identically distributed random sequences. *Stochastic processes and their applications*, 107(1):1–28, 2003.

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. Relative mutability of amino acids. *Atlas of Protein Sequence and Structure*, 5:suppl. 3, 1978.

Amir Dembo and Samuel Karlin. Strong limit theorems of empirical functionals for large exceedances of partial sums of iid variables. *The Annals of Probability*, pages 1737–1755, 1991a.

Amir Dembo and Samuel Karlin. Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of markov variables. *The Annals of Probability*, pages 1756–1767, 1991b.

Hugo Devillers, Hélène Chiapello, Sophie Schbath, and Meriem El Karoui. Robustness assessment of whole bacterial genome segmentations. *Journal of Computational Biology*, 18(9):1155–1165, 2011.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

María Inés Fariello, Simon Boitard, Sabine Mercier, David Robelin, Thomas Faraut, Cécile Arnould, Julien Recoquillay, Olivier Bouchez, Gérald Salin, Patrice Dehais, et al. Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. *Molecular Ecology*, 2017.

Joseph Glaz, Joseph Naus, and Sylvan Wallenstein. *"Introduction". Scan Statistics*. Springer Series in Statistics, 2001.

Simona Grusea and Sabine Mercier. Improvement on the distribution of maximal segmental score in a Markovian sequence. *Jour. Applied Prob.*, 57.1, 2020 2020.

Mickael Guedj, David Robelin, Mark Hoebeke, Marc Lamarine, Jérôme Wojcik, Gregory Nuel, et al. Detecting local high-scoring segments: A first-stage approach for genome-wide association studies. *Statistical applications in genetics and molecular biology*, 5(1): 1192, 2006.

Laurent Guéguen. Sarment: Python modules for HMM analysis and partitioning of sequences. *Bioinformatics*, 21:3427–3428, 2005.

Claudie Hassenforder and Sabine Mercier. Exact distribution of the local score for markovian sequences. *Annals of the Institute of Statistical Mathematics*, 59(4):741–755, 2007.

Samuel Karlin. Statistical signals in bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13355–13362, 2005.

Samuel Karlin and Stephen F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268, 1990.

Samuel Karlin and Amir Dembo. Limit distributions of maximal segmental score among markov-dependent partial sums. *Advances in Applied Probability*, 24(01):113–140, 1992.

Samuel Karlin and Friedemann Ost. Maximal length of common words among random letter sequences. *The Annals of Probability*, pages 535–563, 1988.

Jonathan M Keith. Sequence segmentation. *Bioinformatics: Data, Sequence Analysis and Evolution*, pages 207–229, 2008.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Anders Krogh, Bjorn Larsson, Gunnar von Heijne, and Erik L.L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *J. Mol. Biol.*, 305:567–580, 2001.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

Agnès Lagnoux, Sabine Mercier, and Pierre Vallois. Probability that the maximum of the reflected brownian motion over a finite interval $[0, t]$ is achieved by its last zero before $t$. *Electronic Communications in Probability*, 20, 2015.

Agnès Lagnoux, Sabine Mercier, and Pierre Vallois. Statistical significance based on length and position of the local score in a model of iid sequences. *Bioinformatics*, page btw699, 2017.

Agnès Lagnoux, Sabine Mercier, and Pierre Vallois. Probability density function of the local score position. *Stochastic Processes and their Applications*, 129:3664–3689, 2019.

Alexandra Lefebvre, Sabine Mercier, and Gregory Nuel. Unsupervised learning with confidence intervals of a scoring function with constrained hidden markov models. *Submitted*, 2020.

The Minh Luong, Yves Rozenholc, and Gregory Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. *Computational Statistics & Data Analysis*, 68:129–140, 2013.

Sabine Mercier and Jean-Jacques Daudin. Exact distribution for the local score of one iid random sequence. *Journal of Computational Biology*, 8(4):373–380, 2001.

Sabine Mercier, Dominique Cellier, and D Charlot. An improved approximation for assessing the statistical significance of molecular sequence features. *Journal of applied probability*, 40(02):427–441, 2003.

Alexander Yu Mitrophanov and Mark Borodovsky. Statistical significance in biological sequence analysis. *Briefings in Bioinformatics*, 7(1):2–24, 2006.

Kasper Munch and Anders Krogh. Automatic generation of gene finders for eukaryotic species. *BMC bioinformatics*, 7(1):263, 2006.

Joseph I. Naus. Approximations for distributions of scan statistics. *Journal of the American Statistical Association*, 77 (377):177–183, 1982.

Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48 (3):443–453, 1970.

Grégory Nuel. Effective p-value computations using finite markov chain imbedding (fmci): application to local score and to pattern statistics. *Algorithms for molecular biology*, 1 (1):5, 2006.

William R Pearson. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, pages 3–1, 2013.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Gesine Reinert and Michael S Waterman. On the length of the longest exact position match in a random sequence. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(1):153–156, 2007.

David Robelin. *Détection de courts segments inversés dans les génomes-méthodes et applications*. PhD thesis, Université Paris Sud-Paris XI, 2005.

Stéphane Robin, François Rodolphe, and Sophie Schbath. *DNA, words and models: statistics of exceptional words*. Cambridge University Press, 2005.

Walter L Ruzzo and Martin Tompa. A linear time algorithm for finding all maximal scoring subsequences. In *ISMB*, volume 99, pages 234–241, 1999.

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Daiya Takai and Peter Jones. Takai, d. and jones, p.a. comprehensive analysis of cpg islands in human chromosomes 21 and 22. proc. natl. acad. sci. usa 99, 3740-3745. *Proceedings of the National Academy of Sciences of the United States of America*, 99:3740–5, 03 2002.

Daiya Takai and Peter A. Jones. The cpg island searcher: A new www resource. *In silico biology*, 3(3):235–240, 2003.

Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13, 1967.

Michael S. Waterman. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall, 1995.

Stefan Wolfsheimer, Alexander Hartmann, Ralf Rabus, Gregory Nuel, et al. Computing posterior probabilities for score-based alignments using ppalign. *Stat. Appl. Genet. Mol. Biol.*, 11:Article1, 2012.

Kyoung-Jae Won, Thomas Hamelryck, Adam Prügel-Bennett, and Anders Krogh. An evolutionary method for learning hmm structure: prediction of protein secondary structure. *BMC bioinformatics*, 8(1):357, 2007.

Robin Xavier, Turck Natacha, and Hainard et al. Alexandre. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 7(77):12–77, 2011.

Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415, 2009.

Bo Zhao and Joseph Glaz. Scan statistics for detecting a local change in variance for two dimensional normal data. *Communications in Statistics-Theory and Methods Ser. A.*, 46(11):5517–5530, 2017.

Gang Zhao and Erwin London. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein science*, 15(8):1987–2001, 2006.