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

Sabine Mercier, Grégory Nuel

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

Probabilizing the segmentation space in local score approaches

S. Mercier · G. Nuel

Received: date / Accepted: date

Abstract We are interested here in 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 the local score approach 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 to manage ambiguous or uncertain sequence letters and to construct relevant scoring schemes. We illustrate our approach by considering the TM-tendency scoring function.

Keywords Local score · 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: 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, the profile-HMMs, pair-HMMs, and context-sensitive models. Note that HMM are 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 J. (1993) for a HMM modeling protein-coding gene. Note that there also exist a large number of HMM variants to meet the needs of various applications (see Yoon (2009) for more explanation).

Score-Based approaches: Score-Based approaches have also been extensively 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 for example. 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 ℓ and calculating a cumulative score for each window. This leads to a graphical representation where the maximal region of length ℓ can easily be observed. While this approach is appealing for sequence analysis, its major drawback is that it needs to make a choice on ℓ.

The alternative to sliding windows, when there is no "natural" value for ℓ, is the local score. The local score, defined for the first time in 1990s (see Karlin and Altschul, 1990, or Eq. (1) in Section 2), classically denoted H_n for a sequence of length n, 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. Basically, all sliding windows 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). It is most unfortunate that sliding window approaches are

still nowadays often preferred in online software to local score in the context of bioinformatics, see for example https://web.expasy.org/protscale/ (Gasteiger et al, 2005).

In this paper we prove, that under specified hypotheses, HMMs and Score-Based approaches are in fact equivalent in the sense that they provide the same segmentation distributions. We present in detail the connections between the both approaches. We shall then demonstrate how we have derived several interesting probabilistic results for the local score approach, including scoring function rescaling and posterior distribution of the segmentation space.

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.

Our contribution

The main purpose of the present work is to establish the duality between a generative constrained HMM framework and a Score-Based one verifying given hypotheses. This main result brings very interesting corollaries and applications using the fact that results from both approaches can be transferred to the other. We present here how an adaptation of the forward and backward algorithm allows to compute the full posterior distribution (including localization, 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 are also possible but will not take place in this article but in further works focusing more on the statistical aspects of the problem.

Section 2 is devoted to the local score with a state of the art and a proposition of a probabilization of the segmental space. 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 scheme.

2 Local score

2.1 State of art for the local score

The local score of random sequences 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 Markov chain). Each context uses very different mathematical tools to establish results: Markov Chain theory, suitable for relatively short sequences (Mercier and Daudin, 2001; Nuel, 2006; Hassenforder and Mercier, 2007); renewal theory for long sequences under the critical assumption that $\mathbb{E}[f] < 0$) (Karlin and Dembo, 1992; Mercier et al, 2003); Brownian motion theory for very long sequences and $\mathbb{E}[f] = 0$ (Daudin et al, 2003; Chabriac et al, 2014); or Chen-Stein Theory (Waterman and Vingron, 1994; Hansen, 2006) when considering the local score of pairwise comparison.

Many articles deal with the statistical significance of the local score. 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 scheme in (Karlin and Dembo, 1992). An improvement of this last result in proposed in Grusea and Mercier (2018).

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 of 0 and 1 with a proportion of 1 being larger than a given threshold. Dembo and Karlin (1991b,a) proved an asymptotic behavior of the length of this optimal segment when the length of the sequence goes to infinity and $\mathbb{E}[X] < 0$ (with X the random variable taking value in $\{0, 1\}$). 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 localisation 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 precise enough to know exactly where the local score segment begins. In Lagnoux et al (2018), an asymptotic density for the index position of the local score in the sequence is exposed. These results based on the Brownian motion theory can be useful for sequence length 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) 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 $H_f(\mathbb{A}) \geq 0$).

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. We can define a bijection between $\mathcal{S}$ and $I \subset \{1, \dots, n\}$. For all $S \in \mathcal{S}$, we can define $I = I_S \in \mathcal{I} = \{I = i, j : 1 \leq i \leq j \leq n\}$, such $I_S = \{i \in \{1, \dots, n\}, S_i = 2\}$. Conversely, for any interval (possibly empty) $I = [a, b] \subset \{1, \dots, n\}$, $\exists S_I \in \mathcal{S}$ such as: $S_k = 1 \times \mathbf{1}_{k<a} + 2 \times \mathbf{1}_{k \in [a,b]} + 3 \times \mathbf{1}_{k>b}$. In the following example, $I = [2, 4]$ and $n = 5$.

$$\begin{array}{rl} \text{Index} = & 1\ 2\ 3\ 4\ 5 \\ S \quad = & 1\ 2\ 2\ 2\ 3 \end{array}$$

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 approach focuses on the segment that realizes the maximum accumulated score. But several segments can realize the local score and moreover biological approach can also be interested in the suboptimal segments. Let consider all the different possible segments of the sequence. Let propose the following Gibbs distribution on the space $\mathcal{I}$ of possible segmentations:

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

where 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 local score approach that focus only on the segments that realizes the local score 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 in the four nucleotide alphabet $\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ and the corresponding scores $f(\mathtt{A}) = f(\mathtt{C}) = -2$ 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 significant segment. In top right panel, suboptimal segments appear. For a larger T, segments become more equiprobable (bottom panels with a change of x axis scale). The question is then to chose 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 hypothesis.

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.

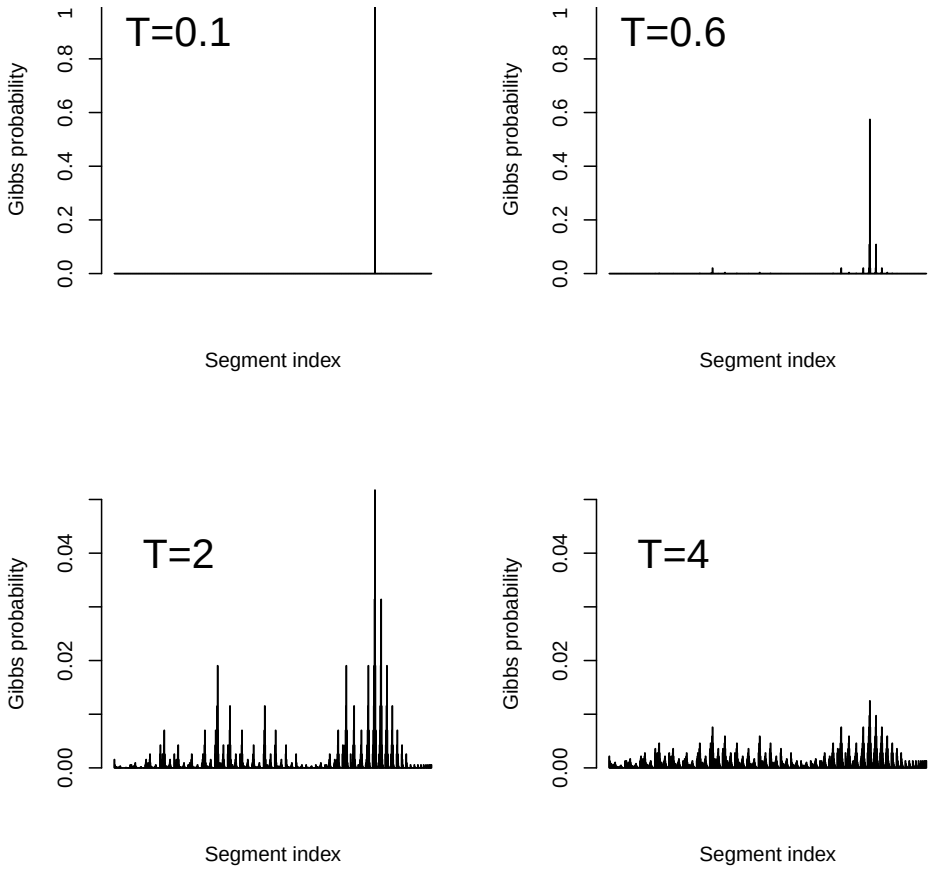

Fig. 1 Gibbs distributions of the 821 segments of a length $n = 40$ sequence for different temperature : Top left panel with $T = 0.1$; Top right panel with $T = 0.6$; Bottom left panel with $T = 2$ and bottom right with $T = 4$. Note the scale change of the x axis from the top panels to the bottom panels.

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:

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

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 $S_i = 2$).

If we assume that all $S \in \mathcal{S}$ are equiprobable then

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

where the normalization factor Z is given by:

$$Z \stackrel{\text{def}}{=} \sum_{S \in \mathcal{S}} \mathbb{P}^{q_0,q_1}_{\text{HMM}}(\mathbb{A}|S) = \sum_{S \in \{1,2,3\}^n} \mathbf{1}_{\{S \in \mathcal{S}\}} \mathbb{P}^{q_0,q_1}_{\text{HMM}}(\mathbb{A}|S) \tag{4}$$

When non ambiguous, we will drop "HMM" and "q_0, q_1" for the sake of simplification.

The Markov chain defined by $(\mathbb{A}, S)$ is constrained due to the definition of the set $\mathcal{S}$. It is a very simple HMM model with only three states that differs from classical HMM. More details are given in Appendix A.

Posterior probabilities

Inspired from classical HMM inference (see Rabiner, 1989; Durbin et al, 1998), we adapted the so-called forward and backward quantities noted $F_j(k)$ and $B_j(k)$ for $j = 1, \ldots, n$ and $k \in \{1, 2, 3\}$ (see Appendix A.2 for definitions). We can also 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 in $j = n$ since it provides the posterior probability of having or not an empty 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:

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

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

Different implementation 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 have a given length. Illustration of the computation of those quantities is given in Section 5.1.

4 Equivalence between the two probability distributions

Let consider a component sequence $\mathbb{A} \in \mathcal{A}^n$. Let denote $\mathcal{M}_\mathcal{A}$ the set of multinomial distributions on $\mathcal{A}$.

Theorem 1

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

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 that $\mathbb{P}(S) = 1/|\mathcal{S}|$ for all $S \in \mathcal{S}$, we then have with Z defined in (4)

$$\mathbb{P}_{\text{HMM}}^{q_0,q1}(S|\mathbb{A}) = \frac{1}{Z}\mathbb{P}(\mathbb{A}|S) \propto \frac{\mathbb{P}(\mathbb{A}|S)}{\mathbb{P}(\mathbb{A}|S = 1 \dots 1)}$$

and

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

The function defined as $\sigma(a) \overset{\forall a \in \mathcal{A}}{=} \log(q_1(a)/q_0(a))$ verifies the assumption. 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 assumption 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* $(\exists a, f(a) > 0)$,

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

and $q_1(a) \overset{\forall a \in \mathcal{A}}{=} q_0(a) \exp(f(a)/T)$.

Proof Let consider the following Lemma.

Lemma 1 *For any scoring function* $f : \mathcal{A} \to \mathbb{R}$*, 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 Let 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 1 – *The two assumptions,* $\sum_a q_0(a) f(a) < 0$ *and* $\exists a, f(a) > 0$ *are classical in the context of local score study (see Karlin and Altschul, 1990) and are usually verified in practice using 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* σ *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 establishes the adapted scores for ambiguous and uncertain components. An application of Theorem 1 allows also to learn scoring function from data. This will be illustrated in a forthcoming work.

An example of application of Theorem 2 is proposed on real proteins in Section 5.3. In this example, we consider the TM-tendency scale proposed in (see Zhao and London, 2006) and 56 human soluble and helical transmembrane (TM) proteins. This score function, proposed in 2006, is presented as a refined "hydrophobicity"-type TM sequence prediction scale that should approach the theoretical limit of accuracy, and seems more accurate that the well-known Kyte and Doolittle hydrophobic scale (see Kyte and Doolittle, 1982). We rescale it and compare the deduced scoring function to a more recent hydrophobic scoring function.

5 Applications

Three applications are presented here to illustrate the results in the previous section. Illustration of 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 long sequence and possible overflows is also illustrated and practical computation methods are proposed. Section 5.2 presents how the HMM approach can allow to compute adapted scores for ambiguous components. We illustrate in Subsection 5.3 how the TM-tendency scoring function of Zhao and London can be rescaled in order to make the two probability spaces to be equivalent and to give a better mathematical interpretability to the result.

5.1 Posterior probabilities on simulated sequences

5.1.1 A binary toy example

Here, we consider for simplification a sequence $\mathbb{A} = 1011011001$ in $\{0, 1\}^{10}$ to illustrate the computation. Appendix B presents simulation details and imple-

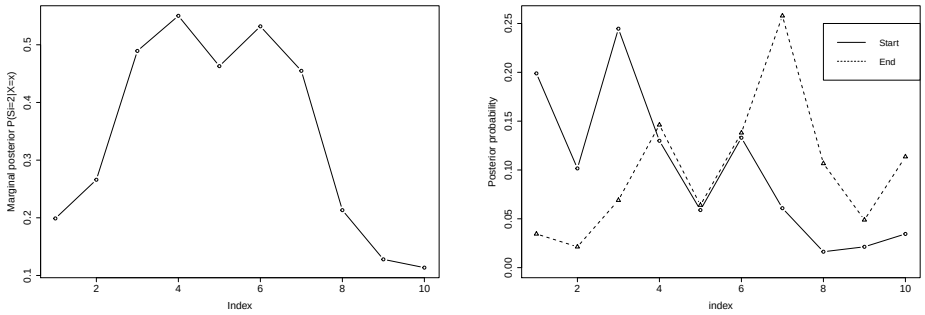

Fig. 2 Toy example ($\mathcal{A} = \{0, 1\}$, $n = 10$, $\mathbb{A} = 1011011001$, $f(0) = -2$ and $f(1) = 1$): Marginal posterior distribution of segmentation State 2 (left panel) and Segment start/end posterior distribution (right panel).

Table 1

	$P(S_i = 1)$	$P(S_i = 2)$	$P(S_i = 3)$
$i = 1$	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

mentation. The simulated sequence length is only $n = 10$ for simplicity. Score scheme has been chosen as follows: $f(0) = -2$ and $f(1) = 1$. Figure 2 (left panel) represents the probabilities of the sequence positions to be in State 2 ; computation for posterior probabilities for starting and ending index is presented in the right panel. Table 1 give the values of the posterior probabilities $\mathbb{P}(S_i = k|X)$ 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.

5.1.2 A practical implementation for long sequence

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

The simulated 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

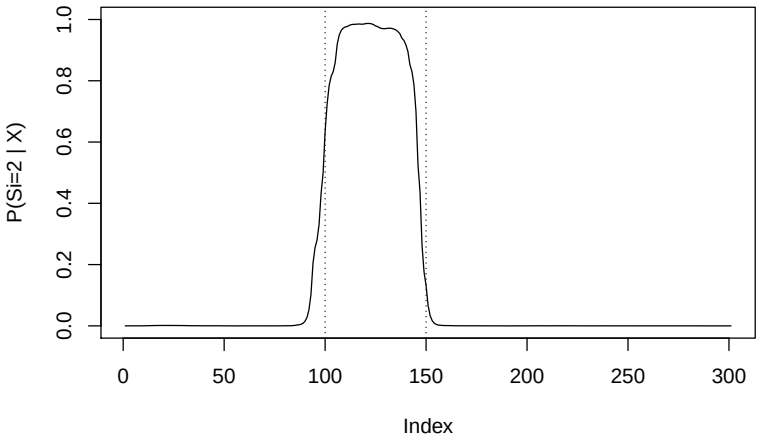

Fig. 3 Marginal posterior distribution of segmentation State 2 for the simple simulation with $n = 300$ and $I = [150, 200]$, $q_0 = (0.4, 0.2, 0.2, 0.2)$, $q_1 = (0.2, 0.4, 0.3, 0.1)$ and $f = \log(q_0/q_1)$.

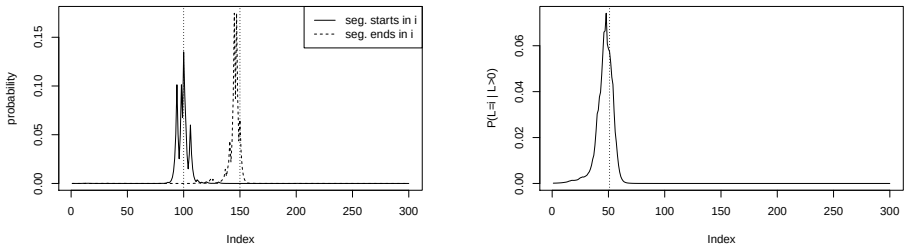

Fig. 4 Segment start/end posterior distribution (left panel) and length (right panel) for the simple simulation with $n = 300$ and $I = [100, 150]$, $q_0 = (0.4, 0.2, 0.2, 0.2)$, $q_1 = (0.2, 0.4, 0.3, 0.1)$ and $f = \log(q_0/q_1)$.

inserted in $I = [100, 150]$. The score scheme used corresponds to $\log(q_0/q_1)$ that leads to: $f(1) = -0.6931472$, $f(2) = 0.6931472$, $f(3) = 0.4054651$, $f(4) = -0.6931472$. Figure 3 represents the probabilities of the sequence component to be in State 2, that is in the inserted segment, $\mathbb{P}(S_j = 2|X) = F_j(2)B_j(2)/Z$. In this figure, the inserted segment is clearly highlighted, and position and length seem correct. Computation for posterior probabilities for starting and ending index 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 sequence length greater than 10^4 are also very fast (the complexity is in $\mathcal{O}(n)$) and recover correctly small atypical segment of hundred components. Figure 5 illustrates the computation for $n = 40000$ and a true segment inserted in $I = [10000, 11000]$.

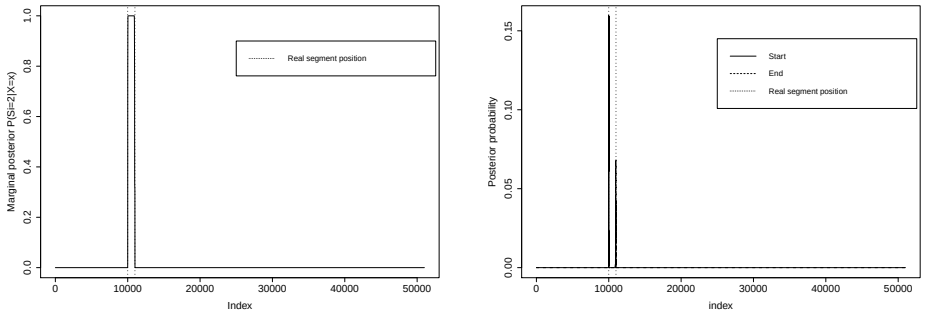

Fig. 5 Long sequence example: Marginal posterior distribution of segmentation State 2 (left panel) and Segment start/end posterior distribution (right panel).

5.2 IUPAC ambiguous DNA code

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

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, under assumption, there exists a unique rescaling T such that $q_1(a) = q_0(a)\exp(f(a)/T)$ defines a probability distribution. And Corollary 1 defines 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 corresponding scores $f(\mathtt{A}) = -2$, $f(\mathtt{C}) = -1$, $f(\mathtt{G}) = 0$, $f(\mathtt{T}) = +1$ and two different background distributions $q_0^U = (0.25, 0.25, 0.25, 0.25)$ and $q_0^{\mathtt{GC}} = (0.1, 0.4, 0.4, 0.1)$. These two distributions verifies the two assumptions of a non positive average score and possible non negative scores. Thanks to Theorem 2, we know that there exists a unique rescaling T such that $q_1(a) = q_0(a)\exp(f(a)/T)$ defines a probability distribution. 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 σ values for both background distributions q_0^U and $q_0^{\mathtt{GC}}$. Table 2 gives how should be the corresponding scores for the four nucleotides and the ambiguous components for each background distribution. For ambiguous letters, we have for example $\sigma(\mathtt{M}) = \sigma(\mathtt{A} \text{ or } \mathtt{C}) = -1.23 \neq -1.76 + (-0.88)$.

Moreover, considering the usual approach of highlighting atypical segments using local score approach, we simulate $5 \cdot 10^3$ sequences of 100 components of the IUPAC alphabet under the following distribution $q_0(x) = 0.213$ for $x = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$, $q_0(x) = 0.021$ for $x = \mathtt{M}, \mathtt{R}, \mathtt{W}, \mathtt{S}, \mathtt{Y}, \mathtt{K}$, $q_0(x) = 0.004$ for $x = \mathtt{V}, \mathtt{H}, \mathtt{D}, \mathtt{B}$ and $q_0(\mathtt{N}) = 0.006$. We insert a segment of 20 components using $q_1(x) = 0.056$ 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 the scoring function $\sigma(x) = \log(q_1(x)/q_0(x))$ for $x = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ and we calculate σ for the other components using Corollary 1. We also calculate the score $\tilde{\sigma}$ of 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 highlight that the scores can be quite different. We calculate the local score of each sequence with both scoring function and we highlight the segment which realizes the local score in both case. Allowing an error of 3 (*resp.* 5) positions for the beginning and the ending index position, the scoring function σ detect 10% (*resp.* 22%) of the correct inserted segments and $\tilde{\sigma}$ only detect 1% (*resp.* 3%) of them. Figure 8 shows an example for one sequence. In this example, the inserted segment is correctly detect by the local score calculated with σ scoring function, but the segment realizing the local score for $\tilde{\sigma}$ is not correct. We use the Lindley process defined as follows, $U_0 := 0$ and $U_{k+1} := \max(0, U_k + f(A_k))$, and the fact that $H_f(\mathbb{A}) = \max_{0 \leq k \leq n} U_k$ (see (Mercier and Daudin (2001) for more explanation) to visualize the segments realizing the two local scores.

Remark 2 *Uncertainty (*e.g. *in protein crystallisation, 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:*

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

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 align-

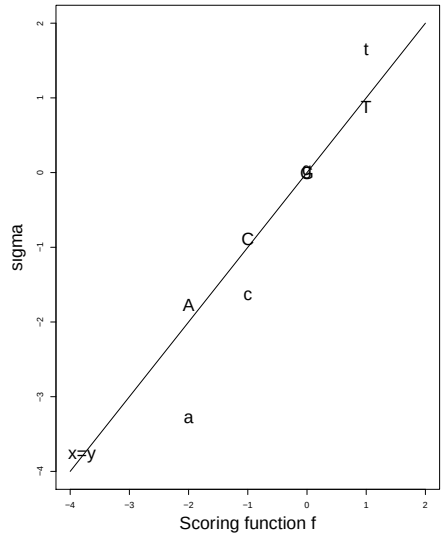

Fig. 6 Initial scoring values for the four nucleotides versus the σ values for both background distributions q_0^U (A, C, G, T) and q_0^{GC} (a, c, g, t)

Table 2

IUPAC Code	A	C	G	T/U		
Meaning	A	C	G	T		
f	-2	-1	0	+1		
σ for q_0^U	−1.76	−0.88	0.00	0.88		
σ 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
σ for q_0^U	-1.23	-0.54	0.26	-0.35	0.34	0.54
σ 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	anyone	
σ for q_0^U	-0.64	0.00	0.18	0.24	0.00	
σ for q_0^{GC}	-0.63	0.00	0.43	0.1	0.00	

ment. 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 scheme 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) given in Table 3.

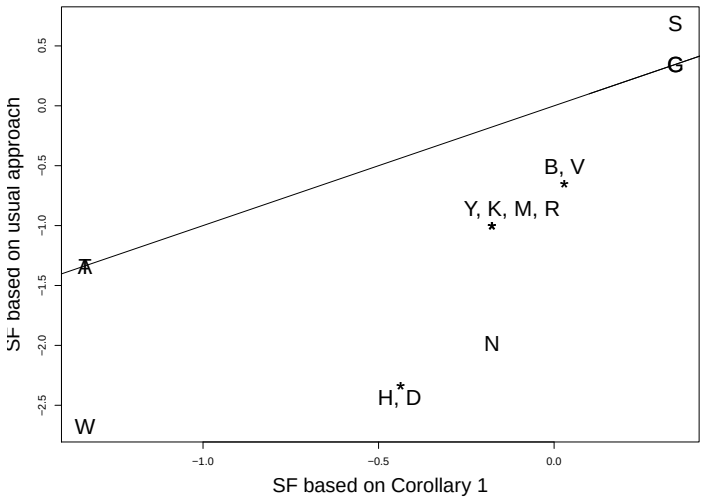

Fig. 7 Scoring function (SF) σ using Theorem 1 for A, C, G, T and Corollary 1 for the ambiguous components, and $\tilde{\sigma}$ using Theorem 1 for $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ and the usual approach for the other componenets.

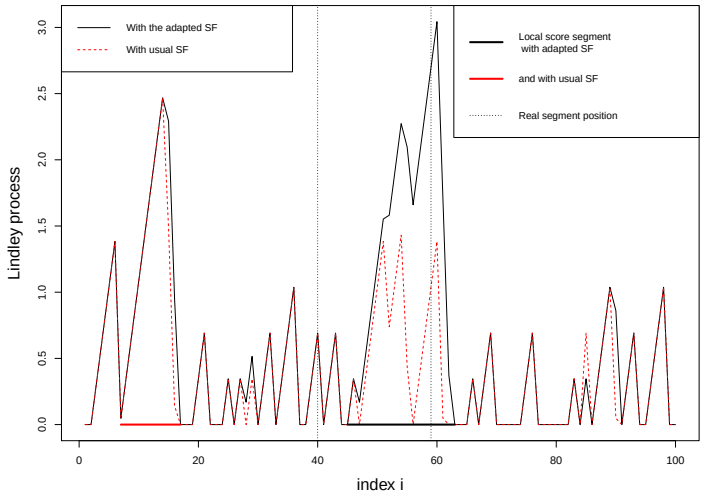

Fig. 8 Local score segments realization on a sequence example for both the adapted scoring function σ using Theorem 1 for A, C, G, T and Corollary 1 for the ambiguous components, and $\tilde{\sigma}$ using Theorem 1 for A, C, G, T and the usual approach for the other components. The adapted Sf detect correctly the inserted segment whereas the usual one not.

Table 3 Zhao and London TM-tendency scale (ZL) Zhao and London (2006). Distribution q_0 (in %) for 56 TM proteins without their TM regions and Zhao and London score rescaled (ZLr) using q_0 distribution.

	A	C	D	E	F	G	H	I	K	L
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
ZLr	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
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
ZLr	0.472	-0.546	-0.485	-0.620	-0.866	-0.179	-0.108	0.492	0.516	0.165

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 $\mathbb{E}[X] \simeq -0.50 < 0$.

We find that $T = 0.34$ that leads to the rescaled scoring function 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.

Let us now consider sp|O15393|TMPS2_HUMAN, a supplementary human TM protein with a transmembrane segment at position 85-105. Let us apply the generating model using the rescaled scoring function of Zhao and London. Figure 9 gives the posterior probabilities that the components of the sequence are in State 2. The upper probabilities are around $\simeq 0.75$ and correspond to the TM segment. The start of the TM region is precise and its end is very close to the real one. The probability that there exists a segment of a different distribution than q_0 is equal to 99.99%. Figure 10 gives segment start/end posterior probabilities (left panel) and posterior probabilities for length (right panel). The length is relatively accurate even if the probabilities is 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).

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 local score. This probabilization of the segmentation

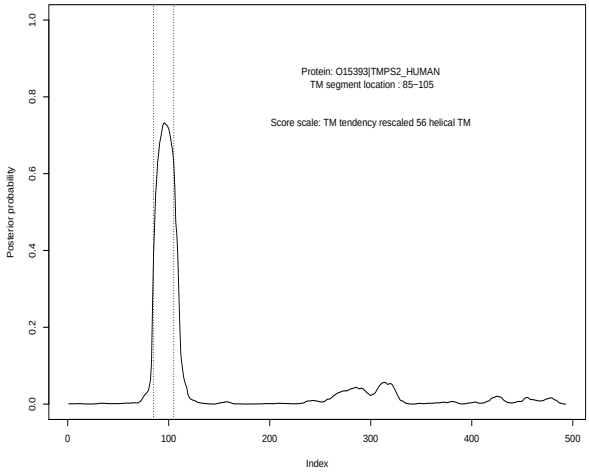

Fig. 9 Posterior probability that S_i in Sate 2 using the rescaled Zhao and London score function.

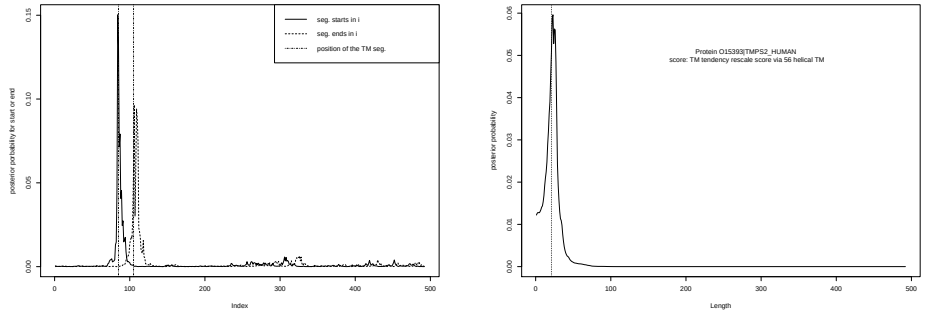

Fig. 10 Segment start/end posterior probabilities (left panel) and posterior probability there exists a segment of a given length (right panel) using the Zhao and London rescaled scores.

space allows us to establish a duality between the HMM approach and the scoring function approach 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: can we say that there is an atypical segment in the sequence and if yes, where is it ? This approach can also provide the probability that the sequence contains an atypical segment, the probabilities for the position of the atypical segment and other information, on the length of

the segment. The computation are easy and fast even for very long sequences with length greater than 10^5 or more.

Biologists can adapt the classical scoring schemes by rescaling them. By 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 $A_{i_0}, \ldots, A_{j_0}$ are distributed with a distribution $q_1 \neq q_0$." (with H_0: "$A_1, \ldots, A_n$ are q_0 distributed.") is verified.

Moreover, this can be applied to sequences with uncertain position (which is often the case in the New Generation Sequencing era) for which canonical score can be derived from the original scoring scheme.

One important limitation of our model is that it is assuming the presence of exactly one segment of interest inserted in each sequence. Extending our model to situations where there is more than one atypical segment is possible and this would clearly be 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 computational complexity would simply increase linearly with the number of atypical segments.

Another perspective of this work consists in learning scoring schemes: 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 maximum likelihood computation (EM or direct optimization) both allowing to estimate the new scoring scheme but also establish confidence intervals, hypothesis testing etc... Unsupervised learning can also be considered. If using HMM method to learn 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 $i = 2, \ldots, n$. 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:

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

A.2 Forward and backward

Classical uses of forward and backward quantities correspond to HMM probabilities (see (see Durbin et al, 1998)). Here, we adapt the forward and backward definition using Bayesian network potentials as it allows us to take into account constraints given in the transition π. 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 $F_1 \stackrel{\text{def}}{=} \phi_1$ and for $i = 2, \ldots, n$: $F_i(k) := F_i(S_i = k|\theta) \stackrel{\text{def}}{=} \sum_{S_{1:i-1}, S_i = k} \prod_{u=1}^i \phi_u(C_u|\theta)$. We can easily prove by induction that:

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

with our specific structure of π 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: $B_n \stackrel{\text{def}}{=} 1$ and for $i = n, \ldots, 2$:

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

Again, by induction we easily get: $B_{i-1}(S_{i-1} = j|\theta) = \sum_k \phi_i(S_{i-1} = j, S_i = k|\theta) B_i(S_i = k|\theta)$. With our specific structure of π 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 $\phi_1(S_1)B_1(S_1 = k) = \mathbb{Q}(S_1 = k, \mathbb{A} = a|\theta)$. 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

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

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 $S = 1, \ldots, 1$ is not possible anymore. Its 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 $\mathbb{P}(\text{ev}) = \sum_k F_n(k) \times B_n(k)$:

$$\begin{aligned}
\mathbb{P}(\text{segment starts at index 1}) &= \mathbb{P}(S_1 = 2) = \frac{F_1(2) \times B_1(2)}{\mathbb{P}(\text{ev})} \\
\mathbb{P}(\text{segment starts at index i}) &= \mathbb{P}(S_{i-1} = 1, S_i = 2) \\
&= \frac{\sum_{S_{i-1}=1} F_{i-1}(S_{i-1} = 1) \times \phi_i(S_{i-1} = 1, S_i = 2) \times B_i(S_i = 2)}{\mathbb{P}(\text{ev})} \\
&= \frac{F_{i-1}(1) \times \phi_i(1,2) \times B_i(2)}{\mathbb{P}(\text{ev})}
\end{aligned}$$

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

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

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 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) # init
# Bk[n,]=c(1.0,1.0,1.0)
Bk[n,]=c(0.0,1.0,1.0) # alternative Bk[n,] for exactly one segment
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_start=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. One way to overcome this problem is to rescale the forward and backward quantities such that each of them sums to 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 - try changing
    score or q0")
# 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)
cat('q1=', q1)
  sigma=log(q1/q0)
    # rbind(sigma,log(q1/q0))
  return(list(q0=q0,score=score,temp=temp,q1=q1,sigma=sigma))
}
```

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215(3):403–410

Arratia R, Waterman MS (1989) The erdos-rényi strong law for pattern matching with a given proportion of mismatches. The Annals of Probability pp 1152–1169

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

Borodovsky M, J M (1993) Genemark: Parallel gene recognition for both DNA strands. Computers and Chemistry 17(2):123–133

Chabriac C, Lagnoux A, Mercier S, Vallois P (2014) 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

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

Daudin JJ, Etienne MP, Vallois P (2003) Asymptotic behavior of the local score of independent and identically distributed random sequences. Stochastic processes and their applications 107(1):1–28

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

Dembo A, Karlin S (1991a) Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of markov variables. The Annals of Probability pp 1756–1767

Dembo A, Karlin S (1991b) Strong limit theorems of empirical functionals for large exceedances of partial sums of iid variables. The Annals of Probability pp 1737–1755

Devillers H, Chiapello H, Schbath S, Karoui ME (2011) Robustness assessment of whole bacterial genome segmentations. Journal of Computational Biology 18(9):1155–1165

Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press

Fariello MI, Boitard S, Mercier S, Robelin D, Faraut T, Arnould C, Recoquillay J, Bouchez O, Salin G, Dehais P, et al (2017) Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. Molecular Ecology

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M, Appel R, Bairoch A (2005) Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press

Glaz J, Naus J, Wallenstein S (2001) "Introduction". Scan Statistics. Springer Series in Statistics

Grusea S, Mercier S (2018) Improvement on the distribution of maximal segmental score in a Markovian sequence. Submitted

Guedj M, Robelin D, Hoebeke M, Lamarine M, Wojcik J, Nuel G, et al (2006) Detecting local high-scoring segments: A first-stage approach for genome-wide association studies. Statistical applications in genetics and molecular biology 5(1):1192

Hansen NR (2006) Local alignment of markov chains. The Annals of Applied Probability pp 1262–1296

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

Karlin S (2005) Statistical signals in bioinformatics. Proceedings of the National Academy of Sciences of the United States of America 102(38):13,355–13,362

Karlin S, Altschul SF (1990) 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

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

Karlin S, Ost F (1988) Maximal length of common words among random letter sequences. The Annals of Probability pp 535–563

Keith JM (2008) Sequence segmentation. Bioinformatics: Data, Sequence Analysis and Evolution pp 207–229

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. J Mol Biol 305:567–580

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. Journal of molecular biology 157(1):105–132

Lagnoux A, Mercier S, Vallois P (2015) 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

Lagnoux A, Mercier S, Vallois P (2017) Statistical significance based on length and position of the local score in a model of iid sequences. Bioinformatics p btw699

Lagnoux A, Mercier S, Vallois P (2018) Probability density function of the local score position. Accepted to Stochastic Processes and their Applications

Luong TM, Rozenholc Y, Nuel G (2013) Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. Computational Statistics & Data Analysis 68:129–140

Mercier S, Daudin JJ (2001) Exact distribution for the local score of one iid random sequence. Journal of Computational Biology 8(4):373–380

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

Mitrophanov AY, Borodovsky M (2006) Statistical significance in biological sequence analysis. Briefings in Bioinformatics 7(1):2–24

Munch K, Krogh A (2006) Automatic generation of gene finders for eukaryotic species. BMC bioinformatics 7(1):263

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

Needleman SB, Wunsch CD (1970) 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

Nuel G (2006) 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

Pearson WR (2013) An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics pp 3–1

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

Reinert G, Waterman MS (2007) 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

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

Robin S, Rodolphe F, Schbath S (2005) DNA, words and models: statistics of exceptional words. Cambridge University Press

Ruzzo WL, Tompa M (1999) A linear time algorithm for finding all maximal scoring subsequences. In: ISMB, vol 99, pp 234–241

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Molecular systems biology 7(1):539

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. Journal of molecular biology 147(1):195–197

Waterman MS, Vingron M (1994) Sequence comparison significance and poisson approximation. Statistical Science pp 367–381

Wolfsheimer S, Hartmann A, Rabus R, Nuel G, et al (2012) Computing posterior probabilities for score-based alignments using ppalign. Stat Appl Genet Mol Biol 11:Article1

Won KJ, Hamelryck T, Prügel-Bennett A, Krogh A (2007) An evolutionary method for learning hmm structure: prediction of protein secondary structure. BMC bioinformatics 8(1):357

Yoon BJ (2009) Hidden markov models and their applications in biological sequence analysis. Current genomics 10(6):402–415

Zhao B, Glaz J (2017) 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

Zhao G, London E (2006) 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