Local stacks in a Markov chain

Niels Richard Hansen

University of Copenhagen

A stack is a subunit of the secondary structure of an RNA-molecule that forms as the molecule folds onto itself and the nucleotides pair up through hydrogen bonds. For a specific scoring of the nucleotide pairing we consider the distribution of local stack scores when regarding the RNA-molecule as a Markov chain on a finite alphabet, and we derive sufficient conditions for the number of essentially different local stacks with a score exceeding a high threshold to be asymptotically Poisson distributed. From the Poisson approximation a Gumbel approximation of the maximal local stack score is obtained.

1 Introduction

An RNA-molecule is a polymer consisting of monomer building blocks known as nucleotides. There are four different RNA nucleotides, denoted $a$, $c$, $g$, and $u$, that by covalent bonding form a polymer as a linear sequence of letters from the four letter alphabet $\{a, c, g, u\}$. In addition to the linear structure the molecule typically folds into a three dimensional structure, which plays a crucial role for the function of the molecule. The structure is formed primarily by hydrogen bonds between $a$ and $u$ and between $c$ and $g$. Such hydrogen bonded pairs of nucleotides are called Watson-Crick pairs. For an RNA-molecule given as a sequence $x_1, \ldots, x_n$ of nucleotides the set of pairs $(i, j)$ with $i > j$ that form hydrogen bonds in the three dimensional structure is known as the secondary structure of the molecule. There are some physical constraints upon which pairs that can enter the secondary structure, e.g. one nucleotide can only make one hydrogen bond and two covalently bonded nucleotides can not form a hydrogen bond, i.e. $(i + 1, i)$ can never occur in the secondary structure, for more details see e.g. (Hofacker et al. 1998).

There is a clear interest in predicting the structure – specifically the secondary structure – from the nucleotide sequence alone. This is typically done by optimisation algorithms like maximising the number of Watson-Crick pairs (Nussinov et al. 1978) or by minimising a free energy function defined on the set of secondary structures, see (Zuker et al. 1999) and the references therein. These algorithms aim at producing a single global structure for the whole sequence. Instead we focus upon local structural

AMS 1991 subject classification. Primary 60G70; secondary 60F10.

Key words and phrases. Extreme value theory, local stacks, Markov chains, Poisson approximation, RNA-structure.
A stack is defined as a (subset of a) secondary structure of the form

$$(i + 1, j - 1), (i + 2, j - 2), \ldots, (i + \Delta, j - \Delta)$$

with $i \geq j$, cf. (Hofacker et al. 1998). It is easy to verify that a secondary structure can always be partitioned into (maximal) stacks. We are then interested in locating potential stacks within a long sequence. In particular, the sequence may be so long that trying to predict a secondary structure for the whole sequence makes little sense, e.g. if the sequence is an entire genome, whereas certain parts of the sequence can show significant potential for forming stacks.

For a specific scoring of local stacks in a long sequence we show how the distribution of the stack scores behave asymptotically when the sequence is a Markov chain. Such results are necessary when searching a long sequence for potential stacks for assessing the significance of the findings. The results are highly similar to those obtained for local alignments of two Markov chains as treated in (Hansen 2004a), but having to deal with one Markov chain that folds onto itself instead of two independent Markov chains that are aligned presents some special difficulties that we need to handle. The main result is Theorem 3.1. In case the expected stack score is negative it provides sufficient conditions for the number of essentially different stacks with a score exceeding a high threshold to be asymptotically Poisson distributed and the maximal stack score to be asymptotically Gumbel distributed. In Remark 3.3 we discuss a modified version of Theorem 3.1 where we restrict the stacks to be contained within a window – something that is of particular practical value.

Theorem 3.1 provides to some extend an affirmative answer to a conjecture by Xiong \& Waterman (1997) but in a slightly different setup though, cf. the discussion in Section 5. Xiong \& Waterman (1997) show limit results for the maximal score of general RNA-structures using a simple scoring method with two parameters under the assumption that the RNA-sequence is a sequence of iid random variables. They then conjecture that for parameter values restricted to a certain subset of the parameter space an asymptotic Gumbel approximation of the distribution of the maximal score is valid. Our result is more general in the sense that we deal with Markov Chains instead of iid sequences but more restrictive in the sense that we only consider stacks. Their parametric restriction corresponds to our assumption that the expected score is negative.

The setup and statement of the main results are self-contained but the proofs rely heavily upon the results developed in (Hansen 2004a). The reader who wants to penetrate the proofs is therefore urged to consult (Hansen 2004a). The main purpose of this paper is to show how the development of local alignment theory for independent random sequences can be modified to be of use in the search for local structural features in long sequences.
2 Local stack structures

Let \((X_k)_{k \geq 1}\) be a sequence of stochastic variables taking values in a finite set \(E\), and let, for \(n \geq 1\) given, \(Y_k = X_{n-k+1}\), \(k = 1, \ldots, n\). Thus \((Y_k)_{1 \leq k \leq n}\) is the time reversion of \((X_k)_{1 \leq k \leq n}\). We assume a score function \(f : E \times E \to \mathbb{Z}\) given and define for \(i, j, \Delta \geq 0\) satisfying \(j \leq j + \Delta \leq n\) and \(n - j + 1 \leq i \leq i + \Delta \leq n\) the stochastic variable

\[
S_{i,j}^\Delta = \sum_{k=1}^{\Delta} f(X_{i+k}, Y_{j+k}) = \sum_{k=1}^{\Delta} f(X_{k+i}, X_{n-j+1-k}).
\]

The interpretation of \(S_{i,j}^\Delta\) is that we fold the sequence such that \(X_{n-j+1-\Delta}, \ldots, X_{n-j}\) forms a stack with \(X_{i+1}, \ldots, X_{i+\Delta}\) and we score the stack using the function \(f\), cf Figure 1.

\[
\cdots X_i \quad X_{i+1} \quad X_{i+2} \quad \cdots \quad X_{i+\Delta-1} \quad X_{i+\Delta} \quad X_{i+\Delta+1} \quad \cdots \\
\cdots X_{n-j+1} \quad X_{n-j} \quad X_{n-j-1} \quad \cdots \quad X_{n-j+2-\Delta} \quad X_{n-j+1-\Delta} \quad X_{n-j-\Delta} \quad \cdots
\]

Figure 1: For \(i, j, \Delta\) chosen such that \(n + 1 \leq i + j\) the figure shows how \(X_{n-j+1-\Delta}, \ldots, X_{n-j}\) is folded down under \(X_{i+1}, \ldots, X_{i+\Delta}\) to form a stack. The ‘hydrogen bonds’ are represented by the bullets ⋄.

We introduce the set

\[
\mathcal{H}_n = \{(i, j, \Delta) \mid 0 \leq j \leq j + \Delta \leq n, n - j + 1 \leq i \leq i + \Delta \leq n\},
\]

which we call the set of stack structures. An element \((i, j, \Delta) \in \mathcal{H}_n\) corresponds to the stack

\[
(i + 1, n - j), (i + 2, n - j - 1), \ldots, (i + \Delta, n - j + 1 - \Delta).
\]

The structures given by \(\mathcal{H}_n\) are closely related to the alignments defined in (Hansen 2004a) with the only difference being the restriction \(n + 1 \leq i + j\). This reflects the physical restriction of folding a sequence onto itself to form a stack.

We can efficiently summarise the scores in a score matrix \((T_{i,j})\) – still reflecting the constraints imposed on \(\mathcal{H}_n\) – defined by \(T_{i,j} = 0\) for \(i + j \leq n + 1\) and recursively

\[
T_{i,j} = \max \{0, T_{i-1,j-1} + f(X_i, Y_j)\}
\]

for \(i + j > n + 1\). Note that the score matrix is only non-zero below the ‘anti’-diagonal, that is, below the diagonal going from the lower left corner to the upper
right corner. It is quite common in the literature on RNA-structures to rotate the score matrix 90 degrees counter clockwise, so that it becomes an upper triangular matrix. We choose the definition presented here to make it directly comparable with the local alignment score matrix introduced in (Hansen 2004a).

By (Hansen 2004a, Remark 3.7)

\[
M_n = \max_{(i,j,\Delta) \in \mathcal{H}_n} S_{i,j}^{\Delta} = \max_{(i,j)} T_{i,j}.
\]  

(2)

**Definition 2.1.** A stack structure \((i, j, \Delta) \in \mathcal{H}_n\) is called an excursion if

\[
T_{i,j} = 0, \quad S_{i,j}^{\delta} > 0 \text{ for } 0 < \delta < \Delta
\]

and either \(S_{i,j}^{\Delta} \leq 0, \ i + \Delta = n, \ j + \Delta = n\)

Let \(\mathcal{E}_n\) denote the set of all excursions.

The maximum over an excursion \(e = (i, j, \Delta) \in \mathcal{E}_n\) is denoted by

\[
M_e = \max_{0 < \delta \leq \Delta} S_{i,j}^{\delta} = \max_{0 < \delta \leq \Delta} T_{i+\delta,j+\delta}.
\]

**Definition 2.2.** The essentially different excesses over \(t\) is defined as

\[
C_n(t) = \sum_{e \in \mathcal{E}_n} 1(M_e > t),
\]  

(3)

hence from (2) we get \((C_n(t) = 0) = (M_n \leq t)\).

### 3 Stacks in a Markov chain

We assume that the process \((X_k)_{k \geq 1}\) is a stationary Markov chain with the matrix \(P\) of transition probabilities being irreducible and aperiodic and with invariant measure \(\pi\). It plays a role that the Markov chain is assumed stationary since the time reversion \((Y_k)_{1 \leq k \leq n}\) is then also a time-homogeneous Markov chain. In fact, it is a stationary Markov chain with an irreducible and aperiodic transition probability matrix \(\overrightarrow{P}\) given by

\[
\overrightarrow{P}(x_0, x_1) = \frac{\pi(x_1)P(x_1, x_0)}{\pi(x_0)}.
\]

With

\[
\mu = \pi_P \otimes \pi_P(f) = \sum_{x,y \in E} f(x,y)\pi_P(x)\pi_P(y)
\]
the mean score we will assume throughout that \( \mu < 0 \).

We assume that there exist cycles \( x_1, \ldots, x_n \) (w.r.t. \( P \), i.e. \( P(x_i, x_{i+1 \mod n}) > 0 \)) and \( y_1, \ldots, y_n \) (w.r.t. to \( \overrightarrow{P} \)) such that

\[
\sum_{k=1}^{n} f(x_k, y_k) > 0,
\]

and in addition that for all \( T \geq 1 \) there exist cycles \( x_1, \ldots, x_n \) (w.r.t. \( P \)) and \( y_1, \ldots, y_n \) (w.r.t. to \( \overleftarrow{P} \)) such that

\[
\sum_{k=1}^{n} f(x_k, y_k) \neq \sum_{k=1}^{n} f(x_k, y_{k+T \mod n}).
\]

We define the \( E^2 \times E^2 \) matrix \( \Phi(\theta) \) by

\[
\Phi(\theta)(x,y),(x',y') = \exp(\theta f(x', y')) P_{x,x'} \overrightarrow{P}_{y,y'}
\]

and denote by \( \varphi(\theta) \) the spectral radius of \( \Phi(\theta) \). Since \( \partial_\theta \varphi(0) = \mu < 0 \) and since \( \varphi \) is convex there is a unique solution \( \theta^* > 0 \) to the equation \( \varphi(\theta) = 1 \).

We also need to introduce the following four \( \Phi \)-matrices. Define for \( \theta \in \mathbb{R} \)

\[
\begin{align*}
\Phi_1(\theta)(x,y,z),(x',y',z') &= \exp(\theta f(x', y') + \theta f(x', z')) P_{x,x'} \overrightarrow{P}_{y,y'} \overleftarrow{P}_{z,z'} \\
\Phi_2(\theta)(x,w,y),(x',w',y') &= \exp(\theta f(x', y') + \theta f(w', y')) P_{x,x'} P_{w,w'} \overrightarrow{P}_{y,y'} \\
\Phi_3(\theta)(x,y,z),(x',y',z') &= \exp(\theta f(x', y') + \theta f(z', x')) P_{x,x'} \overrightarrow{P}_{y,y'} \overleftarrow{P}_{z,z'} \\
\Phi_4(\theta)(x,w,y),(x',w',y') &= \exp(\theta f(x', y') + \theta f(y', w')) P_{x,x'} P_{w,w'} \overrightarrow{P}_{y,y'}.
\end{align*}
\]

Note the subtle difference between \( \Phi_1 \) and \( \Phi_3 \) and between \( \Phi_2 \) and \( \Phi_4 \) respectively. The change is a permutation of the arguments to the second \( f \), hence if \( f \) is symmetric there is no difference, but we don’t want to assume symmetry of \( f \) in general, and we need to take all four \( \Phi \)-matrices into account. Let \( \varphi_i(\theta) \) for \( i = 1, 2, 3, 4 \) denote the spectral radii of the \( \Phi \)-matrices.

If \( (\tilde{Y}_k)_{k \geq 1} \) is a Markov chain, independent of \( (X_k)_{k \geq 1} \), with transition probabilities \( \overrightarrow{P} \) we define a stochastic process by \( S_0 = 0 \) and for \( n \geq 1 \)

\[
S_n = \sum_{k=1}^{n} f(X_k, \tilde{Y}_k).
\]

A constant, \( K^* \), is defined by (1.26) in Theorem B in (Karlin & Dembo 1992) in terms of this process. We refer to (Karlin & Dembo 1992) for details.
Theorem 3.1. Assume that \( \mu < 0 \), that the regularity conditions given by (4) and (5) are fulfilled, and that \( \theta^* \) and \( K^* \) are the constants defined above. Define for \( x \in \mathbb{R} \)
\[
  t_n = \frac{\log K^* + \log(n(n-1)/2) + x}{\theta^*}
\]
and \( x_n \in [0, \theta^*) \) by \( x_n = \theta^*(t_n - \lfloor t_n \rfloor) \). Then if
\[
  \varphi_i \left( \frac{3\theta^*}{4} \right) < 1
\]
for \( i = 1, 2, 3, 4 \) it holds that
\[
  ||D(C_n(t_n)) - \text{Poi}(\exp(-x + x_n))|| \to 0
\]
for \( n \to \infty \). In particular
\[
  \mathbb{P}(M_n \leq t_n) - \exp(-\exp(-x + x_n)) \to 0
\]
for \( n \to \infty \).

The Remarks presented after Theorem 3.1 in (Hansen 2004a) hold almost verbatim for the present Theorem. We have only two additional remarks to make.

Remark 3.2. When \( E = \{a, c, g, a\} \) we should choose the score function \( f \) to reflect that Watson-Crick pairs are favourable (positive score) and non Watson-Crick pairs are unfavourable (negative score). In contrast to local alignment scoring the score \( f(x, y) \) is therefore negative when \( x = y \) and positive for some \( x \neq y \). Precisely how to choose the score function may depend on the type of stacks we search for and the distribution of the Markov chain. A log-likelihood ratio type of score seems, as for local alignments, to be a quite reasonable choice, cf. the discussion in Remark 3.5 in (Hansen 2004a). In particular, one should note the possibility of using a pair transition score function as described in (Hansen 2004a, Remark 3.5), which corresponds closely to stacking energies as used by most structure prediction programs, cf. (Zuker et al. 1999).

Remark 3.3. In Theorem 3.1 the length of the sequence \( n \) enters the threshold \( t_n \) through the term \( \log(n(n-1)/2) \) compared to the term \( \log n^2 \) in (Hansen 2004a, Theorem 3.1). This reflects that we essentially only use \( n(n-1)/2 \) entries in the score matrix in the present setup whereas in the local alignment setup we use the whole \( n^2 \) score matrix. If we further restrict our attention to a subset of the score matrix that corresponds to restricting the stacks to occur within a window of a given size, this should likewise affect the threshold scaling. Indeed, with \( w \) the window size, we define the threshold as
\[
  t_{n,w} = \frac{\log K^* + \log(n(w-1) - w(w-1)/2) + x}{\theta^*}
\]
With
\[ C_{n,w}(t,n) = \sum_{e=(i,j,\Delta) \in E_n: i+j+2\Delta \leq n+w} 1(\mathcal{M}_e > t) \]
and with
\[ \mathcal{M}_{n,w} = \max_{(i,j,\Delta) \in H_n} S_{i,j}^\Delta = \max_{i+j \leq n+w} T_{i,j}, \]
the results in Theorem 3.1 hold for \( C_{t,w}(t,n) \) and \( \mathcal{M}_{n,w} \) when \( n, w \to \infty \) in a suitable way. Suitable means that \( w \) cannot grow arbitrarily slowly compared to \( n \). Certainly \( w \) cannot be fixed as \( \mathcal{M}_{n,w} \) then converges almost surely to a finite value and \( C_{n,w}(t,n) \) converges to zero. It is sufficient that \( w \sim cn \) for some constant \( c > 0 \), but if one scrutinizes the proofs they work if just \( w^{-1}(\log n^2)^3 \to 0 \) for \( n, w \to \infty \). A simulation study (not shown) reveals that for a concrete choice of \( f \) and \( P \) and with \( w = 200 \) the asymptotic results works quite well for \( n \) ranging from 400 to 15 millions, but for \( n \) larger than 50.000 the asymptotic probability \( \exp(-\exp(-x+x_n)) \) starts to underestimate the true probability \( P(\mathcal{M}_{n,w} \leq t,n) \) slightly.

A method somewhat similar to the windowed version of stack scoring as described above was developed for general structures by Hofacker et al. (2004) and implemented in a program called RNALfold. The difference is in the scoring methods chosen, and even though the method by Hofacker et al. (2004) provides local folding of a long RNA-sequence within a window, the reported findings are more dependent upon the chosen window size and the asymptotic behaviour of e.g. the maximal score reported by RNALfold differ in an essential way from the result obtained in Theorem 3.1. The differences and similarities will be treated more thoroughly elsewhere.

4 Proofs

We define the index set \( I_0 \) as follows
\[ I_0 = \left\{ (k,ql) \mid k \in \{0, \ldots, n\}, \ q \in \left\{ 0, \ldots, \left[ \frac{n}{l} \right] \right\}, \ n + 1 + 3l \leq k + ql \right\}. \]

As for alignments we introduce for \( t > 0 \) and \( a = (k,r) \in I_0 \) the stochastic variable
\[ V_a = V_a(t) = 1(\max_{1 \leq h \leq \Delta} \sum_{h=\delta}^\Delta f(X_{k+h}, Y_{r+h}) > t). \]

There is a very subtle reason for restricting our attention to indices \( a = (k,r) \) satisfying \( n + 1 + 3l \leq k + r \) and not just \( n + 1 \leq k + r \). It corresponds to ignoring a sub-diagonal strip of width \( 3l \) in the score matrix. A strip that is asymptotically
Figure 2: An example of the neighbourhood of strong dependence for the diagonal-within-a-strip, \(a = (k, r)\), in the case of folding a sequence to form stacks. The neighbourhood becomes slightly more complicated in this case compared to aligning two independent sequences, because the reuse of variables is more pronounced.

Negligible for \(n \to \infty\) under the condition that \(l \sim (\log n^2)^3\). The reason is to get rid of some very nasty dependencies close to the diagonal, which arise due to the folding of the sequence onto itself.

The neighbourhood of strong dependence also needs to be redefined for this setup. For \(a = (k, r)\) let

\[
B^1_a = \{k - l, \ldots, k + 2l\} \times \{0, l, 2l, \ldots, \lfloor \frac{n}{l} \rfloor l\} \cap I_0
\]

\[
B^2_a = \{0, \ldots, n\} \times \{r - l, r, r + l\} \cap I_0
\]

\[
B^3_a = \{0, \ldots, n\} \times \{n - k + 1 - l, n - k + 1, n - k + 1 + l\} \cap I_0
\]

\[
B^4_a = \{n - r + 1 - l, \ldots, n - r + 1 + 2l\} \times \{0, l, 2l, \ldots, \lfloor \frac{n}{l} \rfloor l\} \cap I_0
\]

and put \(B_a = B^1_a \cup B^2_a \cup B^3_a \cup B^4_a\). Intersecting with \(I_0\) in the definitions above is a restriction of the four sets to the triangular part of the score matrix we are interested in. On Figure 2 we see an example of the neighbourhood of strong dependence.

**Lemma 4.1.** If we, for some \(x \in \mathbb{R}\), let

\[
l = l_n \sim (\log n^2)^3 \quad \text{and} \quad t = t_n = \frac{\log K^* + \log(n(n - 1)/2) + x}{\theta^*}
\]
and define $x_n \in [0, \theta^*)$ by $x_n = \theta^*(t_n - \lfloor t_n \rfloor)$, then under the assumptions in Theorem 3.1, the conditions in (Hansen 2004a, Theorem 4.1) are fulfilled with

$$
\lambda_n = \exp(-x + x_n).
$$

That is

$$
\left| D \left( \sum_{a \in I_0} V_a \right) - \text{Poi}(\exp(-x + x_n)) \right| \to 0.
$$

**Proof:** We proceed as in the proof of Lemma 6.13 in (Hansen 2004a) adapting the arguments to the present situation when it is necessary.

First of all observe that $|I_0| \sim (n(n - 1)/2)^{-1}$. Since

$$
\mathbb{P}(V_a > t_n) = \mathbb{P}(V_a > t_n - x_n/\theta^*) \sim (n(n - 1)/2)^{-1}l \exp(-x + x_n)
$$

as follows from (Hansen 2004a, Lemma 6.12) (the convergence is uniform in $a$), we have that

$$
\sum_{a \in I_0} \mathbb{E}(V_a) \sim |I_0|((n(n - 1)/2)^{-1}l \exp(-x + x_n) \sim \exp(-x + x_n),
$$

or, since $\exp(-x + x_n)$ is bounded, $|\sum_{a \in I_0} \mathbb{E}(V_a) - \exp(-x + x_n)| \to 0$

Furthermore, $\mathbb{E}(V_a)\mathbb{E}(V_b) \sim (n(n - 1)/2)^{-2}l^2 \exp(-2(x + x_n))$ and $\max_a |B_a| = O(n) = o(|I_0|)$, so

$$
\sum_{a \in I, b \in B_a} \mathbb{E}(V_a)\mathbb{E}(V_b) \leq K|I_0|\max_a |B_a|((n(n - 1)/2)^{-2}l^2 \to 0
$$

for $n \to \infty$.

We divide the neighbourhood of strong dependence $B_a$ into five disjoint sets:

$$
C_a = B_a^1 \cap B_a^2, \quad D_a^1 = B_a^1 \setminus C_a, \quad D_a^2 = B_a^2 \setminus C_a
$$

and $B_a^3$ and $B_a^4$. By arguments identical to those in the proof of Lemma 6.13 in (Hansen 2004a), it follows that as $n \to \infty$

$$
\sum_{a \in I_0, b \in C_a, a \neq b} \mathbb{E}(V_aV_b) \to 0,
$$

and using that $\varphi_i(3\theta^*/4) < 1$ for $i = 1, 2$ also

$$
\sum_{a \in I_0, b \in D_a^i} \mathbb{E}(V_aV_b) \to 0
$$

for $i = 1, 2$. 

for $i = 1, 2$. To take care of $\mathbb{E}(V_aV_b)$ for $a \in I_0$ and $b \in B_i^a$ for $i = 3, 4$ we need a conclusion similar to Lemma 6.4 in (Hansen 2004a), which follows from the additional assumptions made that $\varphi_i(3\theta^*4) < 1$ for $i = 3, 4$. Hence for some $\varepsilon > 0$ the estimate

$$\mathbb{E}(V_aV_b) \leq Kl^4 \exp(-3/2\theta^*(1+\varepsilon)l)$$

holds for $a \in I_0$ and $b \in B^i_a$ for $i = 3, 4$. Therefore

$$\sum_{a \in I_0, b \in B^i_a} \mathbb{E}(V_aV_b) \leq Kl^4(n(n-1)/2)^{-3/2(1+\varepsilon)} \sum_{a \in I_0} |B^i_a|$$

$$\leq Kl^4(n(n-1)/2)^{-3/2(1+\varepsilon)} \times K_1(n(n-1)/2)^{3/2} \to 0.$$ 

Finally, we extend $(X_k)_{k \geq 1}$ to a doubly infinite, stationary Markov chain $(X_k)_{k \in \mathbb{Z}}$, which is exponentially $\beta$-mixing, thus

$$\beta(k) = \mathbb{E}\left(\sup_{A \in \mathcal{F}_{[-\infty,0]}} |\mathbb{P}(A|\mathcal{F}_{-\infty,0}) - \mathbb{P}(A)|\right) \leq K_1 \exp(-K_2k).$$

For $a = (k,r)$ we let $m = n - r + 1 + 3l \leq k$ and define $I_1 = (-\infty, m-l]$, $I_2 = [m+1, m+l]$, $I_3 = [m+2l+1, k-l]$, $I_4 = [k+1, k+l]$ and $I_5 = [k+2l+1, \infty)$. With $I = I_1 \cup I_3 \cup I_5$ and $J = I_2 \cup I_4$ we have that $\mathcal{F}_a \subseteq \mathcal{F}_I = \sigma(X_{n \mid n \in I})$ and $V_a$ is $\mathcal{F}_J$ measurable. Hence from (Hansen 2004a, Lemma 5.1 and Lemma 5.2) we obtain that

$$\mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \leq 2\alpha(\mathcal{F}_I, \mathcal{F}_J) \leq 14\beta(l+1) \leq K \exp(-K_2l).$$

This implies that

$$\sum_{a \in I_0} \mathbb{E}|\mathbb{E}(V_a|\mathcal{F}_a) - \mathbb{E}(V_a)| \leq K(n(n-1)/2)^2 \exp(-K_2(\log n^2)^3) \to 0$$

for $n \to \infty$. \qed

**Proof of Theorem 3.1** We need to show that

$$\mathbb{P}\left(\sum_{a \in I_0} V_a(t_n) \neq C(t_n)\right) \to 0.$$

The proof is a copy of the similar proof of display (56) in (Hansen 2004a) with some obvious modifications and a redefinition of $\mathcal{E}_n \subseteq \mathcal{E}_{n+l}$ to

$(i, j, \Delta) \in \mathcal{E}_n$ if $(i, j, \Delta) \in \mathcal{E}_{n+l}$ and either $i + j \leq n + 3l$, $i \geq n - l$, or $j \geq n - l$.

This does not change the conclusion and by Lemma 4.1 and by (Hansen 2004a, Corollary 4.2) we have proved Theorem 3.1. \qed
5 Concluding remarks

Although Xiong & Waterman (1997) conjectured results similar to Theorem 3.1 for structure scoring, it seems that the idea to use local structure scoring methods to search for structural features within longer sequences has not caught on. As showed here and by Xiong & Waterman (1997) there are many similarities between alignment and structure scores and considering the usefulness and importance of local alignment tools it is surprising that similar methods have not been developed when it comes to structures. The present paper provides a theoretical underpinning for such methods.

It should also be remarked that whereas Xiong & Waterman (1997) consider entire structures we consider only local stacks. This means that for a given stack

\[(i + 1, j - 1), (i + 2, j - 2), \ldots, (i + \Delta, j - \Delta)\]

the corresponding loop given by \(j, j + 1, \ldots, i - 1, i\) is penalised in (Xiong & Waterman 1997) by a penalty function that is linear in the loop size \(i - j\). As a result the asymptotic growth of the maximal score is like \(\log n\) and not as obtained in Theorem 3.1 like \(\log(n(n - 1)/2)\). It is ongoing work to handle structure scoring with general penalty functions of the loop size, but some results in this direction have already been obtained in (Hansen 2004b).

References


DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS
UNIVERSITETSPARKEN 5
DK-2100 COPENHAGEN Ø
DENMARK
richard@math.ku.dk