Simultaneous Occurrences of Runs in Independent Markov Chains

by

Stphane Robin and Valery Stefanov

Research Report No. 11
February 2008

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

Simultaneous Occurrences of Runs in Independent Markov Chains

S. Robin[1] and V. Stefanov[2]

([1]) UMR 518 AgroParisTech / INRA Appl. Math. & Comput. Sci.
16, rue Claude Bernard, 75005 Paris, FRANCE
robin@agroparistech.fr

([2]) School of Mathematics and Statistics
The University of Western Australia
Crawley (Perth) 6009, W.A., AUSTRALIA
stefanov@maths.uwa.edu.au

February 5, 2008

We observe m independent and identically distributed binary Markov chains and look for simultaneous occurrences of runs in several of them. We are interested in the distribution of the maximum number of simultaneous runs on finite time intervals. First we introduce a natural exact approach and also explain why it fails to calculate the required probabilities. Then we find exact upper and lower bounds for the probability of interest. We apply these results to detect genomic deletions in cancer patients.

Keywords: Comparative genomic hybridization, Exact bounds, Markov chains, Simultaneous occurrences, Runs

1 Introduction

1.1 Motivations

Occurrences of patterns or motifs on strings generated by Markov chains is now an old problem in applied probability. The simplest patterns are runs. Molecular biology is an area which provides a variety of problems on occurrences of patterns (see Robin *et al.* (2005)), in particular in DNA or protein sequence analyses. A huge literature exists on both exact and asymptotic results regarding waiting times, distance between occurrences, scans, counts, etc. (see Reinert *et al.* (2000) for a review in the biological context). The occurrence properties of complex patterns have also been studied (Stefanov *et al.* (2007), Nuel (2006)). Most of these works deal with one Markov chain at a time, i.e. with only one sequence of letters.

The problem of simultaneous occurrences in several sequences has been rarely addressed. On the other hand, the ever increasing amount of data provided by high throughput technologies makes it quite interesting. Comparative Genomic Hybridization (CGH) is one of these techniques that allows to detect losses or amplifications of the genetic

material in a given patient. Our work is motivated by the simultaneous analysis of the CGH profiles of a set of patients having the same disease, namely bladder cancer. These profiles can be viewed as first-order Markov chains on the 3-letter alphabet {'lost', 'normal', amplified' }, rewritten as $\{-1, 0, 1\}$. The loss (resp. amplification) of a large region is characterised by a succession of '-1' in the profile. The detection of region lost by a large proportion of patient will help in understanding which region are related to the development of the disease.

Standard techniques, such as Chen-Stein approximations can not be applied here for two reasons. The first one is that the conditions for the bounds to converge to 0 are not fulfilled in CGH profile, due to heavy diagonal terms in the transition matrix. The second and most important one is that, contrarily to DNA sequences, CGH profiles are not very long, so asymptotic results are not satisfying. We will typically deal with sequences of few hundreds of letters, observed in about one hundred patients and look for runs of few letters.

1.2 Model and Notations

We consider m independent binary (the two states are denoted by 0 and 1) Markov chains (MC) $\mathbf{X}_i = \{X_{i,t}\}_{t\geq 1}$ $(i = 1..m)$ with the same transition intensity matrix $\mathbf{\Pi}$ and stationary distribution $\boldsymbol{\mu}$. All MC's are supposed to be stationary, i.e., for all i, $X_{i,1} \sim \boldsymbol{\mu}$. State 1 is supposed to be a 'rare' state, that is $\mu_0 > \mu_1$.

Run occurrences. We are interested in occurrences of runs of 1's with length ℓ ('ℓ-runs') in the $\mathbf{X}_i$'s. For such runs we define the binary process $\mathbf{Y}_i^\ell = \{Y_{i,t}^\ell\}_{t\geq \ell}$ which equals 1 every time ℓ-run is completed in $\mathbf{X}_i$, that is

$$Y_{i,t}^\ell = \prod_{u=1}^{\ell} X_{i,t-u+1}.$$

Note that, according to this definition, $(\ell + 1)$ successive 1's achieve two ℓ-runs. For the sake of brevity, we shall drop the superscript ℓ in most of the following formulas.

Simultaneous occurrences of runs. The aim of this paper is to study the number of $\mathbf{X}_i$'s which comple a run at the same position. We therefore define the process $\mathbf{Y}_+ = \{Y_{+,t}\}_{t\geq \ell}$ counting simultaneous run occurrences:

$$Y_{+,t} = \sum_{i=1}^{m} Y_{i,t}.$$

1.3 Problem

Maximum number of simultaneous runs. We are interested in the maximum value of $\mathbf{Y}_+$ when observed between positions 1 and n. Typically, for a given threshold M^*, we want to evaluate

$$\Pr\left\{\max_{\ell\leq t\leq n} Y_{+t} \geq M^*\right\}. \tag{1}$$

Outline. In Section 2, we present a general method that gives the exact value of this probability but leads to intractable computations in practice. We then use part of the exact results to derive a first upper bound (Section 3). In Section 4 we derive a lower bound of the probability of interest, based on a different srtrategy.The results are applied to CGH array data in Section 5. In the last section, we discuss the possible generalization to other motifs.

2 Exact Results

2.1 General case

The exact distribution of $\mathbf{Y}_+$ can be determined theoretically in two ways but these do not lead to evaluation of the probabilities of interest. The first way is based on considering $\mathbf{Y}_+$ as an homogenous Markov chain of order ℓ with state space $\{0, \dots m\}$. The second one, which is equivalent to the first, deals with a first order Markov chain with larger state space. To mark occurrences of runs in each $\mathbf{X}_i$, we have to to consider the following $\ell + 1$ quantities:

$$\begin{aligned} M_{0,t} &= \textstyle\sum_i (1 - X_{it}), \\ M_{s,t} &= \displaystyle\sum_i (1 - X_{i,t-s}) \prod_{u=1}^{s} X_{i,t-u+1}, \quad \text{for } 1 \leq s \leq \ell - 1, \\ M_{\ell,t} &= \displaystyle\sum_i \prod_{u=1}^{\ell} X_{i,t-u+1}. \end{aligned}$$

$M_{0,t}$ counts the number of $\mathbf{X}_i$'s in state 0 at position t. $M_{s,t}$ counts the number of the $\mathbf{X}_i$ in which a run of exactly s 1's (not more) is ending at position t. $M_{\ell,t}$ counts the number of the $\mathbf{X}_i$'s in which an ℓ-run has been achieved before (or at) position t and has not been interrupted up to position t. These counts summarize the memory of the process $\mathbf{Y}_+$.

The process $\{M_t\}_t$ where $M_t = (M_{0,t}, ...M_{\ell,t})$, is a first order Markov chain. The conditional distribution of M_{t+1} given M_t is a simple combination of binomial distributions that is straightforward to write. On the other hand the number of states of this chain is huge for the cases of interest and probability computations are not viable. More precisely, the number of states is equal to the number of possible repartitions of m objects into $\ell+1$ (possibly empty) sets. This is related to the Stirling number of second kind. In our case, for $m = 100$ and $\ell = 10$, this number is about 10^{93}, so the calculation of exact results seems out of reach.

2.2 Runs of length one

The case of 1-runs is trivial but useful for treating ℓ-runs for $\ell > 1$. Actually, the bound introduced in Section 4 is based on properties of 1-runs. Consider the process $\mathbf{X}_+ = \{X_{+t}\}_t$ where $X_{+t} = \sum_i X_{i,t}$. Also X_{+t} counts the number of the $\mathbf{X}_i$'s that are in state 1 at position t; that is $\mathbf{X}_+$ is equal to $\mathbf{Y}^1_+$.

Lemma 1 *The process $\mathbf{X}_+$ is a homogeneous first order Markov chain. The conditional distribution of $X_{+,t+1}$, given X_{+t}, equals the sum of two independent binomial variables $\mathcal{B}(m - X_{+t}, \pi_{01})$ and $\mathcal{B}(X_{+t}, \pi_{11})$, respectively.*

Proof. The process $\mathbf{X}_i$ is in state 1 at position $t+1$ through either moving from state 0 at position t (with probability π_{01}), or moving from state 1 at position t (with probability π_{11}). The result follows, recalling that the $\mathbf{X}_i$'s are independent and that there are exactly X_{+t} of the $\mathbf{X}_i$ in state 1 at position t, and the remaining $m - X_{+t}$ are in state 0. ■

In the following, $\mathbf{\Gamma}$ denotes the one-step transition probability matrix of $\mathbf{X}_+$, and $\gamma(x, y)$ is its (x, y)-entry. According to Lemma 1, we have

$$\gamma(x, y) = \sum_{u=0}^{y} b(u; m - x, \pi_{01}) b(y - u; x, \pi_{11}).$$

with $b(x; n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$.

3 Upper Bound

In the perspective of calculating p-values, we are interested in an upper bound of the probability (1).

Proposition 1 *Let*

(i) $\mathbf{Y}^*$ *be a Markov chain over the state space with* $[(\ell - 1)m - \ell M^* + \ell]$ *dimensions:*

$$\underbrace{0, 1, \ldots, M^* - 1}_{\textit{lower states}}, \underbrace{[M^*]_1, \ldots, [m]_1}_{\textit{1-st excess}}, \underbrace{[M^*]_2, \ldots, [m]_2}_{\textit{2-nd excess}}, \ldots, \underbrace{[M^*]_{\ell-1}, \ldots, [m]_{\ell-1}}_{\textit{($\ell-1$)-th excess}}, [m]_\infty;$$

(ii) $\mathbf{\Gamma}^*$ *be the transition matrix on this space state with all zeros elements except for:*

$$\begin{aligned}
\gamma^*(y, y') &= \gamma(y, y'), \\
\gamma^*(y, [u]_1) &= \gamma(y, u), \\
\gamma^*([u]_k, [u']_{k+1}) &= \gamma(u, u'), \\
\gamma^*([u]_k, y') &= \gamma(u, y'), \\
\gamma^*([u]_{\ell-1}, [m]_\infty) &= \sum_{u'=0}^{r^*} \gamma(u, u'),
\end{aligned}$$

for $0 \le y, y' < M^*$ *and* $M^* \le u, u' \le m$*;*

(iii) $\boldsymbol{\mu}_1^*$ *be a* $[(\ell-1)m - \ell M^* + \ell]$*-dimensional row vector with all coordinates zero except for* $\mu_{1,y}^* = b(y; m, \mu_1)$ *for* $0 \le y \le m$*.*

Then we have

$$\Pr\left\{ \max_{\ell \le t \le n} Y_{+t} \ge M^* \right\} \le \boldsymbol{\mu}_1^* (\mathbf{\Gamma}^*)^{n-1}.$$

Proof. To observe M^* simultaneous ℓ-runs, we need at least M^* of the $\mathbf{X}_i$'s to be in state 1 in the last ℓ positions, that is for any t,

$$\Pr\{Y_{+t} \geq M^*\} \leq \Pr\{X_{+,t-\ell+1}, \dots X_{+,t-1}, X_{+,t} \geq M^*\}.$$

Note that, in the right-hand term, the $\mathbf{X}_i$'s are not required to be the same along the ℓ positions, while they are in the left-hand term. When considering the maximum of Y_{+t}, we get an upper bound for (1):

$$\Pr\left\{\max_{\ell \leq t \leq n} Y_{+t} \geq M^*\right\} \leq \Pr\{\exists t \in \{\ell, \dots, n\} : X_{+,t-\ell+1}, \dots X_{+,t-1}, X_{+,t} \geq M^*\}. \quad (2)$$

The calculation of the right-hand term of (2) requires to follow-up the excesses of $\mathbf{X}_+$ above M^*. We denote by ξ_t the excess indicator: $\xi_t = \mathbb{I}\{X_{+t} \geq M^*\}$ and define the Markov chain $\mathbf{Y}^* = \{Y_t^*\}$:

$$\begin{array}{rcll} Y_t^* & = & X_{+t} & \text{if } X_{+t} < M^*, \\ & = & [X_{+t}]_s & \text{if } (1-\xi_{t-s}) \prod_{v=1}^{s} \xi_{t-s+1} = 1 \\ & = & [m]_\infty & \text{if } X_{+t} \geq M^* \text{ and } \prod_{v=1}^{\ell} \xi_{t-\ell+1} = 1 \end{array}$$

The state space of $\mathbf{Y}^*$ is the one given in i above. Since the $\mathbf{X}_i$'s are independent and stationary, Y_1^* has a binomial distribution $\mathcal{B}(m, \mu_1)$. The vector $\boldsymbol{\mu}_1^*$ given in iii above provides this distribution.

The process $\mathbf{Y}^*$ is a Markov chain whose transition matrix $\mathbf{\Gamma}^*$ is given in ii: Y_t^* is in state $M^* + u_s$ if X_{+t} equals $M^* + u$ and if $X_{+,t-k+1}, \dots X_{+,t-1}$ (but not $X_{+,t-k}$) already exceeded M^*. X_{+t} constitutes then the kth successive excess above M^*. State $[m]_\infty$ is reached when ℓ successive excesses are observed. So, the upper bound (2) is the probability for $\mathbf{Y}^*$ to reach the absorbing state $[m]_\infty$ before the end (position n), and the result follows. ■

4 Lower Bound

We are not able to derive any theoretical measure of the quality of the upper bound given in the preceeding section. Therefore we propose here a lower bound of (1). This bound is based on exact evaluation assuming that the maximum of $\mathbf{Y}_+^\ell$ is reached exactly $\ell - 1$ positions after $\mathbf{X}_+$ has reached its own maximum.

Proposition 2 *Denote by $B(x; \nu, p)$ the probability for a binomial $\mathcal{B}(\nu, p)$ random variable to be larger than or equal to x. Then*

$$\Pr\left\{\max_{\ell \leq t \leq n} Y_{+t} \geq M^*\right\} \geq \sum_{u=M^*}^{m} \Pr\left\{\max_{1 \leq t \leq n-\ell+1} X_{+t} = u\right\} B(M^*; u, \pi_{11}^{\ell-1}). \quad (3)$$

Proof. To observe M^* ℓ-runs at position t, we first need to observe u ones ($u \geq M^*$) at position $t - \ell + 1$, and then to complete at least M^* runs amongst these u sequences. Since the probability for a 1 to give birth to an ℓ-run is $\pi_{11}^{\ell-1}$, we have (for $\ell \leq t \leq n$)

$$\Pr\{Y_{+t} \geq M^*\} = \sum_{u=M^*}^{m} \Pr\{X_{+,t-\ell+1} = u\} B(M^*; u, \pi_{11}^{\ell-1}).$$

We get a lower bound of $\Pr\{\max Y_{+T} = M^*\}$, applying a constraint on the relative positions of the maximum of $\mathbf{X}_+$ and $\mathbf{Y}_+$. Denoting $T^\ell = \arg\max_{\ell \leq t \leq n} Y_{+t}$ and $T^1 = \arg\max_{1 \leq t \leq n-\ell+1} X_{+t}$, we have

$$\Pr\left\{\max_{\ell \leq t \leq n} Y_{+t} \geq M^*\right\} \geq \Pr\left\{\left(\max_{\ell \leq t \leq n} Y_{+t} \geq M^*\right) \cap \left(T^\ell = T^1 + \ell - 1\right)\right\}.$$

Since the positions of the two maximums are linked, the maximum of Y_{+t} in the right-hand term can be replaced by a maximum of X_{+t} and the result follows. ■

5 Application

5.1 CGH profiles

CGH technology. Chromosomal aberration, i.e. deletions or amplifications of genomic regions are associated with many diseases such as cancer or mental retardation. Comparative Genomic Hybridization (CGH) experiments aim at detecting and mapping chromosomal imbalances, via the hybridization of targets of genomic DNA between a test and a reference genome. After some post-processing (see Hupe *et al.* (2004)), a status is associated to each position: deletion / normal / amplification. Such a sequence of status is called a CGH profile. In large studies, the profiles of many patient having the same disease are collected in order to detect region commonly deleted or amplified Rouveirol *et al.* (2006)).

Data. We deal here with a set of $m = 84$ patients ($i = 1..m$) with bladder cancer followed at Institut Curie in Paris (France). Each profile $\mathbf{X}_i$ is made of $n = 2360$ positions spread along the 24 chromosomes (22 non-sexual + X + Y chromosomes). We are mainly concern with losses of genetic material, so we consider binary profiles where 0 stand for normal (or amplified) and 1 for deletion. The estimated transition matrix and stationary distribution are (in %)

$$\mathbf{\Pi} = \begin{pmatrix} 99.72 & 0.28 \\ 2.26 & 97.74 \end{pmatrix}, \qquad \boldsymbol{\mu} = (88.98 \quad 11.08).$$

Biological question. Among these profiles, some commonly deleted regions appear, for example, $M^* = 18$ patients present a succession of $\ell = 22$ deletions ending at position $t = 1340$ (chromosome 10). Another run of $\ell = 9$ deletions is observed in $M^* = 11$ patients at position $t = 1191$ (chromosome 8). We want to assess the significance of such regions.

Handling several chromosomes. As explained above, the signal is spread along $K = 24$ chromosomes. Although it seems reasonable to assume that the sequence of status is a Markov chain, transition from one chromosome to another are meaningless. Hence the probability (1) has to be separated into terms corresponding to each chromosome. Note that, in the shortest chromosome, only few tens of positions are observed. This motivates our choice not to consider asymptotic results for this application.

Let n_k denote the number of positions in chromosome k ($k = 1..K, \sum_k n_k = n$), $\mathbf{Y}$ the complete process and $\mathbf{Y}^k$ its reduction to chromosome k. Assuming that independence between the chromosomes, we have

$$\Pr\left\{\max_{\ell \leq t \leq n} Y_{+t} \geq M^*\right\} = 1 - \prod_k \left(1 - \Pr\left\{\max_{\ell \leq t \leq n_k} Y^k_{+t} \geq M^*\right\}\right).$$

Plugging a lower (resp. upper) bound in the right-hand terms provides a lower (resp. upper) bound of the left-hand term.

5.2 Results

Table 1 displays the results for deleted regions observed in the data. The significance is measured by the p-values. We see that it mainly depends on the number of patients M^* in which the run is observed, more than on the length ℓ of the run. This is due to the high value of π_{11} (97.74%): as soon as a ℓ-run is observed in a large number of patients, a $(\ell+1)$-run is likely to be observed in the same patients.

t^*	chrom.	ℓ	M^*	$\Pr\{X_+ \geq M^*\}$	p(upper)	$\Pr\{\max X_+ \geq M^*\}$	p(lower)
1340	10	22	18	8.57 e–8	5.75 e–9	6.69 e–5	1.24 e–11
2347	24	13	17	4.62 e–7	1.80 e–6	3.73 e–4	1.21 e–7
320	2	3	16	2.32 e–6	6.85 e–4	1.93 e–3	5.93 e–4
1161	9	2	16	2.32 e–6	1.13 e–3	1.95 e–3	1.09 e–3
1387	11	3	15	1.08 e–5	3.20 e–3	8.25 e–3	2.80 e–3
1152	9	2	15	1.08 e–5	5.07 e–3	8.33 e–3	4.87 e–3
996	8	7	13	1.88 e–4	1.42 e–2	1.05 e–1	7.15 e–3
1413	11	2	13	1.88 e–4	7.52 e–2	1.11 e–1	7.28 e–2
1690	13	2	13	1.88 e–4	7.52 e–2	1.11 e–1	7.28 e–2
1430	11	3	12	6.90 e–4	1.72 e–1	3.13 e–1	1.57 e–1
1688	13	2	12	6.90 e–4	2.32 e–1	3.15 e–1	2.26 e–1
1455	11	5	11	2.32 e–3	2.94 e–1	6.56 e–1	2.28 e–1
1187	9	2	11	2.32 e–3	5.53 e–1	6.67 e–1	5.41 e–1
1880	16	2	10	7.12 e–3	8.88 e–1	9.43 e–1	8.80 e–1
584	4	6	9	1.98 e–2	9.09 e–1	9.98 e–1	8.12 e–1
2072	18	3	9	1.98 e–2	9.87 e–1	9.99 e–1	9.80 e–1
1696	13	2	9	1.98 e–2	9.95 e–1	9.99 e–1	9.94 e–1

Table 1: Significance of observed runs of deletions. $p(\cdot)$: upper and lower bounds of the p-value $\Pr\{\max_t Y_{+t} \geq M^*\}$.

For most regions, the upper and lower bounds derived in Sections 3 and 4 are close. The relative difference between them is about few percents. The difference is large for long regions. At this time, we have no argument to decide which bound is the most affected when ℓ increases. In a statistical perspective, the significance is assessed by the upper bound. For long runs, the results tend to be conservative.

Several non-overlapping short regions are detected in chromosome 9; for two of them, the upper bound of the p-value is smaller than 1 percent. This is consistent with previously

available knowledge: deletions in chromosome 9 are known to be associated with the development of bladder cancer. Further investigations are leaded at this time for the other regions. Very long regions ($\ell = 13$ or 22) may correspond to complete losses of one chromosome arm.

6 Discussion

We addressed here the problem of simultaneous runs in independent Markov chains. The problem could be generalized to any motif, i.e. any small sequence $\mathbf{w} = (w_1, \ldots, w_\ell) \in \{0,1\}^\ell$. The method we proposed in Section 2 could probably be generalized to any motif using the FMCE technique, but the size of the state space would still increase. Again, this approach is likely to fail in practice.

The upper bound proposed in Section 3 could be generalized in the following way. Denoting now Y_{+t} the number of processes $\mathbf{X}_i$ in which an occurrence of $\mathbf{w}$ ends at position t, the probability $\Pr\{Y_{+t} \geq M^*\}$ is smaller than

$$\Pr\left\{ \bigcap_{u:w_u=1} (X_{+,t-\ell+u} \geq M^*) \bigcap_{u:w_u=0} (m - X_{+,t-\ell+u} \geq M^*) \right\}.$$

However, this approach works well in the case of runs of 1, 1 being the rare state. For a general motif, the quality of this bound will be reduced in case of numerous 0 in $\mathbf{w}$.

The lower bound proposed in Section 4 may also be generalized to other motifs. Its quality will mainly depends on the first state. If the motif start with a frequent state (e.g. 0), the constraint on the relative positions of the maximum of $\mathbf{X}_+$ and $\mathbf{Y}_+$ will turned to be very restrictive and the bound is expected to be poor. For motifs starting with a rare state (e.g. 1), the quality should be comparable with the case studied in this paper.

Acknowledgments. We are grateful to F. Radvanyi (Institute Curie) and G. Rouveirol (Univ. Paris XI, Orsay) who provided the problem and the relevant CGH data. They also helped in the interpretation of the results. We also held helpful discussions with S. Schbath (INRA-MIG).

References

Hupe, P., Stransky, N., Thiery, J., F.Radvanyi and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*. **20 (18)** 3413–3422.

Nuel, G. (2006). Effective p-value computations using finite markov chain imbedding (fmci): application to local score and to pattern statistic. *Algo. Mol. Biol.* **1 (5)** 1–14. doi:10.1186/1748-7188-1-5.

Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words. *J. Comp. Biol.* **7** 1–46.

Robin, S., Rodolphe, F. and Schbath, S. (2005). *DNA, words and models*. Cambridge University Press.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E. and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*. **22 (7)** 849–856.

Stefanov, V., Robin, S. and Schbath, S. (2007). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Appl. Math.* **(155)** 868–80.