The Annals of Statistics
2003, Vol. 31, No. 3, 1027–1031

CORRECTION

APPROXIMATE *p*-VALUES FOR LOCAL SEQUENCE ALIGNMENTS[1]

BY DAVID SIEGMUND AND BENJAMIN YAKIR

The Annals of Statistics (2000) **28** 657–680

1. Introduction. It has been pointed out to us by D. Metzler (University of Frankfurt) that the proof of Theorem 3 of Siegmund and Yakir (2000) is incomplete. In addition, S. Grossman (Frankfurt) and A. Dembo (Stanford) have observed that some conditions are required in order for the proofs of Theorems 1 and 2 (in particular the proof of Lemma 1) to hold. In this note we give appropriate additional conditions and complete the proof of Theorem 3. We use the notation of our earlier paper.

To describe the conditions for Theorems 1 and 2, we let Q_0 denote the "null" probability given by $Q_0(\alpha, \beta) = \mu(\alpha)\nu(\beta)$ and let $Q_1(\alpha, \beta) = \exp[\theta^* K(\alpha, \beta)] \times Q_0(\alpha, \beta)$ denote the implied "alternative." Also let $Q_{i,j}$ be defined to be the product probability that gives x the marginal distribution it has under Q_i and y the marginal distribution it has under Q_j. We assume that

$$\mathrm{E}_1 K(x, y) - \mathrm{E}_{i,j} K(x, y) > 0 \tag{1}$$

for all i, j. This assumption will legitimize the application of a large deviation bound for additive functionals of finite state Markov chains in the proof of Lemma 1, since the total length of all gaps is small compared to the number of aligned pairs and hence essentially all terms forming the additive functionals have negative expectation. (However, the alternative suggestion to apply the Azuma–Hoeffding inequality does not work.)

For Theorem 3 a convenient condition will involve computations that build on the parameter θ^*. Thus, for example, we define

$$\psi_y(\theta, \eta) = \log \mathrm{E}_0\big[\exp\{\theta K(x, y_1) + \eta K(x, y_2)\}\big],$$

with y_1, y_2 independent copies of y. Note that θ^* is the unique positive solution of the equation

$$\psi_y(\theta, 0) = 0.$$

Under the additional assumption that

$$\mathrm{E}_{1,0} K(x, y) = \mathrm{E}_0\big[\exp\{\theta^* K(x, y_1)\} K(x, y_2)\big] < 0,$$

Received September 2002.
[1]Supported by the NSF and by the Israel–U.S. Binational Science Foundation.

there exists a unique positive η such that

$$\psi_y(\theta^*, \eta) = 0.$$

Denote by θ_y^* the unique positive solution of this equation. In a similar way one can define ψ_x (based on x_1, x_2 and y) and hence θ_x^*—the unique positive solution of $\psi_x(\theta^*, \eta) = 0$.

The conditions we will use involve the relation between m and n, the lengths of the x and the y sequences respectively, and between θ^*, θ_x^* and θ_y^*.

THEOREM 3. *Suppose the conditions of Theorem* 2 *hold, but that* $mn\exp(-a)$ *converges to a finite, positive limit. Assume that the parameters* θ_y^*, θ_x^* *defined above exist and that*

$$\limsup_{m,a\to\infty} \frac{\log m}{a} < \min\left(1, \frac{\theta_x^*}{\theta^*}\right) \quad \textit{and} \quad \limsup_{n,a\to\infty} \frac{\log n}{a} < \min\left(1, \frac{\theta_y^*}{\theta^*}\right).$$

Let Q *denote the right-hand side of display* (4) *of Siegmund and Yakir* (2000). *Then*

$$P_0\left(\max_{\mathbf{z}\in\mathcal{Z}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\right) - [1 - \exp(-Q)] \to 0$$

as $a \to \infty$.

2. A proof of the theorem. To prove Theorem 3, suppose that $mne^{-a} \to x$. Recall that $\mathcal{Z}_j \subset \mathcal{Z}$ is the collection of all $\mathbf{z} \in \mathcal{Z}$ for which the k aligned pairs satisfy $(1 - \varepsilon_1)a/I < k < (1 + \varepsilon_1)a/I$, the number of gaps is *exactly* j, and the overall number of unaligned letters, l, is bounded by $\varepsilon_2 a^{1/2}$ for some small $\varepsilon_2 > \varepsilon_1 > 0$. Let $\tilde{\mathcal{Z}} = \bigcup_{j\le j_1} \mathcal{Z}_j$, where j_1 is a large but fixed integer. By essentially trivial modifications in the proofs of Lemmas 9–12 in the Appendix, we see that for all sufficiently large j_1, $\mathrm{P}(\bigcup_{\mathbf{z}\in\mathcal{Z}\setminus\tilde{\mathcal{Z}}}\{\ell_{\mathbf{z}} - g(\mathbf{z}) \geq a\}) \leq \varepsilon$, so we can confine our attention to the set $\tilde{\mathcal{Z}}$. Note that the elements of this set, represented as paths in a two-dimensional grid, are of restricted dimensions: a path that begins at the point (i_1, i_2) is contained in the square $\{(i_1, i_2), (i_1 + ca, i_2), (i_1 + ca, i_2 + ca), (i_1, i_2 + ca)\}$, where $c = I^{-1}(1 + \varepsilon_1) + \varepsilon_2$.

For a candidate alignment in $\tilde{\mathcal{Z}}$, the dot matrix representation between the two sequences is contained in a rectangle of size $m \times n$ in the two-dimensional lattice, which can be sub-divided into squares of size $a^2 \times a^2$. Let α be the index of a typical square and denote by $\mathcal{Z}_\alpha$ the set of candidate alignments which intersect with the square α. Note that though the $\mathcal{Z}_\alpha$'s are not disjoint, $\mathcal{Z}_\alpha$ can have common elements with $\mathcal{Z}_\beta$ only if $\mathcal{Z}_\beta$ is an immediate neighbor. Denote by B_α the neighborhood of dependence of α. (The neighborhood of dependence contains α, its 8 immediate neighbors, the strip of width of 3 squares that runs parallel to the x sequence, and the perpendicular strip that runs parallel to the y sequence. Elements in the first strip can have common y variables with elements in $\mathcal{Z}_\alpha$.

Likewise, elements in the second strip can have common x variables with elements in $\mathcal{Z}_\alpha$.) Let X_α be the indicator of the event $\{\max_{\mathbf{z}\in\mathcal{Z}_\alpha}[\ell_{\mathbf{z}} - g(\mathbf{z})] \geq a\}$. It follows that X_α and X_β are independent, provided that $\beta \notin B_\alpha$.

Consider $W = \sum_\alpha X_\alpha$ and note that $\mathrm{E}_0 W \to x\lambda$ as $a \to \infty$. According to Arratia, Goldstein and Gordon (1989), the difference between the probability of the event $\{W > 0\} = \{\max_{\mathbf{z}\in\tilde{\mathcal{Z}}}(\ell_{\mathbf{z}} - g(\mathbf{z}) \geq a\}$ and the quantity $1 - \exp(-\mathrm{E}_0 W)$ is bounded by $2(b_1 + b_2)$, where

$$b_1 = \frac{mn}{a^4}\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{Z}_\alpha}\big(\ell_{\mathbf{z}} - g(\mathbf{z})\big) \geq a\Big)^2,$$

$$b_2 = \frac{mn}{a^4}\sum_{\beta\in B_\alpha\setminus\{\alpha\}}\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{Z}_\alpha}\big(\ell_{\mathbf{z}} - g(\mathbf{z})\big) \geq a, \max_{\mathbf{u}\in\mathcal{Z}_\beta}\big(\ell_{\mathbf{u}} - g(\mathbf{u})\big) \geq a\Big).$$

From Theorem 2 we see that $b_1 = O(a^4 e^{-a})$.

For the following calculation we assume that $\theta_x^* \leq \theta^*$ and $\theta_y^* \leq \theta^*$. The other cases are treated similarly and are slightly simpler. In order to bound b_2, consider first some β which is part of the strip parallel to the y sequence and is not in the immediate neighborhood of α. It follows from the argument given in Section 3 below that

$$\begin{aligned}
&\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{Z}_\alpha}\big(\ell_{\mathbf{z}} - g(\mathbf{z})\big) \geq a, \max_{\mathbf{u}\in\mathcal{Z}_\beta}\big(\ell_{\mathbf{u}} - g(\mathbf{u})\big) \geq a\Big)\\
&\quad\leq \sum_{\mathbf{z}\in\mathcal{Z}_\alpha}\sum_{\mathbf{u}\in\mathcal{Z}_\beta}\mathrm{P}_0\big(\ell_{\mathbf{z}} - g(\mathbf{z}) \geq a, \ell_{\mathbf{u}} - g(\mathbf{u}) \geq a\big)\\
&\quad\leq 2\sum_{\mathbf{z}\in\mathcal{Z}_\alpha}\sum_{\mathbf{u}\in\mathcal{Z}_\beta}\exp\{-a - g(\mathbf{z})\}\exp\big\{-(\theta_y^*/\theta^*)a - (\theta_y^*/\theta^*)g(\mathbf{u})\big\}\\
&\quad\leq 2a^d e^{-a-(\theta_y^*/\theta^*)a},
\end{aligned} \tag{2}$$

for some finite constant d.

Consider next a β that is an immediate neighbor of α. Abusing notation, let $\mathcal{Z}_{\alpha\cap\beta}$ be the set of patterns which intersect with a strip of width $2ca$ on the boundary between α and β. Note that

$$\begin{aligned}
&\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{Z}_\alpha}\big(\ell_{\mathbf{z}} - g(\mathbf{z})\big) \geq a, \max_{\mathbf{u}\in\mathcal{Z}_\beta}\big(\ell_{\mathbf{u}} - g(\mathbf{u})\big) \geq a\Big)\\
&\quad\leq \mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{Z}_\alpha\setminus\mathcal{Z}_{\alpha\cap\beta}}\big(\ell_{\mathbf{z}} - g(\mathbf{z})\big) \geq a, \max_{\mathbf{u}\in\mathcal{Z}_\beta\setminus\mathcal{Z}_{\alpha\cap\beta}}\big(\ell_{\mathbf{u}} - g(\mathbf{u})\big) \geq a\Big)\\
&\qquad + 3\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{Z}_{\alpha\cap\beta}}\big(\ell_{\mathbf{z}} - g(\mathbf{z})\big) \geq a\Big).
\end{aligned}$$

The first term is again bounded by $2a^d e^{-a-(\theta_y^*/\theta^*)a}$, whereas the second term is asymptotically proportional to $a^3 e^{-a}$.

Equivalent derivation can be conducted for the strip parallel to the x sequence. Collecting the terms we get the inequality

$$b_2 \le \frac{mn}{a^4}\left[27a^3e^{-a} + \frac{6m}{a^2}a^d e^{-a-(\theta_x^*/\theta^*)a} + \frac{6n}{a^2}a^d e^{-a-(\theta_y^*/\theta^*)a}\right],$$

which converges to zero as $a \to \infty$.

3. Large deviations in two dimensions. The following observation justifies inequality (2) used above. Assume $\mathbf{u}$ and $\mathbf{z}$ share common x variables but no y variables. Let $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, where $\mathbf{u}_1$ involves the x terms that $\mathbf{u}$ shares with $\mathbf{z}$ and $\mathbf{u}_2$ does not. Recall that we have assumed $\theta_y^* \le \theta^*$. Note that both $\ell_{\mathbf{z}} + \sum_{(i,j)\in\mathbf{u}_1} \theta_y^* K(x_i, y_j)$ and $\ell_{\mathbf{z}} + \sum_{(i,j)\in\mathbf{u}_1} \theta_y^* K(x_i, y_j) + \sum_{(i,j)\in\mathbf{u}_2} \theta^* K(x_i, y_j)$ are log-likelihood ratios. Hence, by considering the disjoint possibilities that $\ell_{\mathbf{u}_2}$ is positive or negative, we find that

$$\begin{aligned}
&\mathrm{P}_0(\ell_{\mathbf{z}} \ge s, \ell_{\mathbf{u}} \ge t)\\
&\quad\le \mathrm{P}_0\left(\ell_{\mathbf{z}} + \sum_{(i,j)\in\mathbf{u}_1} \theta_y^* K(x_i, y_j) + \sum_{(i,j)\in\mathbf{u}_2} \theta^* K(x_i, y_j) \ge s + (\theta_y^*/\theta^*)t\right)\\
&\qquad + \mathrm{P}_0\left(\ell_{\mathbf{z}} + \sum_{(i,j)\in\mathbf{u}_1} \theta_y^* K(x_i, y_j) \ge s + (\theta_y^*/\theta^*)t\right)\\
&\quad\le 2\exp\{-s - (\theta_y^*/\theta^*)t\}.
\end{aligned}$$

4. Verification of the condition and remarks. The condition (1) requires, quite reasonably, that the change of measure from Q_0 to Q_1 increase the dependence between x and y, and not simply change the marginal distributions. In particular, the function $K(\alpha, \beta) = f(\alpha) + g(\beta)$, for which Q_1 is also a product measure, is excluded.

The condition of Theorem 3 is somewhat clumsy, but we are unable to identify a simpler general condition that implies it. The following argument suggests that the condition is often satisfied by reasonable scoring matrices $K(\alpha, \beta)$, which inter alia take on positive values on the diagonal and negative values, or at least smaller positive values, off the diagonal.

According to our model the null distribution that the two sequences are unrelated is the product measure Q_0. The implied alternative is Q_1. Large false positive scores behave empirically as if they come from Q_1; and power to detect a genuine alignment will be maximized if the scoring matrix is chosen so that Q_1 is the true joint distribution. The marginals of Q_1 should be similar to those of Q_0, namely μ and ν; otherwise a high score may occur because of differences in marginal frequencies without the clustering along the diagonal that is the goal of alignment. If the marginals of Q_1 are in fact μ and ν, then $\theta_x^* = \theta_y^* = \theta^*$, and the condition of Theorem 3 is easily satisfied for essentially any growth rates

of m and n. For a numerical example, for the BLOSUM62 scoring matrix and the marginal distribution of amino acid frequencies determined by Robinson and Robinson (1991), $\theta^* = 0.318$, while $\theta_x^* = \theta_y^* = 0.316$. Very similar results hold for the BLOSUM50 and PAM250 scoring matrices. [See Storey and Siegmund (2001) for a more detailed numerical examination of these three cases.]

For an artificial example of no biological significance, where one can carry out all calculations analytically, suppose that the null model has (x, y) bivariate standard normal and $K(\alpha, \beta) = c - (\alpha - \beta)^2/2$, where $0 < c < 1$. It is easy to see that the marginal distributions of Q_1 are normal with mean 0 and variance $(1 + \theta^*)/(1 + 2\theta^*) < 1$. Some straightforward calculations show that $\mathrm{E}_{i,j} K(x, y) < 0$ for all i, j, so both (1) and the new condition for Theorem 3 are satisfied.

Acknowledgment. The authors thank Niels Hansen for his careful reading of an earlier draft of this manuscript and several helpful suggestions, in particular a correction to our argument in Section 3.

REFERENCES

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximation: The Chen–Stein method. *Ann. Probab.* **17** 9–25.

ROBINSON, A. B. and ROBINSON, L. R. (1991). Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci. USA* **88** 8880–8884.

SIEGMUND, D. and YAKIR, B. (2000). Approximate p-values for local sequence alignments. *Ann. Statist.* **28** 657–680.

STOREY, J. and SIEGMUND, D. (2001). Approximate p-values for local sequence alignments: Numerical studies. *J. Computational Biology* **8** 549–556.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: dos@stat.stanford.edu

DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM
ISRAEL
E-MAIL: msby@mscc.huji.ac.il