Ann. I. H. Poincaré – PR 38, 6 (2002) 1087–1092

S0246-0203(02)01136-6/FLA

AN ASYMPTOTIC TEST FOR QUANTITATIVE GENE DETECTION

UN TEST ASYMPTOTIQUE POUR LA DETECTION DE GÈNES QUANTITATIFS

Jean-Marc AZAÏS [a,*], **Christine CIERCO-AYROLLES** [a,b]

[a] *Laboratoire de statistique et probabilités, U.M.R. C5583, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 4, France*

[b] *Institut national de la recherche agronomique, unité de biométrie et intelligence artificielle, B.P. 27, chemin de Borde-Rouge, 31326 Castanet-Tolosan cedex, France*

Received 10 April 2001, revised 13 May 2002

ABSTRACT. – The problem of detecting the presence of a quantitative gene using a great number of markers in a backcross genetic scheme is addressed.

An asymptotic test based on the maximum of a differentiable stochastic process is constructed. Bounds for threshold and power calculation are presented. Simulations and numerical experiments illustrate the convergence towards the asymptotic distribution and the sharpness of the bounds.

MSC: 60G15; 62F05

Keywords: Genetic markers; Haldane mapping function; Ornstein–Uhlenbeck process; QTL; Rice formulae

RÉSUMÉ. – Nous étudions le problème de la détection d'un gène quantitatif sur un chromosome à partir d'un grand échantillon d'un population rétrocroisée et en utilisant un grand nombre de marqueurs. Le test asymptotique proposé est basé sur la distribution du maximum de processus stochastiques à trajectoires dérivables ; nous donnons des bornes pour le niveau et la puissance du test. Les simulation illustrent la convergence vers le régime asymptotique ainsi que la précision des bornes.

* Corresponding author.
E-mail addresses: azais@cict.fr (J.-M. Azaïs), cierco@toulouse.inra.fr (C. Cierco-Ayrolles).

1. Model

In a genetic problem, studying a backcross population: $A \times (A \times B)$, Azaïs and Cierco-Ayrolles [1] adress the problem of detecting a gene influencing some quantitative trait on a given chromosome. They consider the following process, depending on the position d on the chromosome

$$S_n(d) := \frac{2}{n}\sum_{k=1}^{n}(Y_k - \overline{Y}_n)\mathbb{I}_{[X_k(d)=1]} - \frac{2}{n}\sum_{k=1}^{n}(Y_k - \overline{Y}_n)\mathbb{I}_{[X_k(d)=-1]}, \tag{1}$$

where

– Y_k is the observed quantitative variable on the individual k, $k = 1, n$.

– $X_k(d)$ is the genotypic composition of the individual k at location d on the chromosome, $d \in [0, L]$. In a backcross crossing scheme it can only take two values (AB or AA) that are denoted $+1$ or -1. This information is given by a genetic marker.

– d is the genetic distance from the origin of the chromosome, it is defined as a function of the probability of existence of crossing-overs. It is measured in Morgan (M);

– $\overline{Y}_n$ is the general mean of the data.

– $\mathbb{I}_{[E]}$ is the indicator function of the event E.

We use a model in which the true position of the gene is at location d_0 and its influence on the quantitative response of individuals is modelled by

$$Y_k = \mu + X_k(d_0)a/2 + \varepsilon_k,$$

with the usual analysis of variance assumptions. We set $\sigma^2 := \mathrm{Var}(\varepsilon_k)$. We assume that the genetic composition of each individual is observed only at locations $d_1, \ldots, d_M$ where some genetic markers exist. So that the full observation is

$$\big\{\big(Y_k, X_k(d_1), \ldots, X_k(d_M)\big),\ k = 1, \ldots, n\big\}.$$

When the putative location d and the true location d_0 agree, excepted minor modifications, the quantity estimated by formula (1) is equal to the analysis of variance estimator or the Gaussian maximum likelihood estimator of the gene effect a. Moreover $S_n(d)$ can be actually computed only if d is a genetic marker position d_i. Between such two positions, a linear interpolation is performed.

We use now a local asymptotic framework in which (a) the number n of observed individuals tends to infinity, (b) the number of genetic markers M_n tends to infinity with n, their locations being denoted by $d_{i,n}$; $i = 1, M_n$ (c) the size a of the gene effect is small; $a = \delta n^{-1/2}$.

Under this framework, and assuming a genetical model with crossing-over following a standard Poisson point process model, Cierco [4] studied the normalized process

$$X_n(d) := S_n(d)\big(\widehat{\mathrm{Var}}\big(S_n(d)\big)\big)^{-1/2},$$

where $\widehat{\mathrm{Var}}$ is obtained from the estimator $\hat{\sigma}^2$ function of the residual sum of squares. She proved that this process converges in distribution to an Ornstein–Uhlenbeck process with

a drift: $(X(d))_{d\in[0,L]}$ which is a Gaussian process with: $\mathbb{E}(X(d)) = \frac{\delta}{2\sigma}\exp(-2|d_0 - d|)$ and $\mathrm{Cov}(X(d), X(d+t)) = \exp(-2|t|)$.

2. Smoothing the detection test process

The problem is to test the null hypothesis $\delta = 0$ against $\delta \neq 0$. The classical approach would be to use the test statistic $T_n = \sup_{d\in[0,L]} |X_n(d)|$ which corresponds to a likelihood ratio test in the case of Gaussian observations. This is inconvenient for two reasons.

(i) The limit process has irregular sample paths, the distribution of its supremum is known only in some cases. In the other cases, existing bounds are not very sharp.

(ii) It does not take into account that the presence of a gene at d_0 modifies the expectation of the limit process in a neighbourhood of d_0.

For these two reasons, we have decided to smooth the detection test process $(X_n(d))_{d\in[0,L]}$. For calculations simplicity, we use a centred Gaussian kernel of varying variance ε^2 denoted φ_ε. Let $(X_n^\varepsilon(d))_{d\in[0,L]}$ be the smoothed process $(X_n * \varphi_\varepsilon)(d)$.

We considered the following test statistic $T_n^\varepsilon = \sup_{d\in[0,L]} |X_n^\varepsilon(d)|$. Property of weak convergence of processes (Billingsley, 1968) implies that the limit of $(X_n^\varepsilon(d))_{d\in[0,L]}$ is the smoothed version of the limit process, the characteristics of which can be easily computed. Note that since we work on asymptotic distribution, our results are free from the markers locations.

2.1. Bound for threshold and power calculation

Bounds are described for a generic process that will be denoted $(Y(d))_{d\in[0,L]}$. In practice, this process is the limit process $(X^\varepsilon(d))_{d\in[0,L]}$. So we consider a Gaussian process $(Y(d))_{d\in[0,L]}$ with $\mathcal{C}^1$ sample paths and we assume that for every $t_1, t_2; s_1, s_2 \in [0, L]$, $t_1 \neq t_2$, $s_1 \neq s_2$, the distribution of $Y(t_1), Y(t_2); Y'(s_1), Y'(s_2)$ is nondegenerate, (Y' is the derivative). In our particular case, this condition is met because the spectrum of $(Y(d) - \mathbb{E}(Y(d)))_{d\in[0,L]}$ has a continuous component [5].

For threshold or power calculations, we are interested in the distribution function of the random variable $|Y|^\star = \sup_{d\in[0,L]} |Y(d)|$. We use the following event equality which is a particular case of the general method described by Azaïs and Wschebor [2]:

$$\forall u \geqslant 0, \quad \mathbb{P}\{|Y|^\star > u\} = \mathbb{P}\big(\{|Y(0)| > u\} \cup \{|Y(0)| \leqslant u \ ; \ (U_u + D_{-u}) \geqslant 1\}\big), \quad (2)$$

where ";" denotes the intersection, U_u and D_{-u} are respectively the number of upcrossings of u and of downcrossings of $-u$ by the process Y on the interval $[0, L]$, defined as

$$U_u := \#\{d \in [0, L]; Y(d) = u; Y'(d) > 0\};$$
$$D_{-u} := \#\{d \in [0, L]; Y(d) = -u; Y'(d) < 0\}.$$

Our method is based on the double inequality below. If ξ is a random variable with non-negative integer values, then:

$$\mathbb{E}(\xi) - \frac{1}{2}\big[\mathbb{E}\big(\xi(\xi-1)\big)\big] \leqslant \mathbb{P}(\xi \geqslant 1) \leqslant \mathbb{E}(\xi). \quad (3)$$

Applying (3) with $\xi = (U_u + D_{-u})\mathbb{I}_{|Y(0)|\leqslant u}$, we have the fundamental inequality:

$$\mathbb{P}\{|Y(0)| > u\} + \mathbb{E}\big((U_u + D_{-u})\mathbb{I}_{|Y(0)|\leqslant u}\big) - \frac{\mathbb{E}[(U_u + D_{-u})(U_u + D_{-u} - 1)]}{2}$$
$$\leqslant \mathbb{P}\{|Y|^* > u\} \leqslant \mathbb{P}\{|Y(0)| > u\} + \mathbb{E}\big((U_u + D_{-u})\mathbb{I}_{|Y(0)|\leqslant u}\big). \tag{4}$$

Expectations involved in the above inequality can be evaluated by Rice's formulae [5] and expressions more adapted to numerical computation may be found in [3].

Remarks. –

– Relation (4) is a refinement of Davies' method [6]. Davies worked with the random variable $Y^\star = \sup_{d\in[0,L]} Y(d)$ and, instead of (4), he used the relation:

$$\mathbb{P}(Y^\star > u) \leqslant \mathbb{P}\big(Y(0) > u\big) + \mathbb{P}(U_u \geqslant 1) \leqslant \mathbb{P}\big(Y(0) > u\big) + \mathbb{E}(U_u).$$

Besides the fact that we work with $|Y|^\star$ instead of $Y^\star$, the upper bound is very similar to Davies' one, except for the small improvement due to the event $\{|Y(0)| \leqslant u\}$.

– By simulation, it has been verified that for centred process and rather large values of u, the lower bound is more accurate than the upper one. This is because of the use of the second order factorial moment. So, in the following, for threshold calculations, we will use the lower bound.

– This inequality cannot be applied directly to the original limit process for it has nondifferentiable sample paths.

3. Simulation study

This section presents the results of a Monte Carlo experiment to evaluate the quality of the proposed method under a variety of conditions. Our aim was to study (a) the relationship between the value of the smoothing parameter and the validity of the asymptotic approximation for reasonable numbers of markers and individuals, (b) the sharpness of the bounds given by the "fundamental inequality" for various values of the smoothing parameter.

Table 1 displays empirical levels for smoothed and unsmoothed procedures with thresholds calculated under the asymptotic distribution.

– For the unsmoothed process ($\varepsilon = 0$), the threshold is calculated using Table II of [7]. For this reason, the chromosome length, 0.98 M (Morgan) has been chosen to correspond to an entry of DeLong's table, and to be close to lengths encountered for several vegetal species.

– For the smoothed process, we used the lower bound in the "fundamental inequality".

Simulations have been performed for two values of the smoothing parameter and three markers densities: a marker every each icM with $i = 1, 2, 7$. The number of individuals has been chosen equal to 500. The crossing-overs were simulated according to a standard Poisson process. We performed 10000 simulations, so that the 5% confidence interval for the empirical levels associated to the theoretical ones are indicated.

Table 1
Threshold and empirical level (in %) of test using the unsmoothed detection test process ($\varepsilon = 0$) $(X_n(d))_{d\in[0,L]}$ and the smoothed detection process $(X_n^{\varepsilon}(d))_{d\in[0,L]}$. The chromosome length is equal to 0.98 M, and the number of individuals is equal to 500. The second line of the table gives a confidence interval for the empirical proportion related to the nominal level over 10^4 simulations

	Nominal level of the test								
	10%			5%			1%		
5% confidence interval for the emp. level	9.41–10.59			4.57–5.43			0.80–1.19		
Threshold $\varepsilon = 0$	2.74			3.01			3.55		
Threshold $\varepsilon^2 = 10^{-2}$	2.019			2.276			2.785		
Threshold $\varepsilon^2 = 10^{-3}$	2.321			2.593			3.128		
Marker density	1 cM	2 cM	7 cM	1 cM	2 cM	7 cM	1 cM	2 cM	7 cM
Emp. level for $\varepsilon = 0$	7.37	6.67	4.99	3.91	3.42	2.4	0.77	0.67	0.43
Emp. level for $\varepsilon^2 = 10^{-2}$	12.17	12.17	11.82	6.75	6.69	6.53	1.76	1.72	1.77
Emp. level for $\varepsilon^2 = 10^{-3}$	10.84	10.66	9.71	5.63	5.55	5.02	1.34	1.32	1.04

Table 2
Power in % associated to the detection test in the case of a gene of size $\delta = 6$, located at a distance $d_0 = 0.4$ from the origin of a chromosome of length 1 M. The value of σ is equal to 1. The empirical powers are calculated over 10^4 simulations and the corresponding 95% confidence intervals are given

	$\varepsilon^2 = 10^{-2}$	$\varepsilon^2 = 10^{-3}$	Unsmoothed process
5% threshold	2.281	2.599	3.02
Lower bound	69.84	69.05	–
Upper bound	71.27	82.11	–
Empirical power	71.37 ± 0.88	72.53 ± 0.87	68.99 ± 0.91

Table 2 presents the power associated to the detection test in the case of a gene of size $\delta = 6$ located at the position $d_0 = 0.4$. The length of the chromosome is 1 M, calculations are made under the asymptotic distribution, using a test with nominal level equal to 5%.

- For the unsmoothed detection test process, the threshold is calculated via DeLong's table and the power by the only possible method which is a Monte Carlo method. 10^4 simulations have been used.
- For the smoothed process, the threshold is calculated as above using the lower bound and the power is calculated by three manners: using the upper bound in the "fundamental inequality", using the lower bound in the "fundamental inequality", by a Monte Carlo method.

4. Discussion

Table 1 clearly indicates that the unsmoothed procedure is very conservative. We have checked by simulation that this is not due to a typo in DeLong's table.

The empirical level given by the smoothed procedure is close to the nominal value. For $\varepsilon^2 = 10^{-3}$, it is nearly inside the confidence interval.

Table 2 shows clearly that smoothing at size $\varepsilon^2 = 10^{-2}, 10^{-3}$ does not diminish power on the asymptotic distribution.

It is also clear in Table 2 that at the size $\varepsilon^2 = 10^{-2}, 10^{-3}$, the lower bound is almost exact.

In conclusion, the procedure we advocate is the use of the asymptotic test after smoothing with $\varepsilon^2 = 10^{-3}$ and with thresholds and powers calculated using the lower bound in the "fundamental inequality". The corresponding thresholds are given by Azaïs and Cierco-Ayrolles [1].

The validity of this procedure has been justified, in conditions near to practice, by the following statements that have been illustrated by the Monte Carlo experiment: (a) smoothing improves the convergence to the asymptotic. In condition near to practice the asymptotic behaviour is met, (b) smoothing does not diminish power on the asymptotic distribution, (c) the lower bound is almost exact for threshold and power calculation.

Acknowledgement

We thank C. Delmas for computational assistance.

REFERENCES

[1] J.-M. Azaïs, C. Cierco-Ayrolles, An asymptotic test for quantitative gene detection, Prépublication du laboratoire de statistique et probabilités, 2002, http://www.lsp.ups-tlse.fr/Azais/publis.html.

[2] J.-M. Azaïs, M. Wschebor, The distribution of the maximum of a Gaussian process: Rice method revisited. In and out of equilibrium: probability with a physical flavour, in: V. Sidoravicius (Ed.), Progress in Probability, Birkhauser, 2002, pp. 321–348.

[3] J.-M. Azaïs, C. Cierco-Ayrolles, A. Croquette, Bounds and asymptotic expansions for the distribution of the maximum of a smooth stationary Gaussian process, ESAIM: Probability and Statistics 3 (1999) 107–129.

[4] C. Cierco, Asymptotic distribution of the maximum likelihood ratio test for gene detection, Statistics 31 (3) (1998) 261–285.

[5] H. Cramér, M.R. Leadbetter, Stationary and Related Stochastic Processes, Wiley, New York, 1967.

[6] R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 64 (1977) 247–254.

[7] D.M. DeLong, Crossing probabilities for a square root boundary by a Bessel process, Communications in Statistics – Theory and Methods A10 (1981) 2197–2213.