Hum Genet (1999) 104 : 350–355

ORIGINAL INVESTIGATION

Chia-Ding Hou · Jengtung Chiang · John Jen Tai

Testing the nonrandomness of chromosomal breakpoints using highest observed breakages

Received: 27 August 1998 / Accepted: 18 December 1998

Abstract To determine whether a chromosomal band is a fragile site rather than a spontaneous breakpoint, an essential step is to test the nonrandomness of breakage at the region. In this paper, the nonapplicability of the testing procedure introduced by Bohm et al. is discussed, and a new detection procedure is proposed. This new procedure considers the relations of one site with the others, and can be applied to tests of the nonrandomness of breakpoints under either the proportional probability model, or the equiprobability model. A data set for Chinese patients with colorectal carcinoma is analyzed as an illustration of the proposed method.

Introduction

Fragile sites are an important issue in human genetics, and have attracted attention because of their apparent association with the origin of chromosomal rearrangements in cancer (Hecht and Glover 1984; Hecht and Sutherland 1984; Le Beau and Rowley 1984; Le Beau 1986; De Braekeleer 1987; Ardisia et al. 1993; Ohta et al. 1996; Sozzi et al. 1996). To determine whether a chromosomal band is a fragile site, rather than a spontaneous breakpoint, an essential step is to test the nonrandomness of breakage at the region. All the published methods test the nonrandomness either following the proportional probability model (PPM) or the equiprobability model (EPM). The PPM assumes that the probability of a random break at a band is proportional to the band width, whereas the EPM assumes that the probability of a random break is independent of the band width. Basically, the current methods can be divided into two major types: one considers that, either based on EPM or PPM, determination of the nonrandomness of breakage of a band is only related to its theoretical breakage proportions over all bands, and nothing to do with the observed band orders (Smith 1986; De Braekeleer and Smith 1988; Mariani 1989; Tarone 1989; Vasarhelyi and Friedman 1989; Jordan et al. 1990; Tai et al. 1993, 1998); the other considers that, also based on EPM or PPM, the observed band orders should be pooled into analysis, such that testing results of random or nonrandom breakpoints coincide with the observed band orders (Bohm et al. 1995). The two types of analytical method are derived from two different ways of thinking about the reasonable definition of a nonrandom breakpoint. There is no theoretical or applied evidence to show that the latter is better than the former, simply because the latter uses a more complicated multinomial distribution as the basis for statistical analysis and the former uses the binomial assumption. In this paper, we shall first demonstrate that it is not correct as Bohm et al. (1995) contended that their test procedures can be directly modified to scale the multinomial-homogeneity expectations to reflect band width. Second, a new detection procedure that detects nonrandom breakpoints using the highest observed breakage is proposed. To demonstrate the applicability of our method, a real data set for Chinese patients with colorectal carcinoma is analyzed for illustration.

C.-D. Hou
Department of Statistics, Fu Jen Catholic University, Taipei, Taiwan, ROC

J. Chiang
Department of Statistics, National Chengchi University, Taipei, Taiwan, ROC

J. J. Tai (✉)
Institute of Epidemiology, National Taiwan University, 1 Jen-Ai Road, Section 1, Taipei, 100 Taiwan, ROC
Fax: +886-02-23511955

Nonapplicability of the procedures of Bohm et al. under PPM

Let k be the number of all bands investigated and m the number of cell metaphases observed in a study. For each band in a metaphase, two observations of gaps or breaks may be detected because there are two homologous chromosomes corresponding to a band. Let N_{ij} be the number of breaks observed at the two homologous chromosomes of

the *i*th band of the *j*th metaphase, N_{ij} = 0, 1 or 2, where $i = 1,2,\ldots k$, $j = 1,2,\ldots m$, and the marginal total $N_i = \Sigma_{j=1}^{m} N_{ij}$ be the total number of breaks observed at the *i*th band over *m* metaphases. The total number of breaks detected is $n = \Sigma_{i=1}^{k} N_i$. Denote the proportion of breaks occurring at the *i*th band to the total breaks in a haploid set by P_i for $i = 1, 2, \ldots k$, then the vector of the observed number of breaks ($N_1, N_2, \ldots, N_k$) is multinomially distributed as:

$$(N_1, N_2,\ldots, N_k) \sim mult(n,k,\underline{P}), \tag{1}$$

where $\underline{P} = (P_1, P_2,\ldots, P_k)$. Based on this distribution, Bohm et al. (1995) thought that a nonfragile site has a small and essentially equal probability of breakage, and a fragile site has a large and not necessarily equal probability of breakage. Under the EPM point of view, they assumed that the *k* chromosomal sites can be indexed according to their order in probability of breakage, i.e.,

$$P_1 \le P_2 \le \ldots\ldots \le P_k$$

The first k_1 ($\le k$) sites are defined to be nonfragile and the remaining $k - k_1$ sites are defined to be fragile if probabilities of breakage satisfy

$$P_1 = P_2 \ldots = P_{k_1} < \tfrac{1}{K} < P_{k_1+1} \le P_{K_1+2} \le \ldots \le P_k$$

They test incrementally smaller subsets of the data for homogeneity under model (1), which assigns equal probabilities to a maximal set of nonfragile and unrestricted probabilities to the remaining fragile sites with significantly higher number of breaks, i.e., test the null hypothesis

$$H_0 : P_1 = \tfrac{1}{k_1}, P_2 = \tfrac{1}{k_1}, \ldots P_{k_1} = \tfrac{1}{k_1} \tag{2}$$

stepwise at significance level $\frac{\alpha}{t+1}$ at the t^{th} iteration [the use of significance level $\frac{\alpha}{t+1}$ at the t^{th} iteration is an application of the Bonferroni approach, see Seber (1977)] using the Pearson's χ^2 statistic

$$\chi^2 = \sum_{i=1}^{k_1} N_i \left(\frac{N_i}{n_1 / k_1} - 1 \right) \tag{3}$$

or the likelihood ratio statistic

$$G^2 = \sum_{i=1}^{k_1} N_i \ln\left(\frac{N_i}{n_1 / k_1} \right) \tag{4}$$

where $n_1 = \sum_{i=1}^{k_1} N_i$

Bohm et al. (1995) concluded that their testing procedure can be directly modified to scale the multinomial-homogeneity expectations to reflect band width and hence their procedure can be used under the PPM assumption. Let w_i be the width of the *i*th band in a haploid and $W = \Sigma_{i=1}^{k} w_i$ be the total width of all bands. Let $\underline{P}^0 = (P_1^0, \ldots, P_k^0) = (\frac{w_1}{W},\ldots,\frac{w_k}{W})$. According to their contention, under the PPM assumption, we can replace hypothesis (2) by

$$H_0 : P_1 = P_1^0, \ldots\ldots, P_{k_1} = P_{k_1}^0 \tag{5}$$

and replace statistic (3) by

$$\chi^2 = \sum_{i=1}^{k_1} N_i \left(\frac{N_i}{n_1 P_i^0} - 1 \right)$$

or replace statistic (4) by

$$G^2 = \sum_{i=1}^{k_1} N_i \ln\left(\frac{N_i}{n_1 P_i^0} \right)$$

and perform the above prodecure to identify fragile sites iteratively. Imposing the observed orders on the null hypothesis of EPM, it is acceptable using the method of Bohm et al. to exclude a band of the highest observed rank stepwise if the testing result is significant at some iterations, But, obviously, since the observed band orders of a set of breakage data cannot reflect the true orders under PPM, imposing the observed orders on the null hypothesis of PPM for testing, their method is not applicable. The following two examples as listed in Tables 1 and 2 are used for illustration of this point.

Table 1 Analysis of artificial data using the procedure of Bohm et al. (1995). ($\alpha = 0.05$), $\underline{B} = (B_1, B_2,\ldots,B_7)$, $\underline{N} = (50, 17, 10, 8, 7, 1, 7)$, $\underline{W} = (50, 17, 10, 8, 7, 6, 2)$. B_i, band i; $\underline{N}$, vector of observed breakages; $\underline{W}$, vector of band widths; $\underline{R} = (R_1,\ldots,R_k)$, $R_i = \tilde{P}_i / P_i^0$, $\tilde{P}_i = N_i / \sum_1^k N_i$, $i = 1,\ldots,k$; $\alpha^* = \frac{\alpha}{t+1}$; $\underline{P}^0$, breakage proportion under PPM; $\underline{\tilde{P}}$ observed breakage proportion

Iteration *t*	$\underline{P}^0$	$\underline{\tilde{P}}$	$\underline{R}$	χ^2	$\chi^2_{\alpha^*,(k-1)}$	Conclusion
1	(0.5,0.17,0.1 0.08,0.07, 0.06,0.02)	(0.5,0.17,0.1 0.08,0.07 0.01,0.07)	(1,1,1, 1,1, 0.163,3.5)	16.67	14.45	B_1 is fragile
2	(0.34,0.2, 0.16,0.14 0.12,0.04)	(0.34,0.2, 0.16,0.14, 0.12,0.04)	(1,1, 1,1, 0.167,3.5)	16.67	13.84	B_2 is fragile
3	(0.303,0.242, 0.212,0.182, 0.091)	(0.303,0.242, 0.212,0.03, 0.212)	(1,1, 1,0.167, 3.5)	16.67	12.76	B_3 is fragile
4	(0.348,0.304, 0.261,0.087)	(0.348,0.304, 0.043,0.304)	(1,1, 0.167,3.5)	16.67	11.34	B_4 is fragile
5	(0.467, 0.4,0.133)	(0.467, 0.067,0.467)	(1,0.167, 0.467)	16.67	9.58	B_5 is fragile
6	(0.75,0.25)	(0.125,0.875)	(0.167,3.5)	16.67	7.24	B_7 is fragile
7	1	1	1	0	–	B_6 is nonfragile

Table 2 Analysis of an artificial data set using the procedure of Bohm et al. (1995). ($\alpha = 0.05$), $\mathcal{B} = (B_1, B_2,...,B_7)$, $\mathcal{N} = (25, 20, 15, 14, 10, 9, 7)$, $\mathcal{W} = (50, 20, 10, 7, 6, 5, 2)$. B_i, band I; $\mathcal{N}$, vector of observed breakages; $\mathcal{W}$ vector of band widths; $\mathcal{R} = (R_1,...,R_k)$, $R_i = \tilde{P}_i / P_i^0$, $\tilde{P}_i = N_i / \sum_1^k N_i$, $i = 1,...,k$, $\alpha^* = \frac{\alpha}{t+1}$; $\mathcal{P}^0$, breakage proportion under PPM; $\tilde{\mathcal{P}}$ observed breakage proportion

Iteration t	$\mathcal{P}^0$	$\tilde{\mathcal{P}}$	$\mathcal{R}$	χ^2	$\chi^2_{\alpha^*,(k-1)}$	Conclusion
1	(0.5,0.2,0.1 0.07,0.06, 0.05,0.02)	(0.25,0.2,0.15, 0.14,0.1, 0.09,0.07)	(0.5,1,1.5, 2,1.67, 1.8,3.5)	40.37	14.45	B_1 is fragile
2	(0.4,0.2, 0.14,0.12, 0.1,0.04)	(0.27,0.2, 0.19,0.13, 0.12,0.09)	(0.67,1, 1.33,1.11, 1.2,2.33)	10.24		$B_2 \sim B_7$ are nonfragile

In the first example in Table 1, by examining the ratio between the two breakage proportions under either the observed distribution or the PPM, the result indicates nonfragility at band 1 to band 6 ($\tilde{P}_i / P_i^0 = 1$, $i = 1,..., 5$; $\tilde{P}_6 / P_6^0 = 0.16$) but fragility at band 7 ($\tilde{P}_7 / P_7^0 = 3.5$). However, following the procedure of Bohm et al., except for band 6, all the others are declared fragile. In the second example in Table 2, the observed distribution indicates nonfragility at band 1 ($\tilde{P}_1 / P_1^0 = 0.5$) but fragility at band 7 ($\tilde{P}_7 / P_7^0 = 3.5$). However, a converse result is obtained for the two bands if the procedure of Bohm et al. is followed. The anomalous conclusions of the above examples show the nonapplicability of their procedure under PPM due to failure adequately to detect outlying cells simultaneously.

An alternative procedure using highest observed breakages

Bohm et al. (1995) mentioned that their procedure does not circumvent the problem inherent with the sparse contingency tables obtained from chromosomal breakage data for single individuals. Koehler and Larntz (1980) concluded from simulation that the χ^2 approximation to Pearsons's χ^2 or likelihood ratio test statistics tends to be poor for sparse tables containing both small and moderately large expected frequencies. Vasarhelyi and Friedman (1989) and Tarone (1989) also concluded that a χ^2 test is inadequate for localization of preferential sites of breakage at the level of chromosomal bands because the calculations involve a large number of chromosomal bands and a small number of breakpoints in each band.

In addition, calculating the Pearson's χ^2 (or likelihood ratio) statistics iteratively for testing nonrandomness using the procedure of Bohm et al. (1995) requires computer assistance, because n and k are usually large in real data. This is inconvenient for a cytogeneticist who does not have much training in statistical computation. Therefore, developing a method that can avoid using a computer should be useful. In the following, a new detection procedure is proposed. This new procedure detects the nonrandomness of breakpoints using the highest observed breakage. This procedure can be directly modified to reflect band width and hence can be used to identify fragile sites under the PPM assumption. Moreover, the accuracy of this new procedure can be evaluated using the lower and upper bounds for the critical value of the M test proposed by Fuchs and Kenett (1980). Calculation of this procedure can be programmed on a hand calculator, so this new procedure is simpler than the approach proposed by Bohm et al. (1995), particularly under the EPM.

The M test was developed to detect outlying cells in the multinomial distribution and to test the null hypothesis (5) as well. To detect positive outliers (i.e., fragile sites), we can apply the M test and use the largest order statistic of the standardized observation (standardized observed breakage in the problem of detecting fragile sites), $\max_{1 \le i \le k_1} N_i^*$, where

$$N_i^* = \frac{N_i - n_1 P_i^0}{\sqrt{n_1 P_i^0 (1 - P_i^0)}}$$

as our test statistic. Sharp lower and upper bounds for the critical value of the M test were proved to be

$$Z_{1-\alpha_1},\ Z_{1-\alpha_2} \tag{6}$$

where

$$\alpha_1 = \frac{k_1 - \sqrt{k_1^2 - 2\alpha k_1 (k_1 - 1)}}{k_1 (k_1 - 1)}$$

$$\alpha_2 = \frac{\alpha}{k_1}$$

and $Z_{1-\alpha_1}$ and $Z_{1-\alpha_2}$ are the $100 \times (1-\alpha_1)$th and $100 \times (1-\alpha_2)$th percentiles of the standard normal distribution, respectively. Since the upper bound on the critical value of the M test is simple to compute and use of it results in a conservative test, we can use the upper bound as our critical value (by computing the lower bound, the accuracy of the upper bound can be evaluated).

To detect the chromosomal fragile sites [i.e., the positive outlying cells under multinomial model (1)] under the PPM assumption, we can apply the M test and create the following procedure:

(i) Let $t = 1$

(ii) Let $\alpha^* = \frac{\alpha}{t+1}$

Table 3 Analysis of the artificial data set given in Table 1 using our procedure. ($\alpha = 0.05$), $\mathcal{B} = (B_1, B_2,...,B_7)$, $\mathcal{N} = (50, 17, 10, 8, 7, 1, 7)$, $\mathcal{W} = (50, 17, 10, 8, 7, 6, 2)$. B_i, band I; $\mathcal{N}$, vector of observed breakages; $\mathcal{W}$ vector of band widths; $\mathcal{R} = (R_1,...,R_k)$, $R_i = \tilde{P}_i / P_i^0$, $\tilde{P}_i = N_i / \sum_1^k N_i$, $i = 1,...,k$, $\alpha^* = \frac{\alpha}{t+1}$; $\mathcal{P}^0$, breakage proportion under PPM; $\tilde{\mathcal{P}}$ observed breakage proportion

Iteration t	$\mathcal{P}^0$	$\tilde{\mathcal{P}}$	$\mathcal{R}$	$\max_{1\le i\le k} R_i$	$Z_{1-\alpha^*/k}$	Conclusion
1	(0.5,0.17,0.1, 0.08,0.07, 0.06,0.02)	(0.5,0.17,0.1, 0.08,0.07, 0.01,0.07)	(1,1,1,1,1 0.16,3.5)	3.57	2.69	B_7 is fragile
2	(0.51,0.173, 0.102,0.082, 0.071,0.061)	(0.538,0.183, 0.108,0.086 0.075,0.011)	(1.05,1.06, 1.06,1.05, 1.06,0.18)	0.53	2.77	B_1, B_2, B_3, B_4, B_5, B_6 are nonfragile

Table 4 Analysis of the artificial data set given in Table 2 using our procedure. ($\alpha = 0.05$), $\mathcal{B} = (B_1, B_2,...,B_7)$, $\mathcal{N} = (25,20, 15,14,10,9,7)$, $\mathcal{W} = (50,20,10, 7,6,5,2)$. B_i, band I; $\mathcal{N}$, vector of observed breakages; $\mathcal{W}$ vector of band widths; $\mathcal{R} = (R_1,...,R_k)$, $R_i = \tilde{P}_i / P_i^0$, $\tilde{P}_i = N_i / \sum_1^k N_i$, $i = 1,...,k$, $\alpha^* = \frac{\alpha}{t+1}$; $\mathcal{P}^0$, breakage proportion under PPM; $\tilde{\mathcal{P}}$ observed breakage proportion

Iteration t	$\mathcal{P}^0$	$\tilde{\mathcal{P}}$	$\mathcal{R}$	$\max_{1\le i\le k_1} R_i$	$Z_{1-\alpha^*/k}$	Conclusion
1	(0.5,0.2,0.1, 0.07,0.06 0.05,0.02)	(0.25,0.2,0.15 0.14,0.1, 0.09,0.07)	(0.5,1,1.5 2,1.67, 1.8,3.5)	3.57	2.69	B_7 is fragile
2	(0.51,0.204 0.102,0.071, 0.061,0.051)	(0.269,0.215, 0.161,0.151, 0.108,0.097)	(0.527, 1.054,1.578, 2.127,1.77, 1.902)	2.96	2.77	B_4 is fragile
3	(0.549,0.22, 0.11,0.066, 0.055)	(0.316,0.253, 0.19,0.127, 0.114)	(0.576,1.15, 1.727,1.924, 2.073)	2.30	2.81	B_1, B_2, B_3, B_5, B_6 are nonfragile

(iii) Test the null hypothesis

$H_0: P_1 = P_1^0, ..., P_k = P_k^0$

using the test statistic $\max_{1\le i\le k1} N_i^*$ (i.e., the highest observed "standardized breakage number") and the critical value $Z_{1-\alpha^*/k}$.

(iv) If the hypothesis in step (iii) is not rejected at the α^* significance level, then CONCLUDE that all the remaining sites are not fragile sites and STOP.

(v) If the hypothesis in step (iii) is rejected at the α^* significance level, then EXCLUDE the site with the highest observed standardized breakage number (i.e., the site corresponding to $\max_{1\le i\le k1} N_i^*$), let $t = t + 1$, $k = k - 1$, and let $(w_1,..., w_k)$ and $(P_1^0, ..., P_k^0) = (\frac{w_1}{\sum w_i},...,\frac{w_k}{\sum w_i})$ be the vectors of band widths and probabilities of breakage under random breakage of the remaining sites, respectively. RETURN to step (ii).

Continue the above steps iteratively until we obtain a subset of the data for which we are not able to reject the hypothesis in step (iii). The sites in this set are considered as nonfragile sites. The other sites are fragile sites.

Let us consider the examples given in the previous section again. The data sets given in Tables 1 and 2 are reanalyzed using the new procedure introduced in this section. Results of this analysis are listed in Tables 3 and 4. From the results given in these tables, it is obvious that the new procedure can adequately detect the positive outlying cells (i.e., fragile sites) simultaneously with the rejection of the null hypothesis (5) under the PPM assumption.

Fuchs and Kenett (1980) concluded in their paper that, among all the cases, the maximum difference between the bounds on the power of the M test, as calculated by using the upper and the lower bound in equation (6), was less than 0.004. Therefore, the upper bound turns out to be very accurate. Furthermore, for the problem of testing the null hypothesis (5), the computed lower bounds for the asymptotic power of the M test exceed the power of the χ^2 test, particularly when the number of outliers is limited (to less than 5–10% of the number of cells) and the deviations are fairly unequal.

Since the EPM can be viewed as a special case of the PPM, under the EPM assumption, we can let $(P_1^0, P_2^0,..., P_k^0) = (\frac{1}{k}, \frac{1}{k},...,\frac{1}{k})$ and directly modify the procedure proposed above as follows

(i) Let $t = 1$

(ii) Let $\alpha^* = \frac{\alpha}{t+1}$

(iii) Test the null hypothesis

$H_0: P_1 = \frac{1}{k}, P_2 = \frac{1}{k}, ... P_k = \frac{1}{k}$

by using the test statistic $\max_{1\le i\le K} N_i$ (i.e., the highest observed breakage) and the critical value $\frac{n}{k} + Z_{(1-\alpha^*/K)}\sqrt{\frac{n(k-1)}{k^2}}$.

(iv) If the hypothesis in step (iii) is not rejected at the α^* significance level, then CONCLUDE that all the remaining sites are not fragile sites and STOP.

(v) If the hypothesis in step (iii) is rejected at the α^* significance level, then EXCLUDE the site with the highest

Table 5 Chromosomal sites classified as fragile under the proportional probability model (PPM) using our procedure in 30 Chinese patients with colorectal carcinoma. (Dotted vertical lines indicate omitted data)

Site	Frequency n_i	Band width[a] w_i	Test statistic $\max_{1\le i\le k} N^*_i$	Bounds for[b] critical value $[Z_{1-\alpha_1}, Z_{1-\alpha_2}]$	HGM10[c]
1p21	19	16	10.4496***	[4.3353,4.3355] [4.6772,4.6772] [5.1293,5.1293]	C
1p22	15	17	8.2332***	[4.3747,4.3749] [4.7139,4.7140] [5.1630,5.1631]	C
1p31	18	29.5	7.4731***	[4.4067,4.4069] [4.7438,4.7438] [5.1905,5.1905]	C
1p32	9	11	6.7425***	[4.4180,4.4181] [4.7543,4.7543] [5.2002,5.2002]	C
1q25	8	9	6.8022***	[4.4233,4.4235] [4.7593,4.7593] [5.2048,5.2048]	C
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
22q12	20	7	15.7136***	[4.2461,4.2465] [4.5942,4.5942] [5.0533,5.0533]	C
Xp22	149	23	51.0815***	[3.9471,3.9486] [4.3180,4.3183] [4.8018,4.8018]	C
Xq22	31	9	20.1852***	[4.1789,4.1794] [4.5318,4.5319] [4.9962,4.9962]	C

*, **, and *** represent significant fragility at $\alpha = 0.05$, 0.01 and 0.001, respectively

[a] The relative width of each of the 320 bands was measured using the banding diagram of the International System for Chromosome Nomenclature (ISCN 1981)

[b] The three intervals are the bounds for the critical value of the M test at $\alpha = 0.05$, 0.01 and 0.001, respectively, where $[Z_{1-\alpha_1}, Z_{1-\alpha_2}]$ is defined as in equation (6)

[c] C, P, and T represent the sites that are identified as fragile by the Tenth International Workshop on Human Gene Mapping (HGM10) (Sutherland and Ledbetter 1989)

Table 6 Chromosomal sites classified as fragile under the equiprobability model (EPM) using our procedure in 30 Chinese patients with colorectal carcinoma. (Dotted vertical lines indicate omitted data)

Site	Frequency (test statistic) n_i	Bounds for critical value[a] $[Z_{1-\alpha_1}, Z_{1-\alpha_2}]$	HGM10[b]
1p21	19***	[7.8535,7.8539] [8.3280,8.3281] [8.9550,8.9550]	C
1p22	15***	[7.2487,7.2489] [7.6865,7.6865 [8.2658,8.2658]	C
1p31	18***	[7.7214,7.7218] [8.1879,8.1879] [8.8044,8.8044]	C
1p32	9***	[6.4998,6.5000] [6.8927,6.8927] [7.4133,7.4133]	C
1q25	8***	[6.0896,6.0897] [6.4577,6.4577] [6.9460,6.9460]	C
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
22q12	20***	[8.1155,8.1159] [8.6063,8.6063] [9.2544,9.2544]	C
Xp22	149***	[11.4986,11.5015] [12.2214, 12.2219] [13.1639,13.1640]	C
Xq22	31***	[9.3008,9.3016] [9.8677,9.8678] [10.6137,10.6137]	C

*, **, and *** represent significant fragility at $\alpha = 0.05$, 0.01 and 0.001, respectively

[a] The three intervals are the bounds for the critical value of the M test at $\alpha = 0.05$, 0.01 and 0.001, respectively, where $[Z_{1-\alpha_1}, Z_{1-\alpha_2}]$ is defined as in equation (6)

[b] C, P, and T represent the sites that are identified to be fragile by the Tenth International Workshop on Human Gene Mapping (HGM10) (Sutherland and Ledbetter 1989)

observed breakage, let $t = t + 1$, $k = k - 1$, and RETURN to step (ii).

Continue the above steps iteratively until we obtain a subset of the data for which we are not able to reject the hypothesis in step (iii). The sites in this set are considered as nonfragile sites. The other sites are fragile sites. Obviously, our test procedure is simpler than the one proposed by Bohm et al. (1995) in calculation and can be programmed on a hand calculator.

Numerical studies

In this section, we reanalyzed the data set given in Wang et al. (1992) using the new procedure introduced in the previous section. The data set involves the frequency and spectrum of both common and rare fragile sites in 30 Chinese patients with colorectal carcinoma. [A brief description of this real data set was given in Tai et al. (1993)]. The results under the EPM and the PPM assumption are given in Tables 5 and 6, respectively.

As shown in Table 5, employing our approach, 41 fragile sites were detected at a significance level of 0.001 under the PPM assumption. Among these 41 sites 35 are listed in HGM10 (Sutherland and Ledbetter 1989). To save space we omit a proportion of the detected fragile sites in Table 5.

According to Table 6, employing our procedure, 49 fragile sites were detected at a significance level of 0.001 under the EPM assumption. Among these 49 sites, 43 are listed in HGM10 (Sutherland and Ledbetter 1989). Again, a proportion of the detected results in Table 6 are omitted to save space. Employing the procedure of Bohm et al. (1995), 74 fragile sites were detected at a significance level of 0.001 under the EPM assumption. (The same

conclusion is reached no matter whether Pearson's χ^2 is standardized or not). Each site with a number of breakages greater than or equal to 4 will be identified as fragile by using their procedure.

Conclusion

Bohm et al. (1995) concluded that their testing procedure can be used to detect fragile sites under PPM assumption. However, as we comment in the second section, their conclusion is incorrect. Their procedure cannot be used to detect fragile sites under the PPM assumption. Porfirio et al. (1987) showed that the number of breakage events found on a chromosomal band is related to the length of the band; the length positively influences the probability of showing a break. If we follow the discovery of Porfirio et al. (1987) and adopt the null hypothesis, which considers the expected breakage frequencies to be proportional to the band widths under random breakage, then the procedure proposed by Bohm et al. (1995) remains problematic and cannot be used to detect fragile sites.

In this article, we introduce a new procedure that detects fragile sites using the largest order statistics. This new procedure can be applied to tests of the nonrandomness of breakpoints under either the PPM or the EPM. Moreover, the accuracy of the new procedure can be evaluated via the lower bounds on the critical value of the M test proposed by Fuchs and Kenett (1980). Fuchs and Kenett (1980) concluded in their paper that, among all cases, the maximum difference between the bounds on the power of the M test, as calculated by using the upper and the lower bound in equation (6), was less than 0.004. Therefore, the upper bound turns out to be very accurate. Furthermore, for the problem of testing the null hypothesis (5), the computed lower bounds for the asymptotic power of the M test exceed the power of the χ^2 test, particularly when the number of outliers is limited (to less than 5–10% of the number of cells) and the deviations are fairly unequal.

It is natural to ask what the total significance level of the new procedure introduced here is [this problem is also applicable to the procedure introduced by Bohm et al. (1995) and most statistical methods in this area, too]. This question is obviously a difficult one and the answer is not at all immediate since the test statistics used at each iteration are correlated. More research is needed to solve this problem.

References

Ardisia C, Venti G, Colozza MA, Breschi C, Porfirio B, Davis S, Tonato M, Donti E (1993) Expression of aphidicolin-induced fragile sites in lymphocytes of patients with breast cancer. Cancer Genet Cytogenet 67:113–116

Beau MM Le (1986) Chromosomal fragile sites and cancer-specific rearrangements. Blood 67:849–858

Beau MM Le (1984), Rowley JD (1984) Heritable fragile sites and cancer. Nature 308:607–608

Böhm U, Dahm PF, McAllister BF, Greenbaum IF (1995) Identifying chromosomal fragile sites from individuals: a multinomial statistical model. Hum Genet 95:249–256

Braekeleer M De (1987) Fragile sites and chromosomal structure rearrangements in human leukemia and cancer. Anticancer Res 7:417–422

Braekeleer M De, Smith B (1988) Two methods for measuring the non-randomness of chromosome abnormalities. Ann Hum Genet 52:63–67

Fuchs C, Kenett R (1980) A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. J Am Stat Assoc 75:395–398

Hecht F, Glover TW (1984) Cancer chromosome breakpoints and common fragile sites induced by aphidicolin. Cancer Genet Cytogenet 13:185–188

Hecht F, Sutherland GR (1984) Fragile sites and cancer breakpoints. Cancer Genet Cytogenet 12:179–181

ISCN (1981) An international system for human cytogenetic nomenclature – high resolution banding. Cytogenet Cell Genet 31:1–23

Jordan DK, Burns TW, Divelbiss JE, Woolson RF, Patil SR (1990) Variability in expression of common fragile sites : in search of a new criterion. Hum Genet 85:462–466

Koehler K, Larntz K (1980) An empirical investigation of goodness-of-fit statistics for sparse multinomials. J Am Stat Assoc 75:336–344

Mariani T (1989) Fragile sites and statistics. Hum Genet 81:319–322

Ohta M, Inoue H, Cotticelli MG, Kastury K, Baffa R, Palazzo J, Siprashvili Z, Mori M, McCue P, Druck T, Croce CM, Heubner K (1996) The FHIT gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. Cell 84:587–597

Porfirio B, Dallapiccolla B, Terrenato L (1987) Breakpoint distribution in constitutional chromosome rearrangements with respect to fragile sites. Ann Hum Genet 51:329–336

Seber GA (1977) Linear regression analysis. Wiley, New York

Smith CAB (1986) Chi-square tests with small numbers. Ann Hum Genet 50:163–167

Sozzi G, Veronese ML, Negrini M, Baffa R, Cotticelli MG, Inoue H, Tornielli S, Pilotti S, Gregorio L De, Pastorino U, Pierotti MA, Ohta M, Heubner K, Croce CM (1996) The FIHT gene at 3p14.2 is abnormal in lung cancer. Cell 85:17–26

Sutherland GR, Ledbetter DH (1989) Report of the committee on cytogenetic markers. (Tenth International Workshop on Human Gene Mapping) Cytogenet Cell Genet 51:452–458

Tai JJ, Hou C-D, Wang-Wuu S, Wang C-H, Leu S-Y, Wuu K-D (1993) A method for testing the nonrandomness of chromosomal breakpoints. Cytogenet Cell Genet 63:147–150

Tai JJ, Hou C-D, Wang-Wuu S (1998) A confirmation analysis method for identification of chromosomal fragile sites. Cancer Genet Cytogenet 105:1–5

Tarone RE (1989) Testing for nonrandomness of events in sparse data situations. Ann Hum Genet 53:381–387

Vasarhelyi K, Friedman JM (1989) Analyzing rearrangement between breakpoint distributions by means of binomial confidence intervals. Ann Hum Genet 3:375–380

Wang C-H (1992) Chromosomal fragile sites expression in lymphocytes of patients with colorectal carcinoma and of healthy controls. MS thesis, National Yang-Ming University, Taipei, Taiwan