BIOMETRICS 57, 435–440
June 2001

Identifying Chromosomal Fragile Sites from a Hierarchical-Clustering Point of View

Chia-Ding Hou,[1,*] Jengtung Chiang,[2,] and John Jen Tai[3]**

[1]Department of Statistics, Fu Jen Catholic University,
Taipei Hsien, 242 Taiwan, R.O.C.
[2]Department of Statistics, National Chengchi University,
Taipei, Taiwan, R.O.C.
[3]Division of Biostatistics, Institute of Epidemiology, National Taiwan University,
1 Jen-Ai Road, Section 1, Taipei, 100 Taiwan, R.O.C.
*email: stat0002@mails.fju.edu.tw
**email: chiangj@nccu.edu.tw

SUMMARY. Identification of fragile sites is a way to investigate the genetic abnormalities that are the hallmark of cancer and play an important role in carcinogenesis. Manifestation of nonrandom breakage at a chromosome band is an essential criterion for determination of the fragility of the band. In this article, a new detection procedure is proposed. This new procedure takes the relationship of one site with the others into consideration and can be applied to tests of the randomness of breakpoints under either the proportional probability model (PPM) or the equiprobability model (EPM). The procedure can form a grouping structure that classifies all sites into several clusters. It is applied to identification of fragile sites for a real data set for Chinese patients with colorectal carcinoma for illustration of the proposed method.

KEY WORDS: Equiprobability model; Fragile site; Hierarchical clustering; Multiple hypothesis testing; Proportional probability model.

1. Introduction

Chromosomes are the threadlike packages of DNA in the nucleus of a cell that carry portions of the hereditary information of an organism. Chromosome band refers to a narrower portion of a chromosome that has been darkened by interaction with a dye. Each human chromosome displays a pattern of bands and can be identified by its pattern. Fragile sites on chromosomes are points at which the chromosome is liable to break. Identification of fragile sites may contribute to the identification of genetic abnormalities that are the hallmark of cancer and play an important role in tumorigenesis (Hecht and Sutherland, 1984; Le Beau, 1986; Sozzi et al., 1996; Chen et al., 1998; Tai and Hou, 1999). From the molecular genetic aspect, identification of fragile sites may also lead to the identification of genes responsible for cancers such as lung, stomach, colon, and ovary (cf., Ried et al., 2000). To determine whether or not a chromosome band containing a number of breaks is a fragile site, an essential criterion is to find statistical evidence for nonrandom breakage at that region. Several statistical methods have been proposed to detect fragile sites. Basically, all the published methods test the randomness under either the proportional probability model (PPM; Smith, 1986; De Braekeleer and Smith, 1988; Tai et al., 1993; Tai, Hou, and Wang-Wuu, 1998) or the equiprobability model (EPM; Mariani, 1989; Tai et al., 1993, 1998). The PPM assumes that the probability of a random break at a band is proportional to the bandwidth, whereas the EPM assumes that the probability of a random break is independent of the bandwidth. Bohm et al. (1995) proposed a procedure that takes the relationship among sites into consideration. However, we have shown that it is not as Bohm et al. (1995) contended that their test procedures can be directly modified to scale the multinomial-homogeneity expectations (viz. the equiprobability model) to reflect bandwidth (Hou, Chiang, and Tai, 1999). Their procedure cannot be used to detect fragile sites under the proportional probability model. In this article, a new detection procedure that detects fragile sites from a hierarchical-clustering point of view is developed. This procedure takes the relationship of one site with the others into consideration and can be applied to tests of the randomness of breakpoints under either the PPM or the EPM. To demonstrate the applicability of our method, a data set of Chinese patients with colorectal carcinoma is analyzed for illustration.

2. Review of Bohm et al.'s (1995) Procedures

In cytogenetic studies, usually blood samples from a number of individuals (e.g., 30 patients) are obtained. For each blood sample, a number of cells (e.g., 50 cells) are further sampled for investigation of the chromosomal fragility under appropriate conditions of induction using chemicals (e.g., aphidicolin;

Table 1
Chromosomal sites classified as fragile (under the PPM or EPM) using our procedure in 30 Chinese patients with colorectal carcinoma

Model	Site	Frequency n_i	Bandwidth[a] W_i	Cluster	HGM10[b]
PPM	1p21	19	16	2	C
	1p22	15	17	2	C
	1p31	18	29.5	2	C
	1p32	9	11	3	C
	1q25	8	9	3	C
	1q44	42	7	1	C
	—	—	—	—	—
	—	—	—	—	—
	—	—	—	—	—
	22q12	20	7	1	C
	Xp22	149	23	1	C
	Xq22	31	9	1	C
EPM	1p21	19		1	C
	1p22	15		2	C
	1p31	18		2	C
	1p32	9		3	C
	1q44	42		1	C
	2p16	32		1	P
	—	—		—	—
	—	—		—	—
	—	—		—	—
	22q12	20		1	C
	Xp22	149		1	C
	Xq22	31		1	C

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

[b] C, P, and T represent the status of fragile sites that are determined as stipulated by Tenth International Workshop on Human Gene Mapping (HGM10). C (Confirmed status) represent "Reported by more than one laboratory with irrefutable evidence that it is a fragile site"; P (Provisional status) represent "Reported with considerable evidence for existence by one laboratory"; T (Tentative status) represent "Reported by one or more laboratories but with insufficient evidence to be sure that it really exists as a fragile site" (Sutherland and Ledbetter, 1989).

see Sutherland and Hecht, 1985). The data structure generated from this type of cytogenetic studies are as shown in Table 1 and will be defined as follows. In analysis of these cytogenetic data, independence is usually assumed among all individuals and all cells by all methods (Tai et al., 1993).

Let k be the number of chromosomal bands and m the number of observed cells in a study. Because breaks may occur at one or both homologous chromosomes of a band, two observations of gaps or breaks may be detected for each band in a cell. Denote the number of breaks observed at the two homologous chromosomes of the ith band of the jth cell by N_{ij}, $N_{ij} = 0, 1$, or 2, where $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, m$, and let the marginal total $N_i = \Sigma_{j=1}^{m} N_{ij}$ be the total number of breaks observed at the ith band over m cells. The total number of breaks detected in the study is $n = \Sigma_{i=1}^{k} N_i$. Let P_i be the probability of a break, conditional on the event that a break occurs in the ith band in a cell, where $i = 1, 2, \ldots, k$. Then the vector of the observed number of breaks $(N_1, N_2, \ldots, N_k)$ is multinomially distributed as

$$(N_1, N_2, \ldots, N_k) \sim \text{mult}\,(n, k, \mathbf{P}), \qquad (1)$$

where $\mathbf{P} = (P_1, P_2, \ldots, P_k)$. Based on this distribution, Bohm et al. (1995) assumed 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, the k chromosomal sites can be indexed according to their orders in probability of breakage. If the first k_1 ($\leq k$) sites are defined to be nonfragile and the remaining $k - k_1$ sites are defined to be fragile, then probabilities of breakage satisfy

$$P_1 = P_2 \cdots = P_{k_1} < P_{k_1+1} \leq P_{k_1+2} \leq \cdots \leq P_k. \qquad (2)$$

Therefore, testing the probability orders in (2) is equivalent to that of testing the hypothesis

$$\text{H}_0\text{: } P_1 = \frac{1}{k_1}, P_2 = \frac{1}{k_1}, \ldots, P_{k_1} = \frac{1}{k_1}$$

stepwisely at significance level $\alpha/(t+1)$ at the tth iteration (the use of significance level $\alpha/(t+1)$ at the tth iteration is an application of the Bonferroni approach; see Seber, 1977) using the Pearson's chi-square statistic or the likelihood ratio statistics. With these statistics, at each iteration, if this hypothesis of homogeneity is rejected, the site with the highest observed breakage is excluded and the remaining sites are tested for homogeneity. One continues iteratively excluding those sites with the highest number of breaks until a subset

of the data for which the hypothesis of homogeneity cannot be rejected is obtained.

3. An Alternative Procedure

Based on Hochberg's (1988) method, we will develop a new procedure for testing the randomness of chromosomal breakpoints. Part of this new procedure is related to the multiple hypothesis testing.

3.1 *Hochberg's* (1988) *Step-Up Multiple Hypothesis Testing Algorithm*

Let p_i be the p-values of the test for testing $H_{0i}, i = 1, \ldots, k$. For the problem of simultaneously testing k univariate null hypotheses $H_{01}, H_{02}, \ldots, H_{0k}$, another choice is to use the sequential multiple hypothesis testing procedure proposed by Hochberg (1988) as follows:

(i) Order the p-values

$$p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(k)}$$

and label the corresponding hypotheses as

$$H_{0(1)}, H_{0(2)}, \ldots, H_{0(k)}.$$

(ii) Let $i = 1$.

(iii) If $p_{(i)} \leq \alpha/i$, then reject all the remaining hypotheses $H_{0(i)}, \ldots, H_{0(k)}$ and stop.

(iv) If $p_{(i)} > \alpha/i$, then accept $H_{0(i)}$ and add one to i. Return to step (iii).

(A brief description of Hochberg's algorithm is also given in Troendle (1995).)

It is known that Hochberg's procedure can control the probability of a type I error at a predetermined level. Moreover, in dealing with the multiple hypothesis testing problem, Hochberg's procedure is uniformly more powerful than Holm's (1979) procedure and Bonferroni's procedure.

3.2 *An Algorithm for Identifying Chromosomal Fragile Sites*

In this section, we will give an alternative procedure for clustering data with multinomial distribution. The current hierarchical clustering methods can be divided into two major types: agglomerative hierarchical methods and divisive hierarchical methods (Hair et al., 1998; Johnson and Wichern, 1998). Basically, the procedure introduced here can be regarded as a special type of divisive hierarchical methods, which can produce a hierarchical group structure from a complex data set involving the vector of multinomial proportions. Such a grouping method can provide a means for reflecting the relative fragility among all sites and can be applied to other areas.

To simplify our presentation, we start with the equiprobability model. Assume that the k chromosomal sites can be indexed according to the orders of their probabilities of breakage, i.e.,

$$P_1 \leq P_2 \leq \cdots \leq P_k.$$

Intuitively, we may think that fragility is a relative concept and hence the definition of fragile sites defined in (2) can be generalized by setting t groups as follows:

$$\begin{aligned} P_1 = P_2 = \cdots = P_{G_1} \quad & \text{(group 1)} \\ < P_{G_1+1} = P_{G_1+2} = \cdots = P_{G_2} \quad & \text{(group 2)} \\ < P_{G_2+1} = P_{G_2+2} = \cdots = P_{G_3} \quad & \text{(group 3)} \\ \vdots \\ < P_{G_{t-1}+1} = P_{G_{t-1}+2} = \cdots = P_k \quad & \text{(group t)} \end{aligned} \tag{3}$$

The first G_1 sites are defined to be nonfragile sites and the remaining $k - G_1$ sites of the $t - 1$ groups can be considered to be fragile. Among the fragile sites, the sites corresponding to the $(s+1)$th group of probabilities of breakage,

$$\left\{P_{G_s+1}, P_{G_s+2}, \ldots, P_{G_{s+1}}\right\},$$

are more fragile than the ones corresponding to the sth group of probabilities of breakage,

$$\left\{P_{G_{s-1}+1}, P_{G_{s-1}+2}, \ldots, P_{G_s}\right\}.$$

Hence, there exist differences in the measure of fragility among these groups but there is no difference within each group. To search for a grouping structure for a set of breakage data, the following testing procedure is proposed:

(i) Let $t = 1$.

(ii) Let $\alpha^* = \alpha/(t+1)$.

(iii) Simultaneously test the null hypotheses

$$H_{0i}: P_i \leq \frac{1}{k}, \qquad i = 1, 2, \ldots, k,$$

using Hochberg's (1988) algorithm at α^* significance level (for testing $H_{0i}: P_i \leq 1/k$ under a binomial model, the exact p-value $P[N_i \geq n_i \mid P_i = 1/k]$ can be easily obtained using any statistical software).

(iv) If all of the simultaneous hypotheses in step (iii) are not rejected at α^* significance level, then conclude that all the remaining sites are not fragile sites and stop.

(v) If the simultaneous hypotheses in step (iii) are rejected at α^* significance level, then exclude the sites corresponding to the univariate hypotheses that are rejected by using Hochberg's (1988) algorithm. Let h be the number of sites that have been excluded. Set $t = t + 1$, $k = k - h$, and return to step (ii).

Continue the above steps iteratively until a subset of the data is obtained for which we are not able to reject the multiple hypotheses in step (iii) simultaneously. The sites in this set are considered to be nonfragile sites. The other sites are fragile sites. The set of fragile sites is separated into several groups by using the above steps iteratively, and a grouping structure can be obtained. These groups can then be further divided into several subgroups by continuing the same procedure introduced above until we obtain a more explicit grouping structure since, if the random vector $(N_1, N_2, \ldots, N_k)$ is multinomially distributed, the vector composed by any subset of $N_1, N_2, \ldots, N_k$ is still multinomially distributed. Continue the above procedure until no subgroups can be divided; then a grouping structure can be obtained for the sites with corresponding groups of probabilities of breakage as in (3). Obviously, the above procedure can hierarchically classify the set of all sites into several clusters. Among these clusters, there exist differences in the measure of fragility, but within each group, there is no difference. Such a hierarchical clustering technique can be

applied not only to detection of chromosomal fragile sites but also to any statistical problem involving the vector of multinomial proportions.

Let W_i be the width of the ith band in a haploid and let $W = \Sigma_{i=1}^{k} W_i$ be the total width of all bands. Let $\mathbf{P}^0 = (P_1^0, \ldots, P_k^0) = (W_1/W, \ldots, W_k/W)$. We can directly modify the procedure proposed above by replacing $(1/k, 1/k, \ldots, 1/k)$ with $(P_1^0, P_2^0, \ldots, P_k^0)$ to reflect bandwidth and using the following procedure to detect fragile sites under the proportional probability model:

(i) Let $t = 1$.
(ii) Let $\alpha^* = \alpha/(t+1)$.
(iii) Simultaneously test the null hypotheses

$$H_{0i}: P_i \leq P_i^0, \qquad i = 1, 2, \ldots, k,$$

using Hochberg's (1988) algorithm at α^* significance level (the exact p-value $P[N_i \geq n_i \mid P_i = P_i^0]$ can be easily obtained using any statistical software).
(iv) If all of the simultaneous hypotheses in step (iii) are not rejected at α^* significance level, then conclude that all the remaining sites are not fragile sites and stop.
(v) If the simultaneous hypotheses in step (iii) are rejected at α^* significance level, then exclude the sites corresponding to the univariate hypotheses that are rejected by using Hochberg's (1988) algorithm. Let h be the number of sites that have been excluded from the set of nonfragile sites. Set $t = t + 1$ and $k = k - h$ and recalculate $(W_1, \ldots, W_k)$ and

$$(P_1^0, \ldots, P_k^0) = \left(\frac{W_1}{\Sigma W_i}, \ldots, \frac{W_k}{\Sigma W_i}\right).$$

Return to step (ii).

4. Simulation Studies

Imposing the observed orders on the null hypothesis of EPM, it is acceptable using Bohm et al.'s method to exclude a band of the highest observed rank stepwisely 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 in their method is not applicable (Hou et al., 1999). In order to investigate the performance of correct identification of Bohm et al.'s method and our method under PPM, in this section, we will perform a series of simulation studies. Here we define the correct identification rate (CIR) as

CIR = (the total number of sites that are correctly detected as fragile or nonfragile)
÷ (the total number of sites)

to compare the two methods. The procedure to generate 10,000 multinomial samples under different models for estimating the rate of correct identification is as follows:

(i) For any data set $(N_1, N_2, \ldots, N_k)$, estimate P_j's by N_j/n and use them as if they were the true population proportions.
(ii) Simulate 10,000 multinomial samples from this population and, for each sample, obtain the rate of correct identification. Calculate the mean of these rates.

Table 2
Data sets of band structures

Data set	Band	Group structure
(1) Two-bands	$\mathbf{B} = (B_1, B_2)$	$P_1 \leq P_2$
(2) Four-bands	$\mathbf{B} = (B_1, \ldots, B_4)$	$P_1 = P_2 = P_3 \leq P_4$
(3) Four-bands	$\mathbf{B} = (B_1, \ldots, B_4)$	$P_1 = P_2 \leq P_3 = P_4$
(4) Four-bands	$\mathbf{B} = (B_1, \ldots, B_4)$	$P_1 \leq P_2 = P_3 = P_4$
(5) Ten-bands	$\mathbf{B} = (B_1, \ldots, B_{10})$	$P_1 = \cdots = P_5 \leq P_6 = \cdots = P_{10}$
(6) Fifty-bands	$\mathbf{B} = (B_1, \ldots, B_{50})$	$P_1 = \cdots = P_{25} \leq P_{26} = \cdots = P_{50}$

It is not feasible to study all the possible data structures, so here we consider six data sets of different band structures depending on the number of bands and group structures. They are summarized in Table 2. For each data set, 10,000 multinomial samples of either 100 or 500 sample sizes are generated. The computer codes are all written in FORTRAN V. The simulated multinomial samples are generated by the IMSL subroutine RNMTN. Results of this analysis are listed in Table 3.

In Table 3, it can be found that the average CIRs of our new procedure introduced here are always near one, but those of the method proposed by Bohm et al. (1995) are often near zero. The conclusion results from the failure of their procedure to adequately detect positive outlying cells (i.e., fragile sites) simultaneously with the rejection of the null hypothesis under the PPM since, at each iteration, they test the null hypothesis $H_{0i}: P_i \leq P_i^0$, $i = 1, 2, \ldots, k$, using Pearson's chi-square statistic or the likelihood ratio statistics and, if this hypothesis is rejected, declare that the site with the highest observed breakage is a fragile site. However, when the null hypothesis is rejected using their procedure, the only conclusion that can be made is that the event $\cap_{i=1}^{k}\{P_i \leq P_i^0\}$ is not true. We cannot pinpoint which of the k univariate events $\{P_i \leq P_i^0\}$, $i = 1, 2, \ldots, k$, are the cause and thus it is not reasonable to declare the site with the highest observed breakage as a fragile site using their procedure.

5. A Real Example

In this section, we reanalyzed the data set given in Wang (1992) using the new procedure introduced in Section 3. 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 is given in Tai et al. (1993)). The results under the EPM and the PPM assumption are given in Table 1.

As shown in Table 1, employing our approach, 37 fragile sites were detected at a predetermined significance level of 0.05 under the PPM assumption. Of these sites, 33 are listed in HGM10 (the report of the Tenth International Workshop on Human Gene Mapping; see Sutherland and Ledbetter, 1989). Employing our procedure, 34 fragile sites were detected at a significance level of 0.05 under the EPM assumption. Among these 34 sites, 32 are listed in HGM10. To save space, we omit part of the detected fragile sites in Table 1. Employ-

Table 3

Numerical comparisons of the correct identification rate (CIR) among two different procedures[a]

Data structure	Description of data set	α	Average CIR: Bohm et al.	Average CIR: Our method
(1) Two-bands $P_1 \le P_2$	$K=2$, $\mathbf{B}=(B_1,B_2)$; $\mathbf{N}=(10,90)$; $nt=100$; $\mathbf{P}=(0.1,0.9)$; $\mathbf{W}=(1,99)$; $\mathbf{P^0}=(0.01,0.99)$; $\mathbf{R}=(10,0.09)$; $F=\{B_1\}$, $NF=\{B_2\}$	0.01	0.0036	0.9708
		0.05	0.0036	0.9881
		0.10	0.0010	0.9965
(2) Four-bands $P_1=P_2=P_3 \le P_4$	$K=4$, $\mathbf{B}=(B_1,\ldots,B_4)$; $\mathbf{N}=(24,24,24,28)$; $nt=100$; $\mathbf{P}=(0.24,0.24,0.24,0.28)$; $\mathbf{W}=(10,10,10,70)$; $\mathbf{P^0}=(0.1,0.1,0.1,0.7)$; $\mathbf{R}=(2.4,2.4,2.4,.4)$; $F=\{B_1,B_2,B_3\}$, $NF=\{B_4\}$	0.01	0.4250	0.9942
		0.05	0.4438	0.9988
		0.10	0.4217	0.9993
(3) Four-bands $P_1=P_2 \le P_3=P_4$	$K=4$, $\mathbf{B}=(B_1,\ldots,B_4)$; $\mathbf{N}=(10,10,40,40)$; $nt=100$; $\mathbf{P}=(0.1,0.1,0.4,0.4)$; $\mathbf{W}=(1,1,49,49)$; $\mathbf{P^0}=(0.01,0.01,0.49,0.49)$; $\mathbf{R}=(10,10,40/49,40/49)$; $F=\{B_1,B_2\}$, $NF=\{B_3,B_4\}$	0.01	0.2500	0.9699
		0.05	0.2500	0.9844
		0.10	0.2500	0.9822
(4) Four-bands $P_1 \le P_2=P_3=P_4$	$K=4$, $\mathbf{B}=(B_1,\ldots,B_4)$; $\mathbf{N}=(10,30,30,30)$; $nt=100$; $\mathbf{P}=(0.1,0.3,0.3,0.3)$; $\mathbf{W}=(1,33,33,33)$; $\mathbf{P^0}=(0.01,0.33,0.33,0.33)$; $\mathbf{R}=(10,10/11,10/11,10/11)$; $F=\{B_1\}$, $NF=\{B_2,B_3,B_4\}$	0.01	0.0180	0.9854
		0.05	0.0121	0.9910
		0.10	0.0070	0.9873
(5) Ten-bands $P_1=\cdots=P_5 \le P_6=\cdots=P_{10}$	$K=10$, $\mathbf{B}=(B_1,\ldots,B_{10})$; $N_1=\cdots=N_5=5, N_6=\cdots=N_{10}=15$; $nt=100$; $P_1=\cdots=P_5=0.05$, $P_6=\cdots=P_{10}=0.15$; $W_1=\cdots=W_5=0.1$, $W_6=\cdots=W_{10}=19.9$; $P_1^0=\cdots=P_5^0=0.001$, $P_6^0=\cdots=P_{10}^0=0.199$; $\mathbf{R}=(50,\ldots,50,150/199,\ldots,150/199)$; $F=\{B_1,\ldots,B_5\}$, $NF=\{B_6,\ldots,B_{10}\}$	0.01	0.3996	0.9409
		0.05	0.3997	0.9491
		0.10	0.3999	0.9795
(6) Fifty-bands $P_1=\cdots=P_{25} \le P_{26}=\cdots=P_{50}$	$K=50$, $\mathbf{B}=(B_1,\ldots,B_{50})$; $N_1=\cdots=N_{25}=5, N_{26}=\cdots=N_{50}=15$; $nt=500$; $P_1=\cdots=P_{25}=0.01$, $P_{26}=\cdots=P_{50}=0.03$; $W_1=\cdots=W_{25}=0.1$, $W_{26}=\cdots=W_{50}=19.9$; $P_1^0=\cdots=P_{25}^0=0.0002$, $P_{26}^0=\cdots=P_{50}^0=.0398$; $\mathbf{R}=(50,\ldots,50,300/398,\ldots,300/398)$; $F=\{B_1,\ldots,B_{25}\}$, $NF=\{B_{26},\ldots,B_{50}\}$	0.10	0.0000	0.9381
		0.05	0.0000	0.9380
		0.10	0.0000	0.9380

[a] B_i = band i; $nt=\Sigma_{i=1}^{k} N_i$; $\mathbf{P}$ = vector of breakage proportions; $\mathbf{W}$ = vector of bandwidth; $\mathbf{P^0}$ = breakage proportion under PPM; $\mathbf{R}=(R_1,\ldots,R_k)$, $R_i=P_i/P_i^0$, $i=1,2,\ldots,k$; F = set of fragile sites; NF = set of nonfragile sites.

ing the procedure of Bohm et al. (1995), 74 fragile sites were detected at a significance level of 0.05 under the EPM assumption. Each site with the number of breakages larger than or equal to four will be identified as fragile using their procedure.

6. Conclusion

Bohm et al. (1995) concluded that their testing procedure can be applied to detect fragile sites under the PPM assumption. However, as Hou et al. (1999) point out, their conclusion is incorrect. Their procedure cannot be utilized to detect fragile sites under the PPM assumption. Furthermore, 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 chi-square approximation to Peason's chi-square or likelihood ratio test statistics tends to be poor for sparse tables containing both small and moderately large expected frequencies.

In this article, we introduce a new procedure that detects fragile sites from a hierarchical-clustering point of view. This new procedure can be applied to tests of the nonrandomness of breakpoints under either the proportional probability model or the equiprobability model. Moreover, it is known that Hochberg's (1988) procedure used in each step of our procedure can control the probability of a type I error at a predetermined significance level, which is always true regardless of the sample size and the data structure involved in a study. Our method, introduced in Section 3, can be used not only to identify fragile sites but also to classify the breakage data into several clusters and to form a meaningful group

structure. Such a grouping technique provides a tool to separate all categories of a multinomially distributed population into several clusters and produces a group structure. It can be applied in any statistical problem involving multinomially distributed population.

Acknowledgements

The authors wish to thank a reviewer and the associate editor for their helpful comments that improved this article. This study was supported by grant NSC 89-2118-M-030-006 (HCD) and NSC 89-2118-M-002-012 (JJT) from the National Science Council, Taiwan, R.O.C.

Résumé

L'identification des sites fragiles est une façon d'explorer les anomalies génétiques qui sont la marque du cancer et qui jouent un rôle important dans la carcinogénèse. L'observation d'une cassure non aléatoire sur la bande d'un chromosome est un critère essentiel pour déterminer la fragilité de la bande. Dans ce papier, une nouvelle procédure de détection est proposée. Cette nouvelle procédure prend en considération la relation d'un site avec les autres sites, et peut être appliquée aux tests de survenue aléatoire des points de cassure sous un modèle de probabilité proportionnelle ou sous un modèle d'équiprobabilité. La procédure peut former une structure de regroupement qui classifie les sites en plusieurs clusters. La méthode proposée est illustrée par son application à l'identification des sites fragiles sur des données réelles portant sur des patients chinois atteints de cancer colo-rectal.

References

Bohm, U., Dahm, P. F., McAllister, B. F., and Greenbaum, I. F. (1995). Identifying chromosomal fragile sites from individuals: A multinomial statistical model. *Human Genetics* **95,** 249–256.

Chen, C.-H., Shih, H.-H., Wang-Wuu, S., Tai, J. J., and Wuu, K.-D. (1998). Chromosomal fragile site expression in lymphocytes from patients with schizophrenia. *Human Genetics* **103,** 702–706.

De Braekeleer, M. and Smith, B. (1988). Two methods for measuring the non-randomness of chromosome abnormalities. *Annals of Human Genetics* **52,** 63–67.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate Data Analysis.* London: Prentice-Hall.

Hecht, F. and Sutherland, G. R. (1984). Fragile sites and cancer breakpoints. *Cancer Genetics and Cytogenetics* **12,** 179–181.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

Hou, C.-D., Chiang, J., and Tai, J. J. (1999). Testing the nonrandomness of chromosomal breakpoints using highest observed breakage. *Human Genetics* **104,** 350–355.

ISCN. (1981). An international system for human cytogenetic nomenclature—High resolution banding. *Cytogenetics and Cell Genetics* **31,** 1–23.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis.* London: Prentice-Hall.

Koehler, K. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* **75,** 336–344.

Le Beau, M. M. (1986). Chromosomal fragile sites and cancer-specific rearrangements. *Blood* **67,** 849–858.

Mariani, T. (1989). Fragile sites and statistics. *Human Genetics* **81,** 319–322.

Ried, K., Finnis, M., Hobson, L., et al. (2000). Common chromosomal fragile site FRA16D sequence: Identification of the FOR gene spanning FRA16D and homozygous deletions and translocation breakpoints in cancer cells. *Human Molecular Genetics* **9,** 1651–1663.

Seber, G. A. (1977). *Linear Regression Analysis.* New York: Wiley.

Smith, C. A. B. (1986). Chi-squared tests with small numbers. *Annals of Human Genetics* **50,** 163–167.

Sozzi, G., Veronese, M. L., Negrini, M., et al. (1996). The FIHT gene at 3p14.2 is abnormal in lung cancer. *Cell* **85,** 17–26.

Sutherland, G. R. and Hecht, F. (1985). *Fragile Sites on Human Chromosomes.* Oxford: Oxford University Press.

Sutherland, G. R. and Ledbetter, D. H. (1989). Report of the committee on cytogenetic markers. Tenth International Workshop on Human Gene Mapping. *Cytogenetics and Cell Genetics* **51,** 452–458.

Tai, J. J. and Hou, C.-D. (1999). Methodological development in analysis of cytogenetic data. *Journal of Genetics and Molecular Biology* **10,** 113–118.

Tai, J. J., Hou, C.-D., Wang-Wuu, S., Wang, C.-H., Leu, S.-Y., and Wuu, K.-D. (1993). A method for testing the nonrandomness of chromosomal breakpoints. *Cytogenetics and Cell Genetics* **63,** 147–150.

Tai, J. J., Hou, C-D., and Wang-Wuu, S. (1998). A confirmation analysis method for identification of chromosomal fragile sites. *Cancer Genetics and Cytogenetics* **105,** 1–5.

Troendle, J. F. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association* **90,** 370–378.

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.

Received April 1999. *Revised October* 2000.
Accepted November 2000.