*Statistics in Molecular Biology*
IMS Lecture Notes – Monograph Series (1999) Volume 33

# A CONDITIONAL APPROACH TO THE DETECTION OF CORRELATED MUTATIONS

BY MAHA C. KARNOUB, FRANCOISE SEILLIER-MOISEIWITSCH[1] AND PRANAB K. SEN

*Glaxo Wellcome Inc., Research Triangle Park, NC and University of North Carolina at Chapel Hill*

Some genomes mutate quickly. Studying their mutation process may allow us to understand the selection pressures these genomes are undergoing. Considering simultaneous mutations at several sites may provide insights into protein structure. We consider the situation where the frequency table for the amino-acid pairs can be summarized by a $2 \times 2$ table. We develop a test for mutational linkage between two positions conditionally on the consensus pair. We illustrate the use of this test with sequences from the V3 loop of the envelope gene from the human immunodeficiency virus.

**1. Introduction.** The genome of a retrovirus like the human immunodeficiency virus (HIV) evolves at a fast pace: the high mutation rate (due to the error-prone reverse transcriptase and recombination between the two RNA strands) is compounded to the high rate of replication ($\sim$ 300 cycles of replication per year). Some of these substitutions confer a survival advantage by enabling some mutants to escape the immune system. Some may cause a phenotypic change (such as cell tropism and virulence). These random mutations thus persist in the viral population. Some subsist only when a substitution at another position occurs. These *linked* mutations may simply maintain structure (and thus viability) or may again be beneficial to the virus.

When the structure of a viral protein is unknown, detecting such double mutations may help in inferring pairs of amino acids that interact and are therefore more likely to be in close spatial proximity. Further, for sequence data, more often than not the probabilistic model underlying analytical methods relies on the assumption that positions undergo independent mutation processes. The methodology introduced in this paper can serve to check this assumption. Its violation leads to the overestimation of genetic distances and thus to erroneous phylogenetic reconstructions [Seillier-Moiseiwitsch et al. (1998)].

In Section 2, we consider specific pairs of sites and test the number of dou-

[1]Her research was funded in part by the National Science Foundation (DMS-9305588), the American Foundation for AIDS Research (70428-15-RF) and the National Institutes of Health (R29-GM49804 and P30-HD37260).

*AMS* 1991 *subject classifications.* Primary 62F03 ; secondary 62P10, 92D20.
*Key words and phrases.* Correlation, independence, mutation, phylogeny.

TABLE 1
*Contingency table for sequence data.*

| | | Position 2 | | |
|---|---|---|---|---|
| | | $V$ | $W$ | |
| Position 1 | $D$ | $n_{11}$, $p_{11}$ consensus | $n_{12}$ $p_{12}$ | $n_{1.}$ $p_{1.}$ |
| | $E$ | $n_{21}$ $p_{21}$ | $n_{22}$ $p_{22}$ | $n_{2.}$ $p_{2.}$ |
| | | $n_{.1}$ $p_{.1}$ | $n_{.2}$ $p_{.2}$ | $n$ |

ble mutations away from the *consensus* (i.e., the most frequent configuration) conditioning on the total number of sequences and on the consensus pair; a conventional test for independence is appraised in this context. The methodological developments are presented in Sections 3, 4, and 5, with some details relegated to the Appendix. We present some simulation results in Section 5 and analyze a set of HIV-1 sequences (Section 6).

**2. The set-up.** Consider a specific pair of sites along the sequences, say Position 1 and Position 2. Assume that at Position 1, across all sequences, amino acids $D$ and $E$ are present while Position 2 exhibits $V$ and $W$. In the contingency table summarizing the data (Table 1), the $(1,1)$-cell contains the number of sequences with the consensus configuration. Let $n_{ij}$ be the number of sequences in row $i$ (Position 1) and column $j$ (Position 2), and $p_{ij}$ the probability of having this configuration. Any sequence in the first row or the first column but not in the $(1,1)$-cell sustained a single mutation away from the consensus. The others had two mutations.

In the context of viral sequences, it is reasonable to treat these sequences as independent, at least as a first assumption. Indeed, replication cycles are short, and each cycle generates a number of substitutions. Hence, when all sequences are sampled from different individuals, many rounds of replication separate any two viruses. Each position had many opportunities to mutate. Also, functionality of the resulting protein drives viral survival. Thus, whether a specific position is allowed or required to change depends on the amino-acid composition at other locations. In a sense, the consensus is "rediscovered" after

each alteration through structural linkage. This selection pressure and the high viral turn-over overwhelm ancestral relationships.

We are interested in testing excess or paucity of double mutations under the assumption of independence of mutations at the two positions. The random variable of interest is $N_{22}$, the number of sequences with double mutations. In Fisher's exact (conditional) test [Bishop et al. (1975), p.364], the marginal totals $n_{1.}$, $n_{.1}$, $n_{2.}$, and $n_{.2}$ are conditioned upon and the conditional probability of the observed counts is given by the hypergeometric law

$$Pr\,(N_{ij} = n_{ij},\, i,j = 1,2 \mid n_{.1}, n_{1.}, n) = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \quad .$$

The exact one- and two-sided tests using the collection of all possible tables with these marginal totals and this probability distribution can be constructed. Alternatively, a large-sample (normal) approximation can be applied as follows.

The estimate of the expected value of $N_{22}$ is

$$\widehat{E}\,(N_{22}) = n\,\widehat{p}_{22}\,.$$

Under the null hypothesis of independence among positions,

$$\widehat{p}_{22} = \frac{n_{2.}\, n_{.2}}{n^2} \quad \text{and} \quad \widehat{E}\,(N_{22}) = \frac{n_{2.}\, n_{.2}}{n}\,.$$

The exact variance conditional on the marginal totals is

$$Var\left(N_{22} - \widehat{E}\,(N_{22})\right) = \frac{n_{1.}\, n_{.1}\, n_{2.}\, n_{.2}}{n^2\,(n-1)}\,.$$

The test statistic is thus

$$T_n = \left(N_{22} - \widehat{E}\,(N_{22})\right) \Big/ \sqrt{\frac{n_{1.}\, n_{.1}\, n_{2.}\, n_{.2}}{n^2\,(n-1)}}\,.$$

Under $H_0$, $\qquad T_n \sim \mathcal{N}\,(0,1)$,
and the critical level of this test statistic is close to the normal percentile point corresponding to the significance level.

In the present context, the marginal totals are not fixed as we do not have any knowledge about the total number of mutations with a specific nucleotide or amino acid at any one of the positions. Since in our set-up, we condition on the consensus cell frequency $N_{11}$ $(= n_{11})$ (see Section 3 for further motivation); hence, Fisher's exact (conditional) test is not appropriate here. However, we may add that as the conditional distribution of $T_n$, given the marginal totals, is asymptotically $\mathcal{N}\,(0,1)$, in probability (under $H_0$), and thus free of the marginal

totals, the asymptotic unconditional null distribution of $T_n$ is also $\mathcal{N}(0,1)$. Consequently, a test based on $T_n$ will have a specified level of significance if the $N_{ij}$'s are all large. Nevertheless, in terms of power properties it might not compare favourably with alternative tests based on the conditional distribution given $N_{11}$ and $N$. In passing, we may remark that the traditional optimality (viz., UMP) property that may be attributed to Fisher's exact test (against some specific parametric alternative) may not be tenable in our set-up where the alternative hypotheses are not only of more complex nature but also not entirely of a parametric flavour. In fact, in our case, a UMP test may not exist. For this reason, we have recourse to a direct approach based on the conditional distribution, given $N_{11}$ and $N$ along with the additional information that $N_{11}$ is the maximum cell-count among the four cells. This is the main theme of the current study.

**3. Distribution of the cell counts.** The standard distributions associated with cell counts in contingency tables are [Bishop et al. (1975)]: the Poisson model, obtained with a sampling plan that has no restrictions on the total sample size, the multinomial model with a fixed total sample size, and independent multinomial distributions for the rows (with fixed row totals) or independent multinomial distributions for the columns (with fixed column totals). For these sampling models, the marginal totals are sufficient statistics for testing the independence of two factors. Further, under the assumption of independence of the factors, the maximum-likelihood estimates under the above sampling processes exist, are unique and, if none of the marginal totals is 0, they are equal. In fact, when the total sample size is fixed, the multinomial and the Poisson schemes are equivalent [Bishop et al. (1975)]. The Poisson model is usually prefered when some of the events are rare, i.e. some of the cell counts are small. In view of the nature of our data, we adopt this model here. Indeed, the measured average error rate per site for HIV-1 reverse transcriptase is between $10^{-4}$ and $10^{-3}$.

We let $\mathbf{N} = (N_{11}, N_{12}, N_{21}, N_{22})'$, and make the following assumptions:

1. the cell counts $N_{ij}$ are independent and have Poisson distributions with parameters $\lambda_{ij}$, i.e.,

$$P\{\mathbf{N} = \mathbf{n}\} = \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{e^{-\lambda_{ij}}\ \lambda_{ij}^{n_{ij}}}{n_{ij}!} \qquad \mathbf{n} \geq \mathbf{0} \tag{3.1}$$

2. $\lambda_{11} \gg \lambda_{ij}$ for all $(i,j) \neq (1,1)$ and the $\lambda_{ij}$'s are large (e.g., greater than 5).

These two assumptions basically ensure that (i) $N_{11}$ is a maximum with probability close to 1, and (ii) the multinormality approximations hold for the distribution of the $N_{ij}$'s.

Let $\lambda = \sum_{i,j} \lambda_{ij}$. We consider the null hypothesis (of independence of mutations)

$H_0 : \lambda_{ij} = \lambda\, \alpha_i\, \beta_j$ for all $(i,j)$, $\alpha_1 + \alpha_2 = 1$, $\beta_1 + \beta_2 = 1$,
against the alternative hypothesis

$H_a$ : $\lambda_{ij}$'s are not factorizable this way,
i.e., the mutations at the two positions are not stochastically independent.

We test the above hypothesis conditionally on the consensus pair. Conditioning on the consensus pair allows us to identify the polymorphisms (i.e., the pairs made up of the consensus amino acid at one position and a mutation at the other position) and the double mutations. The goal of the test is to identify pairs of positions that show a propensity to change together. The reason is two-fold. First, the consensus pair codes for a part of the protein that plays its function well. Departures at one or the other of the two amino acids (but not at both simultaneously) indicate changes that do not disturb the structure so much that the product is no longer a functioning protein. These single departures can be viewed as "noise". Double mutations, on the other hand, come about either as a rescue mechanism for a slightly deleterious single mutation or as a change in phenotype (acquisiton of a different cell tropism, for example). Hence, identifying correlated substitutions serves as an exploratory investigation of structure modifications and thus of alterations in phenotypes. Second, usually, positions that interact are in the same three-dimensional vicinity. Hence, from these correlated pairs, one can infer some information about the structure of the protein. For these two purposes determining the consensus pair is crucial to identifying changes. Further, due to genetic drift and independent selection pressures at each position, $\alpha_1$ and $\beta_1$ are not constant over time. Thus, conditioning on $N$ and $N_{11}$ enable us to separate these effects from selection that act on both positions simultaneously.

Let $n_* = n - n_{11}$ and $\lambda_* = \lambda - \lambda_{11}$. Under Assumptions 1 and 2 above, we show that if the $\lambda_{ij}$'s are large

$$P\{N_{11} > \max(N_{12}, N_{21}, N_{22})\} \to 1 \tag{3.2}$$

(see Appendix). Next, consider the joint distribution of the cell counts, given $N_{11} = n_{11}$ and the fixed total sample size $n$:

$$\begin{aligned} &P\{N_{11} = n_{11},\ N_{12} = n_{12},\ N_{21} = n_{21},\ N_{22} = n_{22} \mid N_{11} = n_{11}\} \\ &\quad = P\{N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}\} \\ &\quad = \quad e^{-\lambda_*} \left( \frac{\lambda_{12}^{n_{12}}\, \lambda_{21}^{n_{21}}\, \lambda_{22}^{n_{22}}}{n_{12}!\, n_{21}!\, n_{22}!} \right), \end{aligned}$$

which follows readily from (3.1). As a result, the distribution of $N_* = N_{21} + N_{12} + N_{22}$, given $N_{11} = n_{11}$, is

$$P\{N_* = n_* \mid N_{11} = n_{11}\} = \frac{e^{-\lambda_*}\,\lambda_*^{n_*}}{n_*!}\,.$$

Further, note that given $N_{11} = n_{11}, N_* = n_*$, we have $N = n = n_* + n_{11}$. Hence, from the above two equations, we obtain that
$P\{N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11}\ \&\ N = n\}$

$$= \frac{n_*!}{n_{12}!\,n_{21}!\,n_{22}!}\left(\frac{\lambda_{12}}{\lambda_*}\right)^{n_{12}}\left(\frac{\lambda_{21}}{\lambda_*}\right)^{n_{21}}\left(\frac{\lambda_{22}}{\lambda_*}\right)^{n_{22}}. \tag{3.3}$$

This is a trinomial distribution for $\mathbf{N}_* = (N_{12}, N_{21}, N_{22})'$, with parameters $n_*$ and $\frac{\lambda_{12}}{\lambda_*}$, $\frac{\lambda_{21}}{\lambda_*}$, and $\frac{\lambda_{22}}{\lambda_*}$. Let $\nu_{ij} = \frac{\lambda_{ij}}{\lambda_*}$ for $(i,j) \neq (1,1)$, $\boldsymbol{\nu} = (\nu_{12}, \nu_{21}, \nu_{22})'$ and $\boldsymbol{D} = \mathrm{Diag}(\nu_{12}, \nu_{21}, \nu_{22})$. Then

$$\begin{aligned} E(\boldsymbol{N}_* \mid N_{11} = n_{11}\ ,\ N_* = n_*) &= n_*\,\boldsymbol{\nu} \\ Var(\boldsymbol{N}_* \mid N_{11} = n_{11}\ ,\ N_* = n_*) &= n_*\,[\boldsymbol{D} - \boldsymbol{\nu}\boldsymbol{\nu}'] \end{aligned} \tag{3.4}$$

At this stage, we invoke (3.2) and (3.3), and claim that as $n$ increases,

$$\begin{aligned} &P\{N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} | N_{11} = n_{11}, N_{11} = \max \text{ and } N = n\} \\ &\approx P\{(N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11}, N_* = n_*\}. \end{aligned} \tag{3.5}$$

Therefore, the moment results in (3.4) can also be shown to be good approximations for the conditional case where in addition $N_{11}$ is the maximum.

**4. Parameter estimation.** By virtue of (3.5), under Assumptions 1 and 2, we shall work with the approximate likelihood function (given $N_{11} = n_{11}$, $N_* = n_*$ and that $N_{11}$ is the maximum cell count among the four cells):

$$\begin{aligned} &P\{N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22} \mid N_{11} = n_{11} = \max,\ N = n\} \\ &\equiv \frac{n_*!}{n_{12}!\,n_{21}!\,n_{22}!}\,\nu_{12}^{n_{12}}\,\nu_{21}^{n_{21}}\,\nu_{22}^{n_{22}}. \end{aligned} \tag{4.1}$$

Recall that, under $H_0$, $\lambda_{ij} = \lambda\,\alpha_i\,\beta_j$ for $i = 1, 2$, and $j = 1, 2$. Without any loss of generality, we set $\alpha_1 + \alpha_2 = 1 = \beta_1 + \beta_2$. With this simplification, we can express $\nu_{ij} = \lambda_{ij}/\lambda_*$ in terms of the two unknowns, $\alpha_1$ and $\beta_1$. The trinomial law in (4.1) has effectively two degrees of freedom (DF) for a goodness-of-fit (GOF) test statistic. The conventional Pearsonian GOF test for the conditional law in

(4.1) would result in 0 DF, and hence, would not be usable. To eliminate this impasse, we recall that $N_{11}$ is held fixed (along with other statistical information contained in its marginal distribution), and further that

$$P\{N_{11} = n_{11} \mid N = n\} = \binom{n}{n_{11}} \left(\frac{\lambda_{11}}{\lambda}\right)^{n_{11}} \left(1 - \frac{\lambda_{11}}{\lambda}\right)^{n - n_{11}},$$

which under the null hypothesis $H_0$ reduces to

$$P\{N_{11} = n_{11} \mid N = n, H_0\} = \binom{n}{n_{11}} (\alpha_1\beta_1)^{n_{11}} (1 - \alpha_1\beta_1)^{n_*}. \tag{4.2}$$

Therefore, letting $\theta = \alpha_1\beta_1$, from (4.2), we obtain a MLE or BAN estimator of $\theta$ :

$$\hat{\theta}_n = \frac{n_{11}}{n}.$$

As such, we work with the conditional model in (4.1) incorporating the additional restraint that

$$\alpha_1\beta_1 = \hat{\theta}_n = \frac{n_{11}}{n}. \tag{4.3}$$

Now we have effectively only one unknown parameter, $\alpha_1$ say, and hence, the classical inference theory for categorical models can be called upon.

The log-likelihood of $\alpha_1$, based on (4.1) and the constraint (4.3), is

$$L_0(\alpha_1) = C + n_{12} \log \alpha_1 + n_{12} \log \left(1 - \frac{\hat{\theta}_n}{\alpha_1}\right) + n_{21} \log (1 - \alpha_1)$$

$$+ n_{21} \log \frac{\hat{\theta}_n}{\alpha_1} + n_{22} \log (1 - \alpha_1) + n_{22} \log \left(1 - \frac{\hat{\theta}_n}{\alpha_1}\right),$$

where $C$ comprises the terms that do not depend on $\alpha_1$. Then

$$\frac{\partial L_0(\alpha_1)}{\partial \alpha_1} = -\frac{n_{2.}}{\alpha_1} + \frac{n_{.2}}{\alpha_1 - \hat{\theta}_n} - \frac{n_{2.}}{1 - \alpha_1} = \frac{n_{.2}}{\alpha_1 - \hat{\theta}_n} - \frac{n_{2.}}{\alpha_1 (1 - \alpha_1)},$$

$$\frac{\partial^2 L_0(\alpha_1)}{\partial {\alpha_1}^2} = \frac{n_{2.}}{{\alpha_1}^2} - \frac{n_{.2}}{\left(\alpha_1 - \hat{\theta}_n\right)^2} - \frac{n_{2.}}{(1 - \alpha_1)^2}. \tag{4.4}$$

In the region where $\frac{n_{2.}}{n} = \alpha_2 + O_p\left(1/\sqrt{n}\right)$ and $\frac{n_{.2}}{n} = \beta_2 + O_p\left(1/\sqrt{n}\right)$, it follows by routine computations that

$$\frac{\partial^2 L_0(\alpha_1)}{\partial {\alpha_1}^2} < 0, \text{ in probability as } n \to \infty.$$

We obtain the solution to the estimating equation in (4.4) as (in probability)

$$\tilde{\alpha}_1 = \frac{1}{2}\left\{1 - \frac{n_{2.}}{n_{.2}} + \sqrt{\left(1 - \frac{n_{2.}}{n_{.2}}\right)^2 + 4\hat{\theta}_n \frac{n_{2.}}{n_{.2}}}\right\}. \tag{4.5}$$

Further,

$$\tilde{\alpha}_2 = 1 - \tilde{\alpha}_1 \ , \ \tilde{\beta}_1 = \frac{\hat{\theta}_n}{\tilde{\alpha}_1} \ , \ \tilde{\beta}_2 = 1 - \tilde{\beta}_1.$$

We incorporate these estimators in our proposed test statistic in Section 5.

**5. The new binomial test.** Our proposed test statistic, based on the restraint that

$$\tilde{\alpha}_1\tilde{\beta}_1 = \hat{\theta}_n = \frac{n_{11}}{n},$$

is

$$\tilde{Z}_{22} = \frac{1}{\sqrt{n}}\left(N_{22} - n_* \frac{\tilde{\alpha}_2\tilde{\beta}_2}{\left(1 - \tilde{\alpha}_1\tilde{\beta}_1\right)}\right) = \frac{1}{\sqrt{n}}\left(N_{22} - n\,\tilde{\alpha}_2\tilde{\beta}_2\right) \ .$$

Note that the original 2×2 table has 3 DF, and hence, having the estimators $\hat{\theta}_n$ and $\tilde{\alpha}_1$ ($\tilde{\beta}_1 = \hat{\theta}_n/\tilde{\alpha}_1$), we have effectively one DF. For this reason, it suffices to use only either $\tilde{Z}_{22}$ or $\tilde{Z}_{12}$ or $\tilde{Z}_{21}$, defined analogously. However, since we are interested in double mutations, $\tilde{Z}_{22}$ is intuitively more appealing.

We may provide a natural interpretation of $\tilde{Z}_{22}$ in terms of the $\kappa$ coefficient for agreement (no mutation or double mutation, in our set-up) for categorical data. Following Cohen (1960) [see also Landis & Koch (1977)], we define

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_o}$$

where $\pi_o$ is the probability of no or double mutation and $\pi_e$ is the hypothetical probability of the same under the baseline constraints that $p_{11} = \alpha_1\beta_1$ and $p_{22} = (1 - \alpha_1)(1 - \beta_1)$ $(= \alpha_2\beta_2)$. Thus, here

$$\pi_o = p_{11} + p_{22} \quad \text{and} \quad \pi_e = p_{11} + (1 - \alpha_1)(1 - \beta_1) \ ,$$

so that $1 - \pi_e = \alpha_1 + \beta_1 - 2\alpha_1\beta_1 = \alpha_1\alpha_2 + \beta_1\beta_2 + (\alpha_1 - \beta_1)^2$ . Note that by virtue of the consensus pairing, $p_{11} = \alpha_1\beta_1$, $\tilde{Z}_{22}$ is the sample counterpart of the numerator of $\kappa$. As we shall see later, though we have not included the

denominator, the factor $(\alpha_1 + \beta_1 - 2\alpha_1\beta_1)$ shows up in the sampling variance of $\widetilde{Z}_{22}$. The main reason for not including the denominator in $\widetilde{Z}_{22}$ is that $(\alpha_1 + \beta_1 - 2\alpha_1\beta_1)$ can be very small when both $\alpha_1$ and $\beta_1$ are close to 1 (as is the case here), which might make the $\kappa$ coefficient look much inflated.

In order to find the critical level for the test statistic, we need to obtain an expression for its sampling variance under $H_0$. Let

$$Z_{ij} = \frac{1}{\sqrt{n}} \left( N_{ij} - n \frac{\lambda_{ij}}{\lambda} \right), \quad i = 1,\ 2, \quad j = 1,\ 2.$$

Note that under $H_0$

$$Z_{ij} = \frac{1}{\sqrt{n}} \left( N_{ij} - n\, \alpha_i\, \beta_j \right) .$$

Hence, appealing to the multinomial law in (3.1) , we have under $H_0$,

$$Z_{ij} \sim \mathcal{N}\left(0, \alpha_i\, \beta_j\, (1 - \alpha_i\, \beta_j)\right), \quad i = 1,\ 2, \quad j = 1,\ 2.$$

Let $U_n = \dfrac{N_{2.}}{N_{.2}}$. Noting that by (4.5), $U_n \widetilde{\beta}_2 = \widetilde{\alpha}_2$, we rewrite

$$\widetilde{Z}_{22} = \frac{1}{\sqrt{n}} (N_{22} - n U_n^{-1} \widetilde{\alpha}_2^2).$$

Also, we show in Appendix 2 that

$$\frac{N_{2.}}{N_{.2}} = \frac{\alpha_2}{\beta_2} + \frac{1}{\sqrt{n}\, {\beta_2}^2} \left( Z_{2.}\, \beta_2 - Z_{.2}\, \alpha_2 \right) + O_p(n^{-1}) \tag{5.1}$$

with $\mu = \alpha_2/\beta_2$. Therefore, we obtain from the above equation that

$$U_n^{-1} = \frac{1}{\mu} - \frac{1}{\sqrt{n}\mu^2\beta_2} (Z_{2.} - \mu Z_{.2}) + O_p(n^{-1}) .$$

We also rewrite $\widetilde{\alpha}_1$ as

$$\widetilde{\alpha}_1 = \frac{1}{2} \left\{ 1 - U_n + \sqrt{(1 - U_n)^2 + 4\widehat{\theta}_n U_n} \right\} = g(U_n, \widehat{\theta}_n), \text{ say,}$$

where by direct substitution, it follows that

$$g(\mu, \widehat{\theta}_n) = \alpha_1,$$

$$g'(\mu, \widehat{\theta}_n) = -\frac{\alpha_1(1 - \beta_1)^2}{\alpha_1 + \beta_1 - 2\alpha_1\beta_1}.$$

From the last four equations, we obtain

$$\widetilde{Z}_{22} = \frac{1}{\sqrt{n}}\Big\{N_{22} - \frac{n(1-\alpha_1)}{\mu}\Big([1 - \frac{(U_n - \mu)}{\mu}][1 + \alpha_1 - 2g'(\mu, \widehat{\theta}_n)(U_n - \mu)] + O_p(n^{-1})\Big)\Big\}$$

$$= d_{12}Z_{12} + d_{21}Z_{21} + d_{22}Z_{22} + O_p(n^{-1/2}),$$

where

$$d_{12} = -\frac{(\beta_1 - \alpha_1)\,(1-\alpha_1)}{\alpha_1 + \beta_1 - 2\,\alpha_1\,\beta_1}\ , \quad d_{21} = \frac{(\beta_1 - \alpha_1)\,(1-\beta_1)}{\alpha_1 + \beta_1 - 2\,\alpha_1\,\beta_1}$$

$$d_{22} = \frac{\alpha_1\,(1-\alpha_1) + \beta_1\,(1-\beta_1)}{\alpha_1 + \beta_1 - 2\,\alpha_1\,\beta_1}\,.$$

Let $E_0$ denote the expected value and $Var_0$ the variance under $H_0$. Then

$$E_0\left(\widetilde{Z}_{22}\right) = 0 + O(n^{-1/2}) = o(1)\ ,$$

$$Var_0\left(\widetilde{Z}_{22}\right) = \sum_{(i,j)\neq(1,1)} (d_{ij})^2\,\alpha_i\,\beta_j - \frac{n}{n_*}\,(\alpha_2\,\beta_2)^2 = \gamma_{22}\ (\text{say})\,.$$

We may note that when $\alpha_1 = \beta_1$,

$d_{12} = d_{21} = 0$ and $Var_0\left(\widetilde{Z}_{22}\right) = \alpha_2^2(1 - \alpha_1^2 - \alpha_2^2)/(1 - \alpha_1^2)$.

In summary, based on the consensus count and the resulting 2×2 table, we propose the following test statistic

$$\frac{N_{22} - n\,\widetilde{\alpha}_2\,\widetilde{\beta}_2}{\sqrt{n\,\widehat{\gamma}_{22}}} \tag{5.2}$$

where $\widehat{\gamma}_{22}$ is obtained from $\gamma_{22}$ by substituting the estimates of the $\alpha_i$'s and $\beta_j$'s as obtained earlier into the $d_{ij}$'s.

Table 2 shows the results of simulations performed to assess the empirical validity of the above statistic. For each sample size and underlying probability distribution, 1,000 contingency tables were generated under the hypothesis of independence of rows and columns. $\underline{r}$ and $\underline{c}$ denote the row and column probability vector, respectively. From the data in each of the simulated contingency tables, the above test statistic was computed. Then for each of these tables, $\underline{r}$ and $\underline{c}$ were estimated, and these estimates were utilized to construct 1,000 bootstrap further tables, keeping the (1,1)-cell fixed. The empirical distribution of the test statistics calculated from the bootstrap tables is used as the reference

TABLE 2

*Results of 1,000 simulations under the assumption of independence between rows and columns. × * indicates that a percentage is two standard deviations away from its expected value and ** that it is three standard deviations away.*

| Sample Size | 1 | 5 | 10 | 90 | 95 | 99 |
|---|---|---|---|---|---|---|
| $\underline{r}$=(.60,.40) , $\underline{c}$=(.60,.40) | | | | | | |
| 100 | 0.1 ** | 0.3 ** | 1.7 ** | 0.1 ** | 0 ** | 0 ** |
| 500 | 0 | 0.1 ** | 0.4 ** | 0.1 ** | 0 ** | 0 ** |
| 1,000 | 0 ** | 0 ** | 0 ** | 0.1 ** | 0 ** | 0 ** |
| $\underline{r}$=(.70,.30) , $\underline{c}$=(.60,.40) | | | | | | |
| 100 | 0.6 | 3.1 * | 7.3 * | 13.2 ** | 7.8 ** | 2.0 ** |
| 200 | 0.8 | 5.5 | 10.9 | 12.5 * | 7.5 ** | 2.0 ** |
| 300 | 0.5 | 4.1 | 9.8 | 13.3 ** | 7.7 ** | 2.1 ** |
| 400 | 0.8 | 5.2 | 9.4 | 9.5 | 5.8 | 1.2 |
| $\underline{r}$=(.80,.20) , $\underline{c}$=(.70,.30) | | | | | | |
| 100 | 0.2 * | 3.5 * | 9.0 | 12.3 * | 7.3 ** | 2.2 ** |
| 200 | 0.3 * | 4.8 | 10.0 | 10.8 | 6.6 * | 1.9 * |
| 300 | 0.6 | 4.3 | 8.5 | 9.2 | 5.3 | 1.4 |

distribution of the observed statistic. The entries of Table 2 are the percentages of the test statistics falling below the 1st, 5th and 10th percentiles and above the 90th, 95th and 99th percentiles of the boostrap reference distribution distribution. These simulations show that, though the (1,1)-cell is assigned the largest frequency, the test is not appropriate for the scenario where $\underline{r} = \underline{c} = (.60, .40)$. For the other probabilistic set-ups, the observed numbers get closer to their expected values as the sample size increases. Hence, the asymptotic result derived in this paper is achieved with sample sizes of 300 to 400, depending on the underlying distribution.

**6. Data analysis.** We consider 141 HIV-1 T-cell-adapted sequences. Each sequence is 35 amino acids long and spans the V3 loop of the envelope protein gene. This region varies highly in composition across individuals and within individuals, and has been found to be functionally important [Korber et al. (1993), Potts et al. (1993)]. This data set does not contain duplicate sequences and a person contributes a single sequence. To avoid detecting linkages that, while declared statistically significant, have little scientific importance, we only considered pairs of positions for which the double-mutation count is at least 4.

For illustration purposes, we show in Table 3 the counts for positions 5 and 10. These data were aggregated into a 2 × 2 table (Table 4). The results appear in Table 5. The bootstrap distribution for the test statistics is computed by the $BC_a$ method [Efron & Tibshirani (1993)].

TABLE 3
*Contingency table for positions 5 and 10.*

| | | Position 10 | | | | | |
|---|---|---|---|---|---|---|---|
| | | K | R | N | S | Q | G |
| Position 5 | N | 87 | 27 | 1 | 2 | 2 | 1 |
| | H | 0 | 3 | 0 | 0 | 0 | 0 |
| | S | 4 | 4 | 0 | 0 | 0 | 0 |
| | G | 4 | 0 | 0 | 0 | 0 | 0 |
| | K | 0 | 1 | 0 | 0 | 0 | 0 |
| | D | 1 | 1 | 0 | 0 | 0 | 0 |
| | Y | 1 | 0 | 0 | 1 | 1 | 0 |

TABLE 4
*Aggregated table for positions 5 and 10.*

| | K | & |
|---|---|---|
| N | 87 | 33 |
| & | 10 | 11 |

The global variation of the virus is divided into *groups* and *clades* within a group (groups O, M and N; clades A-J within group M). Each clade/group exhibits a different consensus sequence. Sequences from the same clade cluster in the same geographic region. These clades may result from major natural selection pressures. The above test is helpful in detecting parallel evolution within a clade (here, the sequences belong to clade B which covers North America, Western Europe and Thailand). Were one to analyze sequences from different clades together, the linkages would reflect correlations in the consensus sequences. Then a binomial test like (5.2) is of little value as it merely consider the number of double mutations, i.e. the number in each clade (and thus depends on the sampling procedure and not on a biological mechanism). More relevant would be $\chi^2$-square-type tests which consider specific amino-acid pairings.

**7. Discussion.** Our interest lies in investigating whether departures from the consensus amino acids at two positions are correlated. The assessment of whether there has been, at a specific position, a substitution away from the consensus clearly relies on the knowledge of this consensus. This is essentially the reasoning behind our conditioning on the number of pairs in the consensus cell. This conditioning affects the probability law for the test statistic by basically reducing its variance. Further, there are two statistical arguments in favour of conditioning on $N_{11}$. First, in a large $r \times c$ contingency table (potentially 20 $\times$ 20) with many empty and low-frequency cells, the exact test conditional on

TABLE 5
*Results for 141 HIV-1 sequences.*

| Positions | Test statistics | p-values |
|---|---|---|
| 5 , 10 | 2.442 | $.005 < p < .01$ |
| 7 , 22 | 0.324 | $> .1$ |
| 8 , 10 | 5.291 | $< .0005$ |
| 10, 23 | 1.354 | $.05 < p < .1$ |
| 10 , 24 | 2.833 | $.001 < p < .005$ |
| 10 , 34 | 3.243 | $.001 < p < .005$ |
| 12 , 19 | 1.926 | $.01 < p < .05$ |
| 12 , 22 | 1.245 | $> .1$ |
| 12, 29 | 1.340 | $> .1$ |
| 14 , 19 | 9.210 | $< .0005$ |
| 14 , 20 | 4.481 | $< .0005$ |
| 19 , 20 | 6.647 | $< .0005$ |
| 19 , 22 | -0.898 | $> .1$ |
| 19 , 32 | 2.513 | $.005 < p < .01$ |
| 21 , 22 | 2.781 | $.001 < p < .005$ |
| 21 , 24 | 4.306 | $< .0005$ |
| 22 , 23 | 3.296 | $< .0005$ |
| 22 , 24 | 5.548 | $< .0005$ |
| 22 , 34 | 0.536 | $> .1$ |
| 23 , 24 | 5.721 | $< .0005$ |

fixed marginals has little power. The table thus needs to be reduced. Here, to do so, we use the consensus cell. Hence, since this consensus is data dependent, it affects the resulting structure of the table and probability model. Second, the size of $N_{11}$ affects the power of the test. It is indeed of a different order of magnitude to the other entries (in the example of Section 6, $N_{11}$ is 87 and the next largest entry is 33). By conditioning on $N_{11}$, we gain power.

Through a simple test for the equality of two binomial proportions, one can verify that the consensus pair is indeed truly maximal. When $\lambda_{11}$ is not very large compared to the other $\lambda_{ij}$'s, there is no obvious consensus pair and thus this conditioning cannot be applied. One then needs to resort to the usual $\chi^2$-test for independence in contingency tables. However, the absence of a clear consensus often indicates, in our experience, that two or more populations of sequences are represented in the sample. Then the analysis proceeds by considering each subpopulation separately.

The theory developed in this paper relies on the independence of the sequences. This assumption is violated in many instances. However, it would be straightforward to alter the reference distribution so that it would take into

account the evolutionary process undergone by the sequences under study. This could be done by simulating the evolutionary process and generating a large number of sets of sequences. The test statistic is computed on each set, to construct a reference distribution. For the sequences utilized in Section 6, the replication rate for HIV is very high (once every one to two days). Each replication introduces one to ten substitutions along the whole genome. As many rounds of replication are likely to separate two sequences (i.e., the total time to their most recent ancestor along the two branches), sharing the same amino acid at a polymorphic site is due to structure restriction and not phylogenetic relationships. Thus, these sequences can be regarded as independent, and the test introduced here can be applied.

**Acknowledgements.** We thank the referees for helpful comments.

## REFERENCES

BISHOP, Y., FIENBERG, S. and HOLLAND, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge, Massachusetts.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements* **20** 37–46.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

KARNOUB, M. (1997). Understanding Dependencies among Mutations along the HIV Genome, PhD thesis, Department of Biostatistics, University of North Carolina at Chapel Hill.

KORBER, B.T.M., FARBER, R.M., WOLPERT, D.H. and LAPEDES, A.S. (1993). Covariation of mutations in the V3 loop of the human immunodeficiency virus type 1 envelope protein: An information-theoretic analysis. *Proceedings of the National Academy of Sciences U.S.A.* **90** 7176–7180.

LANDIS, J.R. and KOCH, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33** 159–174.

POTTS, K.E., KALISH, M.L., BANDEA, C.I., ORLOFF, G.M., ST. LOUIS, M., BROWN, C., MALANDA, N., KAVUKA, M., SCHOCHETMAN, G., OU, C. and HEYWARD, W.L. (1993). Genetic diversity of human immunodeficiency virus type 1 strains in Kinshasa, Zaïre. *AIDS Research and Human Retroviruses* **9** 613–618.

SEILLIER-MOISEIWITSCH, F., PINHEIRO, H., KARNOUB, M. and SEN, P.K. (1998). Novel methodology for quantifying genomic heterogeneity. In *Proceedings of the Joint Statistical Meetings, Anaheim, California, August 1997.*

# Appendix 1: Proof of (3.2)

Note that

$$
\begin{aligned}
&P\{N_{11} \text{ is maximum}\} \\
&\quad = P\{N_{ij} < N_{11} \quad \forall\,(i,j) \neq (1,1)\} \\
&\quad = E\left\{P\{N_{ij} < N_{11} \quad \forall\,(i,j) \neq (1,1) \mid N_{11}\}\right\} \\
&\quad = E\left\{P\{N_{12} < N_{11} \mid N_{11}\}\, P\{N_{21} < N_{11} \mid N_{11}\}\, P\{N_{22} < N_{11} \mid N_{11}\}\right\}
\end{aligned}
$$

where for large $\lambda_{ij}$'s, we use the square-root transformation on the Poisson variates. Thus, under Assumption 2 in Section 3, for each $(i,j) \neq (1,1)$,

$$\begin{aligned} P\{N_{ij} < N_{11} | N_{11}\} \\ &= P\{\sqrt{N_{ij}} - \sqrt{\lambda_{ij}} < \sqrt{N_{11}} - \sqrt{\lambda_{ij}} | N_{11}\} \\ &= P\{W_{ij} < W_{11} + 2(\sqrt{\lambda_{11}} - \sqrt{\lambda_{ij}}) \mid W_{11}\} \end{aligned}$$

where the $W_{ij}$ are independent and asymptotically normally distributed with zero mean and unit variance, and therefore are all bounded in probability. On the other hand, under Assumption 2, $\sqrt{\lambda_{11}} - \sqrt{\lambda_{1j}}$ is large and positive for any $(i,j) \neq (1,1)$. Hence, the above probability converges to 1 as $\lambda_{11}$ increases, which satisfies Assumption 2. Therefore, their product also converges to 1, and hence, being a bounded random variable, their expectation has the same limit, i.e. 1.

## Appendix 2: Proof of (5.1)

$$\begin{aligned} \frac{N_{2.}}{N_{.2}} &= \frac{\sqrt{n}\, Z_{2.} + n\,\alpha_2}{\sqrt{n}\, Z_{.2} + n\,\beta_2} \\ &= \frac{\alpha_2}{\beta_2} + \frac{1}{\sqrt{n}\,{\beta_2}^2} \frac{Z_{2.}\,\beta_2 - Z_{.2}\,\alpha_2}{1 + \dfrac{Z_{.2}}{\sqrt{n}\,\beta_2}}, \end{aligned}$$

where noting that $Z_{.2} = O_p(1)$, we have

$$\left\{1 + \frac{Z_{.2}}{\sqrt{n}\beta_2}\right\}^{-1} = 1 - \frac{Z_{.2}}{\sqrt{n}\beta_2} + O_p(n^{-1}),$$

and hence

$$\frac{N_{2.}}{N_{.2}} = \frac{\alpha_2}{\beta_2} + \frac{1}{\sqrt{n}\,{\beta_2}^2}\,(Z_{2.}\,\beta_2 - Z_{.2}\,\alpha_2) + O_p(n^{-1}).$$

GLAXO WELLCOME INC.
FIVE MOORE DRIVE
P.O.BOX 13398
RESEARCH TRIANGLE PARK, NC 27709-3398
MAC31876@GLAXOWELLCOME.COM

DEPARTMENT OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-7400
SEILLIER@BIOS.UNC.EDU
PKSEN@BIOS.UNC.EDU