Detecting Linked Mutations in Fast Evolving Genomes

Chaehyung Ahn
The University of North Carolina, Biostatistics
Campus Box #7400, Chapel Hill, NC 27516, USA
cahn@bios.unc.edu

Francoise Seillier-Moiseiwitsch
The University of North Carolina, Biostatistics
Campus Box #7400, Chapel Hill, NC 27516, USA
seillier@bios.unc.edu

1. Introduction. The genome of the human immunodeficiency virus (HIV) mutates rapidly enough to deplete the immune system of the host. To make an effective vaccine against HIV, it is important to understand its mutation process. The high error rate of the reverse transcriptase and high turnover rate *in vivo* generate vast numbers of different viral mutants (1). Some of these substitutions help mutants to escape immune system, while others prevent mutants from generating functional proteins, which are vital to their survival. However, if mutations at another position help these proteins to keep their structure, they would be favorable for their survival (2).

Table 1. Contingency table for nucleotide sequence data

Position i	Position j: A	G	C	T
T	n_{11}, $\alpha_1\beta_1$ *consensus*	n_{12}, $\alpha_1\beta_2$	n_{13}, $\alpha_1\beta_3$	n_{14}, $\alpha_1\beta_4$
C	n_{21}, $\alpha_2\beta_1$	n_{22} $\alpha_2\beta_2$	n_{23}, $\alpha_2\beta_3$	n_{24} $\alpha_2\beta_4$
G	n_{31}, $\alpha_3\beta_1$	n_{32}, $\alpha_3\beta_2$	n_{33}, $\alpha_3\beta_3$	n_{34}, $\alpha_3\beta_4$
A	n_{41}, $\alpha_4\beta_1$	n_{42} $\alpha_4\beta_2$	n_{43}, $\alpha_4\beta_3$	n_{44} $\alpha_4\beta_4$
	Polymorphisms outer table	Double mutations inner table		

2. The Set-up. For illustration (Table 1), consider two loci, i and j, and assume the nucleotides T and A are predominant at loci i and j, respectively. Hence, the (1,1)-cell is the consensus (the most frequent configuration). The outer table, except for the (1,1)-cell, contains single mutations away from the consensus, and the inner table pertains to double mutations. We are interested in testing whether mutations at locus i are independent of those at locus j.

We are going to use only the inner table to test this hypothesis, but, in estimating α_i's and β_j's, it would be reasonable to use only the outer table, because the outer table contains the pure information about single mutation rates away from the consensus, while the inner table might be contaminated by possible correlations between the two loci.

3. The 2´2 Case. In this case, the test statistic pertains to the total number of mutations. Assume that all

cell counts have independent Poisson distribution with different means. If we fix the sum of outer table ($n_{11}+n_{12}+n_{21}$), we can get the following multinomial likelihood function

$$L(\alpha_1,\beta_1;n_{11},n_{12},n_{21}) = C\,\frac{\alpha_1^{n_{11}+n_{12}}\beta_1^{n_{11}+n_{21}}(1-\alpha_1)^{n_{21}}(1-\beta_1)^{n_{12}}}{(\alpha_1+\beta_1+\alpha_1\beta_1)^{n_{11}+n_{12}+n_{21}}}$$

From the above likelihood, we can get MLEs $\hat{\alpha}_1 = n_{11}/(n_{11}+n_{21})$ and $\hat{\beta}_1 = n_{11}/(n_{11}+n_{12})$. We verify that $\frac{1}{\sqrt{n}}\left(N_{22} - n\hat{\alpha}_2\hat{\beta}_2\right)$ is asymptotically normally distributed with mean 0 and variance γ_{22}, which is a function of α_1 and β_1. We propose the following test statistic.

$$\frac{N_{22} - n\hat{\alpha}_2\hat{\beta}_2}{\sqrt{n\hat{\gamma}_{22}}}$$

where $\hat{\gamma}_{22}$ is obtained from γ_{22} by substituting $\hat{\alpha}_1$ and $\hat{\beta}_1$ for α_1 and β_1.

Simulations show that this test statistic behaves reasonably well as a normally distributed with mean 0 and unit variance.

4. The r´c Case. In this case, MLEs are $\hat{\alpha}_i = n_{i1}/\sum_{i=1}^{r} n_{i1}$ and $\hat{\beta}_j = n_{1j}/\sum_{j=1}^{c} n_{1j}$, and the following test statistic is proposed.

$$\tilde{Z}_v^T\{\hat{V}_o(\tilde{Z}_v)\}^{-1}\tilde{Z}_v$$

where $\tilde{Z}_v = vec(\tilde{Z}^T)$, $\tilde{Z}_{ij} = \frac{1}{\sqrt{n}}(N_{ij} - n\hat{\alpha}_i\hat{\beta}_j)$ and $\hat{V}_o(\tilde{Z}_v)$ is the covariance matrix of $\tilde{Z}_v$ under the null hypothesis, which is obtained by substituting $\hat{\alpha}_i$ and $\hat{\beta}_j$ for α_i and β_j. The above statistic follows asymptotic χ^2-distribution with (r-1)(c-1) degrees of freedom.

5. Data Analysis. This methodology is used to analyze 356 HIV sequences of the V3 loop of the envelope protein, which spans 35 amino acids.

REFERENCES

Bonhoeffer, S., Holmes, E. C. and Nowak, M. A. (1995). Causes of HIV diversity. Nature 376:125.

Seillier-Moiseiwitsch, F., Pinheriro, 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.

RESUME

A Mutation may help a virus escape from the immune system, but it can also be deleterious to the virus, because the mutation drastically alters the structure of a core protein. However, this structure may be preserved, if a specific mutation occurs at another position. We construct a test statistic to detect such kind of mutational linkage between two loci.