Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 3227–3239
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Log-linear models for mutations in the HIV genome☆

C. Ahn[a], G.G. Koch[a,*], L. Paynter[a], J.S. Preisser[a], F. Seillier-Moiseiwitsch[b]

[a]*Department of Biostatistics, School of Public Health, Campus Box 7420, Chapel Hill, NC 27599-7420, USA*
[b]*Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC 20057-1484, USA*

Available online 31 March 2007

Abstract

We discuss a general application of categorical data analysis to mutations along the HIV genome. We consider a multidimensional table for several positions at the same time. Due to the complexity of the multidimensional table, we may collapse it by pooling some categories. However, the association between the remaining variables may not be the same as before collapsing. We discuss the collapsibility of tables and the change in the meaning of parameters after collapsing categories. We also address this problem with a log-linear model. We present a parameterization with the consensus output as the reference cell as is appropriate to explain genomic mutations in HIV. We also consider five null hypotheses and some classical methods to address them. We illustrate methods for six positions along the HIV genome, through consideration of all triples of positions.

MSC: 62F03; 62H17; 62H20; 62J12; 62P10

Keywords: HIV genome; Consensus; Correlated mutation; Collapsibility; Conditional association; Marginal association; Log-linear models

1. Introduction

Nucleotides and amino acids can be viewed as categorical data. In this paper, we are interested in mutations along the HIV genome. Like other DNA or RNA sequences, HIV sequences are coded by means of four nucleotides or 20 amino acids. These are just nominal variables. We consider amino-acid sequences, since the information contained in amino acids is more important from a functional/structural point of view.

Mutations are usually deleterious to any organism. However, in the case of HIV, a mutation can provide a virus with a better chance to escape the host immune system, which helps the virus survive and pass on its genetic material. We believe mutations at specific positions are correlated, and tendencies for some combinations of mutations to be observed more frequently than expected by chance provide proof. These double mutations either maintain the structure of a vital protein (possibly when a single mutation would destabilize this structure) or yield a viable structural form not recognized by the host.

Several positions along the HIV genome are simultaneously represented by a multidimensional contingency table. High-dimensional tables can have complicated structures and interpretations for model parameters. Roy and Mitra (1956), Roy and Kastenbaum (1956), Roy (1957), Roy and Bhapkar (1960) and Bhapkar and Koch (1968) describe the structure and several hypotheses of interest for three-way tables. Agresti (2002) and Imrey and Koch (2005) describe

☆ This research was partially supported by Grant R01 AI47068 from the National Institutes of Health.
* Corresponding author.
E-mail address: bcl@bios.unc.edu (G.G. Koch).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.007

those hypotheses in terms of a hierarchical log-linear model. Other related discussion is provided in Bishop (1971), Imrey et al. (1981, 1982, 1996), and Imrey (2000).

Following S.N. Roy's nomenclature, we treat positions as "variates" and do not condition on marginal totals. This approach is motivated by the random nature of the substitution process. In Roy and Kastenbaum (1956) and Roy and Mitra (1956), hypotheses relevant to this set-up were formulated. We will consider, among these, those that are interpretable in our context, namely the hypotheses of conditional and multiple independence (H_{03} and H_{04}, respectively in Section 2.2).

A multidimensional contingency table may contain many zeroes or very small cell counts, and this sparseness of data may undermine test statistics with distributional properties based on large-sample sizes. To use such test statistics, we may have to pool some categories to ensure cell counts are adequate for a large-sample approximation to be applicable. However, the meaning of the parameters may not remain the same. Nevertheless, it will be the same in cases where some collapsibility conditions are satisfied.

In the subsequent sections of this paper, we use the reference-cell parameterization for multidimensional contingency tables since it is appropriate to explain mutations in the HIV genome. We initially consider a model for HIV mutations resulting in a 2×2 table, and then extend it to the $r \times c$ and $2 \times 2 \times 2$ situations. Instead of presenting a $I \times J \times K$ contingency table, we describe it with a log-linear model for simplicity. We investigate the collapsibility of a three-way table by considering its conditional and marginal associations. We identify five hypotheses which might be appropriate for HIV mutations. One of these hypotheses corresponds to an interpretable non-hierarchical model.

For hypothesis testing, we simply use a Wald statistic and/or a deviance test statistic. We illustrate methods with results from an analysis involving six positions. We consider all triples of positions, and we present some of them as examples to explain the five hypotheses of interest.

1.1. Parameterization and hypotheses

We consider models for multinomially distributed cell counts rather than cell probabilities. Based on the biological meaning of mutations and the concept of consensus, we present a reference-cell coding for the model with the consensus as the reference cell.

The usual issue of interest in a contingency table is the independence among two or more categorical variates. However, for our consensus-referencing parameterization, the main objective is the investigation of whether double mutations provide a survival advantage to a virus. Such structure will be considered for the 2×2 table and then extended to the more general $r \times c$ table and the $2 \times 2 \times 2$ table as the simplest three-way table.

1.2. Modelling HIV mutations: the 2×2 *case*

Fig. 1 shows the parameterization which will be used throughout this paper. For simplicity, we consider two positions with two categories at each position: A_1 and A_2 at the first position and B_1 and B_2 at the second position. If (A_1, B_1) is the original sequence, there are four possible progenies, (A_1,B_1), (A_1,B_2), (A_2,B_1) and (A_2,B_2). (A_1,B_2) has a mutation at the second position, (A_2,B_1) has a mutation at the first position, and (A_2,B_2) has mutations at both positions. We assume that the mutation at the first position is independent of that at the second position, since mutations in HIV are

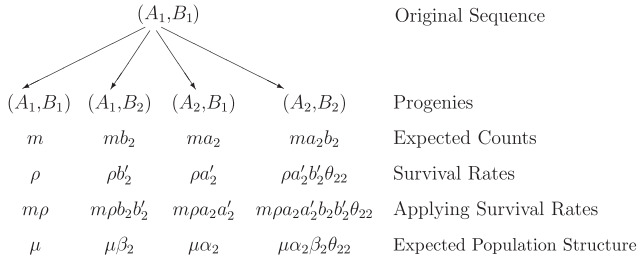

Fig. 1. Parameterization.