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

LARGE COMPOUND POISSON APPROXIMATIONS FOR OCCURRENCES OF MULTIPLE WORDS

BY GESINE REINERT[1] AND SOPHIE SCHBATH[2]

Department of Statistics, UCLA and Unité de Biométrie, INRA, 78352 Jouy-en-Josas, France.

A compound Poisson process approximation for the number of occurrences of multiple words in a sequence of letters is derived, where the letters are assumed to be independent and identically distributed. Using the Chen-Stein method, a bound on the error in the approximation is provided. For rare words, this error tends to zero as the length of the sequence increases to infinity. As an application the efficiency of the approximation for the number of occurrences of rare stem-loop motifs in DNA sequences is illustrated.

1. Introduction. When searching a database for the occurrence of a combination of several words within a sequence, the typical Poisson approximation used by programs like BLAST is no longer valid, as overlapping words may be dependent on each other. Here a compound Poisson approximation for the multiple occurrences of short words within a sequence is derived. Using the Chen-Stein method for Poisson process approximation, an explicit error bound for the approximation is given, improving those obtained by Schbath (1995a) for a single rare word. The approximation error increases with the amount of overlap between the words. The results are applied to the occurrences of stem-loop motifs. Another application might be a set of words coding for the same amino-acid sequence.

In general, consider a finite sequence S of letters chosen independently from a finite alphabet $\mathcal{A}$. The main example will be the four-letter DNA alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ but the results are valid for general finite alphabets such as $\{0, 1\}$ or the 20-letter amino acid alphabet. An abundant literature exists on the asymptotic distribution of the number of occurrences of a single word in such a sequence S. A normal approximation, valid for frequent words, is presented by Prum et al. (1995). A compound Poisson approximation is obtained in Arratia et al. (1990), Geske et al. (1995) and Schbath (1995a) for the number of occurrences of a rare word, whereas the number of clumps of a rare word is approximated by a Poisson variable (as a rule of thumb, a word is rare if its length

[1] This work was partially supported by NSF grant DMS-9505075.

[2] This work was partially supported by NATO and by NSF grant BIR-9504393.

AMS 1991 *subject classifications.* Primary 60F05 ; secondary 92D20.

Key words and phrases. Chen-Stein method, stem-loop motifs, compound Poisson approximation, occurrences of multiple words.