Journal of Automata, Languages and Combinatorics **20** (2015) 1, 53–65

ON THE NUMBER OF MANY-TO-MANY ALIGNMENTS OF MULTIPLE SEQUENCES

Steffen Eger
Computer Science Department, Goethe University Frankfurt am Main
Robert-Mayer-Straße 10, Frankfurt, Germany
e-mail: steeger@em.uni-frankfurt.de

ABSTRACT

We count the number of alignments of $N \geq 1$ sequences when match-up types are from a specified set $S \subseteq \mathbb{N}^N$. Equivalently, we count the number of non-negative integer matrices whose rows sum to a given fixed vector and each of whose columns lie in S. We provide a new asymptotic formula for the case $S = \{(s_1, \ldots, s_N) \mid 1 \leq s_i \leq 2\}$.

Keywords: alignment, composition, sum of discrete random variable, lattice path

1. Introduction

Alignments of sequences arise in computational biology and in computational linguistics. In computational biology, aligning DNA sequences is a standard task. In computational linguistics, aligning (historical) variants of linguistic forms is a field of study (see [8]). In addition, alignments of sequences arise in computational linguistics either in (machine) translation, where words from different languages are matched up, or in related string-to-string translation tasks such as letter-to-sound conversion, where the task is to translate a letter string into a phonetic representation, or in lemmatization, where the task is to translate a word form into a canonical lexicon representation. Traditionally, an alignment of N (for an integer $N \geq 2$) sequences of various lengths is defined as a manner of inserting blanks into the N sequences such that all have equal length. For example, given $\mathbf{x} = x_1$, $\mathbf{y} = y_1y_2$ and $\mathbf{z} = z_1z_2z_3$, three (out of 239 possible) alignments of $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ are:

x_1	-	-		x_1	-	-	-		-	-	x_1
y_1	y_2	-		y_1	-	y_2	-		y_1	y_2	-
z_1	z_2	z_3		z_1	z_2	-	z_3		z_1	z_2	z_3

In computational linguistics, alignments in which subsequences (of length ≥ 1) from the different sequences are matched up with each other ('many-to-many matches') are oftentimes more plausible and also more frequently made of use of (see [15, 19]). When we allow, for example, in addition to the above specification, matches-up of length up to 2, there are several further alignments of $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$, including: