Improving Calculations of the Number of Distinct Alignments of Two Strings

Øystein J. Rødseth

Department of Mathematics, University of Bergen,

Johs. Brunsgt. 12, N–5008 Bergen, Norway

E-mail: rodseth@mi.uib.no

James A. Sellers

Department of Mathematics, Penn State University,

University Park, PA 16802 USA

E-mail: sellersj@math.psu.edu

Abstract

In a recent work, M. Covington discusses the enumeration of two different sets of alignments of two strings of symbols using elementary combinatorial techniques. He defines two functions $a(m, n)$ and $\mathrm{A}(m, n)$ to count the number of two–string alignments in his "small" and "middle" sets of alignments (respectively). He provides a recurrence for each of these functions which allows for the calculation of values of $a(m, n)$ and $\mathrm{A}(m, n)$. In this note, we obtain generating functions for each of these functions. With the generating functions in hand, we provide improvements on Covington's recurrences, making the calculation of $a(m, n)$ and $\mathrm{A}(m, n)$ much more efficient.

Keywords: alignments, combinatorics, recurrences, generating functions
2000 Mathematics Subject Classification: 05A10, 05A15

1 Introduction

In his recent work on distinct alignments of two strings, M. Covington [1] defines two functions which count certain types of alignments. We briefly describe the two functions here. Covington defines the function $a(m, n)$ as the number of distinct alignments of two strings of letters (one of length m and the other of length n) such that alternating skips are not allowed. That is, as we align the two strings, we are not allowed places in the alignment where a skip in one string is immediately followed by a skip in the other string.

In contrast, Covington defines the function $\mathrm{A}(m, n)$ to be the number of distinct alignments of two strings of length m and n respectively such that mismatches or double skips are allowed. In [1], he states that the set of alignments enumerated by $\mathrm{A}(m, n)$ is "the best model of the search space for string matching as generally conceived [2]."

After defining these functions, Covington states the following recurrences that are satisfied by $a(m, n)$ and $\mathrm{A}(m, n)$ respectively:

$$(1) \qquad a(m,n) = a(m-1,n-1) + \sum_{i=0}^{n-2} a(m-1,i) + \sum_{i=0}^{m-2} a(i,n-1)$$

with initial values $a(0, n) = a(m, 0) = 1$, and

$$(2) \qquad \mathrm{A}(m,n) = 2\left(\mathrm{A}(m-1,n-1) + \sum_{i=0}^{n-2} \mathrm{A}(m-1,i) + \sum_{i=0}^{m-2} \mathrm{A}(i,n-1)\right)$$

with initial values $\mathrm{A}(0, n) = \mathrm{A}(m, 0) = 1$

While these recurrences certainly allow for the calculation of many values of $a(m, n)$ and $\mathrm{A}(m, n)$, they are by no means the most efficient recurrences possible. Our goal in this work is to simplify these two recurrences. This will provide dramatic improvements in the calculation of $a(m, n)$ and $\mathrm{A}(m, n)$. In order to find these simplified recurrences, generating functions for $a(m, n)$ and $\mathrm{A}(m, n)$ are developed.

In Section 2, we determine a closed form for the generating function for $a(m,n)$ and obtain the following simplified recurrence: For all $m,n \geq 2$,

$$a(m,n) = a(m-1,n) + a(m,n-1) - a(m-2,n-2)$$

with the initial conditions $a(m,0) = a(0,n) = 1$ for all $m,n \geq 0$, $a(m,1) = m$ for $m \geq 1$, and $a(1,n) = n$ for $n \geq 1$

In Section 3, we complete a similar analysis for $\mathrm{A}(m,n)$, ultimately proving the following recurrence: For all $m,n \geq 2$,

$$\begin{aligned}\mathrm{A}(m,n) = {} & \mathrm{A}(m-1,n) + \mathrm{A}(m,n-1) \\ & + \mathrm{A}(m-1,n-1) - 2\mathrm{A}(m-2,n-2)\end{aligned}$$

with the initial conditions $\mathrm{A}(m,0) = \mathrm{A}(0,n) = 1$ for all $m,n \geq 0$, $\mathrm{A}(m,1) = 2m$ for $m \geq 1$, and $\mathrm{A}(1,n) = 2n$ for $n \geq 1$

These two simplified recurrences provide very rapid means for computing the values of $a(m,n)$ and $\mathrm{A}(m,n)$.

2 Covington's "Small Set" and $a(m,n)$

We open this section with the well-known sum of an (infinite) geometric sequence.

Lemma 1

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Remark. For those concerned about issues of convergence, Lemma 1 is true only for those values of x such that $|x| < 1$. However, throughout this work, we treat all power series as formal so that questions of convergence will not be considered.

Proof. We have

$$(1-x)\sum_{n=0}^{\infty} x^n = \sum_{n=0}^{\infty} x^n - \sum_{n=0}^{\infty} x^{n+1} = \sum_{n=0}^{\infty} x^n - \sum_{n=1}^{\infty} x^n = 1. \qquad \blacksquare$$

As a consequence of Lemma 1 we see that

$$\sum_{n=0}^{\infty}\sum_{i=0}^{n} b(i)x^n = \sum_{j=0}^{\infty} x^j \sum_{i=0}^{\infty} b(i)x^i = \frac{1}{1-x}\sum_{i=0}^{\infty} b(i)x^i \tag{3}$$

for any sequence $b(0), b(1), \ldots$ of real numbers.

Now to the generating function for $a(m,n)$. The starting point is recurrence relation (1) above. We note in passing that $a(m,n)$ is symmetric in m and n, that is

$$a(m,n) = a(n,m) \qquad \text{for all } m,n \geq 0. \tag{4}$$

We can utilize (1) to determine the generating function $f(x,y)$ of $a(m,n)$. The function $f(x,y)$ is a formal power series in the two variables x and y, and is defined as

$$f(x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a(m,n)x^m y^n.$$

Thanks to (4), we note that

$$f(x,y) = f(y,x).$$

We shall prefer to rewrite (1) in the form

$$a(m,n) = -a(m-1,n-1) + \sum_{i=0}^{n-1} a(m-1,i) + \sum_{i=0}^{m-1} a(i,n-1) \tag{5}$$

and will use this version of the recurrence in the work below.

We now find a closed form for the generating function $f(x,y)$. By (5), we have

$$\begin{aligned} f(x,y) &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a(m,n)x^m y^n \\ &= -a(0,0) + \sum_{m=0}^{\infty} a(m,0)x^m + \sum_{n=0}^{\infty} a(0,n)y^n \\ &\quad + \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\left(-a(m-1,n-1) + \sum_{i=0}^{n-1} a(m-1,i) + \sum_{i=0}^{m-1} a(i,n-1)\right)x^m y^n. \end{aligned}$$

Using the initial values $a(m,0) = a(0,n) = 1$ and Lemma 1, we have

$$\begin{aligned} f(x,y) &= -1 + \frac{1}{1-x} + \frac{1}{1-y} \\ &\quad + \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\left(-a(m,n) + \sum_{i=0}^{n} a(m,i) + \sum_{i=0}^{m} a(i,n)\right) x^{m+1}y^{n+1} \\ &= -1 + \frac{1}{1-x} + \frac{1}{1-y} - xy\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a(m,n)x^m y^n \\ &\quad + xy\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{i=0}^{n} a(m,i)x^m y^n + xy\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{i=0}^{m} a(i,n)x^m y^n. \end{aligned}$$

Moreover, using (3), we obtain

$$\begin{aligned} f(x,y) &= -1 + \frac{1}{1-x} + \frac{1}{1-y} - xyf(x,y) \\ &\quad + xy\sum_{m=0}^{\infty}\frac{1}{1-y}\sum_{i=0}^{\infty} a(m,i)x^m y^i + xy\sum_{n=0}^{\infty}\frac{1}{1-x}\sum_{i=0}^{\infty} a(i,n)x^i y^n \\ &= -1 + \frac{1}{1-x} + \frac{1}{1-y} - xyf(x,y) + \frac{xy}{1-y}f(x,y) + \frac{xy}{1-x}f(x,y). \end{aligned}$$

Our task is almost complete. We now perform some routine algebraic simplifications to obtain

$$\left(1 + xy - \frac{xy}{1-x} - \frac{xy}{1-y}\right) f(x,y) = -1 + \frac{1}{1-x} + \frac{1}{1-y},$$

and multiplying through by $(1-x)(1-y) = 1 - x - y + xy$, we get

$$(1 - x - y + x^2y^2)f(x,y) = 1 - xy.$$

This leads to our desired generating function result for $a(m,n)$.

Theorem 1 *The generating function* $f(x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} a(m,n)x^m y^n$ *is given by*

$$f(x,y) = \frac{1-xy}{1-x-y+x^2y^2}.$$

As was stated above, we clearly see from this closed form that $f(x, y)$ is symmetric in x and y; that is, $f(x, y) = f(y, x)$.

Note that $f(x, y)$ can be expanded via a computer algebra system in order to determine the value of $a(m, n)$ for various choices of m and n. For example, using `MAPLE`, we can determine the values of $a(10, n)$ for $0 \leq n \leq 49$ as follows:

```
gf:=(1-x*y)/(1-x-y+x^2*y^2):
s1:=series(gf, x, 50):
series(coeff(s1, x, 10), y, 50);
```

The output generated by these commands looks like the following:

$$
\begin{aligned}
&1 + 10\,y + 47\,y^2 + 149\,y^3 + 386\,y^4 + 899\,y^5 + 1948\,y^6 + 3989\,y^7 \\
&+7804\,y^8 + 14698\,y^9 + 26797\,y^{10} + 47491\,y^{11} + 82081\,y^{12} + 138709\,y^{13} \\
&+229675\,y^{14} + 373276\,y^{15} + 596340\,y^{16} + 937674\,y^{17} + 1452700\,y^{18} \\
&+2219618\,y^{19} + 3347511\,y^{20} + 4986895\,y^{21} + 7343318\,y^{22} + 10694727\,y^{23} \\
&+15413452\,y^{24} + 21993802\,y^{25} + 31086431\,y^{26} + 43540813\,y^{27} \\
&+60457365\,y^{28} + 83250977\,y^{29} + 113727949\,y^{30} + 154178598\,y^{31} \\
&+207488084\,y^{32} + 277268314\,y^{33} + 368014118\,y^{34} + 485287252\,y^{35} \\
&+635932171\,y^{36} + 828327931\,y^{37} + 1072681024\,y^{38} + 1381364425\,y^{39} \\
&+1769308636\,y^{40} + 2254451050\,y^{41} + 2858250529\,y^{42} + 3606274695\,y^{43} \\
&+4528868073\,y^{44} + 5661909901\,y^{45} + 7047671135\,y^{46} + 8735780928\,y^{47} \\
&+10784313652\,y^{48} + 13261008362\,y^{49} + \mathrm{O}(y^{50})
\end{aligned}
$$

The coefficients of the terms y^0 through y^{10} above correspond to the values in the $m = 10$ row in Covington [1, Table 2].

Even so, a simplified version of (1) would be much preferred in calculating $a(m, n)$ for large values of m and n. With Theorem 1 in hand, we can obtain such a simplified recurrence for $a(m, n)$.

Theorem 2 *For all $m, n \geq 2$,*

$$a(m,n) = a(m-1,n) + a(m,n-1) - a(m-2,n-2) \tag{6}$$

with the initial conditions $a(m,0) = a(0,n) = 1$ for all $m, n \geq 0$, $a(m,1) = m$ for $m \geq 1$, and $a(1,n) = n$ for $n \geq 1$.

Proof. By Theorem 1, we know

$$(1 - x - y + x^2y^2) \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a(m,n) x^m y^n = 1 - xy,$$

and it follows that

$$\begin{aligned} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a(m,n) x^m y^n &- \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} a(m-1,n) x^m y^n \\ &- \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} a(m,n-1) x^m y^n + \sum_{m=2}^{\infty} \sum_{n=2}^{\infty} a(m-2,n-2) x^m y^n \\ &= 1 - xy. \end{aligned}$$

Comparing coefficients of the monomial $x^m y^n$ for $m, n \geq 2$ on both sides of the above equation yields (6). The conditions $a(m,0) = a(0,n) = 1$ are known from above. Comparing coefficients of xy, we get $a(1,1) = 1$. Looking at $x^m y$ for $m \geq 2$, we get $a(m,1) = a(m-1,1) + a(m,0) = a(m-1,1) + 1$, and induction gives $a(m,1) = m$ for $m \geq 1$. By symmetry, $a(1,n) = n$ for $n \geq 1$. ∎

Next, we find a closed formula for $a(m,n)$. We define the function

$$\mu(m,n) = \min\{\lfloor m/2 \rfloor, \lfloor n/2 \rfloor\},$$

where $\lfloor k \rfloor$ equals the largest integer less than or equal to k (and is sometimes called the *floor function*). We then apply Lemma 1 to

$$f(x,y) = \frac{1 - xy}{1 - (x + y - x^2y^2)}$$

and find

$$\begin{aligned}
f(x,y) &= (1-xy)\sum_{r=0}^{\infty}(x+y-x^2y^2)^r \\
&= (1-xy)\sum_{r=0}^{\infty}\sum_{\substack{i+j+k=r\\ i,j,k\geq 0}}(-1)^k\frac{r!}{i!j!k!}x^{i+2k}y^{j+2k} \\
&= (1-xy)\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{k=0}^{\mu(m,n)}(-1)^k\frac{(m+n-3k)!}{k!(m-2k)!(n-2k)!}x^my^n \\
&= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{k=0}^{\mu(m,n)}(-1)^k\frac{(m+n-3k)!}{k!(m-2k)!(n-2k)!}x^my^n \\
&\quad -\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\sum_{k=0}^{\mu(m-1,n-1)}(-1)^k\frac{(m+n-2-3k)!}{k!(m-1-2k)!(n-1-2k)!}x^my^n.
\end{aligned}$$

Thus we have

$$\begin{aligned}
(7)\qquad a(m,n) &= \sum_{k=0}^{\mu(m,n)}(-1)^k\frac{(m+n-3k)!}{k!(m-2k)!(n-2k)!} \\
&\quad -\sum_{k=0}^{\mu(m-1,n-1)}(-1)^k\frac{(m+n-2-3k)!}{k!(m-1-2k)!(n-1-2k)!},
\end{aligned}$$

where, as usual, an empty sum is taken as zero. This formula is not practical from a computational perspective, but it does serve as an analogue of Covington's formula for $A(m,n)$, the function which enumerates the distinct alignments in what he calls the "large set" [1, p. 176].

3 Covington's "Middle Set" and $\textsc{a}(m,n)$

We next turn our attention to the generating function for $\textsc{a}(m,n)$. We start with recurrence relation (2) above. We note, as in the case of $a(m,n)$, the obvious symmetry $\textsc{a}(m,n) = \textsc{a}(n,m)$. Putting

$$F(x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\textsc{a}(m,n)x^my^n,$$

we thus have the generating function symmetry in x and y:

$$F(x,y) = F(y,x)$$

We shall prefer to rewrite (2) in the form

$$(8)\quad \mathrm{A}(m,n) = 2\left(-\mathrm{A}(m-1,n-1) + \sum_{i=0}^{n-1} \mathrm{A}(m-1,i) + \sum_{i=0}^{m-1} \mathrm{A}(i,n-1)\right).$$

We now find a closed form for $F(x,y)$. By (8), we have

$$\begin{aligned} F(x,y) &= -\mathrm{A}(0,0) + \sum_{m=0}^{\infty} \mathrm{A}(m,0)x^m + \sum_{n=0}^{\infty} \mathrm{A}(0,n)y^n \\ &\quad + \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} 2\left(-\mathrm{A}(m-1,n-1) + \sum_{i=0}^{n-1} \mathrm{A}(m-1,i) + \sum_{i=0}^{m-1} \mathrm{A}(i,n-1)\right) x^m y^n. \end{aligned}$$

Using the initial values $\mathrm{A}(m,0) = \mathrm{A}(0,n) = 1$ and Lemma 1, we have

$$\begin{aligned} F(x,y) &= -1 + \frac{1}{1-x} + \frac{1}{1-y} - 2\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \mathrm{A}(m,n)x^{m+1}y^{n+1} \\ &\quad + 2\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\sum_{i=0}^{n} \mathrm{A}(m,i)x^{m+1}y^{n+1} + 2\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\sum_{i=0}^{m} \mathrm{A}(i,n)x^{m+1}y^{n+1}. \end{aligned}$$

Moreover, using (3), we obtain

$$\begin{aligned} F(x,y) &= -1 + \frac{1}{1-x} + \frac{1}{1-y} - 2xyF(x,y) \\ &\quad + 2\sum_{m=0}^{\infty} \frac{1}{1-y} \sum_{i=0}^{\infty} \mathrm{A}(m,i)x^{m+1}y^{i+1} + 2\sum_{n=0}^{\infty} \frac{1}{1-x} \sum_{i=0}^{\infty} \mathrm{A}(i,n)x^{i+1}y^{n+1} \\ &= -1 + \frac{1}{1-x} + \frac{1}{1-y} - 2xyF(x,y) + \frac{2xy}{1-y}F(x,y) + \frac{2xy}{1-x}F(x,y). \end{aligned}$$

Thus we have

$$\left(1 + 2xy - \frac{2xy}{1-x} - \frac{2xy}{1-y}\right) F(x,y) = -1 + \frac{1}{1-x} + \frac{1}{1-y},$$

and multiplying through by $(1-x)(1-y) = 1 - x - y + xy$, we have

$$(1 - x - y - xy + 2x^2y^2)F(x,y) = 1 - xy.$$

This leads to the following result.

Theorem 3 *The generating function* $F(x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \mathrm{A}(m,n)x^m y^n$ *satisfies*

$$F(x,y) = \frac{1-xy}{1-x-y-xy+2x^2y^2}.$$

As noted above, it is clear from this result that $F(x,y) = F(y,x)$.

With this closed form of $F(x,y)$ in hand, we can prove the following simplified recurrence for $\mathrm{A}(m,n)$:

Theorem 4 *For all* $m,n \geq 2$,

$$\begin{aligned}\mathrm{A}(m,n) &= \mathrm{A}(m-1,n) + \mathrm{A}(m,n-1) \\ &\quad + \mathrm{A}(m-1,n-1) - 2\mathrm{A}(m-2,n-2)\end{aligned}$$

with the initial conditions $\mathrm{A}(m,0) = \mathrm{A}(0,n) = 1$ *for all* $m,n \geq 0$, $\mathrm{A}(m,1) = 2m$ *for* $m \geq 1$, *and* $\mathrm{A}(1,n) = 2n$ *for* $n \geq 1$.

Proof. By Theorem 3, we have

$$(1-x-y-xy+2x^2y^2)\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathrm{A}(m,n)x^m y^n = 1-xy$$

and it follows that

$$\begin{aligned}&\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathrm{A}(m,n)x^m y^n - \sum_{m=1}^{\infty}\sum_{n=0}^{\infty}\mathrm{A}(m-1,n)x^m y^n - \sum_{m=0}^{\infty}\sum_{n=1}^{\infty}\mathrm{A}(m,n-1)x^m y^n \\ &\qquad - \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\mathrm{A}(m-1,n-1)x^m y^n + 2\sum_{m=2}^{\infty}\sum_{n=2}^{\infty}\mathrm{A}(m-2,n-2)x^m y^n \\ &\qquad = 1-xy.\end{aligned}$$

Comparing coefficients of the monomial $x^m y^n$ for $m,n \geq 2$ gives the recurrence relation stated in the theorem. The conditions $\mathrm{A}(m,0) = \mathrm{A}(0,n) = 1$ are known from above. Comparing coefficients of xy, we get $\mathrm{A}(1,1) = 2$. Looking at $x^m y$ for $m \geq 2$, we have $\mathrm{A}(m,1) = \mathrm{A}(m-1,1) + \mathrm{A}(m,0) + \mathrm{A}(m-1,0) = \mathrm{A}(m-1,1) + 2$, and induction gives $\mathrm{A}(m,1) = 2m$ for $m \geq 1$. By symmetry, $\mathrm{A}(1,n) = 2n$ for $n \geq 1$. ∎

Although it is true that we can obtain a closed formula for $\mathrm{A}(m,n)$ analogous to (7), we note that such a closed formula is more complicated than that in (7) and so is omitted here.

References

[1] M. Covington, The number of distinct alignments of two strings, *J. Quantitative Linguistics* **11**, no. 3 (2004), 173-182.

[2] E. Ukkonen, Algorithms for approximate string matching, *Information and Control* **64** (1985), 100-118.