

10-1986

## Sequence Alignment with Matched Sections

Jerrold R. Griggs

*University of South Carolina - Columbia*, griggs@math.sc.edu

Philip J. Hanlon

Michael S. Waterman

Follow this and additional works at: [https://scholarcommons.sc.edu/math\\_facpub](https://scholarcommons.sc.edu/math_facpub)



Part of the [Genetics and Genomics Commons](#), and the [Mathematics Commons](#)

---

### Publication Info

*Siam Journal on Algebraic and Discrete Methods*, Volume 7, Issue 4, 1986, pages 604-608.

© 1986 by Society for Industrial and Applied Mathematics

This Article is brought to you by the Mathematics, Department of at Scholar Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of Scholar Commons. For more information, please contact [digres@mailbox.sc.edu](mailto:digres@mailbox.sc.edu).



to estimate the complexity of this task it is of interest to count the number of alignments for two sequences of two given lengths. There are previous results on this problem. H. T. Laquer [4] solves a more general recursion equation and relates the number of sequence alignments to the Stanton-Cowan numbers.

Frequently biologists find an alignment more believable when the matches occur in larger blocks. We will represent alignments as rows of 0's and 1's where a 1 indicates presence of a letter or nucleotide and a 0 indicates a gap. Figure 1(c) and 1(d) convert the alignments of Fig. 1(a) and 1(b) into these 0-1 rows. In this paper we count the alignments where the matching 1's must occur in blocks of  $b$  or more. In Fig. 1(a) and 1(c),  $b = 1$  while in Fig. 1(b) and 1(d),  $b \leq 3$ .

Let  $g(b, n)$  denote the number of alignments of two sequences of size  $n$  in which matching sections have size at least  $b$ . Equivalently,  $g(b, n)$  is the number of  $(0, 1)$ -matrices with 2 rows and an unspecified number of columns such that both rows contain precisely  $n$  1's, each column contains at least one 1, and columns with two 1's occur in adjacent sections of size  $b$  or more. We are interested in the asymptotic behavior of  $g(b, n)$  for fixed  $b$  as  $n \rightarrow \infty$ , as a function of  $b$ .

Observe that alignments where no column sum equals 2 are simply permutations of  $n$  columns with a single 1 in row 1 and  $n$  columns with a single 1 in row 2. Those are satisfactory for any  $b$ . Thus for all  $b$  and  $n$ ,

$$(1) \quad g(b, n) \cong \binom{2n}{n}.$$

Applying Stirling's formula as  $n \rightarrow \infty$  with  $b$  fixed,

$$(2) \quad g(b, n) \cong ((\pi n)^{-1/2})(4^n + o(1)) \quad \text{as } n \rightarrow \infty.$$

Further, note that  $g(1, n)$  counts the total number of 2-sequence alignments. A generating function approach is successful for the general problem of  $b \geq 1$ .

**THEOREM 1.** *Let  $b \geq 1$ . Define*

$$h(x) = (1 - x)^2 - 4x(x^b - x + 1)^2$$

and let  $\rho$  be the smallest positive real root of  $h(x) = 0$ . Then

$$g(b, n) \sim (\gamma_b n^{-1/2}) D_b^n \quad \text{as } n \rightarrow \infty,$$

where  $D_b = \rho^{-1}$  and

$$\gamma_b = (\rho^b - \rho + 1)(-\pi \rho h'(\rho))^{-1/2}.$$

*Proof.* Assume that  $b$  is fixed,  $b \geq 1$ . Let  $G(x) = \sum_{n \geq 0} g(b, n)x^n$  denote the ordinary generating function for the numbers  $g(b, n)$ . In order to obtain  $G(x)$  we first form the generating function  $\phi_m(x)$  for the numbers of 2-sequence alignments in which there are precisely  $m$  columns each of the forms  $\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}$  and  $\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}$  and in which the columns  $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}$  come in sections of at least  $b$ . As noted above, there are  $\binom{2m}{m}$  ways to order the  $2m$  columns with sum 1. This contributes a factor of  $\binom{2m}{m}x^m$  to  $\phi_m(x)$  since each row gets  $m$  1's from these  $2m$  columns. Next observe that there are  $2m + 1$  slots into which may be inserted either no  $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}$  columns or at least  $b$   $\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}$  columns. These slots precede, go between, and follow the  $2m$  columns with one 1. So each such slot contributes a factor, call it  $y = y(x)$ , to  $\phi_m(x)$ , where

$$\begin{aligned} y &= y(x) = 1 + x^b + x^{b+1} + \dots \\ &= 1 + (x^b)/(1 - x) \\ &\Rightarrow y = (x^b - x + 1)/(1 - x). \end{aligned}$$

Hence,

$$(3) \quad \phi_m(x) = \binom{2m}{m} x^m y^{2m+1}.$$

We obtain (3) since each alignment coded by  $\phi_m(x)$  is determined completely by the permutation of its columns with sum 1 and by the number  $s$  of  $\frac{1}{4}$  columns inserted into each slot. Such an alignment of size  $n$  contributes a term  $x^n$  to the sum  $\phi_m(x)$ .

The set of all 2-sequence alignments with columns  $\frac{1}{4}$  in groups of size at least  $b$  is the union over  $m \geq 0$  of the alignments enumerated by the series  $\phi_m(x)$ . Hence we obtain:

$$\begin{aligned} G(x) &= \sum_{m \geq 0} \phi_m(x) \\ &= \sum_{m \geq 0} \binom{2m}{m} x^m y^{2m+1} \\ &= y \sum_{m \geq 0} \binom{2m}{m} (xy^2)^m. \end{aligned}$$

Applying the Binomial theorem,

$$G(x) = y(1 - 4xy^2)^{-1/2}.$$

Plugging in for  $y$ , we obtain

$$G(x) = (x^b - x + 1)(h(x))^{-1/2},$$

where

$$h(x) = (1-x)^2 - 4x(x^b - x + 1)^2$$

or

$$h(x) = 1 - 6x + 9x^2 - 4x^3 - 8x^{b+1} + 8x^{b+2} - 4x^{2b+1}.$$

Observe that  $h(0) = 1$  and  $h(\frac{1}{4}) = (\frac{3}{4})^2 - ((\frac{1}{4})^b + \frac{3}{4})^2 < 0$ , so that  $h$  has a real root in  $(0, \frac{1}{4})$ . Let  $\rho$  be the smallest such root of  $h$ . The radius of convergence of  $G(x)$  is determined by the roots of  $h(x)$ , so the following lemma implies that  $G(x)$  has radius of convergence  $\rho$ .

**LEMMA.** *The unique root of  $h(x)$  with the smallest modulus is  $\rho$ , and  $\rho$  is a single root of  $h(x)$ .*

*Proof of Lemma.* Let  $z \in C$ ,  $|z| \leq \rho$ , be a root of  $h(z)$ . We first show that in fact  $|z| = \rho$  must hold. We have that

$$h(z) = (1-z)^2 4z \left( \frac{1}{4z} - \frac{(z^b - z + 1)^2}{(1-z)^2} \right) = 0.$$

Since  $0 < z < \frac{1}{4}$ , it follows that

$$\frac{1}{4z} = \left( 1 + \frac{z^b}{1-z} \right)^2,$$

so that

$$\begin{aligned} \frac{1}{4\rho} &\leq \frac{1}{4|z|} = \left| 1 + \frac{z^b}{1-z} \right|^2 \\ &\leq \left( 1 + \frac{|z|^b}{1-|z|} \right)^2 \\ &\leq \left( 1 + \frac{\rho^b}{1-\rho} \right)^2. \end{aligned}$$

Next we observe that because  $\rho$  is a root of  $h$ ,

$$\frac{1}{4\rho} = \left( 1 + \frac{\rho^b}{1-\rho} \right)^2,$$

which implies that the inequalities above are all equalities. It follows that  $|z| = \rho$ . (This could have been deduced instead from the well-known fact that a series  $f(z) = \sum_{n=0}^{\infty} a_n z^n$  with real coefficients  $a_n \geq 0$  and with radius of convergence  $\rho > 0$  has a singular point at  $z = \rho$  ([5]; confer, e.g., [3]).)

We next observe that

$$\left| 1 + \frac{z^b}{1-z} \right| = 1 + \frac{|z|^b}{1-|z|},$$

where  $|z| = \rho \in (0, \frac{1}{4})$  forces

$$1 + \frac{z^b}{1-z} = 1 + \frac{|z|^b}{1-|z|},$$

so that  $1/4z = 1 + (z^b/(1-z))$  is real and positive. Hence  $z$  itself is real and positive, which implies that  $z$  must be  $\rho$ . Thus  $\rho$  is the unique root with the smallest modulus.

One can then calculate that

$$h'(\rho) = (1-\rho)(-1-\rho^{-1}-4b\rho^{(2b-1)/2}+4\rho^{1/2}).$$

It follows easily from  $\rho \in (0, \frac{1}{4})$  that  $h'(\rho) < 0$ . Therefore  $\rho$  is only a single root of  $h(z)$ . This completes the proof of the lemma.

Returning to the theorem, we define functions  $s(x)$ ,  $A(x)$ ,  $B(x)$  by:

$$\begin{aligned} h(x) &= (\rho - x)s(x), \\ A(x) &= (x^b - x + 1)(s(x))^{-1/2}, \\ B(x) &= (\rho - x)^{-1/2}. \end{aligned}$$

Then we have that

$$G(x) = A(x)B(x).$$

Here  $A(x)$  has radius of coverage  $> \rho$  since it follows from the lemma that  $s(x)$  has not root  $z$  with  $|z| \leq \rho$ . Also,  $B(x)$  has radius of convergence  $\rho$ . Again by the binomial theorem,

$$B(x) = (\rho - x)^{-1/2} = \rho^{-1/2} \left( 1 - \frac{x}{\rho} \right)^{-1/2} = \rho^{-1/2} \sum_{n \geq 0} \binom{2n}{n} \left( \frac{x}{4\rho} \right)^n,$$

so that

$$B(x) = \sum_{n \geq 0} b_n x^n$$

where

$$b_n = \rho^{-1/2} \binom{2n}{n} (4\rho)^{-n}.$$

It remains to observe that  $(b_{n-1}/b_n) \rightarrow \rho$  as  $n \rightarrow \infty$  to apply a theorem of Bender [1, Thm. 2] to  $G(x) = A(x)B(x)$  to deduce that

$$g(b, n) \sim A(\rho)b_n \text{ as } n \rightarrow \infty.$$

Of course, to calculate  $A(\rho)$ , we are taking  $s(\rho) = \lim_{x \rightarrow \rho} (h(x)/(\rho - x)) = -h'(\rho)$ . The theorem now follows immediately.

Table 1 lists some values of  $D_b$  and  $\gamma_b$  to 4 or more places. These were computed on a hand computer, using Newton's method to find the root  $\rho$  for each  $b$ .

$$g(b, n) \sim (\gamma_b n^{-1/2}) D_b^n \text{ as } n \rightarrow \infty,$$

where  $D_b = \rho^{-1}$  and  $\gamma_b = (\rho^b - \rho + 1)(-\pi\rho h'(\rho))^{-1/2}$ .

For comparison, recall that from (2), for all  $b$ ,  $g(b, n) \cong \binom{2n}{n} \sim (.5641896)n^{-1/2}4^n$  as  $n \rightarrow \infty$ . Table 1 also suggests what happens to  $D_b$  and  $\gamma_b$  as  $b \rightarrow \infty$ , which is straightforward to derive from the observation that as  $b \rightarrow \infty$  the smallest root of  $h(x)$ ,  $\rho$ , increases and approaches  $\frac{1}{4}$ :

TABLE 1

$b$	$D_b$	$\gamma_b$
1	5.8284	.57268
2	4.5189	.53206
3	4.1489	.54290
4	4.0400	.55520
5	4.0103	.56109
10	4.00001	.564183

COROLLARY. As  $b \rightarrow \infty$ ,  $D_b \rightarrow 4$  and  $\gamma_b \rightarrow \pi^{-1/2}$ .

Acknowledgment. We thank David Richman for his helpful observations.

REFERENCES

[1] E. A. BENDER, *Asymptotic methods in enumeration*, SIAM Rev., 16 (1974), pp. 485-515.  
 [2] W. FITCH AND T. SMITH, *Optimal sequence alignments*, Proc. National Academy of Science, 80 (1983), pp. 1382-1386.  
 [3] K. KNAPP, *Problem book in the theory of functions I*, Dover, New York, 1948, problem 11-3, p. 30.  
 [4] H. T. LAQUER, *Asymptotic limits for a two-dimensional recursion*, Stud. Appl. Math., 64 (1981), pp. 271-277.  
 [5] R. P. STANLEY, personal communication.  
 [6] R. G. STANTON AND D. D. COWAN, *Note on a 'square functional' equation*, SIAM Rev., 12 (1970), pp. 277-279.  
 [7] M. S. WATERMAN, *General methods of sequence comparison*, Bull. Math. Biol., 46 (1984), pp. 473-500.