SIAM J. ALG. DISC. METH.
Vol. 7, No. 4, October 1986

SEQUENCE ALIGNMENTS WITH MATCHED SECTIONS*

JERROLD R. GRIGGS†, PHILIP J. HANLON‡ AND MICHAEL S. WATERMAN¶

Abstract. In molecular biology, two finite sequences are compared by displaying one sequence written over another in an alignment. The number of alignments of two sequences is related to the Stanton-Cowan numbers. This paper gives asymptotics for the number of alignments of two sequences of length n with *matching sections of size at least* b.

Key words. sequence alignments, generating functions, Stanton-Cowan numbers

AMS(MOS) subject classifications. 05A15, 92A10

Mathematics has played an important role in modern molecular biology in the area of sequence comparison. When nucleic acid (DNA or RNA) or protein sequences are determined, the question of relationships between sequences arises. Frequently two (or more) sequences are compared by dynamic programming or other methods to produce one or more sequence alignments which display one sequence written over another. When one letter (nucleotide in DNA) is written above another, they are presumed to have a common evolutionary ancestor. When a gap appears above or below a letter, the evolutionary event of insertion or deletion is assumed to have taken place. A review of methods to perform this analysis appears in Waterman [7].

An example of two different alignments of two sequences appears in Fig. 1(a) and 1(b) (Fitch and Smith [2]). The upper sequence is chicken β-hemoglobin messenger RNA (mRNA), nucleotides 115–171, and the lower sequence is chicken α-hemoglobin mRNA, nucleotides 118–156. These mRNA sequences are transcribed into hemoglobin protein molecules and are well known to have arisen from a common ancestor. In fact so many hemoglobin sequences are known that the alignment is presumed known, and the paper of Fitch and Smith is a study of the ability of various alignment algorithms to produce correct results.

As is easy to imagine, many ad hoc methods have arisen to align sequences. The most naive simply look at the sequences and perform the alignment visually. In order

(a) UUUGCGUCCUUUGGAAC CUCUCCAGCCCCA CUG C CAUCCUUGGCAA CC C CAUGG UC
UUU C CC CACU UC G AUCUGUCACA C GGC UCCGCUCA AAUC

(b) UUUGCGUCCUUUGGAACCUCUCCAGCCCCAGUGCCAUCCUUGGCAACCCCAUGGUC
UUUCCCCACUUCG AUCU GUCACACGGCUCCGCU CAAAUC

(c) 11111111111111111011111111111111011101011111111111101101011111 0011
11101001100000001111011001000011111111110100011100111111110001111

(d) 11111111111111111111111111111111111111111111111111111111111111
11111111111100111100000111111111111111110001110000000111

FIG. 1. (a) *and* (b) *are two alignments of nucleotides* 115–171 *of chicken* β*-hemoglobin* mRNA (*upper*) *and nucleotides* 118–156 *of chicken* α*-hemoglobin* mRNA (*lower*). (c) *and* (d) *are* 0–1 *representations of* (a) *and* (b), *respectively.*

* Received by the editors July 29, 1985, and in revised form February 14, 1986.

† Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208. This research was supported in part by the National Science Foundation.

‡ Department of Mathematics, California Institute of Technology, Pasadena, California 91125. This research was supported in part by the National Science Foundation.

¶ Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089. This research was supported in part by the System Development Foundation.

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), $b\leqq 3$.

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 $n\to\infty$, 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,

$$g(b, n)\geqq\binom{2n}{n}. \tag{1}$$

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

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

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 $b\geqq 1$.

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

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

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

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

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

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

Proof. Assume that b is fixed, $b\geqq 1$. Let $G(x)=\sum_{n\geqq 0} g(b, n)x^n$ 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 $\binom{1}{0}$ and $\binom{0}{1}$ and in which the columns $\binom{1}{1}$ 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 $\binom{1}{1}$ columns or at least b $\binom{1}{1}$ 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

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

Hence,

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

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 $\frac{1}{1}$ 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 $\frac{1}{1}$ in groups of size at least b is the union over $m \geqq 0$ of the alignments enumerated by the series $\phi_m(x)$. Hence we obtain:

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

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 ρ 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 ρ.

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

Proof of Lemma. Let $z \in C$, $|z| \leqq \rho$, 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

$$\frac{1}{4\rho} \leqq \frac{1}{4|z|} = \left|1+\frac{z^b}{1-z}\right|^2$$
$$\leqq \left(1+\frac{|z|^b}{1-|z|}\right)^2$$
$$\leqq \left(1+\frac{\rho^b}{1-\rho}\right)^2.$$

Next we observe that because ρ 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 $a_n \geqq 0$ 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 ρ. Thus ρ 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 ρ 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:

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

Then we have that

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

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

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

so that

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

where

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

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

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

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

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

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

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, $g(b, n) \geqq \binom{2n}{n} \sim (.5641896)n^{-1/2}4^n$ as $n \to \infty$. Table 1 also suggests what happens to D_b and γ_b as $b \to \infty$, which is straightforward to derive from the observation that as $b \to \infty$ the smallest root of $h(x)$, ρ, increases and approaches $\frac{1}{4}$:

TABLE 1

b	D_b	γ_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* $b \to \infty$, $D_b \to 4$ *and* $\gamma_b \to \pi^{-1/2}$.

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.