Graphs and Combinatorics 6, 133–146 (1990)

Graphs and Combinatorics

On the Number of Alignments of k Sequences

J.R. Griggs[1]*†, P. Hanlon[2]*, A.M. Odlyzko[3] and M.S. Waterman[4]*†‡

[1] Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA
[2] Department of Mathematics, Caltech, Pasadena, CA 91125, USA, *current address*: Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA
[3] AT&T Bell Laboratories, Murray Hill, NJ 07974, USA
[4] Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, CA 90089, USA

Abstract. Numerous studies by molecular biologists concern the relationships between several long DNA sequences, which are listed in rows with some gaps inserted and with similar positions aligned vertically. This motivates our interest in estimating the number of possible arrangements of such sequences. We say that a k sequence alignment of size n is obtained by inserting some (or no) 0's into k sequences of n 1's so that every sequence has the same length and so that there is no position which is 0 in all k sequences. We show by a combinatorial argument that for any fixed $k \geq 1$, the number $f(k,n)$ of k alignments of length n grows like $(c_k)^n$ *as* $n \to \infty$, where $c_k = (2^{1/k} - 1)^{-k}$. A multi-dimensional saddle-point method is used to give a more precise estimate for $f(k,n)$.

1. Introduction

Researchers in molecular biology are determining or reading DNA sequences at an increasing rate. As of spring 1986, over 6×10^6 letters of DNA are known and organized in a data base, GenBank. Biology is concerned with the inference of biological properties from these sequences. This is sometimes accomplished by study of the relationships between several sequences suspected of having common function or evolutionary history. These studies are performed by listing the sequences in rows, one sequence per row, with the sequences arranged so that similar positions of each sequence are aligned vertically. Gaps are sometimes inserted into some of the sequences to bring the letters into the desired alignment. These gaps greatly increase the computational complexity of the task. Mathematics has contributed to these analyses with the invention of efficient algorithms for sequence comparisons. See Waterman [5] for a review of these methods.

To give an example of these alignments we turn to a recent study by Ullrich et al [4] which presented some surprising and important findings. These workers determined the amino acid sequence of the human insulin receptor precursor, a

* Research supported in part by the National Science Foundation
† Research supported in part by the System Development Foundation
‡ Research supported in part by the National Institute of Health

protein. As is now common they found the sequence of the DNA encoding the protein instead of directly sequencing the protein. Once the sequence was known they then, by sequence comparison, found unexpected relationships with the human epidermal growth factor receptor and the members of the *src* family of oncogene products. Oncogenes and these proteins in particular are implicated in certain human cancers. Some of the sequence alignment is given below, with the human insulin receptor sequence first, followed by the growth factor sequence and then three oncogene protein sequences. The dashes represent gaps inserted into the sequences to achieve the alignment.

```
...LGQGSFGMVYEGNARDIIKGEAETRVAYKT...

...LGSGAFGTVYKGLWIPEGE-KVKIPVAIKE...

...LGGGQYGEVYEGVWKKYSL-----TVAVKT...

...LGQGCFGEVWMGTW--ND----TTRVAIKT...

...IGRGNFGEVFSGRL--RAD---NTLVAVKS...

...LGTGAFGKVVEATAFGLGKEDAVLKVAVKM...
```

It is natural to ask how many such arrangements of the sequences there are. For the case of two sequences, this question has been studied, beginning with Laquer [2]. He solved a general recursion and related the number of sequence alignments to the Stanton-Cowan numbers [3]. More recently Griggs et al [1] counted the number of alignments of two sequences of length n with matching sections of size at least b. In the present paper we study the number $f(k,n)$ of alignments of k sequences of length n. In section 2 we show by a combinatorial argument that the exponential growth rate of $f(k,n)$ is $(2^{1/k}-1)^{-k}$. In section 3 a multi-dimensional saddle-point method is used to give a more precise estimate.

2. An Exponential Growth Rate

We make the simplifying assumption here that all sequences have the same length n. We count only the possible alignments with respect to their relative positions, ignoring the actual elements in the sequences (biologically, the nucleotide bases or the amino acids in the genetic sequences).

We start with k sequences of 1's, with n 1's each. A *k sequence alignment of size n* is obtained by inserting some (or none) 0's in each sequence so that every sequence has the same *length*, call it L, and so that there is no position which is 0 in all k sequences. Thus, a k sequence alignment of size n corresponds to a $k \times L$ $(0,1)$-matrix with all row sums n and no column sums 0. The latter condition, no column sums 0, ensures that for given k and n the number of alignments is finite, and is motivated by the biological consideration that alignments are not allowed which have gaps (0's) in all k sequences. We see that an alignment has length at least n (when every column is all 1's) and at most kn (when each column has just a single 1). We wish to study asymptotically the number of k sequence alignments of size n, denoted $f(k, n)$.

We first consider how $f(k, n)$ behaves as $n \to \infty$ for fixed k. An indication of how quickly the number of alignments grows with the size n is given by merely counting the number of alignments with the maximum length, kn. In such alignments, there are, for each i, $1 \leq i \leq k$, n columns with a single 1 in row i and 0's in the other rows. Thus, the number of such alignments is the number of ways of ordering these kn columns, which equals the multinomial coefficient $\binom{kn}{n, n, \ldots, n}$. Applying Stirling's formula, one obtains that $f(k, n) \geq \binom{kn}{n, n, \ldots, n}$ which grows like $\frac{\sqrt{k}}{(2\pi n)^{(k-1)/2}}(k^k)^n$ as $n \to \infty$.

In this section, we describe the leading asymptotic behavior of $f(k, n)$, which means we find constants c_k such that $f(k, n)$ grows like $(c_k)^n$ as $n \to \infty$. The proof of Theorem 1 actually gives $f(k, n)$ up to a factor which is polynomial in n. Here we state the following weaker result.

Theorem 1. *For fixed* $k \geq 1$

$$\lim_{n \to \infty} \frac{\ln f(k, n)}{n} = \ln c_k$$

where $c_k = (2^{1/k} - 1)^{-k}$.

We can now describe neatly how the base c_k in the exponential term describing $f(n, k)$ grows with k. Let $d_k = \frac{1}{\sqrt{2}}\left(\frac{k}{\ln 2}\right)^k$

Proposition. *For all* $k \geq 1$, $c_k < d_k$, *and*, $c_k \sim d_k$ *as* $k \to \infty$.

We tabulate values of c_k for selected k, to four or more places (Table 1). For comparison, we tabulate also the lower bound k^k and the upper bound $d_k = 2^{-1/2}$ $(\ln 2)^{-k}\, k^k$. We see that d_k approximates c_k quite closely, even for k very small.

Proof of Theorem 1. Let $[k]$ denote the set $\{1, 2, \ldots, k\}$. A column C in the matrix corresponding to a k-alignment of sequences of size n has at least one 1, so the set S of indices of rows of C containing 1 is a nonempty subset of $[k]$. We call S the *type of* C. A k-alignment of size n is then constructed by taking some number A_S of columns of type S, for each $\varnothing \neq S \subseteq [k]$, and then arbitrarily permuting these columns of various types, where the A_S satisfy the conditions that for each row $i \in [k]$,

$$\sum \{A_S \mid i \in S\} = n. \qquad \langle 2.1 \rangle$$

Table 1

k	c_k	k^k	$d_k = \frac{1}{\sqrt{2}}(k/ln\,2)^k$
1	1	1	1.020
2	5.828	4	5.887
3	56.95	27	57.33
4	780.3	256	784.2
5	13,755.3	3125	13810.4
6	2.965×10^5	4.666×10^4	2.974×10^5
7	7.554×10^6	8.235×10^5	7.575×10^6
8	2.221×10^8	1.678×10^7	2.226×10^8
9	7.401×10^9	3.874×10^8	7.417×10^9
10	2.757×10^{11}	1.000×10^{11}	2.762×10^{11}
20	1.130×10^{29}	1.049×10^{26}	1.131×10^{29}

The number of k-alignments with A_S columns of type S, $\varnothing \neq S \subseteq [k]$, for A_S satisfying $\langle 2.1 \rangle$, is thus the number of permutations of these columns, the multinomial coefficient $\binom{\Sigma \{A_S | \varnothing \neq S \subseteq [k]\}}{\{A_S | \varnothing \neq S \subseteq [k]\}}$. In this symbol, the lower row denotes the multiset listing all sizes A_S, repetitions allowed, e.g., for $k = 2$, $n = 8$, $A_{\{1\}} = A_{\{2\}} = 5$ and $A_{\{1,2\}} = 3$, we have $\binom{13}{5,5,3} = \left(\frac{13!}{5!5!3!}\right)$. It follows that the total number of alignments is given by

$$f(k,n) = \sum \binom{\sum \{A_S\}}{\{A_S\}}, \qquad \langle 2.2 \rangle$$

where the sum is over all possible parameters A_S satisfying $\langle 2.1 \rangle$.

With k and n fixed, the number of terms in this summation, i.e., the number of solutions to $\langle 2.1 \rangle$, is no more than, say $(n+1)^{2^k-1}$, because each of the $2^k - 1$ parameters A_S is restricted to the range $0 \leq A_S \leq n$ by $\langle 2.1 \rangle$.

We will find the largest term in the sum $\langle 2.2 \rangle$, i.e., the largest exponential growth from among all the terms in the sum $\langle 2.2 \rangle$. The asymptotic behavior of $f(n,k)$ is at most n^{2^k-1} times the largest such term, and this polynomial factor is dominated asymptotically for fixed k as $n \to \infty$ by the exponential. This reduces the whole problem to locating the maximum value of $\binom{\Sigma \{A_S\}}{\{A_S\}}$ subject to $\langle 2.1 \rangle$. Assume henceforth that we are at such a maximum.

Without loss of generality, we may assume that for $0 < j \leq k$ the parameters A_S with $|S| = j$ are approximately equal. That is, for n large, we may assume there exists some number $\alpha_j \in [0,1]$ such that for all $|S| = j$, $|A_S| \sim \alpha_j n$. The reason is that if $\{A_S | \varnothing \neq S \subseteq [k]\}$ is a maximum, then we can replace each size A_S, $|S| = j$, by the average $(\sum \{A_S \,||S| = j\})/\binom{k}{j}$. The principle being used here is that a multinomial coefficient $\binom{N}{t_1, \ldots, t_m}$ is increased by making the numbers $t_1, \ldots, t_n$ more nearly equal. These new parameters satisfy $\langle 2.1\rangle$ easily, because each row has the same number of 1's, by symmetry, which must be n. (We actually require all of these averages $\sum\{A_S \mid |S| = j\}/\binom{k}{j}$ to be integers, but only really require below that the A_S, $|S| = j$, be equal asymptotically, so that integer values of such A_S, nearly equal, can be arranged.)

Thus we must now maximize $\binom{\Sigma\{\alpha_j n\}}{\{\alpha_j n\}}$, where $\{\alpha_j n\}$ denotes the multiset with values $\{\alpha_j n\}$ repeated $\binom{k}{j}$ times each, and the $\alpha_j \in [0,1]$ satisfy

$$\sum_{j=1}^{k} \binom{k-1}{j-1} \alpha_j = 1, \qquad \langle 2.3\rangle$$

since for any row, say row 1, there are $\binom{k-1}{j-1}$ choices for $S \subseteq [k]$ with $1 \in S$ and $|S| = j$, and for each such S there are $\alpha_j n$ 1's in row 1 which contribute to the total number or 1's.

Again consider a collection of α_j's and corresponding columns which maximize the total $\binom{\Sigma\{\alpha_j n\}}{\{\alpha_j n\}}$. For any s, $2 \le s \le k$ such that $\alpha_s > 0$, we take a column with s 1's and replace it by two columns, one with one 1 and one with $(s-1)$ 1's and count the number of alignments with the new set of parameters A_S. This new value must be no more than the one before, by the optimality hypothesis. Let $l = \sum_{j=1}^{k} \binom{k}{j} \alpha_j$, that is, $l = L/n$ where L is total length of the alignments for the optimal selection. Then the ratio of the number of the alignments at optimality divided by the number obtained after replacing the one column by two must be at least one. Most of the factorials are identical and cancel out. However, the new collection has one more column total (which contributes a factor $\sim ln$): one more column each of size 1 (contributing a factor $\sim \alpha_1 n$) and size $s-1$ (contributing a factor $\sim \alpha_{s-1} n$), and one less of size s (losing a factor $\sim \alpha_s n$). Considering where the factors came from, one obtains for large n:

$$\frac{(\alpha_1 n)(\alpha_{s-1} n)}{(ln)(\alpha_s n)} \ge 1. \qquad \langle 2.4\rangle$$

Conversely, if $\alpha_1 > 0$ and $\alpha_{s-1} > 0$, $2 \le s \le k$, if we replace two columns with 1 and $s-1$ 1's, respectively, from disjoint sets of rows, we obtain the reverse of the above:

$$\frac{(\alpha_1 n)(\alpha_{s-1} n)}{(ln)(\alpha_s n)} \le 1. \qquad \langle 2.5\rangle$$

Since some value of s has $\alpha_s > 0$ at optimality, $\langle 2.4\rangle$ forces $\alpha_1 > 0$, which then implies by repeated application of $\langle 2.5\rangle$ that all $\alpha_j > 0$. Then by simplifying and combining $\langle 2.4\rangle$ and $\langle 2.5\rangle$, we obtain

$$\frac{\alpha_{s-1}}{\alpha_s} = \frac{l}{\alpha_1}, \qquad s = 2, \ldots, k.$$

Letting r denote this ratio, l/α_1, we obtain:

$$\begin{aligned} \alpha_j &= r^{k-j}\alpha_k, \qquad j = 1, \dots, k, \\ l &= r^k \alpha_k. \end{aligned} \tag{2.6}$$

Since l is defined in terms of the α_j's, plugging in with ⟨2.6⟩ yields:

$$r^k \alpha_k = \sum_{j=1}^{k} \binom{k}{j} r^{k-j} \alpha_k.$$

So

$$\begin{aligned} r^k &= \sum_{j=1}^{k} \binom{k}{j} r^{k-j} \\ &= (1+r)^k - r^k. \end{aligned}$$

Thus

$$\begin{aligned} 2r^k &= (1+r)^k, \\ r &= (2^{1/k}-1)^{-1}. \end{aligned} \tag{2.7}$$

Next we apply the equations ⟨2.3⟩, ⟨2.6⟩, and ⟨2.7⟩ to obtain

$$\begin{aligned} \alpha_j &= 2^{-(k-1)/k}(2^{1/k}-1)^{j-1}, \qquad j = 1, \dots, k, \\ l &= 2^{-(k-1)/k}(2^{1/k}-1)^{-1}. \end{aligned} \tag{2.8}$$

Now insert these parameters in terms of α_k and r into $\binom{\sum \{\alpha_j n\}}{\{\alpha_j n\}}$ and apply Stirling's formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$ as $n \to \infty$:

$$\begin{aligned} \binom{r^k \alpha_k n}{\{r^{k-j}\alpha_k n\}} &= \frac{(r^k \alpha_k n)!}{\prod_{j=1}^{k} ((r^{k-j}\alpha_k n)!)^{\binom{k}{j}}} \\ &\sim \frac{(r^k \alpha_k n)^{r^k \alpha_k n}}{\prod_{j=1}^{k} (r^{k-j}\alpha_k n)^{\binom{k}{j} r^{k-j}\alpha_k n}}. \end{aligned}$$

Here the powers of e all cancelled and we disregarded the factor of $e^{O(\ln n)}$. Then the powers of $\alpha_k n$ all cancel, due to conditions on the α_j's, and it all reduces to a power of the ratio r. Then realizing that we are really looking at $f(k,n)$, taking logs, and dividing by n, we obtain

$$\frac{\ln f(k,n)}{n} \sim (\ln r)\left(k r^k \alpha_k - \sum_{j=1}^{k} (k-j)\binom{k}{j} r^{k-j}\alpha_k\right). \tag{2.9}$$

Next use $\sum_{j=1}^{k} (k-j)\binom{k}{j} r^{k-j} = \sum_{j=1}^{k} \binom{k-1}{j} r^{k-j} = kr((1+r)^{k-1} - r^{k-1})$, and plug in ⟨2.7⟩ for r and ⟨2.8⟩ for α_k, and simplify to obtain

$$\frac{\ln f(k,n)}{n} \to k \ln(2^{1/k}-1)^{-1},$$

which proves the theorem. □

Proof of Proposition. The assertion $c_k \le d_k$ follows by showing that for all $x > 0$.

$$\left((2^{1/x} - 1)\frac{x}{\ln 2}\right)^x \ge \sqrt{2},$$

which is equivalent to

$$(2^{1/x} - 1)\frac{x}{\ln 2} > 2^{1/(2x)} \qquad (x > 0)$$

Substituting $u = \dfrac{\ln 2}{2x}$, so that $e^u = 2^{1/(2x)}$, we must show that

$$(e^{2u} - 1)\frac{1}{2u} > e^u \qquad (u > 0)$$

or

$$e^{2u} - 1 > 2ue^u \qquad (u > 0)$$

It is simple to verify either by taking derivatives or by looking at the series expansion.

To verify that $c_k \sim d_k$, we look at the exponential series:

$$\begin{aligned} c_k &= (2^{1/k} - 1)^{-k} \\ &= (e^{\ln 2/k} - 1)^{-k} \\ &= \left(\frac{\ln 2}{k} + \frac{(\ln 2)^2}{2!k^2} + O(k^{-3})\right)^{-k} \\ &= \left(\frac{\ln 2}{k}\right)^{-k}\left(1 + \frac{\ln 2}{2k} + O(k^{-2})\right)^{-k}. \end{aligned}$$

This line is equal to $\left(\dfrac{\ln 2}{k}\right)^{-k} 2^{-1/2}(1 + O(k^{-1}))$ which completes the proof.

3. A More Precise Estimate for $f(k, n)$

Recall that $f(k, n)$ is the number of 0,1 matrices with no column of all 0's and with every row sum equal to n. In the last section we used a combinatorial argument to show that exponential growth rate of $f(k, n)$ is $(2^{1/k} - 1)^{-k}$. In this section we use analytic approximation methods to give a more precise estimate for $f(k, n)$. We will prove the following result.

Theorem 2. *For a positive integer k, let $\rho = 2^{1/k} - 1$ and let $r = \rho^{-k} = (2^{1/k} - 1)^{-k}$. Then*

$$f(k, n) = \left(\frac{r^n}{n^{(k-1)/2}}\right)\left((\rho\pi^{(k-1)/2}k^{1/2})^{-1}2^{(k^2-1)/2k} + O(n^{-1/2})\right).$$

The proof of this result, which uses the multi-dimensional saddle-point method, will take the rest of this section.

Definition 3.1. For $r_1, \ldots, r_k$ non-negative integers, define $N(r_1,\ldots,r_k)$ to be the number of 0,1 matrices with no column of all 0's and with the jth row sum equal to r_j.

The reader should note that $f(k,n) = N(n,n,\ldots,n)$.

Proposition 3.1. *With notation as above we have*

$$\sum_{r_1,\ldots,r_k} N(r_1,\ldots,r_k)z_1^{r_1}\ldots z_k^{r_k} = \left(2 - \prod_{j=1}^{k} (1+z_j)\right)^{-1}. \qquad \langle 3.1\rangle$$

Proof. Let $N_u(r_1,\ldots,r_k)$ be the number of k by u 0,1 matrices having no column of all 0's and with row sums $r_1, \ldots, r_k$. It is easy to see that

$$\sum_{r_1,\ldots,r_k} N_u(r_1,\ldots,r_k)z_1^{r_1}\ldots z_k^{r_k} = \left(\prod_{j=1}^{k} (1+z_j) - 1\right)^{u}.$$

As $N(r_1,\ldots,r_k) = \sum_{u=0}^{\infty} N_u(r_1,\ldots,r_k)$ we have

$$\sum_{r_1,\ldots,r_k} N(r_1,\ldots,r_k)z_1^{r_1}\ldots z_k^{r_k} = \left(1 - \left(\prod_{j=1}^{k} (1+z_j) - 1\right)\right)^{-1}$$

$$= \left(2 - \prod_{j=1}^{k} (1+z_j)\right)^{-1}. \qquad \square$$

We want an estimate for $N(n,n,\ldots,n)$. We begin by estimating $N(r_1,\ldots,r_k)$ using Cauchy's Theorem and $\langle 3.1\rangle$. Although Cauchy's Theorem could be applied directly to $\langle 3.1\rangle$ it is more convenient to peel off one variable z_k. For notational convenience we let $A(z) = A(z_1,\ldots,z_{k-1})$ be the polynomial

$$A(z) = \prod_{j=1}^{k-1} (1+z_j).$$

With this notation, the left hand side of $\langle 3.1\rangle$ can be rewritten as

$$\frac{1}{2 - A(z)(1+z_k)} = \sum_{r=0}^{\infty} \frac{A(z)^r}{(2-A(z))^{r+1}} z_k^r. \qquad \langle 3.2\rangle$$

So for fixed r we have

$$\sum_{r_1,\ldots,r_{k-1}} N(r_1,\ldots,r_{k-1},r)z_1^{r_1}\ldots z_{k-1}^{r_{k-1}} = \frac{A(z)^r}{(2-A(z))^{r+1}}. \qquad \langle 3.3\rangle$$

Applying Cauchy's Theorem to $\langle 3.3\rangle$ we have

$$N(r_1,\ldots,r_{k-1},r) = \frac{1}{(2\pi i)^{k-1}} \int \cdots \int \frac{A(z)^r}{(2-A(z))^{r+1}} \frac{dz_1}{z_1^{r_1+1}} \cdots \frac{dz_{k-1}}{z_{k-1}^{r_{k-1}+1}}. \qquad \langle 3.4\rangle$$

In $\langle 3.4\rangle$, the integrals are taken around circles $|z_j| = \rho_j$ where the ρ_j are chosen so that $A(z)^r(2 - A(z))^{-(r+1)}$ is analytic in the region $|z_j| < \rho_j$.

Let $\rho = 2^{1/k} - 1$ and choose all the ρ_j equal to ρ. This choice of the ρ_j is motivated by the fact that the integrand in $\langle 3.4\rangle$ has a saddle-point at $(z_1,\ldots,z_{k-1}) =$

$(\rho, \rho, \ldots, \rho)$. We need to verify that $A(z)^r(2 - A(z))^{-(r+1)}$ is analytic in the region $|z_j| \leq \rho$. To see this note that if $|z_j| \leq \rho$ then $|1 + z_j| \leq 2^{1/k}$ so $|A(z)| \leq 2^{(k-1)/k}$. Hence for $(z_1, \ldots, z_{k-1})$ in the region $|z_j| \leq \rho$ we have

$$|2 - A(z)| \geq 2 - 2^{(k-1)/k} = \rho(1 + \rho)^{k-1}. \qquad \langle 3.5\rangle$$

This proves analyticity in the desired region. Setting all the r_j equal to n we obtain the following expression for $f(k, n)$.

$$f(k,n) = \frac{1}{(2\pi i)^{k-1}} \underset{|z_j| = \rho}{\int \cdots \int} \frac{A(z)^n}{(2 - A(z))^{n+1}} \frac{dz_1}{z_1^{n+1}} \cdots \frac{dz_{k-1}}{z_{k-1}^{n+1}}. \qquad \langle 3.6\rangle$$

Make the substitution $z_j = \rho e^{i\theta_j}$, $j = 1, 2, \ldots, k - 1$. The range of integration on θ_j will be $-\pi \leq \theta_j \leq \pi$ and we have

$$\frac{dz_j}{z_j^{n+1}} = i\rho^{-n} e^{-in\theta_j} d\theta_j.$$

Substituting in $\langle 3.6\rangle$ we obtain

$$f(k,n) = \frac{\rho^{-(k-1)n}}{(2\pi)^{k-1}} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \left(\frac{\prod_{j=1}^{k-1} (1 + \rho e^{i\theta_j}) e^{-i\theta_j}}{2 - \prod_{j=1}^{k-1} (1 + \rho e^{i\theta_j})} \right)^n \frac{d\theta_1 \ldots d\theta_{k-1}}{2 - \prod_{j=1}^{k-1} (1 + \rho e^{i\theta_j})}. \qquad \langle 3.7\rangle$$

The integrand in $\langle 3.7\rangle$ takes on its largest absolute value when all the θ_j's are 0. More generally we will show the bulk of the contribution to the integral on the right hand side of $\langle 3.7\rangle$ occurs in a small region near $(\theta_1, \ldots, \theta_{k-1}) = (0, \ldots, 0)$ (for n large).

Let $\delta = (\ln n)/\sqrt{n}$. Write the right hand side of $\langle 3.7\rangle$ as $I_1 + I_2$ where I_1 involves the integral over that part of the region where some θ_j is in the range $\delta \leq |\theta_j| \leq \pi$ and I_2 involves the integral over the region $-\delta \leq \theta_j \leq \delta$ $(j = 1, 2, \ldots, k - 1)$. We will show that for n large I_1 is small compared to I_2 and we will compute the asymptotic value of I_2. Our first step is to bound the absolute value of I_1.

Note that

$$|1 + \rho e^{i\theta}|^2 = (1 + \rho\cos\theta)^2 + \rho^2 \sin^2\theta = 1 + 2\rho\cos\theta + \rho^2. \qquad \langle 3.8\rangle$$

If $0 \leq \varepsilon \leq \pi$ then $1 + 2\rho\cos\theta + \rho^2 \leq 1 + 2\rho\cos(\varepsilon) + \rho^2$ for $\varepsilon \leq |\theta| \leq \pi$. So there is a positive real $c_\varepsilon \neq 0$ satisfying

$$1 + 2\rho\cos\theta + \rho^2 \leq e^{-2c_\varepsilon\theta^2}(1 + \rho)^2$$

for $\varepsilon \leq |\theta| \leq \pi$. Also by comparing the first two derivatives of $1 + 2\rho\cos\theta + \rho^2$ and $e^{-2c\theta^2}(1 + \rho)^2$ at 0 (with respect to θ) we have there exists an ε', $0 < \varepsilon' \leq \pi$ and a positive $c_{\varepsilon'} \neq 0$ such that

$$1 + 2\rho\cos\theta + \rho^2 \leq e^{-2c_{\varepsilon'}\theta^2}(1 + \rho)^2$$

for $0 \leq |\theta| \leq \varepsilon'$.

Choosing $\varepsilon = \varepsilon'$ and $c = \min\{c_\varepsilon, c_{\varepsilon'}\}$ we obtain a positive c such that

$$|1 + \rho e^{i\theta}|^2 \leq e^{-2c\theta^2}(1 + \rho)^2. \qquad \langle 3.9 \rangle$$

for $-\pi \leq \theta \leq \pi$. Applying this estimate to the integrand of I_1 we obtain

$$|I_1| \leq \frac{\rho^{-(k-1)n}}{(2\pi)^{k-1}} \int\limits_{\substack{|\theta_j| \leq \delta \\ \text{some } j}} \cdots \int \left| \frac{\prod_j (1 + \rho e^{i\theta_j}) e^{-i\theta_j}}{2 - \prod_j (1 + \rho e^{i\theta_j})} \right|^n \frac{d\theta_1 \ldots d\theta_{k-1}}{\left| 2 - \prod_j (1 + \rho e^{i\theta_j}) \right|}$$

$$\leq \frac{(\rho^{-1}(1 + \rho))^{(k-1)n}}{(2\pi)^{k-1} \rho^{n+1} (1 + \rho)^{(k-1)(n+1)}} \int\limits_{\substack{|\theta_j| \leq \delta \\ \text{some } j}} \cdots \int e^{-nc(\theta_1^2 + \cdots + \theta_{k-1}^2)} \, d\theta_1 \ldots d\theta_{k-1}. \qquad \langle 3.10 \rangle$$

In the range $|\theta| \geq \delta$ the function $e^{-c\theta^2}$ has maximum value $e^{-c\delta^2}$. So

$$e^{-nc(\theta_1^2 + \cdots + \theta_{k-1}^2)} \leq e^{-nc\delta^2} = e^{-c(\ln \mathrm{n})^2}$$

for $(\theta_1, \ldots, \theta_{k-1})$ in the region of integration in $\langle 3.10 \rangle$. Hence

$$|I_1| \leq \frac{1}{(2\pi)^{k-1} \rho^{kn+1} (1 + \rho)^{k-1}} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} e^{-c(\ln \mathrm{n})^2} \, d\theta_1 \ldots d\theta_{k-1}$$

which gives

$$I_1 = O\left(\rho^{-kn} e^{-c(\ln \mathrm{n})^2}\right). \qquad \langle 3.11 \rangle$$

We now estimate the integral I_2. Note that

$$\ln\left(\frac{1 + \rho e^{i\theta}}{1 + \rho}\right) = \ln\left(1 + \frac{\rho(e^{i\theta} - 1)}{1 + \rho}\right)$$

$$= \ln\left(1 + \frac{\rho\left(i\theta - \frac{\theta^2}{2} + O(|\theta|)^3\right)}{1 + \rho}\right)$$

$$= \frac{\rho\left(i\theta - \frac{\theta^2}{2}\right)}{1 + \rho} + \frac{\rho^2 \theta^2}{2(1 + \rho)^2} + O(|\theta|^3).$$

Simplifying we obtain

$$\ln\left(\frac{1 + \rho e^{i\theta}}{1 + \rho}\right) = \frac{\rho i\theta}{1 + \rho} - \frac{\rho\theta^2}{2(1 + \rho)^2} + O(|\theta|^3). \qquad \langle 3.12 \rangle$$

Before proceeding, it will be convenient to establish some notation. For α a positive integer, define $p_\alpha(\theta)$ and $P_\alpha(\theta)$ by

$$p_\alpha(\theta) = \sum_{j=1}^{k-1} \theta_j^\alpha,$$

$$P_\alpha(\theta) = \sum_{j=1}^{k-1} |\theta_j|^\alpha.$$

Now note that

$$\prod_{j=1}^{k-1}(1+\rho e^{i\theta_j})e^{-i\theta_j}=(1+\rho)^{k-1}\prod_{j=1}^{k-1}\left(\frac{1+\rho e^{i\theta_j}}{1+\rho}\right)e^{-i\theta_j}$$
$$=(1+\rho)^{k-1}\exp\left(\sum_{j=1}^{k-1}\left\{\ln\left(\frac{1+\rho e^{i\theta_j}}{1+\rho}\right)-i\theta_j\right\}\right). \qquad \langle 3.13\rangle$$

In equation $\langle 3.13\rangle$, substitute the right hand side of $\langle 3.12\rangle$ for each $\ln((1+\rho e^{i\theta_j})/(1+\rho))$. Doing so and simplifying we obtain

$$\prod_{j=1}^{k-1}(1+\rho e^{i\theta_j})e^{-i\theta_j}=(1+\rho)^{k-1}\exp\left(\frac{-i}{1+\rho}p_1(\theta)-\frac{\rho}{2(1+\rho)^2}p_2(\theta)+O(P_3(\theta))\right). \qquad \langle 3.14\rangle$$

Also

$$2-\prod_{j=1}^{k-1}(1+\rho e^{i\theta_j})$$
$$=2-(1+\rho)^{k-1}\prod_{j=1}^{k-1}\left(\frac{1+\rho e^{i\theta_j}}{1+\rho}\right)$$
$$=2-(1+\rho)^{k-1}\exp\left(\frac{i\rho}{1+\rho}p_1(\theta)-\frac{\rho}{2(1+\rho)^2}p_2(\theta)+O(P_3(\theta))\right). \qquad \langle 3.15\rangle$$

Expanding the exponential on the right hand side of $\langle 3.15\rangle$ and simplifying we get

$$2-\prod_{j=1}^{k-1}(1+\rho e^{i\theta_j})$$
$$=(1+\rho)^{k-1}\rho\exp\left(\frac{-i}{1+\rho}p_1(\theta)+\frac{1}{2(1+\rho)^2}p_2(\theta)+\frac{2^{-(k-1)/k}}{(1+\rho)^2}p_1^2(\theta)+O(P_3(\theta))\right).$$

Hence

$$\frac{\prod_{j=1}^{k-1}(1+\rho e^{i\theta_j})e^{-i\theta_j}}{2-\prod_{j=1}^{k-1}(1+\rho e^{i\theta_j})}=\rho^{-1}\exp\left(-\frac{1}{2(1+\rho)}p_2(\theta)-\frac{2^{-(k-1)/k}}{(1+\rho)^2}p_1^2(\theta)+O(P_3(\theta))\right). \qquad \langle 3.16\rangle$$

For our purposes, the important aspect of equation $\langle 3.16\rangle$ is that in the exponential on the right hand side, the coefficient of $p_1(\theta)$ is 0. Substituting $\langle 3.16\rangle$ into the integrand of I_2 we obtain

$$I_2=\frac{\rho^{-kn}}{(2\pi)^{k-1}}\int_{-\delta}^{\delta}\cdots\int_{-\delta}^{\delta}$$
$$\frac{\exp\left(-\frac{n}{2(1+\rho)}p_2(\theta)-\frac{n_2^{-(k-1)/k}}{(1+\rho)^2}p_1^2(\theta)+O(n)P_3(\theta))\right)}{2^{(k-1)/k}\rho\exp(-O(P_1(\theta)))}d\theta_1\ldots d\theta_{k-1}$$

Let $F(\theta_1,\dots,\theta_{k-1})$ be the quadratic form

$$F(\theta_1,\dots,\theta_{k-1}) = \frac{1}{2(1+\rho)}p_2(\theta) + \frac{2^{-(k-1)/k}}{(1+\rho)^2}p_1(\theta)^2$$
$$= 2^{-(k-1)/k}(p_2(\theta) + p_1(\theta)^2).$$

In terms of this form, the last expression for I_2 can be written as

$$I_2 = \frac{\rho^{-kn-1}}{(2\pi)^{k-1}}2^{-(k-1)/k}\int_{-\delta}^{\delta}\cdots\int_{-\delta}^{\delta}\exp(-nF(\theta))\exp(O(n)P_3(\theta) + P_1(\theta))\,d\theta_1\dots d\theta_{k-1}. \tag{3.17}$$

Note that $nP_3(\theta)+P_1(\theta) = \mathcal{O}\left(\frac{ln^3 n}{n^{1/2}}\right)$ so $\exp(nP_3(\theta)+P_1(\theta))=1+O(nP_3(\theta)+P_1(\theta))$. Hence $I_2 = S_1 + S_2 + S_3$ where

$$S_1 = \frac{\rho^{-kn-1}}{(2\pi)^{k-1}}2^{-(k-1)/k}\int_{-\delta}^{\delta}\cdots\int_{-\delta}^{\delta}\exp(-nF(\theta))\,d\theta_1\dots d\theta_{k-1},$$

$$S_2 = O\left(\rho^{-kn}n\int_{-\delta}^{\delta}\cdots\int_{-\delta}^{\delta}(P_3(\theta))\exp(-nF(\theta))\,d\theta_1\dots d\theta_{k-1}\right),$$

and

$$S_3 = O\left(\rho^{-kn}\int_{-\delta}^{\delta}\cdots\int_{-\delta}^{\delta}P_1(\theta)\exp(-nF(\theta))\,d\theta_1\dots d\theta_{k-1}\right).$$

Note that

$$\int_{-\delta}^{\delta}\cdots\int_{-\delta}^{\delta}\exp(-nF(\theta))\,d\theta_1\dots d\theta_{k-1}$$
$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-nF(\theta))\,d\theta_1\dots d\theta_{k-1} + O(e^{-ln^2 n}).$$

Substituting $u_j = n^{1/2}\theta_j$ so $\frac{du_j}{n^{1/2}} = d\theta_j$ we have

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-nF(\theta))\,d\theta_1\dots d\theta_{k-1}$$
$$= n^{-(k-1)/2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp(-F(u))\,du_1\dots du_{k-1}.$$

It is well-known that any positive definite quadratic form $E(u_1,\dots,u_{k-1})$ is congruent over $\mathbf{R}$ to the sum of u_i^2. Using this fact it is easy to see that the integral of $\exp(-E(u))$ over $\mathbf{R}^{k-1}$ is $\pi^{(k-1)/2}|\det(C_E)|^{-1/2}$ where C_E is the matrix of the quadratic form E. In our case

$$C_F = \left(\frac{1}{2(1+\rho)}\right)I + \frac{2^{-(k-1)/k}}{(1+\rho)^2}J,$$

where J is the matrix of all 1's. Recall that the $(k-1)$ by $(k-1)$ matrix $XI + YJ$ has determinant $X^{k-2}(X + (k-1)Y)$. So

$$\det(C_F) = k2^{-(k^2-1)/k}.$$

Thus

$$S_1 = \frac{k^{-1/2}\rho^{-kn-1}2^{-(k^2-1)/(2k)}}{\pi^{(k-1)/2}n^{(k-1)/2}} + O(\rho^{-kn}e^{-c(\ln n)^2}). \qquad \langle 3.18\rangle$$

We now obtain bounds for S_2 and S_3. Observe that

$$\exp(-nF(\theta)) \le \exp\left(-\frac{n}{2(1+\rho)}\right)p_2(\theta).$$

So

$$S_2 = O\left(n\rho^{-kn}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}P_3(\theta)\exp\left(-\frac{n}{2(1+\rho)}\right)p_2(\theta)\,d\theta_1\ldots d\theta_{k-1}\right). \quad \langle 3.19\rangle$$

By symmetry of $P_3(\theta)$ in the variables $\theta_1, \ldots, \theta_{k-1}$ we can rewrite $\langle 3.19\rangle$ as

$$S_2 = O\left((k-1)n\rho^{-kn}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}|\theta_1|^3\exp\left(-\frac{n}{2(1+\rho)}\right)p_2(\theta)\,d\theta_1\ldots d\theta_{k-1}\right). \qquad \langle 3.20\rangle$$

Recall that $\int_{-\infty}^{\infty} e^{-L\theta^2}\,d\theta = \frac{\sqrt{\pi}}{\sqrt{L}}$. We use this to integrate in $\langle 3.20\rangle$ with respect to $\theta_2, \ldots, \theta_{k-1}$. Doing so we obtain

$$S_2 = O\left(n^{-(k-4)/2}\rho^{-kn}\int_0^{\infty}\theta_1^3\exp\left(\frac{-n}{2(1+\rho)}\theta_1^2\right)d\theta_1\right). \qquad \langle 3.21\rangle$$

Integration by parts gives

$$\int_0^{\infty}\theta^3\exp\left(-\frac{n}{2(1+\rho)}\theta^2\right)d\theta = O\left(\frac{1}{n^2}\right).$$

Hence

$$S_2 = O(\rho^{-kn}n^{-k/2}). \qquad \langle 3.22\rangle$$

The derivation of a bound on S_3 is similar. We have

$$\begin{aligned}
S_3 &= O\left(\rho^{-kn}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}P_1(\theta)\exp\left(-\frac{n}{2(1+\rho)}p_2(\theta)\right)d\theta_1\ldots d\theta_{k-1}\right),\\
&= O\left(\rho^{-kn}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}|\theta_1|\exp\left(-\frac{n}{2(1+\rho)}\right)p_2(\theta)\,d\theta_1\ldots d\theta_{k-1}\right),\\
&= O\left(\rho^{-kn}n^{(k-2)/2}\int_0^{\infty}\theta_1\exp\left(-\frac{n}{2(1+\rho)}\theta_1^2\right)d\theta_1\right).
\end{aligned}$$

It is easy to check that $\int_0^\infty \theta_1 \exp\left(-\frac{n}{2(1+\rho)}\theta_1\right) d\theta_1 = O\left(\frac{1}{n}\right)$. Hence

$$S_3 = O(\rho^{-kn} n^{-k/2}). \qquad \langle 3.23 \rangle$$

Combining the estimates ⟨3.11⟩, ⟨3.18⟩, ⟨3.22⟩ and ⟨3.23⟩ we obtain

$$f(k,n) = S_1 + (I_1 + S_2 + S_3) = \left(\frac{\rho^{-kn}}{n^{(k-1)/2}}\right)\left(\left(\rho\pi^{(k-1)/2} k^{1/2}\right)^{-1} 2^{-(k^2-1)/2k} + O(n^{-1/2})\right)$$

which was the statement of the main theorem of this section.

Acknowledgement. We thank Kevin McCurley for his assistance with computer simulations during the early stages of this work.

References

1. Griggs, J.R., Hanlon, P., Waterman, M.S.: Sequence alignments with matched sections. SIAM J. Alg. Disc. Meth. **7**, No. 4, 604–608 (1986)
2. Laquer, H.T.: Asymptotic limits for a two-dimensional recursion. Stud. Appl. Math. **64**, 271–277 (1981)
3. Stanton, R.G., Cowan, D.D.: Note on a square functional equation. SIAM Rev. **12**, 277–279 (1970)
4. Ullrich, A., Bell, J.R., Chen, E.Y., Herrera, R., Petruzzelli, L.M., Dull, T.J., Gray, A., Coussens, L., Liao, Y.C., Tsubokawa, M., Mason, A., Seeburg, P.H., Grunfeld, C., Rosen O.M., Ramachandran, J.: Human insulin receptor and its relationship to the tyrosine kinase family of oncogenes. Nature **313**, 756–761 (1985)
5. Waterman, M.S.: General methods of sequence comparison. Bulletin of Math. Bio. **46**, 473–500 (1984)

Received: July 27, 1986