Statistical analysis of nucleotide sequences

Erika E.Stückle[1,2], Claudio Emmrich[2], Ulrich Grob[1] and Peter J.Nielsen[1,*]
Max-Planck-Institut für Immunbiologie, Stübeweg 51, D-7800 Freiburg and Albert-Ludwigs-Universität Freiburg, Fakultät für Physik, Hermann-Herder-Straße 3, D-7800 Freiburg, FRG

Received July 4, 1990; Revised and Accepted September 27, 1990

ABSTRACT

In order to scan nucleic acid databases for potentially relevant but as yet unknown signals, we have developed an improved statistical model for pattern analysis of nucleic acid sequences by modifying previous methods based on Markov chains. We demonstrate the importance of selecting the appropriate parameters in order for the method to function at all. The model allows the simultaneous analysis of *several* short sequences with *unequal* base frequencies and Markov order $k \neq 0$ as is usually the case in databases. As a test of these modifications, we show that in E.coli sequences there is a bias against palindromic hexamers which correspond to known restriction enzyme recognition sites.

INTRODUCTION

The rapid growth of DNA-sequence databases seen in the last few years (1,2,3,4,5,6) makes it increasingly possible to identify biologically interesting sequence motifs by statistical methods. In 1982, Dumas and Ninio (7) and Stormo, Schneider and Gold (8) introduced the concept of 'words' or k-tuples for sequence analysis in molecular biology. Thus, a nucleic acid sequence can be viewed as consisting of overlapping 'words' instead of one continuous stretch of information. Soon it was recognized that not only the four monomers (A,C,G,T) but also higher oligomers are nonrandomly distributed (9,10,11,12).

Two general ideas emerged from these observations. First, relatively large functional domains may show distinct oligomer distributions and thereby can be distinguished from each other. This was shown to be true by Smith et al. (13) who found that exons and introns can be separated on the basis of their dinucleotide frequencies. Second, there may be a bias against or towards sequence motifs which are used as regulatory or recognition signals. Indeed, it was shown that highly recurring oligomers in eukaryotic DNA often correspond to regulatory sequences or protein binding sites (14). Dinucleotide distribution has also been used to predict the frequencies of restriction endonuclease recognition sequences (15).

Meanwhile there are several investigations published which have used and expanded the concept of k-tuples (16,17,18). In a next step, Markov chain methods were used to address statistical analysis of biological sequences(19,20,23). These studies showed that the probability of finding a particular base at one position can depend not only on the immediately adjacent bases but also on several more distant bases upstream or downstream. It was also shown that within a sequence this dependency can vary. These methods could not be used to predict expected oligonucleotide frequencies. Recently more rigorous statistical models based on Markov chains have been developed (21,22,24,25,26) which address this problem. There the expected number of occurences of each oligomer of length L can be calculated from observed frequencies of oligomers of length less than L.

However, as we show in this paper, the algorithms for calculating the expected frequencies of oligomers given in (24) and (26) are not applied appropriately. In addition, we improve the statistical method of Pevzner et al. (26) to permit the simultaneous analysis of **several** linear sequences with **unequal** base frequencies and Markov order $\mathbf{k \neq 0}$. These are conditions of real databases.

To test this modified statistical model, we examined all E.coli sequences stored in the GenBankTM database for the occurrence of palindromic hexamers. In general agreement with previous results (21,22), we find that palindromic hexamers are less frequent than expected. In addition, those hexamers which correspond to restriction enzyme recognition sites are often those hexamers which deviate most strongly from the expected frequencies.

MATERIAL

The sequences were obtained from the GenBank™ database, version 60 and computed on a microVAX II from Digital Equipment Corporation. For the investigation of the hexameric palindroms we used 797 E.coli sequences with a total length of about 1.2×10^6 bases. Some programs of the GCG (Genetics Computer Group) program package have been used to access the sequence data (27).

METHOD AND RESULTS

We define the sequence S by:

$$S = S_1S_2S_3\ldots S_n;\ S_i \in \{A,C,G,T\},\ i=1,\ldots n.$$

$s = s_1s_2s_3\ldots s_L$; $s_i \in \{A,C,G,T\}$, $i=1,\ldots L$ describes an oligomer of length L.

In this paper we treat nucleic acid sequences as Markov chains which implies that the chain has a finite 'memory'. A stationary

[*] To whom correspondence should be addressed

Markov chain with Markov order (Mo) k is defined such that the conditional probabilities satisfy

$$p(s_N|s_{N-1}\ldots s_1) = p(s_N|s_{N-1}\ldots s_{N-k}).$$

Border effect

For the statistical analysis of biological sequences stored in computer databases we were concerned that the ends of the sequences might not be of random composition. To test this we choose the set of all human immunoglobulin gene sequences stored in the database. These 207 entries include both germ line and rearranged sequences. We counted the oligomers of length L=1 to L=5 by one base shift (overlapping oligomers). The oligomers derived from the ends of these sequences showed an unusual distribution. For the 5′ end of the sequences:

84	began	with	G
38	,,	,,	GA
11	,,	,,	GAT
11	,,	,,	GATC
7	,,	,,	GATCA.

Assuming that the oligomers are Poisson distributed, the probability that a specific oligomer of length L=5 occurs seven times at a given site is $p=2.29\times10^{-9}$. Therefore, it is unlikely that the above listed oligomers occurred by chance. Accumulation of these oligomers at the endpoints of the sequences can be explained by cloning effects. Cutting DNA with Bgl II, Bam HI, Xho II, Sau 3AI and several other restriction enzymes gives the 5′ start oligomer 'GATC'. Cutting DNA with Bcl I produces the 5′ oligomer 'GATCA'. We have also observed this effect in sequences which have nothing to do with immunoglobulins (see the palindromic hexamer analysis below).

After removing the first and last eight bases (most restriction enzymes recognize four or six base sites) of each sequence, the above effect could no longer be seen.

As long as one is not interested in recognition oligomers for restriction enzymes, it is certainly better for sequence analysis by any statistical model to remove the first and last eight bases of each sequence.

Choosing the correct parameters

Given the observed frequencies f of oligomers of length L−2 and L−1 in a sequence S, the expected frequencies E(s) of oligomers s of length L are calculated in (24) by:

$$E(s_1s_2s_3\ldots s_L) = \frac{f(s_1s_2\ldots s_{L-1})f(s_2s_3\ldots s_L)}{f(s_2s_3\ldots s_{L-1})}. \qquad [1]$$

This formula is based on a Markov chain of the order k=L−2. This is not suitable as will be seen later.

The oligomer $s = (s_1s_2\ldots s_L)$ is assumed to behave statistically (that means its frequency of appearance is that which is expected from a random distribution), if $|std(s)| < \alpha$ where

$$std(s_1s_2\ldots s_L) = \frac{f(s_1s_2\ldots s_L)-E(s_1s_2\ldots s_L)}{(E(s_1s_2\ldots s_L))^{1/2}}. \qquad [2]$$

std is called the normalized deviation (it gives the number of standard deviations the observed frequency differs from the expected frequency). α is the threshold value which depends on the length of oligomers and the length of the sequences. The occurrences of oligomers s with $|std(s)| \geq \alpha$ are considered statistically significant. Such oligomers are potential candidates for biologically meaningful motifs.

The published models (24,26) and the one described here are based on a Gaussian distribution and thus are applicable only if the expected frequencies of oligomers are considerably greater than one. Furthermore, since the possible number of different oligomers of a given length increases exponentially with the length, a greater threshold value a must be chosen for L>4. For example, from the 16384 different possible heptamers, about 50 are expected not to fulfil $|std(s)| < \alpha$ by chance for $\alpha=3$ (at Markov order k=0).

From the intergral of the Gaussian distribution, it can be calculated that no heptamers are expected to show $|std(s)| \geq 5$ by chance. The advantage of choosing $\alpha=5$ is that statistical oligomers are eliminated with the disadvantage that some signals could be lost. For oligomers with lengths between 7 and 10 inclusive, $\alpha=5$ should be used. For penta- and hexamers a similar stringency requires that $\alpha=4$ and for tetramers $\alpha=3$. It is common to choose the Markov order as k=L−2. This is acceptable for L<4 but when L>4, this is not appropriate in algorithm [1]. This is because, for example, when a heptamer (L=7) occurs very often, the corresponding hexamers (L=6) and pentamers (L=5) as suboligomers of this heptamer also occur often, even if they rarely occur outside of the heptamer. Therefore, the frequency E(s) is overestimated and consequently the normalized deviation std(s) is too small. Such a heptamer may show no significant deviation from statistical behaviour even if it is not statistical.

To show the importance of choosing proper a and k values we simulated a random sequence S with Markov order (Mo) k=0 consisting of n=179951 bases. By using $\alpha=5$, k=0 we calculated that a heptamer must occur about twenty times to give std(s)>5. Consequently we modified S by inserting four test heptamers each twenty times. This was done by randomly replacing different statistical heptamers with the test oligomers in order not to change the length of the sequence.

We inserted the following heptamers:

AGCCATC
ATGACGC
GTCATTG
TGACATG

and analyzed this modified sequence S′ with the algorithm [1] and in addition with Mo k=0 to k=4.
The generalized formula is:

$$E(s_1s_2\ldots s_L) = \frac{f(s_1\ldots s_{k+1})f(s_2\ldots s_{k+2})\ldots f(s_{L-k}\ldots s_L)}{f(s_2\ldots s_{k+1})\ldots f(s_{L-k}\ldots s_{L-1})}. \qquad [3]$$

Using this modification with k=0 we obtained, as expected, 51 heptamers with $|std(s)| \geq 3$. This shows that $\alpha=3$ is not stringent enough to eliminate statistical oligomers from the set of oligomers with nonrandom frequencies:

for the inserted heptamers:

	k=0	*k=1*	*k=2*	*k=3*	*k=4*	*k=5*
std(AGCCATC) =	5.28	5.21	5.20	4.99	3.96	1.84
std(ATGACGC) =	6.66	6.50	6.02	5.92	5.16	3.10
std(GTCATTG) =	6.41	6.13	5.97	5.75	4.46	2.42
std(TGACATG) =	6.11	5.77	5.33	4.97	3.47	2.50

for several statistical heptamers arbitrarily chosen from the 51 heptamers with $|std(s)| > 3$ for MO $k=0$:

std(AATTAAT) =	3.32	3.41	3.64	3.38	2.66	1.13
std(AGATTCC) =	3.21	3.23	3.33	2.94	3.03	1.02
std(CGCACGT) =	4.46	4.29	4.23	4.03	3.17	1.76
std(GTCGGCA) =	3.16	3.05	2.87	3.16	2.15	0.89
std(TGACACA) =	3.21	2.94	2.89	2.56	1.01	1.05

For Markov order $k=5$ (as used in the model of (24)) all heptamers behave statistically ($|std(s)| < 5$) even those which are not statistical by construction. Only for Mo $k < 3$ can a significant deviation from statistical behaviour be seen for all four inserted heptamers and not for the statistical heptamers.

This effect could be avoided if one could determine the true statistical frequencies f_s of suboligomers for each of the potential signal oligomers (which is possible only when one already knows the signals). From f_s one could then calculate the expected frequency E(s) at Mo $k=L-2$ (which corresponds to a longer correlation between bases).

For the simulated case we computed the frequencies f_s of the statistical suboligomers of sequence S. With these values we calculated for all Markov orders the expected frequencies of oligomers s of the modified sequence S′. The normalized deviation std is calculated as before in [2] where the expected frequency E is given by [3] with f replaced by f_s.

For heptamers calculated with this method (second line, the first line is calculated as before) we obtained:

$k=0$ $k=1$ $k=2$ $k=3$ $k=4$ $k=5$

for the inserted heptamers:

std(AGCCATC) =	5.28	5.21	5.20	4.99	3.96	1.84
	5.29	5.24	5.37	5.45	5.66	5.75
std(ATGACGC) =	6.66	6.50	6.02	5.92	5.16	3.10
	6.66	6.56	6.26	6.61	6.92	7.75
std(GTCATTG) =	6.41	6.13	5.97	5.75	4.46	2.42
	6.41	6.19	6.17	6.27	6.26	6.61
std(TGACATG) =	6.11	5.77	5.33	4.97	3.47	2.50
	6.11	5.89	5.64	5.59	5.02	6.63

for the same randomly chosen statistical heptamers shown above:

std(AATTAAT) =	3.32	3.41	3.64	3.38	2.66	1.13
	3.32	3.39	3.61	3.31	2.68	1.19
std(AGATTCC) =	3.21	3.23	3.33	2.94	3.03	1.02
	3.22	3.23	3.28	2.89	3.05	1.02
std(CGCACGT) =	4.46	4.29	4.23	4.03	3.17	1.76
	4.46	4.31	4.23	3.99	3.23	1.83
std(GTCGGCA) =	3.16	3.05	2.87	3.16	2.15	0.89
	3.16	3.06	2.88	3.17	2.10	0.91
std(TGACACA) =	3.21	2.94	2.89	2.56	1.01	1.05
	3.21	3.01	3.01	2.83	1.33	1.05

Obviously, now for all Mo $k=0$ to $k=5$, the inserted heptamers do not behave statistically because $|std(s)| > 5$ in all orders. As expected for the statistical heptamers, the values are essentially the same.

These examples show that only in ideal cases where one already knows the nonstatistical signals is the statistical analysis of sequences relatively independent of k for $0 \leq k \leq L\text{-}2$. This is because one can calculate the expected suboligomer frequencies from reference sequences where the signal oligomer is not a signal.

This method is not ideal for real sequences because the signals are unknown and it is not clear which sequences should be used for calculating the suboligomer frequencies. Therefore, it is necessary to turn to smaller Markov orders for calculating the expected frequencies E(s).

Like Brendel et al.(24) we also analyzed hexamer frequencies for phage T7 in order to show the effect of improper k and a values. In (24) they used Mo $k=4$, $\alpha=3$, we use the more stringent conditions of Mo $k=2$, $\alpha=4$. The results of both methods are completely different (data not shown).

The improved statistical method

Taking into account the importance of k and a as shown above, we then extended the previously published model of Pevzner et al.(26) to the calculation of the variance of oligomer frequencies in **several** linear sequences with **unequally** distributed bases and Mo $\mathbf{k \neq 0}$. In Pevzner et al.(26) a formula for *one* linear sequence with *e*qually distributed bases and Mo $k=0$ is given. In real databases there are many short sequences and therefore the method has to be extended to apply it for real databases.

Previously (26), the normalized deviation std was computed by:

$$std(s_1s_2\ldots s_L) = \frac{f(s_1s_2\ldots s_L) - E(s_1s_2\ldots s_L)}{(V(s_1s_2\ldots s_L))^{1/2}} \qquad [4]$$

with E(s) from Brendel et al. (24).

The difference between the two previous models (24) and (26) is, that oligomer frequencies in the model given by the latter are not assumed to be Poisson distributed. As the oligomers overlap, they are not independent and therefore this is certainly a better analysis than the earlier model (24).

In the following, the formula for the variance V is taken from Pevzner et al.(26), corrected for several typographical errors, and extended.

In (26) an autocorrelation polynomial

$$K_s(x) = \sum_{r=0}^{L-1} k_r x^r$$

with the coefficient

$$k_r = \begin{cases} 1, \text{ the first and last } L-r \text{ bases agree} \\ 0, \text{ otherwise} \end{cases}$$

was introduced. The frequency of an oligomer s is counted by the random variable X:

$$X = \sum_{i=1}^{n} x_i, \quad \text{with } x_i = \begin{cases} 1, \text{ oligomer is beginning at position i} \\ 0, \text{ otherwise} \end{cases}$$

One circular sequence with unequal base distribution and Mo $\mathbf{k \neq 0}$

First we consider a circular sequence. The mathematical expectation and variance are calculated by:

$$E(x_i) = p(x_i=1) = p_s;$$

$$EX = E\sum_{i=1}^{n} x_i = np_s \qquad [5]$$

$$V(x_i) = E(x_i^2) - E^2(x_i) = p_s(1-p_s)$$

$$VX = EX^2 - E^2X = \sum_{i,j=1}^{n} (E(x_ix_j) - E(x_i)E(x_j)).$$

where for unequally distributed bases and Mo $k \neq 0$ the probability p_s is given by:

$$p_s = p(s_1 s_2 \ldots s_k) \prod_{\beta=1}^{L-k} p(s_{\beta+k} | s_{\beta+k-1} \ldots s_\beta),$$

where $p(u_n | u_{n-1}, u_{n-2} \ldots u_1)$ is the probability of state u_n given $u_{n-1} \ldots u_1$.

Denoting by d(i,j) the shortest distance between position i and j, the variance can be devided into the following three terms:

$$VX = \sum_{\substack{i,j=1 \\ d(i,j)>L}}^{n} (E(x_i x_j) - E(x_i)E(x_j)) + \qquad (a)$$

$$\sum_{\substack{i,j=1 \\ d(i,j)=0}}^{n} (E(x_i x_j) - E(x_i)E(x_j)) + \qquad (b)$$

$$\sum_{\substack{i,j=1 \\ 0<d(i,j)\leq L}}^{n} (Ex_i x_j) - E(x_i)E(x_j)). \qquad (c)$$

For Mo $k \neq 0$, oligomers which do not overlap can be correlated and therefore it is not apparent that term (a) may be neglected. We suggest later that term (a) can be neglected. Contrary to Pevzner et al. (26), we have included the case where d(i,j) = L in term (c).

Term (b) yields:

$$\sum_{i=1}^{n} V(x_i) = np_s(1-p_s).$$

As

$$E(x_i x_j) = \begin{cases} p_s \prod_{\beta=1}^{r} p(s_{\beta+k} | s_{\beta+k-1} \ldots s_\beta); & k_r = 1 \\ 0; \text{ otherwise} & \end{cases}$$

we obtain for term (c):

$$\sum_{\substack{i,j=1 \\ 0<d(i,j)\leq L}}^{n} (E(x_i x_j) - E(x_i)E(x_j)) = np_s(2\tilde{K}_s - 2 - 2Lp_s) \qquad [6]$$

with the modified autocorrelation polynomial:

$$\tilde{K}_s = 1 + \sum_{r=1}^{L} k_r \prod_{b=1}^{r} p(s_{\beta+k} | s_{\beta+k-1} \ldots s_\beta),$$

where for all s indices $\gamma > L$, γ has to be replaced by $\gamma - r$. The variance is:

$$\mathbf{VX = np_s(2\tilde{K}_s - 1 - (2L+1)p_s).} \qquad [7]$$

According to this model we calculated the tetramer std values of phage T7 and compared this to the values obtained previously (24,26)(Table 1). The effect of considering d(i,j) = L on std is minimal. Since the contribution from term (a) is expected to be at most of the same order of magnitude, it seems plausible that term (a) may be neglected.

In the databases, many primarily short sequences with unequally distributed bases and Mo $k \neq 0$ are stored. The model above has to be extended to accomodate this.

Table 1. Tetramers of phage T7 with |std(s)| > 3, by model of (24)(std-(24)), model of (26) (std-(26)) and model of (24) with extension d(i,j) = L (std):

	std-(24)	std-(26)	std
GGTT	−3.2302	−3.2918	−3.2833
GAGC	−3.3968	−3.4467	−3.4499
GATG	4.6592	4.6421	4.6386
GATT	3.0101	3.0504	3.0481
GATC	−11.3559	−11.4982	−11.5082
GTTC	4.2241	4.2954	4.2960
GCTG	5.0594	5.0503	5.0429
GCTT	−4.0502	−4.1236	−4.1217
AGCG	3.0769	3.1095	3.1087
AGCT	−5.4575	−5.5663	−5.5821
AAAA	−3.2854	−2.7182	−2.7208
AATT	−3.8376	−3.8789	−3.8764
AATC	5.3794	5.4336	5.4294
TGAA	−3.6421	−3.7294	−3.7340
TAGC	3.0981	3.1172	3.1138
TATC	3.1762	3.2119	3.2104
TACG	3.0363	3.0766	3.0727
TTGA	3.1676	3.2241	3.2257
TTTC	−3.2535	−3.2803	−3.2787
CGGT	3.2634	3.2985	3.3003
CGTC	−3.2858	−3.2807	−3.2789
CGCG	−3.8523	−3.7018	−3.7019
CGCT	3.9327	4.0009	4.0038
CAGG	−3.0007	−3.0374	−3.0391
CAAT	3.0169	3.0533	3.0527
CATA	−3.2940	−3.3224	−3.3212
CATC	3.9302	3.9153	3.9143
CTAG	−3.9564	−3.9912	−3.9899
CTTT	4.5553	4.5947	4.5930
CCTG	−4.5076	−4.5770	−4.5815
CCTT	3.8355	3.8793	3.8730
CCCT	3.4521	3.4768	3.4760

One linear sequence with unequal base distribution and Mo $k \neq 0$

For a linear sequence with unequal base distribution, Mo $k = \neq 0$, d(i,j) = L considered in Term (c'), the variance is:

$$VX = \sum_{\substack{i,j=1 \\ d(i,j)>L}}^{n-L+1} (E(x_i x_j) - E(x_i)E(x_j)) + \qquad (a')$$

$$\sum_{\substack{i,j=1 \\ d(i,j)=0}}^{n-L+1} (E(x_i x_j) - E(x_i)E(x_j)) + \qquad (b')$$

$$\sum_{\substack{i,j=1 \\ 0<d(i,j)\leq L}}^{n-L+1} (E(x_i x_j) - E(x_i)E(x_j)). \qquad (c')$$

Term (a') is neglected for the same reason as before.

Term (b') yields:

$$\sum_{\substack{i,j=0 \\ d(i,j)=0}}^{n-L+1} (E(x_i x_j) - E(x_i)E(x_j)) =$$

$$\sum_{i=1}^{n-L+1} V(x_i) = (n-L+1)p_s(1-p_s),$$

For term (c') we obtain:

$$\sum_{\substack{i,j=1 \\ 0<d(i,j)\leq L}}^{n-L+1} (E(x_i x_j) - E(x_i)E(x_j)) =$$

$$\sum_{i=1}^{n-L+1} \sum_{r=1}^{L} \sum_{\substack{j=1 \\ d(i,j)=r}}^{n-L+1} (E(x_i x_j) - E(x_i)E(x_j)) =$$

$$\sum_{\substack{i=1 \\ r<\min\{i,n-i-L+2\}}}^{n-L+1} \sum_{r=1}^{L} 2(E(x_ix_j)-E(x_i)E(x_j)) \; +$$

$$\sum_{\substack{i=1 \\ r\geq\min\{i,n-i-L+2\}}}^{n-L+1} \sum_{,\, r\neq 1}^{L} (E(x_ix_j)-E(x_i)E(x_j)) \; =$$

$$(n-L+1)p_s(2\tilde{K}_s-2-2Lp_s)-\sum_{i=1}^{n-L+1} \sum_{\substack{r=1 \\ r>\min\{i,n-i-L+2\}}}^{L} (E(x_ix_j)-E(x_i)E(x_j)) \; =$$

$$(n-L+1)p_s(2\tilde{K}_s-2-2Lp_s)-$$

$$p_s\left(2\sum_{r=1}^{L} rk_r \prod_{\beta=1}^{r} p(s_{\beta+k}|s_{\beta+k-1}\ldots s_\beta)-L(L+1)p_s\right),$$

by using equation [6].

For one linear sequence of length n_i the expected frequency and the variance is:

$$EX_i = \sum_{i=1}^{n_i-L+1} E(x_i) = (n_i-L+1)p_s \qquad [8]$$

$$VX_i = (n_i-L+1)p_s(2\tilde{K}_s-1-(2L+1)p_s) \; -$$

$$2p_s\sum_{r=1}^{L} rk_r \prod_{\beta=1}^{r} p(s_{\beta+k}|s_{\beta+k-1}\ldots s_\beta) \; + \; L(L+1)p_s^2. \qquad [9]$$

Several linear sequences with unequal base distribution and Mo $k\neq 0$

As the sequences are independent, we obtain for the expected frequency and the variance of N linear sequences with total length m:

$$EX = \sum_{i=1}^{N} EX_i = \sum_{i=1}^{N} (n_i-L+1)p_s = (m-N(L-1))p_s, \quad [10]$$

$$VX = (m-N(L-1))p_s(2\tilde{K}_s-1-(2L+1)p_s)-$$

$$2Np_s\sum_{r=1}^{L} rk_r \prod_{b=1}^{r} p(s_{\beta+k}|s_{\beta+k-1}\ldots s_\beta) \; + \; NL(L+1)p_s^2. \qquad [11]$$

Verification of the method

With this statistical model we examined the 797 E.coli (about 1.2×10^6bp) sequences present in the GenBank™ database (release 60) with respect to the occurrence of all possible hexameric palindroms. The result using Mo k=2 and $\alpha=4$ are shown in Table 2. As also seen previously (22), there is a general, obvious underrepresentation of hexameric palindroms in E.coli sequences. One important reason for this could be that many such hexamers are recognition signals for restriction endonucleases. This would predict that palindromic hexamers not recognized by any known restriction enzymes should behave statistically. On the average this is indeed the case (see Table 3) where the average score of std(s) for nonrestriction site hexamers is −2.12 (well within the $-4\leq\alpha\leq4$ boundary chosen) and for restriction site hexamers the average score of std(s) = −5.26. The underrepresentation of restriction site hexamers in E.coli sequences is even more striking if only those hexamers are examined for which a corresponding enzyme in E.coli has been identified. Then the average score of std(s) = −8.68 (Table 3). For the results shown we did not remove the first and last eight nucleotides of each sequence (see 'Border effect' described above)

Table 2. Recognition hexamers for restriction enzymes and palindromic hexamers not recognized by any known restriction enzyme

std(GAATTC)	=	−5.14*	EcoR I
std(CACGTG)	=	−12.63	PmaC I
std(GACGTC)	=	−0.05	Aha II, Aat II
std(TACGTA)	=	−1.51*	SnaB I
std(AAGCTT)	=	−6.44*	Hind III
std(CAGCTG)	=	−10.15	Pvu II, NspB II
std(GAGCTC)	=	−8.05*	Sac I, Ban II, HgiA I, Bsp 1286
std(AATATT)	=	8.28	Ssp I
std(CATATG)	=	−4.33	Nde I
std(GATATC)	=	9.33*	EcoR V
std(ACATGT)	=	−0.47	Afl III, Nsp 7524 I
std(CCATGG)	=	−6.56	Nco I, Sty I
std(GCATGC)	=	−12.70	Sph I, Nsp 7524 I
std(TCATGA)	=	−2.64	BspH I
std(CCCGGG)	=	−0.52	Xma I, Ava I, Sma I
std(GCCGGC)	=	−19.71*	Nae I
std(TCCGGA)	=	1.61	BspM II
std(ACGCGT)	=	−3.49	Mlu I, Afl III
std(CCGCGG)	=	−14.55*	NspB II, Sac II, Ksp I
std(GCGCGC)	=	−6.69	BssH II
std(TCGCGA)	=	−0.71	Nru I
std(ACTAGT)	=	0.62	Spe I
std(CCTAGG)	=	−4.51	Avr II, Sty I
std(GCTAGC)	=	−6.24	Nhe I
std(TCTAGA)	=	−5.46	Xba I
std(AGATCT)	=	−4.26	Bgl II, BstY I
std(CGATCG)	=	−5.07	Pvu I
std(GGATCC)	=	−6.47	BamH I, BstY I
std(TGATCA)	=	−5.16	Bcl I
std(AGCGCT)	=	−8.08*	Eco 47 III, Hae II
std(GGCGCC)	=	−19.08*	Ban I, Nar I,Aha II, Bbe I, Hae II
std(TGCGCA)	=	−3.24	Fsp I
std(AGGCCT)	=	−1.96*	Stu I
std(CGGCCG)	=	−18.48*	Eag I, Eae I, Gdi II
std(GGGCCC)	=	−10.25	Apa I, Ban II, Bsp 1286
std(TGGCCA)	=	−13.11	Eae I, Bal I
std(AGTACT)	=	−3.16	Sca I
std(GGTACC)	=	−8.44*	Asp 718, Ban I, Kpn I
std(GTATAC)	=	−7.27	Acc I, Xca I
std(ATCGAT)	=	−1.05	Cla I
std(CTCGAG)	=	−4.22	Xho I, Ava I
std(GTCGAC)	=	−2.89	Sal I, Acc I, Hinc II
std(TTCGAA)	=	−5.04	BstB I
std(ATGCAT)	=	−8.12*	Nsi I
std(CTGCAG)	=	−11.26*	Pst I
std(GTGCAC)	=	−4.75	ApaL I, Sno I, Bsp 1286, HgiA I
std(ATTAAT)	=	6.41	Ase I
std(CTTAAG)	=	−3.60	Afl II
std(GTTAAC)	=	4.04	Hpa I, Hinc II
std(TTTAAA)	=	−5.99	Dra I, Aha III
std(AAATTT)	=	−4.33	
std(CAATTG)	=	−1.66	
std(TAATTA)	=	1.46	
std(AACGTT)	=	−0.36	
std(TAGCTA)	=	2.58	
std(TATATA)	=	−2.91	
std(ACCGGT)	=	3.45	
std(CGCGCG)	=	−8.65	
std(CGTACG)	=	−7.19	
std(TGTACA)	=	−0.72	
std(ATATAT)	=	−2.86	
std(CTATAG)	=	0.31	
std(TTATAA)	=	−5.04	
std(TTGCAA)	=	−3.70	

*Recognition hexamers for restriction enzymes isolated from E.coli (28).

Table 3

		Frequency			
	Number of sites	rare (std < −4)	random (−4 < std < 4)	frequent (std > 4)	average score
all hexameric palindroms	64	35	25	4	−4.58
restriction site hexamers	50	31	15	4	−5.26
nonrestriction site hexamers	14	4	10	0	−2.12
sites recognized by E.coli restriction enzymes	14	11	2	1	−8.68

since this could lead to an underrepresentation of some restriction enzyme sites. In fact, analysis of E.coli sequences with and without these end nucleotides had a considerable effect only on hexamers corresponding to those restriction enzymes often used for cloning (e.g. EcoR I, Hind III, BamH I). We assume these end hexamers represent natural restriction sites rather than being the result of synthetic linker addition during cloning.

It should be kept in mind that $|\mathrm{std}(s)| < \alpha$ does not imply that the oligomer s is not biologically meaningful. The frequency with which an oligomer appears is determined by many biological necessities. For example, superimposed on the general reduced frequency of hexameric restriction site palindroms described above is a strong bias against those hexamers with a high G/C content. Since G/C rich sequences would be expected to form more stable duplexes, the paucity of G/C rich hexamers may reflect the need for DNA strand separation in such processes as chromosomal replication and RNA transcription. Oligomer biases correllating with codon usage have also been reported (22).

DISCUSSION

In this paper we derived a rigorous statistical model for analysis of nucleic acid sequences. Our method represents an improvement of previously published algorithms (24,25,26). We are able to analyze databases of many linear sequences with unequal base frequencies and Markov order $k \neq 0$. These are necessary prerequisites to analyze existing sequence databases.

We also demonstrate by using simulated sequences that choosing appropriate values for Markov order k and threshold value α is essential to obtain correct results. As the bases are correlated, it is surely not optimal to take Markov order $k=0$. On the other hand, we showed that the maximal possible value $k=L-2$ for the Markov order is not appropriate either. For all oligomers with length $L>4$ it is the best compromise to choose Markov order $k=2$. In this way, correlation between bases is considered without being so stringent that, as shown in the simulation, significant oligomers remain undetected. When $L \leq 4$, then $k=L-2$ can be used. A completely different algorithm would be required for Markov order $k=L-2$ for all L. However it is not clear that this would give more accurate results than our model.

In contrast to another method (24), we adapt the threshold value to the length of the oligomers. As the oligomer increases in length, the absolute number of possible oligomers of this length also increases. Thus, for a given α, longer oligomers are expected statistically to have more members which fall outside the threshold boundary even though they are randomly distributed. We show in the results that for a given oligomer length, increasing α is equivalent to increasing the stringency.

The extended method was verified by the investigation of the frequencies of hexameric palindroms in E.coli nucleic acid sequences. As reported by others (21,22), we find that in general they are underrepresented. We also find, as shown in Tables 2 and 3, that most hexamers which are restriction enzyme recognition sites are significantly underrepresented in E.coli sequences. This bias is even more striking when only recognition sites for restriction enzymes known to exist in E.coli are considered. In contrast, non-restriction site palindroms are not significantly under- or overrepresented.

Since some databases contain duplicated sequence entries we checked the effect that duplications would have on our results. The std-values were shifted only by about 5% if one sequence of approximately 700bp was copied five times within the 1.2×10^6bp. The experiment was repeated for three different sequences with the same result. Therefore we assume that our results are not significantly influenced by possible duplicated sequences in the database.

There is one other phenomenon which requires attention. We call it 'border effect'. The very first and last oligomers of the sequences are heavily biased. This is probably due to cloning effects since many of the sequences in data banks are determined from cloned fragments generated by restriction enzyme digestion. Indeed, when we examined all E.coli sequences present in the GenBankTM database both with and without the last 8 nucleotides at each end of the sequences, the only major difference in the calculated palindromic hexamer frequencies corresponded to those hexamers recognized by restriction enzymes frequently used for cloning (e.g. EcoR I, Hind III, BamH I). Thus, when restriction enzyme recognition sites are not being investigated, it is generally better to remove the first and last eight bases of each sequence file before one starts with statistical investigations.

With the statistical model presented here, oligomers can be found whose frequencies show a significant deviation from statistical (random) behaviour. These oligomers are candidates for biologically relevant patterns. Caution must be exercised with this method however for long oligomers or small sequence databases since the expected frequencies of oligomers E(s) is not substantially greater than one. For such cases we are currently developing a simpler but less informative approach involving calculating whether the oligomer occurs at all.

ACKNOWLEDGEMENTS

We thank Christoph Gartmann and Jörg Oertel for their assistance in writing and running the programs and the reviewers for helpful criticism.

REFERENCES

1. Bilofsky, H.S. and Burks, C. (1988) Nucl. Acids Res. 16,1861−1863.
2. Cameron, G.N. (1988) Nucl. Acids Res. 16, 1865−1867.
3. Sidman, K.E., George, D.G., Barker, W.C. and Hunt, L.T. (1988) Nucl. Acids Res. 16, 1869−1871.
4. Burks, C. et al. (1990) Methods in Enzymology 183, 3−22.
5. Kahn, P. and Cameron, G.N. (1990) Methods in Enzymology 183,23−31.
6. Barker, W.C, George D.G. and Hunt, L.T. (1990) Methods in Enzymology 183, 31−49.
7. Dumas, J.P. and Ninio, J. (1982) Nucl. Acids Res. 10, 197−207
8. Stormo, G.D., Schneider, T.D. and Gold, L.M. (1982) Nucl. Acids Res. 10, 2971−2996.
9. Nussinov, R. (1984) Nucl. Acids Res. 12, 1749−1763.
10. Santibanez-Koref, M. and Reich, J.G. (1986) Biomed. Biochim. Acta 45, 737−748.
11. Nussinov, R. (1987) J. Theor. Biol. 125, 219−235.
12. Nussinov, R. (1987) DNA 6, 13−22.
13. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucl. Acids Res. 11, 2205−2220.
14. Bodnar, J.W. and Ward, D.C. (1987) Nucl. Acids Res. 15,1835-1851.
15. Peterson, R.C. (1988) BioTechniques 6, 34−40.
16. Claverie, J.-M. and Bougueleret, L. (1986) Nucl. Acids Res. 14, 179−196.
17. Volinia, S., Bernardi, F., Gambari, R. and Barrai, I. (1988) J. Mol. Biol. 203, 385−390.
18. Claverie, J.-M., Sauvaget, I. and Bougueleret, L. (1990) Methods in Enzymology 183, 237−252.
19. Almagor, H. (1983) J. Theor. Biol. 104, 633−645.
20. Blaisdell, B.E. (1985) J. Mol. Evol. 21, 278−288.
21. Phillips, G.J., Arnold, J. and Ivarie, R. (1987) Nucl. Acids Res. 15, 2611−2626.
22. Phillips, G.J., Arnold, J. and Ivarie, R. (1987) Nucl. Acids Res. 15, 2627−2638.
23. Arnold, J., Cuticchia, A.J., Newsome, D.A., Jennings III, W.W. and Ivarie, R. (1988) Nucl. Acids Res. 16, 7145−7158.
24. Brendel, V. Beckmann, J.S. and Trifonov, E.N. (1986) J. Biomol. Struct. Dyn. 4, 11−21.
25. Trifonov, E.N. (1988) Math. Biosci. 90, 507−517.
26. Pevzner, P.A., Borodovsky, M.Y. and Mironov, A.A. (1989) J. Biomol. Struct. Dyn. 6, 1013−1026.
27. Devereux, J. Haeberli, P. and Smithies, O. (1984) Nucl. Acids Res. 12, 387−395
28. Roberts, R.J. (1985) Nucl. Acids Res. 13, r165−r200.