Volume 11 Number 24 1983 Nucleic Acids Research

Frequencies of restriction sites

Michael S.Waterman

Departments of Mathematics and Biology, University of Southern California, Los Angeles, CA 90089, USA

Received 28 July 1983; Revised and Accepted 11 November 1983

ABSTRACT

Restriction sites or other sequence patterns are usually assumed to occur according to a Poisson distribution with mean equal to the reciprocal of the probability of the given site or pattern. For situations where non-overlapping occurrences of patterns, such as restriction sites, are the objects of interest, this note shows that the Poisson assumption is frequently misleading. Both the case of base composition (independent bases) and of dinucleotide frequencies (Markov chains) are treated. Moreover, a new technique is presented which allows treatment of collections of patterns, where the departure from the Poisson assumption is even more striking. This later case includes double digests, and an example of a five enzyme digest is included.

INTRODUCTION

The analysis is based on counting non-overlapping patterns occurrence. That is, if the pattern is GCGC (HhaI) and a portion of the sequence is ...TGCGCGCT.., exactly one occurrence of GCGC would be counted. The probability theory for such situations, known as discrete renewal theory, was worked out by W. Feller (1968) more than thirty years ago but has not been applied to genetic sequences until now. Overlapping repeats, where two occurrences of GCGC would be counted in the above sequence, are not covered in the present analysis, but they do present special and unexpected difficulties which have been previously overlooked. We have developed quite different techniques to handle overlapping repeats and will present them elsewhere.

SINGLE PATTERN: BASE COMPOSITION

For simplicity, a DNA sequence with equally likely nucleotides is first analyzed. (Any base composition can be used.) A specific pattern is defined, TAGCTA, for example, and the first renewal is said to occur first instance of the pattern, moving 5' to 3'. Then the search begins anew, as if the sequence were starting over at the first base 5' from the pattern occurrence. It is this device, renewal, on which the analysis is based.

If u_n is the probability of a renewal at the nth position, then an occurrence of TAGCTA constitutes a renewal unless one occured a position n-4:

```
- - - - TAGCTA
             n
```

Therefore

$$P(\text{TAGCTA}) = (1/4)^6 = u_n + u_{n-4} \left(\frac{1}{4}\right)^4 ,$$

where $\left(\frac{1}{4}\right)^4 = P(\text{GCTA})$ needed to fill in the remainder of the pattern. Feller proves $u_n \to 1/\mu$, where μ is the mean repeat time for the pattern. This yields the equation

$$(1/4)^6 = 1/\mu + 1/\mu \ (1/4)^4$$

and

$$\mu = 4096 + 16 = 4112.$$

In other words, TAGCTA will, on the average, occur every 4112 bases in random sequences of equal nucleotide frequency.

For sequences of length 6, the mean repeat time is as high as 5460 for sequences of the same base, TTTTTT for example. The associated renewal equation is

$$(1/4)^6 = u_n + u_{n-1} \ (1/4) + \ldots + u_{n-5} \ (1/4)^5$$

The lowest mean repeat time is 4096 for sequences such as TATCAC which have the renewal equation

$$(1/4)^6 = u_n \ .$$

For a purine/pyrimidine (R/Y) alphabet, we analyze the pattern RY with P(R) = p and P(Y) = q. Here

$$pq = u_n$$

and

$$\mu_{RY} = 1/pq \ .$$

Therefore if $p = q = 1/2$, $\mu_{RY} = 4$. However if the pattern is RR

$$p^2 = u_n + u_{n-1}p$$

and

$$\mu_{RR} = (1+p)/p^2$$

so, for $p = 1/2$, $\mu_{RR} = 6$.

In general, for fixed length patterns, the ratio of smallest to largest repeat time is approximately 4/3. This implies that all formulae which are not pattern specific are not exactly correct. This is dependent on our definition of repeat, "overlapping repeats" do not have the same property.

In the case of purine-pyrimidine patterns, the "alphabet" has two letters, and the ratio of largest to smallest repeat times is approximately 2. This is contrasted with the amino acid case where the ratio is approximately 20/19. In other words, the effect of the specific pattern is greater for smaller alphabets.

MULTIPLE PATTERNS: BASE COMPOSITION

The situation is more complex for a collection of several patterns and, while we solve the problem, the technique does not appear in Feller and seems to have been overlooked. The idea is similar: a pattern at position n constitutes a renewal unless it or another pattern in the collection has an earlier renewal intersecting the pattern in question.

To be specific, we are interested in two patterns RY and YR with $P(Y) = P(R) = 1/2$. Let u be associated with RY and v with YR. An occurrence of RY at position n is a "u" renewal at n unless there was a "v" renewal at position n-1. This gives the equation

$$(1/2)^2 = u_n + v_{n-1}\frac{1}{2},$$

while an occurrence of YR at position n gives the equation

$$(1/2)^2 = v_n + u_{n-1}\,1/2 .$$

The limiting equations are

$$(1/2)^2 = 1/\mu_u + 1/\mu_v \; 1/2$$

$$(1/2)^2 = 1/\mu_v + 1/\mu_u \; 1/2$$

which solve to yield $\mu_u = \mu_{RY} = 6 = \mu_{YR} = \mu_v$. The mean repeat time for RY without YR was shown above to equal 4. The addition of YR eliminates some of those renewals.

The question of interest is the mean repeat time μ_* of the collection which satisfies

$$u_n + v_n \to \frac{1}{\mu_*} ,$$

and we have shown $u_n \to 1/6$ and $v_n \to 1/6$.

Therefore $\mu_* = \dfrac{1}{1/6 + 1/6} = 3$.

For another example consider the collection of five restriction sites in Table I. (Equally likely nucleotides are again assumed.) The renewal system is, with $p = .25$,

$$p^4 = u_n + x_{n-2}p^2 + y_{n-2}p^2 + x_{n-3}p^3 + w_{n-3}p^3 + y_{n-3}p^3$$

$$p^4 = v_n + w_{n-1}p + v_{n-2}p^2 + w_{n-3}p^3 + x_{n-3}p^3 + y_{n-3}p^5$$

$$p^4 = w_n + v_{n-1}p + w_{n-2}p^2 + u_{n-3}p^3 + v_{n-3}p^3$$

$$p^4 = x_n + u_{n-2}p^2 + u_{n-3}p^3 + v_{n-3}p^3$$

$$p^6 = y_n + u_{n-4}p^4 + u_{n-5}p^5 + v_{n-5}p^5$$

The resulting system can be written in matrix form.

$$\begin{pmatrix} p^4 \\ p^4 \\ p^4 \\ p^4 \\ p^6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & p^3 & p^2+p^3 & p \\ 0 & 1+p^2 & p+p^3 & p^3 & p^2 \\ p^3 & p+p^3 & 1+p^2 & 0 & p^3 \\ p^2+p^3 & p^3 & 0 & 1 & p^3 \\ p^3 & p^4 & p^5 & p+p^5 & 1+p^4 \end{pmatrix} \begin{pmatrix} \frac{1}{\mu_u} \\ \frac{1}{\mu_v} \\ \frac{1}{\mu_w} \\ \frac{1}{\mu_x} \\ \frac{1}{\mu_y} \end{pmatrix}$$

TABLE I

COLLECTION OF RESTRICTION SITES

Restriction Enzyme	Restriction Site	Associated Symbol
Hpa II	CCGG	u
Fnu D II	CGCG	v
Hha I	GCGC	w
Hae III	GGCC	x
Bam H I	GGATCC	y

$$\left(\frac{1}{\mu_u}, \frac{1}{\mu_v}, \frac{1}{\mu_w}, \frac{1}{\mu_x}, \frac{1}{\mu_y}\right) = (.00356, .00290, .00290, .00358, .00022)$$

Since the mean repeat time for the collection, μ_*, satisfies

$$(u_n + v_n + w_n + x_n + y_n) \to \frac{1}{\mu_*},$$

so that

$$\mu^* = \frac{1}{\frac{1}{\mu_u} + \frac{1}{\mu_v} + \frac{1}{\mu_w} + \frac{1}{\mu_x} + \frac{1}{\mu_y}} = 75.93 \ldots$$

This answer should be compared to the naive calculation of $(4p^4+p^6)^{-1} = 63.02 \ldots$ which is 83% of the correct value.

The above procedure for collections of restriction sites can easily be programmed on a computer and, with an application of an equation solver, μ_* results.

When an analysis of mean repeat time is desired for restriction sites in double stranded DNA, it is a simple matter to increase the list of restriction sites (patterns). In our example, no new patterns result as our sites are all palindromes of even length. For palindromes of odd length, a single base change is necessary: Ara II (GGTCC) adds GGACC. For non-palindromes, simply add the reverse complement: Mnℓ I (GAGG) adds CCTC.

DINUCLEOTIDE FREQUENCES

DNA is not a sequence of independent bases but can be better described by higher order dependencies such as dinucleotide frequencies (2). Fortunately, the work of Feller (1) will accomodate such relevant data. Let p_{IJ} denote the frequency of I to J transitions (5' to 3') relevant to our particular DNA and P_I is the frequency of nucleotide I in the sequence.

To illustrate let us redo the TAGCTA example considered above

$$P(\text{TAGCTA}) = p_T p_{TA} p_{AG} p_{GC} p_{CT} p_{TA} = u_n + u_{n-4} \cdot \ (p_{AG} p_{GC} p_{CT} p_{TA})$$

and

$$\mu = \frac{1 + p_{AG} p_{GC} p_{CT} p_{TA}}{p_T p_{TA} p_{AG} p_{GC} p_{CT} p_{TA}} \ .$$

This approach will yield better answers and is recommended. Higher order dependencies can be used if they are known and thought to be relevant.

CONCLUSION

It is possible to apply generating functions to these sequences and obtain higher order moments (1), even for collections of patterns. The calculations are more involved than those performed above. This will allow normal approximations for the sequence length required for a given number of pattern occurrences as well as the number of occurrences in a given length of sequence.

ACKNOWLEDGEMENT

This work was partially supported by a grant from System Development Foundation.

REFERENCES

1. Feller, W., An Introduction to Probability Theory and its Applications 3rd ed. (Wiley, New York, 1968).
2. Smith, T.F., J.R. Sadler, and M.S. Waterman (1983) *Nuc. Acids Res. 11*, 2205.