EXPECTED FREQUENCIES OF DNA PATTERNS USING WHITTLE'S FORMULA

by Richard Cowan
Department of Statistics, University of Hong Kong.

Personal archive version of a paper which appeared in *J. Appl. Prob.* **28**, 886-892 (1991). The author is now in the School of Mathematics and Statistics, University of Sydney. email: *richard.cowan@sydney.edu.au*

SUMMARY: Given a realisation of a Markov chain, one can count the numbers of state transitions of each type. One can ask how many realisations are there with these transition counts and the same initial state. Whittle (1955) has answered this question, by finding an explicit though complicated formula; he has also shown that each realisation is equally likely. In the analysis of DNA sequences which comprise letters from the set {A,C,G,T}, it is often useful to count the frequency of a pattern, say ACGCT, in a long sequence and compare this with the expected frequency for all sequences having the same start letter and the same transition counts (or 'dinucleotide counts' as they are called in the molecular biology literature). To date, no exact method exists; this paper rectifies that deficiency.

KEYWORDS: Markov chains, DNA, dinucleotide, transition counts.

1. Introduction

It has been common practice, in the analysis of DNA sequences, to tabulate the frequency of 'dinucleotides' in a given sequence (Nussinov (1981), Avery (1987), Bulmer (1987), Gardiner-Garden and Frommer (1987), Smith et al (1983)). DNA consists of a long chain of four different nucleotides A, G, C and T. The rules which determine the order of nucleotides are not well understood. They are, however, somewhat stochastic in nature due to the random genesis and mutation of DNA sequences. A 'dinucleotide' is a pair of consecutive nucleotides, so there are 16 possible dinucleotides. Their observed frequency in a given sequence is conveniently tabulated in a 4×4 matrix, M.

$$M = \begin{bmatrix} n_{AA} & n_{AC} & n_{AG} & n_{AT} \\ n_{CA} & n_{CC} & n_{CG} & n_{CT} \\ n_{GA} & n_{GC} & n_{GG} & n_{GT} \\ n_{TA} & n_{TC} & n_{TG} & n_{TT} \end{bmatrix}$$

where, for example, n_{GC} is the frequency of the dinucleotide GC. If the sequence is of length n, the elements of M sum to $n-1$.

Calculations of 'expected frequencies' of dinucleotides under the '4-sided die model' of DNA are often made in the molecular biology literature. In this model the sequence is assumed to be generated by n independent throws of a 4-sided die. The expected frequencies used are those conditional upon the *single* nucleotide counts n_A, n_C, n_G and n_T. For example, the conditional expected frequency of GC is $n_G\, n_C/n$.

Comparisons in the literature between observed and expected dinucleotide counts have highlighted many features, for example (in eukaryotic species), markedly lower than expected frequencies of CG, TA and (to a lesser extent) GT and AT, with elevated frequencies of TG, CT and CA (Nussinov (1981)). These features have helped to focus research on possible biochemical explanations. As a result, molecular mechanisms have been proposed for the depression of CG (Bird (1980)), mechanisms which also account

for the elevation of CA and TG. Usually, comparisons of observed and expected have not been formalised as rigorous statistical tests, but, because of the very long sequences (and large numbers of sequences), the simple *ad hoc* statistic 'observed/expected' has been informative. As a result it is now clearly established that the 4-sided die model is invalid. At least first-order Markov dependency is needed.

Molecular biologists need, on a day-to-day basis, a simple working tool to assess whether the observed/expected statistic of a given pattern, say $CTAG$, is unusual in some way. By default, the 4-sided die model is used as null hypothesis to calculate the expectation conditional upon $n_A,\ n_C,\ n_G$ and n_T. This is now inappropriate given the state of knowledge on dinucleotide frequencies.

The current paper finds the expected frequency of any nominated pattern of length ≥ 2, given the starting nucleotide and the matrix M of dinucleotide counts. Thus a useful tool for molecular biology is provided. In doing so, the paper provides some interesting applications and extensions of Whittle's powerful combinatoric formula (Whittle (1955)). The results, whilst presented in the setting of the DNA problem which motivated them, are applicable to all Markov chain applications.

2. Whittle's formula and some new arrangements

Let the row-sums of M be denoted by $a,\ c,\ g$ and t respectively and the column-sums by $a^*,\ c^*,\ g^*$ and t^*. A sequence which begins and ends with the same letter has a matrix M with

$$a - a^* = c - c^* = g - g^* = t - t^* = 0 \tag{1}$$

A sequence commencing with one letter, G say, and terminating with another, T say, has

$$g - g^* = 1; \qquad t - t^* = -1; \qquad a - a^* = c - c^* = 0 \tag{2}$$

M must satisfy either (1) or equations in the generic form of (2). Thus M and the starting letter determine the end letter (though often M alone suffices).

Whittle (1955) has derived a formula for the number of sequences conforming (i) to the counts given in M, and (ii) to a start letter (and implied end letter) consistent with M. For example, the number $W_{GT}(M)$ of sequences commencing with G, terminating with T and having transition counts conforming to (2), is

$$W_{GT}(M) = K_M H_{GT}(M), \qquad \text{where} \tag{3}$$

$$K_M = \frac{a!\,c!\,g!\,t!}{\prod_{i,j} n_{ij}!} \tag{4}$$

and where i and j index the set $\{A, C, G, T\}$ and $H_{GT}(M)$ is the (4,3)th cofactor (4 for T, 3 for G) of

$$\begin{bmatrix} 1 - n_{AA}/a & -n_{AC}/a & -n_{AG}/a & -n_{AT}/a \\ -n_{CA}/c & 1 - n_{CC}/c & -n_{CG}/c & -n_{CT}/c \\ -n_{GA}/g & -n_{GC}/g & 1 - n_{GG}/g & -n_{GT}/g \\ -n_{TA}/t & -n_{TC}/t & -n_{TG}/t & 1 - n_{TT}/t \end{bmatrix}. \tag{5}$$

Whittle also covers the case, unlikely with long DNA sequences, where a row-sum is zero; a ratio in (5) involving such a row-sum is defined as zero. It is also shown by Whittle, that if a Markov chain generates

the said M from the given start letter, then each of the sequences conforming to M and the start letter (and implied end letter) are equally likely. It is helpful to simplify $H_{GT}(M)$ by algebraically evaluating the appropriate cofactor in (5). This yields the formula

$$H_{GT}(M) = \frac{1}{acg}\big(A\,n_{GC}\,n_{CT} + C\,n_{GA}\,n_{AT} + A\,C\,n_{GT} + n_{GA}\,n_{AC}\,n_{CT} + n_{GC}\,n_{CA}\,n_{AT} - n_{AC}\,n_{CA}\,n_{GT}\big) \quad (6)$$

where we adopt the notation $A = a - n_{AA},\ C = c - n_{CC},\ G = g - n_{GG}$ and $T = t - n_{TT}$. If a row-sum, say c, is zero then (6) is still valid with the convention stated above that ratios such as n_{CA}/c are zero and consequently $C/c \equiv 1 - n_{CC}/c = 1$. For example, $H_{GT}(M) = (n_{GA}\,n_{AT} + A\,n_{GT})/(a\,g)$ if $c = 0$ and $H_{GT}(M) = 0$ if $g = 0$.

Formulae (3) and (5), based on G-to-T sequences, have direct analogies for other start and end letters. For example, with an A-to-A sequence, $W_{AA}(M) = K_M H_{AA}(M)$. Here $H_{AA}(M)$ is the (1,1)th cofactor of (5) which, upon expansion, yields for $a > 0$

$$H_{AA}(M) = \frac{1}{cgt}\,(C\,G\,T - C\,n_{GT}\,n_{TG} - G\,n_{CT}\,n_{TC} - T\,n_{CG}\,n_{GC}\ - n_{CG}\,n_{GT}\,n_{TC} - n_{CT}\,n_{TG}\,n_{GC}) \quad (7)$$

with $H_{AA}(M) = 0$ if $a = 0$. All other examples can be found by permutation symmetry from either (6), when 'start $\neq$ end', or (7), when 'start = end'. In (6), the denominator is without the row-sum corresponding to the *end* letter. Two examples are:

$$H_{CA}(M) = \frac{1}{cgt}\,(G\,n_{CT}\,n_{TA} + T\,n_{CG}\,n_{GA} + G\,T\,n_{CA} + n_{CG}\,n_{GT}\,n_{TA} + n_{CT}\,n_{TG}\,n_{GA} - n_{GT}\,n_{TG}\,n_{CA})\,.$$

$$H_{GG}(M) = \frac{1}{act}\,(A\,C\,T - A\,n_{CT}\,n_{TC} - C\,n_{AT}\,n_{TA} - T\,n_{AC}\,n_{CA} - n_{AC}\,n_{CT}\,n_{TA} - n_{AT}\,n_{TC}\,n_{CA})\,. \quad (8)$$

Incidentally, Cowan (1992) shows that, for an M consistent with sequences that start and end with the same letter, that is satisfying (1),

$$\frac{H_{AA}(M)}{a} = \frac{H_{CC}(M)}{c} = \frac{H_{GG}(M)}{g} = \frac{H_{TT}(M)}{t}.$$

Thus the apparent difference between the bracketed terms in (7) and (8) is non-existent.

3. Expected pattern frequency

Suppose that our interest focuses on the expected frequency of a certain pattern π, say $\pi = CTTGCTA$ or $\pi = GAGA$, amongst the equally-likely sequences which conform to a given M matrix and start letter denoted by $S \in \{A, C, G, T\}$. The frequency, n_π say, includes separate counting of overlapping occurrences; for example, $\pi = GAGA$ occurs 3 times in $CGAGATGAGAGAC$ at positions 2, 7 and 9. (We say that π occurs *at* position k if it starts at k.)

Let ℓ be the length of π. Clearly $n_\pi = \sum_{k=1}^{n-\ell+1} I_k(\pi)$, where $I_k(\pi),\ (k = 1, 2, \cdots, n-\ell+1)$, is defined by $I_k(\pi) = 1$ if π appears *at* position k; $I_k(\pi) = 0$ otherwise. Thus

$$E(n_\pi|M, S) = \sum_{k=1}^{n-\ell+1} EI_k(\pi) = \sum_{k=1}^{n-\ell+1} P\big\{I_k(\pi) = 1\big\}. \quad (9)$$

Also,

$$P\{I_k(\pi) = 1\} = \frac{\#\text{ conforming sequences with } \pi \text{ at position } k}{\text{total } \#\text{ conforming sequences}}. \tag{10}$$

For definiteness, suppose that the sequences must start with G and end with T. Thus the denominator of (10), and hence of each term in (9), is $W_{GT}(M)$. The numerator of (10), which we denote by $\beta(k, M, \pi)$, is more difficult to evaluate. There is, however, an equivalent combinational entity that is amenable to analysis. We illustrate with $\pi = CTTGCTA$. In Figure 1(a), π is placed at position k.

(a) $\underbrace{G\ldots\ldots\ldots\ldots CTTGCT}_{k}\underbrace{A\ldots\ldots\ldots\ldots T}_{n-\ell-k+2}$

(b) $\underbrace{G\ldots\ldots\ldots\ldots C}_{k}\underbrace{A\ldots\ldots\ldots\ldots T}_{n-\ell-k+2}$

Figure 1: Equivalent combinations in 2 distinct problems.

Let α be the dinucleotide formed from the first and last letters of π; for example, $\alpha = CA$. For every sequence like 1(a), one can write a sequence such as 1(b), without the central letters of π, having α at position k. It is clear that the number of sequences, like 1(a), conforming to M and having π at position k, *equals* the number of sequences like 1(b), conforming to the matrix $M(CTTGCTA)$ given below and having α at position k.

$$M(CTTGCTA) = \begin{bmatrix} n_{AA} & n_{AC} & n_{AG} & n_{AT} \\ n_{CA}+1 & n_{CC} & n_{CG} & n_{CT}-2 \\ n_{GA} & n_{GC}-1 & n_{GG} & n_{GT} \\ n_{TA}-1 & n_{TC} & n_{TG}-1 & n_{TT}-1 \end{bmatrix}.$$

In general, $M(\pi)$ is a variant of M depleted by the dinucleotides in π and supplemented by a 'one' added to n_α, the count of α. Thus $\beta(k, M, \pi)$ equals $\beta(k, M(\pi), \alpha)$. Therefore, from (9),

$$E(n_\pi | M, G) = \frac{\sum_{k=1}^{n-\ell+1} \beta(k, M, \pi)}{W_{GT}(M)} = \frac{\sum_{k=1}^{n-\ell+1} \beta(k, M(\pi), \alpha)}{W_{GT}(M)}. \tag{11}$$

It is equally clear, from Figure 1(b), that $\sum_{k=1}^{n-\ell+1} \beta(k, M(\pi), \alpha)/W_{GT}(M(\pi))$ is the expected number of CA's in the G-to-T sequences which conform to $M(\pi)$. This expected number is *known*; it is the entry for CA in $M(\pi)$, namely $n_{CA} + 1 - n_{CA}(\pi)$, where $n_{CA}(\pi)$ is the number of CA dinucleotides in π. Therefore, in general,

$$\sum_{k=1}^{n-\ell+1} \beta\big(k, M(\pi), \alpha\big) = (n_\alpha + 1 - n_\alpha(\pi)) W_{GT}(M(\pi))$$

for our illustrative case of G-to-T sequences. From (11), we have the final result, expressed in the following theorem.

THEOREM: *Let S, F be letters in the 'nucleotide set' $\mathcal{X} = \{A, C, G, T\}$ and M be a matrix of 'dinucleotide counts' consistent with sequences that start with S and end with F. Let π be a sequence of letters*

from $\mathcal{X}$ of length ≥ 2 and let α be the dinucleotide formed from the first and last letters of π. Let n_α be the count of α dinucleotides in the matrix M and $n_\alpha(\pi)$ be the said count in π. If all S-to-F sequences consistent with M are equally likely, and if n_π is the number of occurrences of π in a randomly chosen such sequence, then

$$E(n_\pi|M,S) = \frac{(n_\alpha + 1 - n_\alpha(\pi))W_{SF}(M(\pi))}{W_{SF}(M)},$$

where $M(\pi)$ is a variant of M depleted by the dinucleotides in π and supplemented by a 'one' added to n_α. It takes little imagination to see that this theorem holds for a general finite set $\mathcal{X}$. So our result is relevant to the general theory of Markov chains. Also, when M determines the start letter S, $E(n_\pi|M)$ is a sufficient notation (see example below).

4. An example

What is the expected frequency of $\pi = CGAAATGCT$ in G-to-T sequences consistent with the M shown below? The matrix $M(\pi)$ is also shown and $\alpha = CT$.

$$M = \begin{bmatrix} 6 & 3 & 8 & 10 \\ 2 & 5 & 5 & 7 \\ 11 & 4 & 6 & 4 \\ 8 & 7 & 5 & 3 \end{bmatrix} \begin{matrix} a = 27 \\ c = 19 \\ g = 25 \\ t = 23 \end{matrix} \qquad M(\pi) = \begin{bmatrix} 4 & 3 & 8 & 9 \\ 2 & 5 & 4 & 7 \\ 10 & 3 & 6 & 4 \\ 8 & 7 & 4 & 3 \end{bmatrix}$$

Firstly, let us find $W_{GT}(M)$ from (3), (4) and (6). We can calculate that $H_{GT}(M) = 7/25$ and $K_M = 2^{16}3^{13}5^67^511^613^517^419^423^3$, so $W_{GT}(M)$ is their product. Similarly $H_{GT}(M(\pi)) = 1435/4968$ and $K_{M(\pi)} = 2^{16}3^{10}5^67^511^713^417^419^323^2$. Also $n_\alpha = 7$ and $n_\alpha(\pi) = 1$, so $E(n_{CGAAATGCT}|M) = 394625/762026616 = 0.000517862$.

Using the same M, what is $E(n_\pi|M)$ if $\pi = GAT$? Now $H_{GT}(M(\pi)) = 1736/6175$ and $K_{M(\pi)} = 2^{17}3^{10}5^67^511^713^517^419^423^3$. Also $n_\alpha \equiv n_{GT} = 4$ and $n_\alpha(\pi) = 0$, so $E(n_{GAT}|M) = 27280/6669 = 4.09057$.

Further examples can be found in Gardiner-Garden and Frommer (1987). In general one finds that, for larger n_{ij} values, K_M and $K_{M(\pi)}$ are not easy to calculate individually but their ratio causes no difficulties due to the cancellation of most factoral terms. Also $H_{SF}(M(\pi))/H_{SF}(M)$ tends to one as n gets larger for fixed π, providing further simplification.

5. Discussion

We have provided a simple, exact formula for expected pattern frequencies in sequences that conform to given transition (dinucleotide) counts. The formula can be evaluated using a pocket calculator.

The formula is somewhat more complicated then one might at first expect, due to the subtlety of Whittle's formula. The K_M-part of his formula is deceptively simple. The 16 possible dinucleotides can be divided into 4 *classes* depending on the *first* of the two letters involved. Each class can be further subdivided into 4. Within a given class, say dinucleotides commencing with G, there are $g!/(n_{GA}!n_{GC}!n_{GG}!n_{GT}!)$ distinct orderings of the class members. K_M is the product of such terms and so is the total number of distinct orderings of all 4 classes. For every conforming sequence, there corresponds one ordering of the 4 classes and this ordering conversely determines the sequences. Some orderings do *not*, however, correspond

to a full-length conforming sequence; they terminate prematurely with the unused dinucleotides being of the wrong class to continue the sequence. The H-term in Whittle's formula gives the proportion of the 4-class orderings which successfully produce a full-length sequence.

We consider that the *conditional* expectation, $E(n_\pi|M,S)$ is the appropriate entity to use when comparing 'observed' with 'expected'. Furthermore this is consistent with the approach currently used in molecular biology when the 4-sided die model of DNA is employed. Under a Markov Chain model of DNA, with the further assumption of *stationarity*, the unconditional expectation of pattern frequency for a π such as $CGAAATGCT$ is $En_\pi = (n-\ell+1)p_C p_{CG} p_{GA} p_{AA}^2 p_{AT} p_{TG} p_{GC} p_{CT}$, where p_{ij} is the transition probability from letter i to letter j and p_C is the equilibrium probability of letter C. Avery (1987) has used these unconditional expectations in his study of intron sequences, after estimating the various p-terms. In effect, he uses an estimated, unconditional expectation which we denote as $\widehat{En_\pi}$. He estimates a transition probability, say p_{GA}, by $\hat{p}_{GA} = n_{GA}/(n_{GA}+n_{GC}+n_{GG}+n_{GT})$, which is the maximum likelihood estimator based on the *conditional* likelihood given the start letter S. (It remains valid even if stationarity is not assumed.) Avery estimates equilibrium probabilities such as p_C by $\tilde{p}_C = n_C/n$, this being a moment estimator under the assumption of stationarity. The estimator $\tilde{p}_C$ is not strictly compatible with estimators such as $\hat{p}_{GA}$; the compatible estimator of p_C comes from the solution of a set of linear equations involving the estimators such as $\hat{p}_{GA}$. (Alternatively, one could derive a compatible set of estimators for all transition and equilibrium probabilities by maximising the unconditional likelihood under a *stationary* Markov chain model.)

As $n \to \infty$ with π fixed, Avery's $\widehat{En_\pi}$ and our $E(n_\pi|M,S)$ become equal for a number of reasons: (a) his estimates for the p- terms converge to the true values; (b) the incompatibility mentioned above disappears; (c) the distinction between our conditional expectation and Avery's unconditional expectation diminishes; (d) the effects of any transient phase on the validity of Avery's formula, itself dependent on a stationarity assumption, become negligible.

In short, one would expect Avery's approach to give a good approximation to ours for n relatively large compared with ℓ, the length of π. We conclude by presenting Avery's method applied to our earlier examples. We find that $\widehat{En}_{GAT} = (95-2)\frac{25}{95} \cdot \frac{11}{25} \cdot \frac{10}{27} = 3.988$ whilst by a similar method $\widehat{En}_{CGAAATGCT} = 0.0005048$.

REFERENCES

Avery, P.J. (1987) The Analysis of Intron Data and Their Use in the Detection of Short Signals. *J. Mol. Evol.* **26,** 335–340

Bird, A.P. (1980) DNA methylation and the frequency of C_pG in animal DNA. *Nucleic Acids Res.* **8,** 1499–1504

Bulmer, M. (1987) A Statistical Analysis of Nucleotide Sequences of Introns and Exons in Human Genes. *Mol. Biol. Evol.,* **4(4),** 395–405

Cowan, R. (1992) Whittle's formula on a circle. In preparation

Gardiner-Garden, M. and Frommer, M. (1987) C_pG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282

Nussinov, R. (1981) The Universal Dinucleotide Asymmetry Rules in DNA and the Amino Acid Codon Choice. *J. Mol. Evol.,* **17,** 237–244

Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Statistical characterization of nucleic acid sequence functional domains. *Nucleic Acids Res.,* **11,** 2205–2220

Whittle, P. (1955) Some Distribution and Moment Formulae for the Markov Chain. *J. Roy. Statist. Soc. B,* **17,** 235–242